

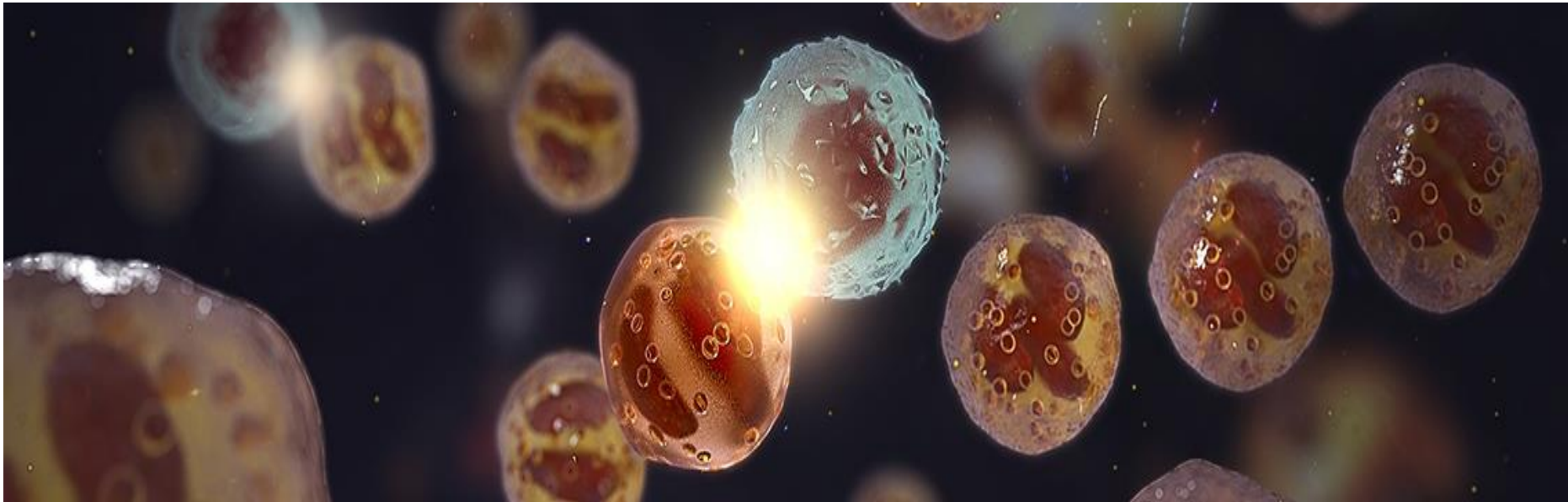
Deriving Knowledge from Real-World Evidence Using Large-Scale Analytics

Nirmal Keshava, Ph.D.

Senior Principal Informatics Scientist
Research & Development – Information (RDI)
AstraZeneca, PLC
Waltham, MA
Nirmal.Keshava@astrazeneca.com

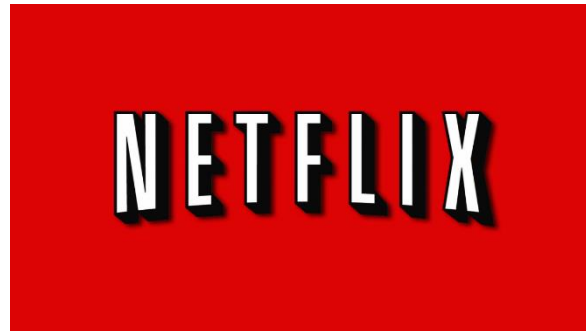
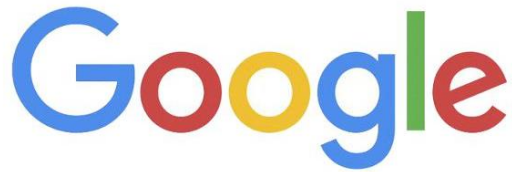
Bio-IT World

7 April 2016



The Rest of the World is Changing

Or, Has Already Changed



Introduction

Data Science

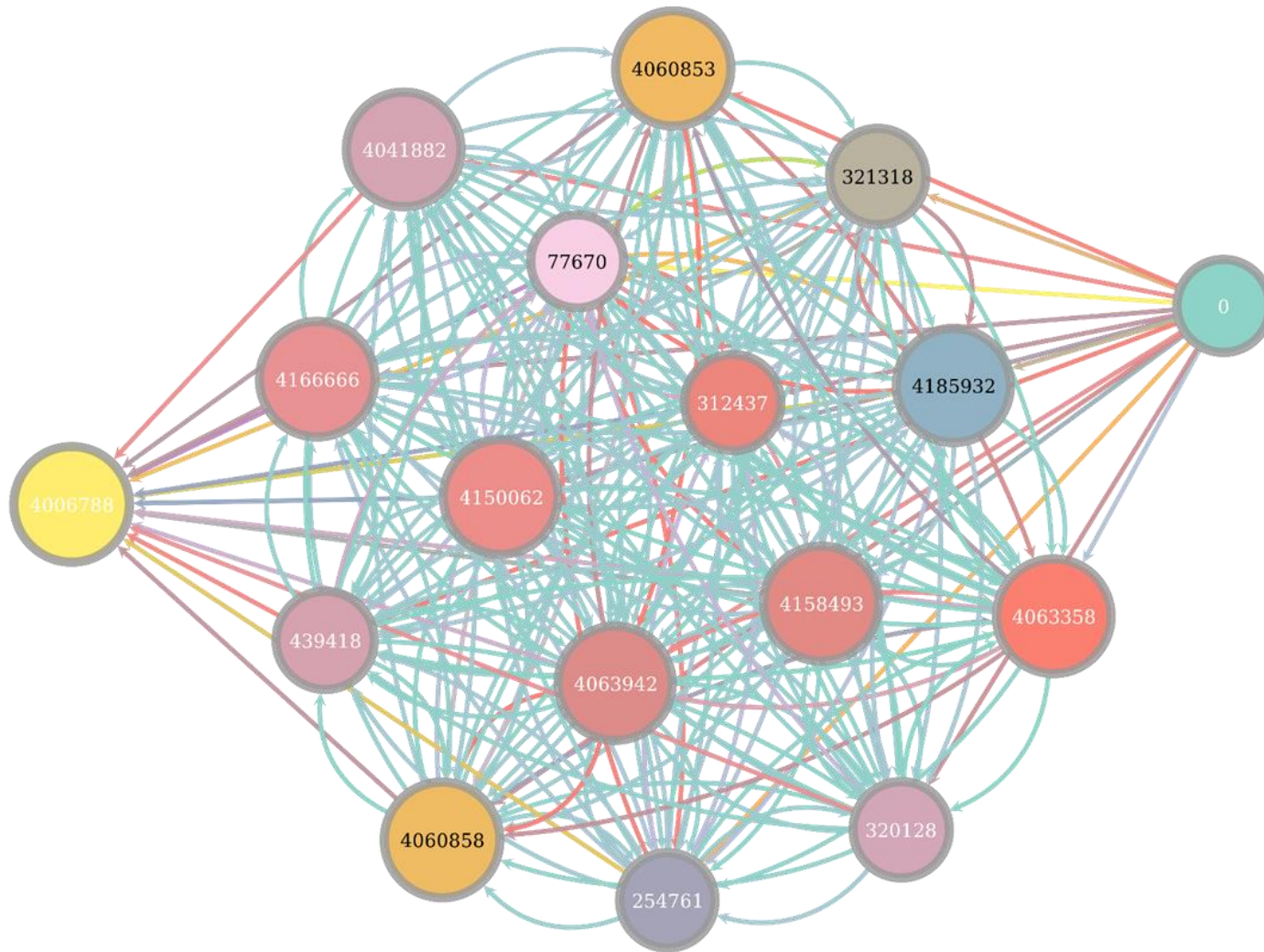
Large-Scale Analytics

Summary



1. Data Science

Creating Structure Out of Complexity



Data Science

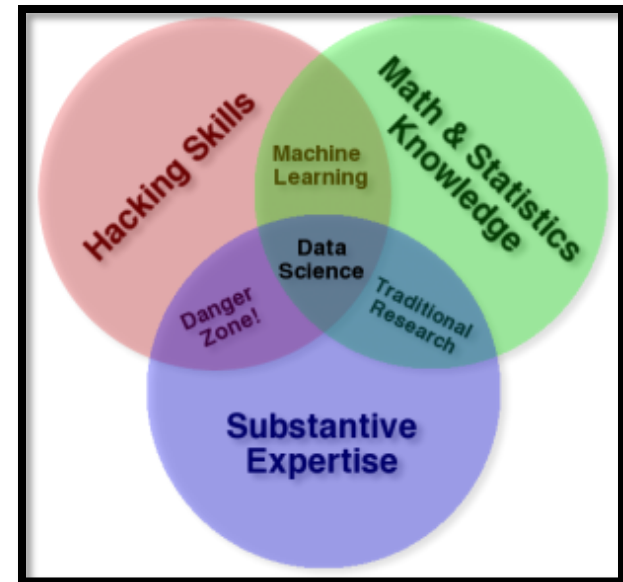
Definitions & Applications

This emerging field was nicely characterized as “the story of the coupling of the mature discipline of statistics with a very young one—computer science.”*

In fact, engineering has been an existing intersection of these fields for decades.

It's critical to understand the different classes of problems being addressed by data science:

- Business Intelligence
- Retail
- Targeted Advertising
- Financial Industry



* <http://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#75c9e3f69fd2>



Challenges to Realizing Value from Data

- 1. The Data Science “Risk Curve”**
- 2. Fear of Failure (Single Elimination)**
- 3. Mgmt.: What Language are You Speaking?**
- 4. Data Scientist: I Don’t Speak Their Language**
- 5. In-House vs. Externalization**
- 6. Blind Shopping (& Shoppers)**
- 7. It Was Harder Than I Thought**
- 8. Selling Upwards / Looking for an Advocate**
- 9. Immediate Obsolescence**
- 10. Let’s Integrate It First**



The Solution Begins with the Question

Does Anyone Really Know What It Is?

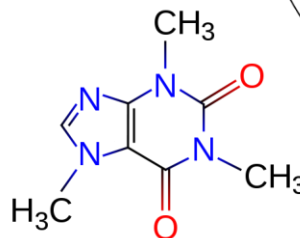
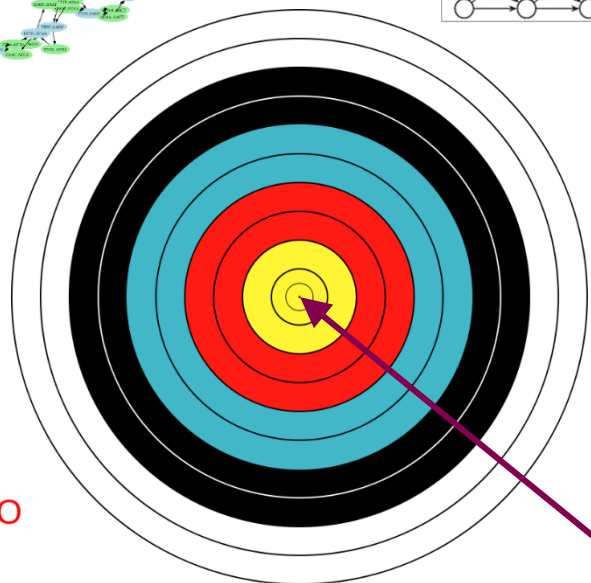
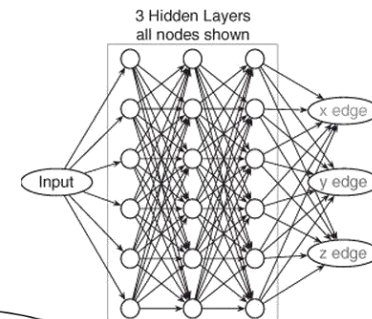
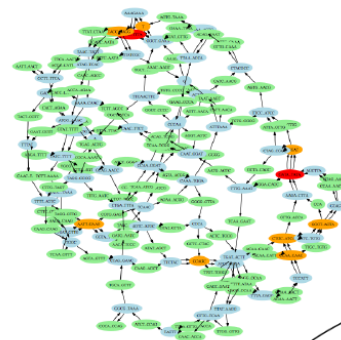
What is the problem aiming to achieve?

Minimizing risk and maximizing success involves scoping the real problem from the outset

- In (excruciating) technical detail
- Unknowns are OK as long as they are accounted for

Entering at highest technical level defines downstream requirements

- Aiming elsewhere only invites risk and increases the chance of failure



And the Problem Begins with the People *...All Around the Same Table*



Who Should Get a Seat at the Table?

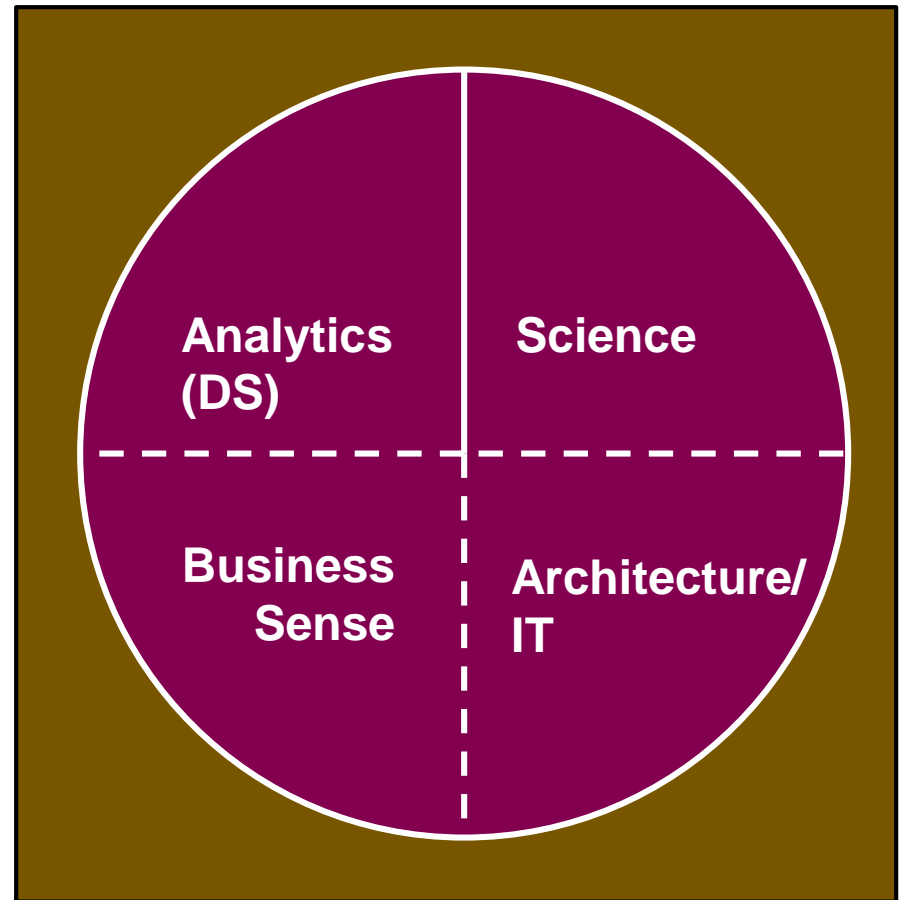
4 Principal Roles

Data science in pharma is an inherently multi-disciplinary process that integrates analytics, science, architecture and business sense

Individual players can usually see multiple parts of the problem, but not all of it

The data scientist is critical for “engineering” solutions and exposing value in a very hands-on way

- Must be extremely facile and flexible in analytics
- Arbitrating a common language across disciplines



Organizing a Business Around Data

Is Pharma Organized for This?

Pharma organizations are grappling with different models for how to bring data science into their pipelines

- Centralized informatics organizations
- De-centralized (embedded) informaticists
- External vendors
- Internal analytics development
- Curation
- Hosting on-site vs. cloud (regulatory issues)

Different organizations (i.e., therapeutic areas, pipeline stages) have their own imperatives, and they frequently conflict in their goals and the means by which they achieve them



2. Examples of Large-Scale Analytics

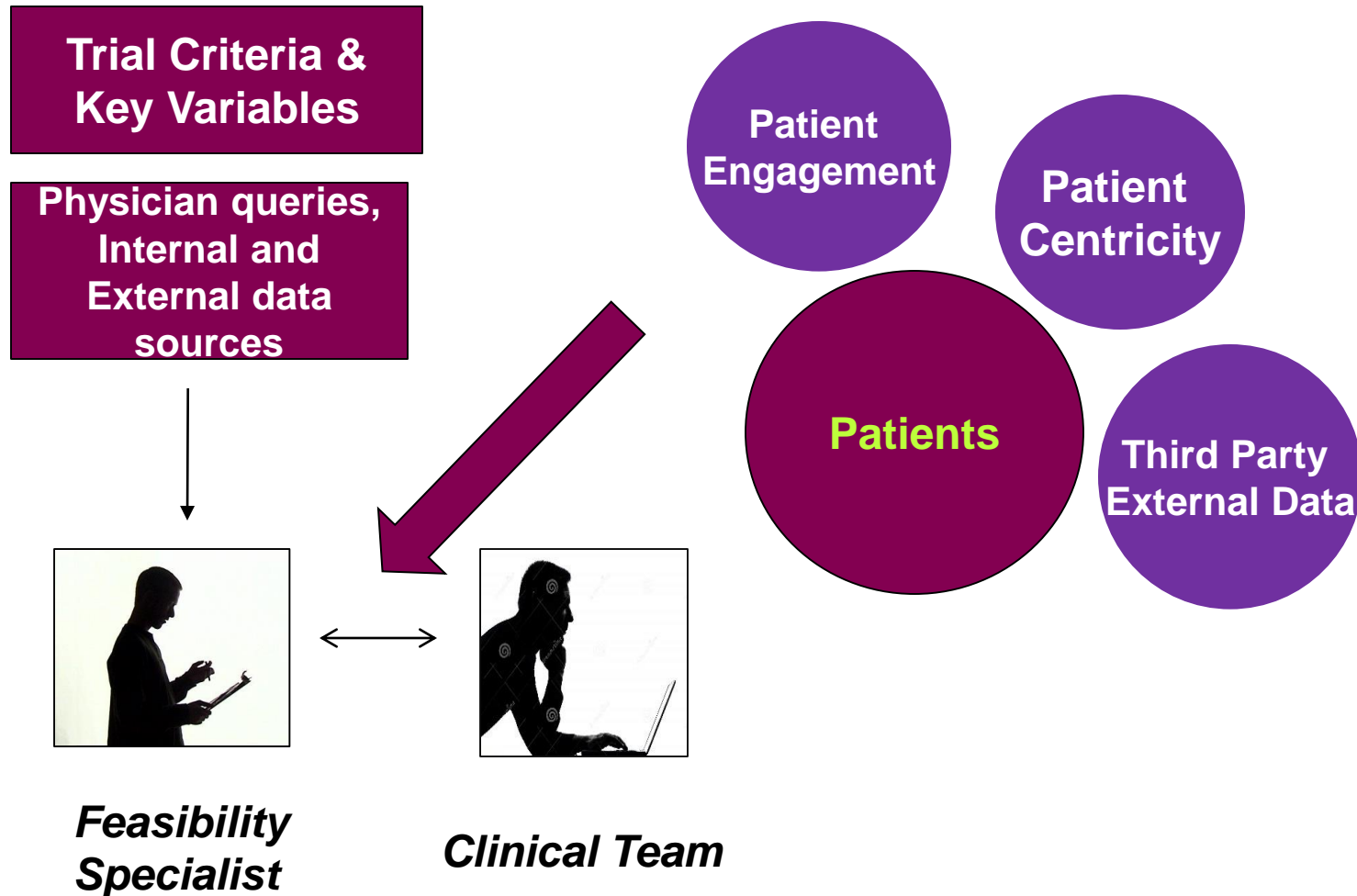
Improving Situational Awareness from Trial Feasibility Informatics

Building Structure and Knowledge from Patient-Centric Data



Example 1: Estimating Trial Feasibility

Highly Manual Process for Targeting Patients



Using Informatics to Expedite Clinical Trials

Can Digital Assets Decrease Uncertainty?

PoC Goal: Use in-house digital assets to improve understanding of trial populations

- Specifically RA and COPD patient populations

Resources

- Payer claims databases
- Patient-centric databases

Long-term Goal: Establish a new feasibility workflow to include third-party data analyses

- Identify cohorts in closed or ongoing clinical trials to get a sense of how this patient population has been identified previously
- Identify cohorts in electronic health records and/or claims data
- Identify cohorts in patient-centric social media forums
- Build a user-interface to facilitate this process and incorporate into the current feasibility process flow



Building Insight on Disease Populations from RWE Databases

Payer claims data and electronic medical records (EMRs) are real-world evidence (RWE) databases that contain records of medical encounters for specific patient populations

- Claims: transactional records intended to document reimbursement
- EMR: contain clinical information and data, possibly across specialties

RWE databases provide insight on patient populations based on disease state, age, gender, etc.

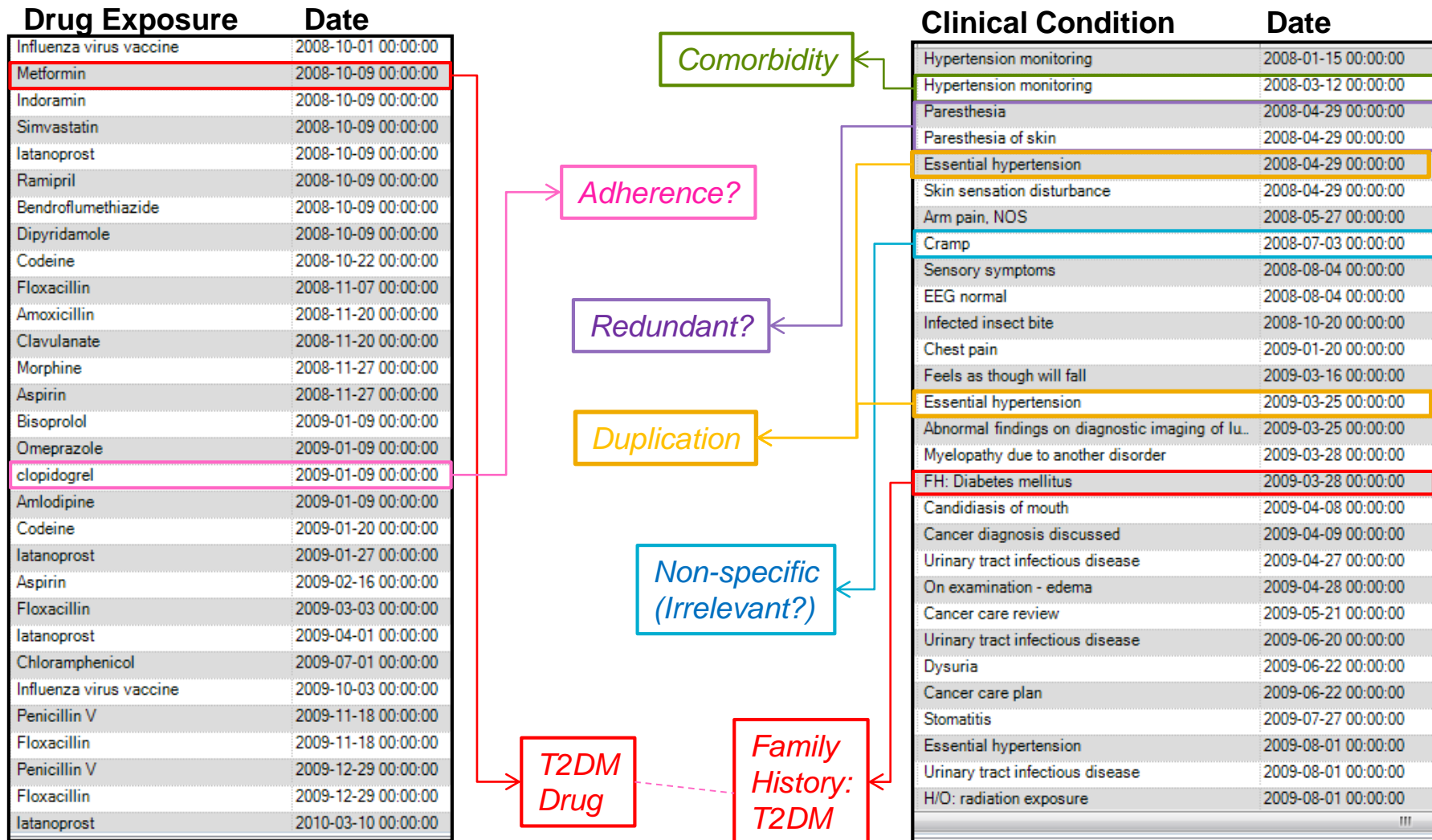
RWE DBs can assess the feasibility of prospective clinical trials by looking at characteristics of a comparable population in the real world

RWE DBs, however, can be extremely noisy, and care must be taken in interpreting the results



Finding the Signal Amid Noise in RWE

Duplication, Irrelevance, Redundancy, & Lack of Standardization



Curation, Curation, Curation

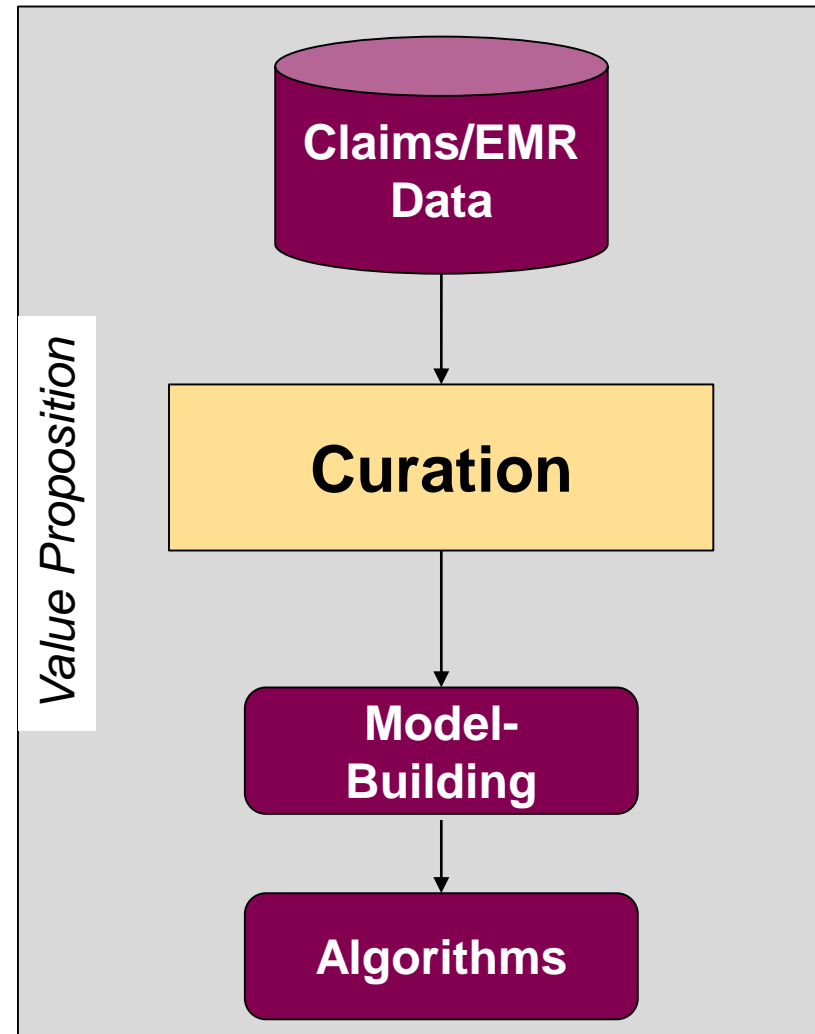
Minimizing Noise While Preserving Signal

Data curation is a critical component for exploiting data that was never intended for computation

Curation is noise reduction and in claims data involves addressing the following:

- Redundant/repeated information
- Censoring
- Irregular sampling
- Irrelevant/non-specific information

The data scientist imports domain knowledge into data to improve its interpretation



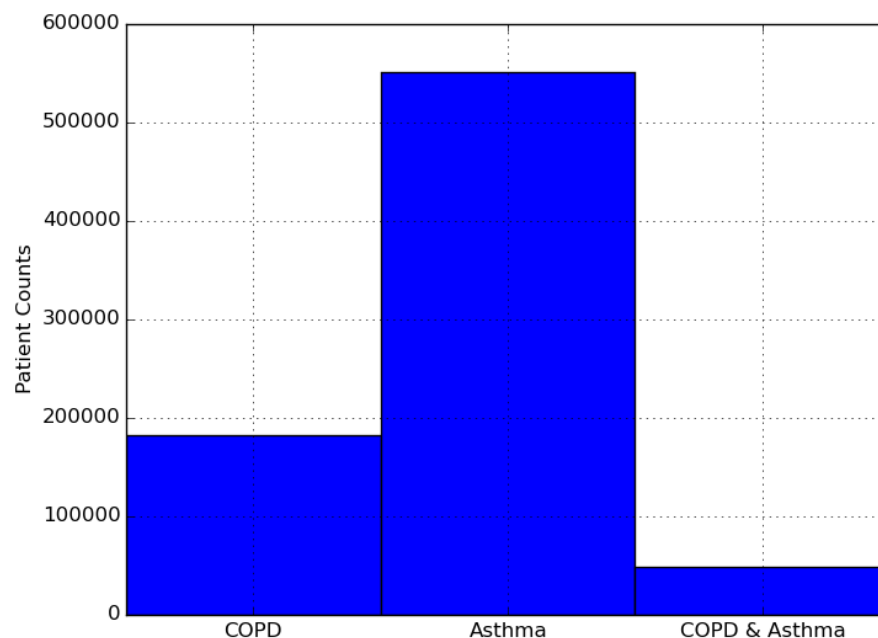
COPD Populations in MSLR RWE DB

Analyzing Primary Care RWE Data

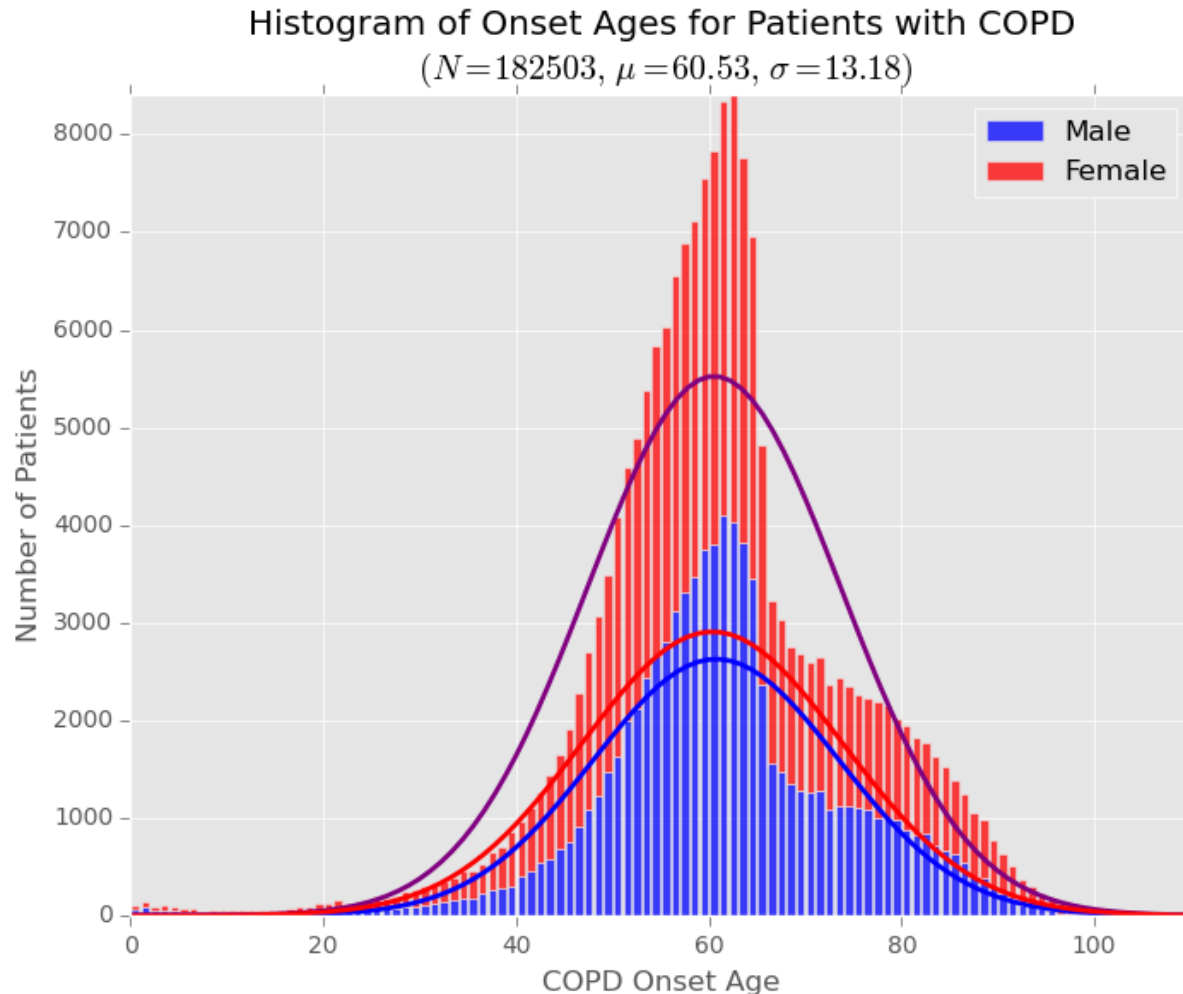
MSLR is a US commercial claims RWE DB with records for 6 M+ patients

Disease-specific condition codes identify patients having COPD, asthma, or both

Further detail on patient populations can be derived from records on drug exposures, procedures, visits, mortality, etc.

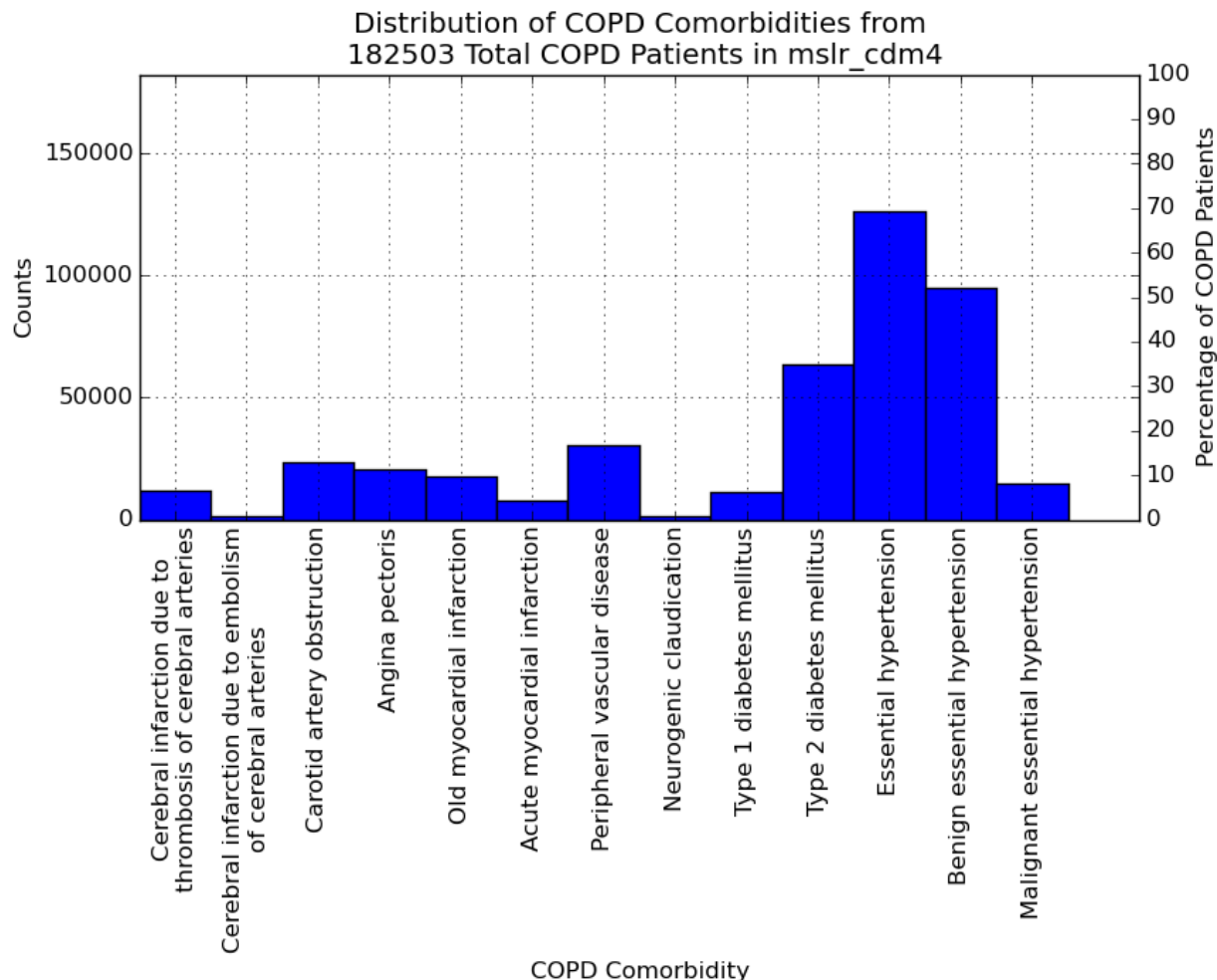


COPD Population Age of Onset By Gender



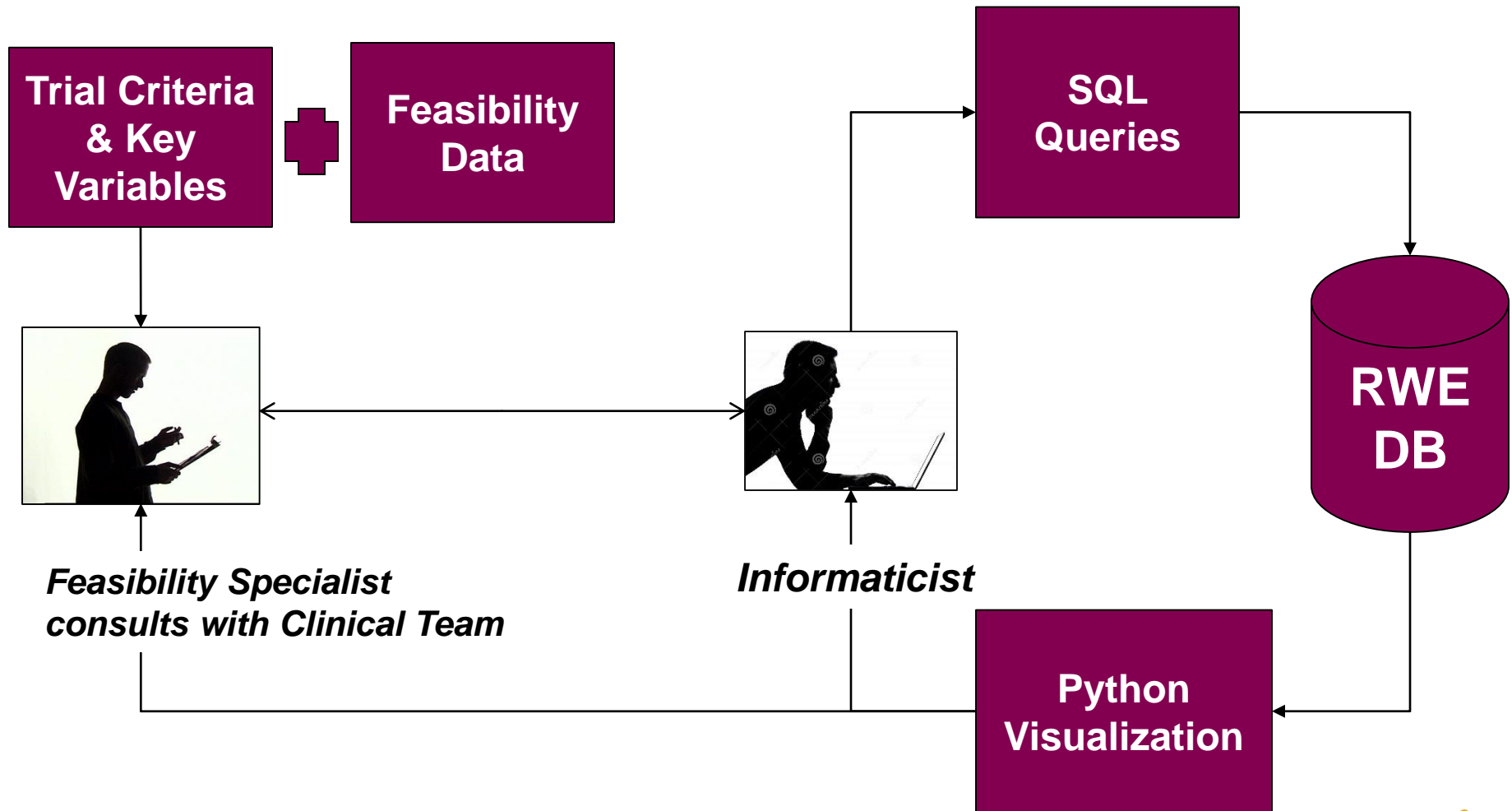
COPD Populations with CVD

Relative Rates of CVD Comorbidities



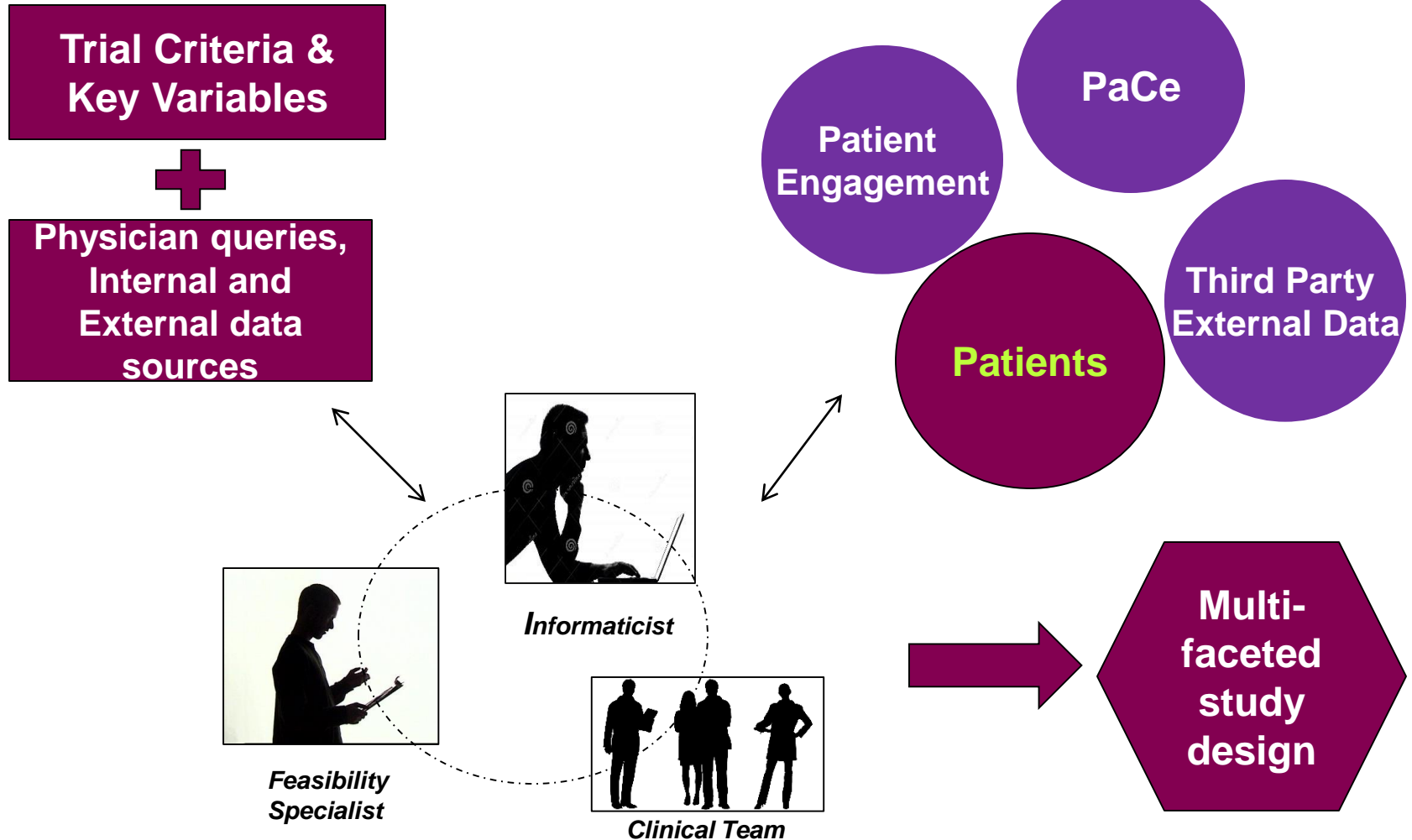
New Workflow for Trial Feasibility

Bringing Informatics Closer to the Front Line



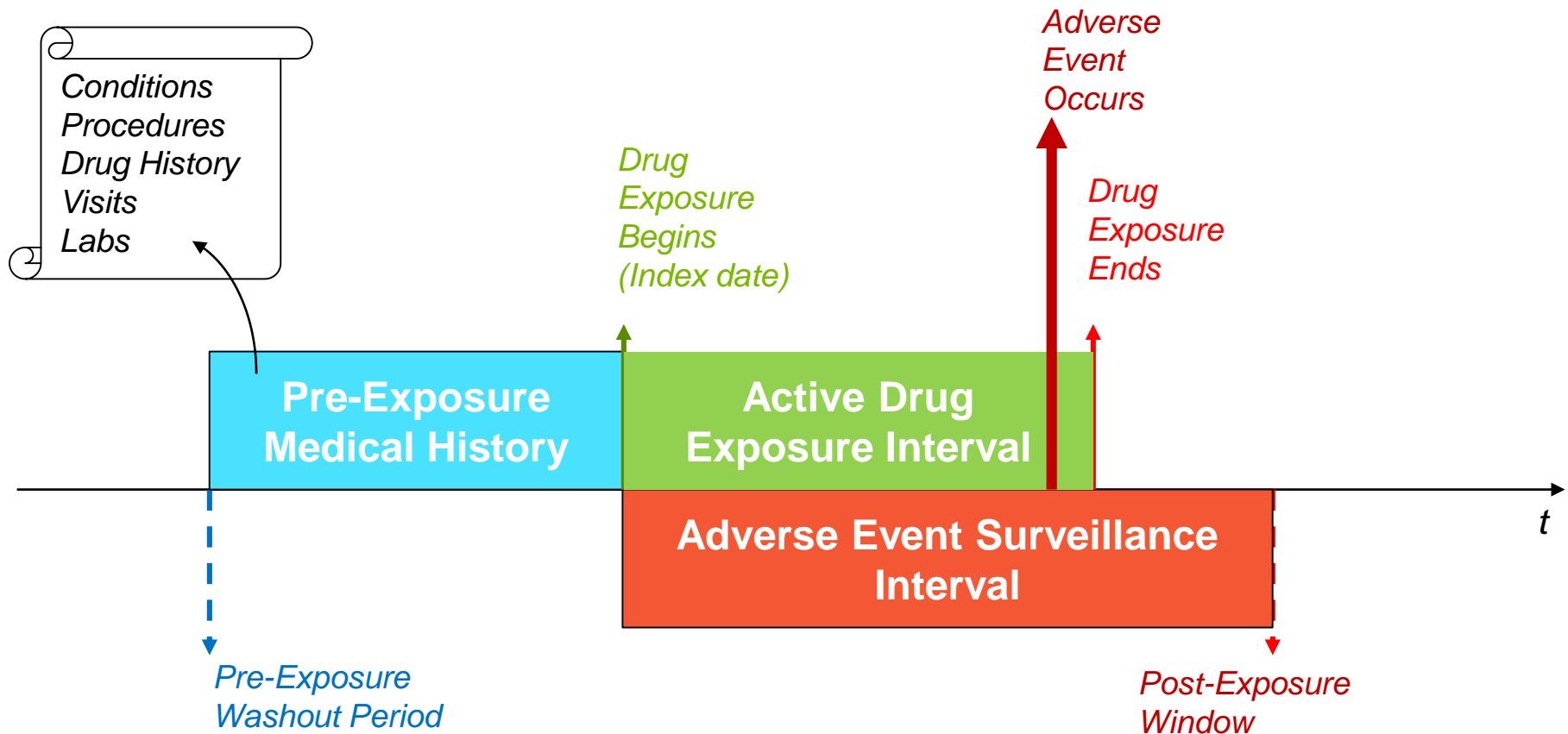
Future Feasibility Workflow

And Evolving Even Further



Example 2: Looking for At-Risk Populations

Adverse Events & Pharmacovigilance



Characterizing Patient Populations

Are There Enrichments in Occurrence of ADEs

Drug: Warfarin

- Primary use as a blood thinner
- Known to cause different adverse events

Many categories of covariates are available in real-world evidence (RWE)

- Demographic
- Conditions, drug history, and procedures
- Labs/clinical

Key Questions:

1. Are there sub-populations where ADEs occur in higher / lower rates?
2. How can we use the diverse set of RWE measurements?



Integrating & Analyzing Diverse Data

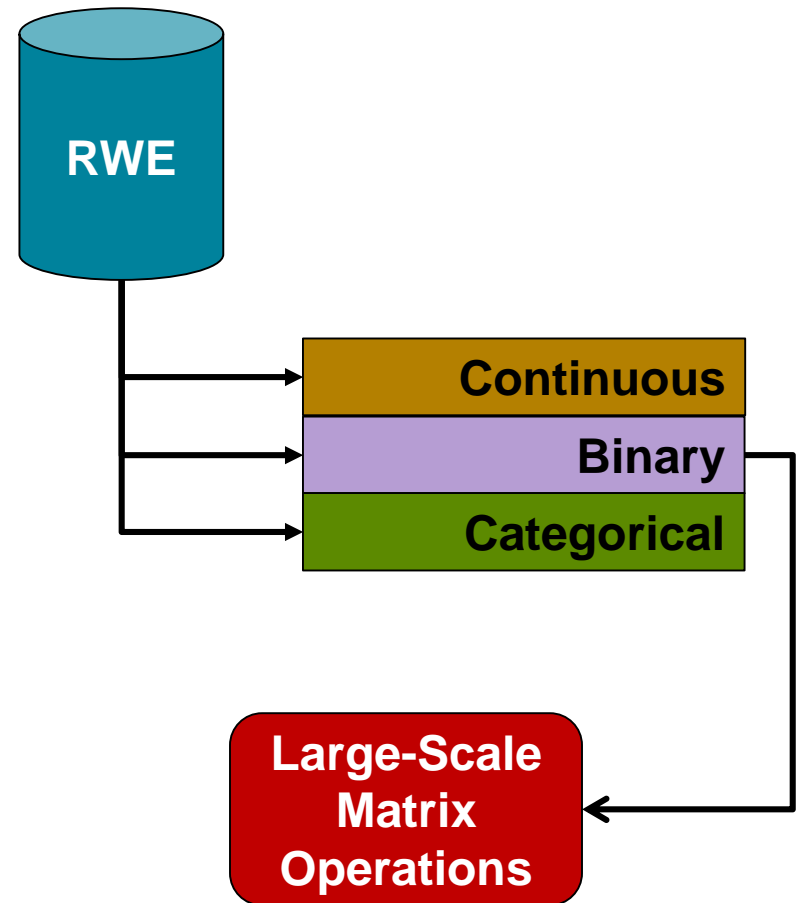
Leveraging Techniques from Other Industries

We want a method to integrate any type of data:

- Continuous-valued variables (e.g., age, physiological msmts.)
- Binary variables (e.g., gender, medical procedures/Dx)
- Categorical variables (e.g., mutation/CNVs, subjective assessments, measurement ranges)

We want to take advantage of large-scale *matrix-based* approaches developed in other fields, that create **structure**:

- Netflix challenge
- Linked In
- Amazon



Summary

There are enormous opportunities to extract value from data in pharma by capitalizing upon the full potential of the data scientist

Building multi-disciplinary teams (and organizations) around data is a critical component for extracting value

The opportunities to drive science and development of drugs is just beginning



Acknowledgments

RDI

- Tom Plasterer
- Kathy Reinhold
- Sara Dempster
- Jonathan Herz
- Vishant Prabhakharan

Biometrics & Information

- Lisa Bennett

Global Medical Affairs

- Robert LoCasale
- Xiongkan Ke

Clinical Operations

- Michele Teufel
- Ed Mannello
- Mark Cooper



Characterizing Patient Populations

Are There Enrichments in Occurrence of AMIs

Drug: Nifedipine

- Primary indication is for hypertensive heart disease
- Listed as a positive control for acute myocardial infarction (AMI) in observational medicine literature (OMOP)

Many categories of covariates are available

- Demographic
- Conditions, drug history, and procedures
- Labs/clinical

How can we make use of disparate variables?

