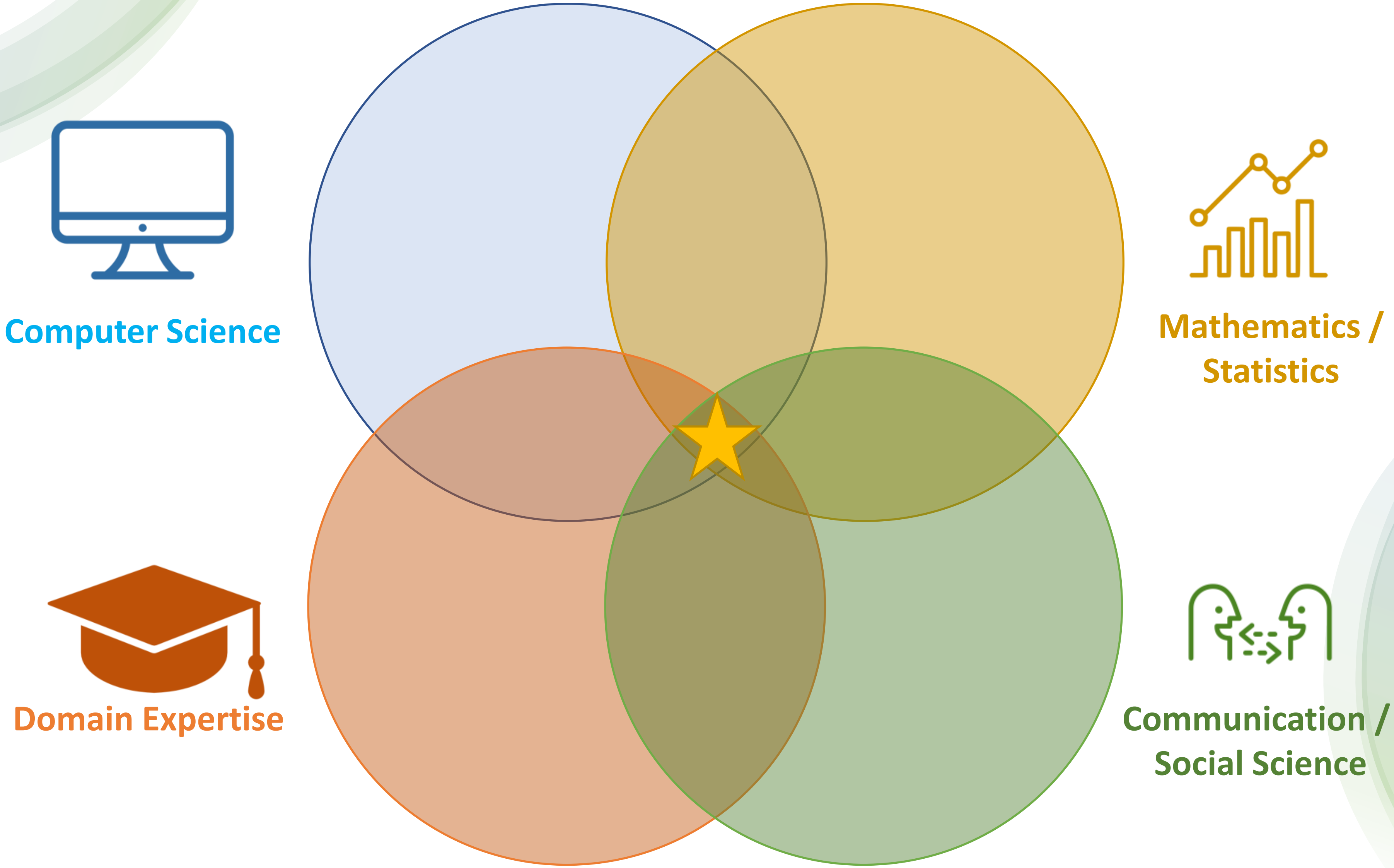
Statistical Data Science

Thammakorn Saethang

Data Science

- **Data Science is interdisciplinary, covering computer science, machine learning, mathematics, statistics, and domain knowledge from application areas.**
- **Statistics plays important roles in data science.**
- **This course explores the nature of the relationship between statistics and data science, suggesting state-of-the-art reasoning from both areas, and developing a synergistic path forward.**

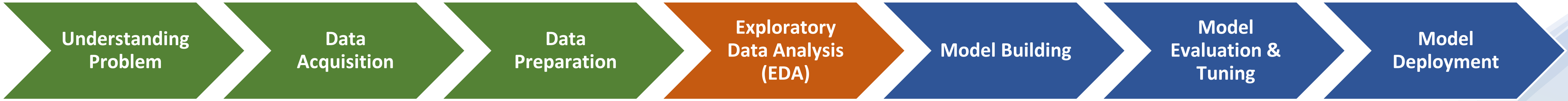
What is Data Science?



Core Processes in Data Science

- Statistical Analyses
- Data Visualization
- Data Summarization

- Cross-Validation
- Parameter Tuning
- Feature Selection
- Refining Algorithm

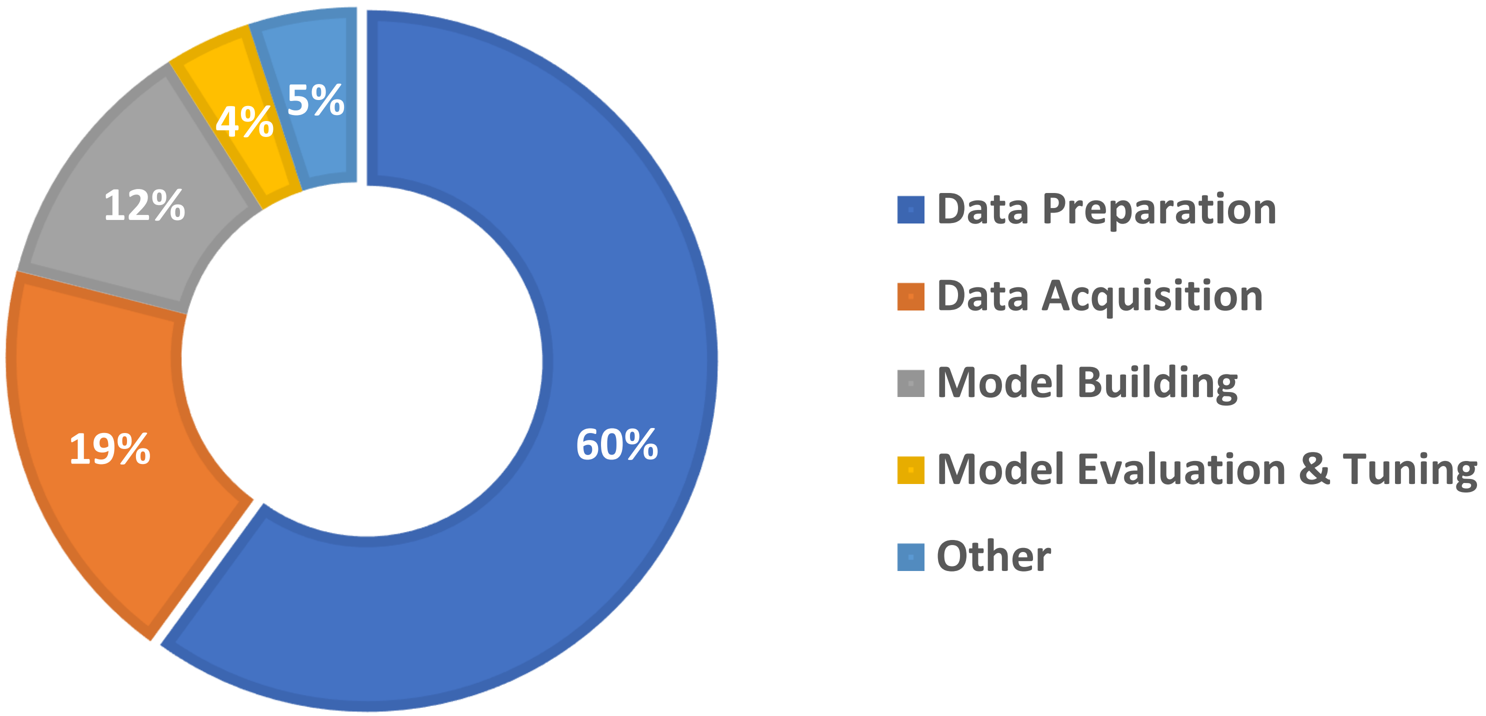


- Data Cleaning
- Data Organizing
- Data Transformation

- Algorithm Selection
- Identify Data Pattern
- Model Training

- Designing Pipelines
- Updating Model with new data

How do data scientists spend their time?



Data Preparation for Machine Learning (ML)

Data preparation process:

- Cleaning
- Organizing
- Removing outliers
- Denoising
- Validation
- Standardization
- Transformation*
- etc.

Importance of data preparation:

- ML algorithms require specific data format*
- Garbage in, garbage out (GIGO)
- Modern ML models requires data normalization
- etc.

Statistical Data Science

- Statistical DS = DS with modern statistics
  - Statistics is the study of the collection, analysis, interpretation, presentation of data.
  - Modern statistics / statistical learning = statistics + modeling and understanding complex datasets.
    - It is a recently developed area in statistics and blends with parallel developments in computer science and, in particular, machine learning.

Statistical Data Science Applications

Spotify – Music Recommendation & Personalized Playlists

- Analyzes audio signals and song features using statistical and ML methods.
- Uses clustering to group users with similar music tastes.“
- Discover Weekly” playlist is a major success driven by data.

Amazon – Recommendation System & Dynamic Pricing

- Uses machine learning to recommend products to individual users.
- Conducts massive A/B testing to optimize the website.
- Employs dynamic pricing that changes based on demand, timing, and user behavior.

Facebook / TikTok – Continuous A/B Testing

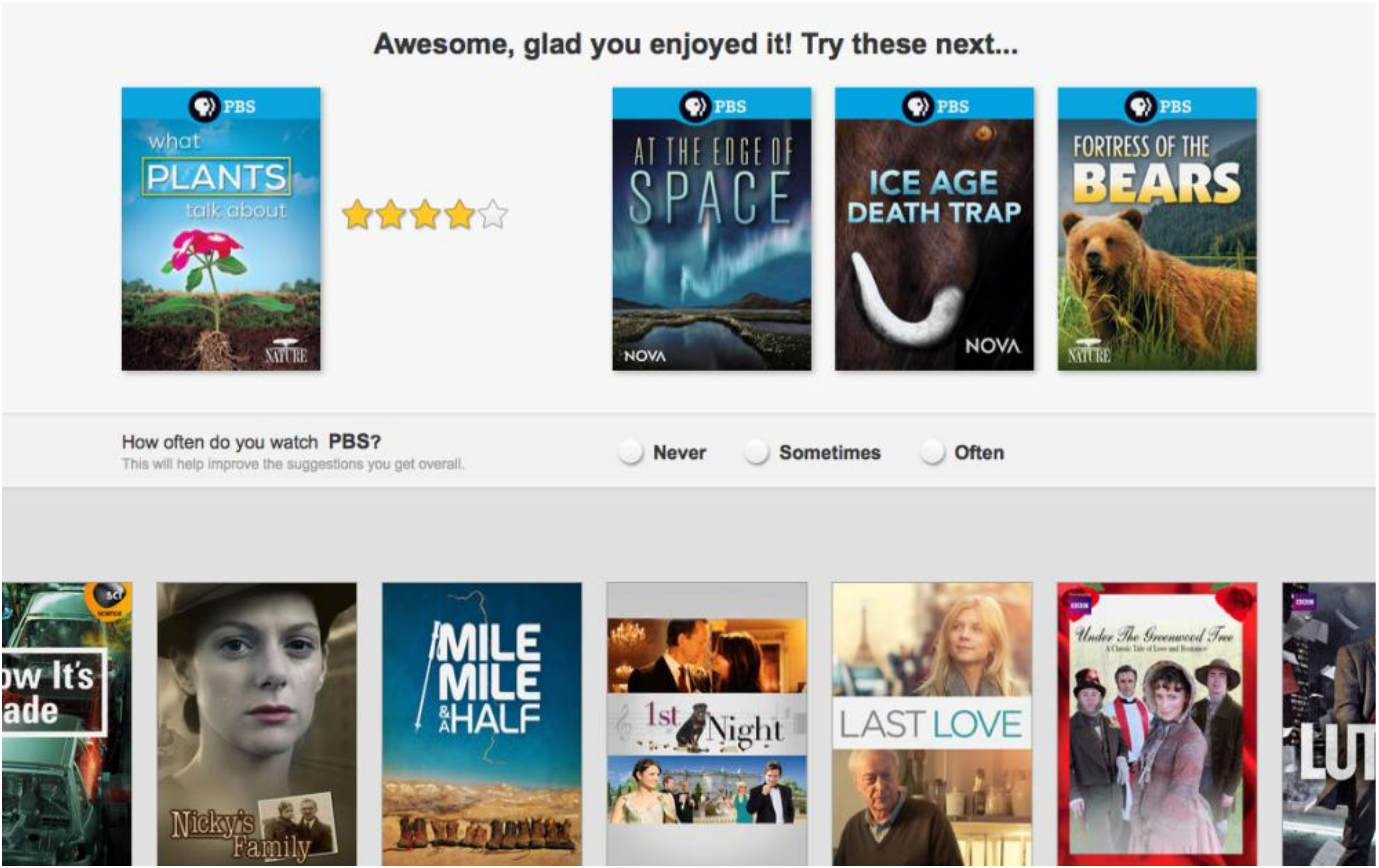
- Runs thousands of A/B tests daily to test UI/UX changes.
- Uses statistical significance testing and power analysis.
- Applies causal inference to identify true drivers of user engagement

Healthcare – Predictive Analytics

- Hospitals predict risk of stroke, sepsis, and patient deterioration.
- During COVID-19, forecasting models were used for bed and supply planning.
- Applications include classification, forecasting, and survival analysis.

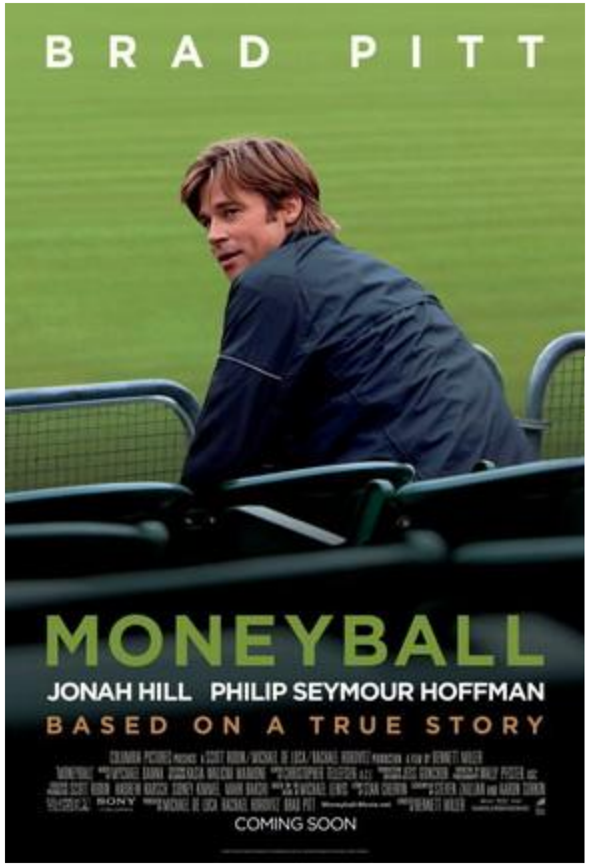
Fraud Detection – Banking / E-Commerce

- Detects unusual spending patterns in real time.
- Employs logistic regression, random forest, and anomaly detection.
- Reduces losses from fraudulent transactions significantly.

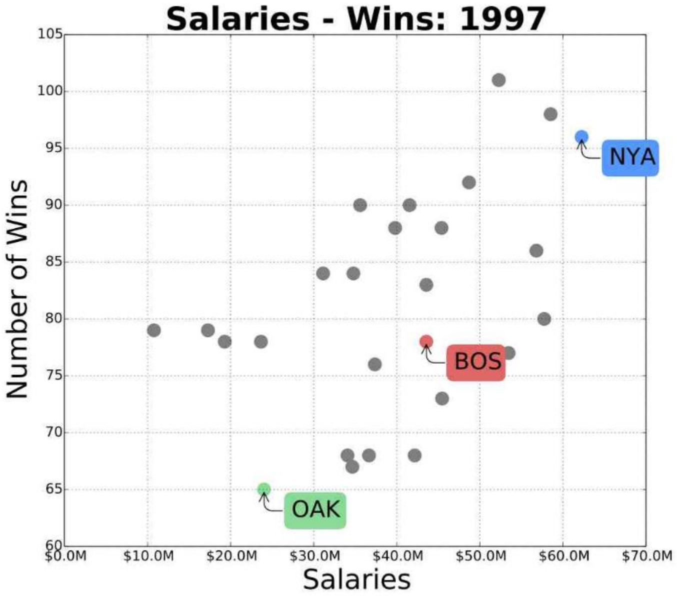


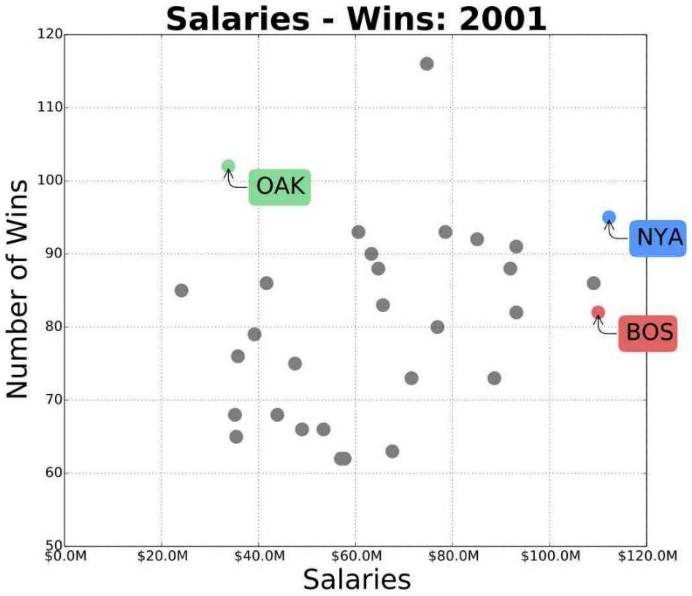
The Netflix Recommender

Statistical Learning in Sports



Using data analytics and Moneyball theory, Beane hired the best players he could with an extremely limited budget for payroll. With approximately $41 million in salary, the Oakland Athletics ultimately competed with larger market teams such as the Yankees, who spent over $125 million in payroll during the 2002 baseball season.



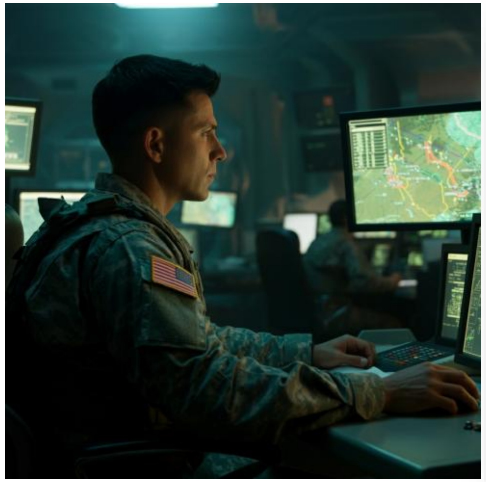


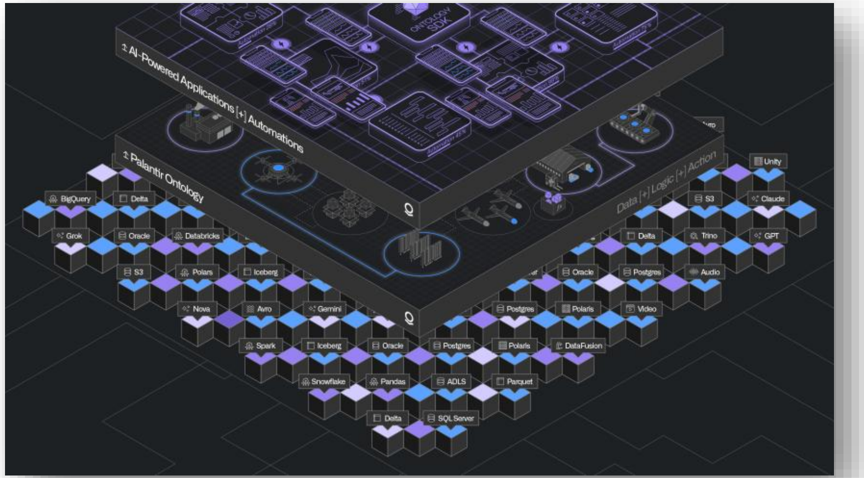
Palantir

A Real-World Example of Large-Scale Data Science & Analytics Platforms

Enterprise-scale data platforms that help organizations:

- Integrate massive datasets
- Analyze and visualize data
- Deploy AI/ML securely
- Make mission-critical decisions





Palantir's Main Platforms:

Gotham
Used by defense, intelligence, and law enforcement
Supports data integration for national security
Enables link analysis, pattern detection, mission planning

Foundry
Used by corporations (Airbus, Merck, energy, finance)
Centralizes all organizational data
Creates pipelines, dashboards, simulations, digital twins
Supports large-scale AI/ML workflows

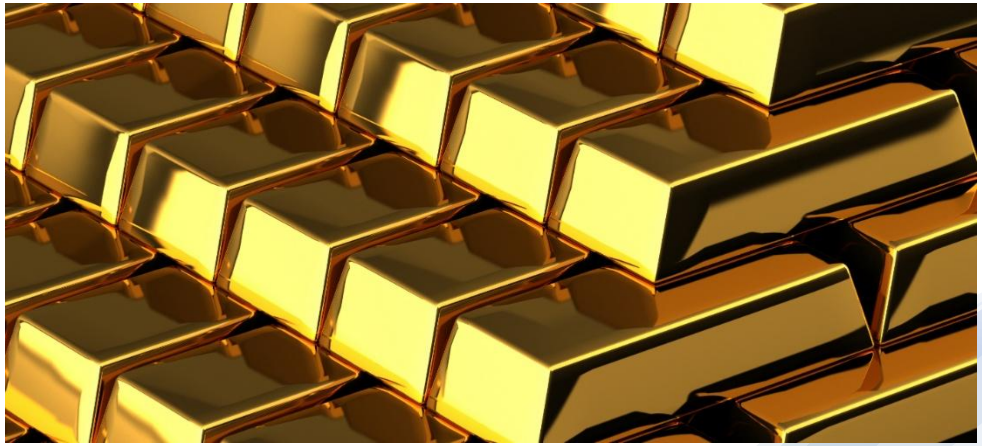
AIP (Artificial Intelligence Platform)
Integrates enterprise data with LLMs and AI agents
Designed for secure AI deployment
Used in manufacturing, logistics, defense, healthcare

Other Applications

- **SCB / KBank – Credit Scoring and Fraud Detection**
- **Grab – Dispatch Optimization + Demand Forecasting**
- **Shopee / Lazada – Recommender + Dynamic Pricing**

“What problems would the world face if DS / statistics did not exist?”

Need Stats / DS?




rime] 20 Wicks 1.1 Close Dashed
der Blocks [JacobMagleby] 7 10 Close Engulfs 100% Of Order Block
Sep
Oct
Nov
Dec

Inference vs. Prediction

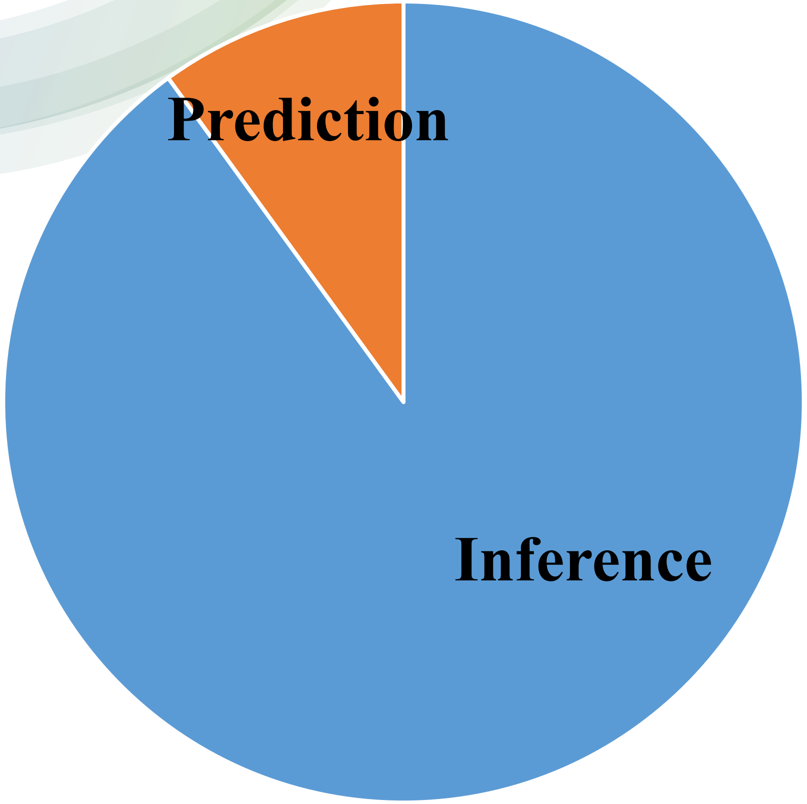
- **Inference:**
  - given a set of data you want to *infer* how the output is generated as a function of the data

    Ex. You want *to understand* how ozone levels are influenced by temperature, solar radiation, and wind.

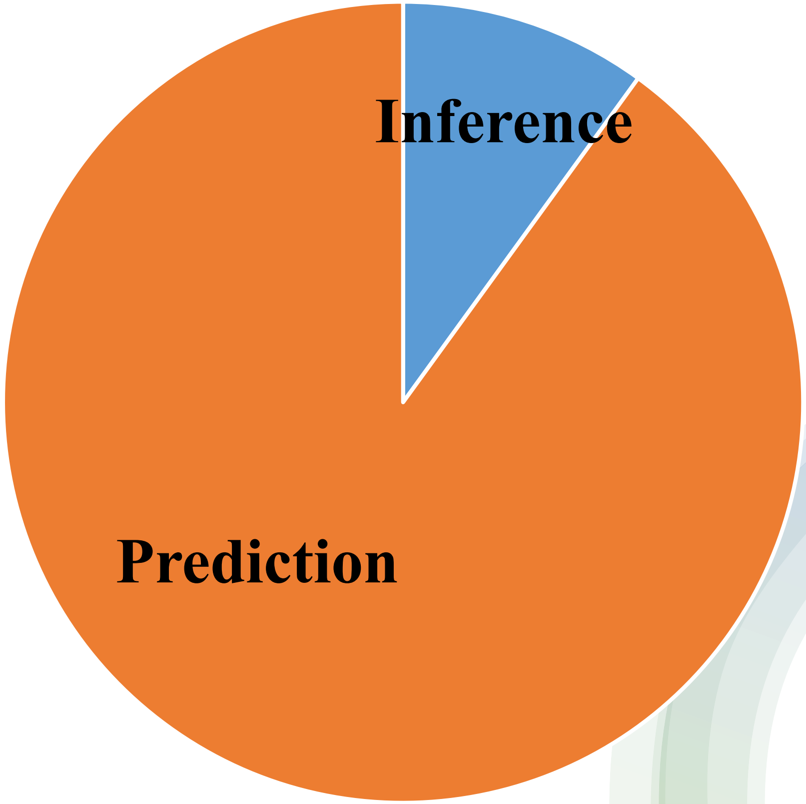
- **Prediction:**
  - Given a new data point, you want to use an existing data set to build a model that reliably *predicts* the outcome.

    Ex. You want to predict future ozone levels using historic data.

Statistics vs. ML



How statisticians see the world

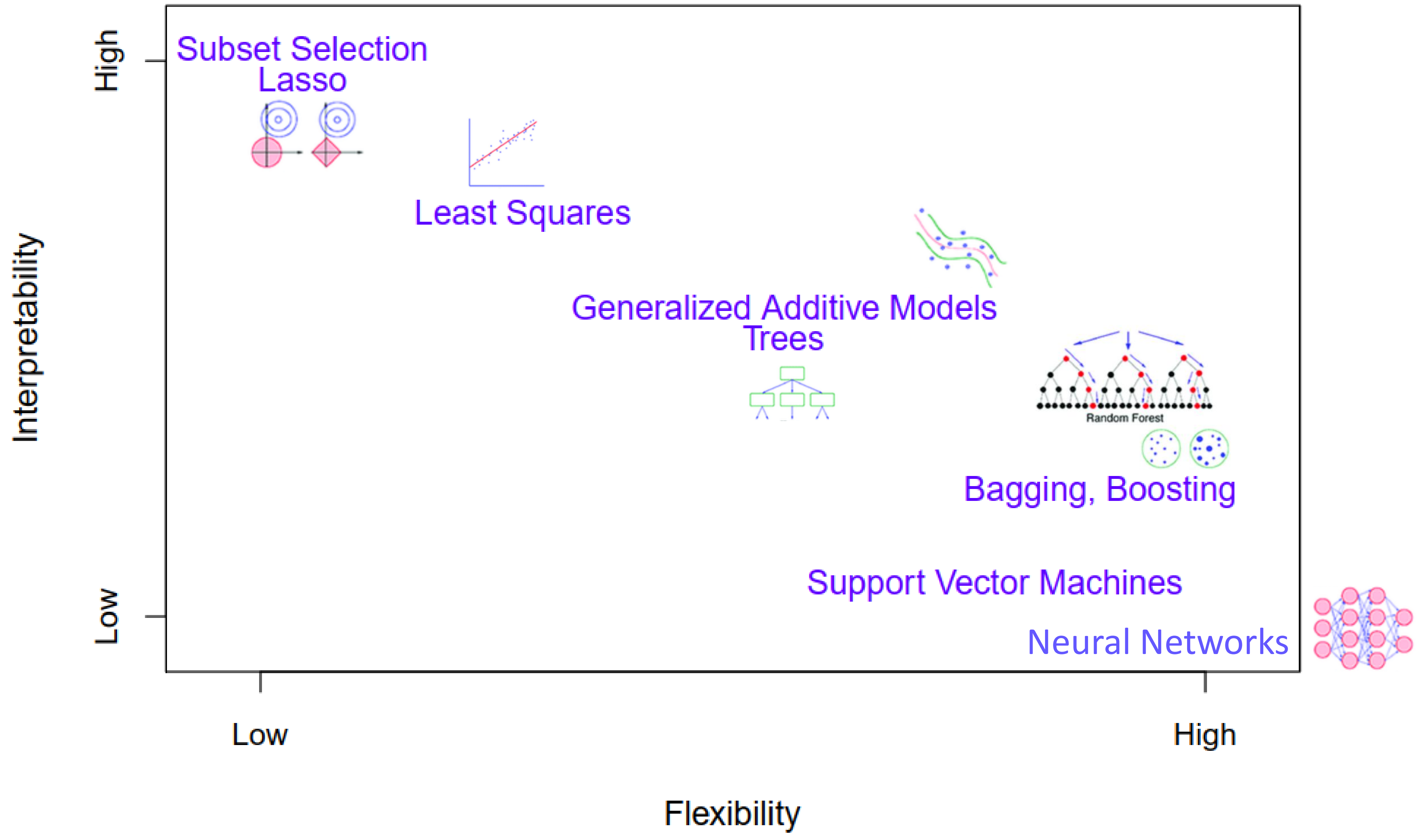


How machine learners see the world

Interpretability is a necessity for inference

- Interpretable:
  - Generalized linear models (e.g. linear regression, logistic regression), linear discriminant analysis, linear support vector machines (SVMs), decision trees

- Less interpretable:
  - Neural networks, SVM with non-linear kernels, random forests



https://www.ncbi.nlm.nih.gov/pubmed/29354080

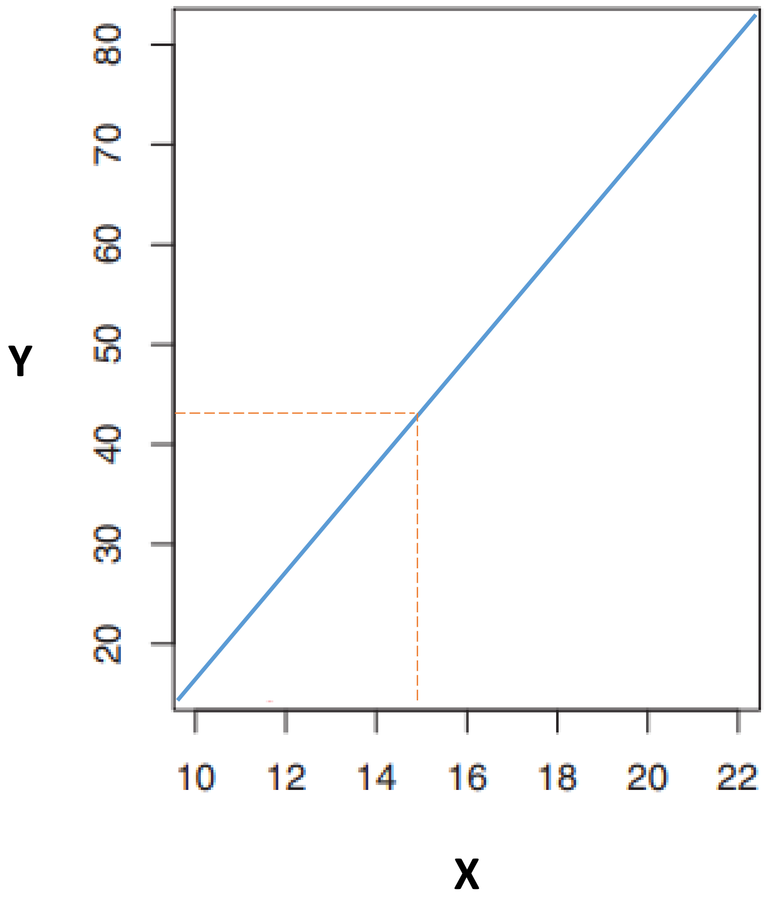
Regression Versus Classification Problems

- Quantitative outcome variable
  - Regression

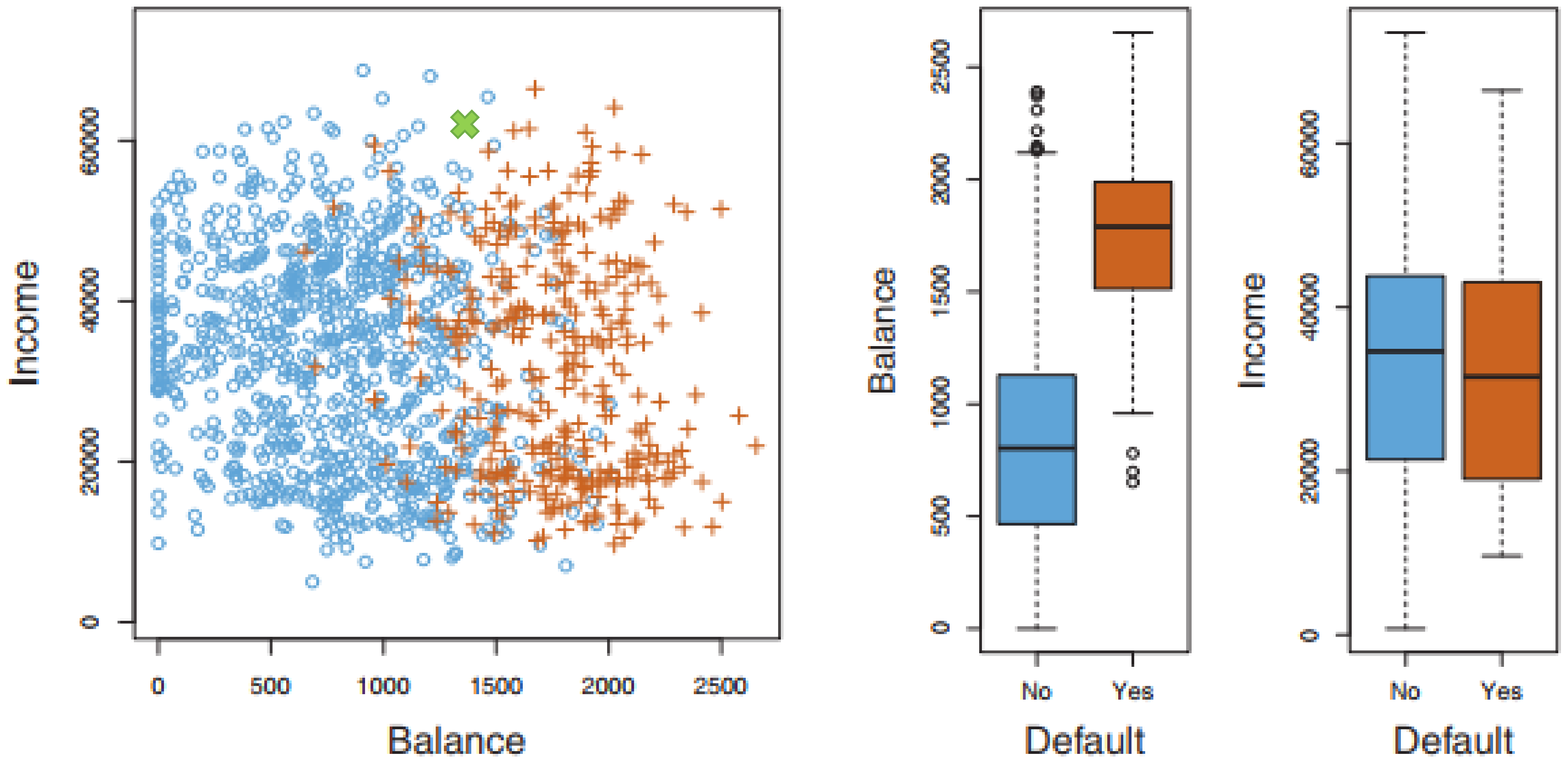
- Qualitative outcome variable
  - Classification

Suppose we already have the best model

If someone has 15 years of education → his income?



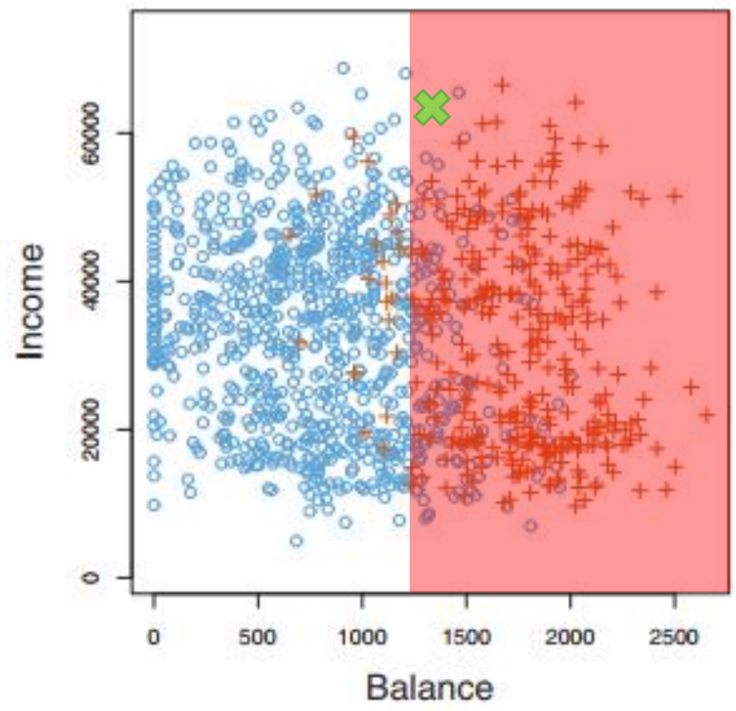
Regression problem



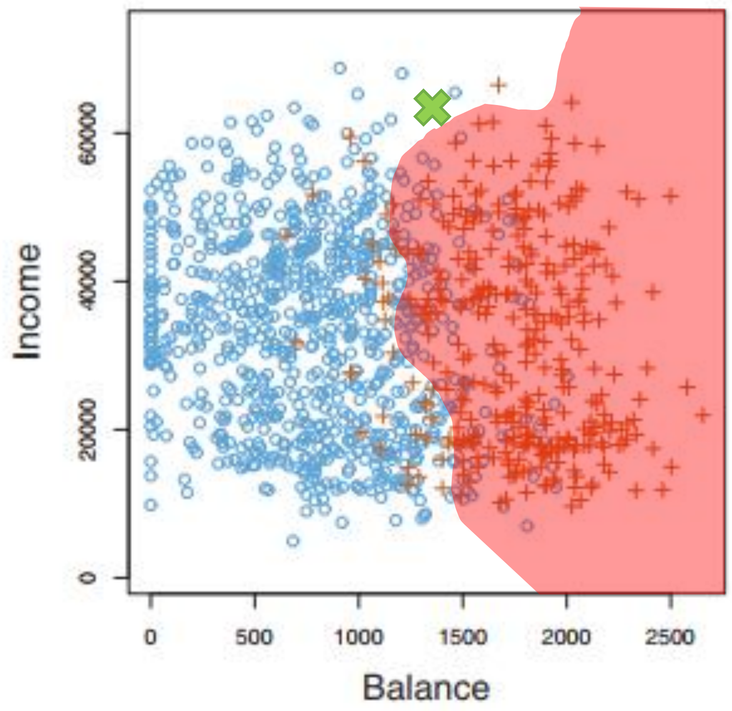
1. **What will be our model to explain this data?**
2. **If we have a new data, can we make a prediction for that data?**
   (✖)

Classification Problem

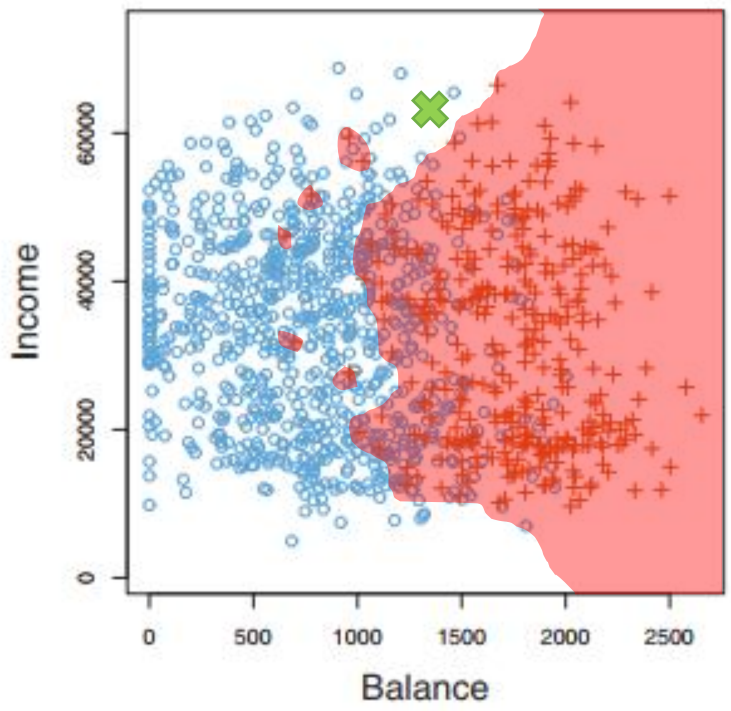
Model 1



Model 2



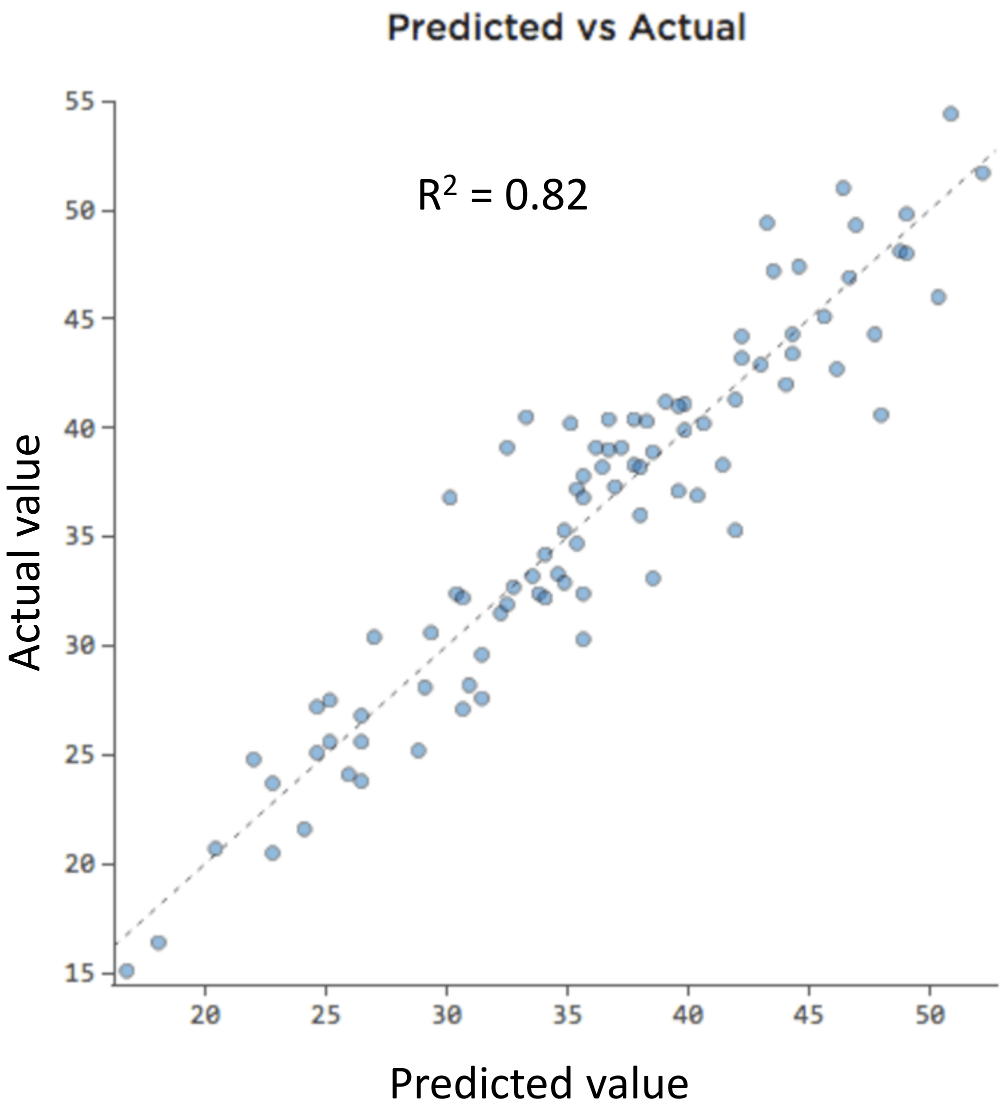
Model 3

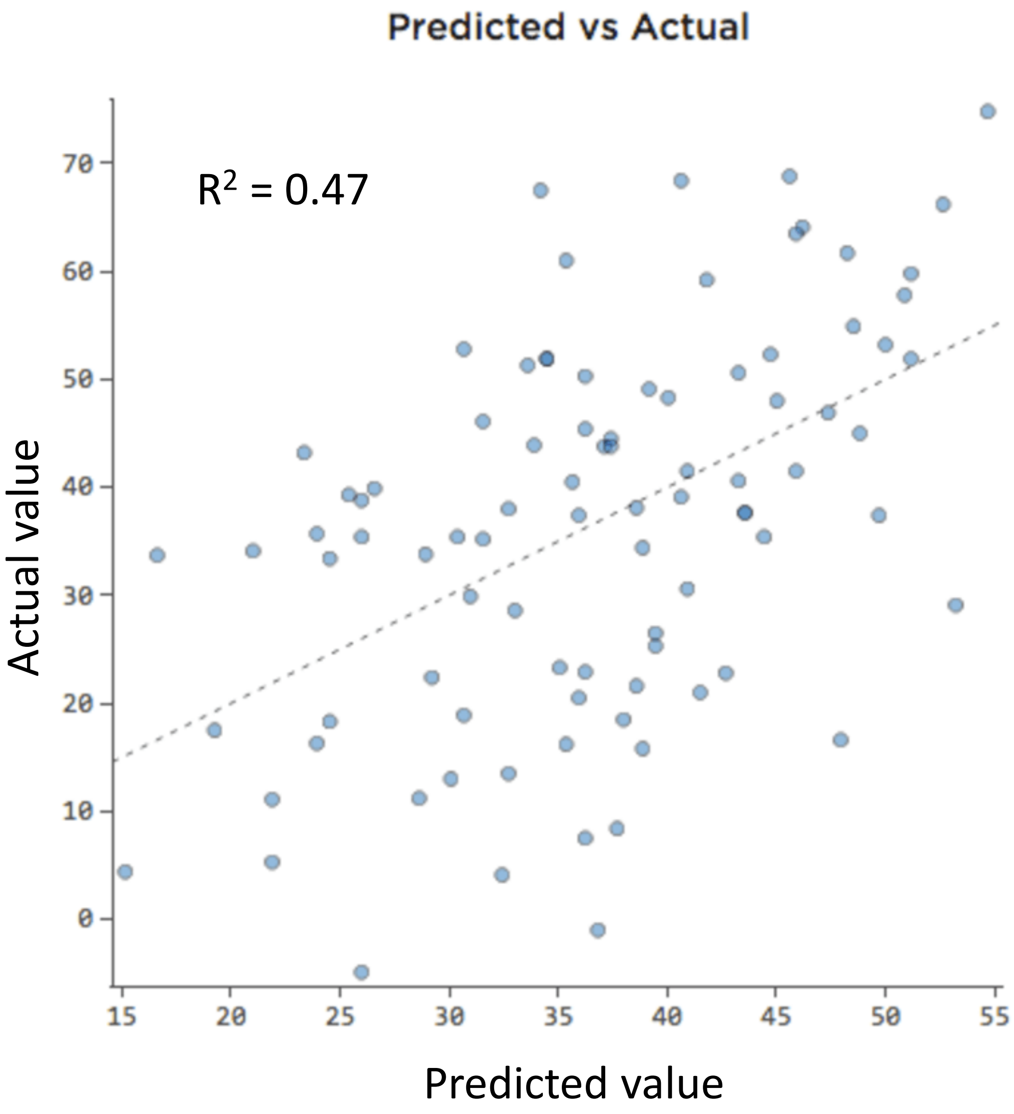


Evaluating Model Performance

- In regression problems
  - Mean squared error
  - Root mean squared error
  - R-squared

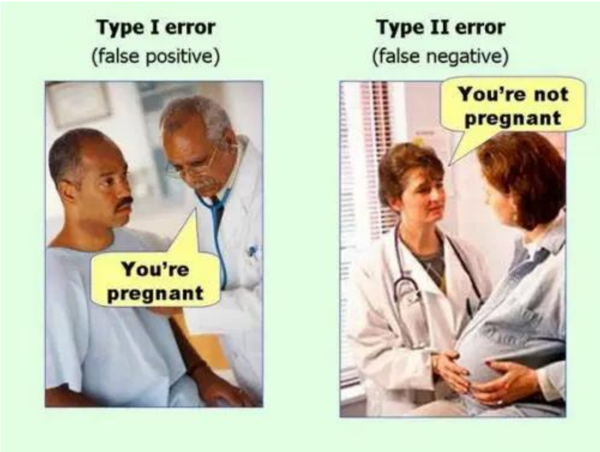
- In classification problems
  - Accuracy
  - Sensitivity
  - Specificity
  - Area under receiver operating characteristic curve


Predicted vs Actual
R^2 = 0.82
Actual value
Predicted value
55
50
45
40
35
30
25
20
15
20
25
30
35
40
45
50


Predicted vs Actual
R^2 = 0.47
Actual value
Predicted value
70
60
50
40
30
20
10
0
15
20
25
30
35
40
45
50
55

Confusion matrix

		Predicted class	
		P	N
Actual Class	P	True Positives (TP)	False Negatives (FN) Type 2 error
	N	False Positives (FP) Type 1 error	True Negatives (TN)



Accuracy (ACC)

$$ACC = (TP + TN)/(P + N)$$

Balanced Accuracy (BACC)

$$BACC = (TP/P + TN/N)/2$$

F1 Score

is the harmonic mean of Precision and Sensitivity

$$F1 = 2TP/(2TP + FP + FN)$$

Matthews Correlation Coefficient (MCC)

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Sensitivity or True Positive Rate (TPR)

eqv. with hit rate, recall

$$TPR = TP/P = TP/(TP + FN)$$

Specificity (SPC) or True Negative Rate (TNR)

$$SPC = TN/N = TN/(FP + TN)$$

Precision or Positive Predictive Value (PPV)

$$PPV = TP/(TP + FP)$$

Negative Predictive Value (NPV)

$$NPV = TN/(TN + FN)$$

Fall-out or False Positive Rate (FPR)

$$FPR = FP/N = FP/(FP + TN) = 1 - TNR$$

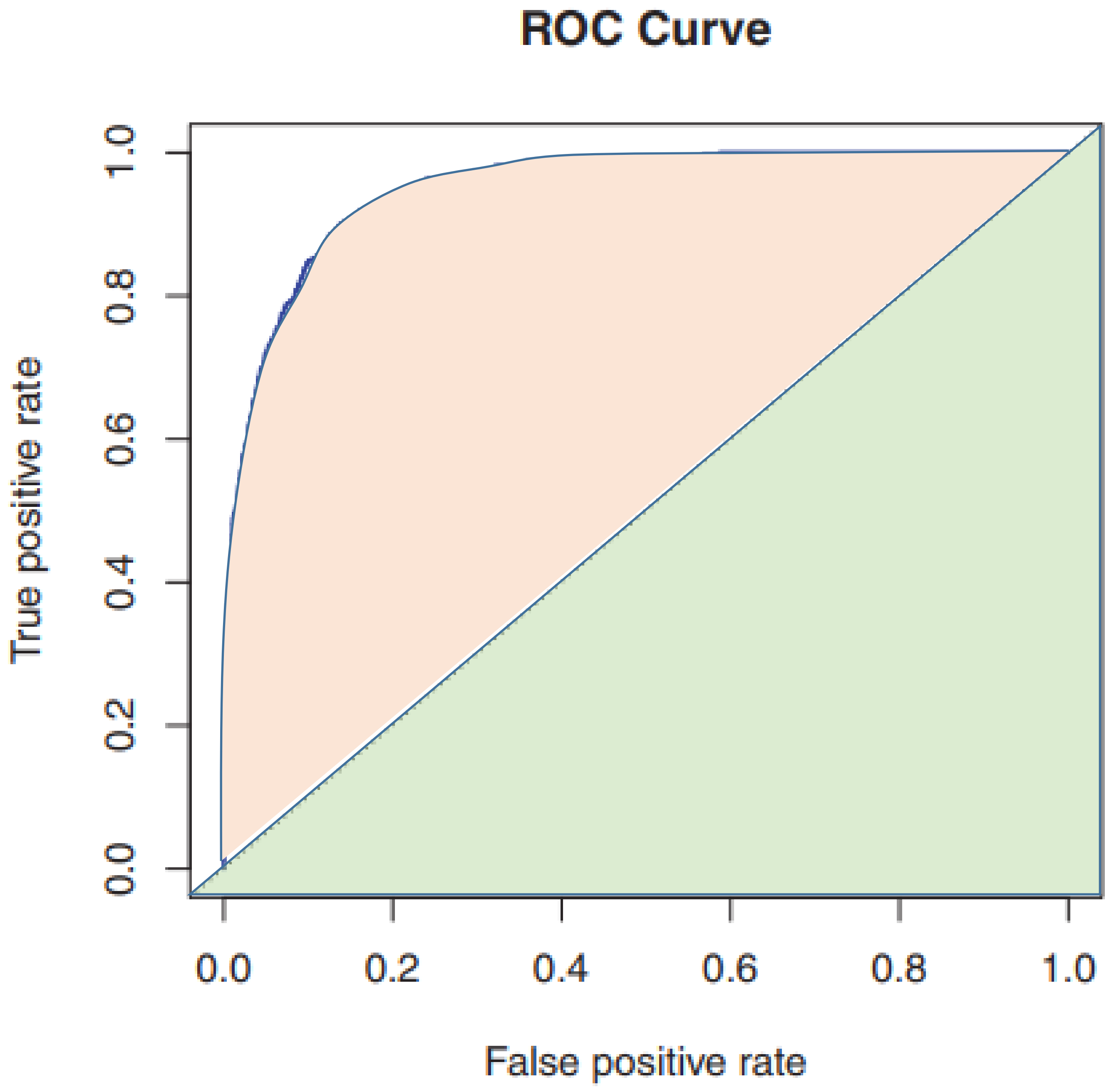
False Discovery Rate (FDR)

$$FDR = FP/(FP + TP) = 1 - PPV$$

Miss Rate or False Negative Rate (FNR)

$$FNR = FN/(FN + TP) = 1 - TPR$$

Area Under the roc Curve (AUC)

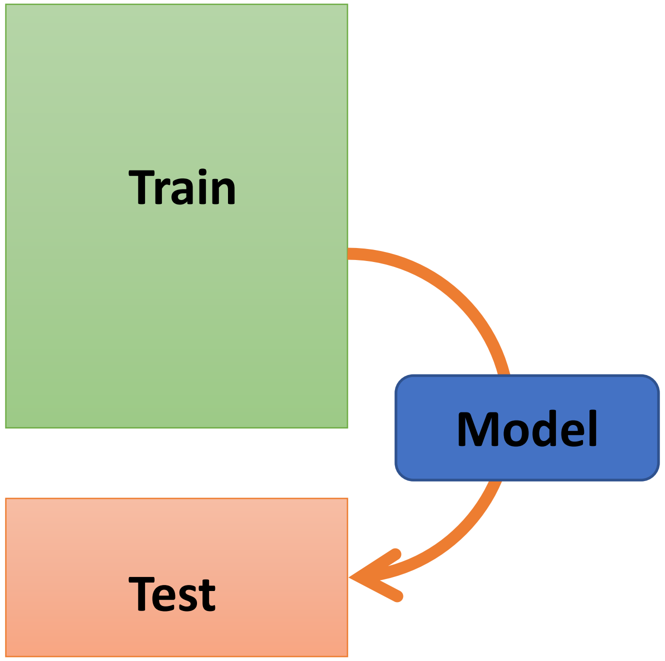


Random guess

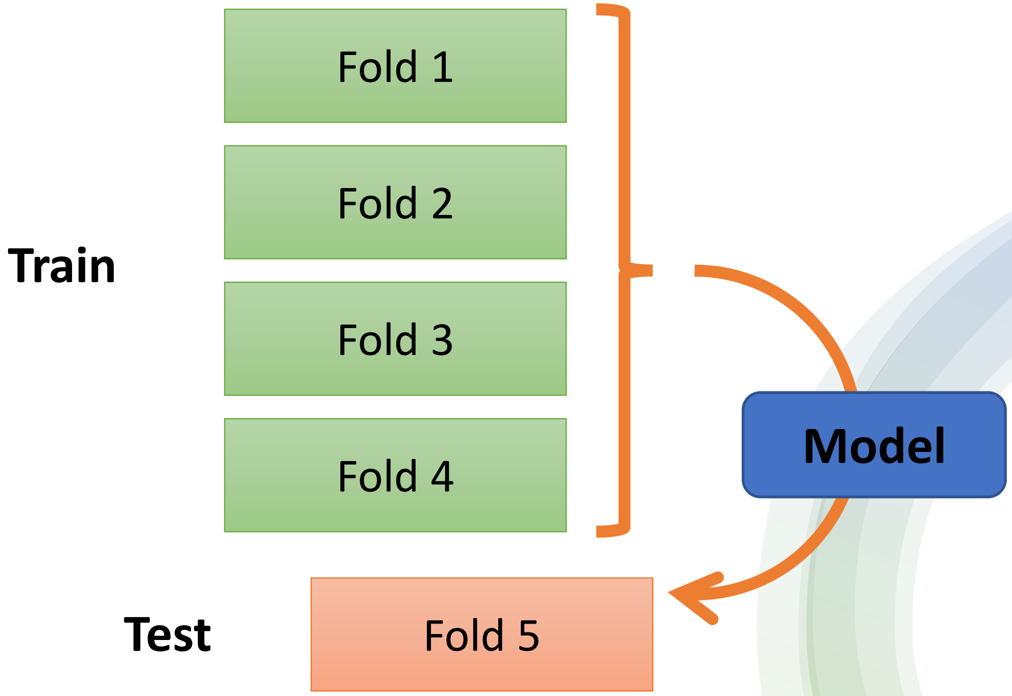
Better than random guess

Model Evaluation Technique Using Cross-Validation

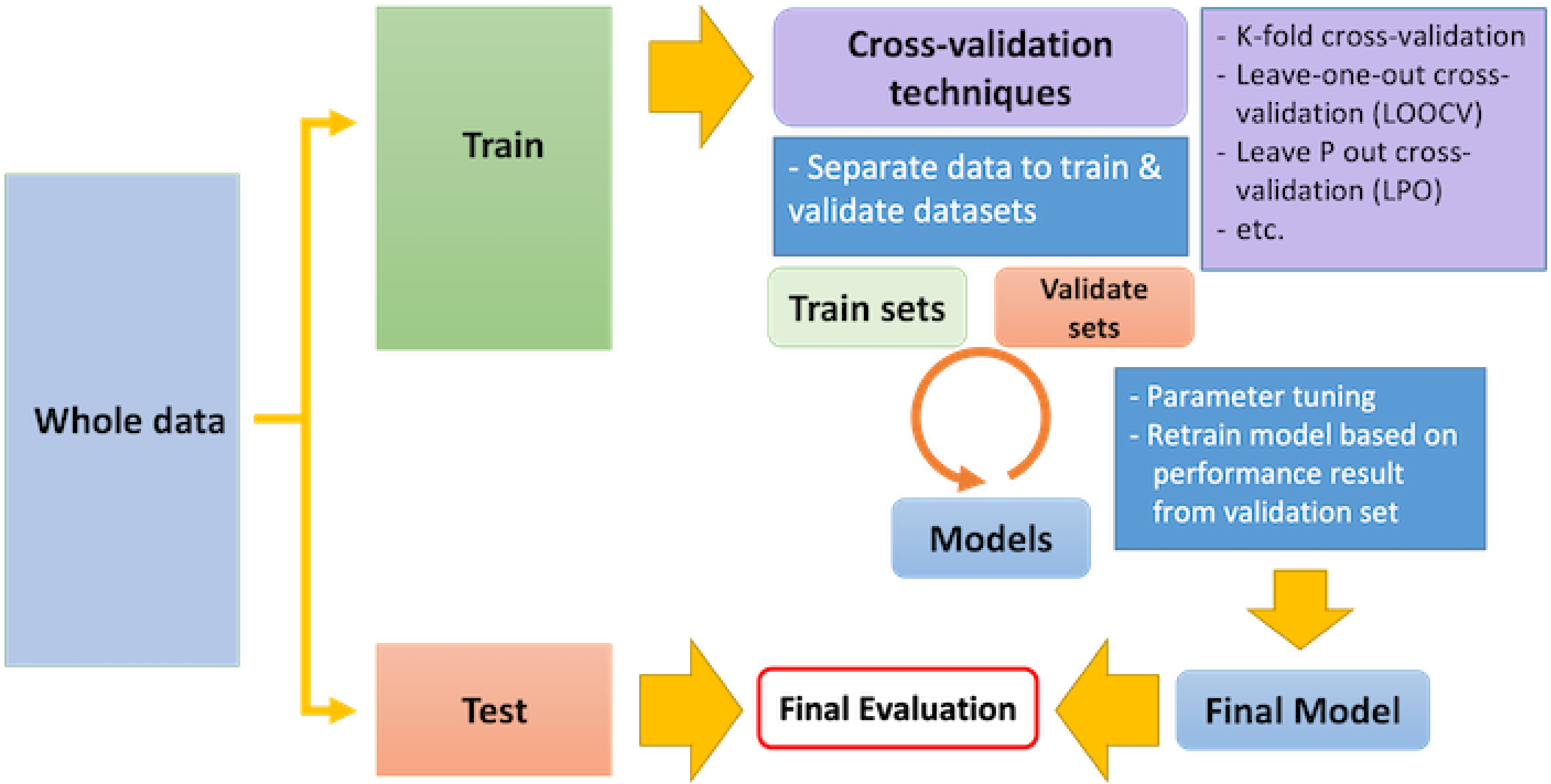
Hold-out



K-fold cross-validation (CV)



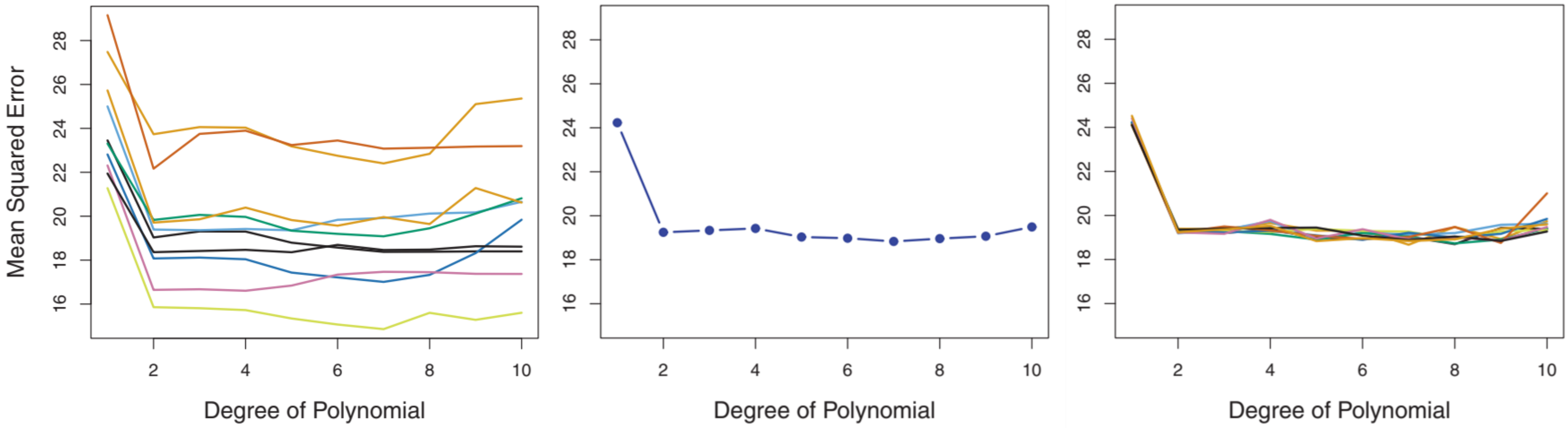
	Fold1	Fold2	Fold3	Fold4	Fold5
Iteration 1	Test	Train	Train	Train	Train
Iteration 2	Train	Test	Train	Train	Train
Iteration 3	Train	Train	Test	Train	Train
Iteration 4	Train	Train	Train	Test	Train
Iteration 5	Train	Train	Train	Train	Test


Whole data
Train
Test
Cross-validation techniques
- K-fold cross-validation
- Leave-one-out cross-validation (LOOCV)
- Leave P out cross-validation (LPO)
- etc.
- Separate data to train & validate datasets
Train sets
Validate sets
Models
- Parameter tuning
- Retrain model based on performance result from validation set
Final Model
Final Evaluation

The hold-out method was repeated ten times, each time using a different random split

LOOCV

10-fold CV



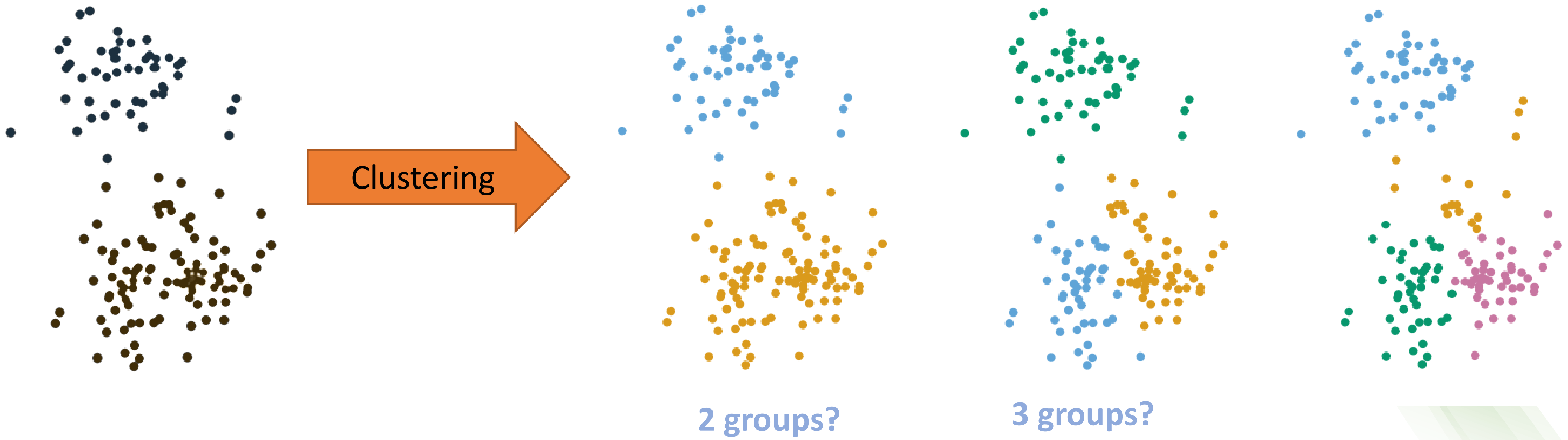
Supervised & Unsupervised Learning

- **Supervised learning**
  - Output/outcome/label variable is available
  - Involves building a statistical model for predicting, or estimating, an *output* based on one or more *inputs*.
    - Regression models
    - Classification models

- **Unsupervised learning**
  - No output/outcome/label variable
  - There are inputs but no supervising output; nevertheless, we can learn relationships and structure from such data.

Example of Unsupervised Learning Problem

- We want to find customers with common interests and group them together
- For simplicity, suppose we use PCA (principal component analysis) on the data and plot the first 2 components



Rython (R & Python) for Data Science