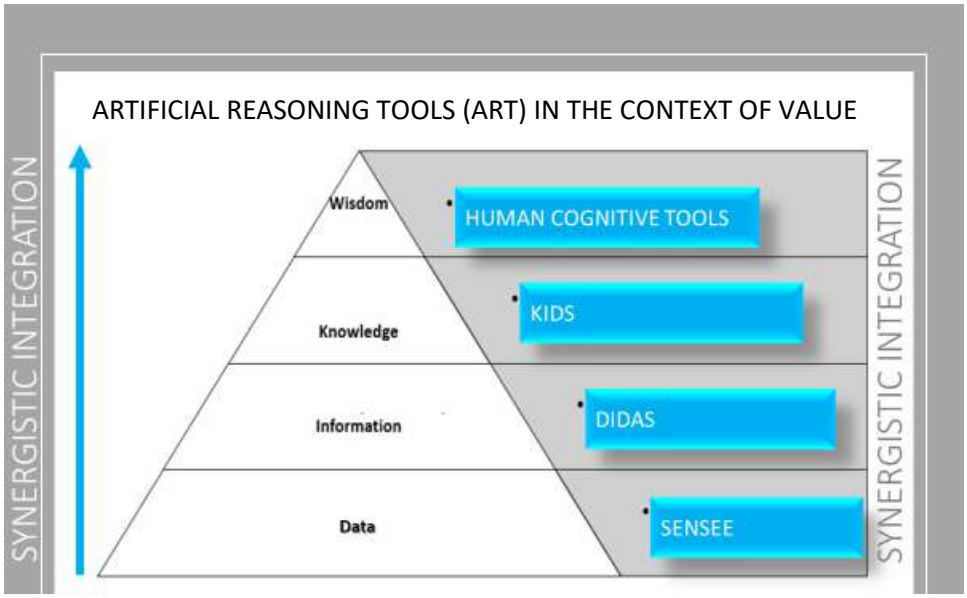
The **ART** of transforming these ideas into reality?

Shoumen Datta, MIT



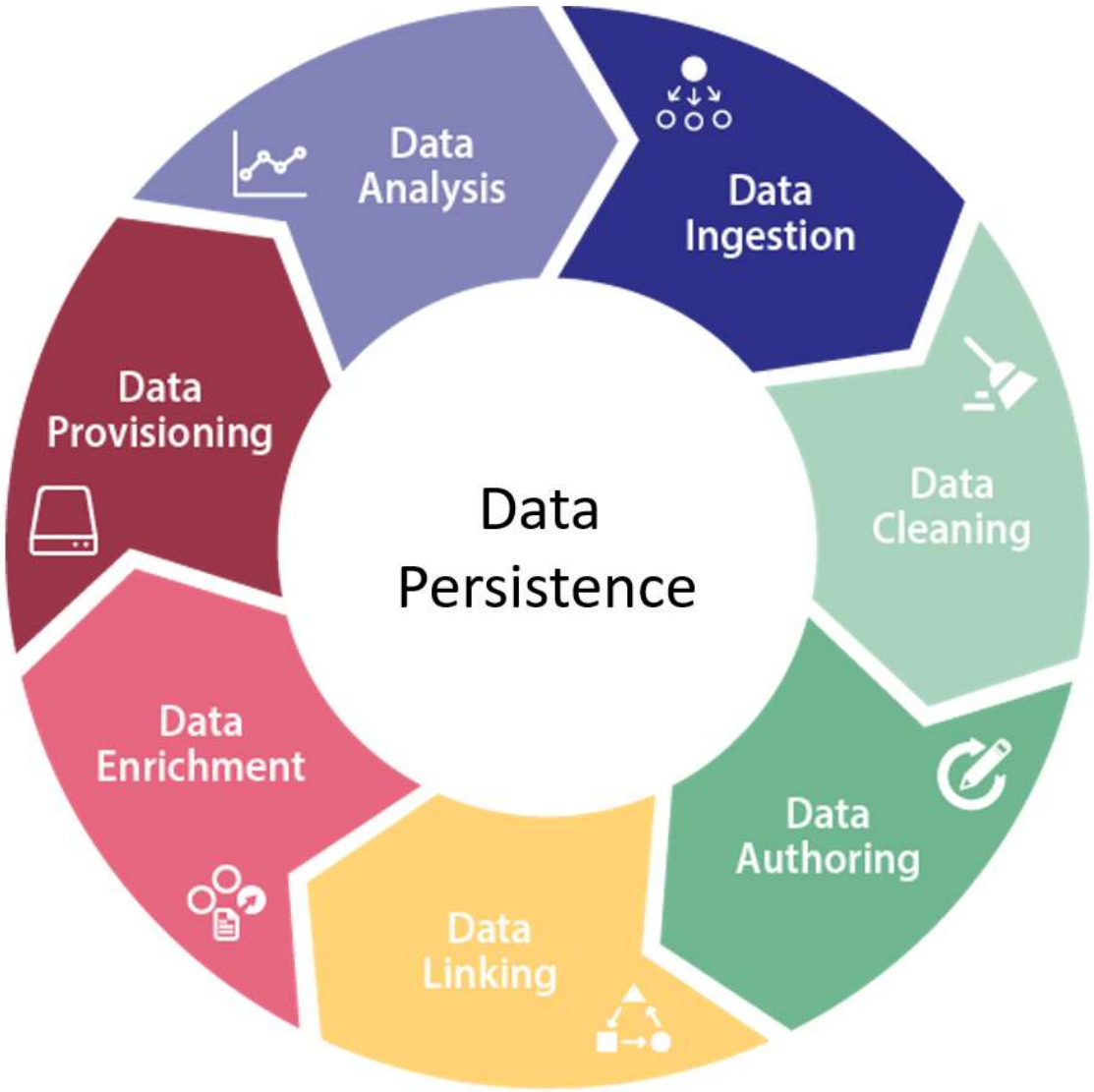
PEAS and **SIGNALS**

For context of this discussion, download “PEAS” & “SIGNALS”

- PDFs from https://dspace.mit.edu/handle/1721.1/111021
- Alternate http://bit.ly/SIGNALS-SIGNALS

1 ◆ Exploring how sensor repositories may be helpful. See Figure 11 on page 28 – http://bit.ly/SIGNALS-SIGNALS

The hypothetical ideas in this document outlines data related processes, **not** under implementation, yet. It suggests how we may start phase one (1.0): to create a proof of concept for SENSEE (**SEN**sor **SE**arch **E**ngine). *Draft will be frequently updated and may be downloaded from AWS using this short URL* http://bit.ly/SENSEE-PoC



2 ◆ Exploring how sensor repositories may be helpful. See Figure 11 on page 28 – http://bit.ly/SIGNALS-SIGNALS

This project is a mini proof of concept (PoC) for "SENSEE 1.0" (excludes DIDA'S KIDS).

See "*PEAS Platform for the Agro-Ecosystem*" and essays http://bit.ly/SIGNALS-SIGNALS

What do we want for SENSEE 1.0 demonstration purposes?

A short URL, which enables an app equivalent (web-service) on a smartphone. This app, when it opens on the phone, will reveal a dialog box. Users will type questions in the dialog box.

What is the purpose of the app for SENSEE 1.0 demonstration purposes?

Reply to queries related to the questions the user types in the dialog box. The questions will be of the type suggested (below and elsewhere). The answers will be sourced from the xl spread sheet, indicated below as "source" (please download xl from URL provided below). Decisions about sensor types is the expected outcome (at this phase, SENSEE 1.0 may only help sensor experts).

Type of questions that users may wish to ask using the dialog box (on a mobile device):

Deliverable for the mini proof of concept (PoC): answer natural language questions, for example:

[a] what is the LOD score for ionic mercury / mercury ?

[b] can I use graphene paper to detect E. coli ?

[c] is the ammonium ionophore liquid or solid ?

[d] what type of recognition tool do you have at hand for detecting imidacloprid ?

[e] what is the response time for superoxide dismutase ?

[f] what is the phase of the nitrate ionophore ? *Where semantics will come in (not now, in future).*

A detailed data dictionary (semantics) is not required. Query language will be restricted, for PoC.

Answers are in this spread sheet. The "raw data" pertains to sensor categories, attributes.

SOURCE - download xl from - http://bit.ly/SENSOR-LIBRARY-ERIC-MCLAMORE

FAQ - What is considered the primary key for the dataset?

This is NOT a data set. This is a reference for type of "tools for detection of molecules" that we refer to as sensors. Therefore, in database terms, the entire set of columns uniquely identifies rows in the table (but we can drop a few less populated columns, in the initial response. It is going to be useful when we move to knowledge graphs). At this time we are not presenting any data for the type of sensors in the xl sheet. When we have a "data set" which represents logged data from a specific sensor (SENSEE 2.0), then the ***time stamp*** on that data logger may be one of the primary keys to create an unique identifier. As this point we have a set of columns which are ***all*** key attributes for a sensor (perhaps think column headings as "features" for future phases.)

FAQ - Is there a codebook for column name headings (definitions / standards understood by non-expert users)?

No. This is not the vernacular that farm-users are likely to know or use. For the deliverable proof of concept, ignore the semantics of the column headings (it is a future task when we begin to use graph databases / semantic data catalogs - Fig 12 on page 29 http://bit.ly/SIGNALS-SIGNALS).

3 ◆ Exploring how sensor repositories may be helpful. See Figure 11 on page 28 – http://bit.ly/SIGNALS-SIGNALS

FAQ - Is there a specific metadata standard for this community of practice (e.g. common naming for data elements)?

[a] Explore SensorML and StarFL – review – https://pubag.nal.usda.gov/catalog/1229146

[b] Early XML implementations
https://www.isprs.org/proceedings/xxxv/congress/comm4/papers/516.pdf
https://www.iitk.ac.in/nicee/wcee/article/13_956.pdf

[c] STANDARDS
https://geo-ide.noaa.gov/wiki/index.php?title=SensorML_and_ISO_Metadata
https://www.w3.org/TR/vocab-ssn/

[d] REVIEW
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.562.8057&rep=rep1&type=pdf
https://publik.tuwien.ac.at/files/PubDat_208567.pdf
https://www.tandfonline.com/doi/pdf/10.1080/17538940902866195?needAccess=true
https://pdfs.semanticscholar.org/0999/630530af38c85f31ef2a6e2bb3f701da582b.pdf

[e] THE ROAD AHEAD – FUTURE CONSIDERATIONS
https://www.w3.org/2018/03/wot-f2f/slides/Mdata-WoT-2018-03-26MM.pdf
https://www.usgs.gov/land-resources/nli/landsat/science

FAQ - How to document missing data/null values in columns MW [Da], LOD [M], Max range [M], etc. (e.g. "9999" or N/A)?

When SENSEE 1.0 is a working platform, these values may be contributed. The "open platform" and "open port" approach necessary in the SENSEE information architecture, in order that data can be ingested when we have raw data. SENSEE 1.0 is a database ***for sensor device*** reference, at this time. In SENSEE 2.0, we will offer a library of interfaces (APIs) that a 3rd party can download to upload ***sensor-specific data***. SENSEE 2.0 expects to ingest ***case-specific*** sensor data (not there yet). In the context of SENSEE 2.0, feature engineering and feature selection ("selecting a few things that are most important, given that only a few can be sustained") will guide selectivity of data ingestion depending on compatibility between SENSEE 2.0 metadata vs external sources. Publishing "libraries of tools to enable interoperability between databases and data formats" as downloadable tools from the SENSEE portal may help to ensure that we are providing users the ***ability to collaborate*** even if [a] they are in Cairo, Cardiff, Cali or Calcutta and [b] their data schema, style sheets and data holders may not match SENSEE. Open APIs, open platforms and interoperability must be an integral part of the data management plan. Open information architecture may catalyze distributed data collection to strengthen SENSEE, ART and build toward a "Google of ag" approach, as we move beyond SENSEE (1.0, 2.0) and ART to meet the challenges of DIDA'S KIDS (please see "PEAS Platform for the Agro-Ecosystem").

Additional information to address potential question - What are the common data elements for column standardization [e.g. Is the LOD [M] the same as Range (LOD)?]

The value is a concentration expressed in the standard form. Range (xl sheet) should be bounded or display the lowest detectable. Limit of detection (LOD) is defined as the lowest concentration (hence, value in nanomoles, micromoles, millimoles) at which 95% of positive samples are detected. LOD is not necessarily within the linear range of an assay. LOD can be lower than Lower Limit of Quantification (LLOQ), defined as the lowest standard on the calibration curve. For further details, explore: https://www.fda.gov/downloads/Drugs/Guidances/ucm070107.pdf

ELEMENTS OF A BIGGER PICTURE – IDEAS BEYOND THE PROOF OF CONCEPT

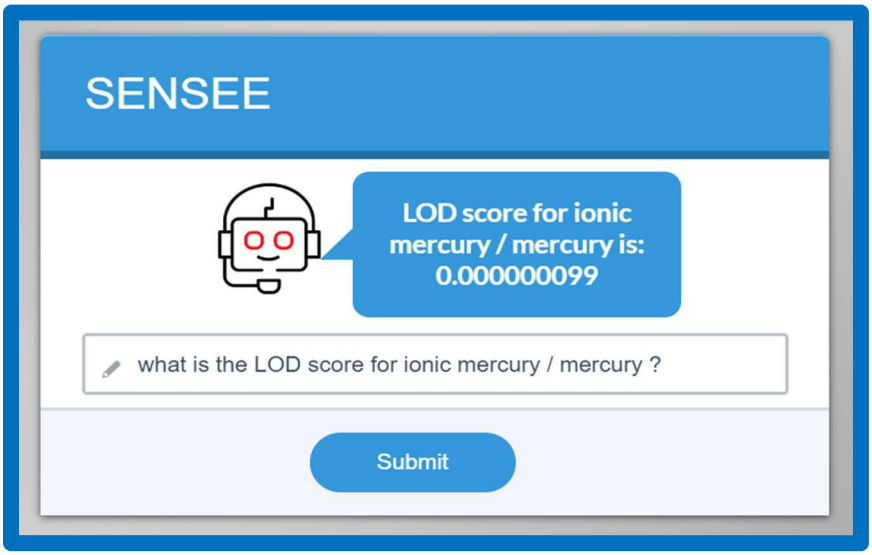
Information arbitrage (PEAS Platform for the Agro-Ecosystem), ART, data-informed decision support for the agro-ecosystem (DIDA'S) and the food industry are expected outcomes. A few steps of this scenario are outlined. Target (?) is to deliver [I] through [V] for FY 2019-2020.

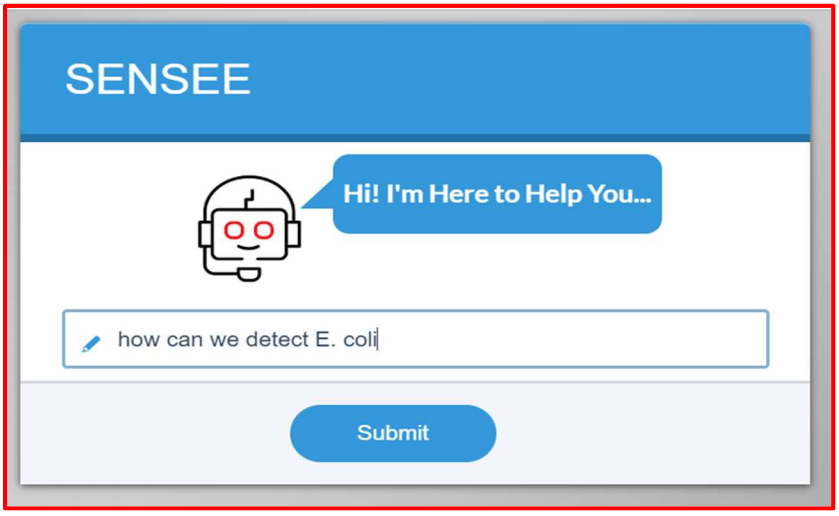
[I] PoC delivers app to return limited number of queries based on sensor description (xl sheet)
Result: http://146.185.133.187/SENSEE1/ ● http://139.162.7.63/SENSEE/
Task: App responds to query about a few sensor descriptions. Host and maintain app and DB. Provide URL to download web service and continue to bolster search functions (SENSEE 1.0).
Comments: Hard coding exact questions (syntax) is inadequate. Elasticsearch and NLP, for example, BERT (Bidirectional Encoder Representations from Transformers) is preferred and/or necessary (https://arxiv.org/pdf/1706.03762.pdf, https://github.com/google-research/bert and https://arxiv.org/abs/1905.05950). NLU/NLP engine may be trained to search keywords in the user's question and may eliminate the need for users to abide by restrictive syntax. At this time, the extent of the library and framework is extremely limited (only one xl sheet provided, others are expected). Therefore, the demand for NLP techniques may be rudimentary and limited to effective text representations and extraction of keywords from natural language queries. Advanced NLP expertise, semantic extraction techniques, data structures and modeling, will be required when volume and variety of sensor descriptions are likely to increase. The UI for Q&A using a web service (http://appinventor.mit.edu) is expected to remain simple. For more on BERT NLP explore: http://mlexplained.com/2017/12/29/attention-is-all-you-need-explained/

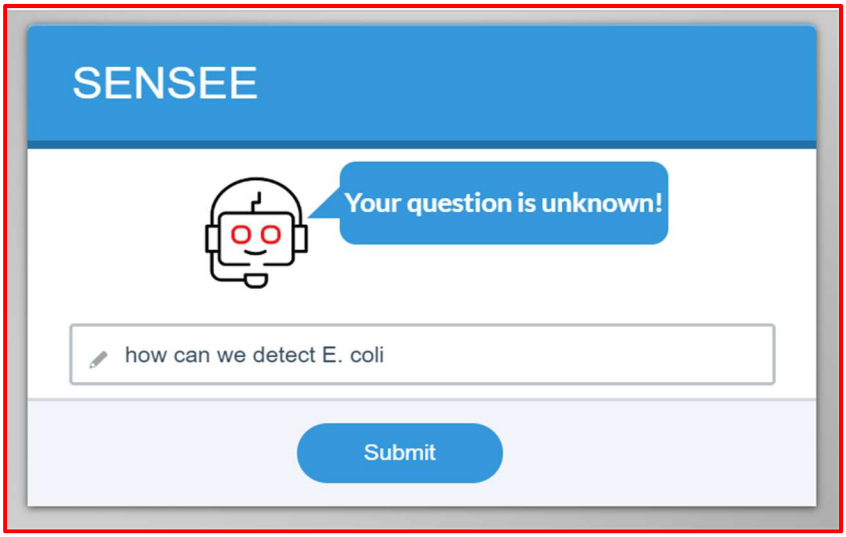
[II] Create auto-config xl tool which can be downloaded to upload sensor details (100-1000 labs)
Target: 2019
Task: Expand and scale the system to contain other types of sensors created by other labs.
Comments: How will these other labs accomplish this task? How will the other labs add to the DB? How will they create and populate new columns if the descriptor/characteristic is not present in the current DB? Most labs maintain sensor type descriptions as tables (CSV, xl). Provide a short URL which will lead to a "tool" which can be downloaded and serve as a document management system (DMS) to accept the xl document with sensor descriptions. The uploaded (ingested) document will be parsed and analyzed by system "software" using keywords (think search engine optimization). The existing DB will be updated if the table/column headings are a match. Consider these tasks: [i] Online job application sites where the resume is uploaded to a site and the portal populates its "boxes" (fields) by extracting information from the uploaded resume using metadata tools. The vast number of errors in this process often requires editing (by the user) because the rules of the parser are shoddy.

5 ◆ Exploring how sensor repositories may be helpful. See Figure 11 on page 28 – http://bit.ly/SIGNALS-SIGNALS

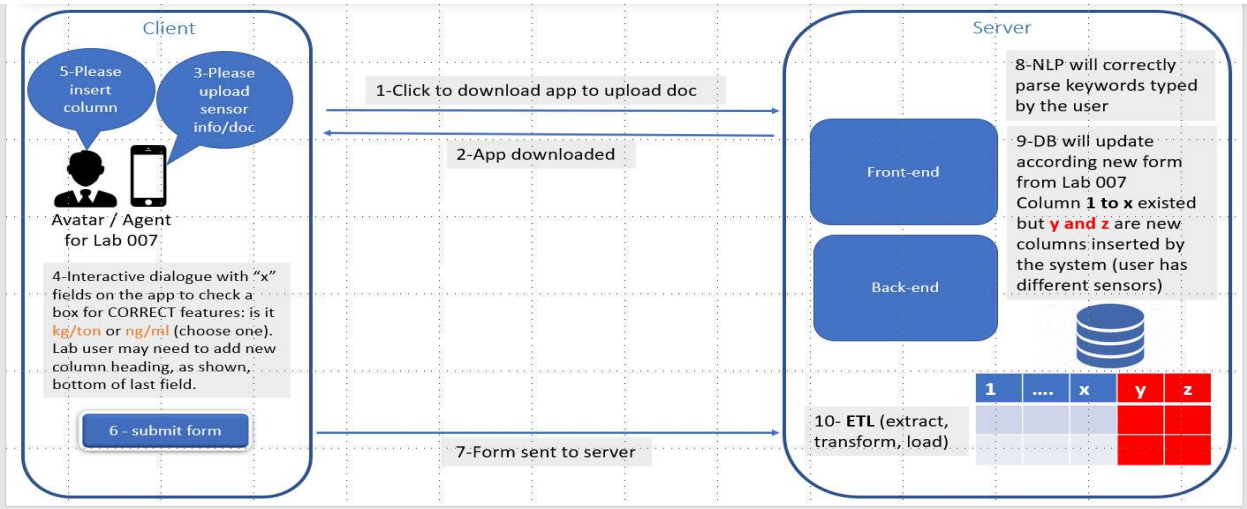
Step I – Dialog Box – http://146.185.133.187/SENSEE1/ ● http://139.162.7.63/SENSEE/

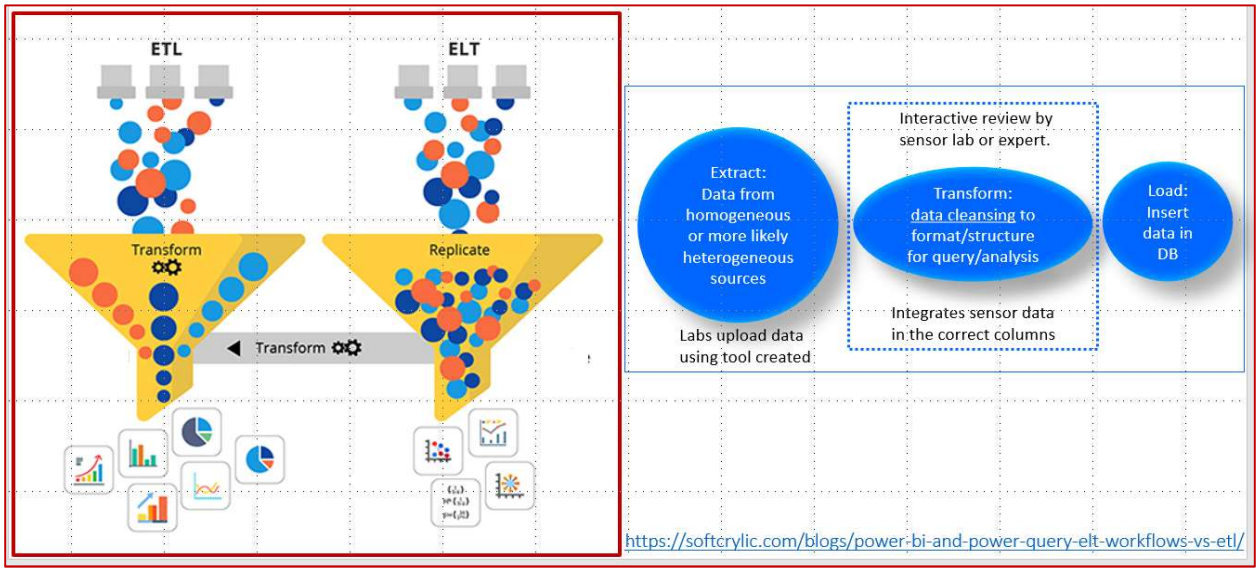


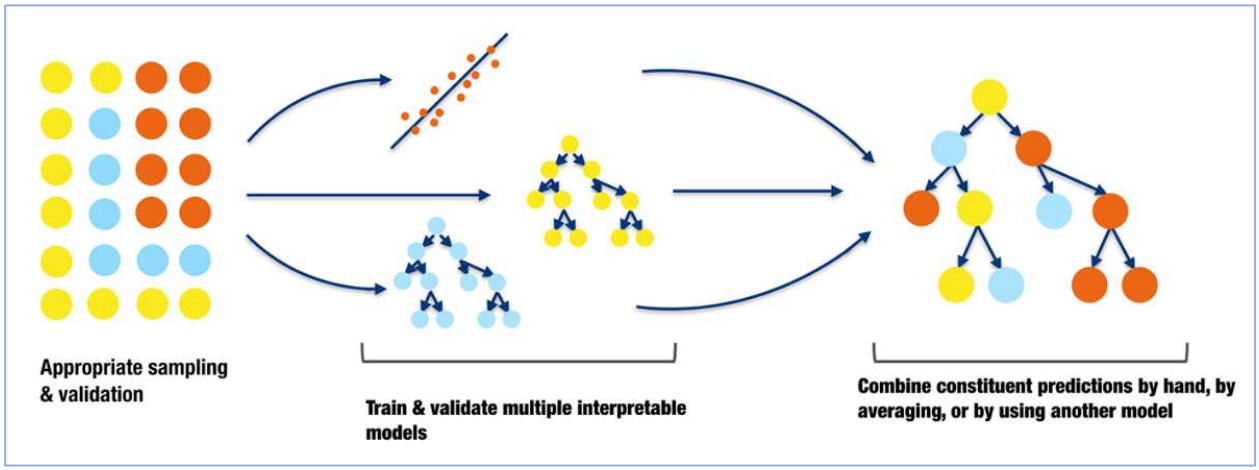




Suggested cartoon for Step II where "tool" may be downloaded via web service/mobile app



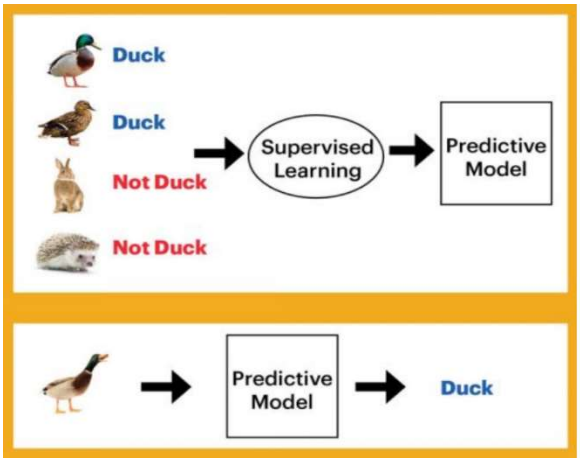


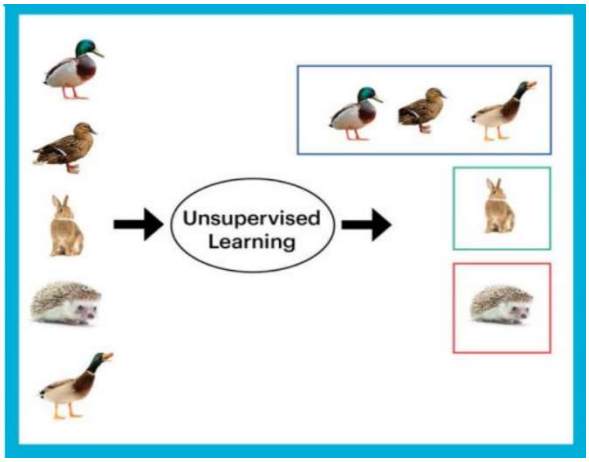


[ii] Online tax preparation systems cannot afford to be shoddy and has developed far better protocols to insert parsed data from uploaded documents (eg W2, W4, 1099 forms). In addition, the software (eg https://turbotax.intuit.com/) may serve as an example of (workflow) interacting with automatically-generated features and interpretation of incoming unstructured (data) sensor description. The value of this software is in the underlying dependencies and UI prompting users to accept or reject the suggestions (see https://www.youtube.com/watch?v=WLaXFYIF4x8).

SENSEE 1.0 tool in step II must be better than resume uploaders and approach the sophistication of the tax preparation packages. The example of tax prep software is relevant for the application in this PoC because it will allow labs to verify uploaded descriptions and characteristic values. It will be helpful (and perhaps necessary) if the DB can be auto "expanded" to create/include new column headings if a new criteria/characteristic is uncovered.

The interactive Q&A type "accept/reject" option in the tax prep software, if implemented in the tool to ingest sensor description data (in SENSEE 2.0 we will ingest and aggregate actual sensor data), will reveal a dataset of "wrong" answers when the user rejects the suggested field value. It is crucial to capture this "what was predicted" vs "what was accepted" "what was rejected" because the correct, and the incorrect input, ***both***, are useful to "classify" what is correct and what is incorrect. Hence, the strength of this approach is in creating datasets for supervised learning (classification algorithms, for example, in the illustration below – duck / not duck).





When using this tool to aggregate SENSEE 1.0 sensor description data (1,000 labs?), it will be important that the application captures every recommendation shown to a user, and the outcome. The major challenge in this approach is the demand for domain specific knowledge (in this case, sensor engineering) and the ability of the individual(s) with domain knowledge to work with a software specialist who has some understanding about the domain. In order to recommend options to the user, the software must rely on a set of rules, dependencies and logic structures, which are relevant to the context of the use (in this case, sensors). For example, the units for the concentration of mercury in a sample may be in ng/ml or in mg or ppb but not in kg/ton or cubic feet. To reduce computational load, perhaps kg/ton or cubic feet, will be excluded as options, to reduce search space, and compress time to search (these metrics will be useful for evaluation).

8 ◆ Exploring how sensor repositories may be helpful. See Figure 11 on page 28 – http://bit.ly/SIGNALS-SIGNALS

It may be informative to review Claude Shannon's "Programming a Computer to Play Chess" (https://vision.unipv.it/IA1/aa2009-2010/ProgrammingaComputerforPlayingChess.pdf) where he coded into the program, knowledge of "weak positions" to limit the search space. Could we use techniques with unstructured documents with sensor information and limit the search space, for example, using topic modeling? In trying to implement these techniques, one cannot escape the need for convergence of domain knowledge with software and principles of machine learning, in tool design. (www.historyofinformation.com/detail.php?entryid=4364 and https://journals.sagepub.com/doi/10.1177/0306312711424596).

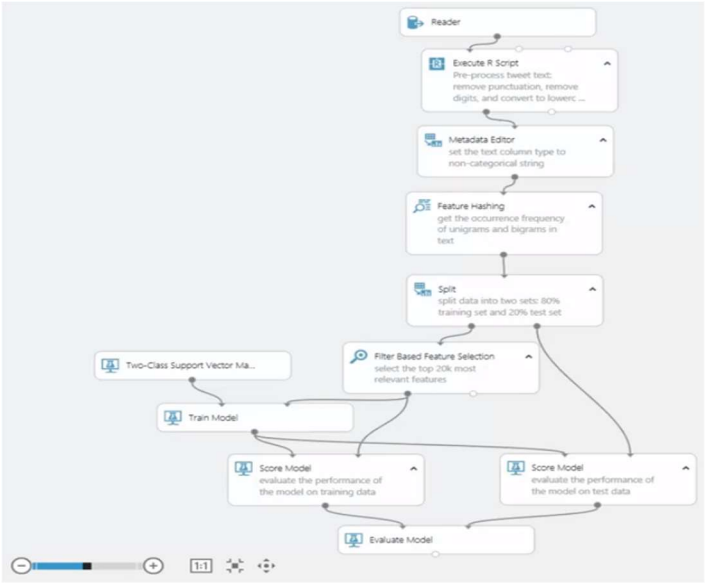
Recommended skill development: To combine sensor data domain expertise with machine learning. Train sensor experts to understand software and machine learning principles and vice versa (machine learning and programmers to grasp the basic tenets of sensor engineering). An exercise for students at UC Berkeley uses resume parsing. The task was to take "pasted-in" text from resumes (just normal ASCII text, no rich PDFs or other formats) to extract "skills" from the content. Using this knowledge, the task was to recommend one skill to a person, which, if acquired, may lead to a recommended job (the skill set the person already possess is one skill away from the recommended job). The goal was to provide guidance for veterans entering civilian life (https://www.shift.org/). Topic modeling was accomplished using genism (https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/ and also https://github.com/RaRe-Technologies/gensim). See https://github.com/jameslamb/skills and view http://bit.ly/JAMES-UC-BERKELEY

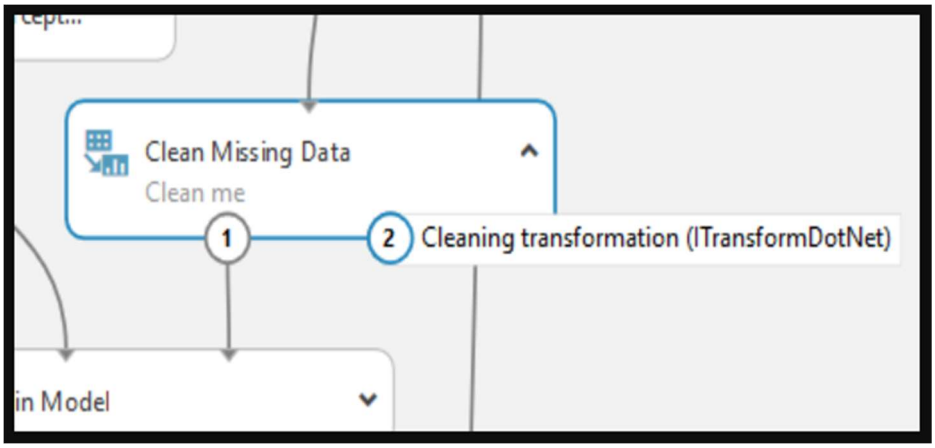
[III] Create 'feature' library / downloadable 'feature tool' to extract sensor details (>1,000 labs)
Target: 2019 ● http://www.jmaxkanter.com/static/papers/DSAA_DSM_2015.pdf
Task: Accomplish the same as [II] but for >1,000 labs.
Comment: Using repository SENSEE 1.0, select features that can accommodate variety of possibilities (expected sensor types from >1,000 labs). Feature selection and feature engineering must serve at least two purposes, the software must be able to operate as a search engine and extract the feature from publications/papers related to sensors (search engine, ML, limited to the library of features). In SENSEE 2.0, feature-guided ingestion of sensor-specific data (pH, salts, temperature, analyte) where the brand of the sensor may be different (different manufacturers) but the pH data from all/any pH sensor, can be pooled into a database using the selected feature (populate the data, that is the pH data, for selected feature pH). The aim of this PoC (at this time) is to create SENSEE 1.0 as a repository for different types of sensors by aggregating ***sensor descriptions*** from labs, worldwide (see examples in xl sheet, link on page 3). When completed, combined SENSEE 1.0 and SENSEE 2.0 is expected to fuel ***artificial reasoning tools*** (**ART**).

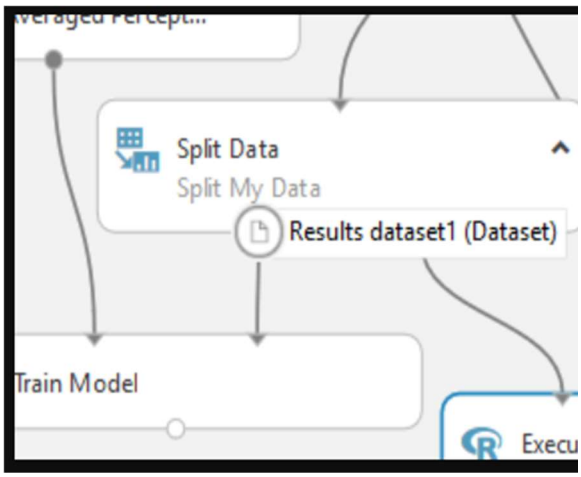
[IV] Automated feature tool linked to **sen**sor **se**arch **e**ngine SENSEE 1.0 (scale to 10,000 labs?)
Target: 2020 ● https://blog.featurelabs.com/deep-feature-synthesis/
Task: "Drag & Drop" tool to gather sensor type (***later,*** sensor data in SENSEE 2.0). Distributed tools: feature automation ● https://people.eecs.berkeley.edu/~dawnsong/papers/icdm-2016.pdf
Comment: Build feature engineering "engines" to process unstructured data and automate tools to arrange it "meaningfully" to serve queries. Idiot-proof "for dummies" interface is provided to users who may rearrange "modules" (data description, knowledge extraction) using ***drag & drop*** tools, perhaps similar to Lego Mindstorm (www.lego.com/en-us/mindstorms/learn-to-program).

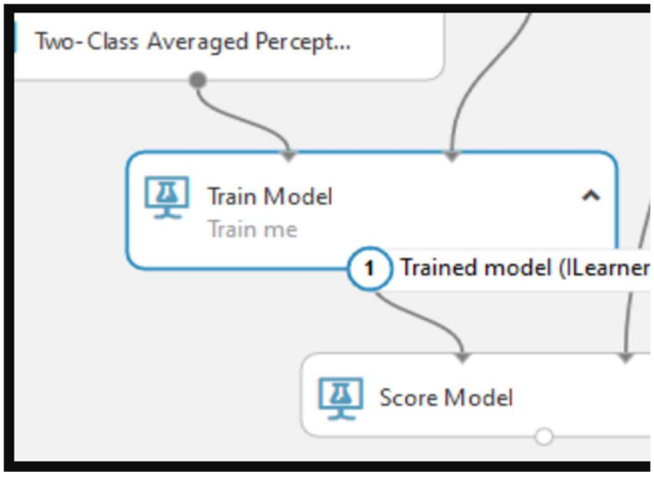
DRAG AND DROP USER INTERFACES - AGNOSTIC ABOUT USER'S KNOWLEDGE OF PROGRAMMING

NEXT GEN LEGO MINDSTORM - EXAMPLE FROM MICROSOFT AZURE MACHINE LEARNING STUDIO









Source: https://github.com/hning86/azuremlps

In this approach, users are **not** required to understand programming, logic and underlying processes. If the masses are not inhibited from using the tool, it will accelerate the diffusion of the tool. Democratization of access through lego-esque, modular "drag and drop" user friendly interfaces, will catalyze adoption of the tool, not only in the agro-ecosystem (the discussion in http://bit.ly/SIGNALS-SIGNALS) but in any domain, for example, healthcare, manufacturing (think digital twins https://arxiv.org/abs/1610.06467), finance, oil & gas, logistics and transport.

Effective and efficient auto-generation and selection of features by modeling information about features (Dawn Song - https://people.eecs.berkeley.edu/~dawnsong/) is a milestone development, we wish to embrace. Using machine learning to ***describe features of features*** and form better expectations of which features might be worth generating and testing, is an incisive advance. ExploreKit feature automation approach is a positive evolution from the brute-force, opaque, model-based approaches to data transformation, which are still a part of machine learning (ML), for example, back propagation (https://www.nature.com/articles/323533a0) and random forests (https://link.springer.com/content/pdf/10.1023%2FA%3A1010933404324.pdf).

In FeatureLabs (https://idss.mit.edu/staff/kalyan-veeramachaneni/), feature engineering is performed relative to richer descriptions of input data and successfully applied for commercial purposes by corporations (Einstein AutoML https://www.salesforce.com/video/1776007/ and https://github.com/salesforce/TransmogrifAI/tree/master/features/src/main/scala/com/salesforce/op/features; also see – improving AutoML transparency – https://arxiv.org/pdf/1902.05009.pdf).

One aim of this project is to partner with bonafide experts to create open feature automation tools. The principles, as indicated above, will be applicable to a broad spectrum of applications.

[V] Gift SmartPath SENSEE tool (USDA/NIFA/NSF) - national/global sensor data repository
Target: 2020
Task: Monitor and Model ➔ Detect and Predict ➔Diagnose and Explain ➔ Decide and Act (actions generate outcomes, which are monitored, and the data-informed process is repeated).
Comment: Open tool will democratize access to data and can be adapted for other domains.

Future Scope (FS)

[FS-VI] Development of SERVICE – "PEAS Platform for Agro-Ecosystem" for users. Farmers and growers (food) seeking solutions from logic tools and help from **ART**. (Target FY 2021)

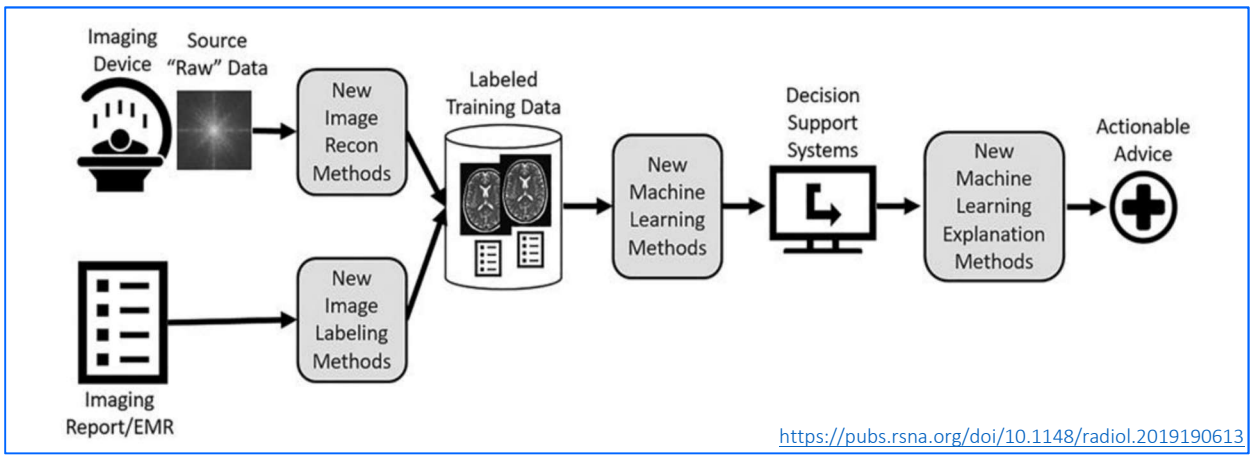
[FS-VII] Pay-A-Penny-Per-Use (**PAPPU**) data-informed decision as a service (**DIDA'S**) is a far-fetched fusion between data, science, systems and social business. Convergence of [V] and [VI] creates value from uber connectivity between sensors and may provide "meaningful" outcomes. The "meaning" will evolve if we develop tools (II, III, IV) to get the data into a form where it is anomaly-free and structured for use by logic tools in **ART**. Later, use of math-stat functions and ML. Turning statistical functions (or mathematical formulae) into algorithms, is not trivial. The rate limiting step is human resource. Teams must possess the confluence of skills to understand system science (sensor engineering), data science (stat, math, programming, computation) and *converge* it with their depth in *domain knowledge*, to address and solve, real-world problems.

11 ◆ Exploring how sensor repositories may be helpful. See Figure 11 on page 28 – http://bit.ly/SIGNALS-SIGNALS

This PoC is not a “software job” and professional programmers are unlikely to succeed because very few classical programmers can generate an/any optimizing sort algorithm, from scratch. What is referred to as “data science programming” is a type of app development which uses a set of pre-defined frameworks along with specific domain knowledge (in this case statistics) to create solutions. Before R and Python, "data science" was executed on spreadsheets, followed by statistical software (SAS, SPSS, statistical package for social sciences) in the era of desktops.

Data-informed service for the agroecosystem must be hyper-mobile, with high fault tolerance, operate close to the point of action (edge) in near real-time, accommodate engineering elements which are diverse, operate seamlessly agnostic of data standards or structure and must service consumer/user demands which may change often or fluctuate rapidly. Dynamic composability of data and synthesis of information, relevant to the context, is the desired outcome, even for **ART**.

[FS-VIII] Entrepreneurial Innovation – Users may pay for this service (see PAPPU DIDA’S) if we create a visual tool for non-expert end-users to grasp the curated information in SENSEE and how it may integrate with sensor data (**ART**) and machines (digital twins). To democratize access, this PoC advocates intuitive/cognitive maps (re-think Lego Mindstorms and topology optimization software, for example, see www.ansys.com/) which offers “drag and drop” icons (tactile, haptics) to orient users and catalyze the connectivity and complexity with lucidity, clarity and brevity. **Educating** and enabling users to make sense of the organization of unstructured knowledge will immensely demystify “blackboxes” and aid in the diffusion of **ART** tools, leading to reasonable adoption. By providing a mechanism to represent existing information, knowledge graphs in logic tools describe and enable access to other information sets. Diffusion of **ART** may lead to a better educated crowd and improve crowd-sourced (farmers, growers) **architecture to access knowledge.** It is a prelude to the development of an open platform for convergence of data and information, in a meaningful context, to move beyond logic tools in **ART** to data-informed **DIDA’S** and then to ***knowledge-informed*** decision as a service (**KIDS**). In 1980’s decision science, DIDAS (https://link.springer.com/chapter/10.1007/978-3-662-21637-8_2) was a control theory concept. The use of DIDA’S in this document is also about decisions and, in principle, it may resonate with “DIDAS Family” (for automatic control). The sense of **DIDA’S** in this PoC is a step after **ART** and before **KIDS**. Beyond knowledge, the extraction of ‘experience’ may enrich the outcome from **KIDS**, but it will be difficult and must include agent-based selection (**ABS**).



In the short term, to deploy **ART**, we expect to create SENSEE 1.0 and improve its ability to deal with a broad range of questions, before sourcing sensor-specific data for SENSEE 2.0 PoC.

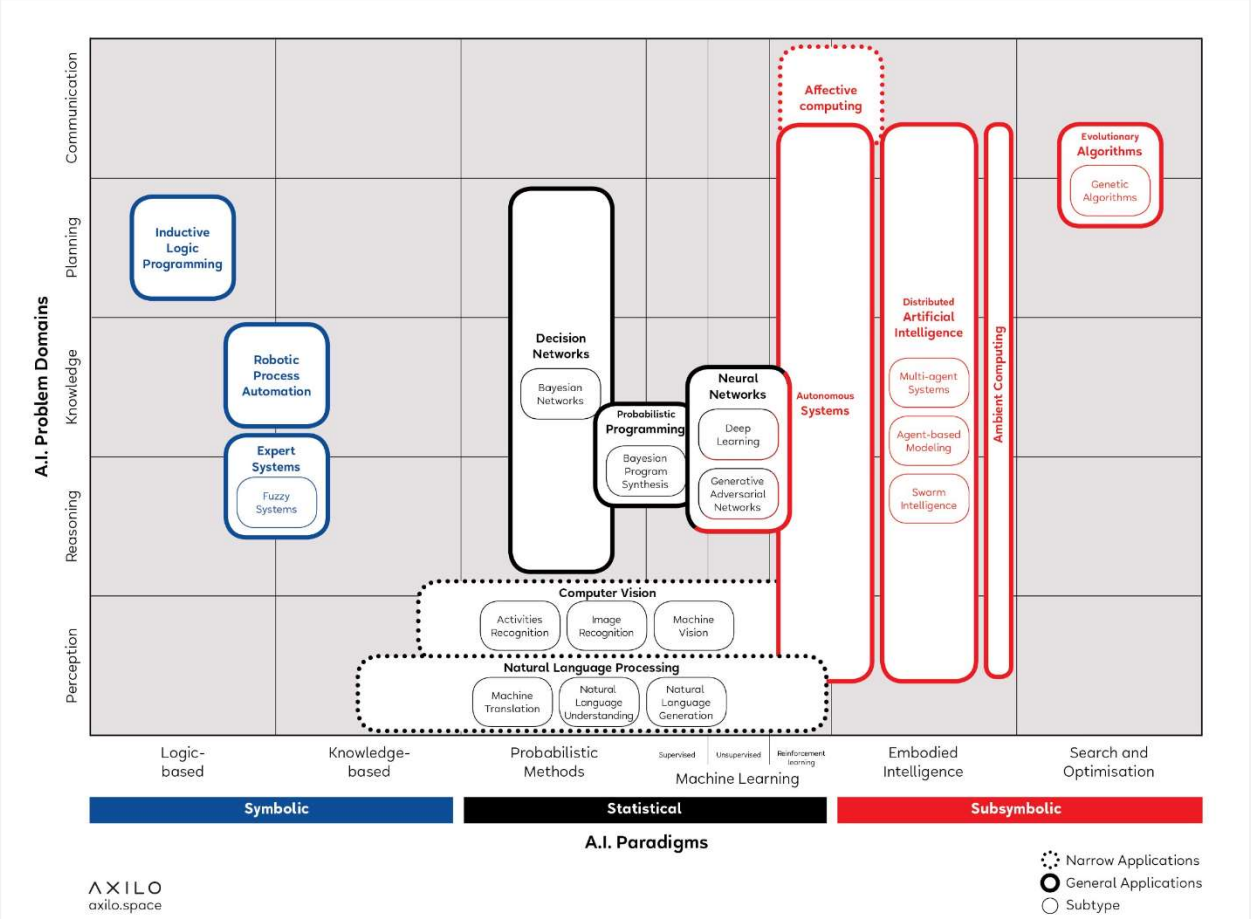
1) Which sensor in the McLamore lab has the highest sensitivity?
2) Which sensor in the McLamore lab has the lowest LOD?
3) Which sensor in the McLamore lab has the highest selectivity?
4) Which sensor in the McLamore lab has the fastest response time?
5) Which sensor in the McLamore lab has the highest durability?
6) What is the most durable glass capillary sensor?
7) What sensors can be fabricated on conductive paper?
8) What sensors can be made with nanocellulose?
9) What sensors can be made with cabbage extract/anthocyanin?
10) What sensors are used for hydroponics research?
11) What sensors are used for irrigation water research?
12) What sensors are used for cell culture research?
13) What sensors are used for lake water research?
14) What sensors are used for wastewater research?
15) What sensors are used for plant roots research?
16) What sensors are used for coastal monitoring/seawater research?
17) What sensors are used for tissue culture research?
18) What sensors are used for stem cell development research?
19) What sensors are used for differentiated stem cells/neurons research?
20) What sensors are used for wound dressings research?
21) What sensors are used for osteoblast/osteoclast research?
22) What sensors are used for INS1 cell research?
23) What sensors are used for blood research?
24) What sensors are used for human tears research?
25) What sensors are used for mouse pancreas research?
26) What sensors are used for honeybee wax research?
27) What sensors are used for honeybee honey research?
28) What sensors are used for saliva research?
29) What sensors are used for food product research?
30) What sensors are used for food packaging research?
31) What sensors are used for juice research?
32) What sensors are used for soup/broth research?
33) What sensors are used for ice cream research?
34) What sensors are used for drinking water research?
35) How many sensors measure H+/hydronium ion/hydrogen?
36) How many sensors measure NH4+/ammonium ion?
37) How many sensors measure NO/ nitrogenous radical/nitrous oxide?
38) How many sensors measure H2O2/O3 oxygen radical/hydrogen peroxide?
39) How many sensors measure DO/dissolved oxygen?

40) How many sensors measure K+/potassium ion?
41) How many sensors measure Ca2+/calcium ion?
42) How many sensors measure CH3COO-/acetate?
43) How many sensors measure NO2-/nitrite?
44) How many sensors measure NO3-/nitrate?
45) How many sensors measure Ag+/silver?
46) How many sensors measure histamine?
47) How many sensors measure glutamate?
48) How many sensors measure catecholamine?
49) How many sensors measure indole acetic acid?
50) How many sensors measure malate?
51) How many sensors measure glucose?
52) How many sensors measure ionic mercury?
53) How many sensors measure methyl mercury?
54) How many sensors measure paraoxon?
55) How many sensors measure ATP/adenosine triphosphate?
56) How many sensors measure MBF1/multi bridging factor 1?
57) How many sensors measure interferon gamma?
58) How many sensors measure superoxide dismutase?
59) How many sensors measure E. coli?
60) How many sensors measure E. coli O157:H7?
61) How many sensors measure Salmonella?
62) How many sensors measure Listeria monocytogenes?
63) How many sensors measure Campylobacter?
64) How many sensors use carbon nanotubes?
65) How many sensors use graphene?
66) How many sensors use graphene oxide/ GOx?
67) How many sensors use graphite?
68) How many sensors use glassy carbon?
69) How many sensors use liquid ionophore membrane?
70) How many sensors use solid state ionophore membrane?
71) How many sensors use nanoplatinum?
72) How many sensors use nanoceria?
73) How many sensors use nano titanium dioxide/nTiO2?
74) How many sensors use nano zinc dioxide/nZnO2?
75) How many sensors use platinum porphyrin dye?
76) How many sensors use fractal materials?
77) How many sensors use nano palladium/nPd?
78) How many sensors use diamine oxidase?
79) How many sensors use aptamer?

80) How many sensors use antibody?
81) How many sensors use lectin?
82) How many sensors use phage?
83) How many sensors use alkanethiol?
84) How many sensors use nano copper/nCu?
85) How many sensors use copper oxide/Cu2O?
86) How many sensors use phosphotriesterase?
87) How many sensors use chitosan/CHI?
88) How many sensors use PNIPAAm/ poly(N-isopropylacrylamide)?
89) How many sensors use hydrogel?
90) How many sensors measure CIP2A/ Cell Proliferation Regulating Inhibitor of Protein Phosphatase 2A?
91) How many sensors measure internalin A/ InLA?
92) How many sensors use concanavalin A / ConA?
93) How many sensors use mannose-binding lectin ?
94) How many sensors use C-type lectin?
95) How many sensors use specific intercellular adhesion molecule-3-grabbing nonintegrin / SIGN-R1?
96) How many sensors use wheat germ agglutinin N-type lectin?
97) How many sensors use lectin for N-acetyl-D-glucosamine (NAG)?
98) How many sensors use F-type lectin?
99) How many sensors use fucose binding lectin / FUC?
100) What sensors can be fabricated with glass capillary?
101) What sensors can be fabricated with graphene paper?
102) What sensors can be fabricated with laser scribed graphene/laser inscribed graphene?
103) What sensors can be fabricated with a platinum/iridium electrode?
104) What sensors can be fabricated with a 96 well microtiter plate?
105) What sensors can be fabricated with Pt/Ir microelectrode wire?
106) What sensors can be fabricated with a gold IDE?
107) What sensors can be fabricated with DropSense IDE?
108) What sensors can be fabricated with Zensor SPE?
109) What sensors have the lowest LOD **AND** the highest selectivity?
110) What sensors have the highest LOD **AND** the lowest selectivity?
111) What sensors have the lowest LOD **AND** the lowest selectivity?
112) What is the cheapest **AND** most reliable sensor for measuring mercury in water?
113) Which sensors can I use to detect pathogens in juice?
114) What is the smallest amount of glucose that a sensor can detect in tears?
115) Are paraoxon sensors commercially available in Canada?
116) How long does it take to detect salmonella with a biosensor?

117) How stable are LSG sensors?
118) What chemical components of hydroponic media can be measured with sensors?
119) Which sensors can be used to detect toxins in breast milk?
120) What is the simplest **AND** most cost-effective sensor platform for detecting glucose in buffer?
121) What sensors are available for testing pollutants in lake water?
122) What is the most common material used for fabricating hydrogen sensors?
123) What is the most versatile glucose sensor?
124) What's the most popular material platform for building ammonium sensors?
125) What is the lowest detection limit for ATP sensors?
126) How specific are calcium sensors?
127) Do listeria sensors have a high chance of producing false positive results?
128) Are wastewater sensors reusable?
129) What is the saturation concentration for glutamate biosensors?
130) Which kinds of sensors are selective towards E. coli O151:H7?
131) What's the widest operating range for a H2O2 sensor?
132) Which sensor has the lowest durability?
133) Which sensor has the highest durability?
134) What is the lowest LOD among all sensors?
135) How many different kinds of platforms have been adopted in sensor testing?
136) What is the best LOD for H2O2 testing?
137) What type of H2O2 sensor has the best selectivity?
138) Which platform should we use when we test the NH_4^+ regardless of price?
139) Does the liquid ionophore always have better performance than the solid ionophore?
140) Which type of sensor has the longest response time?
141) What is the fastest sensor in small molecule testing?
142) What is the most commonly used sample in the test?
143) Which sensor has the largest range in Glucose test?
144) Which platform is most commonly used in Mclamore's lab?
145) Is it possible to test H2O2 in ocean water?
146) What recognition-transduction scheme do we use to detect glucose?
147) Is it possible for a sensor to have a response time lower than 0.1 sec?
148) Which platform should we use when we want to detect the NH4+ in two seconds?
149) Could we use the LSG to detect H+?
150) Can we make a bacteria sensor whose response time is <500 seconds?
151) What is the response time for liquid K+ ionophore in detecting K+?
152) What is the most popular platform for hydrogen peroxide biosensors?
153) Which targets can be determined using diamine oxidase sensors?
154) In which samples potassium ions can be detected?

155) What molecules can be detected in breast milk using biosensors?
156) What is the difference in sensitivity between glucose biosensors based on graphene or platinum foil?
157) What is the most sensitive biosensor based on carbon nanotubes?
158) How many biosensors have been proposed for glucose determination?
159) Anthocyanin is used as a target for which biosensor?
160) Which biosensors can be used for hydroponic medium?
161) In which samples, glutamate was determined using biosensors?
162) Which biosensors were proposed for catecholamine determination?
163) What is the lowest limit of detection for graphene-based biosensors?
164) What is the maximal range for nitrate biosensors?
165) What platforms can be used for ammonium detection?
166) Most durable recognition-transduction scheme for interferon gamma biosensors?
167) Best limit of detection achieved with phosphotriesterase-based biosensors?
168) How many biosensors were described for ATP determination?
169) What platforms were proposed for ATP-sensitive biosensors?
170) What is the average LOD of K+ sensors?
171) Which platform could be used for selective glutamate analysis?
172) What is largest analyte/molecule for which there is a sensor in the database?
173) Is there any cost associated with any type of sensor?
174) How many labs are making sensors to detect lead in water?
175) Are there sensors to detect air-borne viruses in the air?



DEMYSTIFYING THE ANALYTICS BLACKBOX – A PREREQUISITE FOR ADOPTION?
https://arxiv.org/pdf/1902.05009.pdf

Snake oil sales of "intelligent" tools in the name of "AI" is an anathema to those who respect, appreciate and understand, albeit in part, the immense contribution of scientists and engineers who delve into details in quest of robust, evidence-based, numerically-supported, i/o, even for the basic form of artificial reasoning. Even then, it is not a general "one-shoe-fits-all" app that can be peddled willy-nilly. If users are better equipped to ask probing and precise questions, the tools and systems (eg in this cartoon) can serve the users, perhaps with greater accuracy and precision, before the information perishes. The most common question "what happened" (descriptive analytics), may lead to "why did it happen" (diagnostic analytics). The collection of logic tools (ART) we have discussed in this document and the cartoon (above) must work in confluence to respond to the most likely follow-up question "what is going to happen, next" (predictive analytics) and then the obvious: "what is your recommendation, what should I do" (prescriptive analytics). In PAPPU DIDA'S concept of data-informed decision as a service, prescriptive analysis may suffice for human-in-the-loop systems where the actuation is human-controlled. With increasing scale of concurrent levels of operation and improving control of automation (think about 0-5 levels of autonomy, think OODA concept by John Boyd and PEAS paradigm in Agent based systems), it is not impossible that users may eventually trust and enable systems with case-specific permissions to "take action, execute" (automated analytics). Auto-actuation, in the SARS to SARA paradigm, is discussed in the essay with the title SARS♦AG.

18 ◆ Exploring how sensor repositories may be helpful. See Figure 11 on page 28 – http://bit.ly/SIGNALS-SIGNALS

PAPPU DIDA'S – BEFORE WE ASPIRE FOR KIDS

Irrespective of the strength of ideas, in this and other essays, the path to adoption is fraught with challenges. Therefore, ideas are often useless and without value unless we tackle the hardest questions, first, which defines the pragmatic aspect: will anybody pay to use this idea in reality? Ultimately it is the economics of technology which defines and controls the diffusion of ideas.

The complicated answer has several moving parts and none may be fully correct or incorrect. One canonical response is that adoption will be determined by the cost versus value paradigm. In a global economy where products are receding in the background and services (including those services which are based on products, eg, washing machines) are gaining momentum, the idea of "PEAS Platform for the Agro-Ecosystem" begs to ask whether one expects that users will buy sensors, of different types, in bulk. How will the non-expert users collect, analyze and integrate the data from the sensors. to help them in their real task, the task of food production?

The physical product (sensor) must deliver value (data, decision) which will inform responses and lead to actual work (actuation) to improve ag systems and help to increase food production.

One option is an age-old, time-tested, solution where lowered cost to the user (opex) is a function of the frequency of use and generally free from sunk or capital costs (capex). In the last century, this model was epitomized by POTS, the plain old telephone system, where the user paid only the "charge per call" which was reasonably affordable even when the per capita income was low.

Pay a penny per use (PAPPU) re-invents POTS with the qualifier that the user pays a penny (US) for each use (perhaps unwise to restrict it to one penny). The "use" may not be a thing, object or tangible product but rather a "process" which we refer to as data-informed decision as a service (DIDA'S). The "penny per use" idea may draw scorn from certain segments of investors and corporate leaders because the idea does not support the "next quarter" earnings (greed) report.

A version of PAPPU (pay a penny per **unit**) could stretch to fit "99 cents hamburger" model evident in PayPal's 2018 revenue ($16 billion from 12 billion transactions, $1.25 / transaction). The "unit" view of PAPPU may be applicable in transport, energy, water (as **units** delivered).

SENSEE, ART, DIDA'S and other data-informed decision-support on an open-platform calls for synergistic systems integration. The value is realized at the "end" when systems data may be synthesized to provide meaningful use in the context of the problem. It delivers information at the point of use, at the edge. Is this of actionable value for the user? Consumers may pay only for the desired outcome. Transaction cost economics is perhaps key to this *modus operandi.*

If the outcomes are dependent on a plethora of sequences in the operational process, then each process is a "profit center" and may generate a penny in revenue each time the user "touches" the system to extract information (or knowledge). If the economy can bear the economics of PAPPU then systems diffusion and adoption will continue to grow (decades) based on the economy (until saturation, when demand plateaus irrespective of cost). The number of sensors, and other data, are likely to intersect with vast number of decisions (ART, DIDA'S, KIDS). The actual *transactional volume* of payments, from 'micro' or 'nano' payments, are potentially gargantuan.

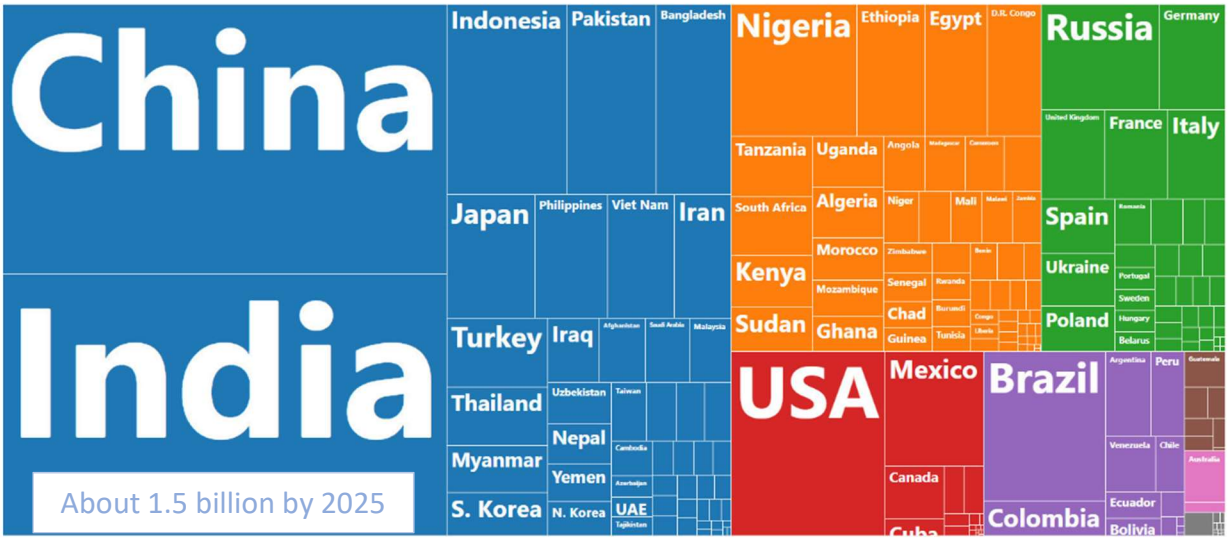
19 ◆ Exploring how sensor repositories may be helpful. See Figure 11 on page 28 – http://bit.ly/SIGNALS-SIGNALS

Documenting that the system was "touched" and billing/collecting that one penny is a technical challenge which requires tracking events (think IPv6, as an "indicator" for ***system*** activity). The task of segmenting that one penny revenue, between several service providers, may be a massive challenge in "weighted" decomposition/recomposition of events, to distribute earnings based on the degree of contribution of the provider who executed that step/event (for example, sensor manufacturer, systems integrator, platform provider, software vendor, analytics, mobile fintech).

Since no new "physics" is necessary to delineate these processes, it is safe to state that these can be accomplished without any invention but with forward thinking imagination and innovation. It is a déjà vu scenario from the "Store of the Future" (2000-2001, RFID track and trace) which sputtered and asphyxiated in the face of systems integration challenges, only to be resurrected by Amazon, which, finally, implemented the retail concept in Amazon's GO (September 2018).

Increasingly, PAPPU (DIDA'S) will be the monetization mantra for the ART-IoT generation and the future where equality, equity and égalité may re-claim its rightful place in society striving for ethical profitability. It may take 20-30 years to overcome the resistance from despots, investors and corporate behemoths, but eventually the infectious spread of this concept may succeed in sowing a critical-mass of practitioners. The concomitant growth of infrastructure (for example, affordable access to low latency, reduced jitter, high bandwidth wireless telecommunications, 5G, trusted mobile banking) may be necessary to pave the road for PAPPU in ART and DIDA'S.

The ability to escape the dead weight of old technology (eg Africa, Asia) may accelerate the implementation of ***pay a penny per unit*** (PAPPU) as an integral part of the socio-economic fabric of a product-less, service-based economy, which may exclude the tiny population residing in OECD nations and/or the red and green zones in the cartoon show below.



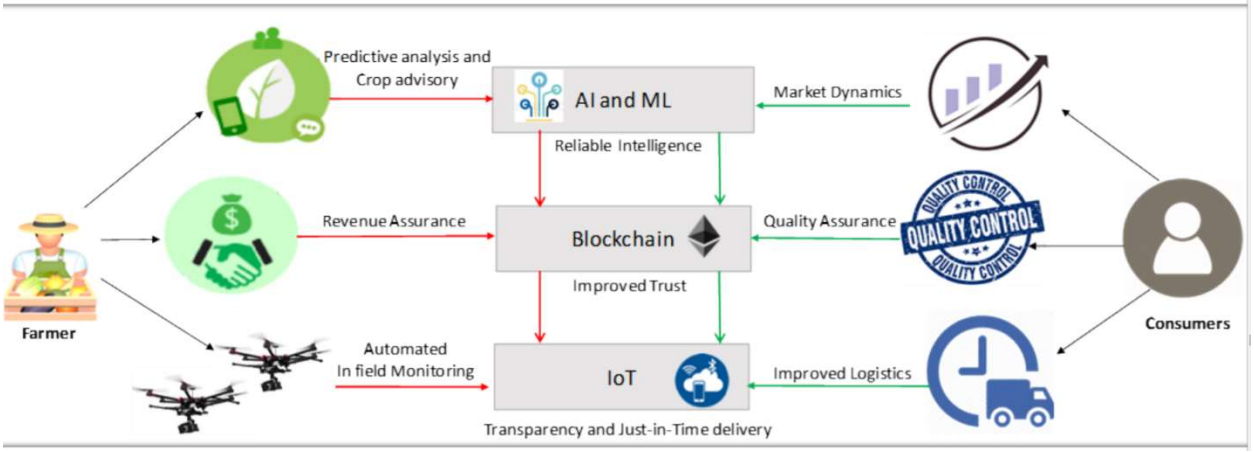
PAPPU may evolve as a preferred business model for the global economy by lowering the barrier to entry into markets where people are surviving on about $2 per day. The impact may be especially profound on healthcare and the agroecosystem for production of food.

Because PAPPU is inextricably linked to mobile payments, distributed banking and digital finance, the pundits of social media will jump on this discussion to claim PAPPU is incomplete unless "blockchain" is integrated in the process. Blockchain hype-mongers are worse than snake oil sales and the adage or aphorism "hammer in search of a nail" seems too respectful in view of the torrent of garbage that is spewed in the name of blockchain. However, trust in transaction is undeniably central and an age-old concept (https://www.jstor.org/stable/20752121).

Therefore, it is important for PAPPU to provide tools to ensure safety of the payment system and other steps where verification guarantees are related to the service or product (for example, food safety). But, informed organizations may not, blindly, consider blockchain security for PAPPU.

Whether and how and in what form the concepts in blockchain may be helpful, remains to be seen. It is not entirely useless and such "solutions on steroids" deserves a place in society to counter the unethical practices that rapidly multiply in financial operations. However, such specific examples of use, and value of blockchain, may not be ***generalized*** as a solution for ***all*** levels of transactions. It is deceitful and malicious for blockchain proponents to tarnish all verticals and industries using the broad brush of finesse that is rampant in the financial industry.

Blockchain is erupting into an euphemism for avarice, for the sector of people involved in the process of marketing tools for blockchain. It is an anathema for >80% of the world trying to survive beyond the gluttonous grip of tools and technologies of dubious value. Blockchain is certainly not a panacea. There may be a few other low-cost ways to achieve safety, security, identification and authorization (for example, https://dspace.mit.edu/handle/1721.1/102893).



Chacun voit midi à sa porte – hammer in search of a nail: peddling the "blockchain" at the "center of the world view" of operations. It is not necessary for individuals in trains, planes and automobiles to wear an armor-suit. The safety belt is sufficient, although it may not be enough, in certain instances. The latter is the risk that emanates from the rewards due to progress, which society has, and will continue to, shoulder. Rather than feeding people, the burden of blockchain will starve the hungry, where food is most needed, by increasing cost of operations. Imposing rules and regulations will secure profit for the blockchain industry, deliver little for food safety and deprive nations from food. (https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8598784)

FAAQ - FREQUENTLY ASKED ANGRY QUESTIONS

■ Is this really a PoC ?

■ What is the purpose of repository (pg 11) in step V ?

Explanation of the lack of purpose: unclear from SENSEE PoC alone

Steps I and II (see page 5) are indeed proofs of concept to show we can create a simple mobile dialog box to ask questions (examples on page 3) about types of sensors available from a small group of labs (10-100) who are creating sensors. One example from the McLamore Lab is in the xl sheet available from http://bit.ly/SENSOR-LIBRARY-ERIC-MCLAMORE

The important distinction, with respect to this discussion, is that, this is ***not about sensor data alone***. We discuss ***types of sensors*** (SENSEE 1.0), ***sensor-specific data*** (SENSEE 2.0) and phases of decision support (ART, DIDA'S, KIDS – PEAS Platform for the Agro-Ecosystem).

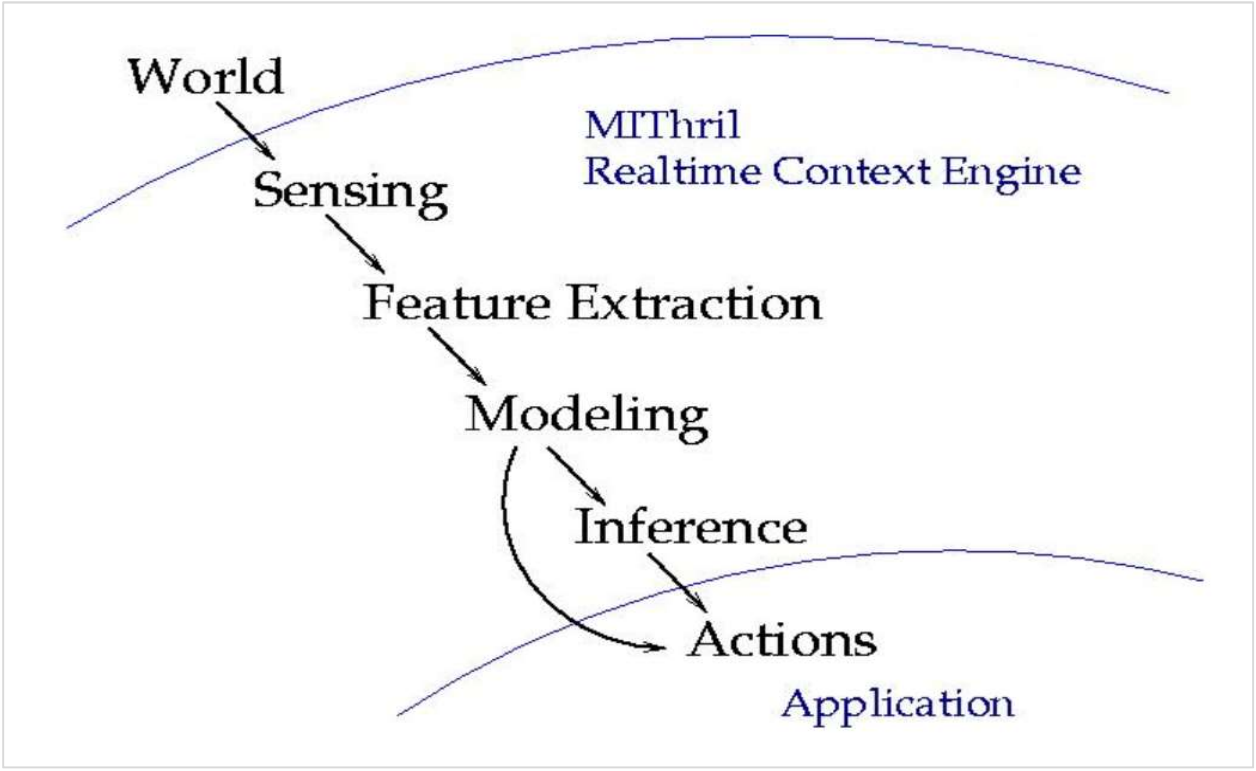
SENSEE 1.0 is the PoC which refers to characteristics, properties and features of sensors, as engineered by academic or industry labs, (tabulated in xl sheet). When we mention "users" typing in a query in the dialog box, the "user" is **NOT** a farmer, grower, consumer. SENSEE 1.0 users are experts and sensor engineers or students exploring the repository (step V). Who has made what sensors? What is the limit of detection? What platforms (impedance spectroscopy, photonics, SERS) were used to capture the signal? Critics may point out that Google has done that job and Google will retrieve millions of references to papers and documents with any key word. In defense of our aspirational step V (the repository), we wish to point out that the queries (examples on page 3) using the SENSEE 1.9 dialog box (step I) may reveal specific information, from a ***curated catalog*** of information, related to sensor science and engineering.

> ***Necessary Digression:*** In the context of the "big picture" it may be worthwhile to understand that this "curated catalog" refers to the xl sheet with sensor specifications (example provided from McLamore Lab). In step II we will have, hopefully, extracted similar information provided by another few labs. This collection of information will be in databases (SQL, noSQL, GraphQL, SPARQL, TSDB) which will serve to retrieve responses to questions (example on page 3). Hence, this "curated catalog" in this discussion may lead to the repository (see step V on page 11). The reason for this preface to the "digression" is to point out how this "curated catalog" may morph into a "semantic data catalog" as explained in **Figure 12** in the **"SITS"** section of the essay series "SIGNALS" which may be downloaded from http://bit.ly/SIGNALS-SIGNALS. In steps I through III we may use elasticsearch and NLP to parse words in a query and match with keywords (in search engine) to retrieve the most likely answers. This *syntax-driven* process is error-prone (depends on how well the human developer has coded/trained the engine). To boost performance, semantics may remove, albeit partially, some errors due to syntax. The latter varies with expertise, mother language and social environment. Hence, the intended move to semantic databases may be facilitated by transforming this "curated catalog" (this discussion) to the future ***semantic data catalog***.

One may conclude this is a repository for experts. At this time, this ***is*** true. This is the beginning of an ambitious attempt toward ***synergistic integration of platforms*** (SIP) which will converge data (**ART**), information (**DIDA'S**), knowledge (**KIDS**) and, eventually, *experience*, to suggest solutions. Users may be experts from industries, farmers, meat packers, distributors, food inspectors, grocery stores, cold chain logistics providers. It may be **any user** who may benefit from data-informed decision as a service (DIDA'S) and may even pay for the service (PAPPU). The future of ***experience as a service*** may be your personal mobile agent, which you lease to a buyer, or air-drop (www.imore.com/airdrop) to a local client, interested to profit from your experience.

Broad spectrum of users (above) may need different types of information, which may be in different databases. Hence, the choice of knowledge graph connectivity to synthesize and deliver a meaningful response, by selecting data and information using distributed architecture to access a multitude of resources (see figure 11 on page 28 in the "SITS" section of "SIGNALS" http://bit.ly/SIGNALS-SIGNALS). Extraction of data and information from the "world" may benefit from **context engine** architecture (cartoon below). The latter may be one way to create knowledge bases without reliance on ontologies, using publish/subscribe (ingest from CSV, xl, relational databases, JSON/XML feeds), perhaps in a manner analogous to CMS (content management software for data). Ontologies may become key to future knowledge extraction.

This endeavor is **ONE of the resources** (SENSEE 1.0) we aim to develop to address that *future* **SIP platform**. We expect that aggregation of **contextual data** and ***curated*** information may further improve the performance of this basic service (ART) through connectivity with other distributed resources (see **"SARS♦AG"** in SIGNALS http://bit.ly/SIGNALS-SIGNALS). To achieve that goal, the open source platform must support data interoperability (for example, DDS) between local and global databases/platforms, enable dynamic composability to pick and choose (drag & drop) data/information from diverse sources, always explore user-friendly tools for synergistic integration with domains of data, information and ***crowd-sourced*** knowledge, which may enable user ***experiences*** from the past to inform the future. Also, we expect ***actual "world" sensor data*** (eg temperature sensor) to be aggregated, ***agnostic*** of the make and model of the sensor (SENSEE 2.0). Sensor *data*, and extracted *information*, may be *more useful* for pragmatic, and profitable applications, in the near-term, that we may deliver through **ART**.



The 'world' in context of applications. www.media.mit.edu/wearables/mithril/context/index.html

Science and engineering have enabled an embarrassing wealth of sensors but without an organizational repository (aspirational step V on page 11) the value of these sensors may remain under-utilized. The proposal to create a World Sensor Organization (**WSO**) to address these issues, remains unexplored (Commentary [C] discusses WSO, in the PDF, **"IoT is a Metaphor"** which may be download from https://dspace.mit.edu/handle/1721.1/111021).

To catalyze science to serve society, in a parallel endeavor (Eric McLamore, personal communication), experts interfacing with the edge, that is, with end-users (farmers, growers, meat packers, aquaculture, retail grocery suppliers), are attempting to harvest questions which end-users may ask or *should* know, in order to better use data to inform and transform their practices (address contamination, understand regulation, use of technology). The measure of success is the outcome (***PEAS Platform for the Agro-Ecosystem***), in terms of food production at an affordable cost, of a better quality, as well as quantity, using ethical tools and practices.

To clarify, this PoC may be divided into an actual proof of concept phase (steps I and II) and a R&D approach in steps III through V (which is no longer just a PoC but a more thoughtful path to step V). The sum of PoC plus R&D is an essential (***but only one***) part of the SIP platform concept discussed in SARS♦AG (see essay **"SARS♦AG"** in http://bit.ly/SIGNALS-SIGNALS).

SARS♦AG combines the tools and sensors discussed in "SITS" (see essay **"SITS"** in SIGNALS http://bit.ly/SIGNALS-SIGNALS) with questions that **end-users** may want to ask. This ***combination*** makes it possible to bridge the wealth of advances in sensor research with the need for tools and technologies, on the ground, at a pragmatic level. Data-informed end-users may meaningfully converge this knowledge, with their experiences, in order to improve the outcomes (food production, food distribution, food safety, prevent food wastage, profit margins).

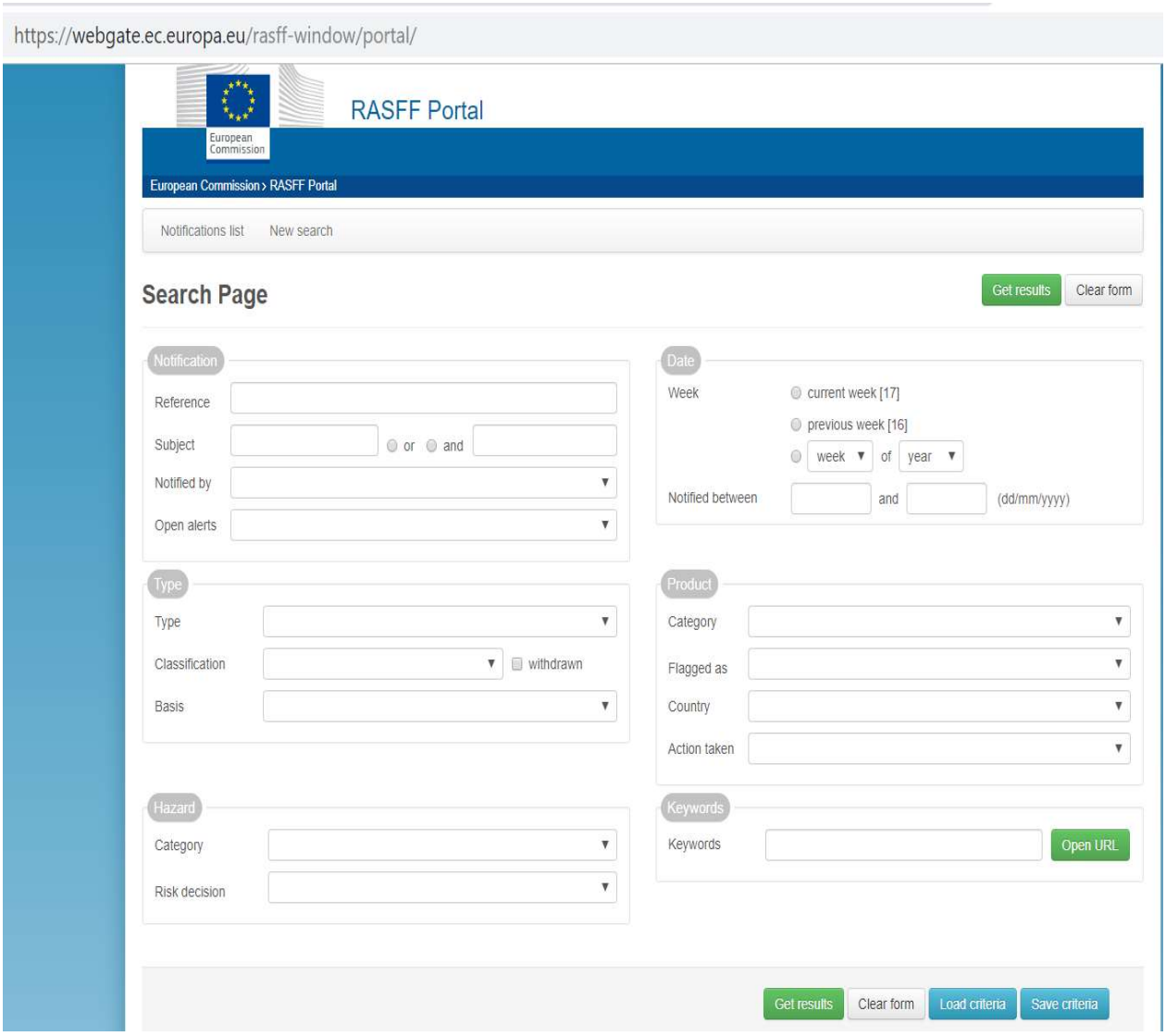
Questions, whose answers may help end-users, are the sign-posts for the development of ***PEAS Platform for the Agro-Ecosystem***. Sensor repositories (SENSEE 1.0, SENSEE 2.0) must meaningfully connect with questions from the field-workers. Some of these questions may have nothing to do with sensor science and engineering. Thus, a tremendous amount of analysis must be invested in understanding, classifying and identifying the nature of the sources we need to connect, in order to answer some of the questions from end-users. It may be clear to the reader why multiple sources of data and connected information (knowledge graphs) may be essential.

The RASFF portal (next page) may be an example which can be adapted, in principle, to guide end-users to ask questions and organize the input in "buckets" or holders. In this approach, the input data may be amenable to classification or clustering algorithms, to help sort out the nature of the questions. If we allow "question collection" to an open format (write down top 10 questions) using an open dialog box, where anybody can ask anything, in any form, using syntax devoid of context, then extracting the key ideas from this unstructured mess (without standard keywords) may be frightfully exhaustive, if not impossible.

The RASFF approach could use keywords and harvested frequency of words or terms from this question-gathering exercise. Using tools like PCA (principal component analysis), it may be relatively easier to identify the topics covering 80% (Pareto principle) of the questions.

Using these topics, as a guide, we can begin to build / connect with contextual domains of data (for example, micro-climate from federated nano-satellite weather channels), information (example, price of bio-diesel) and knowledge (example, crowd-sourced experience of end-users, elsewhere). When synthesized, it may help us to respond, in near real-time, appropriately, to the end-user, delivering ***actionable information***, perhaps 80% of the time, with respect to desired level of relevance, precision, accuracy and value, to reach a certain quality of service (QoS).

Theoretical discussions about questions, data and platforms, using power-point filled with boxes, with arrows and artificial acronyms, is easy. Providing meaningful value to the end-user is not easy. We shall strive to combine data and logic informed approaches in the context of case-specific problem-based artificial reasoning tools (**ART**). The outcome and quality of service (QoS) remains to be determined. The aspiration is to approach DIDA'S after critically evaluating the successes and failures from ART implementations, in real world scenarios with actual clients. The journey to KIDS is still amorphous, as outlined in ***PEAS Platform for the Agro-Ecosystem***.



Challenges in Knowledge Extraction & Application: *KIDS is a journey, not a destination*

The assumption in ***"actionable information"*** is the strength of ***credible content*** which induces humans in the loop to perform a process or execute a specific action. The ***trust*** in this suggestion is dependent on the depth of the connectivity between system of systems and the ability of the tool (ART) to collect, synthesize and propose a meaningful outcome. Hence, the process of delivering ***value*** for the user in terms of "actionable information" is ***not*** an instant step. It may be best described in terms of bio-mimicry. For example, if you ask a 5-year old about "errand" planning (grocery store, library, co-op, laundry), the answer may be correct or incorrect because the 5-year old may not know the locations, what you need at each location, store hours and if the traffic on the road may change while you are between errands. If you ask the same question to a 15-year old, she uses Waze and store hours of operation to customize a Google map with a sequence/pattern you may wish to follow, based on data and information (data-informed decision as a service). The 15-year old has "learned" how to plan and manage time, fit the process to parameters of family's needs, and intuitively, understands semantics.

This PoC, SENSEE, ART, DIDA'S, KIDS, hence, are sign posts on the road ahead. We continue to learn, improve accuracy, precision and credibility, to increasingly gain the trust of the user. We continue to explore tools to address long-term "learning" and apply the results.

In the real world, tools often lead to questions about standards because a tool is not an one-off product. Standardization is viewed as an unifying process (for example, IPv6), which enables creation of tools agnostic of where (location) it is used or manufactured, as long as it is in compliance with standards. However, dynamic systems involved in decision making may be hard to standardize, primarily because of geo-political and socio-economic factors with respect to the decisions and the impact of those decisions.

Ecosystems are in a perpetual quest to develop and deploy advances in standards, which can be driven by adoption (for example, Android and Windows operating system). But, standards of decision making are far from homogeneity. The diffusion of any standard operation procedure in terms of decision making depends on the strength of measurement science (including data), interoperability of information between systems and software tools which can ***combine data, analytics and information with knowledge and experience*** (aspiration in ART, DIDA'S, KIDS).

This PoC "plan" is the ***beginning*** of analytics, which, theoretically embraces experience. ***PEAS Platform for the Agro-Ecosystem*** attempts to bring together the value of aggregating and querying sensor descriptions. Conventional wisdom suggests that sensor data and analytics, as "information" is more useful for end-users. The description of sensor types (PoC for SENSEE 1.0) is a step before jumping into sensor data.

It may help to remind the readers that this ***planning document*** is relevant to one idea from the suggestions in the series of essays (SITS, SIP-SAR, SARS♦AG and PEAS). How many types of sensors are available for any given task? What are the characteristics of the sensors in terms of signal detection, sensitivity and other key attributes? This may not interest the end-user in a farm, but it is critical to the ***design of field deployment*** and those in academia and industry.

This PoC creates a foundation with SENSEE 1.0 that does not contain sensor data but contains sensor descriptions (http://bit.ly/SENSOR-LIBRARY-ERIC-MCLAMORE). In its first version, we aim to collect sensor descriptions from about 1,000 labs using a partly automated document management system to populate the database (SENSEE 1.0). Using a simple mobile web service type app, (please request the IP address of the web service if you wish to test the usability of the app), experts may query SENSEE 1.0 to ask direct/specific questions (examples: at the beginning of this document, page 3) or use Boolean operators (for example, how many sensors use laser scribed graphene *and* plasmon resonance spectroscopy for signal transduction).

It is the aim of step II of this PoC to demonstrate [1] the ability of SENSEE 1.0 to contain a *critical mass* of sensor descriptions in the form of ***curated*** value fields and [2] the ability to answer a variety of questions using the simple app (user interface, dialog box) developed in step I. The ability to answer questions, group values using keywords and other combinations, depend upon the natural language processing (NLP) tools to be developed. Initially, the questions may be limited to those which may include keywords the rudimentary string parser is able to handle. This is a *learning* process which we anticipate will improve over time (if we invest) and presents opportunities to explore creative ideas (https://openreview.net/pdf?id=rJl-b3RcF7).

SENSEE 1.0 may provide an upload/ingestion tool for sensor labs to upload their sensor descriptions (xl file) and use interactive tools to clean/curate the fields if the ingested "data" is in the incorrect field (for example – incorrect column heading – limit of detection score inserted in the column with the heading "molecular weight in Daltons/kiloDaltons" of molecule/analyte). The utility of SENSEE 1.0 may be limited unless we have a critical mass of sensor categories and attributes in the database. Experts and students may find this tool to offer specific answers compared to PubMed or Google search using key terms. It is hard to imagine farmers, growers and meat packers, who may be interested in sensor characteristics, for example, detection of ammonia using microfluidics versus laser scribed graphene (LSG) sensors.

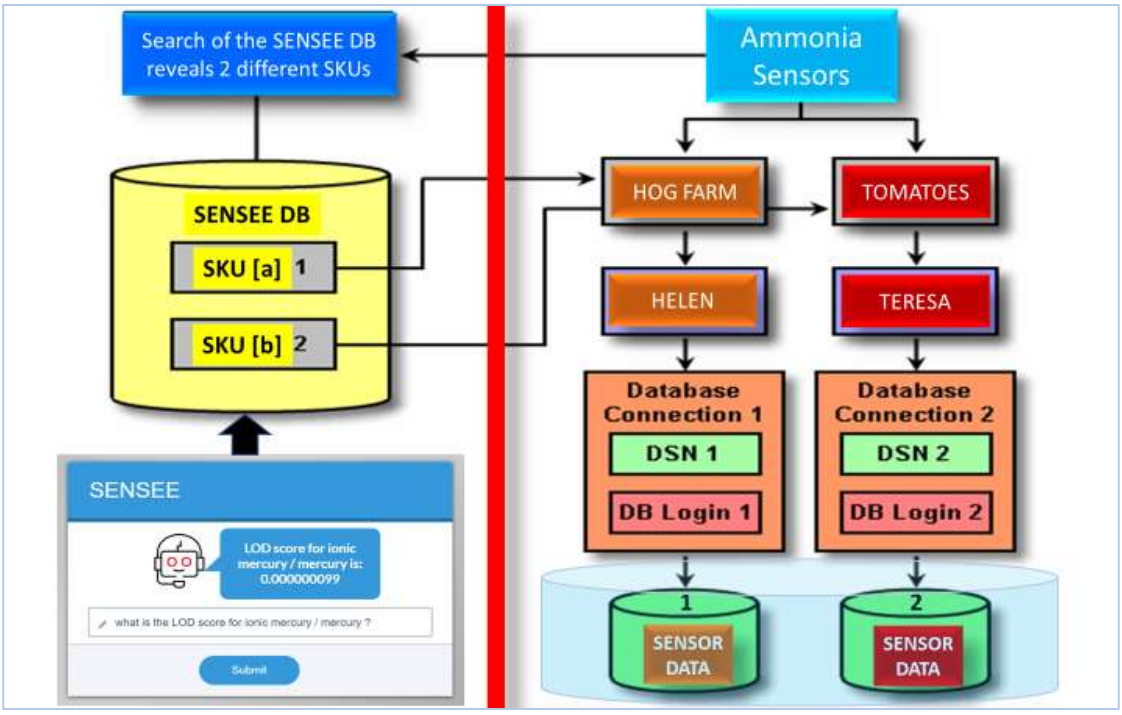
SENSEE 1.0 will continue to strengthen the search function (NLP), automating data ingestion tools, developing curation and data clean-up codes (step III). The research quest is in step IV, where the suggestion is to develop feature automation tools. ***Through all these steps we are still discussing sensor descriptions (types). We have not yet discussed sensor data.***

Each sensor, for example, ammonia sensor using [a] glass capillary or [b] LSG (as platforms) are independent sensors (***different SKUs*** or stock keeping units, different items). If sensor [a] or [b] is manufactured, then it can be used to detect/quantify ammonia gas. If sensor[a] is manufactured in a small batch (100 widgets) but sensor[b] is manufactured in a large batch (10,000 widgets), then SKU[a] may have the serial numbers 1-100 and SKU[b] may have the serial numbers 1-10,000 (hypothetical).

This digression is critical to grasp the distinction between SKU and serialization. It is important for the discussion about data from sensors (SENSEE 2.0) with respect to tracking and tracing (id) of sensors, when they are in the field (often tagged with RFID tags for identification purposes). Data acquisition from sensors in use (in the field) must be specific for sensors which are related to a specific case (real world client) which is a part of a problem that we are trying to solve for the client (test bed, customer). Sensor data is the topic for the next PoC, referred to as SENSEE 2.0 in ***PEAS Platform for the Agro-Ecosystem.***

SENSOR DATA – SENSEE 2.0

At the hand of hog farmer Helen, sensor SKU[a] serial numbers 1-5, is now capable of generating ***sensor data (measuring ammonia)***. At the hand of tomato grower Teresa, sensor SKU[b] serial numbers 10-25, is now capable of generating ***sensor data (measuring ammonia).***



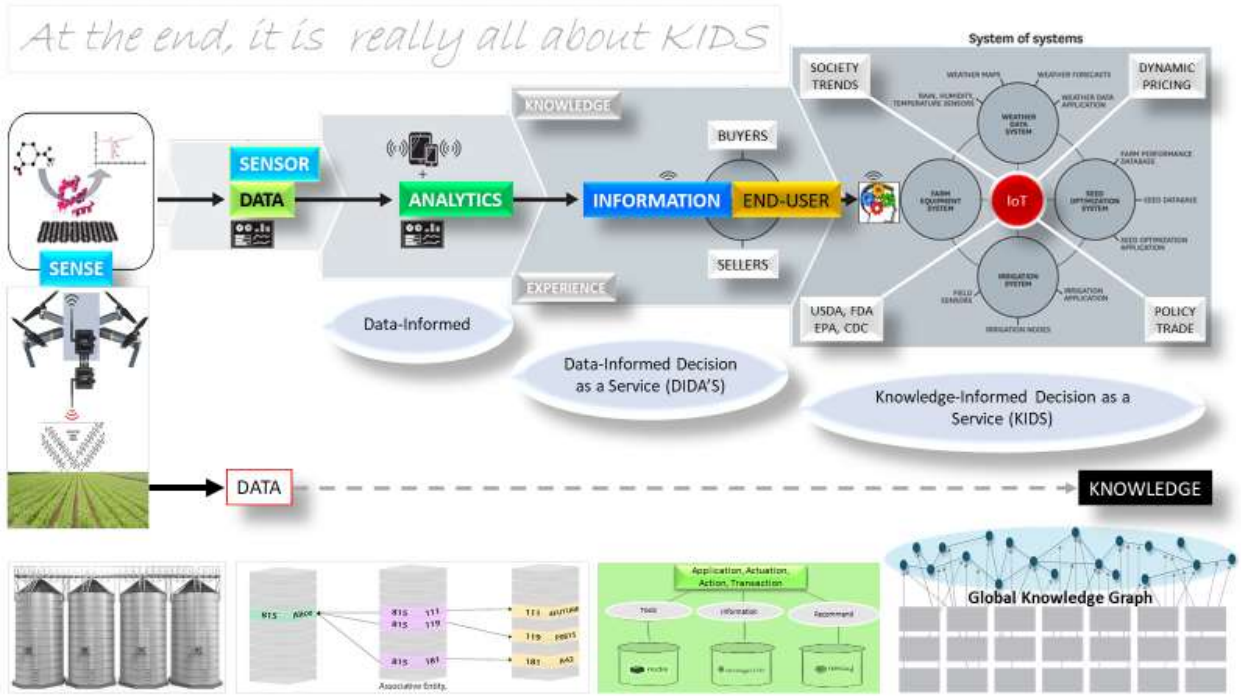
This PoC is about the LEFT side (SENSEE 1.0) where sensor descriptions are in SENSEE DB. Use of the sensors will generate sensor data – sensor data acquisition (right) is for SENSEE 2.0

This PoC does not address the sensor data database. That is the next step (SENSEE 2.0) where data is sensor-specific, case-specific, problem-based, has a purpose (end-user directed). It is used in conjunction with logic tools and data analytics in order to evolve as information expected out of the umbrella of artificial reasoning tools (ART). The latter is a step toward curating information and knowledge with respect to DIDA'S (data-informed decision as a service). ART may "crawl" for a while till it gains strength in its logic spine to walk, albeit slowly, to reach DIDA'S and hope for KIDS (knowledge-informed decision as a service).

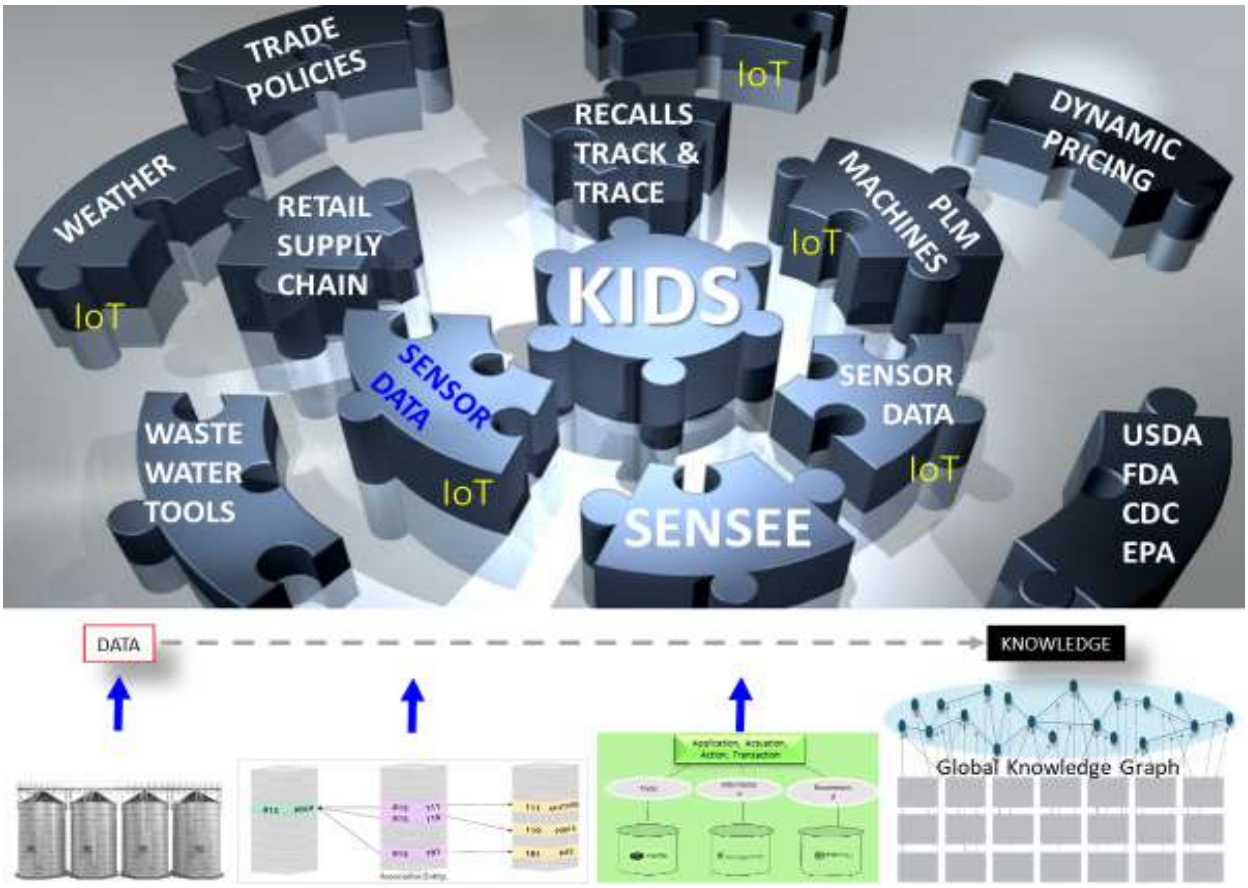
The range of questions in the next phase will include questions about sensor data. For example, the lab or manufacturer of ammonia sensors may query SENSEE 2.0, which sensor is more stable for higher concentrations of ammonia? Helen may ask when are the hogs producing the maximum volume of ammonia? Teresa may ask is there a difference between the ammonia concentrations during dawn versus dusk? **ALL** these questions are different but relying on the data from specific sensors (why data persistence is central to ***different views*** of data analysis and analytics). To provide value to end-users, ***SENSEE 2.0 and SENSEE 1.0 can be independent but connect when graphs "discover" them in the process of identifying resources for a case.***

DATA-INFORMED 2 KNOWLEDGE-INFORMED: KIDS ARE FAR OFF IN THE FUTURE

In the context of delivering real value to end-users, "actionable information" must move from **ART** (buzz word) to outcome. Future tools must be able to extract actionable, computable, contextual domain knowledge from informal sources of data (for example, text-based data). It is ***not an easy task*** in reality to augment user's ability to perform analyses using common models, dynamic composability of modular components, sensor-specific data, environmental data (big data is a bad word) and near real-time data from the edge. In order to capture "experience" even ontological frameworks may be useless. Even if domain experts can capture "experience" in a text-based format, it is doubtful if such text may "conform" to ontological frameworks. The latter is generally useful for text-based data (https://schema.org/).



This cartoon summarizes the journey from data to knowledge (didn't dare to include experience). In our phase one approach (PoC SENSEE 1.0), we focus on sensor characteristics. In phase two (not within the scope of this PoC) sensor data (bottom, left, silos) fuels data analytics and relationships between data (SENSEE 2.0 in cartoon of relations/associations between databases). Information becomes valuable to users if ART can proceed to synthesize and generate the data-informed decision as a service (DIDA'S) model (bottom cartoon, in pale green). As we approach knowledge integration phase, connectivity of local data and information with global system of systems adds value for data-informed policy decisions, understanding local dynamics and pricing in the context of market economics and trade practices. Creating GKG (global knowledge graph, and in future, labeled property graphs, LPG) is an evolutionary task, as we continue to stitch resources that can inform and provide knowledge to end-users. These "end-users" are no longer only farmers, growers, academic or manufacturers, it could include global organizations (FAO, WTO, UNDP) as well as policy forums and politicians (for example, farm bill, world customs organization, public health, other institutions, such as, FDA, CDC, WHO, ADB, EBRD, WTO).



KIDS is an open plan platform concept. Platforms are comprised of multiple applications and integrated solutions with embedded tools and databases that function as complete, seamless environments. Product innovation platforms are intended to support users collaborating across domains and businesses in the ***PEAS Platform for the Agro-Ecosystem***. These capabilities are increasingly needed throughout the entire extended enterprise in almost every vertical, agnostic of the type of application or function or users, including farmers, meat packers, produce growers, retail stores, customers, suppliers, and business partners. Developing open platform tools and technologies are not limited to any one domain because these modular tools can be applied, used and re-configured for re-use, almost anywhere, for example: error correction, graph and search engine algorithms, natural language processing (NLP), automated feature engineering, drag and drop tools, analytics, workflows, and services, such as KIDS, where "open" means 'plug & play' user friendly human-computer interactions and interoperability between system of systems. The monetization of KIDS is key for entrepreneurial innovators and investors seeking ROI. Users will ask how KIDS relevant to ***my*** farm or ***my*** manufacturing operation (think emergence of digital twins, eg, to mimic shop floor for machine tools) or ***my*** healthcare. Dynamic composition of the tools in the KIDS tool kit will be essential for the high degree of differentiation that must be achieved on-demand and served in real-time. To serve individual users or groups, raw data acquisition must be specific for the user's domain and the analytical tools (algorithms) must be contextually relevant. Cybersecurity will demand that user data is sufficiently protected. KIDS will offer different views and instances to different users. Only software can deliver granular services at a feasible transaction cost. For global adoption of ART, DIDA'S, KIDS, and PEAS, embracing the *pay a penny per unit* (PAPPU) pricing model, may be a potent and vital catalyst.

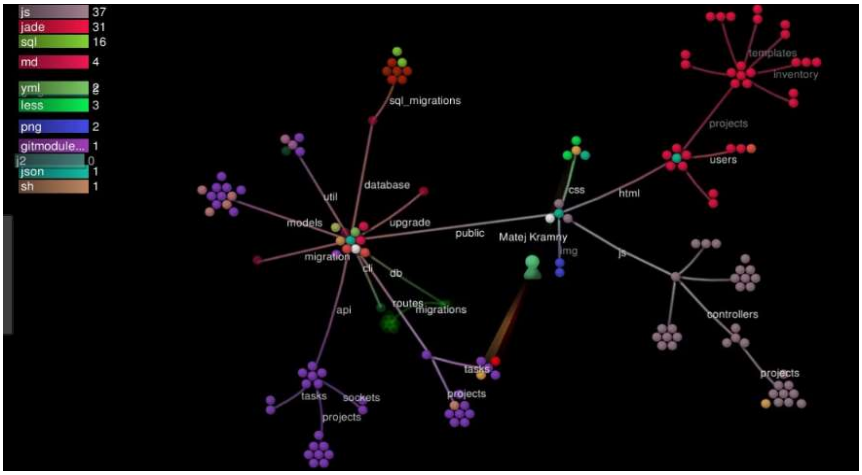
30 ◆ Exploring how sensor repositories may be helpful. See Figure 11 on page 28 – http://bit.ly/SIGNALS-SIGNALS

Human expertise is an **embedded** component of procedures, processes and decisions in any system. However, integration of human knowledge and human-computer interactions in operational analysis, are difficult to execute (why "smart" systems are still dumb). In certain systems, for example, the agro-ecosystem, human-generated knowledge is quintessential, yet it may be captured in text-based documents, which may be inconsistent and contain domain-specific vernacular. Even more challenging are the facts and observations that humans may grasp but *unable to articulate* or capture in writing or in notes. Thus, we have lost that intuitive factoid because of our inability to sufficiently capture what we are thinking.

The adage "hard to express in words" is a normal neurological state. It may be safe to conclude that there is *no intelligence in artificial intelligence* (https://arxiv.org/abs/1610.07862). AR (artificial reasoning) may be the best outcome from machine learning, no matter how "deep" one claims it to be. Hence, a name change from AI to AR is long overdue. The suggestion of **ART** in this document is more appropriate rather than perpetuating the lies and myth of AI.

Integrating ***knowledge and experience*** can better inform (DIDA'S) decisions, rather than relying only on ART. But, operational decisions at the point of use may find it difficult to synthesize the data (for example, streaming sensor data) with knowledge systems, in near real-time, to inform the user. The end-user in the field or farm or manufacturing shop floor, is more interested in the ***integrated information*** rather than data streams on a slick mobile dashboard.

Relationships between knowledge domains may boil down to ontologies. In other words, knowledge extraction must include design and development of taxonomies and metadata strategies for content management. The "fit" of these strategies to text-based data and other forms of unstructured sources of experiences remains to be explored with respect to existing (and/or evolving) vocabulary/taxonomy/knowledge organization system (KOS)/ontology software. The older thesaurus standards (ANSI/NISO Z39.19 or ISO 25964), newer ontology standards (OWL, RDF), and the SKOS (Simple Knowledge Organization System) model for "controlled vocabulary" may be relevant in this context (OWL/RDF is discussed later in this document). Most organizations may need more than one kind of controlled vocabulary. Hence, combined taxonomy, thesaurus, ontology and knowledge graph structures (discussed later) are emerging (for example, graph database-based Synaptica Graphite, cartoon shown below).



Most systems are starved of information, but we have an abundance of data, albeit uncurated data, often replete with noise and/or a poor ratio of signal to noise. Relationships are key to extracting experience and knowledge (previous cartoon) in order to inform and integrate data and analytics. Smartlogic Semaphore (https://github.com/ansible-semaphore/semaphore) provides a good view of these connections but in reality these are rarely "available" for rapid deployment (for example, in the agro-ecosystem). The cartoon is an example of what we ***think*** might be helpful for knowledge supported decision systems (KIDS). Other related software tools include PoolParty, TopBraid Enterprise Data Governance's Vocabulary Manager, Mondeca Intelligent Topic Manager and VocBench, to name a few specific suggestions, from an extensive list discussed in this book: http://www.hedden-information.com/accidental-taxonomist/.

Systems of the future must address this chasm between technical output (for which systems integrators want to charge money) versus the user value (the outcome for which the user is willing to pay). In any vertical, data plays a key role as a business driver. In the knowledge economy, the data analytics business will remain in the doldrums unless tools and technologies can deliver meaningful knowledge extraction mechanisms to support the context of applications.

Most corporations are eager to stop at **ART** rather than invest in DIDA'S and make sense of data in order to synthesize the ***knowledge support*** that end-users in the field *don't even know* that they are missing. This is a systemic problem, not limited to agro-ecosystem. The diabolical claims made by semantic web experts, machine learning and artificial intelligence marketing arms, makes one fearful to suggest that this problem of information extraction and knowledge-informed suggestions, needs, and may benefit from, ontological frameworks, semantics, ML tools and perhaps, artificial neural networks (ANN, CNN, RNN), at some later stage, in KIDS.

We need one or more networked platforms (mobile, edge/mist/fog/cloud, federated learning, distributed sub-domains, high fault tolerance, seamless interoperability between nodes, data distribution services, ontologies) where we "drag and drop" entities to combine, select, push-pull and dynamically hybridize, multiple tools, in order to create application-specific, domain expert-curated methodologies, for classifying, clustering and quantifying data, information and knowledge. Extraction from highly unstructured data calls for advanced Natural Language Processing algorithms (http://bit.ly/Interested-in-NLP) which must work with graph-theoretic methods (see figure 11 on page 28 of "SIGNALS" http://bit.ly/SIGNALS-SIGNALS) and ontology tools, for example, Synaptica Graphite, which offers directed-graph visualizer.

Optimization of value for the user is a continuous process of labelling and analyzing diverse sources of data, before the information perishes. If text-based documents are sources of knowledge and experience, then we must find new ways to harvest that contribution in our attempt to synthesize data (external data, crowd-sourced data), information and knowledge with experience, to aid the gradual transition from artificial reasoning tools (ART) to data-informed decision as a service (DIDA'S) to knowledge-informed decision as a service (KIDS). Continuous improvement will contribute to the domain-specific enrichment (ontologies due to pecan farmers vs tomato growers, agro-ecosystem vs manufacturing). Usability and functional preferences may lead to *de facto* standards and support "organic" growth of open-source toolkits to fine-tune knowledge extraction from distributed domains and improve the value of decision support, in the context of the journey from SENSEE to ART to DIDA'S to KIDS.

32 ◆ Exploring how sensor repositories may be helpful. See Figure 11 on page 28 – http://bit.ly/SIGNALS-SIGNALS

BEYOND KIDS ➔ EXTRACTION OF EXPERIENCE

Decision making is largely driven by human expertise. Automated decision-making works really well during sales presentations, using power-point. Human experts contain a wealth of tacit information that is intuitive, informally captured and explicitly under-utilized due to our inability to capture, catalog and re-use experience. For example, an experienced physician needs to see only the color of the sclera to "know" if the patient has contracted jaundice. Capturing the ontological framework of this knowledge and creating a computational equivalent of this type of "expertise" may be the Holy Grail for the future of advanced decision support. The forthcoming exodus of experienced physicians (due to Brexit?) will leave the NHS (British National Health Service) denuded of a repertoire of critical knowledge which may be irreplaceable. To the best of our knowledge, there are no mechanisms or tools in UK or anywhere else in the world, to capture such knowledge and re-use the experience to support new, or less experienced, employees or re-train other physicians who may have gaps of knowledge in the areas that were covered by the physicians who may be leaving UK.

Even if this tacit knowledge is implicitly or explicitly represented in text-based documents, these documents are not amenable to analysis using the existing tools of knowledge representation, for example, SKOS (Simple Knowledge Organization System) model for "controlled vocabulary". These documents are likely to contain jargon, abbreviations, and domain-specific cryptic remarks, which may be difficult, if not impossible, to analyze with any "controlled vocabulary" commercial solutions. Text-based documents often yield important contextual ***pattern*** of information that is based on recurring experiences, which data alone cannot provide (for example, pathology report of blood cell count or streaming sensor data). ***Contextual information*** is useful when these patterns occur again to inform and guide decision support.

The research question is whether we can create domain-specific methods, guidelines, and toolkits to study and analyze formal and informal, text-based documents to extract patterns and other supporting information to aid future operational decisions? It will ***not*** be a one shoe fits all "AI" solution, rather a portfolio of domain-directed methodologies for transforming documents into a computable format, to augment our ability to integrate their value in future analyses. A group of experts may jump to standardize methods to capture this knowledge. The latter may be an acceptable and even useful approach, but just one part of a multi-part dynamic solution set which may consist of overlapping open source toolkits, and domain specific guidelines, to map unstructured patterns to symptoms, indicators and other detectable parameters (for example, smell of acetone in breath indicates excess of ketone bodies in blood, likely due to diabetes). Another group of experts may clamor to establish ***best practice*** guidelines for analyzing text.

The arm-chair academic and the business strategy version of this discussion will evolve with every reiteration but how closely will it reflect the complex needs of the practitioners? If we were to compile a "to-do" list for the future and compare it with actual examples of questions and problems from end-users in the field, today, we may begin to observe the gaps between the technologies we think we are combining, to answer questions, versus the non-linearity of the real issues. There are no easy solutions except for constant human involvement in decisions. Mining ***user experience*** does not fit the boundaries intrinsic in a scientific tool or pre-set vocabularies.

Experience is not structured to "fit" a tool or tools we may develop (as discussed above). User experience will evolve and *how it is recorded* will evolve, too. Hence the tools we thought could be useful for mining experiences from yesterday may prove to be impotent tomorrow. For example, thesaurus management software (also used for taxonomies), such as Synaptica KMS, and other products, no longer exist. If we improve the tools, combine the functionalities and aggressively pursue concurrent evolution of tools, then, it may provide a second-rate approach to harvest user experiences.

The next task is to ***synthesize*** extracted knowledge with technical information and data, to augment the user experience, in near real-time, at the point of use. It is a ***very difficult*** task.

Mind the Gap – *between tools and technologies versus real-world issues and problems*

A Sense of the Future: To-Do List for Tools and Technologies *(adapted from www.nist.gov)*

■ Develop open source ontologies (schema.org) and tools for curating, cleaning, labeling, feature selection and feature engineering for text-based logs, databases, data, and information.
■ Develop/standardize/innovate NLP techniques for descriptive data from int/ext ecosystems.
■ Integrate extracted data with analytics, workflows, feedback loops.
■ Create tools for knowledge access and visualization of components. Show real-time data (eg streaming sensor data) combining with past patterns to inform decision support / suggestions.
■ Develop metrics for verifying and validating methods and calibration at the granular level of sensors and actuator. This data should aid in diagnostics and prognostics of the system.
■ Data from real world test beds must be accessible by local and global partners (see Figure 16 on page 33 in "SIGNALS" http://bit.ly/SIGNALS-SIGNALS) for distributed learning tools.
■ Focus on ***value delivered*** to the consumer (value is dynamic in context, *never* in equilibrium).

A Sense of the Future: Problems and Issues from Fields and Farms *(source: expert end-users)*

● Average rainfall in a drought prone village decreased from 1000 mm to 500 mm, over the last decade. Can we reduce the annual volatility of water availability for irrigation and drinking, for humans and cattle? Technologies at hand include [1] Improving Ground Water Level (IGWL) [2] Energy-Water nexus technology developed by Datamatrix Infotech and [3] tools of IoT. Key problems to address: sensing and modeling. Do we sufficiently understand the science involved in the physical processes? Without the grasp of the basic tenets, the case-specific outcomes, even if they are successful for the problem at hand, may not be generalized or reproducible, elsewhere.

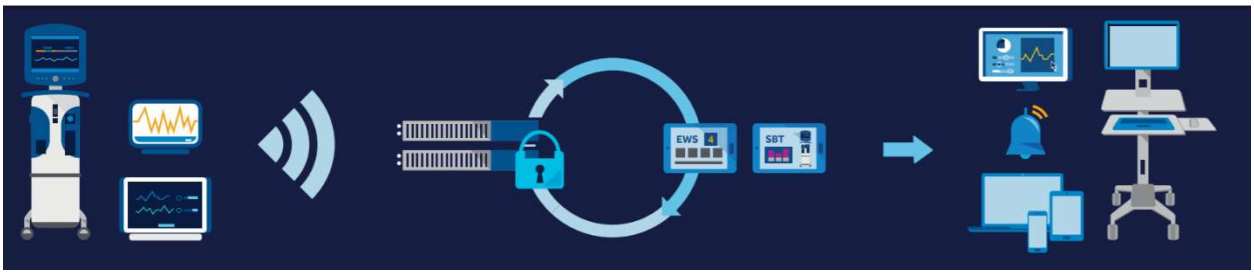
● Using an improved rice growing technique (http://tiny.cc/SAGUNA), 3000 farmers found that in addition to increase in yield, soil carbon has also increased, soil moisture is retained for a longer duration and number of earthworms have increased. Can we study and capture the interactions and dynamics between micro-climates, microbes, minerals, nutrients and signals from the soil?

● Flood forecasting and management model using 0.20m x 020m resolution LiDAR data to build 3D terrain model. Citizens and authorities may use this information to make decisions but micro-variations in rain intensity complicates forecasting. The ripple effect involves irrigation, water management, soil moisture and erosion – the sum of which affects crops and production of food. How do we converge metrics and measurements with knowledge and experience to focus on micro-environments in order to provide operational guidance to those who are in these zones?

Sense of the Future: Meaningful Support**:** *Knowledge Graphs, Knowledge Supported Decisions*

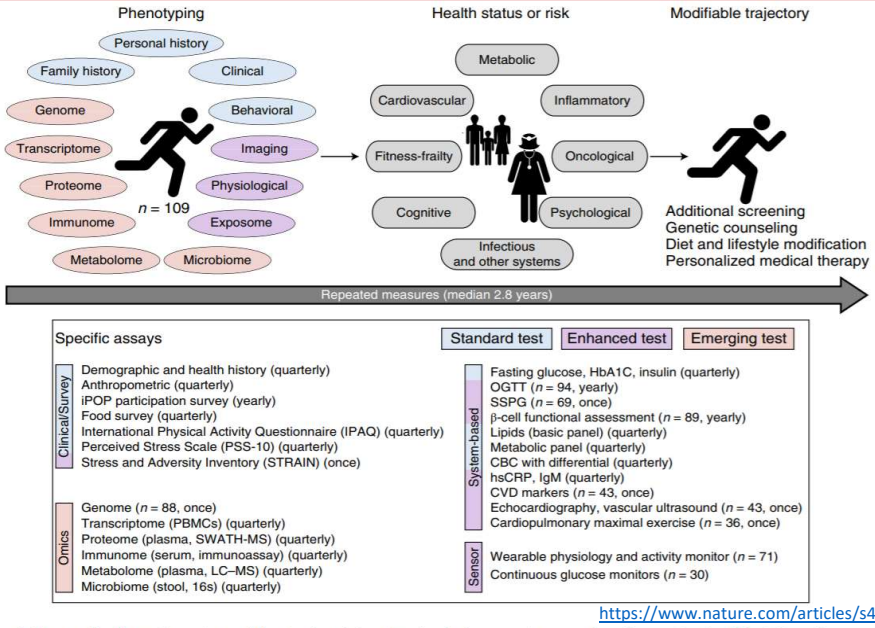
Data fusion may not be information. Data-informed processes are more than **ART**. Evolution from information to knowledge may be a far more ***difficult*** process because the "reason" why the information may become knowledge must be represented. as a part of the general problem-solving logic. This concept from the 1950's is at the heart of AI and integrating "reasoning" remains an unsolved problem. AI was originally referred to as artificial reasoning and shares certain principles with cybernetics (http://bitsavers.informatik.uni-stuttgart.de/pdf/rand/ipl/P-1584_Report_On_A_General_Problem-Solving_Program_Feb59.pdf). Hence, AI to AR.

The current form of data analytics and logic tools may offer a layer of ART before a true data-informed decision making (DIDA'S) system may claim success. The educated customer may benefit from DIDA'S but the less informed clients may find ART adequate for specific cases of "low hanging fruits" which may require lower level of skills, available in ART. Our elusive quest for ***knowledge-supported*** decision making is aspirational. It may ***evolve from DIDA'S to knowledge-informed decision as a service (KIDS), if there is customer demand.***



Application of sensor data and data from devices are not limited to any one vertical. It is a systems approach which transforms data to information for users at the edge. The cartoon (above) is an application of ART to KIDS in health/healthcare where real-time information can save lives and money. The ability to connect various troves of data are critical to the point of use, which indicates a future dominated by knowledge graphs to generate the knowledge. It is not unique to healthcare. It is applicable to agro-ecosystem, energy, finance. For example, Goldman Sachs is creating social graphs which integrates email (who emailed whom), telecommunication (who called whom), trading (who traded what) and linked financial data (who sent money to whom). It has 100 million edges and 2 billion nodes (http://bit.ly/GKG-KIDS).

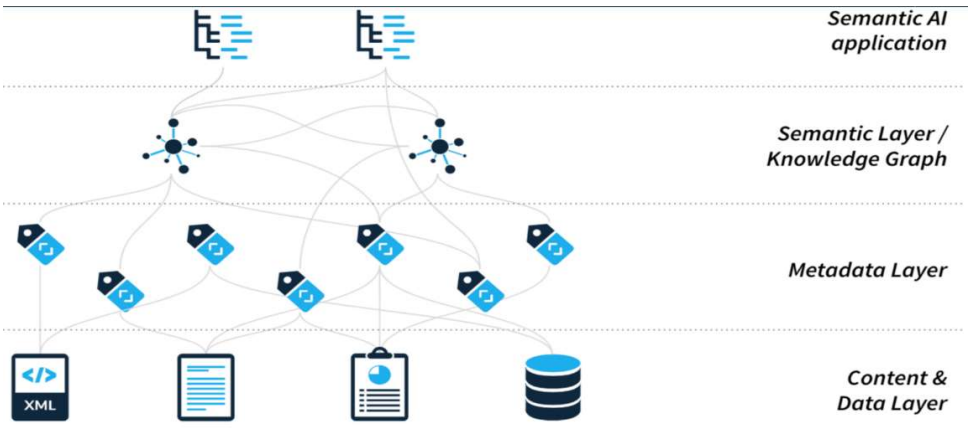
35 ◆ Exploring how sensor repositories may be helpful. See Figure 11 on page 28 – http://bit.ly/SIGNALS-SIGNALS



https://www.nature.com/articles/s41591-019-0414-6

Fig. 1 | Study design and data collection. Overview of the in-depth longitudinal phenotyping used to determine health risk and status. Data types were categorized as: standard (blue), enhanced (purple) and emerging (red) tests. PBMCs, peripheral blood mononuclear cells; HbA1C, glycated hemoglobin; OGTT, oral glucose tolerance test; SSPG, steady-state plasma glucose; CBC, complete blood count; hsCRP, high sensitivity C-reactive protein; CVD,

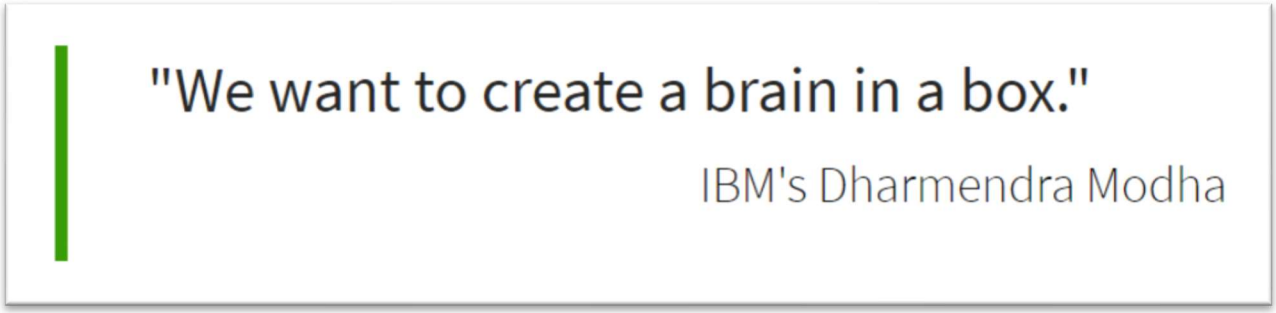
KIDS in Health and Healthcare: From molecular profiling to behavioral changes, which may improve a healthy lifestyle, is a process which involves, for each individual, multi-year deep longitudinal studies, before actionable health discoveries may provide relevant data. If this data is sufficiently analyzed and synthesized, it may provide some information for precision health, of that one person (individual, patient). The challenge in scaling this process is the availability of qualified human resources to deconstruct and reconstruct the data in the process of extracting actionable information. Knowledge about the patient must be stored for use in the future. Barring the hype, this is a potential area for use of machine learning (ML) algorithms that can select the data from distributed databases and using knowledge networks (knowledge graph algorithms) find and weigh the relevant connections and correlations. By selecting weights as an index or metrics, ML algorithm engines may be trained to issue a set of recommendations and indicate the probability of confidence associated with each (suggestion, diagnosis, prognosis). This approach is universally applicable to any domain (data, knowledge graph, decision support) including non-human entities (digital twins). The implementation of knowledge-informed decision as a service (KIDS) starts with data. Bulk of the data from humans and machines originates from sensors.



KIDS WITH ABS ➔ CONNECTING IN CONTEXT: DOTS, DATA AND INFORMATION

Connectivity, in context, may be a basic instinct for all life forms. This generic statement appears less trivial if we consider that plants are designed to seek out sunshine and that ability, is in part, due to fractal patterns in the organization of leaves and the phototropic plant[1] hormone auxin, which induces photomorphogenesis. For animals, the quest for food[2] and flight from predators, are examples of connectivity, in context.

Almost all decisions, in humans, connects various data, information and knowledge stores, in our brain. The pathways remain poorly understood, despite an avalanche of foolish claims and blasphemous stupidity[3] of pompous statements from individuals dyed with hubris.



An elementary form of bio-mimicry of decision systems may be at the heart of KIDS. We must be able to connect, in context, data from different domains, selected to suit the user's query, to synthesize information and knowledge, which will offer value to the user, if delivered, in time.

The connectivity we aspire to extract from global knowledge graphs (GKG) for use in ART, DIDA'S, KIDS, may find analogies from the annals of telecommunications. Networking is the bread and butter of connectivity for the telecom industry. Since GSM was introduced in 1991, the industry has pushed incessantly for increasing bandwidth, speed (data rate) and higher power for fairly expensive devices (iPhone). Circa 2015, telecoms were forced to accommodate, adapt and re-invent its practices, with the diffusion of IoT. Connectivity between vast number of devices, sending data pulses (sensor) or short bursts of data on-demand (user query) may survive on low bandwidth, low data rates, low power for IoT-type connectivity between devices, many of which are low cost devices. In other words, the opposite of the conventional wisdom espoused by the practitioners in the telecom industry of the 1990's.

IoT drove a fork in the telecom industry and non-traditional players invested in low power wide area networks (LPWAN). To counter LPWAN penetration as the key IoT backbone, frantic traditionalists created a partnership (3GPP) and agreed, in haste, on the NB-IoT standard, a mix of NB-LTE and cellular IoT (2016). Thus, emerged agile hybrid networks of traditional cellular, non-cellular, non-traditional mesh and other protocols that can take-over or hand-off any signal (eg WiFi, Bluetooth, WiMax), anywhere, anytime, from any device or any object.

[1] https://www.untamedscience.com/biology/plants/plant-growth-hormones/
[2] http://library.mit.edu/item/002405318
[3] https://www.foxbusiness.com/features/after-watson-ibm-looks-to-build-brain-in-a-box

The agility, with which traditional telecom players could surmount the barriers (due to dead weight of old technology) and embrace new tools, is a lesson for the decision sciences. For the latter, churning "data-driven" into "data-informed" still falls short of customer demand. The knowledge-informed decision is far more valuable. Transition from data-informed to knowledge-informed calls for incisive changes in synthesis of data and information, for KIDS. The caveat in taking this analogy from the telecom industry, too far, is the vastly convoluted pace of creating "standards" in the data and information domain. The operational failure of the semantic web and sluggish progress in creating ontologies are indicative of the challenges facing KIDS.

An example from the telecom industry which may resonate with proponents of KIDS is the case of a "connected" car or future of semi-autonomous or autonomous vehicles. The car needs instructions in real-time with near-zero latency. It must receive software upgrades, bug fixes, send reports from sensors, enable instructions to modulate actuators and maintain constant dialog with control centers using cloud, fog or mist computing. The user and the connected car ***doesn't care*** if data arbitrage is being conducted over a fixed connection, WiFi, WiMax, DSRC (dedicated short-range communications[4]), C-V2X (cellular vehicle to x), SDN (software defined networking[5]) or NFV (network functions virtualization[6]). The vehicle needs ***connecting***, with a certain quality of service (QoS). End-users may not care ***how*** the connectivity is implemented, as long as the ***network-agnostic networks*** can work seamlessly, in harmony, using whatever media is available (copper, fiber, wireless, LTE) to deliver the contracted quality of service, every time.

For users seeking assistance from KIDS, the user does not care which data domains the knowledge graph[7] must connect and whether it is a RDF graph or a labeled property graph (LPG). The ***quality of the outcome*** is the relevant determinant of value from ART, KIDS, for the end-user. The dynamic pricing index (service fee) for KIDS may be linked to the QoS metric-on-delivery. The domains of data, data analytics and information databases, are static "nodes" or resources from the perspective of the end-user. The query from the end-user is the ***trigger*** to instantiate an ***user-centric*** selection of the nodes. In other words, ***the query from the user will drive the connectivity*** between data swamps, analytics and information nodes, necessary to answer ***that specific question*** from the user. The analogy in the networking world is referred to as application driven networking (ADN) or application centric infrastructure (ACI) and are variations of the concept commonly referred to as service-oriented architecture (SOA). The ***service*** call ***shapes the events*** which will ***follow*** in order to respond to the specific request.

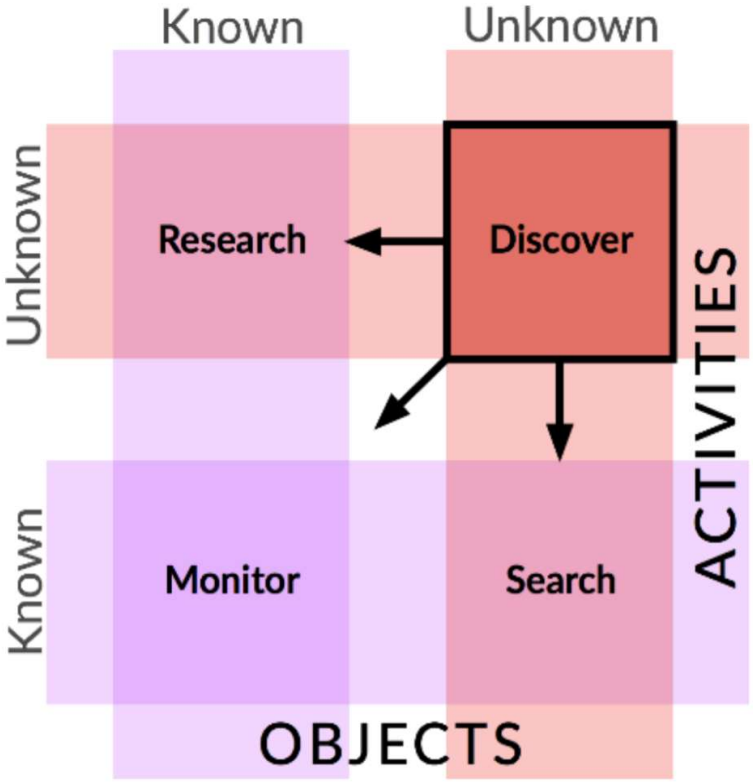
User-centric dynamic composability to create *ad hoc* knowledge graphs will benefit from integrating Agents[8] in the design of KIDS. One role of the Agent will be to parse the query and determine which nodes must be connected for KIDS to attempt to answer the query with a decent QoS metric. To provide a trivial analogy, imagine a busy intersection in Mumbai or Mombasa. The traffic lights aren't working due to brown out and motorists are confused by the DETOUR sign. A traffic policewoman is at the round-about, motorists are driving to the circle and policewoman directs the driver, depending on the driver's question. In KIDS, a software Agent, in the role of an "analyst" may execute the function of the policewoman, when it detects a query from an user.

[4] https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety
[5] http://publications.uni.lu/bitstream/10993/20596/1/survey_on_SDN.pdf
[6] http://www.ttcenter.ir/ArticleFiles/ENARTICLE/3431.pdf
[7] https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html and https://youtu.be/mmQl6VGvX-c
[8] http://ermolayev.com/psi-public/SOTA-TR-PSI-2-2004.pdf

What happens when the query falls in the "unknown unknown" category[9] shown in the cartoon? Although this scenario is more prevalent in the cybersecurity domain, the principle of the solution is similar to the discussion, at hand. The Agent must work as an *analyst* in handling the queries to "discover" the "unknown unknown" concepts in the user's query.

Discovery is a critical part of the knowledge graph future because the query-driven process must have a mechanism to find out ***what*** to connect, to create the knowledge graph. Data domains ***relevant*** and ***relative*** to the ***context*** of the query must be identified and ***connected***. The principle of **R2C2** (relevance, relation, context, connect) may be key to connect correct nodes of the knowledge graph. The graph, thus constructed, and the graph network which will ensue, must be capable of extracting the relationships, correlations or convergence, the queries are seeking. This graph (linked RDF triples, relationships) may be stored in a knowledge graph database and the abstraction may be recycled or the actual instance may be re-used.

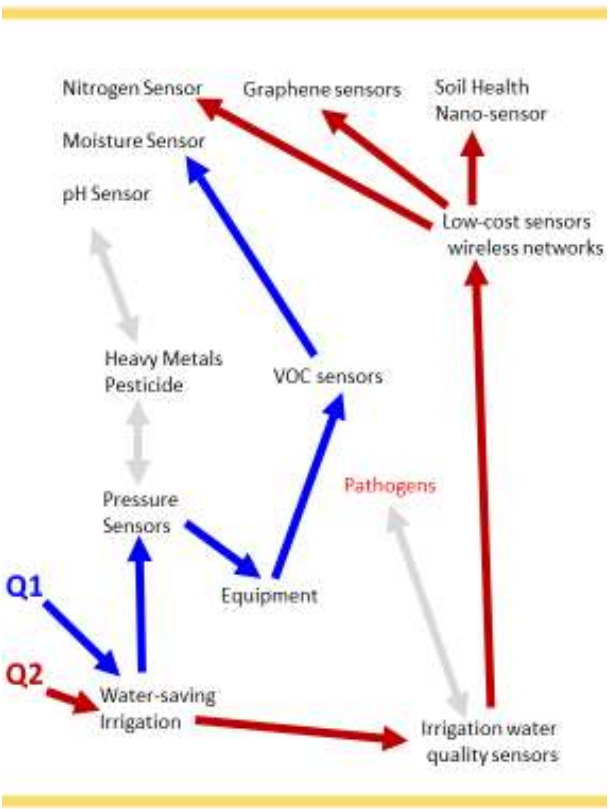
It will be remiss not to mention that one of the most egregious errors in the IoT hype is the idea that billions and trillions of devices and objects will be connected due to IoT. Even if there are trillions of things, the ability to connect is dependent on the ability of one object, with the correct tools to connect, to know and to ***discover***, that there is another object, within its reach, which is safe, compatible and configured to access, and connect. Although the central role of knowledge access[10] and ***discovery*** is an established principle, it is seldom emphasized for IoT.



[9] https://agile-defense.com/wp-content/uploads/DarkLight-Game-Changing-AI-for-Cyber-Security-Brochure.pdf
[10] http://oxygen.csail.mit.edu/KnowledgeAccess.html

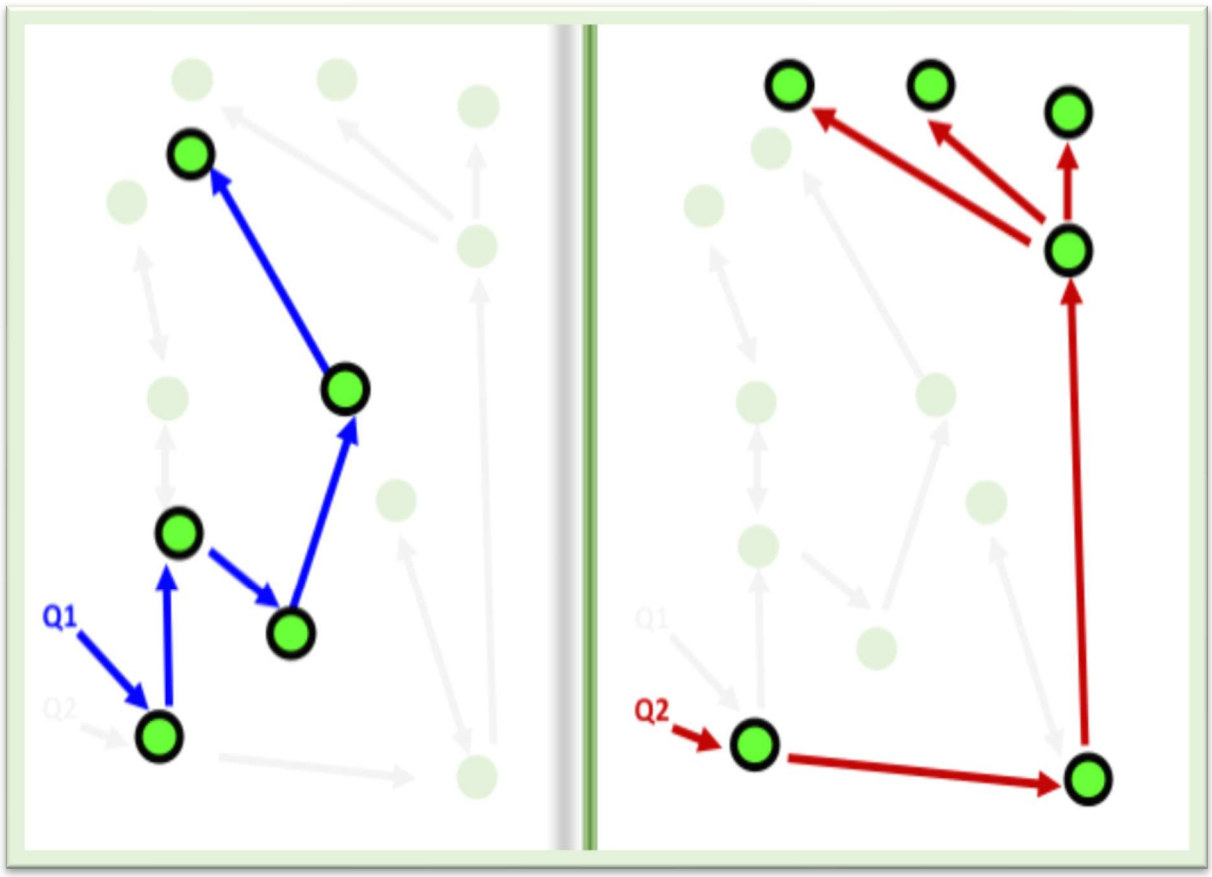
User-centric dynamic composability may also find some parallel concepts in network functions virtualization (NFV), a tool from the networking domain. The key idea of NFV is to ***replace*** dedicated network appliances, that is, ***hardware***, such as routers and firewalls, with ***software*** running on commercial off-the-shelf (COTS) servers. The aim of NFV is to transform the way communication service providers (CSP) architect networks and deliver network services. In the KIDS paradigm, the CSP equivalent may be domains of data, analytics and information stores or *service providers*. In the KIDS model, examples of *service providers* (in terms of data, analytics and information) may be weather data, provided as a publish/subscribe tool from the Weather Channel, prices of commodities (food) from Bloomberg Business, pollution data in terms of particulate matter (PM2.5) in the air from feeds maintained by the EPA. In NFV, network function software is *dynamically instantiated* in various locations in the network *as needed*, without requiring the installation of new equipment. The parallel for KIDS may be the Agent which catalyzes the *dynamic composability of domains, on demand* (triggered by user's query). This *ad hoc* composition is necessary to respond to the query, in the context of the query. The Agent does not install new components. The Agent simply selects the domains which may contain contextual and related resources, which are salient create the graph, to answer the query.

Development of embedded analysts, Agents-based ***selection*** (ABS) in the design of software architecture[11] is an old concept, which is a grand idea still hiding under a bushel. KIDS with embedded ABS may be necessary to navigate available resources and stitch the correct sequence of domains to synthesize knowledge-informed decision as a service, on-demand.



KIDS cartoon illustrates various domains of data and information available to a system. Two incoming queries, both on irrigation, are asking quite different questions. Q1 appears to be interested in saving water. Knowledge graph Q1 connects the nodes which provides information about soil moisture, which may optimize water distribution by the irrigation system. Q2 is also interested in saving water but not before understanding the quality of the water and condition of the soil. Graphs for Q2 connect different domains. KIDS with embedded ABS may be the tool (hypothetical suggestion) necessary to ***understand*** the query (syntax, semantics, ontological schema, unstructured vernacular) and then direct the path the knowledge graph must choose to respond to Q1 and Q2 with sufficiently high QoS metric (value). The latter will allow KIDS to charge users a small fee (exceeding a contractual QoS metric triggers the pay a penny per unit scheme).

[11] https://dl.acm.org/citation.cfm?id=122367



KIDS cartoon illustrates the connectivity between different domains by overlaying a knowledge graph on top of available resources (resource agnostic) in a system, for example, agro-ecosystem. The abstraction demonstrates very different (number of nodes, edges) knowledge graphs, due to queries from end-users (Q1 and Q2 are both related to irrigation water, in this fictional scenario). Dynamic composition of these *ad hoc* knowledge graphs are query-driven, user case specific. A distant analogy from the telecom domain is the creation of virtual private networks (VPN) by building a virtual network overlay on top of multiprotocol label switching (MPLS) network components. The "brains of the network" are managed[12] by software defined networking (SDN) controller platforms, which contains a collection of "pluggable" modules to perform different network tasks. For KIDS, an equivalent "brains" platform may contain modular ABS (Agent-based selection) analysts, to direct the formation of knowledge graphs, by extracting ***connectivity*** structures (paths or graphs between entities, see colored circles in the cartoon) relevant to the ***context*** of the query. By training these tools (eg graph neural networks[13]) to parse the questions, various clusters of ***meta structures*** may be created to facilitate ***knowledge discovery*** tasks to locate[14] ***where*** is the data or information, related and relevant to the query. These algorithms[15] may add value to the embedded ABS analysts and in turn enhance the performance of KIDS.

[12] https://www.sdxcentral.com/networking/sdn/definitions/sdn-controllers/
[13] https://repository.hkbu.edu.hk/cgi/viewcontent.cgi?article=1000&context=vprd_ja
[14] https://blog.cdemi.io/beginners-guide-to-understanding-bgp/
[15] https://iswc2017.semanticweb.org/wp-content/uploads/papers/MainProceedings/272.pdf

Extracting useful knowledge using the graph theoretic approach must be anchored to deliver ***contextual meaning*** of data. Hence, to connect nodes using the R2C2 principle, one must take into consideration the semantic profile of what is being connected. The data intensity of system of systems may be comparable to data intensive science projects, for example, the LHC (Large Hadron Collider) and ASKAP (Australian Square Kilometre Array Pathfinder), which generate petabytes of data, each year. If such vast volumes of data are stored in "data swamps" then we may lose its value unless curating contextual data from data swamps meets a miracle.

To make data relevant and meaningful for end-users, the applications must [1] select and coordinate data and information, [2] provide synergistic data integration between data domains (data from specific sensors or equipment, crowd sourced data), [3] enable visualization (plots, suggestions, recommendations) and/or [4] take action (actuate, execute). All of this is expected to happen in a seamless manner, in near real-time, without the need for users to understand any of the underlying representations and structure of the data.

In this vein, semantic[16] tools provide categorization capabilities and may facilitate machine-encoded definitions of vocabularies (which could be different based on vernacular), concepts and terms. In addition, semantics may explain the interrelationships among them (different vocabularies residing in different documents or repositories). The challenge is (and may always will be) balancing expressivity (of semantic representation) with the complexity of defining terms (used by experts, scientists, engineers) and implementing an end-user-friendly resulting system. This balance is *application-dependent*, for example, in terms of ease of use between tomato growers and nurses. The degree to which the implementation must be user-friendly depends on the intrinsic technical competency of the users. A single solution may not fit all, even *within* groups, for example, pediatric nurse practitioners vs geriatric care nurses.

The success of semantics in this respect will be governed by a very different form of human relationship. **Leadership** in this collaborative approach will determine how the fields may progress in the future (for example, agro-ecosystem vs automotive vs health vs finance). Success of semantic structures necessary for data driven software processes will depend on peer relationships, where [a] domain experts or scientists in specific fields will form co-dependent liaisons with [b] computer scientists, as well as software architects/engineers and [c] data providers, data system administrators and so-called data scientists. Fields which are traditionally "farther" away from computer science and software engineering, for example, agriculture, chemistry, economics, must strive harder to bridge this chasm by co-locating computer science departments with agriculture, chemistry and economics, perhaps in the same building or quadrangle or emulate instances[17] where agriculture is a part of a media laboratory. Without global and cultural cross-fertilization, ontological schemas and semantic catalogs of the future may be anemic, half-baked, sloppy and second grade (yet, masquerading as good enough).

Semantic tools may be a part of data-driven, evidence-driven, reasoning solutions, logic tools (ART). Statistical and mathematical modeling-based ML may be part of logic tools, too. These are different but complementary. The distinctions may get fuzzy when we combine high volume data, analytics, distributed information and several knowledge domains, as in KIDS.

[16] https://pdfs.semanticscholar.org/9256/c883b1ecaea08abc46179e2927302523a66d.pdf
[17] https://www.media.mit.edu/groups/open-agriculture-openag/overview/

- Deductive reasoning, syllogism & categorisation
  (Aristotele, 384 BC – 322 BC)
- Formal logic & calculus rationicator (reasoning, symbol
  (G.W. Leibniz 1646 - 1716)
- „Begriffschrift", technically: predicate logic
  (Gottlob Frege, 1848 – 1925)
- Frames for representing stereotyped situations
  (Marvin Minsky, 1974)
- Rules & expert systems
- Ontologies
  (Leibniz, Kant, Gruber 1994)
- Description Logics
  (Baader & Hollunder, 1991 et al.)
- Semantic Web
  (Berners-Lee, Hendler, Lassila, 2001)
  & Linked Data
  & Knowledge Graphs

The promise of semantic tools and technology may be rooted in its ability to capture the semantics of the data with the data itself. It can capture meta-description of different kind of objects, attributes, associations, and activity into a conceptual model, which can be populated with instances of actual data. Described using OWL/RDF[18] *syntax*, the conceptual model "ontology" represents the data itself in a single, consistent manner that is independent of how it is physically stored. With exceptions, ontologies formally describe taxonomies and classification networks, defining the structure of knowledge for various domains: nouns representing classes of objects and verbs representing ***relationships*** between the objects. Ontologies can represent information coming from heterogeneous data sources, hence, it can deal with structured, semi-structured, and unstructured data. The latter is particularly valuable for diverse end user groups.

When data is mapped against an OWL/RDF ontology, instances of the data are expressed based upon the idea of making statements about resources in the form of **subject–predicate–object** expressions. These expressions, also referred to as S-V-O (subject, verb, and object) are known as *triples* in RDF terminology. The 'Subject' denotes the object, and the predicate (verb) denotes a single semantic trait or aspect of the object that can be a literal value or expressed as a relationship between the subject and another object that is the target of the relationship.

[18] https://www.w3.org/OWL/

For example, "soil pH 8" in RDF triple is **subject** denoting "soil" and **predicate** denoting "pH" and an **object** denoting "8" which is the OWL/RDF take on using the object as the subject from the entity–attribute–value model within object-oriented design: entity (soil), attribute (pH) and value (8). The object (soil) can have another attribute (contains) that points to another object (phosphate). The object (phosphate) might have an attribute (produces) another object (acidity). Yet again, the object (soil) might have an attribute (contains) another object (microbes).

This is why RDF triples, despite their shortcomings, enables the formation to link a series of relationships between two or more objects. A graph, in this context, is a linked set of RDF triples. OWL/RDF-based data model may suit certain kinds of knowledge representation better than relational models because it can fuse data from multiple relationship tables about the same object. It is the foundation on which directed graphs are built. A collection of RDF statements intrinsically represents a directed multigraph.

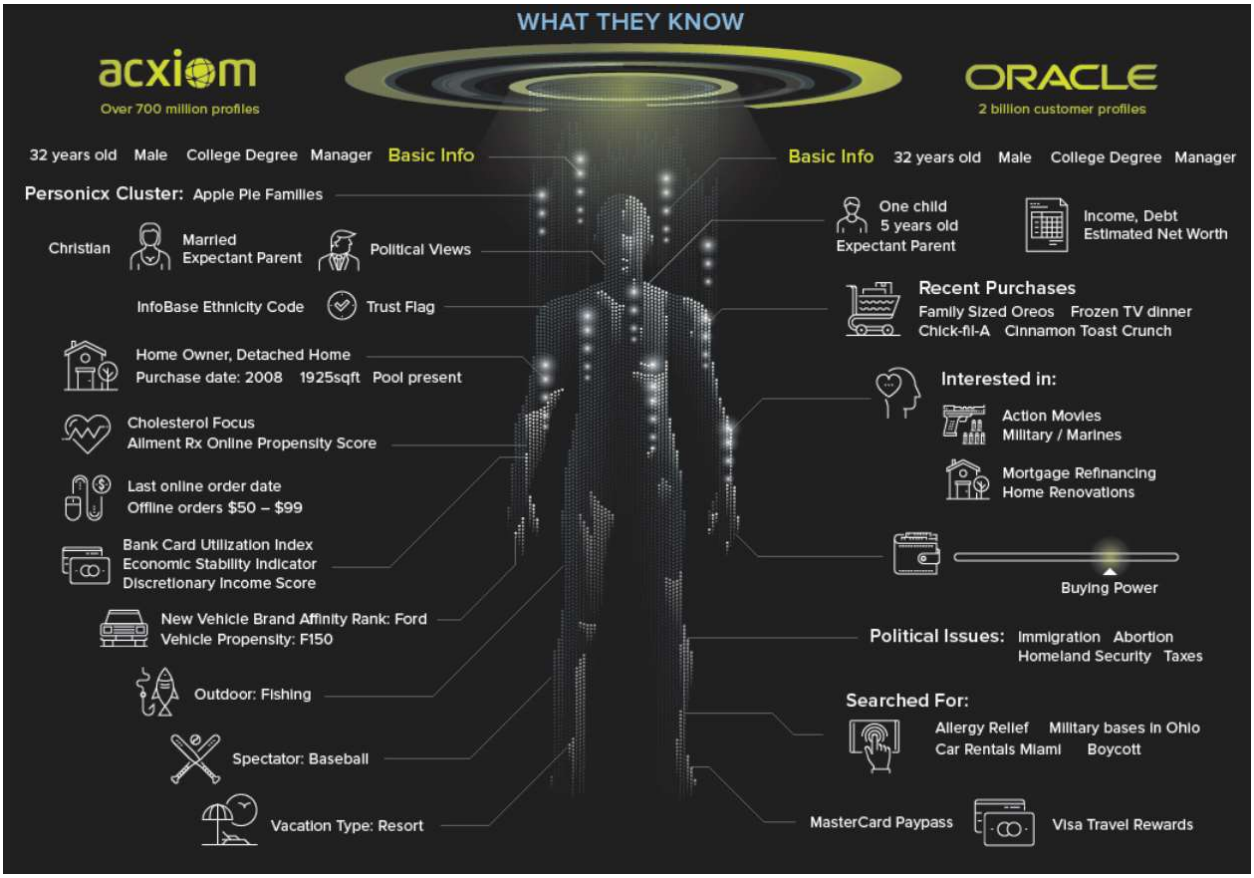
A **knowledge graph** (mentioned often in this document) is a knowledge base that is made machine readable with the help of logically consistent, linked graphs that, taken together, constitute an interrelated group of facts. RDF triple represents human knowledge in standard, machine readable form by linking a subject, predicate/verb and object. RDF representation can be visually displayed as the nodes (subjects and objects) and edges (verbs/predicates) of graphs.

The idea of ***artificial reasoning in logic tools*** relies on machine readable statements of facts. The expectation is that "triples" linked in a logically consistent way, via knowledge graph, will possess ***reasoning*** ability. When logically consistent factual triples are added to the graph, ***machines can infer new links or connections***. Hence, new connections can be discovered by humans due to the ***reasoning power of machines*** (**artificial reasoning tools**). Machines can then gain ***access to the relevant data*** in the ***context*** of these linked triples (knowledge graphs) as part of an ***information service*** (discussed earlier in the section KIDS WITH ABS) provided by **ART.**

For example, **ART** (artificial reasoning tools) may uncover the ***relationship*** between water, pH and Pb. The ***machine reasons*** that if pH of water in distribution pipes is less than pH 7 then the probability increases for metals, such as, lead (Pb), to leach out of the material (alloy) of the pipes (acid leaching) and increase the concentration of Pb ions in domestic drinking water supply (Pb is a neurotoxin). Having uncovered the relationship, the function of the knowledge graph (in ART) is to contribute to a solution, preferably quantitative. ART must ***discover*** and ***locate data*** to synthesize the solution and display the outcome on the end-user's mobile device. ART must discover data for each parameter, analyze and aggregate to create logical fusion. This demands data interoperability and choice of open APIs between systems, for example, the water distribution ***map*** (GIS), water quality ***in*** the distribution system (county public works database), chemistry knowledge for rate of leaching of Pb vs water pH (extract and merge the standard data with the actual pH of water, in this case). The ***predictive analytics tool*** may wish to forecast the (cumulative) increase in the concentration of Pb, with each passing day of inactivity. ART must display the useful version of this outcome and recommend mitigation strategy to reduce the morbidity of neurotoxicity due to Pb ions leaching from pipes into drinking water. This is the expectation from combined SENSEE 1.0 and 2.0 project, in terms of solutions for real world problems. We start with ***logic tools*** to deliver **ART** rather than "boil the ocean" with ML and other tools, which takes years, to reach the "knowledge-informed" quality of service (QoS).

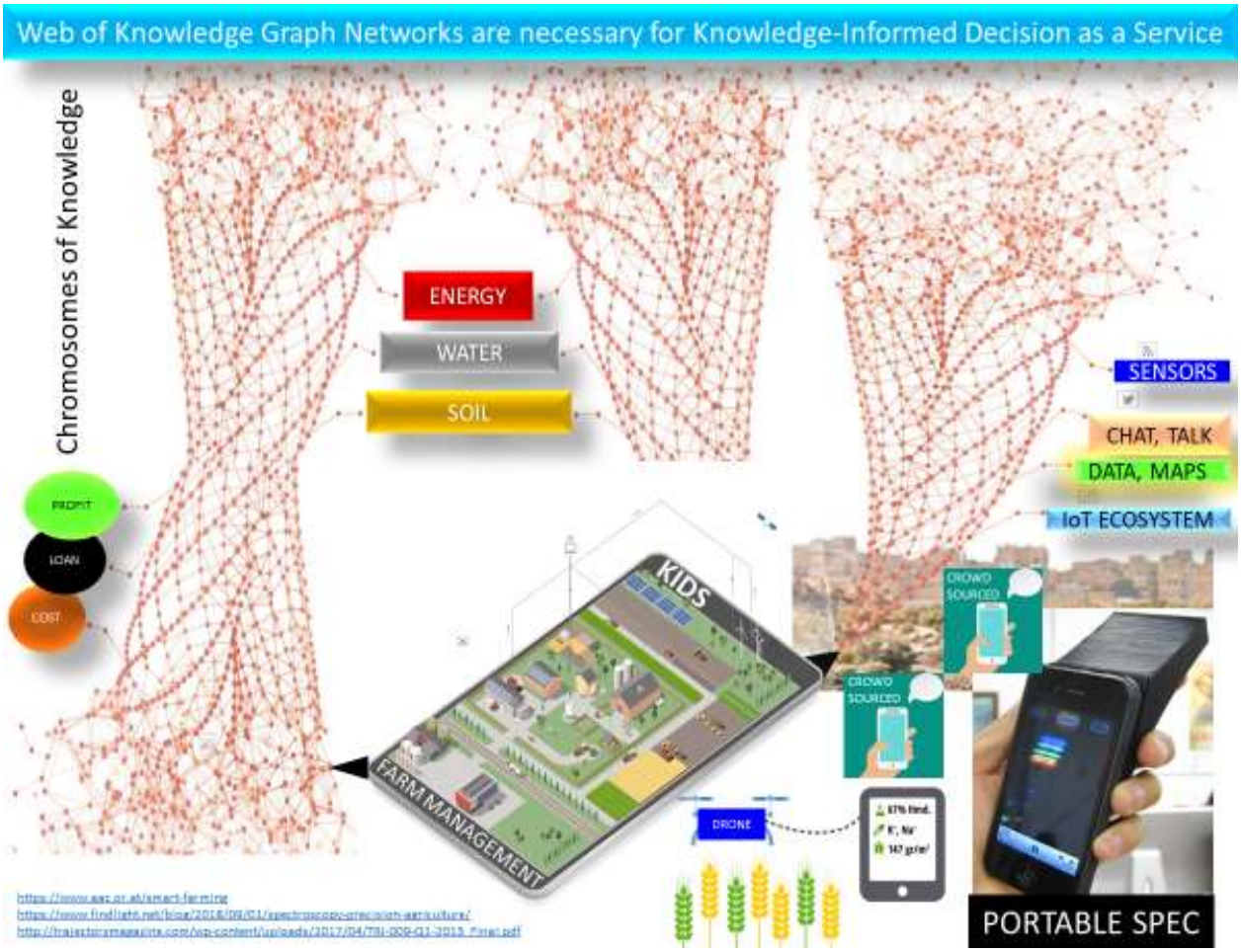
Neither OWL/RDF standards nor graph networks or knowledge graph databases, are a panacea. They may not represent everything and all advantages are temporary. Application of graph theory will not obliterate the role of other architectures and databases. The balance of tools vs interoperability between systems, are central to "understanding and forging relationships" between relevant systems, through contextual combination of tools and confluence of ideas.

Knowledge combination/integration beyond (heterogenous) rules and ontologies are not only difficult[19] but calls for ***new thinking***. ***The semantics of knowledge bases other than rules*** (for example, descriptions of temporal processes like workflows in ART which could logically decide using logic tools when the irrigation system must turn on/off water pumps, or protocols in spatio-temporal logic) ***must be integrated***. We need a higher plane of logic framework in which knowledge modules, with different native semantics, can be overlaid with meaningful semantics, preferably agnostic of linguistic bias, ideally as a "plug and play" operation, graph-friendly "drag and drop" operation for non-expert end-users, who may wish to decompose and/or re-compose the choice of logic and logic tools, based on experience or input from other expert humans in the loop. Chaperoning convergence between distributed knowledge domain(s), operational rules, data, information, and systems science, is a daunting and challenging goal (see cartoon below).



https://www.visualcapitalist.com/personal-data-ecosystem/

[19] http://www.kr.tuwien.ac.at/staff/tkren/pub/2008/rowschool2008.pdf



CAN KIDS UNDERSTAND THE QUESTIONS FROM REAL-WORLD END-USERS?

The list of sensor description related questions that the PoC attempted to answer in Step I was sourced from experts and the queries (list of questions in this document) also originated from experts. Descriptions of sensors from experts or descriptions extracted from web document searches (for example doi:10.4172/2329-6798.1000111) may be in sharp contrast to queries from real-world applications where questions are from users in agro-ecosystem, retail or healthcare.

Unstructured questions from users must be sufficiently understood by KIDS, if we are aiming to provide value for real-world applications where the user may be paying a fee for the service. To make SENSEE useful to end-users, to a limited extent, we have to start with the questions from end users (***PEAS Platform for the Agro-Ecosystem***). The end- user may want to know what types of sensors are available to detect mercury, who are the manufacturers, which brands are highly rated, what is the price, what is the maintenance fee and software licensing cost. These questions may not be answered by SENSEE 1.0 but by **ART**, in future. The query-triggered search must be able to understand the domains that the search engine must connect in order to extract the data and information relevant to the question. The latter is beyond the scope of SENSEE 1.0 and 2.0 but expected to be a building block for artificial reasoning tools (**ART**).

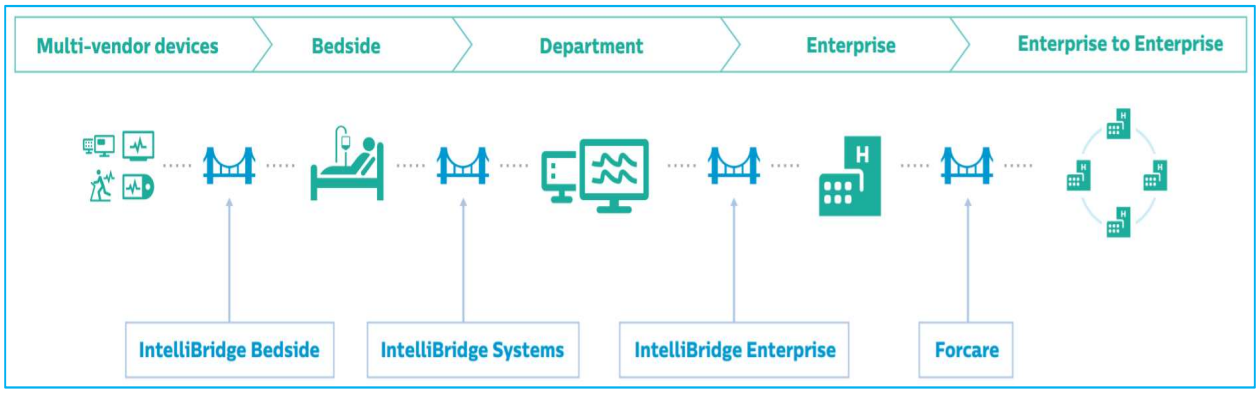
In many instances, user may ask questions about systems and technologies which may not involve sensors. For example, end-users on a farm may have a question about the amount of moisture in the soil vs the volume of water that must be dispensed by the irrigation pump system. Multiple domains must be integrated to address the diverse range of questions expected from end-users in any ecosystem.

The ability of ART, DIDA'S, KIDS, to understand the question and ***relationships in the question*** are critical to the success of PEAS platform. ABS analyst may be critical to evaluate and understand which direction to pursue and which domains to connect, based on the question.

When we move from SENSEE 1.0 (sensor descriptions in sensor search engine) to sensor data (SENSEE 2.0), the source of the data will be ***sensors in use by end-users*** (sensors deployed in farms, stores, shop floors, transportation). The end-users will have to agree to upload sensor data streams to the ***open PEAS platform***. The incentive for the user is the expectation that ART (KIDS) will make sense of the data and provide end-users with actionable information. In future, perhaps, offer knowledge, to improve decision systems or aid human users in decision making.

Participation of manufacturers, who are possessive about data and dissemination from their sensors and equipment, are potential sources of conflict. It is well nigh impossible for any one manufacturer to provide the range of sensors and equipment necessary for all operations. The manufacturer-specific dashboard may always remain a data portal, short on information and devoid of knowledge. Users, however, can change the *status quo*. User-adoption of ART (KIDS) will depend on the critical mass of data and information connectivity, as well as the ability to understand questions from users and answer them with a very high QoS (quality of service).

Aggregation platforms in the agro-ecosystem may share some analogies with the lack of device data interoperability in healthcare (https://mdpnp.mgh.harvard.edu/projects/ice-standard/). Deaths due to lack of interoperability is calling for change (https://www.himss.org/file/1325897) in the healthcare system to aggregate data in the context of the patient and transform the data to information, relevant to the patient and the point of care medical professional, as well as the extended enterprise. In other words, the cartoon below may represent **KIDS in healthcare.**



Aggregation of tools on a platform is an old (https://dspace.mit.edu/handle/1721.1/56251) idea which may find its origins in the "bazaars" of ancient Mohenjo-Darro and Mesopotamia, the "clusters" in town centers and the modern "malls" which are almost universal. Radio, TV and movie halls aggregated music, shows and movies. Digital aggregation pioneered by Amazon, eBay and Napster is evident in the streaming platform ROKU (https://blog.roku.com/oxygen). In healthcare (ICE, clinical environment, www.mdpnp.org/MD_PnP_Program___OpenICE.html) or in the agro-ecosystem, or most other system of systems, data aggregation offers value. KIDS is in good company and not an enigma for end-users, if they have the patience to start with ART.

KIDS can catalyze data fusion by aggregating contextually relevant data from different systems, and provide analytical support, for decision making, in near real-time, on a mobile app through a smartphone or tablet, anytime, anywhere. Bringing the ***algorithms to the data*** at the edge (point of use) is possible by running computation at the edge (http://eyeriss.mit.edu/) and using sparse, trainable artificial neural networks (https://arxiv.org/abs/1803.03635) to help humans make better decisions, at the edge, before the value of the information perishes. By partially automating the system, the actionable information can also actuate sensors or systems (paradigm shift from SARS to SARA – see SARS♦AG here http://bit.ly/SIGNALS-SIGNALS).

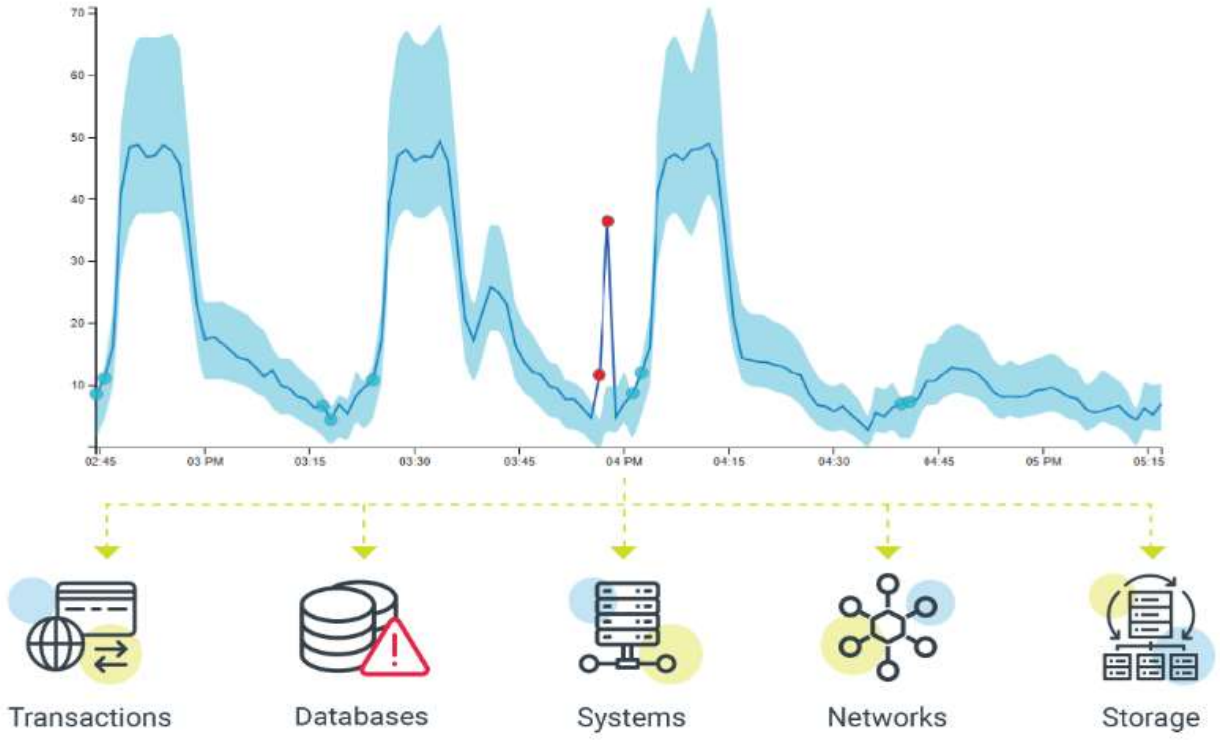
ART is expected to deliver low-risk automation to solve specific problems, for example, based on the situation and feedback from the outcome (control theory feedback optimization loop), turn on/off irrigation water pumps, selectively, by distribution zones, using a GIS map.

The value of ART (logic tools and ART are not unique approaches) and the monetization potential from knowledge-informed analytics, is linked to performance of KIDS. Imagine how agent-based artificial reasoning (ABAR) bots, may continuously seek non-obvious exceptions, non-obvious correlations and non-obvious errors. Creating machine learning algorithms and deep learning tools only to search for anomalies (red dots in the cartoon, next page, example of root cause analysis) is an under-utilization of the benefits from tools of artificial reasoning (AR).

ABAR may be trained to find positive, as well as negative, correlations. Training is still in an enigmatic *black-box* domain. With greater clarity, perhaps, training can harvest crowd-sourced nuggets of knowledge. For example, an apocryphal anecdote from an ALCOA plant describes the breakdown of a chemical processing step, just days after the retirement of an experienced plant operator. When the operator was invited back to help identify the problem, it turned out that the operator used to spit in the smelted ore chamber. After his retirement, no one was spitting during that processing step. The surfactant from the spit was key. Surfactants catalyze chemical purification processes for aluminium. Hence, the value from crowd-sourced information, knowledge, experience and wisdom.

Creativity and innovation will be necessary to capture these occasional unstructured events. The next task is to integrate them with ongoing ML/DL *training tools* "educating" ABAR. How can crowd sourced wisdom train ABAR? For mass adoption, the process must have an ETL type tool (mobile capture and upload, drag and drop) for non-experts (plant managers, transportation planners, *any* end-user).

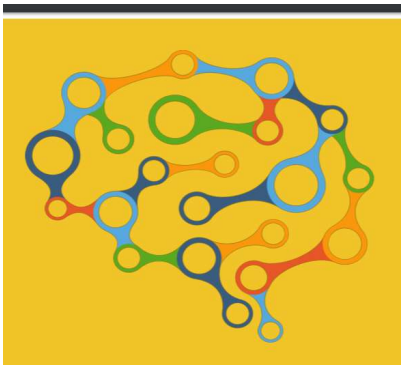
48 ◆ Exploring how sensor repositories may be helpful. See Figure 11 on page 28 – http://bit.ly/SIGNALS-SIGNALS



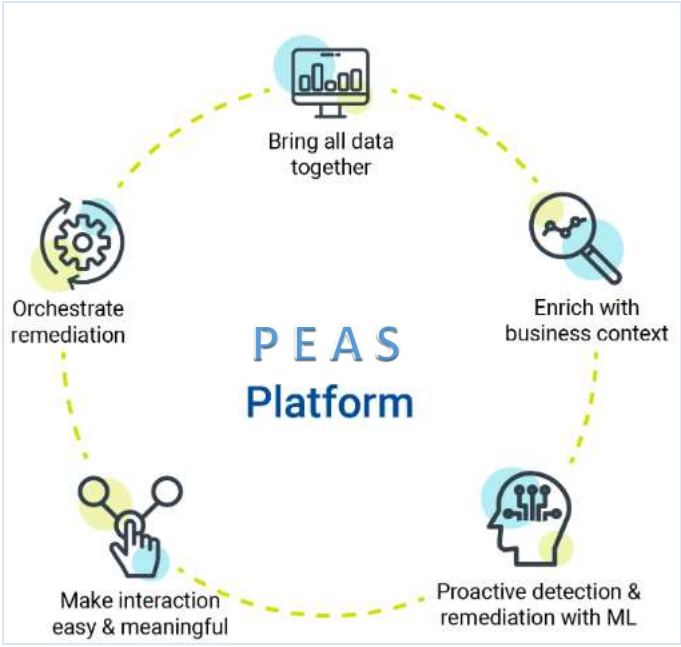
But, how will we know that a nugget of wisdom just hatched during a conversation? That is a very difficult problem. If the "nugget" is captured, how can we add it to training algorithms? One poor analogy is cooking. First, fry onions and garlic, then add spices, followed by the meat, add more aromatic spices. Another analogy is a human attached to an IV drip (intra-venous), which can drip saline, blood, morphine, antibiotics, cyanide. The key is the flexibility and modularity of adding things to a process. Do we need a "funnel" or app or API for delivery?

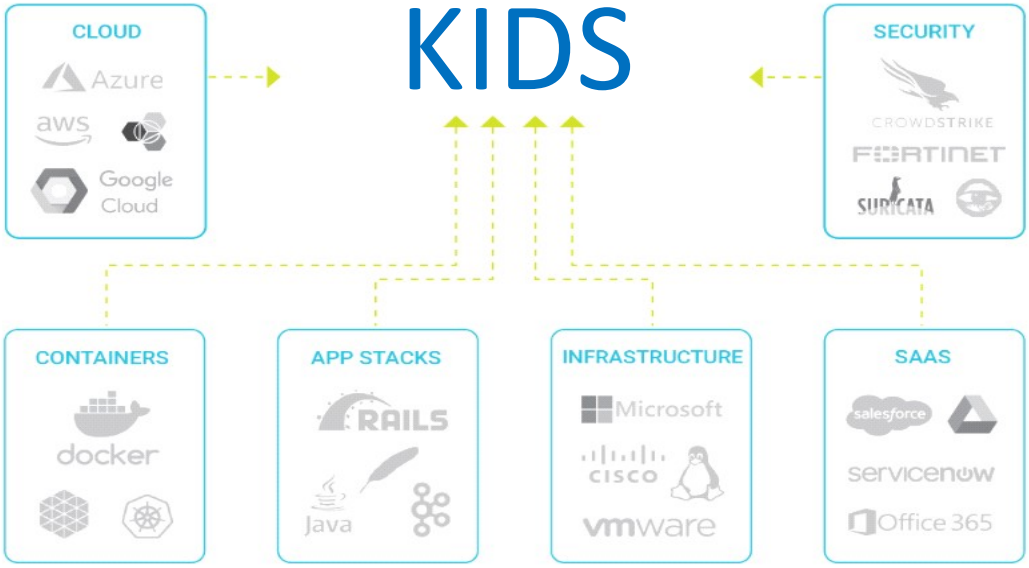
The outcome of training ABAR is to create an army of AKBAR (**a**gents with **k**nowledge-**b**ased **a**rtificial **r**easoning). In "IoT is a Metaphor" the attempt is to transform data/information to knowledge, and find new ways how organizations and enterprises may create value for users, through knowledge-informed decision as a service. KIDS with agent-based selection (ABS) will evolve to include AKBAR (agents with knowledge-based artificial reasoning), where mobile agents can travel *between* networks and cross-pollinate domains with information. Spread of mis-information, hence, raises its ugly head and cybersecurity considerations become central. However, it is still tempting to speculate, as mentioned elsewhere in this document, how we may "air-drop experience" from those who have it, to those who may wish to use it, for a fee.

The future demands we ask different questions, new questions, relevant and contextual questions, obvious and non-obvious questions, incisive and analytical questions. Hopefully, at least some of these next generation questions will also contain a few *correct questions*, to spur new thinking, create tools that are still cryptic among the unknown unknowns, and help us to visualize the possibilities, with *new eyes*.



The cartoon on the left and everything else discussed here is non-linear. Optimization routines, linear programming and static databases are rigid and less useful. All the rage about Hadoop is almost dead (HDFS, without transactions, search, indexing or caching, failed to solve real data problems) even though it was a NoSQL distributed data technology. Lack of semantics (context) turned Data Lakes into Data Swamps. To get out of the "swamp" we need NLP (not LP) and new eyes. We wish to use knowledge graphs (KG) and KG algorithms as a new path in the journey from data to knowledge (KIDS).





Towards building the next generation database query engine

Yesterday's DB engines are incapable of solving today's problems. David Mack.

In the 1970's the relational database was born (www.seas.upenn.edu/~zives/03f/cis550/codd.pdf) and remains a staple in the industry. But, enterprise companies are beginning to explore machine learning, because databases are inadequate for the company's informational needs. Relational databases have been wildly successful, forming an essential piece of almost any application. With this success has brought a rich deluge of data into database systems. Relational databases are great at supporting the developer defined symbolic relationships in the data (e.g. purchase belongs to user), but have barely any support for the noisy, sparse, probabilistic relationships that arise within the data itself (e.g. users with higher disposable income tend to make more purchases). This limitation is reflected in query languages (e.g. SQL) themselves. They are famously unfriendly for non-technical business users, so much so that entire teams of data analysts, BI experts and data scientists are drafted to help non-technical employees access their data. A very simple query such as "get the second highest salary" translates into:

SELECT DISTINCT Salary FROM Employee e1 WHERE 2=Select COUNT(DISTINCT Salary) FROM Employee e2 WHERE e1.salary<=e2.salary

A new generation of database query engine is taking a different approach. At its core, it's very different from current database query engines:

It accepts natural language (e.g. English) instead of SQL for queries

It represents data as a mixture of sparse features instead of as items from fixed categories

In the real world, nothing fits into neat boxes. Words have many meanings. Sentences can be ambiguous. Concepts and thoughts are related to others, in many different nuanced ways. Fall-leaves, tobacco and leather seem to go together, but why exactly? Our data representation supports and embraces this deep interconnectedness. We achieve this by representing data as mixtures of sparse features (i.e. many dimensional vectors). These representations are created using learned embeddings and learned transformation functions. This allows the query engine to better use the nuance of the query's words to find relevant data. It allows it to aggregate and filter data based on learned sub-categories, of which membership is not binary.

It learns multi-step deep algorithms from examples

Many times in life, we can specify the inputs and outputs, however working out how to get between them is hard (for example, try writing a series of rules to tell if a photo is of a hotdog). ML can work out the middle part in the right circumstances. Classical algorithms, the ones readily implemented in traditional database query engines, are very rigid. Each step must be a clear-cut decision with easily specified inputs. In a learned algorithm, each step can incorporate many weak signals to work out what to do next. Furthermore, it can do many different sub-steps in parallel, weaving a much more complex solution than could be written by an engineer. This is like comparing how many people cook in the kitchen vs a recipe book: we measure ingredients by eye, combine ingredients by feel and cook it until it smells and looks good. We improvise. None of which is captured in a recipe.

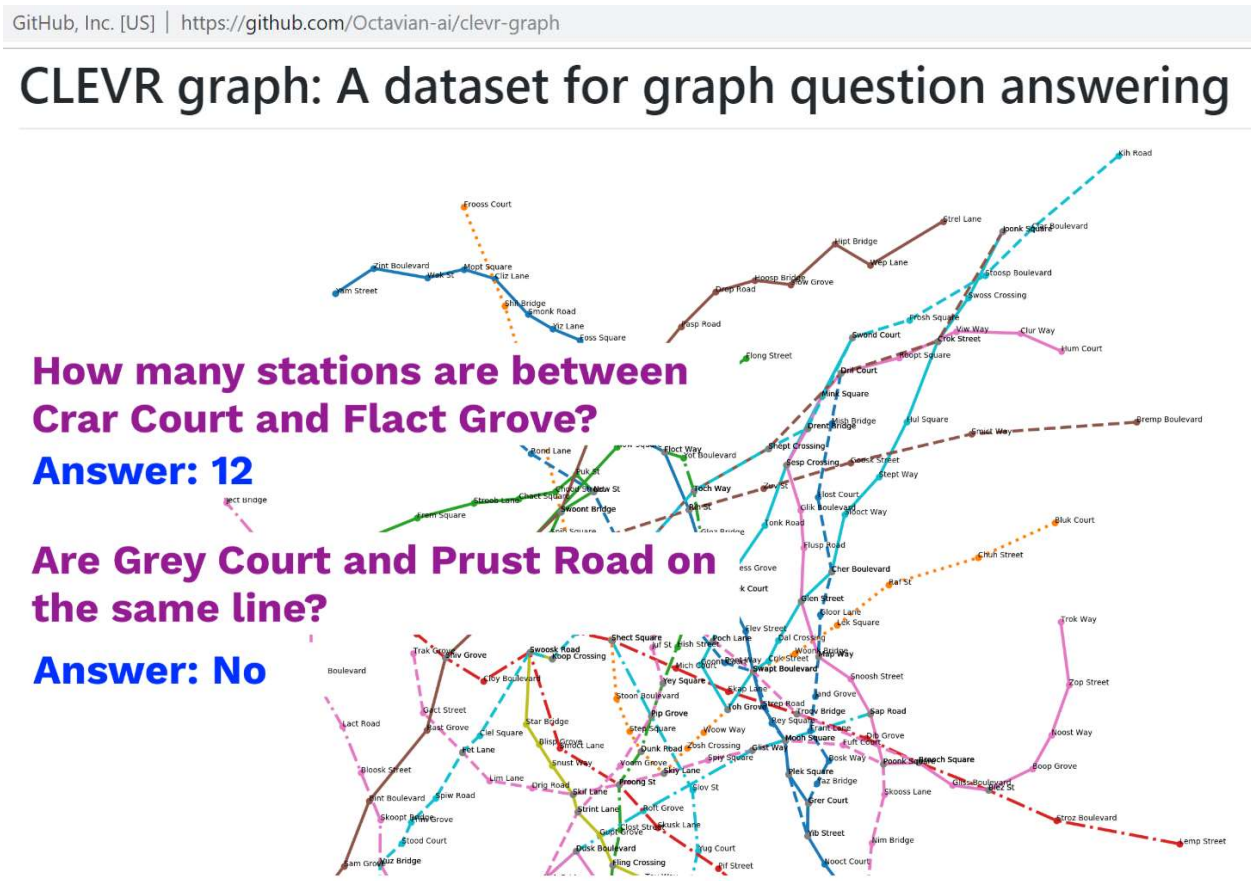
How to build it

Such a radical departure from how current query engines work requires a similar departure in the underlying technology. We're using a neural network as the core of the query engine. We present the database information as tables of data and adjacency matrices (e.g. an array of a connections) to the neural network, and let it process the data and query to produce a result. The network processes the query through an RNN and learned word embedding. This provides both an array of query tokens and also an overall query vector. The data is then processed through a network reminiscent of the Transformer architecture. After applying learned embeddings to data, it is passed through series of attention systems https://arxiv.org/abs/1706.03762 and https://arxiv.org/abs/1902.10186. These allow the network to leverage task-specific sub-networks and to combine earlier calculations together to form complex aggregates. Working example here.

Some of the sub-networks include (for our graph-processing network):
Node property recall
Edge (i.e. relationship) recall
Using previous step's output as addressing instructions for the above
Iterative message passing
Recalling previous step's output and transforming them in a range of ways

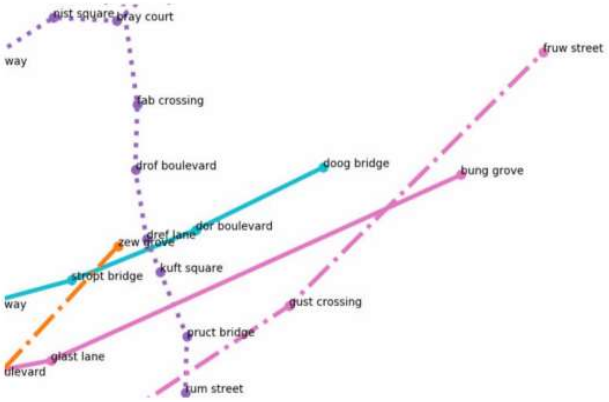
EXPLORE KEY PAPERS in this zipped folder http://bit.ly/ML-MISC-01 and the list provided here:
https://arxiv.org/abs/1706.03762
https://arxiv.org/abs/1806.01261
https://arxiv.org/abs/1803.03067
https://arxiv.org/abs/1711.09846
https://arxiv.org/pdf/1905.12107.pdf
https://github.com/deepmind/graph_nets

Andrew Jefferson Octavian http://bit.ly/GKG-DL-TALK-2018 DB query engine see our recent article.



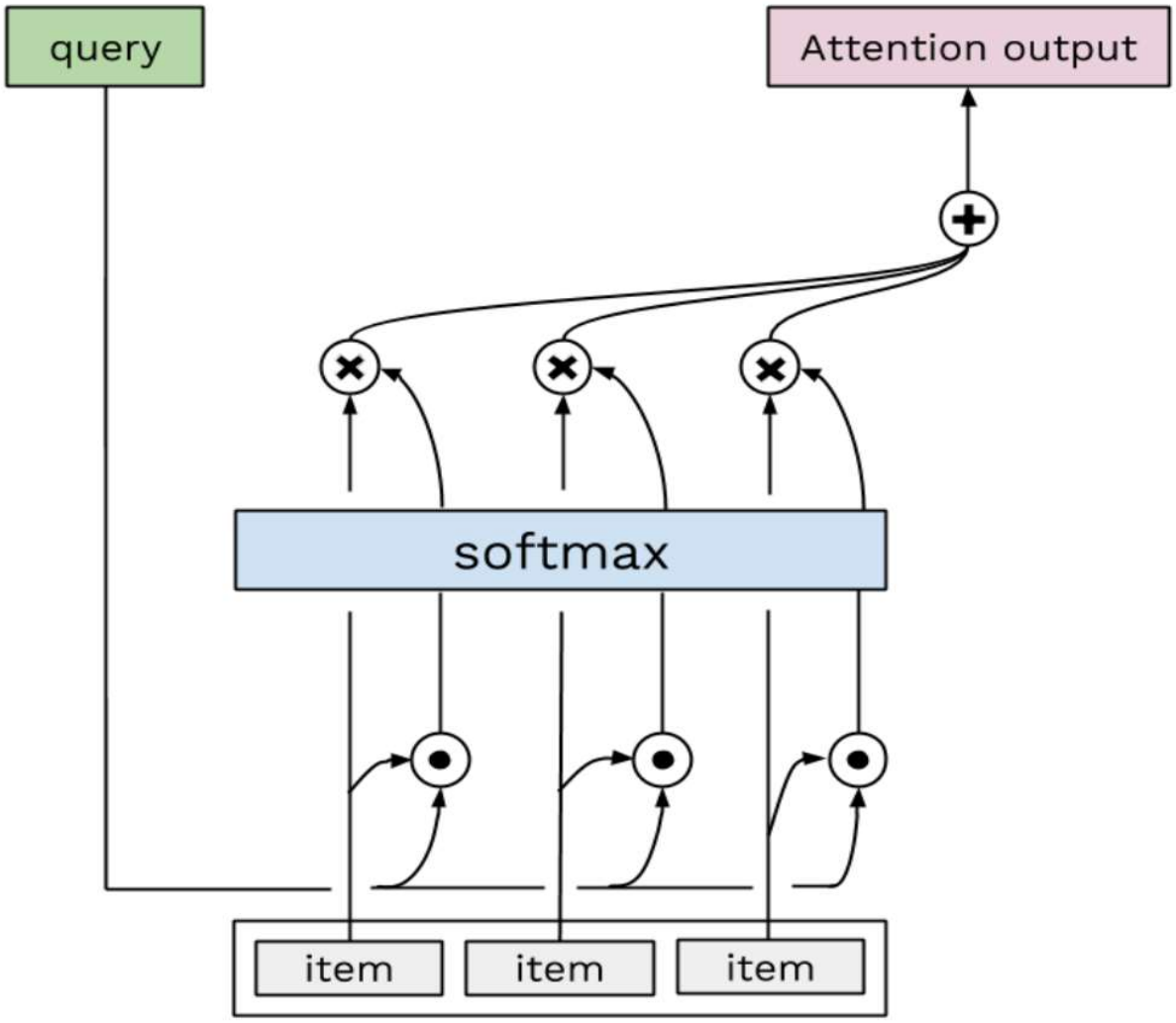
The graph data is modelled on transit networks (London tube and train network). Questions are modelled on questions typically asked by passengers (users) around mass transit (How many stops between? Where do I change?). Aim: solution to this dataset has real world applications.

- Does {Station} have disabled access?
- Is there disabled access at {Station}?
- Does {Station} have rail connections?
- Can you get rail connections at {Station}?
- How many stations are between {Station} and {Station}?
- Are {Station} and {Station} adjacent?
- Which {Architecture} station is adjacent to {Station}?
- Are {Station} and {Station} connected by the same station?
- Is there a station called {Station}?
- Is there a station called {FakeStationName}?
- Which station is adjacent to {Station} and {Station}?
- How many architectural styles does {Line} pass through?
- How many music styles does {Line} pass through?



Is “ATTENTION” insufficient?

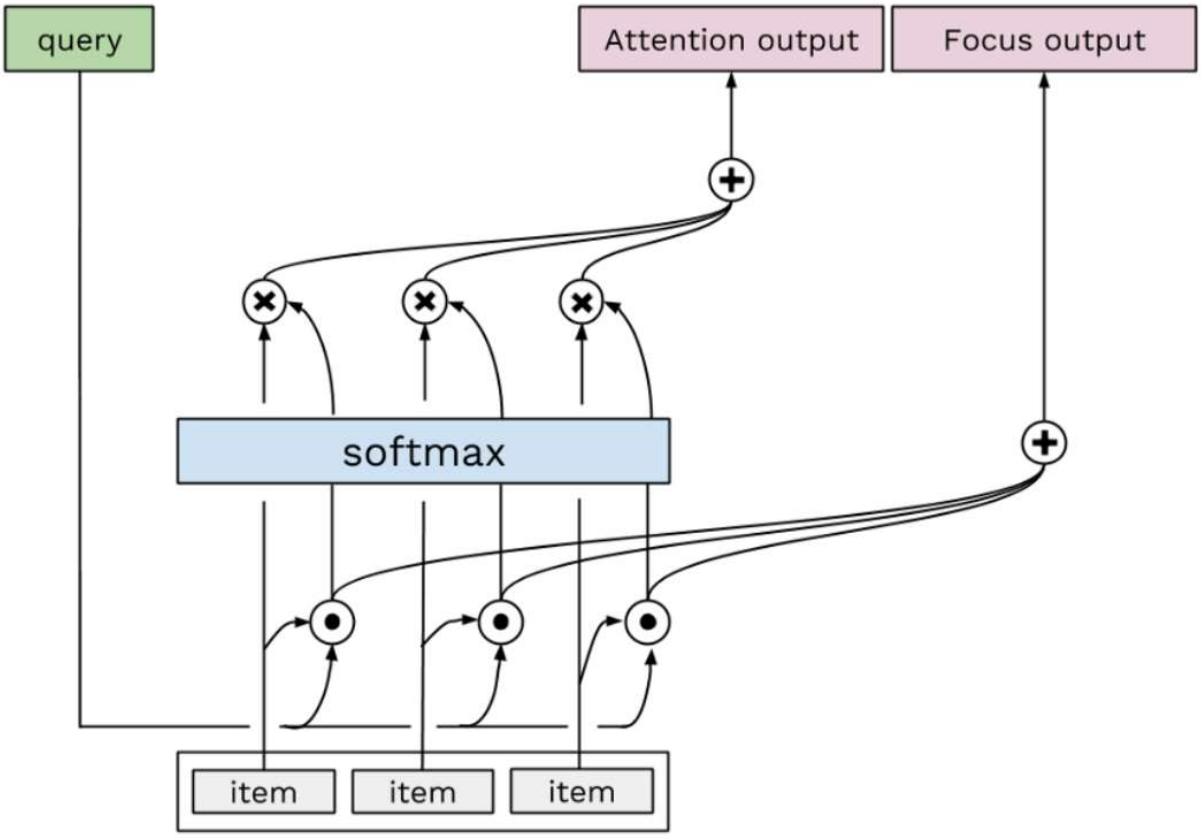
How to show if some things are not present in a list?



A popular neural-network technique for working with lists of items (e.g. translating sentences treating them as lists of words) is to apply “attention”. This is a function where a learnt “query” of what the network is looking for is compared to each item in the list, and a weighted sum of the items similar to the query is output. Attention is the basis of current best-in-class translation models. The mechanism has worked particularly well because tasks can be solved by rearranging and combining list elements to form a new list (e.g. attention models have been important components in best of class translation, question-answering, reasoning models).

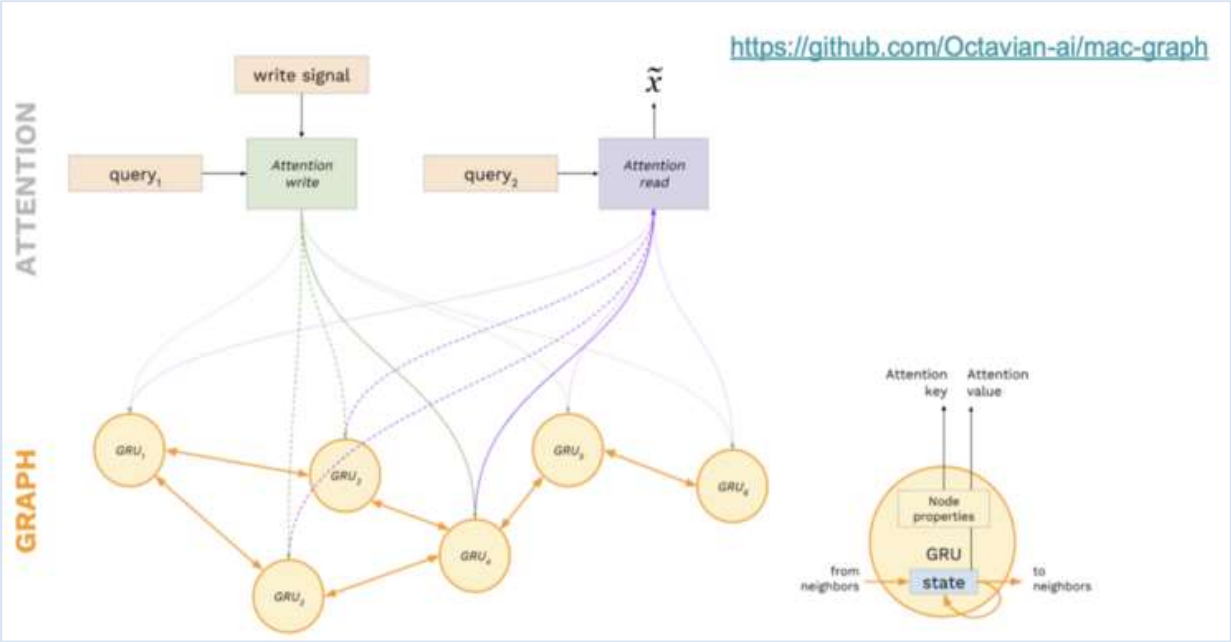
Despite attention's versatility and success, it has a deficiency that plagued our work on graph question answering: **attention does not tell us if an item is present in a list.** This first happened when we attempted to answer questions like "Is there a station called London Bridge?" and "Is Trafalgar Square station adjacent to Waterloo station?". Our tables of graph nodes and edges have all this information for attention to extract, but attention itself was failing to successfully determine item existence. This happens because attention returns a weighted sum of the list. If the query matches (e.g. scores highly) against one item in the list, the output will be almost exactly that value. If the query did not match any items, then a sum of all the items in the list is returned. Based on attention's output, the rest of the network cannot easily differentiate between those two situations.

The simple solution we propose is output a scalar aggregate of the raw item-query scores (e.g. before using softmax). This signal will be low if no items are similar to the query, and high if many items are. In practice this has been very effective (indeed, the only robust solution of the many we've tested) at solving existence questions. From now on we will refer to this signal as "focus".



Attention with focus signal, using a summation for aggregation of raw scores

https://github.com/Octavian-ai/attention-focus



What is the cleanliness level of {Station} station? How big is {Station}? What music plays at {Station}? What architectural style is {Station}? Describe {Station} station's architectural style. Is there disabled access at {Station}? Does {Station} have rail connections? Can you get rail connections at {Station}?	99.9% accuracy after 10k training steps
Are {Station} and {Station} adjacent?	99% accuracy after 20k training steps
Which {Architecture} station is adjacent to {Station}?	98.8% accuracy after 30k training steps
How many stations are between {Station} and {Station}?	98% accuracy up to ~9 apart after 25k training steps
Are {Station} and {Station} connected by the same station? Which station is adjacent to {Station} and {Station}?	98% accuracy after 20k steps
Is there a station called {Station}? Is there a station called {FakeStationName}?	99.9% accuracy after 30k training steps
Are {Station} and {Station} on the same line?	Not yet tested
How many architectural styles does {Line} pass through?	Not yet tested
How many music styles does {Line} pass through?	Not yet tested
How many sizes of station does {Line} pass through?	Not yet tested
How many stations with rail connections does {Line} pass through?	Not yet tested
Which lines is {Station} on?	Not yet tested
How many lines is {Station} on?	Not yet tested
Which stations does {Line} pass through?	Not yet tested
What's the nearest station to {Station} with disabled access?	Not yet tested

Some example questions from CLEVR graph question bank. It's a synthetic (procedurally generated) dataset which consists of 10,000 fictional transit networks modelled on the London underground. For each randomly generated transit network graph we have a single question and correct answer. Each graph used to test the network is one the network has ***never seen before***. Therefore, it ***cannot memorise*** the answers to the questions but must learn how to extract the answer from new graphs.

The data (DIDA'S) to knowledge (KIDS) transition in my essays are eloquently explained by **Dan McCreary.** I am *copying* a few of his articles that are relevant to these essays. I have edited the content and removed leading suggestions (subtle advertising). Please access original versions from www.danmccreary.com/

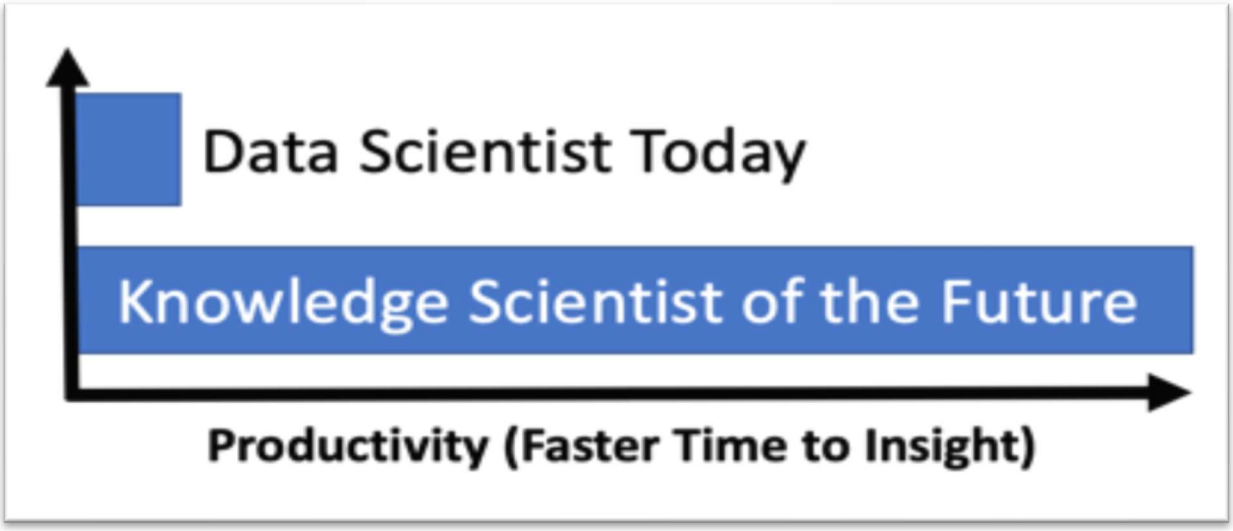
Another resource worth exploring is here http://bit.ly/Yann-LeCun but it contains a myriad of exaggerated claims and ideas of solutions "provided by powerpoint".

"Imitation is the sincerest form of flattery that mediocrity can pay to greatness."

– Oscar Wilde

57 ◆ Exploring how sensor repositories may be helpful. See Figure 11 on page 28 – http://bit.ly/SIGNALS-SIGNALS

From Data Science to Knowledge Science



Knowledge scientists may be more productive than data scientists because they may offer a new set of assumptions about the inputs to their models and store their insights in a knowledge graph. Their input features may remain **highly connected** to other relevant data as such as provenance and lineage metadata. An article pointed out: *Data scientists…spend from 50-80% of their time as mundane laborers, collecting and preparing unruly digital data, before it can be explored for useful nuggets.*

Cleaning up data involves data cleanup code. **Feature Store** is an attempt to build reusable artifacts for data scientists. Google and Uber have discussed their efforts to build tools to reuse features and standardize the feature engineering processes. My big concern is that many of these efforts are focused on building flat files of disconnected data. Once the features have been generated they can easily become disconnected from reality. They quickly start to lose their relationships to the real world.

58 ◆ Exploring how sensor repositories may be helpful. See Figure 11 on page 28 – http://bit.ly/SIGNALS-SIGNALS

An alternative approach is to build a set of tools for analysts to connect directly to a well-formed enterprise scale knowledge graph to get a subset of data and transform it to structures that are immediately useful for analysis. The results of this analysis can be used to enrich a knowledge graph. These Machine Learning approaches can complement the library of turn-key graph algorithms.

Data quality within a knowledge graph: MarkLogic is a document store where native data is stored in either JSON or XML documents. It promotes productivity through [1] document-level data quality score and [2] implicit query language-level validation of both simple and complex data structures. In MarkLogic, a built-in metadata element called the **data quality score** is usually an integer between 1 and 100 that assigned as new data enters the system. A low score (<50), indicates quality problems (missing data elements, fields out of acceptable ranges or corrupt or inconsistent data). A score of 90 may indicate that it could be used for downstream processes. Documents with a score >70 or >80 may improve search or analysis performance. To accomplish this task, a validation schema is built-in to MarkLogic. The concept of valid data is also built into W3C document query language (XQuery). Each document can be associated with a root element (within a namespace) and bound to an implicit set of rules about that document. GUI editor (oXygen XML Schema editor) allows non-programmers to create and audit data quality rules. XML Schema validation can generate a true/false Boolean value as well as a count of the number of errors in the document. Together with tools like Schematron and external data checks, each data steward can determine how to set the data quality score for various documents.

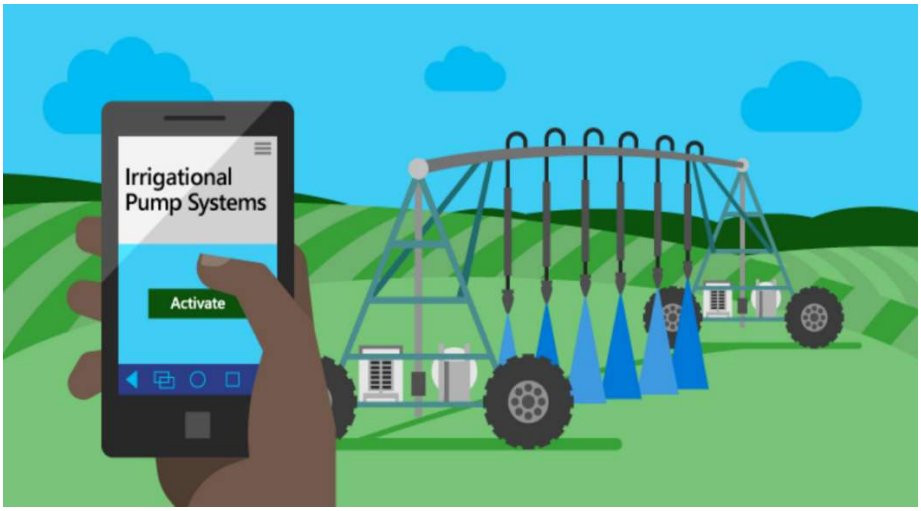
Document-level data quality scores are a natural fit with **events**, as a part of the workflow (for example, inbound call to call center, new customer purchase, subscription renewal, enrollment in a new healthcare plan, a new claim being filed, new sensor installed, new sensor data stream goes live). All of these events can be captured as complete documents, stored in streaming systems like Kafka and ingest the business event data to knowledge graphs. Data quality scores can be included in business event documents and knowledge graph (use it anywhere it is necessary for analysis).

In contrast, dumping table-by-table data from relational databases into CSV files are full of numeric codes that may not have clear meaning. Curating this low-level data from “data lakes” to deliver meaningful connected knowledge is an arduous task. Storing flattened CSV-level data and numeric codes is where features stores fall short. Once the features are extracted and stored in a data lake or object store, they become disconnected from how they were created. A new process might run on the knowledge graph that raises or lowers the score associated with a data item. However, that feature can’t easily be updated to reflect the new score. Feature scores can add latency that will prevent new

data quality scores from reflecting the current status. Perhaps, relational data architects (using tables to store data) tend to under-value document models and associating data quality score with event documents.

In summary, the quality in a graph is different than quality in a document (please explore https://en.wikipedia.org/wiki/SHACL). The connectedness of a vertex in a graph will also determine quality (please explore https://www.w3.org/TR/prov-dm/). Eventually, knowledge graph products may have a metadata layer about how data was collected and transformed during the journey from the source to the knowledge graph. As RDF fades into the annals of W3C, we see a concomitant rise of the LPG (labeled property graph) ecosystem. LPG still lacks mature machine-learning integration tools to enable knowledge science.

Note: *This is a step beyond ART toward knowledge. It signals a move, in principle, from data-informed decision as a service (DIDA'S) to knowledge-informed decision as a service (KIDS). The cartoon below is a stand-alone digital proxy for irrigation pump system. The relevance of the data and information (before pump activation) must be* ***correlated*** *with soil moisture, optimum moisture saturation desired for the crop, the weather (prediction of rain within an acceptable window of time or forecast for even higher temperature which may accelerate loss of moisture from the soil), and other relevant information in the context of this action (activating irrigation pumps). ART can aggregate the information and ABS can direct pump activation (pump speed, volume of water, duration of action, coverage area, energy consumed). Analytical engines at the edge (on mobile phones) running short neural networks (https://arxiv.org/pdf/1803.03635.pdf) may assist ART, ABS, KIDS. This is an observation by the author and* ***not*** *due to Dan McCreary (http://bit.ly/GKG-KIDS).*

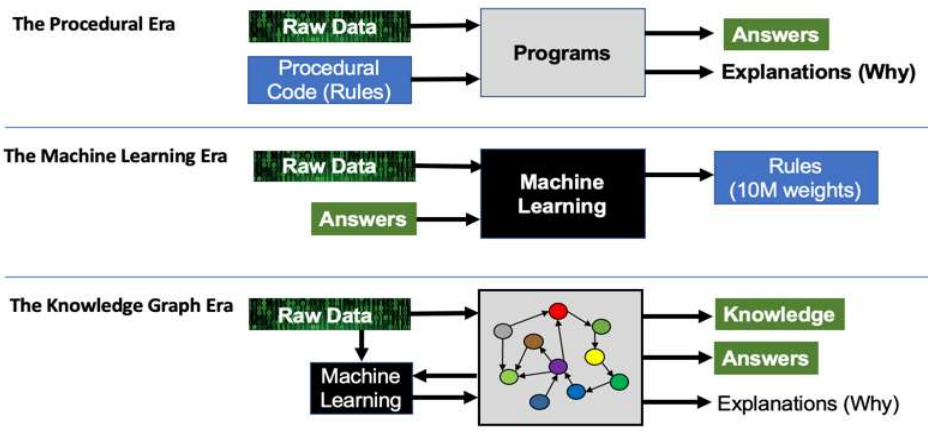


Link: https://cloudblogs.microsoft.com/industry-blog/microsoft-in-business/2019/05/15/how-polaris-energy-services-is-transforming-the-agriculture-industry-in-the-cloud/

Knowledge Graphs: The Third Era of Computing

"When did computing start?" A cuneiform tablet from circa 3,000 BC may hold the answer. Knowledge representation began when we wanted to remember things that were important to us. For example, ledger of financial transactions such as "X owes Y ten baskets of grain." It was natural to store these facts in rows and columns of a table because tables were a good "natural representation" for financial transactions. These transactions records evolved into rows of symbols which represented concepts and gave birth to written languages. These representations continued for 5,000 years. Clay tablets evolved into papyrus scrolls, then Luca Pacioli's double entry bookkeeping system, which eventually became punch cards and then flat files in COBOL, then tables in a row-store popular in relational databases and finally Enterprise Resource Planning (ERP) systems.

These tabular representations worked well when our problem had uniform data sets. By uniform we mean that each record (row) has similar attributes with similar data types. Not all problems fit well into tables. The more tables you have the more expensive the relational joins. How do we store the **analysis** of a patient chart? You might have a so-called AI agent scanning data (eg drug adherence data). What do they produce? The answer is often a list of conditions and the probability of occurrence (diabetes, asthma). Healthcare systems store these concepts in a complex hierarchy (a taxonomy) with many connections between the concepts (ontology). The AI tool may recommend next best actions. This analytical data (outcome) may not fit easily into a table. But it does fit well into a graph of connected concepts, the knowledge graph.



The bulk of developers are still writing PHP/Java/Python over tables (not graphs). In the knowledge graph era, machine learning continuously reads raw data, combines this with existing knowledge and produces new knowledge, answers and explanations. Knowledge graphs are at the core of the third era of computing, aimed to enrich shared knowledge.

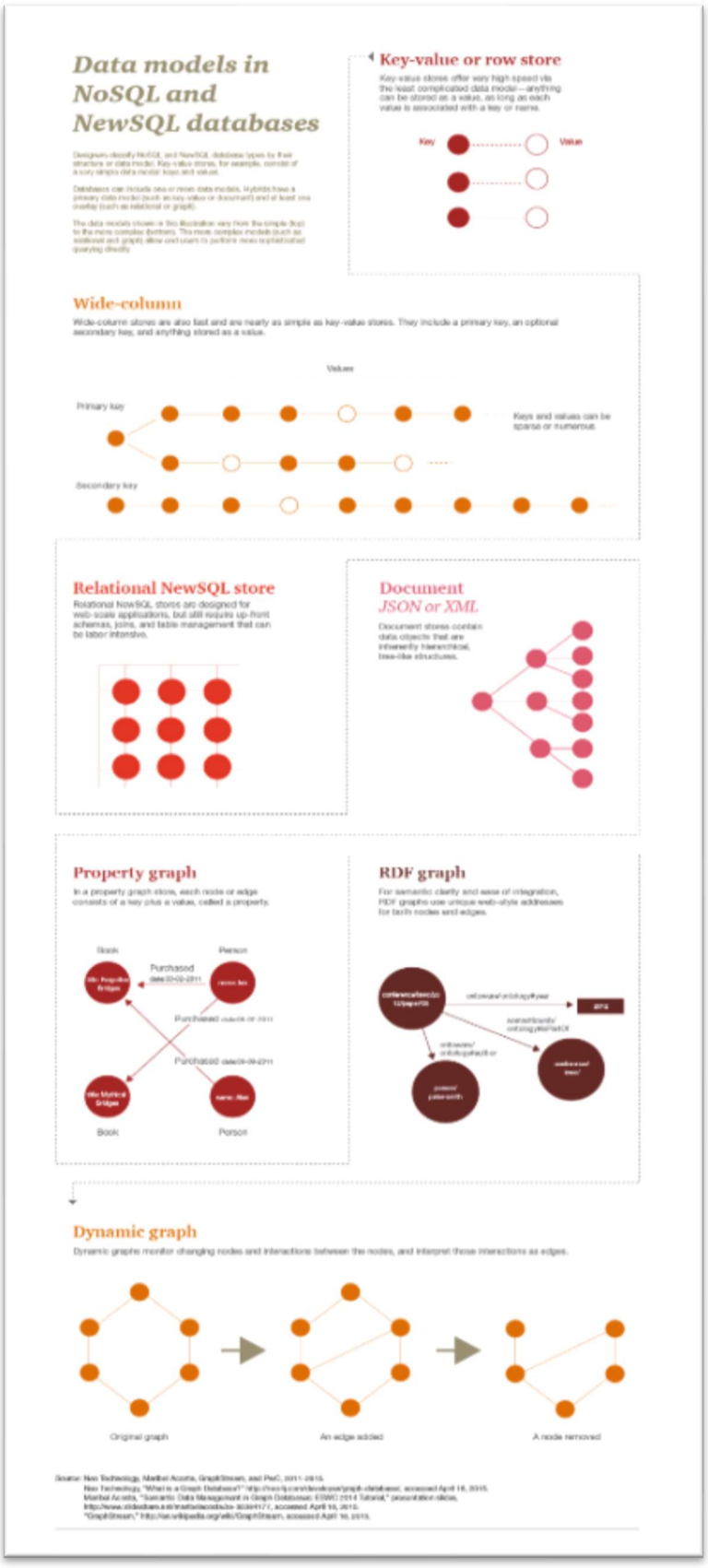
Knowledge graphs combine to produce a system that not only learns from complex data, but it also can explain its decisions. We use machine learning to harvest raw data and look for patterns in this data. Machine learning finds relevant information (people, places and things) in images, texts and sound. ML converts this to new entries in our knowledge graph along with confidence weights. The data can be checked for consistency and quality by graph algorithms. The outcome from the graph is new knowledge, answers and explanations of why we made specific decisions. Our knowledge graph becomes a repository of semantically precise vertices and relationships with confidence weights retained from the machine learning processes.

However, knowledge enrichment processes are not perfect and can easily add false assertions if new facts are not curated by subject matter experts (promotes fake news).

The justified emergence of knowledge graphs as a buzzword surrounds its ability to use very large distributed graph databases to store complex networks of concepts that can be "traversed" using a highly parallel graph query language. Knowledge graphs in Google, Facebook, LinkedIn, Amazon (product graph) and Pinterest (interest graph) have over 100 billion vertices, thus solving the scalability problems for graph databases.

In the past, the predominant way of building knowledge graphs was to use hand coded knowledge and an inference engine that could leverage higher-level RDF-based standards such as RDFS, OWL and SKOS. Now organizations are using machine learning to build seed concept graphs using natural language processing (NLP). There is also a strong shift to use the more flexible labeled-property-graphs (LPGs) to do similar reasoning (http://bit.ly/WHY-ARTIFICIAL-REASONING).

- https://medium.com/@dmccreary/how-knowledge-graphs-promote-fake-news-362947220ea8
- https://medium.com/@dmccreary/blockologies-a-pattern-language-for-ai-data-flows-de9f4507547


Data models in NoSQL and NewSQL databases
Key-value or row store
Key
Value
Wide-column
Values
Primary key
Keys and values can be sparse or numerous
Secondary key
Relational NewSQL store
Document
JSON or XML
Property graph
Book
Person
Purchased
Purchased
Purchased
Book
Person
RDF graph
Dynamic graph
Original graph
An edge added
A node removed

◆ Di ◆ Data-informed ◆ exploration of the tessellated facets of our elusive quest for meaning