The Data Science Process

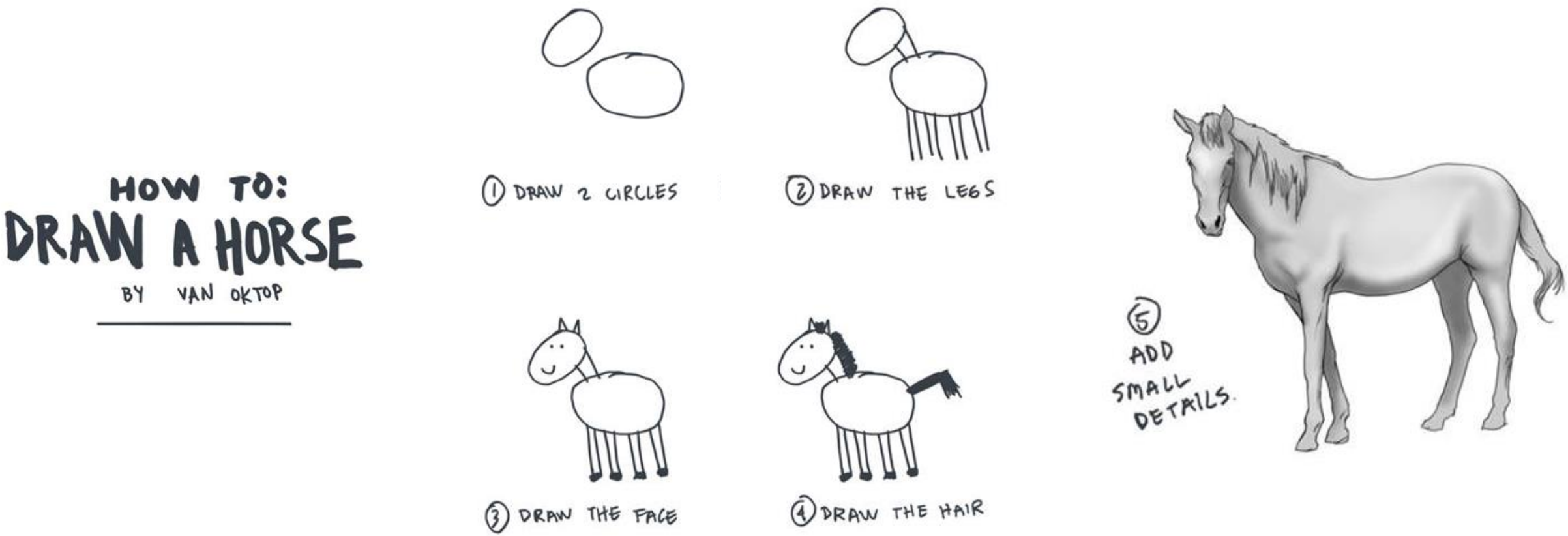
DAPT 631

Data
Science

BUILD
A MACHINE LEARNING MODEL
IN JUST THREE
QUICK AND EASY STEPS
USING [...]!!!

– Most tutorials

How to Become a Data Scientist?



2

50% of analytic projects fail.

– Gartner, 2015

Through 2022, only 20% of analytic insights will deliver business outcomes.

– Gartner, 2019

https://blogs.gartner.com/andrew_white/2019/01/03/our-top-data-and-analytics-predicts-for-2019/

Data + Machine Learning = Profit...?

On September 21, 2009, the grand prize of US$1,000,000 was given to the BellKor's Pragmatic Chaos team which bested Netflix's own algorithm for predicting ratings by 10.06%.

“[T]he additional accuracy gains that we measured did not seem to justify the engineering effort needed to bring them into a production environment.”

Netflix Technology Blog

Learn more about how Netflix designs, builds, and operates our systems and engineering organizations

Apr 5, 2012

Analytic projects fail because…

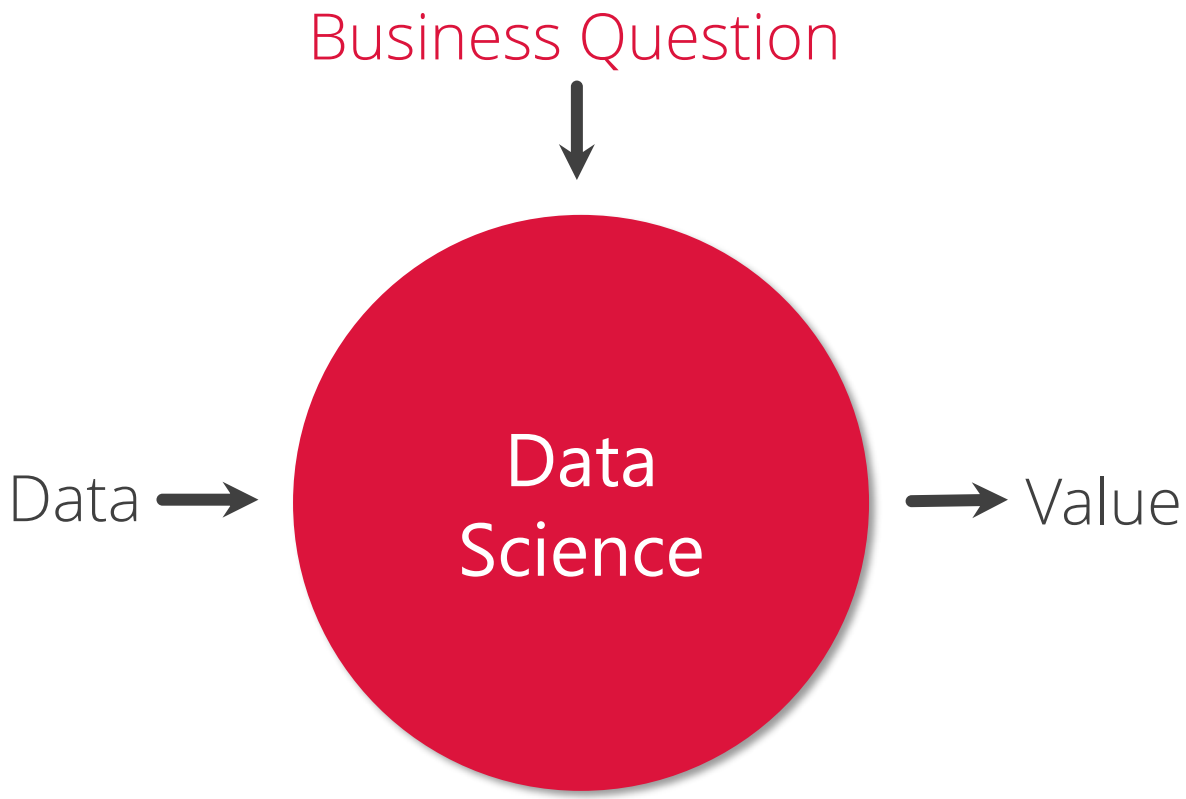
…they aren't completed within budget or on schedule,

or because they fail to deliver the features and benefits

that are optimistically agreed on at their outset.

How to Avoid Failure?

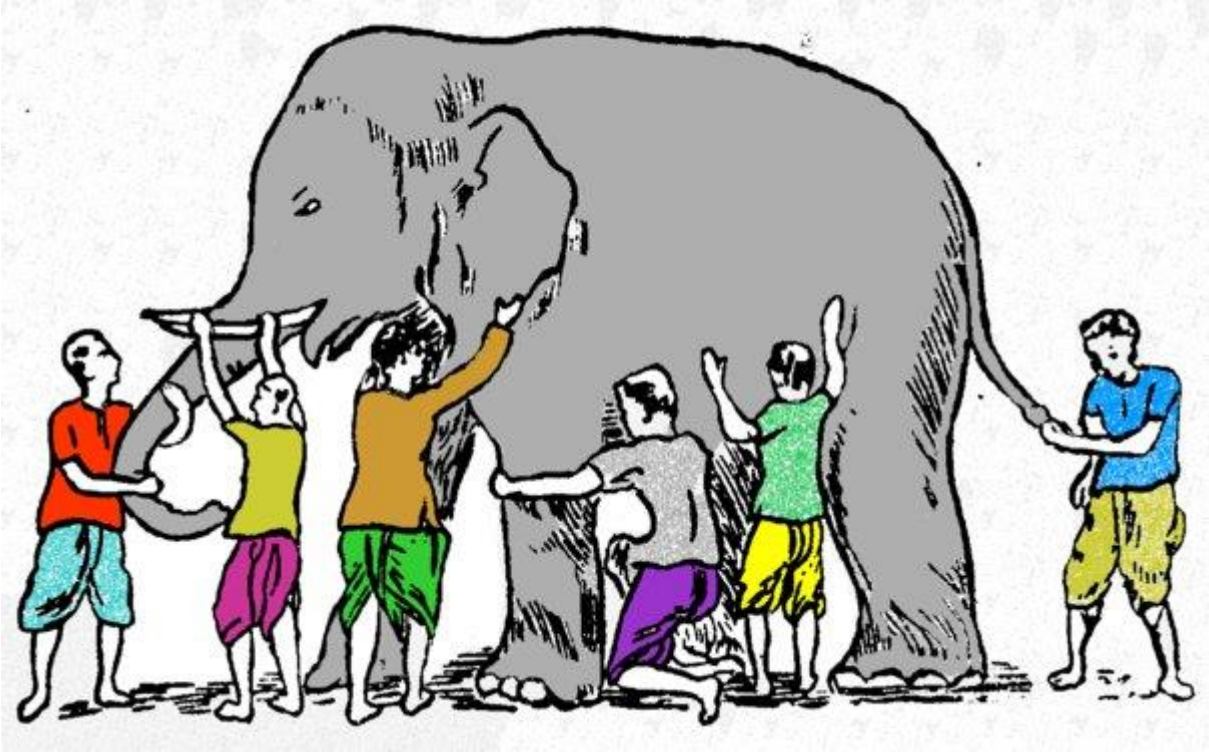
1. Build with Organizational Buy-in
2. Build with End In Mind
3. Build with a Structured Approach


Business Question
Data
Data Science
Value

“The beginning of wisdom is to call things by their proper name.”

– Confucius

The Blind Men and the Elephant

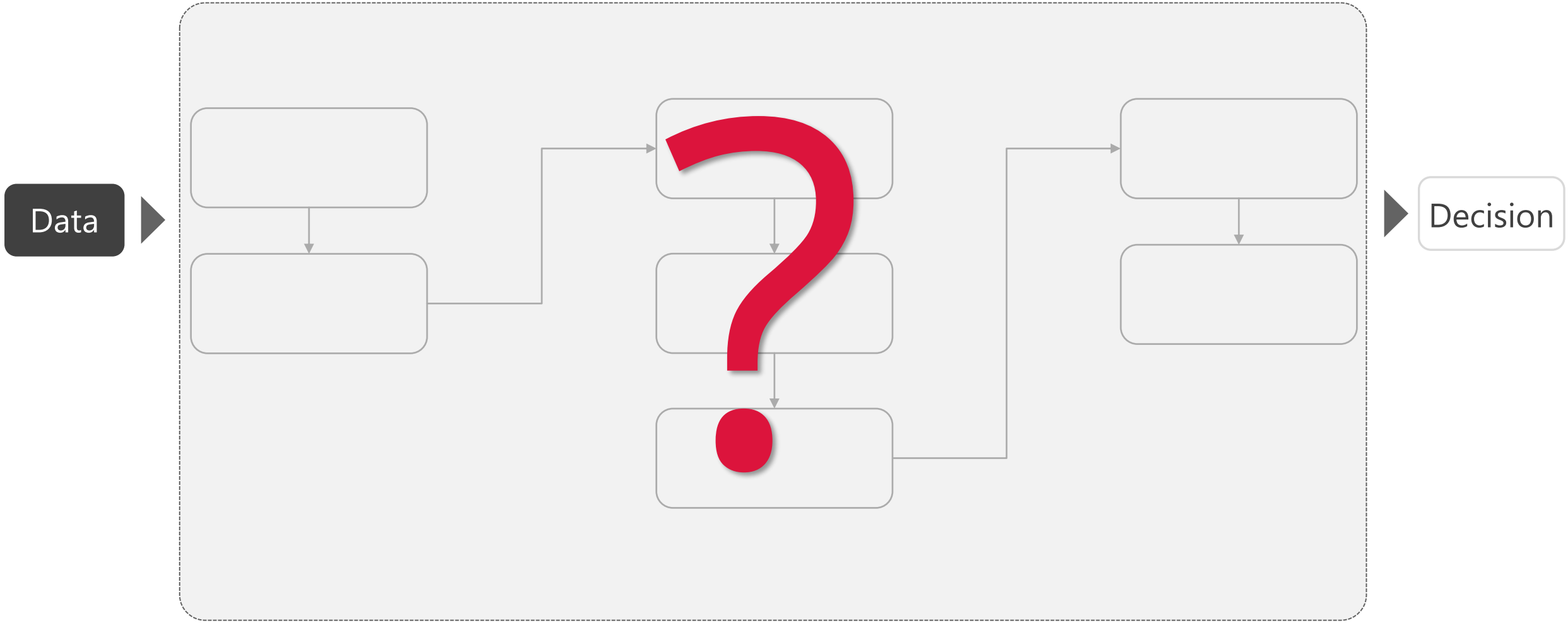


It was six men of Indostan
To learning much inclined,
Who went to see the Elephant
(Though all of them were blind),
That each by observation
Might satisfy his mind.

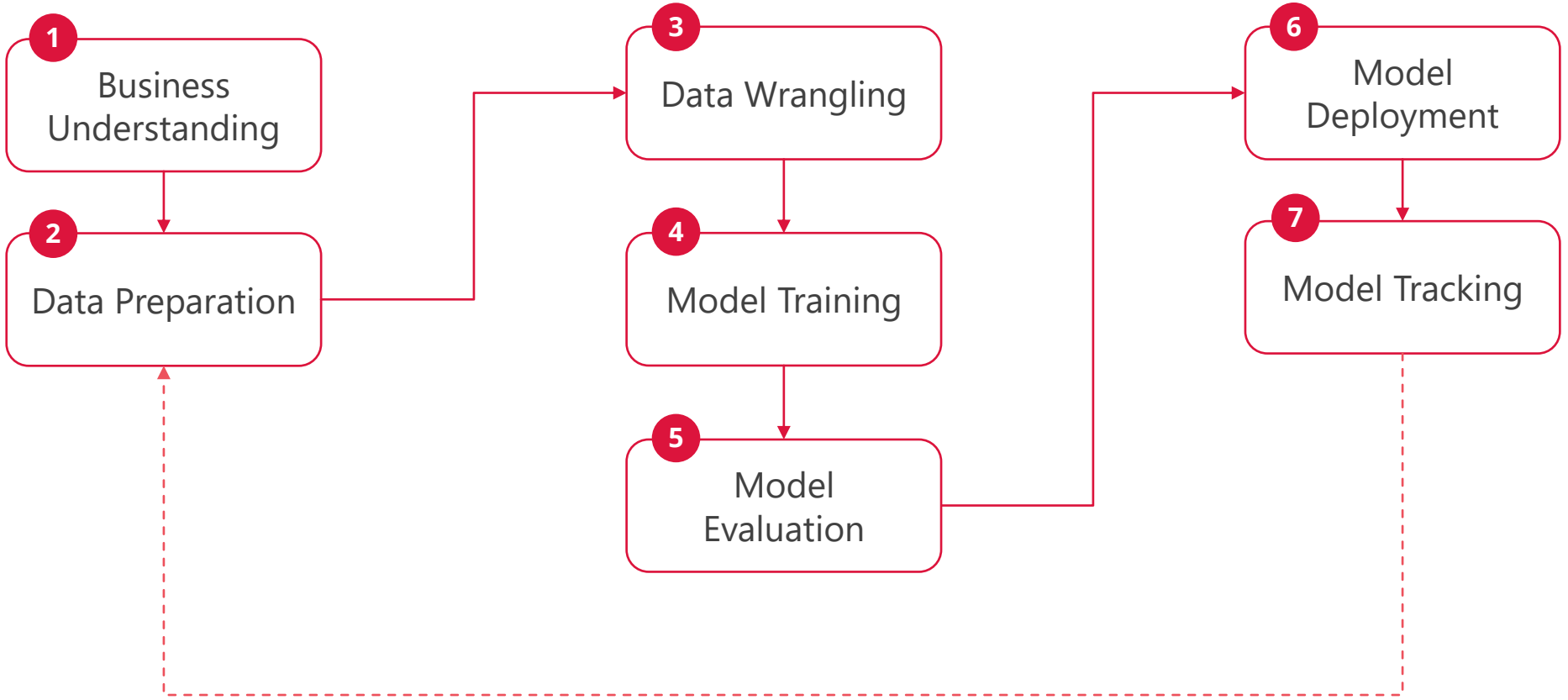
And so these men of Indostan
Disputed loud and long,
Each in his own opinion
Exceeding stiff and strong,
Though each was partly in the right
And all were in the wrong!

– John Godfrey Saxe

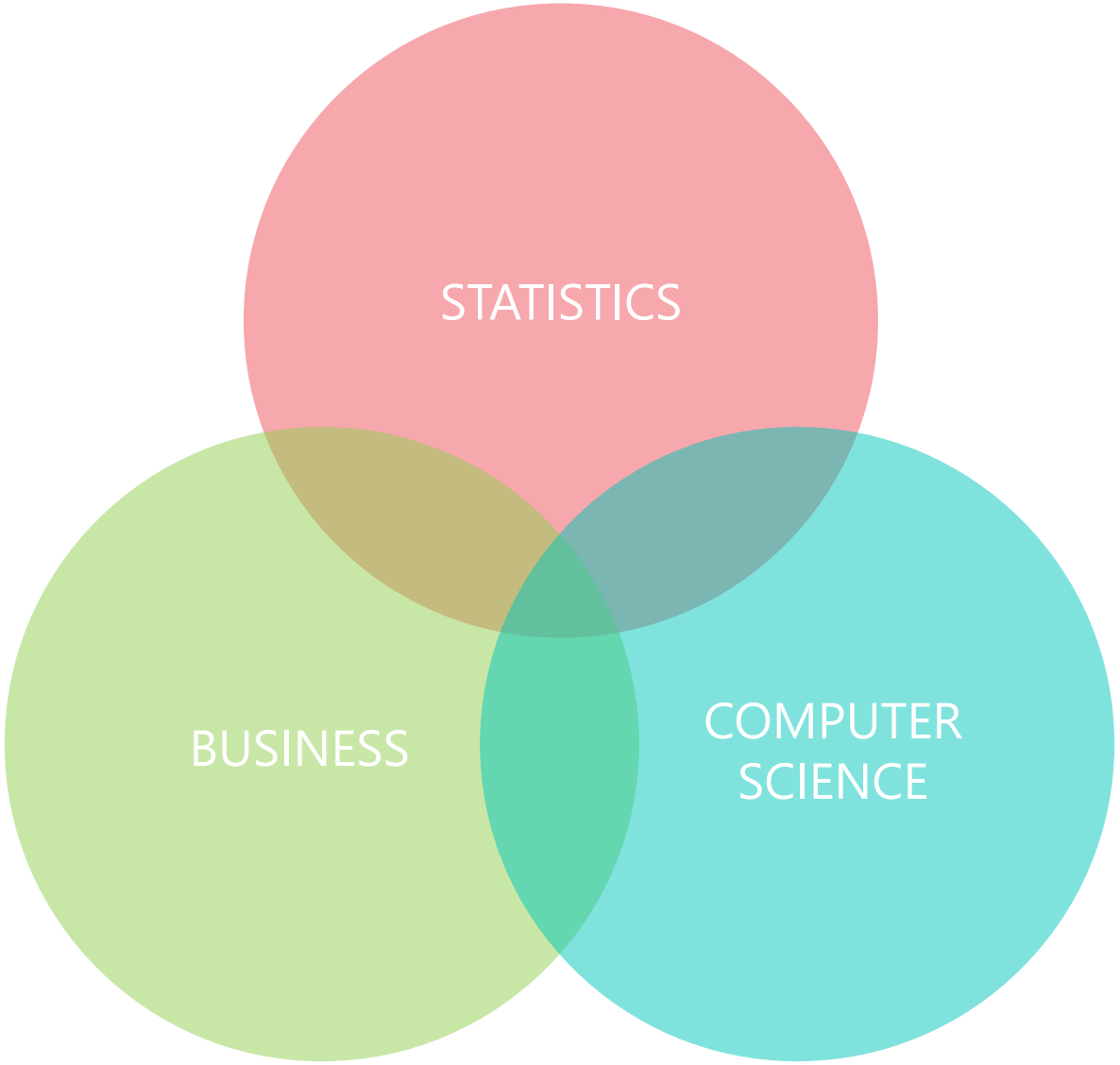
Data
Data Science
Decision
Past
Future


Data
Decision

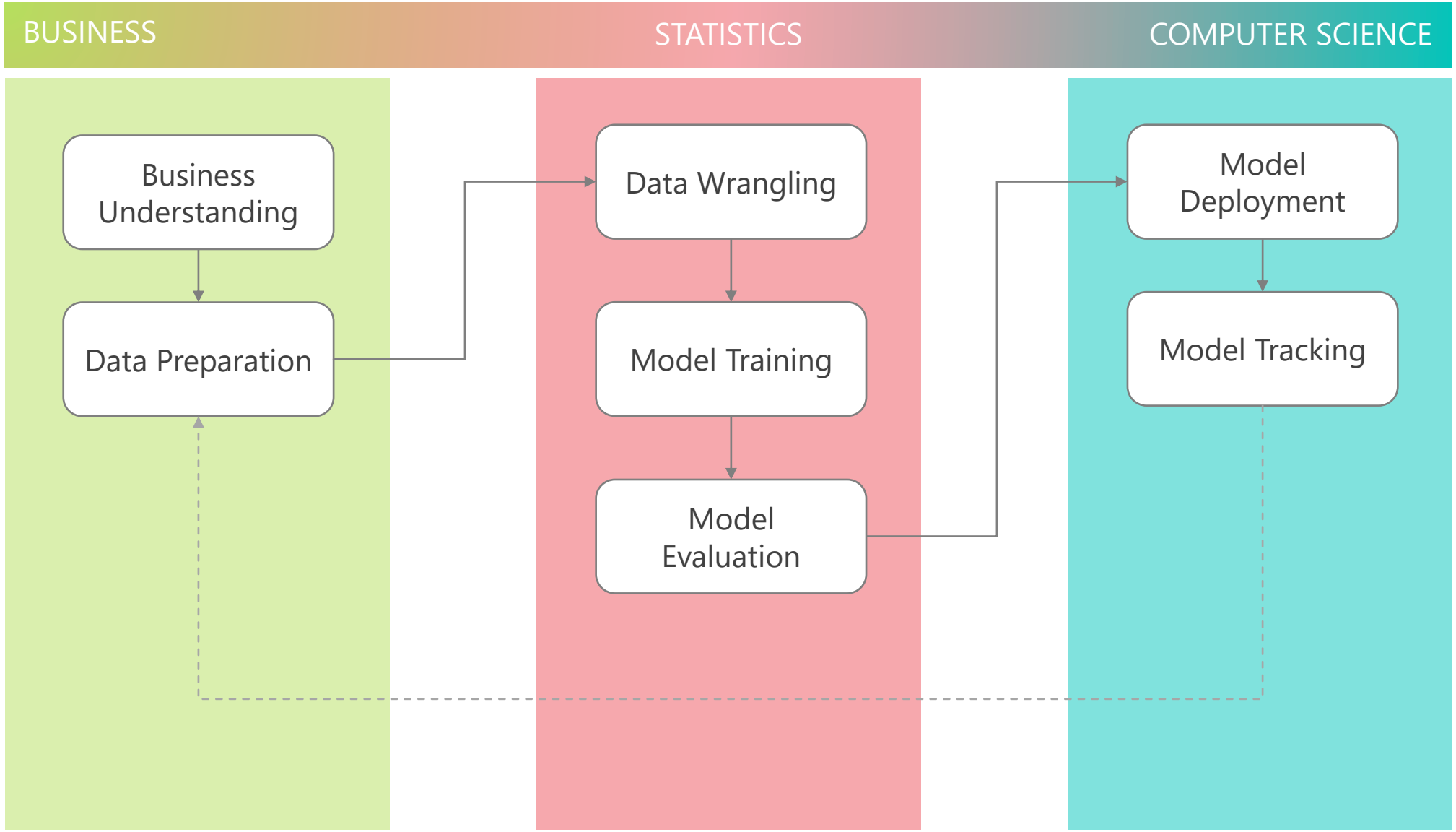
Data Science Process



Data Science



The Data Science Process



The Data Science Process

Business Understanding	Data Preparation	Data Wrangling	Model Training	Model Evaluation	Model Deployment	Model Tracking
Determine	Identify	Impute	Train	Evaluate	Document	Monitor
Understand	Collect	Transform	Assess	Peer Review	Deploy	Maintain
Map	Assess	Reduce	Select	Present		Test
	Vectorize					

Business Understanding	Data Preparation	Data Wrangling	Model Training	Model Evaluation	Model Deployment	Model Tracking

Far better

an approximate answer to the right question

than

an exact answer to the wrong question.

– John Tukey

1 **DETERMINE**

2 **UNDERSTAND**

3 **MAP**

What does the client want to achieve?

1 **DETERMINE**

2 **UNDERSTAND**

3 **MAP**

Primary Objective

- Reduce attrition
- Customized targeting
- Plan future media spend
- Prevent fraud
- Recommend Products

1 DETERMINE

2 **UNDERSTAND**

- Understand **success criteria**
  - Specific, measurable, time-bound
- List **assumptions, constraints,** and **important factors**
- Identify **secondary or competing objectives**
- Study **existing solutions** (if any)

3 MAP

1 DETERMINE

2 UNDERSTAND

3 MAP

Business Objective → Technical Objective

- State the **project objective(s) in technical terms**
- Describe how the data science project will **help solve the business problem**
- Explore **successful scenarios**

A problem well stated

is a problem half-solved.

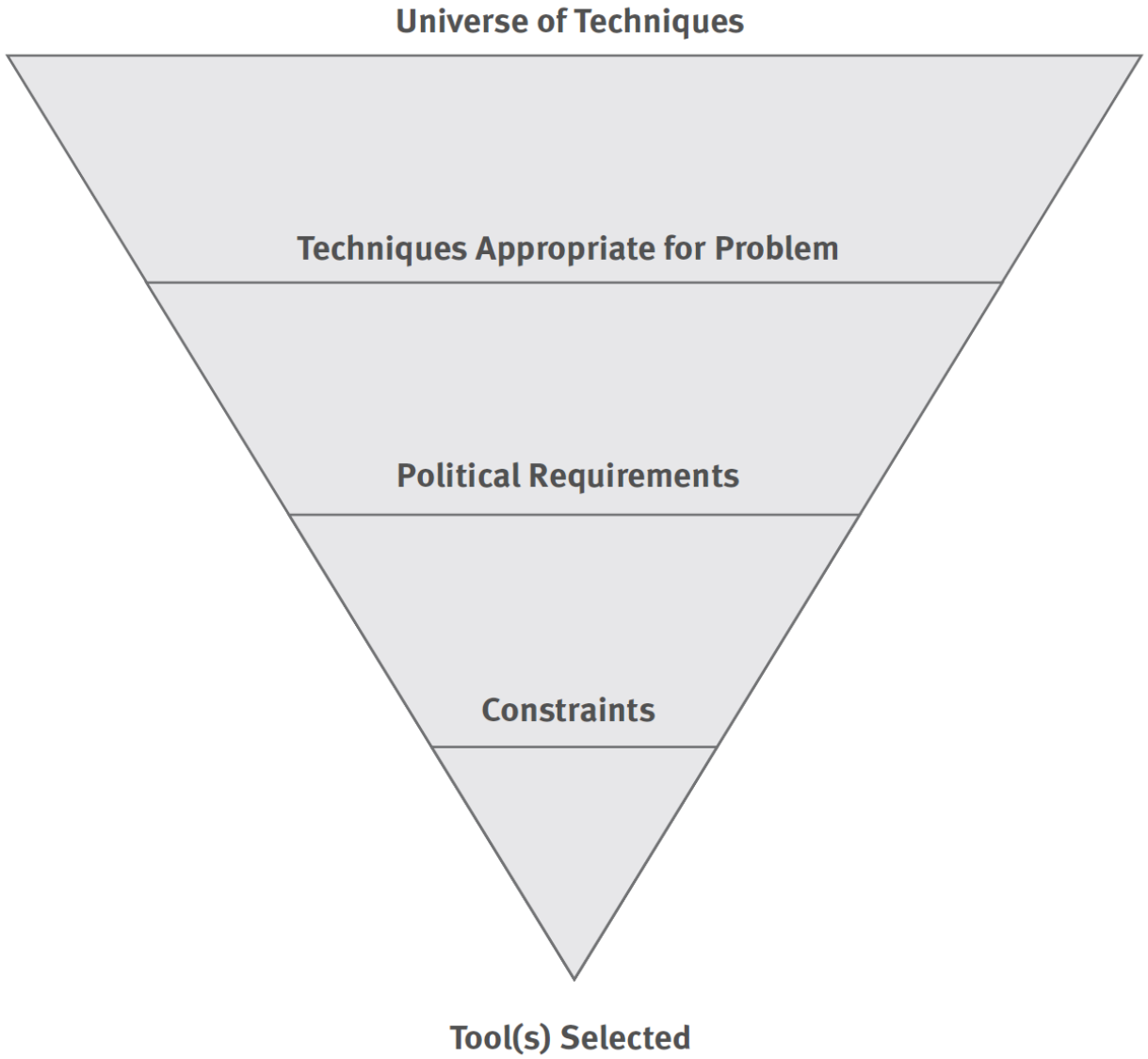
– CF Kettering

1 DETERMINE

2 UNDERSTAND

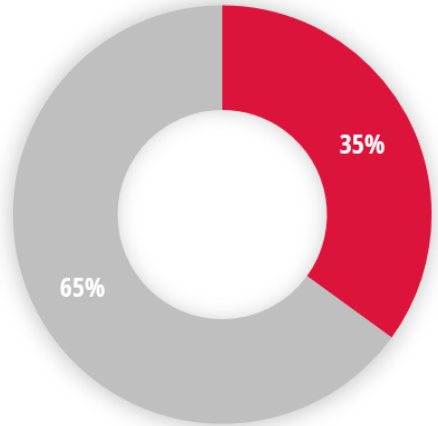
3 MAP

OBJECTIVE →	TECHNIQUE →	EXAMPLES
Predict Values	Regression	Linear regression, Bayesian regression, Decision Trees
Predict Categories	Classification	Logistic regression, SVM, Decision Trees
Predict Preference	Recommender System	Collaborative / Content-based filtering
Discover groups	Clustering	k-means, Hierarchical clustering
Identify unusual data points	Anomaly Detection	k-NN, One-class SVM
...		


Universe of Techniques
Techniques Appropriate for Problem
Political Requirements
Constraints
Tool(s) Selected

If all you have is a hammer
then everything looks like a nail.

- **Primary Objective:** Prevent attrition → Increase subscription renewals
- **Competing Objective:** High value customers are also targeted for up-sell
- **Constraints:** Avoid targeting customers too close to their contract expiration
- **Success Criteria:** Current renewal rate = 65% → Improve by 8%
- **Existing Solution:** Business-rule-based targeting
- **Data Science Objective:** Build a binary classification model to identify customers who are not likely to renew their subscriptions three months in advance of their contract expiration.
- **Success Scenario:** The model correctly identifies 80% of the future attritors, a promotional campaign targets all likely attritors, and successfully converts 19% of them into non-attritors.

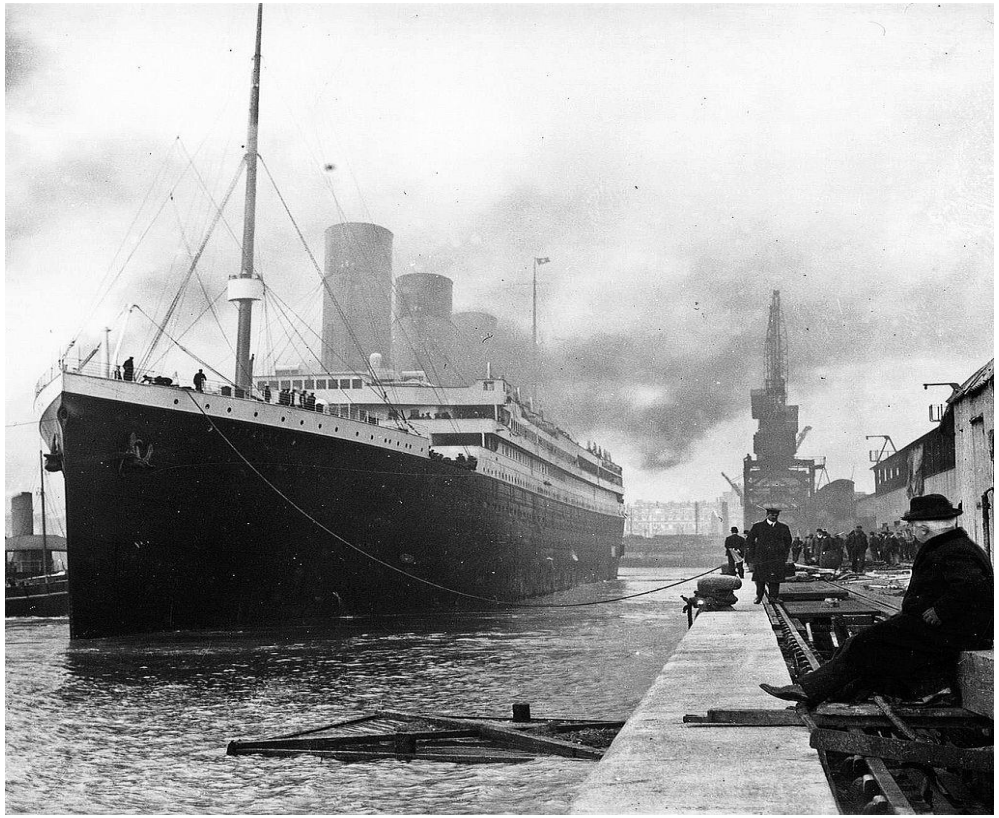


Project Plan

- Duration
- Inventory of resources
- Tools and techniques
- Risks and contingencies
- Costs and benefits
- Milestones

The thought that disaster is impossible often leads to an unthinkable disaster.

– Gerald Weinberg



Titanic at Southampton docks, prior to departure

1 **IDENTIFY**

2 **COLLECT**

3 **ASSESS**

4 **VECTORIZE**

1 **IDENTIFY**

2 COLLECT

3 ASSESS

4 VECTORIZE

- **Data sources, formats**
  - Database, Streaming API's, Logs, Excel files, Websites, etc.
- **Entity Relationship Diagram** (ERD)
- Identify **additional data sources**
  - Demographics data appends,
  - Geographical data,
  - Census data, etc.
- Identify **relevant data**
- Record **unavailable data**
- How long a history is available, and how much of it should be used?

1 IDENTIFY

2 **COLLECT**

- Access or acquire all relevant data in **a central location**
- **Quality control checks and tests**
  - File formats, delimiters
  - Number of records, columns
  - Primary keys

3 ASSESS

4 VECTORIZE

1 IDENTIFY

2 COLLECT

3 **ASSESS**

4 VECTORIZE

First look at the data

- **Get familiar** with the data
- Study **seasonality**
  - Monthly/weekly/daily patterns
  - Unexplained gaps or spikes in the historical data
- Detect **mistakes**
  - Extreme or outlier values
  - Unusual values
  - Special missing values
- Check **assumptions**
- Review **distributions**

tableau SOFTWARE

alteryx

ggplot2

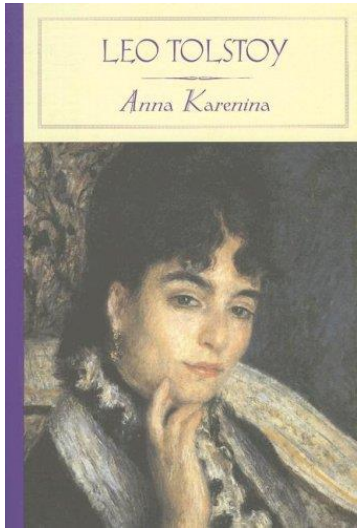
matplotlib

seaborn

Tidy dataset are all alike;

Every messy dataset is messy in its own way.

- Hadley Wickham



There is no substitute for getting to know your data.

– Witten and Frank

1 IDENTIFY

2 COLLECT

3 ASSESS

4 **VECTORIZE**

GOAL: Create the Analysis Dataset

$$y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ . \\ y_n \end{pmatrix}$$

Outcome
Target
Independent Variable

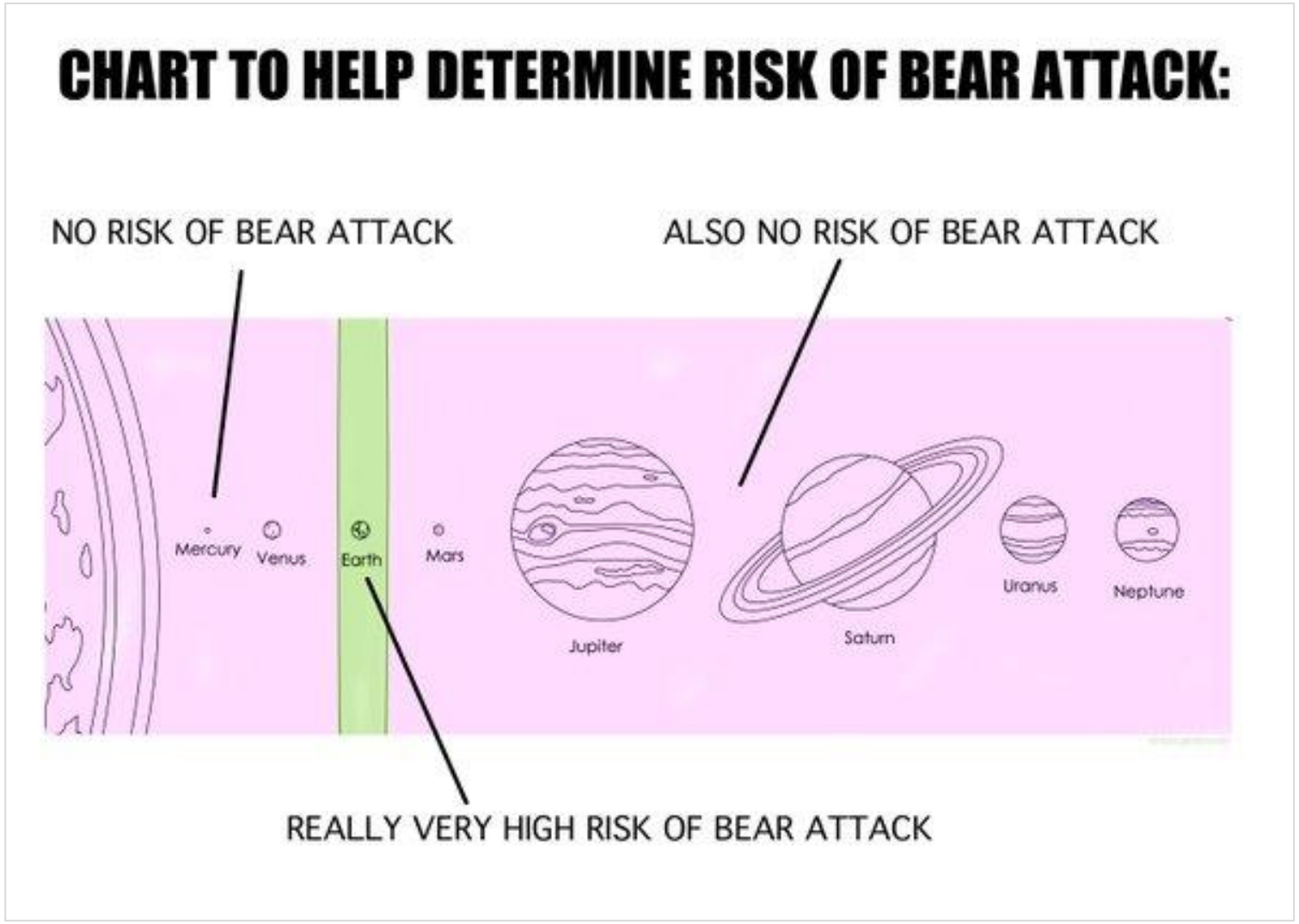
$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1j} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2j} \\ x_{31} & x_{32} & x_{33} & \dots & x_{3j} \\ . & . & . & & . \\ . & . & . & & . \\ . & . & . & & . \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{nj} \end{pmatrix}$$

Inputs
Features / Attributes
Dependent Variables

Target Definition

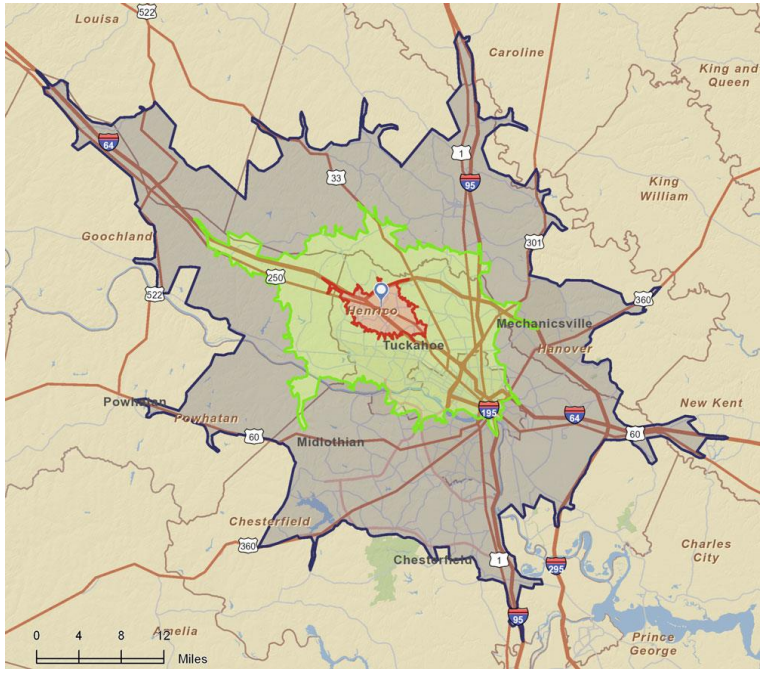
- **Churn = 90 days of consecutive inactivity** (for a pre-paid telecom customer)
- What's **inactivity**?
  - Incoming and outgoing calls
  - Data usage
  - Incoming text
  - Promotional texts
  - Voicemail usage
  - Call forwarding
  - Etc.
- Customers may **change their device** or phone number.
  - Churn at the individual (person) level, or at the device (phone) level?
- Customers may return (become active again) after 90 days of inactivity?
- Prediction window
  - Predict 90 days of consecutive inactivity?
  - Would 10 days of consecutive inactivity suffice?
  - How many customers return after *x* days of inactivity?
- Fraud, Involuntary churn
- …

Accurate but not Precise

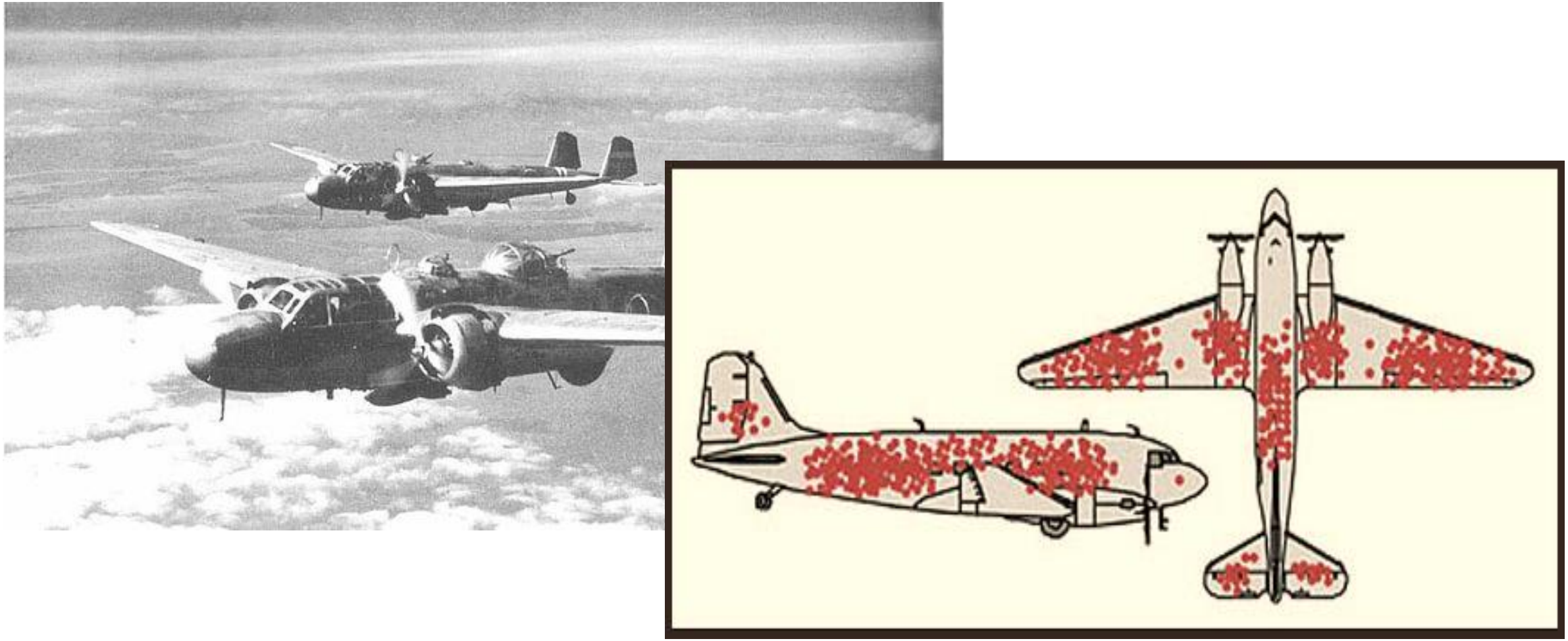


Modeling Sample

- **Historical trends and seasonality**
  - Are there certain timeframes that should be discarded?
  - The model should be generalizable
- **Eligible, relevant population**
  - Must align with the business goals
- **Eligible, relevant markets**
  - Must align with the business goals
  - E.g., within a certain drive-time distance
- **Outdated products or events**



Selection Bias



Abraham Wald's Work on Aircraft Survivability

Journal of the American Statistical Association Vol. 79, No. 386 (June, 1984)

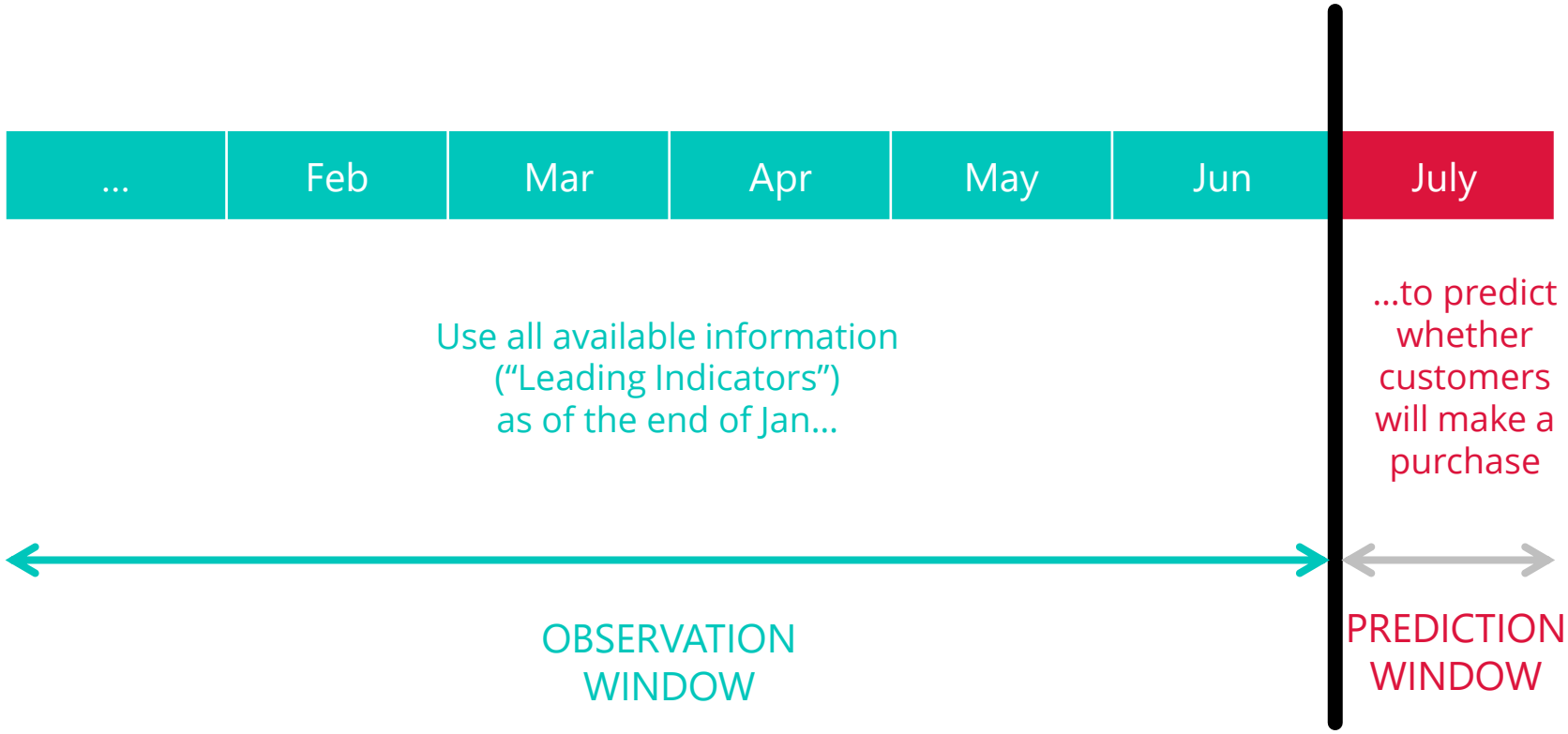
Making Prediction about the Future

Dog or muffin?

vs.

Who is likely to churn?

Information Leakage



- The leading indicators must be calculated from the timeframe *leading up to* the event – it must not overlap with the prediction window.
- Beware of proxy events, e.g., future bookings

Data Aggregation

- **Attribute creation**
  - Derived attributes: Household income / Number of adults = Income per adult
- **Brainstorm with team members** (both technical and non-technical)

$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1j} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2j} \\ x_{31} & x_{32} & x_{33} & \dots & x_{3j} \\ . & . & . & & . \\ . & . & . & & . \\ . & . & . & & . \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{nj} \end{pmatrix}$$

Derived Attributes

CUSTOMER ID	PURCHASE DATE
1001	02-12-2015:05:20:39
1001	05-13-2015:12:18:09
1001	12-20-2016:00:15:59
1002	01-19-2014:04:28:54
1003	01-12-2015:09:20:36
1003	05-31-2015:10:10:02
…	…

CUSTOMER ID	x_1	x_2	…	x_j
1001	…	…		…
1002	…	…		…
1003	…	…		…
…	…	…	…	…

1. Number of transactions (Frequency)
2. Days since the last transaction (Recency)
3. Days since the earliest transaction (Tenure)
4. Avg. days between transaction
5. # of transactions during weekends
6. % of transactions during weekends
7. # of transactions by day-part (breakfast, lunch, etc.)
8. % of transactions by day-part
9. Days since last transaction / Avg. days between transactions
10. …

1 IDENTIFY

2 COLLECT

3 ASSESS

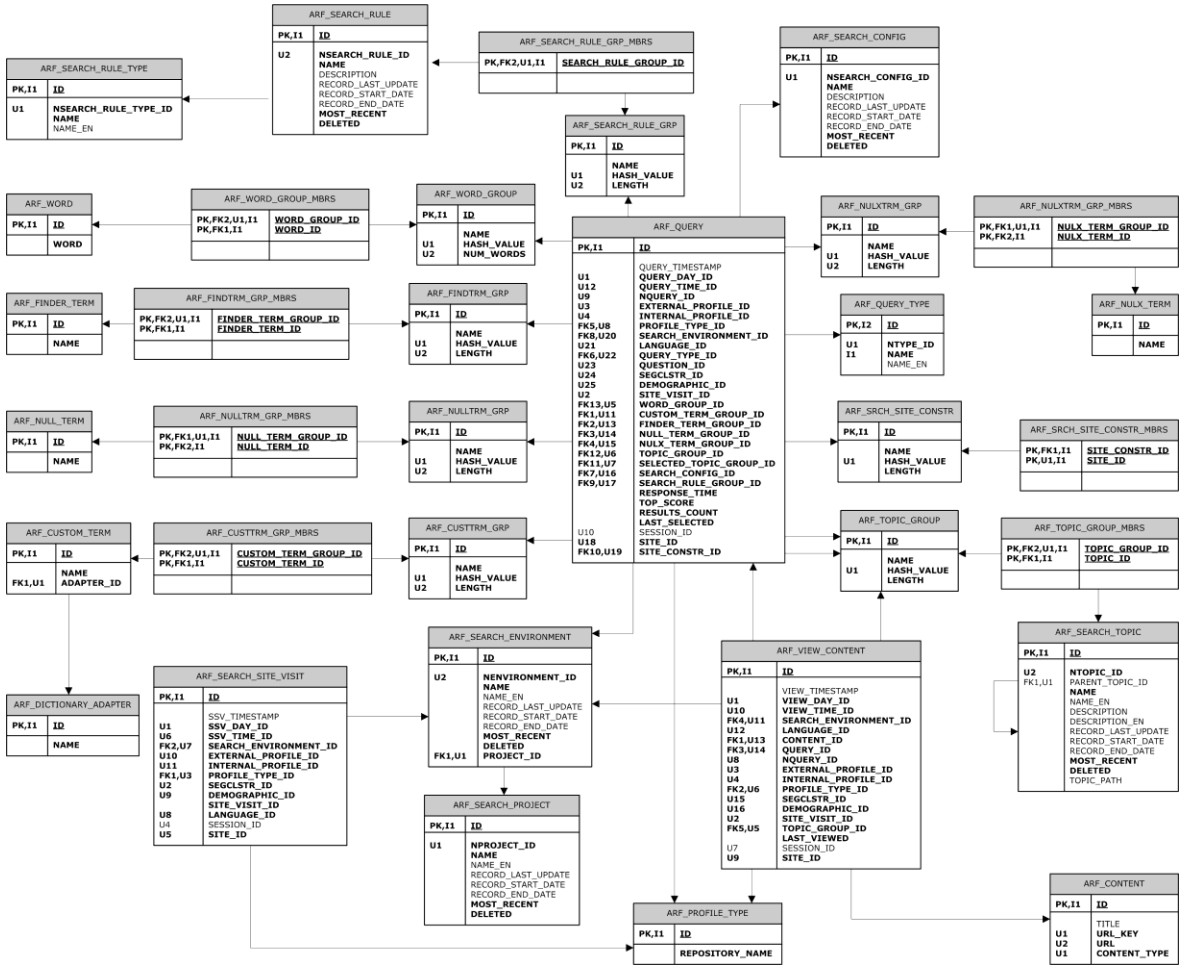
4 **VECTORIZE**

OUTPUT: The Analysis Dataset

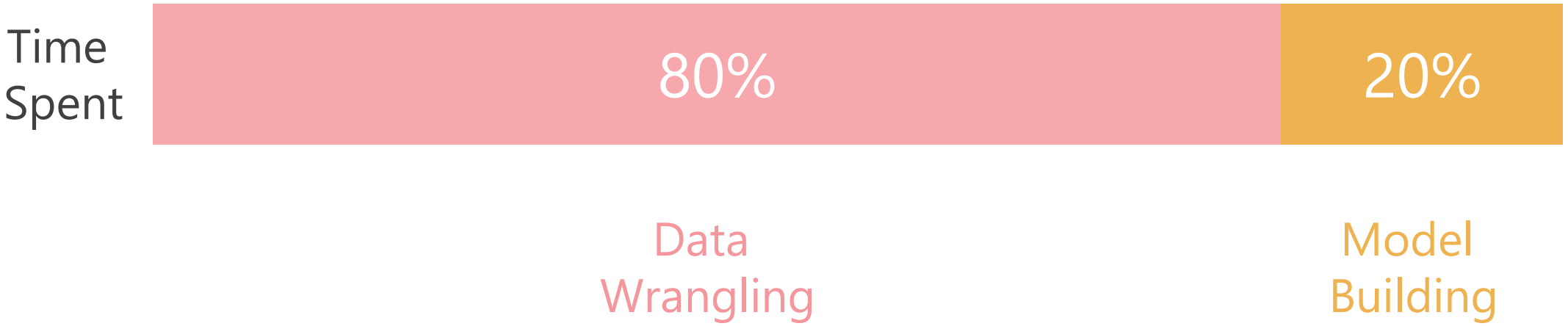
$$y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ . \\ y_n \end{pmatrix} \qquad X = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1j} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2j} \\ x_{31} & x_{32} & x_{33} & \dots & x_{3j} \\ . & . & . & & . \\ . & . & . & & . \\ . & . & . & & . \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{nj} \end{pmatrix}$$

Outcome
Target
Independent Variable

Inputs
Features / Attributes
Dependent Variables


ARF_SEARCH_RULE
PK,I1 ID
U2 NSEARCH_RULE_ID
NAME
DESCRIPTION
RECORD_LAST_UPDATE
RECORD_START_DATE
RECORD_END_DATE
MOST_RECENT
DELETED
ARF_SEARCH_RULE_GRP_MBRS
PK,FK2,U1,I1 SEARCH_RULE_GROUP_ID
ARF_SEARCH_RULE_TYPE
PK,I1 ID
U1 NSEARCH_RULE_TYPE_ID
NAME
NAME_EN
ARF_SEARCH_RULE_GRP
PK,I1 ID
U1 NAME
U2 HASH_VALUE
LENGTH
ARF_SEARCH_CONFIG
PK,I1 ID
U1 NSEARCH_CONFIG_ID
NAME
DESCRIPTION
RECORD_LAST_UPDATE
RECORD_START_DATE
RECORD_END_DATE
MOST_RECENT
DELETED
ARF_WORD
PK,I1 ID
WORD
ARF_WORD_GROUP_MBRS
PK,FK2,U1,I1 WORD_GROUP_ID
PK,FK1,I1 WORD_ID
ARF_WORD_GROUP
PK,I1 ID
U1 NAME
U2 HASH_VALUE
NUM_WORDS
ARF_QUERY
PK,I1 ID
U1 QUERY_TIMESTAMP
U12 QUERY_DAY_ID
U9 QUERY_TIME_ID
U3 NQUERY_ID
U4 EXTERNAL_PROFILE_ID
FK5,U8 INTERNAL_PROFILE_ID
FK8,U20 PROFILE_TYPE_ID
U21 SEARCH_ENVIRONMENT_ID
FK6,U22 LANGUAGE_ID
U23 QUERY_TYPE_ID
U24 QUESTION_ID
U25 SEGCLSTR_ID
U2 DEMOGRAPHIC_ID
FK13,U5 SITE_VISIT_ID
FK1,U11 WORD_GROUP_ID
FK2,U13 CUSTOM_TERM_GROUP_ID
FK3,U14 FINDER_TERM_GROUP_ID
FK4,U15 NULL_TERM_GROUP_ID
FK12,U6 NULX_TERM_GROUP_ID
FK11,U7 TOPIC_GROUP_ID
FK7,U16 SELECTED_TOPIC_GROUP_ID
FK9,U17 SEARCH_CONFIG_ID
SEARCH_RULE_GROUP_ID
RESPONSE_TIME
TOP_SCORE
RESULTS_COUNT
LAST_SELECTED
U10 SESSION_ID
U18 SITE_ID
FK10,U19 SITE_CONSTR_ID
ARF_NULXTRM_GRP
PK,I1 ID
U1 NAME
U2 HASH_VALUE
LENGTH
ARF_NULXTRM_GRP_MBRS
PK,FK1,U1,I1 NULX_TERM_GROUP_ID
PK,FK2,I1 NULX_TERM_ID
ARF_NULX_TERM
PK,I1 ID
NAME
ARF_FINDER_TERM
PK,I1 ID
NAME
ARF_FINDTRM_GRP_MBRS
PK,FK2,U1,I1 FINDER_TERM_GROUP_ID
PK,FK1,I1 FINDER_TERM_ID
ARF_FINDTRM_GRP
PK,I1 ID
U1 NAME
U2 HASH_VALUE
LENGTH
ARF_QUERY_TYPE
PK,I2 ID
U1 NTYPE_ID
I1 NAME
NAME_EN
ARF_NULL_TERM
PK,I1 ID
NAME
ARF_NULLTRM_GRP_MBRS
PK,FK1,U1,I1 NULL_TERM_GROUP_ID
PK,FK2,I1 NULL_TERM_ID
ARF_NULLTRM_GRP
PK,I1 ID
U1 NAME
U2 HASH_VALUE
LENGTH
ARF_SRCH_SITE_CONSTR
PK,I1 ID
U1 NAME
HASH_VALUE
LENGTH
ARF_SRCH_SITE_CONSTR_MBRS
PK,FK1,I1 SITE_CONSTR_ID
PK,U1,I1 SITE_ID
ARF_CUSTOM_TERM
PK,I1 ID
FK1,U1 NAME
ADAPTER_ID
ARF_CUSTTRM_GRP_MBRS
PK,FK2,U1,I1 CUSTOM_TERM_GROUP_ID
PK,FK1,I1 CUSTOM_TERM_ID
ARF_CUSTTRM_GRP
PK,I1 ID
U1 NAME
U2 HASH_VALUE
LENGTH
ARF_TOPIC_GROUP
PK,I1 ID
U1 NAME
HASH_VALUE
LENGTH
ARF_TOPIC_GROUP_MBRS
PK,FK2,U1,I1 TOPIC_GROUP_ID
PK,FK1,I1 TOPIC_ID
ARF_DICTIONARY_ADAPTER
PK,I1 ID
NAME
ARF_SEARCH_ENVIRONMENT
PK,I1 ID
U2 NENVIRONMENT_ID
NAME
NAME_EN
RECORD_LAST_UPDATE
RECORD_START_DATE
RECORD_END_DATE
MOST_RECENT
DELETED
FK1,U1 PROJECT_ID
ARF_SEARCH_SITE_VISIT
PK,I1 ID
U1 SSV_TIMESTAMP
U6 SSV_DAY_ID
SSV_TIME_ID
FK2,U7 SEARCH_ENVIRONMENT_ID
U10 EXTERNAL_PROFILE_ID
U11 INTERNAL_PROFILE_ID
FK1,U3 PROFILE_TYPE_ID
U2 SEGCLSTR_ID
U9 DEMOGRAPHIC_ID
SITE_VISIT_ID
U8 LANGUAGE_ID
U4 SESSION_ID
U5 SITE_ID
ARF_VIEW_CONTENT
PK,I1 ID
U1 VIEW_TIMESTAMP
U10 VIEW_DAY_ID
FK4,U11 VIEW_TIME_ID
U12 SEARCH_ENVIRONMENT_ID
FK1,U13 LANGUAGE_ID
FK3,U14 CONTENT_ID
U8 QUERY_ID
U3 NQUERY_ID
U4 EXTERNAL_PROFILE_ID
FK2,U6 INTERNAL_PROFILE_ID
U15 PROFILE_TYPE_ID
U16 SEGCLSTR_ID
U2 DEMOGRAPHIC_ID
FK5,U5 SITE_VISIT_ID
TOPIC_GROUP_ID
LAST_VIEWED
U7 SESSION_ID
U9 SITE_ID
ARF_SEARCH_TOPIC
PK,I1 ID
U2 NTOPIC_ID
FK1,U1 PARENT_TOPIC_ID
NAME
NAME_EN
DESCRIPTION
DESCRIPTION_EN
RECORD_LAST_UPDATE
RECORD_START_DATE
RECORD_END_DATE
MOST_RECENT
DELETED
TOPIC_PATH
ARF_SEARCH_PROJECT
PK,I1 ID
U1 NPROJECT_ID
NAME
NAME_EN
RECORD_LAST_UPDATE
RECORD_START_DATE
RECORD_END_DATE
MOST_RECENT
DELETED
ARF_PROFILE_TYPE
PK,I1 ID
REPOSITORY_NAME
ARF_CONTENT
PK,I1 ID
TITLE
U1 URL_KEY
U2 URL
U1 CONTENT_TYPE

Business Understanding | Data Preparation | Data Wrangling | Model Training | Model Evaluation | Model Deployment | Model Tracking



Give me six hours to chop down a tree

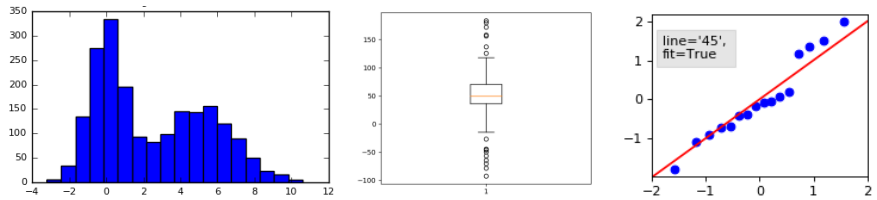
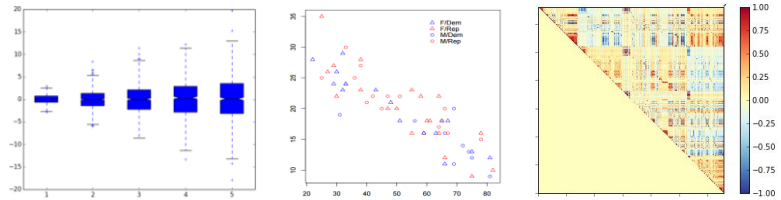
and I will spend the first four sharpening the axe.

– Abraham Lincoln (?)

- **Descriptive statistics**
  - Review with the client
- **Correlation analysis**
  - Review with the client
  - Watch out for data leakage
- **Missing value imputation**
- **Trim extreme values**
- **Process categorical attributes**
- **Transformations** (square, log, etc.)
  - Binning / variable smoothing
- **Multicollinearity**
  - Reduce redundancy
- **Additional feature** (derived variables)
- **Interactions**
- **Normalization** (scaling)

	Univariate	Multivariate

Non-Graphical	○ Categorical: Tabulated frequencies
	○ Quantitative:
	  ○ Central tendency: mean, median, mode
	  ○ Spread: Standard deviation, inter-quartile range
	  ○ Skewness and kurtosis
		○ Cross-tabulation
		○ Univariate statistics by category
		○ Correlation matrices

Graphical	○ Histograms ○ Box plots, stem-and-leaf plots ○ Quantile-normal plots 	○ Univariate graphs by category (e.g., side-by-side box-plots) ○ Scatterplots ○ Correlation matrix plots 

- **Feature Reduction:** The process of selecting a subset of features for use in model construction
  - Useful for both supervised and unsupervised learning problems

Art is the elimination of the unnecessary.

– Pablo Picasso

Feature Reduction: Why

- **True dimensionality <<< Observed dimensionality**
  - The abundance of redundant and irrelevant features
- **Curse of dimensionality**
  - With a fixed number of training samples, the predictive power reduces as the dimensionality increases. [Hughes phenomenon]
  - With d binary variables, the number of possible combinations is $O(2^d)$.
- **Goal of the Analysis**
  - Descriptive → Diagnostic → Predictive → Prescriptive

- **Law of Parsimony** [Occam's Razor]
  - Other things being equal, simpler explanations are generally better than complex ones.
- **Overfitting**
- **Execution time** (Algorithm and data processing)

Feature Reduction Techniques

1. Percent missing values
2. Amount of variation
3. Pairwise correlation
4. Multicolinearity
5. Principal Component Analysis (PCA)
6. Cluster analysis
7. Correlation (with the target)
8. Forward selection
9. Backward elimination
10. Stepwise selection
11. LASSO
12. Tree-based selection

- Try **more than one machine learning technique**
- Fine-tune **parameters** and **hyper-parameters**
- Assess **model performance**
- Avoid **Over-fitting**

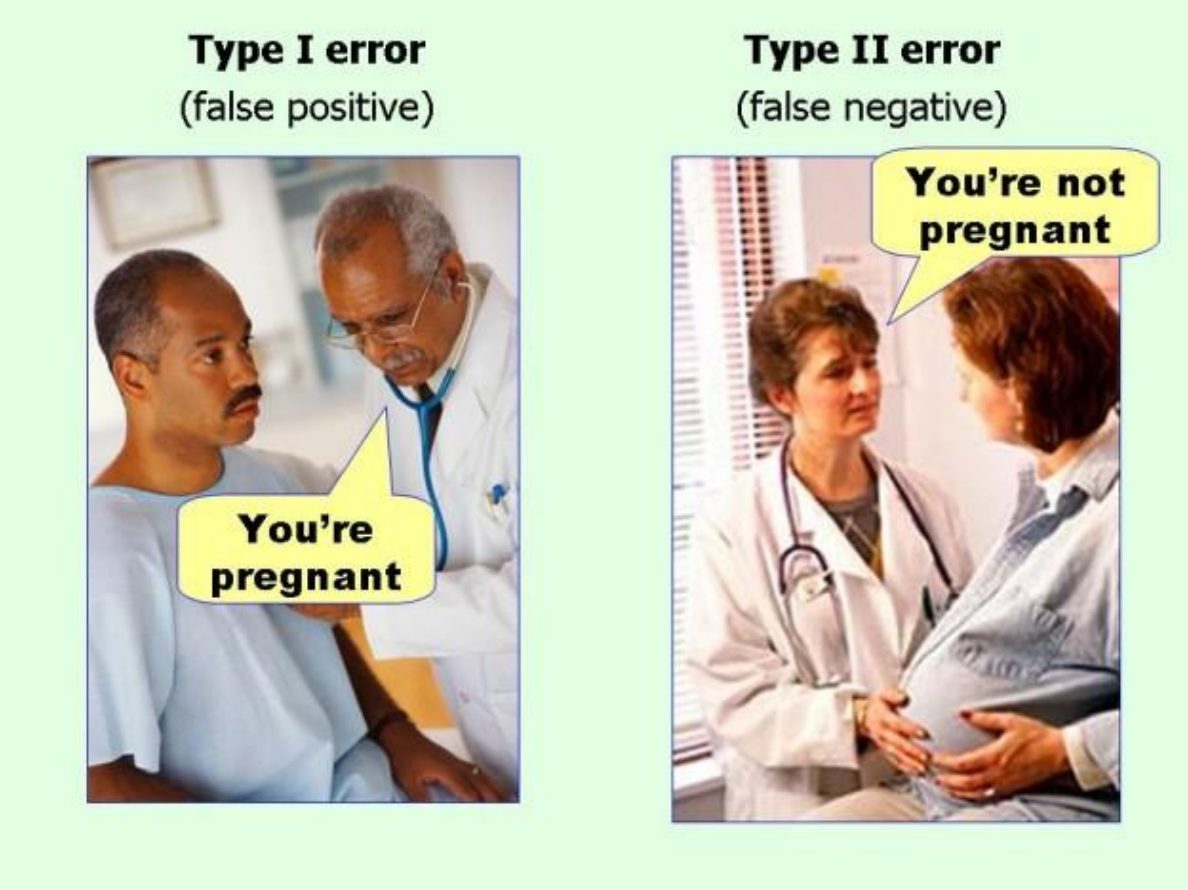
SPSS

sas

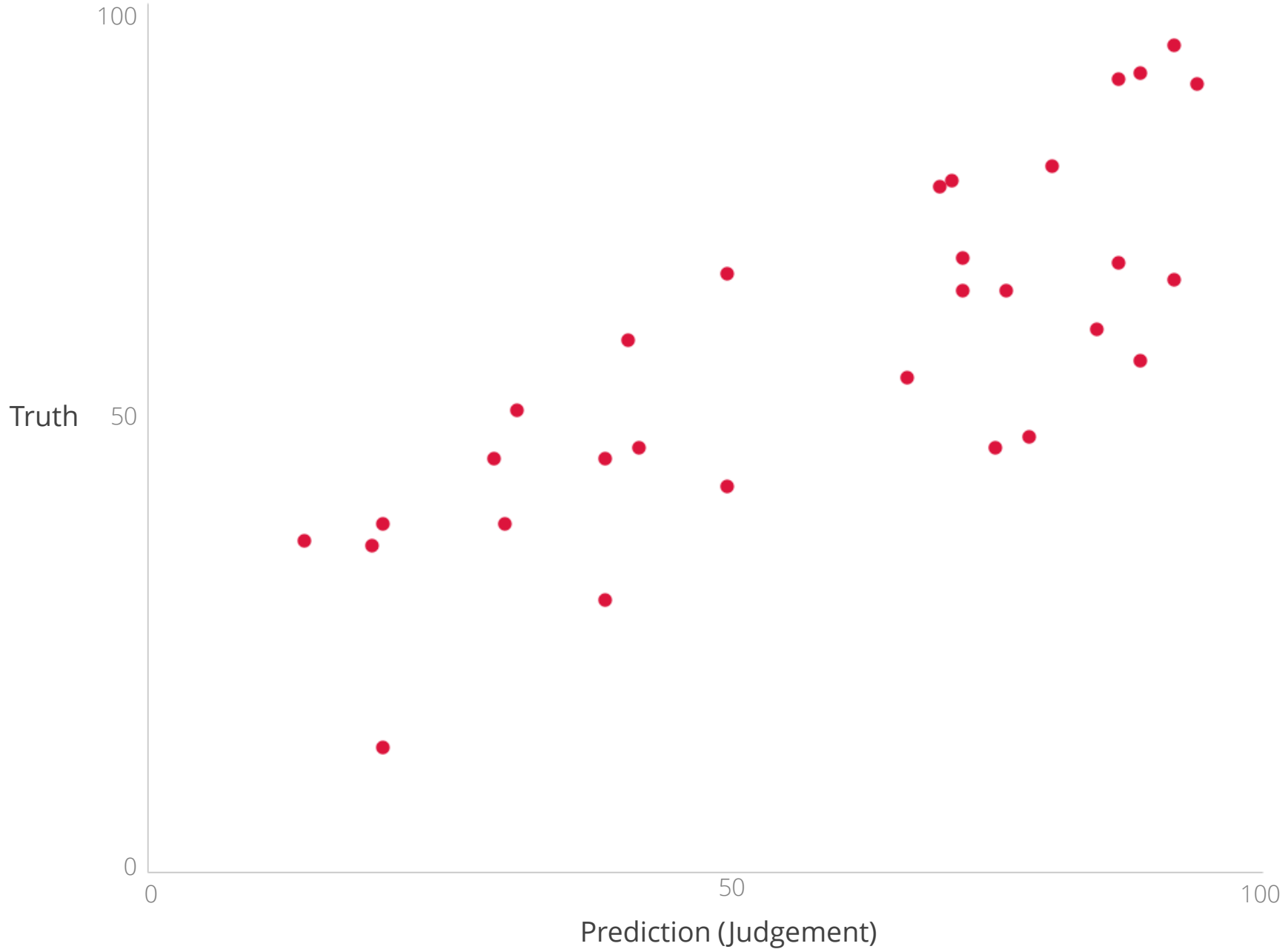
R

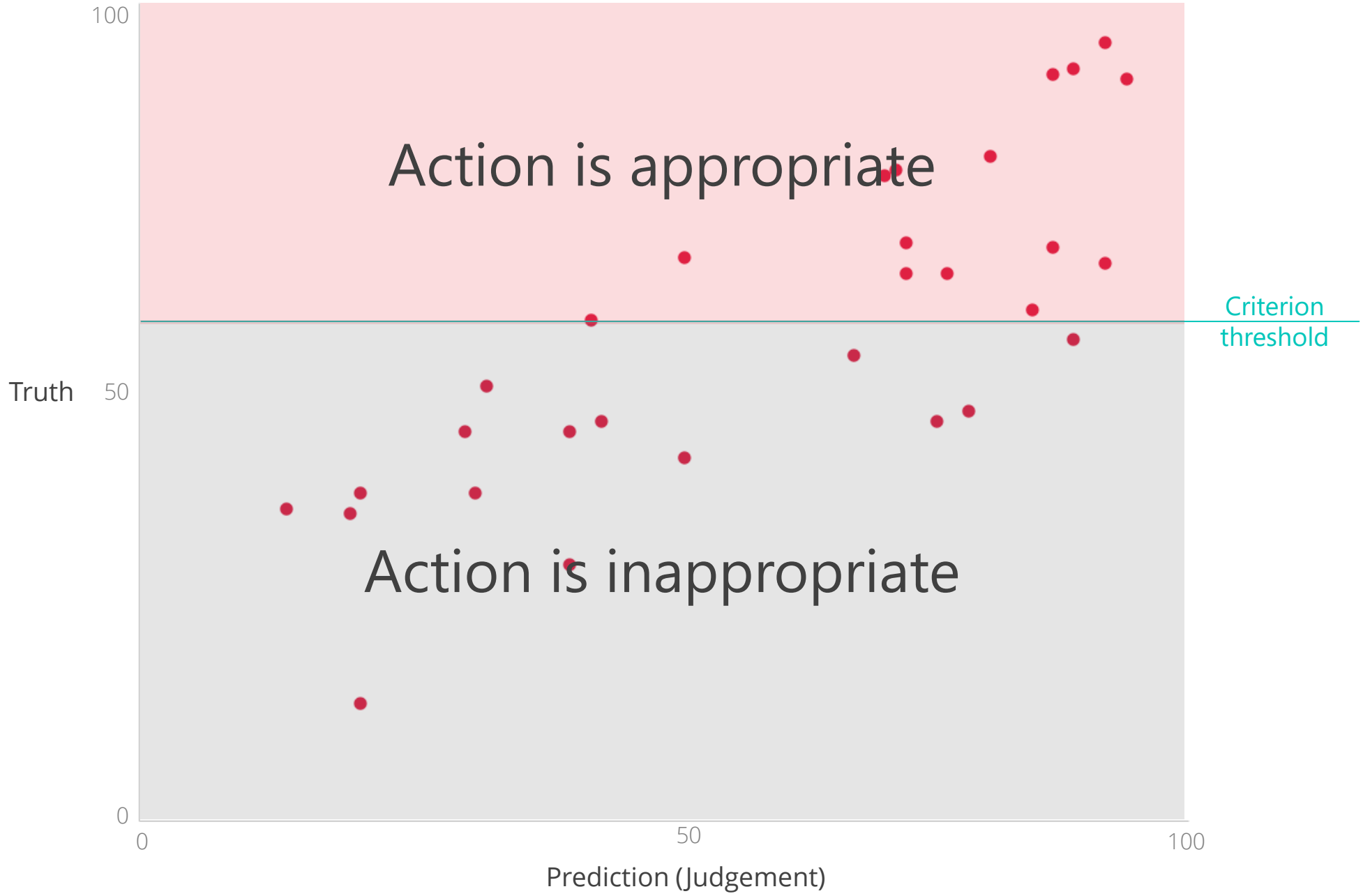
Azure ML

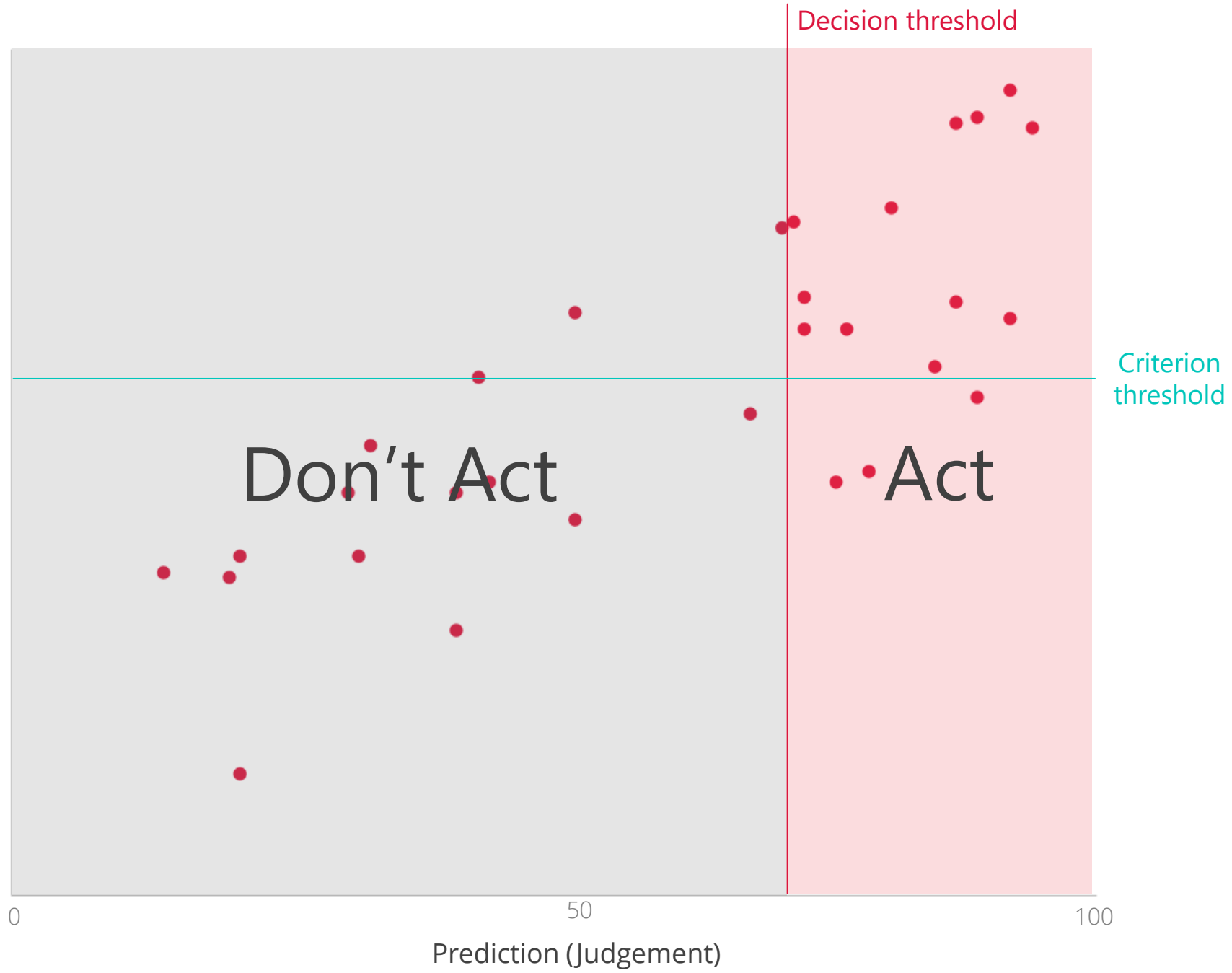
Assess Model Performance

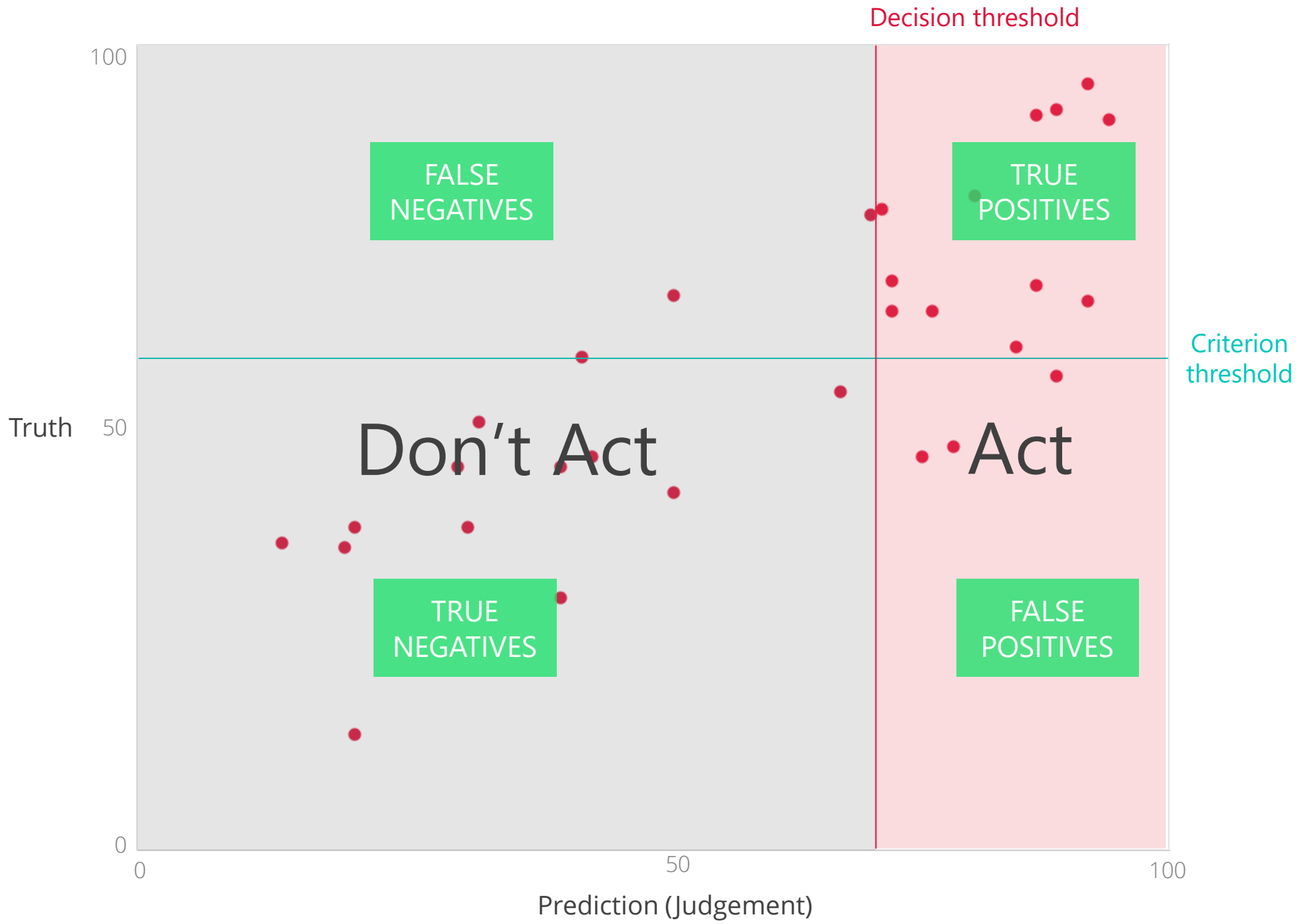


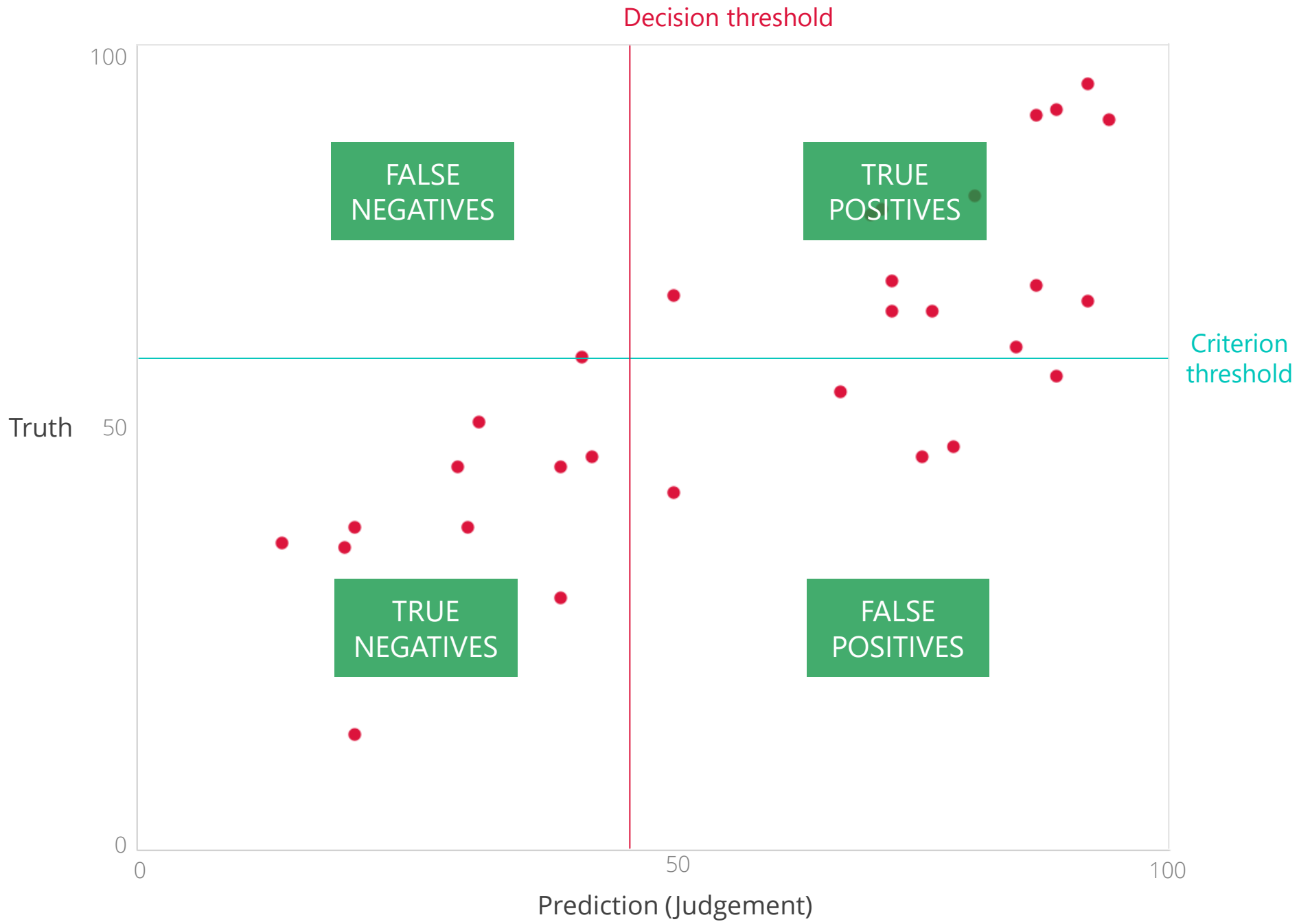
- Area Under the ROC Curve (AUC), Confusion Matrix, Precision, Recall, Log-loss
- Model Lift, Model Gains, Kolmogorov-Smirnov (KS), etc.

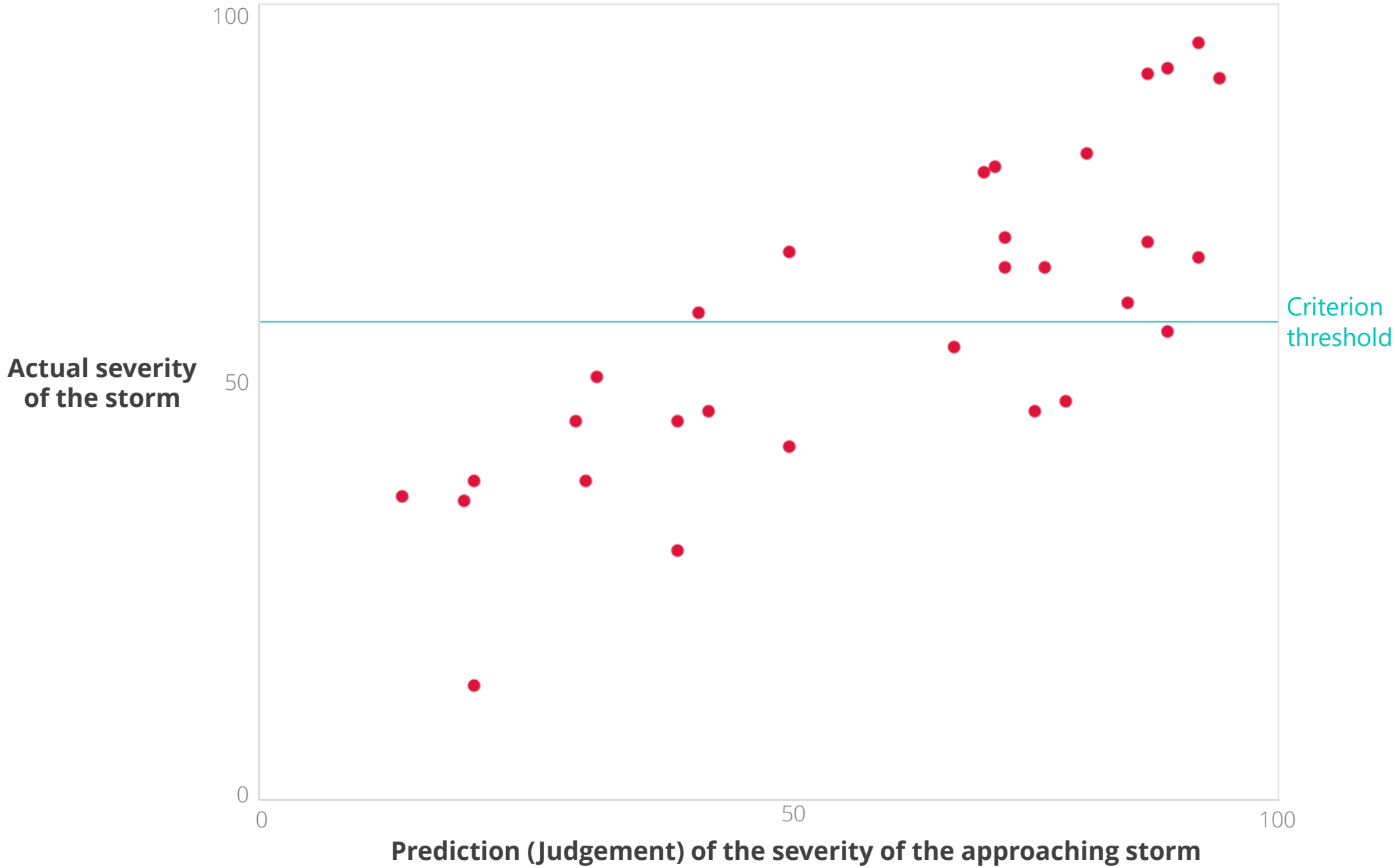

100
Truth
50
0
0
50
100
Prediction (Judgement)

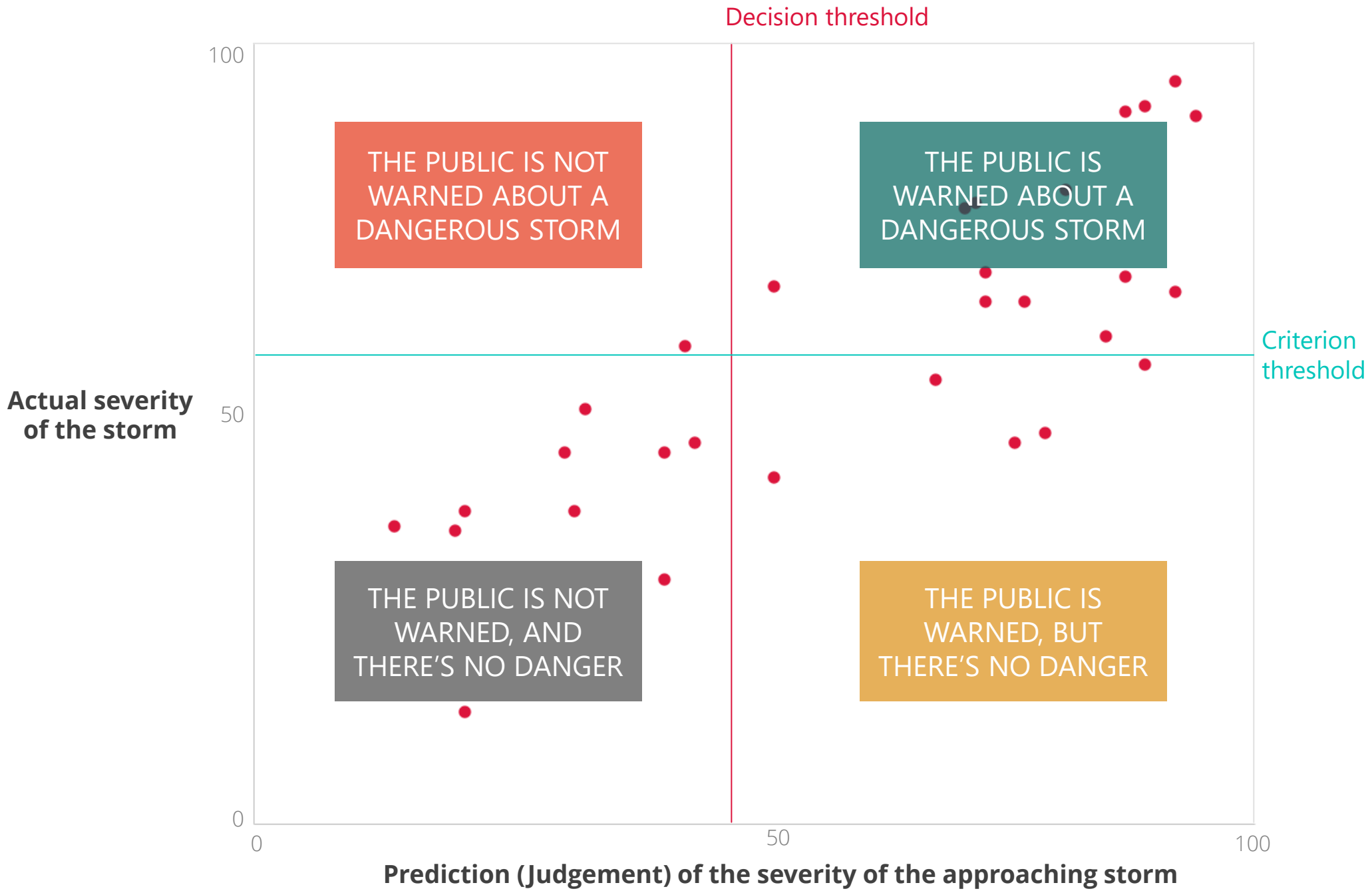

100
Action is appropriate
Criterion threshold
Truth
50
Action is inappropriate
0
0
50
100
Prediction (Judgement)


Decision threshold
Criterion threshold
Don't Act
Act
0
50
100
Prediction (Judgement)


Decision threshold
100
FALSE NEGATIVES
TRUE POSITIVES
Criterion threshold
Truth
50
Don't Act
Act
TRUE NEGATIVES
FALSE POSITIVES
0
0
50
100
Prediction (Judgement)


Decision threshold
100
FALSE NEGATIVES
TRUE POSITIVES
Criterion threshold
Truth
50
TRUE NEGATIVES
FALSE POSITIVES
0
0
50
100
Prediction (Judgement)


100
50
0
0
50
100
Actual severity of the storm
Criterion threshold
Prediction (Judgement) of the severity of the approaching storm

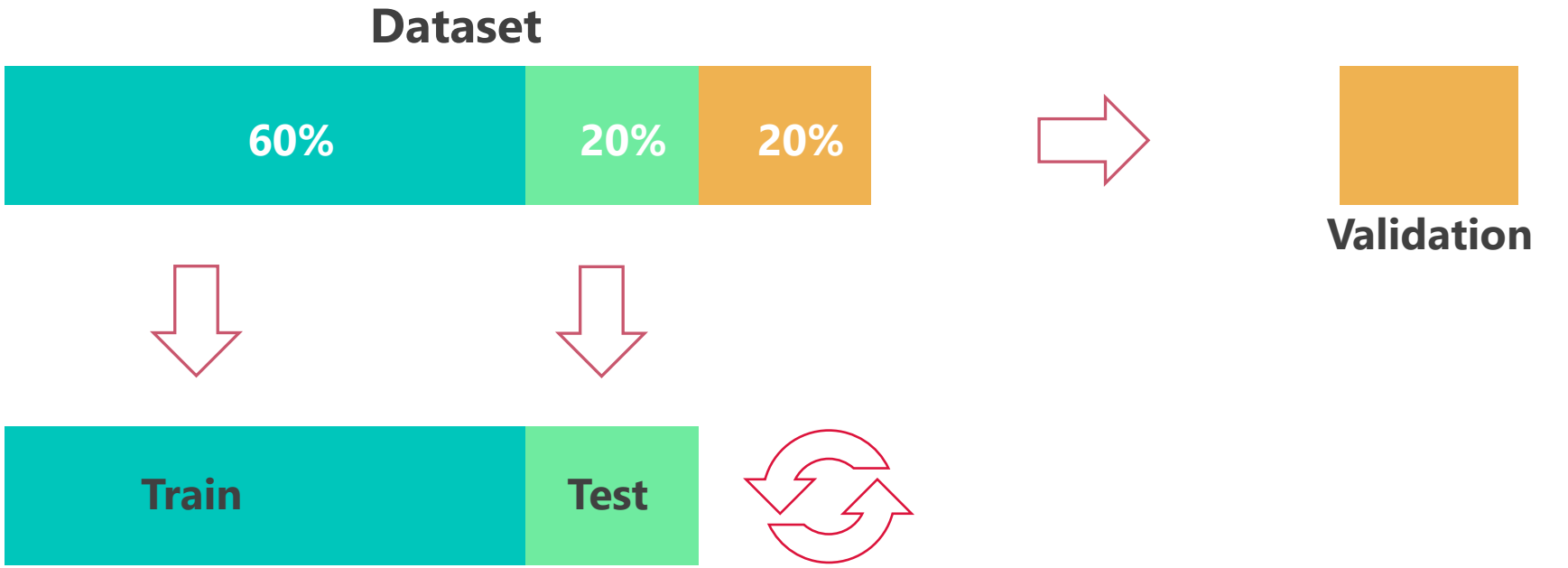

Decision threshold
100
50
0
0
50
100
Criterion threshold
Actual severity of the storm
THE PUBLIC IS NOT WARNED ABOUT A DANGEROUS STORM
THE PUBLIC IS WARNED ABOUT A DANGEROUS STORM
THE PUBLIC IS NOT WARNED, AND THERE'S NO DANGER
THE PUBLIC IS WARNED, BUT THERE'S NO DANGER
Prediction (Judgement) of the severity of the approaching storm

When a measure becomes a target,

it ceases to be a good measure.

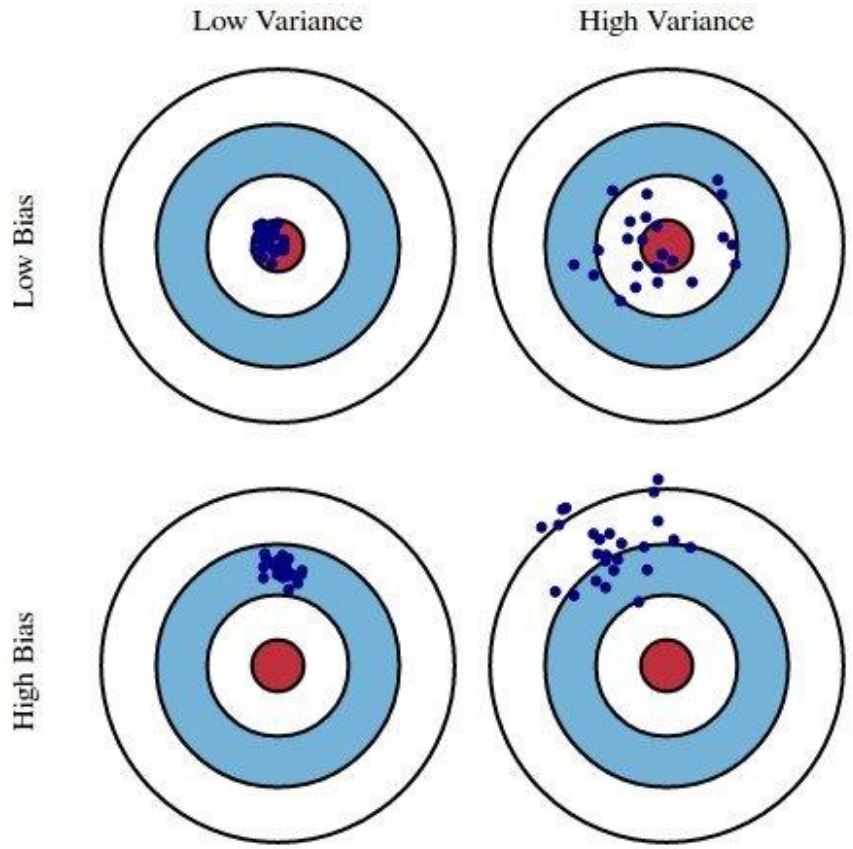
Goodhart's law

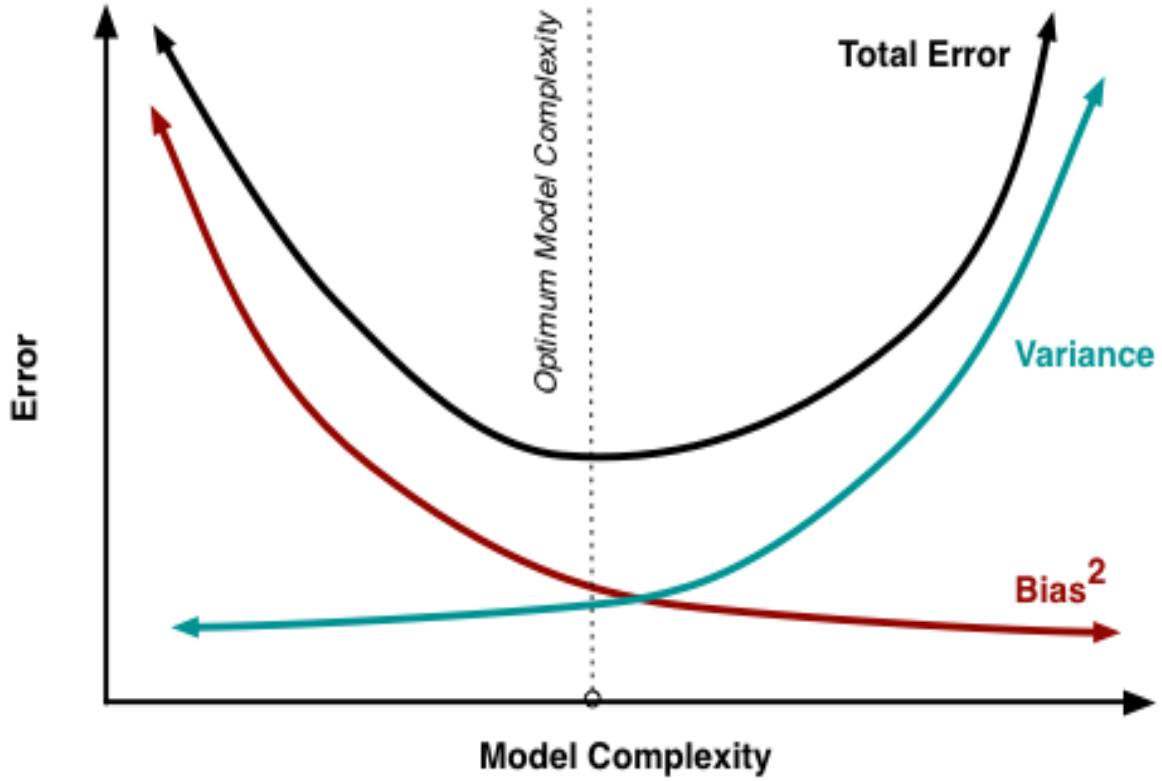
Tri-fold Partition



- **Fine-tune** and **select** the best model based on Train + Test sets.
- **Evaluate** the chosen algorithm on the Validation set (previously unseen data).

Bias-Variance Tradeoff





With four parameters I can fit an elephant,

and with five I can make him wiggle his trunk.

– John von Neumann

1 MODEL SELECTION

2 ASSESSMENT

3 PRESENTATION

1 **MODEL SELECTION**

- **Law of Parsimony** (Occam's Razor)
- **Model execution time**
- **Deployment complexity**

2 ASSESSMENT

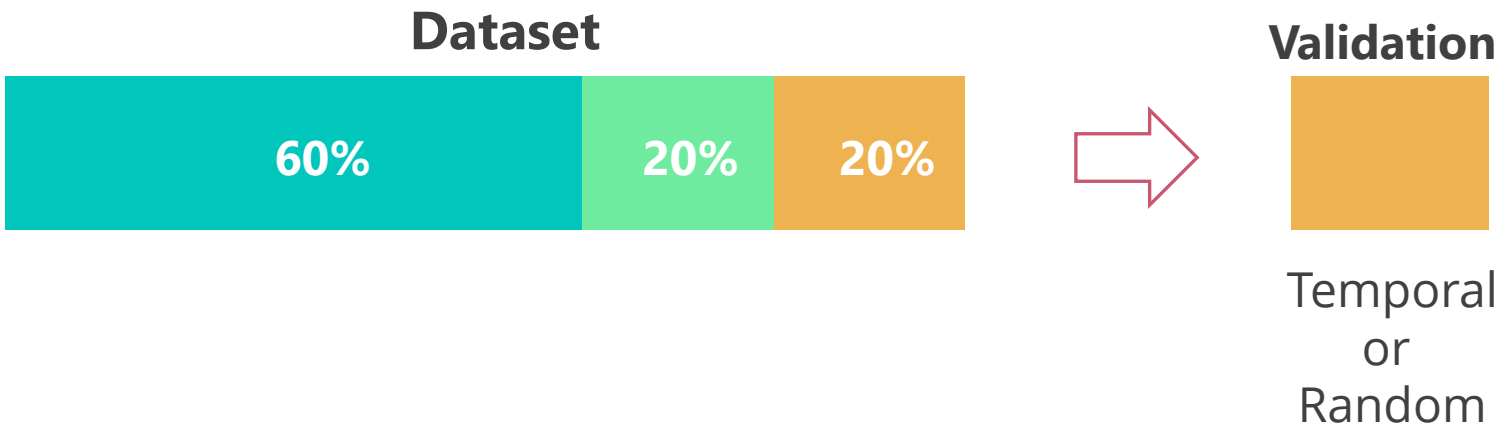
Build the simplest solution that can adequately answer the question.

3 PRESENTATION

1 MODEL SELECTION

2 ASSESSMENT

3 PRESENTATION



1 MODEL SELECTION

2 ASSESSMENT

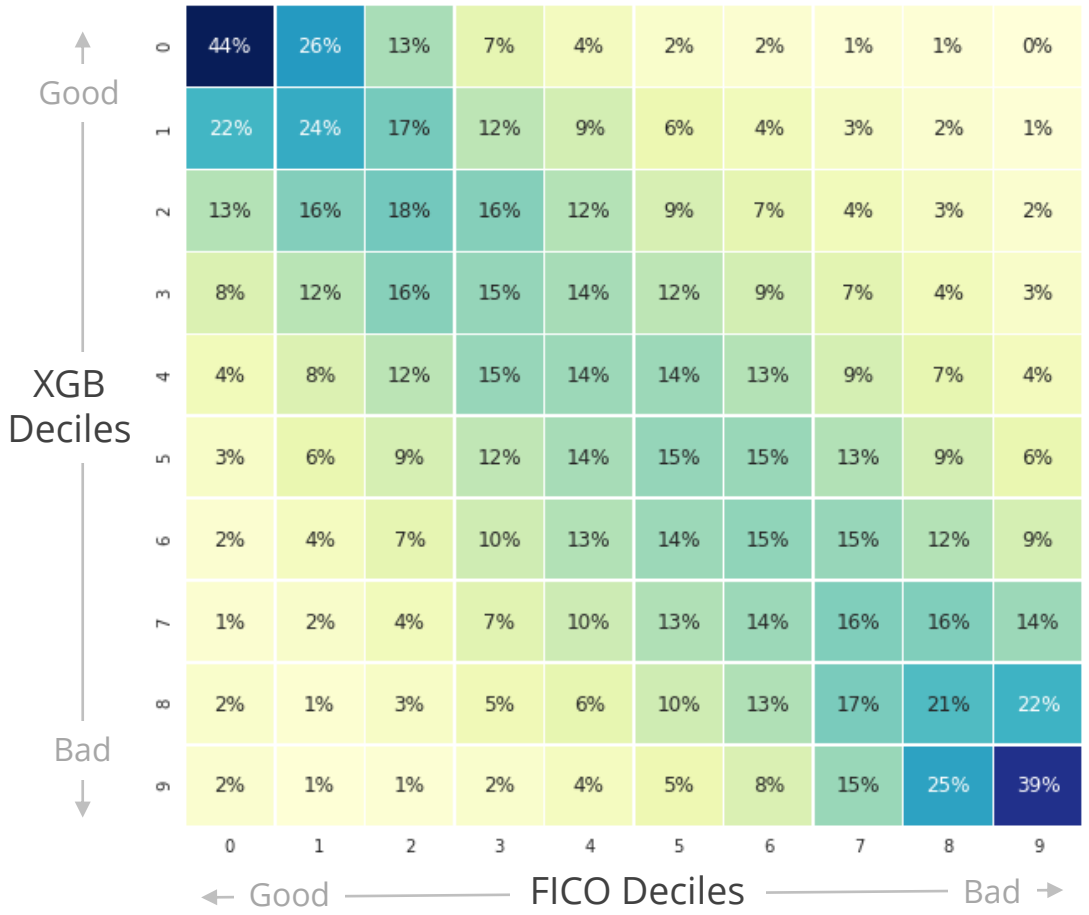
3 **PRESENTATION**

- **AUC, Somer's D, etc.**
- **Cumulative Gains Chart / Lift Chart**
- **Predictor Importance**
- **Each predictor's relationship with the target**
- **Model usage recommendations**
  - Decile reports
  - How many deciles should be targeted?
- **Personify**
- **Compare against existing business rules/model**
- **Model peer-review** (Quality Control)

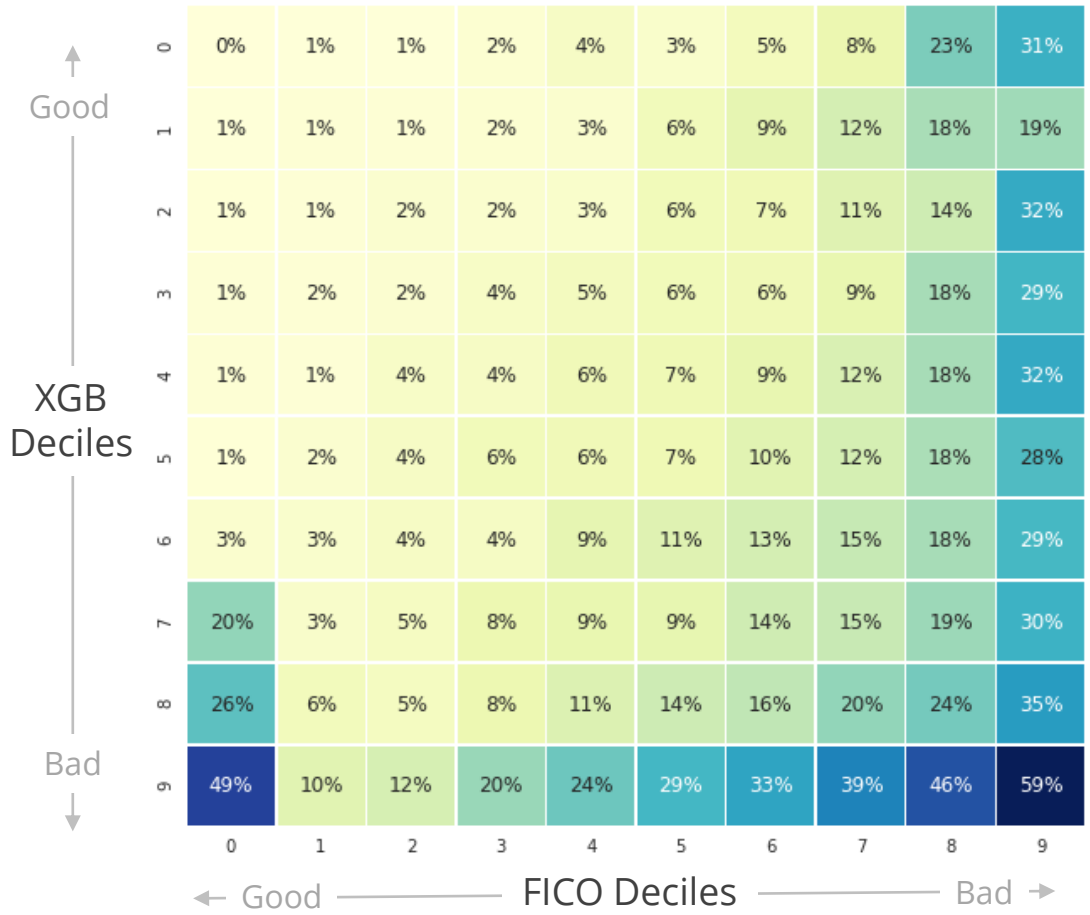
Interpret results as they relate to the business application.

Model Comparisons

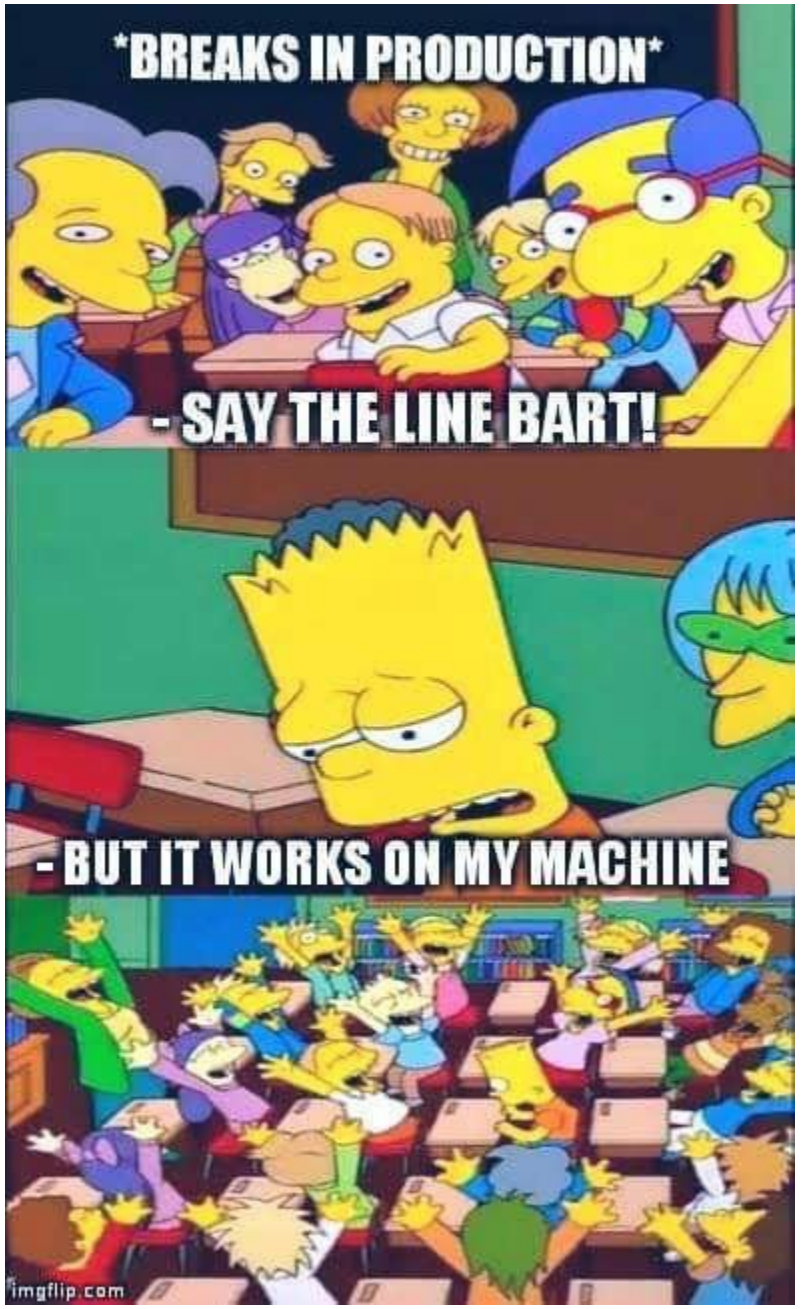
Size (column %s)



Bad Rate



- **Model production cycle**
  - Weekly, monthly, live?
- **Scoring code, or publish model as a web service**
- **Model Documentation** (Technical Specifications)
  - Data preparation, transformations, imputations, parameter settings, etc.
- **Reproducibility**
  - `requirements.txt`, Docker containers


BREAKS IN PRODUCTION
- SAY THE LINE BART!
- BUT IT WORKS ON MY MACHINE
imgflip.com

1 **MONITOR**

2 **MAINTAIN**

3 **TEST**

1 **MONITOR**

2 **MAINTAIN**

3 **TEST**

- **Model decay tracking (monitoring) plan**
  - Model performance over time
  - Predictor distribution
  - Probability Stability Index (PSI)

1 MONITOR

2 MAINTAIN

- **Model maintenance plan**
- **Plan for adding new data sources**
- **Version control**
  - GitHub, DVC

3 TEST

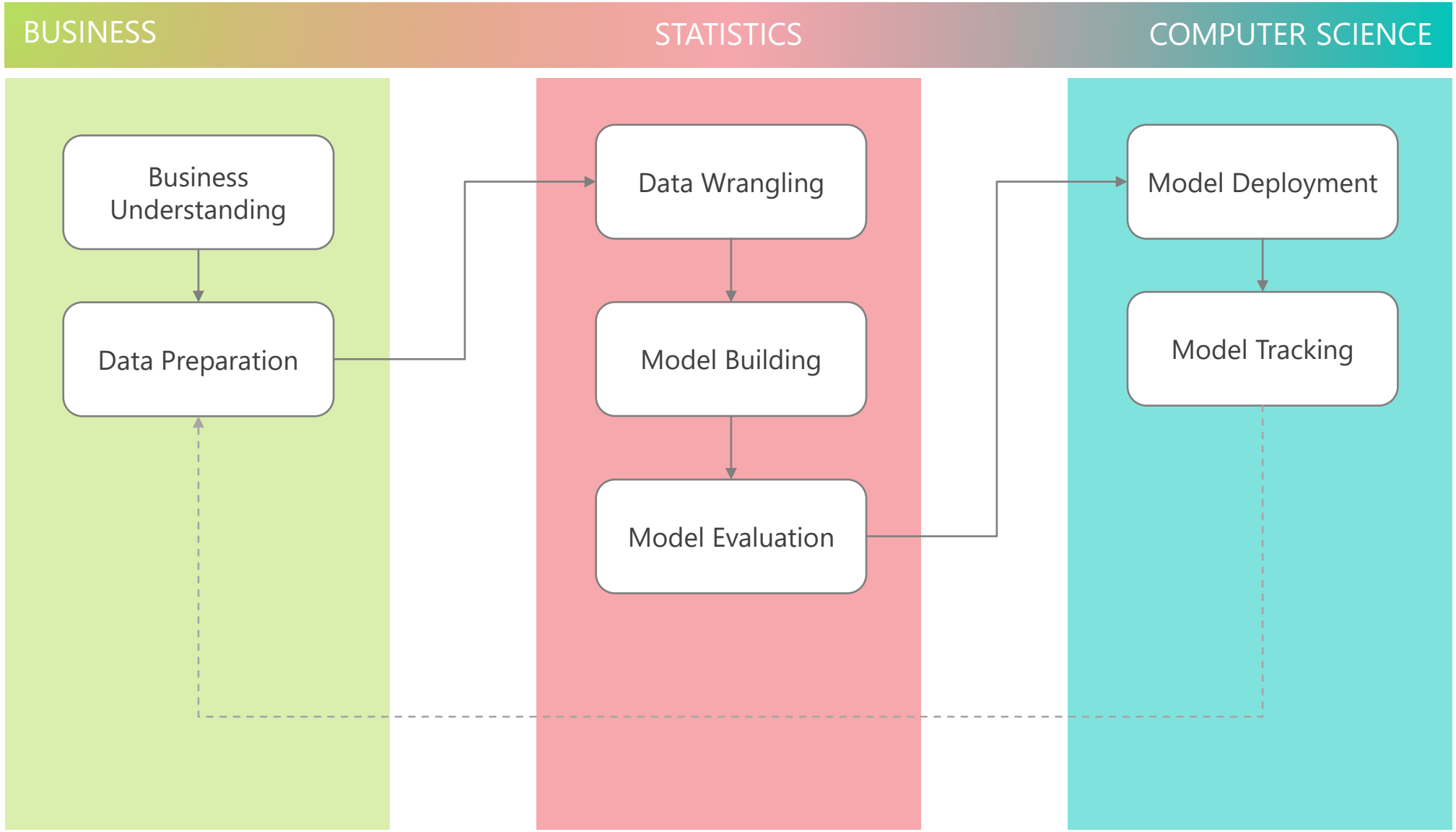
1 MONITOR

2 MAINTAIN

3 **TEST**

- **Campaign Set-up and Execution**
  - Experimental Design (A/B, Fractional Factorial)

Data Science Process: Recap

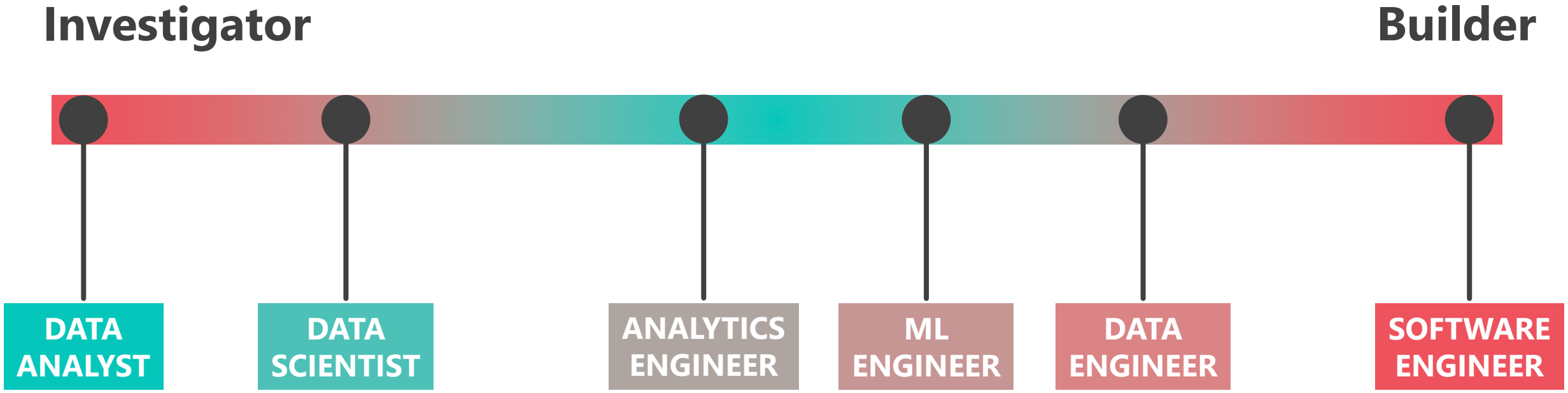


Data Science Process: Recap

Business Understanding	Data Preparation	Data Wrangling	Model Training	Model Evaluation	Model Deployment	Model Tracking
Determine	Identify	Impute	Train	Evaluate	Document	Monitor
Understand	Collect	Transform	Assess	Peer Review	Deploy	Maintain
Map	Assess	Reduce	Select	Present		Test
	Vectorize					

DISCUSS **COLLATE** **WRANGLE** **PERFORM** **COMMUNICATE** **EXECUTE** **TRACK**

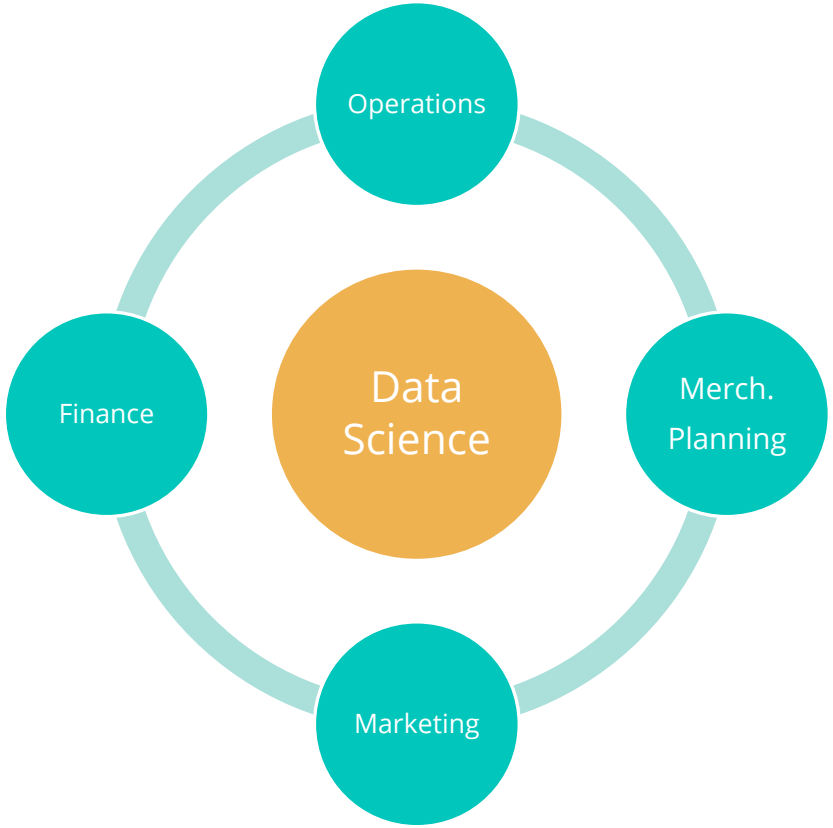
Data Roles


Investigator
Builder
DATA ANALYST
DATA SCIENTIST
ANALYTICS ENGINEER
ML ENGINEER
DATA ENGINEER
SOFTWARE ENGINEER

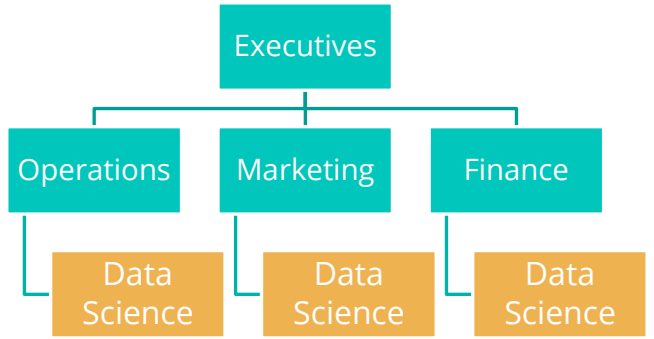
Data Science Team Structures

CENTRALIZED

HUB & SPOKE



DECENTRALIZED



DAPT 631

Assignment #1 (Individual Assignment)

Due by 13-FEB-2024

1. From the following five aspects of a Data Science project, which one do you think is the most important? Please provide a rationale for your answer.
    a. Understanding the business question,
    b. Quality of data,
    c. Statistical rigor and accuracy,
    d. Using the correct tools/techniques, and
    e. Communicating results back to stakeholders.
2. From the following aspects of Data Science, which one are you most interested in? Please elaborate your answer.
    a. Business Requirements,
    b. Exploratory Data analysis (EDA),
    c. Feature Engineering and data pipelines,
    d. Model Building,
    e. Putting the model into production (automation),
    f. Explain model results to non-technical audience,
    g. Develop and implement strategies based on model results,
    h. Other.

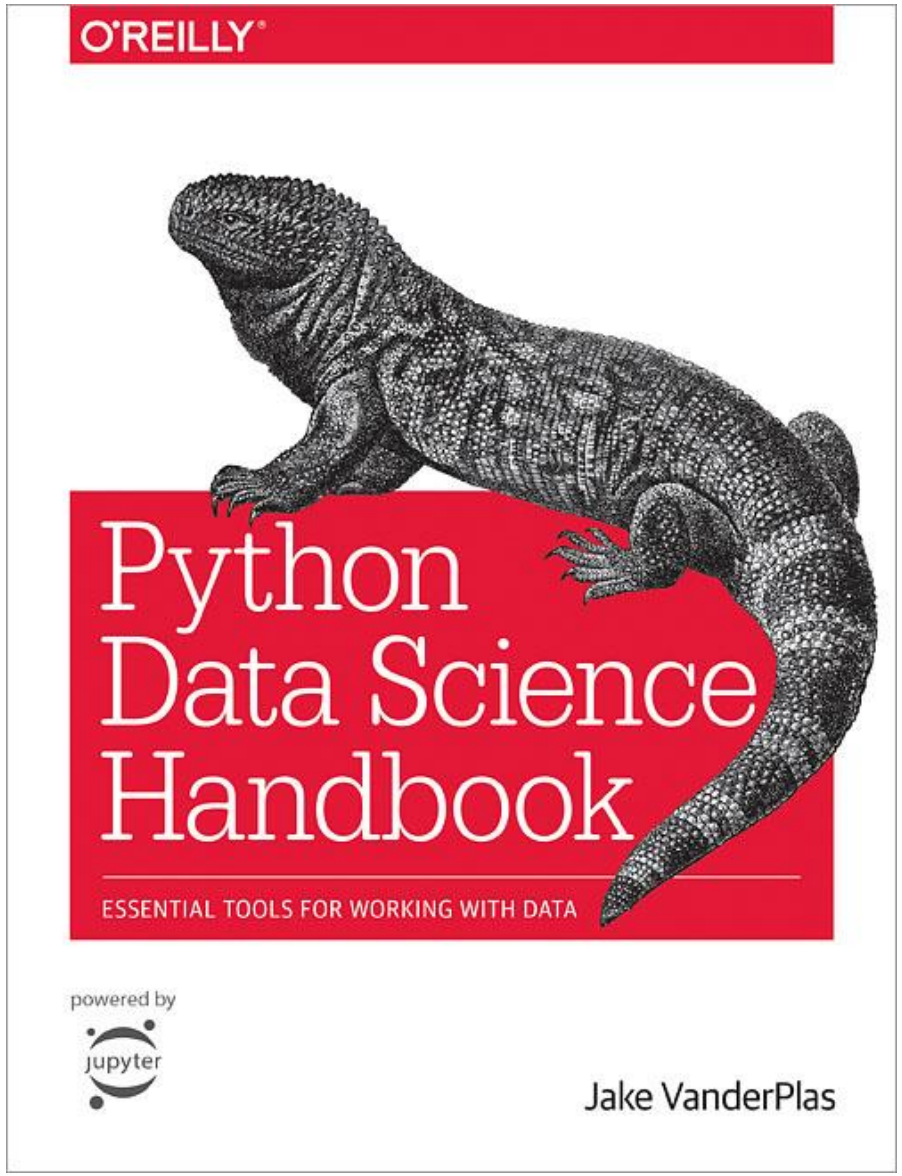
This is an individual assignment, but students are encouraged to discuss these items with each other.

Next Up

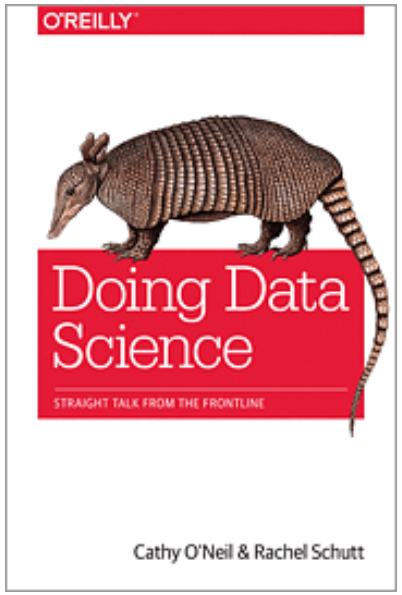
1. ~~Introduction~~
2. ~~The Data Science Process~~
3. **Supervised Learning**
4. Unsupervised Learning
5. The Grunt Work
6. Wrap Up

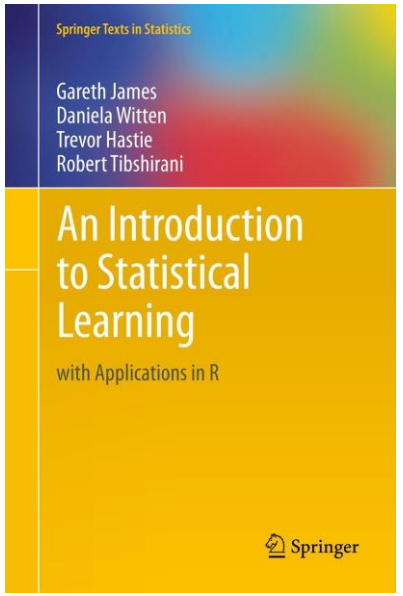
Data Mining Algorithms

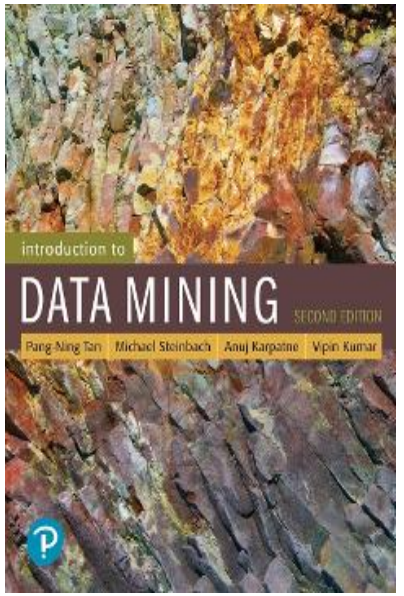
Linear Regression → Decision Trees → Random Forests → Gradient Boosting → …



Chapter 5: Machine Learning







- Chapter 3. **Classification**: *Basic Concepts and Techniques*
- Chapter 5. **Association Analysis**: *Basic Concepts and Algorithms*
- Chapter 7. **Cluster Analysis**: *Basic Concepts and Algorithms*

LEARNDATASCI

100+ Free Data Science Books