Data Mining for Genomic-Phenomic Correlations

Joyce C. Niland, Ph.D.
Associate Director & Chair, Information Sciences

Rebecca Nelson, Ph.D.
Lead, Data Mining Section

City of Hope National Medical Center
Duarte, California, USA

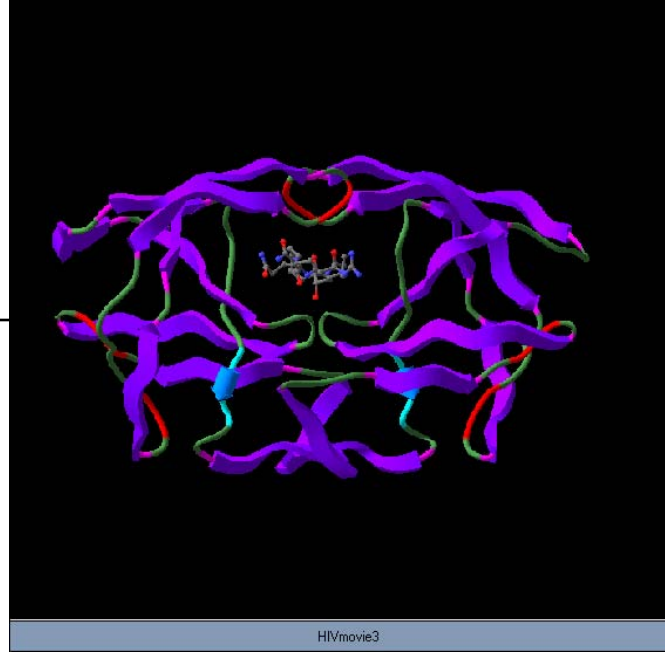

City of Hope National Medical Center, Duarte, California

City of Hope National Medical Center

- Founded in 1913
- State-of-the-art care to patients with cancer & other life-threatening diseases (e.g. diabetes)
- Leading edge research into the causes, prevention, and cure of such diseases
  - Promising new therapeutic agents being taken from “bench” to “bedside” (translational research)
  - Over 400 ongoing clinical trials, 1/3 initiated at City of Hope
- Human genome has expanded scope and objectives of translational research

Genomics…

- Improved diagnosis of disease
- Earlier detection of genetic risk



Gene Therapy…

- Repairing defective genes with healthy ones

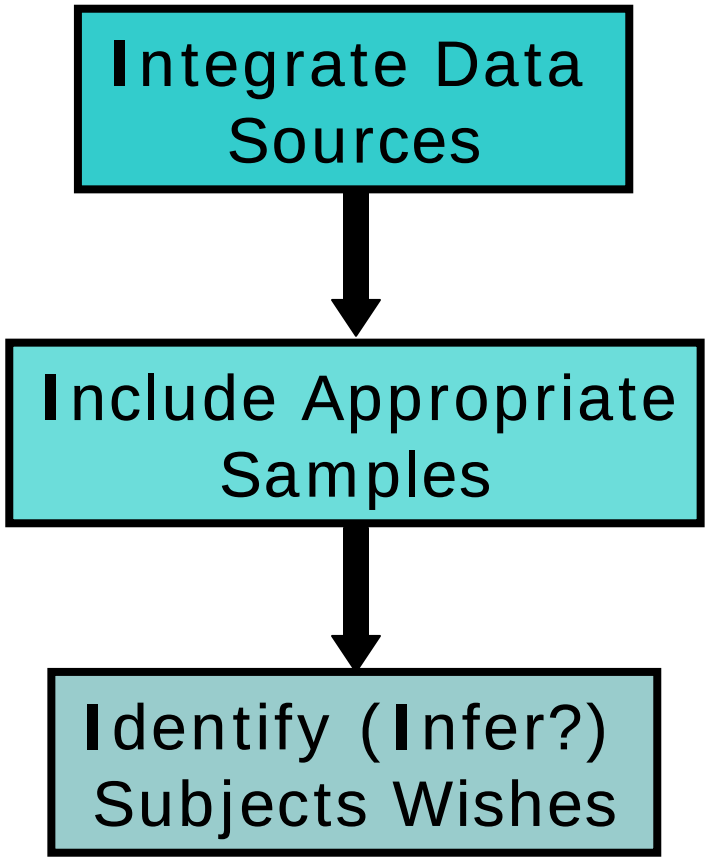
Pharmacogenetics…

- New drugs based on information about genes

Recent Examples of Genomic-Phenomic Data Mining Supported by Biostatistics

- Predictors of Genetic Susceptibility to Heart Disease Post-Bone Marrow Transplant
- Pathogenesis of Radiation-induced Breast Cancer
- Gene Expression in Prostate Cancer Tumors
- Prognostic Biomarkers for Stage I-III Renal Cell Carcinoma
- Validation of Biomarkers for Tumor Initiating Cells in Brain Cancer
- Expression of DNA Repair in Normal versus Tumor Cell Genes

The 3 (4?) I's of Data Mining Process



Data Warehouse

Cohort Assembly

Honest Broker

Integrating Data Systems to Support Biomedical Research

- Biomedical research is an increasingly complex collaborative undertaking
  - Requires integration of data, rules, processes, and vocabularies from many different source systems
- Most information systems developed independently
  - Operational systems, created to meet different functional and departmental needs of an institution

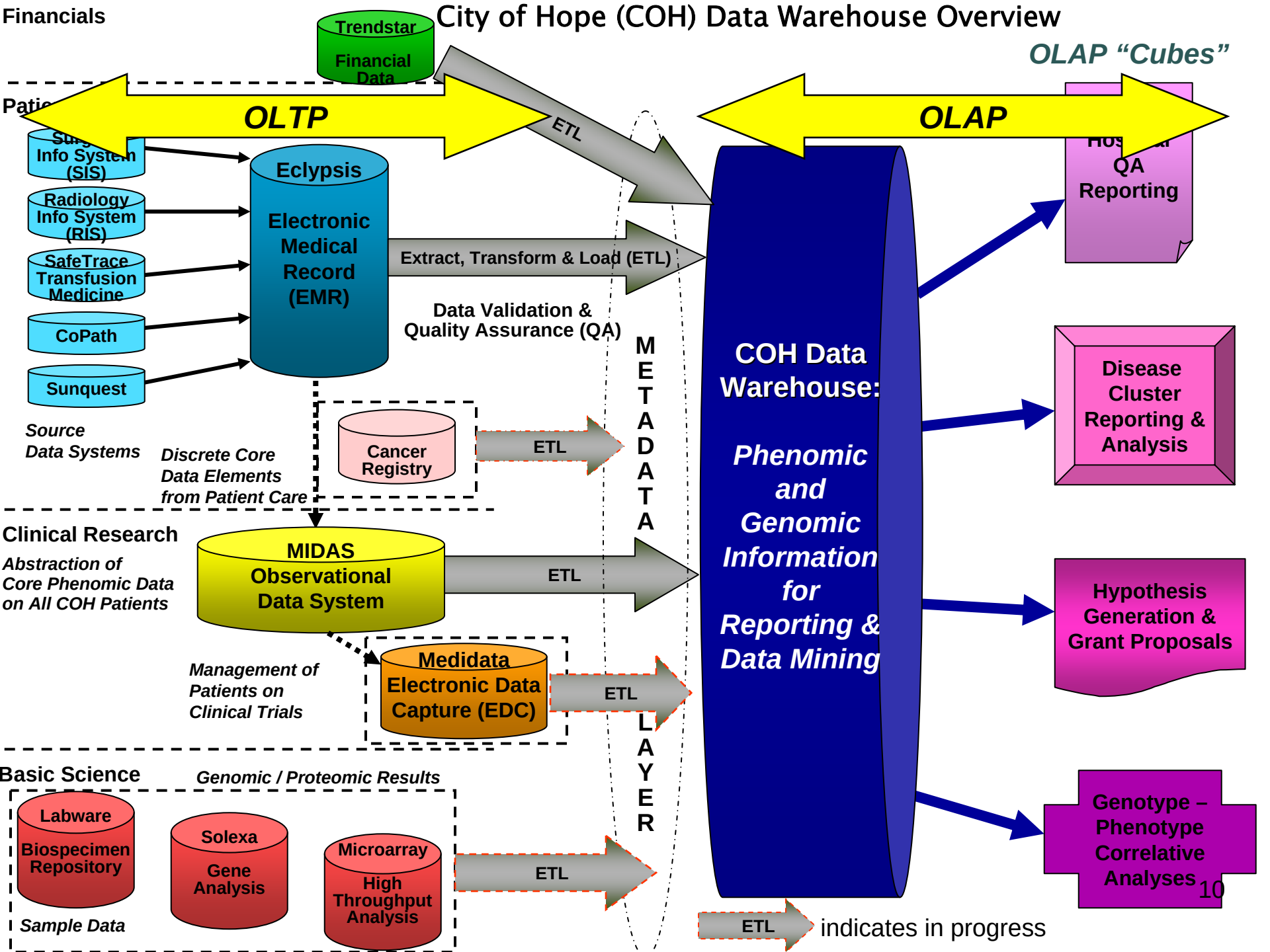
Operational Systems Versus Data Warehousing

- ***On-Line Transaction Processing (OLTP):***
  - Focuses on an organization's day-to-day business needs (electronic medical record, financial systems, clinical trial management systems)

- ***On-Line Analytic Processing (OLAP):***
  - Retrieves, analyzes, reports, and shares data from disparate systems, vendors & departments (DW)

Data Warehousing Concept

- Large, centralized, and longitudinal store of data to facilitate organization-wide consolidated reporting and analysis
- Multiple source databases
- Central coordination and management via metadata repository
- Multiple target “data marts” (aggregated datasets for efficient querying and analysis)

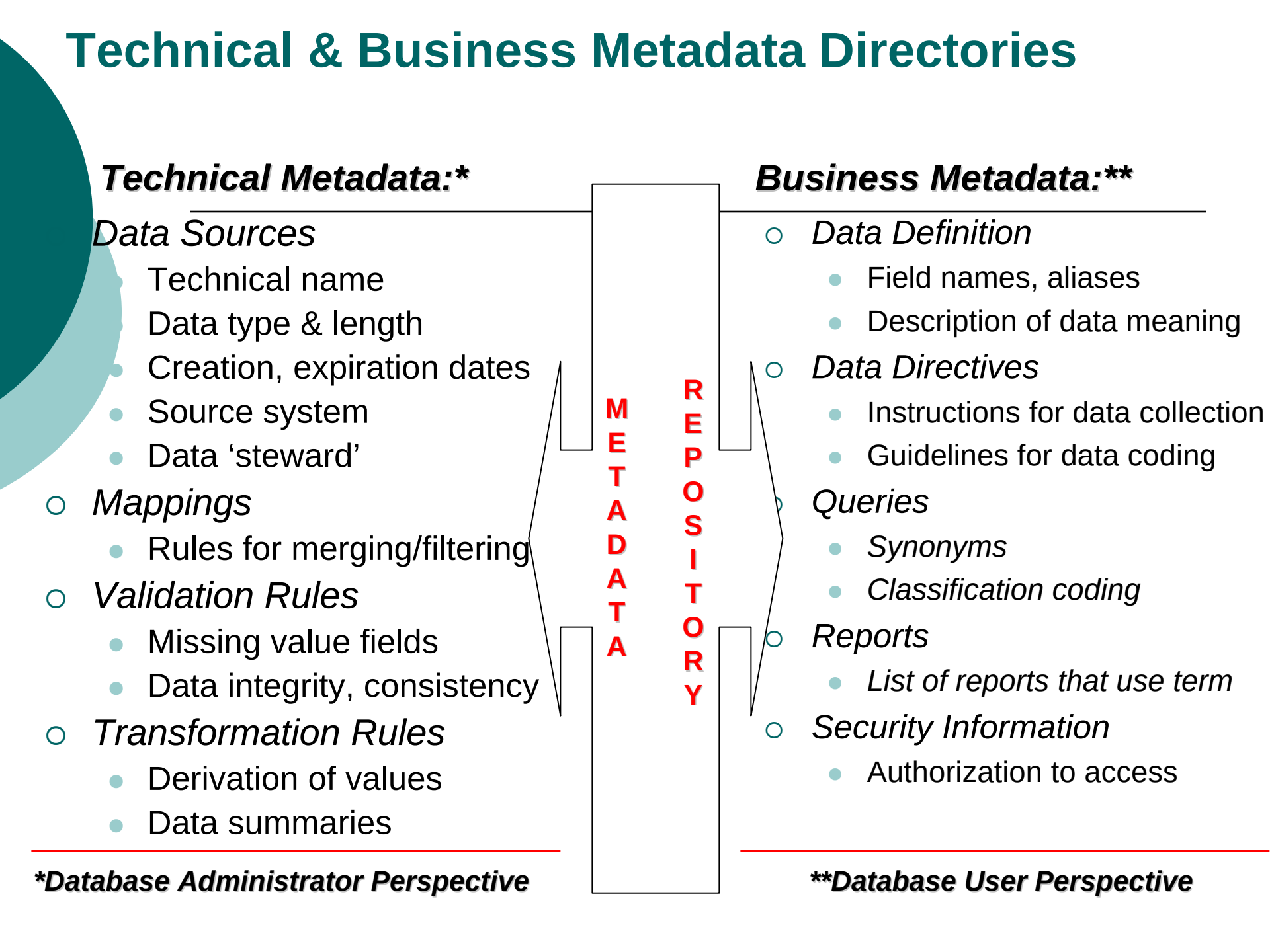

City of Hope (COH) Data Warehouse Overview
Financials
Trendstar
Financial Data
OLAP "Cubes"
OLTP
ETL
OLAP
Info System (SIS)
Radiology Info System (RIS)
SafeTrace Transfusion Medicine
CoPath
Sunquest
Eclypsis
Electronic Medical Record (EMR)
QA Reporting
Extract, Transform & Load (ETL)
Data Validation & Quality Assurance (QA)
METADATA
COH Data Warehouse:
Phenomic and Genomic Information for Reporting & Data Mining
Disease Cluster Reporting & Analysis
Hypothesis Generation & Grant Proposals
Genotype – Phenotype Correlative Analyses
Source Data Systems
Discrete Core Data Elements from Patient Care
Cancer Registry
ETL
Clinical Research
Abstraction of Core Phenomic Data on All COH Patients
MIDAS Observational Data System
ETL
Management of Patients on Clinical Trials
Medidata Electronic Data Capture (EDC)
ETL
LAYER
Basic Science
Genomic / Proteomic Results
Labware Biospecimen Repository
Solexa Gene Analysis
Microarray High Throughput Analysis
ETL
Sample Data
ETL
indicates in progress

Utility of a Data Warehouse

While protecting personally identifiable information and proprietary research data:

- Decision support to administrators
- Screening of patients for eligibility
- Measurement of quality of care & outcomes
- Query capabilities to investigators
- Data mining to generate new hypotheses, facilitate new discoveries

Technical & Business Metadata Directories

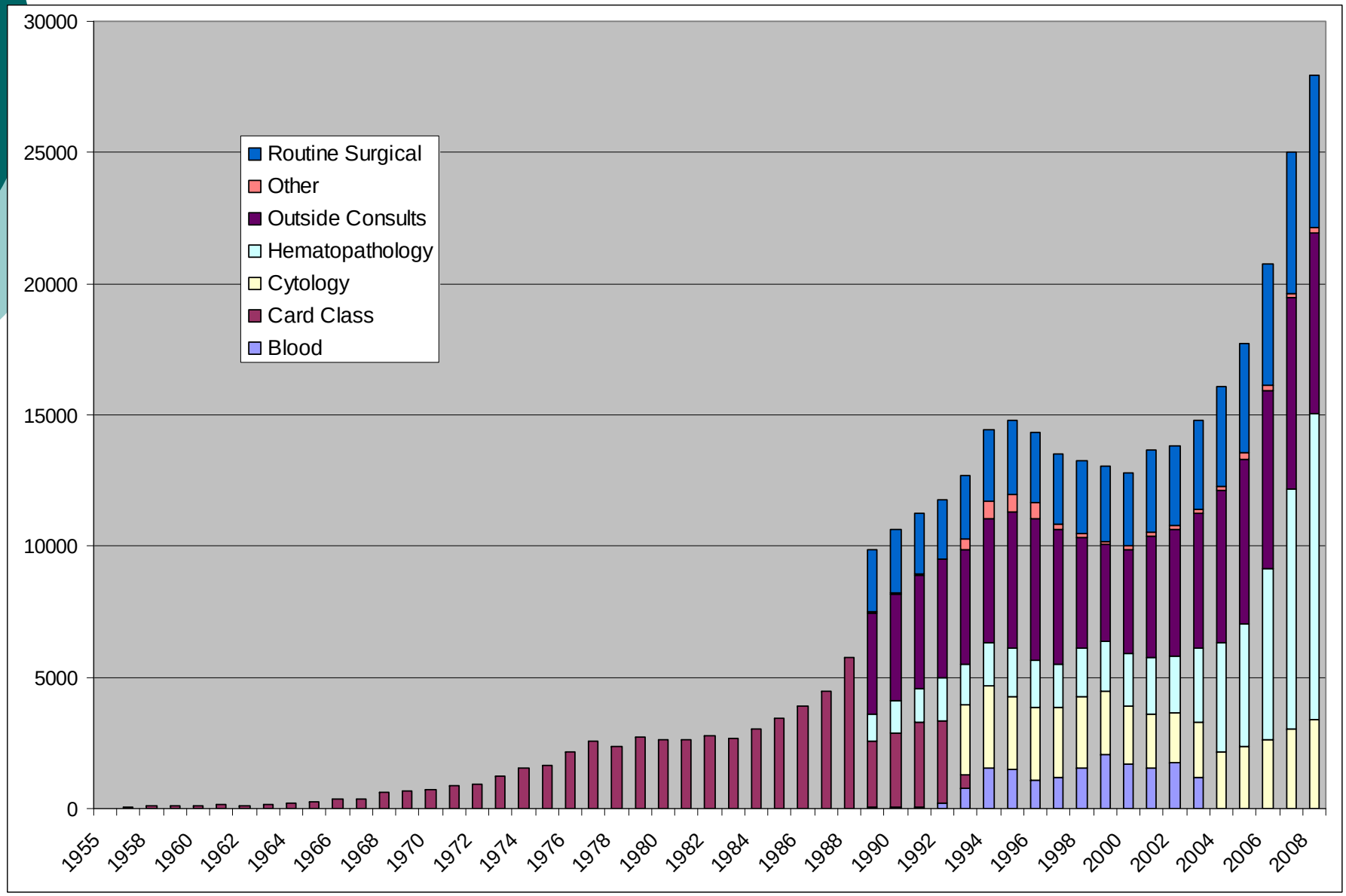


Cohort Assembly

- Inclusion of subjects with appropriate phenomic characteristics AND available tissue

- >360,000 specimens logged in CoPath system, going back to 1955
  - Critical to integrate tissue sample data into the data warehouse
  - Broken down into “Class of Case” to describe type of specimen

CoPath Specimens by Year and Type



Integrate “Best” Source of Tissue Annotation Data

- Standardized formatted path reports now available from pathologist

- ‘Synoptic Report’ includes:
  - Path T Stage
  - Nodes Examined
  - Nodes Positive
  - Path N Stage
  - Path M Stage
  - Margins
  - Histology
  - Grade

Concordance Between CoPath Synoptic Reports & Cancer Registry

Comparing Path Data Sources: Synoptic Report vs. CNExT

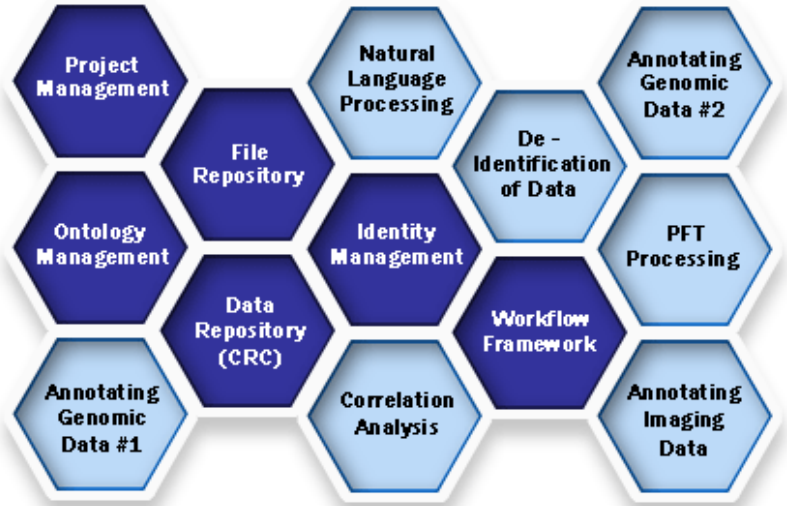
Synoptic Report - Pathology	CNExT – Cancer Registry
○ Surgery specific	○ Patient specific
○ Only 22% of cancer cases	○ 100% of cancer cases
○ Non-strict reporting rules	○ Strict reporting rules
○ No confirmation by treating MD	○ Confirmed by treating MD

Text Mining to Identify Cases

- Text mining may be needed if neither data source specific enough
  - Example: Diagnosis of Unknown Primary Sites for Metastatic Tumors via 92-Gene RT-PCR
  - Needed to find tissues samples labeled as purely metastatic, with no link to original cancer site

Enabling Investigators to Conduct Cohort Searches via i2b2

- **i2b2**: *Integrating Biology and the Bedside*
  - Facile user interface to allow investigators to search for cohorts on their own
    - Executes advanced queries against meta-database to identify available subjects, tissue samples, or biospecimens matching query criteria
    - if feasible *number* of cases returned, then submit IRB protocol for approval

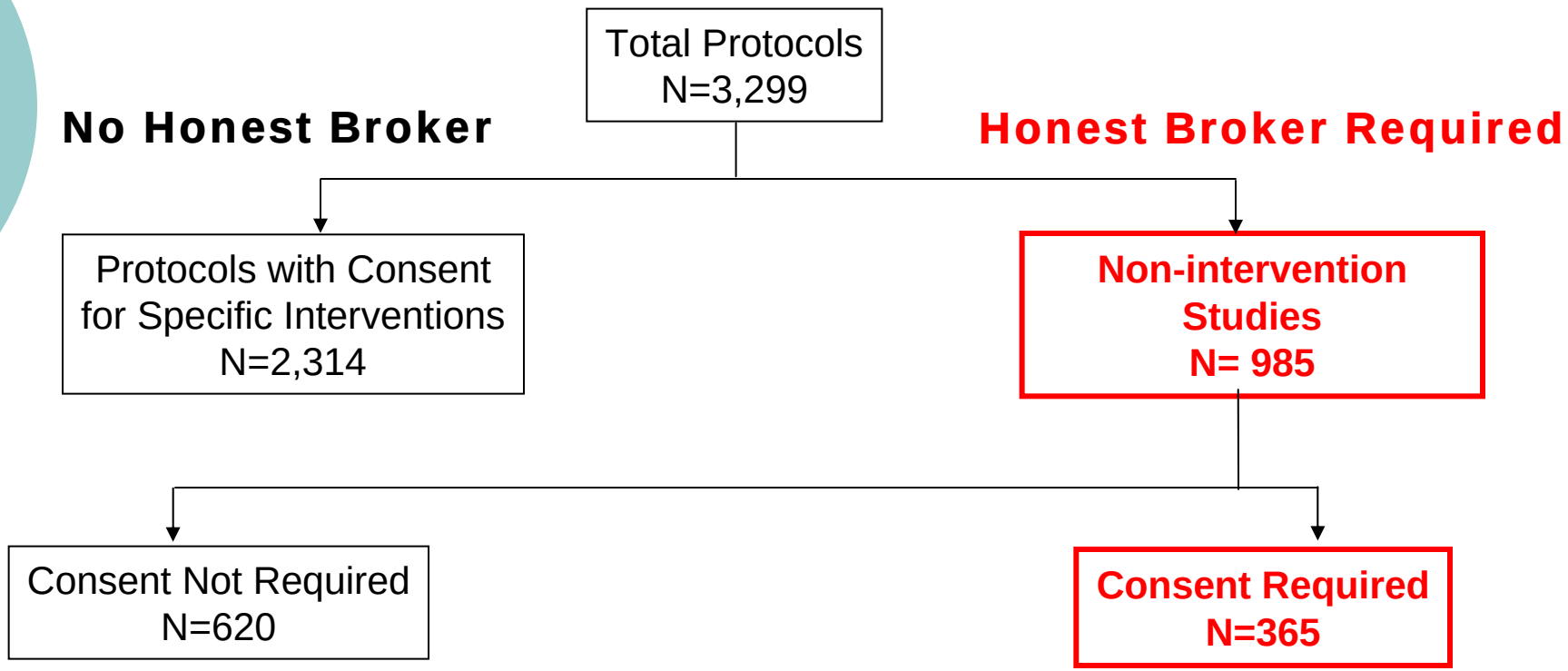


- Biostatistics Division first needs to provide ‘Honest Broker’ service to eliminate any dissenting patients

Definition of “Honest Broker”

- Impartial party and process to determine whether patient’s wishes would be violated by:
  - Analyzing their data arising from standard care, or
  - Studying their discard tissue samples
- Moving to single “General Research Consent” for all patients going forward
  - However many different consents used for various studies in the past, must be considered

Protocols Requiring “Honest Broker” Process



Patient Consent Status

Consent Status	N	Percent
Consented	70,133	92.5
Dissented	5,637	7.4
Consent Withdrawn (Dissented)	25	< 1.0
Total	75,795	

Honest Broker Algorithmic Approach

- Use computerized algorithms to evaluate:
  - Consent Type
  - Participation Type
- Any “no” response to a consent/participation type ***related to objectives*** of the study do ***not*** include in cohort
- Note: Consents change over time
  - May require different algorithms depending on protocol version

Participation Type Field Examples

agree to allow the collection and storage of samples from tissue removed during surgical resection of breast tissue (performed as part of my routine clinical care), and for the specimens to be linked to the clinical data obtained from my medical records for the purposes of cancer research
I agree to allow the collection of clinical data to be stored in a clinical research database. I understand that this clinical database will be updated with data obtained from my ongoing medical care at City of Hope
I agree to have my medical information and my blood stored for future research
I agree to have my specimens stored for future research purposes
I agree to have my specimens stored for future research purposes.
I agree to have my/my child's tissue or specimens stored for future research purposes.
I agree to have my/my child's name and contact information released to an agency under contract with COH, one of which is Examined Management Services, Inc. (EMSI)
I agree to have tissue stored for future research
I agree to have tissue stored for future research.
I agree to participate in the focus group
I agree to provide a blood sample to be used as part of this study
I agree to provide a urine sample to be used as part of this study
I agree to the collection and storage of samples from tissue removed during biopsy or surgical resection of prostate cancer (performed as part of my routine clinical care) for the purposes of cancer research
I agree to the collection of blood samples during routine clinical care that will be stored in blood bank for the purposes of cancer research
I agree to udnergo research blood sample collection at the first 3 proposed time points as part of this study
I agree to undergo research blood sample collection at all proposed time points as part of this study
I do not agree to have my child's name and contact information released to an agency under contract with COH, one of which is Examined Management Services, Inc. (EMSI)
I will allow additional needle biopsy specimens to be obtained during a routine and required diagnostic procedure for the purposes of cancer research

Participation Type Responses

- 218 different participation types across all non-interventional studies
- Multiple participation types within a given consent form require complex coding algorithms:
  - Example of participations within 1 consent form:
    - I allow my data to be collected for research='Yes'
    - I allow a routine sample to be used for research='Yes'
    - I allow an extra sample to be collected for research='No'
    - I allow my clinical and sample data to be linked='No'

Conclusions

- Advances in personalized medicine will require genome-phenome correlations
- Data warehousing is an optimal approach
- Business metadata are critical for valid use of data
- Requires complex computational algorithms to
  - Assemble appropriate cohort
  - Mine data, particularly non-standardized, text
  - Ensure valid linkages and data extraction
  - Protect patient privacy wishes

Thank You