Data Science

Mary Newhauser

Agenda

Intro to data science

About me

Python basics

Exploratory data analysis in Python

Linear regression in Python

The cool stuff…

What is data science?

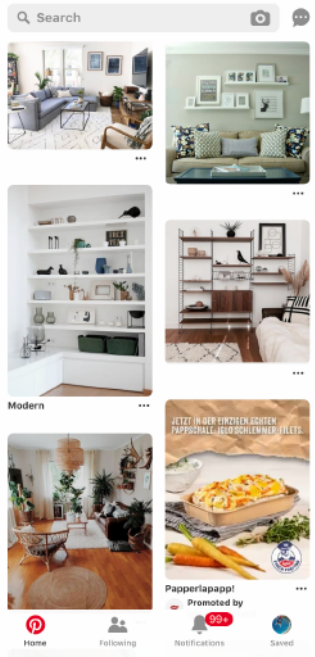
Data science is…

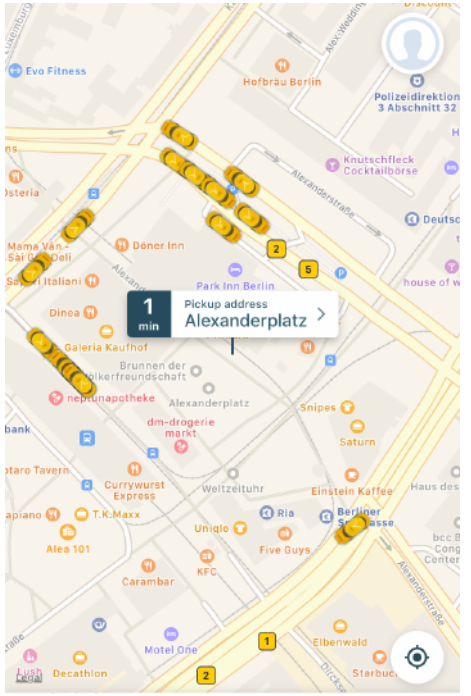
Transforming data into insights (usually using code).

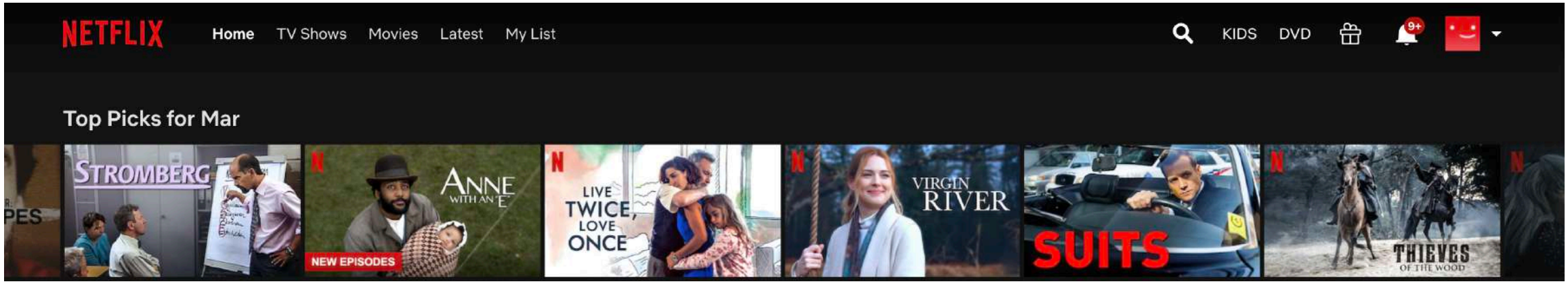
Data

Insights

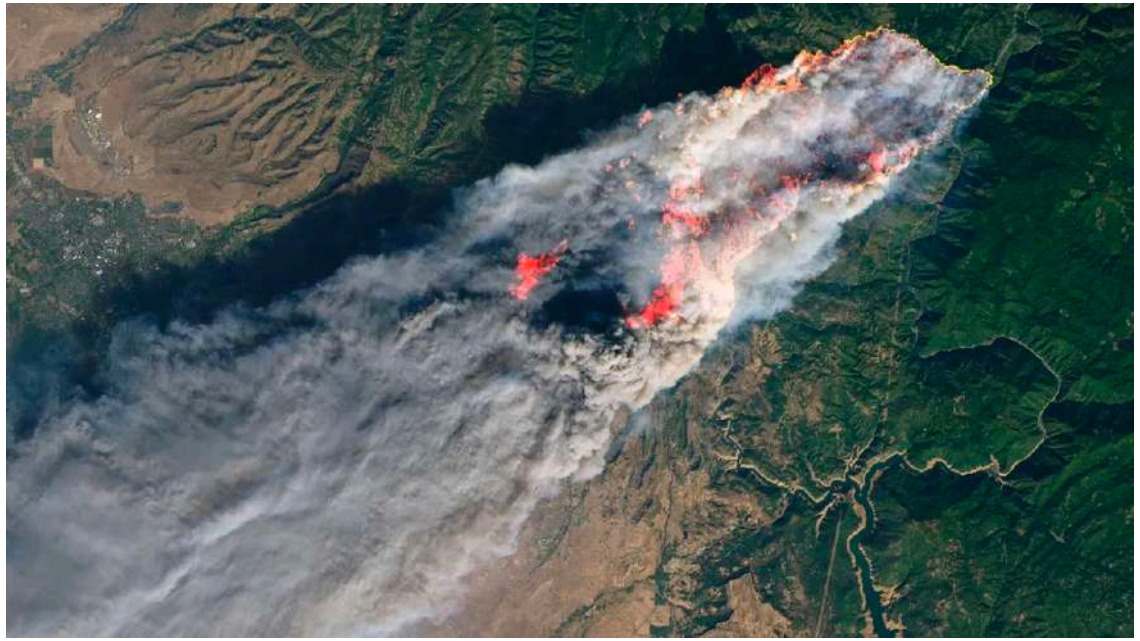
Data science in everyday life

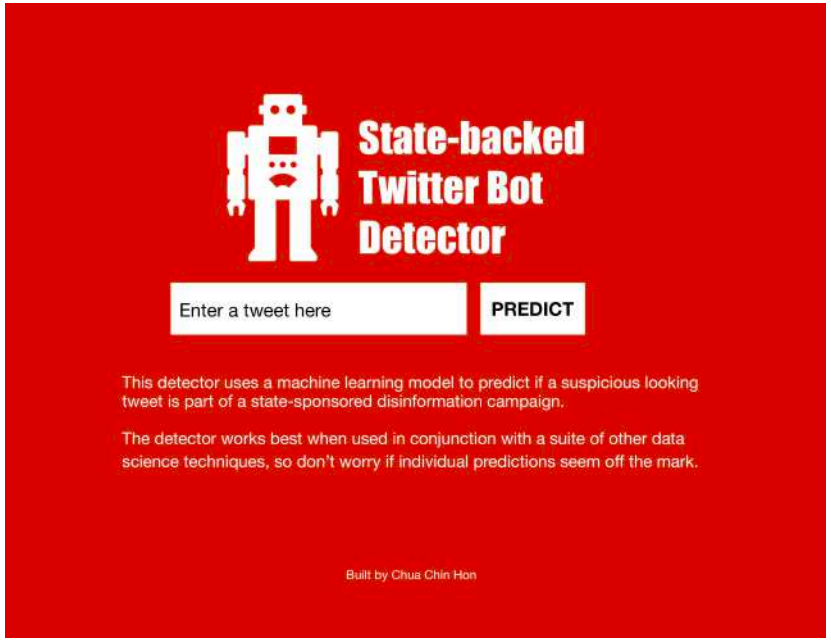


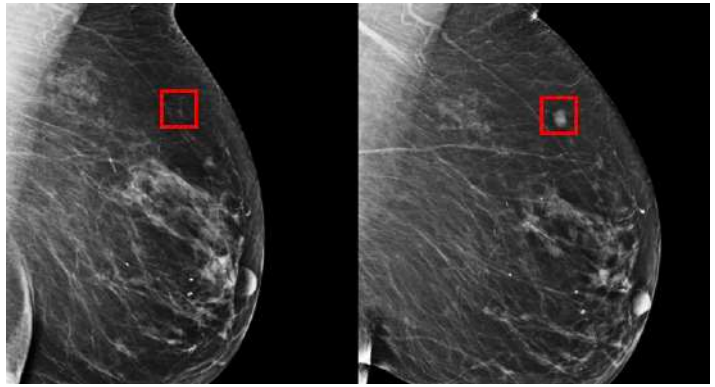


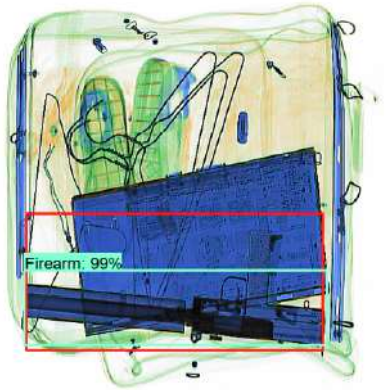


Data science for good









Data science

Predictive modeling

Using statistics to predict future outcomes

- Algorithms
- The fun stuff!

Exploratory data analysis (EDA)

Summarizing main characteristics of a dataset

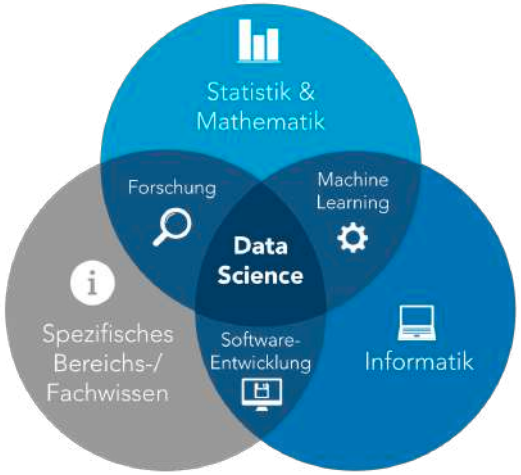
- Descriptive statistics (mean, median, mode)
- Percentages %
- Graphs

What is a data scientist?

A data scientist is…

Someone who translates data into actionable insights.

Part computer scientist,
part mathematician,
part statistician.



Fluent in statistical
programming languages.

python™

R

A PROBLEM SOLVER!

Data scientist skillset

Technical

- **Python**
- R
- SQL
- Database architecture

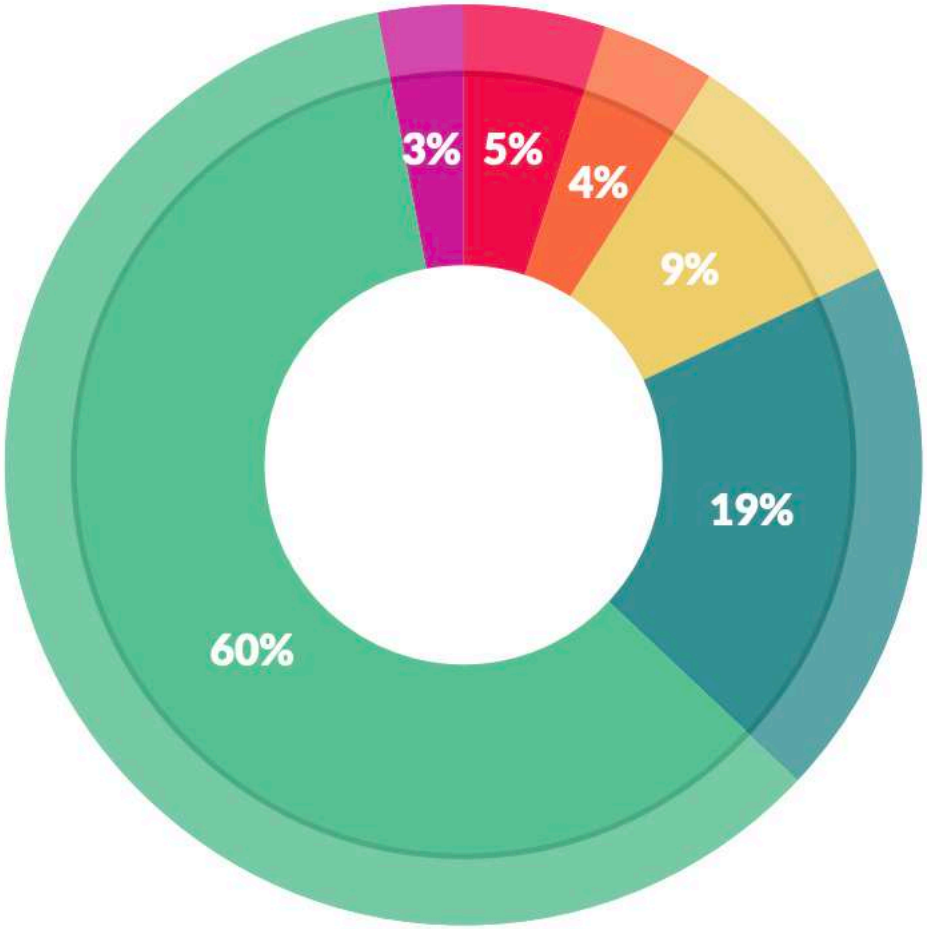
Academic

- Statistics
- Math

Other

- Communication, presentation
- Research
- **Critical thinking**
- Patience
- Data visualization

What does a data scientist do?



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

Source: CrowdFlower

Perks of being a data scientist

High salary

Solve cool problems

Job security

Work from anywhere*

What do I do?

Data Scientist, Intelligent Solutions Team, Wiley-VCH

Hired January 2019

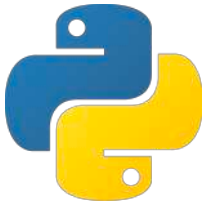
Help make the academic publishing process smarter

Build text classification models

Building a graph database

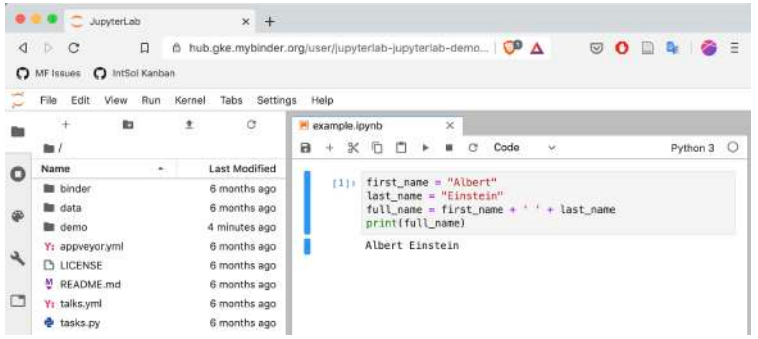
My workflow

The language I code in: Python

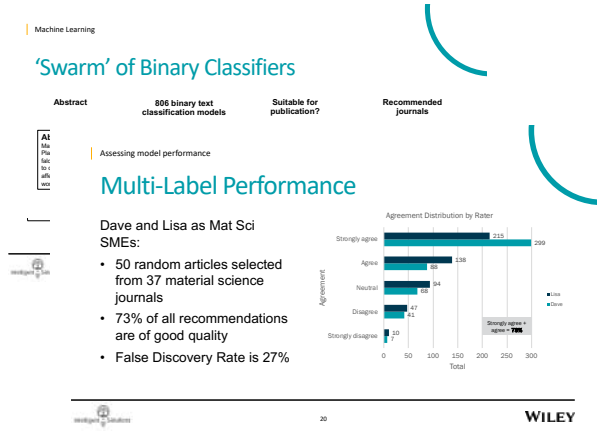


```
first_name = "Albert"
last_name = "Einstein"
full_name = first_name + last_name
print(full_name)
```

Where I write my code: Jupyter Notebook



How I present my results: PowerPoint



My side projects

Scraping and analyzing YouTube comments

Build model to identify people vulnerable to Fake News online

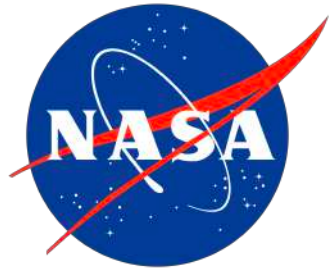
Python basics

What is Python?

A versatile, open source coding language that can be used for website and application development, analytics and much more.

Who uses Python?

Uber



Microsoft

reddit

Why learn Python?

Easy to learn

Readable

Open source

Versatile

Community

Job demand

Basics

Data types

Functions

Packages

Comments

Data types

Strings

```
player = 'Lionel Messi'

print(player)
```

```
Lionel Messi
```

Integers

```
career_goals = 690

print(career_goals)
```

```
690
```

Data types

Lists

- List of related data
- Contain strings or integers

Data types

Dictionaries

- Paired or related data
- Variety of data types


```
{'player': 'Lionel Messi', 'national_team': 'Argentina', 'club': 'FC Barcelona', 'career_goals': 690}
```

Data types

Data frames

- Spreadsheet data
- Variety of data types
- Make to use the Pandas package

Dictionaries

Package

Function

```
# Import Pandas package
import pandas as pd

# Use Pandas' DataFrame function to create a DataFrame
pd.DataFrame([lionel_messi, cristiano_ronaldo, zlatan_ibrahimovic])
```

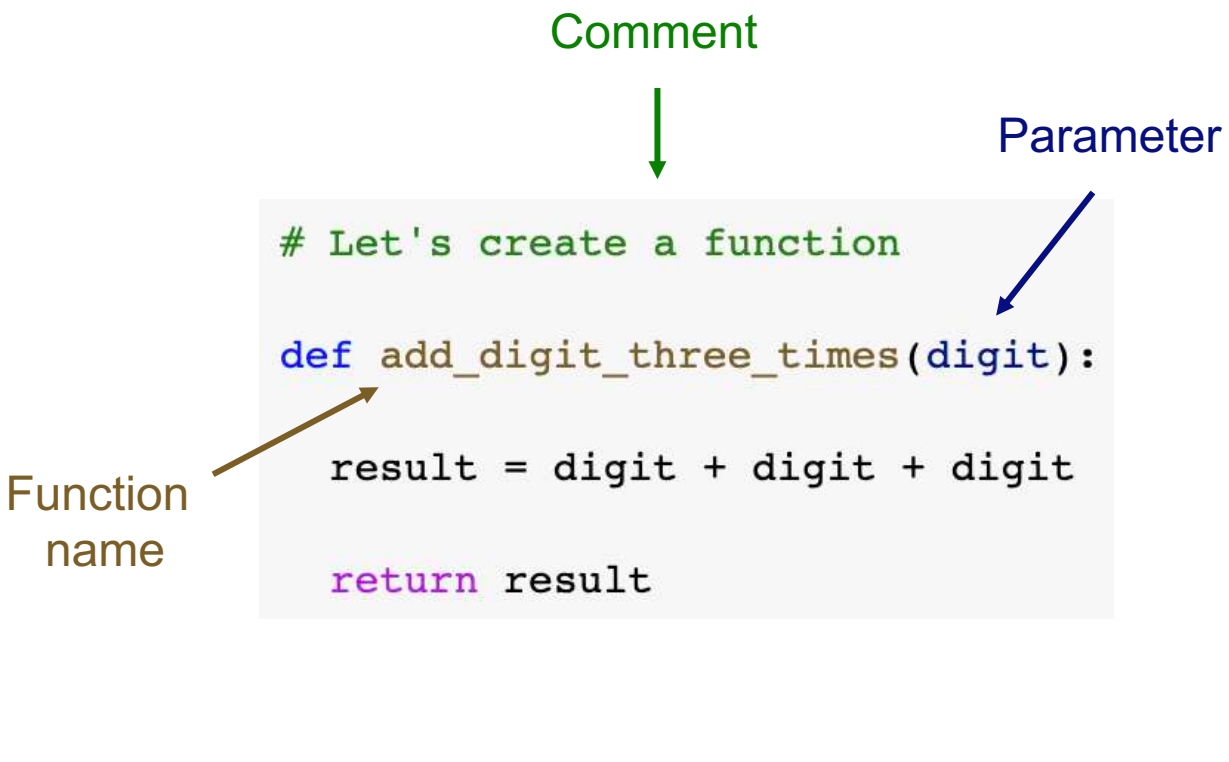
	player	national_team	club	career_goals
0	Lionel Messi	Argentina	FC Barcelona	690
1	Cristiano Ronaldo	Portugal	Juventus FC	720
2	Zlatan Ibrahimovic	Sweden	AC Milan	545

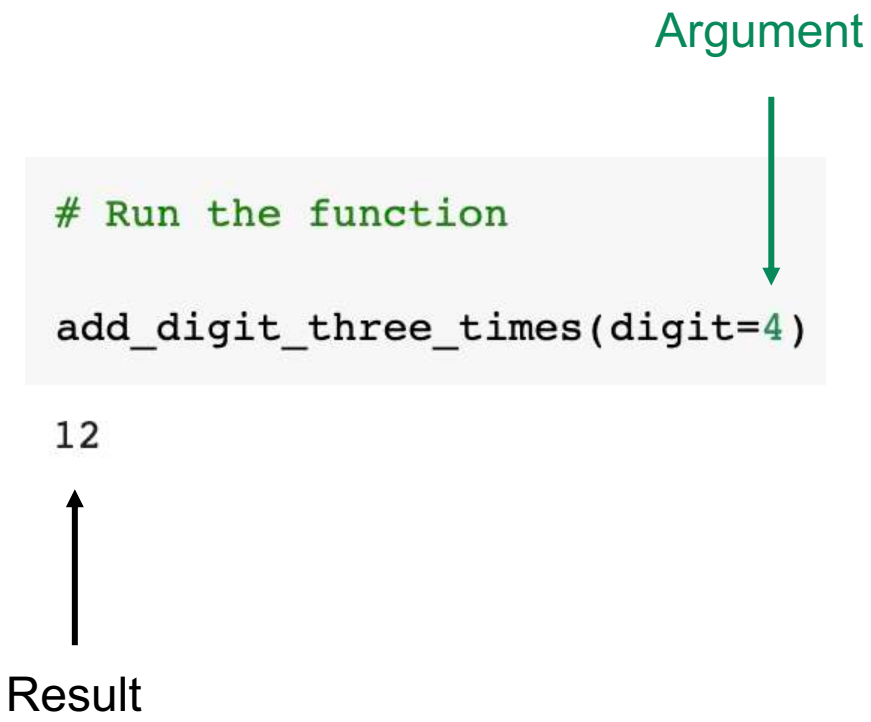
Functions

Help perform tasks we want to repeat

Automate the boring stuff

Anatomy of a function





Functions

So… functions are great!

And the best part is that OTHER people have written them for us!

Let’s say we want to find the average of some sets of numbers:

1, 2, 3, 4, 5
2, 4, 6, 8, 10

We could write our own function… or we could use someone else’s

Packages

Packages contain functions written by other people

Popular data science packages:

Using packages

We need to import functions from packages in order to use them

We use a nickname for packages we use often

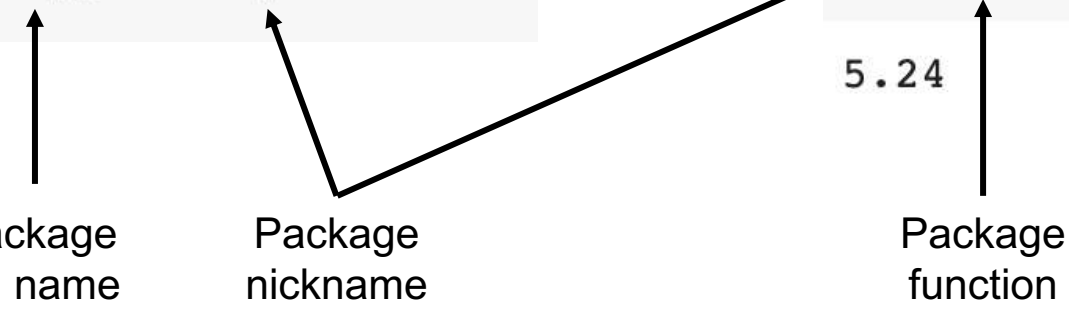
```
# Import the NumPy package

import numpy as np
```

```
# Use NumPy's rounding function

np.round(5.2398, 2)
```

```
5.24
```



Predictive modeling

Predictive modeling is…

Using data and statistics to predict events or outcomes.

Data

$$log(\frac{P}{P-1}) = mx + b + \epsilon$$

Statistics

Predictions

What can models predict?

Types of models

Linear Regression

Logistic Regression

Decision Trees

Generalized Linear Models (GLM)

Random Forests

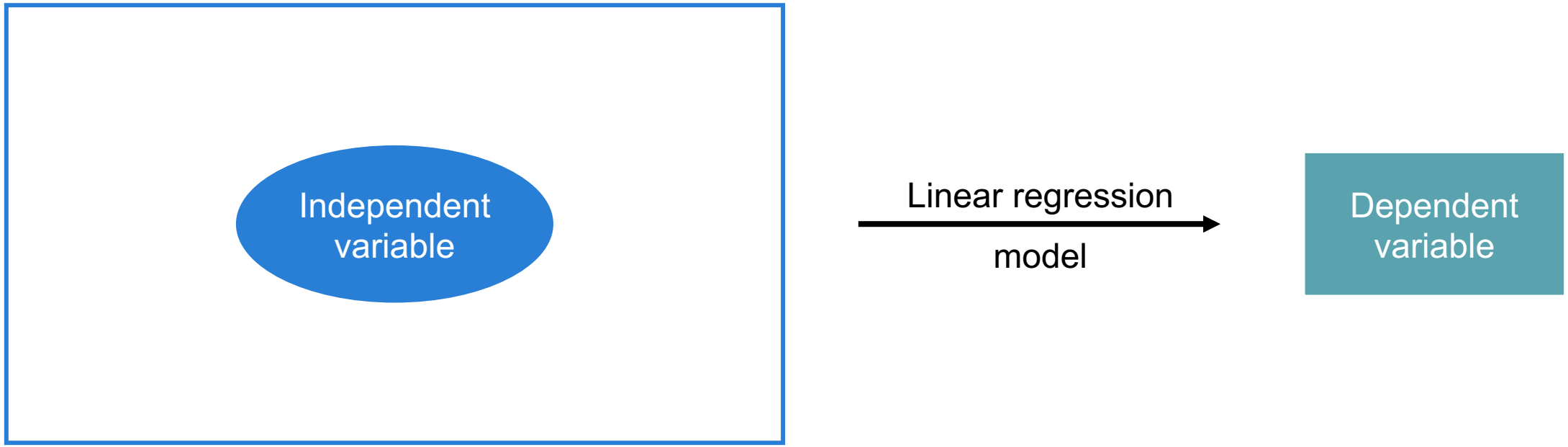
Neural Networks

Multivariate Adaptive Regression Splines (MARS)

Linear regression

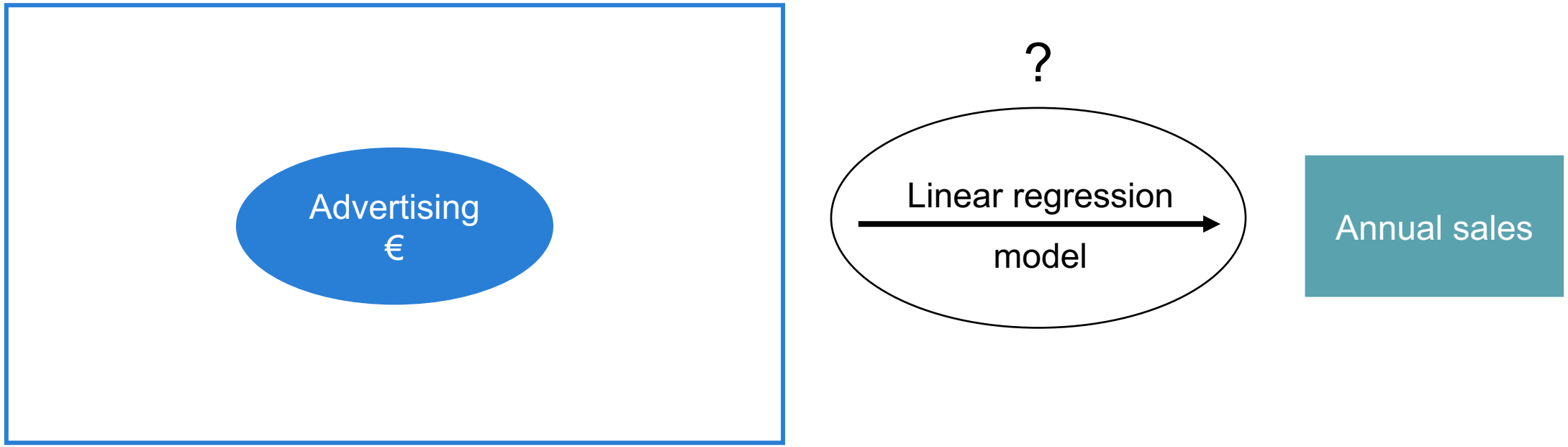
Linear regression

A model that determines the value of one dependent variable from one or more independent variables.



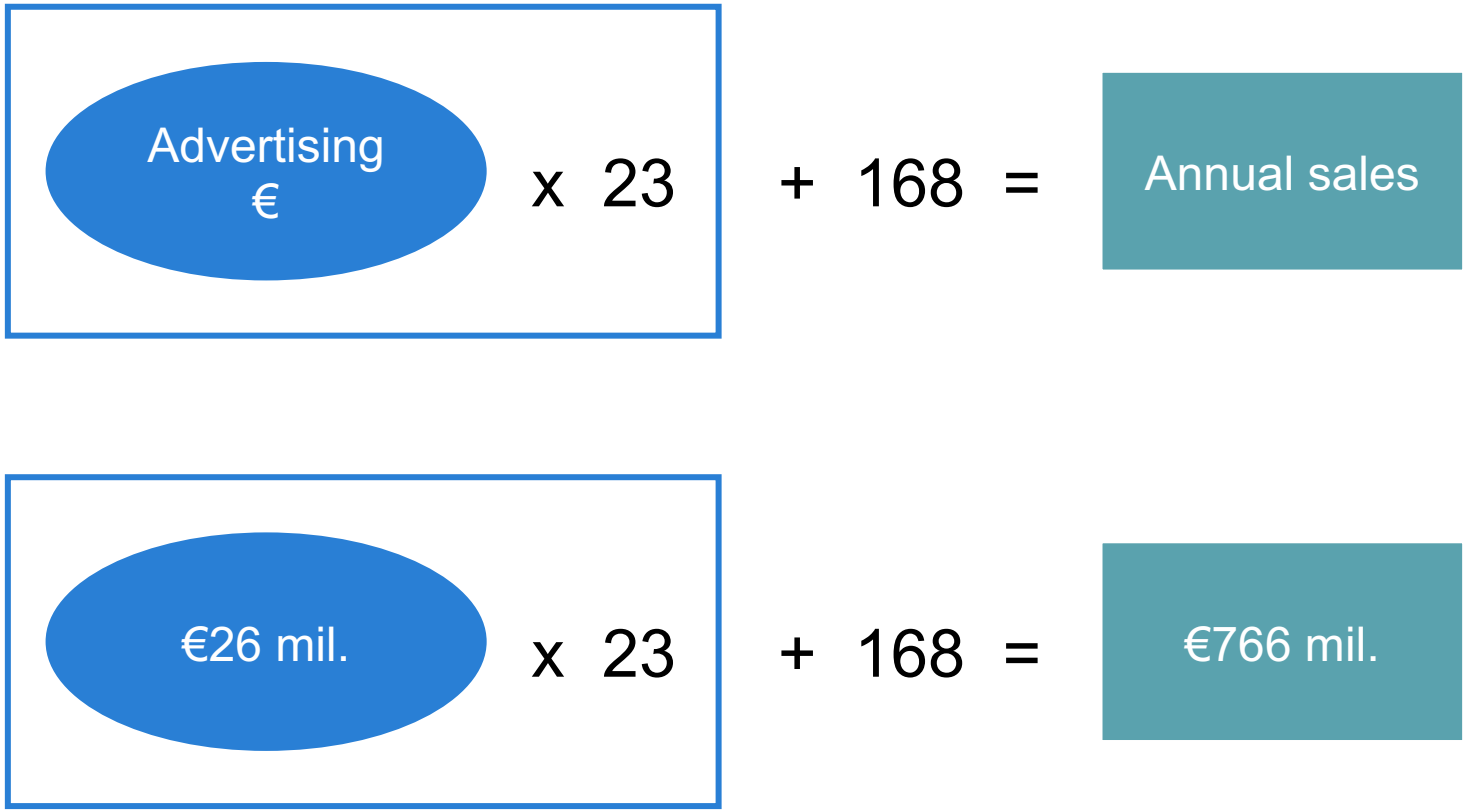
Linear regression

Example: Predicting sales revenue based on advertising budget



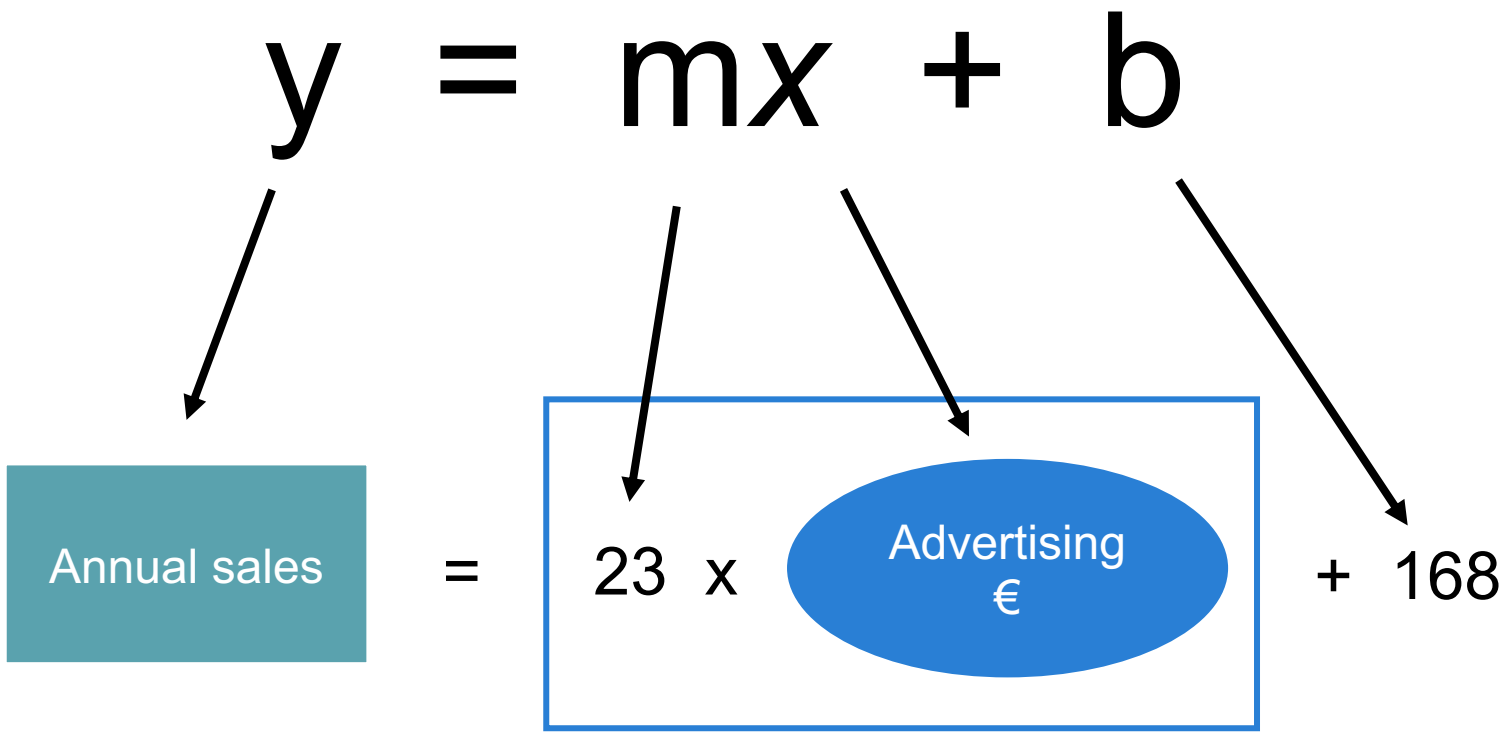
Linear regression

A linear regression model is an equation.



Linear regression

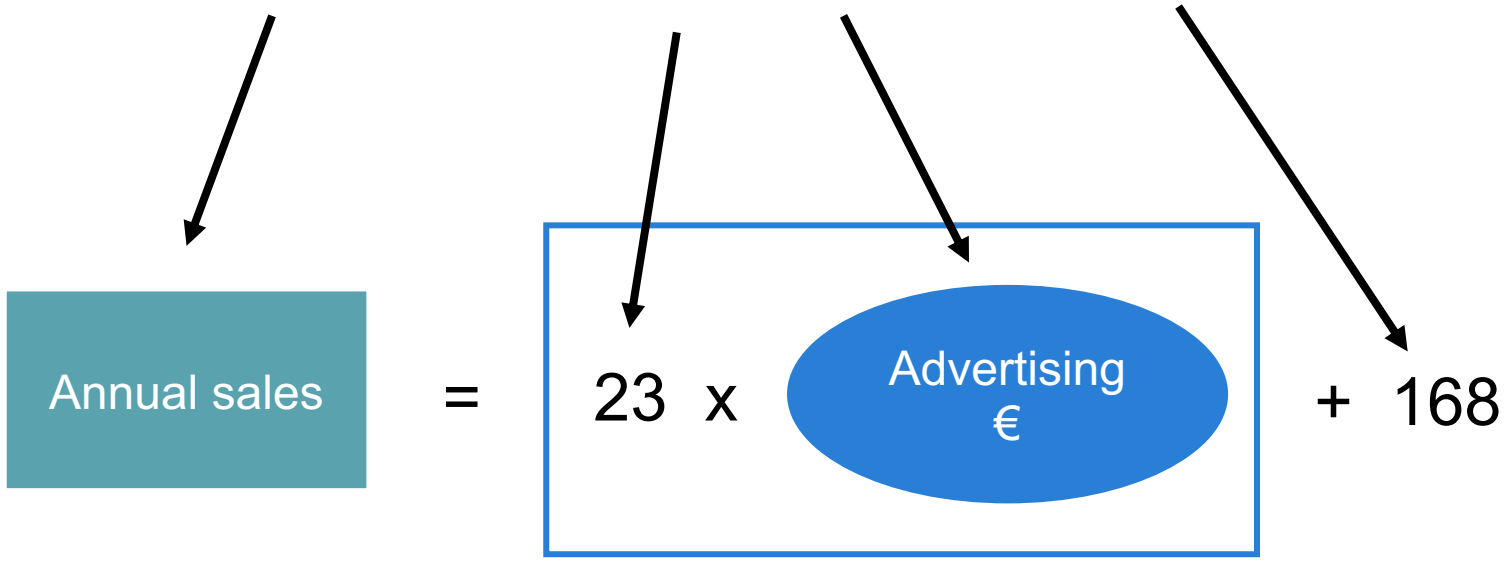
Look familiar?



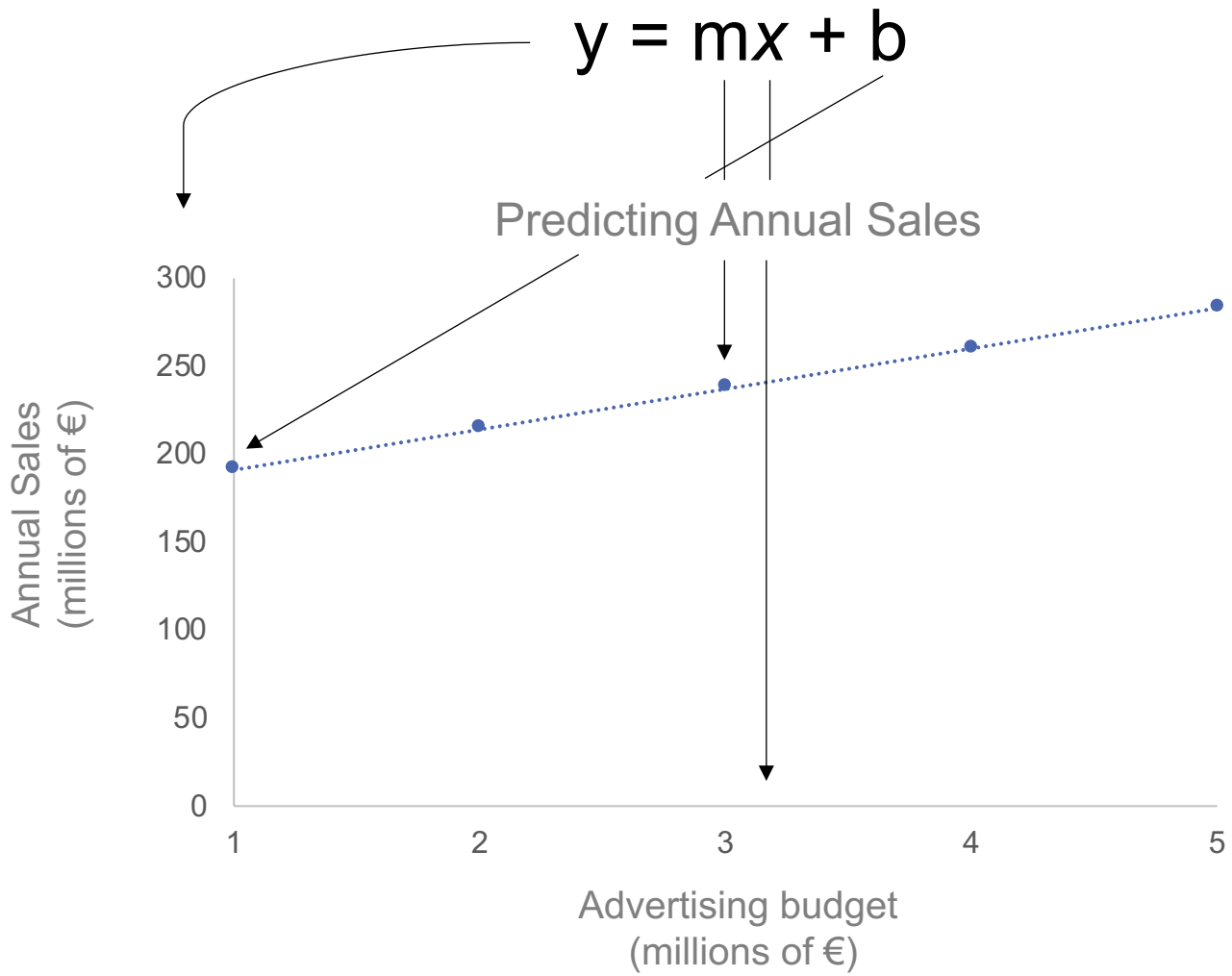
Linear regression

Now, we use the equation to ***predict*** future sales

$$y = mx + b$$

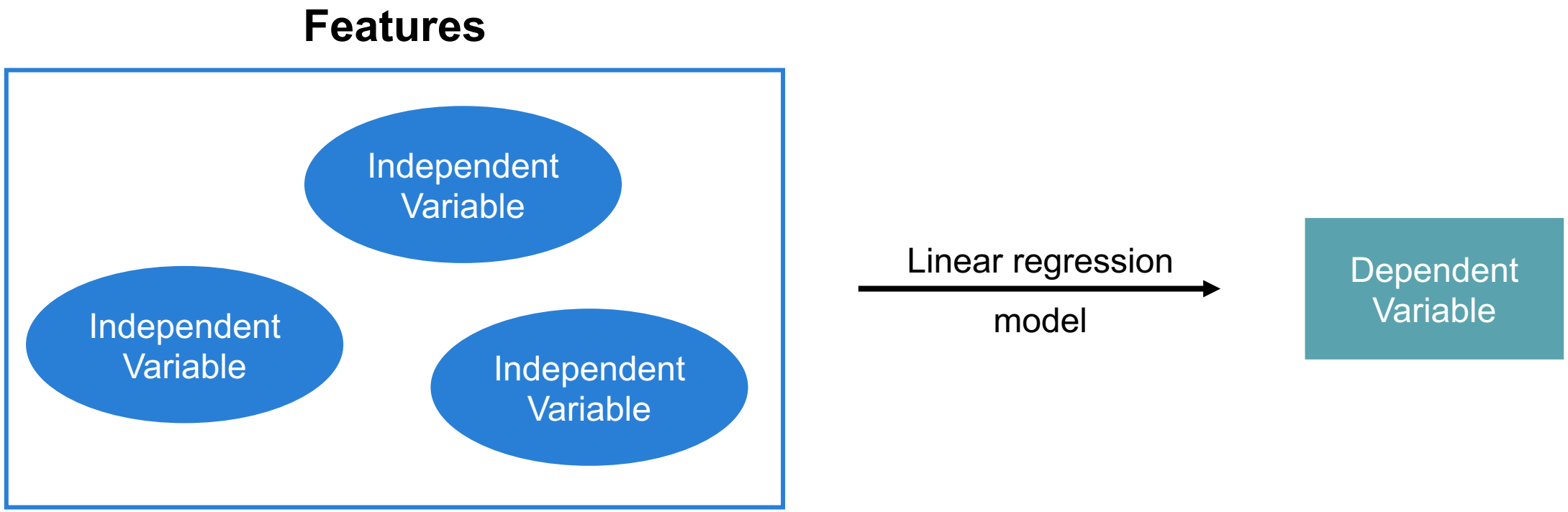


Linear regression



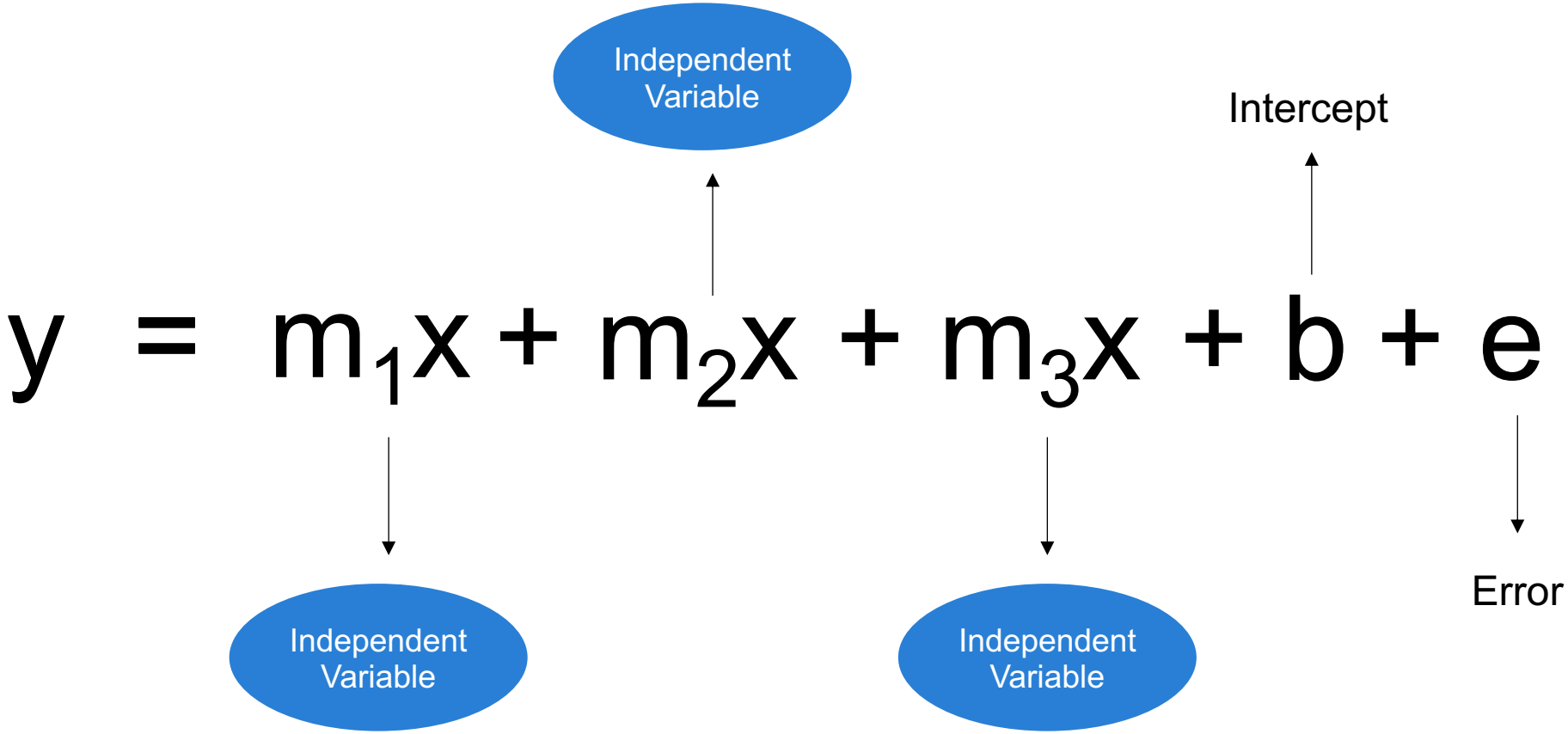
Linear regression

Most models have more than one independent variable.



Linear regression

Equation when we have *multiple* independent variables.



Predicting wine quality

Using linear regression

Background

Exports of *vinho verde* wine from Portugal are increasing.

Quality of wine determines its price.

Quality of wine impacts sales.

How do we measure wine quality?

Physicochemical tests

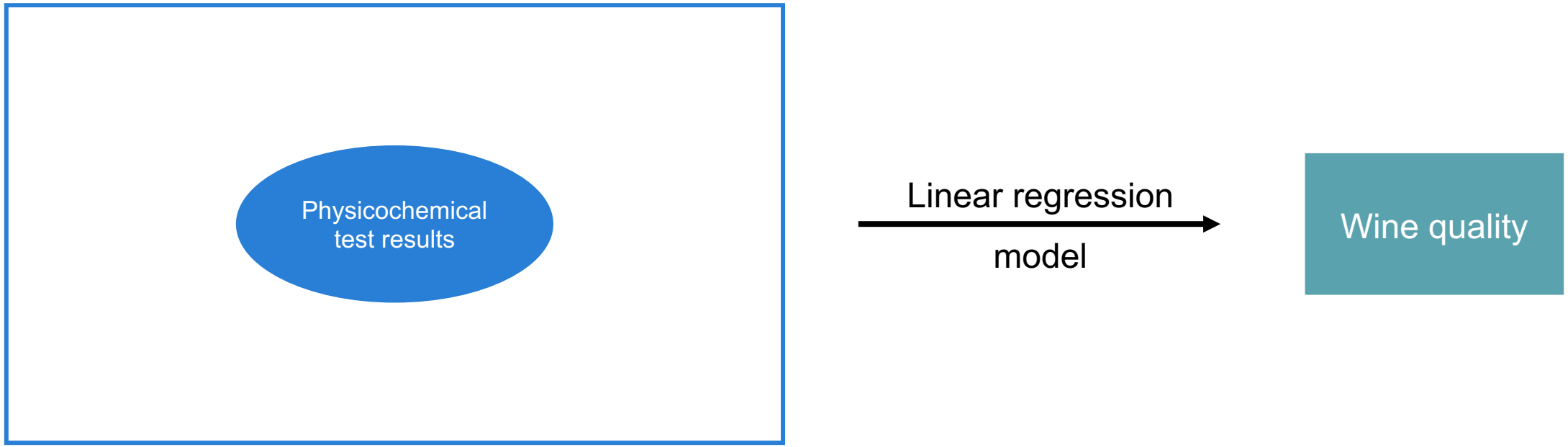
- Density
- Alcohol values
- pH values

Sensory tests

- Human taste testers

Objective

Use data from physicochemical tests to predict how wine will perform during human quality taste tests.



Dataset

Wine samples from May 2004 – February 2007

One dataset for red wine, one for white wine

Lab test data

- Acidities
- Other chemicals
- Density
- pH
- Alcohol (%)

Sensory test data

- Median quality (scale 0-10) given by at least 3 human taste testers

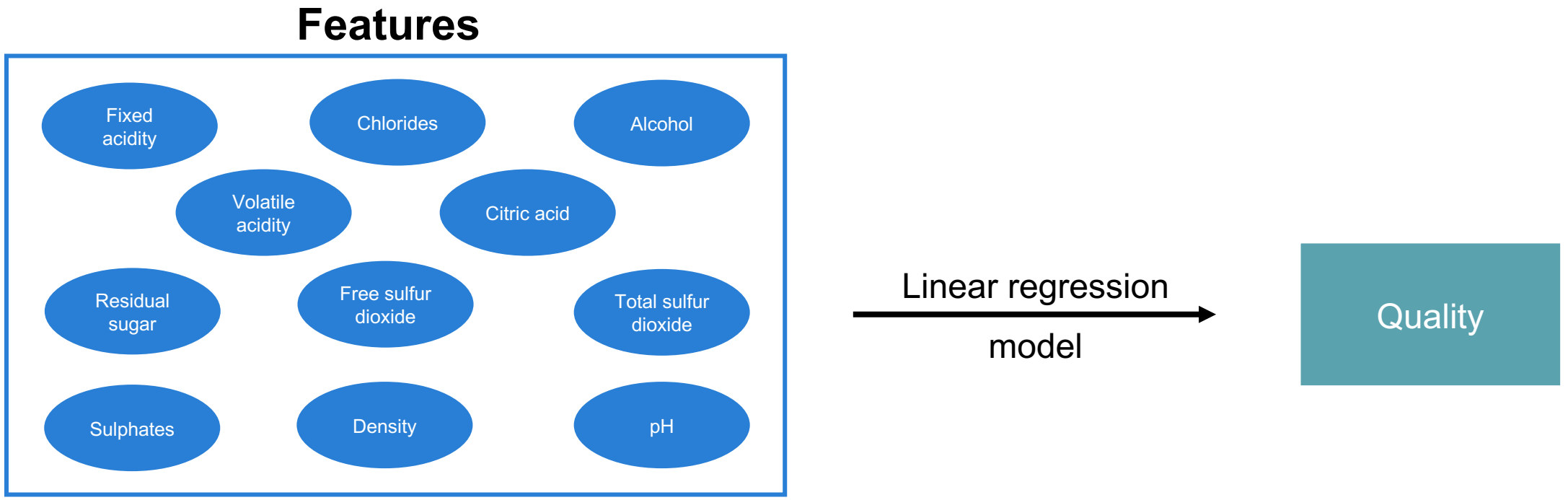
Dataset features

winequality-red.csv

	A	B	C	D	E	F	G	H	I	J	K	L
1	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
2	7.4	0.7	0	1.9	0.076	11	34	0.9978	4	0.56	9.4	5
3	7.8	0.88	0	2.6	0.098	25	67	0.9968	3	0.68	9.8	5
4	7.8	0.76	0.04	2.3	0.092	15	54	0.997	3	0.65	9.8	5
5	11.2	0.28	0.56	1.9	0.075	17	60	0.998	3	0.58	9.8	6
6	7.4	0.7	0	1.9	0.076	11	34	0.9978	4	0.56	9.4	5
7	7.4	0.66	0	1.8	0.075	13	40	0.9978	4	0.56	9.4	5
8	7.9	0.6	0.06	1.6	0.069	15	59	0.9964	3	0.46	9.4	5
9	7.3	0.65	0	1.2	0.065	15	21	0.9946	3	0.47	10	7
10	7.8	0.58	0.02	2	0.073	9	18	0.9968	3	0.57	9.5	7
11	7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3	0.8	10.5	5
12	6.7	0.58	0.08	1.8	0.097	15	65	0.9959	3	0.54	9.2	5
13	7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3	0.8	10.5	5
14	5.6	0.615	0	1.6	0.089	16	59	0.9943	4	0.52	9.9	5
15	7.8	0.61	0.29	1.6	0.114	9	29	0.9974	3	1.56	9.1	5
16	8.9	0.62	0.18	3.8	0.176	52	145	0.9986	3	0.88	9.2	5
17	8.9	0.62	0.19	3.9	0.17	51	148	0.9986	3	0.93	9.2	5

Objective

Use data from physicochemical tests to predict how wine will perform during human quality taste tests.

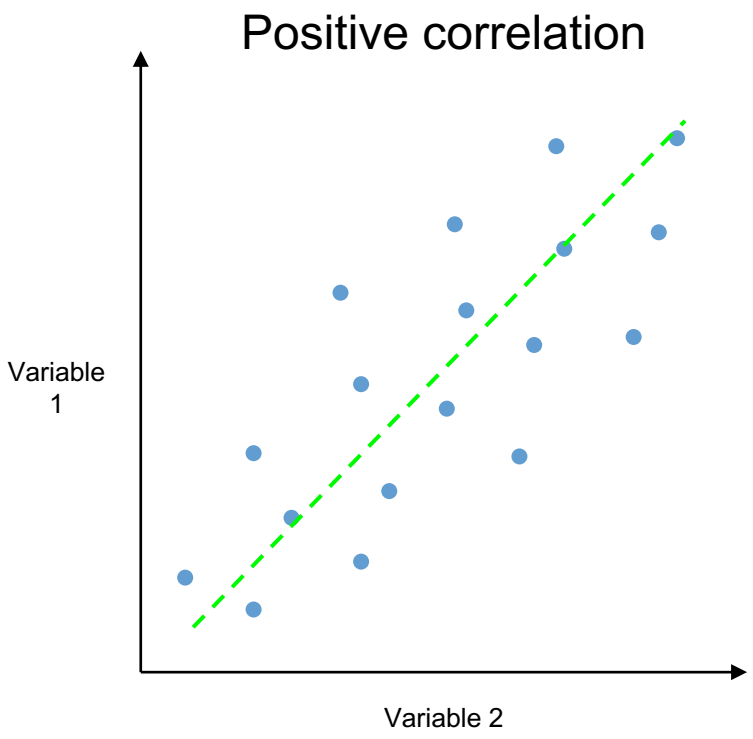


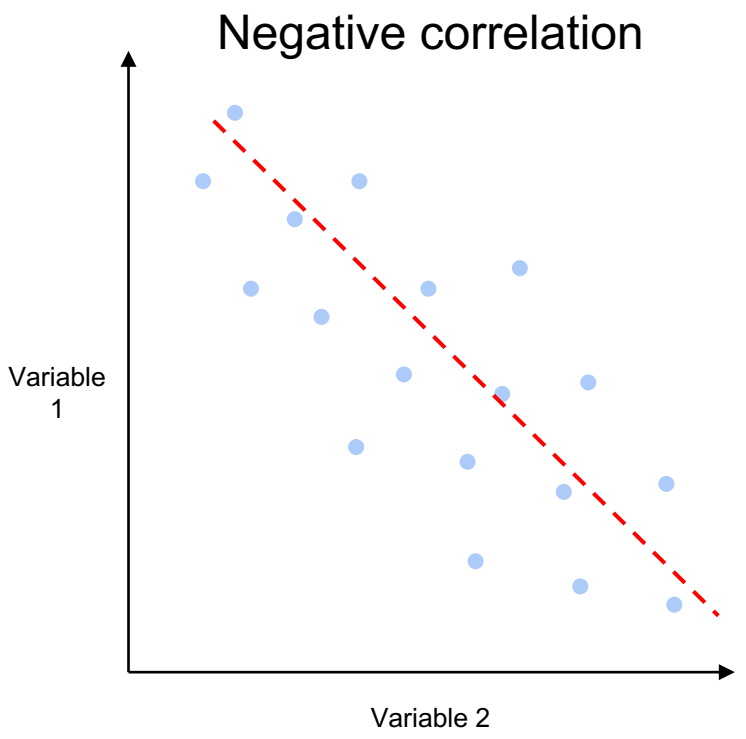
Steps in Python

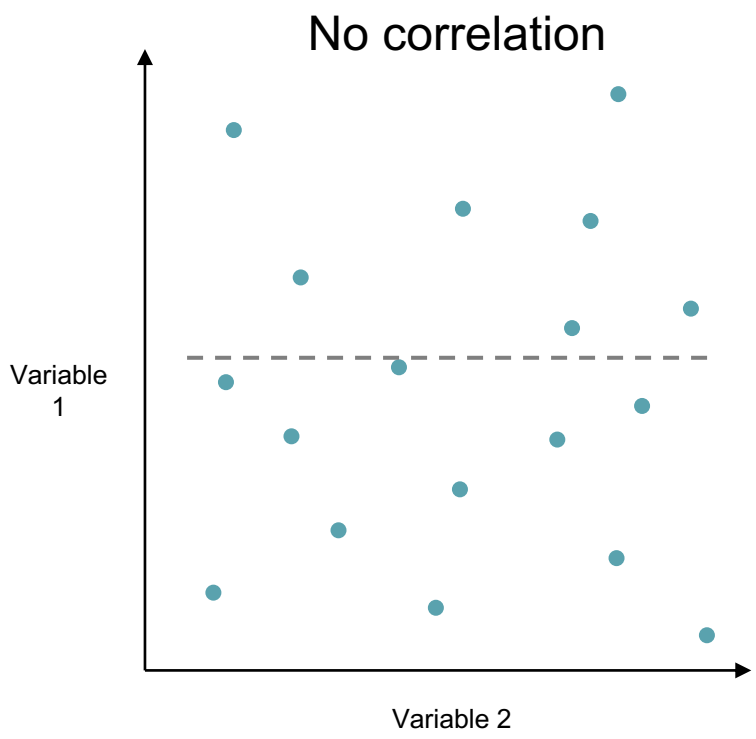
1. Import packages
2. Import data
3. Explore data
4. Preprocess data
5. Build model
6. Make predictions
7. Evaluate model

Correlations

The strength of the relationship between two variables.



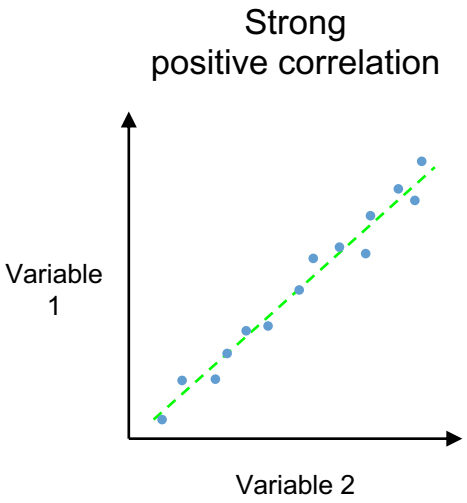




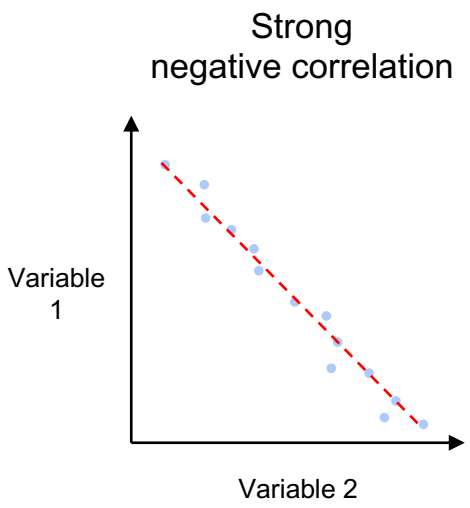
Correlations

Strength is measured by the *correlation coefficient* (r).

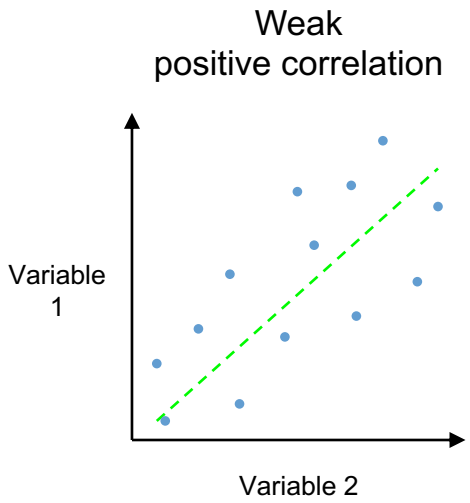
Ranges from [-1, +1].



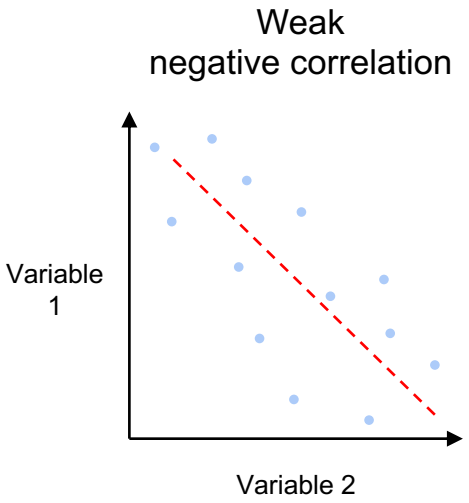
r = 0.9



r = -0.9



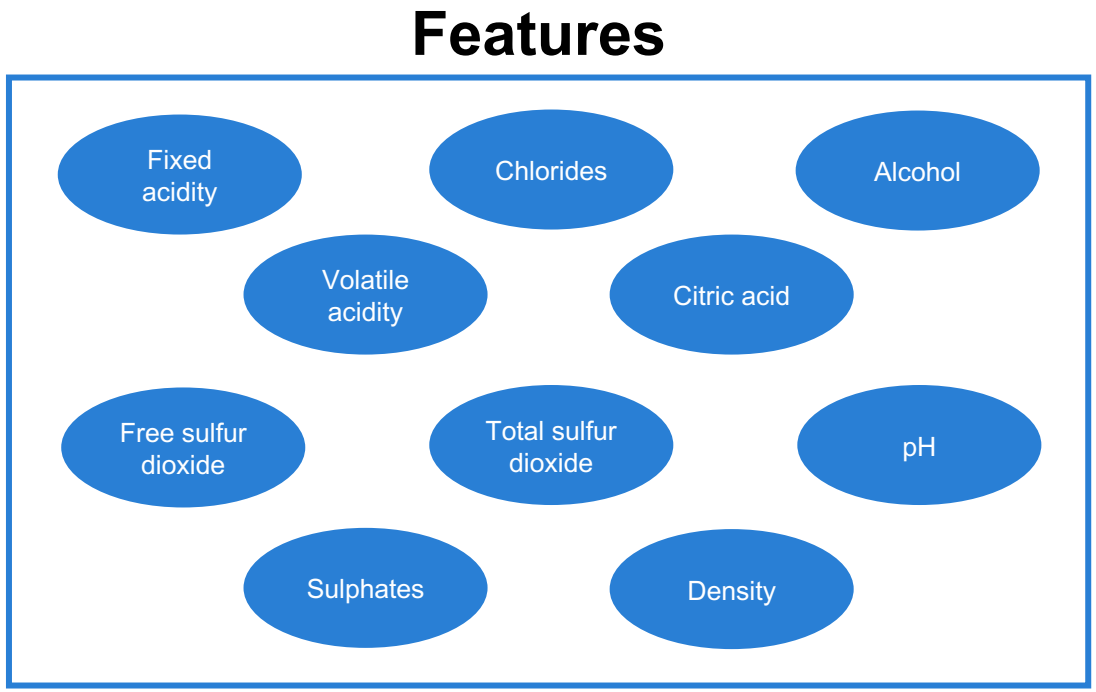
r = 0.3



r = -0.3

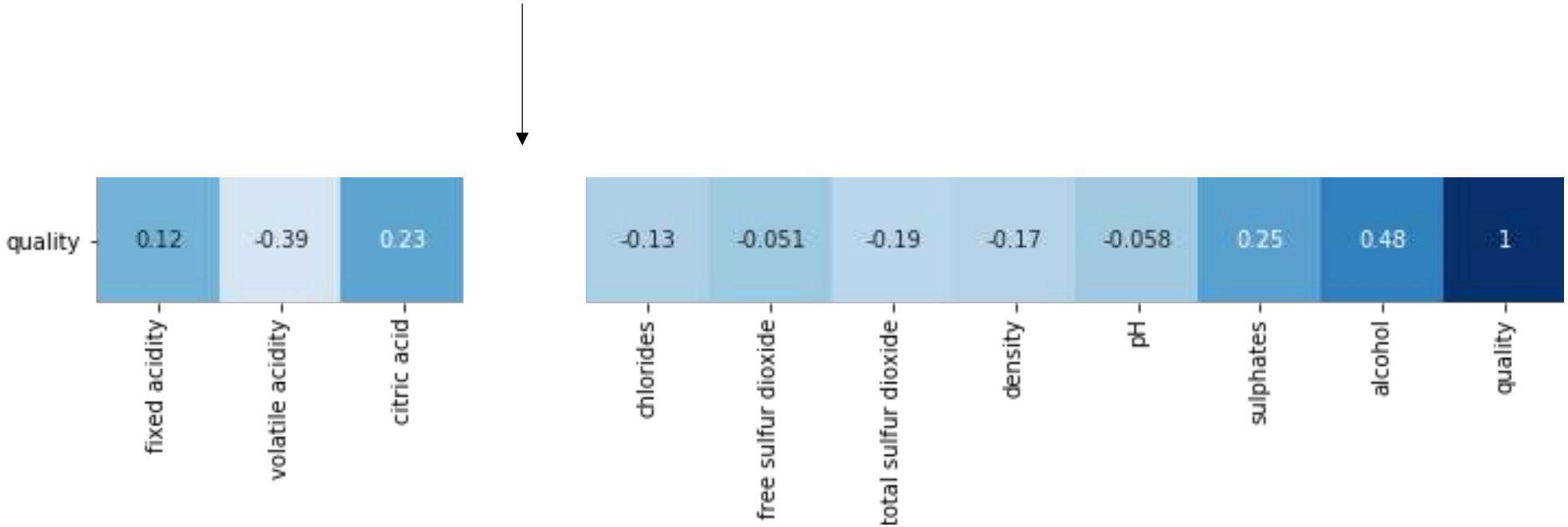
Feature selection

We want our features to have positive correlations with the outcome variable.



Feature selection

Anything with r < 0.05 is probably meaningless.



Feature selection

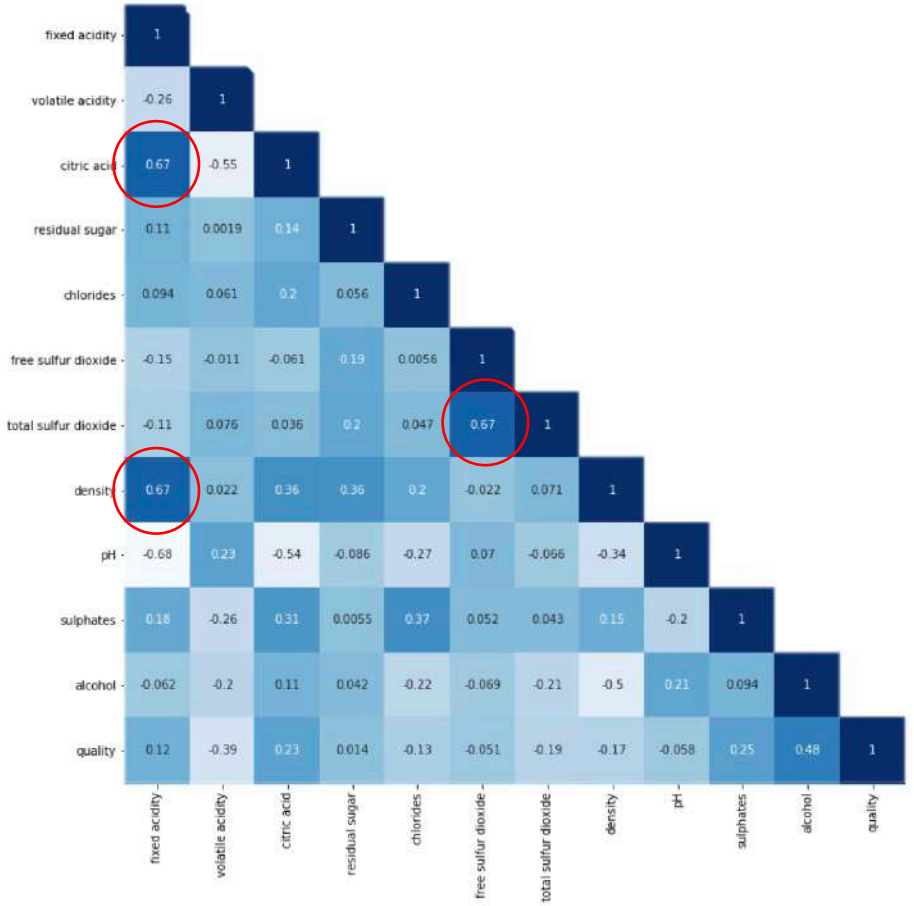
BUT, we don’t want our features to be (too) correlated with each other.

Multicollinearity is bad!

- Hard to determine effect of feature on outcome variable
- Hard to know what the include in model
- Inaccurate final model

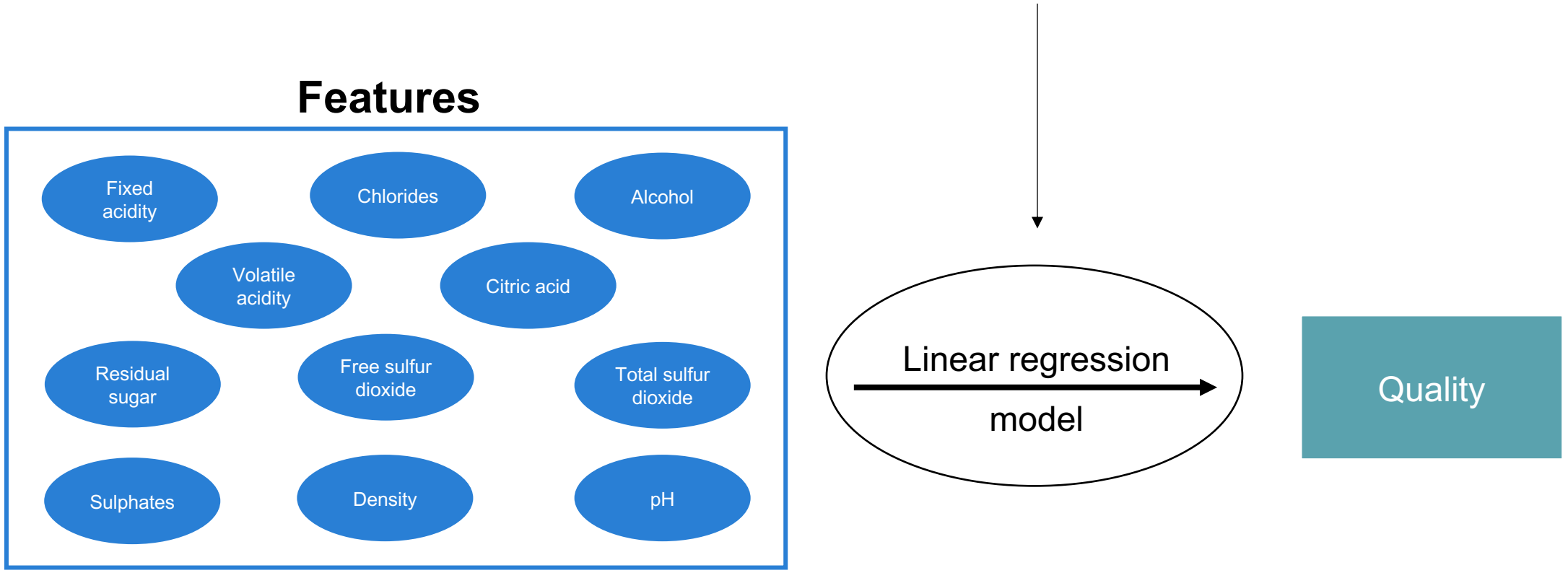
Feature selection

Anything with r > 0.80 is suspicious.



Model output

Model output tells us how to write our linear regression equation.



Model output

$$m_1x + m_2x + m_3x + b + e = y$$

$$(_{\text{Fixed acidity}}x) + (_{\text{Chlorides}}x) + (_{\text{Alcohol}}x) + (_{\text{Volatile acidity}}x) + (_{\text{Citric acid}}x) + (_{\text{Free sulfur dioxide}}x) + (_{\text{Total sulfur dioxide}}x) + (_{\text{Sulphates}}x) + (_{\text{Density}}x) + (_{\text{pH}}x) + b + e = \text{Quality}$$

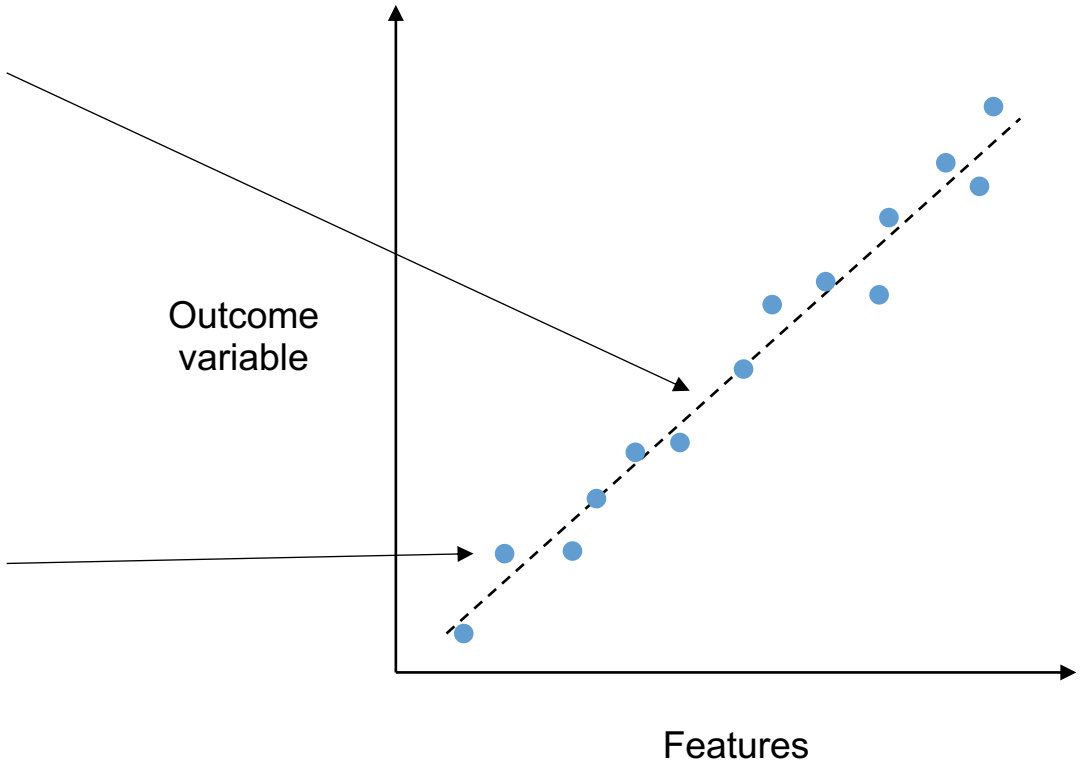
$$(-0.005*x) + (-2.049*x) + (0.288*x) + (-1.293*x) + (-0.251*x) + (0.001*x) + (-0.002*x) + (0.976*x) + (9.833*x) + (-0.587*x) + -4.787 + e = \text{Quality}$$

$$(-0.005*8.0) + (-2.049*0.089) + (0.288*10.0) + (-1.293*0.59) + (-0.251*0.05) + (0.001*12.0) + (-0.002*32.0) + (0.976*0.61) + (9.833*0.99735) + (-0.587*3.36) - 4.787 + 0 = \mathbf{5.45}$$

Error

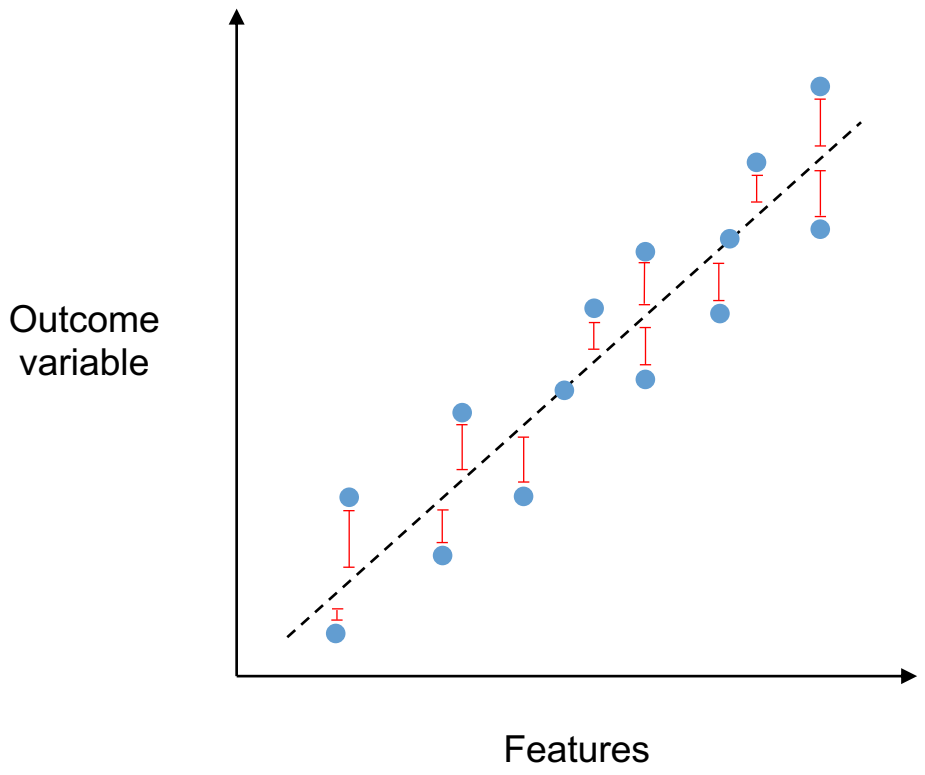
Linear regression predicts values that follow a straight line.

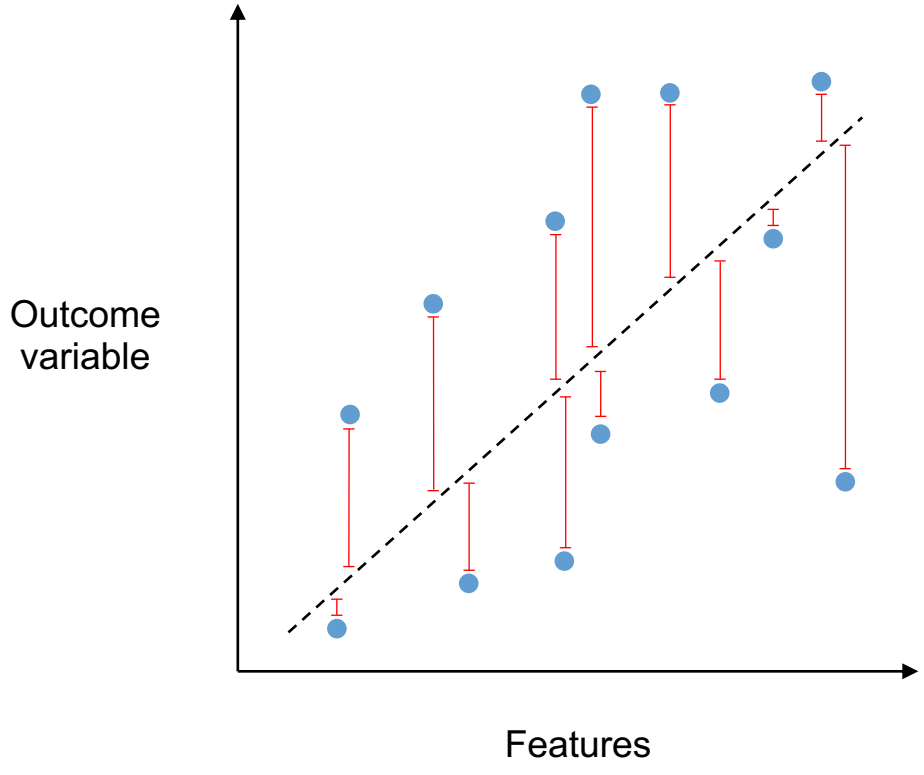
But values in the real world never completely follow a straight line.



Error

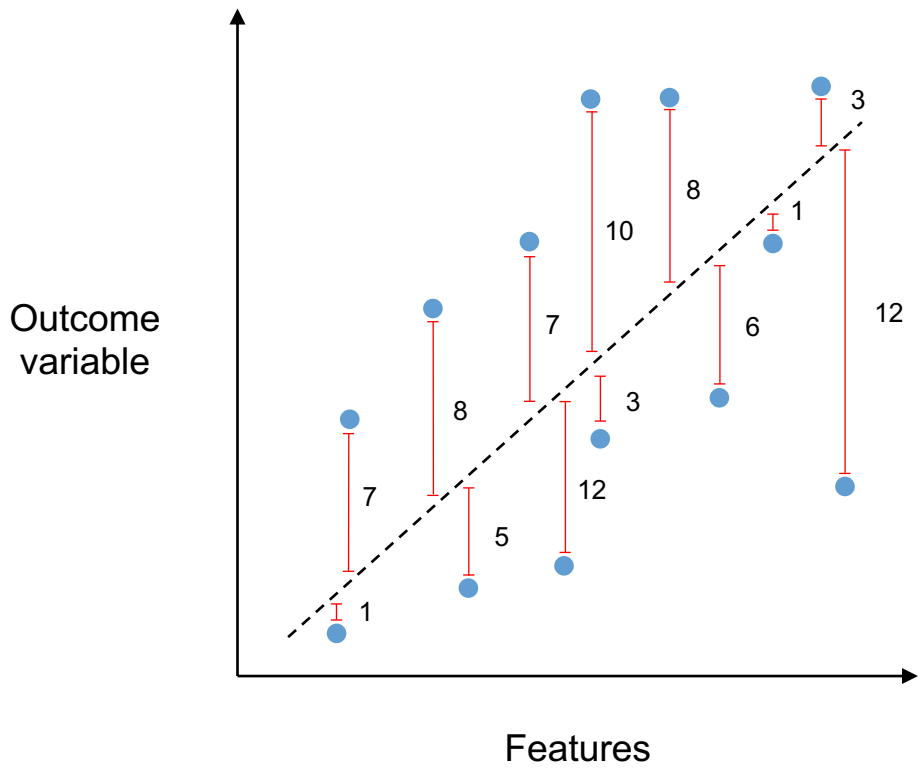
We can measure error by measuring the residuals – the difference between predicted values and actual values.





RMSE

We'll use root mean squared error (RMSE) to estimate the model's error.



$$RMSE = \sqrt{\frac{(Sum\ of\ residuals)^2}{\#\ of\ data\ points}}$$

$$RMSE = \sqrt{\frac{(1+7+8+5+7+12+10+3+8+6+1+3+12)^2}{13}}$$

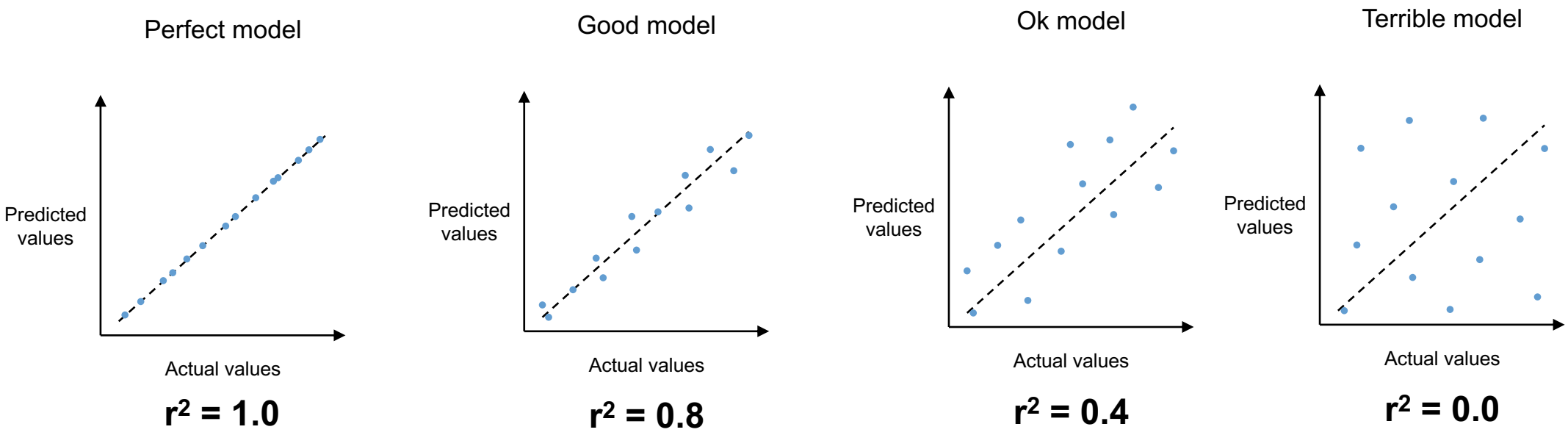
$$RMSE = \sqrt{\frac{(83)^2}{13}} \longrightarrow RMSE = \sqrt{\frac{41.5}{13}} \longrightarrow RMSE = \sqrt{3.19}$$

$$\boxed{RMSE = 1.79}$$

R^2

R^2 represents how well the model (equation) explains the data.

Ranges from [0, 1].

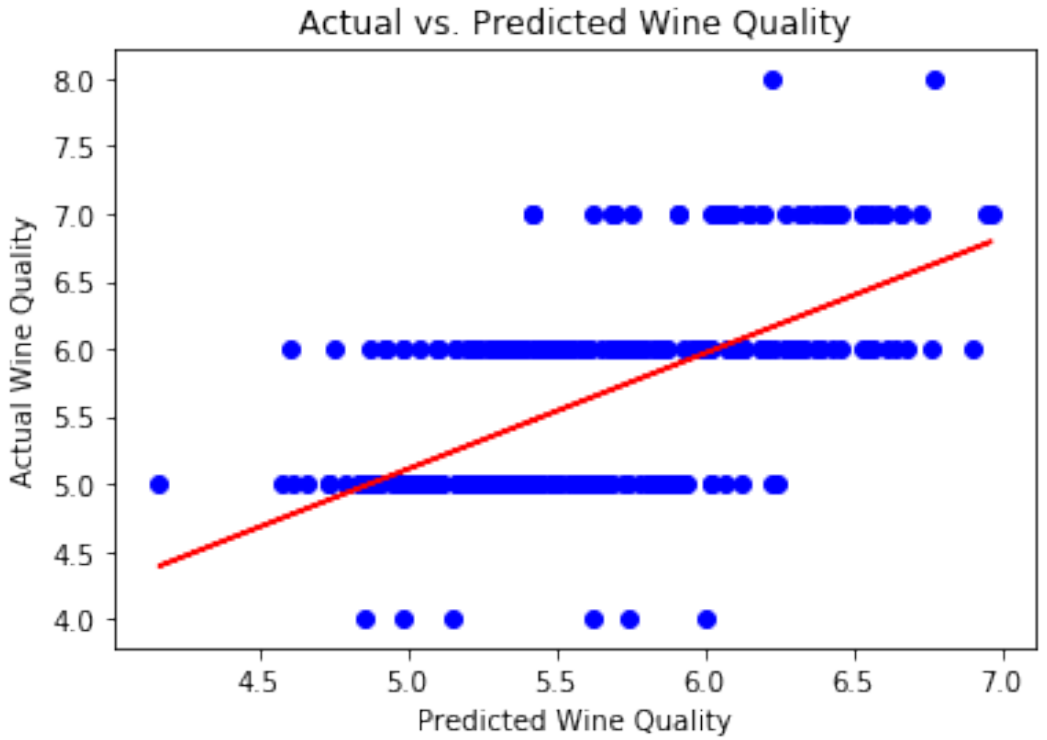


Evaluating our model

RMSE = 0.377

R^2 = 0.362

So our model isn't amazing…



Improving a model

1. Normalize the features

2. Drop some features (density)

3. **Use a different model**

Now for the cool stuff…

Neural networks

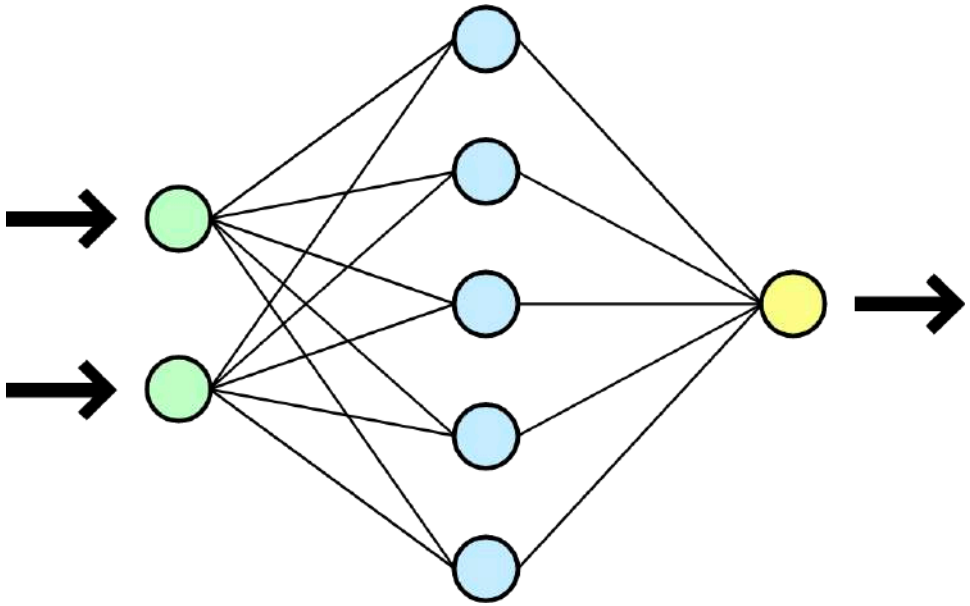
Take input, do calculations and give output.

Work like our brains.

Learn to perform tasks by being fed examples.

Learn features from examples.

Neural networks



Dog

Natural Language Processing

NLP models try to make sense of text without reading it.

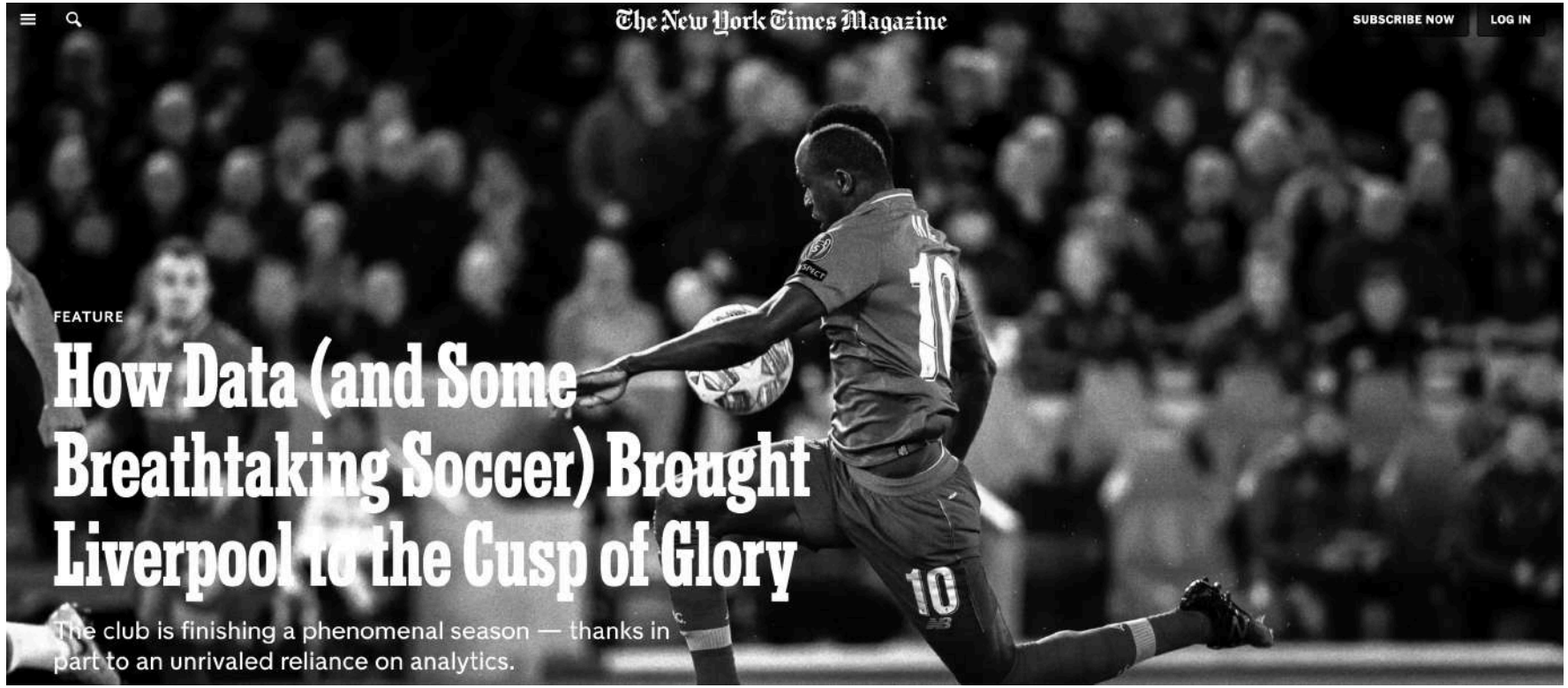
Programming computers to understand human language.

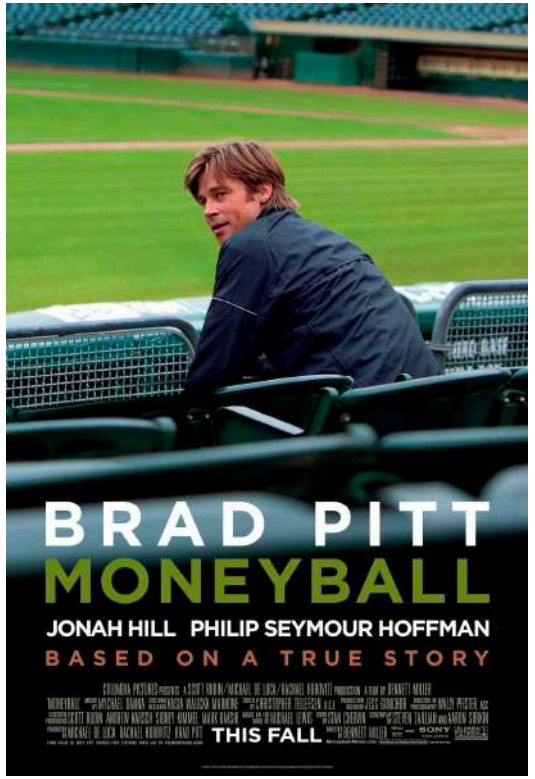
Examples

- Speech recognition
- Predictive text
- Word clouds

Sports

[The Numbers Game: How Data is Changing Football]





Learn more

Python & data science resources

Code Academy – Interactive online courses (free, paid option)

Learnpython.org – Similar to Data Camp but no videos (free)

EdX – Free online courses with great practice projects (free)

Data Camp – Awesome video courses (mostly paid)

Sentdex YouTube channel – Great video tutorials (1,000+)

Thanks for coming!

For career questions or advice:

mary.newhauser@gmail.com