# BIOGRAPHY

Tim Williams

I am a Biologist by training and Computer Systems Analyst by trade, with experience in systems administration, validation, and systems deployment in the pharmaceutical industry. A former SAS programmer, I started learning R coincident with my interest in the Semantic Web and Linked Data in 2012. My experience as a mentor and educator ranges from teaching programmers in Mumbai to leading interactive Knowledge Graph Workshops and multiple Linked Data projects with the PHUSE organization.

Tim.Williams@ucb.com

LinkedDataTim@gmail.com

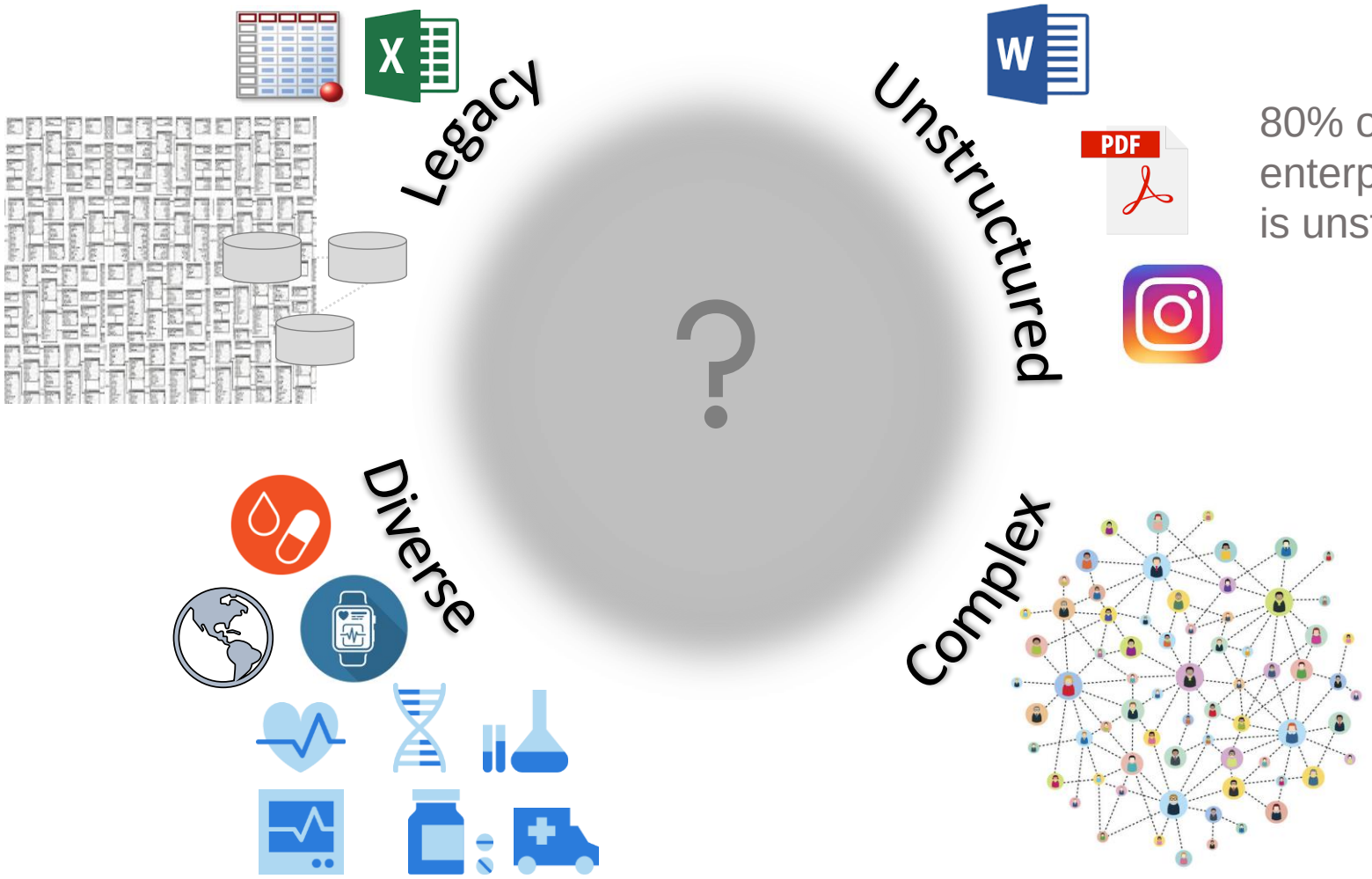
@NovasTaylor

https://www.linkedin.com/in/timpwilliams

# Knowledge Graphs as a Foundation for the Analytics ?evolution

# Why do we need a ?evolution?

# Data Landscape




80% of enterprise data is unstructured*

*Gartner

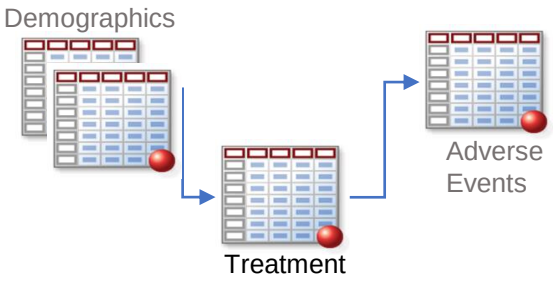
# Industry Standards

## Study Data Tabulation Model

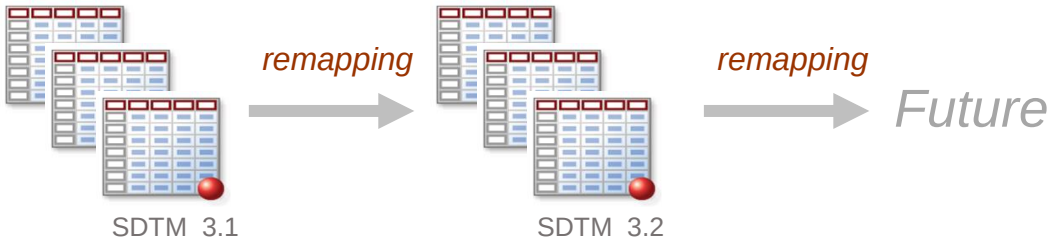
"...defines a standard structure for human clinical trial (study) data tabulations and for nonclinical study data tabulations that are to be submitted as part of a product application to a regulatory authority..."

- en.wikipedia.org/wiki/SDTM

- Domain Silos
- ≠ Clinical Trial Process Model




- $ Millions for version conversion, pooling, data integration, data mining



# Code Landscape

## What does the code do?

- Data Manipulation
  - Merging, subsetting, categorization…
- Statistical Procedures
- Output formatting

## Problems

- Errors. 15-50 defects 1000 lines*
- Application & Skill set lock-in
- Logic and metadata in code
- Add context and complexity

* "Code Complete" – Steve McConnell

# Knowledge Graphs

# Knowledge Graph

- Semantic graph data that is interpretable by both machines and humans

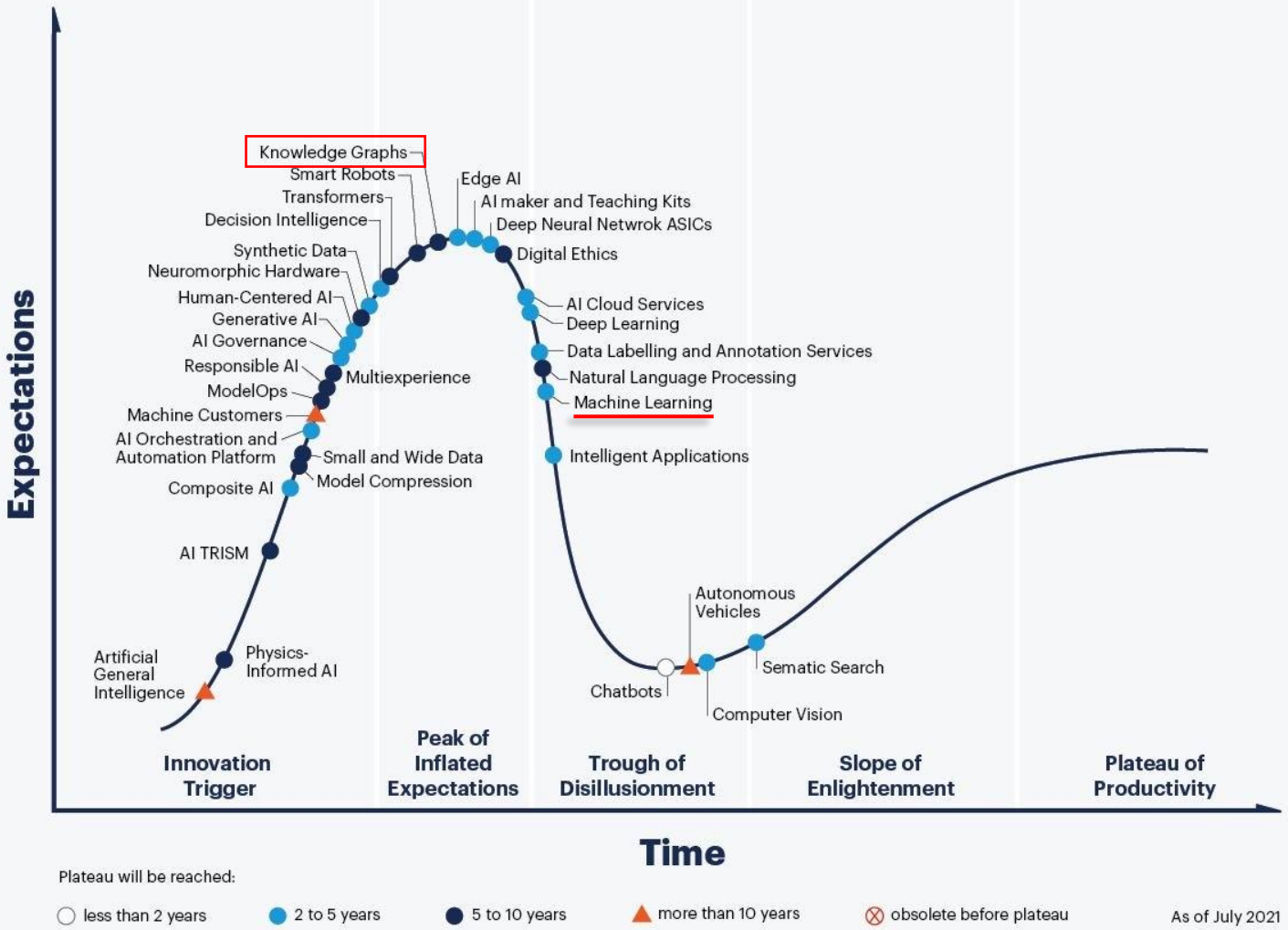
Free e-book

A Data Engineer's Guide to Semantic Modelling

- Ilaria Maresi (June 2020)

https://www.thehyve.nl/articles/semantic-modelling-guide-ebook

# Hype Cycle for Artificial Intelligence, 2021




Plateau will be reached:

○ less than 2 years ● 2 to 5 years ● 5 to 10 years ▲ more than 10 years ⊗ obsolete before plateau

As of July 2021

gartner.com

Source: Gartner

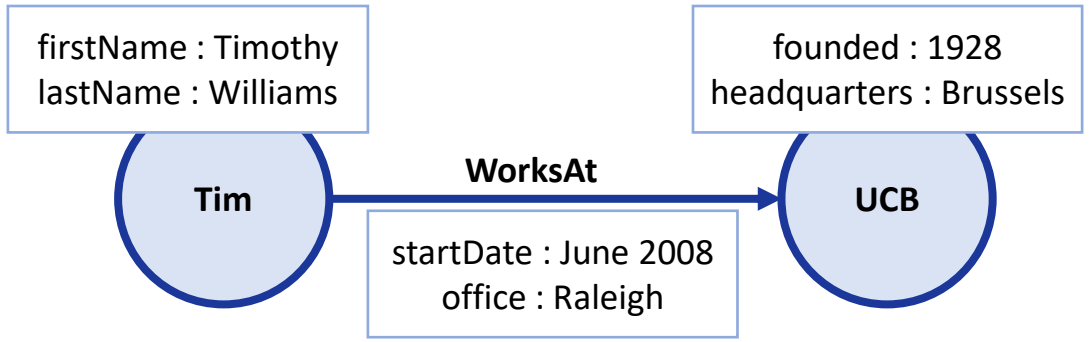


Gartner

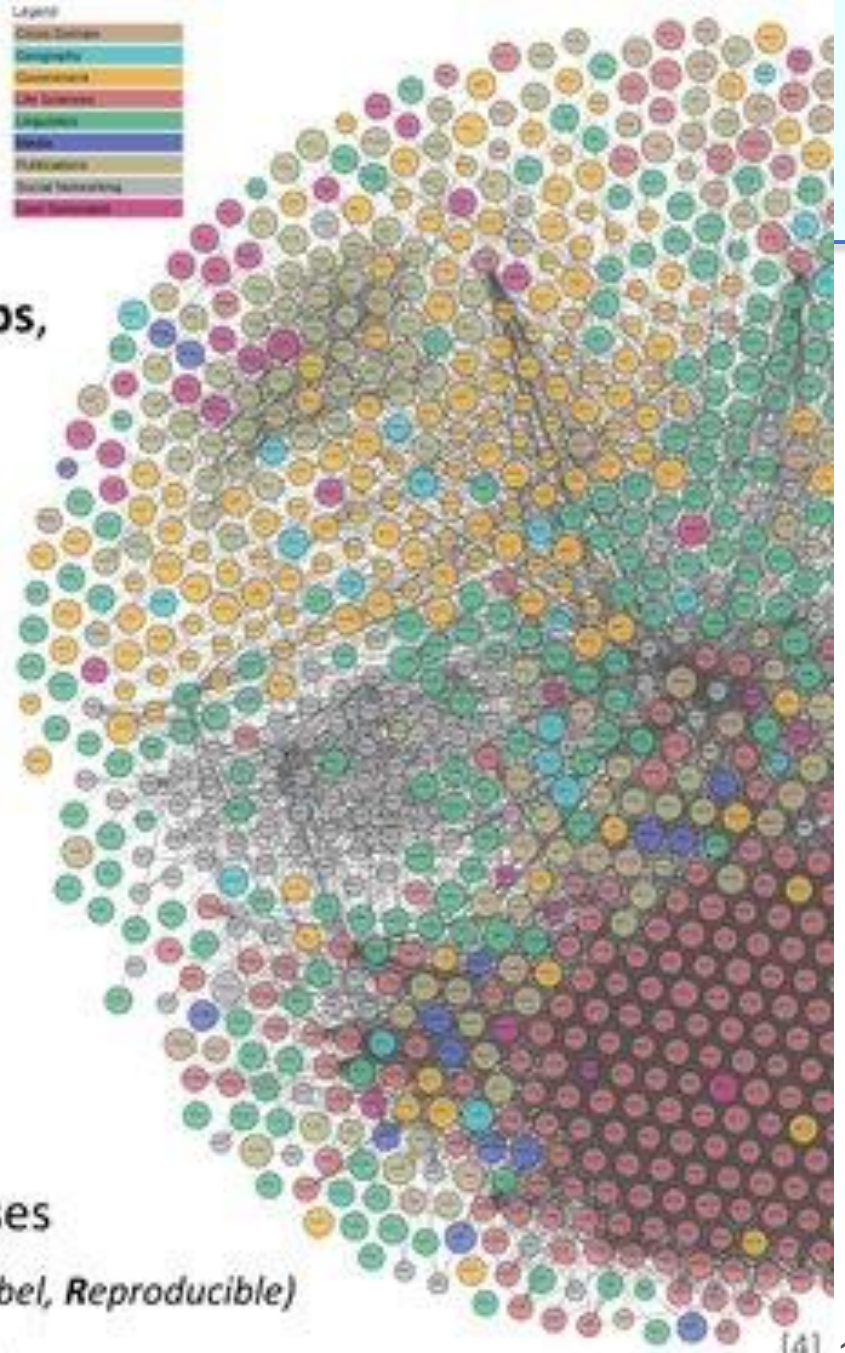
# Different Types of Graphs

## Resource Description Framework (RDF)


## Labeled Property Graph



# Knowledge Graphs - A Definition



- A **Graph** consisting of **concepts**, **classes**, **properties**, **relationships**, and **entity descriptions**
- Based on **formal knowledge representations** (RDF(S), OWL)
- Data can be **open** (e.g. DBpedia, WikiData), **private** (e.g. supply chain data), or **closed** (e.g. product models)
- Data can be **original**, **derived**, or **aggregated**
- We distinguish
  - **instance data** (ground truth),
  - **schema data** (vocabularies, ontologies)
  - **metadata** (e.g. provenance, versioning, licensing)
- **Taxonomies** are used to categorize entities
- **Links** exist between internal and external data
- Including **mappings** to data stored in other systems and databases
- *Fully compliant to* ***FAIR Data principles*** *(Findable, Accessible, Interoperabel, Reproducible)*

[4]

# Working with Knowledge Graphs

# Select KG Concepts and Languages

## General Concepts
- Linked Data
- Semantic Web
- Knowledge Graph
- Reasoner & Inference
- Nodes, Edges
- URIs / IRIs
- Graph Database
- Triplestore
- F.A.I.R. Principles

## Graph Types
- Resource Description Framework (RDF)
  - RDF-Star
- Property Graph

## Rules and Validation Languages
- SPARQL Inferencing Notation (SPIN)
- Semantic Web Rule Language (SWRL)
- Rule Interchange Format (RIF)
- Shape Expressions (ShEx)
- Shape Constraint Language (SHACL)

## Query Languages
- SPARQL
  - SPARQL-Star
- Cypher
- GraphQL

## Mapping Languages
- RML
- R2RML
  - SMS

## Miscellaneous
- LDFlex
- LDN (Linked Data Notifications)

## Languages for Working with Graph Data
- **Haskell**
- **JavaScript**
- **Perl**
- **Python**
- **Ruby** ☆
- **SAS**
- **Java**
- **C++**
- **Scala**
- **Go**
- **Rust**
- **Elixir**
- **Erlang**
- **PHP**
- **R**
- **D**
- **Clojure**

...and many more

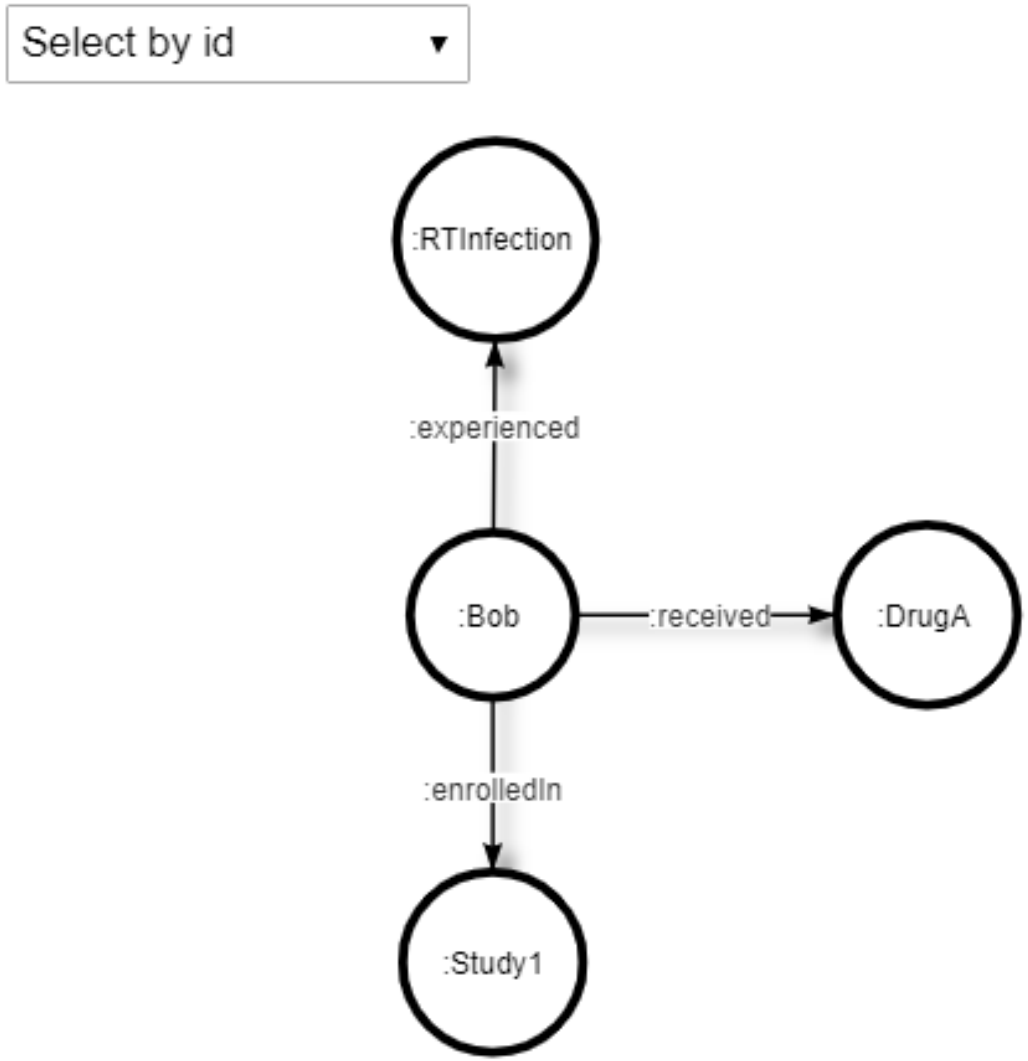
## Ontology and Schema
- Web Ontology Language (OWL)
- Resource Description Framework in Schema (RDFs)
- schema.org
- Ontologies for Life Sciences

## RDF Syntax
- RDF-XML
- RDFa (attributes)
- Notation3 (N3)
- N-Triples
  - Turtle (TTL)

# Graph Whiteboard Application

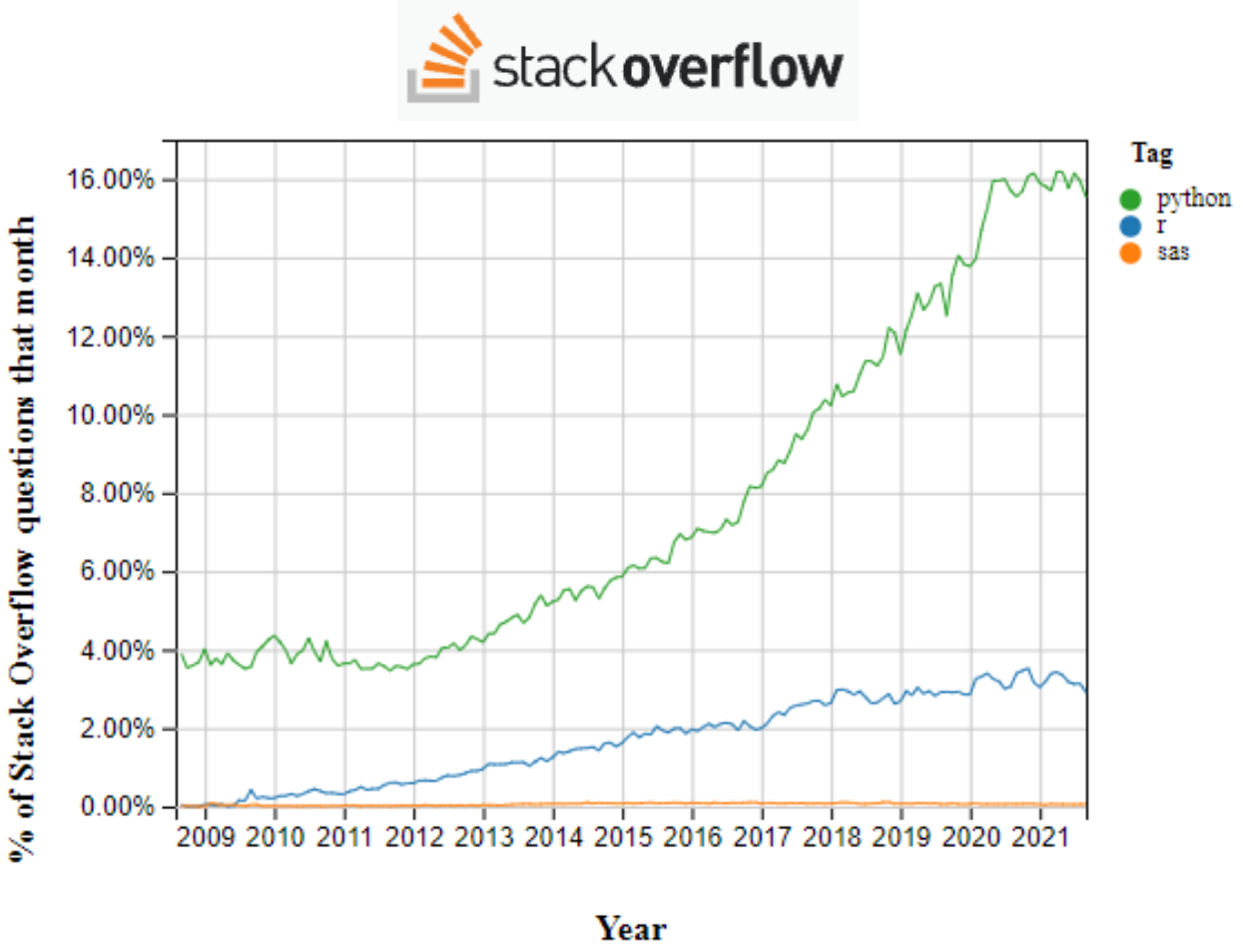
https://arrows.app/

# Screen Capture from R Program



# ? For Creating RDF

The ? in the "*Analytics ?evolution*"




## R Libraries for RDF

- rdflib
- redland
- jsonld
- digest # SHA1 hashing for IRI formation
- tools # md5sum hashing: IRI formation, file content changes

# ? For Creating RDF

1. Data to R data frame
2. Data Massage
   a. Create IRIs in data
   b. Create RDF according to required model

Option 1: Traditional R Dataframe loop

```
srcDf <- patientData
for(i in 1:nrow(srcDf)){
  rdf_add(some_rdf,
    subject   = paste0(STUDY, "PATIENT_", srcDf[i,"patID_h"]),
    predicate = paste0(STUDY, "received"),
    object    = paste0(STUDY, srcDf[i,"drug"])
  )
}
```

Option 2: Tidyverse Approach

"A tidyverse lover's intro to RDF" - Carl Boettiger

3. Write out to file as Terse Triple Language (nodes and relations)
4. Validate TTL file with **SHA**pes **C**onstraint **L**anguage (SHACL)
5. Upload to graph database

# ? For Querying RDF

## Library

- SPARQL

**Option 1: Statement in R Code**

```
sparql <-
  'SELECT  ?patientID ?treatment
   WHERE {
     ?patient :hasTreatment ?treatment .
   }'
foo <- rdf_query(rdf, sparql)
```

**Option 2: Source external .RQ file**

```
foo <- rdf_query(rdf,
                 read_file("C:/myKGProject/PatientTreatments.rq"))
```

# ? For Analyzing Graph Data

Similar to other data analysis workflows

1. Query result to R data frame
2. Analysis like any other R data frame

*"Analysis with benefits"*

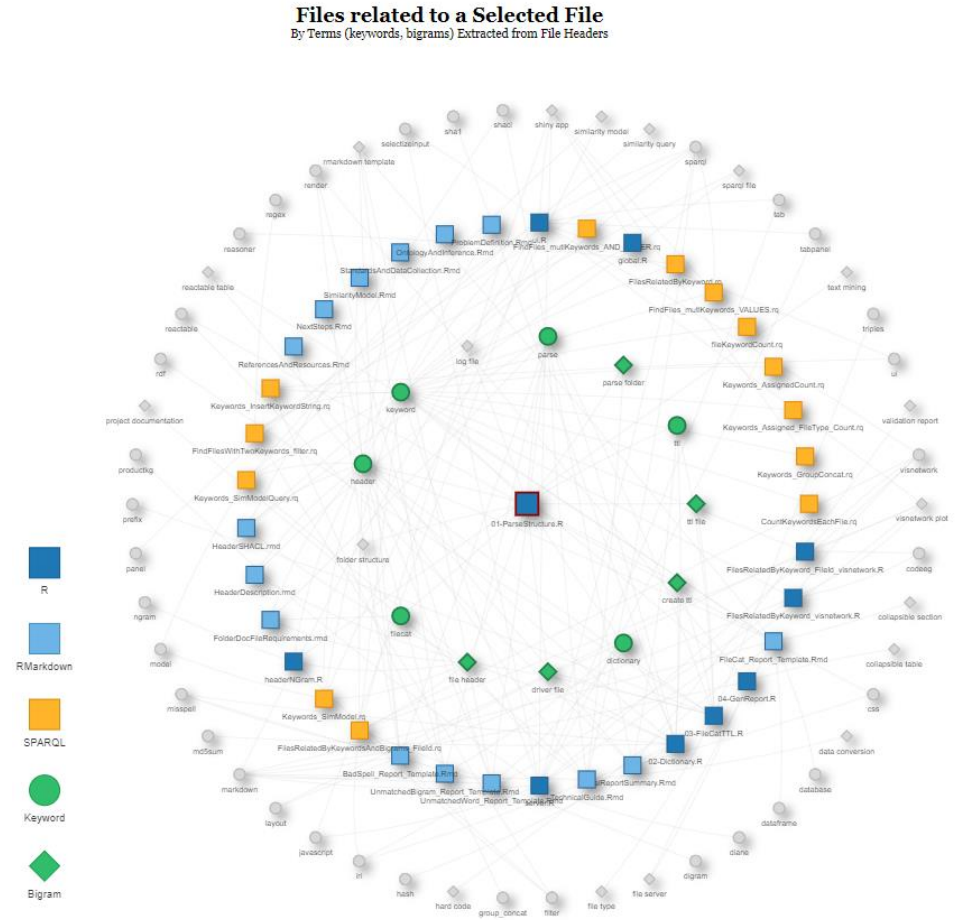
- *Values link to other values*
- *Semantic relationships*
- *Integral metadata*
- *Superior data quality*
  - *Unique identifiers*
  - *SHACL*

# ? For Visualizing RDF

## Libraries

- visNetwork
- ggplot2
- plotly
- DiagrammeR
- r2d3 (for D3JS)
- json, jsonlite

  - Export to JSON for consumption by D3JS and others.

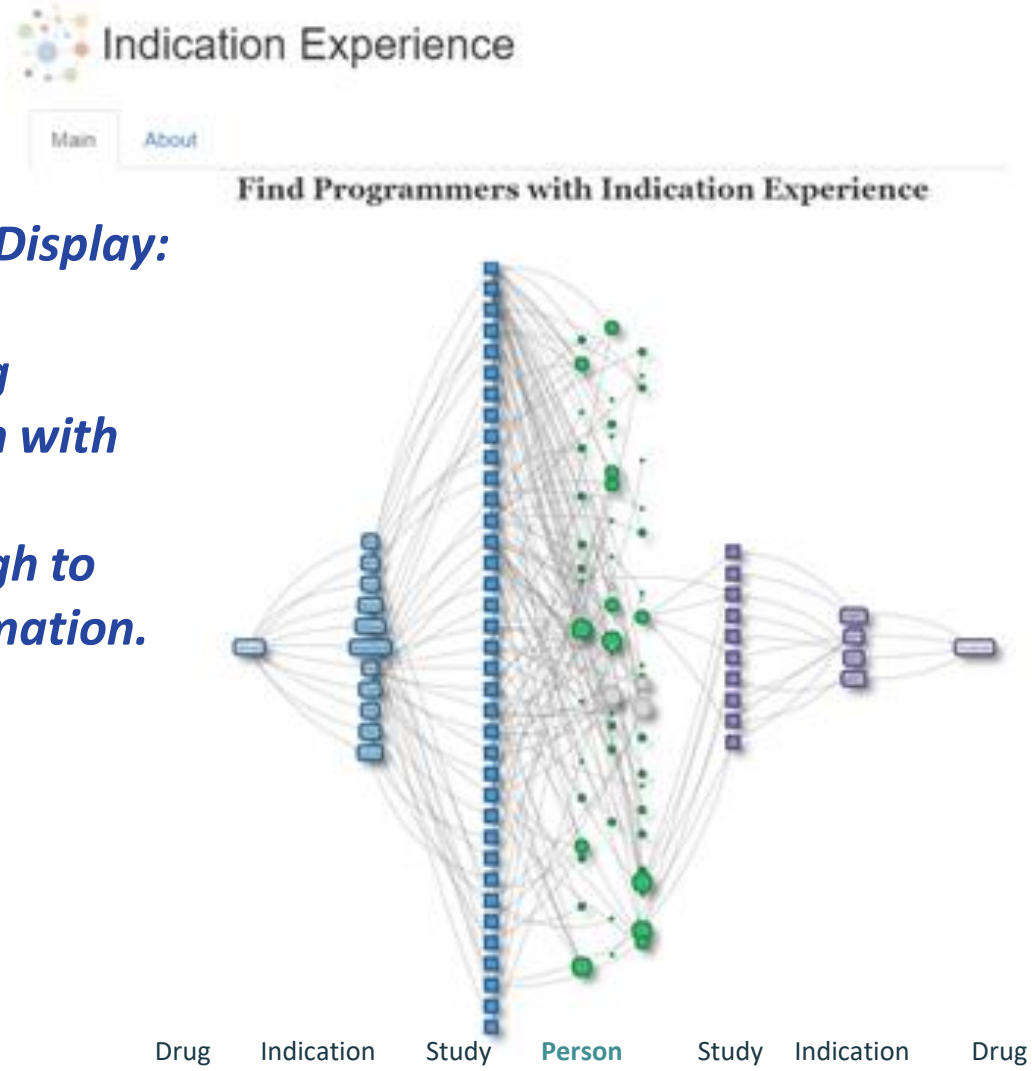


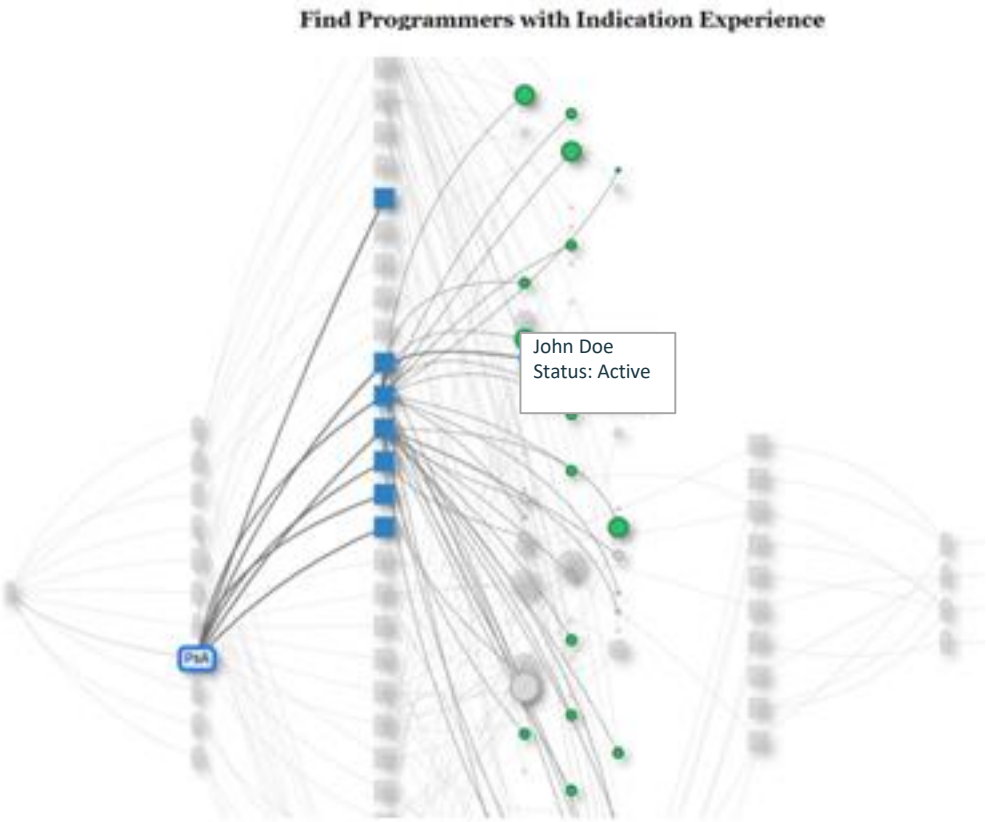

*"Visualizing Knowledge Graphs: Information Solution or Ball of Confusion"*
- Tim Williams, EUConnect21 (November)

# Interactive Displays

- Identify staff with the most programming experience for an indication
- Click-through to contact information

- ***Interactive Display:***
- ***Connection highlighting***
- ***Information with mouseover,***
- ***Click-through to more information.***



# Automation

"We shouldn't be fixated on the graph itself, but on enabling knowledge workflows. Knowledge Graph technologies allow the introduction of automation into information worker workflows, helping them save time on mundane information processing tasks." . "

- Mike Tung
"From Knowledge Graphs to Knowledge Workflows". 2021

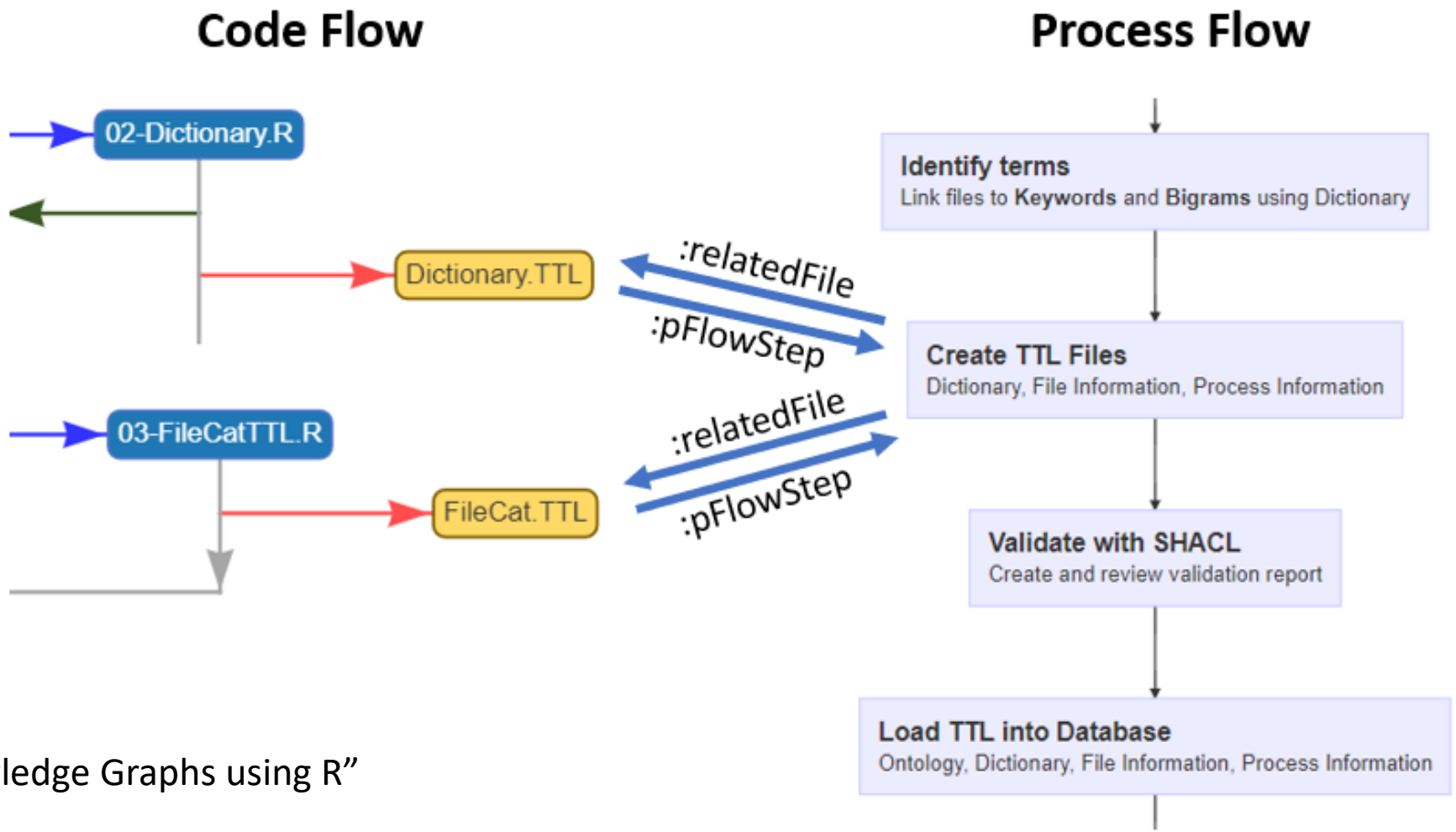
https://bit.ly/3c0XmH6

# Process and Code Flow

A Process Flow Example

A Code Flow Example

- Leads to connected flows
- Facilitates Automation




Excerpt from: "A Gateway to Knowledge Graphs using R"
– Tim Williams

# Data-Centric evolution

# Data as the Foundation

**Evolving**

- Discovery
- Safety
- Marketing

- Approvals
- Regulatory Response
- Publication…

Data

- mean
- p-value
- regression

- Bayesian
- Machine Learning
- Natural Language
- AI …

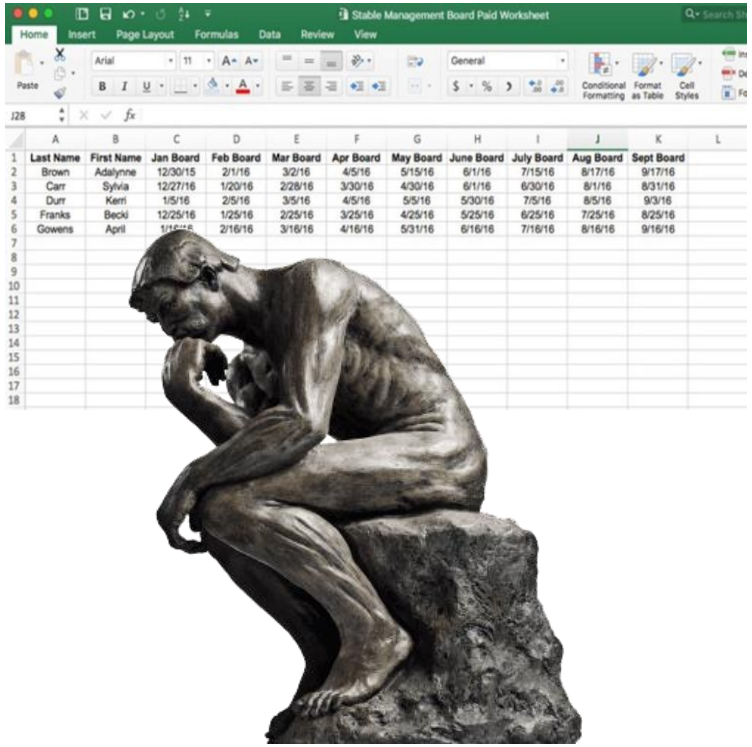
Data

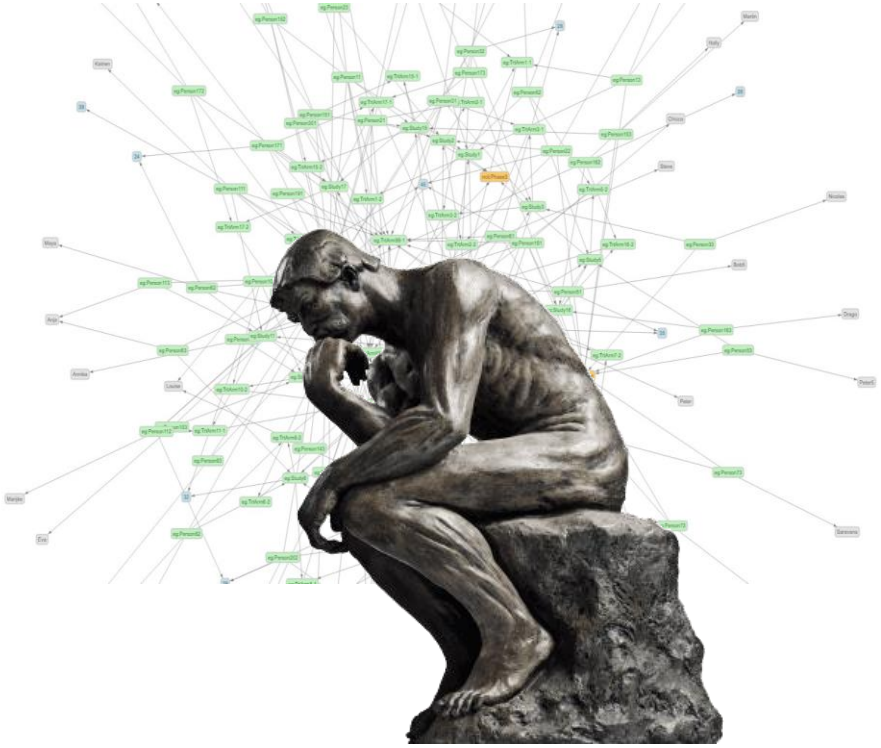
Data

- Legacy Standards
- New versions/ Standards

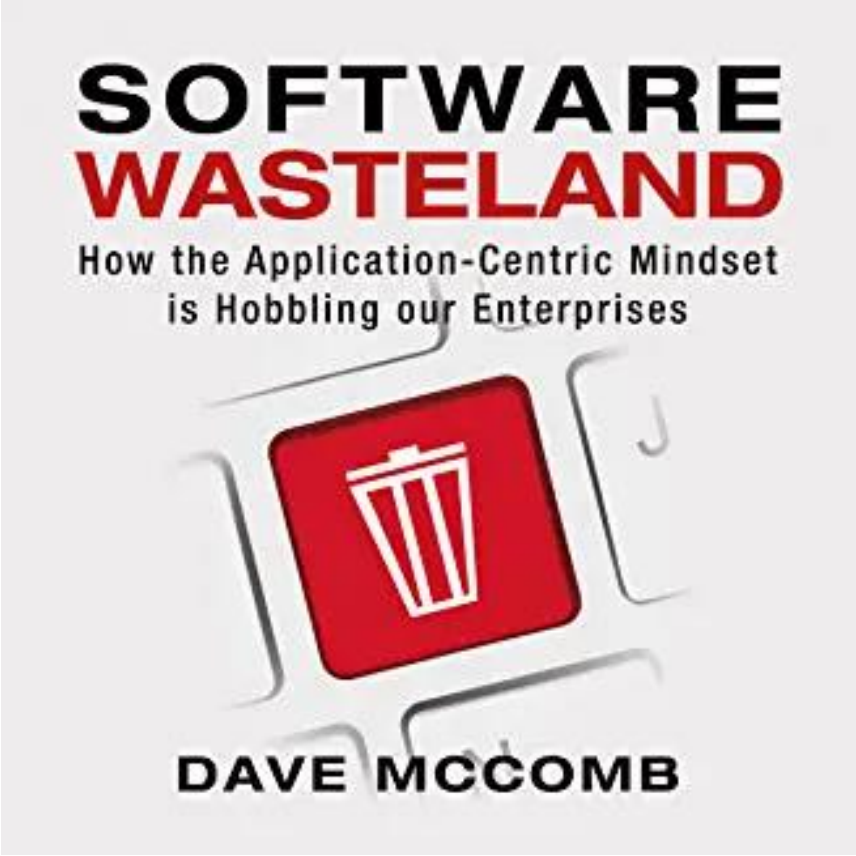
Data

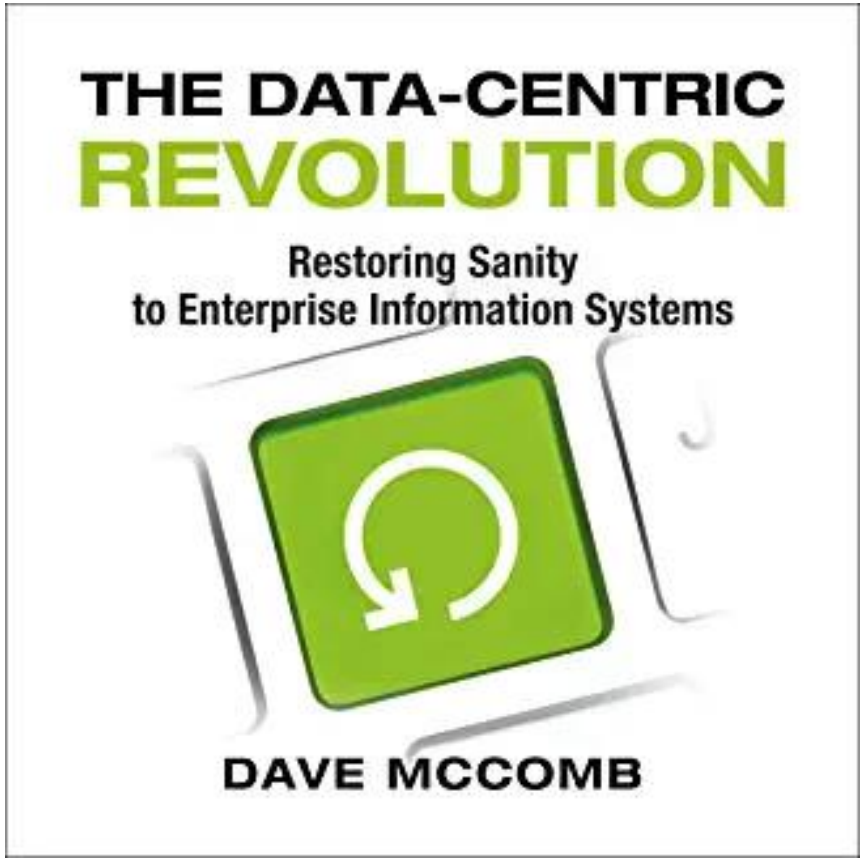
# Thinking about Data



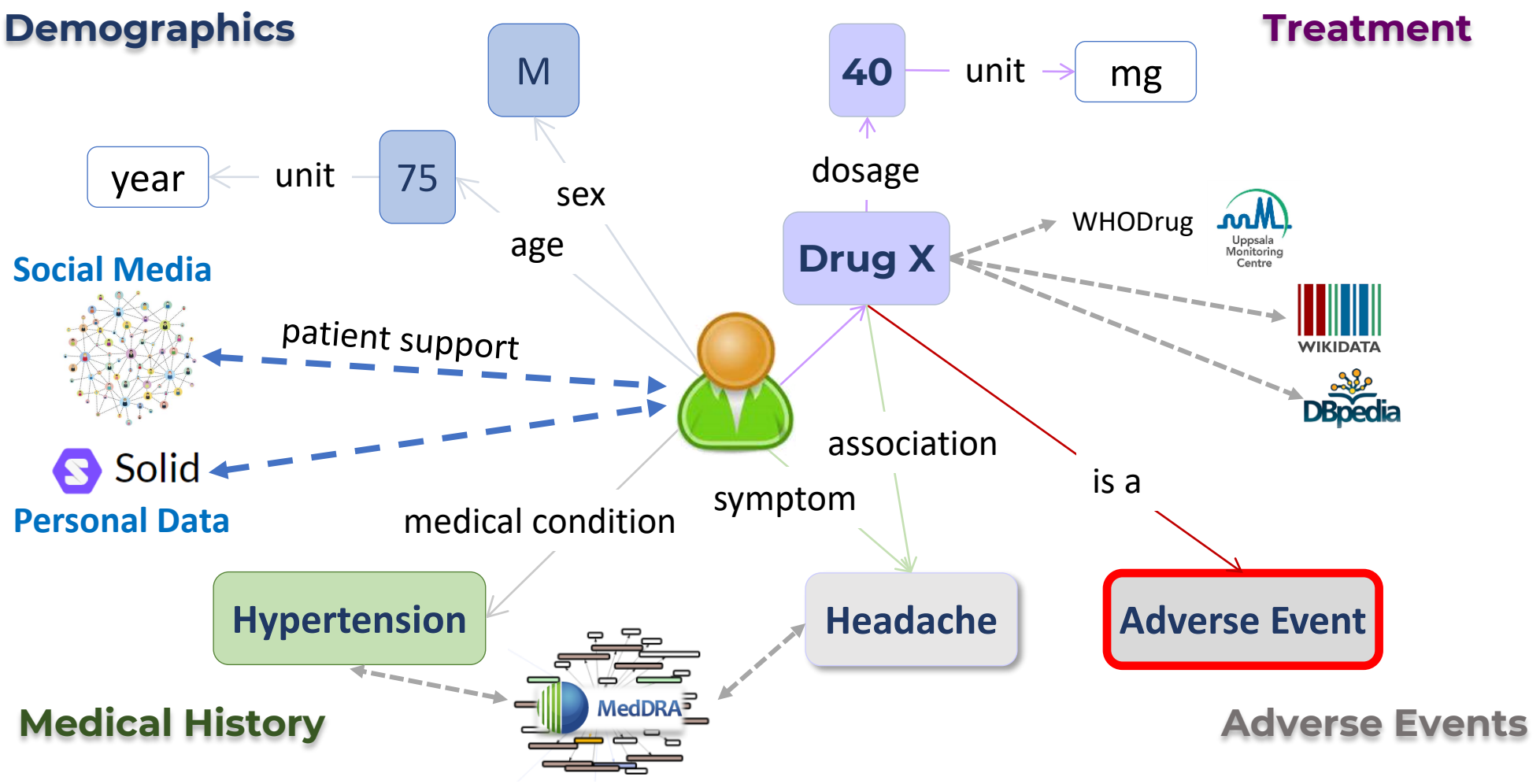



# Data-Centric



# Comprehensive Patient View



# The ?evolution is happening

- Real-time reasoning over formal semantics for self-driving cars.

Real Time Monitoring of Clinical Trials?


"Reasoning On The Road"

https://youtu.be/xIhWVCzQzRM

# Artificial Intelligence

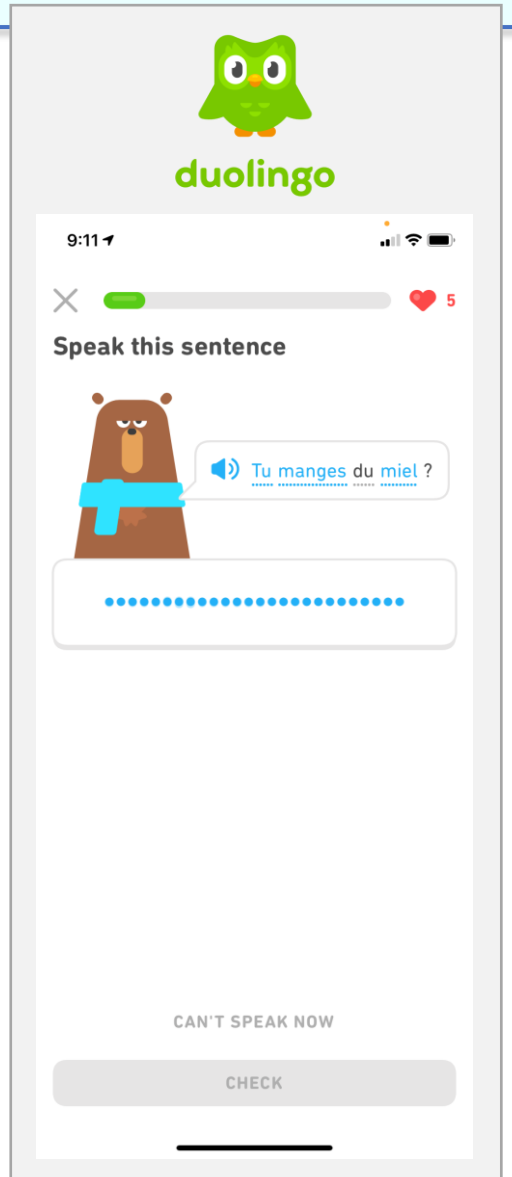
| First Wave | Second Wave | Third Wave |
|---|---|---|
| Handcrafted Knowledge | Statistical Learning | Contextual Adaptation |
| Domain experts enter logic and rules to the computer. | Systems Trained on Data | Machine-driven modeling |
| Still current and relevant. | On the rise, becoming mainstream | Just beginning |

- Schedulers
- Computer Chess

Taxation Software

duolingo

https://venturebeat.com/2020/08/18/how-duolingo-uses-ai-in-every-part-of-its-app/




"A DARPA Perspective on Artificial Intelligence" – John Launchbury
https://www.youtube.com/watch?v=N2L8AqkEDLs

# Machine Learning with ?

~ 100 ML Packages on CRAN!

r-pkg.org/ctv/MachineLearning

METACRAN Packages Authors Services About Search for packages

## Task view: Machine Learning & Statistical Learning

Last updated on 2021-09-21 by Torsten Hothorn

Several add-on packages implement ideas and methods developed at the borderline between computer science and statistics - this field of research is usually referred to as machine learning. The packages can be roughly structured into the following topics:

- *Neural Networks and Deep Learning*: Single-hidden-layer neural network are implemented in package nnet (shipped with base R). Package RSNNS offers an interface to the Stuttgart Neural Network Simulator (SNNS). Packages implementing deep learning flavours of neural networks include deepnet (feed-forward neural network, restricted Boltzmann machine, deep belief network, stacked autoencoders), RcppDL (denoising autoencoder, stacked denoising autoencoder, restricted Boltzmann machine, deep belief network) and h2o (feed-forward neural network, deep autoencoders). An interface to tensorflow is available in tensorflow. The torch package implements an interface to the libtorch library.
- *Recursive Partitioning*: Tree-structured models for regression, classification and survival analysis, following the ideas in the CART book, are implemented in rpart (shipped with base R) and tree. Package rpart is recommended for computing CART-like trees. A rich toolbox of partitioning algorithms is available in Weka, package RWeka provides an interface to this implementation, including the J4.8-variant of C4.5 and M5. The Cubist package fits rule-based models (similar to trees) with linear regression models in the terminal leaves, instance-based corrections and boosting. The C50 package can fit C5.0 classification trees, rule-based models, and boosted versions of these.
Two recursive partitioning algorithms with unbiased variable selection and statistical stopping criterion are implemented in package party and partykit. Function `ctree()` is based on non-parametric conditional inference procedures for testing independence between response and each input variable whereas `mob()` can be used to partition parametric models. Extensible tools for visualizing binary trees and node distributions of the response are available in package

### Packages

**ahaz** — 1.14
8 years ago by Anders Gorst-Rasmussen
Regularization for semiparametric additive hazards regression

**arules** — 1.6-8
4 months ago by Michael Hahsler
Mining Association Rules and Frequent Itemsets

**BART** — 2.9
9 months ago by Rodney Sparapani
Bayesian Additive Regression Trees

**bartMachine** — 1.2.6
9 months ago by Adam Kapelner
Bayesian Additive Regression Trees

**BayesTree** — 0.3-1.4
5 years ago by Robert McCulloch
Bayesian Additive Regression Trees

**BDgraph** — 2.64
5 months ago by Reza Mohammadi
Bayesian Structure Learning in Graphical Models using Birth-Death MCMC

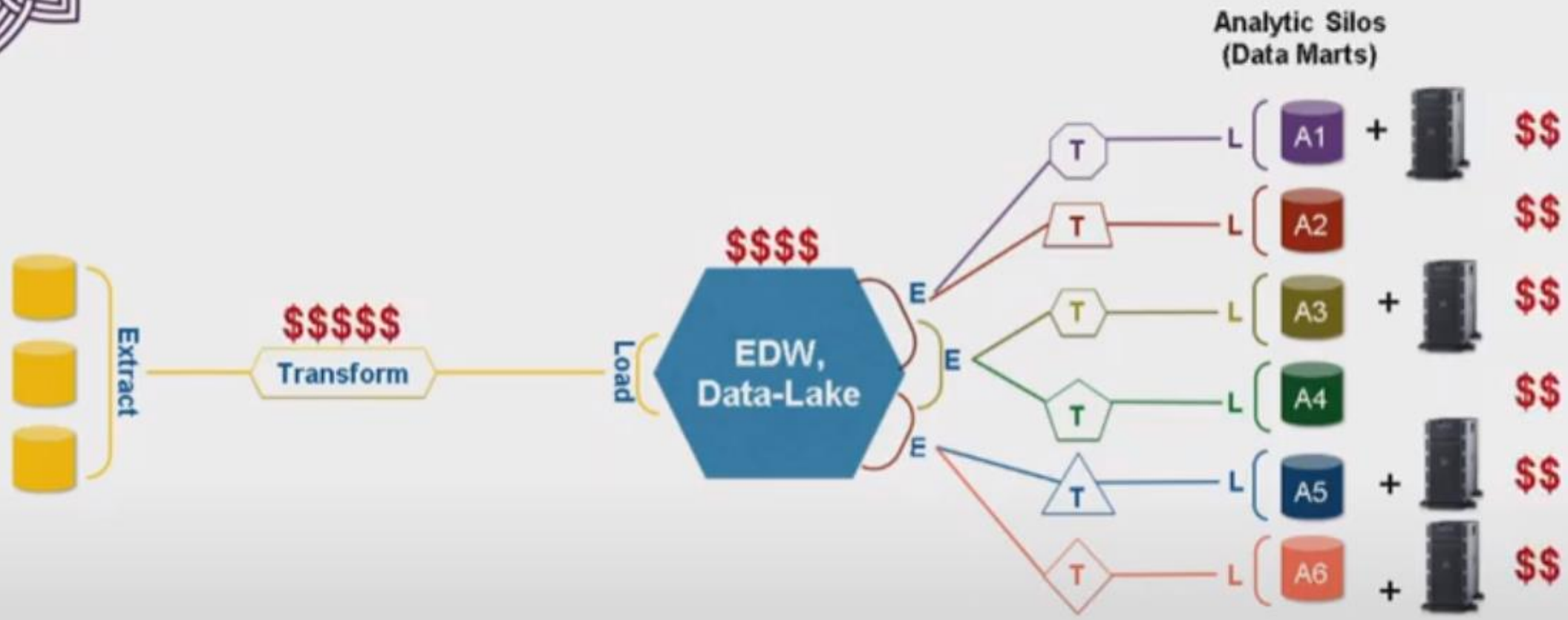
**biglasso** — 1.4.1
8 months ago by Chuyi Wang

**bmrm** — 4.1
2 years ago by Julien Prados
Bundle Methods for Regularized Risk

# Healthcare Example

The Montefiore Health System
Semantic Data Lake

# The State of the Art






Montefiore

Source: “The Chasm of a Million Analytics and How to Bridge It” – Dr. Parsa Mirhaji (2018) https://bit.ly/2OZXM72

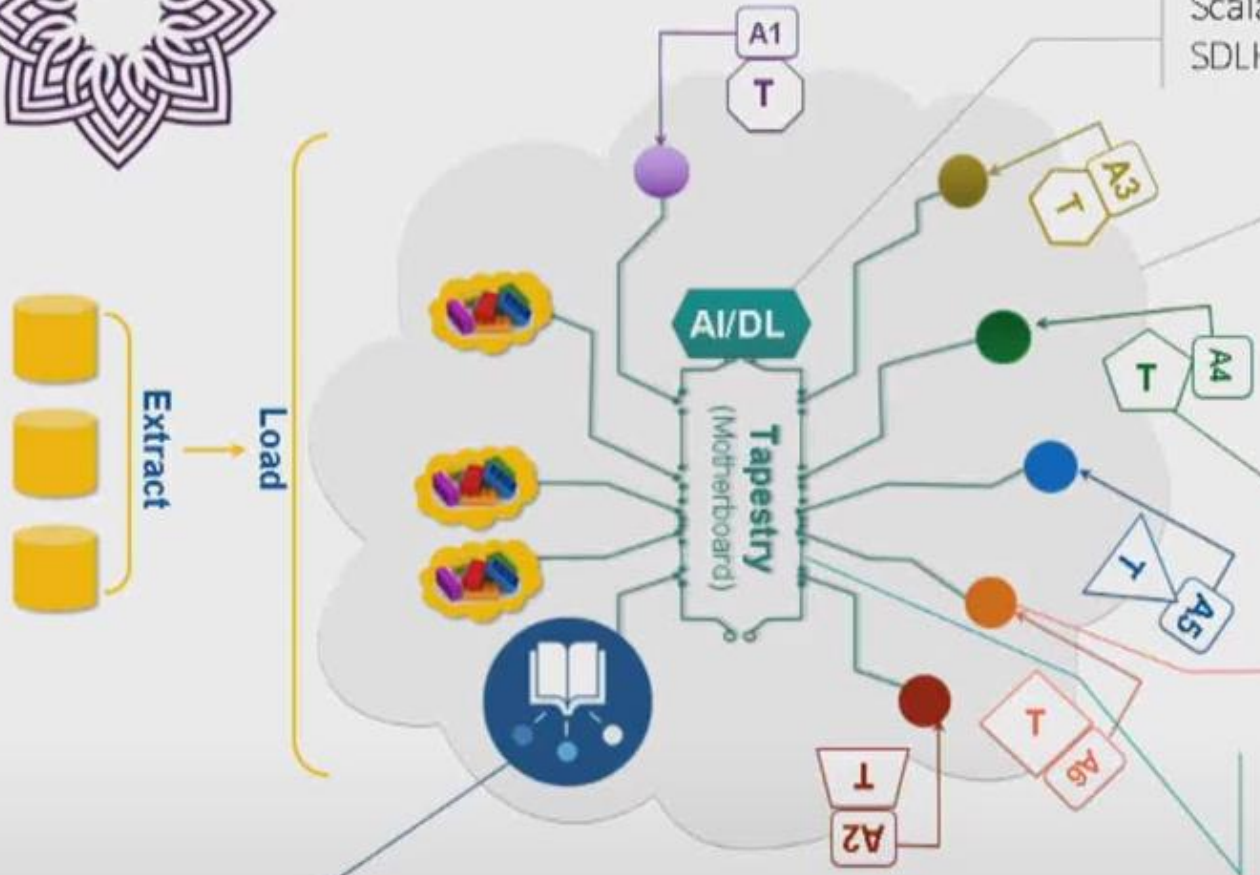
“The way we think about analytics, the way we think about problem solving, creates these bubbles we call analytic bubble making machines… This is not maintainable. ”

“Vendors exploit the situation to compartmentalize and fragment medicine event further by creating this little point solutions that have no idea…about patient's history…genetics…". ”

- Dr. Parsa Mirhaji
“The Chasm of a Million Analytics and How to Bridge It”
2019 Knowledge Graph Conference
https://bit.ly/2OZXM72

Source: “The Chasm of a Million Analytics and How to Bridge It” – Dr. Parsa Mirhaji (2018) https://bit.ly/2OZXM72

# Patient-centered Analytic Learning Machine




Scalable AI/DL/ML continuously learns, automates, and facilitates all SDLH processes

Extract and Load original data into schema-less graph along with rich metadata describing data at the source

Transform on demand for just-in-time analytics without moving data around

Return all analytic results and byproducts back to SDLH to avoid analytic silos

**Analytic Tapestry** integrates all data and analytics across all domains and applications for, audit, AI/ML, meta-analysis, and patient empowerment

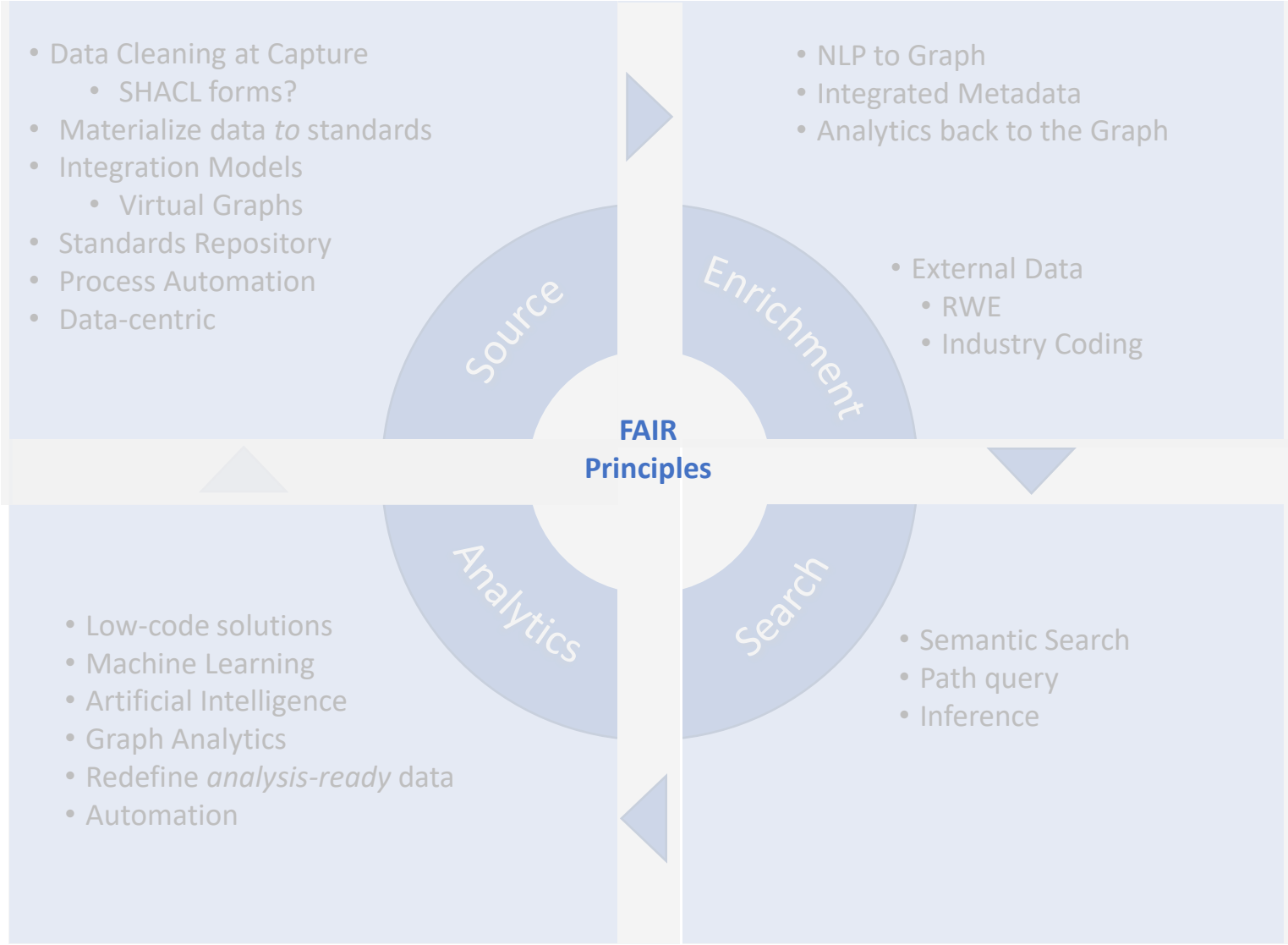
Extensible Biomedical Knowledgebase (patented) empowers contextualized learning and automation

- Key to data standardization, automation, and clinical interoperability as well as 'computer reasoning'



Montefiore

Source: "The Chasm of a Million Analytics and How to Bridge It" – Dr. Parsa Mirhaji (2018) https://bit.ly/2OZXM72




Adapted from:
https://www.slideshare.net/AlanMorrison/datacentric-business-transformation-using-knowledge-graphs

# Contact

## Tim Williams

Tim.Williams@ucb.com

LinkedDataTim@gmail.com

@NovasTaylor

https://www.linkedin.com/in/timpwilliams