Big Data and Graph Analytics in a Health Care Setting

Supercomputing 12

November 15, 2012

Bob Techentin

Mayo Clinic

MAYO CLINIC

What is the Mayo Clinic?

- Mayo Clinic Mission:
  - To inspire hope and contribute to health and well-being by providing the best care to every patient through integrated clinical practice, education and research
- Primary Value:
  - The needs of the patient come first

> The best interest of the patient is the only interest to be considered, and in order that the sick may have the benefit of advancing knowledge, union of forces is necessary.
>
> -- William J. Mayo, June 15, 1910

Mayo Clinic Environment
(2011 Data)

- Not-for-profit foundation for Health Care/Research/Education
- Total Mayo staff size: 58,300 in three locations (Rochester, MN; Jacksonville, FL; and Scottsdale, AZ) plus multiple (60+) regional sites near Rochester
- Mayo – Rochester: 15 million square feet (about 3.5X Mall of America)
- 1,113,000 patient registrations, 123,000 Hospital admissions, 588,000 hospital days, 2,400+ hospital beds, 20 million+ yearly laboratory tests
- Total Mayo research staff size: 4,252 FTEs

Clinical Data Analysis Challenges

- Many clinical analysis problems are virtually intractable by traditional computational means. Mayo Clinic's patient data includes 5,000,000 patient records, exceeding 10 PB of data storage.
- Medical records are a combination of numeric data (e.g., lab work), clinical observations (e.g., medical history), categorical data (e.g., clinical diagnosis), and longitudinal measurements (e.g., annual checkup). Developing query mechanisms for this complex data is challenging.
- Relationships between clinical factors is a found in a very small number of records. This is a low signal-to-noise relationship. Access to a large, well-organized medical record database is critical.
- Analysis approaches differ for various applications
  - Clinical decision support – patient vs. population
  - Research – correlation of disparate factors
  - Monitoring – detect / predict population trends

Clinical Data Analysis Example: Clinical Decision Support – Patient vs. Population

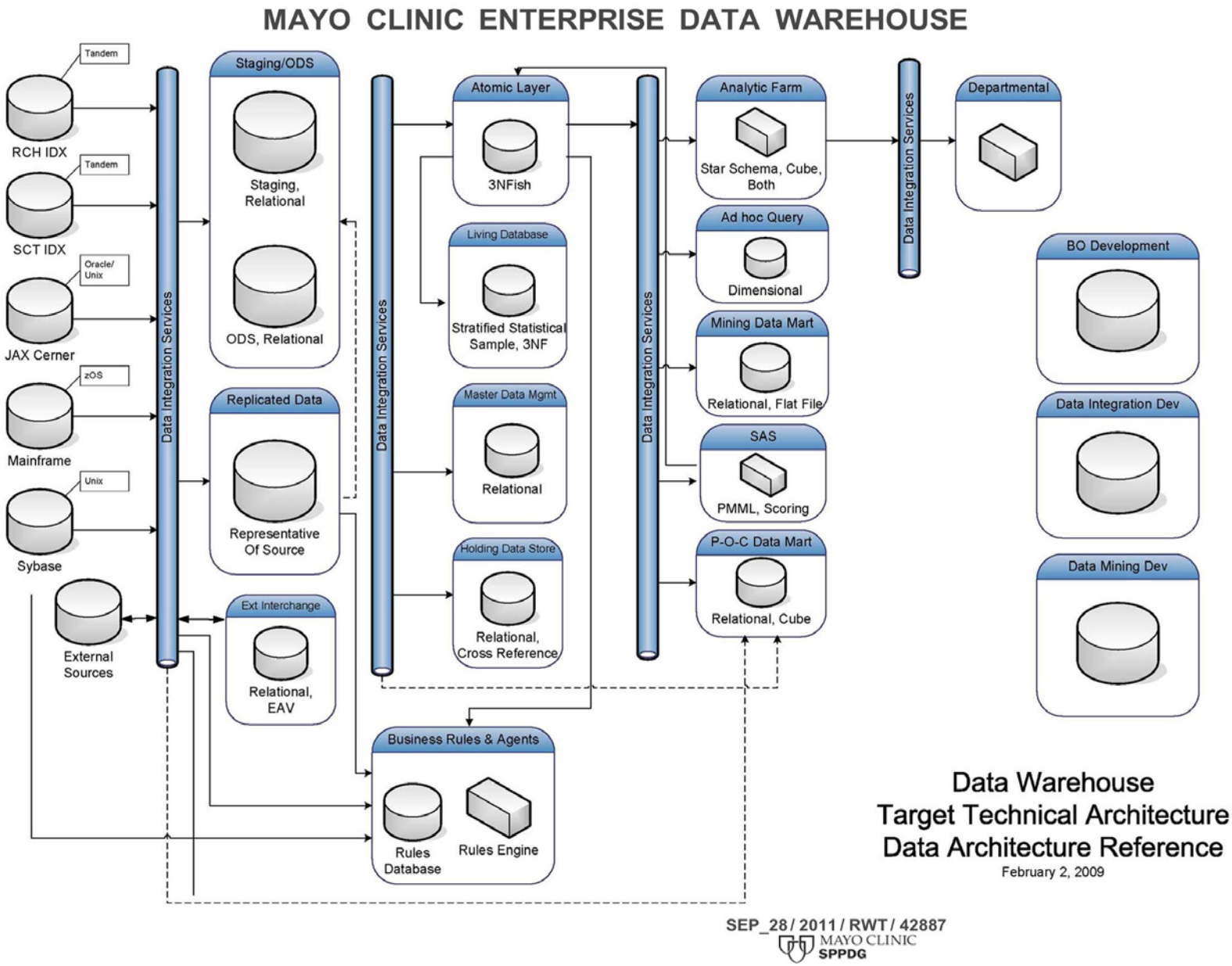
- Clinical Decision Support Rules have been implemented at Mayo Clinic hospitals
  - Rules are laboriously crafted by medical experts, and implemented by information systems staff
  - Rules are back-tested against medical records, manually!
  - Implemented rules “fire” thousands of times per week, with a high false positive rate
- Systems engineering approach now being applied to development of new rules
  - Systematically identify failure modes
  - Analyze the whole system (including Big Data) and connections between processes

Clinical Data Analysis Example: Research – Correlation of Disparate Factors

- Discovery in 2006 that a huge (7-8%) decrease in incidence of breast cancer in one year (2003) was caused by nationwide cessation of hormone supplementation in menopausal women in 2002!
  - First reported at San Antonio Breast Cancer Symposium, December 13-15, 2006
    - Discovered through laborious manual records reviews!
- Clinicians and medical researchers need technology to reduce investigations from years to months or days
  - Many such reviews are simply not done because of limited manual resources

Characteristics of Mayo Enterprise Data Warehouse

- Complex
  - Over 1000 departmental systems
  - Many specialty domains
  - Wide variety of data semantics
- Large
  - Oldest formal medical record
  - Complete conversion to electronic records (medical, surgical, radiology, billing, etc.) over 10 years ago
  - Millions of patients / Terabytes of data
- Significant challenges
  - Data sets are not well connected
  - Noisy, incomplete, low reliability data / Low signal-to-noise ratio



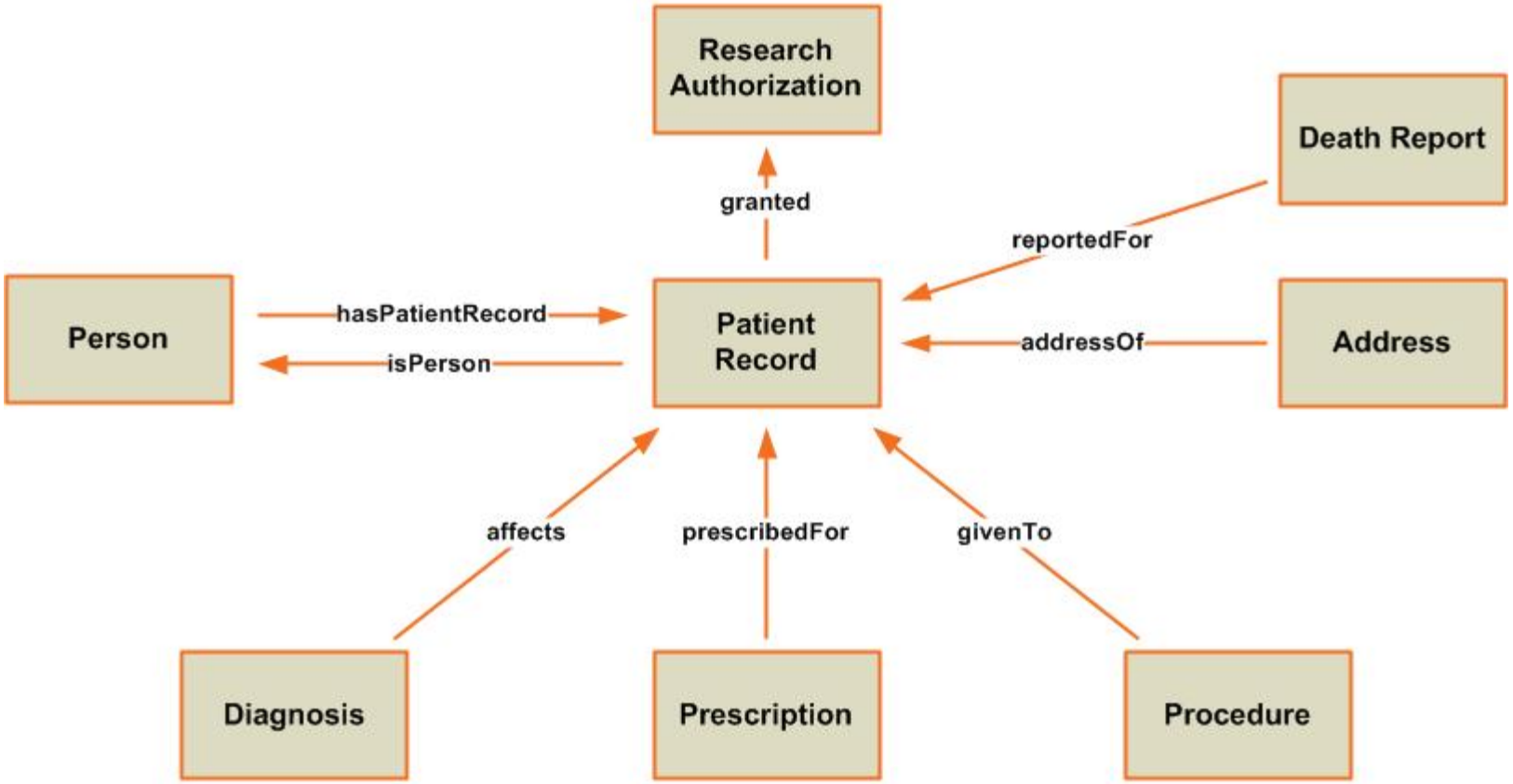
SEP_28 / 2011 / RWT / 42887

MAYO CLINIC
SPPDG

Example Medical Record Database: Rochester Epidemiology Project

- REP data captures a subset of medical record data for a relatively stable population in Southeast Minnesota
  - ~ 500K Individuals, 40 year duration
  - ~ 2 M medical records from many clinics and hospitals
  - Medical record events limited to births, deaths, diagnoses, prescriptions, procedures
- Relatively small and simple
  - 50 GB Sybase database is two orders of magnitude smaller than Mayo Electronic Medical Record
  - Translation to semantic (RDF) model results in dozens of classes and only 3.5 billion triples

ROCHESTER EPIDEMIOLOGY PROJECT SEMANTIC RELATIONSHIPS
(Named Relationships Between Primary Classes are Extracted from Relational Databases and Inference Rules)



SEP_04 / 2012 / RWT / 43600

MAYO CLINIC SPPDG

Early Results From Semantic Analysis of Rochester Epidemiology Project Data

- Translation from multiple relational database sources into single semantic model is relatively straight forward
  - 50 GB of relational data expands to 350 GB in NTRIPLES format
  - Translation and data management can be slow and cumbersome
- Initial queries appear promising for research and clinical practice
  - Cray XMT-2 and YarcData uRiKA
  - Query times of seconds, compared to hours for Sybase

Expanding Clinical Record Analysis Beyond the Test Cases

- The Mayo Clinic Data Warehouse is two orders of magnitude larger than the REP databases, and the EMR is much larger yet
  - Many more patients, across the U.S. and international
  - Much more complex / richer dataset
  - Aggregated data sets might exceed 300 B triples
- New systems supporting clinical practice need to scale in capacity and capability
  - Medical centers might require 10,000-100,000 queries per day
  - Integration of graph analytics with FLOPS - Logical AND numerical!

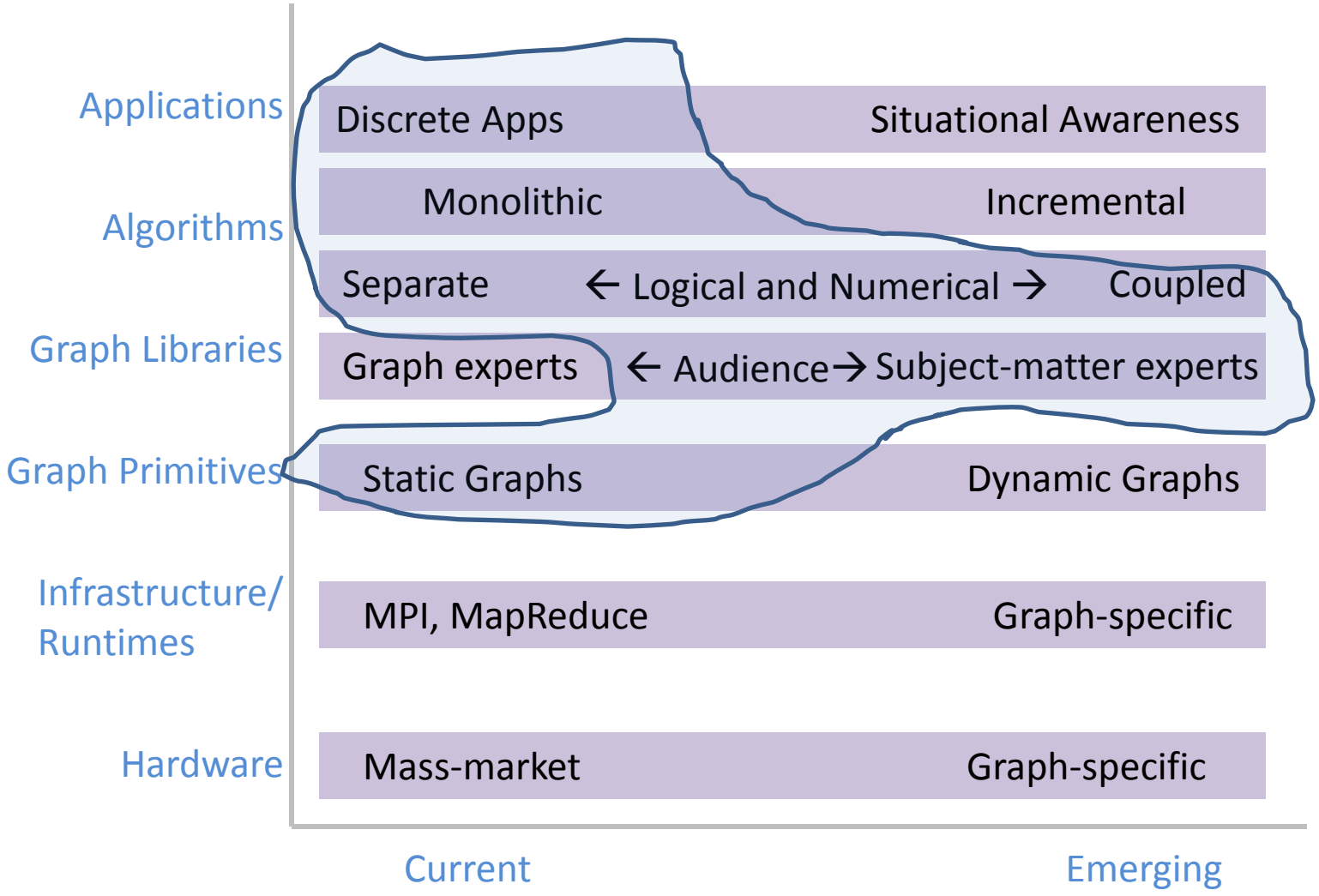
Semantic Analysis Challenges for Clinical Applications

- Annotating “facts” with quality or reliability
  - Reification adds substantial complexity to semantic model
- Identifying complex co-morbidity conditions
  - Simple queries return N^M combinations – cluster analysis may be more effective
- Integration of logical and numerical processes
  - Queries over ranges of values or dates (without FILTER!)
  - Integrated analysis (with more sophistication than COUNT)
- Approximate matching
  - “Find a patient like me” where “me” is 40,000 triples
- Temporal analysis

Summary

- Both “Big Data” and “Graph Analytics” will have a role in future clinical practice and medical research
- Initial work with semantic analysis of medical record data using XMT-2 and uRiKA appear promising
- Beginning to recognize limitations of current systems
  - Scale and performance of future systems
  - Expressing and querying data quality and reliability
  - Integration of true graph analytics
  - Tight coupling of logical and numerical analysis

A Few Dimensions of Evolution TODAY



A Few Dimensions of Evolution
FUTURE