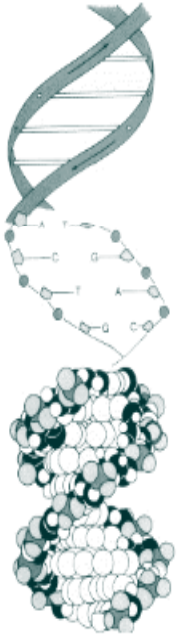


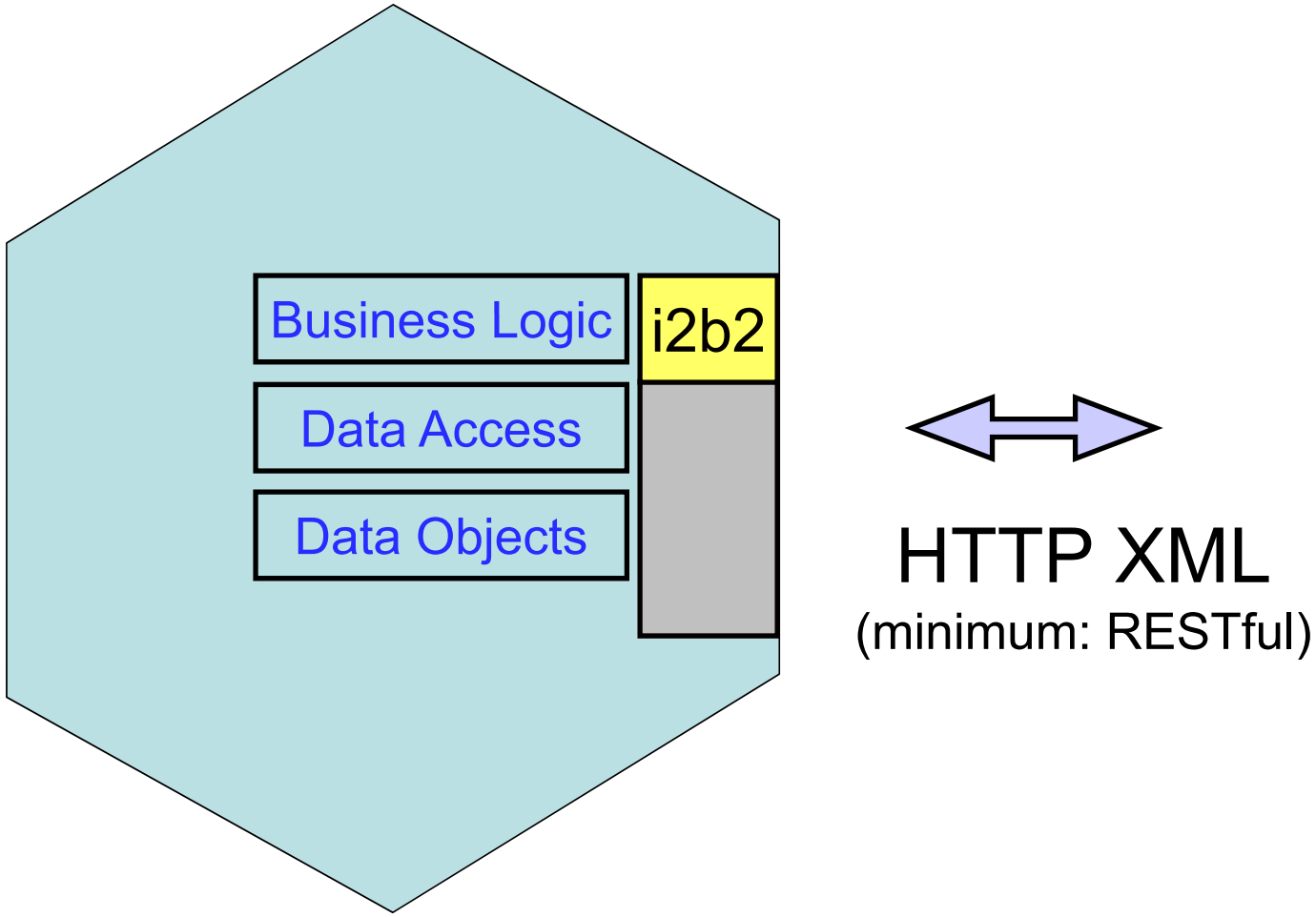
A Translational Engine at the National Scale: i2b2

Shawn Murphy MD, Ph.D.

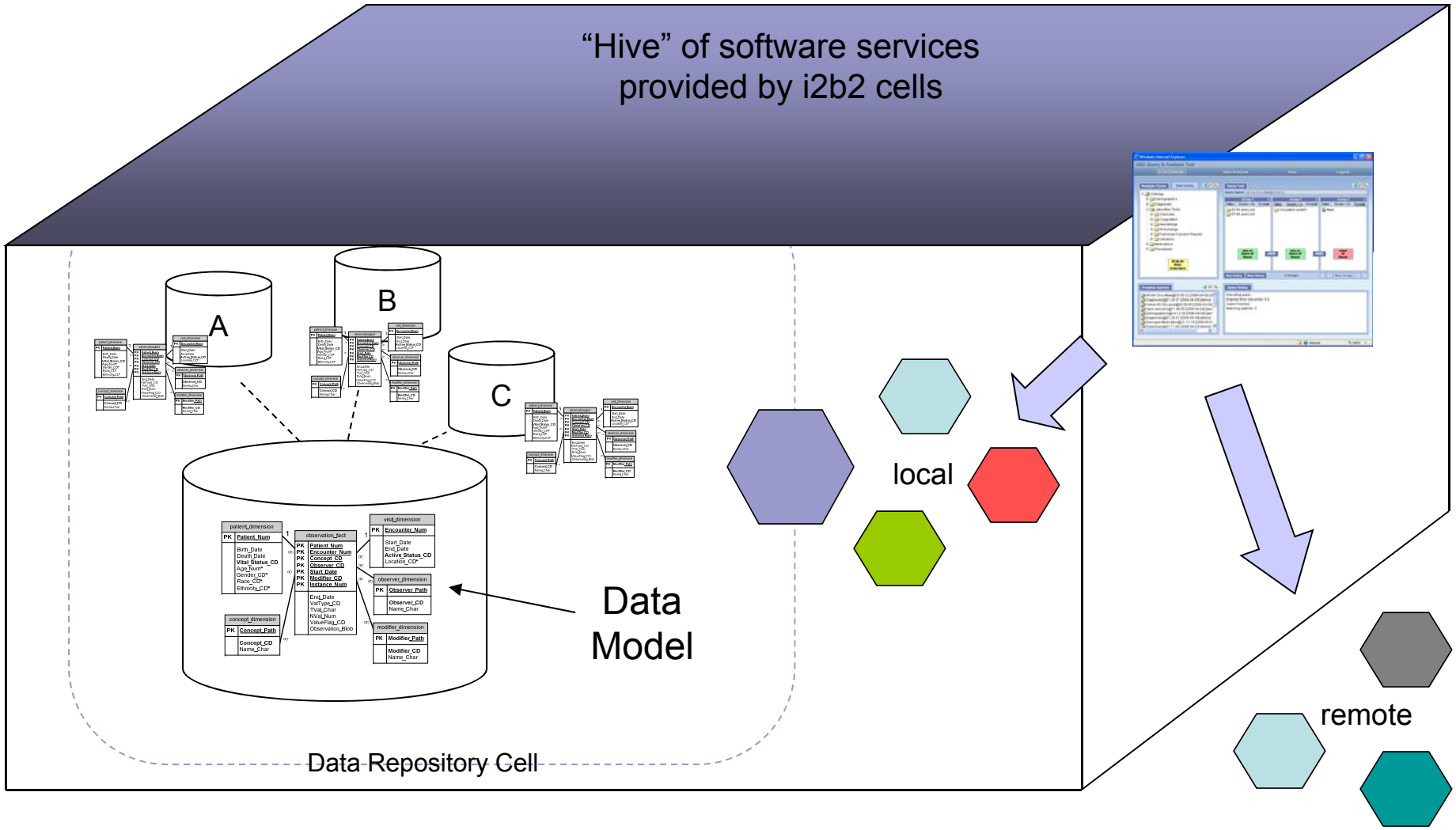
The National Center for Biomedical Computing entitled Informatics for Integrating Biology and the Bedside (i2b2), what is it?

- Software for explicitly organizing and transforming person-oriented clinical data to a way that is optimized for clinical genomics research
  - Allows integration of clinical data, trials data, and genotypic data
- A portable and extensible application framework
  - Software is built in a modular pattern that allows additions without disturbing core parts
  - Available as open source at https://www.i2b2.org

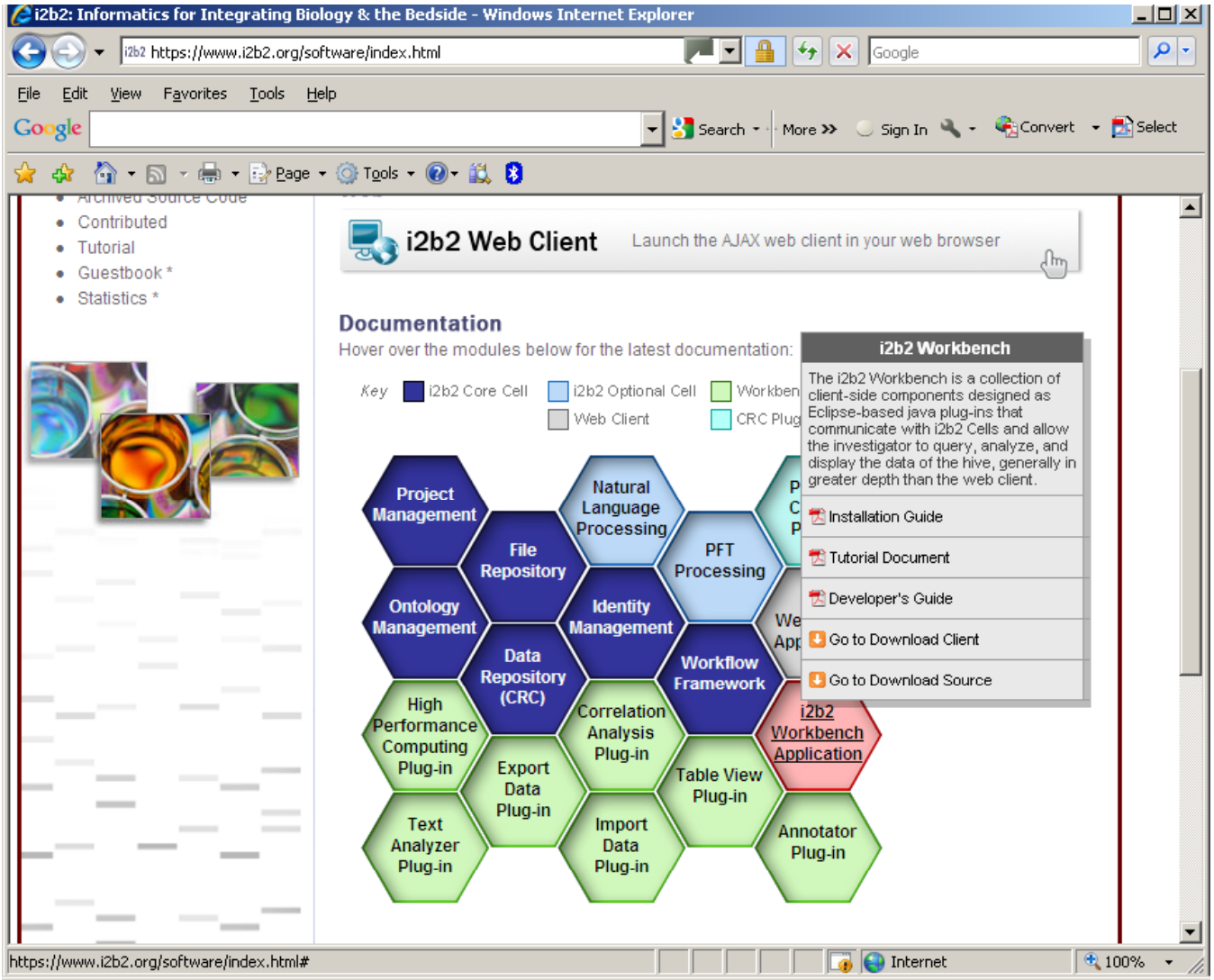
i2b2 Cell: The Canonical Software Module



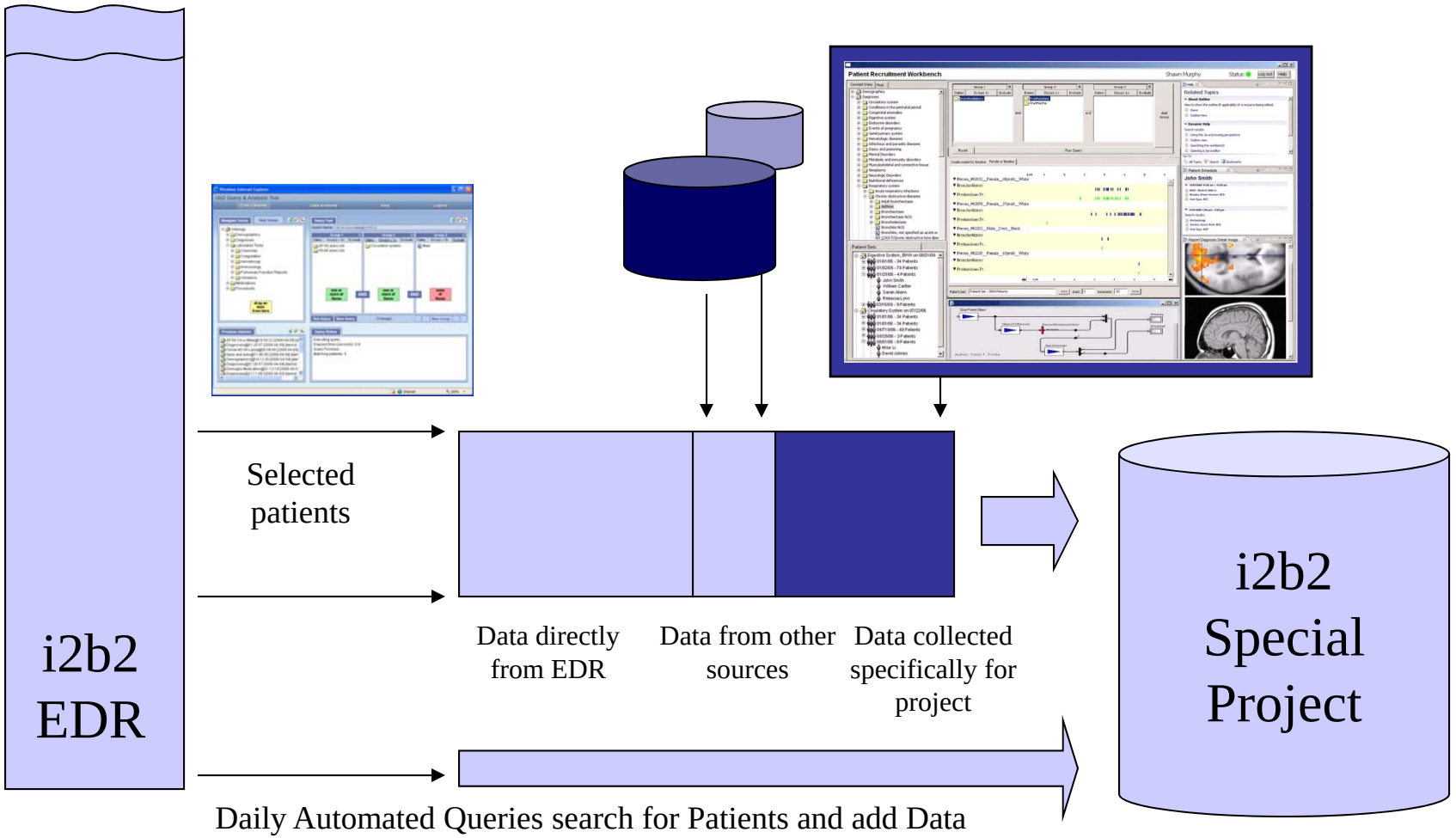
An i2b2 Environment (the Hive) is built from i2b2 Cells



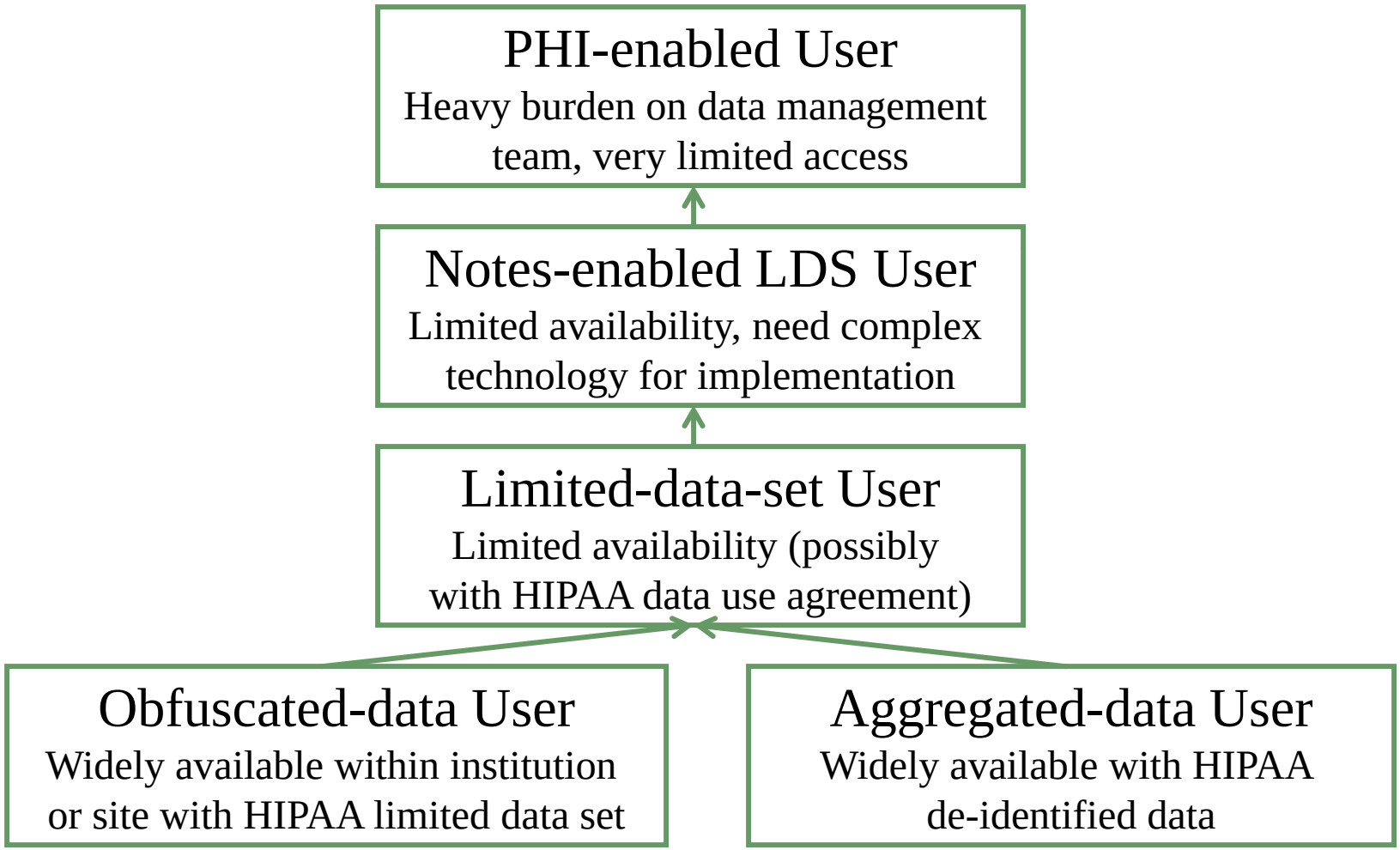
I2b2 Software components are distributed as open source



Perform deep studies with patient sets selected from Enterprise Data Repository



Privacy Levels in i2b2



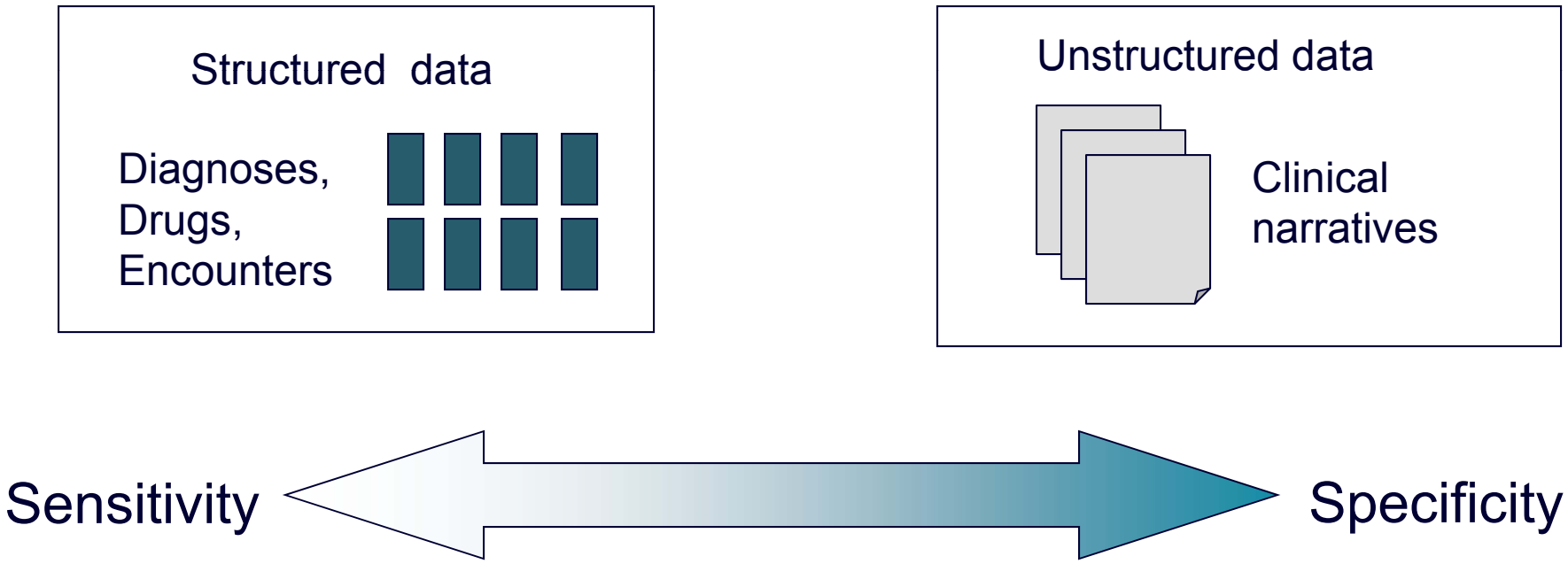
Can We Trust the Phenotypes?

Validation Study (N = 185)

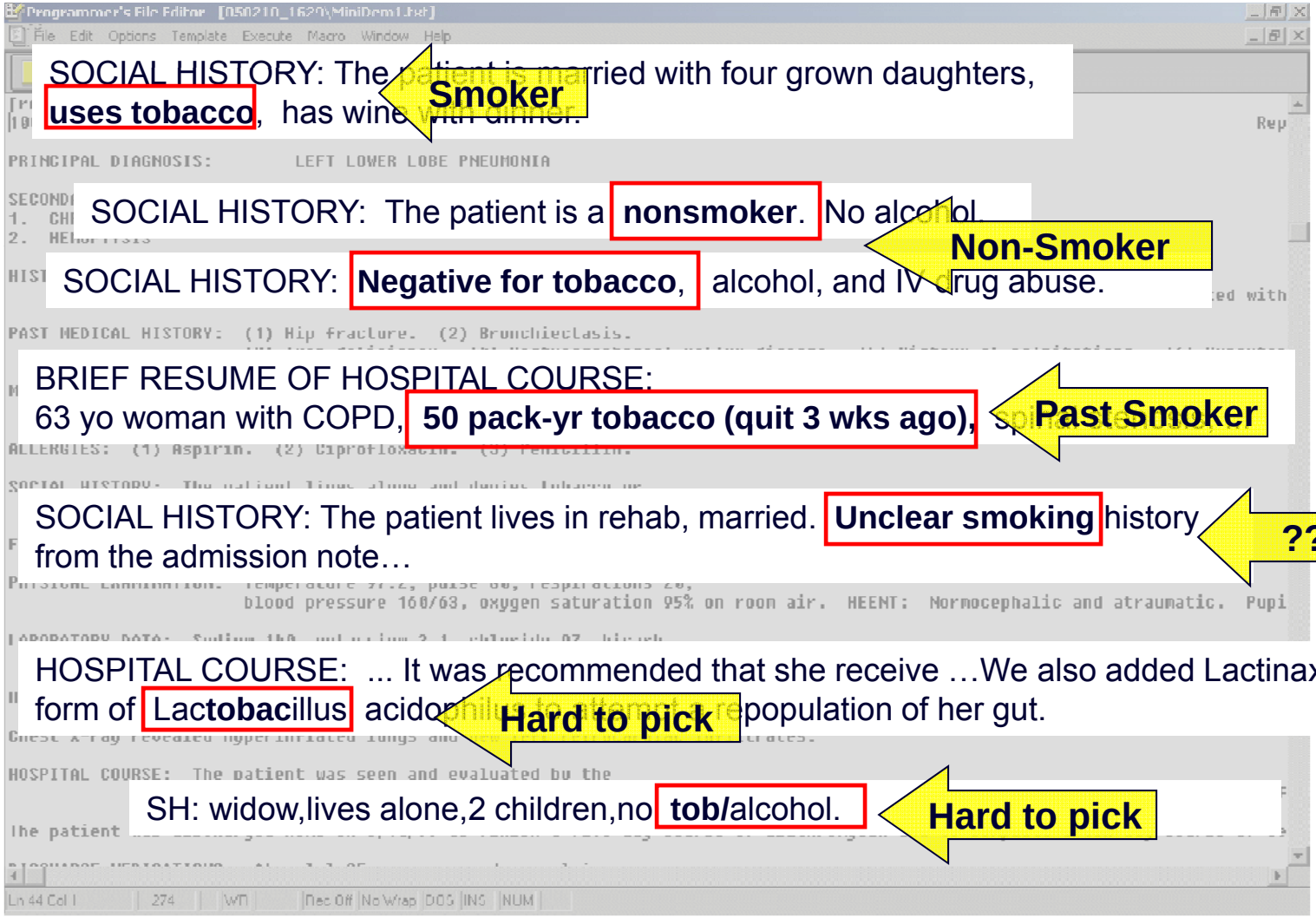
- Evaluate case and control algorithms compared to gold standard of diagnostic interview by expert clinician
- Recruit cases and controls as defined by informatics algorithm
- Interview by clinicians blinded to ascertainment group
- Recruited patients with depression or schizophrenia to enhance blinding

Jordan Smoller MD, ScD and team

Sensitivity vs. Specificity



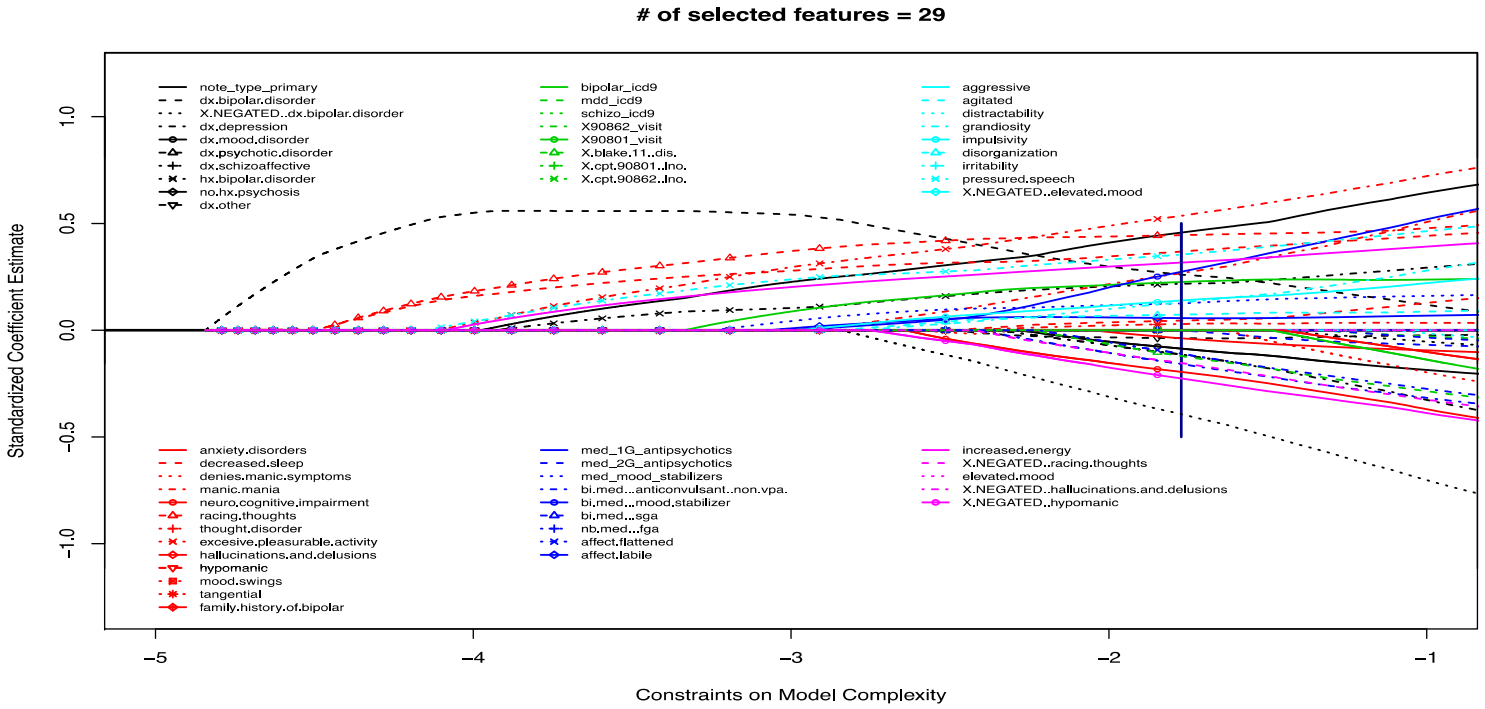
Natural Language Processing



Train classification algorithms

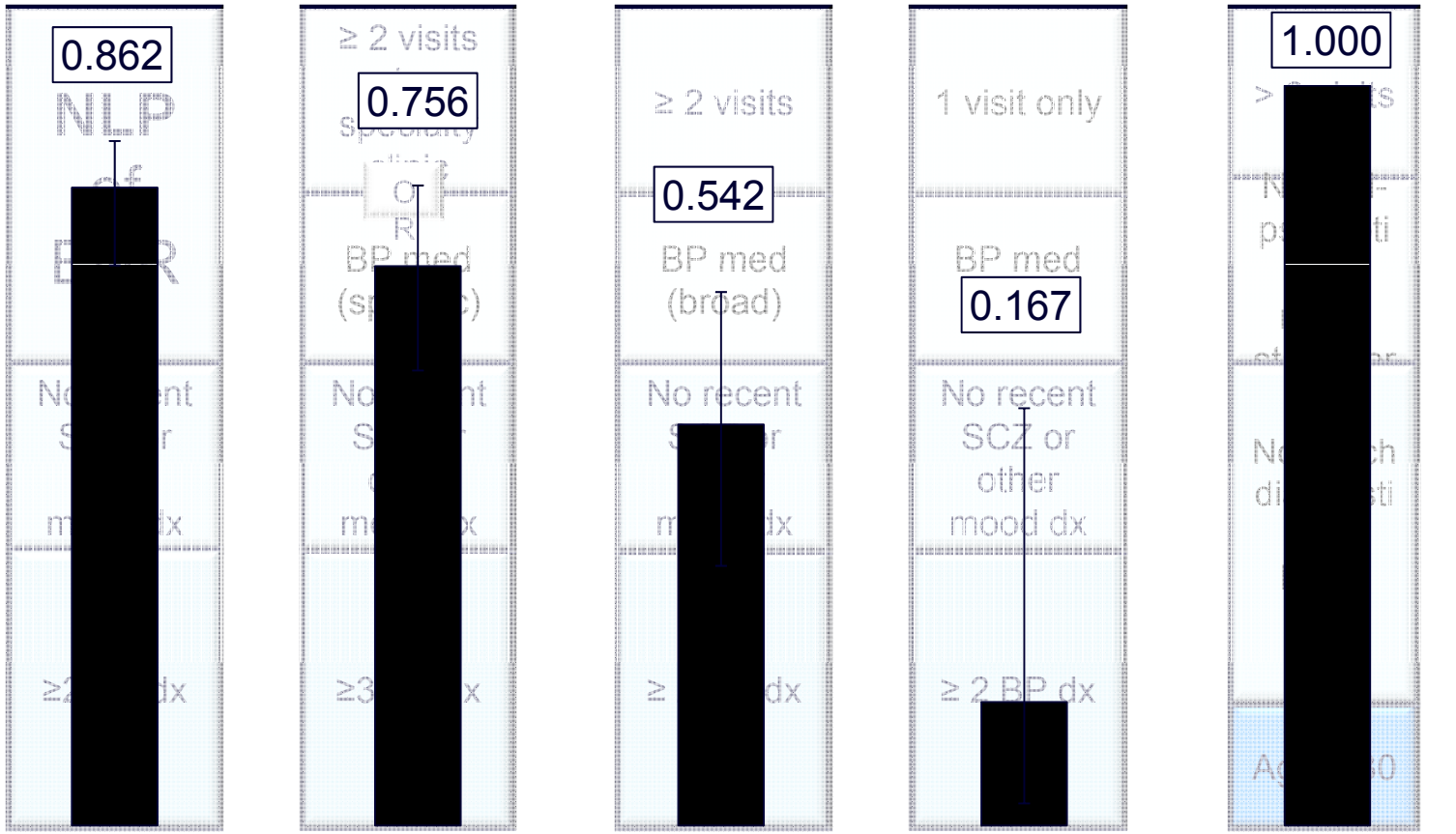
1. Over 300 words/phrases (features) were identified using chart review
2. Important features were selected for model using adaptive LASSO shrinkage

Tianxi Cai PhD and team



Bipolar Cohort Ascertainment

Cases - Positive Predictive Values



NLP
Natural Language Processing

Bipolar Clinic/Med

MRP
Most Recent & Predominant

MRP-SV
Most Recent & Predominant, Single Visit

Healthy Controls

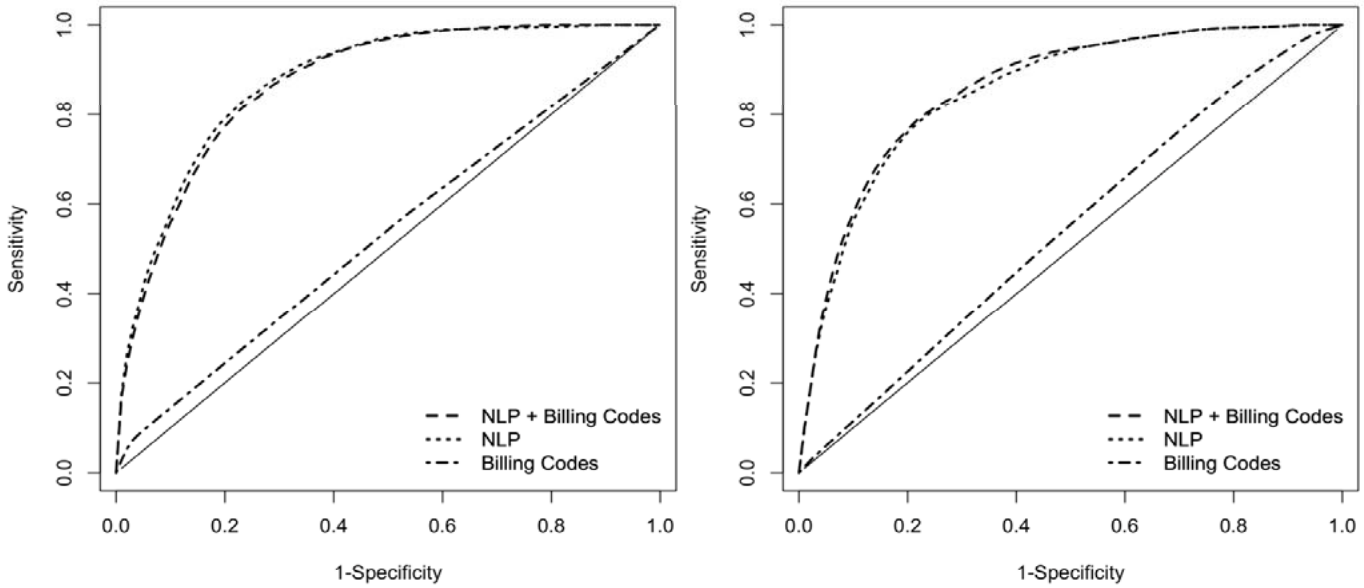
Psychological Medicine (2012), **42**, 41–50. © Cambridge University Press 2011
doi:10.1017/S0033291711000997

ORIGINAL ARTICLE

Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model

R. H. Perlis[1,2*], D. V. Iosifescu[1,3], V. M. Castro[4], S. N. Murphy[5], V. S. Gainer[4], J. Minnier[6], T. Cai[6], S. Goryachev[4], Q. Zeng[7], P. J. Gallagher[2], M. Fava[1], J. B. Weilburg[1], S. E. Churchill[8], I. S. Kohane[9] and J. W. Smoller[2]

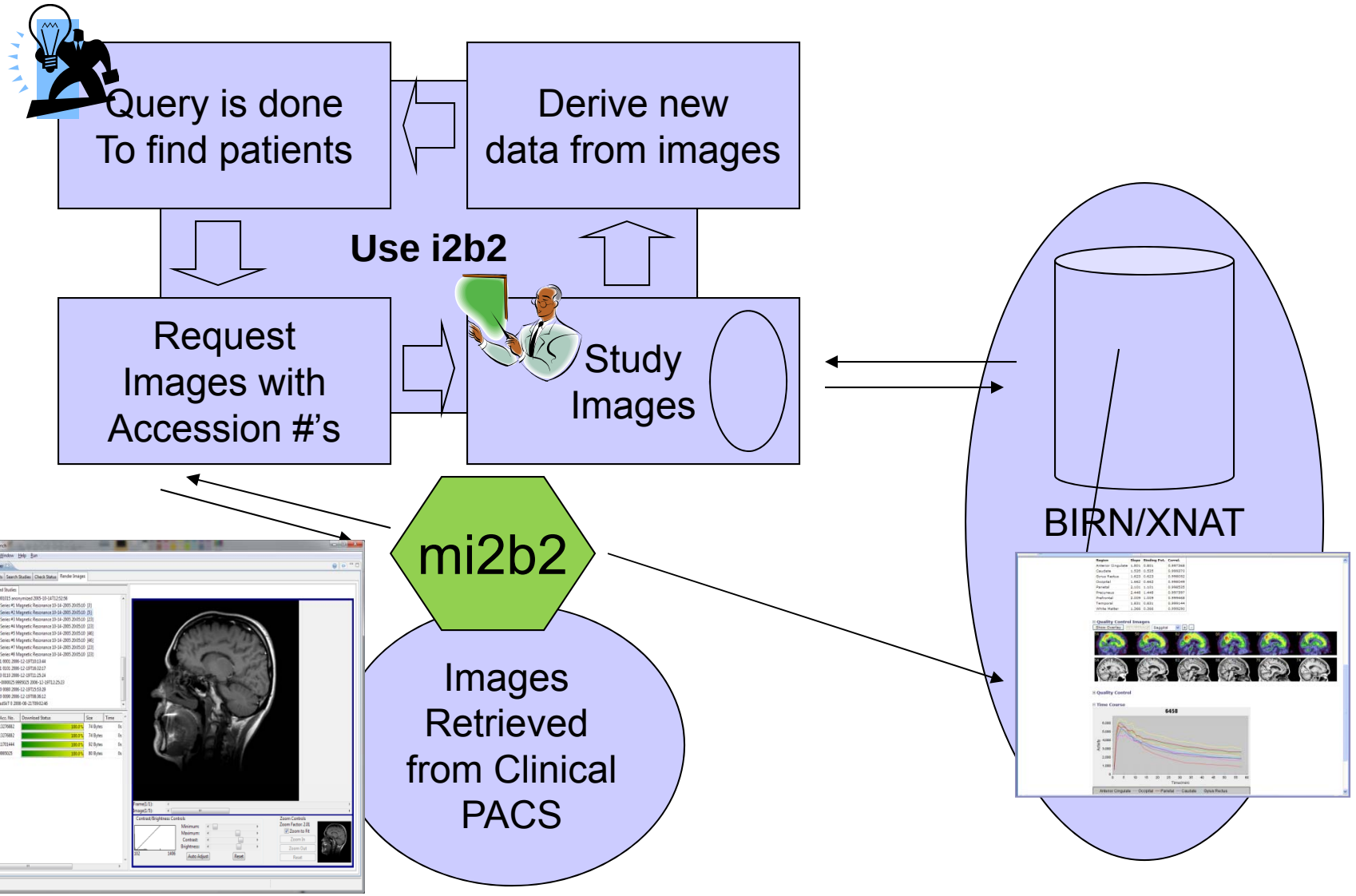
Use NLP to define cohorts of treatment-resistant and treatment-responsive depression



Specificity: 95%
AUC > 85%

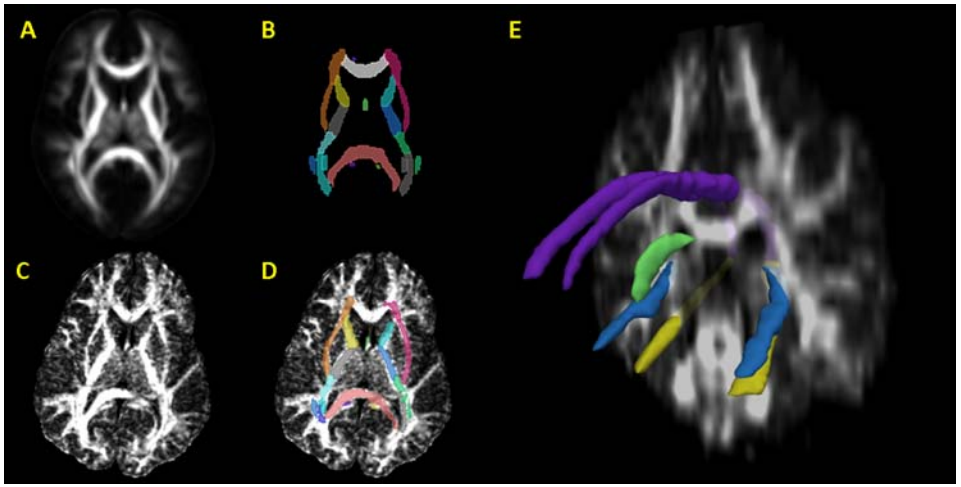
Clinical Status	Model	Specificity	Sensitivity	Precision	AUC
Depressed	Billing Codes	0.95	0.09 (0.03)	0.57 (0.14)	0.54 (0.02)
Depressed	NLP	0.95	0.42 (0.05)	0.78 (0.02)	0.88 (0.02)
Depressed	NLP + Billing Codes	0.95	0.39 (0.06)	0.78 (0.02)	0.87 (0.02)
Well	Billing Codes	0.95	0.06 (0.02)	0.26 (0.27)	0.55 (0.03)
Well	NLP	0.95	0.37 (0.06)	0.86 (0.02)	0.85 (0.02)
Well	NLP + Billing Codes	0.95	0.39 (0.07)	0.85 (0.02)	0.86 (0.02)

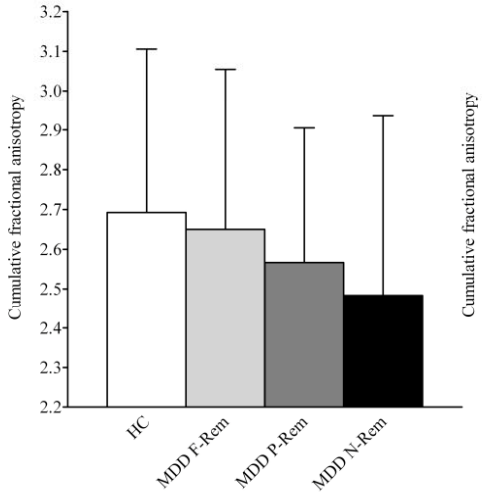
Medical Imaging Cell (mi2b2)

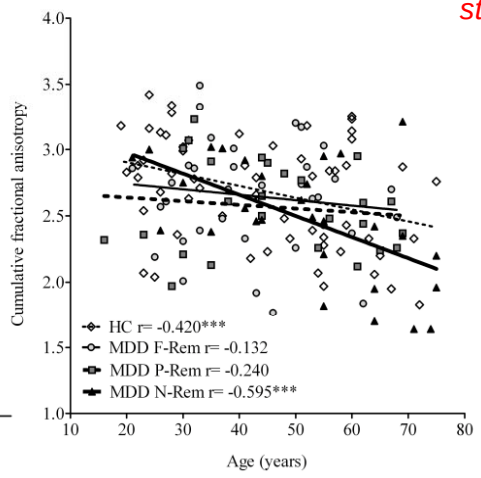


White matter abnormalities associated with treatment-resistant depression

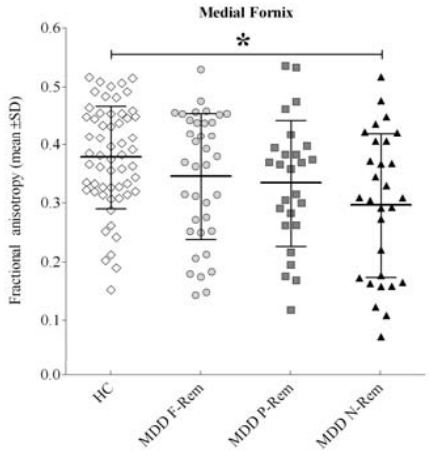
- Scans collected as part of routine clinical care
- Diffusion tensor imaging in 150 pts
- Age-related decline in white matter integrity increases with treatment resistant depression







Medial fornix shows strongest effect



Hoogenboom et al. World J Biol Psychiatry, 2012

Rapid investigation of QTc prolongation

- *FDA warning 2011 for Celexa*

Safety Announcement:
[8-24-2011] "should no longer be used at doses greater than 40 mg per day because it can cause abnormal changes in the electrical activity of the heart."

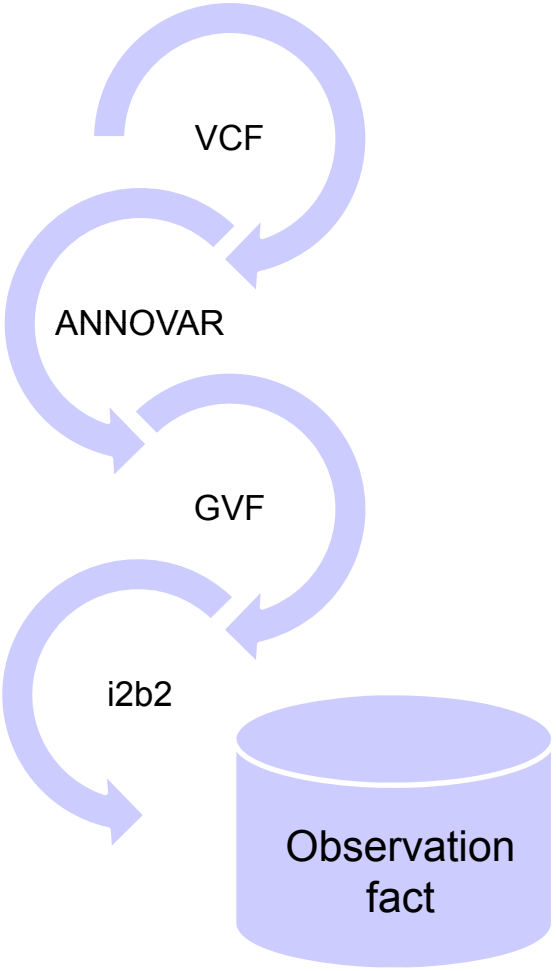
- *But, did NOT include Lexapro (which is active ingredient of Celexa [s-enantiomer])*
- *Shown to be true with RPDR-derived data set with >38,000 EKGs obtained within 14 – 90 day window after medication initiated*

	Adjusted model†	
Anti-depressant	**prolongation**	**p-value**
SSRI		
Citalopram (Celexa)	2.85	0.004
Escitalopram (Lexapro)	3.80	< 0.001
Fluoxetine (Prozac)	1.44	0.150
Paroxetine (Paxil)	0.07	0.943
Sertraline (Zoloft)	0.87	0.383
Other anti-depressants		
Amitriptyline	4.10	< 0.001
Bupropion	-2.15	0.032
Duloxetine	0.60	0.547
Mirtazapine	-1.46	0.145
Nortriptyline	1.23	0.219
Venlafaxine	1.15	0.251
previously known prolonger		
Methadone	5.32	< 0.001

† Adjusted for age, gender, race, type of insurance, history of major depression, history of myocardial infarction and Charlson comorbidity score

Roy Perlis MD, MSc and team

Importing NGS variant output into i2b2

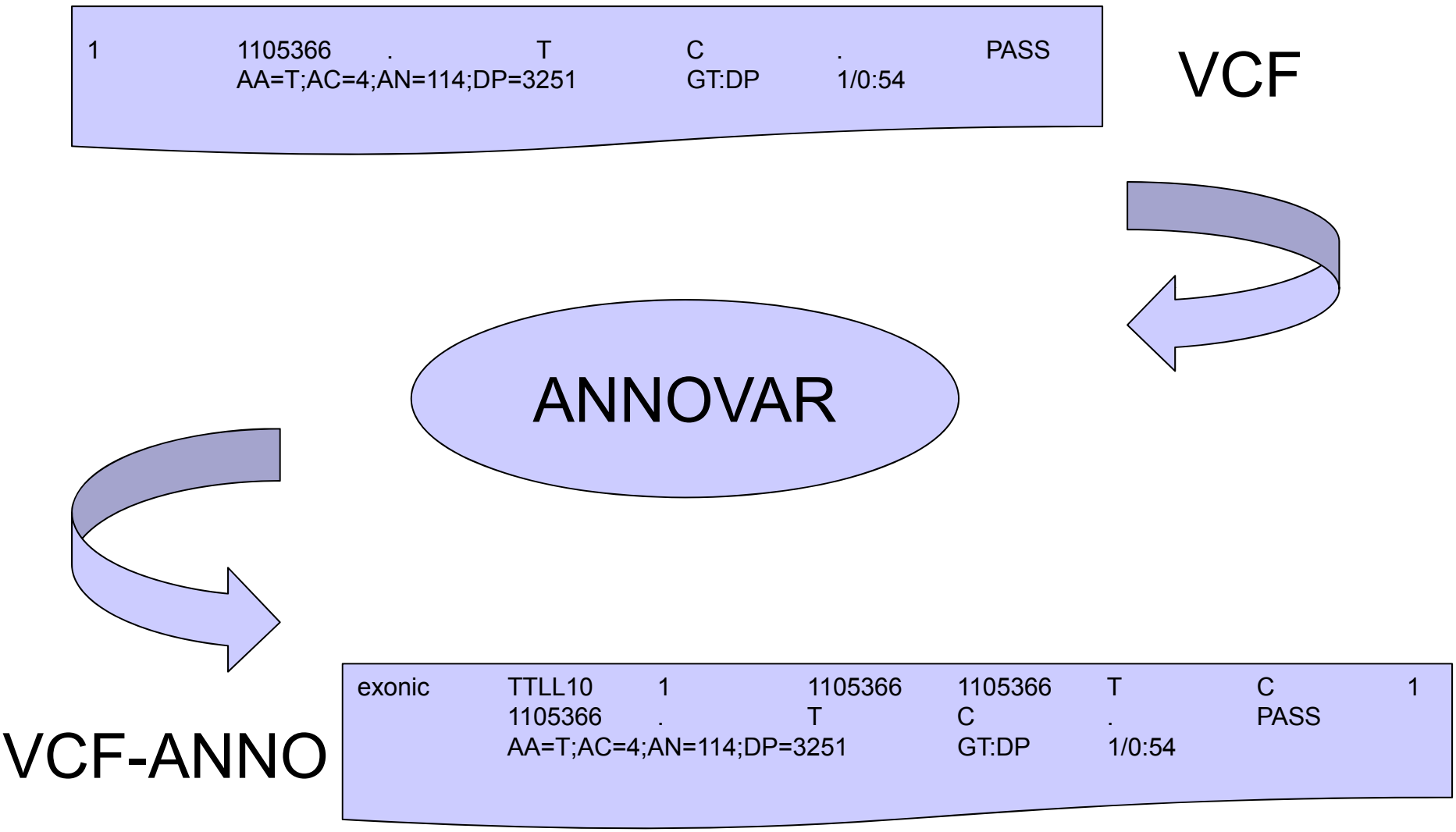


Variant Call Format

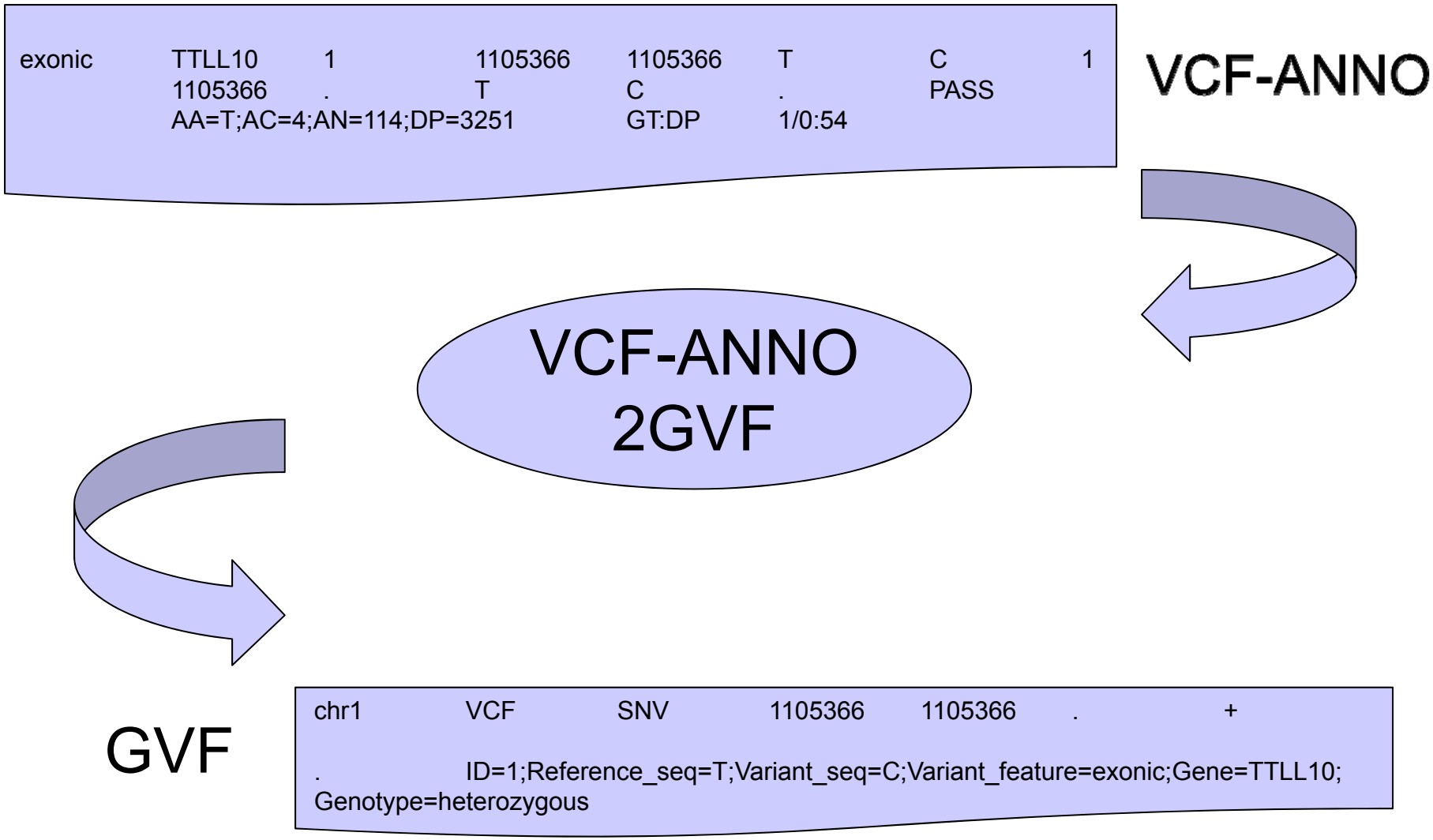
Gene Annotated VCF

Genome Variation Format

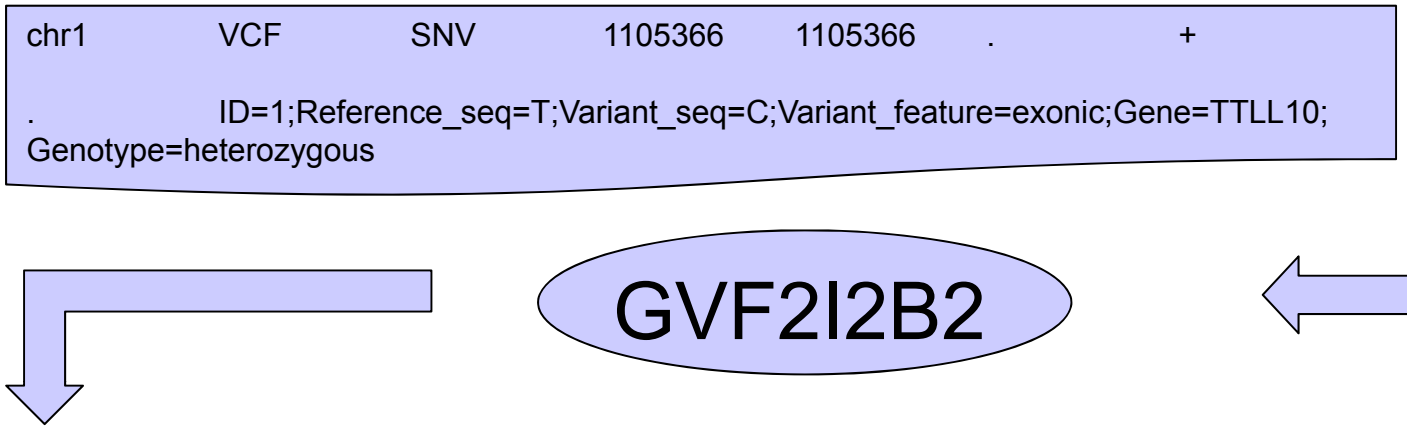
Pipeline - VCF to VCF-ANNO



Pipeline - VCF-ANNO to GVF

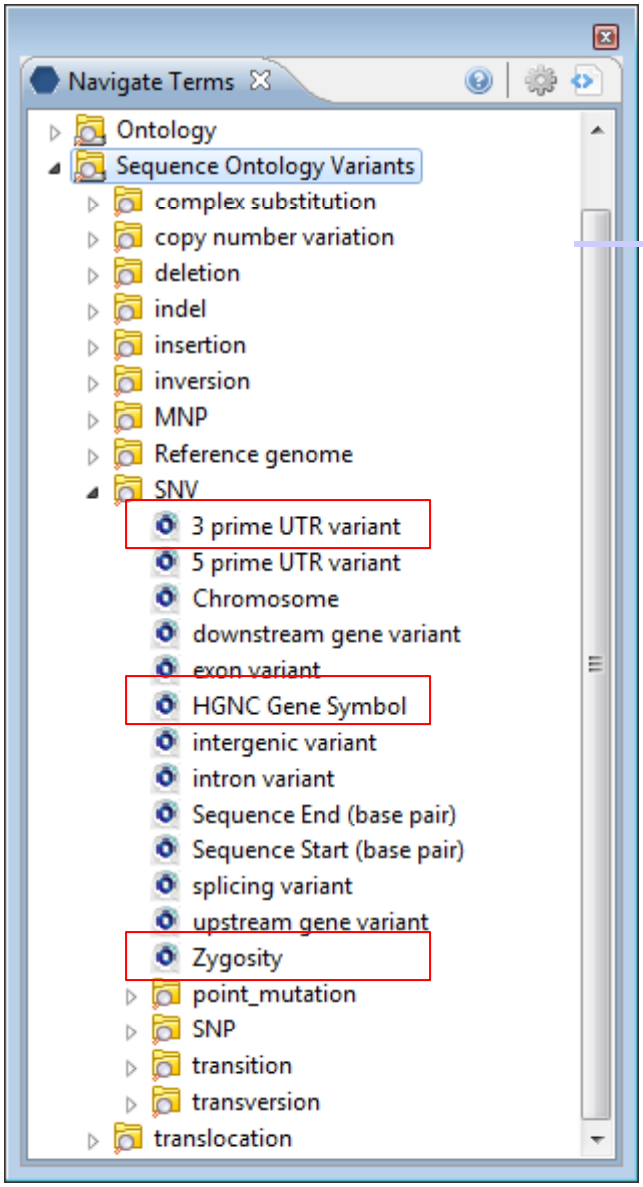


Pipeline – GVF to I2B2 records



- 1880001024|1000000024|"SO:0001483"|"@"|"2010-03-03 00:00:00"|"@"|1|||||||||||||"GVF2I2B2"|
- 1880001024|1000000024|"SO:0001483"|"@"|"2010-03-03 00:00:00"|"SO:0000340"|1|"T"|
"chr1"|||||||||||"GVF2I2B2|
- 1880001024|1000000024|"SO:0001483"|"@"|"2010-03-03 00:00:00"|"SEQ:Start"|1|"N"|"E“|
1105366|||||||||||"GVF2I2B2|
- 1880001024|1000000024|"SO:0001483"|"@"|"2010-03-03 00:00:00"|"SEQ:End"|1|"N"|"E"|
1105366|||||||||||"GVF2I2B2|
- 1880001024|1000000024|"SO:0001483"|"@"|"2010-03-03 00:00:00"|"SO:0001029"|1|"T"|
"+"|||||||||||"GVF2I2B2”|
- 1880001024|1000000024|"SO:0001483"|"@"|"2010-03-03 00:00:00"|"SEQ:Zygosity"|1|"T"|
"heterozygous"|||||||||||"GVF2I2B2”|
- 1880001024|1000000024|"SO:0001483"|"@"|"2010-03-03 00:00:00"|"SEQ:HUGO"|1|"T"|
"TTLL10"|||||||||||"GVF2I2B2"|
- 1880001024|1000000024|"SO:0001483"|"@"|"2010-03-03 00:00:00"|"SO:0001791"|1||
|||||||||||"GVF2I2B2"|

Navigating NGS Variant Data with Sequence Ontology



Combination of concepts and modifiers to identify:

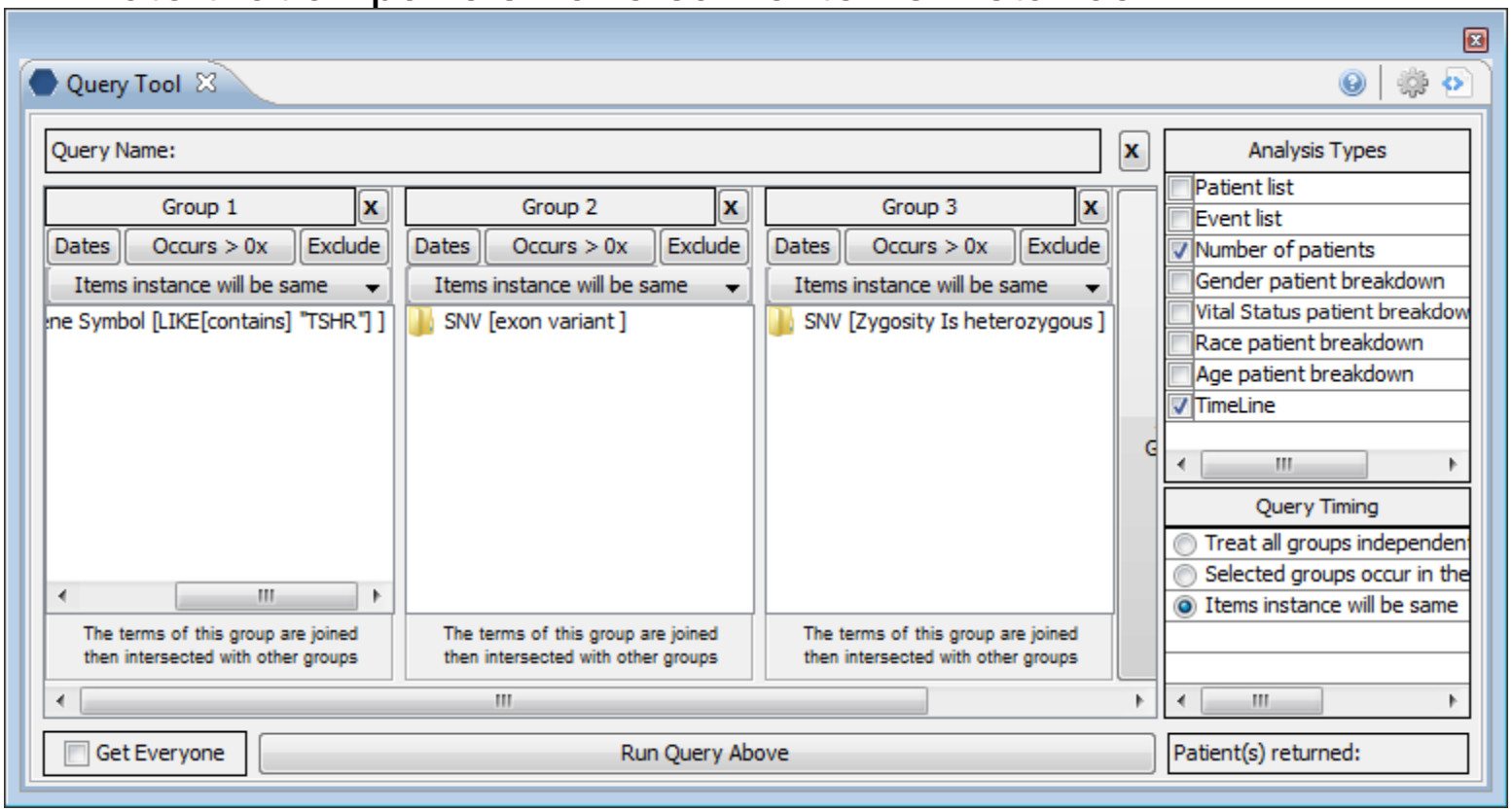
An SNV located on a 3'UTR

An SNV associated with a certain gene

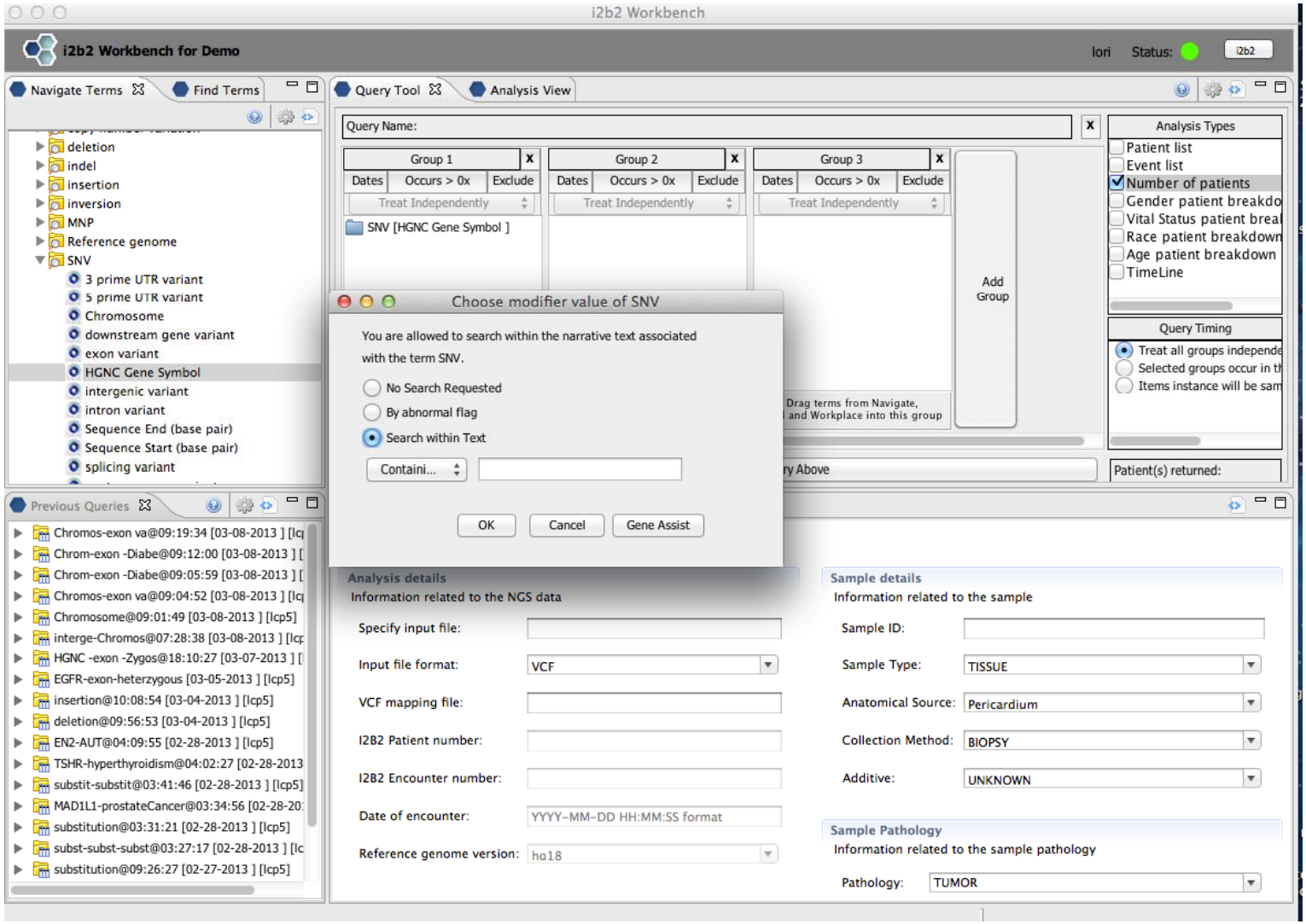
An SNV of specified zygosity

Querying NGS Variant Data

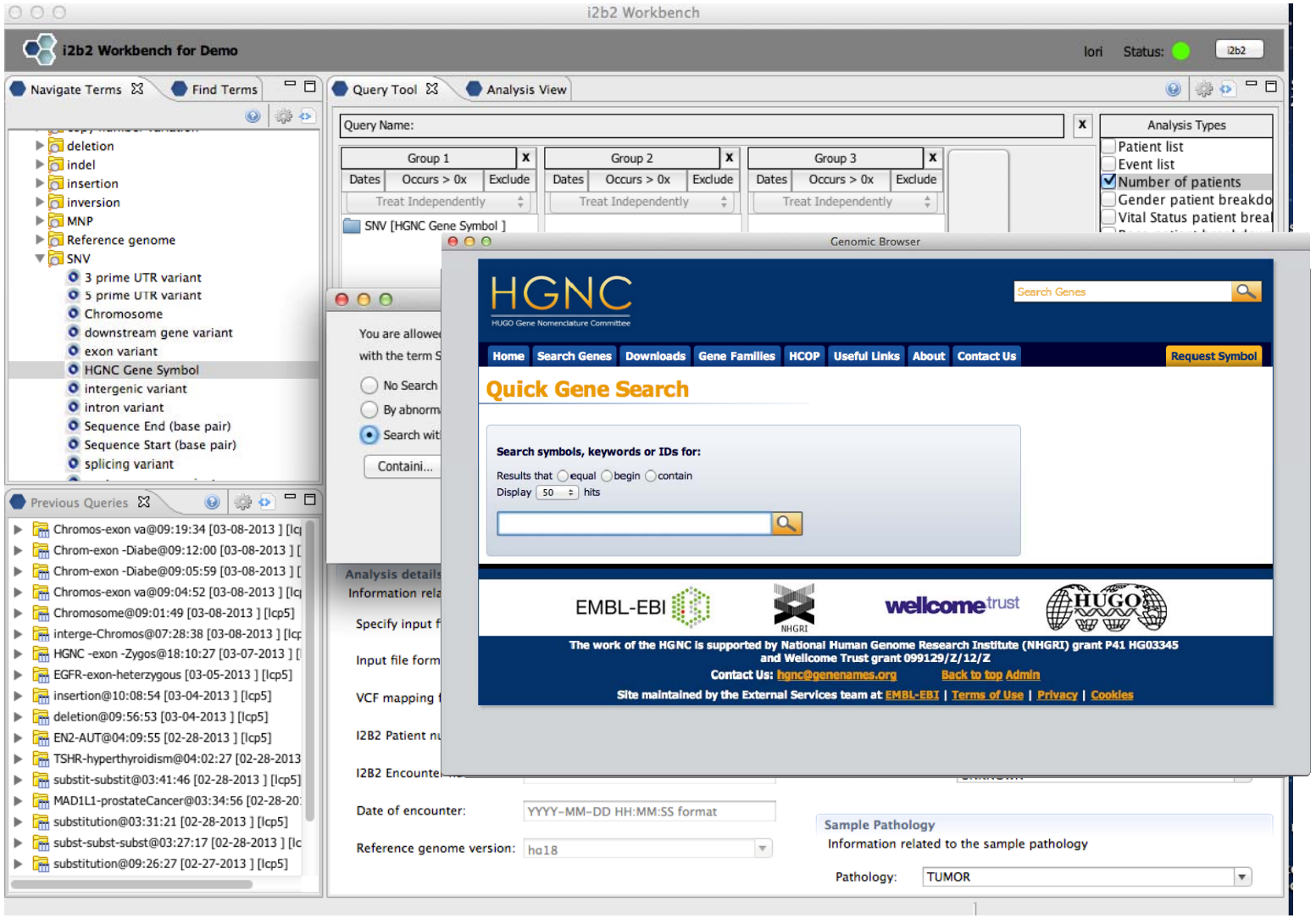
- Querying for a heterozygous SNV on an exon of gene TSHR
  - Note that all panels have same items instance



Gene Assist


i2b2 Workbench
i2b2 Workbench for Demo
Navigate Terms
Find Terms
Query Tool
Analysis View
Query Name:
Group 1
Group 2
Group 3
Dates
Occurs > 0x
Exclude
Treat Independently
SNV [HGNC Gene Symbol]
Add Group
Analysis Types
Patient list
Event list
Number of patients
Gender patient breakdo
Vital Status patient breal
Race patient breakdown
Age patient breakdown
TimeLine
Query Timing
Patient(s) returned:
deletion
indel
insertion
inversion
MNP
Reference genome
SNV
3 prime UTR variant
5 prime UTR variant
Chromosome
downstream gene variant
exon variant
HGNC Gene Symbol
intergenic variant
intron variant
Sequence End (base pair)
Sequence Start (base pair)
splicing variant
Choose modifier value of SNV
You are allowed to search within the narrative text associated with the term SNV.
No Search Requested
By abnormal flag
Search within Text
Containi...
OK
Cancel
Gene Assist
Previous Queries
Analysis details
Information related to the NGS data
Specify input file:
Input file format:
VCF
VCF mapping file:
I2B2 Patient number:
I2B2 Encounter number:
Date of encounter:
YYYY-MM-DD HH:MM:SS format
Reference genome version:
hg18
Sample details
Information related to the sample
Sample ID:
Sample Type:
TISSUE
Anatomical Source:
Pericardium
Collection Method:
BIOPSY
Additive:
UNKNOWN
Sample Pathology
Information related to the sample pathology
Pathology:
TUMOR

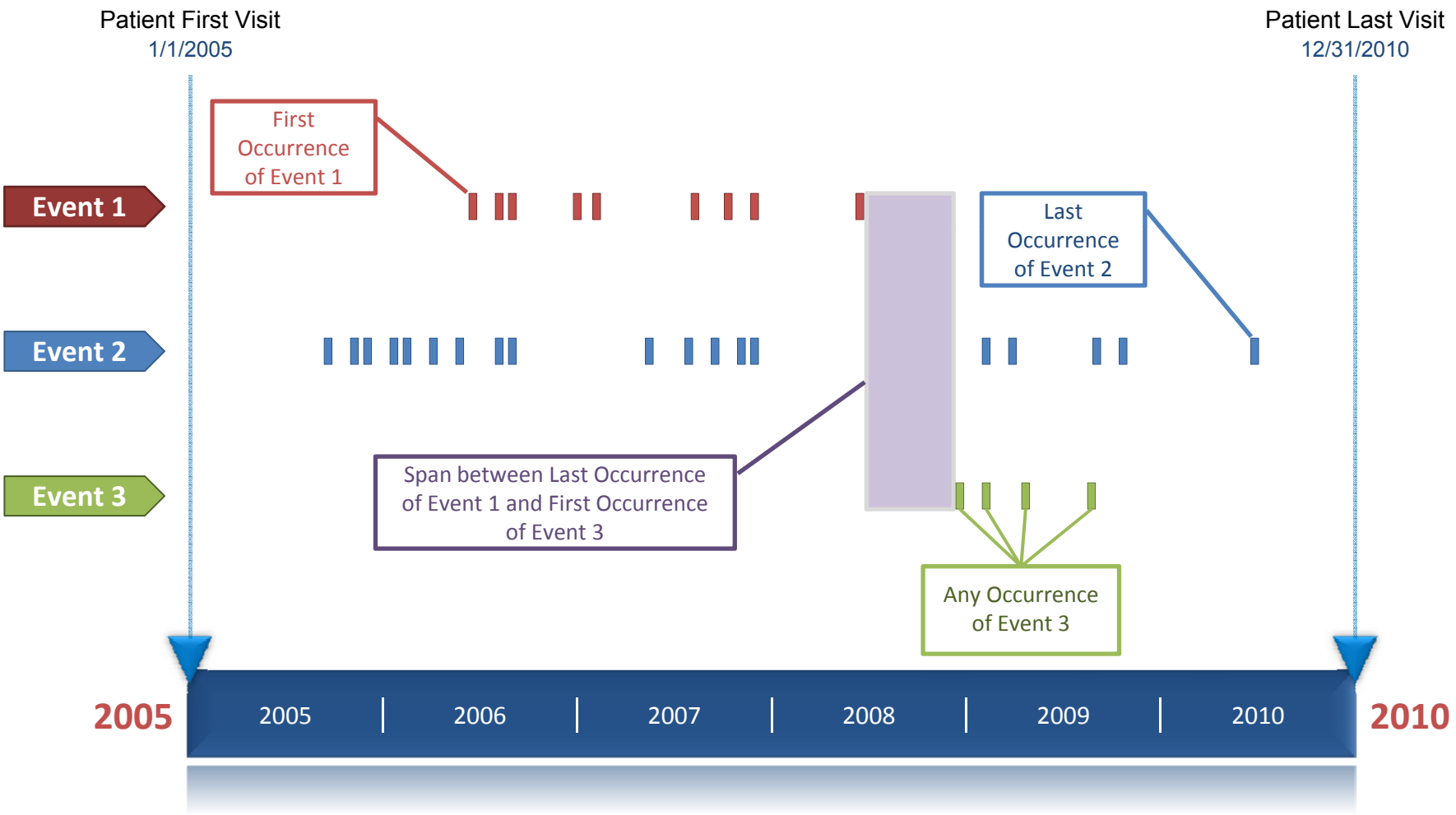
Gene Assist


i2b2 Workbench
i2b2 Workbench for Demo
lori
Status:
Navigate Terms
Find Terms
Query Tool
Analysis View
Query Name:
Group 1
Group 2
Group 3
Dates
Occurs > 0x
Exclude
Treat Independently
SNV [HGNC Gene Symbol]
Analysis Types
Patient list
Event list
Number of patients
Gender patient breakdo
Vital Status patient breal
deletion
indel
insertion
inversion
MNP
Reference genome
SNV
3 prime UTR variant
5 prime UTR variant
Chromosome
downstream gene variant
exon variant
HGNC Gene Symbol
intergenic variant
intron variant
Sequence End (base pair)
Sequence Start (base pair)
splicing variant
Previous Queries
Chromos-exon va@09:19:34 [03-08-2013] [lcp
Chrom-exon -Diabe@09:12:00 [03-08-2013] [
Chrom-exon -Diabe@09:05:59 [03-08-2013] [
Chromos-exon va@09:04:52 [03-08-2013] [lcp
Chromosome@09:01:49 [03-08-2013] [lcp5]
interge-Chromos@07:28:38 [03-08-2013] [lcp
HGNC -exon -Zygos@18:10:27 [03-07-2013] [
EGFR-exon-heterzygous [03-05-2013] [lcp5]
insertion@10:08:54 [03-04-2013] [lcp5]
deletion@09:56:53 [03-04-2013] [lcp5]
EN2-AUT@04:09:55 [02-28-2013] [lcp5]
TSHR-hyperthyroidism@04:02:27 [02-28-2013
substit-substit@03:41:46 [02-28-2013] [lcp5]
MAD1L1-prostateCancer@03:34:56 [02-28-20
substitution@03:31:21 [02-28-2013] [lcp5]
subst-subst-subst@03:27:17 [02-28-2013] [lc
substitution@09:26:27 [02-27-2013] [lcp5]
Genomic Browser
HGNC
HUGO Gene Nomenclature Committee
Search Genes
Home
Search Genes
Downloads
Gene Families
HCOP
Useful Links
About
Contact Us
Request Symbol
Quick Gene Search
Search symbols, keywords or IDs for:
Results that equal begin contain
Display 50 hits
EMBL-EBI
NHGRI
wellcome trust
HUGO
The work of the HGNC is supported by National Human Genome Research Institute (NHGRI) grant P41 HG03345
and Wellcome Trust grant 099129/Z/12/Z
Contact Us: hgnc@genenames.org
Back to top Admin
Site maintained by the External Services team at EMBL-EBI | Terms of Use | Privacy | Cookies
Analysis details
Specify input f
Input file form
VCF mapping
I2B2 Patient nu
I2B2 Encounte
Date of encounter:
YYYY-MM-DD HH:MM:SS format
Reference genome version:
hg18
Sample Pathology
Information related to the sample pathology
Pathology:
TUMOR

New Core i2b2 Features

- Top Level 1.7 features
  - Temporal Queries Enabled
  - Identity Management cell to manage patient participation in i2b2 projects
  - Patient management for Out-of-the-box clinical trials support
  - EMR views of patient for translational medicine and clinical trials support

Temporal Query Formulation



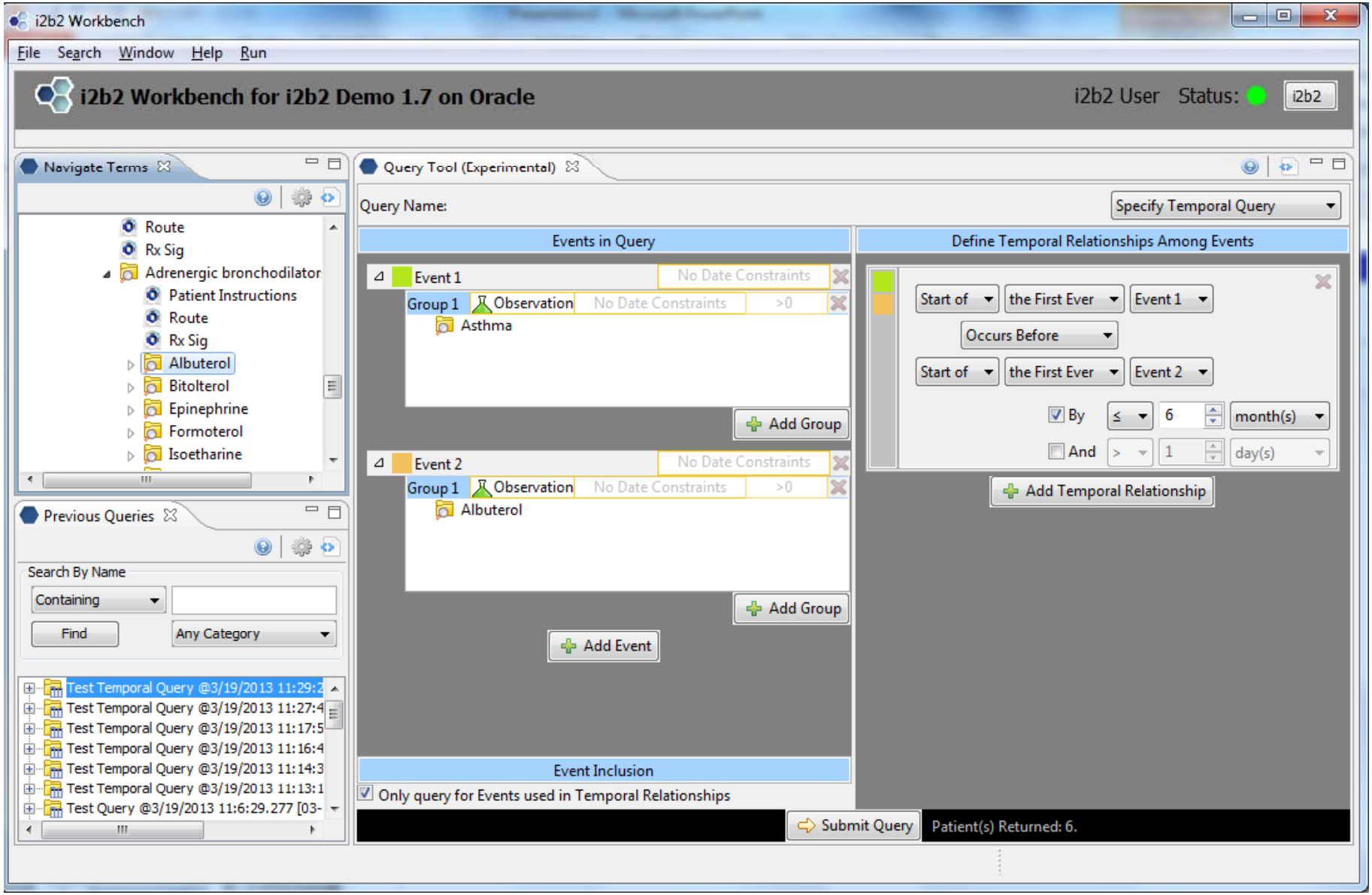
Temporal Query Definition

- Temporal queries are organized around the concept of an event.

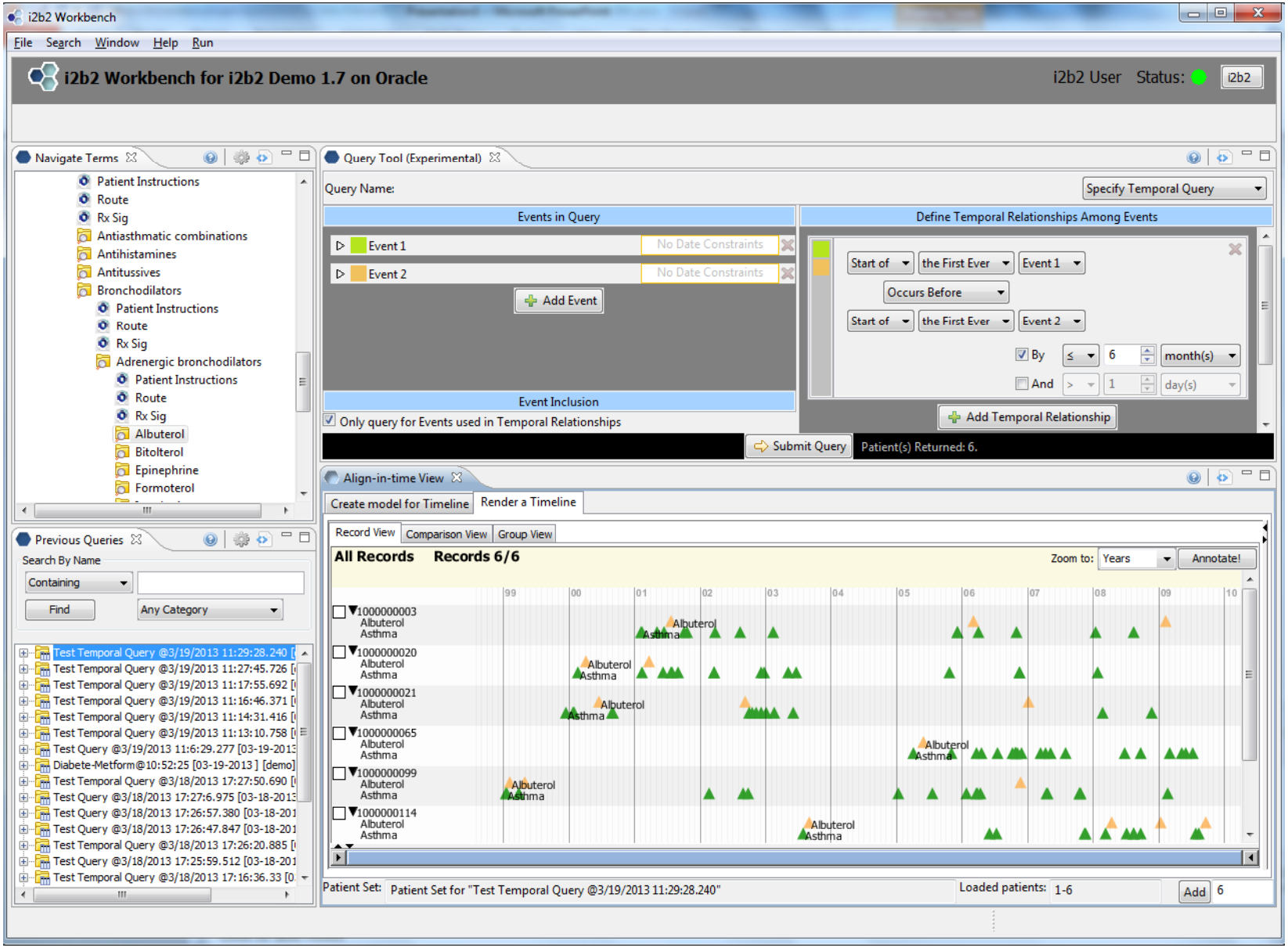
- Events are related to each other through “subquery constraints”
  - An event is defined from an observation
    - Start date or end date
    - First, last, *Any*
  - Events are compared in time to be
    - Equal, Less, Greater, Less Than Equal, Greater Than Equal
  - Within a certain time span
    - Year, Month, Day, Hr, Min

- Each event is defined as a “subquery” within the query definition
  - In the XML is a Recursive definition – a subquery is a query

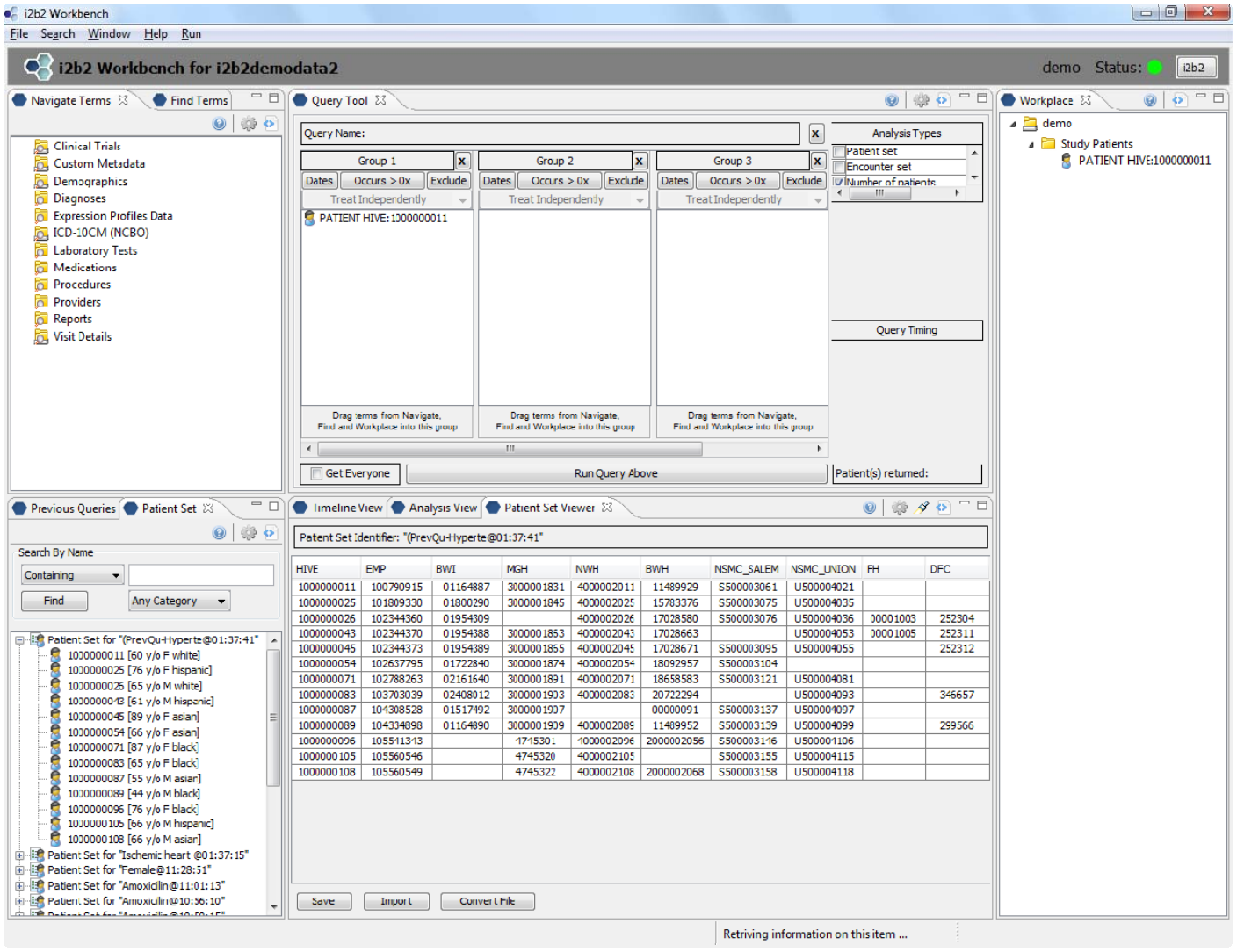
Defining Temporal Relationships



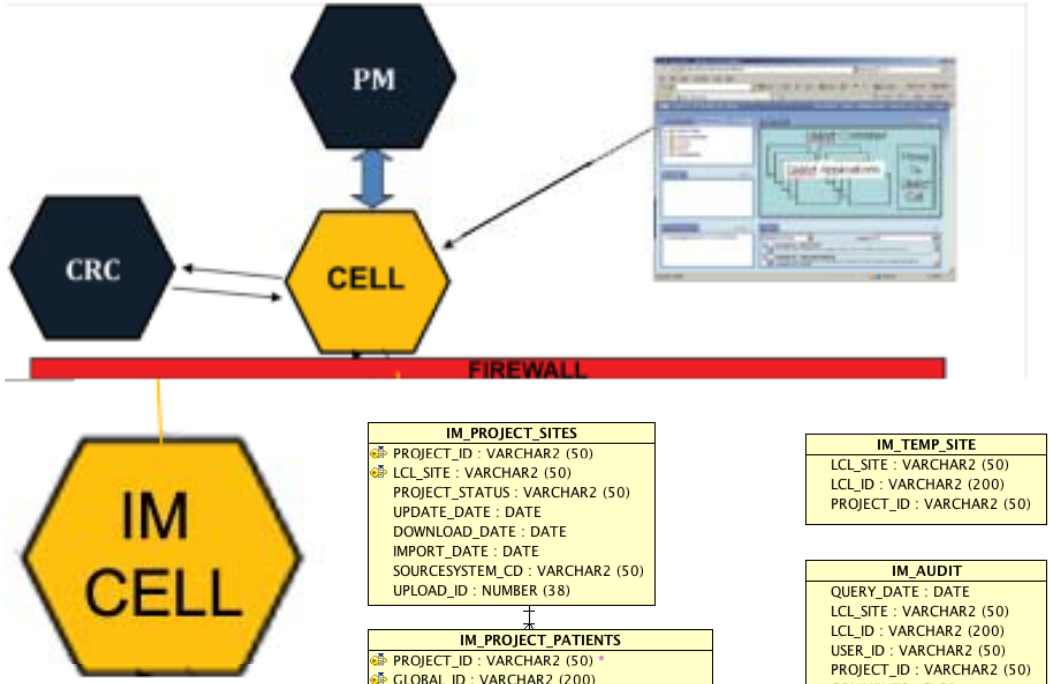
Viewing Temporal Relationships



Patient can be conceived throughout i2b2 as single item



I2B2 IM ARCHITECTURE



IM_PROJECT_SITES
- PROJECT_ID : VARCHAR2 (50)
- LCL_SITE : VARCHAR2 (50)
- PROJECT_STATUS : VARCHAR2 (50)
- UPDATE_DATE : DATE
- DOWNLOAD_DATE : DATE
- IMPORT_DATE : DATE
- SOURCESYSTEM_CD : VARCHAR2 (50)
- UPLOAD_ID : NUMBER (38)

IM_TEMP_SITE
- LCL_SITE : VARCHAR2 (50)
- LCL_ID : VARCHAR2 (200)
- PROJECT_ID : VARCHAR2 (50)

IM_AUDIT
- QUERY_DATE : DATE
- LCL_SITE : VARCHAR2 (50)
- LCL_ID : VARCHAR2 (200)
- USER_ID : VARCHAR2 (50)
- PROJECT_ID : VARCHAR2 (50)
- COMMENTS : CLOB

IM_PROJECT_PATIENTS
- PROJECT_ID : VARCHAR2 (50) *
- GLOBAL_ID : VARCHAR2 (200)
- PATIENT_PROJECT_STATUS : VARCHAR2 (50)
- UPDATE_DATE : DATE
- DOWNLOAD_DATE : DATE
- IMPORT_DATE : DATE
- SOURCESYSTEM_CD : VARCHAR2 (50)
- UPLOAD_ID : NUMBER (38)

IM_MPI_DEMOGRAPHICS
- GLOBAL_ID : VARCHAR2 (200) *
- GLOBAL_STATUS : VARCHAR2 (50)
- DEMOGRAPHICS : VARCHAR2 (400)
- UPDATE_DATE : DATE
- DOWNLOAD_DATE : DATE
- IMPORT_DATE : DATE
- SOURCESYSTEM_CD : VARCHAR2 (50)
- UPLOAD_ID : NUMBER (38)

IM_MPI_MAPPING
- GLOBAL_ID : VARCHAR2 (200) *
- LCL_SITE : VARCHAR2 (50)
- LCL_ID : VARCHAR2 (200)
- LCL_STATUS : VARCHAR2 (50)
- UPDATE_DATE : DATE
- DOWNLOAD_DATE : DATE
- IMPORT_DATE : DATE
- SOURCESYSTEM_CD : VARCHAR2 (50)
- UPLOAD_ID : NUMBER (38)

- Identity Management (IM) cell becomes a primary core cells within the i2b2 hive
- Capable of converting i2b2 patient numbers back into identifying MRNs
- Retrieving and storing Protected Health Information (PHI)
- Critical for investigators who plan to run a study and recruit patients
- Lists of patients with real identifiers are managed and linked to a project
- Allows multiple identifiers to be resolved by an Enterprise Master Patient Indexes

JBoss 7

- ✓Improved Performance
- ✓Better Security
- ✓J2EE 6
- ✓Administration Improvements
- ✓Integrate into Eclipse

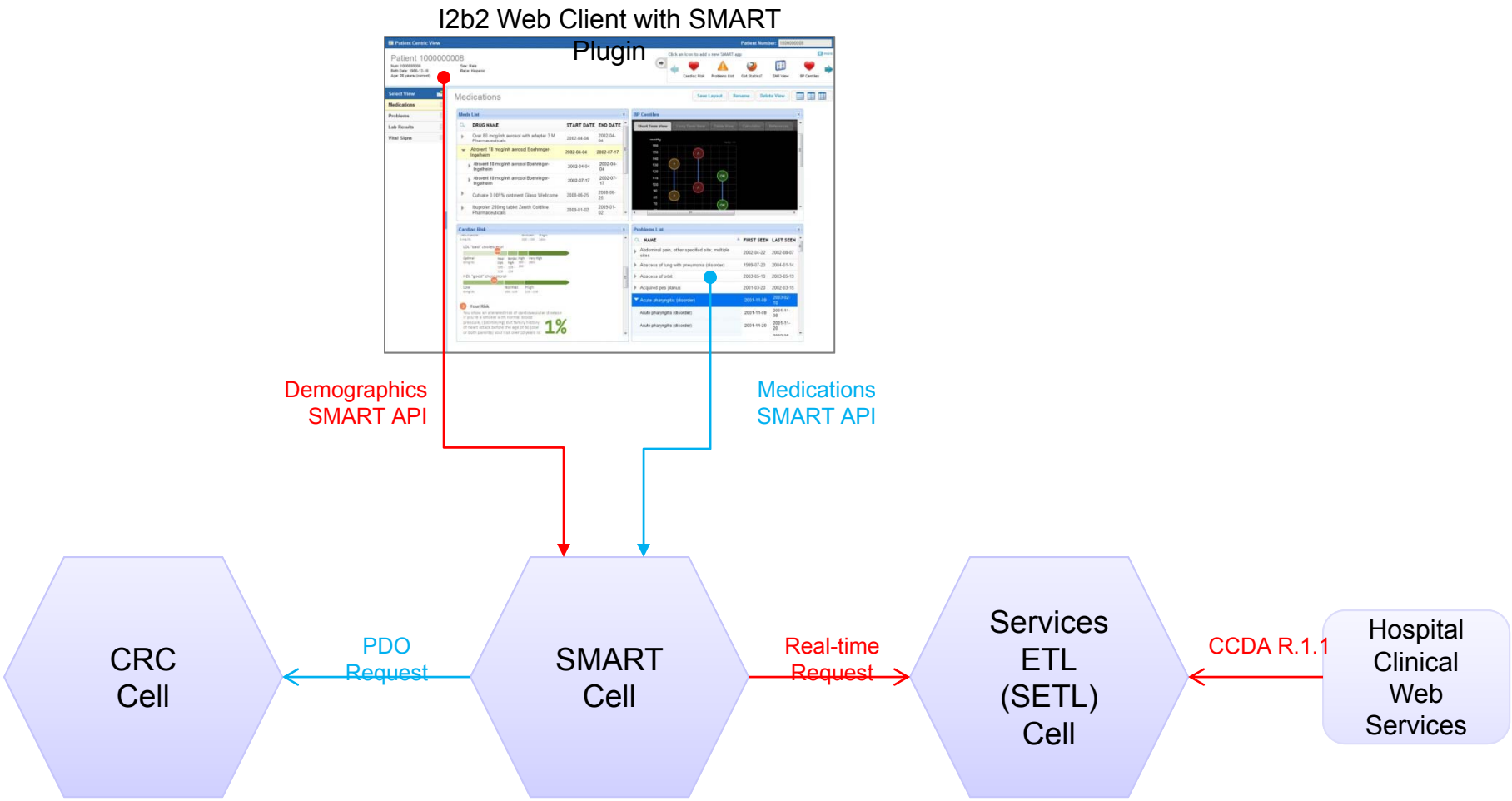
PARTNERS™
HEALTHCARE SYSTEM

Identity Management

- ✓Core Cell
- ✓Encrpytion
- ✓Auditing
- ✓New for 1.7

PARTNERS™
HEALTHCARE SYSTEM

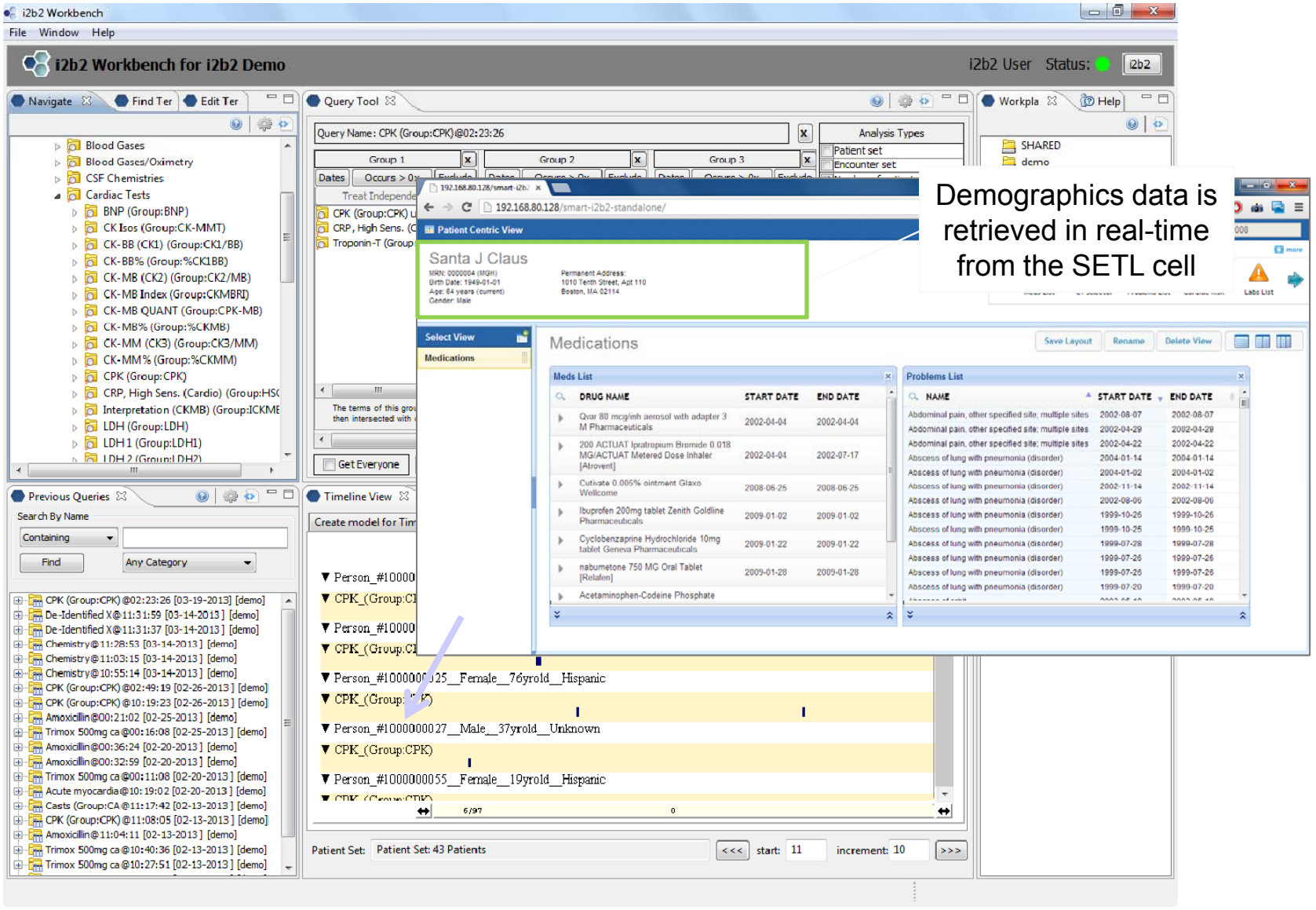
Development of new Services ETL (SETL) Cell



Services ETL: Output PDO *Response*

```
<ns2:patient_set>
    <patient>
        <patient_id source="hive">XXXXXXXXX</patient_id>
        <param column="vital_status_cd" name="vital_status_cd">U</param>
        <param column="birth_date" name="birth_date">19490101</param>
        <param column="sex_cd" name="sex_cd">M</param>
        <param column="language_cd" name="language_cd">SPANISH</param>
        <param column="religion_cd" name="religion_cd">PROTESTANT</param>
        <param column="race_cd" name="race_cd">OTHER</param>
        <param column="ethnicity_cd" name="ethnicity_cd">AFRICAN</param>
        <param column="marital_status_cd" name="marital_status_cd">UNKNOWN</param>
        <param column="legal_first_name" name="legal_first_name">SANTA</param>
        <param column="legal_middle_initial" name="legal_middle_initial">J</param>
        <param column="legal_last_name" name="legal_last_name">CLAUS</param>
        <param column="legal_suffix" name="legal_suffix">JR</param>
        <param column="permanent_line1" name="permanent_line1">1010 TENTH
            STREET</param>
        <param column="permanent_line2" name="permanent_line2">APT 110</param>
        <param column="permanent_city" name="permanent_city">BOSTON</param>
        <param column="permanent_city" name="permanent_city">MA</param>
        <param column="permanent_city" name="permanent_city">02114</param>
        <param column="permanent_city" name="permanent_city">US</param>
        <param column="local_line1" name="local_line1">55 FRUIT ST</param>
        <param column="local_line2" name="local_line2">APT 2</param>
        <param column="local_city" name="local_city">BOSTON</param>
        <param column="local_city" name="local_city">MA</param>
        <param column="local_city" name="local_city">02114</param>
        <param column="local_city" name="local_city">US</param>
        <param column="primary_phone" name="primary_phone">9781231231217</param>
        <param column="work_phone" name="work_phone">781444555553333</param>
        <param column="mobile_phone" name="mobile_phone">6039275569</param>
        <param column="other_phone" name="other_phone">(987)111-1111</param>
    </patient>
</ns2:patient_set>
```

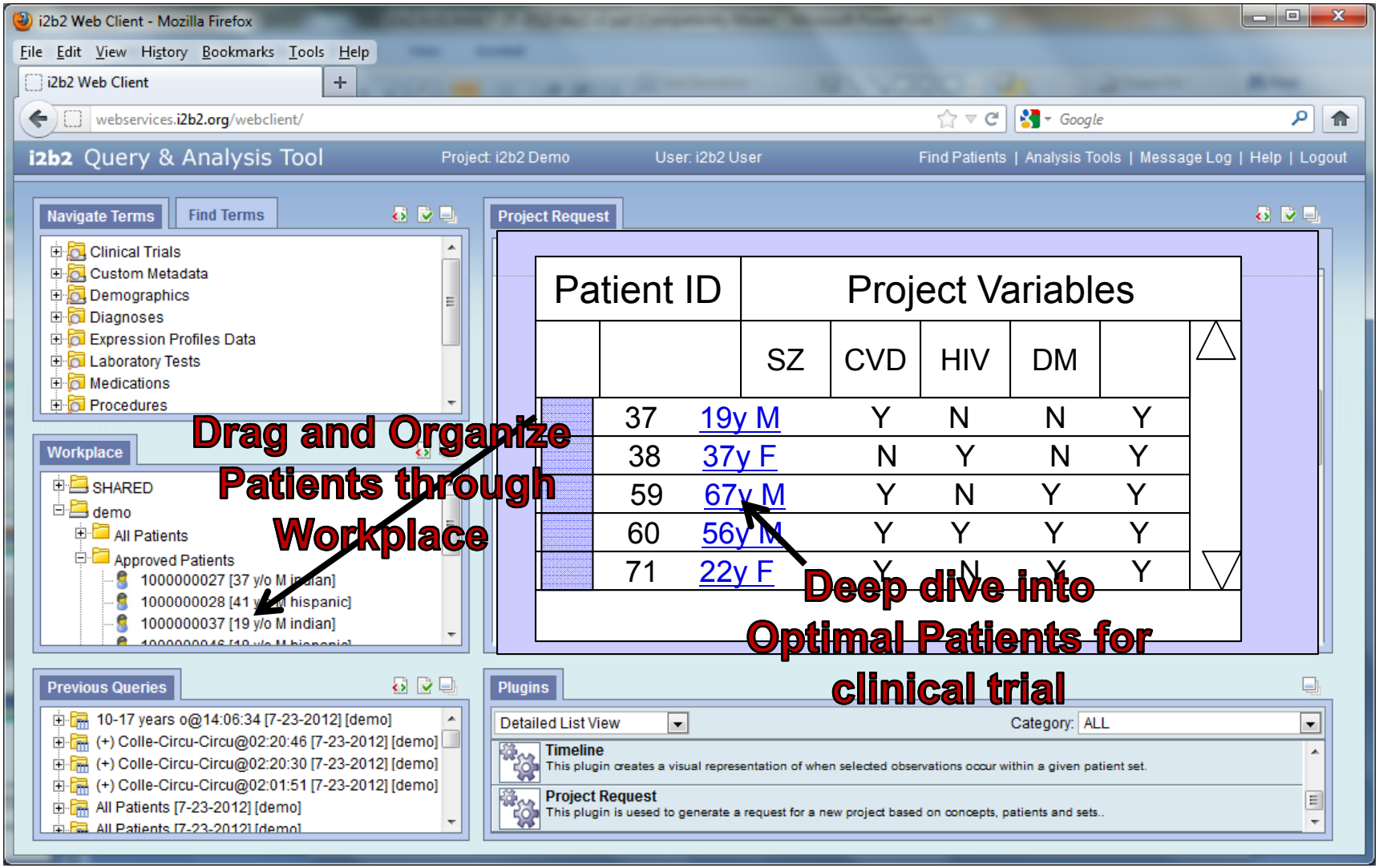
Launching ‘standalone’ SMART-i2b2 views



Supports workflow for Clinical Trials

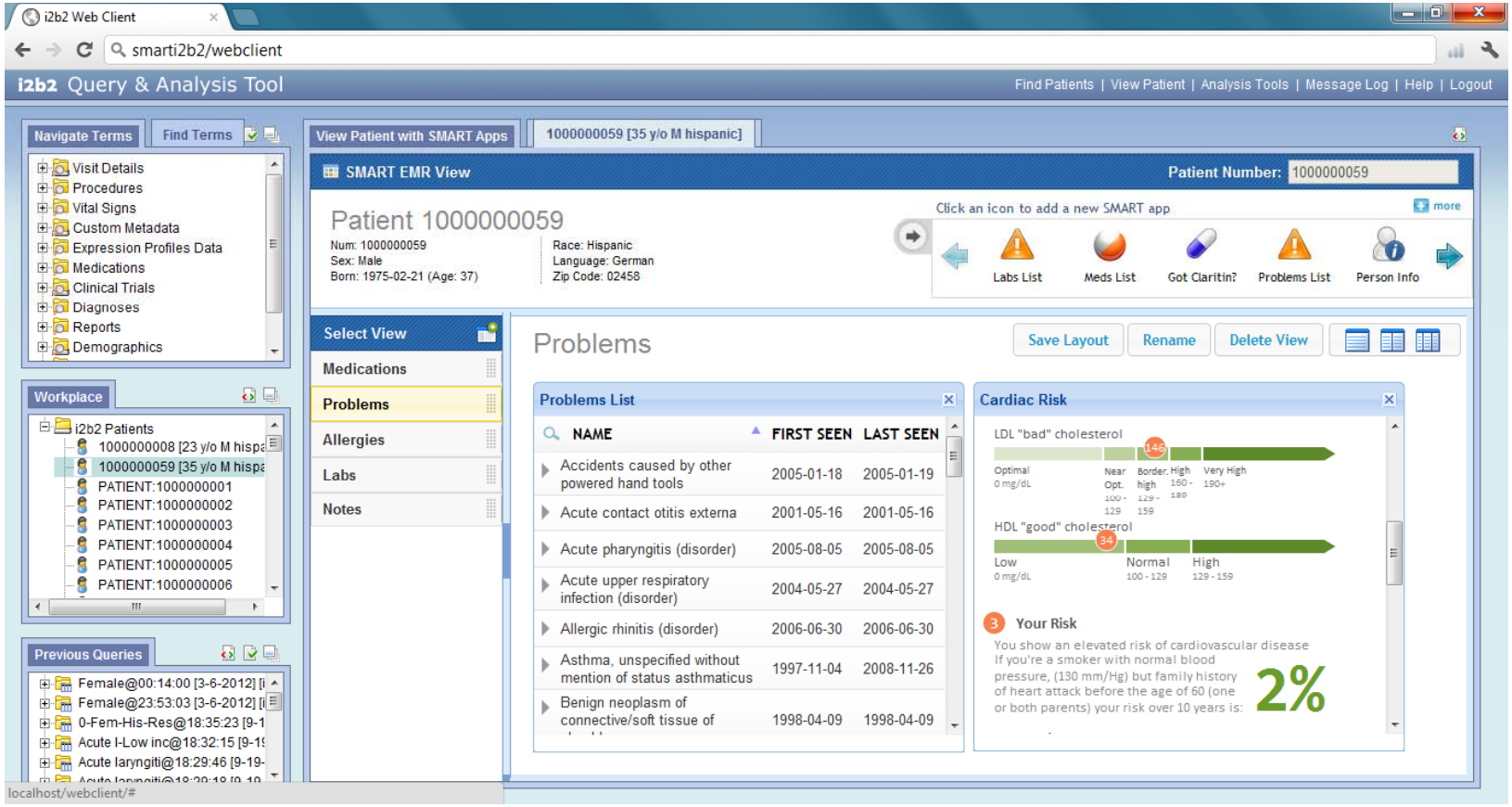
- Person does query as obfuscated user in large data mart
- Optimal query results can be used to create request for approval so that patients can be viewed as a limited data set
- Approval is obtained and a new project is created where those patients in the Optimal patient set can be viewed in plug-ins such as timeline, charts, and de-identified SMART views.
- Patients are chosen that represent truly Optimal patients for the Clinical trial.
- PHI is viewed on the truly Optimal patients in a specially Audited view that may be governed by the local application of “special rules”

Workflow to support Clinical Trials



Workflow to support Clinical Trials

- EMR View that looks familiar to clinicians
- Patients can be screened using SMART App's

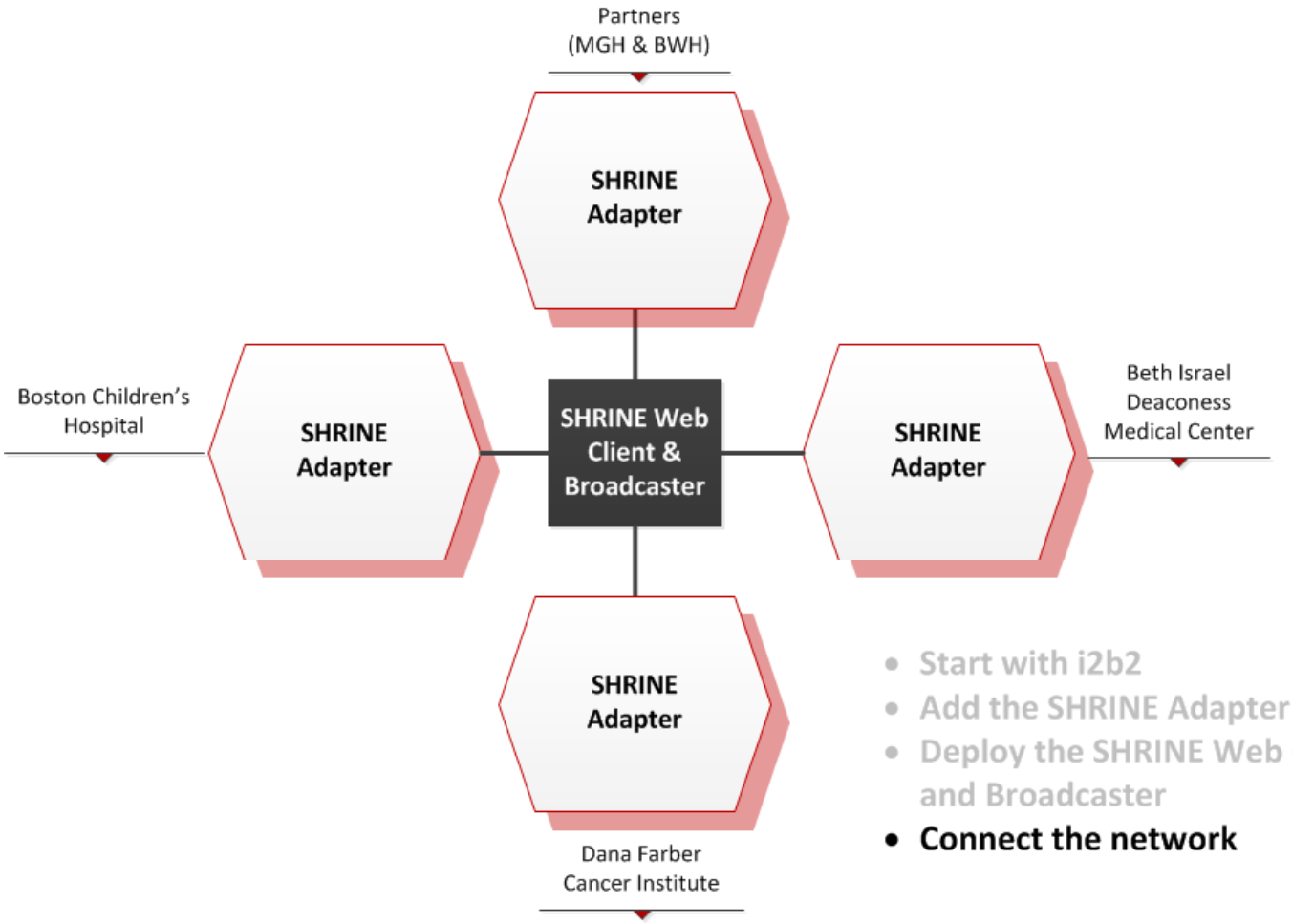


Timelines for new i2b2 Releases

- Release Candidate of Version 1.7 Core in April
- Clinical trials out-of-the-box plug-ins in September
- Next-Generation Sequencing plug-ins in September

HARVARD CATALYST
THE HARVARD CLINICAL AND TRANSLATIONAL SCIENCE CENTER

Harvard SHRINE Architecture



- Start with i2b2
- Add the SHRINE Adapter
- Deploy the SHRINE Web Client and Broadcaster
- **Connect the network**

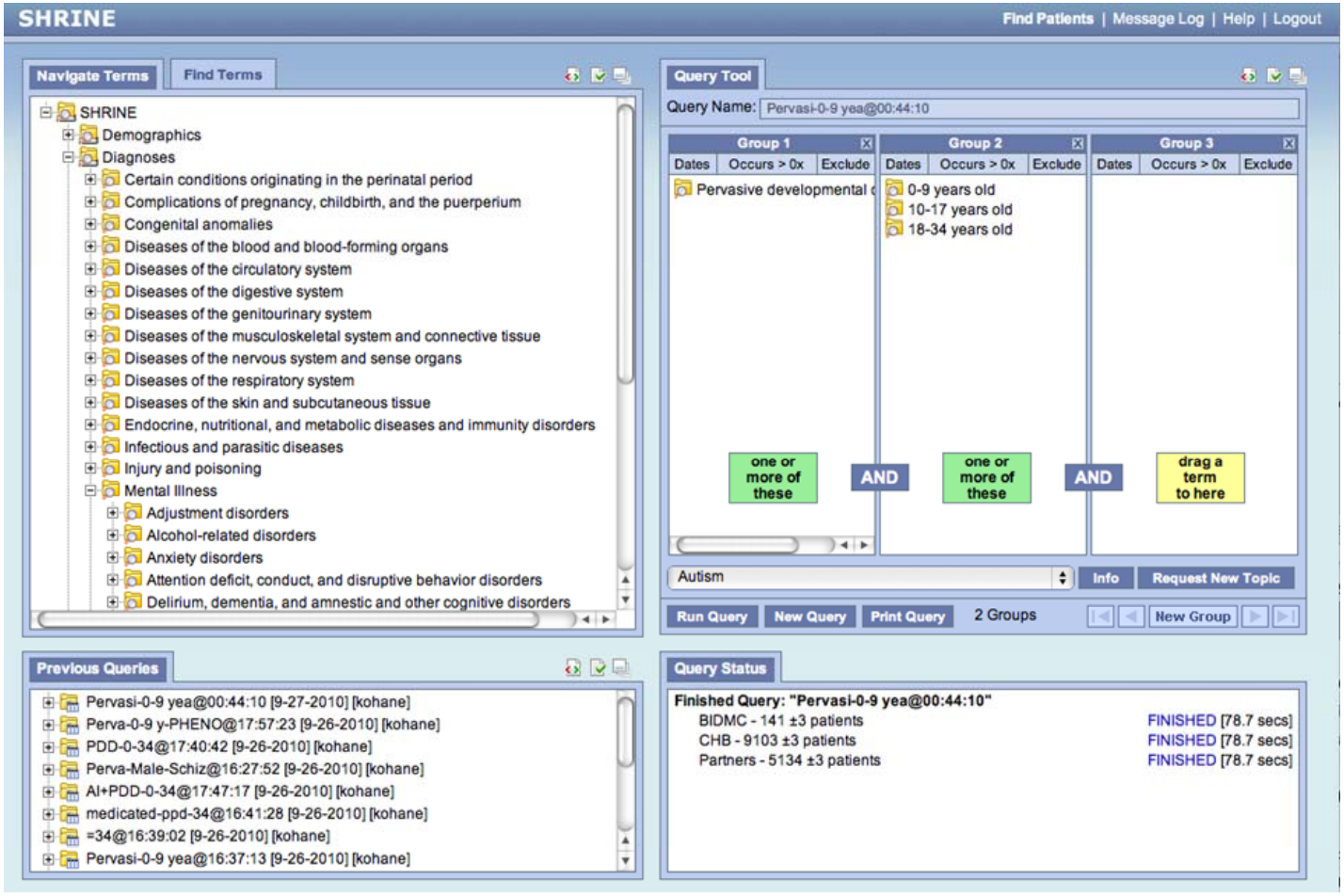
HARVARD CATALYST
THE HARVARD CLINICAL AND TRANSLATIONAL SCIENCE CENTER

Harvard SHRINE Metrics

- **6M patients**
- **1B facts**
- **4 ontology categories, 18K terms**
- **10K potential users**
- **5 IRBs**
- **5 major competing hospitals, 4 sites**
  - Partners HealthCare (BWH, MGH)
  - Boston Children's Hospital
  - Beth Israel Deaconess Medical Center
  - Dana-Farber Cancer Institute

HARVARD CATALYST

THE HARVARD CLINICAL AND TRANSLATIONAL SCIENCE CENTER



HARVARD CATALYST
THE HARVARD CLINICAL AND TRANSLATIONAL SCIENCE CENTER

National SHRINE Demo

- **A demonstration of the feasibility of a national research network**
- **8 selected sites to perform queries on co-morbidities related to the primary diagnoses of autism and diabetes**
  - Patient must be diagnosed with either Autism Spectrum Disorders or Diabetes Mellitus at least once between 6/01/09 – 5/31/11

HARVARD CATALYST
THE HARVARD CLINICAL AND TRANSLATIONAL SCIENCE CENTER

8 Participating Institutions

The Institute of Translational Health Sciences, University of Washington

University of Michigan Institute for Clinical & Health Research

Partners HealthCare

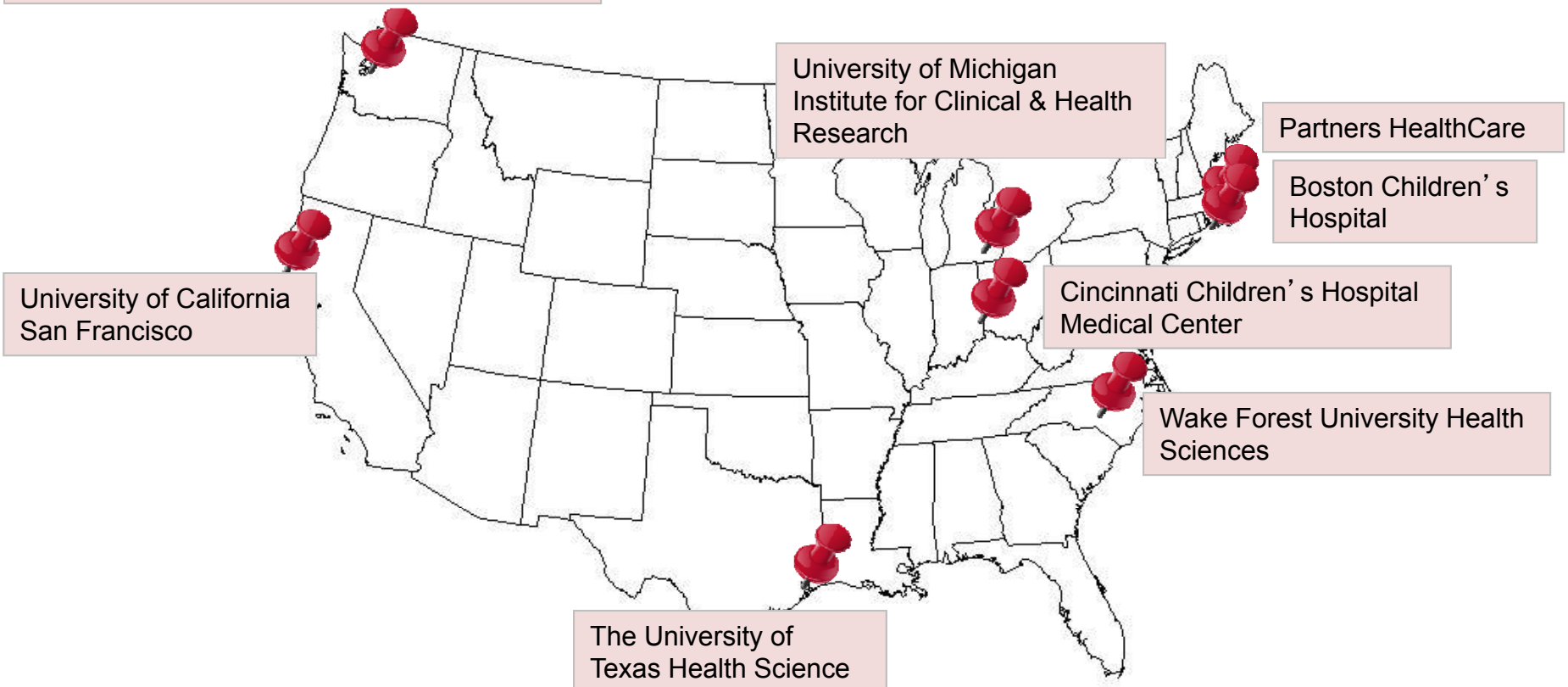
Boston Children's Hospital

University of California San Francisco

Cincinnati Children's Hospital Medical Center

Wake Forest University Health Sciences

The University of Texas Health Science Center at Houston



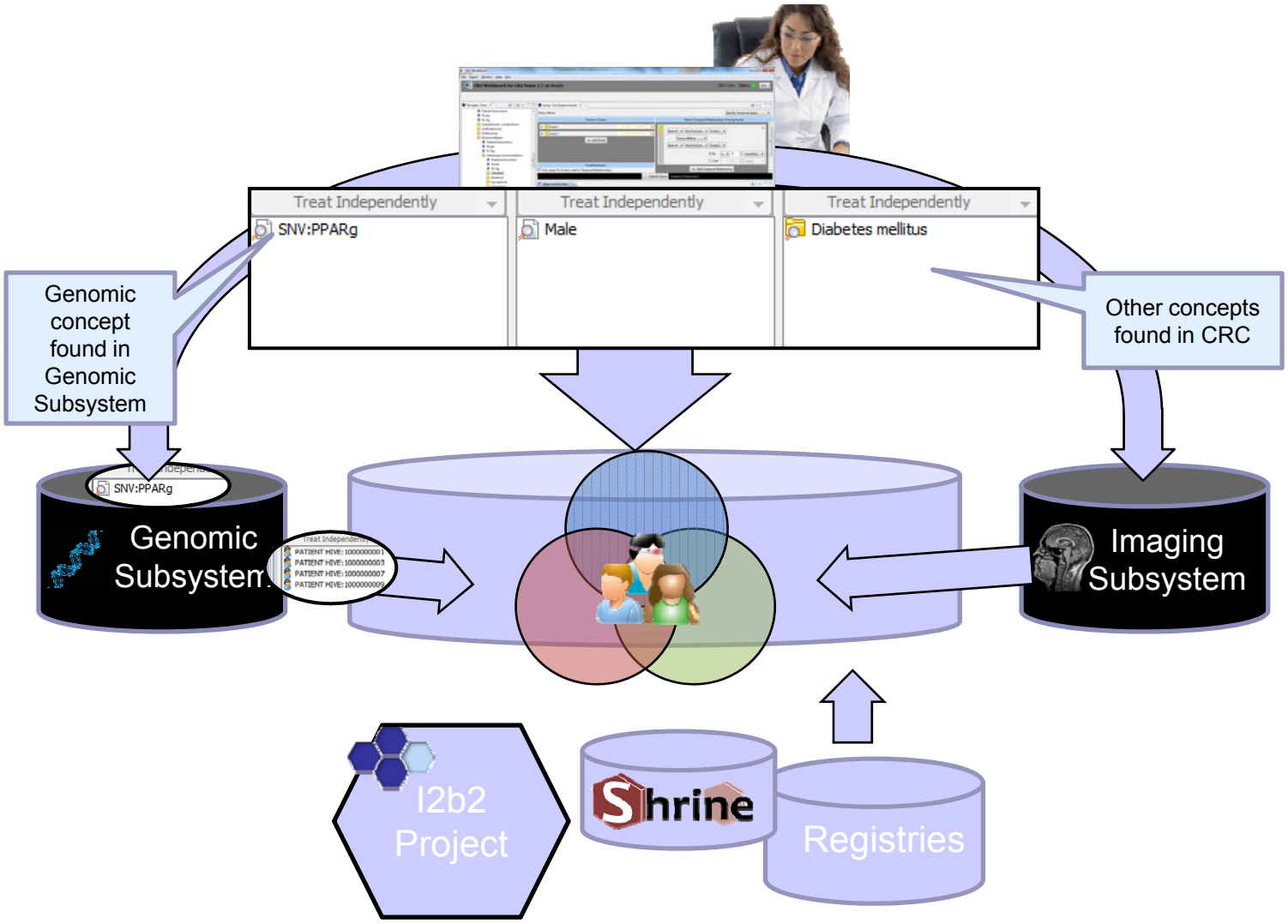
HARVARD CATALYST
THE HARVARD CLINICAL AND TRANSLATIONAL SCIENCE CENTER

SHRINE CT

- **Next major version of SHRINE – focused on supporting clinical trials**

- **Start with SHRINE aggregate query**
- **Narrow/validate set of subjects by incrementally filtering on detailed patient data**
- **Establish a set of subjects for recruitment into a clinical trial**

Inventing new Big Data Architecture



Big Picture - Data flow of next-gen sequencing

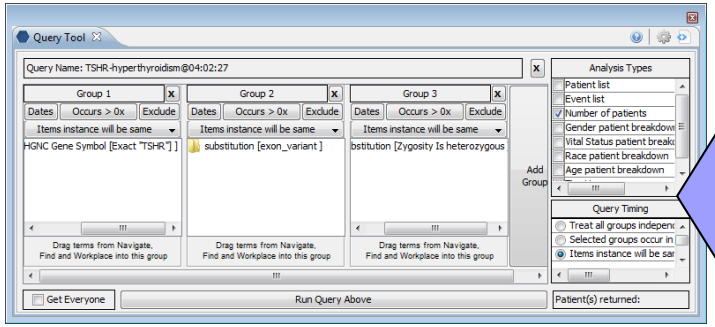
base calls from the sequencer

FASTQ files with base calls

SAM with standard alignment

VCF digests variants

GVF maps to ontologies



De-identified Data Warehouse

Data flow of next-gen sequencing

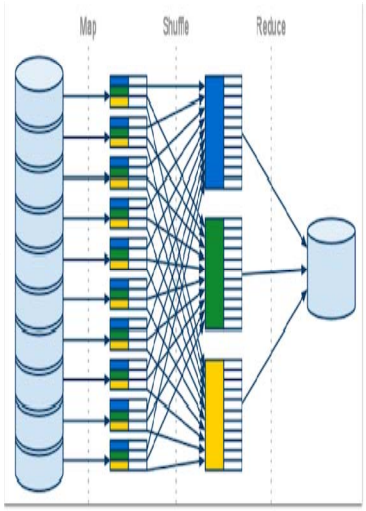
base calls from the sequencer

FASTQ files with base calls

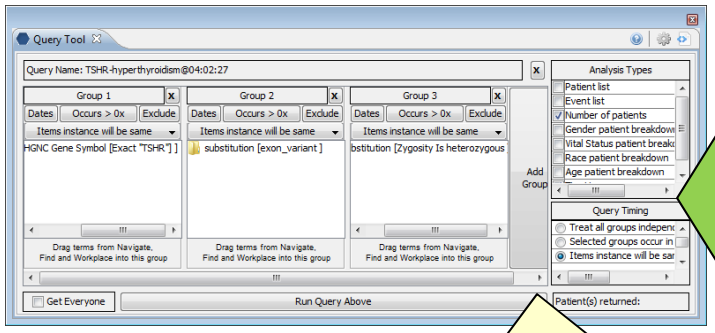
SAM with standard alignment

VCF digests variants

GVF maps to ontologies

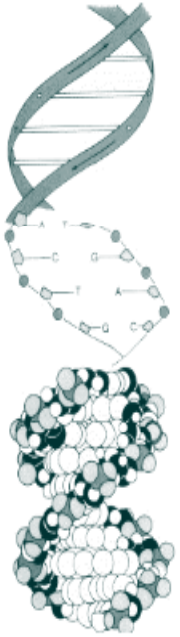


Map-Reduce Queries



De-identified Data Warehouse

Sequence patterns



I2b2, SHRINE, and SMART Information and Software on the Web

i2b2 Homepage (https://www.i2b2.org)

i2b2 Software (https://www.i2b2.org/software)

i2b2 Community Site (https://community.i2b2.org)

SHRINE at Harvard (http://shrine.catalyst.harvard.edu)

SHRINE Software:

(https://open.med.harvard.edu/display/SHRINE/Software)

SMART i2b2 Homepage (http://smarti2b2.org)

SMART Platforms Homepage (http://smartplatforms.org)

References

- Friedman C, Hripcsak G, Johnson SB, Cimino JJ, and Clayton PD. A generalized relational schema for an integrated clinical patient database. In Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care, IEEE Computer Society Press, Los Alamitos, CA, 1990, pp 335--339.
- Kimball, R. The Data Warehousing Toolkit. New York: John Wiley, 1997.
- Murphy, S. N., Morgan, M. M., Barnett, G. O., Chueh, H. C. Optimizing Healthcare Research Data Warehouse Design through Past COSTAR Query Analysis. Proc AMIA Symp. 1999;892-897
- Murphy, S.N., Gainer, V.S., Chueh, H. A Visual Interface Designed for Novice Users to find Research Patient Cohorts in a Large Biomedical Database. AMIA, Fall Symp. 2003: 489-493.
- Murphy, S.N., Weber, G., Mendis, M., Gainer, V.S., Churchill, S., Kohane, I.S. Serving the Enterprise and Beyond with Informatics for Integrating Biology and the Bedside (i2b2). Journal of the American Medical Informatics Association, 2010 March 1; 17(2): 124-130.
- Weber GM, Murphy SN, McMurry AJ, Macfadden D, Nigrin DJ, Churchill S, Kohane, I.S. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. *J Am Med Inform Assoc* 2009;16(5):624-30

Co-investigators, developers, and collaborators

- RPDR
  - Eugene Braunwald
  - John Glaser
  - Diane Keogh
  - Henry Chueh
- I2b2, SHRINE and SMART
  - Isaac Kohane
  - Susanne Churchill
  - Griffin Weber
  - Michael Mendis
  - Vivian Gainer
  - Lori Phillips
  - Rajesh Kuttan
  - Wensong Pan
  - Janice Donahue
  - William Simons (SHRINE)
  - Andy McMurry (SHRINE)
  - Doug MacFadden (SHRINE)
  - Ken Mandl (SMART)
  - Josh Mandel (SMART)
  - Nich Wattanasin (SMART)
- Medical Imaging (mi2b2)
  - Christopher Herrick
  - David Wang
  - Bill Wang
- i2b2 Driving Biology Projects
  - Roy Perlis
  - Jordan Smoller
  - Dan Iosifescu
  - Vivian Gainer
  - Victor Castro
  - Andrew Cagan
  - Robert Plenge
  - Stan Shaw
  - Qing Zeng
  - Guergana Savova
  - Caitlin Clements
  - Wouter Hoogenboom,
  - Tianxi Cai

THE HARVARD CLINICAL
AND TRANSLATIONAL
SCIENCE CENTER

See you on June 18-20, 2013 at the 3rd i2b2 - SHRINE Conference in Boston!