Sirenic

Stream Infrastructure for Real-Time ICU Data Analysis

JOHNS HOPKINS
UNIVERSITY

Yanif Ahmad, Computer Science
Email: yanif@jhu.edu
Collaborators: Raimond Winslow (BME)
Yair Amir (CS), Suchi Saria (CS, MPH)

Long-Lived Data Infrastructure

"Pipeline": workflow-centric data and task management

- Multi-scale problems
- Multi-faceted workloads

Data platform desiderata

- Transparency, flexibility
- Scalability and usability

Diverse userbase

- Core team
- Long tail of users

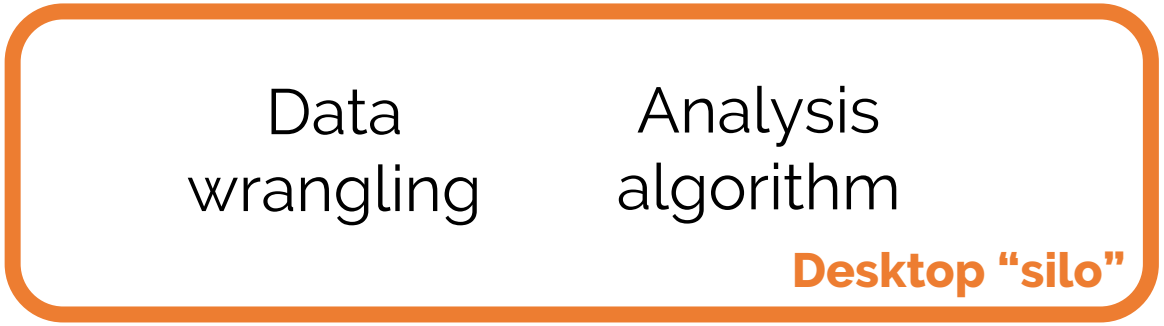
Workload-Driven Database Architectures

	Specialization dimension	Applications	Gains	Enabling mechanism
Query	Transactions	Web services, e-commerce	80x speedup (since 2008)	Coordination restriction
	Batch analytics	Log analysis, BI	10x cheaper (since 2004)	Simpler, open infrastructure
Dataset	Continuous and incremental	Monitoring, finance	100-1,000x speedup (since 2003)	Remove design impedance
	Graphs, NoSQL	Documents, web graphs, brains	10x cheaper 10-100x speedup	Better hardware utilization
Emerging	Approximate and probabilistic	Data cleaning and exploration	New subfield (~2004)	New algorithms and theory
	Workflows and computational	Data-driven science, linear algebra, ML	10-100x speedup (since 2010)	Better hardware utilization and interoperability
	“Wide” data	Statistics, data exploration, ML	TBA (since 2013)	Better hardware utilization

New Tools for Data Systems Design

Key challenges to specialization:

- Implementation bottlenecks
- No short feedback cycle



K3: declarative systems design

- Technical aim: mechanize domain-specific data tools construction
- Usability aim: remove barriers in turning analytics into services

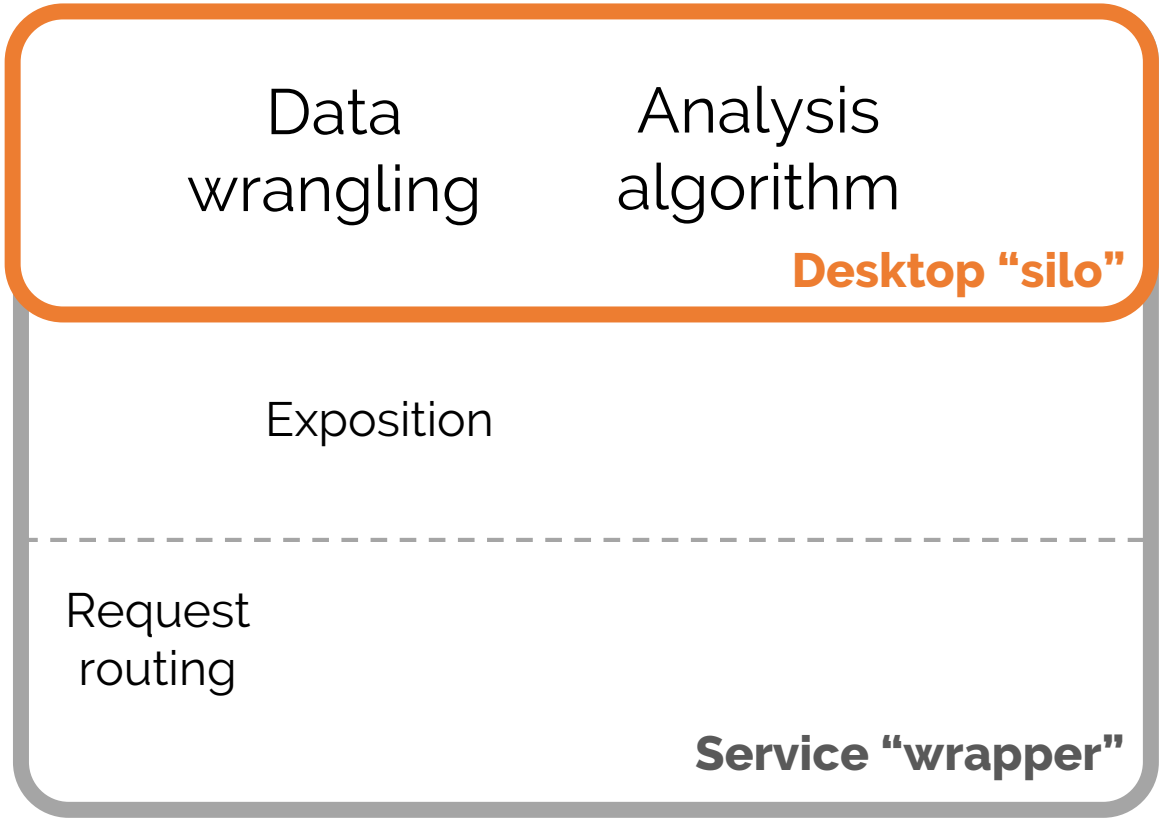
New Tools for Data Systems Design

Key challenges to specialization:

- Implementation bottlenecks
- No short feedback cycle

K3: declarative systems design

- Technical aim: mechanize domain-specific data tools construction
- Usability aim: remove barriers in turning analytics into services



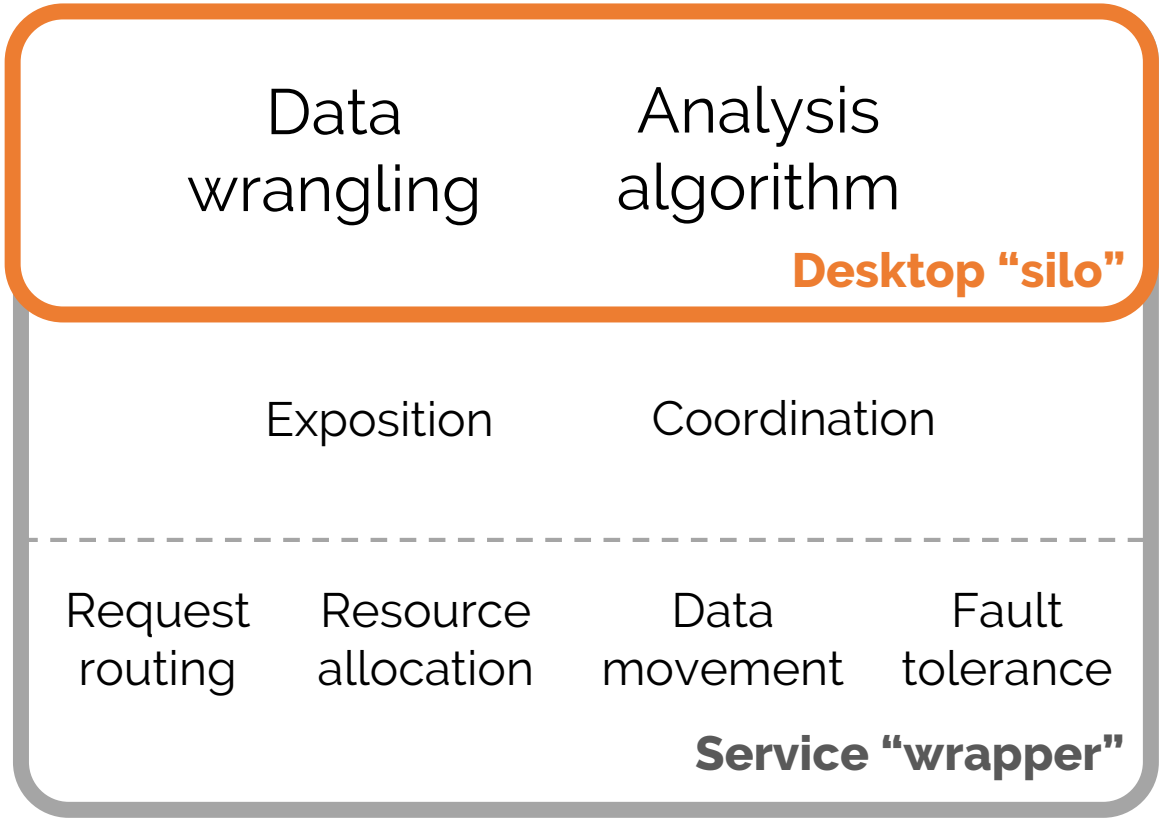
New Tools for Data Systems Design

Key challenges to specialization:

- Implementation bottlenecks
- No short feedback cycle

K3: declarative systems design

- Technical aim: mechanize domain-specific data tools construction
- Usability aim: remove barriers in turning analytics into services



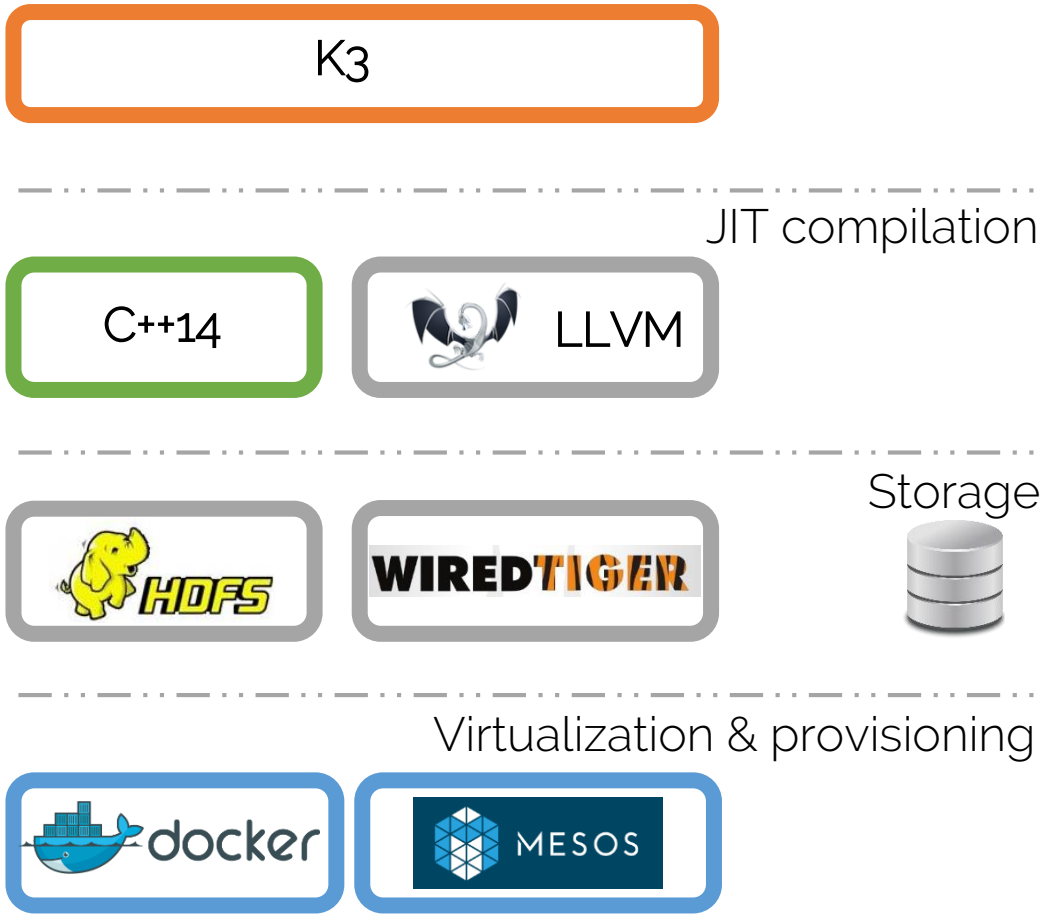
New Tools for Data Systems Design

Key challenges to specialization:

- Implementation bottlenecks
- No short feedback cycle

K3: declarative systems design

- Technical aim: mechanize domain-specific data tools construction
- Usability aim: remove barriers in turning analytics into services



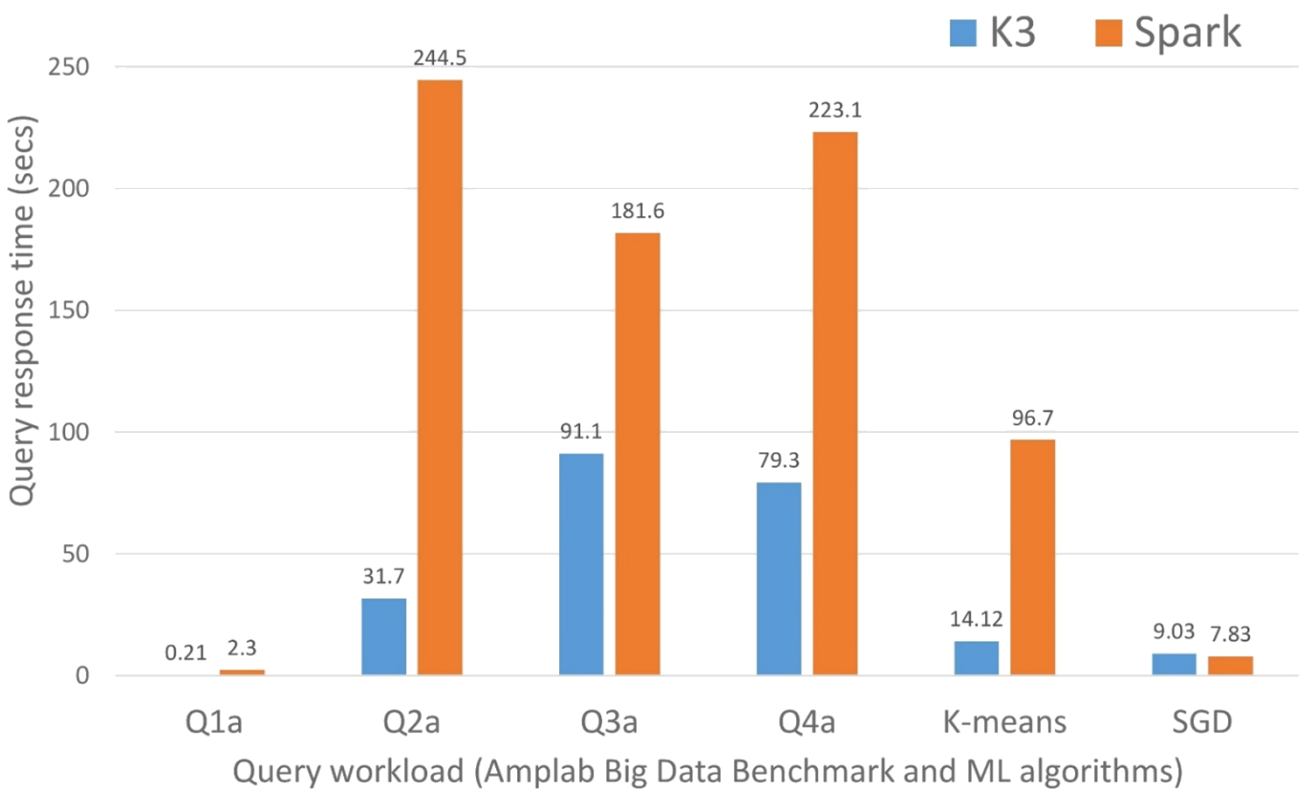
New Tools for Data Systems Design

Key challenges to specialization:

- Implementation bottlenecks
- No short feedback cycle

K3: declarative systems design

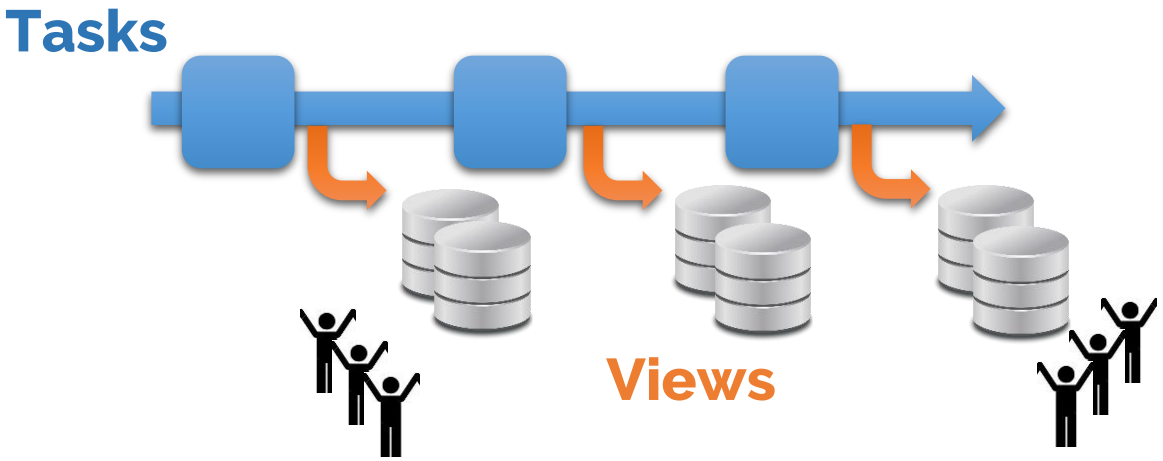
- Technical aim: mechanize domain-specific data tools construction
- Usability aim: remove barriers in turning analytics into services

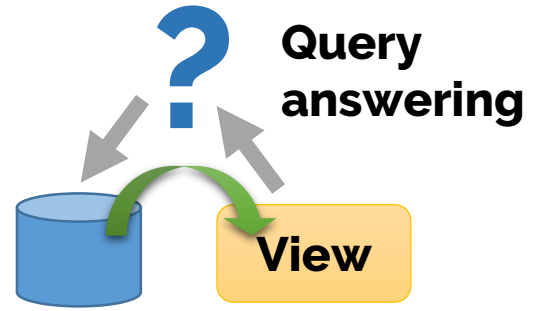


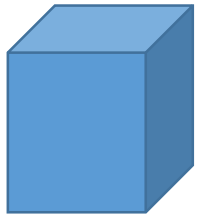
Initial results from July 2014:
20 node, 160-core deployment, ~250GB web log dataset, and ~20GB molecular dynamics dataset

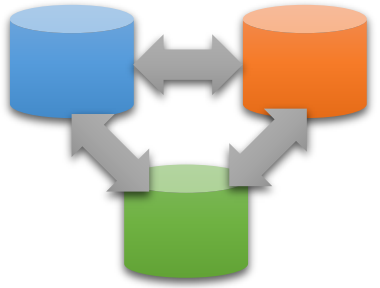
Building Data Pipelines in K3

Goal: enhance usability through **pipeline data products**









Approach: **data views**, for data independence in pipeline data products

- A central logical abstraction mechanism in DBMS

Views naturally model:

- Multi-resolution, and multi-modal data
- Archiving (snapshots/checkpoints) and visualization

Physiological Time Series

Dataset: bedside monitors capturing patient state

- 16 channels per bed (e.g., ECG, NIBP, SpO2), 2kHz sampling rate per channel

Data challenge: alarms detection and patient state prediction

- Particularly for rapid onset diseases (e.g., sepsis, arrhythmias)

Project aims: design and deploy initial platform to facilitate real-time analysis algorithm design

GE Solar 8000i patient monitor

Sirenic Architecture and Data Tasks

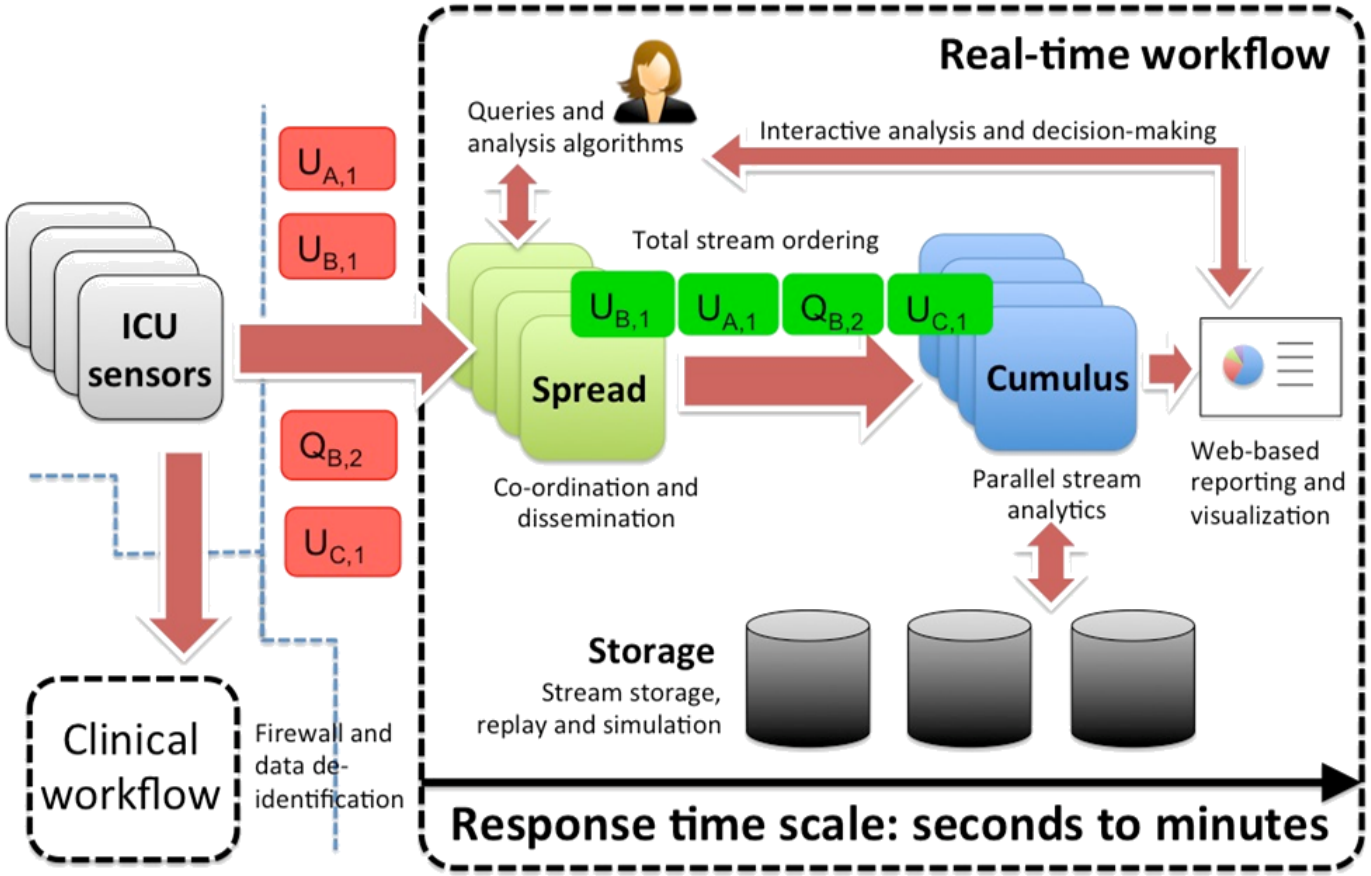
Live data source since mid-August:
JHH PICU (Dr. Mela Bembea, Director)
3 beds, 1610 anonymized series

Data and model-driven analyses:

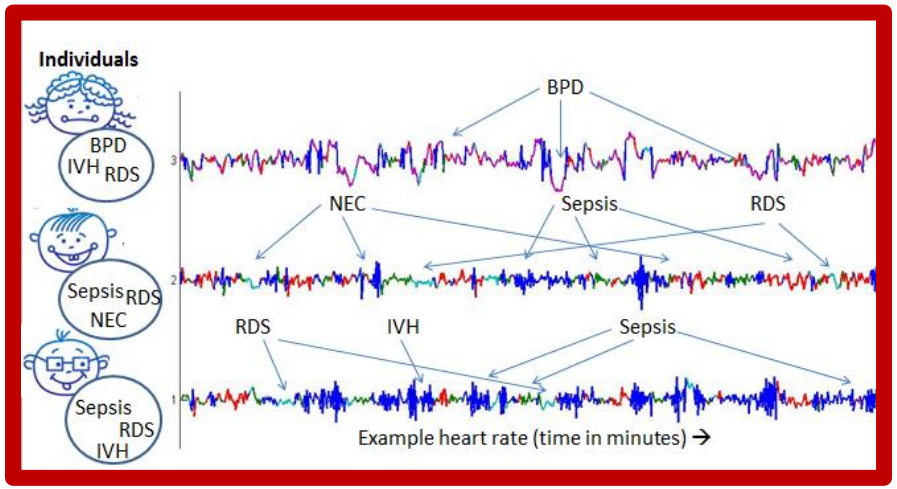
- Signature (motif) identification and extraction
- Latent structure model development

Workflow tasks:

- HL7 parsing & deidentification
- Model-driven simulation
- Drill-down visualization

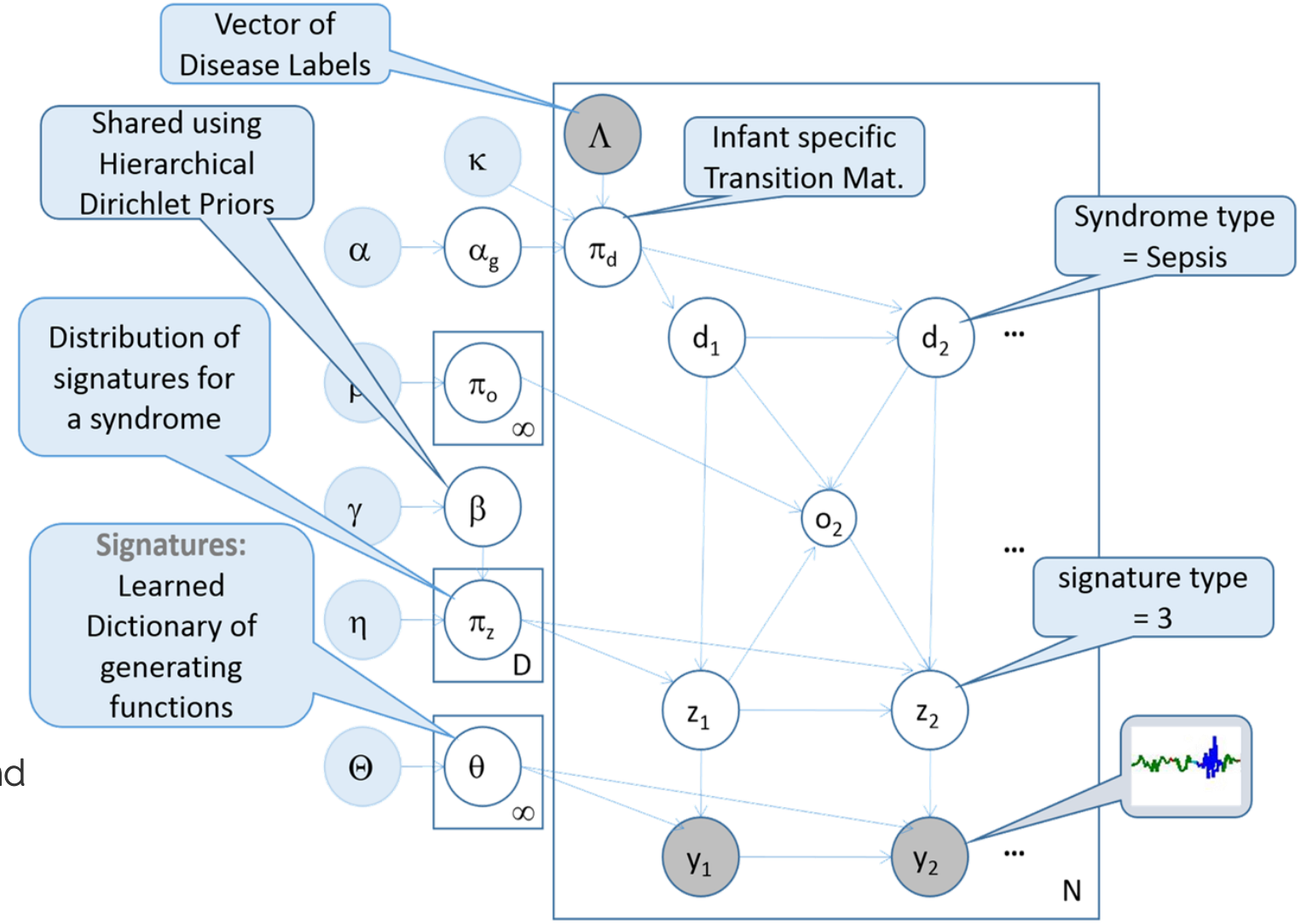


Modelling Health State Over Time



S. Saria, A. Rajani, J. Gould, D. Koller, A. Penn. **Integration of Early Physiological Responses Predicts Later Illness Severity in Preterm Infants**. Science *Translational Medicine*, September 2010. Vol. 2, Issue 48

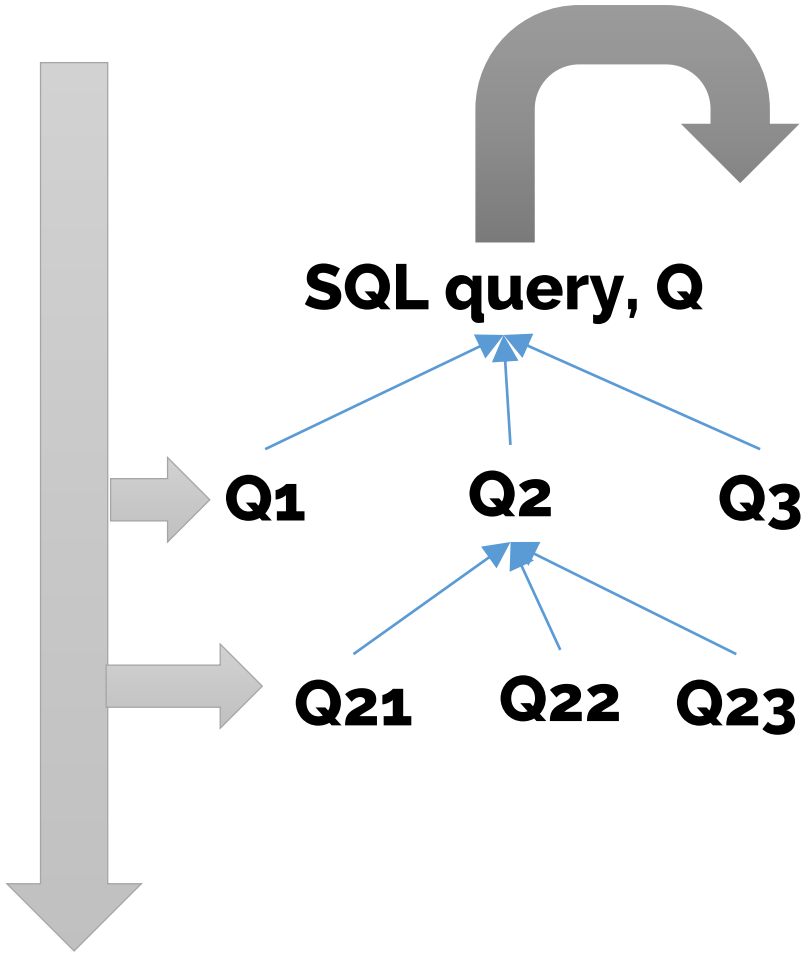
S. Saria, D. Koller, A. Penn. **Discovering shared and individual latent structure in multiple time series** NIPS Predictive Models in Personalized Medicine, August 2010.



Real-time Views in Sirenic

Novel change propagation techniques based on multiple "delta" data representations

- Think "finite differences" for database queries
- Data always "in-flight", facilitating delayed corrections (missing data, improved robustness)
- Clean snapshot semantics, and efficient snapshot management



Real-time Views in Sirenic

Novel change propagation techniques based on multiple "delta" data representations

- Think "finite differences" for database queries
- Data always "in-flight", facilitating delayed corrections (missing data, improved robustness)
- Clean snapshot semantics, and efficient snapshot management

Use cases:

- Scalable inference via distributed, blocked Gibbs sampling
- Multi-resolution aggregate summaries and difference data structures for data exploration

Project status: initial data exploration, and infrastructure deployment

Summary

Data Applications	**Sirenic**
Data Abstraction Layer	**Data Views**
Systems Abstraction Layer	**K3: Programmable Data Systems**

Thank You!

yanif@cs.jhu.edu
http://damsl.cs.jhu.edu