Big Data, Big Graphs, Big Computers

Peter Li
Mayo Clinic
Cohort Data Mining

• Typical Clinical Study
  – Specific Variables and Outcome
  – Hypothesis Driven

• Individualized Medicine
  – Search for a cohort like this patient
  – Need fuzzy, imprecise query

• Traditional Database: query on specific attributes
  – Easy: Optimize queries with indices
  – Hard: Predict indices needed
  – Harder: Complex operators: time differences, value expansion
Drug Ontology

Disease Ontology

Laboratory Ontology

Event Ontology

Time-Invariant e.g. demographics (ontologies not shown)

PaGent A

Time Line

Time = t₁

Time = t₂

Time = t₃

IniGal

Visit

Clinic Visit

Follow Up

In-patient

PaKern of hospitalization with Renal Failure after drug X and Y are given, time = t₂ + t₃ or t₆

Drug Ontology

Patient A

Time Line

Initial Visit

Drug X

Drug Y

Drug Ontology

Patient B

Time Line

Initial Visit

Drug X

Drug Y

Renal Failure

Relationships that mirror Ontology/ontology links

Renal Failure
Relational to Graph (RDB to RDF)

Snapshot query of entire REP database takes about 12 hours

SQL Database Server

SQL Queries

Relational Database (50 GB)

D2R

SQL → RDF Data and Schema Translator

D2R Mapping File

(500 GB)

D2R

RDF Triples

Semantic Database (300 GB)

uRiKA on XMT-2

Semantic Database Server

SPARQL Queries

Web Interface

SPARQL queries submitted over web interface take as little as 0.5 seconds to return

Copying 500 GB RDF triples file to XMT-2, and then building the (binary) uRiKA database takes about 10 hours

Courtesy of Dr. Gilbert, SPPDG, Mayo Clinic
XMT2 Architecture

- Scale out threads, not Hz
  - 1 processor has 128 threads
- Shared common memory
- Network connected nodes

- Mayo XMT2
  - 64 processors has 8,192 threads
  - Shared 2 TB RAM
  - 16 TB Lustre FS
Mouse Genome Chromosome 19
Challenges

• How to convert measurement data to interconnected nodes?
  – Fasting Serum Glucose = 120 mg/dl
  – “fasting serum glucose” is easy to connect (to lab ontology)
  – How to connect the actual value, 120?

• Large databases (preconceived model) \(\rightarrow\) semantic graph
  – 20+ years of electronic medical records, 100+ TB of data
  – What to do with clinical text?
  – How to encode temporal data? Absolute, Relative? Point, Interval?

• Large graphs \(\rightarrow\) how to analyze it with HPC?
  – RDF increase size 6X
  – Clusters rely on data locality
  – Graphs has no locality