

# A/B Testing

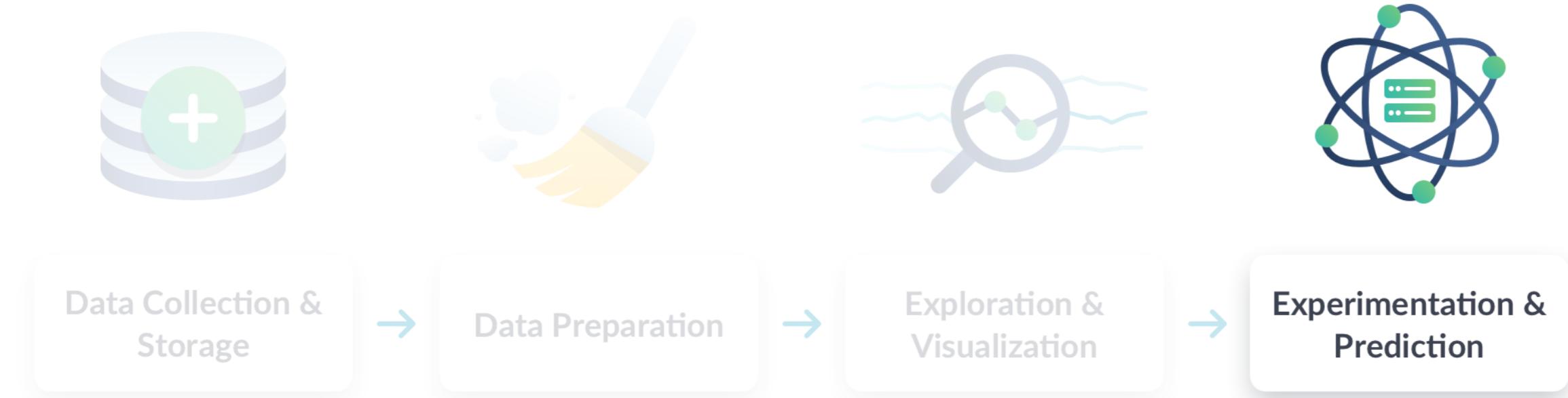
UNDERSTANDING DATA SCIENCE



Lis Sulmont

Curriculum Manager, DataCamp

# Data science workflow



# What are experiments in data science?

*Experiments help drive decisions and draw conclusions*

1. Form a question
2. Form a hypothesis
3. Collect data
4. Test the hypothesis with a statistical test
5. Interpret results

# Case study: which is the better blog post title?

**Form a question:** Does blog title A or blog title B result in more clicks?

**Form a hypothesis:** Blog title A and blog title B result in the same amount of clicks.

**Collect data:**

- 50% users will see blog title A
- 50% users will see blog title B
- Track click-through rate until sample size reached

A

Become an expert Data Scientist with this one weird trick!



B

You won't believe these tips for becoming a Data Scientist!



# Case study: which is the better blog post title?

Test the hypothesis with a statistical test: Is the difference in titles' click-through rates significant?

Interpret results:

- Choose a title
- Or ask more questions and design another experiment!

A

Become an expert Data Scientist with this one weird trick!



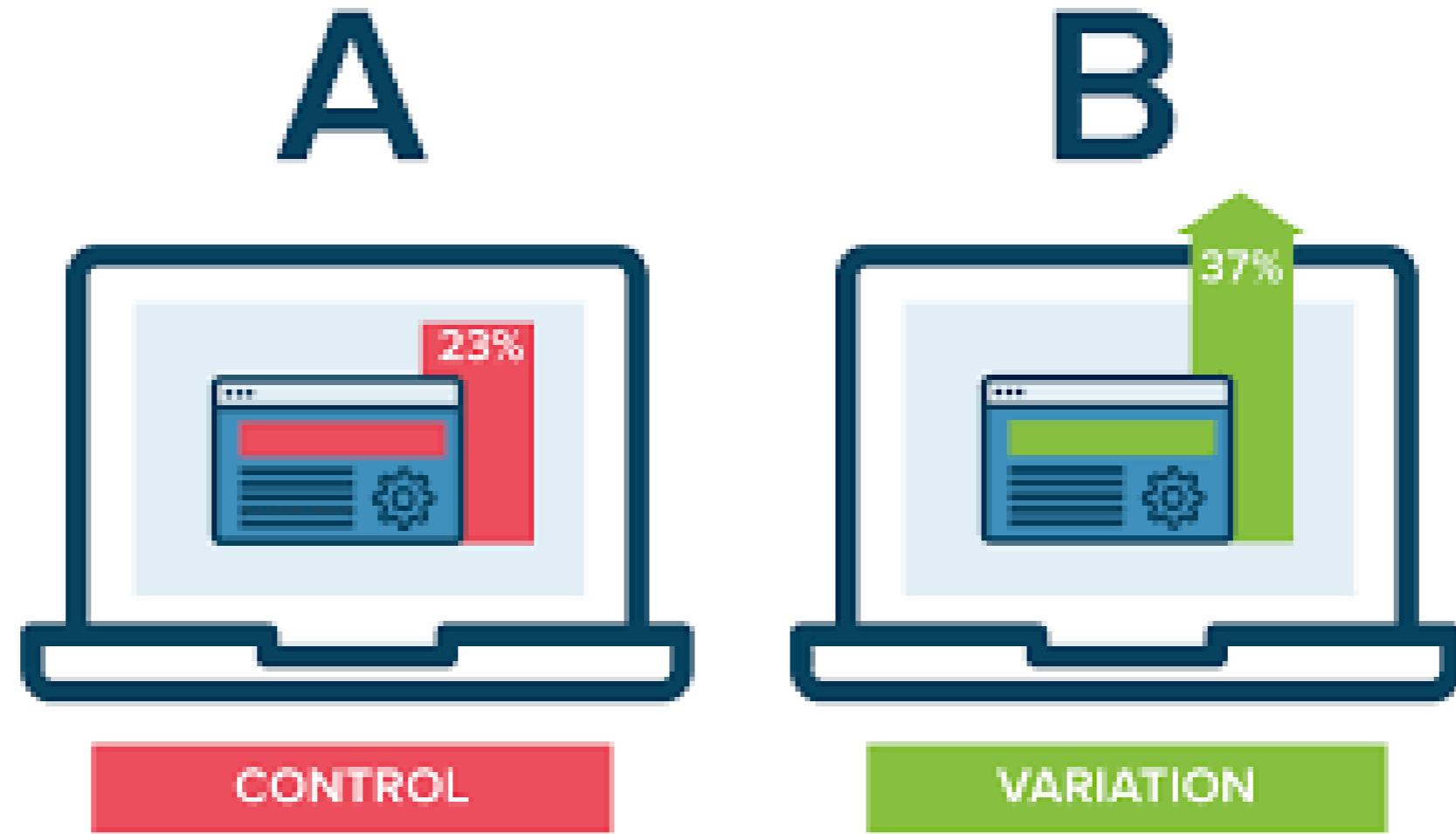
B

You won't believe these tips for becoming a Data Scientist!



# What is A/B Testing?

*AKA Champion/Challenger Testing*



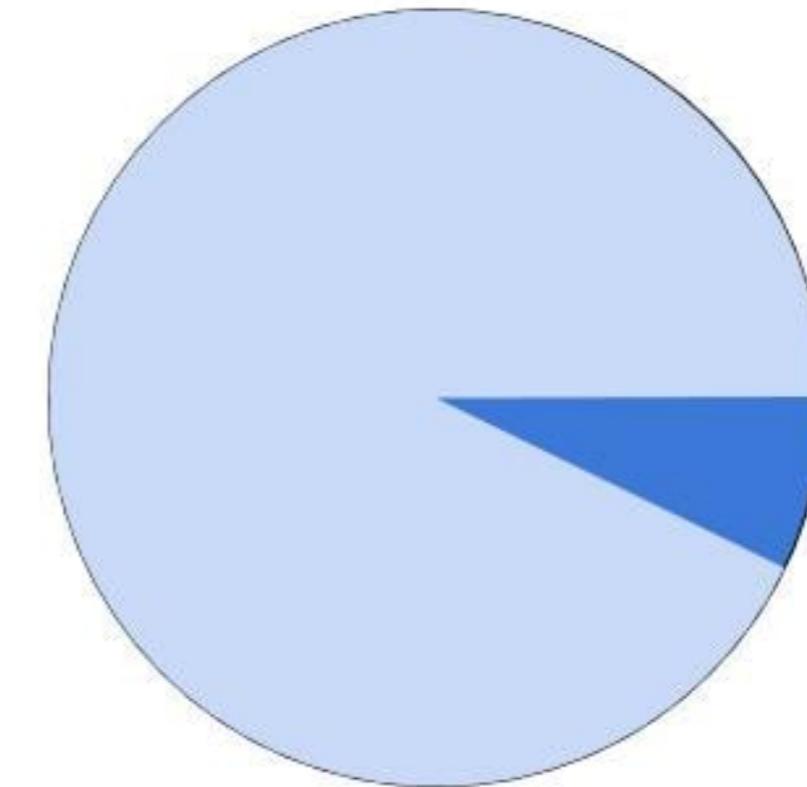
# Terminology Review

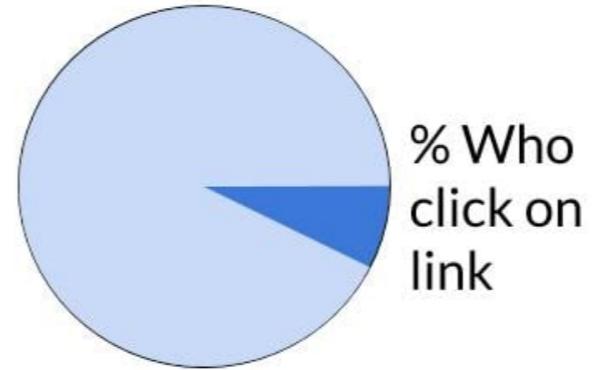
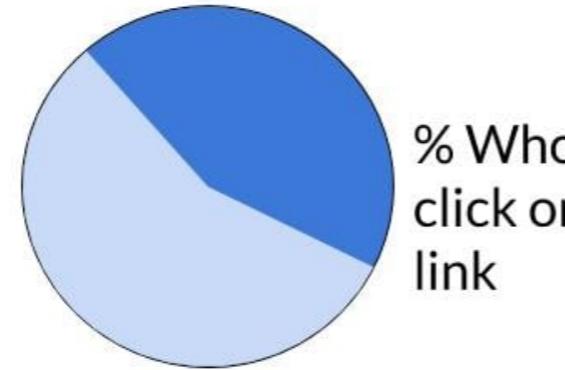
- **Sample size:** number of data points used
- **Statistical significance:** result is likely not due to chance
  - Given assumptions of statistical model
  - Use **statistical tests** to calculate this:
    - e.g., t-test, Z-test, ANOVA, Chi-square test

# A/B Testing Steps

- Picking a metric to track
- Calculating sample size
- Running the experiment
- Checking for significance

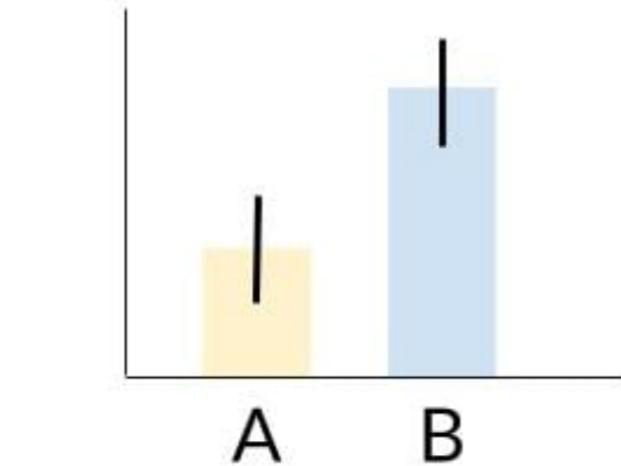
# Pick a metric to track: click-through rate



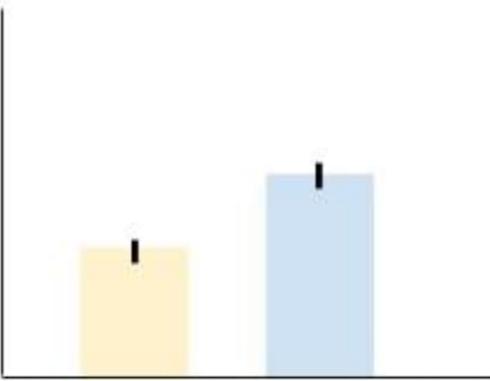


- Baseline metric to gauge any changes
  - *How often people generally click on a link to our blogs*
- If the rate is much larger or smaller than 50%, we need a large sample size
  - Click rate is typically small (<3%)

Low sensitivity, detects  
large differences



High sensitivity, detects  
small differences



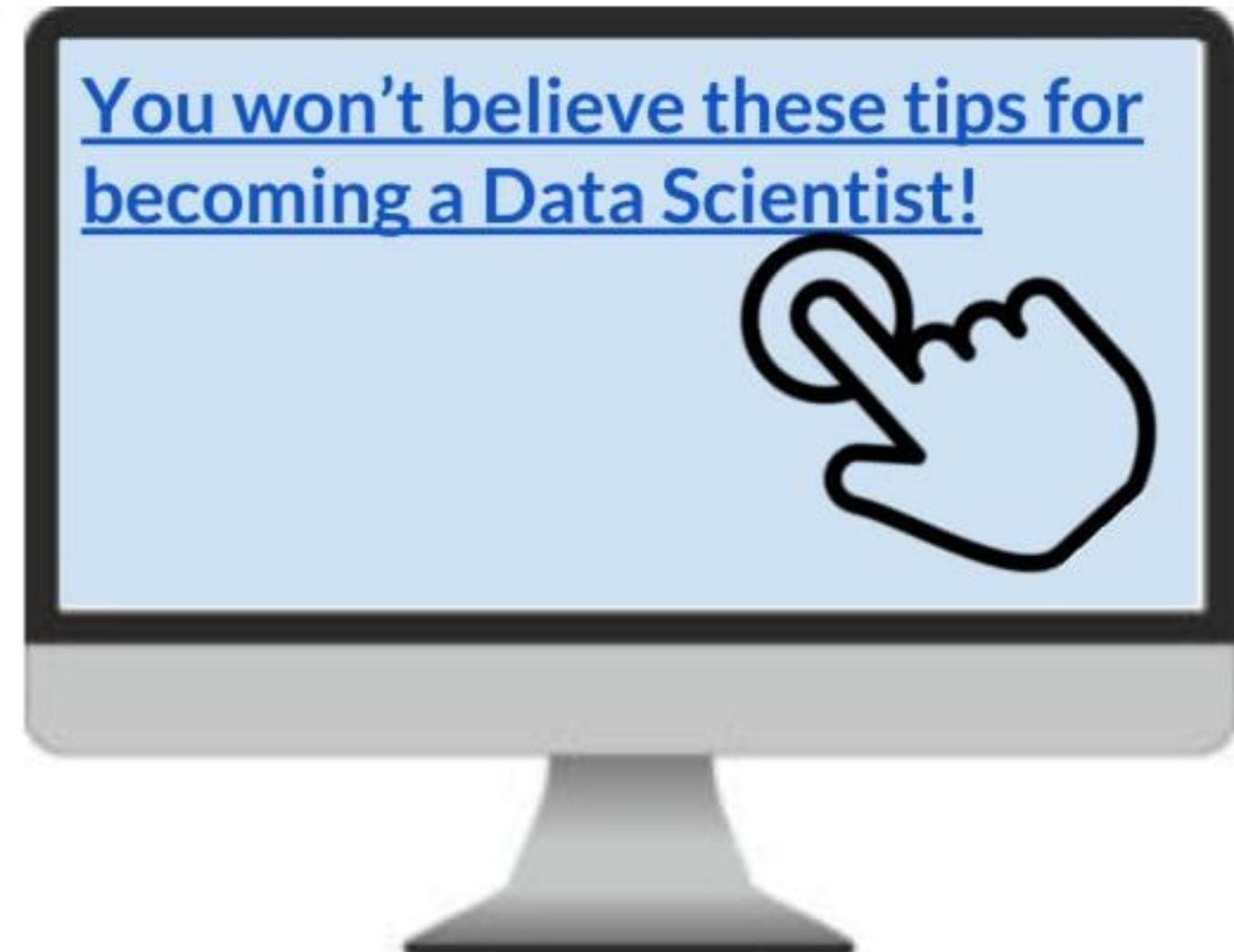
Larger sample sizes allow us to detect smaller changes

# Run your experiment

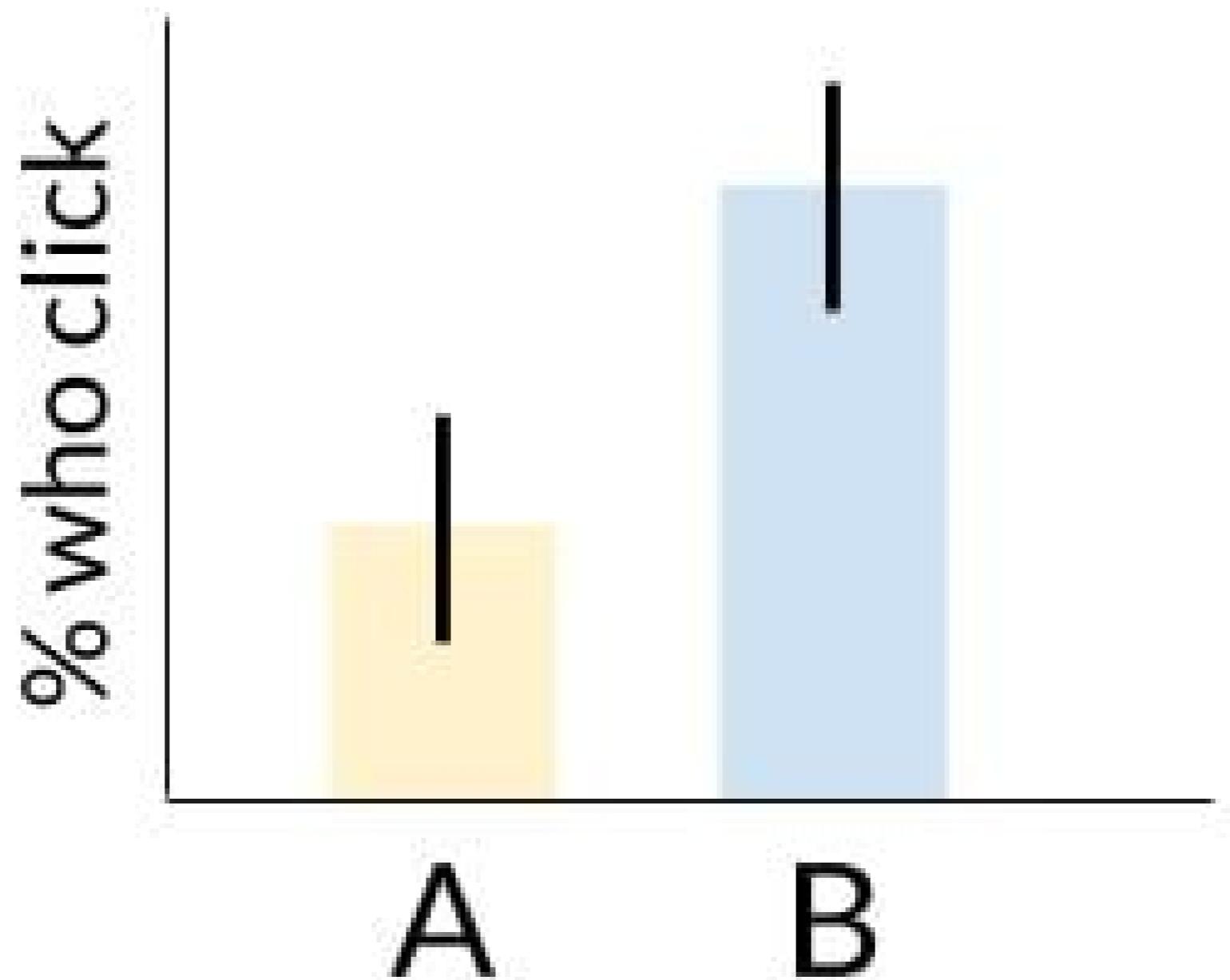
A



B



# Check for significance



# What if the results aren't significant?

- Difference is smaller than the threshold we chose
- Running our test longer won't help
- Still might be a difference; it's just small and insignificant to us

# **Let's practice!**

**UNDERSTANDING DATA SCIENCE**

# Time series forecasting

UNDERSTANDING DATA SCIENCE



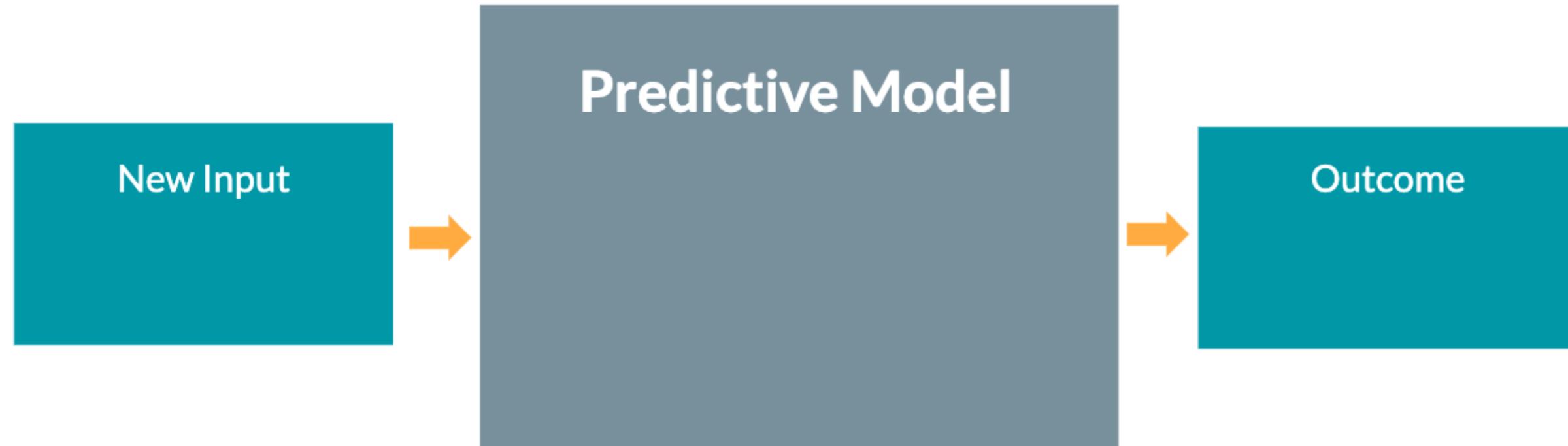
Lis Sulmont  
Curriculum Manager

# Modeling in data science

## What is a statistical model?

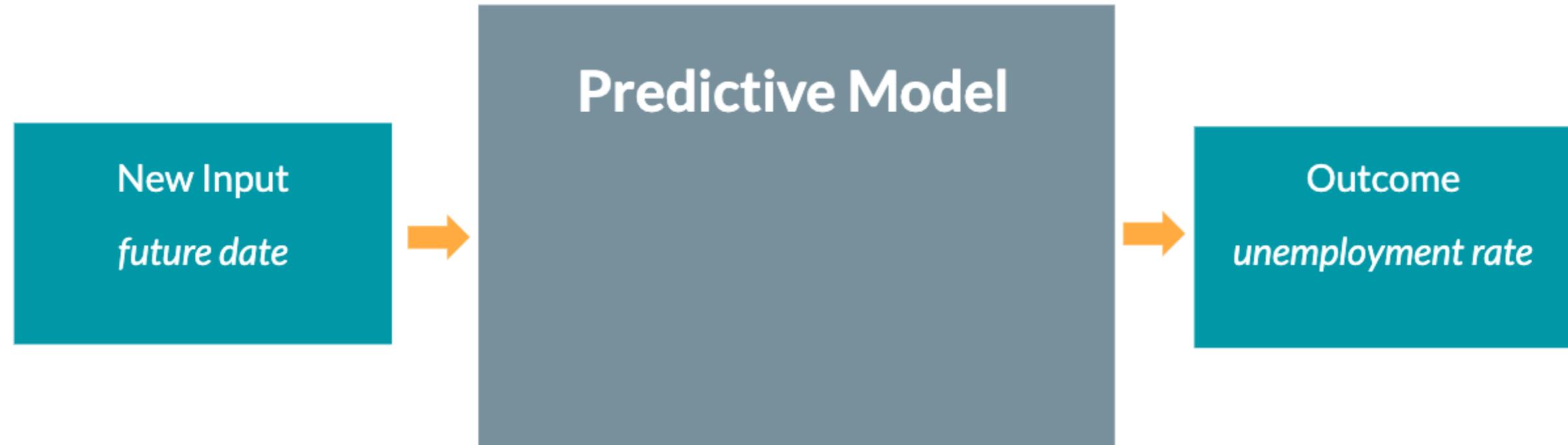
- Represent a real-world process with statistics
- Mathematical relationships between variables, including random variables
- Based on statistical assumptions and historical data

# Predictive modeling



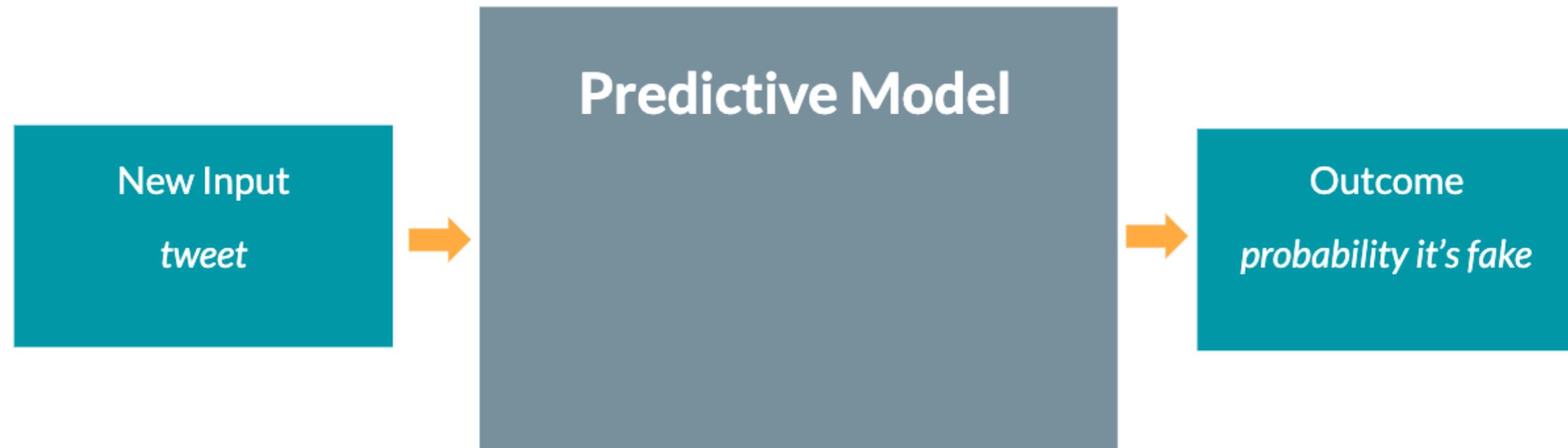
- Enter new input(s) and model predicts an outcome

# Predictive modeling



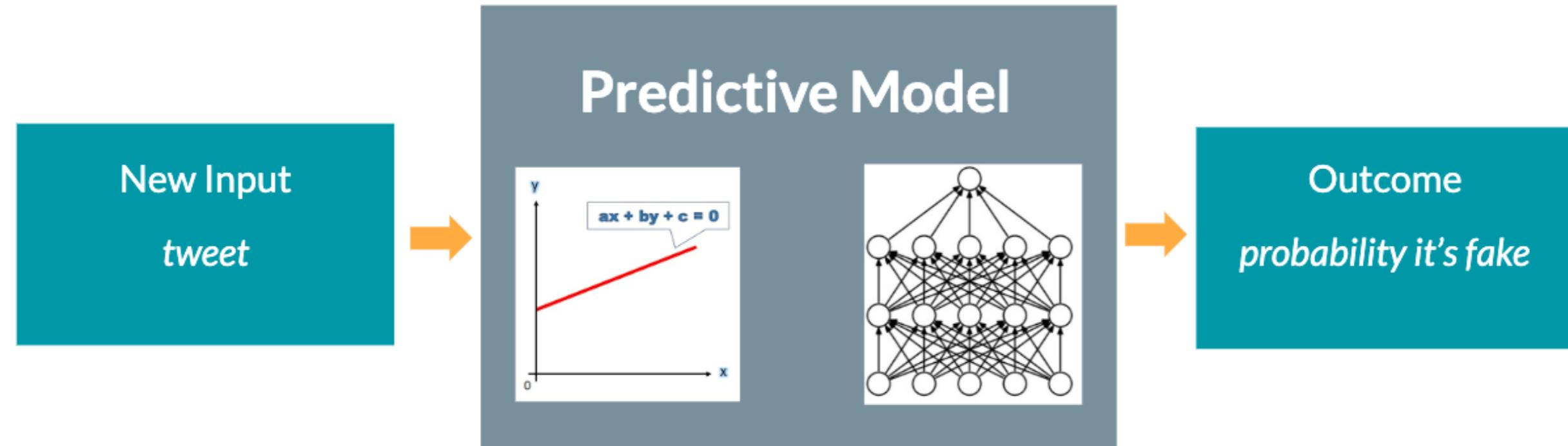
- Enter new input(s) and model predicts an outcome

# Predictive modeling



- Enter new input(s) and model predicts an outcome
  - Probability of an outcome

# Predictive modeling

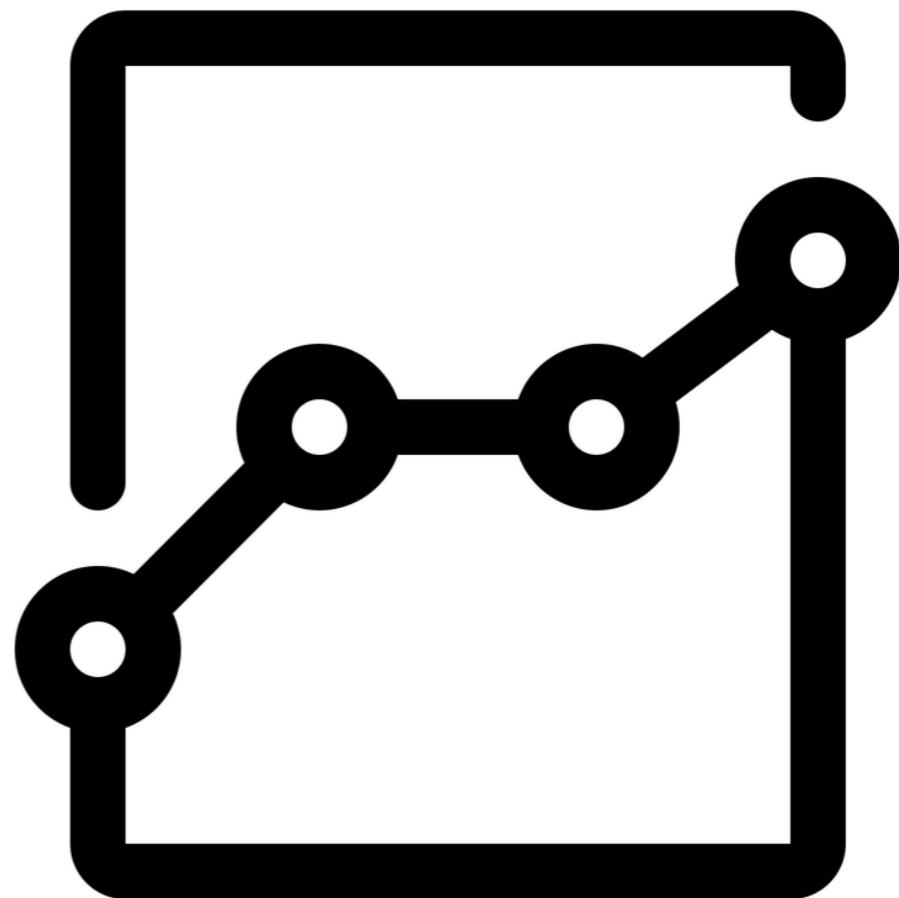


- Enter new input(s) and model predicts an outcome
  - Probability of an outcome
- Ranges in complexity, from a linear equation to a deep learning algorithm

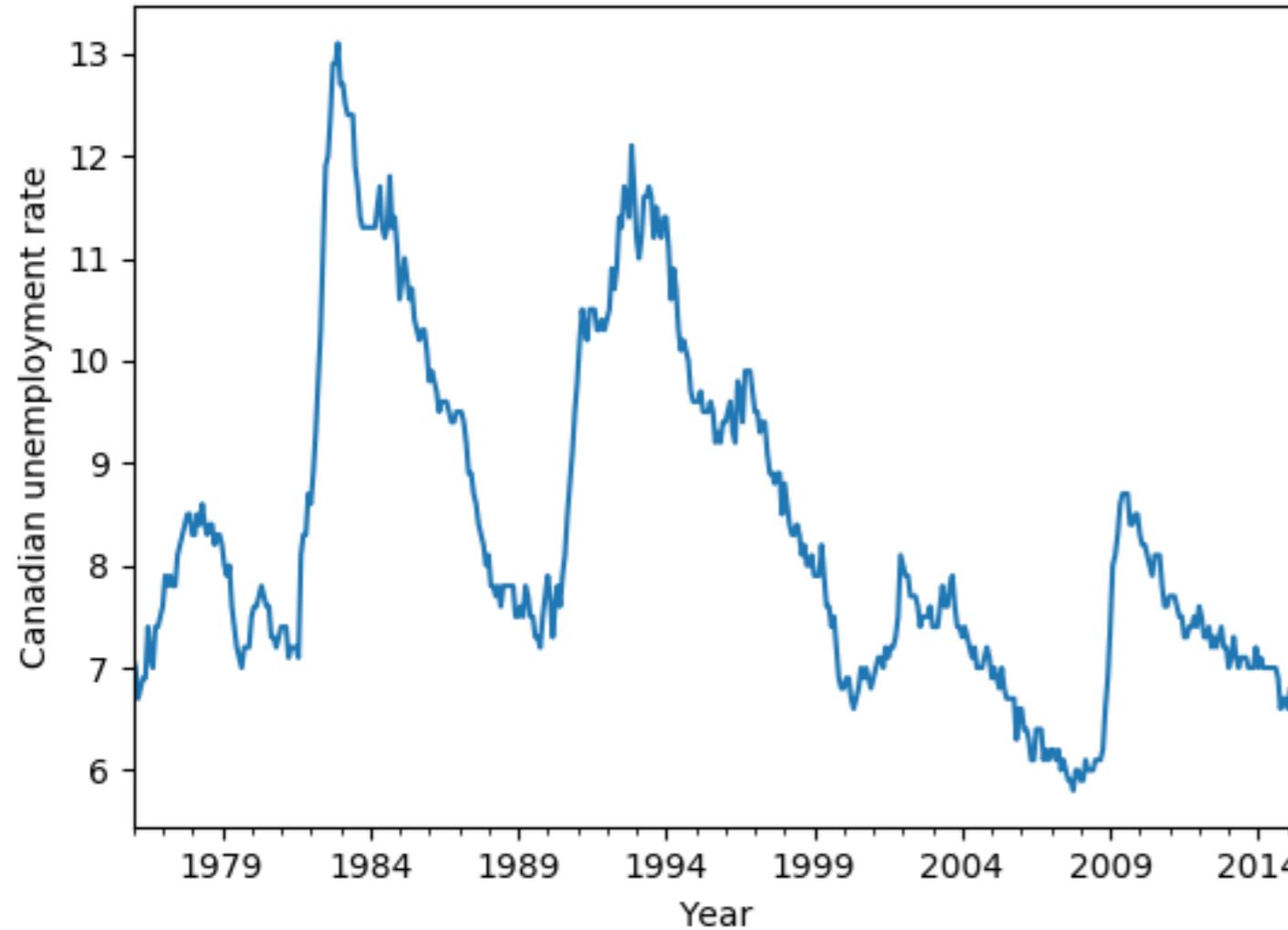
# Time series data

A series of data points sequenced by time

- Stock prices
- Gas prices
- Unemployment rates
- Heart rate
- CO2 levels
- Height of ocean tides

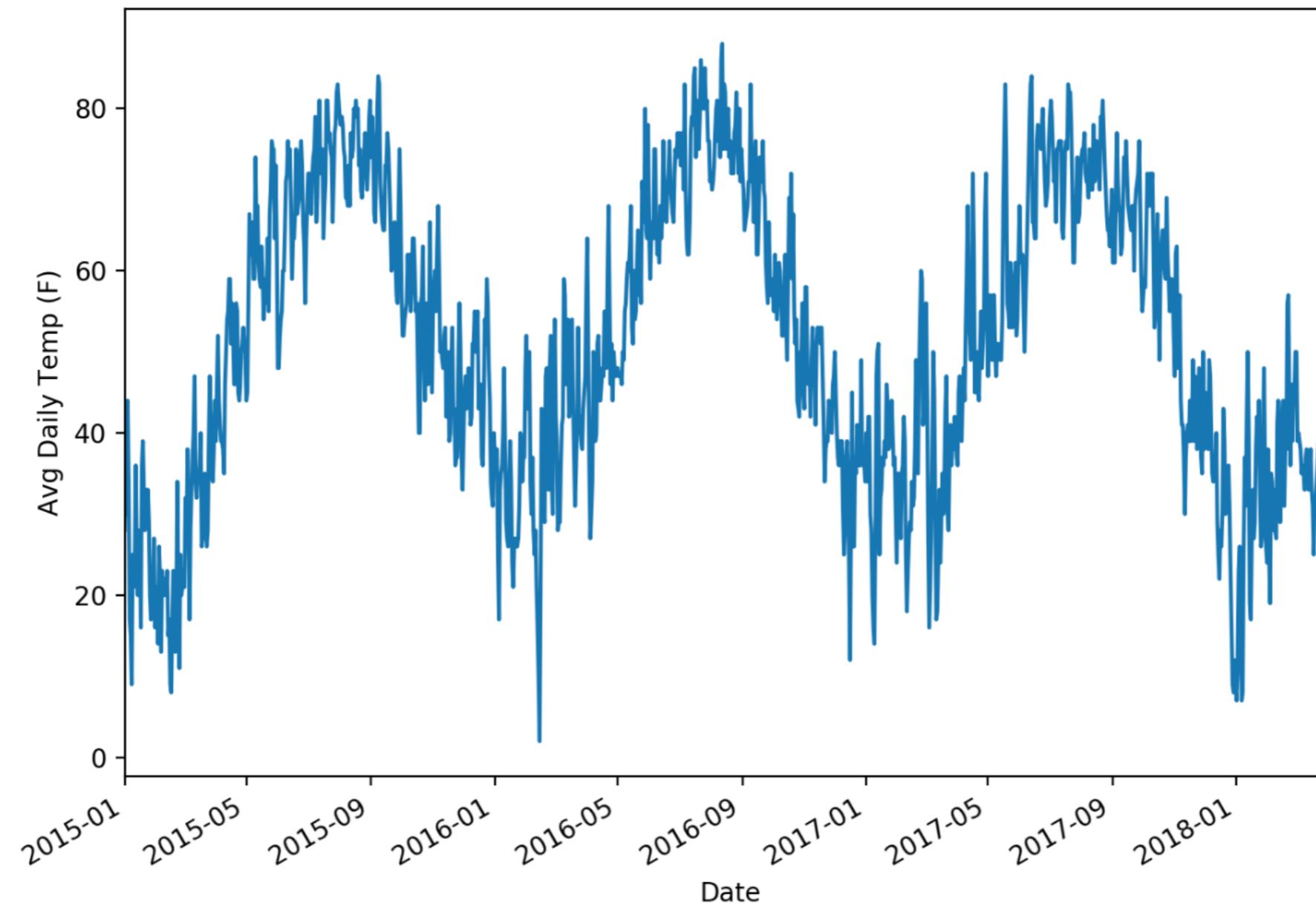


# Plotting time series data

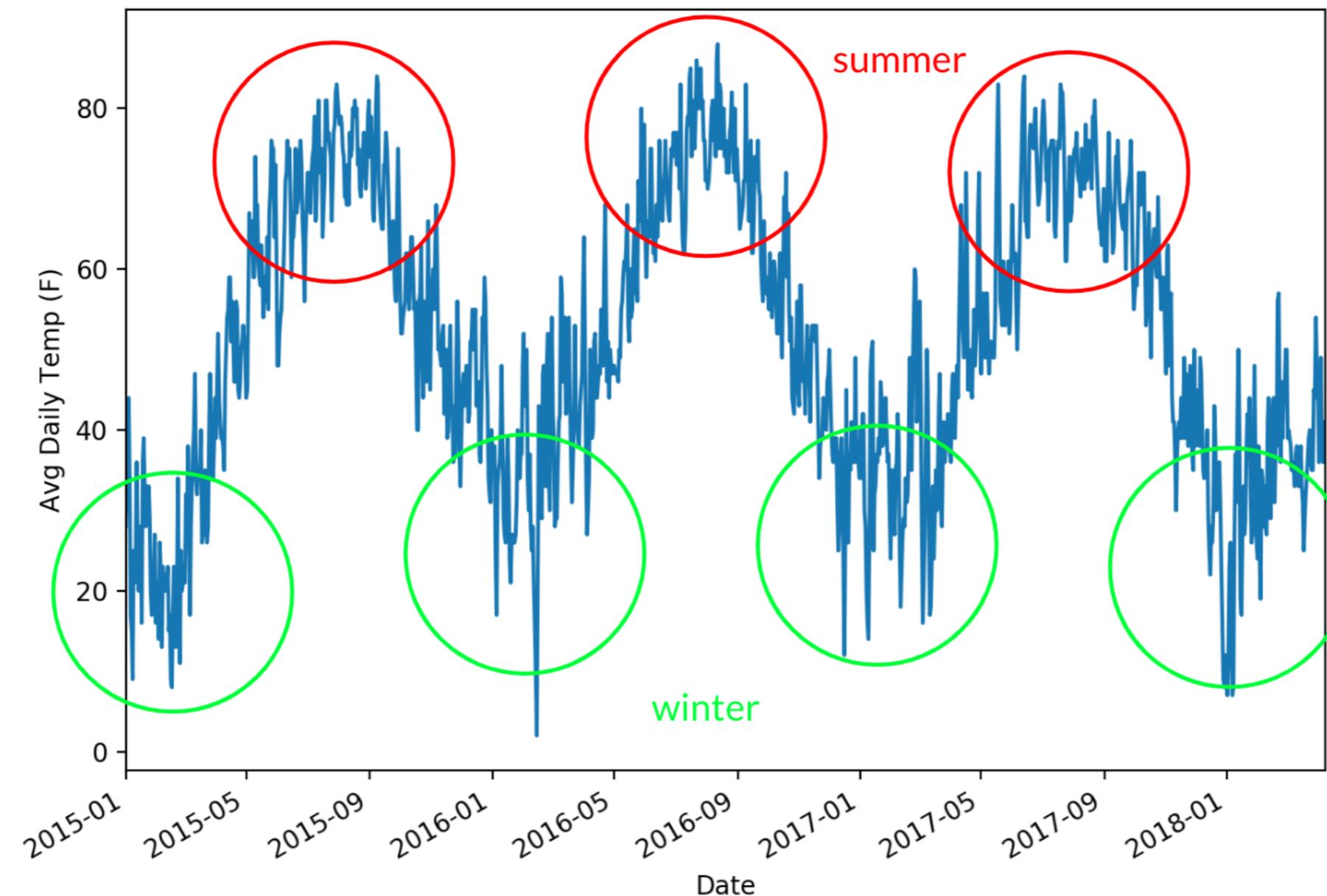


Date	Rate
1976-01-01	7.1
1976-02-01	7
...	
1991-01-01	10.3
...	
2015-04-01	6.8
2015-05-01	6.8

# Seasonality in time series



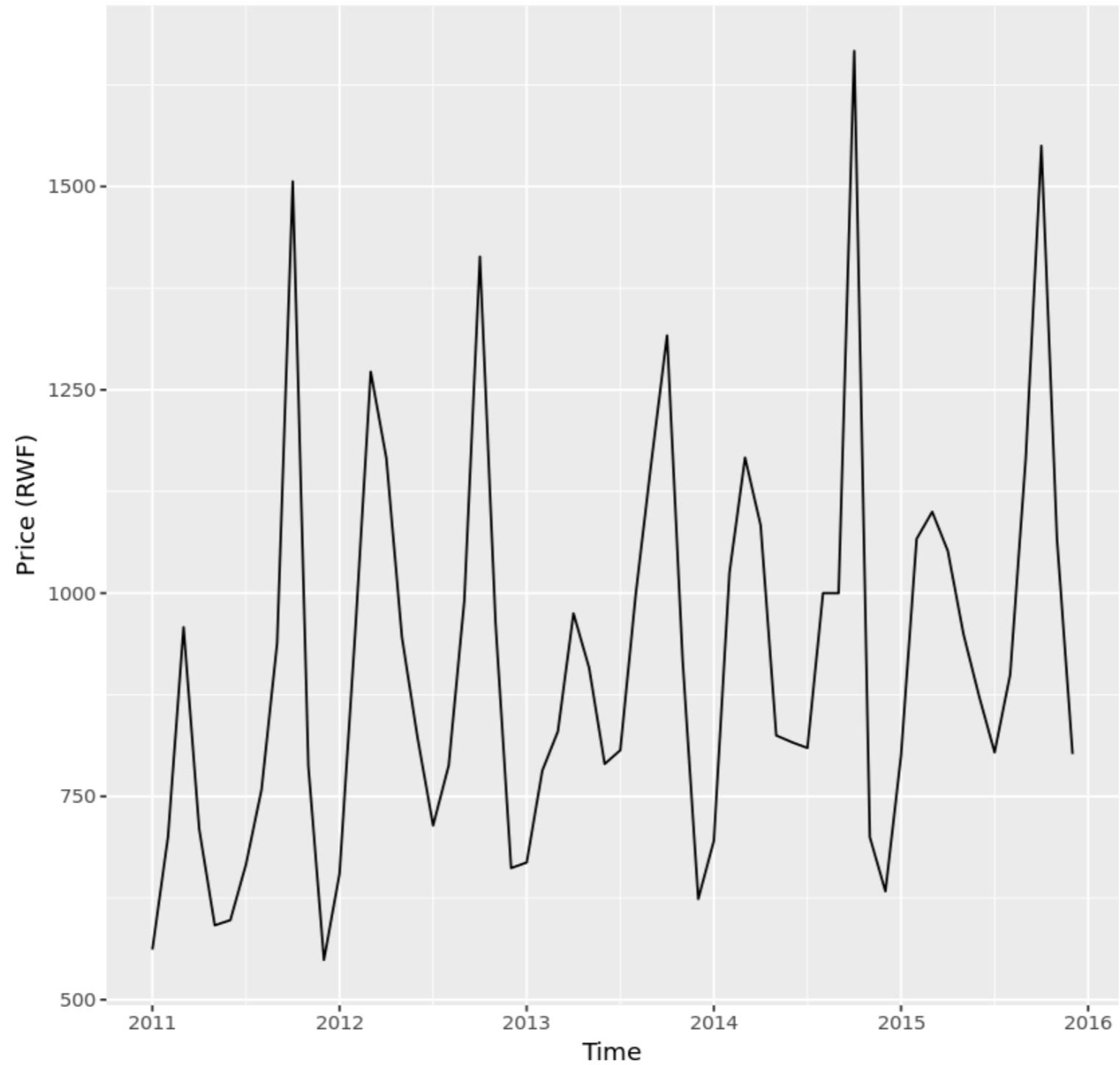
# Seasonality in time series



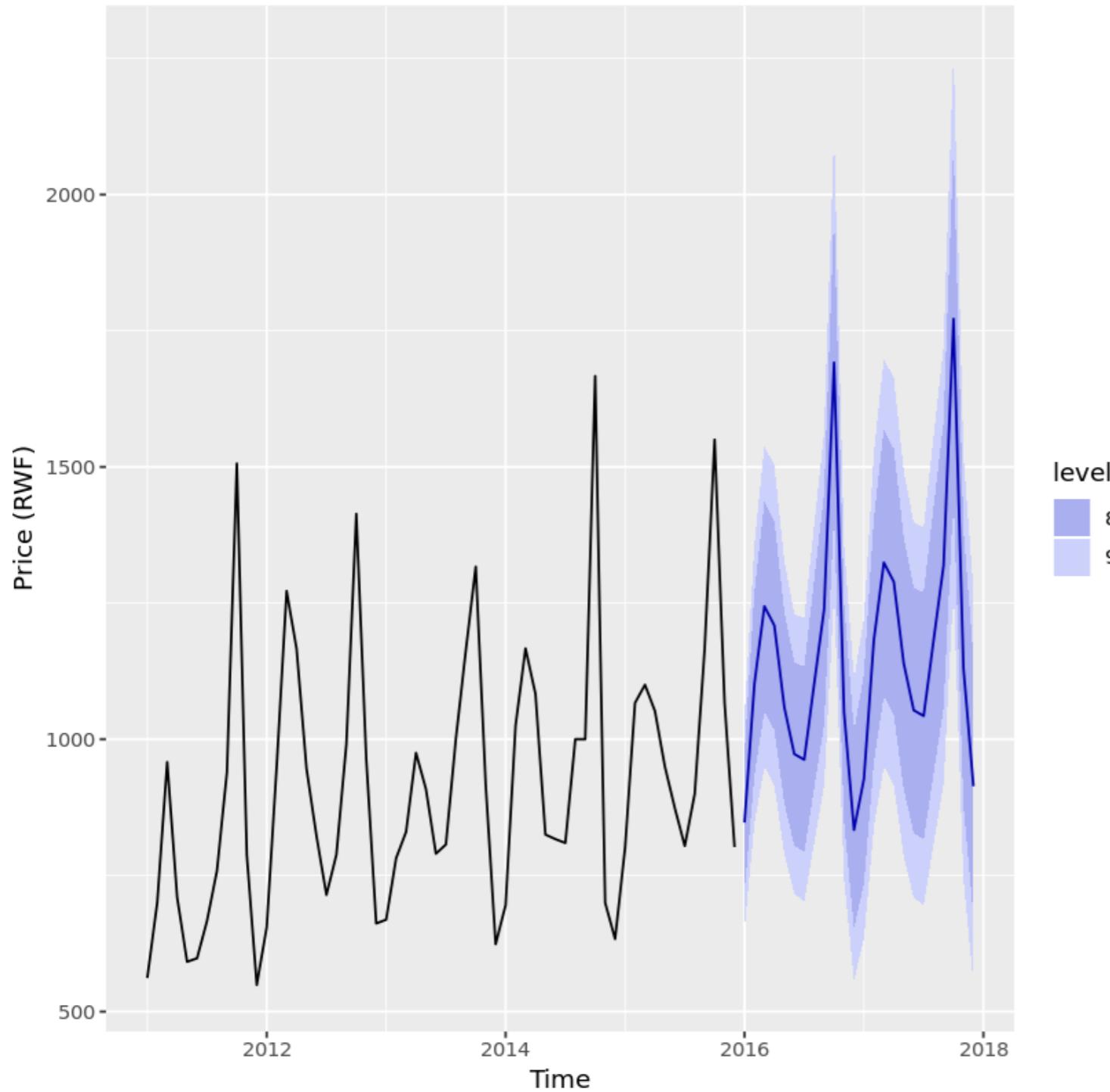
# Forecasting time series

- *Examples:*
  - How much rainfall will we get next month?
  - Will traffic ease up in the next half hour?
  - How will the stock market move in the next six hours?
  - What will be earth's population in 20 years?
- Derive a model from historical data to generate predictions
- Modeling methods use a combination of statistical and machine learning methods

## Pea Prices in Rwanda



## Pea price forecast



## Confidence intervals

Model is X% sure that the true value will fall in this area

level  
80  
95

# **Let's practice!**

**UNDERSTANDING DATA SCIENCE**

# Supervised machine learning

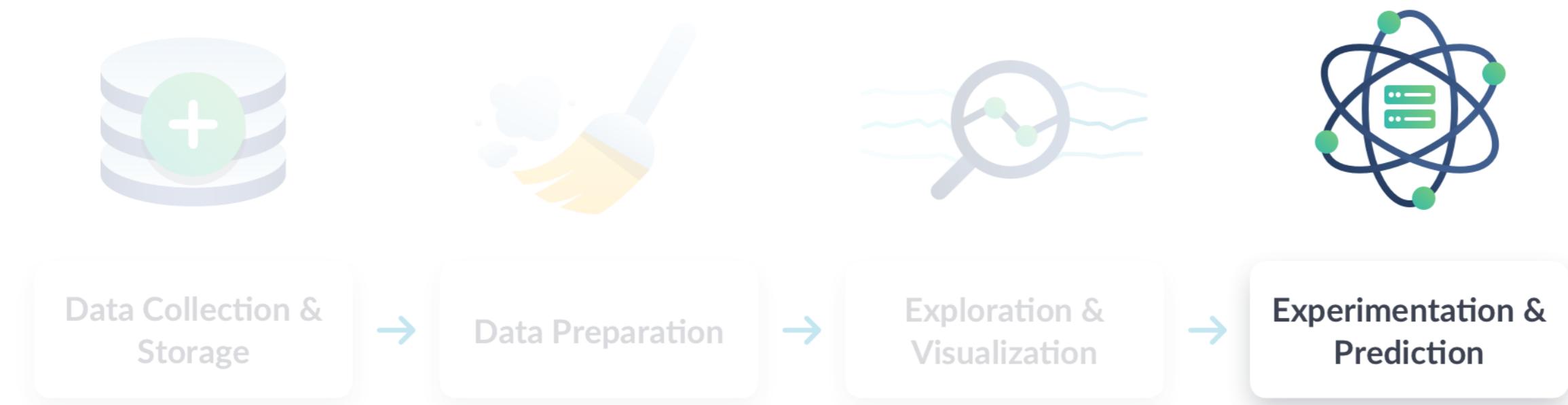
UNDERSTANDING DATA SCIENCE



Lis Sulmont

Curriculum Manager, DataCamp

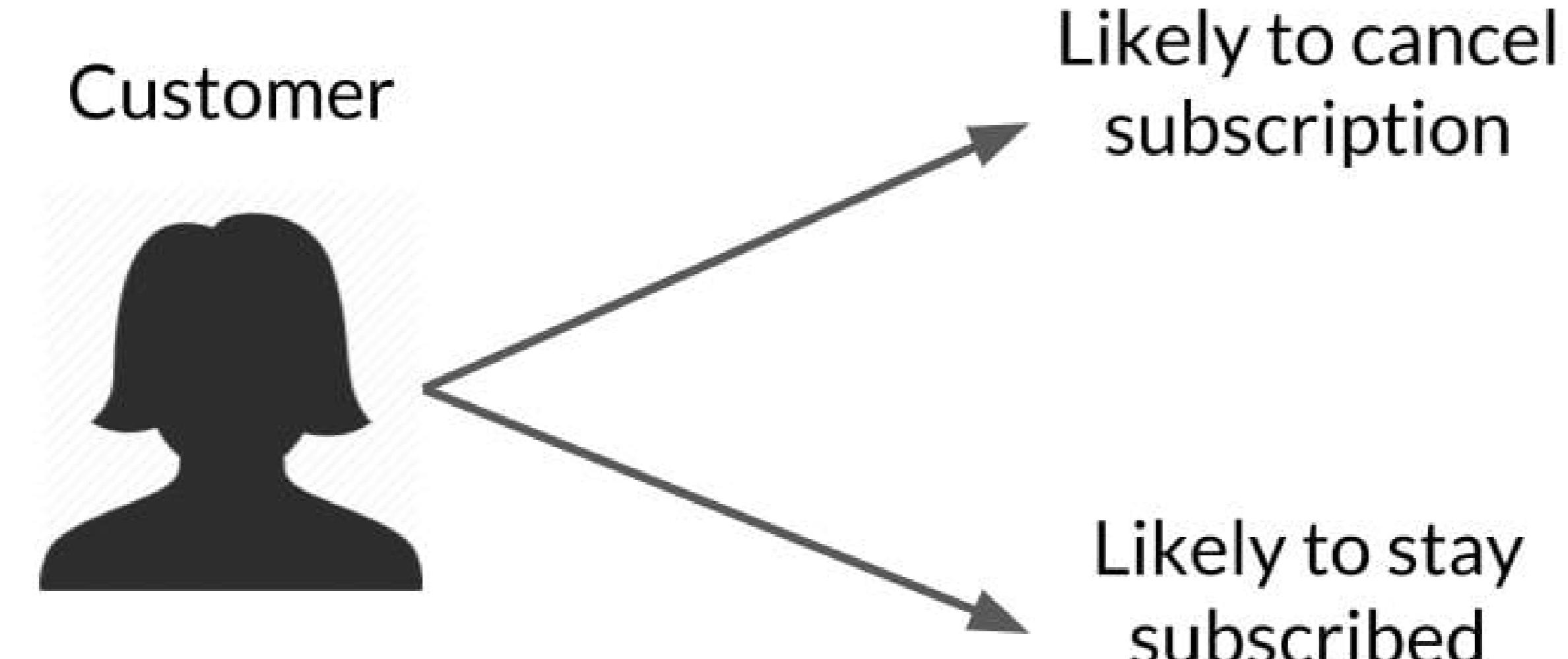
# Data science workflow



# What is supervised machine learning?

- Machine learning: Predictions from data
- *Supervised machine* learning: Predictions from data with *labels* and *features*
  - Recommendation systems
  - Diagnosing biomedical images
  - Recognizing hand-written digits
  - Predicting customer churn

# Case study: churn prediction



# Case study: churn prediction

Training  
Data:  
Customers



# Case study: churn prediction



# Case study: churn prediction

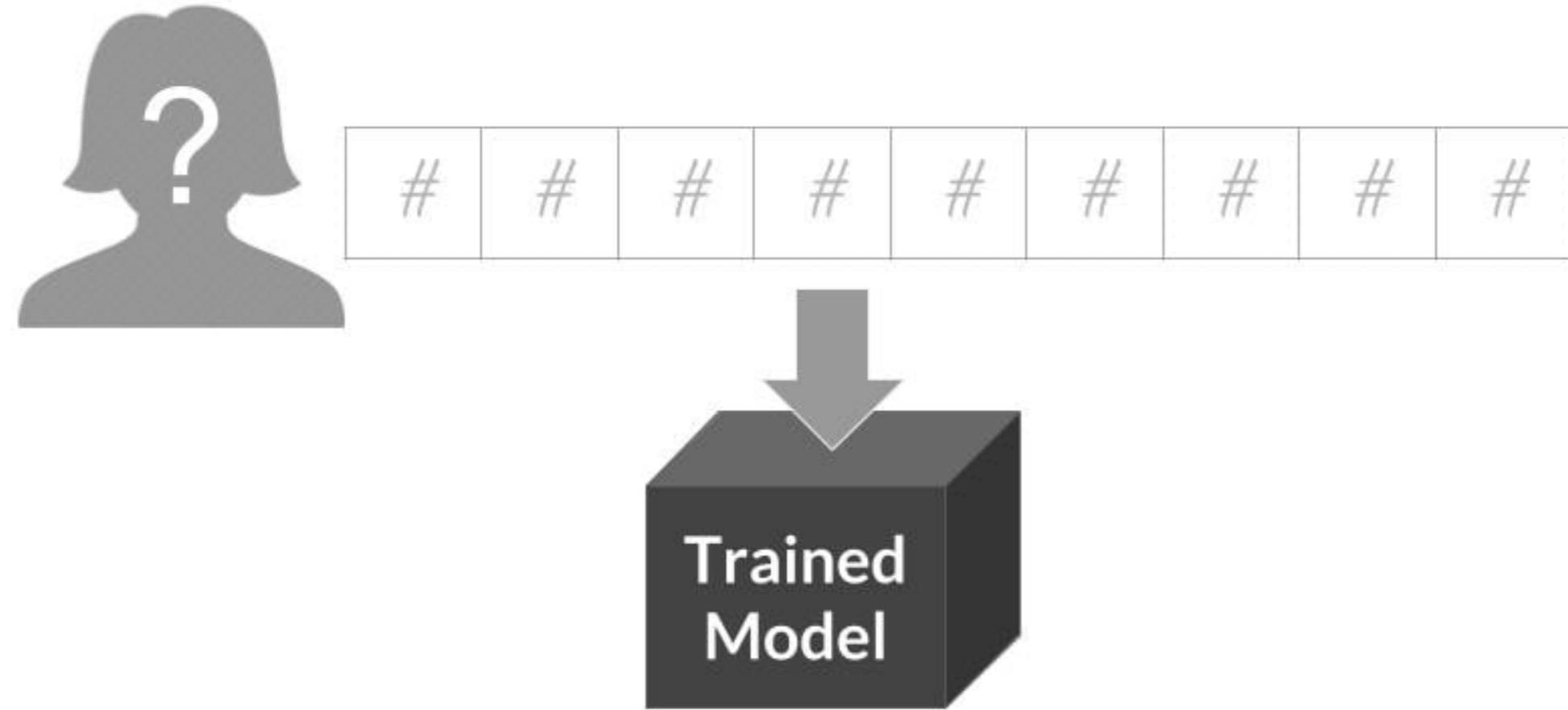
	Age	Gender	Date of last purchase?	Date of last visit?	Likes cats?	Household \$\$	Location	Number of Kids	Profession	
										churn
										subscribe
										subscribe
										churn
										subscribe
										churn

# Case study: churn prediction

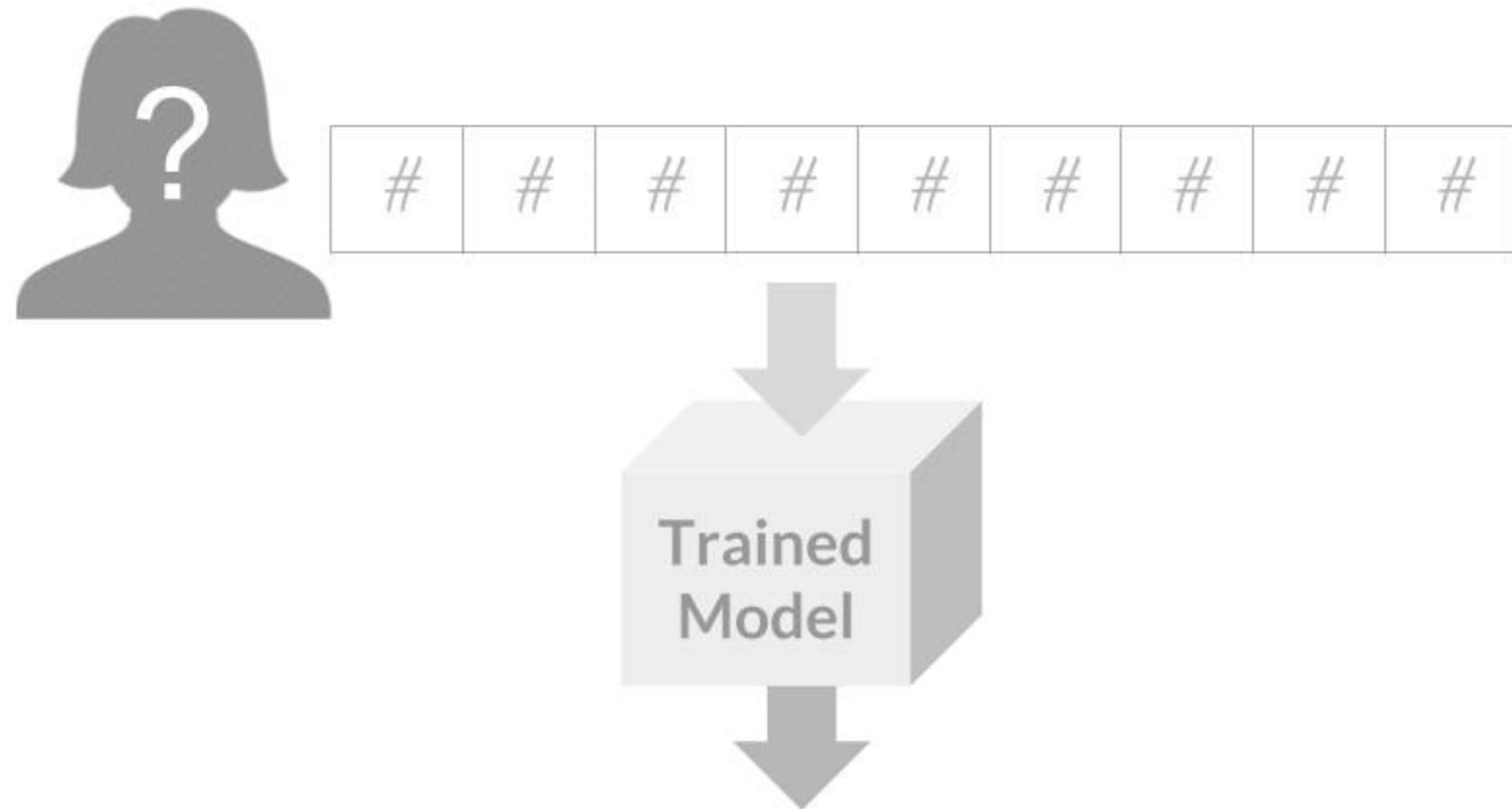


#	#	#	#	#	#	#	#	#
---	---	---	---	---	---	---	---	---

# Case study: churn prediction



# Case study: churn prediction



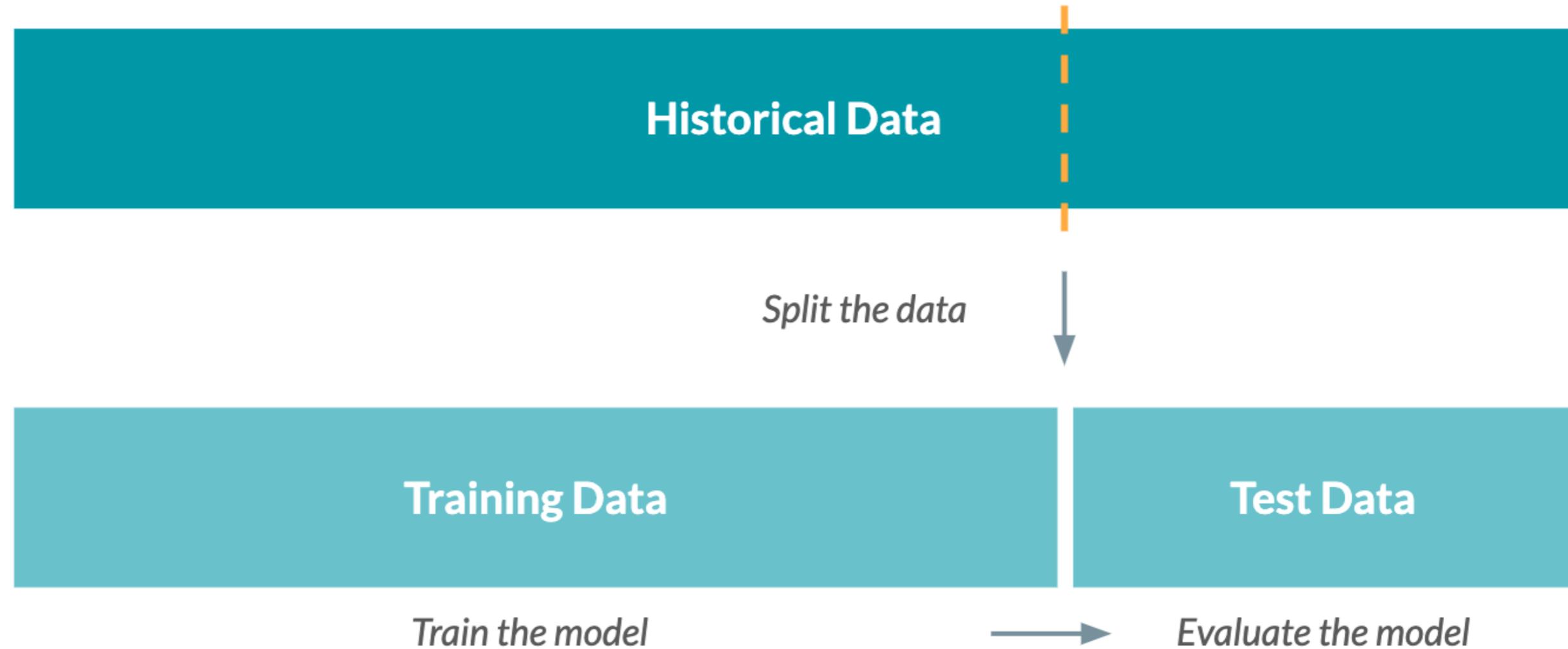
**Prediction: Subscribe**

# Supervised machine learning recap

- Make a prediction based on data
- Data has *features* and *labels*
  - Label: what we want to predict
  - Features: data that might predict the label
- Trained model can make predictions

# Model evaluation

Split historical data into training and testing sets



# Model Evaluation

Possible Labels	True Labels	Model Prediction	Model Accuracy
<i>Customer remains</i>	970		
<i>Customer churns</i>	30		

# Model Evaluation

Possible Labels	True Labels	Model Prediction	Model Accuracy
<i>Customer remains</i>	970	1000	
<i>Customer churns</i>	30	0	

# Model Evaluation

Possible Labels	True Labels	Model Prediction	Model Accuracy
<i>Customer remains</i>	970	1000	# of correct predictions / # of predictions =
<i>Customer churns</i>	30	0	$970/1000 =$ <b>97%</b>

- Checking both outcomes is important for rare events
- Model has 0% accuracy at predicting an actual churn outcome

# **Let's practice!**

**UNDERSTANDING DATA SCIENCE**

# Clustering

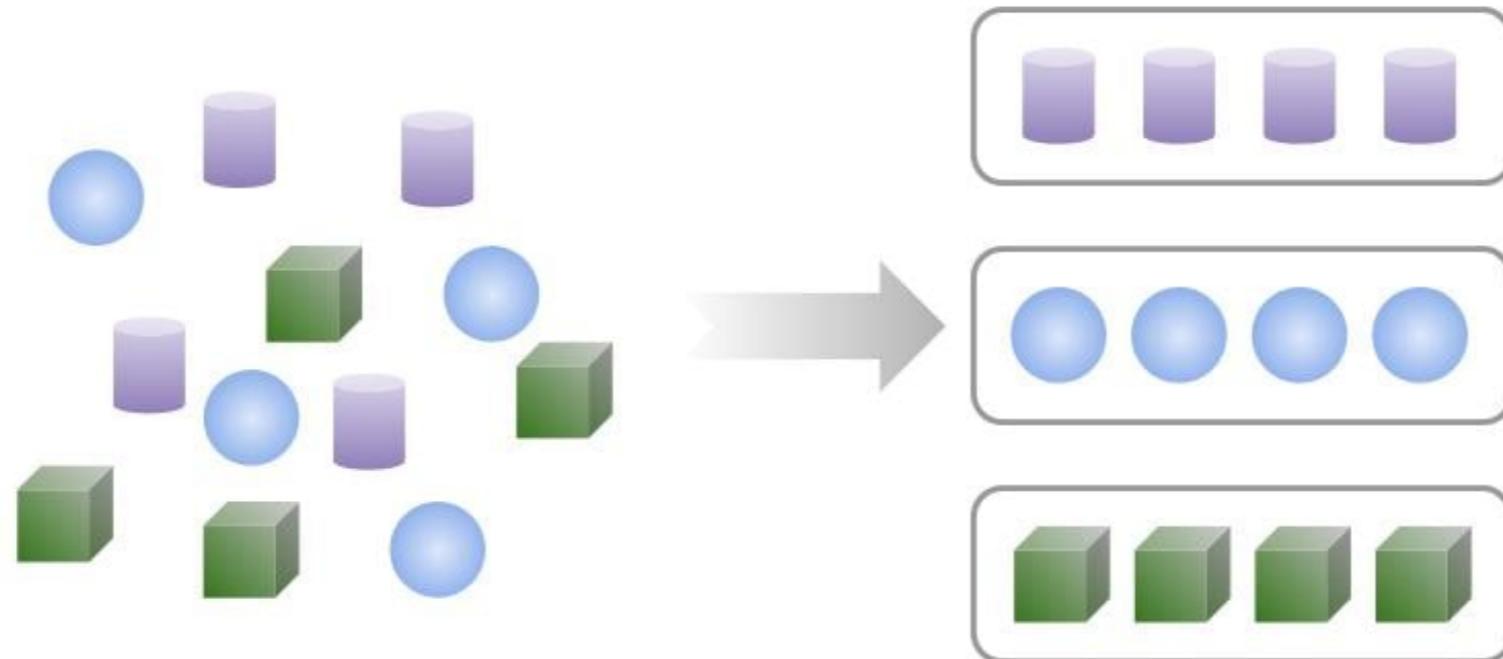
UNDERSTANDING DATA SCIENCE



Lis Sulmont

Curriculum Manager, DataCamp

# What is clustering?



- Divide data into categories
- Use cases
  - Customer segmentation
  - Image segmentation
  - Anomaly detection

# Supervised Machine Learning



# Unsupervised Machine Learning

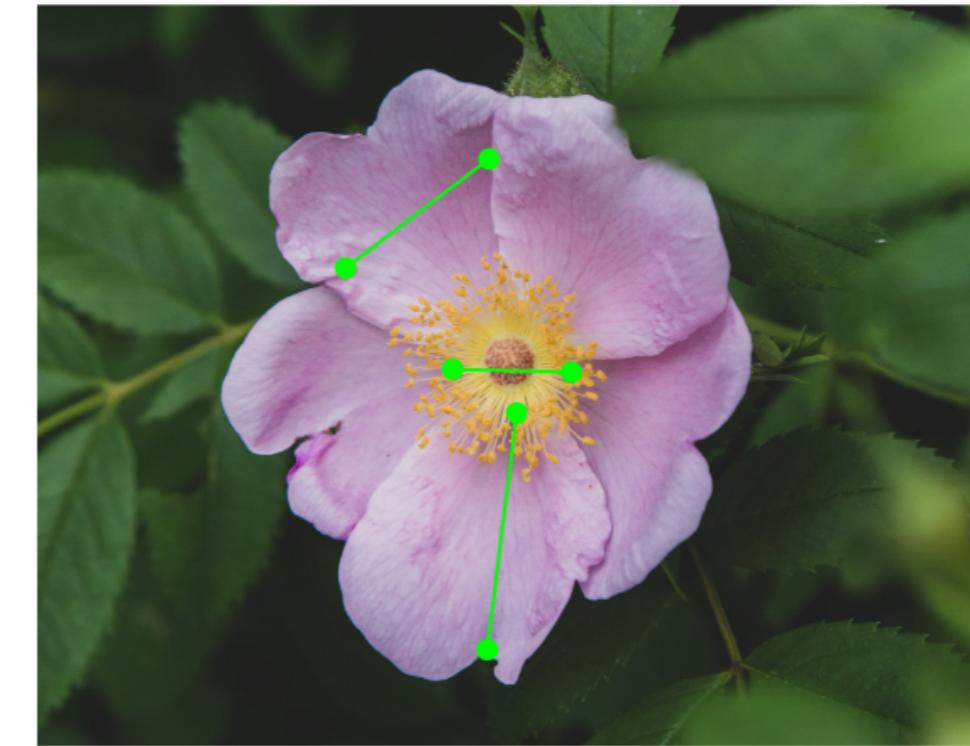


# Case study: discovering new species

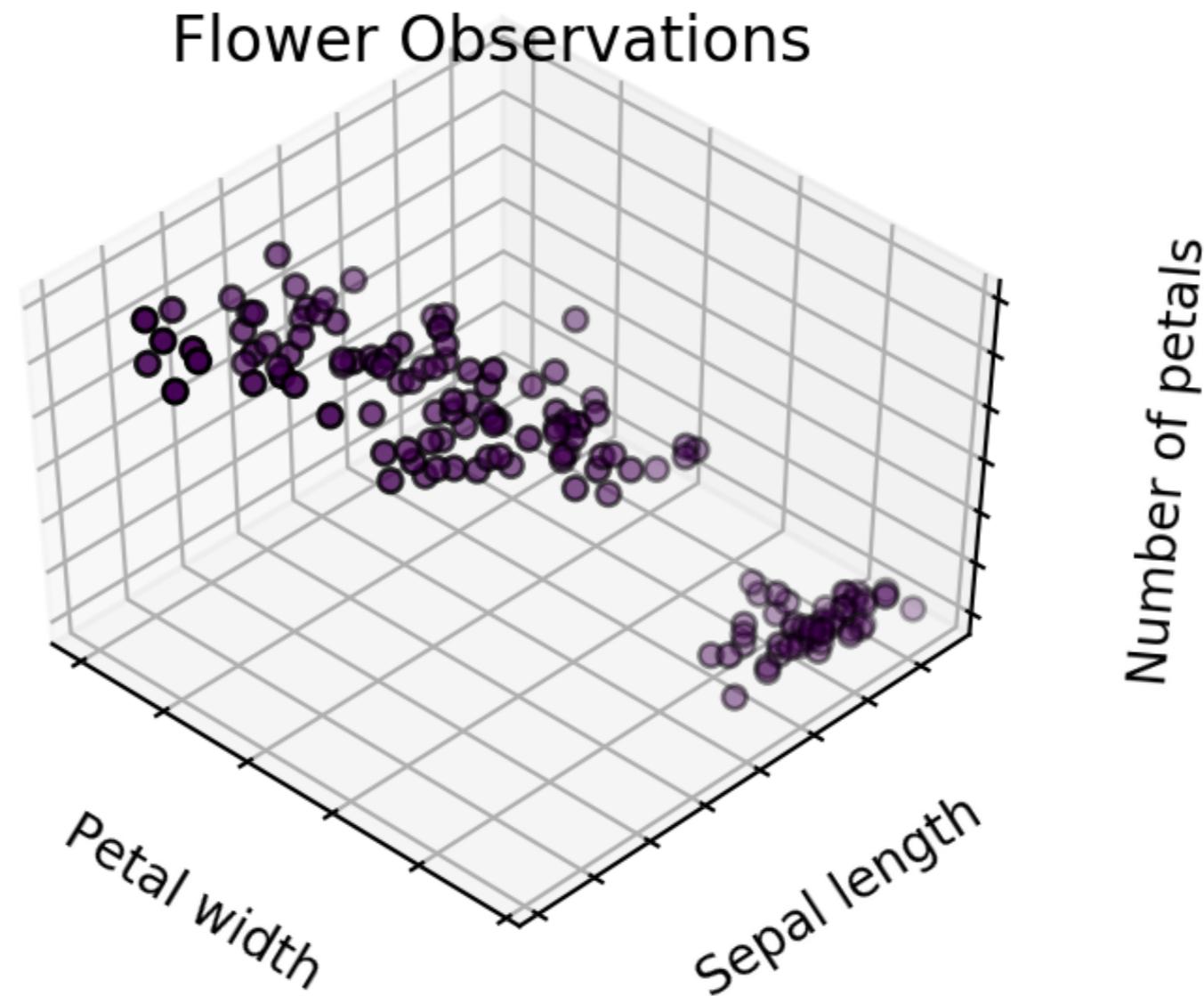


# Defining features

- Flower colors
- Petal length and width
- Sepal length and width
- Number of petals

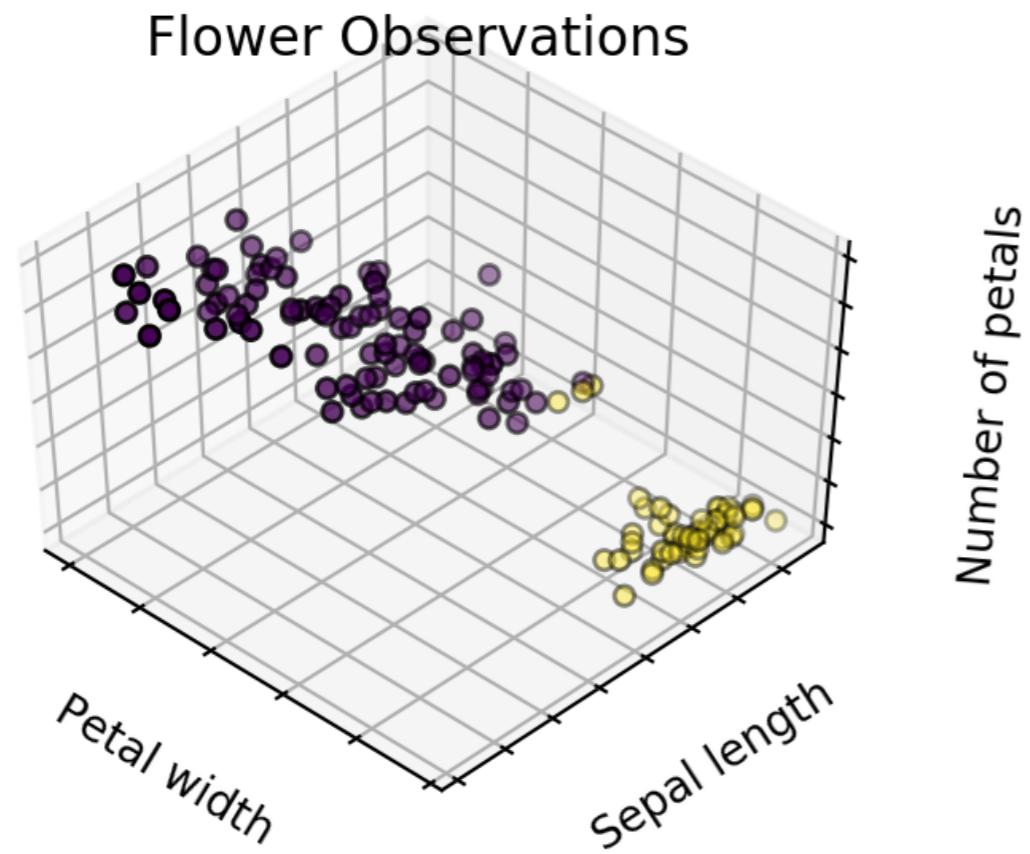


# Defining number of clusters

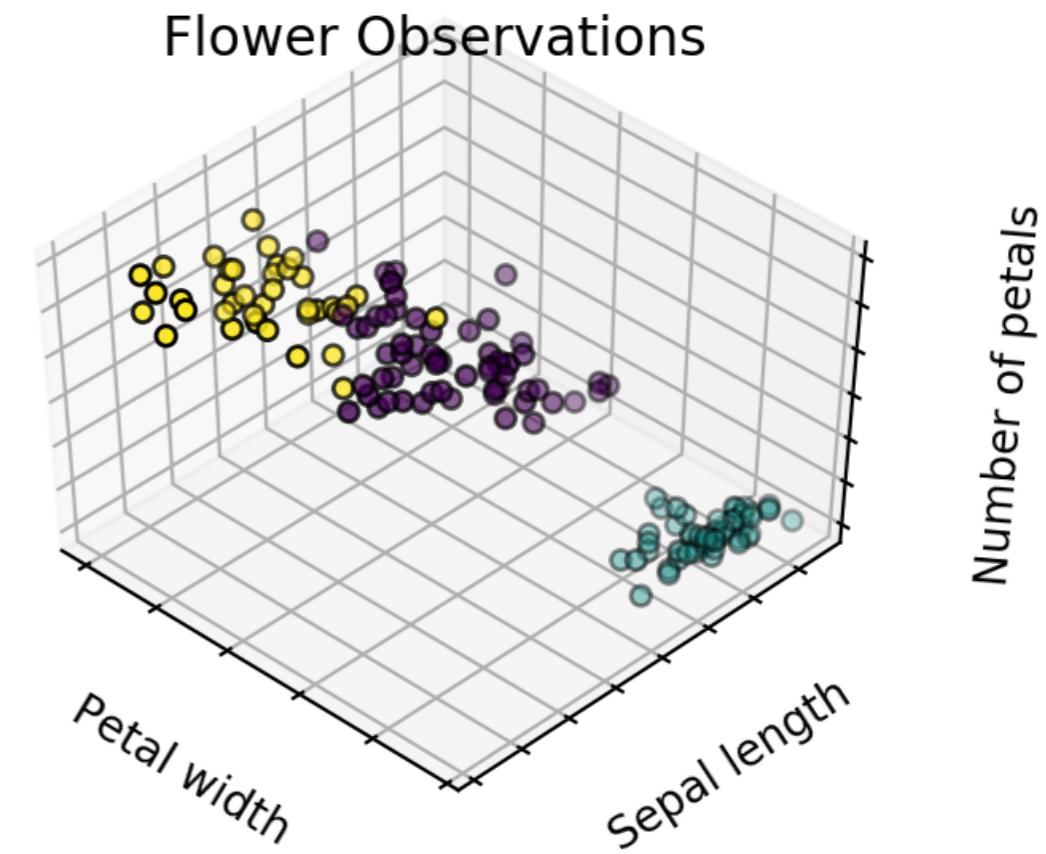


# Comparing number of clusters

Two clusters:

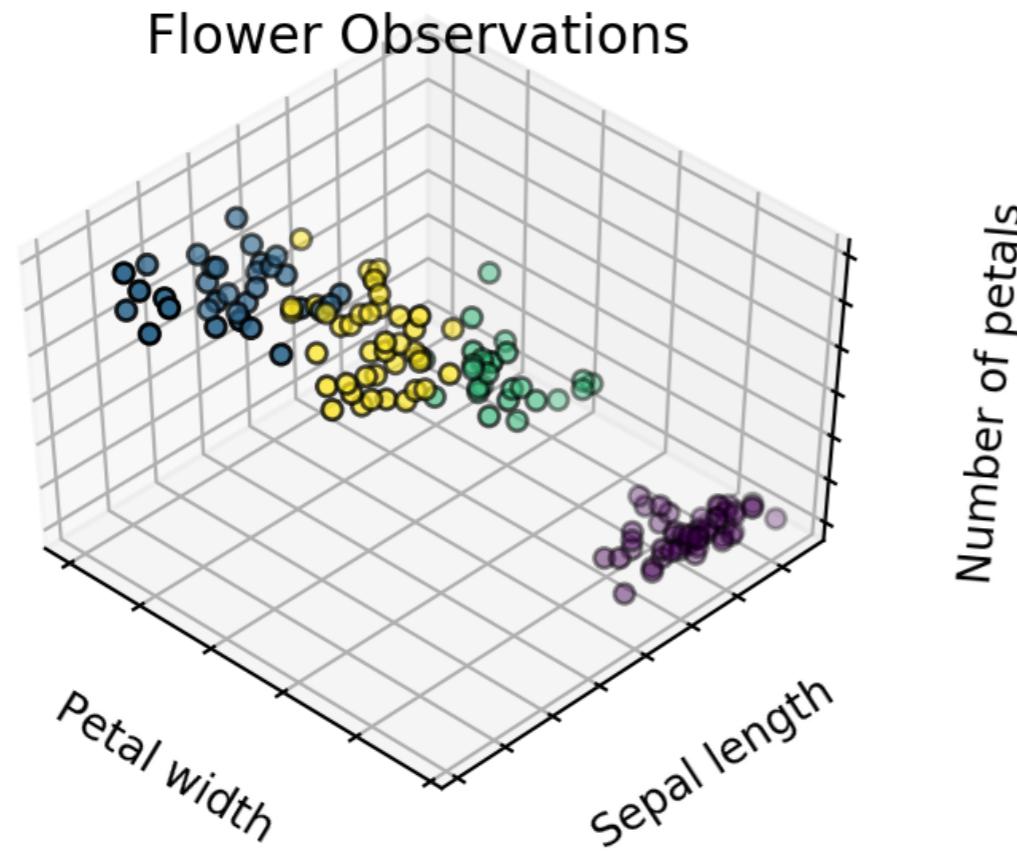


Three clusters:

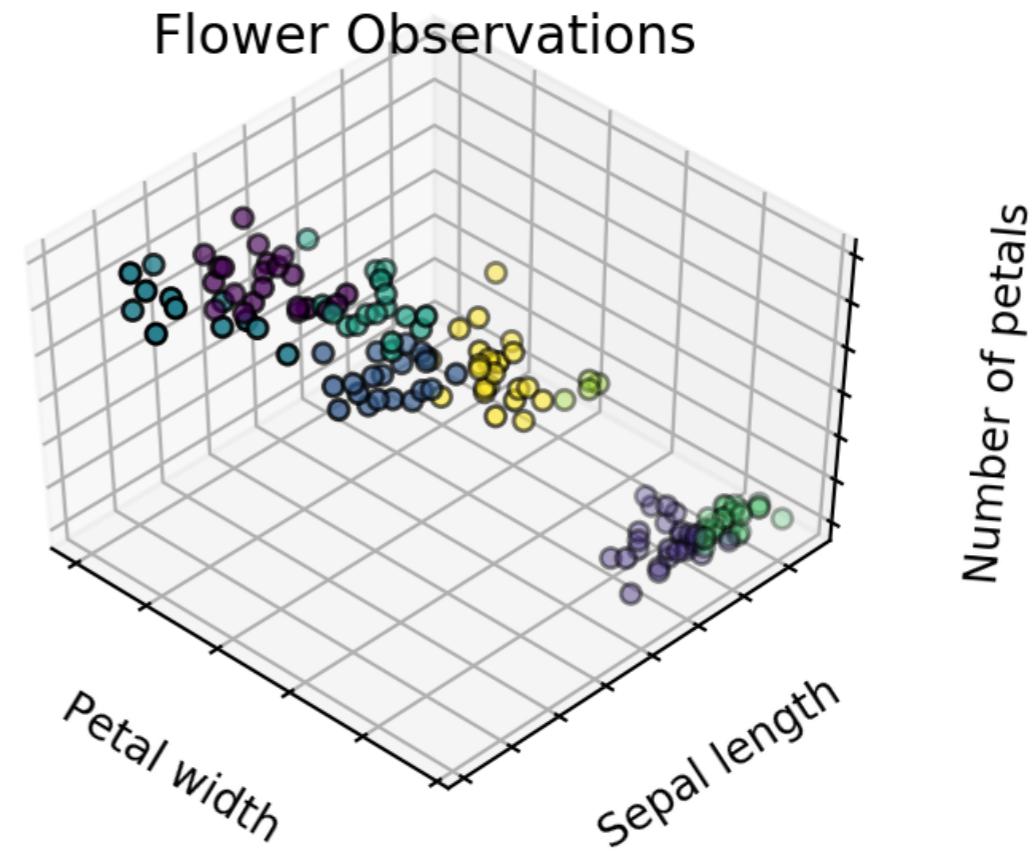


# Comparing number of clusters

Four clusters:



Eight clusters:



# Comparing number of clusters

- Up to you to decide on final number of clusters
- Use domain knowledge to help decide

# Clustering review

## Definition

- Divide unlabeled dataset into different categories

## Steps

- Select features
- Select number of clusters
- Use clusters to solve problems

# **Let's practice!**

**UNDERSTANDING DATA SCIENCE**

# Congratulations!

UNDERSTANDING DATA SCIENCE



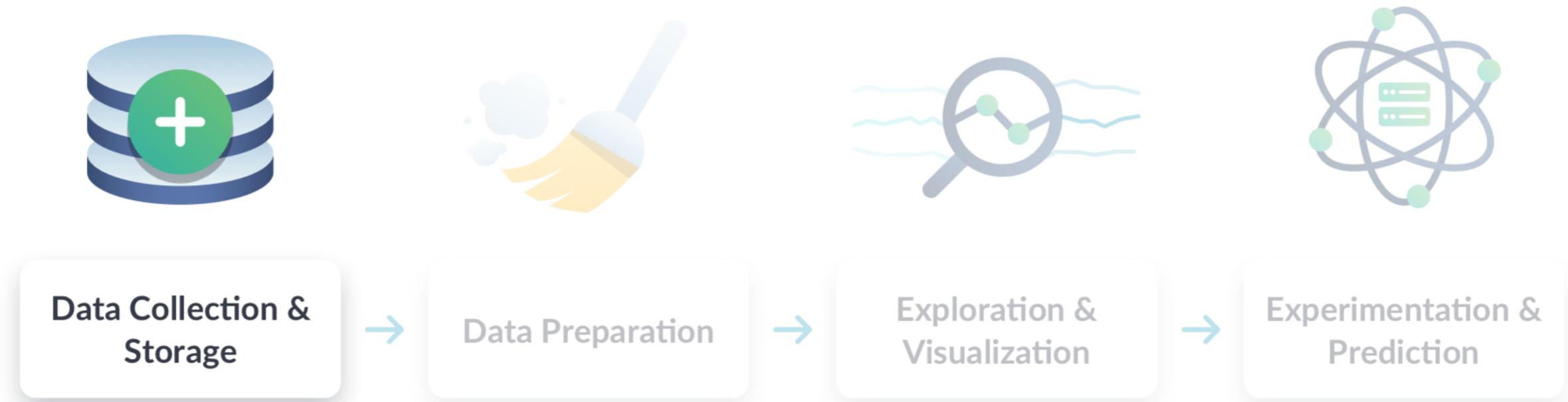
Lis Sulmont

Curriculum Manager, DataCamp

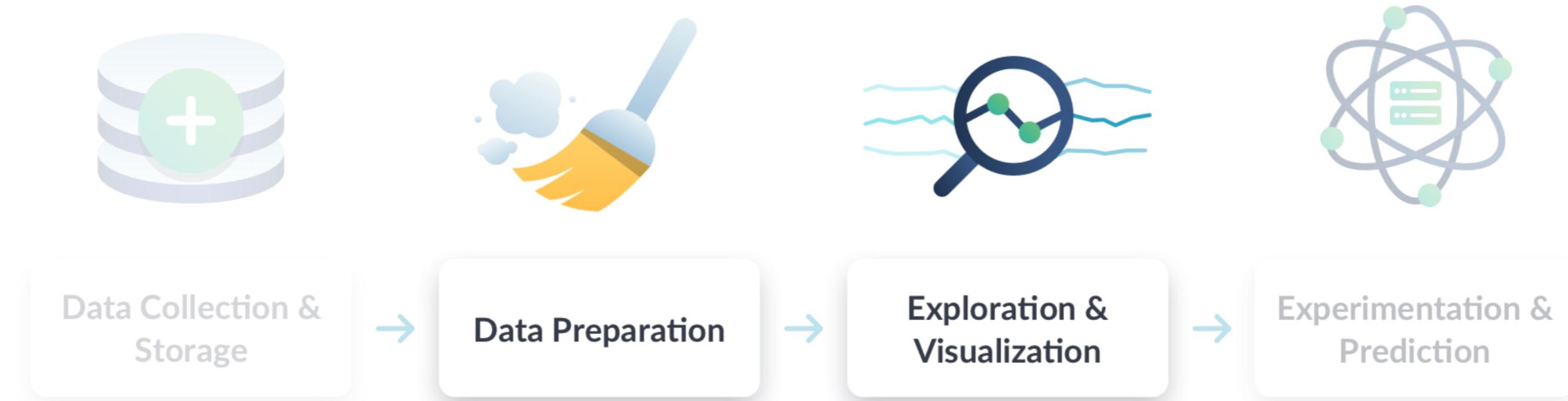
# Chap 1: Introduction to Data Science



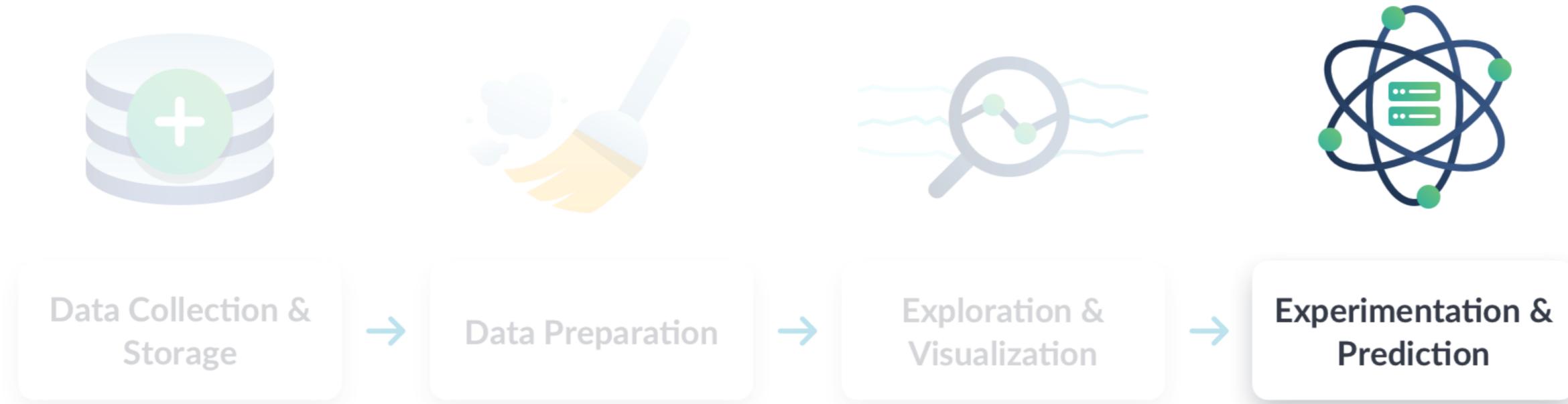
# Chap 2: Data Collection and Storage



# Chap 3: Preparation, Exploration & Visualization



# Chap 4: Experimentation and Prediction



# What's next?

- **Introduction to Python**
  - Code your first lines of Python to do data science
- **Data Literacy Fundamentals Skill Track**
  - More "For Everyone" courses on other data topics



# Congratulations!

UNDERSTANDING DATA SCIENCE