

Leveraging New Technologies and Big Data to Drive Analytics in Clinical Research

Michael Kowolenko, Ph.D.

Director, Institute of Next Generation IT Systems

Who we are

Solutions Based Organization that offers:

Software and systems prototyping

Technical development and commercialization

Data Sciences and Services

Pipeline of skilled, experienced students

What do we build?

- **Attributes**
 - Answer questions in real time
 - “Easy Button”
 - Diverse Data
 - Open Source
 - Infrastructure Sandbox

Key Concepts in Big Data

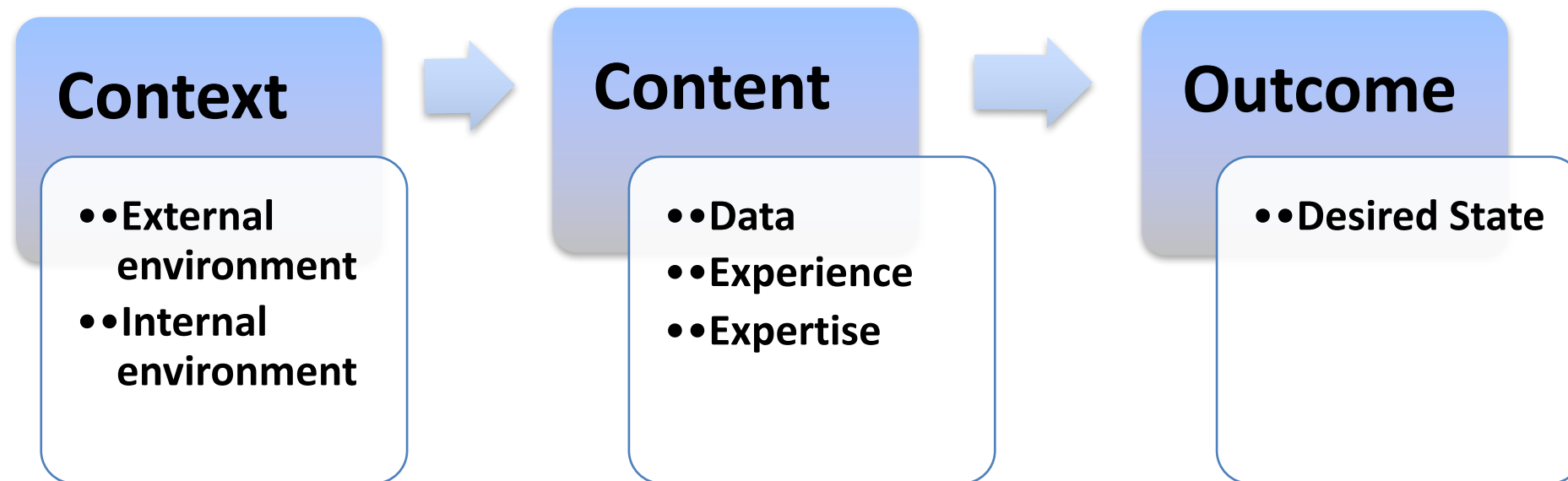
- **Operations vs Strategy**
- **Aggregation vs Isolation**
- **What vs.. Why**

Intelligence Augmentation

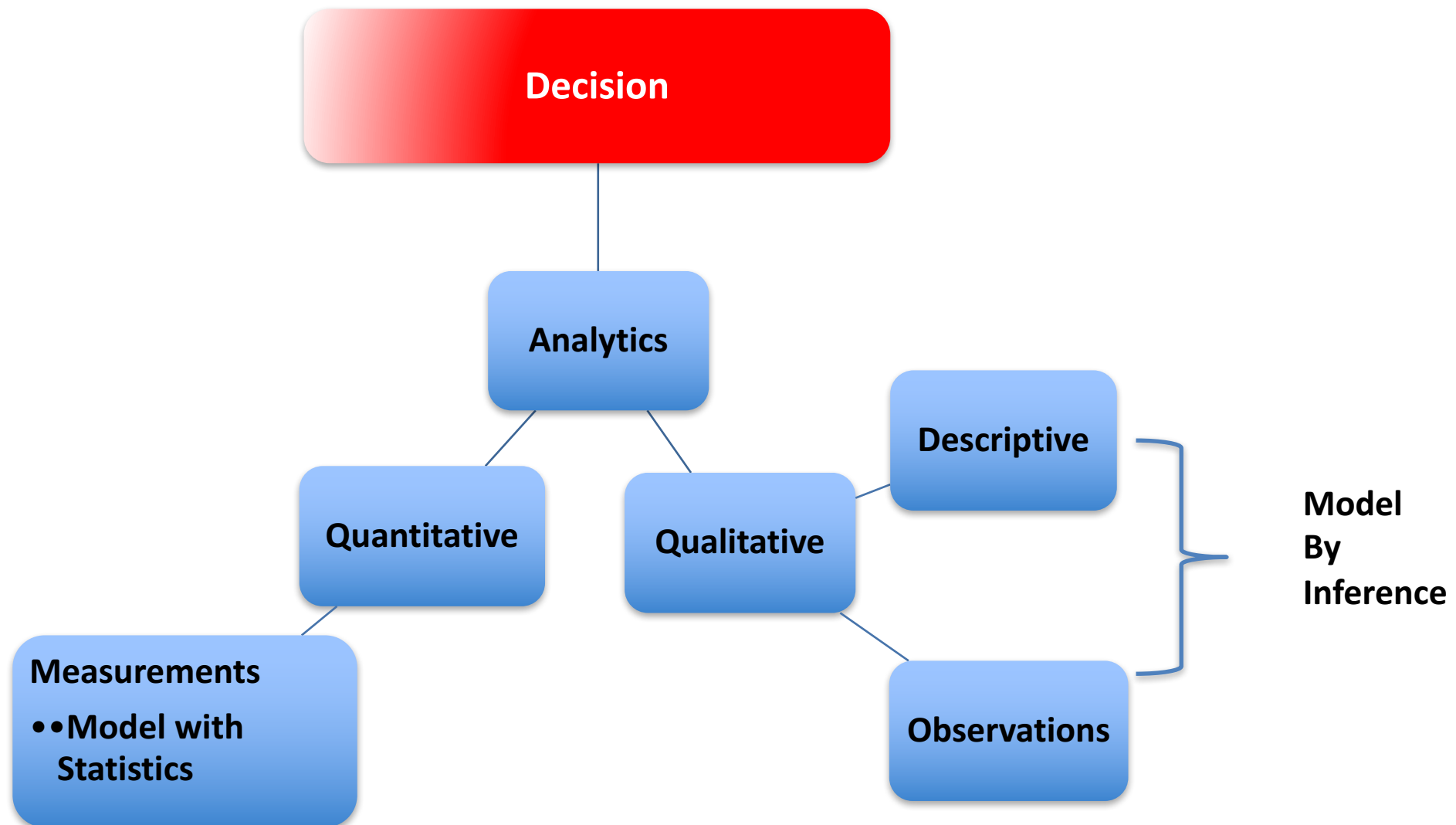
- **Humans (some) are good at ambiguity**
- **Machines aren't; they do what they are told (even with AI)**
- **Humans take shortcuts**
- **Machines don't**

Leverage each for their capabilities

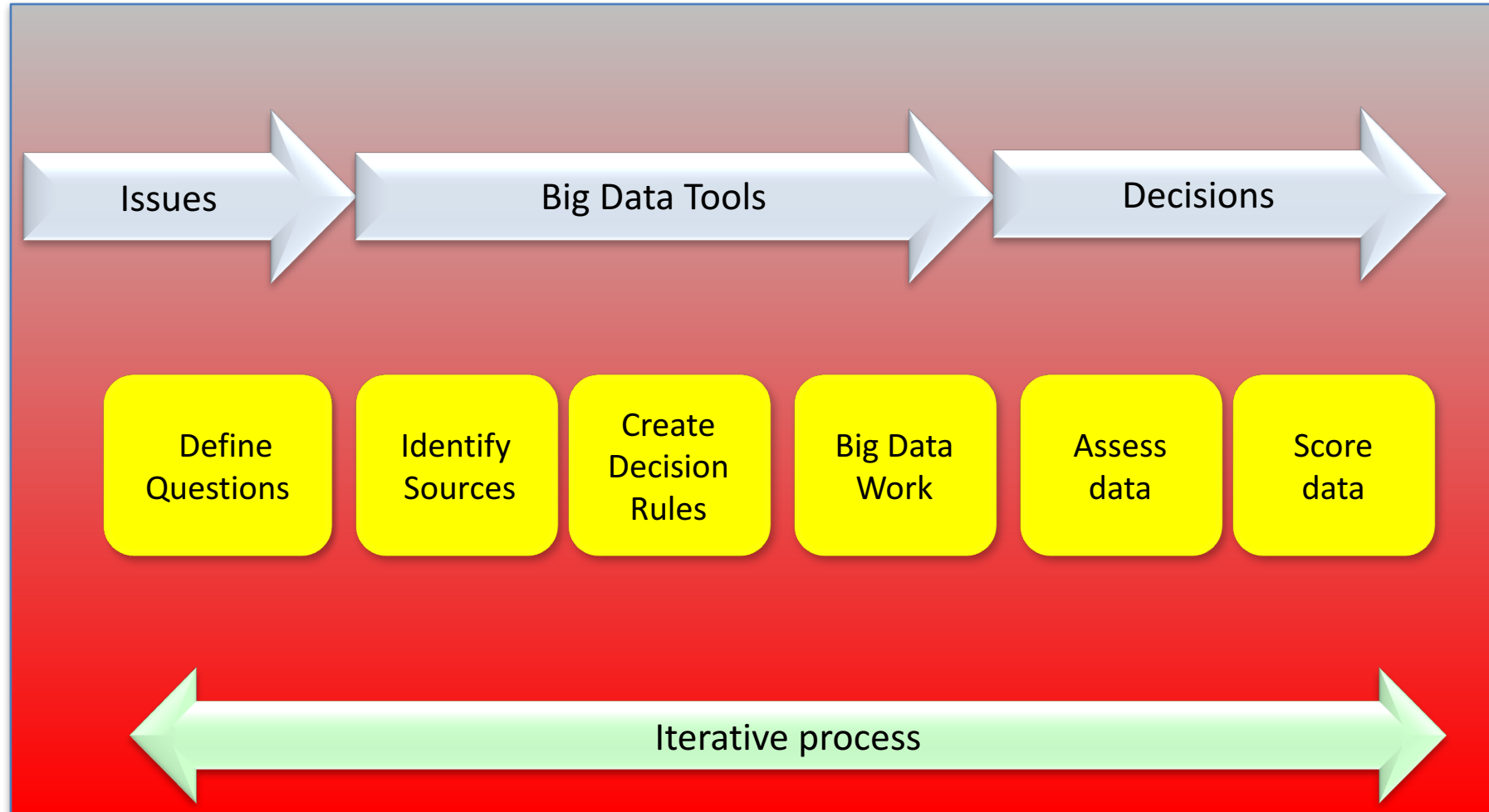
The Decision Process



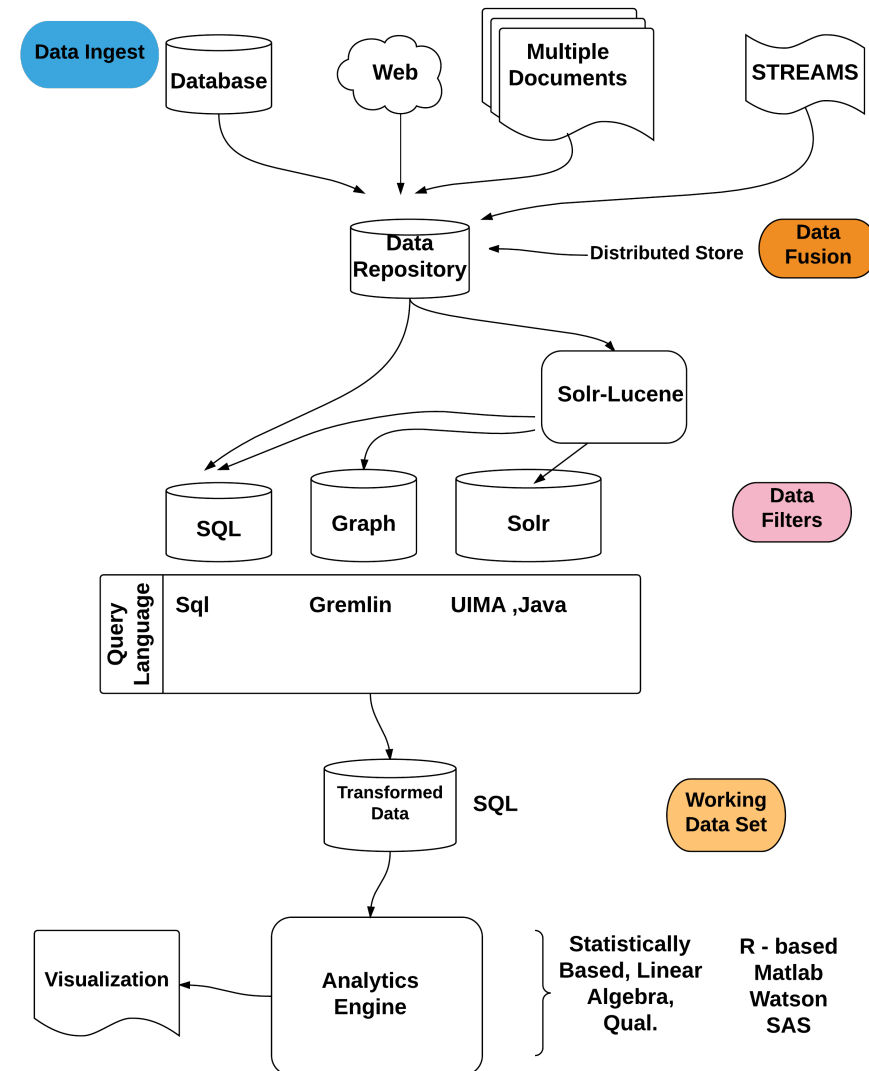
Data Models Required for Decisions



Process



Data Flow



Pharma – Natural Resistance to Change

- Often times focused on form rather than substance
- Standards can be leveraged
- cGxP can be used as a strategic advantage
- Regulatory Authorities
 - Expect you to be the expert
 - Need to understand rationale
 - Must see fact driven decisions
 - Open to new methodologies

What about part 11?

- limiting system access to authorized individuals
- use of operational system checks
- use of authority checks
- use of device checks
- determination that persons who develop, maintain, or use electronic systems have the education, training, and experience to perform their assigned tasks
- establishment of and adherence to written policies that hold individuals accountable for actions initiated under their electronic signatures
- appropriate
- controls over systems documentation controls for open systems corresponding to controls for closed systems bulleted above (§ 11.30)
- requirements related to electronic signatures (e.g., §§ 11.50, 11.70, 11.100, 11.200, and 11.300)
- accuracy, reliability, integrity, availability, and authenticity of required records and signatures

These are not barriers , rather represent good control

Data systems are unforgiving when it comes to manipulation

Pharma must must develop expertise in data systems controls

Databases – one size does not fit all

- **Row store**
 - Joins
 - Pre-defined
 - Arrays
- **Column store**
 - Reads
 - Batch process
 - Tuples
- **Graph**
 - Walks
 - Node jumping (no search-match)

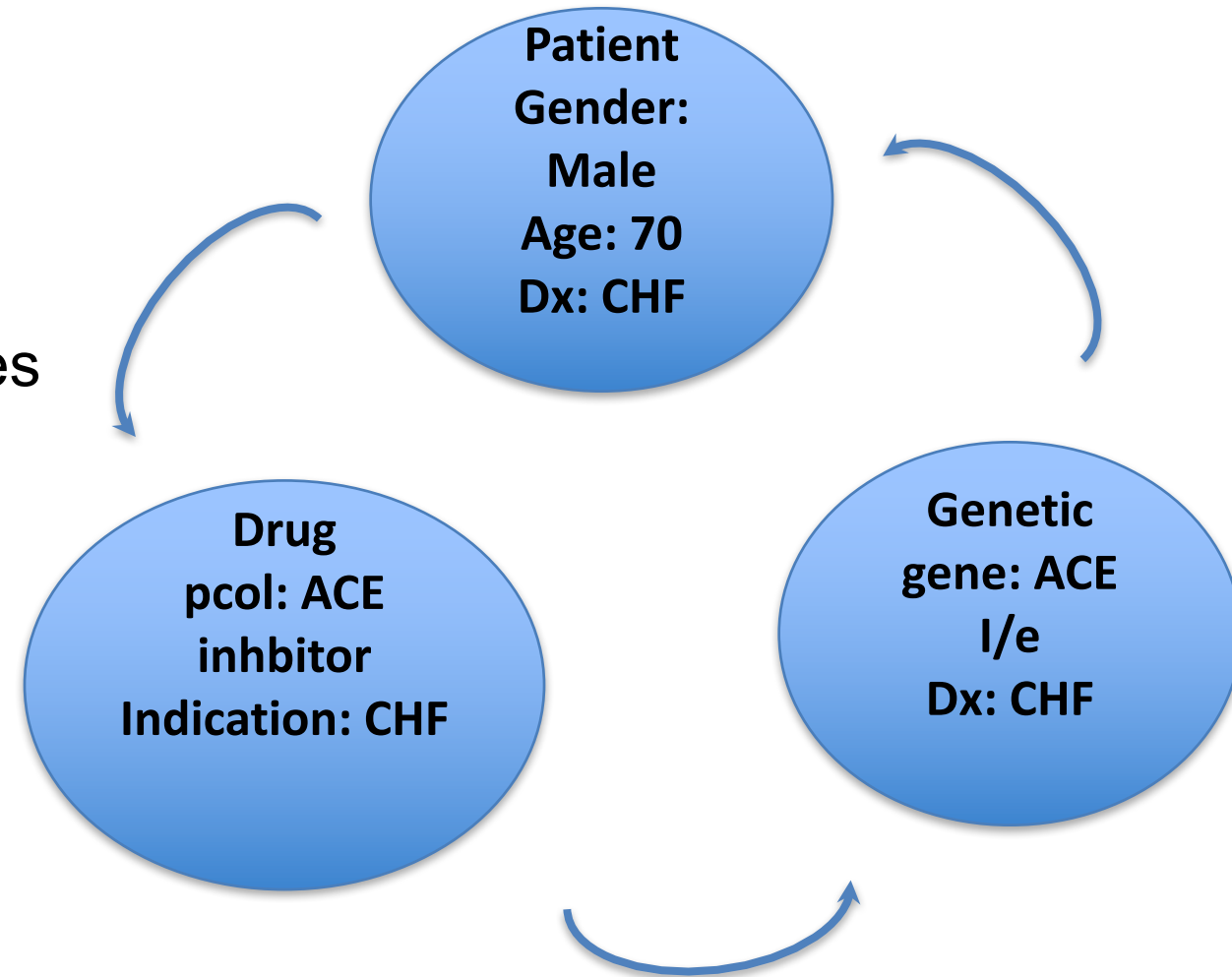
Distributed Store

- Distributed store is not distributed compute (Hadoop 2)
 - MapReduce
 - Spark
- Types of store
 - H-base
 - Cassandra
 - Mongo
- Extraction Language
 - Big table
 - Pig
 - Avero

**Really effective for
pre-processing data
Disambiguation
Filtering**

Graph Data

- Nodes
 - Entities with attributes
 - Edges : the relationship with nodes



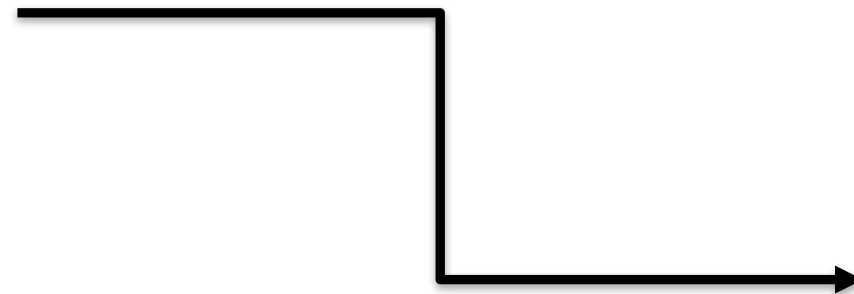
Why Graphs?

- **Very efficient – walks vs joins**
- **No index**
- **No keys**
- **Easy to modify**
- **Good for recommendation systems (think triples)**

Standardization as a Bridge to New Data Models

- A major focus of the organization is standardization
- It helps with structured data approaches for quantitative analysis
- Can be leveraged for qualitative analysis

Unstructured Text Analytics



Regulatory Authorities

Expect you to be the expert
Need to understand rationale
Must see fact driven decisions
Open to new methodologies

Natural Language Processing (NLP)

Language Identification

review text and determine language by pattern

recognition

Segmentation

grammar

parsing

Normalization

abbreviations

Classification

dictionaries

Disambiguation

rules

nouns

verbs

Relationship Extraction

rules

Natural Language Computing is hard because

Natural language is:

- highly ambiguous at all levels
- complex and subtle
- fuzzy, probabilistic
- interpretation involves combining evidence
- involves reasoning about the world
- embedded a social system of people interacting
 - persuading, insulting and amusing them
 - changing over time

Ambiguous Headlines

- **Iraqi Head Seeks Arms**
- **Juvenile Court to Try Shooting Defendant**
- **Teacher Strikes Idle Kids**
- **Stolen Painting Found by Tree**
- **Kids Make Nutritious Snacks**
- **Local HS Dropouts Cut in Half**
- **British Left Waffles on Falkland Islands**
- **Red Tape Holds Up New Bridges**
- **Clinton Wins on Budget, but More Lies Ahead**
- **Ban on Nude Dancing on Governor's Desk**

Andrew McCallum, UMass Amherst,
including material from Chris Manning and Jason Eisner

Use of UTA with EMRs

- Demo

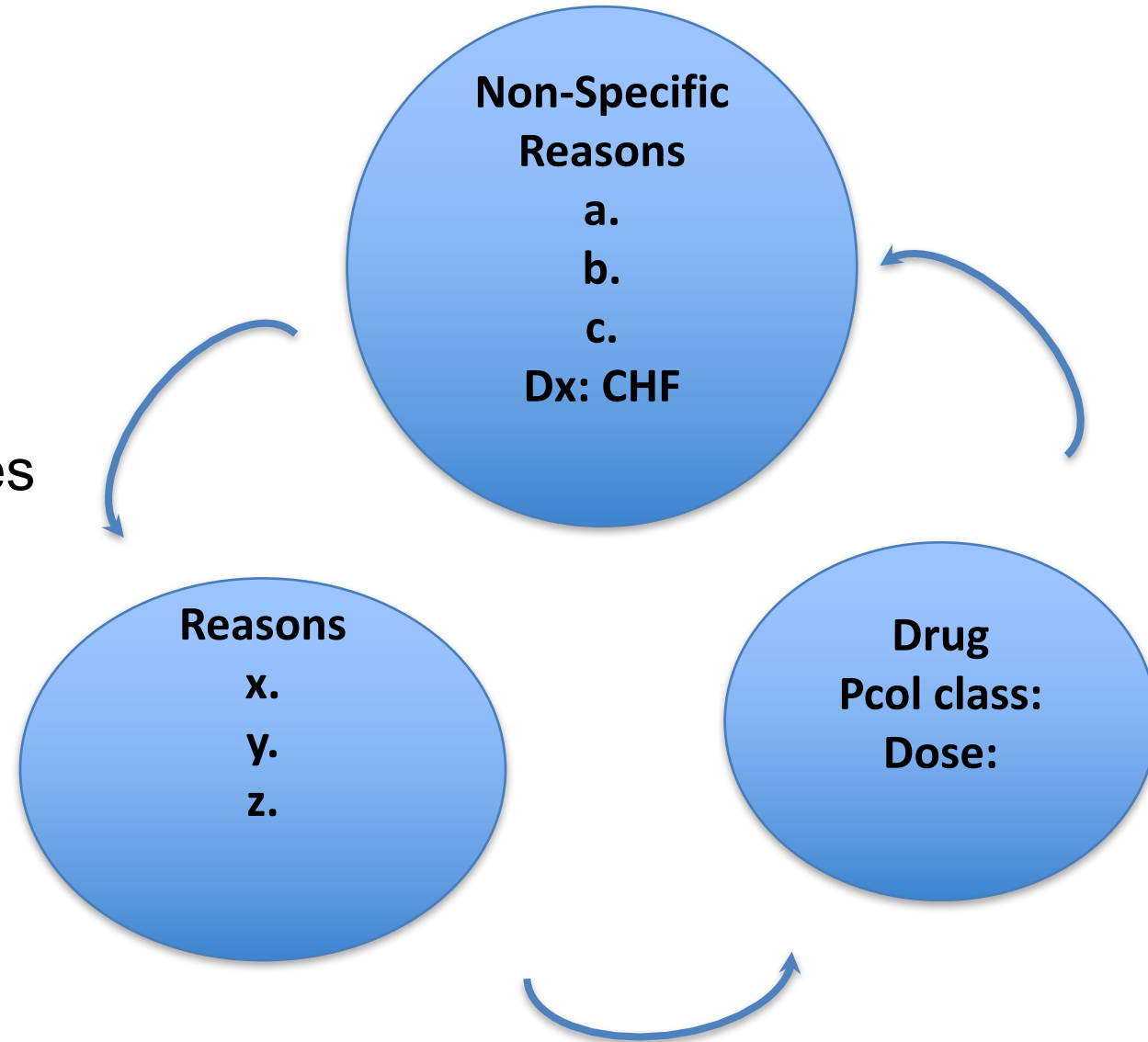
Phuse is in a good place...

- **Process for Standardization**
- **Attributes of Studies Well Characterized**
- **Very amenable to UTA and Graph data tools**

6. General Considerations	
6.1. All Measurement Types	
6.1.1. P-values	
6.1.2. Number of Therapy Groups	
6.1.3. Multi-phase Clinical Trials	
6.1.4. Integrated Summaries	
6.2. Demographic Measurements	
6.2.1. Variables to Display	
6.2.2. Standard for defining Race and Ethnicity	
6.3. Disposition Information	
6.3.1. Grouping of Reasons for Discontinuation	
6.3.2. Non-Specific Reasons	
6.3.3. Multiple Reasons	
6.3.4. Study versus Treatment Disposition	
6.4. Medications	
6.4.1. Collection of Medications	

Graph Data

- Nodes
 - Entities with attributes
 - Edges : the relationship with nodes



What Drives Infrastructure

- How fast do you need an answer?
- What is the analytic performed?
 - A nanosecond = 1 ft. of travel at 1 GHz
 - WiFi carries 1.4 Gb/sec (ideal conditions)
 - Packet assembly and filtering
 - Shorter = Better

Infrastructure Considerations

Compute

Processors
Cores /CPU
Threads
Memory caching
FPGA
GPU

Data

Source
Internal
External
Bus
Bandwidth
Intra machine
Inter machine

Storage

Access requirements
Media
Staging
Software defined

Attributes of Success (and failure)

- Senior Management Support
 - Cultural issues
 - Decision Readiness
- Critical Thinking Skills
 - Understand alternatives
 - Priorities
 - Bias
 - Tactical vs Strategic Questions
- Infrastructure
 - High I/O demand
 - Data transfer/memory
 - Software Integration

It's all about the filter

Specificity and Context

***Driven by the Problem
Statement***