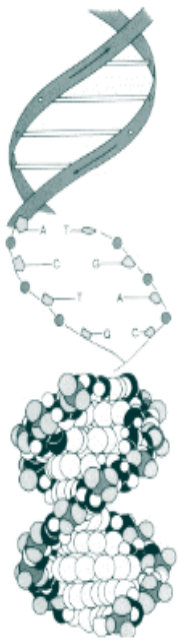
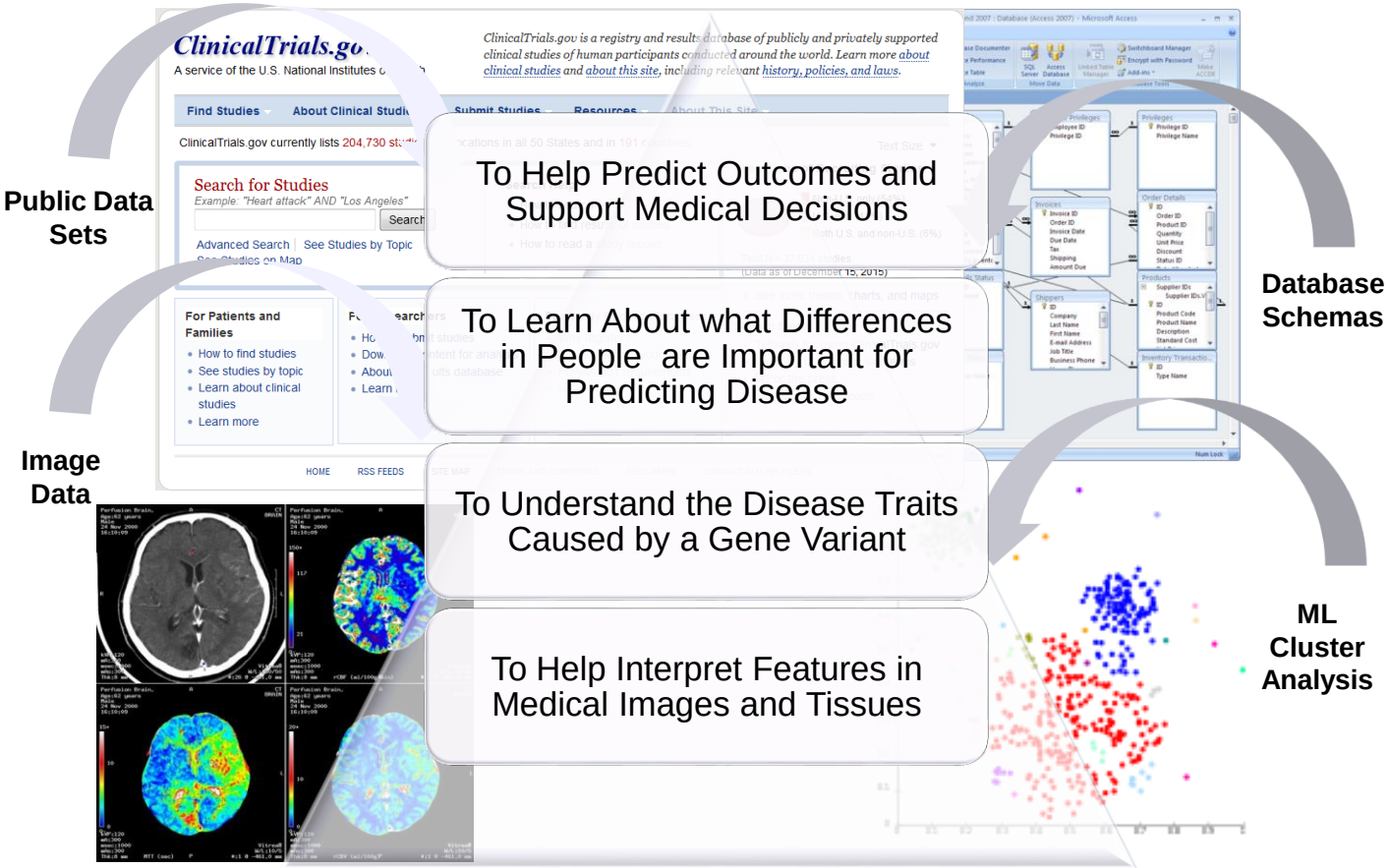


Using Big Data to Create an Information Commons with the i2b2 Infrastructure

Christopher Herrick
Lori Phillips

Information Commons:
Using Big Data to Improve Healthcare



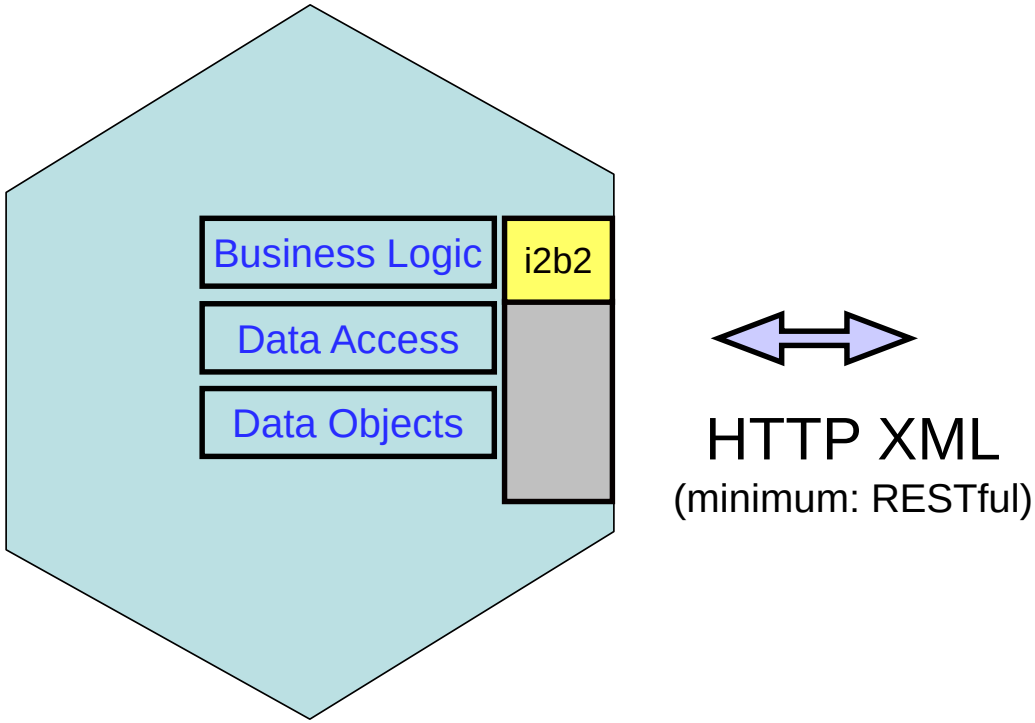
How to Host an Information Commons?

- Goal
  - Hook together repositories of big data to gather patient information, deliver data, and perform analytics

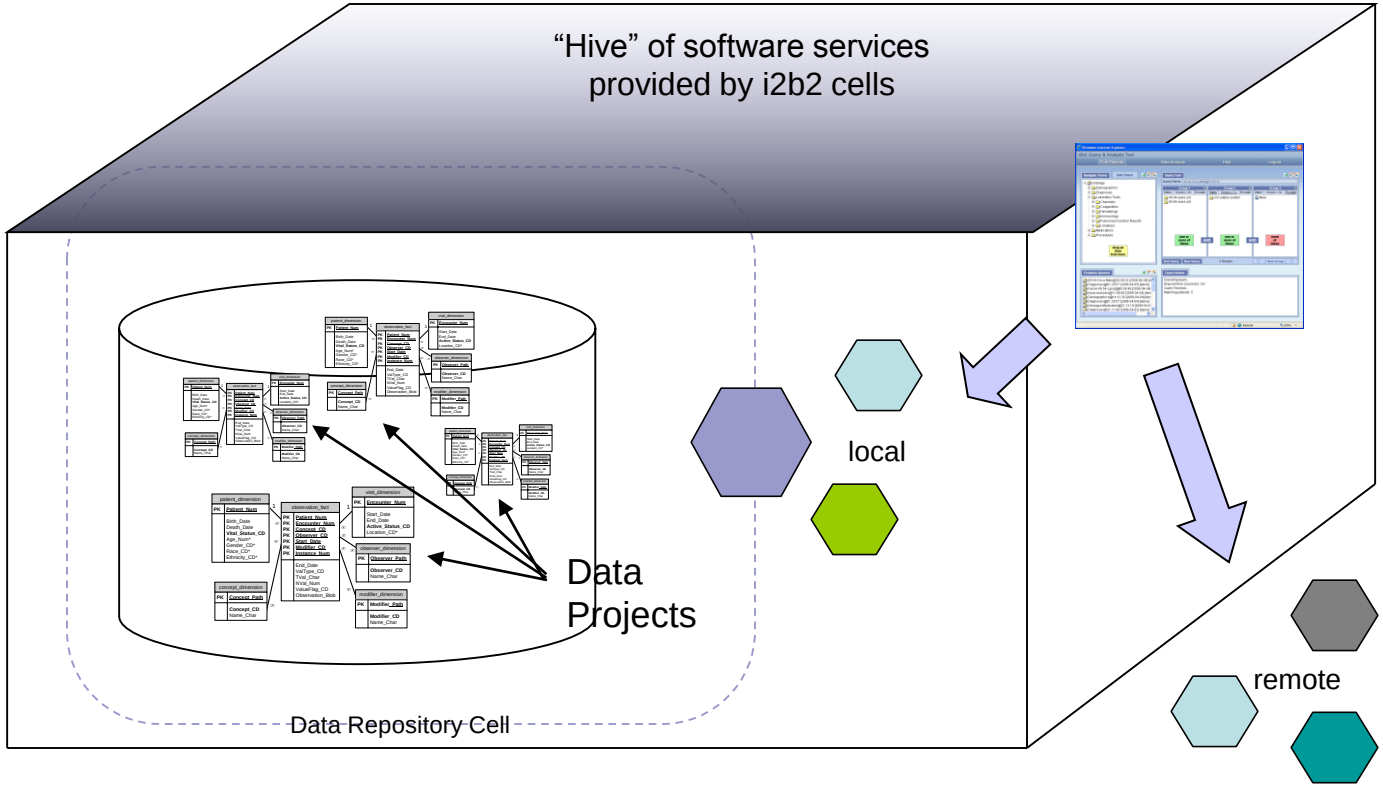
- Problem
  - Big Data cannot be loaded into a common repository
    - Difficult to move
    - No common format for a single database
    - People who work with Big Data are highly specialized

- Approach
  - Big Data stays at source; managed by the specialized teams who own data
  - Publish data through ontologies of available specialized data
  - Provide API to allow patient linkage, querying, and data analyses to be managed

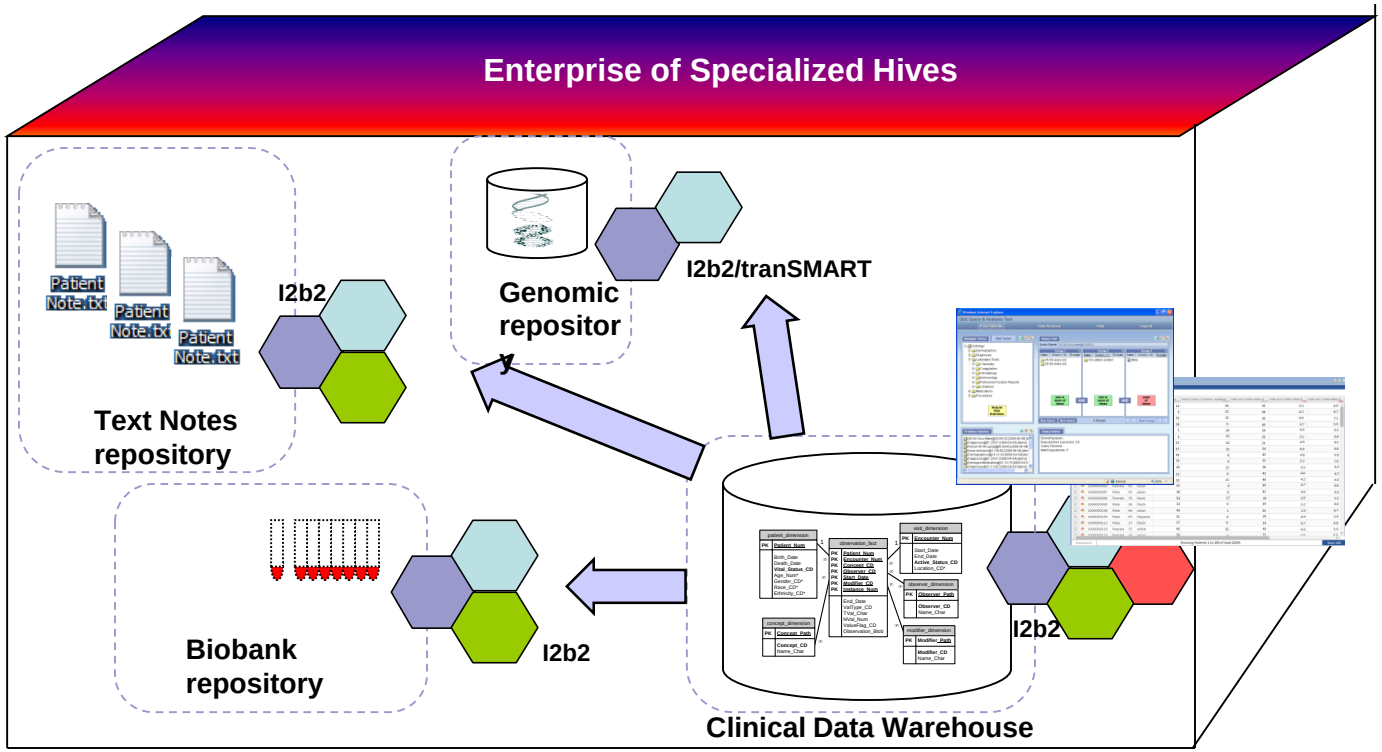
i2b2 Cell: The Canonical Software Module



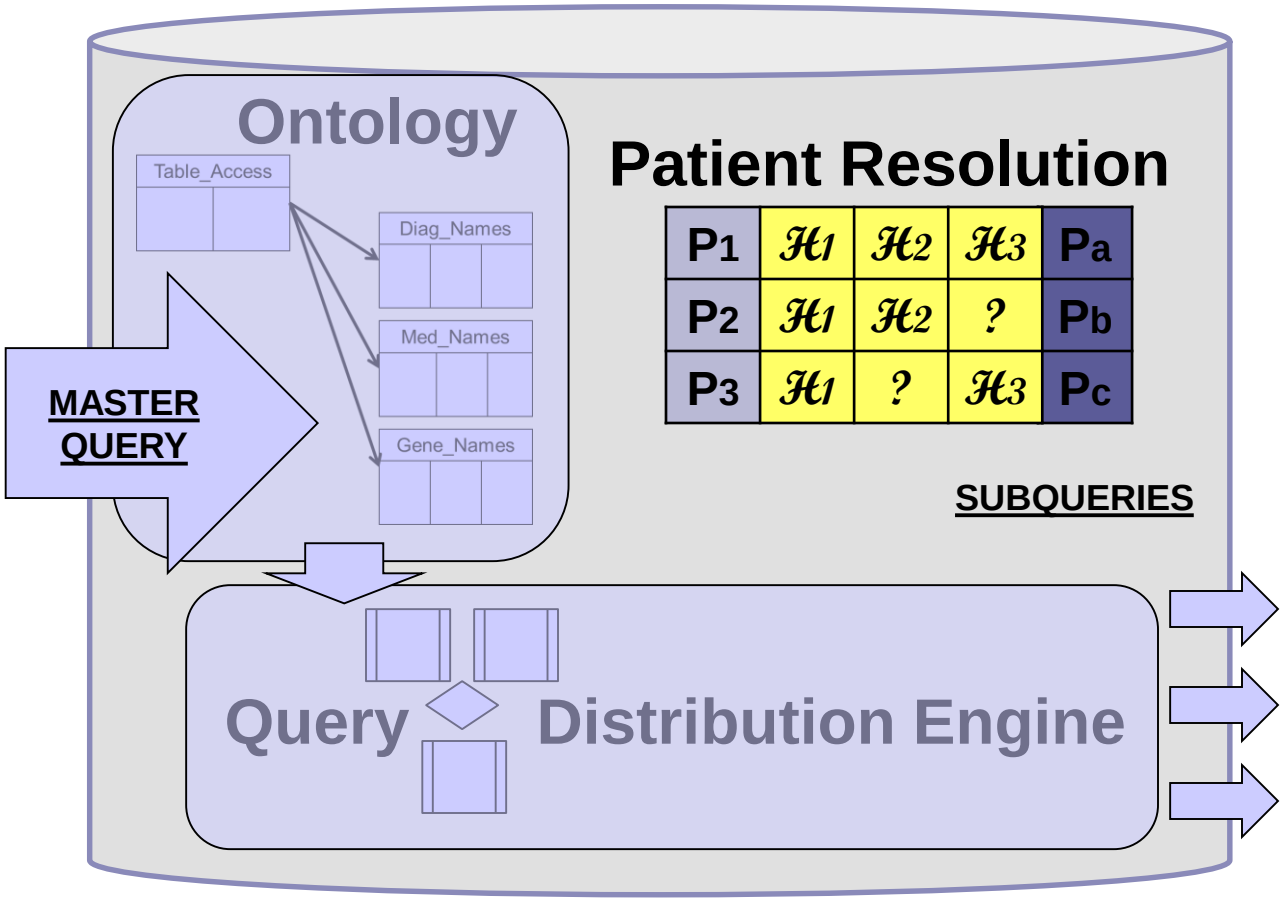
An i2b2 Hive is built from i2b2 Cells which host data "projects"



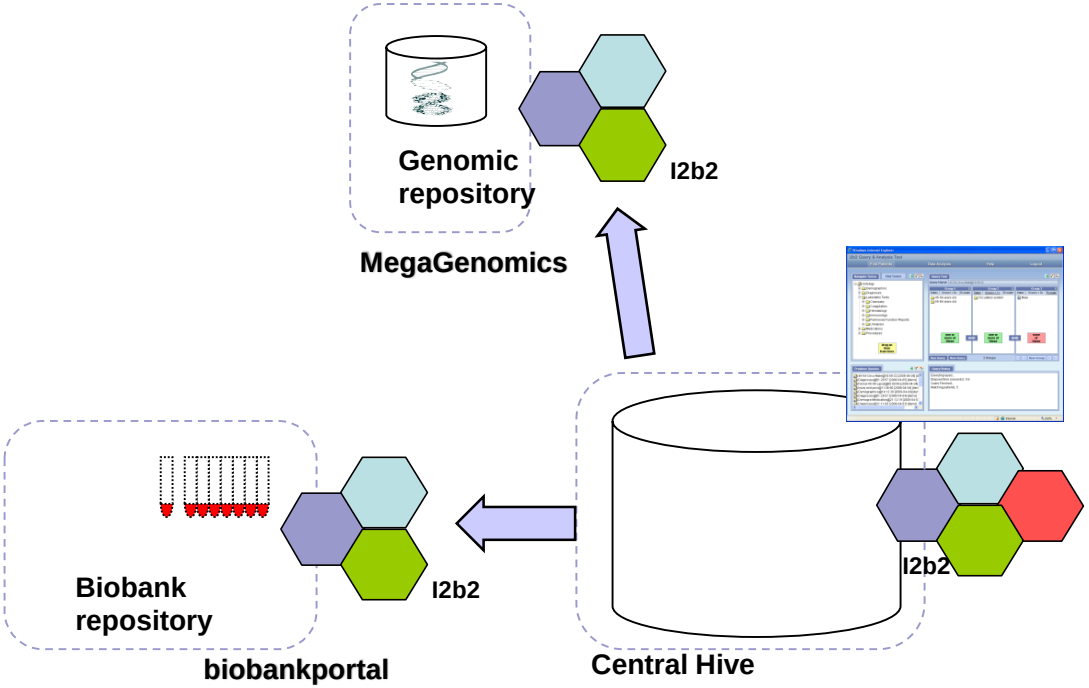
A Patient Information Commons can be built from Specialized i2b2 Hives



How the Central Hive Distributes Queries



Demonstration Hive



PARTNERS HEALTHCARE | BIOBANK PORTAL

Find Patients | Help & Support | Lori Phillips

Navigate Terms | Find

- Biobank Consent Information
- Biobank Demographics
- Biobank Genomics
- Biobank Genomics MEGA Search
- Biobank Health Information Survey
- Biobank Sample Types
- Curated Disease Populations
- Healthcare Data
- Healthy Populations (Controls)

Workplace

- lcp5

Previous Queries | Find

- BD - cu-SNP (PC@14:55:59 [6-20-2016] [lcp5]
- Circulatory sys@13:39:12 [6-16-2016] [lcp5]
- DNA (PC:29465 F@13:34:54 [6-16-2016] [lcp5]
- DISTRIBUTED_QUERY [5-23-2016] [lcp5]

Query Tool

Query Name:

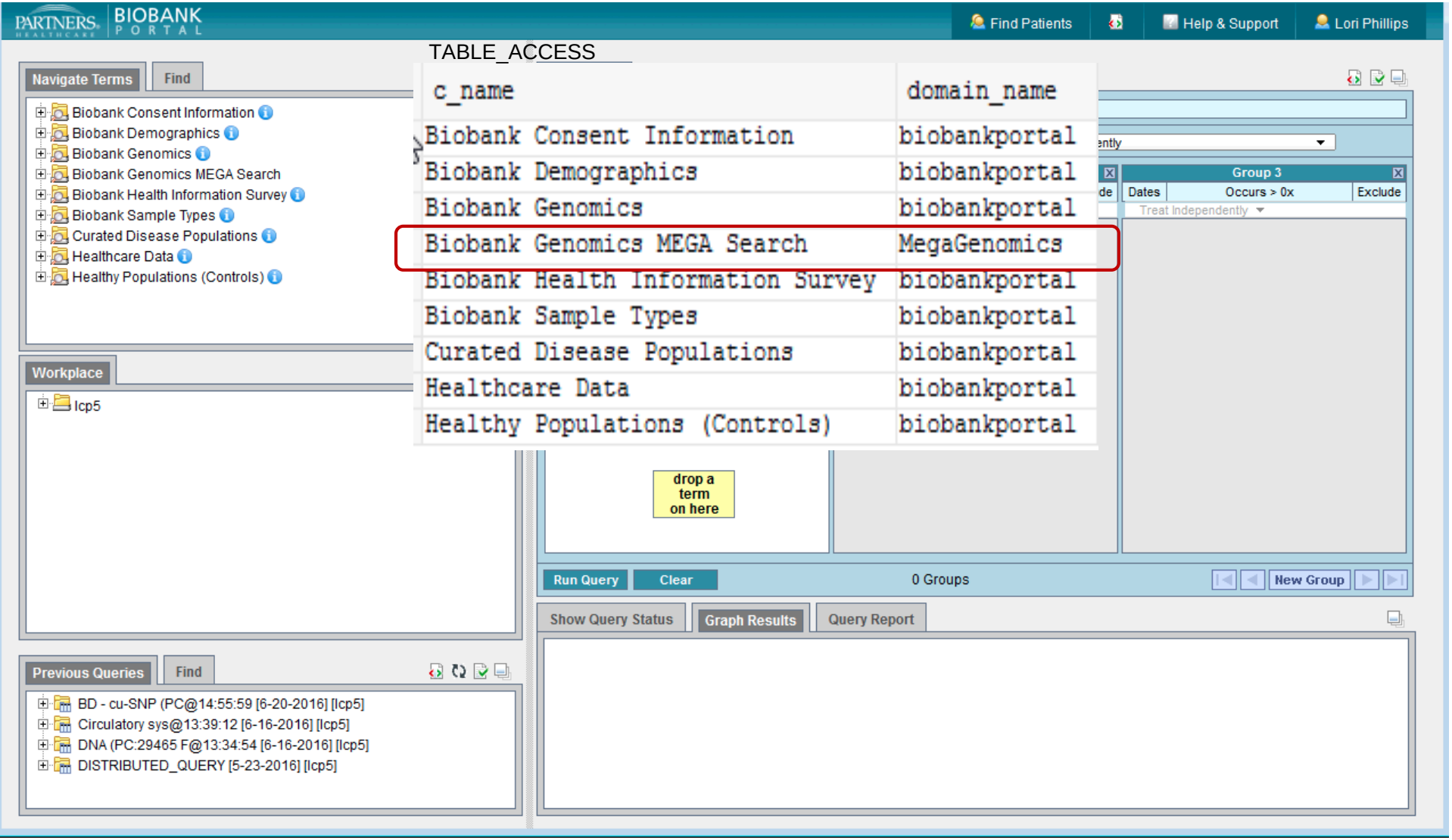
Temporal Constraint: Treat all groups independently

Group 1			Group 2			Group 3		
Dates	Occurs > 0x	Exclude	Dates	Occurs > 0x	Exclude	Dates	Occurs > 0x	Exclude
Treat Independently			Treat Independently			Treat Independently		

drop a term on here

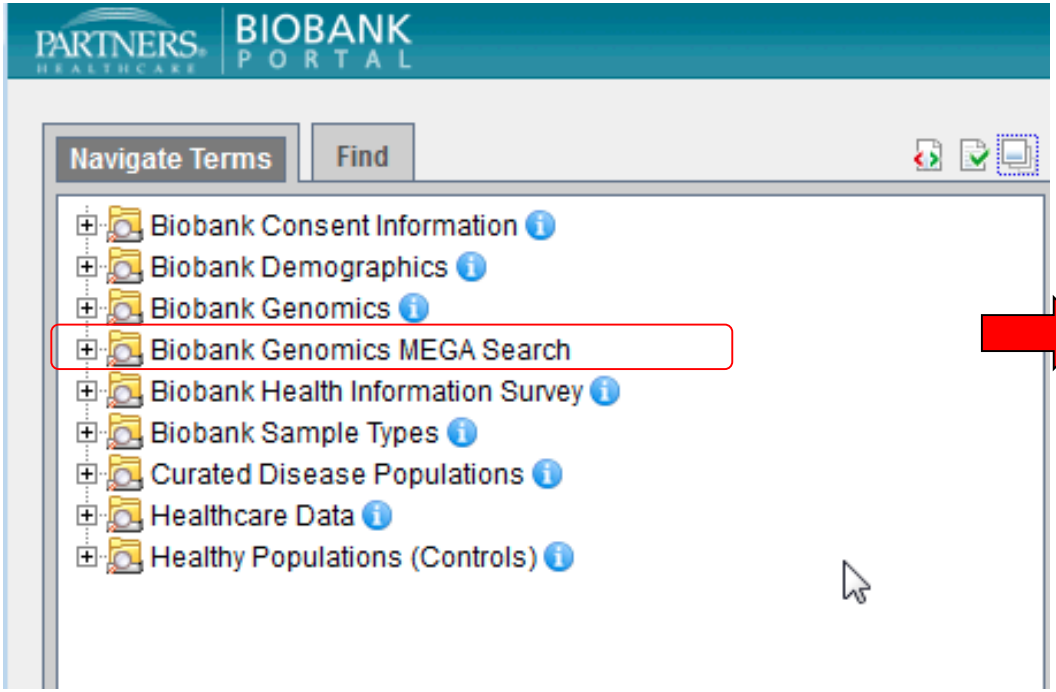
Run Query | Clear | 0 Groups | New Group

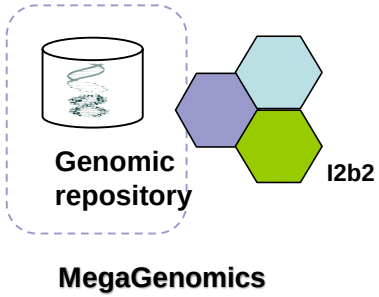
Show Query Status | Graph Results | Query Report

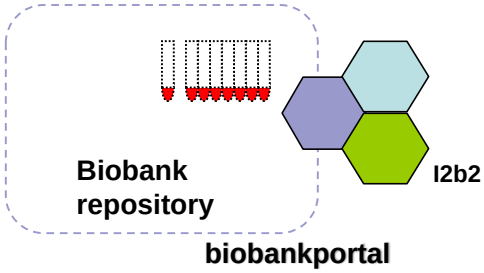


TABLE_ACCESS

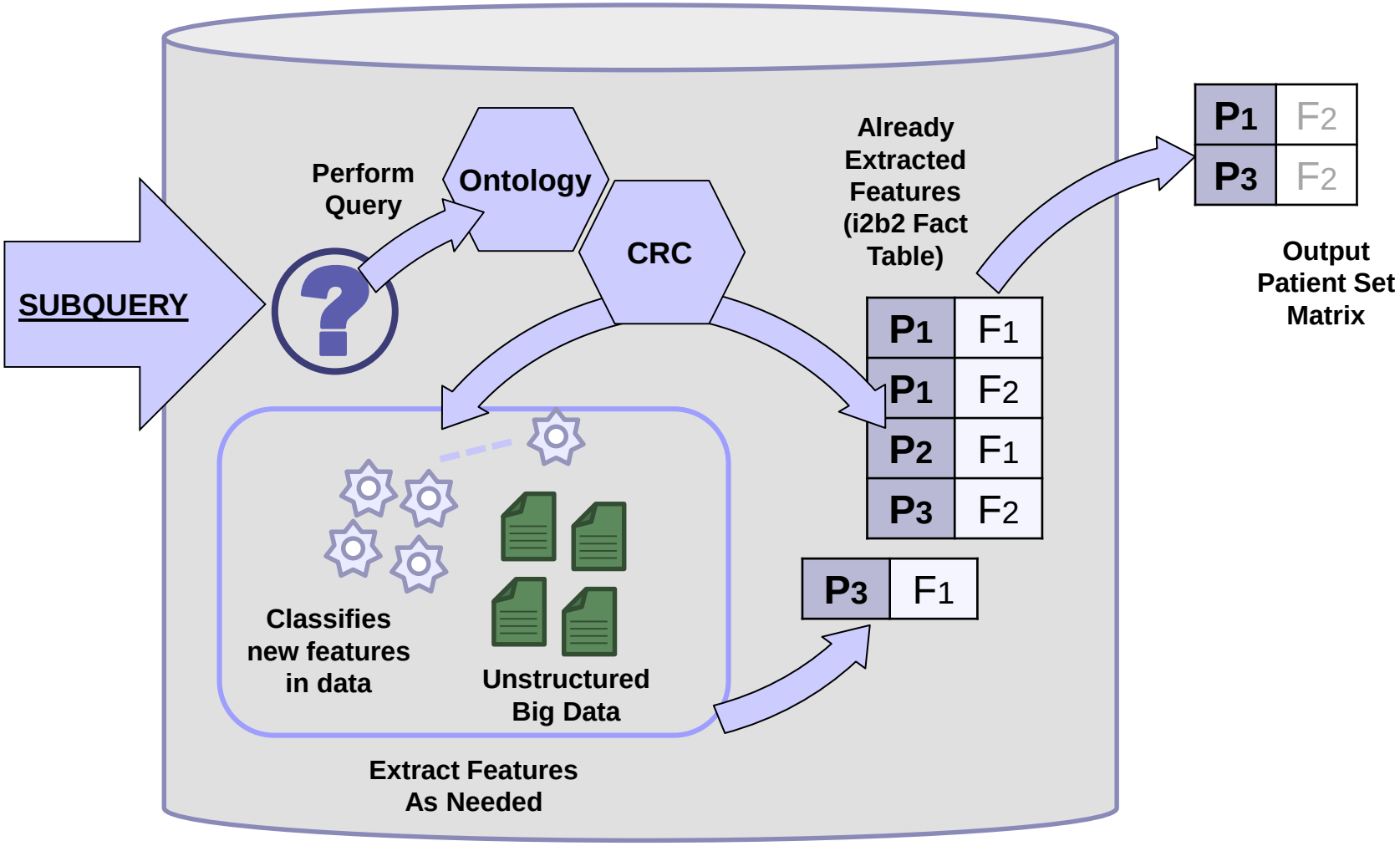
c_name	domain_name
Biobank Consent Information	biobankportal
Biobank Demographics	biobankportal
Biobank Genomics	biobankportal
Biobank Genomics MEGA Search	MegaGenomics
Biobank Health Information Survey	biobankportal
Biobank Sample Types	biobankportal
Curated Disease Populations	biobankportal
Healthcare Data	biobankportal
Healthy Populations (Controls)	biobankportal


PARTNERS HEALTHCARE
BIOBANK PORTAL
Navigate Terms
Find
Biobank Consent Information
Biobank Demographics
Biobank Genomics
Biobank Genomics MEGA Search
Biobank Health Information Survey
Biobank Sample Types
Curated Disease Populations
Healthcare Data
Healthy Populations (Controls)

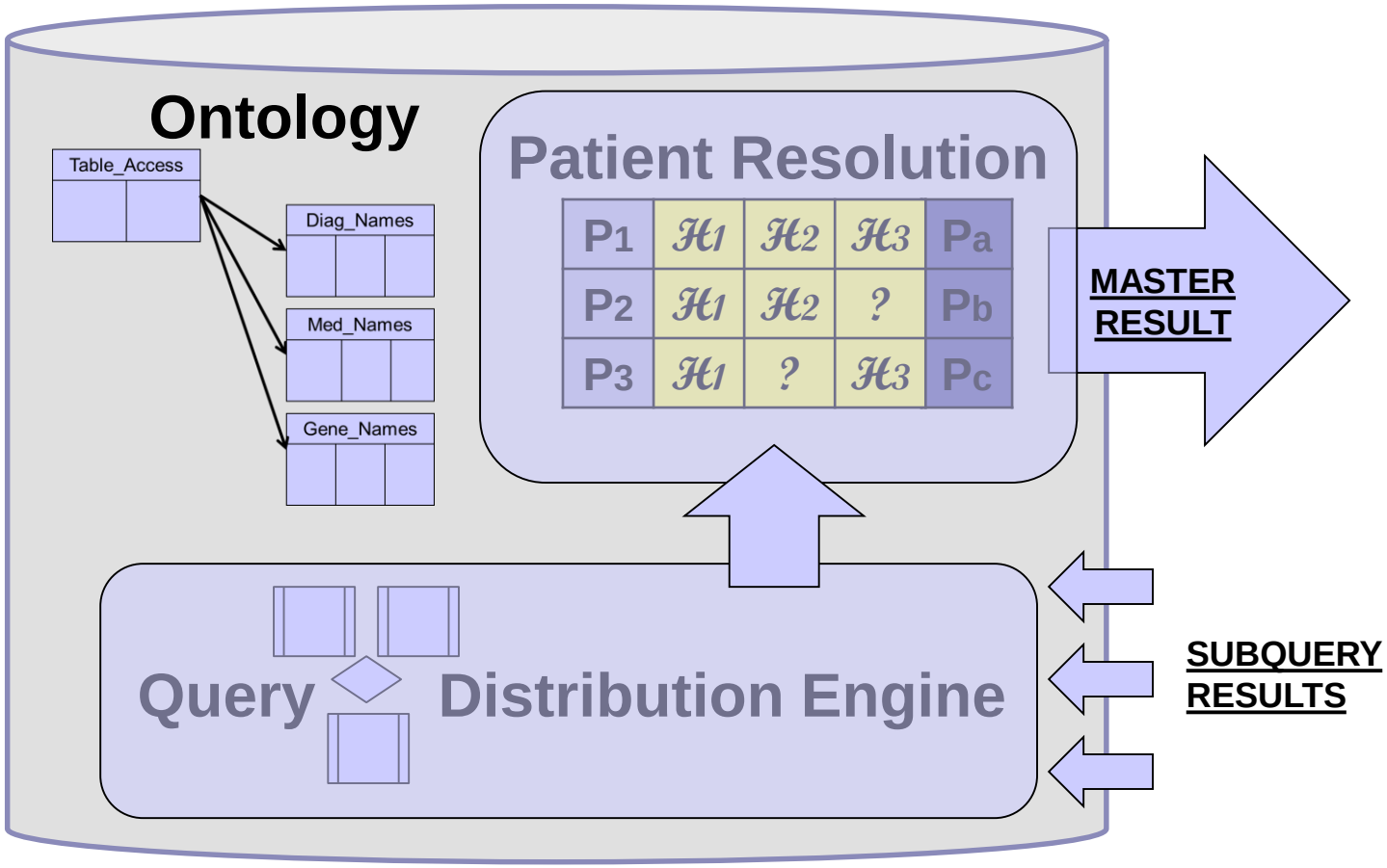

Genomic repository
I2b2
MegaGenomics


Biobank repository
I2b2
biobankportal

How the Remote Hive Returns Queries



How the Central Hive Returns Queries



High-level Overview of Distributed Query Process

1) Query is initiated with regular ontology fields
2) CRC figures out specialized hives to send query from ontology cell
3) Loop processes query into a compendium of patient lists
   a. Panels of items get sent to Remote PICI for querying
   b. Remote PICI sends back a list of patient identifiers
   c. Remote patient identifiers are read back into the Central PICI and mapped to centralized patient identifiers
   d. Set logic occurs (unions/intersections)
   e. Loop continues
4) Finally, the central node completes reduction into a single patient set
5) The patient set is then computed upon by the Central PICI or sent to a Remote PICI for an analysis

- Try it at:
  - Central PICI: http://52.6.250.114/webclient/
  - Remote child 1: http://54.152.187.58/webclient
  - Remote child 2: http://54.152.144.45/webclient

Optimizing Distributed Query Processing

- Optimization will continue to evolve as we roll this out
- Many levers to play with
  - Serial vs. Parallel vs Map/Reduce
  - Combining patient sets that come back
    - Memory
    - Database
    - Hybrid
  - Order of processing
    - Optimize operations
    - Reliance on statistics from remote nodes
    - Learn from previous queries
  - Patient mapping
  - Reducing data sent back and forth
    - Reduced API

QUESTIONS?