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.
Knowledge Graphs as a Foundation for the Analytics Evolution
Why do we need a ?evolution?
Data Landscape

80% of enterprise data is unstructured*
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
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
Different Types of Graphs

Resource Description Framework (RDF)

Tim

WorksAt

UCB

Labeled Property Graph

Tim

WorksAt

UCB

firstName : Timothy
lastName : Williams

founded : 1928
headquarters : Brussels

startDate : June 2008
office : Raleigh
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, Interoperable, Reproducible)
Working with Knowledge Graphs
### 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

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

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

### Mapping Languages
- RML
- R2RML
  - SMS

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

### 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

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

### Miscellaneous
- LDFlex
- LDN (Linked Data Notifications)
https://arrows.app/
The "For Creating RDF"

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

The "Analytics" in the "Analytics "evolution""
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
```r
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 SHApes Constraint Language (SHACL)
5. Upload to graph database
Option 1: Statement in R Code

```r
sparql <-
  'SELECT ?patientID ?treatment
   WHERE {
   }
'
foo <- rdf_query(rdf, sparql)
```

Option 2: Source external .RQ file

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

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

- Identify staff with the most programming experience for an indication
- Click-through to contact information
“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
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 Revolution
Evolving Data as the Foundation

- Discovery
- Safety
- Marketing

- Approvals
- Regulatory Response
- Publication...

- mean
- p-value
- regression

- Bayesian
- Machine Learning
- Natural Language
- AI ...

- Legacy Standards
- New versions/ Standards

Data

Data

Data

Data
Thinking about Data
Comprehensive Patient View

Demographics
- sex: M
- age: 75
- unit: year

Social Media
- patient support

Medical History
- Hypertension

Drug X
- dosage: 40 mg

Adverse Event
- WHODrug

Adverse Events
- is a medical condition

Personal Data
- Solid

Association
- symptom
- trend

Medical condition
- MedDRA

Comprehensive View

Treatment
- WHODrug

Adverse Events
- WHODrug
• Real-time reasoning over formal semantics for self-driving cars.

Real Time Monitoring of Clinical Trials?

“The Reasoning On The Road”

https://youtu.be/xIhWVCzQzRM
<table>
<thead>
<tr>
<th>First Wave</th>
<th>Second Wave</th>
<th>Third Wave</th>
</tr>
</thead>
<tbody>
<tr>
<td>Handcrafted Knowledge</td>
<td>Statistical Learning</td>
<td>Contextual Adaptation</td>
</tr>
<tr>
<td>Domain experts enter logic and rules to the computer.</td>
<td>Systems Trained on Data</td>
<td>Machine-driven modeling</td>
</tr>
<tr>
<td>Still current and relevant.</td>
<td>On the rise, becoming mainstream</td>
<td>Just beginning</td>
</tr>
</tbody>
</table>

- Schedulers
- Computer Chess

**Statistical Learning**

- Taxation Software
  - [Link](https://venturebeat.com/2020/08/18/how-duolingo-uses-ai-in-every-part-of-its-app/)

**Contextual Adaptation**

**“A DARPA Perspective on Artificial Intelligence” – John Launchbury**

[Link](https://www.youtube.com/watch?v=N2L8AqkEDLs)
Machine Learning with Machine Learning & Statistical Learning

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 RNNs offers an interface to the Stuttgart Neural Network Simulator (SNNs). Packages implementing deep learning flavours of neural networks include deepnet (deep, forward neural network, restricted Boltzmann machine, deep belief network, stacked autoencoders), toRNN (denoising autoencoder, stacked denoising autoencoder, restricted Boltzmann machine, deep belief network) and lda (feed-forward neural network, deep autoencoders). An interface to Tensorflow is available in tensorflow. The tfrch package implements an interface to the Rinterface 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 part is recommended for computing CART-like trees. A rich toolbox of partitioning algorithms is available in party, package Rpart 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 identical stopping criterion are implemented in package party and partial. Function chen() is based on non-parametric conditional inference procedures for testing independence between response and each input variable whereas walk() can be used to partition parametric models. Extensible tools for visualizing binary trees and node distributions of the response are available in package partykit.

~ 100 ML Packages on CRAN!
The Montefiore Health System Semantic Data Lake
“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


Patient-centered Analytic Learning Machine

- **Scalable AI/DL/ML** continuously learns, automates, and facilitates all SDLH processes.
- **Extract and Load** original data into a 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’.

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

Tim Williams

Tim.Williams@ucb.com
LinkedInDataTim@gmail.com
@NovasTaylor
https://www.linkedin.com/in/timpwilliams