INOVA™
Join the future of health.

Genomic Studies at Inova

Inova Translational Medicine Institute
Aaron Black PMP
April 1, 2015

INOVA® HEALTH SYSTEM

Organization

- Background
- Data management
- Challenges \ Lessons Learned

Background:
The Healthcare System

Six hospital + ambulatory healthcare system

- Largest healthcare system in Northern VA
- Two million patient visits/year
- 20,000 deliveries/year

Main hospital

- 1,000 beds including Children's hospital
- 10,000 deliveries/year

Background: The Research Institute

- Started: 2010
- Overall goal: research on integration of genomic information into the practice of medicine
- Staffing: 1/3 clinical, 1/3 bioinformatic/IT, 1/3 laboratory

Background: ITMI Studies

Themes

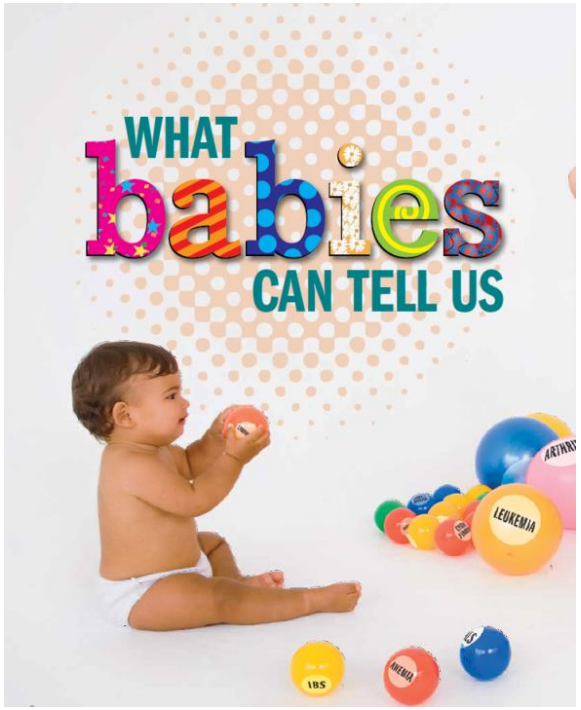
- Trio-based WGS
- Other ‘omics
- Comprehensive clinical data
- Integrated Laboratory
- Unified database

Preterm Birth Study (2011)

- Molecular associations with preterm birth
- ~500 PTB trios, ~500 FT trios
- WGS + ‘omics + clinical data
- Specimens: blood, saliva, cord blood, placenta

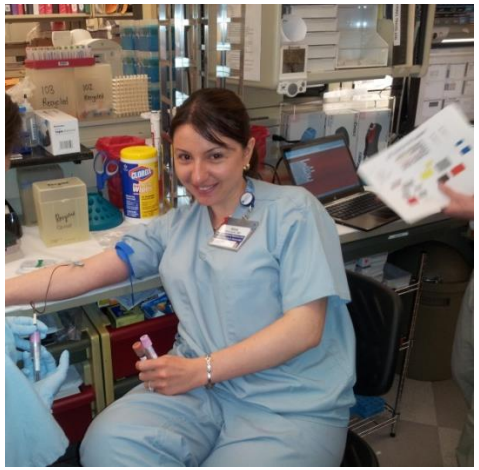
Longitudinal Study (2012)

- **WGS + 'omics + clinical data on 5,000 → 10,000 <u>trio-based</u> families**
- Longitudinal study (≥18 yrs)
- Blood, saliva, urine, cord blood, placenta
- DNA, RNA, protein, epigenetic + clinical data



Congenital Disorders Study (2012)

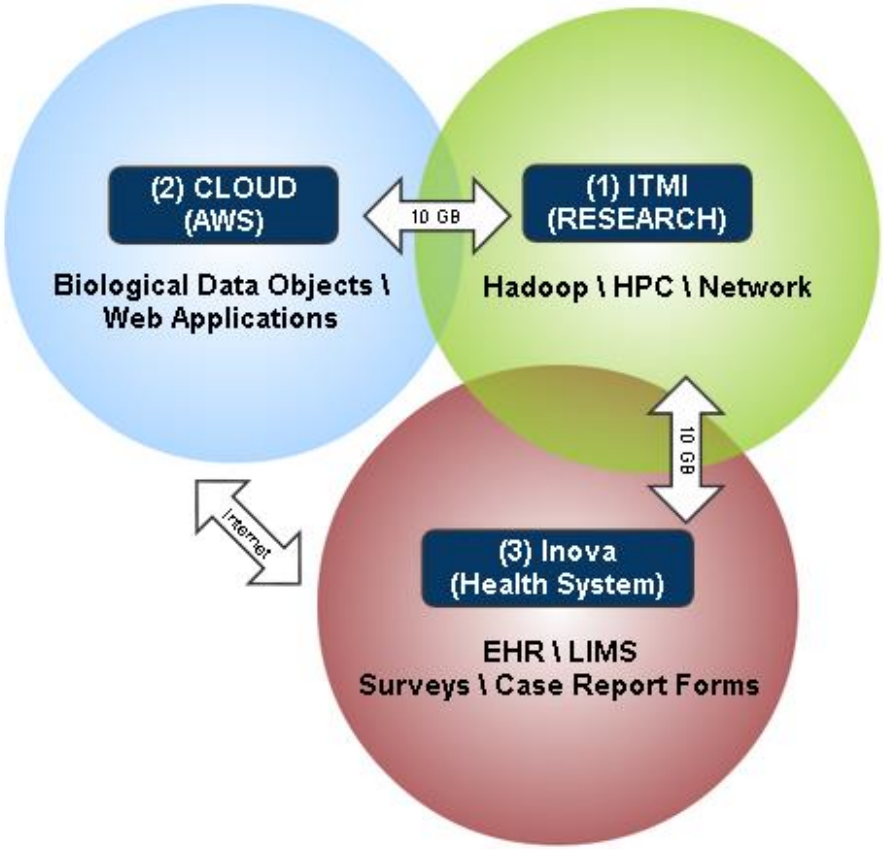
- Mostly NICU-based
- Any other patient with a "congenital/genetic" disorder
- ~2-3 families/week
- Trio-based WGS, etc.



Data Management

Infrastructure: Hybrid Cloud

Cloud



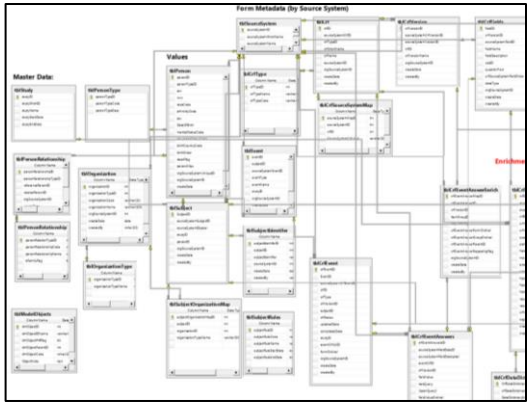
On-Premises

INOVA™
Join the future of health.

Database

Yes, a Hybrid

Relational Databases
Inova Infrastructure



Object Store
AWS

Storage & Content Delivery

S3
Scalable Storage in the Cloud

Storage Gateway
Integrates On-Premises IT Environments with Cloud Storage

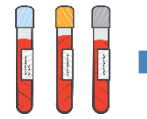
Glacier
Archive Storage in the Cloud

NoSQL \ Hadoop
Value pair \ Graph
Research Network

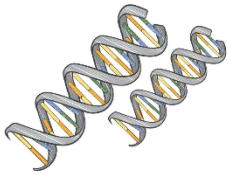
```
{
 "type": "record",
 "name": "HandshakeRequest",
 "namespace":"org.apache.avro.ipc",
 "fields": [
  {"name": "clientHash","type": {"type": "fixed", "name": "MD5", "size": 16}
  {"name": "clientProtocol", "type": ["null", "string"]},
  {"name": "serverHash", "type": "MD5"},
  {"name": "meta", "type": ["null", {"type": "map", "values": "bytes"}]}
 ]
}

{
 "type": "record",
 "name": "HandshakeResponse",
 "namespace": "org.apache.avro.ipc",
 "fields": [
  {"name": "match","type": {
      "type": "enum",
      "name": "HandshakeMatch",
      "symbols": ["BOTH", "CLIENT", "NONE"]}},
  {"name": "serverProtocol","type": ["null", "string"]},
  {"name": "serverHash",
      "type": ["null", {"type": "fixed", "name": "MD5", "size": 16}]},
  {"name": "meta", "type": ["null", {"type": "map", "values": "bytes"}]}
 ]
}
```

INOVA™
Join the future of health.

ITMI Data Collection



Bio Materials



Omic Data

Reporting
& Analysis

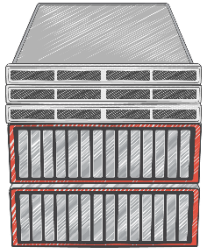
De-identified Data

Meta data

Only As Needed

Hybrid Storage Sync

Consent Patient



On-premises Epic &
Laboratory &
Clinical &
Datawarehouse

External
Collaborators

On-premises Inova
Research staff

On-premises HPC Cluster & Storage

1024 Cores
1.2 PB storage
16 TB Ram

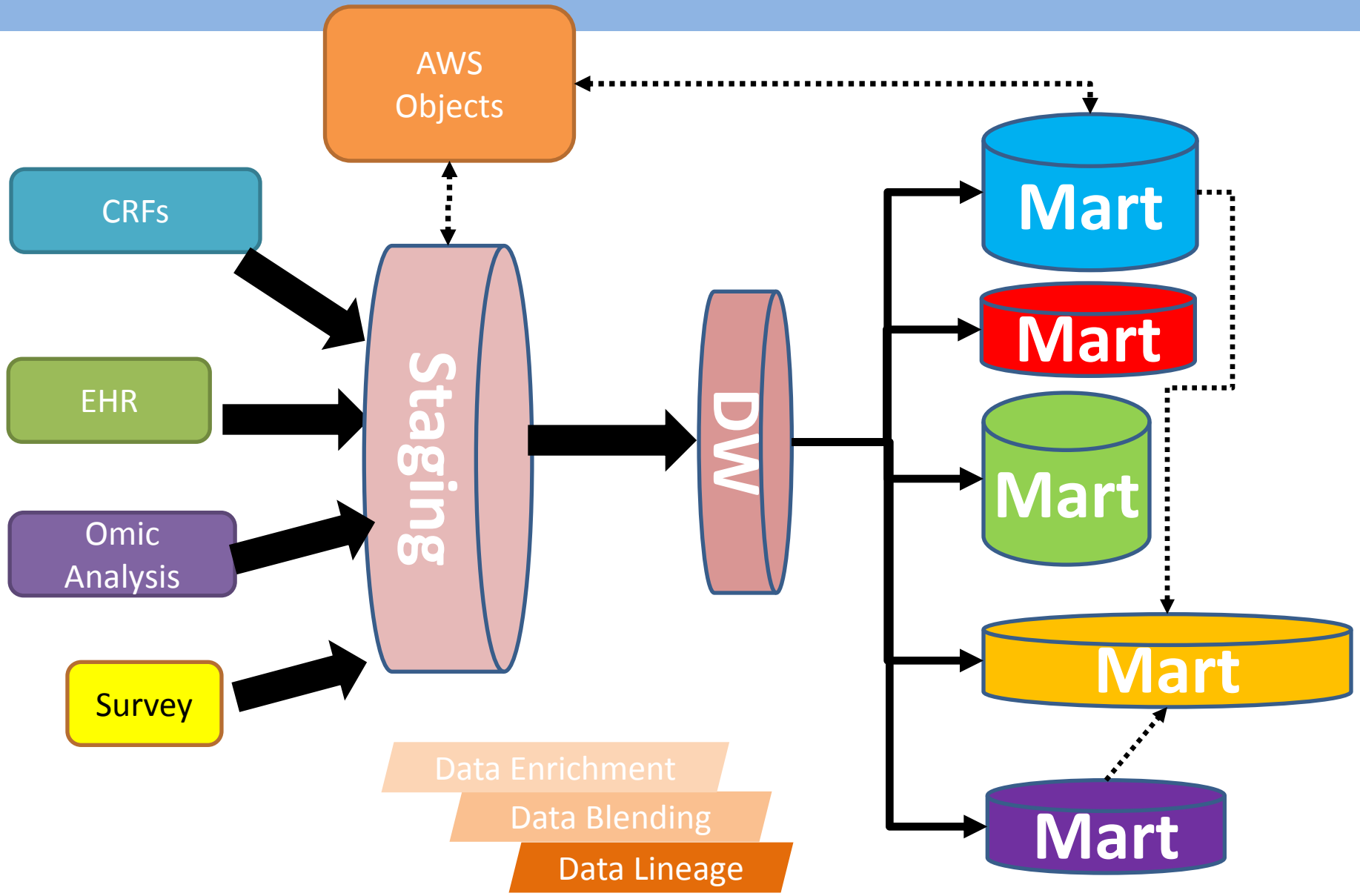
INOVA™
Join the future of health.

The Stats

- **Participants**
  - *9,750* participants (enrolled), *> 110* Different Countries of Birth
- **Files and data**
  - Manage 3+ Petabytes of storage capacity (between cloud and on-premise)
  - ~10,000,000 million files
  - File Sizes Range from Kilobytes to 100+ GB per file
- **Specimens**
  - *>450,000* specimens
- **Whole Genomes Sequences**
  - *7,300 +*
  - *~36,500,000,000 variants!*
  - *Also have epigenetic data*
- **Clinical**
  - *55,000+* Patient Diagnosis (Longitudinal)
  - *110,000+*Surveys and Case Report Forms
- **Labs Results**
  - *2,000,000+* discrete lab results
  - *2,200,000+* discrete variables from Case Report Form and Surveys

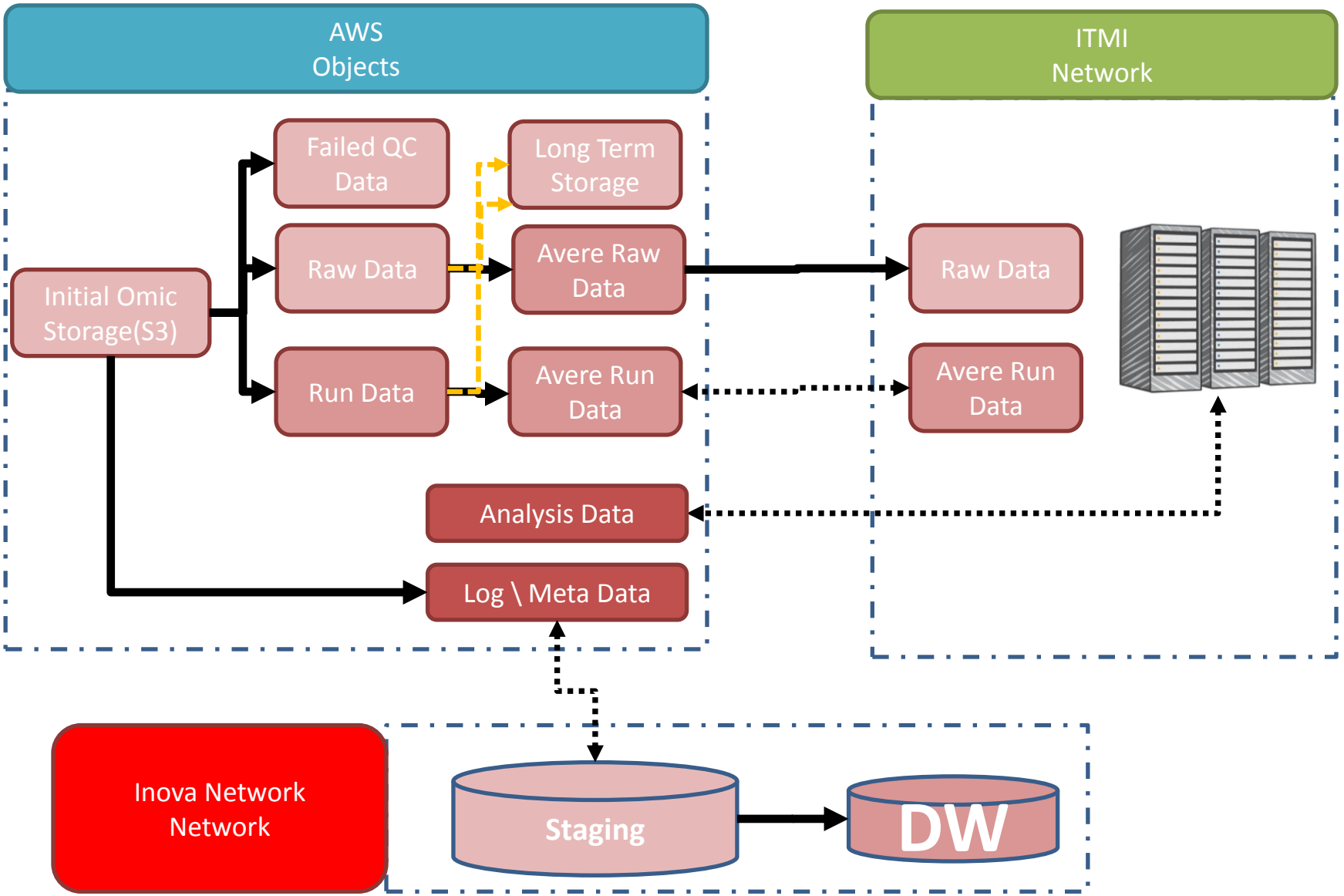
INOVA™
Join the future of health.

General Data Process



INOVA™
Join the future of health.

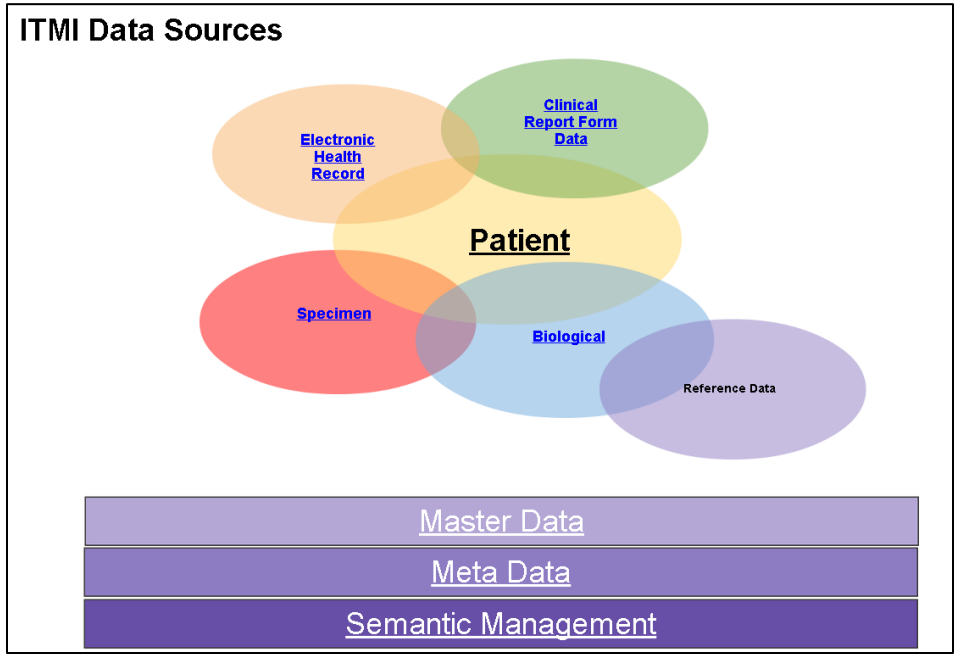
Omic Data Process



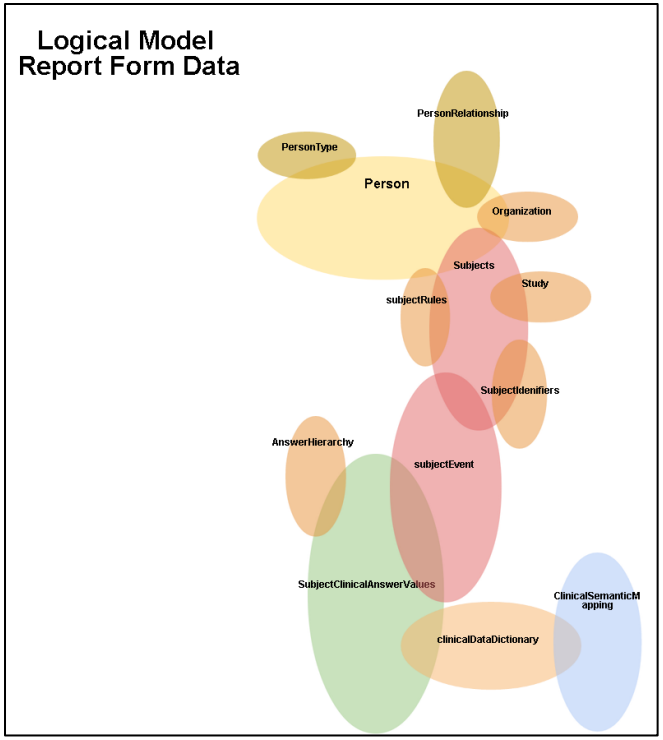
INOVA™
Join the future of health.

Data Model

(1) Logically Model

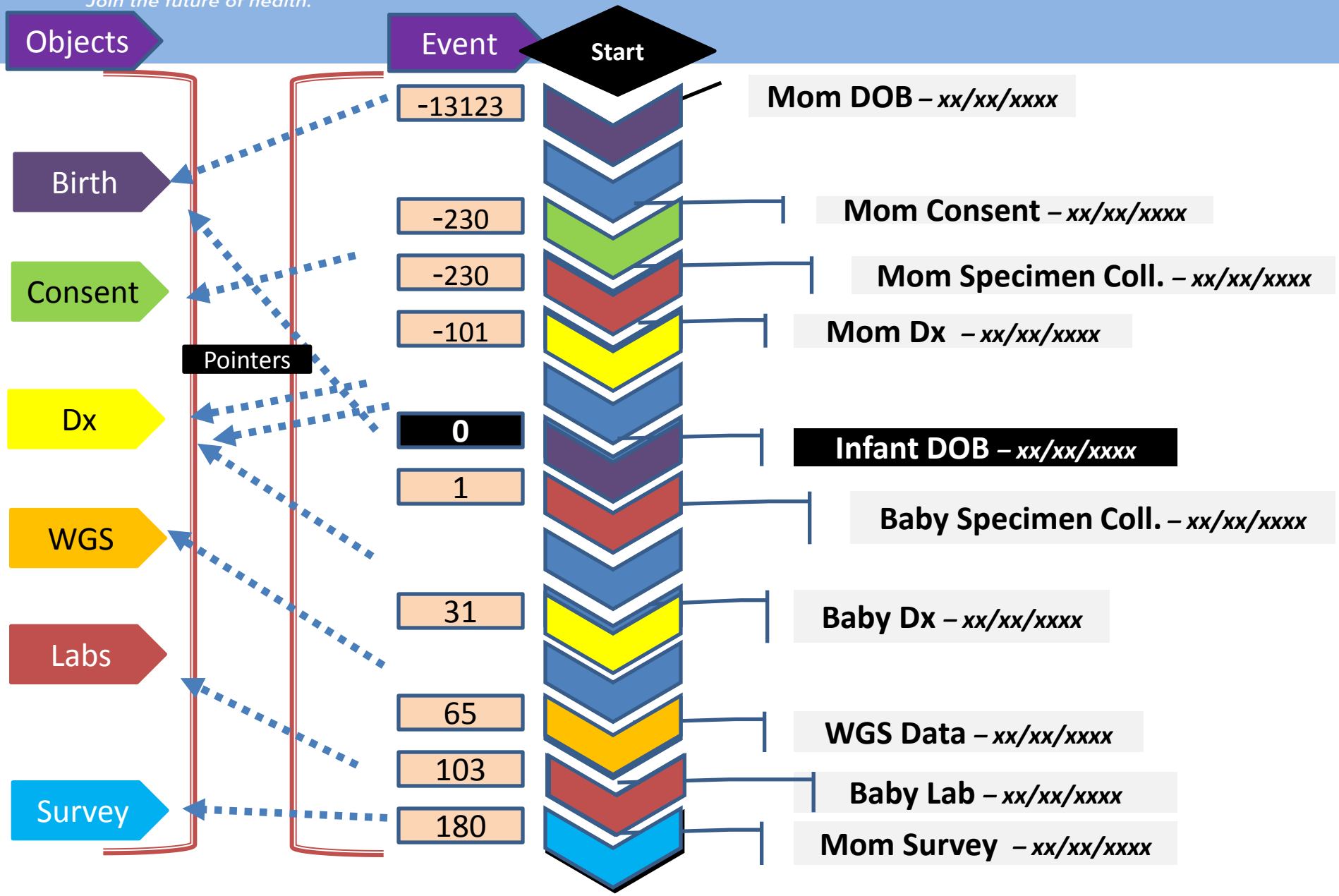


(2) Break down each



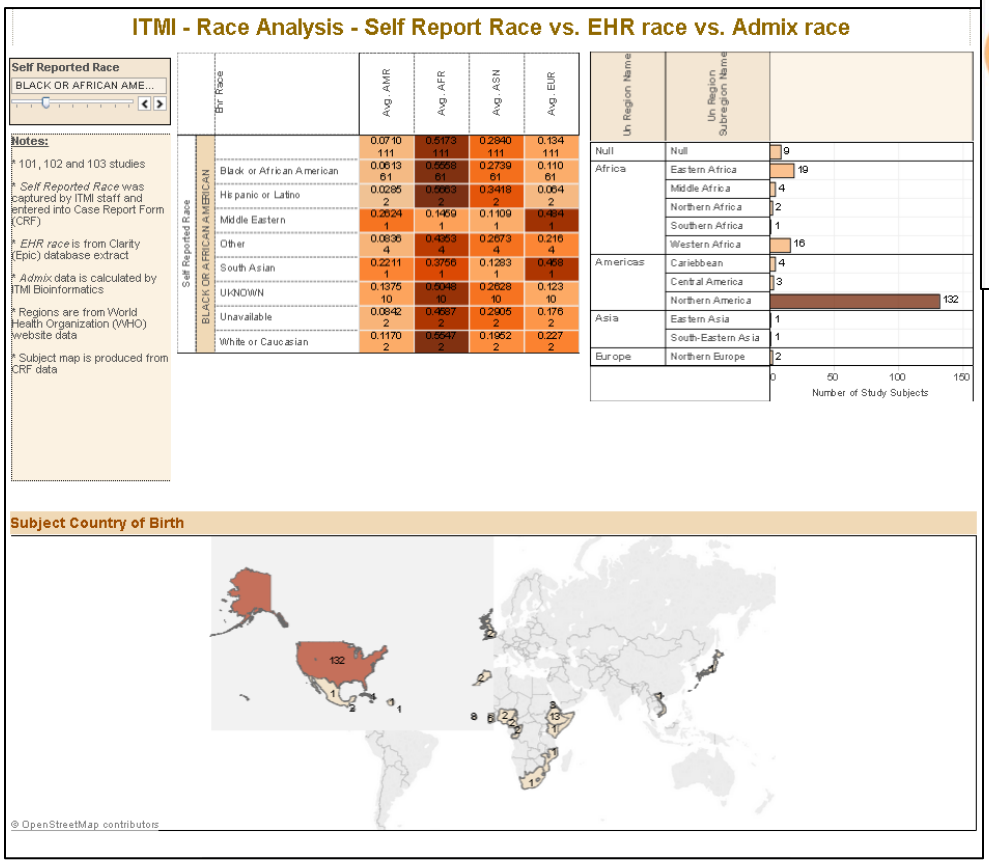
INOVA™
Join the future of health.

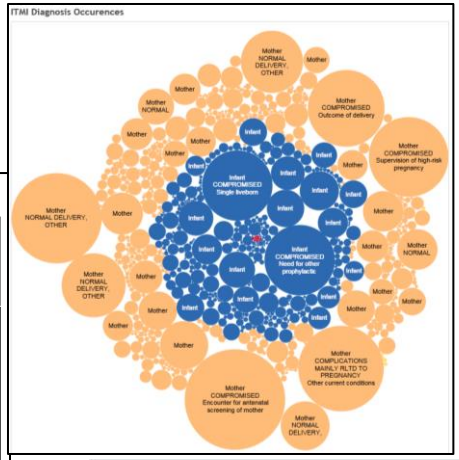
LONGITUDINAL DATA

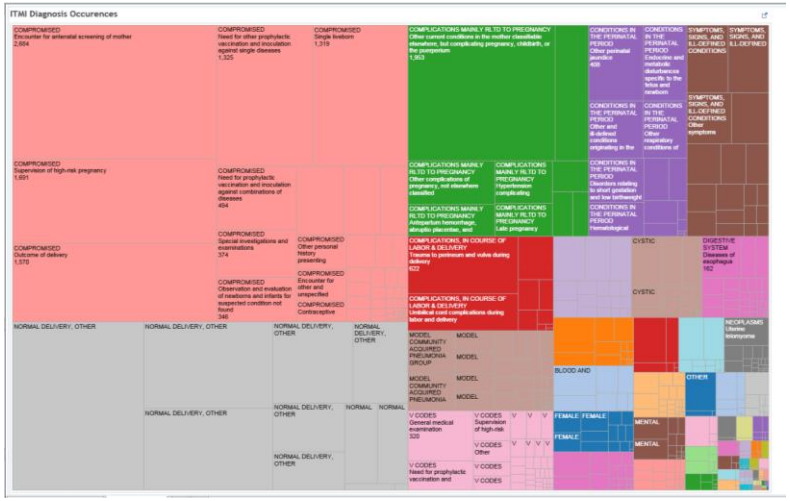


INOVA™
Join the future of health.

Visualization

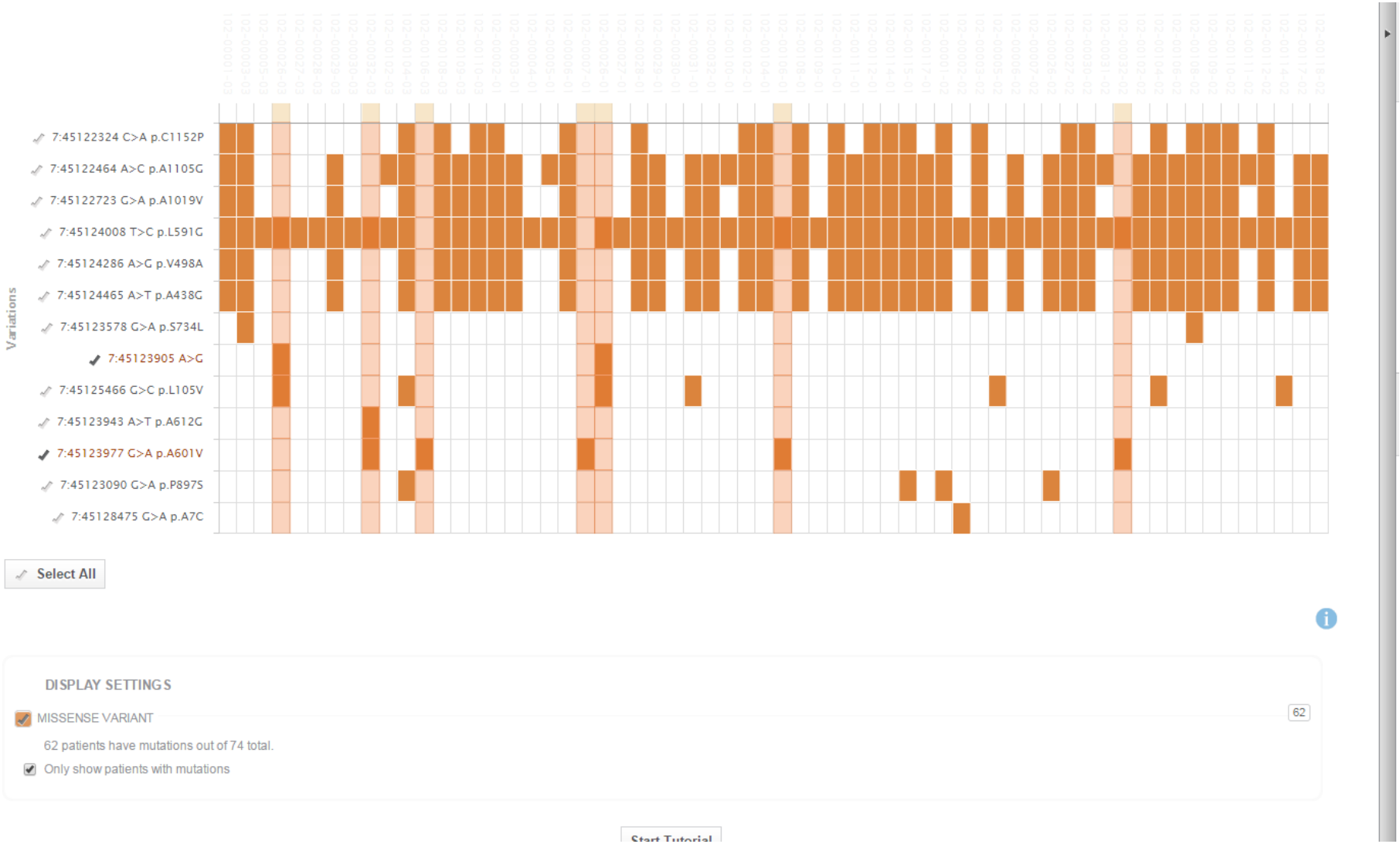






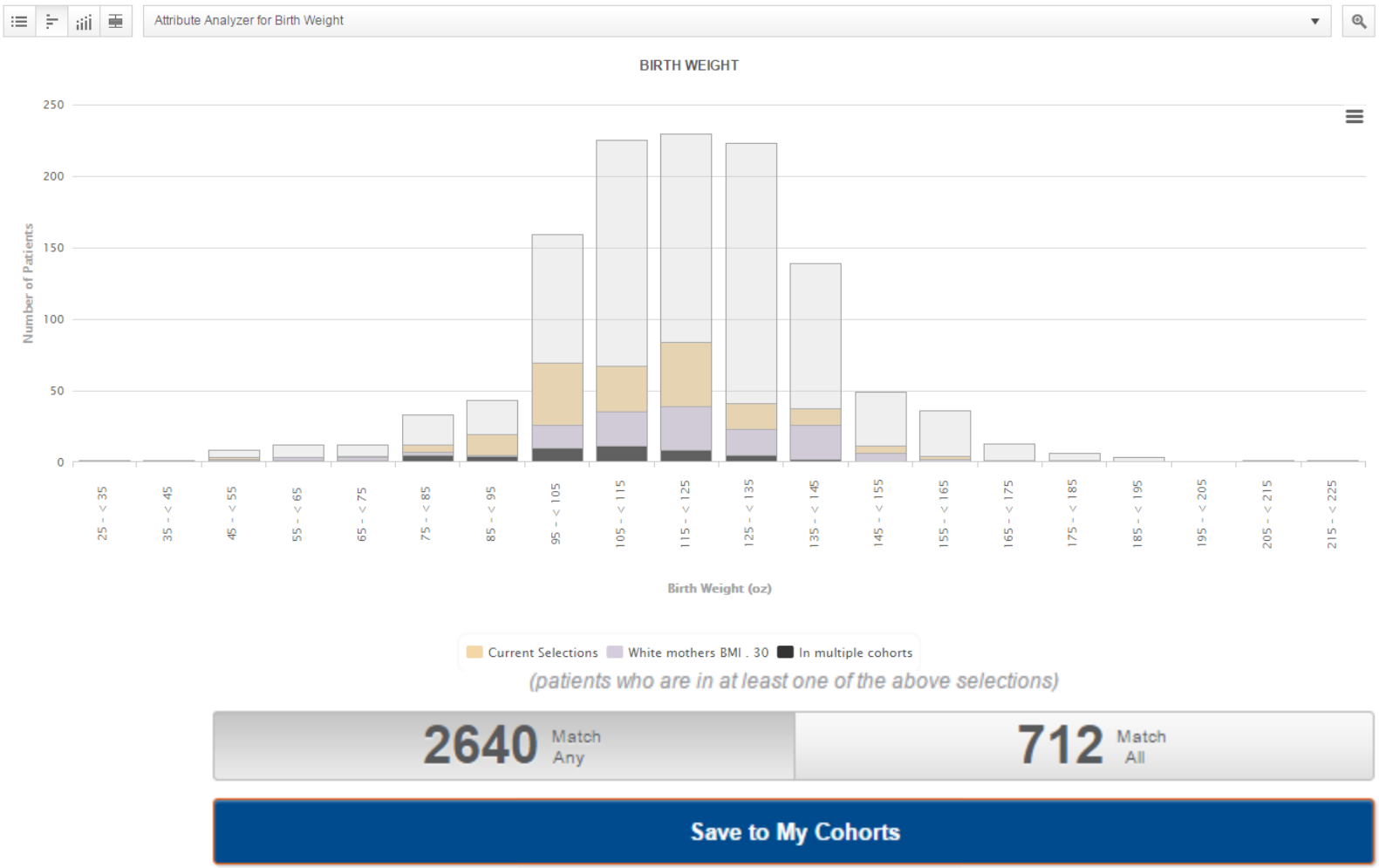
INOVA™
Join the future of health.

Web Portal



INOVA™
Join the future of health.

Web Portal



III. Challenges \ Lessons Learned

INOVA™
Join the future of health.

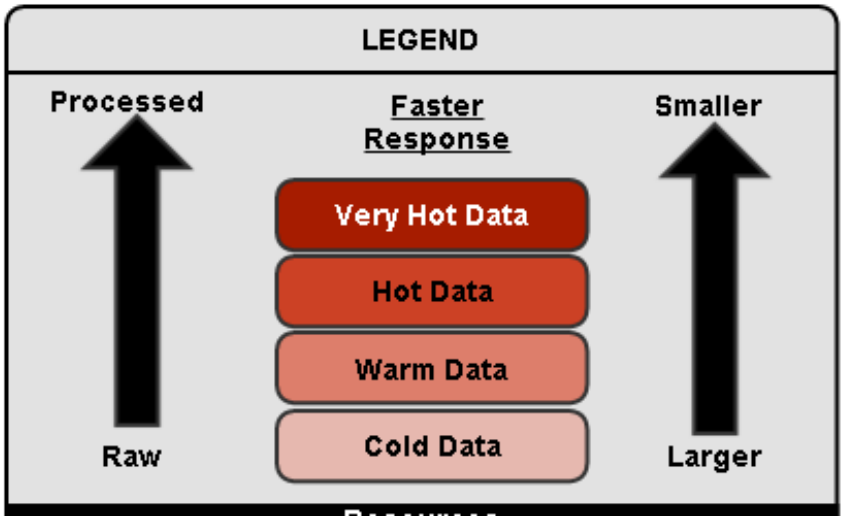
Challenges

- Scalability
  - Storage
  - Compute
- Data Movement
- IT Standards
  - Data
  - Programming

INOVA™
Join the future of health.

Lessons Learned

- Build infrastructure around data
  - Network is bottleneck
  - Hidden costs
- Manage Data Tiers



INOVA™
Join the future of health.

Lessons Learned

- Spend time to model
  - Build strong metadata layer
  - Make it understandable to your team
  - Use best practices
- IT Partners that understand the business

Acknowledgements

Inova Translational Medicine Institute
John Niederhuber MD
Joe Vockley, PhD
Greg Eley, PhD
Aaron Black, PMP
Kathi Huddleston, PhD
Ram Iyer, PhD
Dale Bodian, PhD
Wendy Wong, PhD
Alina Khromykh, MD
Dan Stauffer, PhD
Sarah Ruppert, CGC
Tiffani DeMarco, CGC
Kim Rutledge, CGC
and team

Inova Health System
David Ascher, MD
Larry Maxwell, MD
Al Khoury, MD
George Bronsky, MD
Barbara Nies, MD
and team

Fairfax Neonatal Associates
Robin Baker, MD
Rajiv Baveja, MD
and team

Institute for System Biology
Ilya Shmulevich, PhD
Jared Roach, MD, PhD
Brady Bernard, PhD
Gustavo Glusman, PhD
and team