
BUILT FOR THE SPEED OF BUSINESS

# The Emerging Role of Data Science in Pharma:
## How to Harness this Transformative Practice

Sarah Aerni, PhD
Senior Data Scientist

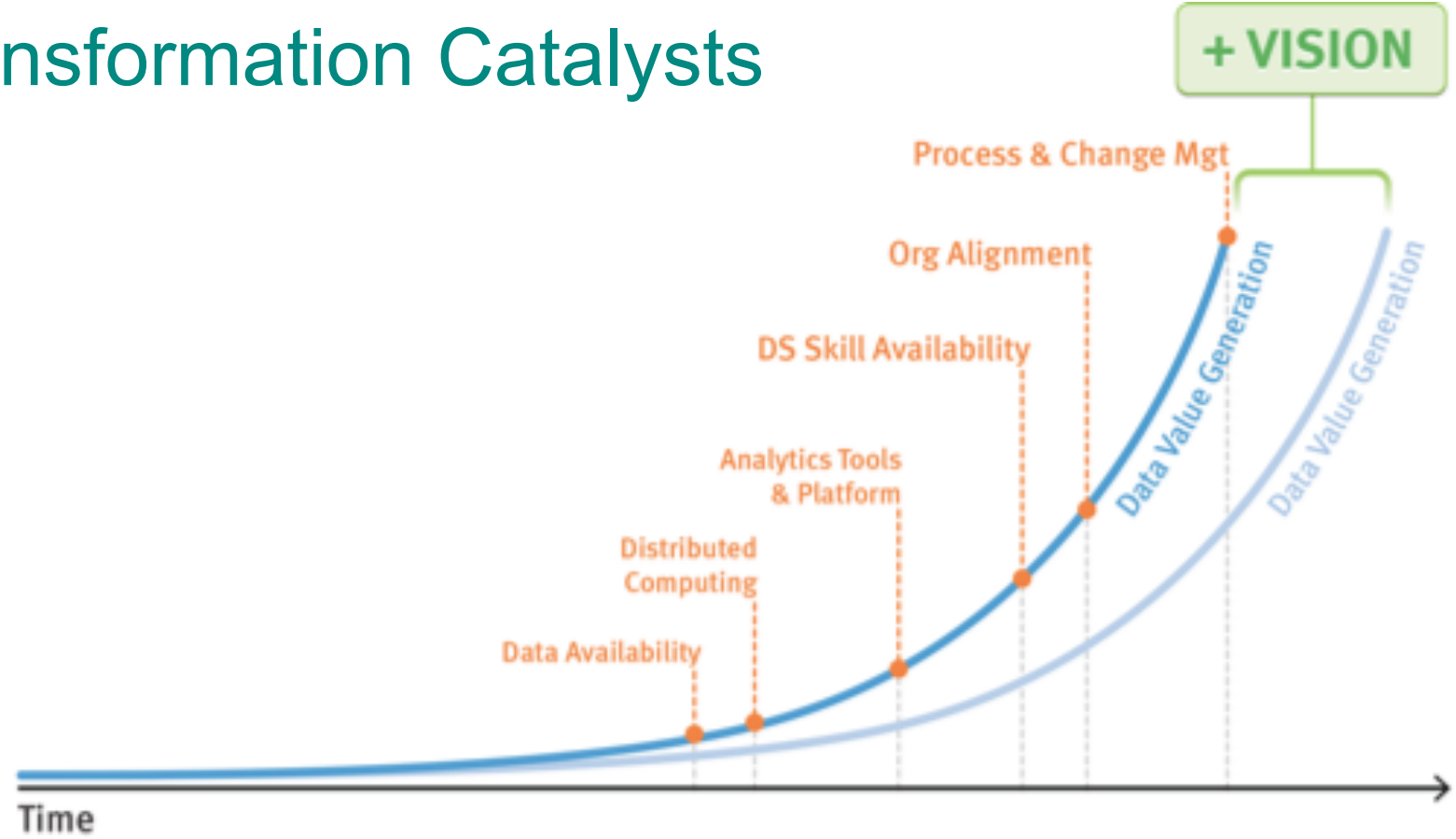
saerni@gopivotal.com
PRISME
May 15, 2014

Pivotal

# Agenda

- Driving Transformation through Data Science
- Pivotal Technology Enabling Innovation and Driving Insights
  - The right technology for the right job
  - Reinventing data processing through paradigm shifts
- Use Case Discussion
  - Predicting pharmaceutical potency

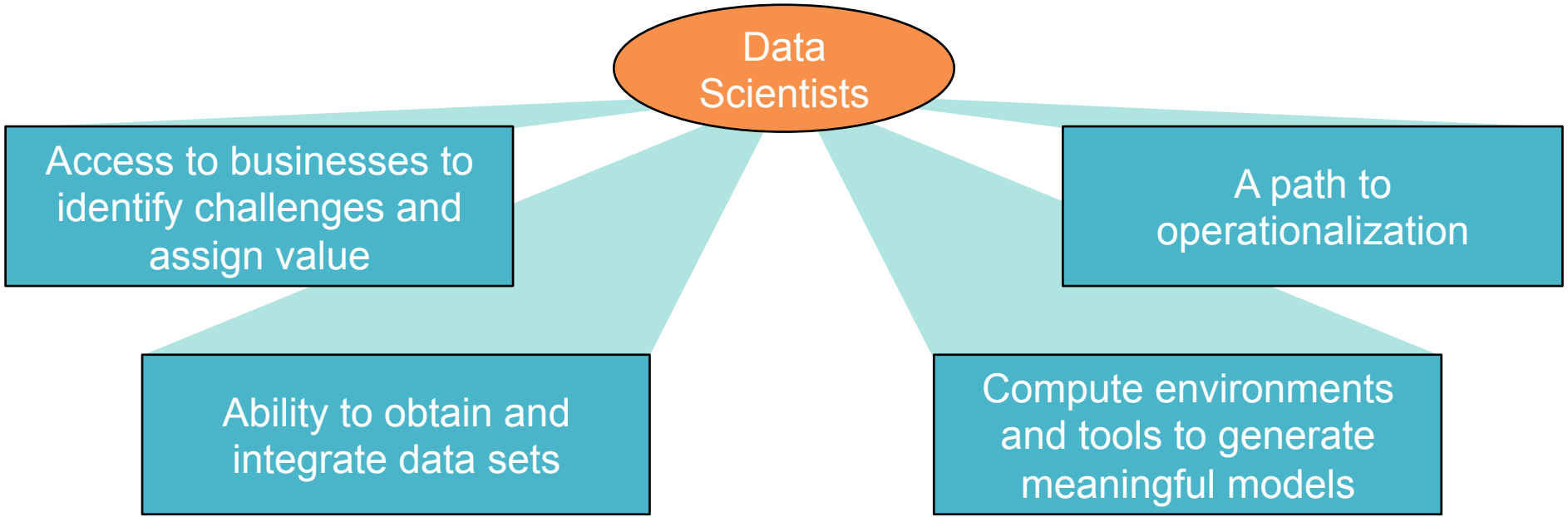
# Transformation Catalysts




Pivotal

# Enabling Disruption with Data Science

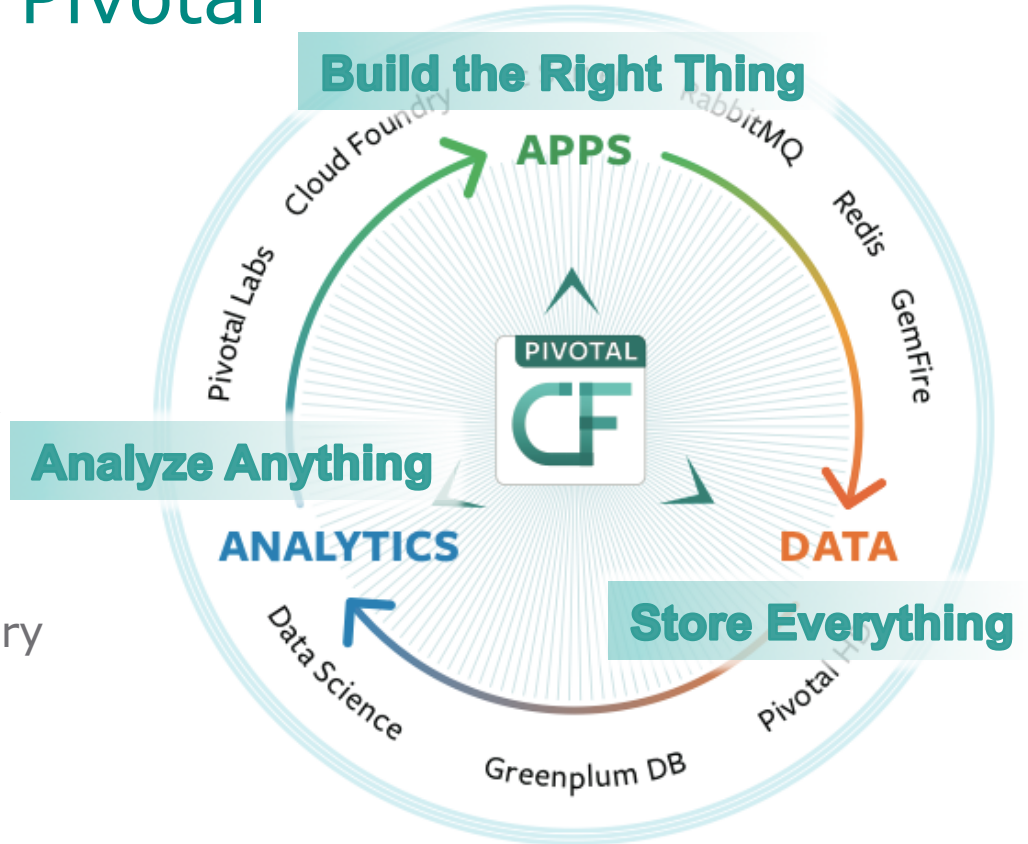
- To successfully build a data-driven enterprise, skilled individuals need to have access to data, tools and channels for operationalization




Pivotal

# Primary Motions for Pivotal

- *Agile:* data-driven apps and rapid time to value
- *Data Lake:* store everything, analyze anything
- *Enterprise PaaS:* revolutionary software development and speed; build the right thing



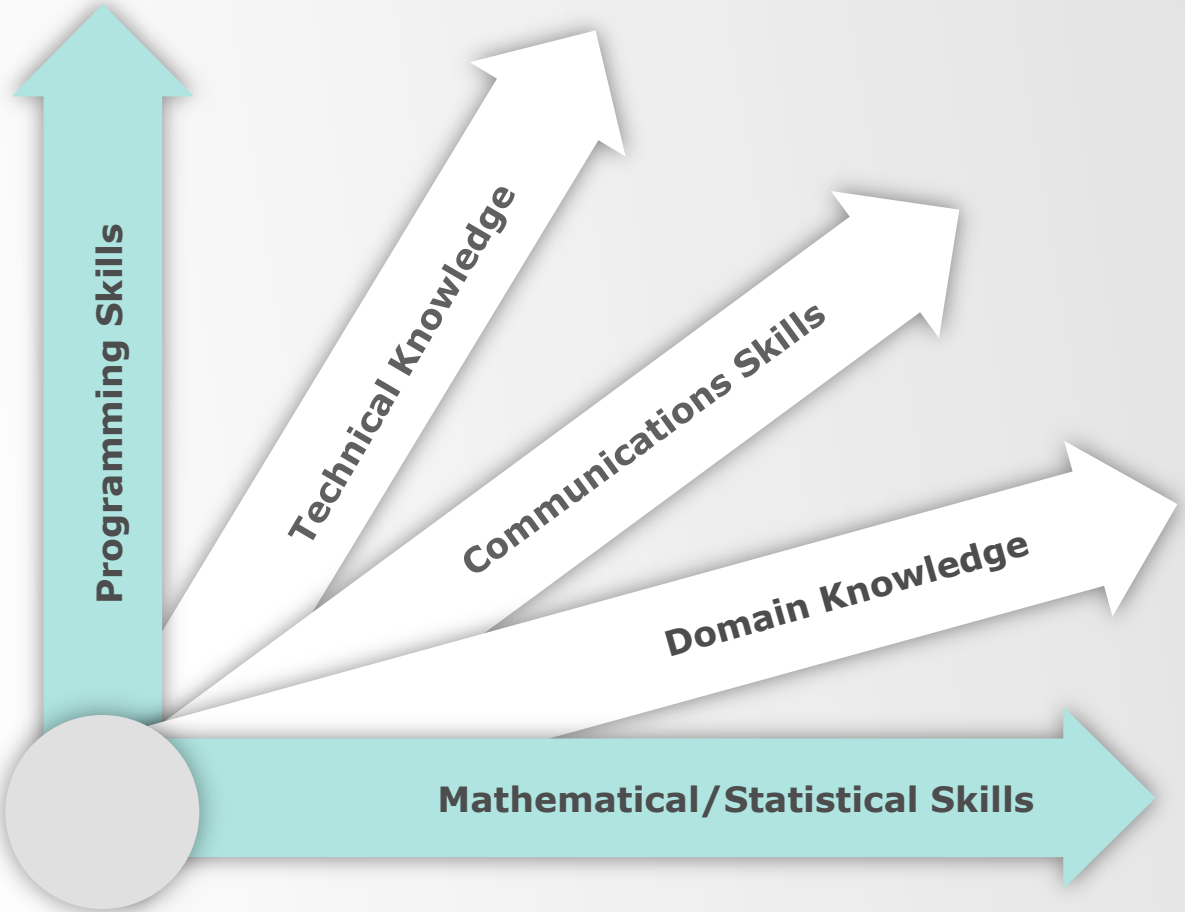
# What Is Pivotal Data Labs?


Data Engineering

+

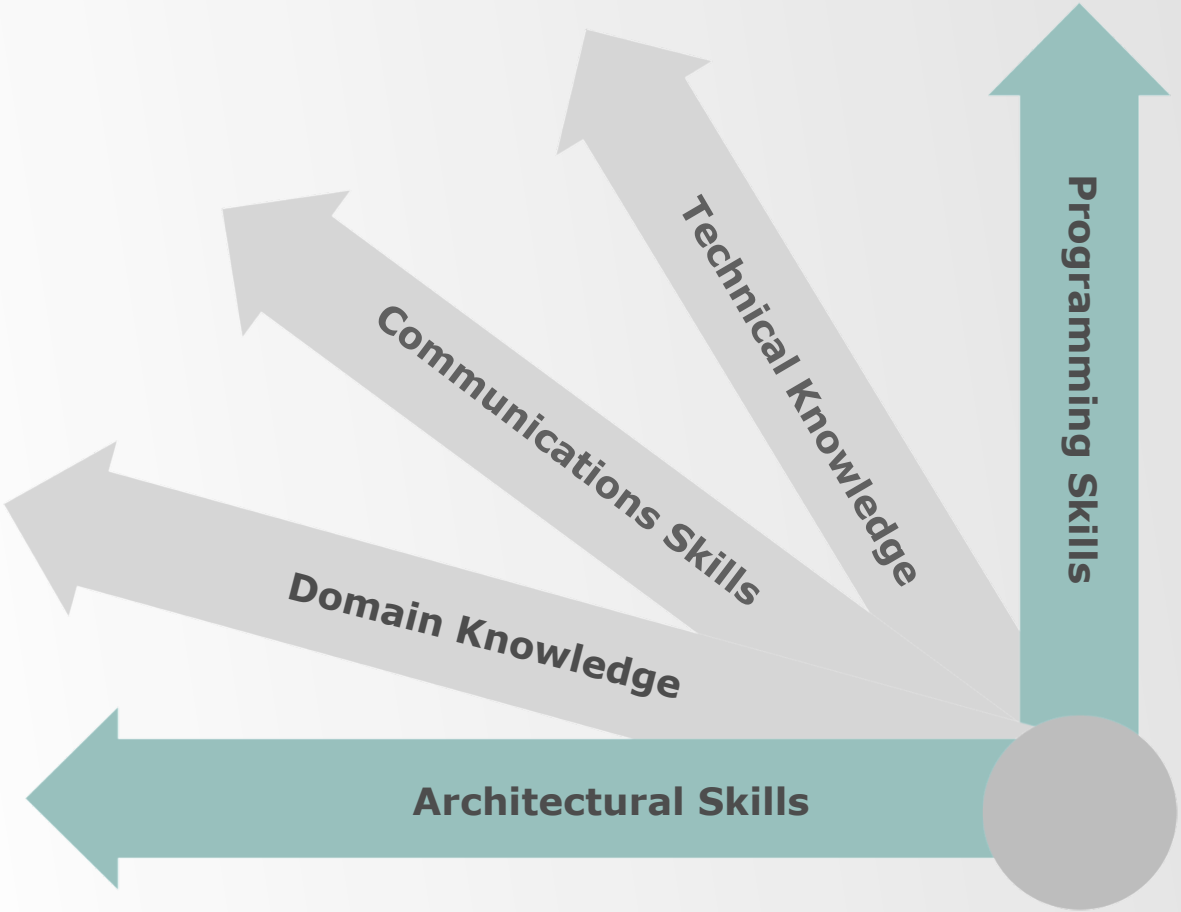
Data Science

# What is a "Data Scientist"?



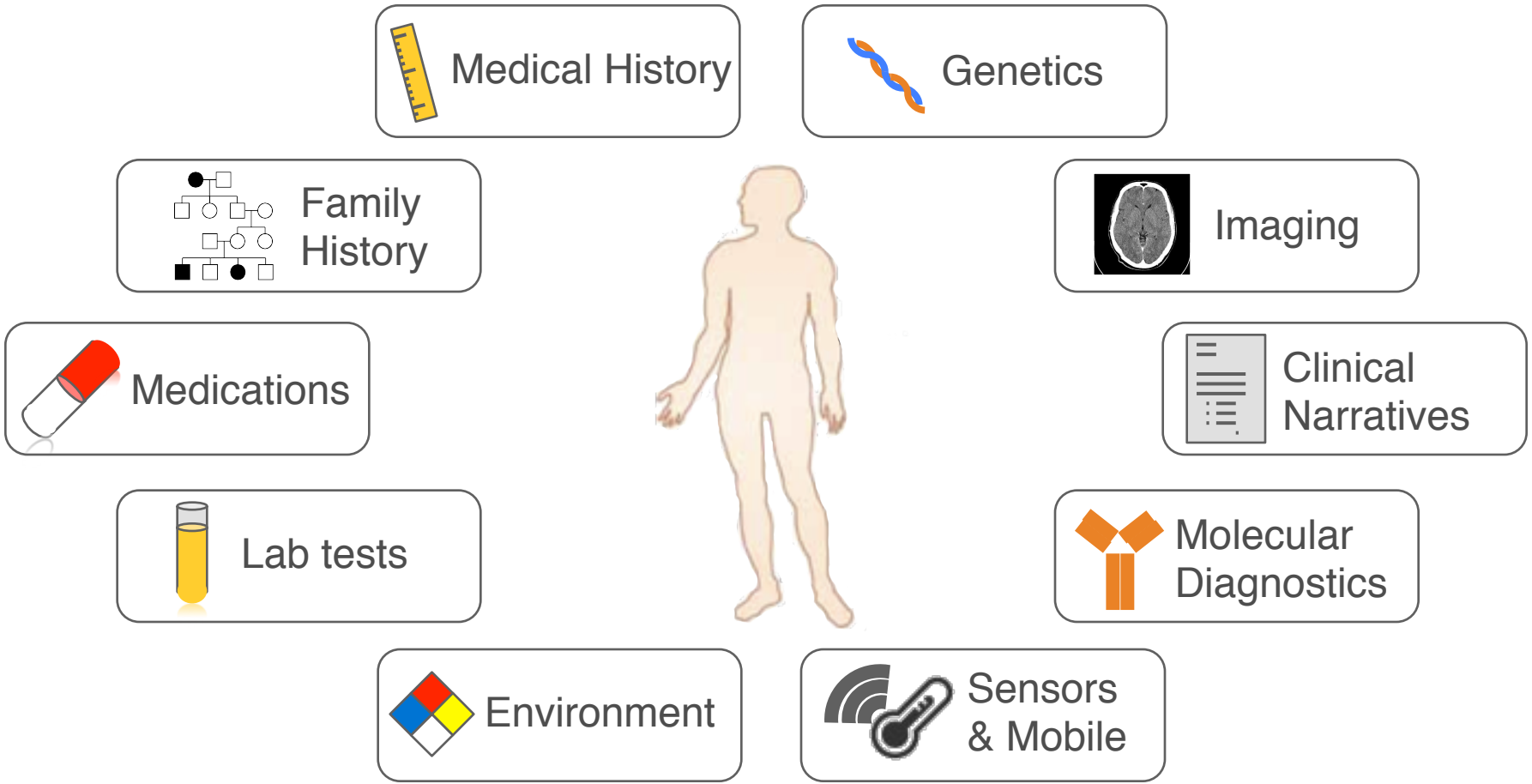

Pivotal

# What is a "Data Engineer"?

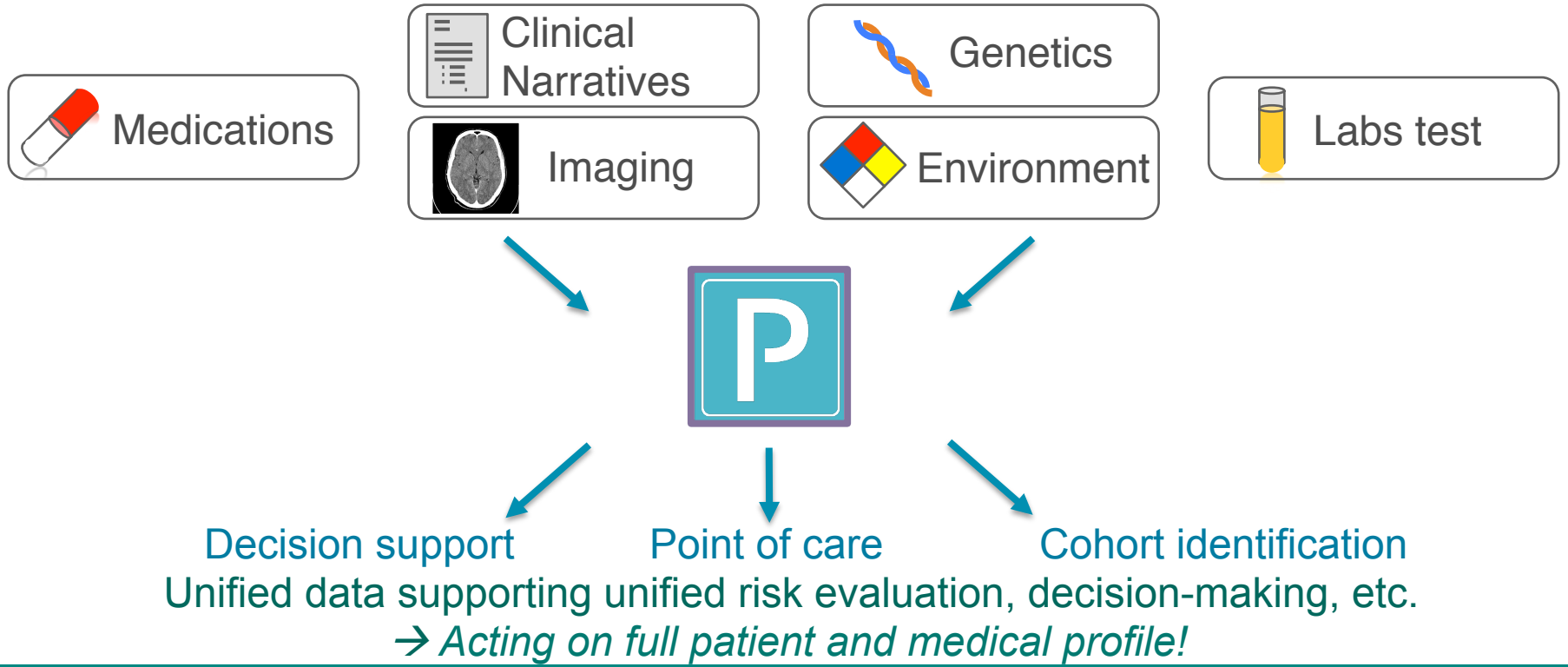



Pivotal

# The Quantified Patient



# Leveraging healthcare data to drive predictive and personalized care

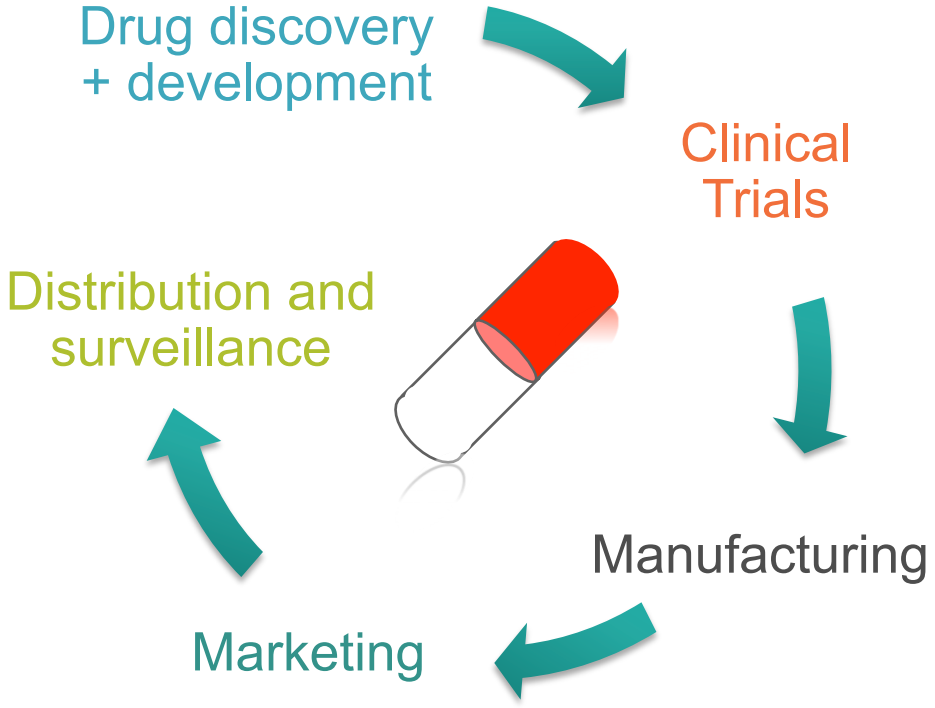



Unified data supporting unified risk evaluation, decision-making, etc.

→ *Acting on full patient and medical profile!*

Pivotal

# Data driven drugs: From discovery to delivery




## RICH DATA SOURCES

- Molecular data
  - Cellular drug screens
  - Animal models
- Clinical data including notes, images, markers (e.g. genomics, lab results)
- Sensor and assay data
- Internal and partner/purchased external data
- Contact center data
- Patient registries, public and federal data, clinical partnerships

Pivotal

# The landscape of technology for big data

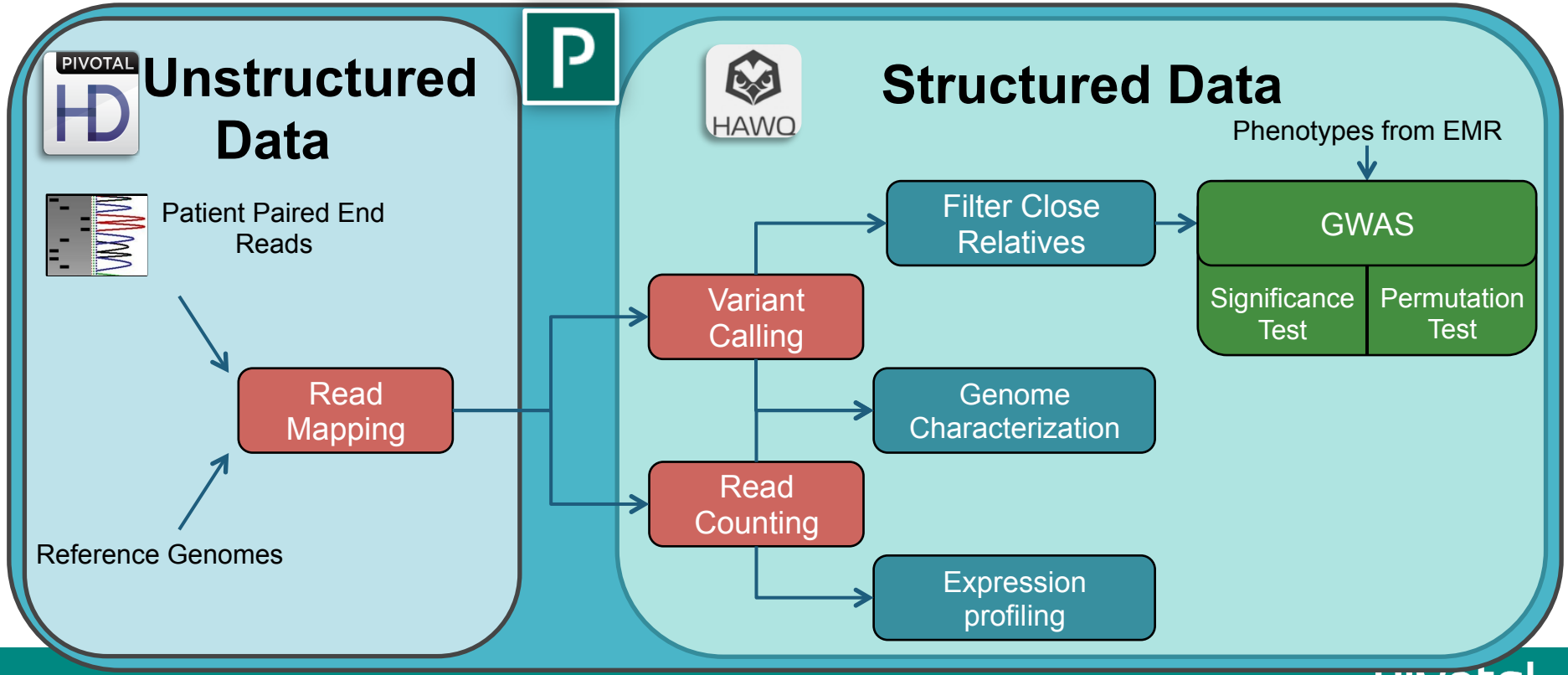
| | Sample Applications | Challenges | Use Cases |
|---|---|---|---|
|   | Batch processing of large volumes of data | Not optimal for highly iterative methods, functions over windows | Word count on tweets |
|   | Analytics on large-scale structured data | Requires restructuring of data to manipulate very large files | Predicting mortality on clinical data from diverse sources |
| HAMSTER/MPI GraphLab  | Operations on very large matrices | Requires knowledge of OpenMP, mis-used for embarrassingly parallel problems | Protein docking, molecular dynamics |

Pivotal

# Choosing the right environment for different analytics challenges

| |  Imaging |  Clinical Narratives |  Genetics |
|---|---|---|---|
| PIVOTAL HD | Good for processing many images rapidly | Many documents with no shared processing | Read mapping |
| HAWQ | In-database processing of very large images stored as a table | Information retrieval | BAM file manipulations, counts |
| HAMSTER/MPI GraphLab  | Processing very large images | | Multiple sequence alignment |

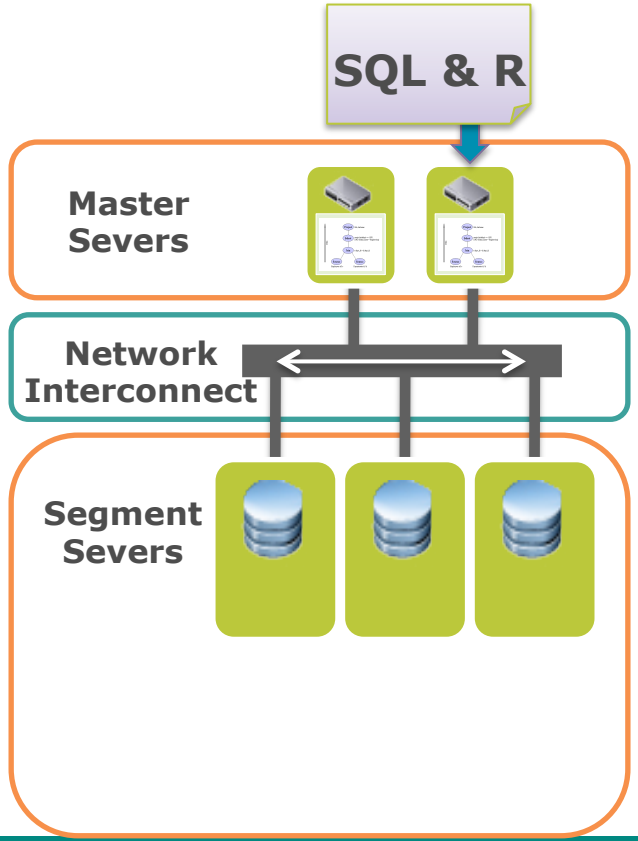
# A new architecture for improved pipeline



# In-database genome-wide association study

**COVARIATES**

| Indiv | Covariates | | | | SNP | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | | 10 | 1 | 2 | | M |
| 1 | F | 23 | | 18 | AA | CC | | TT |
| 2 | M | 39 | | 41 | AT | CG | | TT |
| 3 | M | 50 | | 23 | AA | GG | | TC |
| N | F | 19 | | 24 | TT | CG | | TC |



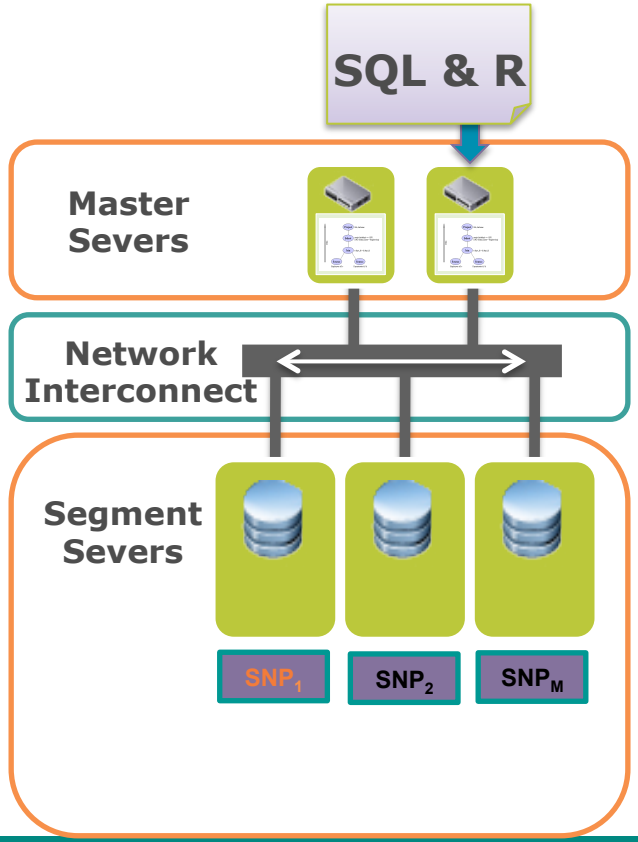
# In-database genome-wide association study

**COVARIATES**

| Indiv | Covariates | | | |
|---|---|---|---|---|
| | 1 | 2 | … | 10 |
| 1 | F | 23 | … | 18 |
| 2 | M | 39 | … | 41 |
| 3 | M | 50 | … | 23 |
| … | … | … | … | … |
| N | F | 19 | … | 24 |

**GENOTYPES**

| Indiv | SNP | Geno |
|---|---|---|
| 1 | 1 | AA |
| 2 | 1 | AT |
| 3 | 1 | AA |
| 1 | 2 | CC |
| 2 | 2 | CG |
| 3 | 2 | GG |
| … | … | … |
| N | M | TC |




Pivotal

# In-database genome-wide association study

**COVARIATES**

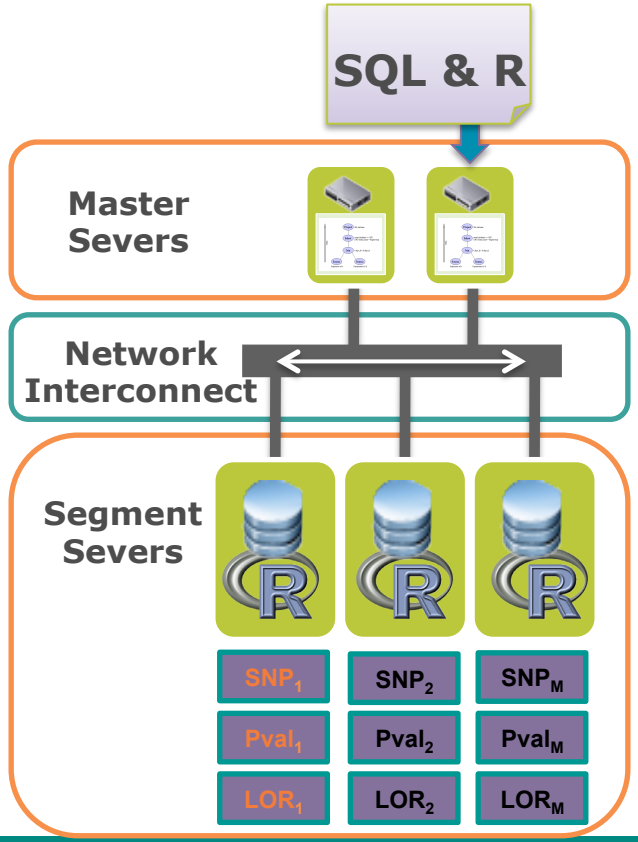
| Indiv | Covariates | | | |
|---|---|---|---|---|
| | 1 | 2 | | 10 |
| 1 | F | 23 | | 18 |
| 2 | M | 39 | | 41 |
| 3 | M | 50 | | 23 |
| N | F | 19 | | 24 |

**GENOTYPES**

| Indiv | SNP | Geno |
|---|---|---|
| 1 | 1 | AA |
| 2 | 1 | AT |
| 3 | 1 | AA |
| 1 | 2 | CC |
| 2 | 2 | CG |
| 3 | 2 | GG |
| N | M | TC |

**RESULTS**

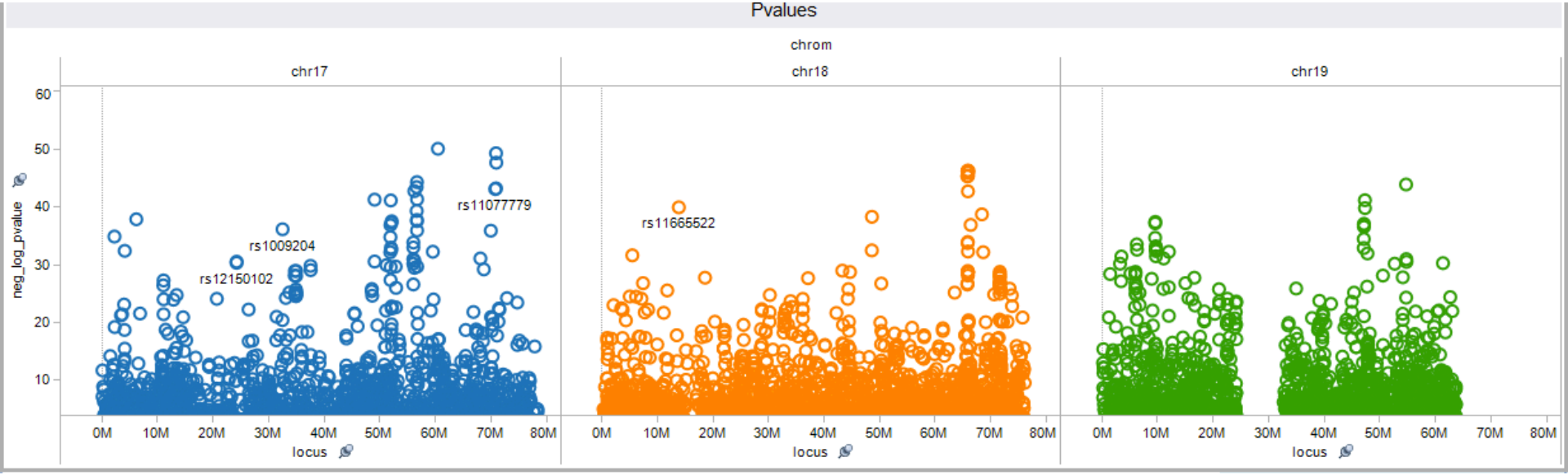
| SNP | P-value |
|---|---|
| 1 | $2.34 \times 10^{-21}$ |
| 2 | 0.395 |
| 3 | $7.15 \times 10^{-17}$ |
| M | 0.000142 |




- In-database computation of ~500,000 loci for thousands of individuals occurs rapidly and in parallel
- Results are easily manipulated and explored

Pivotal

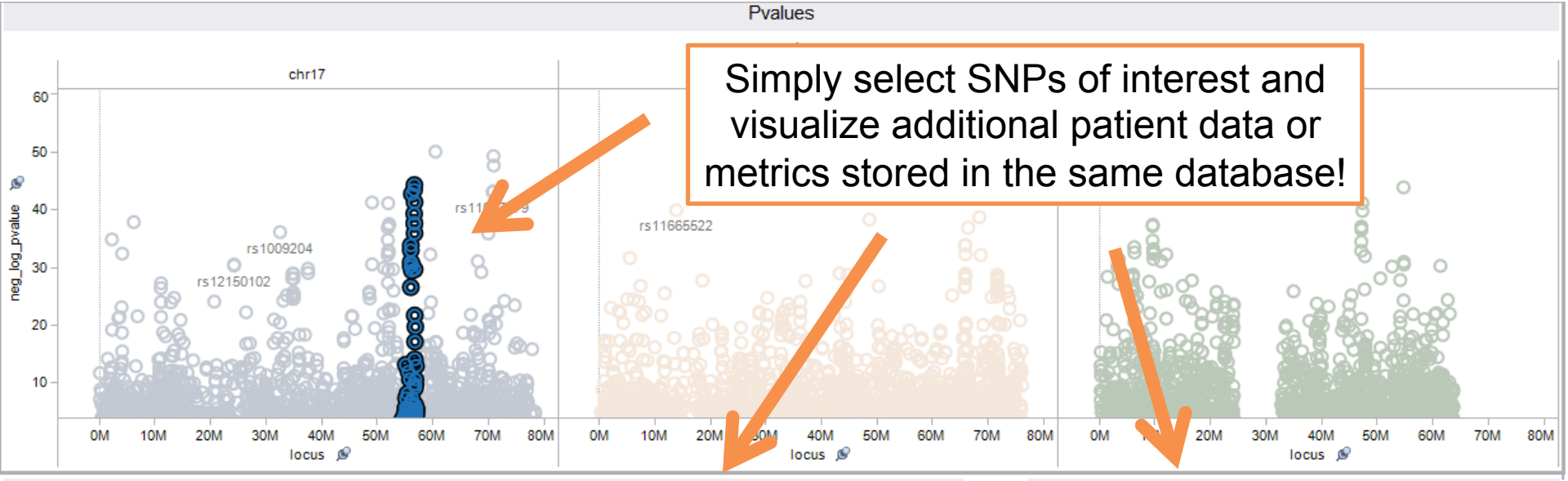
# Visualize and analyze genomics data without movement




Generate relevant plots using tools like Tableau
immediately after parallel statistical analysis in-database
on Pivotal technology

Pivotal

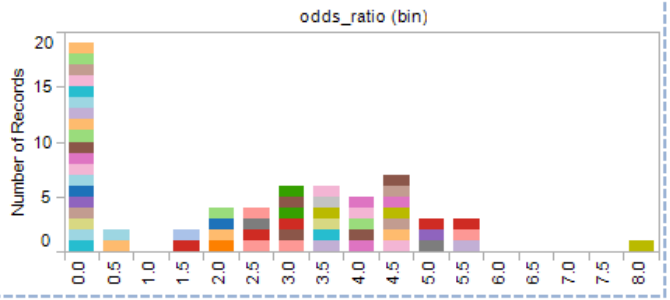
# Visualize and analyze genomics data without movement



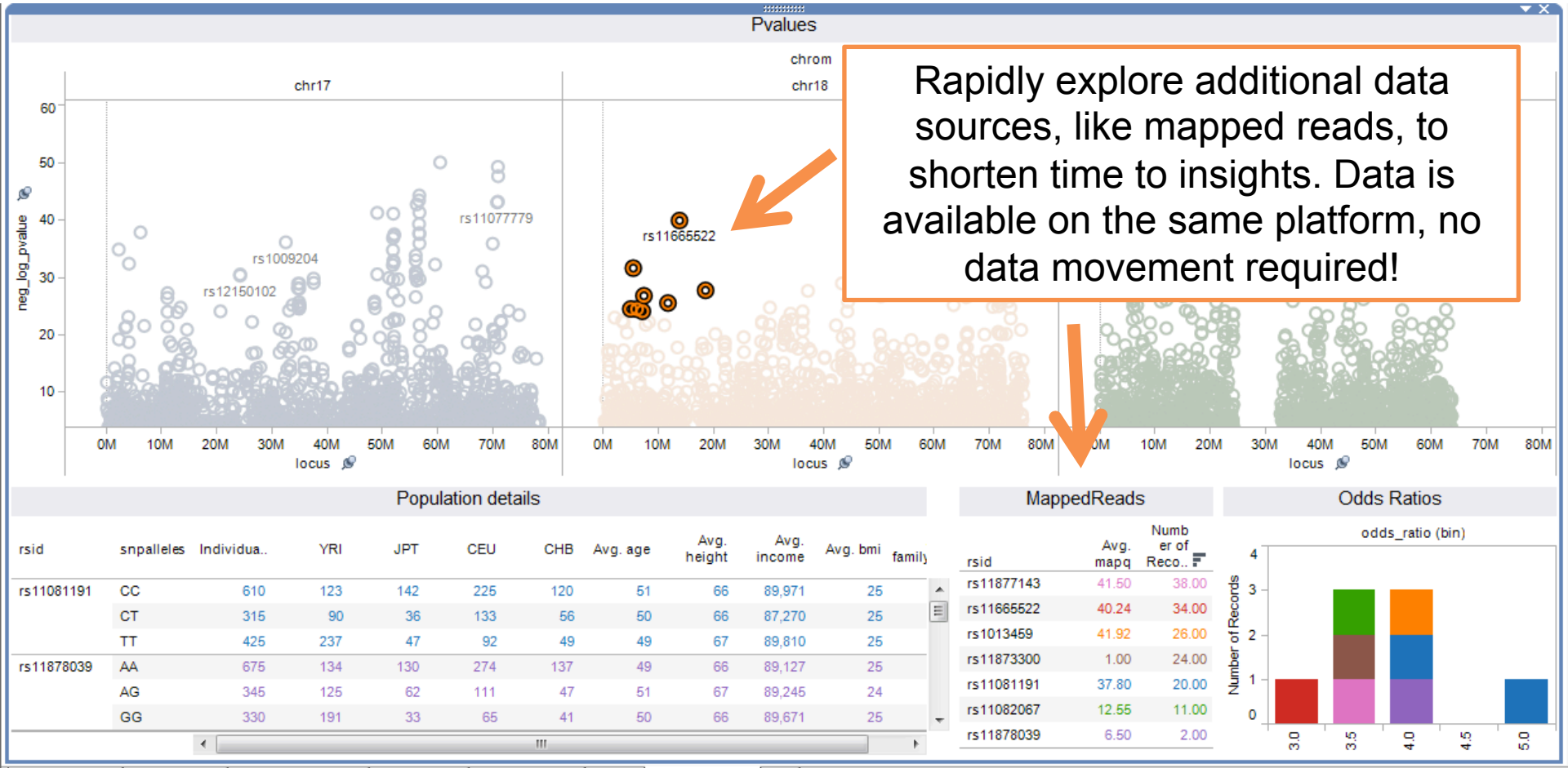

## Population details

| rsid | snpalleles | Individua.. | YRI | JPT | CEU | CHB | Avg. age | Avg. height | Avg. income | Avg. bmi | Avg. familysize |
|---|---|---|---|---|---|---|---|---|---|---|---|
| rs11658440 | TT | 975 | 379 | 175 | 236 | 185 | 50 | 66 | 89,844 | 25 | 5 |
| rs11658956 | AA | 380 | 216 | 38 | 75 | 51 | 51 | 67 | 90,572 | 25 | 4 |
| | AG | 330 | 107 | 71 | 106 | 46 | 49 | 66 | 87,846 | 25 | 5 |
| | GG | 640 | 127 | 116 | 269 | 128 | 50 | 66 | 89,274 | 25 | 4 |
| rs11867179 | CC | 820 | 324 | 105 | 261 | 130 | 50 | 66 | 89,076 | 25 | 4 |
| | CG | 400 | 101 | 86 | 135 | 78 | 49 | 66 | 89,380 | 25 | 5 |
| | GG | 105 | 16 | 32 | 45 | 12 | 52 | 67 | 90,403 | 24 | 5 |

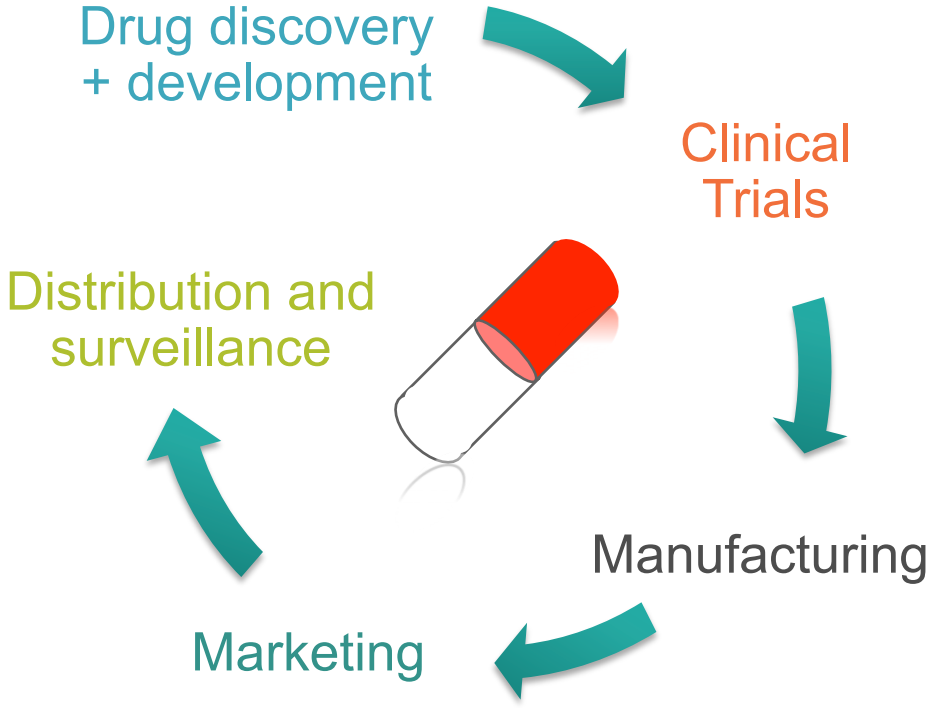
## Odds Ratios



# Visualize and analyze genomics data without movement



# Data driven drugs: From discovery to delivery




## RICH DATA SOURCES

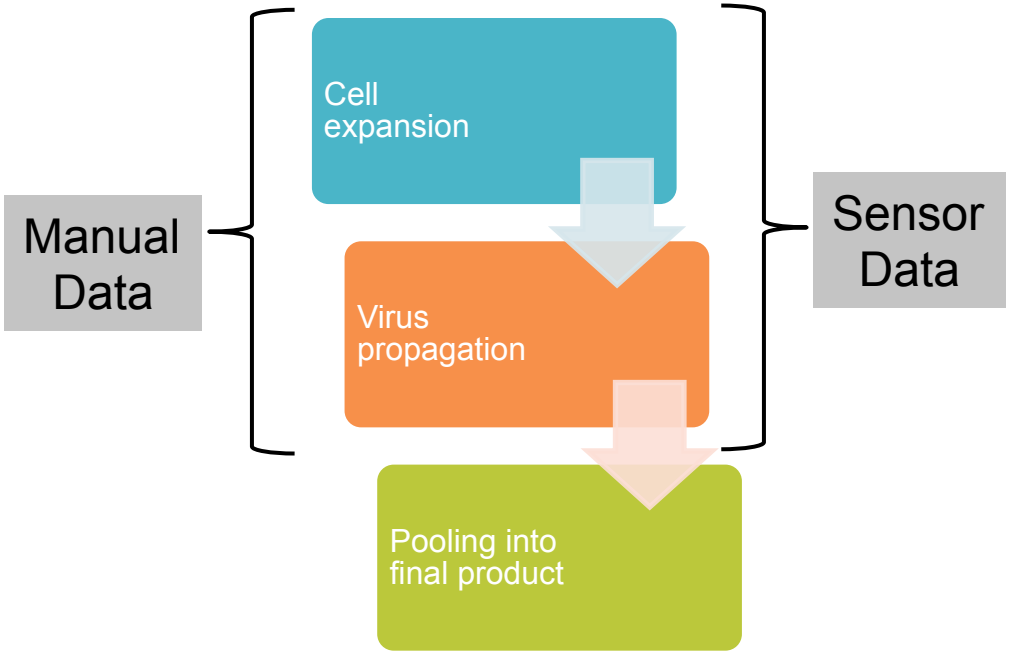
- Molecular data
  - Cellular drug screens
  - Animal models
- Clinical data including notes, images, markers (e.g. genomics, lab results)
- Sensor and assay data
- Internal and partner/purchased external data
- Contact center data
- Patient registries, public and federal data, clinical partnerships

Pivotal

# Vaccine Potency Prediction

## Business Problem

Predict potency and antigen levels of live virus vaccines based on manufacturing sensor data and manual data collected throughout the process.



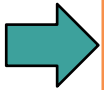

Simplified Vaccine Manufacturing Process

Pivotal

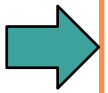
# Enabling predictive models through new architectures

**Challenges**

- Accessibility
  - Some data had never been used in predictive modeling due to poor data models
- Data Integrity
  - Manually entered data is prone to errors. There is no immediate feedback to examine the validity of the values entered
- Data Completeness
  - Manual data entry is time consuming. There is no feedback on what data is most useful in improving the efficiency and quality and hence no prioritization of what data should be collected



Purpose-built data models for rapid data querying and exploration



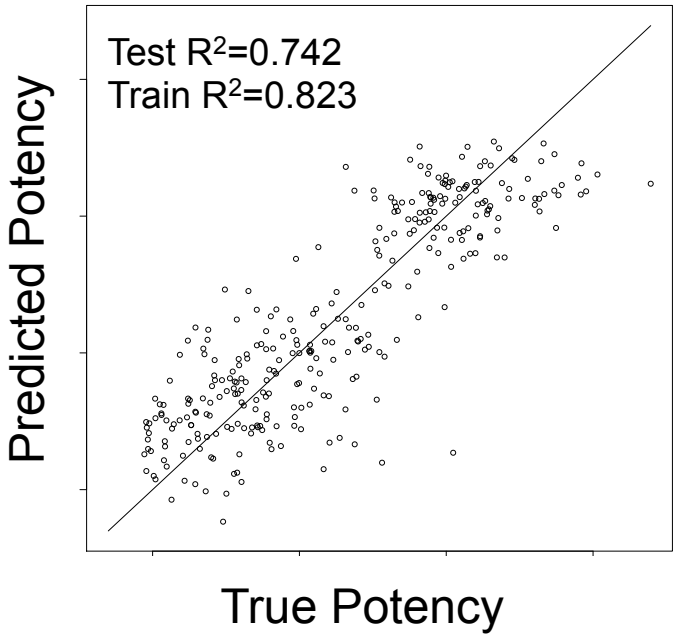
Automated data cleansing techniques

Opportunities to eliminate collection of incomplete or non-predictive data

Pivotal

# Model generation and evaluation

## Predicting vaccine potency using manufacturing data

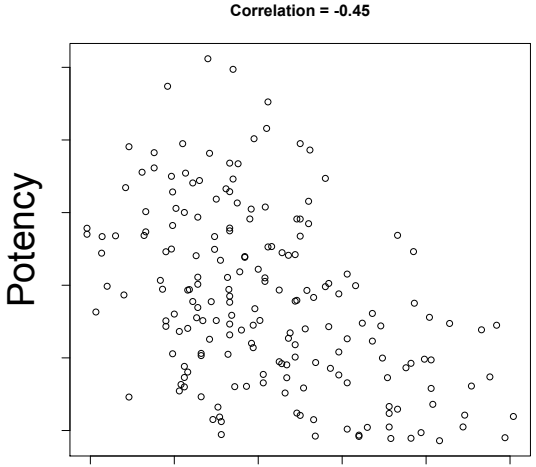



- Feature engineering and transformation
  - Enabled by rapid in-database processing
- Experimentation with model forms
  - Partial least squares
  - Random forest
  - Regularized regression
- Interpretation of model results for insight generation
  - Use cross-validation framework to assess variable importance

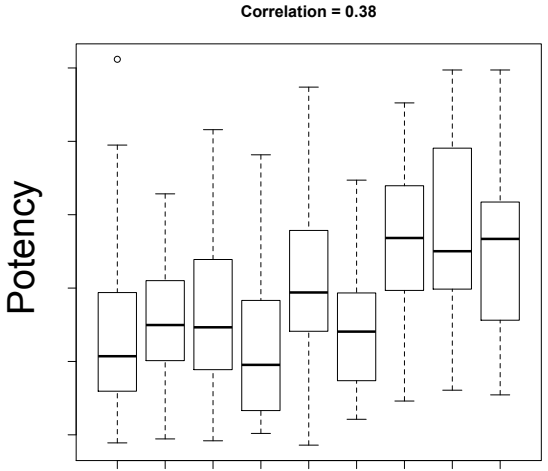
Pivotal

# Sample model insights

Interpreting the utility of a measure obtained during manufacturing based on model outcomes




Assayed value

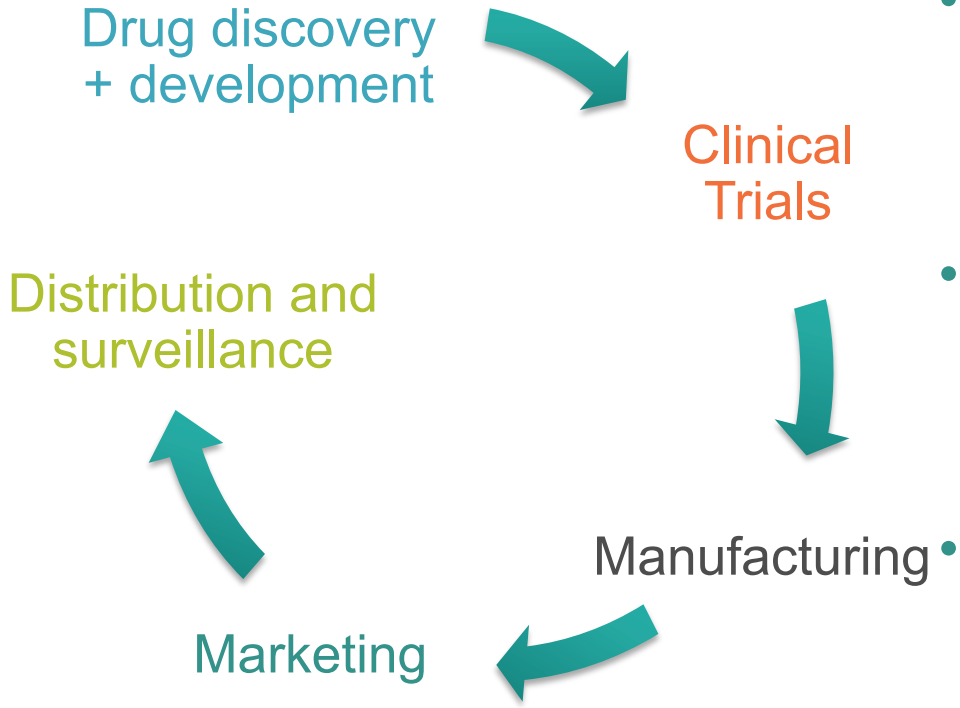



Duration of a step

- Some features may reveal tunable parameters to alter potency, others may simply be markers
- Features consistently absent from models ***may*** be uninformative for predicting potency
- Opportunities to provide real-time feedback on data entry errors and predicted potency outcomes

Pivotal

# Data driven drugs: From discovery to delivery

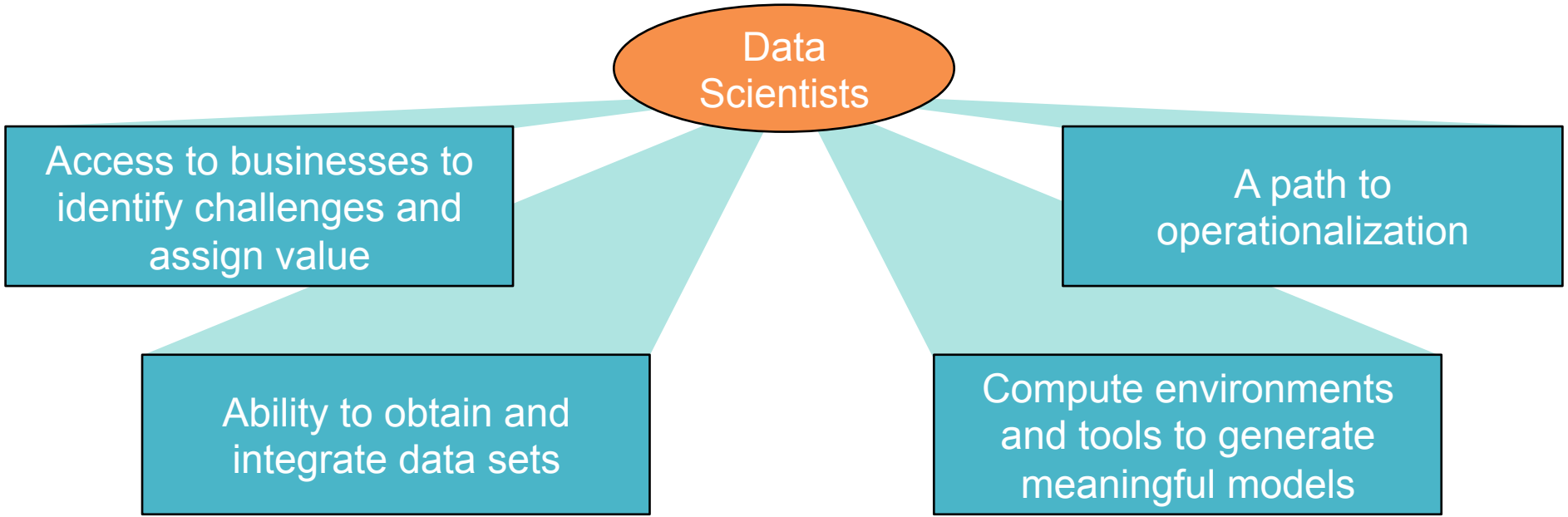



- **Data repurposing**
  New value exists in leveraging historical data across drugs and stages
  **Adverse events for new clinical indications**
- **Data discovery**
  External and publicly available datasets can augment proprietary sources
  **Twitter data to forecast demand**
- **Data collection**
  Obtaining new data from different sources drives additional value
  **Mobile and sensor data to measure patient adherence and outcomes**

Pivotal

# Enabling Disruption with Data Science

- To successfully build a data-driven enterprise, skilled individuals need to have access to data, tools and channels for operationalization




Pivotal


BUILT FOR THE SPEED OF BUSINESS