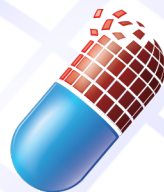


Semantic Information Integration within the Healthcare Sector

David Booth, Ph.D.
SemTech DC 2011*

**Substituting for Jürgen Angele*



PanGenX

Two brief examples of semantic information integration in health care

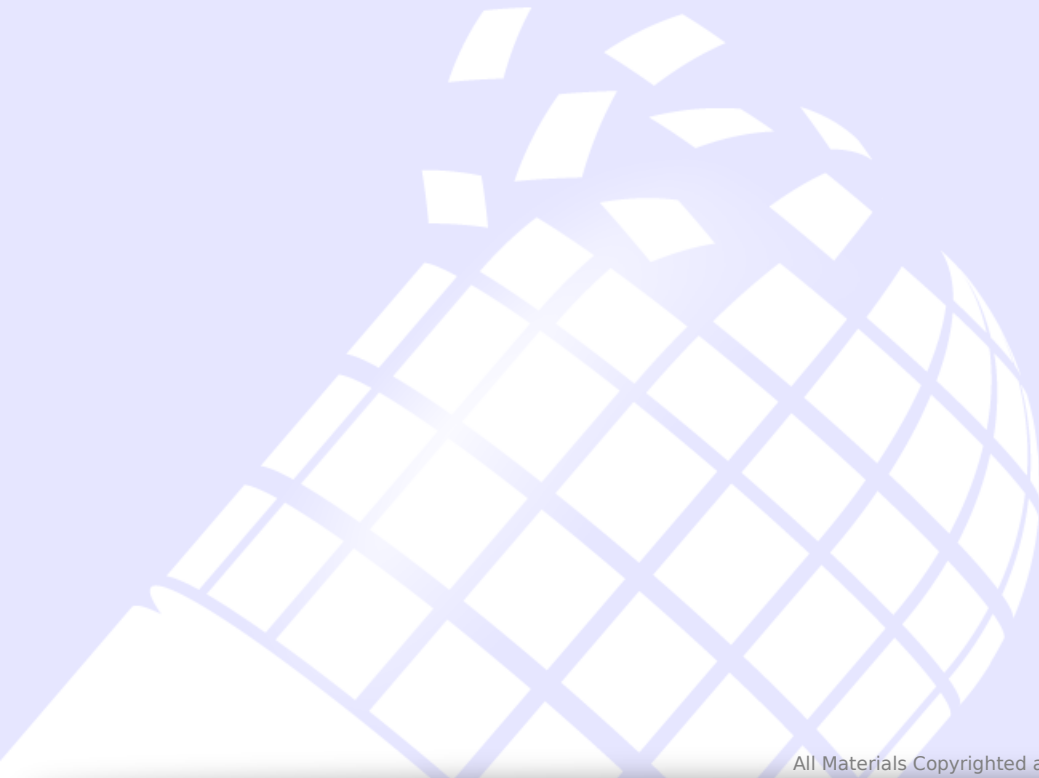


- Cleveland Clinic
 - Outcomes research and reporting
- PanGenX
 - Enabling personalized medicine



Common drivers

- Improve cost, effectiveness, care and safety



Example 1: Cleveland Clinic

Outcomes research and reporting

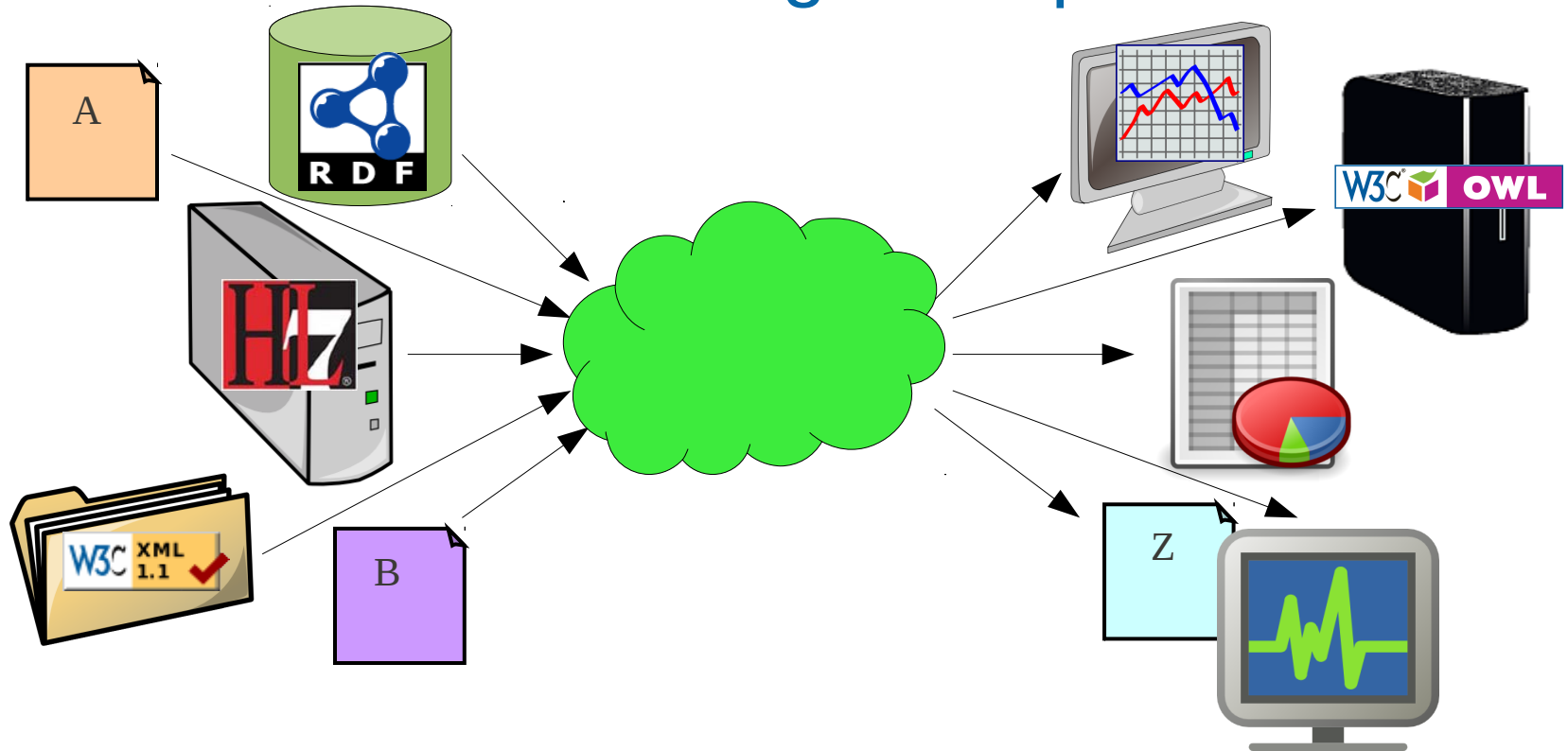
- **Rated #1 in heart care for the past 17 years**
 - *U.S. News and World Report*
- **Heart and Vascular Institute**
- **Maintains a registry of ~200,000 patient records**
- **Semantic web technology used since 2008 in generating:**
 - Research data for ~130 journal articles per year
 - Internal and external reports on quality and volume of care

Current Electronic Health Data

Data Sources:

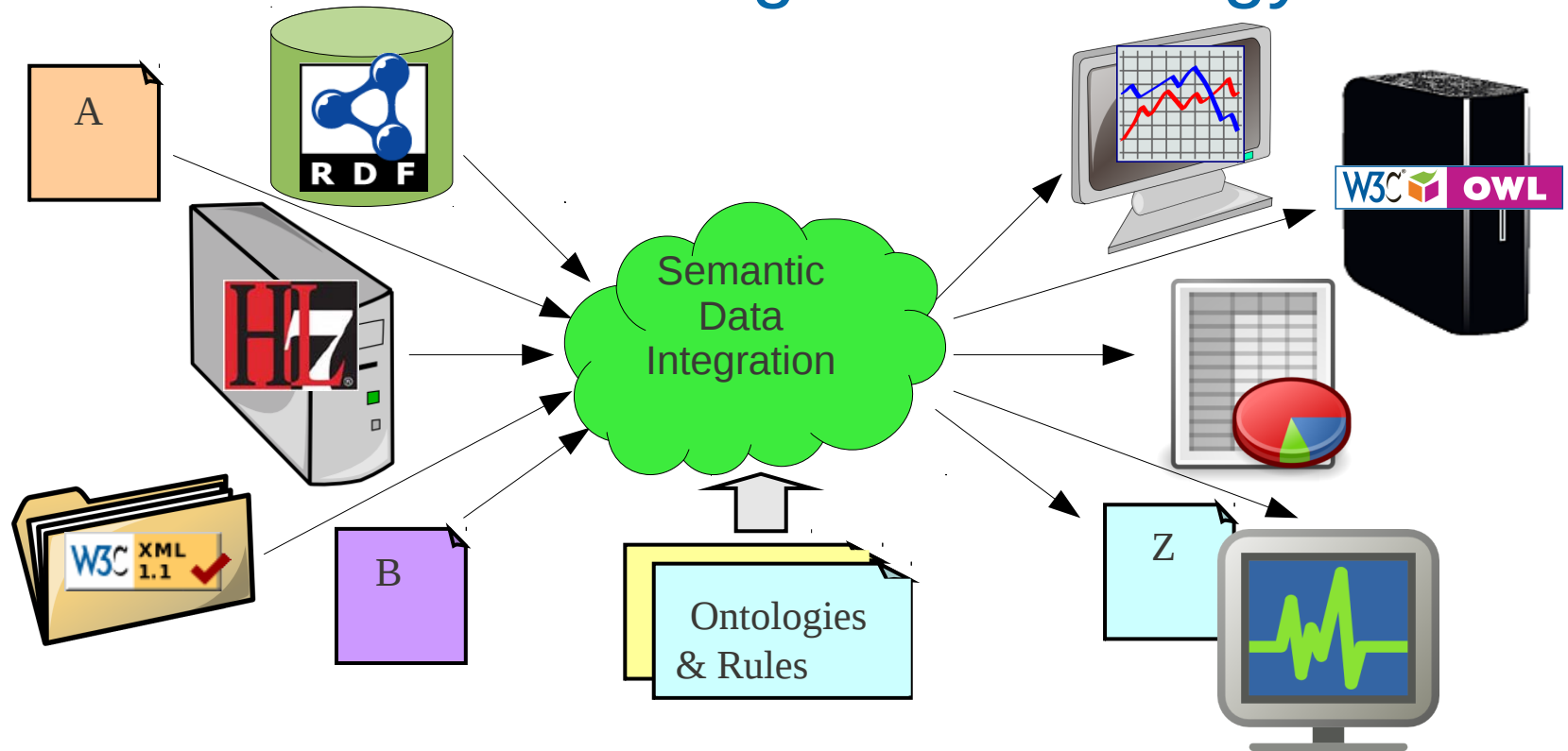
- **Enterprise EMRs**
- **Lab databases**
- **Billing/Claims databases**
- **Research data registries**
- **Reporting databases**

Information integration problem



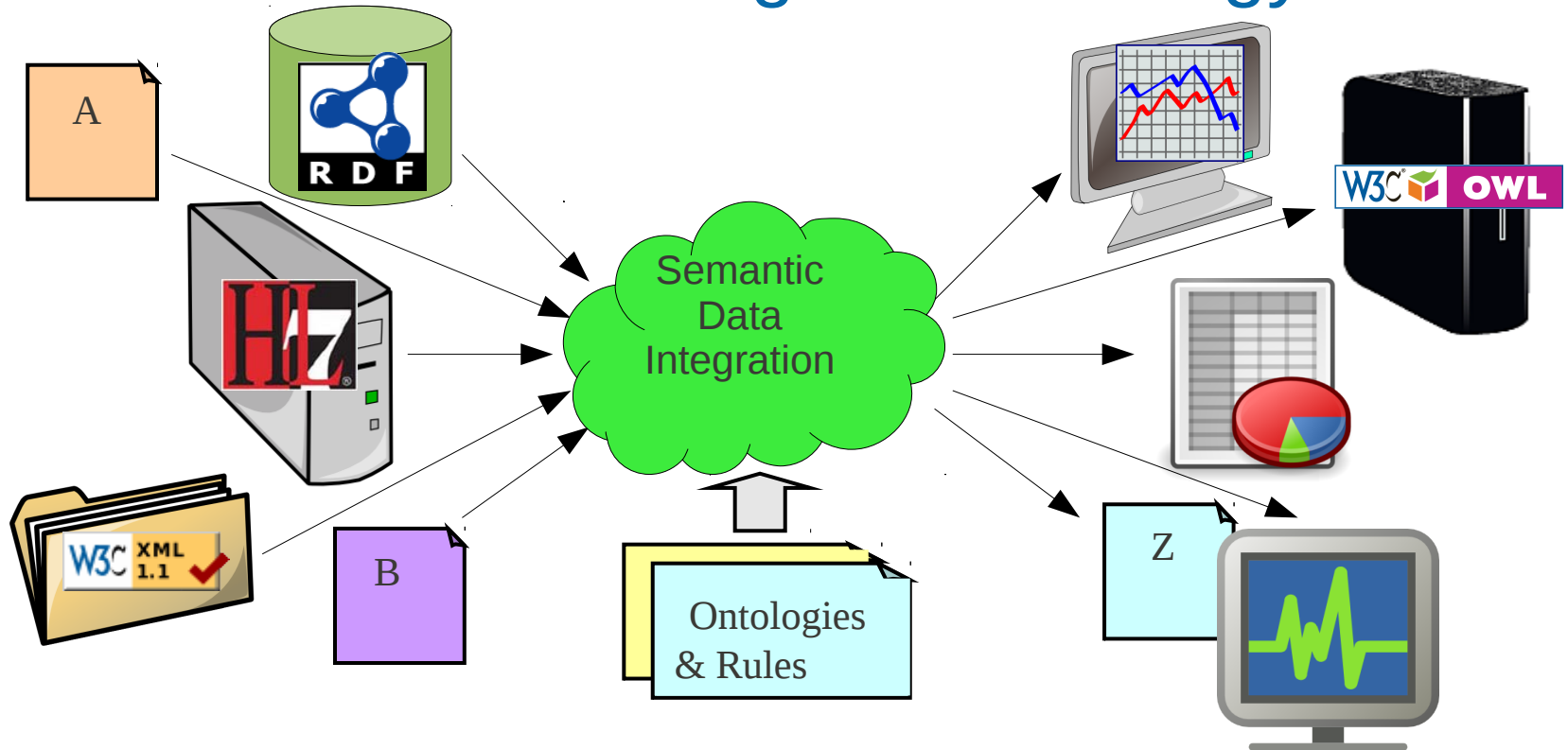
- Many data sources and applications
- Many technologies and protocols
- Goal: Each application wants the illusion of a single, unified data source

Semantic integration strategy



1. Data production pipeline
2. Use RDF in the middle; Convert to/from RDF at the edges
3. Use ontologies and rules for semantic transformations

Semantic integration strategy

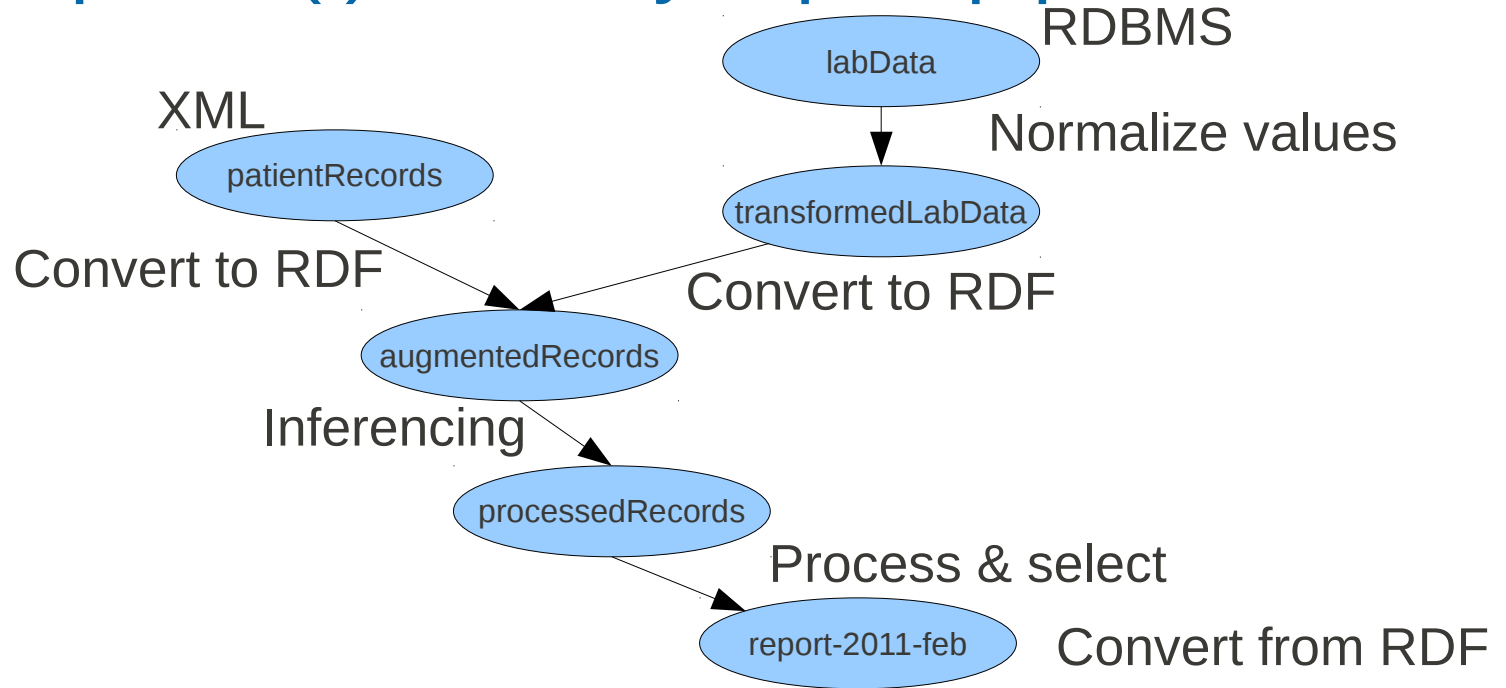


1. Data production pipeline

2. Use RDF in the middle; Convert to/from RDF at the edges

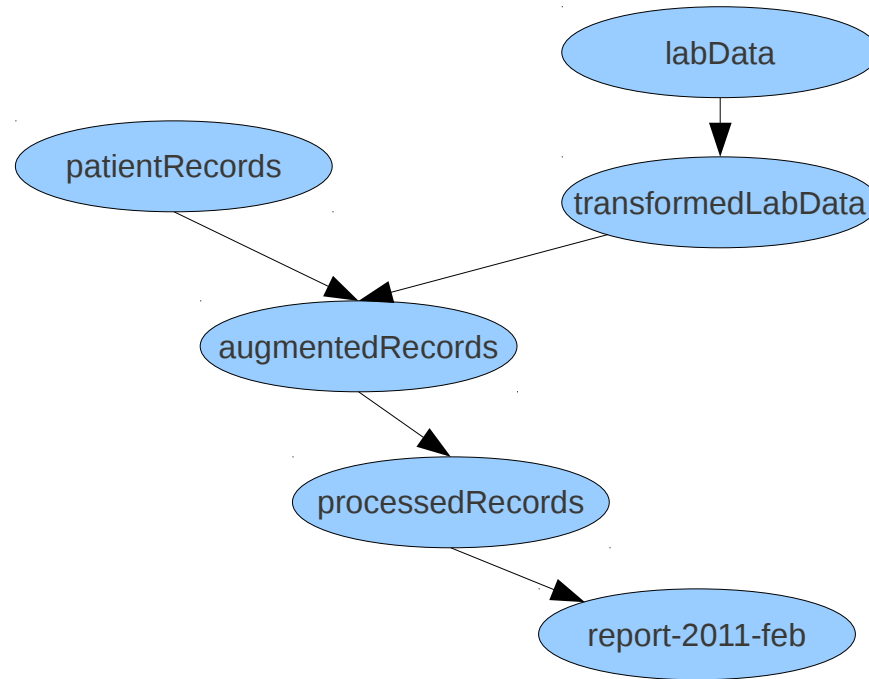
3. Use ontologies and rules for semantic transformations

Simplified(!) monthly report pipeline



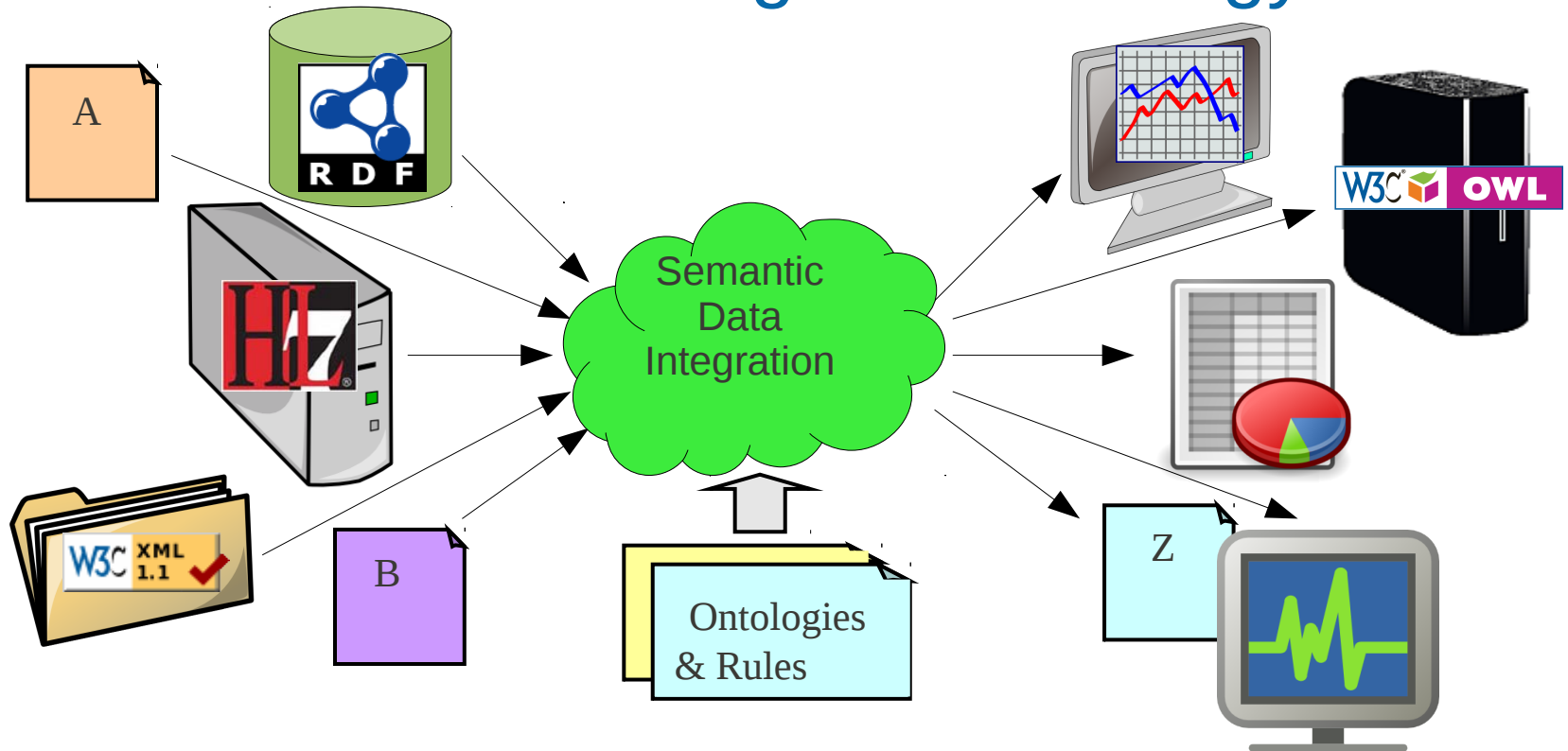
- Pipeline of several data sources and data production stages
 - Many technologies used – built ad hoc
 - E.g., 4Suite, RDFLib, FuXi, MySQL, Cyc, Oracle 11g Spatial, ViaDuct
-

Pipeline lessons learned



- **Pipeline is necessary, but:**
 - Ad hoc becomes complex & hard to maintain
 - **Need better, simpler pipeline mechanisms**
 - **Need better mechanisms for efficient, automated data update**
-

Semantic integration strategy



1. Data production pipeline

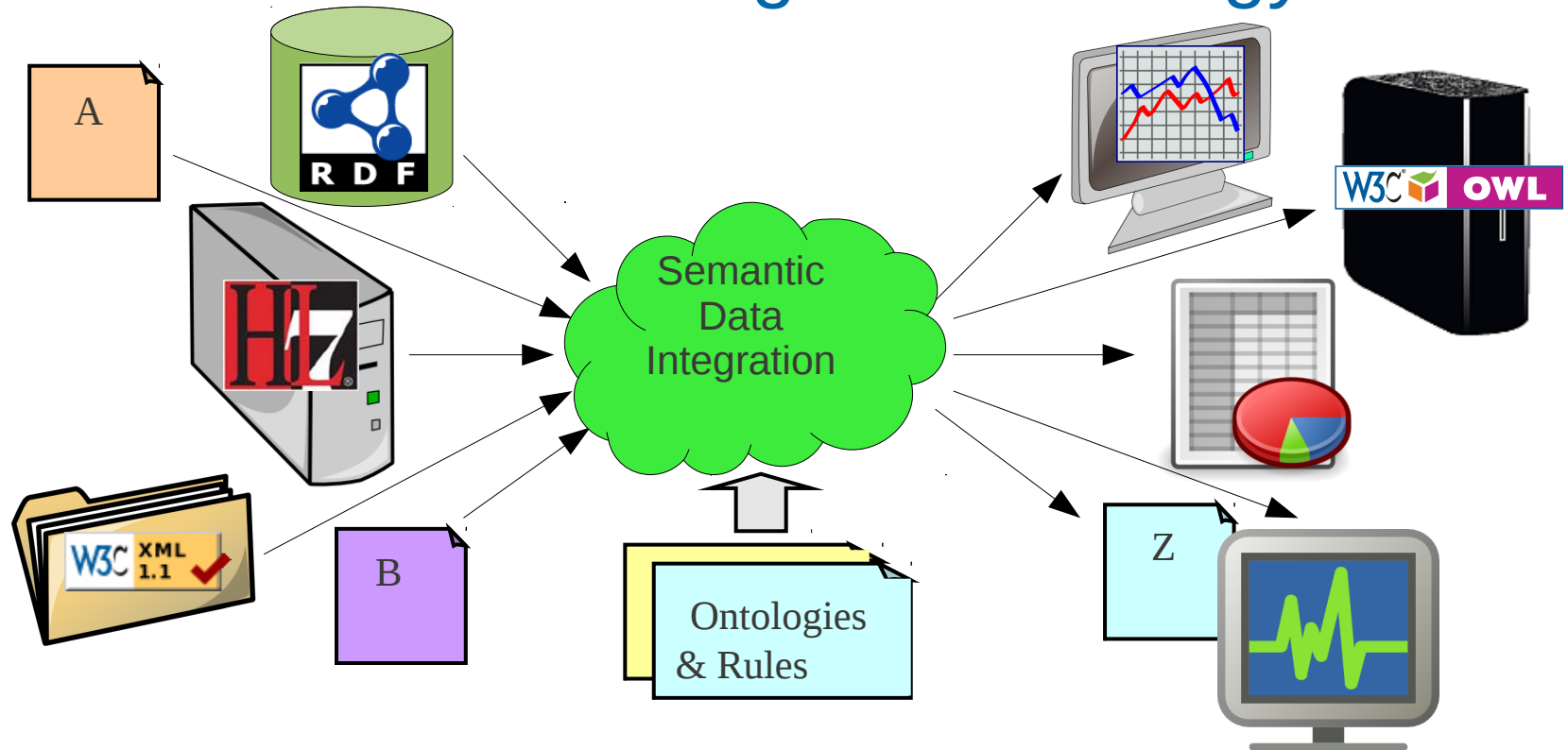
2. Use RDF in the middle; Convert to/from RDF at the edges

3. Use ontologies and rules for semantic transformations

RDF in the middle

- **Allows disparate data to be more easily connected**
 - **Facilitates inference**
 - Some technologies used: RDFLib, N3 rules, FuXi, Cyc
 - **Tools are available for converting to/from RDF at the edges**
 - Some technologies we used: 4Suite, ViaDuct
 - Many others now available, e.g., R2RML for relational mapping
 - **Lessons learned:**
 - Good strategy!
 - Tools are more mature now
-

Semantic integration strategy



1. Data production pipeline
2. Use RDF in the middle; Convert to/from RDF at the edges
3. Use ontologies and rules for semantic transformations

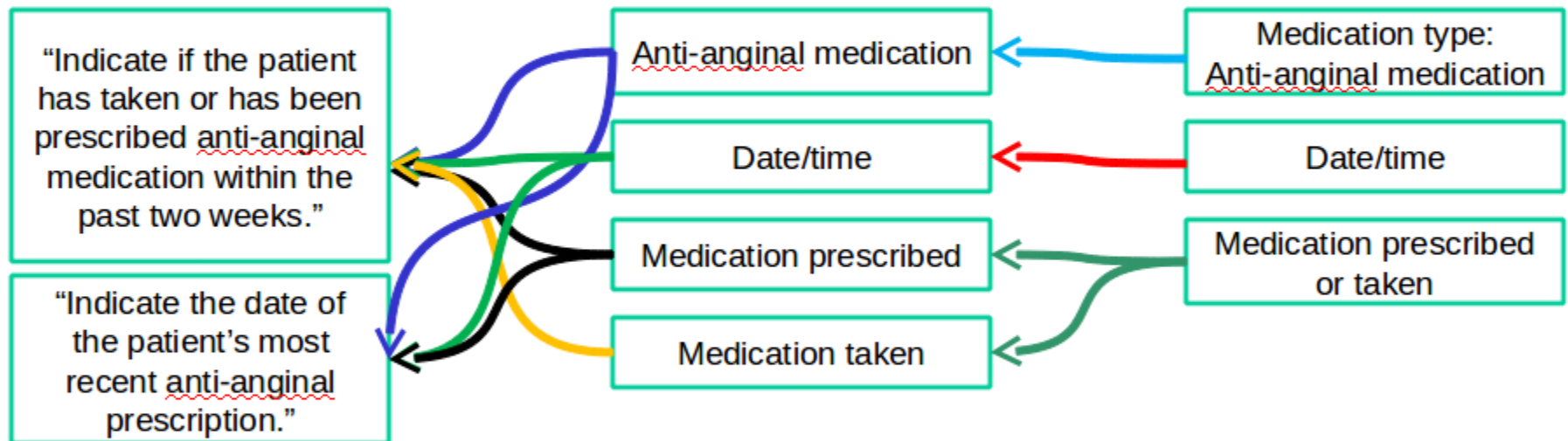
Ontologies

- **Several ontologies used**
 - Cyc, SNOMED, patient record ontology
- **Some used only for certain terms**
- **Others used for inference (e.g., Cyc)**
- **One strategy was to define a hub ontology based on *core data elements* . . .**

Use of Core Data Elements

Disassembling and Reassembling data from Source to Destination

CathPCI v4.3 #5025 "Anti-Anginal Meds" Example



Question to Answer

Core Data Element(s)

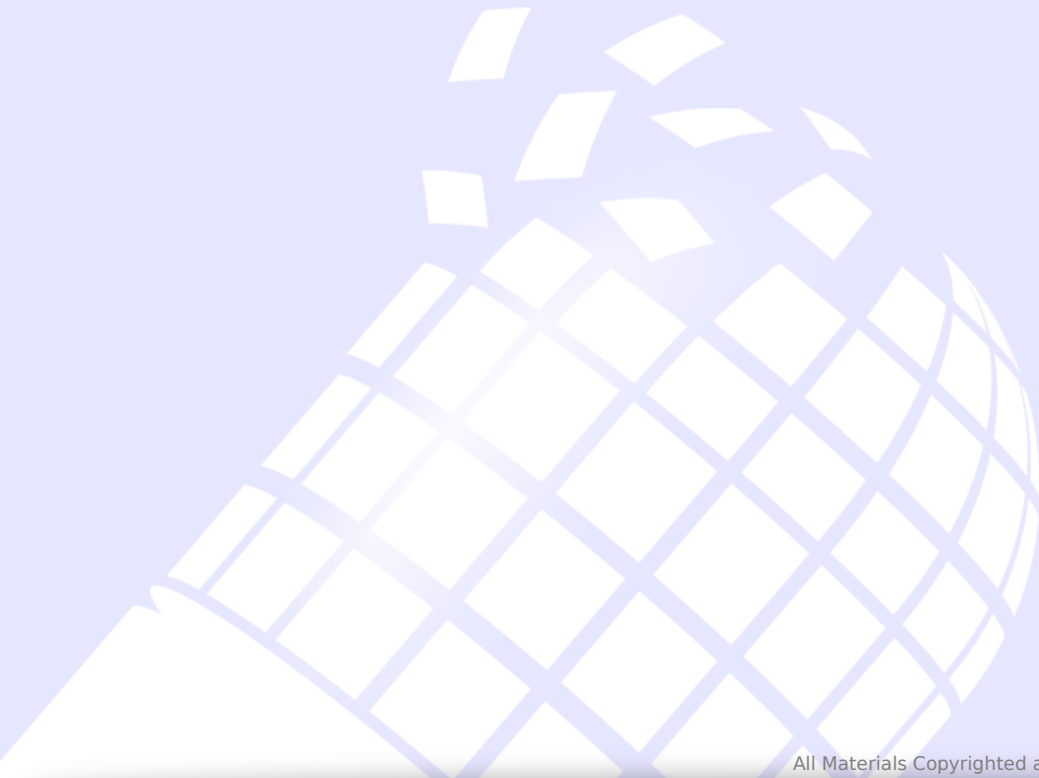
Source Information

Ontologies and rules lessons learned

- **Good strategy!**
- **The ontology will never be perfect . . . but that's okay**
 - You can use other presentation ontologies
- **Ontology versioning can be a challenge**
 - Best to avoid if possible

Example 2: PanGenX

Enabling personalized medicine



The Problem

Not every drug is right for every person

**Personalized Medicine provides the unique patient
with the correct drug at the correct dose**

Toxicity Challenge

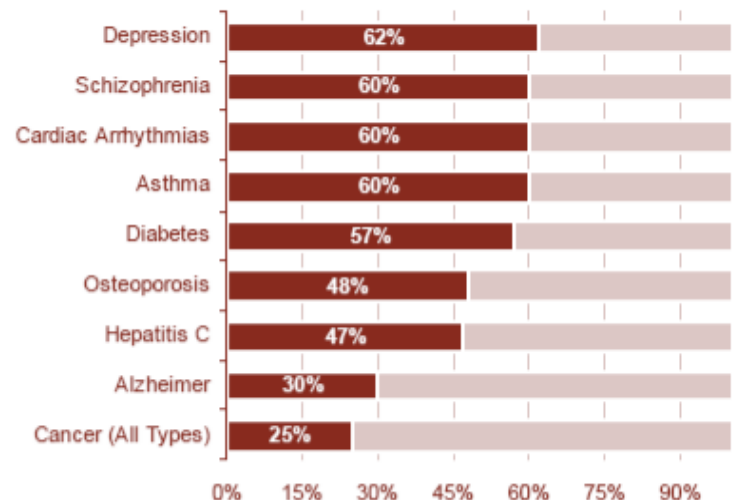
Adverse Drug Reactions :

- 2.2M people affected
- 4th leading cause of death
 - Responsible for 106,000 deaths every year
- Annual costs of \$177B



Efficacy Challenge

Therapeutic Areas | Efficacy Rate w/ Standard Treatment

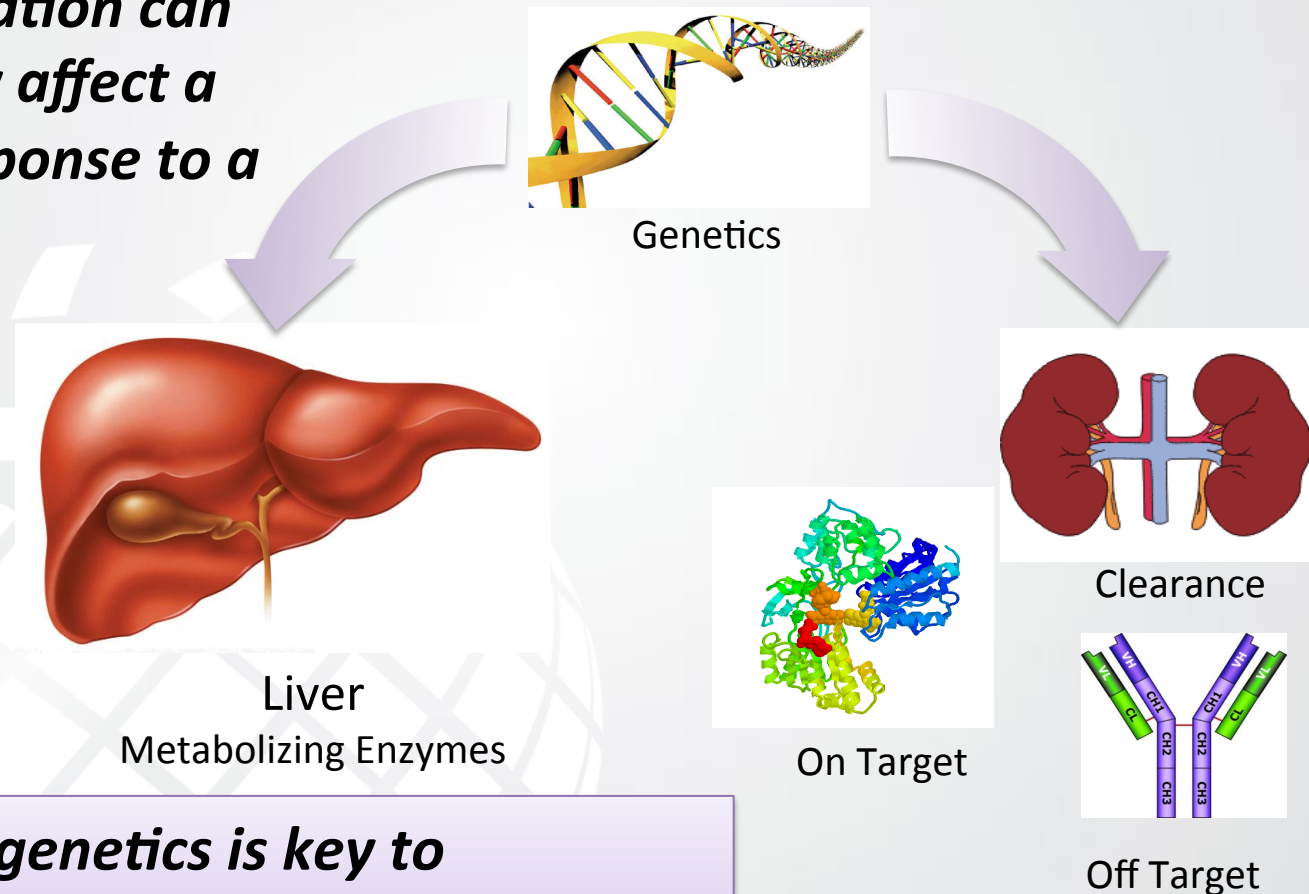


\$600B WW Rx spend, 50% with no efficacy-\$300B wasted?

Pharmacogenetics

Why is Pharmacogenetics relevant?

Genetic variation can dramatically affect a person's response to a drug



Pharmacogenetics is key to facilitating Personalized Medicine

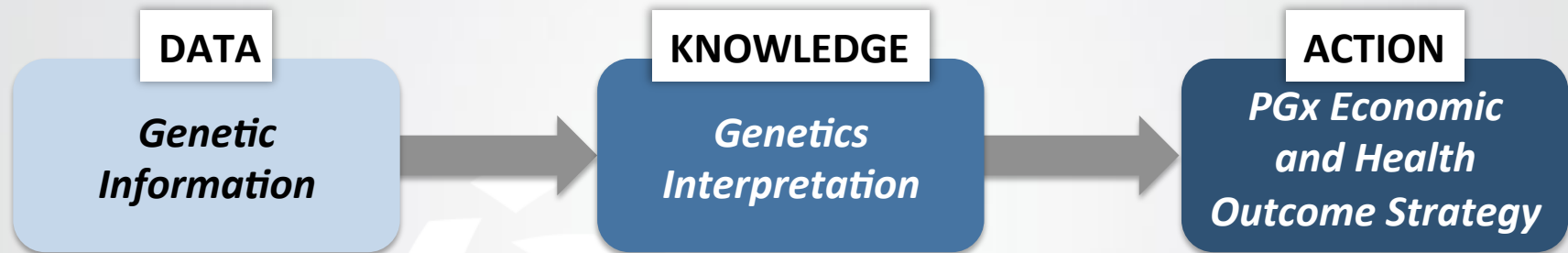
Data vs. Knowledge

Data alone is not enough to make informed decisions



*"We are close to having a **\$1,000 genome sequence**, but this may be accompanied by a **\$1,000,000 interpretation**"*

- Bruce Korf, president, American College of Medical Genetics



➤ The **Four C's** of transforming Data to Knowledge:

- **Comparison**: how does information about this situation compare to others?
- **Consequences**: what are the implications of this information?
- **Connections**: how does this data relate to other data?
- **Conversation**: what do experts think about this information?

Personalized Medicine and Big Data



May 2011 McKinsey report predicts that Big Data is the “next frontier for innovation, competition, and productivity”

For U.S. health care, the report is predicting \$300 billion per year in savings due to utilization of Big Data to drive the execution of strategies proposed by health care experts

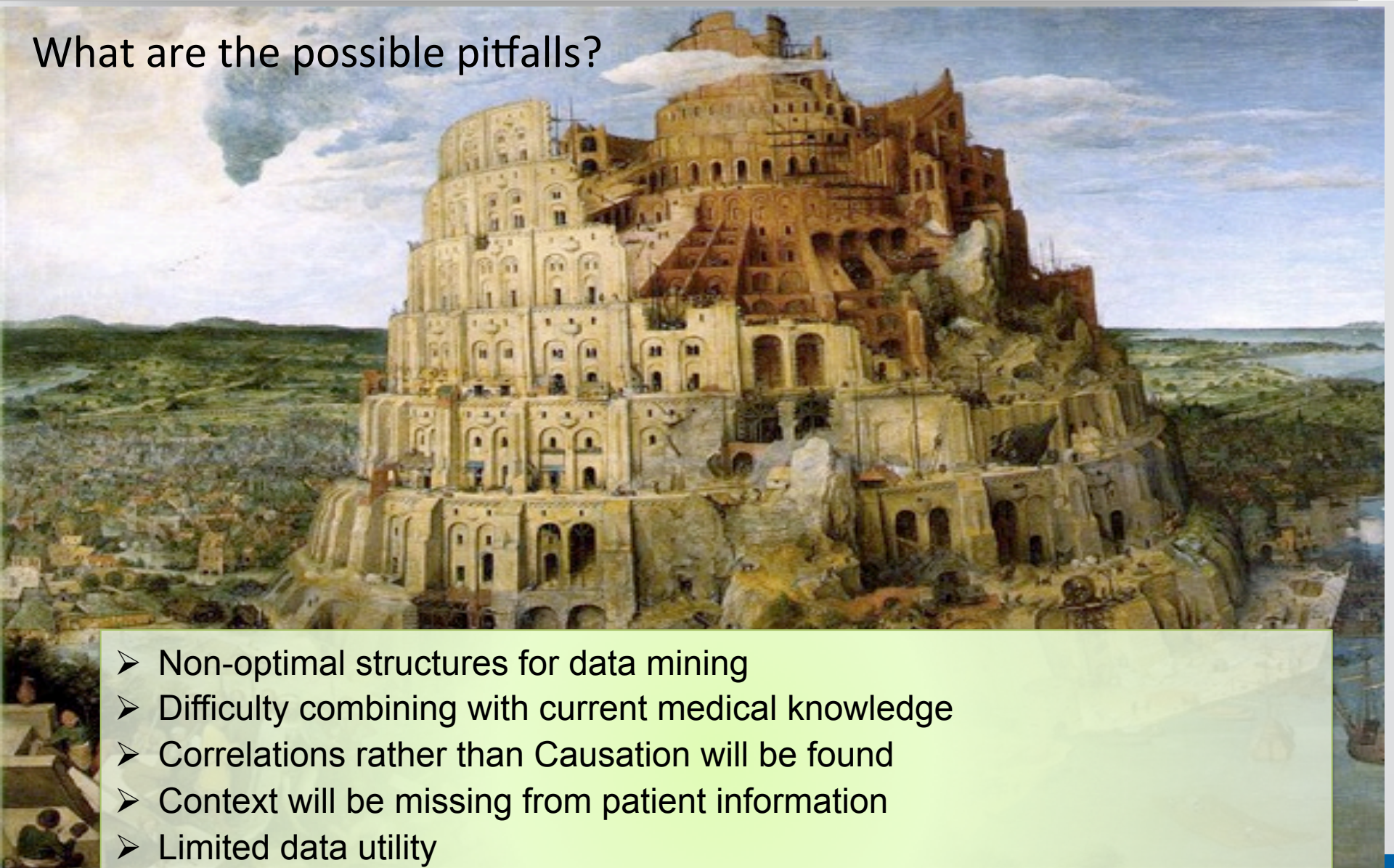
Big Data will become equal to labor and capital in its importance to production

What can “Big Data” promise?

- Identification of the most cost- and clinically-effective treatments to reduce over- and under-treatment.
- Better analyze disease patterns and clinical data to enable personalized medicine.
- New clinical decision-support systems to decrease errors by matching physician orders with best practices.
- Proactively identify beneficial lifestyle changes for certain patients.
- Design better clinical trials.
- Decrease time to bring new drugs to market.

Personalized Medicine and Big Data

What are the possible pitfalls?

- 
- A detailed painting of a massive, multi-tiered stone structure, resembling a giant's foot or a massive staircase, built into a hillside. The structure is composed of many levels of arches and windows, with a small figure of a person visible on one of the lower levels for scale. The background shows a landscape with a body of water and a distant city.
- Non-optimal structures for data mining
 - Difficulty combining with current medical knowledge
 - Correlations rather than Causation will be found
 - Context will be missing from patient information
 - Limited data utility

The PanGenX Solution

PanGenX-KB™ -- Knowledge-as-a-Service



Enabling Informed Decisions and Intelligent Actions

PanGenX KB:

- A proprietary, scalable knowledgebase, analytics engine, and decision-support tool
- Cloud accessible
- One knowledgebase suitable for many applications
- Customizable for each therapeutic area

1. Data production pipeline

- Combination of public and private datasets
- E.g., Genomic, phenotypic, drug, outcomes, etc.

2. Use RDF in the middle; Convert to/from RDF at the edges

- Good for integration, inference and context/provenance (with named graphs)

3. Use ontologies and rules for semantic transformations

- SPARQL is convenient as a rules language

- General strategy is good:
 1. Data production pipeline
 2. Use RDF in the middle; Convert to/from RDF at the edges
 3. Use ontologies and rules for semantic transformations
- Semantic web technology helps!
- Lots more (and better) tools available now than a few years ago

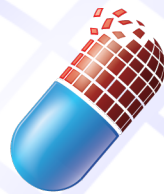
Contact Information

David Booth

david@dbooth.org

Eric Neumann

eneumann@pangenx.com



PanGenX