Precision Medicine *Analytics Platform*

Paul Nagy, PhD, FSIIM
Associate Professor of Radiology
Division of Health Science Informatics
Armstrong Institute for Quality and Patient Safety
Deputy Director of JHM Technology Innovation Center
pnagy@jhu.edu

JOHNS HOPKINS APPLIED PHYSICS LABORATORY

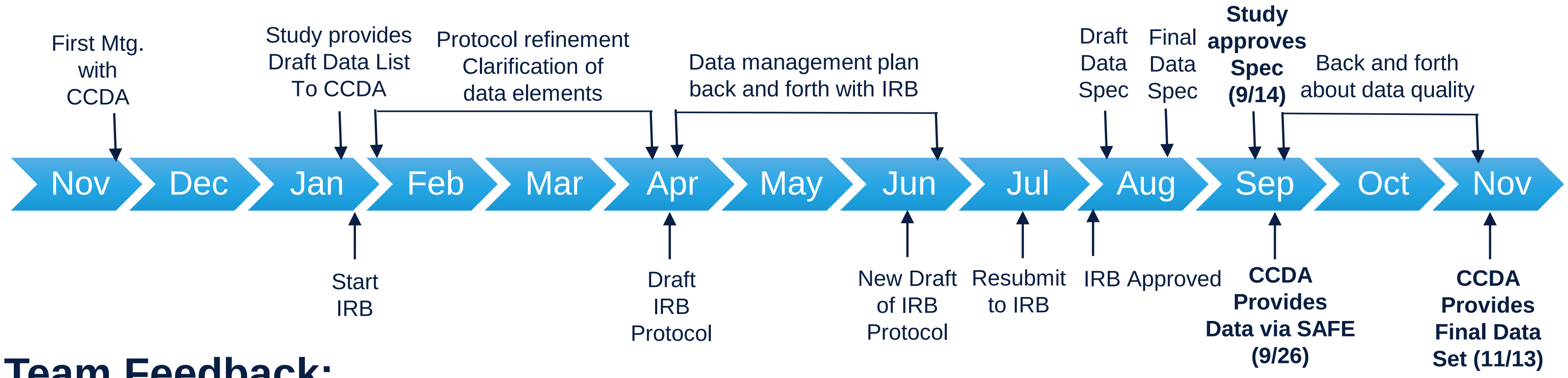
JOHNS HOPKINS UNIVERSITY

JOHNS HOPKINS MEDICINE

On Average
4.5 mo. start-finish
20% on data
extraction

Case Study – Research Data Access

Timeline & Feedback from CCDA Customer


Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
First Mtg. with CCDA
Study provides Draft Data List To CCDA
Protocol refinement Clarification of data elements
Data management plan back and forth with IRB
Draft Data Spec
Final Data Spec
Study approves Spec (9/14)
Back and forth about data quality
Start IRB
Draft IRB Protocol
New Draft of IRB Protocol
Resubmit to IRB
IRB Approved
CCDA Provides Data via SAFE (9/26)
CCDA Provides Final Data Set (11/13)

Study Team Feedback:

1. **No** comprehensive **overview of the entire process**. We stumbled our way through.

2. Not much guidance on **how to complete the various IRB forms**. A **standard template** would help.

3. **Review (of data management plan for IRB**) is a rate-limiting step because it is only done by one person.

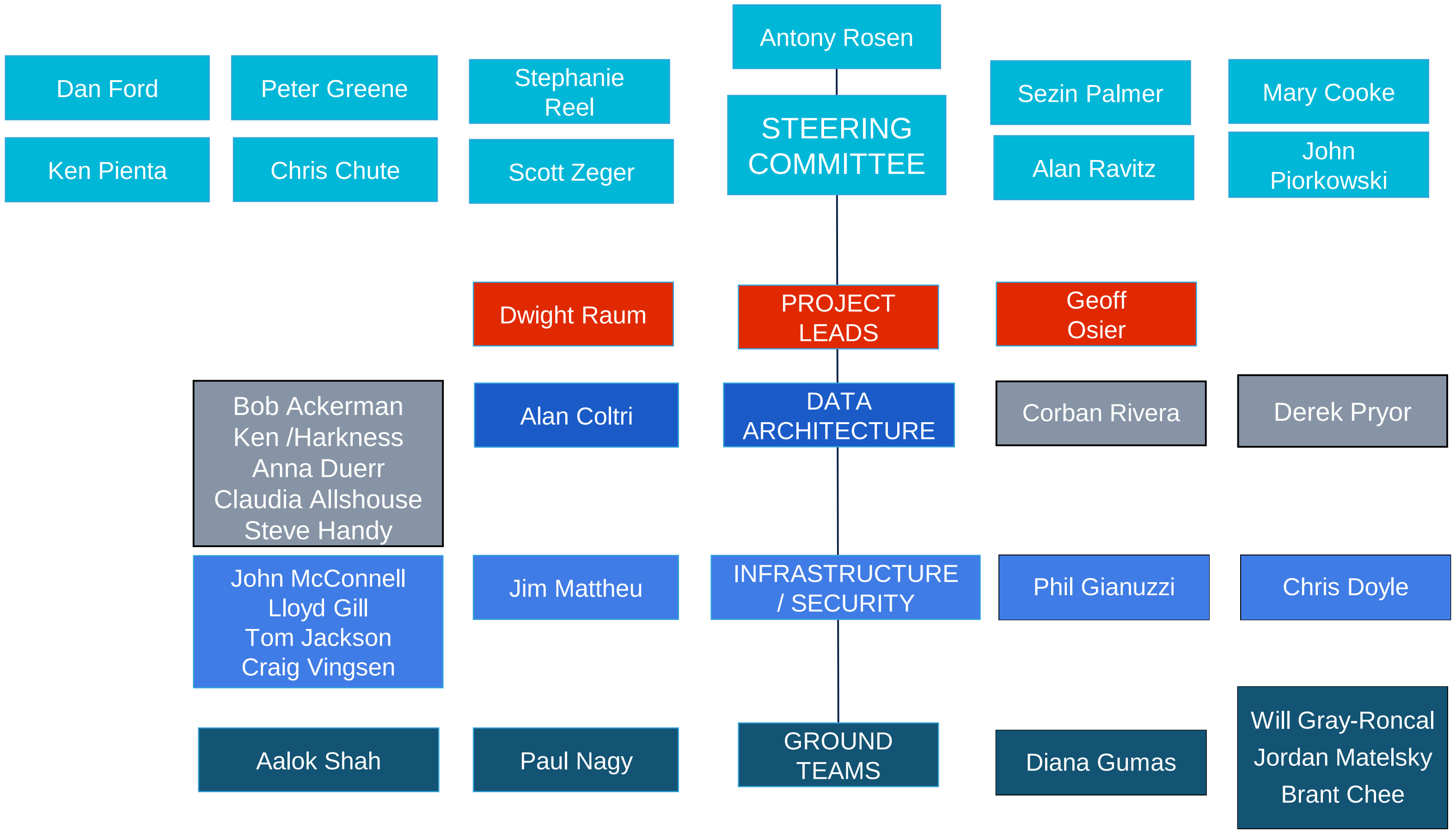
4. Researchers have to do a lot of work trying to figure out exactly what data elements they need and then provide the CDDA. A **meta-thesaurus would be useful**.

5. Researchers do not know **which data can be obtained easily** and those that cannot.

JOHNS HOPKINS inhealth

*in*Health is using revolutionary tools of measurement, data science, and connectivity to discover clinically-relevant and biologically-anchored subgroups at scale, and to deliver what we learn to impact the precision and value of health care

School of Medicine and Applied Physics Laboratory



Antony Rosen

Dan Ford

Peter Greene

Stephanie Reel

STEERING COMMITTEE

Sezin Palmer

Mary Cooke

Ken Pienta

Chris Chute

Scott Zeger

Alan Ravitz

John Piorkowski

Dwight Raum

PROJECT LEADS

Geoff Osier

Bob Ackerman
Ken /Harkness
Anna Duerr
Claudia Allshouse
Steve Handy

Alan Coltri

DATA ARCHITECTURE

Corban Rivera

Derek Pryor

John McConnell
Lloyd Gill
Tom Jackson
Craig Vingsen

Jim Mattheu

INFRASTRUCTURE / SECURITY

Phil Gianuzzi

Chris Doyle

Aalok Shah

Paul Nagy

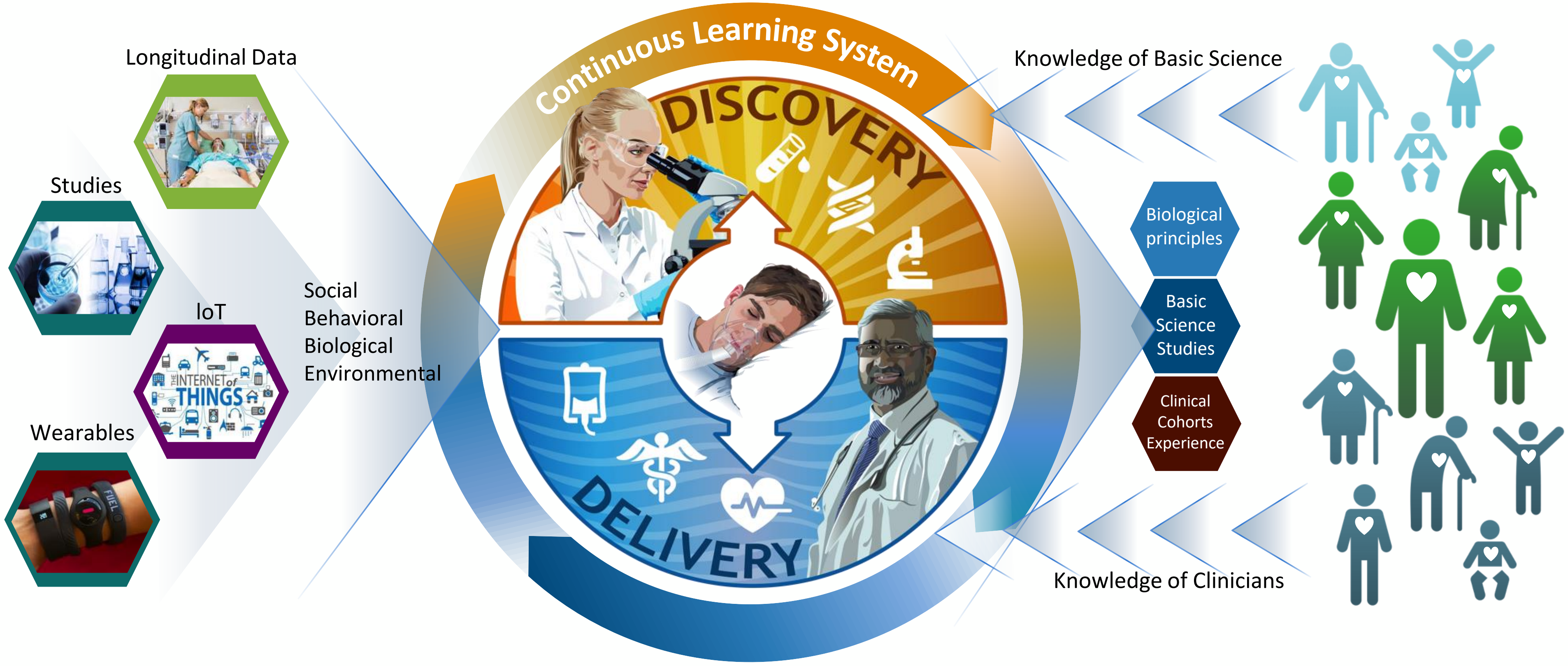
GROUND TEAMS

Diana Gumas

Will Gray-Roncal
Jordan Matelsky
Brant Chee

Large, multidisciplinary, cross-functional team

Changing Face of Medicine: Centers of Excellence



Longitudinal Data

Studies

IoT

Wearables

Social
Behavioral
Biological
Environmental

Knowledge of Basic Science

Biological principles

Basic Science Studies

Clinical Cohorts Experience

Knowledge of Clinicians

Tenets of PMAP

1. Researchers need better access to clinical data.
2. Researchers need environments that ensure data security protecting patient information recognizing full de-identification is difficult
3. Researchers need an environment that is built for machine learning and data science to enable discovery.
4. Combined access to very different data types: EMR, medical imaging, genomics, and physiological monitoring data.
5. Clinical researchers need to bring new discoveries into clinical care.

Platform Components

Data Platform

- Confidentiality
- Integrity
- Availability
- Authentication
- Authorization
- Accounting

Tooling

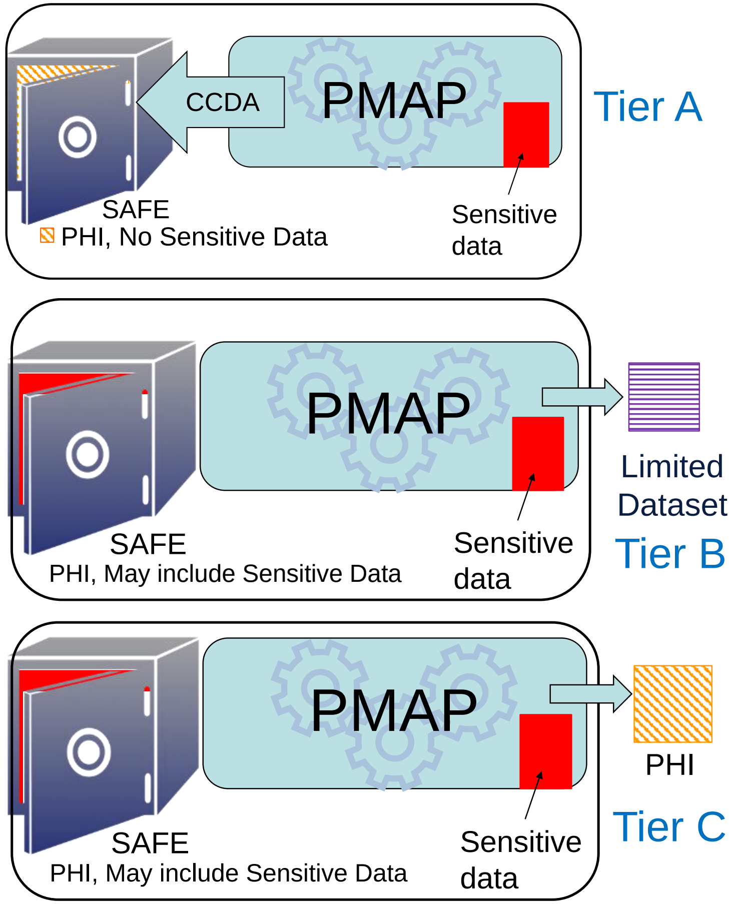
- Patient matching
- Data Catalog
- Cohort Discovery
- Honest Broker
- Annotation
- Preprocessing

Research Projections

- SQL Server
- Cohort Dashboard
- Jupyter Notebooks
- Docker
- Compute

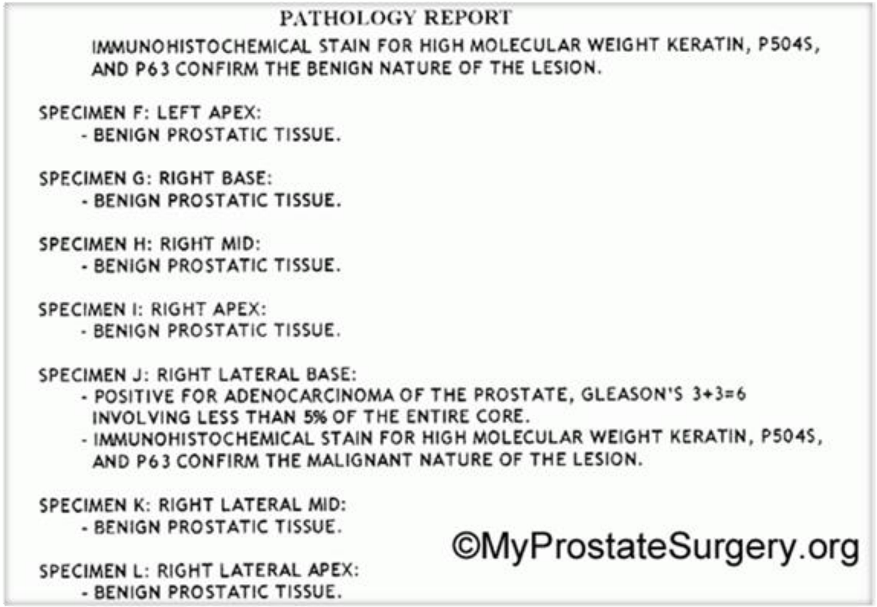
Make JHM Easy

- Risk Tiers – Bioethical Framework
  - Proposed by IRB and Data Trust
  - Created to accommodate data science in PMAP
  - Leverages secure analysis environment (SAFE)
- Tier A Proposals
  - Approved as a class by Data Trust
  - IRB review streamlined
  - Investigators reference data categories rather than data elements
- Positive Investigator Impact
  - Simplified data specification (save hours)
  - Faster review time (save months)



NLP

- Current research being done with Prostate Cancer CoE
- > 95% accurately extracts Gleason scores and anatomical references from notes
- Post-processing on data allows
  - Large-scale inference
  - Error detection and other data quality downstream analytics
- Beginning to evaluate using this for other CoEs



PATHOLOGY REPORT

IMMUNOHISTOCHEMICAL STAIN FOR HIGH MOLECULAR WEIGHT KERATIN, P504S, AND P63 CONFIRM THE BENIGN NATURE OF THE LESION.

SPECIMEN F: LEFT APEX:
- BENIGN PROSTATIC TISSUE.

SPECIMEN G: RIGHT BASE:
- BENIGN PROSTATIC TISSUE.

SPECIMEN H: RIGHT MID:
- BENIGN PROSTATIC TISSUE.

SPECIMEN I: RIGHT APEX:
- BENIGN PROSTATIC TISSUE.

SPECIMEN J: RIGHT LATERAL BASE:
- POSITIVE FOR ADENOCARCINOMA OF THE PROSTATE, GLEASON'S 3+3=6 INVOLVING LESS THAN 5% OF THE ENTIRE CORE.
- IMMUNOHISTOCHEMICAL STAIN FOR HIGH MOLECULAR WEIGHT KERATIN, P504S, AND P63 CONFIRM THE MALIGNANT NATURE OF THE LESION.

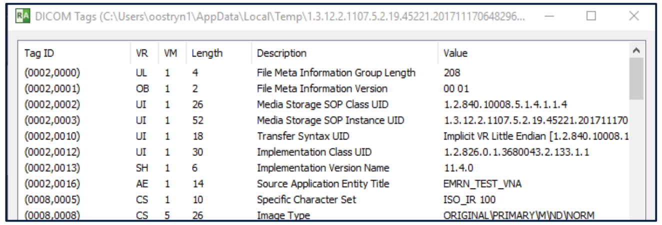
SPECIMEN K: RIGHT LATERAL MID:
- BENIGN PROSTATIC TISSUE.

SPECIMEN L: RIGHT LATERAL APEX:
- BENIGN PROSTATIC TISSUE.

©MyProstateSurgery.org

Medical Images VNA Access

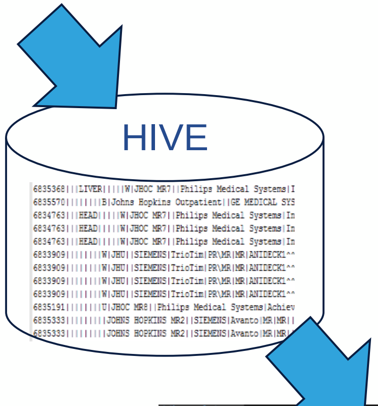
- PMAP Data Commons only stores the imaging metadata (DICOM)
  - Patient, study, series, image
- Users can query DICOM with Hive's SQL-like language
- Users (with appropriate permissions) request images to be fetched for their Projections
- Images pulled from VNA
- Deep learning GPU Compute



DICOM Tags (C:\Users\oostryn1\AppData\Local\Temp\1.3.12.2.1107.5.2.19.45221.201711170648296...

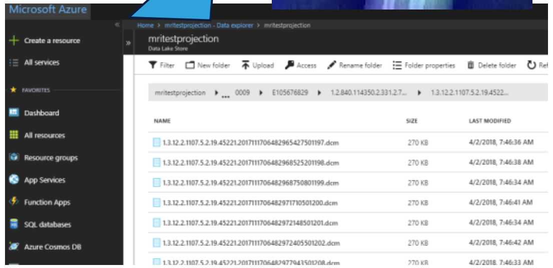
Tag ID	VR	VM	Length	Description	Value
(0002,0000)	UL	1	4	File Meta Information Group Length	208
(0002,0001)	OB	1	2	File Meta Information Version	00 01
(0002,0002)	UI	1	26	Media Storage SOP Class UID	1.2.840.10008.5.1.4.1.1.4
(0002,0003)	UI	1	52	Media Storage SOP Instance UID	1.3.12.2.1107.5.2.19.45221.201711170
(0002,0010)	UI	1	18	Transfer Syntax UID	Implicit VR Little Endian [1.2.840.10008.1
(0002,0012)	UI	1	30	Implementation Class UID	1.2.826.0.1.3680043.2.133.1.1
(0002,0013)	SH	1	6	Implementation Version Name	11.4.0
(0002,0016)	AE	1	14	Source Application Entity Title	EMRN_TEST_VNA
(0008,0005)	CS	1	10	Specific Character Set	ISO_IR 100
(0008,0008)	CS	5	26	Image Type	ORIGINAL\PRIMARY\M\ND\NORM

VNA DICOM Data



HIVE

```
6835368|||LIVER|||||W|JHOC MR7||Philips Medical Systems|I
6835570||||||||B|Johns Hopkins Outpatient||GE MEDICAL SYS
6834763|||HEAD|||||W|JHOC MR7||Philips Medical Systems|In
6834763|||HEAD|||||W|JHOC MR7||Philips Medical Systems|In
6834763|||HEAD|||||W|JHOC MR7||Philips Medical Systems|In
6833909||||||||W|JHU||SIEMENS|TrioTim|PR\MR|MR|ANIDECK1^^
6833909||||||||W|JHU||SIEMENS|TrioTim|PR\MR|MR|ANIDECK1^^
6833909||||||||W|JHU||SIEMENS|TrioTim|PR\MR|MR|ANIDECK1^^
6833909||||||||W|JHU||SIEMENS|TrioTim|PR\MR|MR|ANIDECK1^^
6835191||||||||U|JHOC MR8||Philips Medical Systems|Achiev
6835333|||||||||JOHNS HOPKINS MR2||SIEMENS|Avanto|MR|MR||
6835333|||||||||JOHNS HOPKINS MR2||SIEMENS|Avanto|MR|MR|
```



mritestprojection ▸ ... 0009 ▸ E105676829 ▸ 1.2.840.114350.2.331.2.7... ▸ 1.3.12.2.1107.5.2.19.4522...

NAME	SIZE	LAST MODIFIED
1.3.12.2.1107.5.2.19.45221.2017111706482965427501197.dcm	270 KB	4/2/2018, 7:46:36 AM
1.3.12.2.1107.5.2.19.45221.2017111706482968525201198.dcm	270 KB	4/2/2018, 7:46:38 AM
1.3.12.2.1107.5.2.19.45221.2017111706482968750801199.dcm	270 KB	4/2/2018, 7:46:34 AM
1.3.12.2.1107.5.2.19.45221.2017111706482971710501200.dcm	270 KB	4/2/2018, 7:46:41 AM
1.3.12.2.1107.5.2.19.45221.2017111706482972148501201.dcm	270 KB	4/2/2018, 7:46:34 AM
1.3.12.2.1107.5.2.19.45221.2017111706482972405501202.dcm	270 KB	4/2/2018, 7:46:42 AM
1.3.12.2.1107.5.2.19.45221.2017111706482977943501208.dcm	270 KB	4/2/2018, 7:46:33 AM

Genomics on PMAP

JOHNS HOPKINS inhealth

On platform vs. Federation

- Caching strategy: cache gene and allele level features on the platform. Federate access to raw sequence
- Cached formats: VCF, sample metadata, Gene expression matrices, genome references
- Federated: fastq, bam

Public Annotation Sources

- ClinVar, OMIM, GeneOntology

Functionality Targets

Cohort Discovery

Semantic search that allows the user to discover patients based on variants/ variant sets or genomic alterations (indels, CNV).

Cohort Extraction

A method for displaying/exporting variants and associated metadata for that group of patients returned by a search.

Data Explorer

A genome browser that allows the user to drill down and display the aligned sequence data and associated quality metrics in the chromosomal location defined by the user (around variants).

Data Commons

Genomics derived datasets linked by a patient identifier.

Variant Annotation

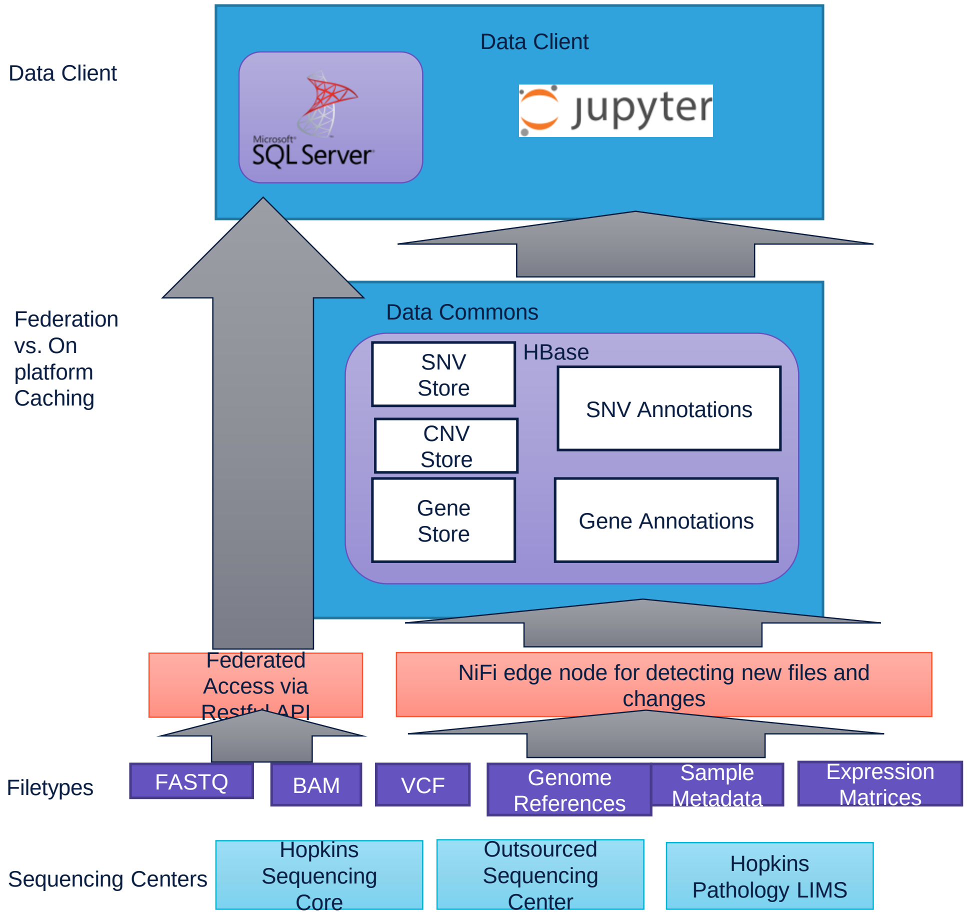
The ability to add characterizations to a variant, according to ACMG scales or other well-defined methods

Variant Filtering

The ability to filter a returned list of variants by criteria or metadata associated with the sequence.

Container Environment for Analysis Scaling

The ability to configure the container environment based upon the study needs



Data Client

Data Client

Federation vs. On platform Caching

Data Commons

HBase

SNV Store

CNV Store

Gene Store

SNV Annotations

Gene Annotations

Federated Access via Restful API

NiFi edge node for detecting new files and changes

Filetypes

FASTQ

BAM

VCF

Genome References

Sample Metadata

Expression Matrices

Sequencing Centers

Hopkins Sequencing Core

Outsourced Sequencing Center

Hopkins Pathology LIMS

Physiological Monitoring

- Hadoop environments are well suited for large streaming data real time analysis environments.
- Twitter Firehouse:
  - 500 Million tweets a day
- PhysioCloud research group.

shutterstock.com · 650081530

APACHE Spark™

Precision Medicine Portal

October 17, 2018

JOHNS HOPKINS
APPLIED PHYSICS LABORATORY

JOHNS HOPKINS
UNIVERSITY

JOHNS HOPKINS
MEDICINE

Overall Layout

Public Resources

Precision Medicine Introduction

What are inHealth, PMCOEs, and the PMAP, and why you would want to use these resources

News and Events: Relevant Conferences, Programs, Newly published resources / code examples

Researcher Resources

Becoming a PMCOE

- PMCOE in a Box Templates
- e.g. Value Matrix, Cohort in a Box

Templates for grants/IRB/Data Trust applications

Research Lifecycle (or link to ICTR resource)

Research Community Resources (e.g. browsing data scientist / faculty profiles, contact ICTR/CCDA/TIC)

PMAP Education

Documentation, Videos on using the technologies

Gallery of code / notebooks

Community Only (JHED Login)

PMAP Administration

- Sign up
- Account Management (Tiering, Usage reporting)
- Payment
- Team Management / Role Assignment
- Define databases to ingest into PMAP
- Request Data Projections
- Help / Bug Reporting / Open Issues

PMAP Tools

- Data Catalog
- Cohort Discovery
- Cohort Dashboard
- Crunchr
  - Access to compute resources, standard containers, shared notebooks, version control through Crunchr command line
- Data Annotation Tools
  - NLP, Imaging, Genomics annotation tools
- Projection Creator (select users only, e.g. CCDA)

CAMP

Center of Excellence Analytics
in Medicine Program

Accelerating team science in biomedical discovery

Course Directors: *Paul Nagy, PhD and Diana Gumas*

JOHNS HOPKINS
MEDICINE

JOHNS HOPKINS
INSTITUTE *for* CLINICAL & TRANSLATIONAL RESEARCH

APL JOHNS HOPKINS
APPLIED PHYSICS LABORATORY

TECHNOLOGY
INNOVATION CENTER

JOHNS HOPKINS inhealth

DHSi

CAMP

Center of Excellence Analytics in Medicine Program

Opportunity for PMAP team to work directly with the JHMI research teams

CAMP - CoE Analytics in Medicine Program

12 week course designed to introduce researchers to PMAP

Opportunity for PMAP team to work directly with the JHMI research teams

Provide an overview of PMAP’s infrastructure, data management, data science capabilities, and clinical applications

Further understanding of what tools will be most useful

Work together on research proposals

Non-PMCoE teams are charged

CAMP - CoE Analytics in Medicine Program

- ✓ Data Needs
- ✓ PMAP Overview
- ✓ Working With IRB / Data Trust
- ✓ Working With CCDA
- ✓ Epic Data Sources
- ✓ Machine Learning For Imaging
- ✓ Derived Features For Genomics
- ✓ Physio/Wearables Data Use
- ✓ Machine Learning For NLP
- ✓ Overview of Partner Institutions
- ✓ Cohort Dashboard
- ✓ Integrate Research Into Clinic

CAMP Team

CAMP
2018 FACULTY

Paul Nagy, PhD
Deputy Director, TIC
CAMP Director

Diana Gumas
Senior IT Director, ICTR
CAMP Director

Ken Pienta, MD
Professor, Urology Research
CAMP PMCOE Lead

Corban Rivera, PhD
APL
CAMP APL Lead

Gregory Kirk, MD
Professor of Epidemiology, SPH
CAMP Research Lead

Aalok Shah
IT Product Development Manager, TIC
CAMP Manager

Emily Marx
Communications, TIC
CAMP Manager

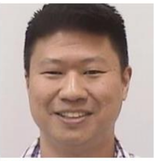
Scott Zeger, PhD
Director, inHealth
CAMP Advisor

Antony Rosen, MD
Vice Dean for Research, SOM
CAMP Advisor

Mary Cooke
Director, inHealth
CAMP Advisor

Chris Chute, MD
Chief Health Research Info Officer, JHM
CAMP Advisor

Alan Coltri
Director, Data Arch + Integration, IT@JH CIO


Brant Chee, PhD
APL

Dwight Raum,
Chief Technology Officer, Johns Hopkins

Ken Harkness
Project Lead, IT@JH Clinical Systems Dev

Patrick Ostendarp
Product Development Lead, TIC

Alex Baras, MD
Assistant Professor, Pathology Informatics

Caitlyn Bishop
APL

Jerry Prince, PhD
Professor, Electrical + Computer Engineering

Lee Watkins
IT Operating Unit Director, CIDR

Rai Winslow, PhD
Director, Biomedical Engineering

Benjamin Smith
Application Coordinator, IT@JH Epic

Chris Doyle
Product Development Lead, TIC

John Scott
Senior Software Engineer, TIC

Manar Alhamdy
Senior Software Engineer, TIC

Stephen Granite
Senior IT Manager, ICM

Bonnie Woods
Senior IT Manager, ICTR

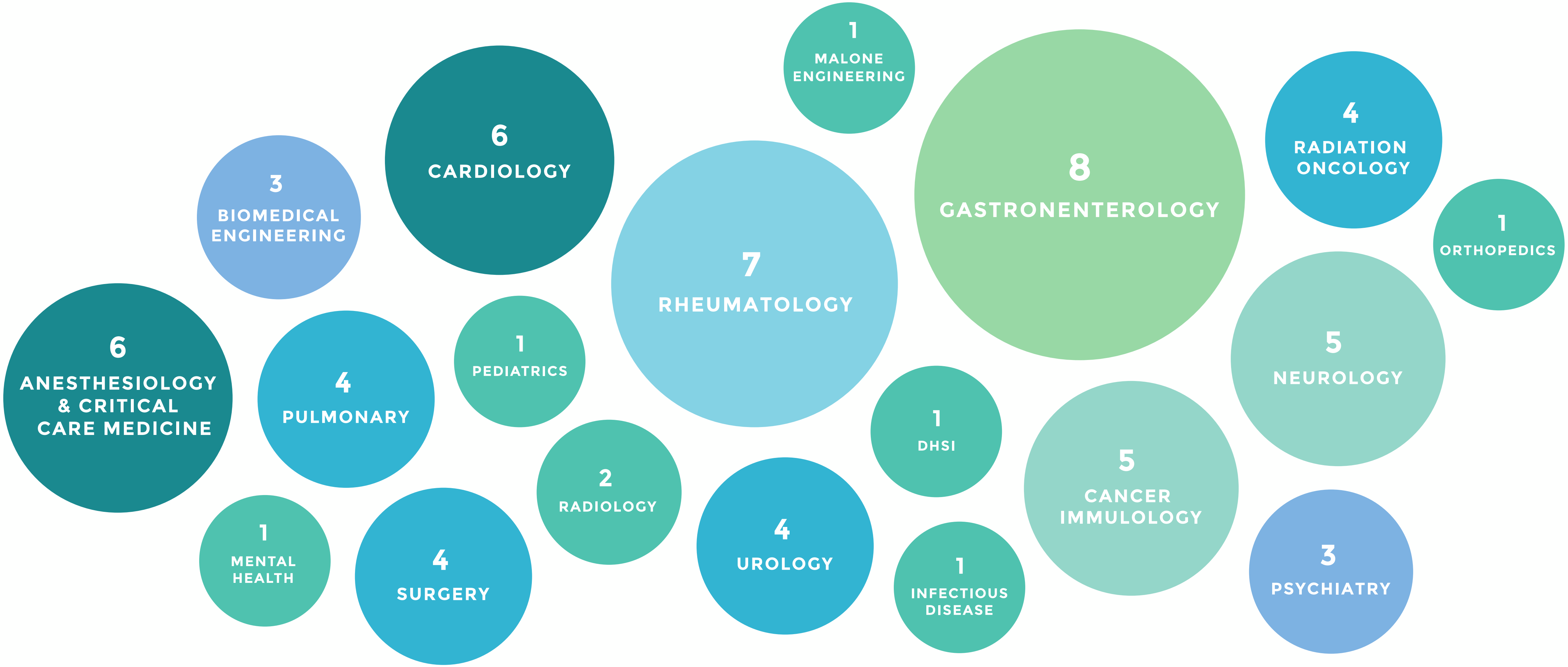
David Li
IT Director, IT@JH Epic

Jordan Matelsky
APL

Masoud Rouhizadeh, MD
Software Engineer, ICTR

Steven Handy
Senior Software Engineer, TIC

Department Distribution



6
ANESTHESIOLOGY & CRITICAL CARE MEDICINE

3
BIOMEDICAL ENGINEERING

4
PULMONARY

1
MENTAL HEALTH

4
SURGERY

6
CARDIOLOGY

1
PEDIATRICS

2
RADIOLOGY

7
RHEUMATOLOGY

4
UROLOGY

1
MALONE ENGINEERING

1
DHSI

1
INFECTIOUS DISEASE

8
GASTRONENTEROLOGY

5
CANCER IMMULOLOGY

4
RADIATION ONCOLOGY

1
ORTHOPEDICS

5
NEUROLOGY

3
PSYCHIATRY

CAMP Makeup (N=70)

12
Full Professors

13
Associate Professors

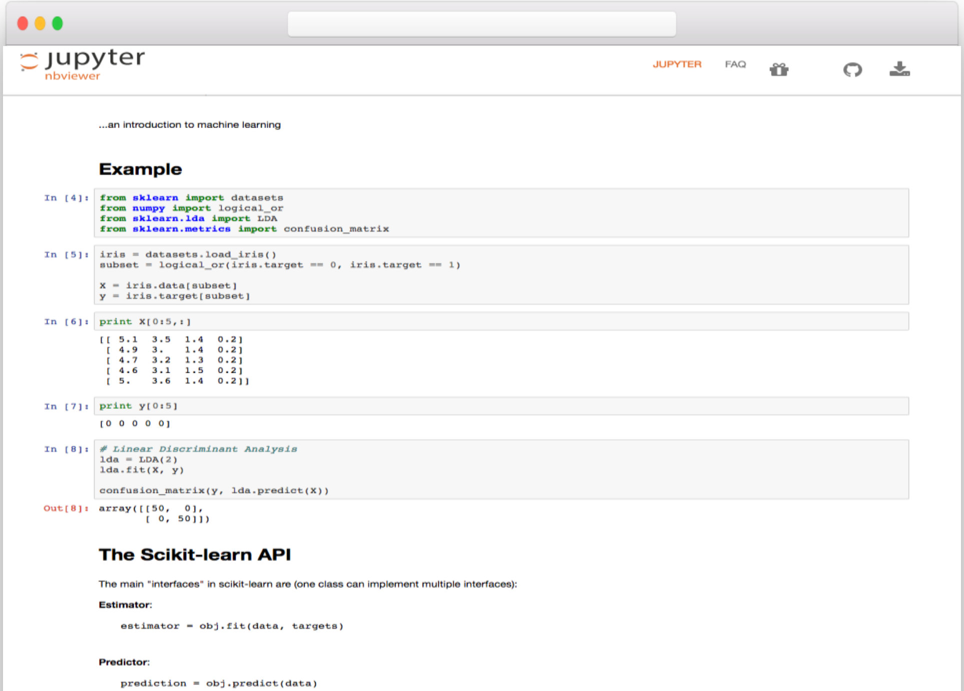
14
Assistant Professors

1
Instructors

10
Fellows/Students

20
Research Staff

Jupyter: Open Data Science at Scale



...an introduction to machine learning

Example

```
In [4]: from sklearn import datasets
        from numpy import logical_or
        from sklearn.lda import LDA
        from sklearn.metrics import confusion_matrix

In [5]: iris = datasets.load_iris()
        subset = logical_or(iris.target == 0, iris.target == 1)

        X = iris.data[subset]
        y = iris.target[subset]

In [6]: print X[0:5,:]

        [[ 5.1  3.5  1.4  0.2]
         [ 4.9  3.   1.4  0.2]
         [ 4.7  3.2  1.3  0.2]
         [ 4.6  3.1  1.5  0.2]
         [ 5.   3.6  1.4  0.2]]

In [7]: print y[0:5]

        [0 0 0 0 0]

In [8]: # Linear Discriminant Analysis
        lda = LDA(2)
        lda.fit(X, y)

        confusion_matrix(y, lda.predict(X))

Out[8]: array([[50,  0],
               [ 0, 50]])
```

The Scikit-learn API

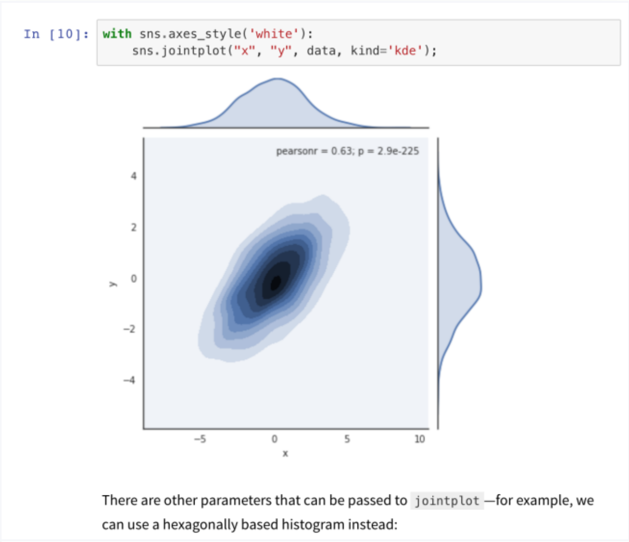
The main "interfaces" in scikit-learn are (one class can implement multiple interfaces):

Estimator:

```
    estimator = obj.fit(data, targets)
```

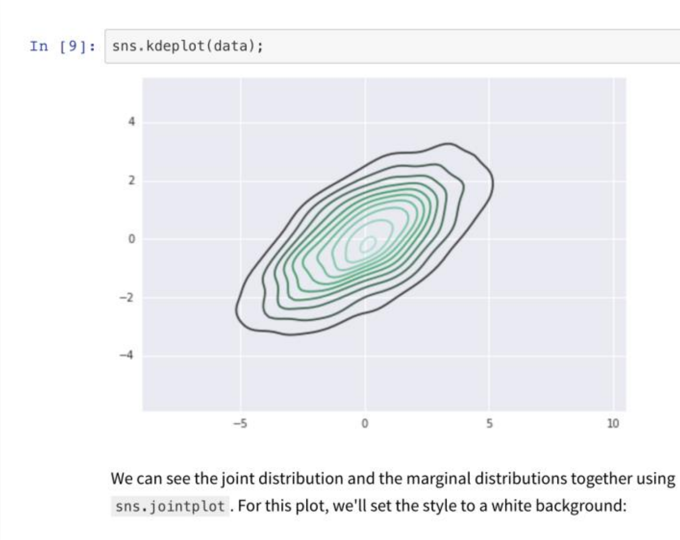
Predictor:

```
    prediction = obj.predict(data)
```



```
In [10]: with sns.axes_style('white'):
             sns.jointplot("x", "y", data, kind='kde');
```

There are other parameters that can be passed to `jointplot`—for example, we can use a hexagonally based histogram instead:

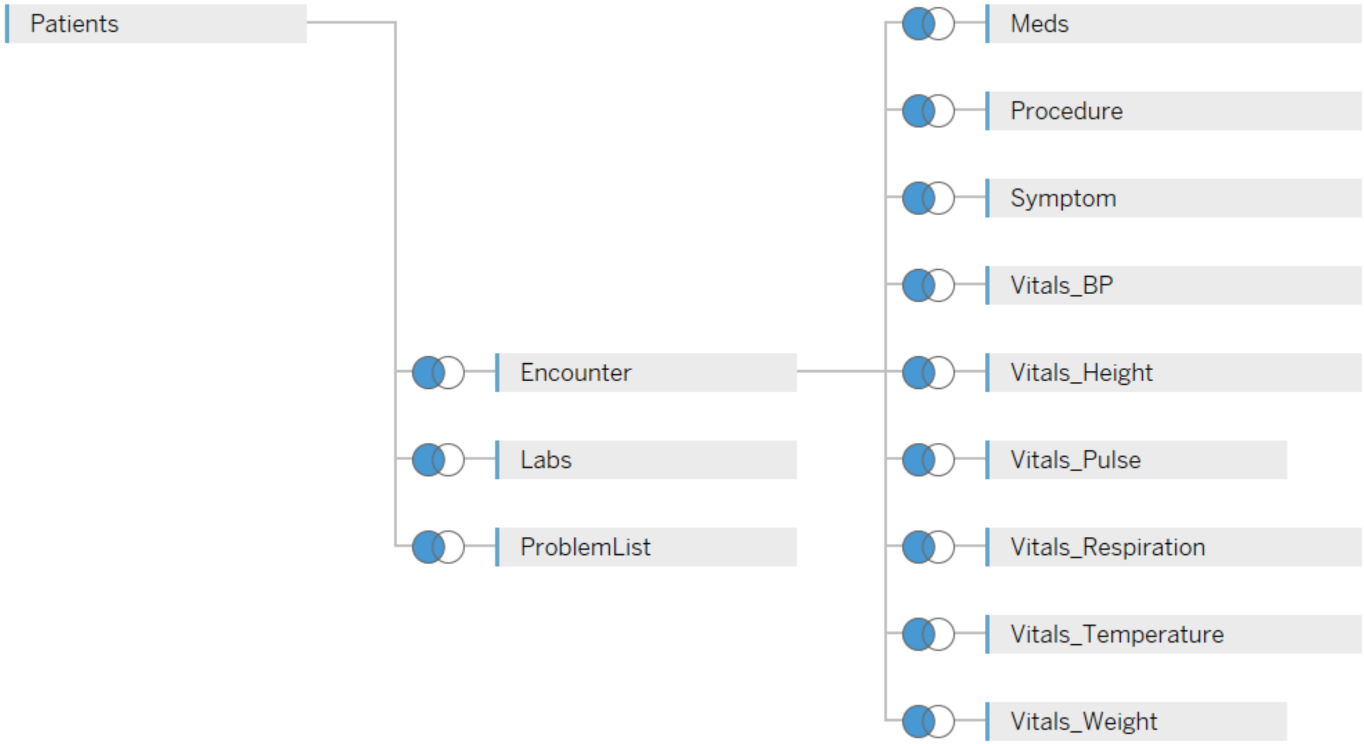


```
In [9]: sns.kdeplot(data);
```

We can see the joint distribution and the marginal distributions together using `sns.jointplot`. For this plot, we'll set the style to a white background:

Rich web client → Text & Math → Code → Results → Share, reproduce

De-identified Epic Dataset



Patients

Encounter

Labs

ProblemList

Meds

Procedure

Symptom

Vitals_BP

Vitals_Height

Vitals_Pulse

Vitals_Respiration

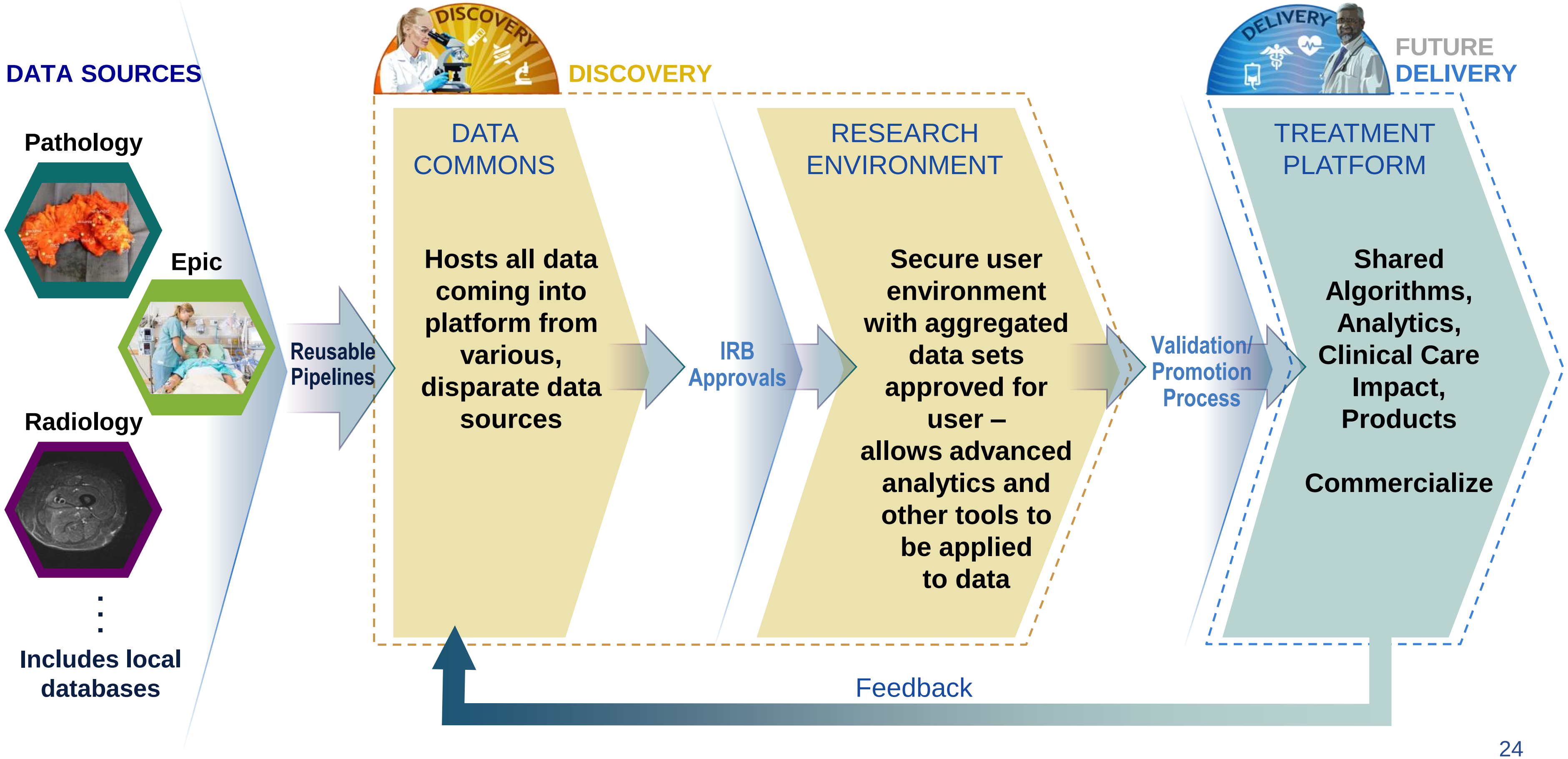
Vitals_Temperature

Vitals_Weight

112 Million Data Elements
No note fields

Table Name	# of Records	# of Columns	Size
Encounter	690,183	4	36.57 MB
Labs	3,486,911	12	403.41 MB
Meds	5,926,733	9	1.86 GB
Patients	60,676	5	7.18 MB
ProblemList	115,162	4	16.99 MB
Procedure	6,449	6	0.727 MB
Symptom	28,056	5	1.711 MB
Vitals_BP	390,181	7	30.58 MB
Vitals_Height	279,288	7	18.96 MB
Vitals_Pulse	388,450	7	65.92 MB
Vitals_Respiration	251,166	7	18.18 MB
Vitals_Temperature	314,571	7	21.87 MB
Vitals_Weight	352,553	7	24 MB
TOTAL	12,290,379	87	2.55 GB

Platform Architecture



DATA SOURCES

Pathology

Epic

Radiology

⋮

Includes local databases

Reusable Pipelines

DISCOVERY

DATA COMMONS

Hosts all data coming into platform from various, disparate data sources

IRB Approvals

RESEARCH ENVIRONMENT

Secure user environment with aggregated data sets approved for user – allows advanced analytics and other tools to be applied to data

Validation/ Promotion Process

FUTURE DELIVERY

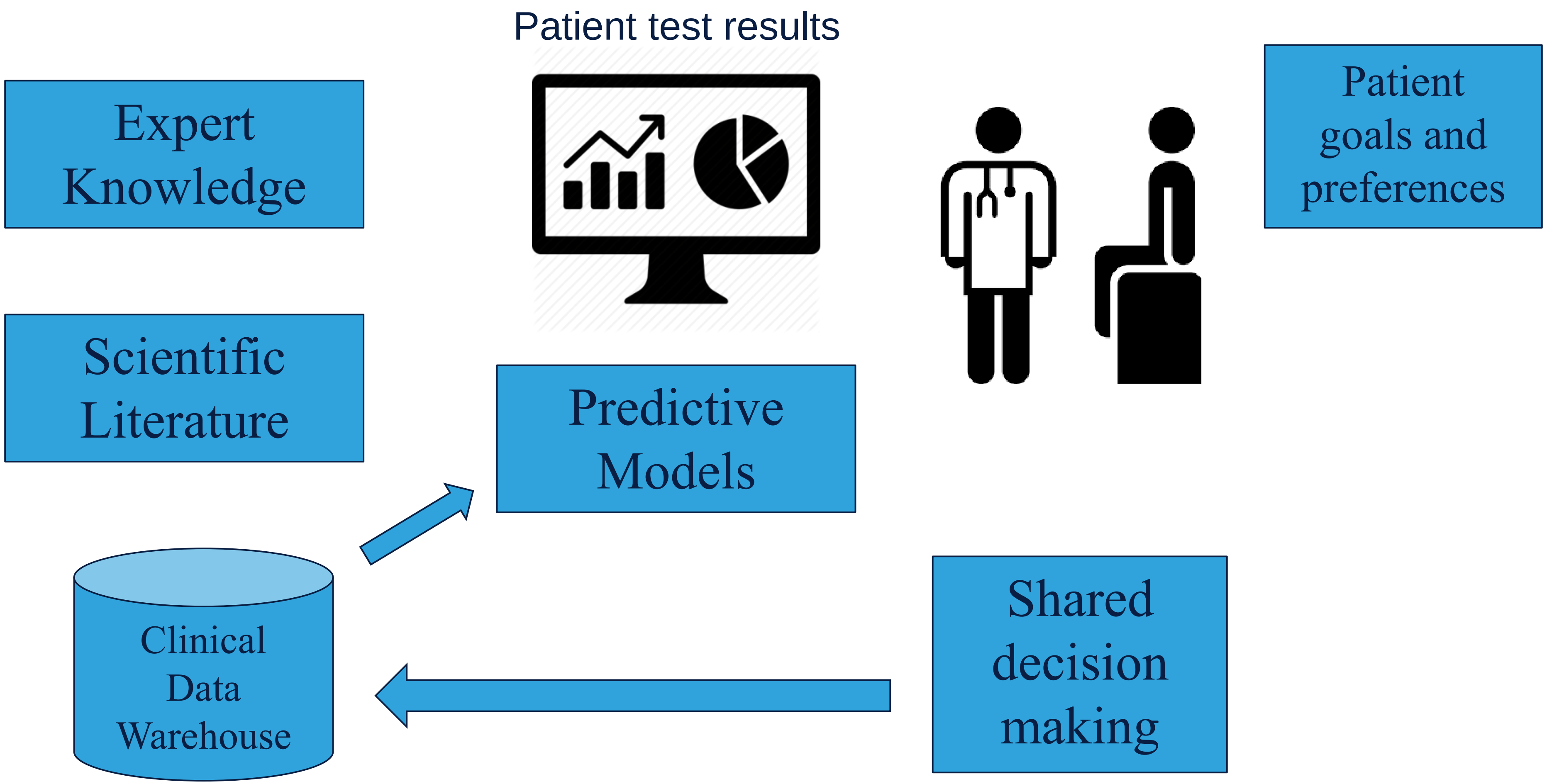
TREATMENT PLATFORM

Shared Algorithms, Analytics, Clinical Care Impact, Products

Commercialize

Feedback

Decision Support



Patient test results

Expert Knowledge

Scientific Literature

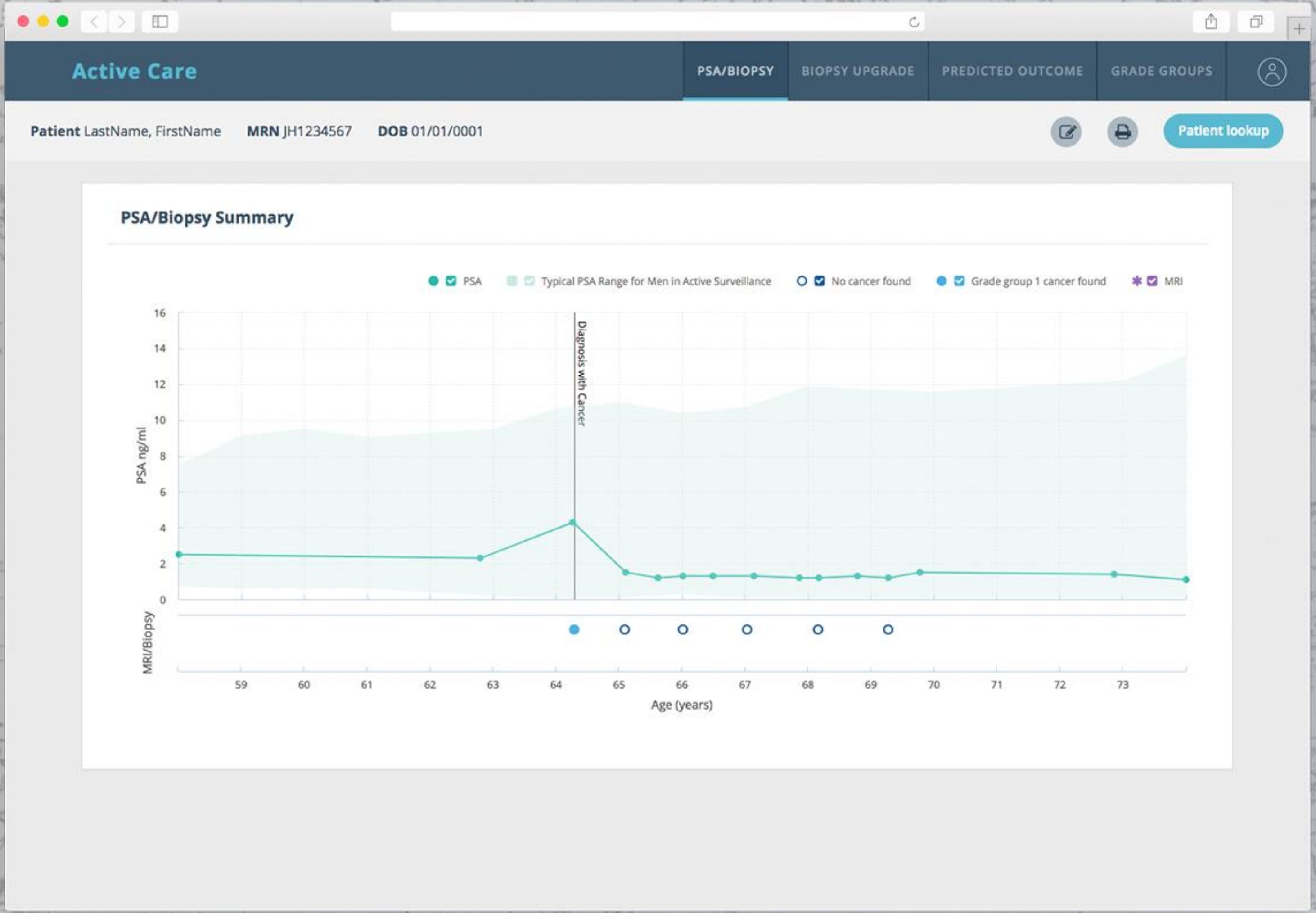
Predictive Models

Patient goals and preferences

Clinical Data Warehouse

Shared decision making

JOHNS HOPKINS inhealth



Patient LastName, FirstName **MRN** JH1234567 **DOB** 01/01/0001

PSA/Biopsy Summary

JOHNS HOPKINS MEDICINE