

Practical Data Science with R

@matthewrenze

#DevoxxMA









The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades, ... because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it.

Hal Varian, Google's Chief Economist
The McKinsey Quarterly, Jan 2009

The Economist

FEBRUARY 27TH - MARCH 5TH 2010

Economist.com

Gordon Brown's pitch
What went wrong at RBS
Genetically modified crops blossom
The EU woos Russia
The right to eat cats and dogs

The data deluge

AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT

A man in a suit holds a large green and yellow umbrella over a small flower.

The New York Times

For Today's Graduate, Just One Word: Statistics

By STEVE LORIN
Published: August 5, 2009

MOUNTAIN VIEW, Calif. — At Harvard, Carrie Grimes majored in anthropology and archaeology and ventured to places like Honduras, where she studied Mayan settlement patterns by mapping where artifacts were found. But she was drawn to what she calls “all the computer and math stuff” that was part of the job.

[TWITTER](#)
[LINKEDIN](#)
[COMMENTS
\(58\)](#)
[SIGN IN TO E-MAIL](#)

Data Scientist: *The Sexiest Job of the 21st Century*

**Meet the people who
can coax treasure out of
messy, unstructured data.**
by Thomas H. Davenport
and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, “It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early.”

Job Postings for Data Scientists



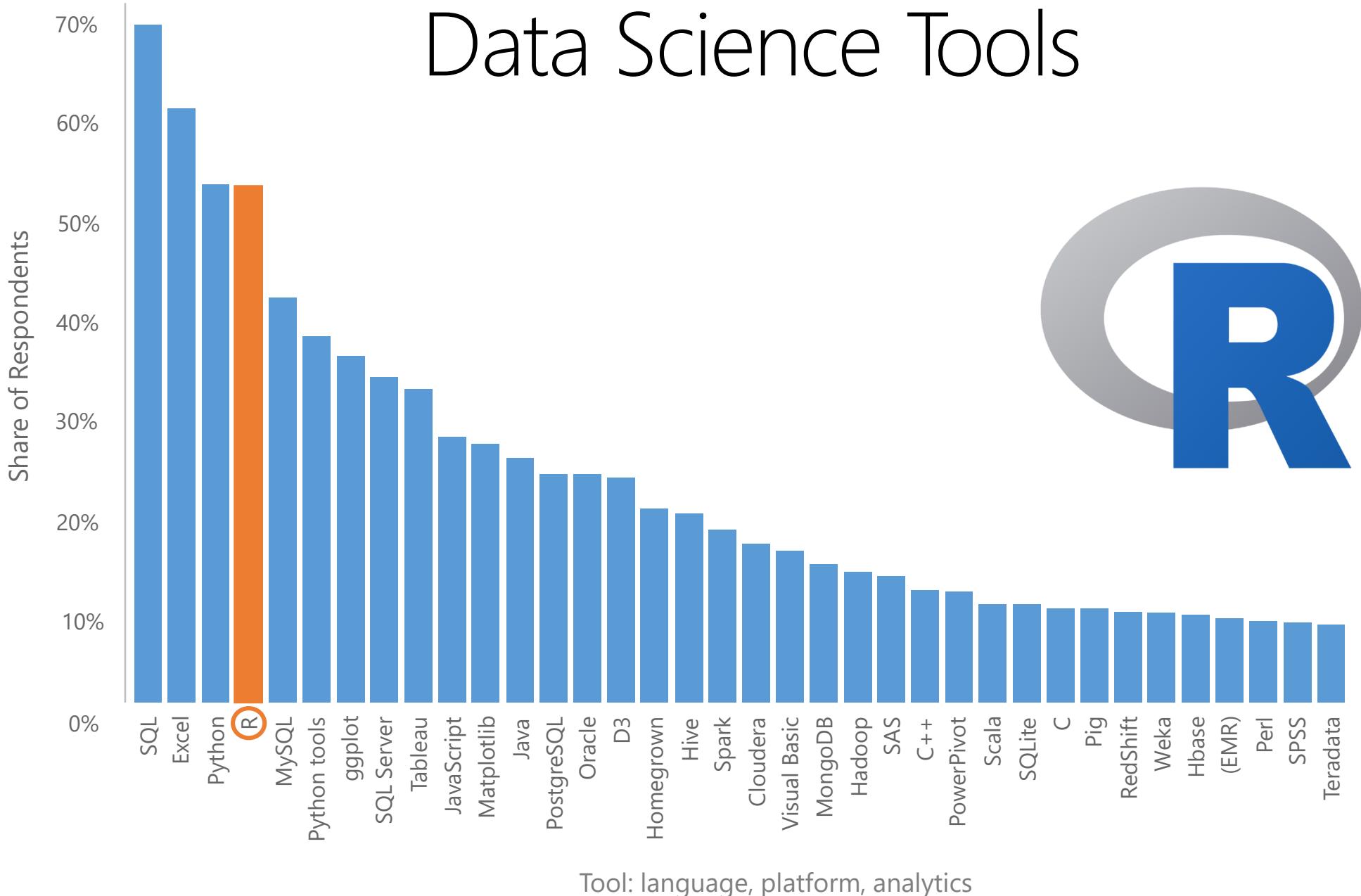
Top-paying Tech Skills

| Skill | 2016 | Change |
|---|------------|--------|
| HANA (High Performance Analytical Application) | \$ 128,958 | -3.3% |
| MapReduce | \$ 125,009 | -0.3% |
| Cloud Foundry | \$ 124,038 | n/a |
| Hbase | \$ 123,934 | 5.7% |
| Omnigraffle | \$ 123,782 | -1.9% |
| Cassandra | \$ 123,459 | 2.2% |
| Apache Kafka | \$ 122,728 | n/a |
| SOA (Service Oriented Architecture) | \$ 122,094 | -1.9% |
| Ansible | \$ 121,382 | n/a |
| Jetty | \$ 120,978 | 1.3% |
| PaaS (Platform as a Service) | \$ 120,403 | -4.4% |
| Elasticsearch | \$ 120,002 | n/a |
| ABAP (Advanced Business Application Programming) | \$ 119,961 | 0.5% |
| NoSQL | \$ 119,498 | 1.3% |
| CMMI (Capability Maturity Model Integration) | \$ 119,466 | -0.6% |
| Amazon Redshift | \$ 119,197 | n/a |
| Pig | \$ 119,118 | -4.2% |
| Solr | \$ 119,032 | 0.1% |
| Cloudera | \$ 118,896 | -9.0% |
| Docker | \$ 118,873 | 0.2% |

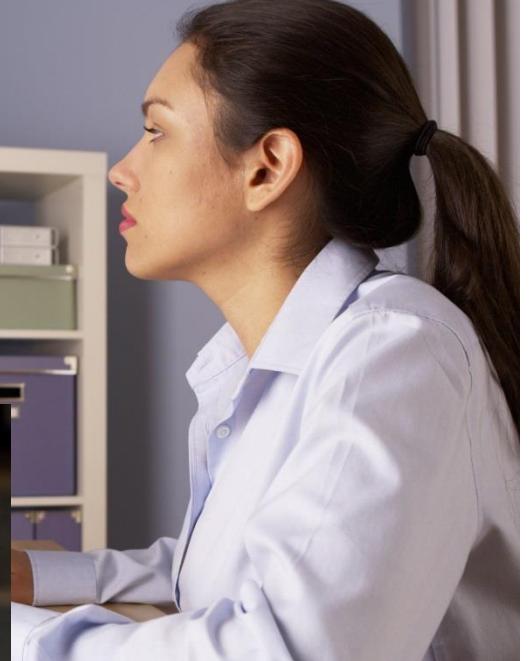
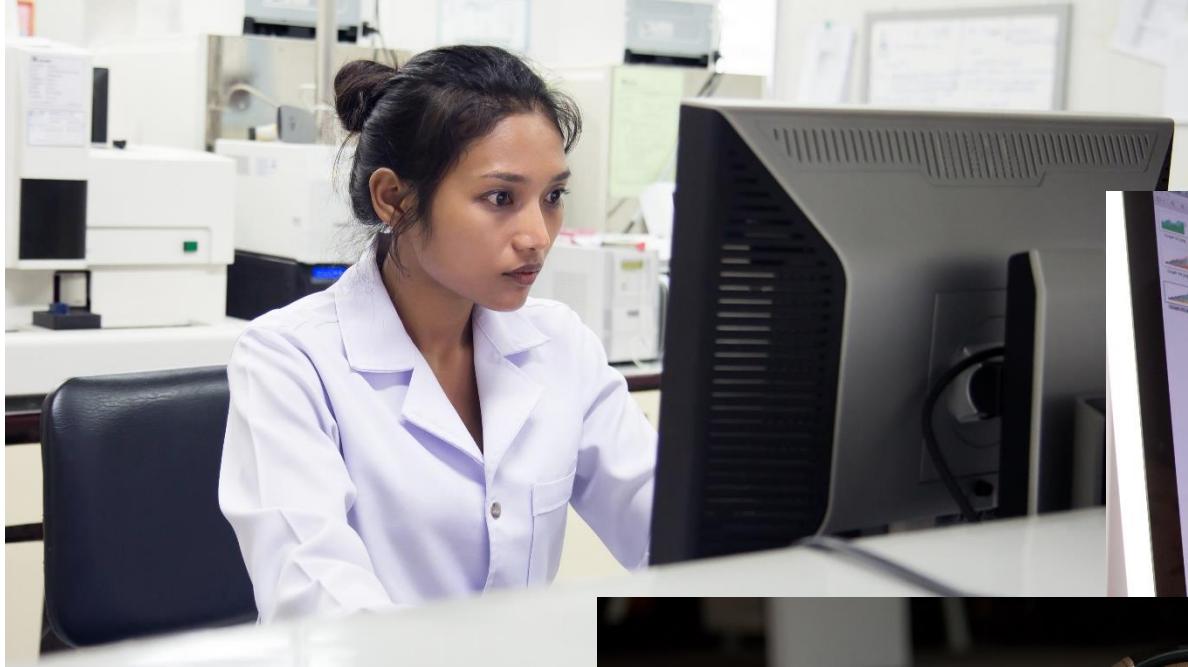
| Skill | 2016 | Change |
|--|------------|--------|
| Amazon Route 53 | \$ 118,828 | n/a |
| Hadoop | \$ 118,625 | -2.5% |
| Hive | \$ 118,589 | -1.3% |
| Korn Shell | \$ 118,273 | 1.4% |
| PMBok (Project Management Body of Knowledge) | \$ 118,233 | 0.7% |
| Dynamo DB | \$ 118,119 | n/a |
| Groovy | \$ 117,897 | -0.1% |
| IaaS (Infrastructure as a Service) | \$ 117,422 | n/a |
| JAX-RS (Java API RestFUL Services) | \$ 116,997 | n/a |
| RabbitMQ | \$ 116,909 | n/a |
| JDBC (Java Database Connectivity) | \$ 116,833 | 2.0% |
| SOX (Sarbanes Oxley) | \$ 116,743 | 0.6% |
| Objective C | \$ 116,667 | 2.5% |
| FCoE (Fibre Channel over Ethernet) | \$ 116,145 | 7.2% |
| UML (Unified Modeling Langauge) | \$ 115,285 | -3.6% |
| XSLT (Extensible Stylesheet Language Transformations) | \$ 115,089 | 3.5% |
| Redis | \$ 114,922 | 2.8% |
| ETL (Extract Transform and Load) | \$ 114,892 | 2.6% |
| SDN (Software Defined Network) | \$ 114,739 | -2.3% |
| Informatica | \$ 114,143 | 1.1% |

Source: Dice Salary Survey 2017

Data Science Tools



Source: O'Reilly 2015 Data Science Salary Survey

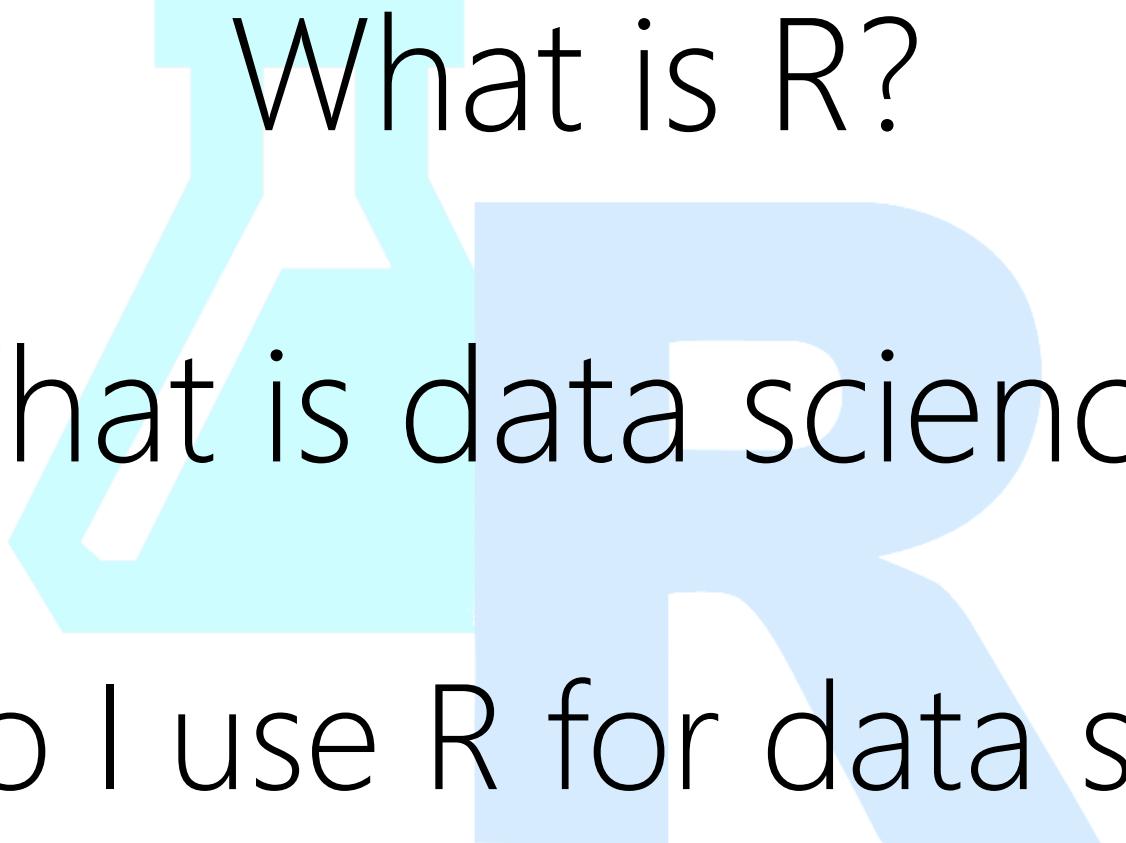








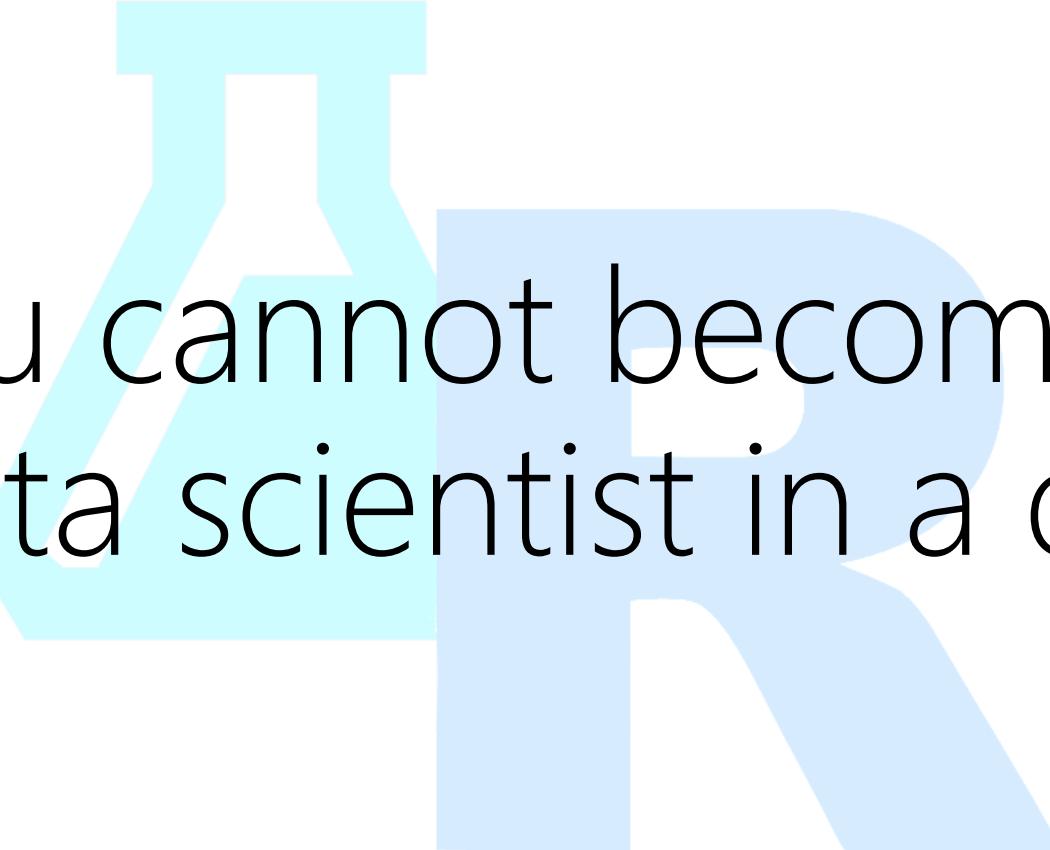




What is R?

What is data science?

How do I use R for data science?



You cannot become a
data scientist in a day.

Overview

1. Introduction
2. Working with Data
3. Descriptive Statistics
4. Data Visualization
5. Statistical Modeling
6. Handling Big Data
7. Machine Learning
8. R in Practice



About Me

Data Science Consultant
Education

B.S. in Computer Science
B.A. in Philosophy

Community

Public Speaker
Pluralsight Author
Microsoft MVP
ASPIInsider
Open-source Software

IOWA STATE
UNIVERSITY



About You

What's your name?

What do you do?

Why did you attend?

Favorite super power?



Source: www.thatsmyface.com

Schedule

Lectures (15 min)

Demos (10 min)

Labs (20 min)

Breaks (5 min)

Logistics

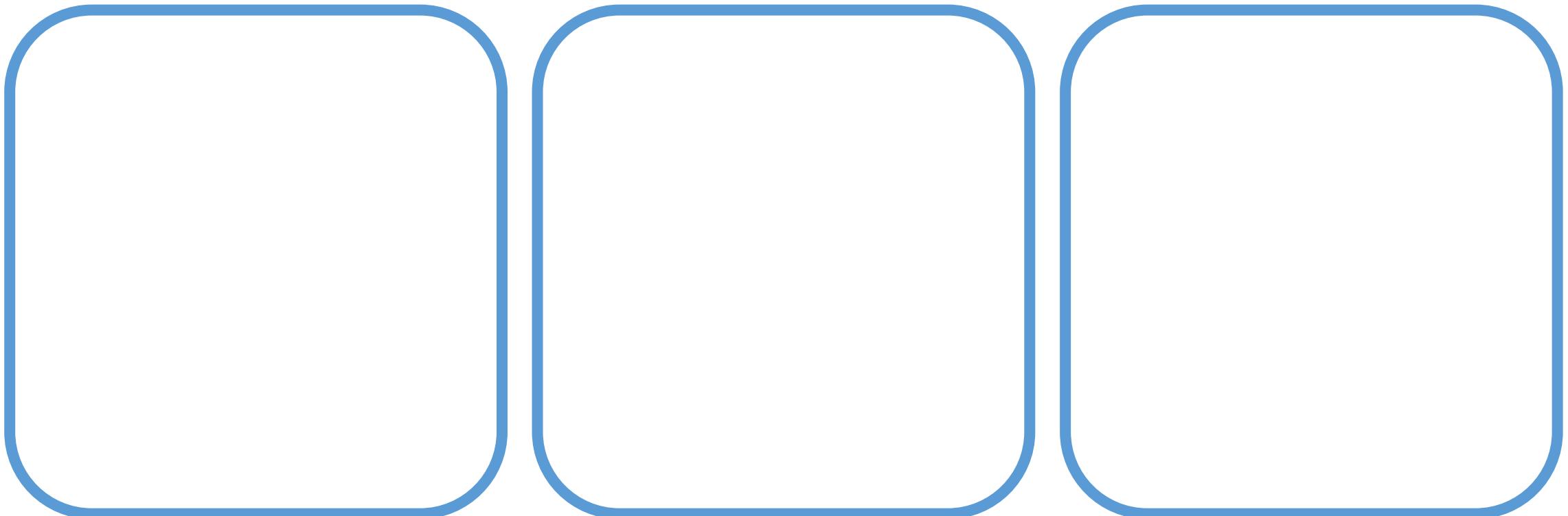
Pairing for labs is optional

Ask questions if needed

Come and go as needed

Feedback forms at the end

Labs



Labs

A
(Easy)

Labs

A

(Easy)

B

(Normal)

Labs

A

(Easy)

B

(Normal)

C

(Hard)

Labs

A

(Easy)

B

(Normal)

C

(Hard)



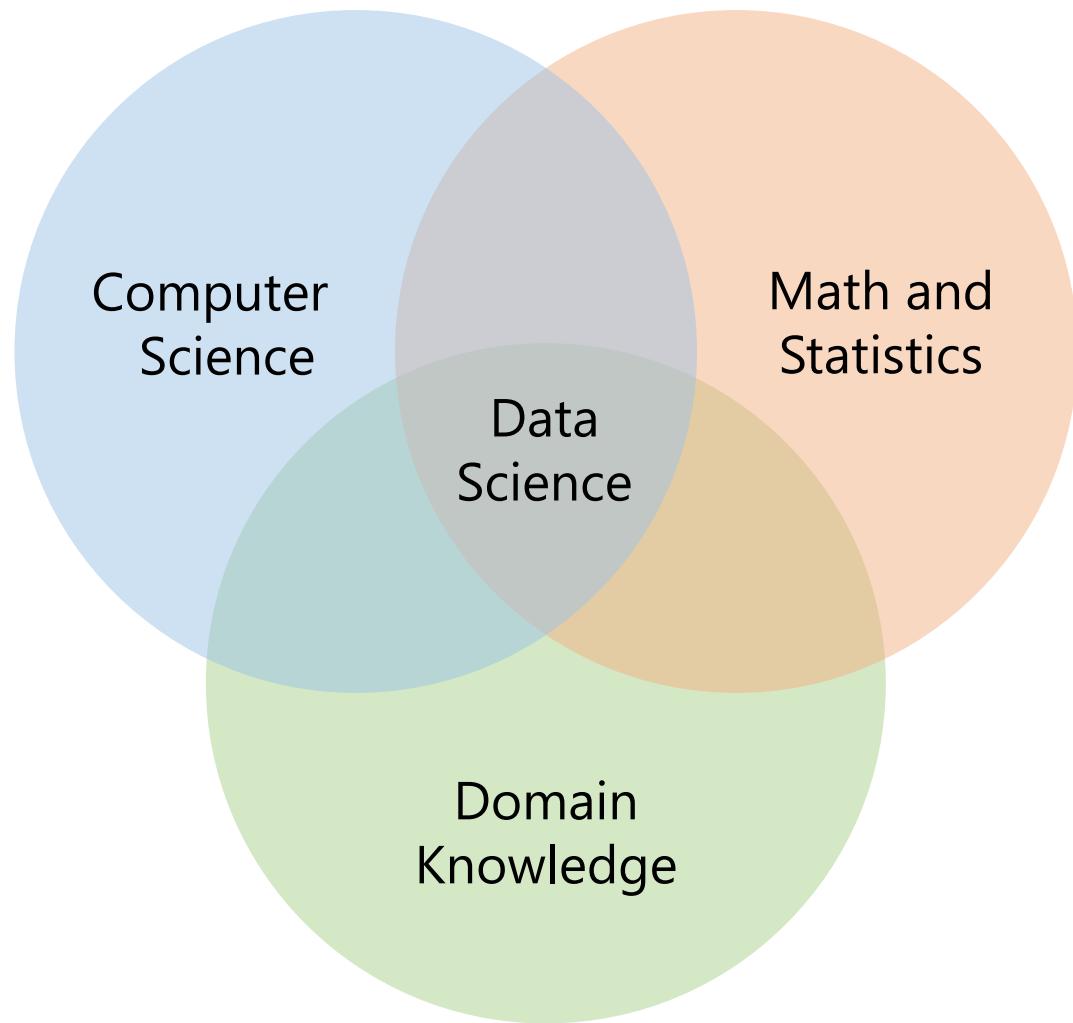
Workshop URL

<http://www.matthewrenze.com/workshops/practical-data-science-with-r/>

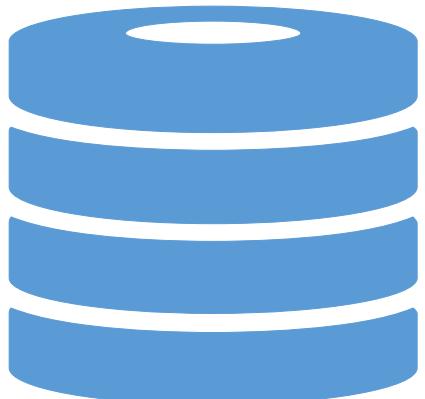
Introduction to Data Science

What is data science?

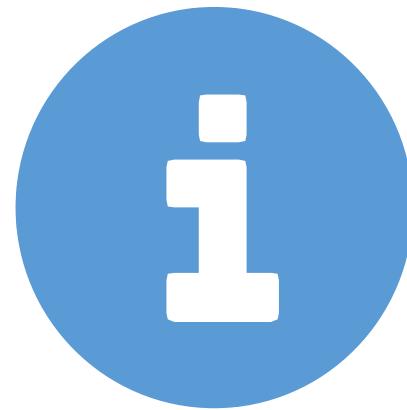
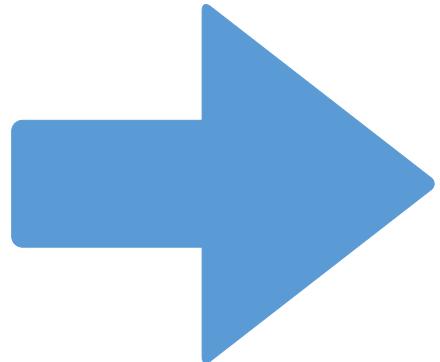
What is Data Science?



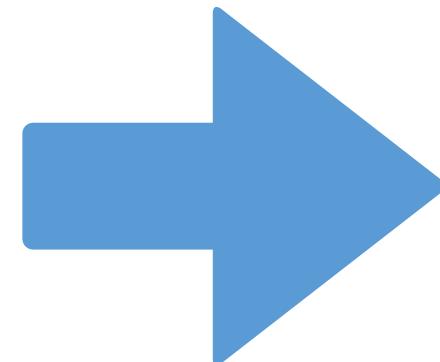
Goal of Data Science



Data



Knowledge



Decision

What is a Data Scientist?



Source: <http://www.clipartpanda.com>

What is a Data Scientist?

Performs data science
Proper accreditations



Source: <http://www.clipartpanda.com>

What is a Data Scientist?

More than a scientist
More than an analyst
More than a developer



Source: <http://www.clipartpanda.com>

Job Postings for Data Scientists



The Data Science Toolkit

Programming

Data manipulation

Descriptive statistics

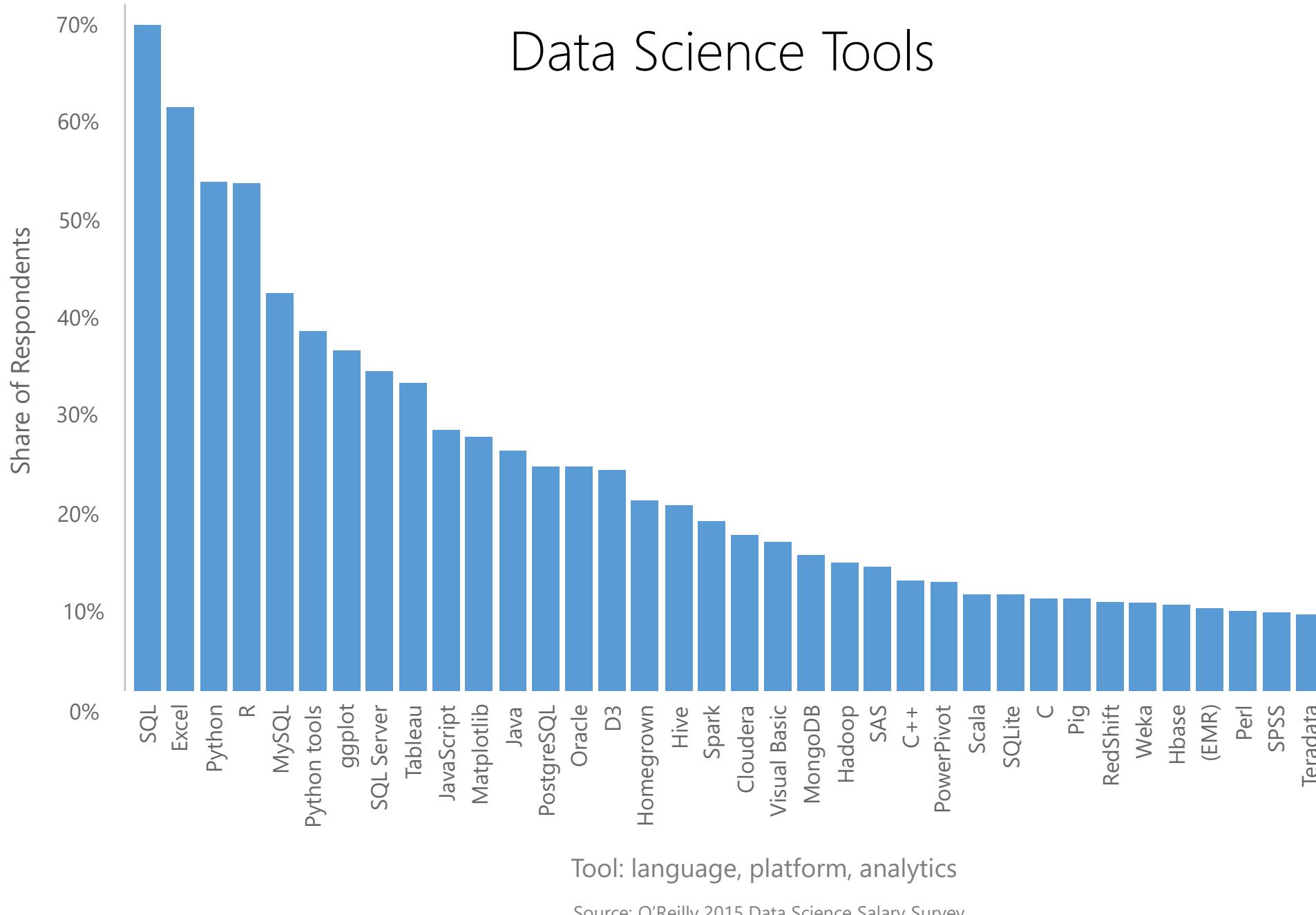
Data visualization

Statistical modeling

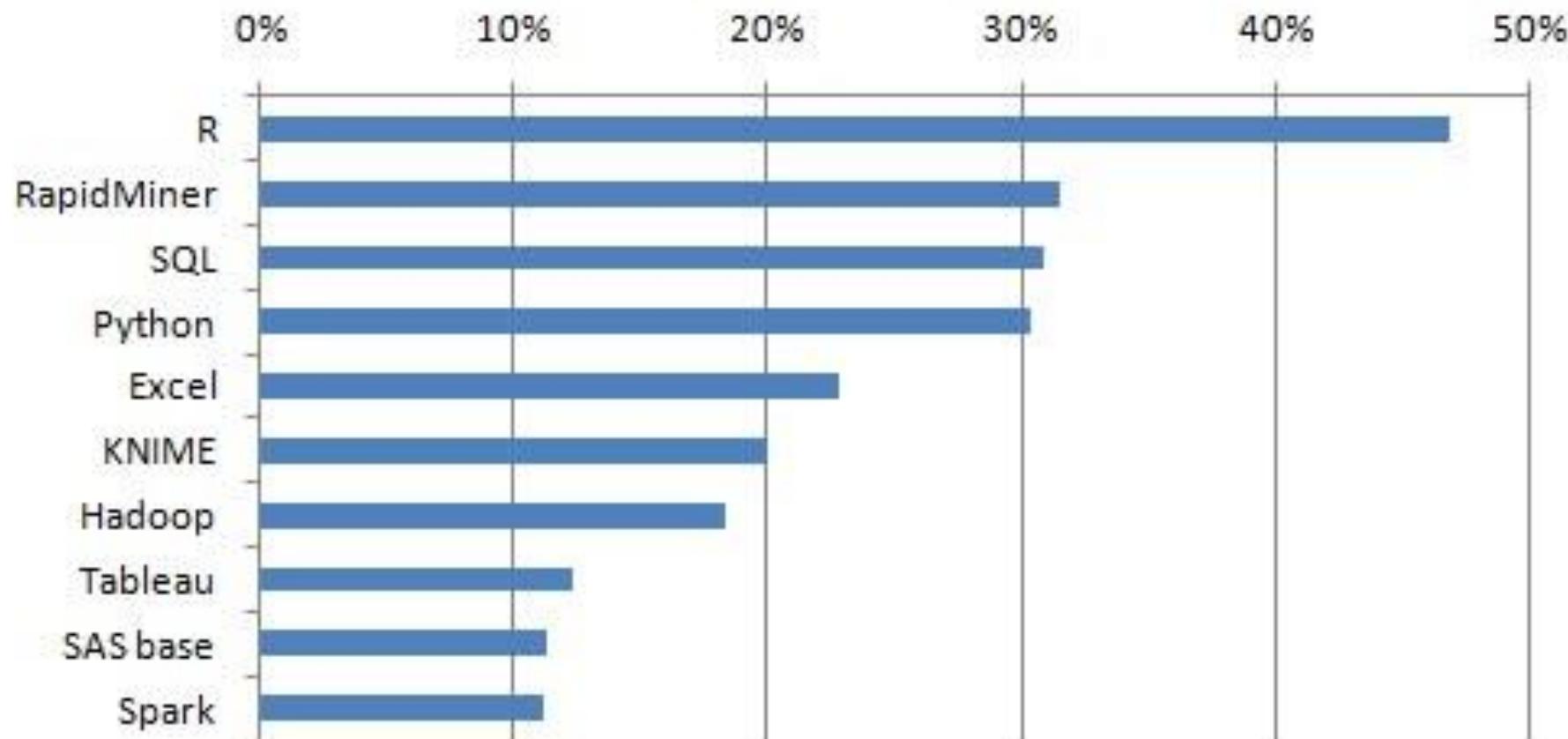
Big Data

Machine learning

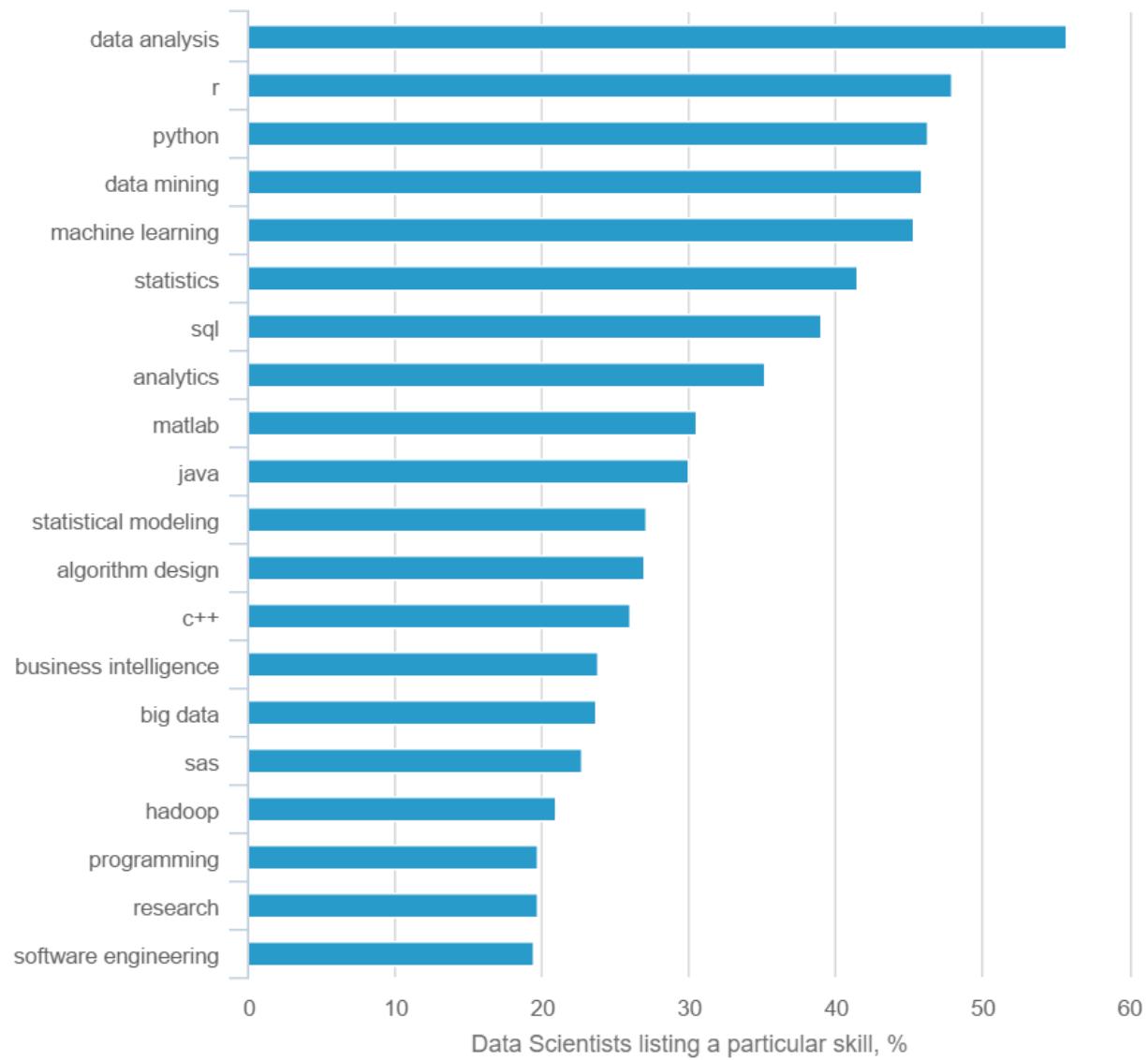
Deploying into production



Top Analytics, Data Mining, Data Science software used, 2015



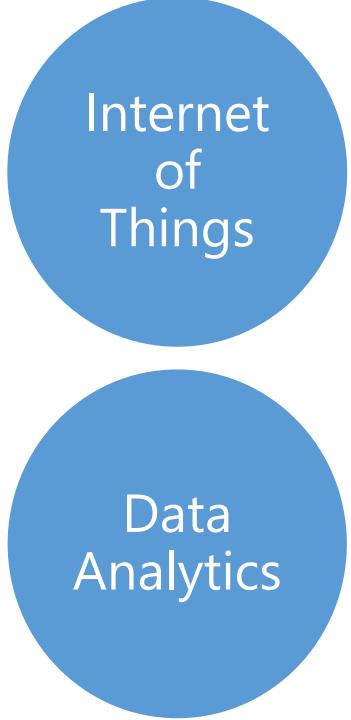
TOP 20 SKILLS OF A DATA SCIENTIST



Why is data science important?

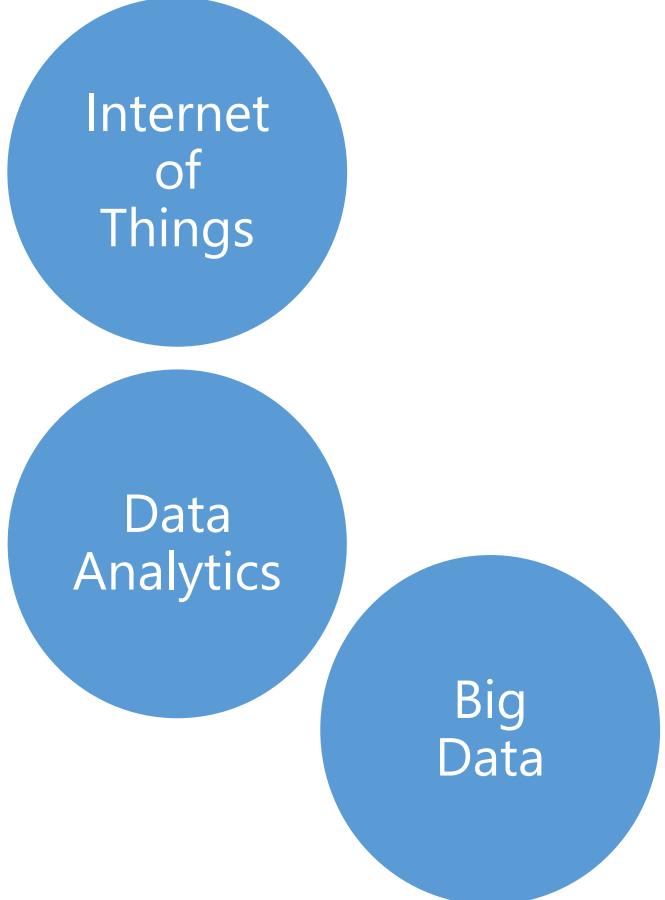


Data
Analytics



Internet
of
Things

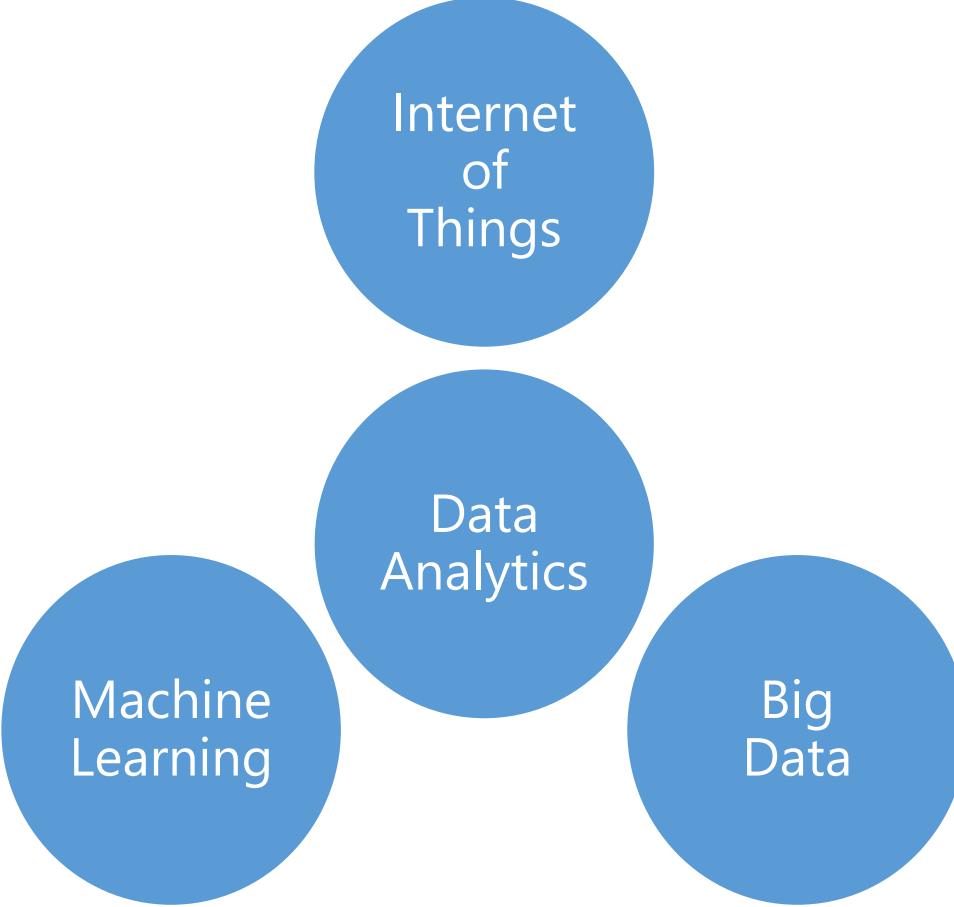
Data
Analytics



Internet
of
Things

Data
Analytics

Big
Data

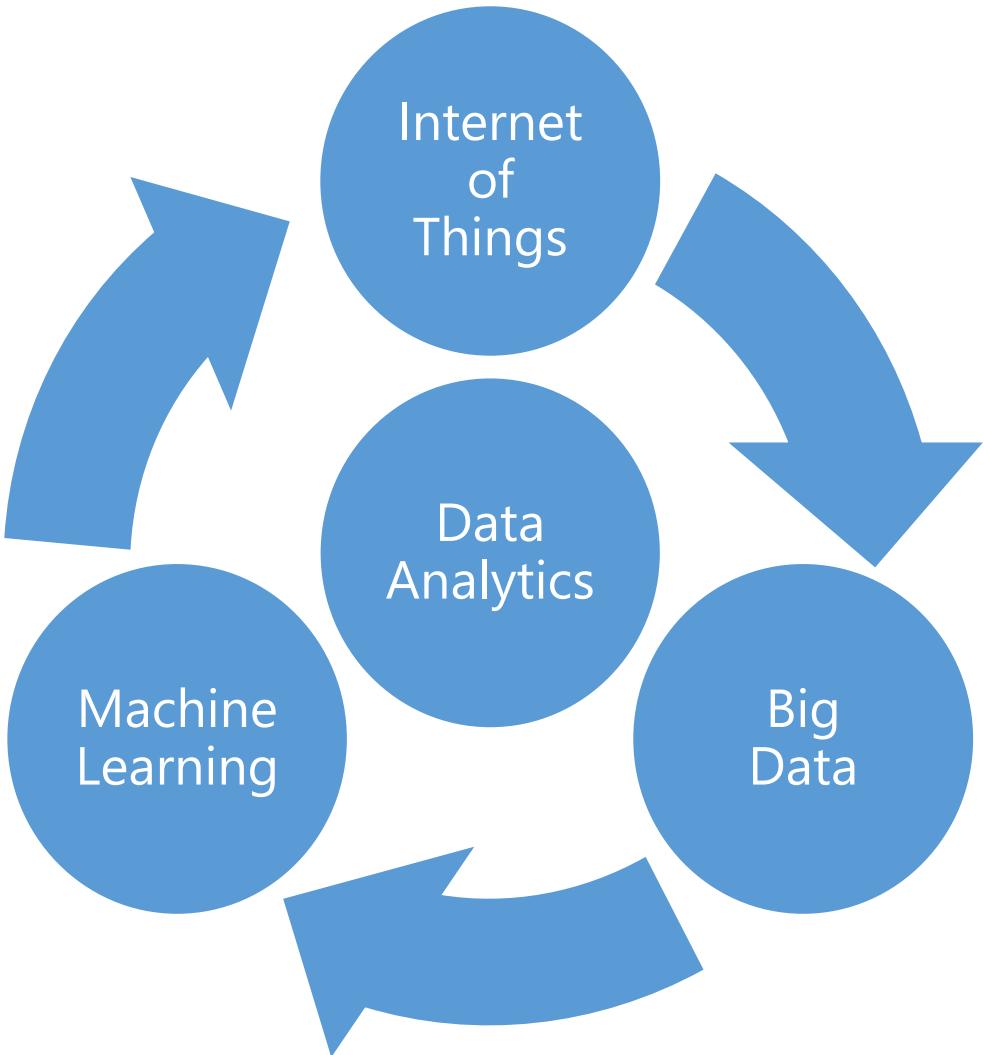


Internet
of
Things

Data
Analytics

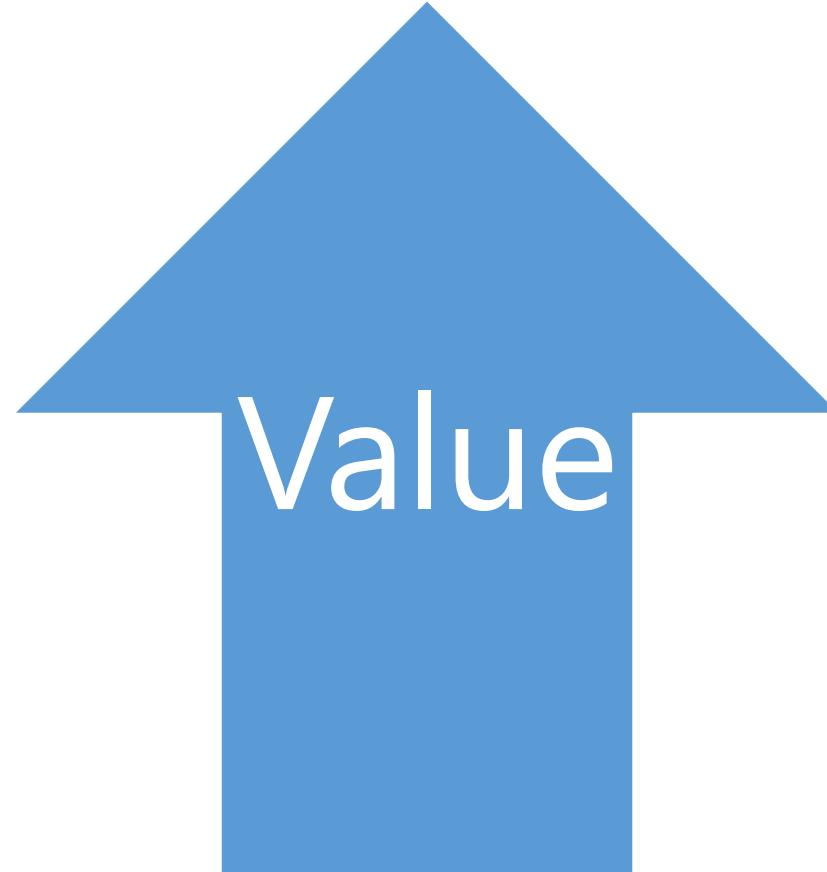
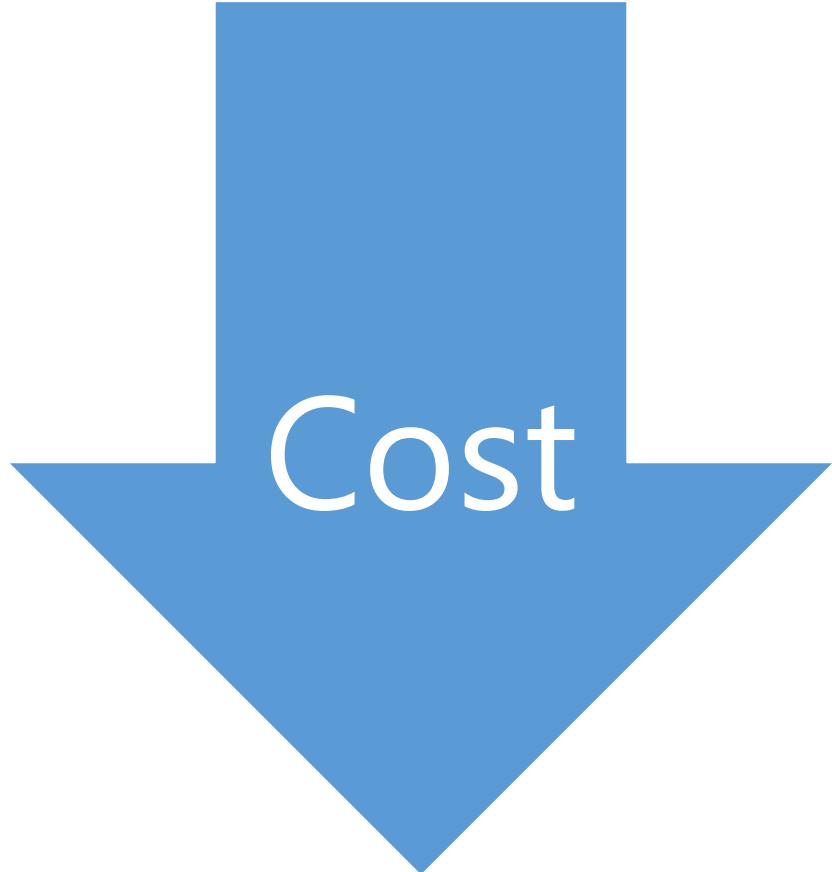
Machine
Learning

Big
Data



Driven by economics

Possible by technology



How is data science done?

The Data Science Process



Data

The Data Science Process

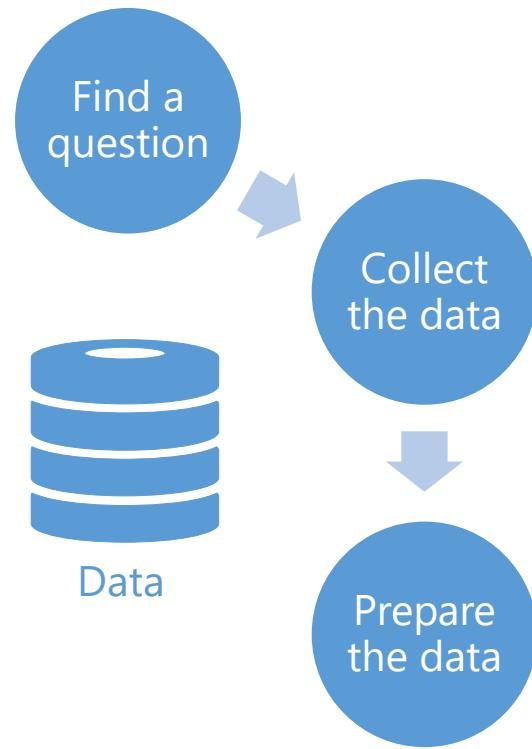


Data

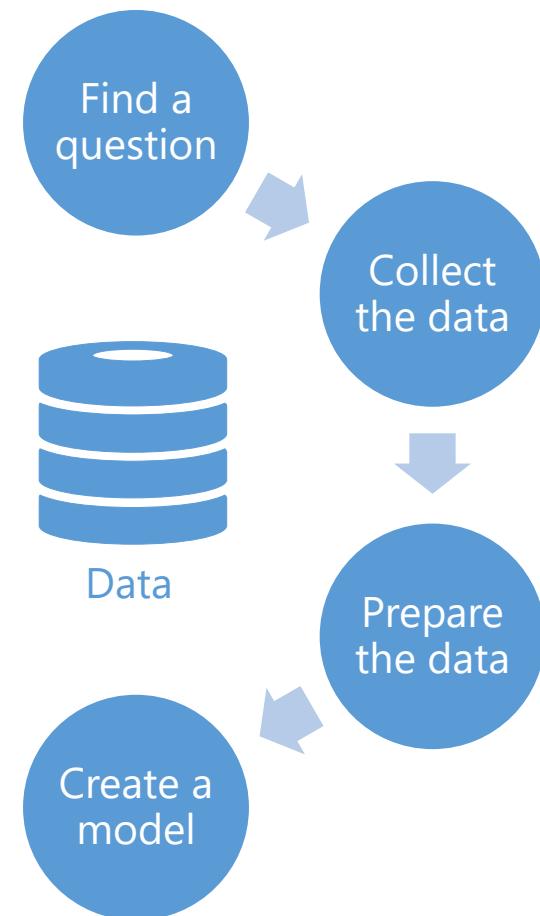
The Data Science Process



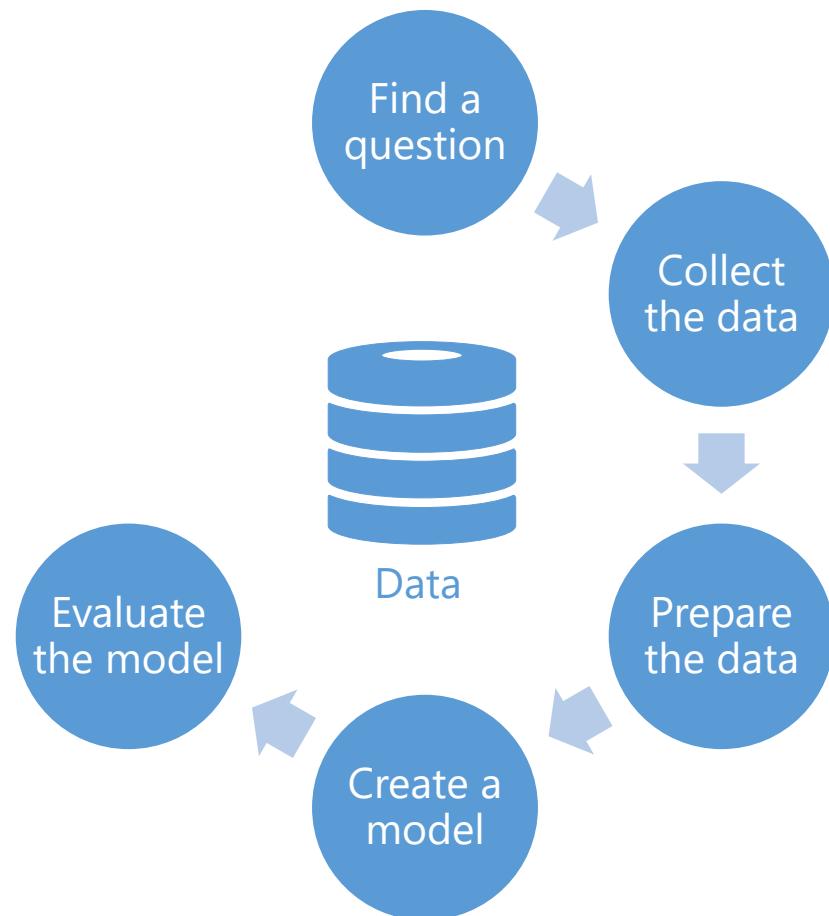
The Data Science Process



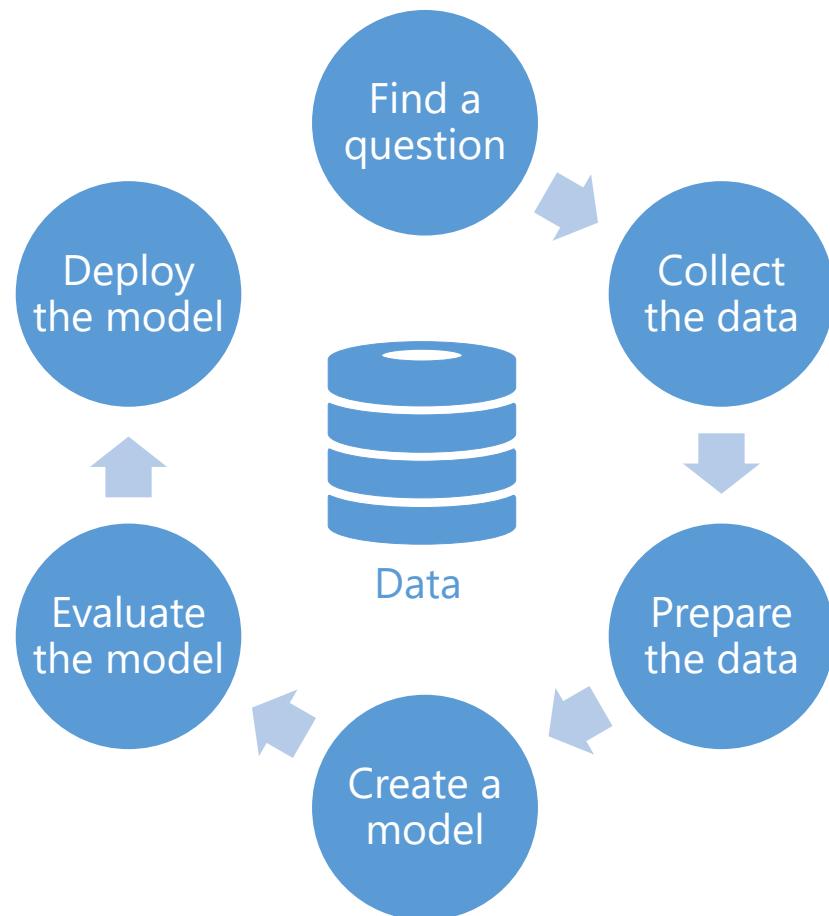
The Data Science Process



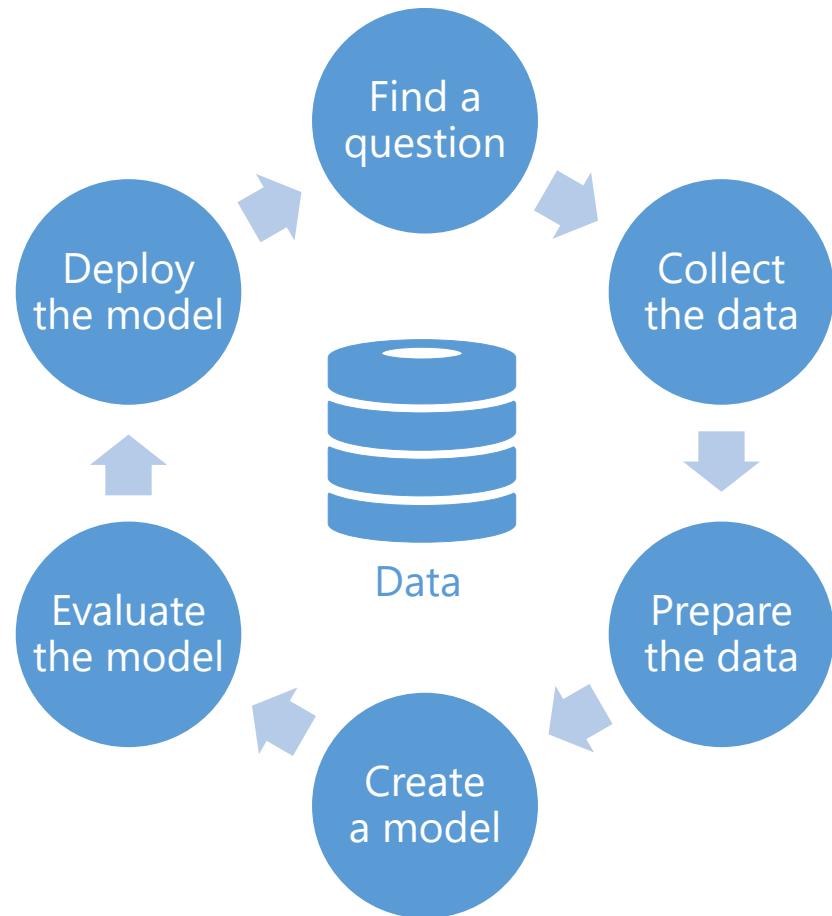
The Data Science Process



The Data Science Process

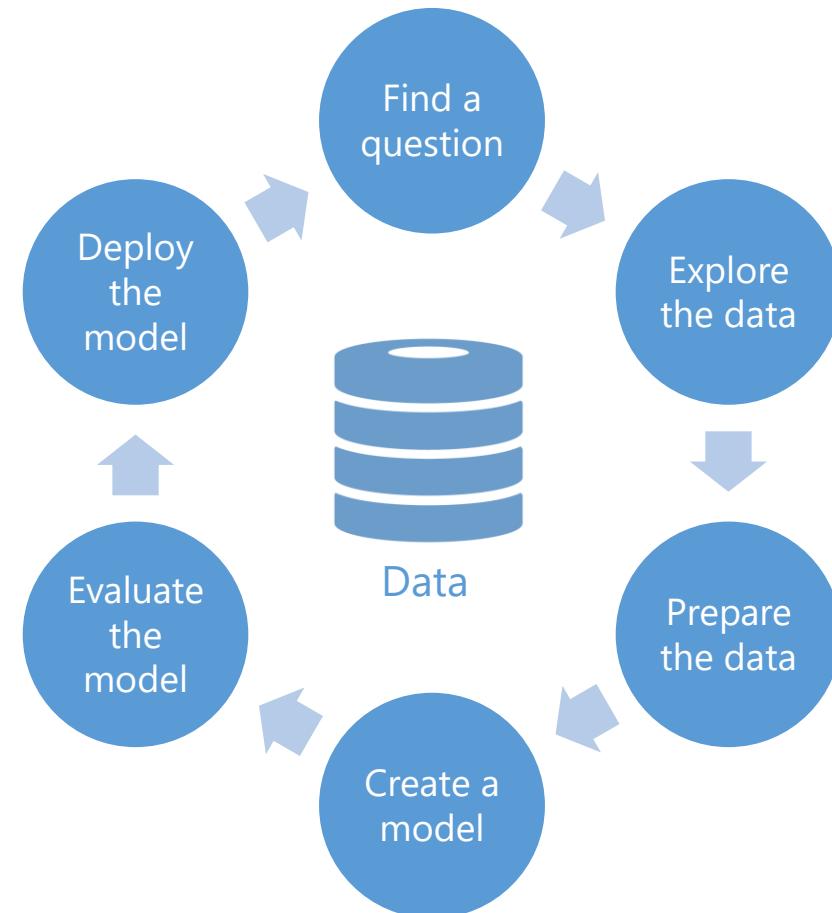


The Data Science Process



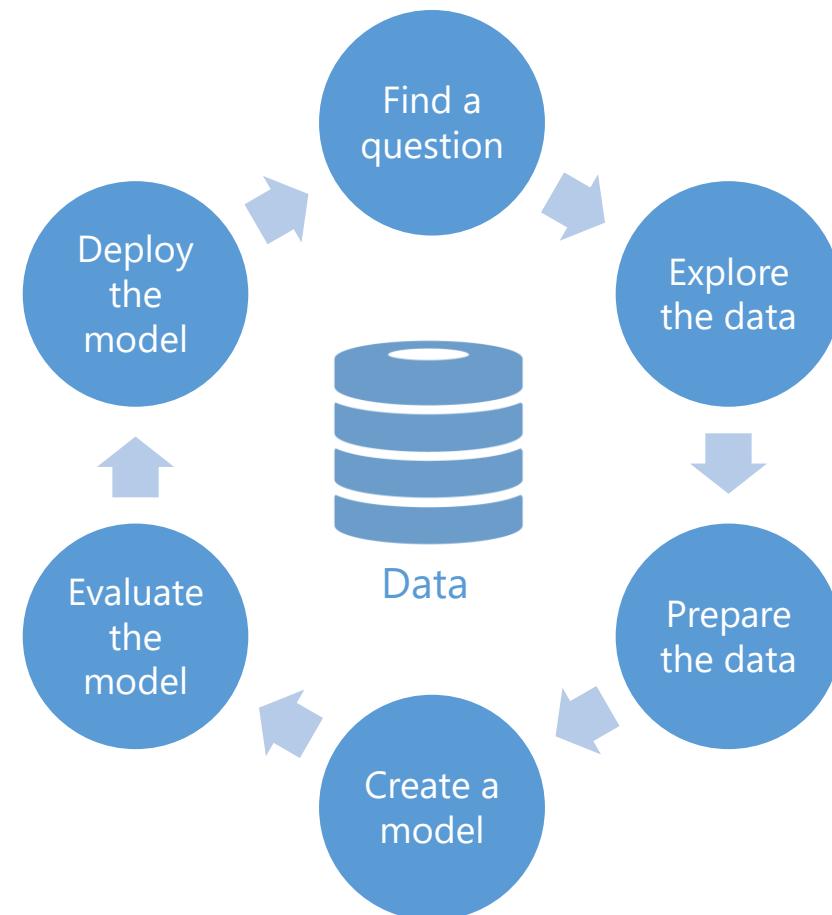
The Data Science Process

Iterative process



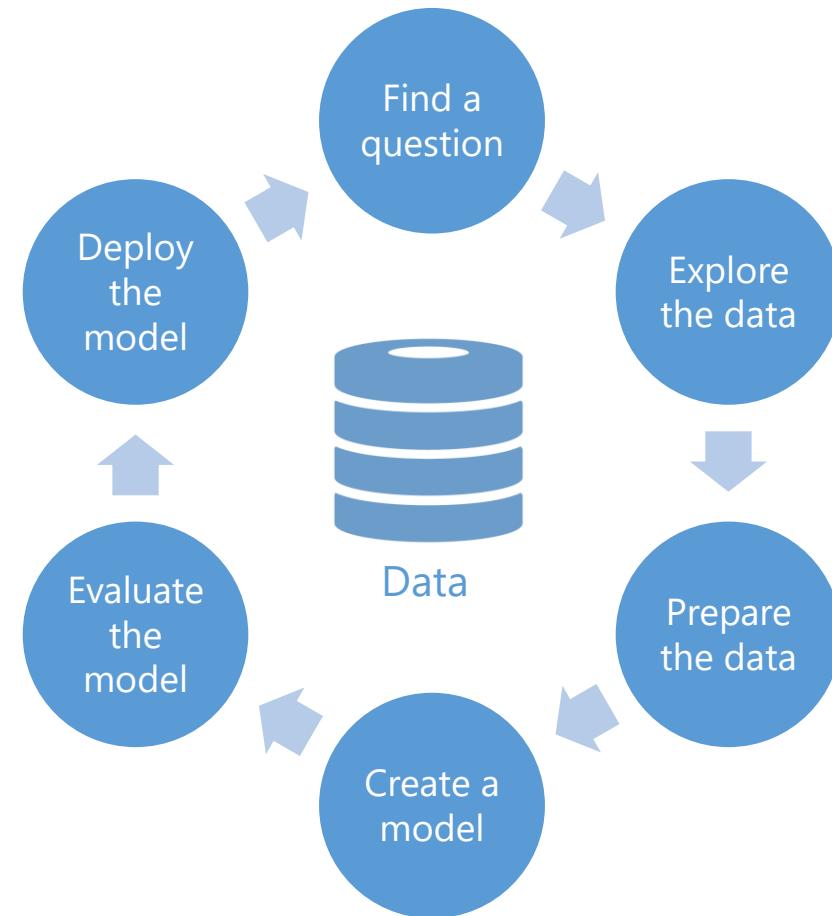
The Data Science Process

Iterative process
Non-sequential



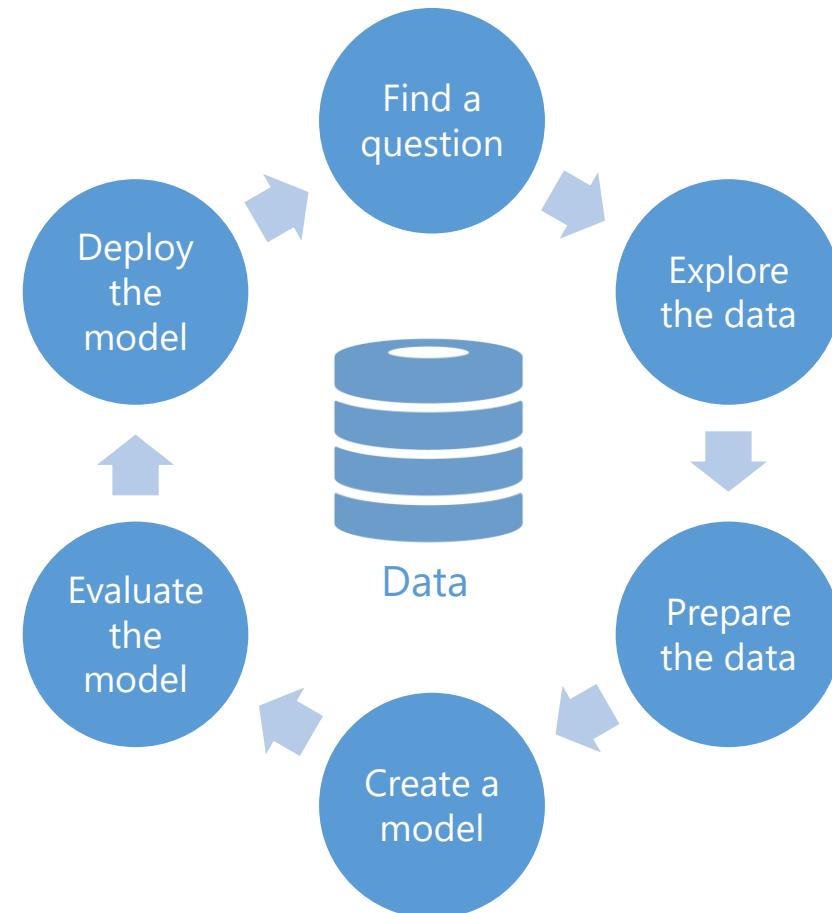
The Data Science Process

Iterative process
Non-sequential
Early termination



The Data Science Process

Iterative process
Non-sequential
Early termination
Established processes



Introduction to R

What is R?

Open source

Language and environment

Numerical and graphical

Cross platform



What is R?

Active development
Large user community
Modular and extensible
10,000+ extensions



A row of four tacos filled with meat, cheese, and cilantro, served in a metal taco holder.

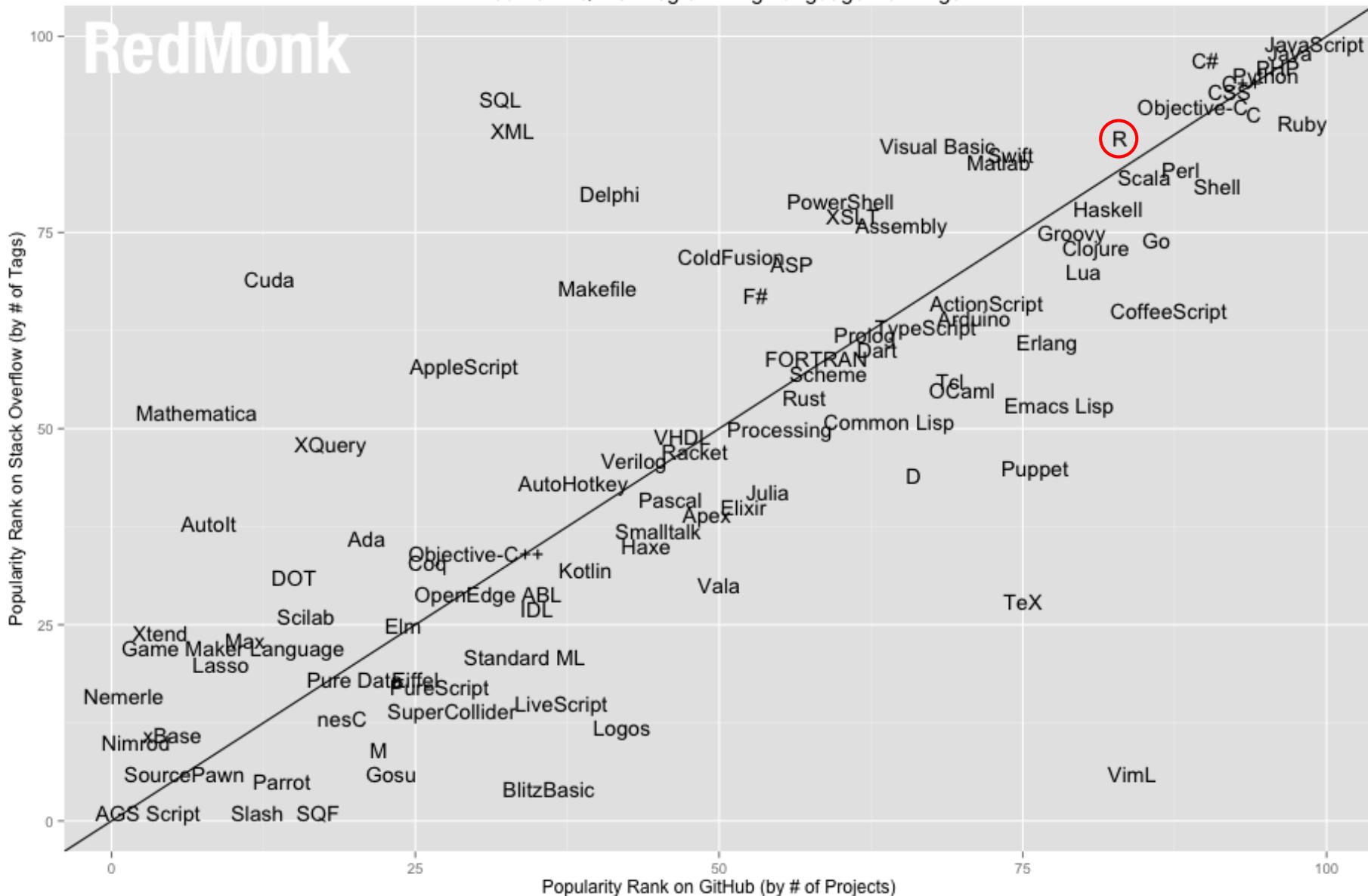
FREE!

A low-angle photograph of the Statue of Liberty against a clear blue sky. Her right arm is raised high, holding the torch aloft. Her left arm is bent, holding a tablet or smartphone that displays the word "FREE".

FREE

RedMonk

RedMonk Q116 Programming Language Rankings



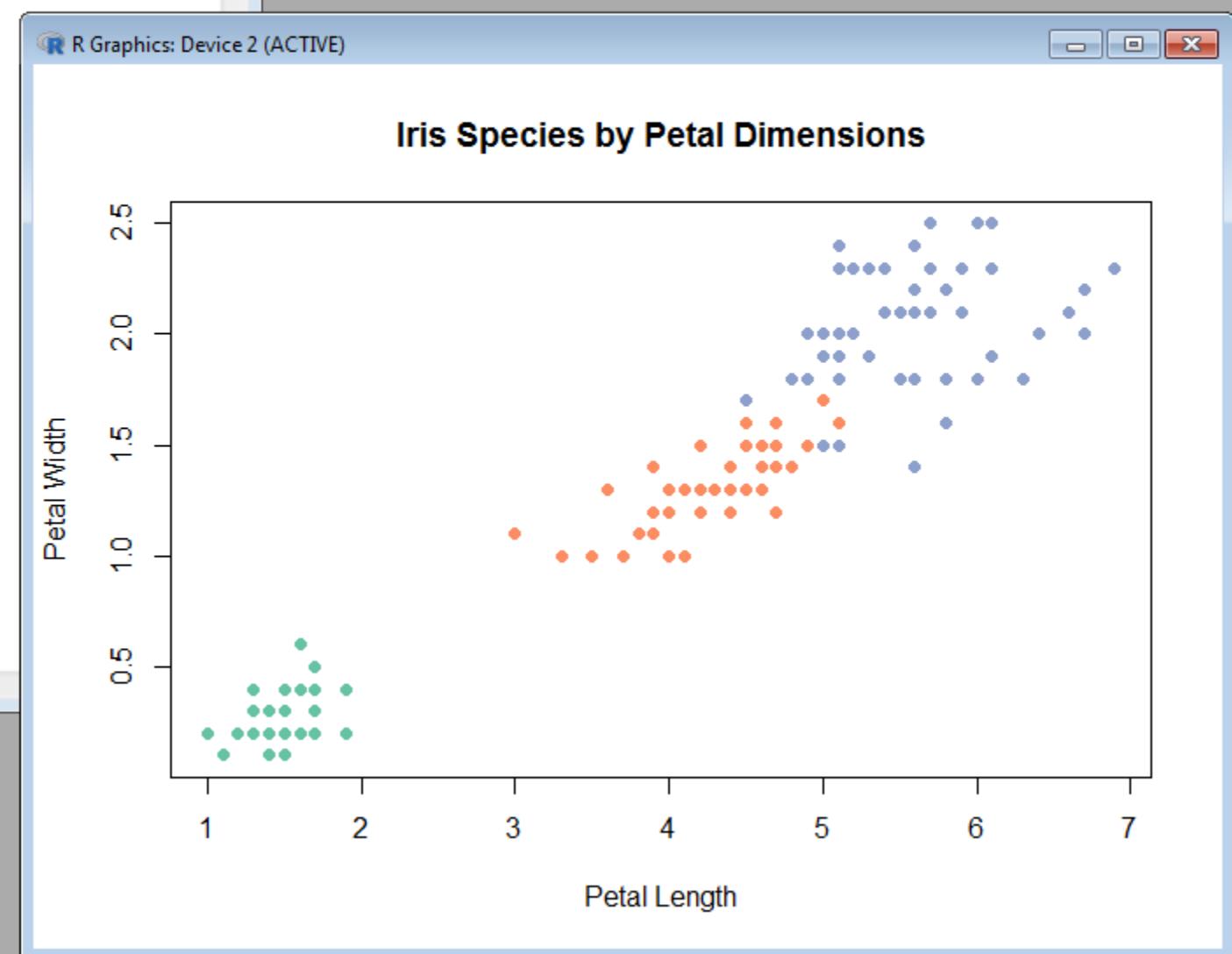


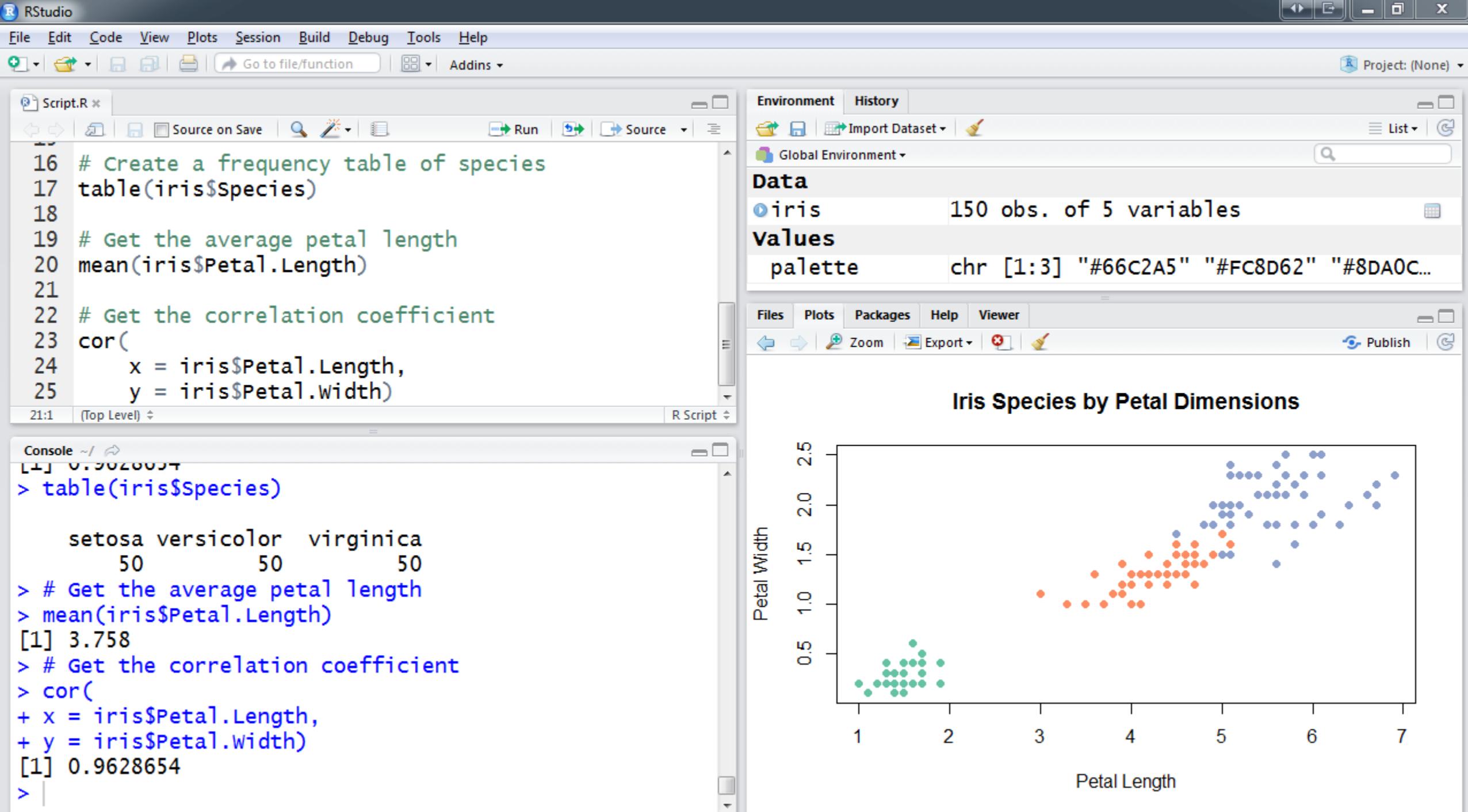
R Console

```
> # Create a plot of species by dimension
> plot(
+   x = iris$Petal.Length,
+   y = iris$Petal.Width,
+   pch = 19,
+   col = palette[as.numeric(iris$Species)],
+   main = "Iris Species by Petal Dimensions",
+   xlab = "Petal Length",
+   ylab = "Petal Width")
>
> # Create a frequency table of species
> table(iris$Species)

  setosa versicolor virginica 
      50       50       50 

>
> # Get the average petal length
> mean(iris$Petal.Length)
[1] 3.758
>
> # Get the correlation coefficient
> cor(
+   x = iris$Petal.Length,
+   y = iris$Petal.Width)
[1] 0.9628654
```





Script.R - Microsoft Visual Studio

File Edit View NCrunch Project Debug Team Tools Architecture Test ReSharper R Tools Analyze Window Help

Matthew Renze

Quick Launch (Ctrl+Q)

Script.R

```
main = "Iris Species by Petal Dimensions",
xlab = "Petal Length",
ylab = "Petal Width")

# Create a frequency table of species
table(iris$Species)

# Get the average petal length
mean(iris$Petal.Length)

# Get the correlation coefficient
cor(
  x = iris$Petal.Length,
  y = iris$Petal.Width)
```

R Interactive

```
> # Create a frequency table of species
> table(iris$Species)

  setosa versicolor virginica
      50          50         50
> # Get the average petal length
> mean(iris$Petal.Length)
[1] 3.758
> # Get the correlation coefficient
> cor(
+   x = iris$Petal.Length,
+   y = iris$Petal.Width)
[1] 0.9628654
>
```

Variable Explorer

| Name | Value | Class | Type |
|---------|---|------------|-----------|
| iris | 150 obs. of 5 variables | data.frame | list |
| palette | chr [1:3] "#66C2A5" "#FC8D62" "#8DA0CE" | character | character |

Variable Explorer R History

R Plot

Iris Species by Petal Dimensions

Petal Width

Petal Length

Solution Explorer R Plot R Package Manager R Help

Error List Output Azure App Service Activity

Ln 30 Col1 Ch1 INS ↑ 7 ↗ 0 ↘ Root ↞ master ↞

Code Demo

Tips for Labs

- Use script pane
- Ctrl + Enter
- Multi-line execution
- Up arrow for history

Lab 1

R Programming Basics

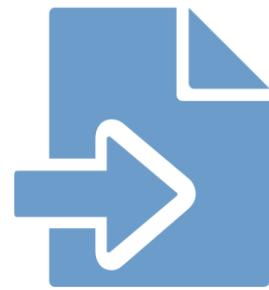
Working with Data

Working with Data

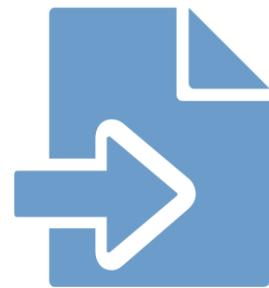
Working with Data



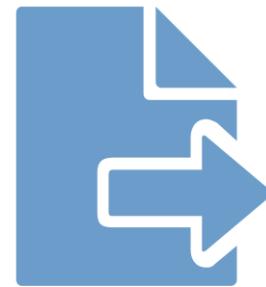
Working with Data



Working with Data



Working with Data



Loading Data in R

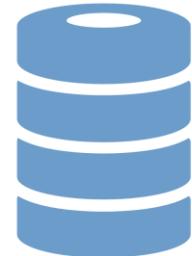
Loading Data in R



Loading Data in R



Loading Data in R



Loading Data in R



Cleaning Data



Cleaning Data

Reshape data



Cleaning Data

Reshape data

Rename columns



Cleaning Data

Reshape data

Rename columns

Convert data types



Cleaning Data

Reshape data

Rename columns

Convert data types

Ensure proper encoding



Cleaning Data

Reshape data

Rename columns

Convert data types

Ensure proper encoding

Ensure internal consistency



Cleaning Data

Reshape data

Rename columns

Convert data types

Ensure proper encoding

Ensure internal consistency

Handle errors and outliers



Cleaning Data

Reshape data

Rename columns

Convert data types

Ensure proper encoding

Ensure internal consistency

Handle errors and outliers

Handle missing values



Tools for Cleaning Data



Tools for Cleaning Data

base



Tools for Cleaning Data

base

tidyverse



Tools for Cleaning Data

base

tidyr

reshape2



Tools for Cleaning Data

base

tidyr

reshape2

stringr



Tools for Cleaning Data

base

tidyr

reshape2

stringr

lubridate



Tools for Cleaning Data

base

tidyr

reshape2

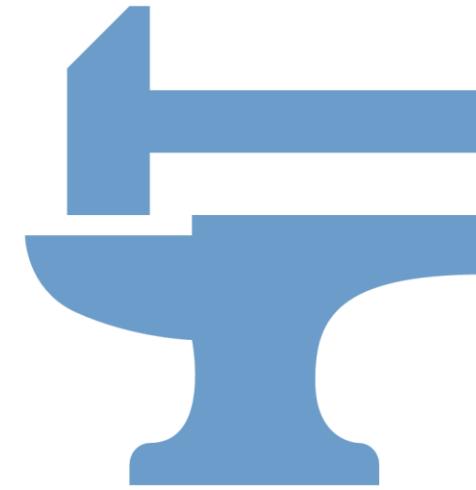
stringr

lubridate

validate

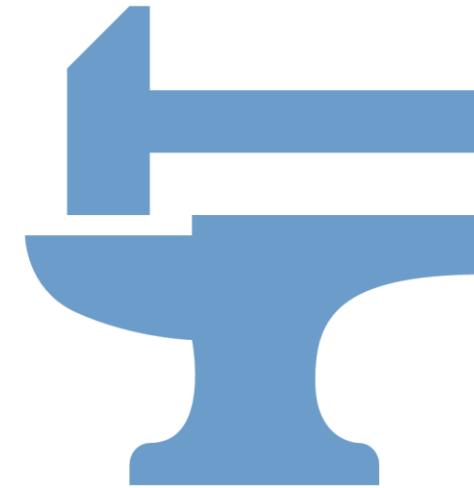


Transforming Data



Transforming Data

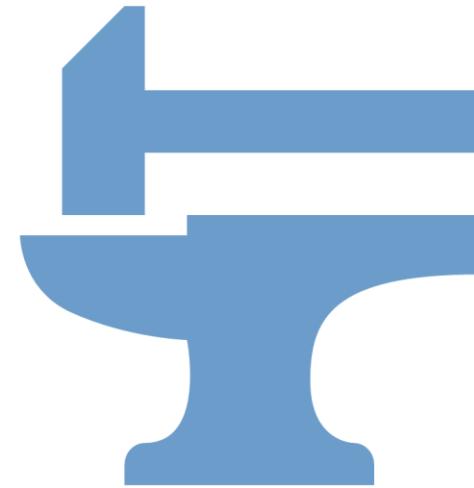
Select columns



Transforming Data

Select columns

Select rows

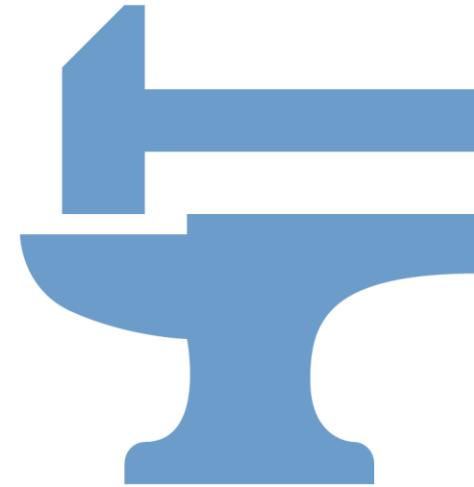


Transforming Data

Select columns

Select rows

Group rows



Transforming Data

Select columns

Select rows

Group rows

Order rows



Transforming Data

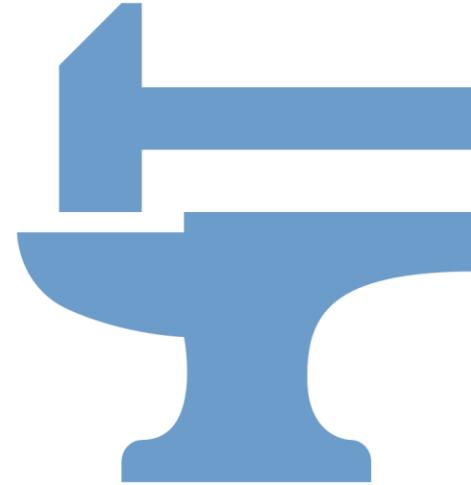
Select columns

Select rows

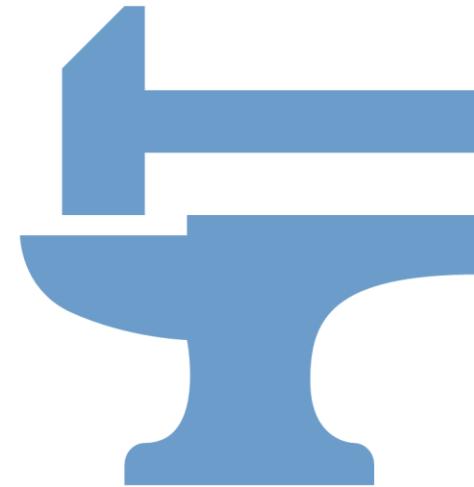
Group rows

Order rows

Merging data sets

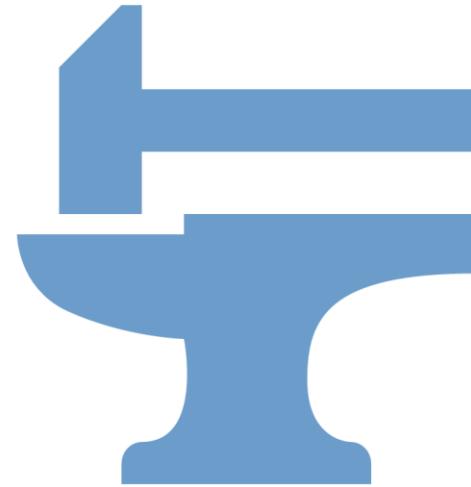


Tools for Transforming Data



Tools for Transforming Data

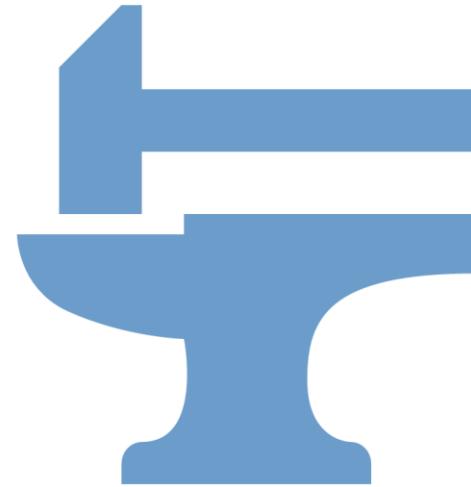
base



Tools for Transforming Data

base

plyr

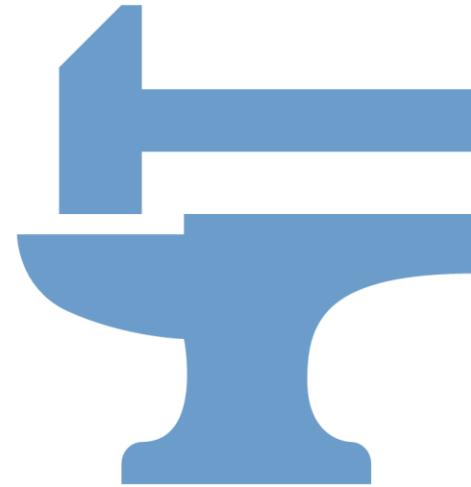


Tools for Transforming Data

base

plyr

dplyr



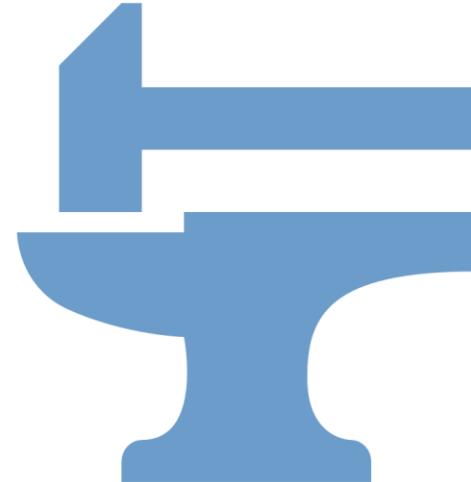
Tools for Transforming Data

base

plyr

dplyr

data.table



Tools for Transforming Data

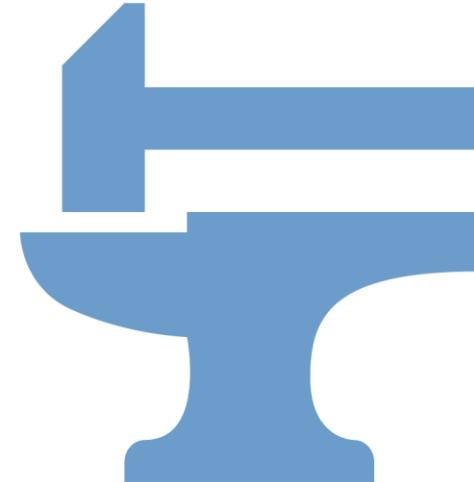
base

plyr

dplyr

data.table

sqldf



Exporting Data

File-based data

Web-based data

Databases

Statistical data



Advice for Working with Data

Data munging is difficult
and time consuming

TIP: Record all steps





PROD. NO.
SCENE

TAKE

ROLL

Open Movies Database

| Movies | | | | | | |
|----------------------|------|--------|----------------------|--------|-----------------|---------------|
| Title | Year | Rating | Runtime (minutes) | Genre | Critic Score | Box Office |
| The Whole Nine Yards | 2000 | R | 98 | Comedy | 45% | \$57.3M |
| Cirque du Soleil | 2000 | G | 39 | Family | 45% | \$13.4M |
| Gladiator | 2000 | R | 155 | Action | 76% | \$187.3M |
| Dinosaur | 2000 | PG | 82 | Family | 65% | \$135.6M |
| Big Momma's House | 2000 | PG-13 | 99 | Comedy | 30% | \$0.5M |



PROD. NO.
SCENE

TAKE

ROLL





1. Column with wrong name
2. Rows with missing values
3. Runtime column has units
4. Revenue in multiple scales
5. Wrong file format

Code Demo

Lab 2

Transforming Data



Descriptive Statistics

Descriptive Statistics

Describe data

Provides a summary

aka: Summary statistics

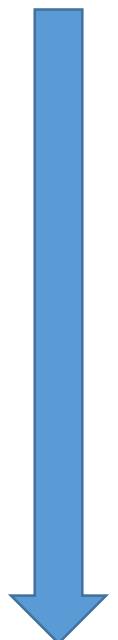
| Movie Runtime | |
|--------------------------|-----------------|
| Statistic | Value (minutes) |
| Minimum | 38 |
| 1 st Quartile | 93 |
| Median | 101 |
| Mean | 104 |
| 3 rd Quartile | 113 |
| Maximum | 219 |

Statistical Terms

| ID | Date | Customer | Product | Quantity |
|-----------|-------------|-----------------|----------------|-----------------|
| 1 | 2015-08-27 | John | Pizza | 2 |
| 2 | 2015-08-27 | John | Soda | 2 |
| 3 | 2015-08-27 | Jill | Salad | 1 |
| 4 | 2015-08-27 | Jill | Milk | 1 |
| 5 | 2015-08-28 | Miko | Pizza | 3 |
| 6 | 2015-08-28 | Miko | Soda | 2 |
| 7 | 2015-08-28 | Sam | Pizza | 1 |
| 8 | 2015-08-28 | Sam | Milk | 1 |

Statistical Terms

Observations



| ID | Date | Customer | Product | Quantity |
|----|------------|----------|---------|----------|
| 1 | 2015-08-27 | John | Pizza | 2 |
| 2 | 2015-08-27 | John | Soda | 2 |
| 3 | 2015-08-27 | Jill | Salad | 1 |
| 4 | 2015-08-27 | Jill | Milk | 1 |
| 5 | 2015-08-28 | Miko | Pizza | 3 |
| 6 | 2015-08-28 | Miko | Soda | 2 |
| 7 | 2015-08-28 | Sam | Pizza | 1 |
| 8 | 2015-08-28 | Sam | Milk | 1 |

Statistical Terms

Observations
Variables



| ID | Date | Customer | Product | Quantity |
|----|------------|----------|---------|----------|
| 1 | 2015-08-27 | John | Pizza | 2 |
| 2 | 2015-08-27 | John | Soda | 2 |
| 3 | 2015-08-27 | Jill | Salad | 1 |
| 4 | 2015-08-27 | Jill | Milk | 1 |
| 5 | 2015-08-28 | Miko | Pizza | 3 |
| 6 | 2015-08-28 | Miko | Soda | 2 |
| 7 | 2015-08-28 | Sam | Pizza | 1 |
| 8 | 2015-08-28 | Sam | Milk | 1 |

Statistical Terms

Observations

Variables

Categorical variables

| ID | Date | Customer | Product | Quantity |
|----|------------|----------|---------|----------|
| 1 | 2015-08-27 | John | Pizza | 2 |
| 2 | 2015-08-27 | John | Soda | 2 |
| 3 | 2015-08-27 | Jill | Salad | 1 |
| 4 | 2015-08-27 | Jill | Milk | 1 |
| 5 | 2015-08-28 | Miko | Pizza | 3 |
| 6 | 2015-08-28 | Miko | Soda | 2 |
| 7 | 2015-08-28 | Sam | Pizza | 1 |
| 8 | 2015-08-28 | Sam | Milk | 1 |

Statistical Terms

Observations

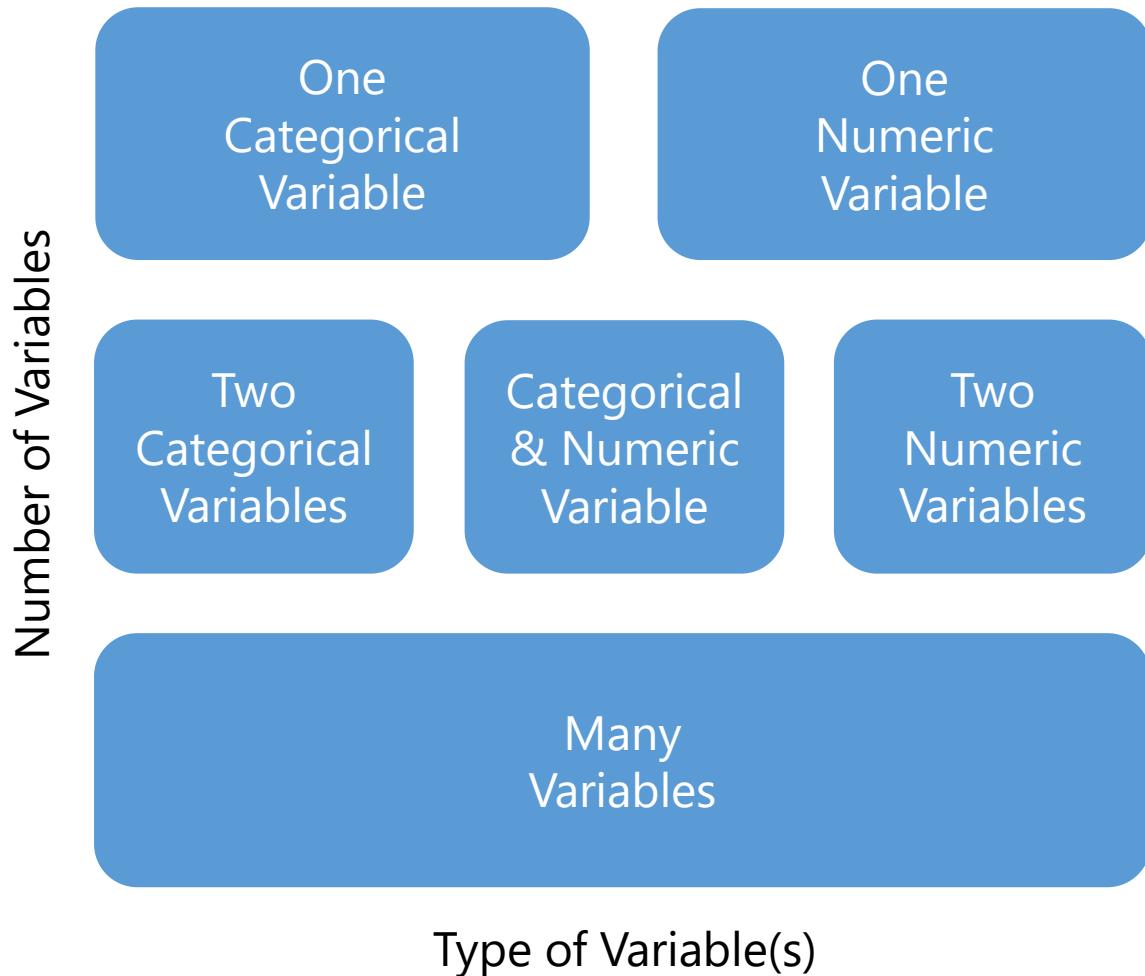
Variables

Categorical variables

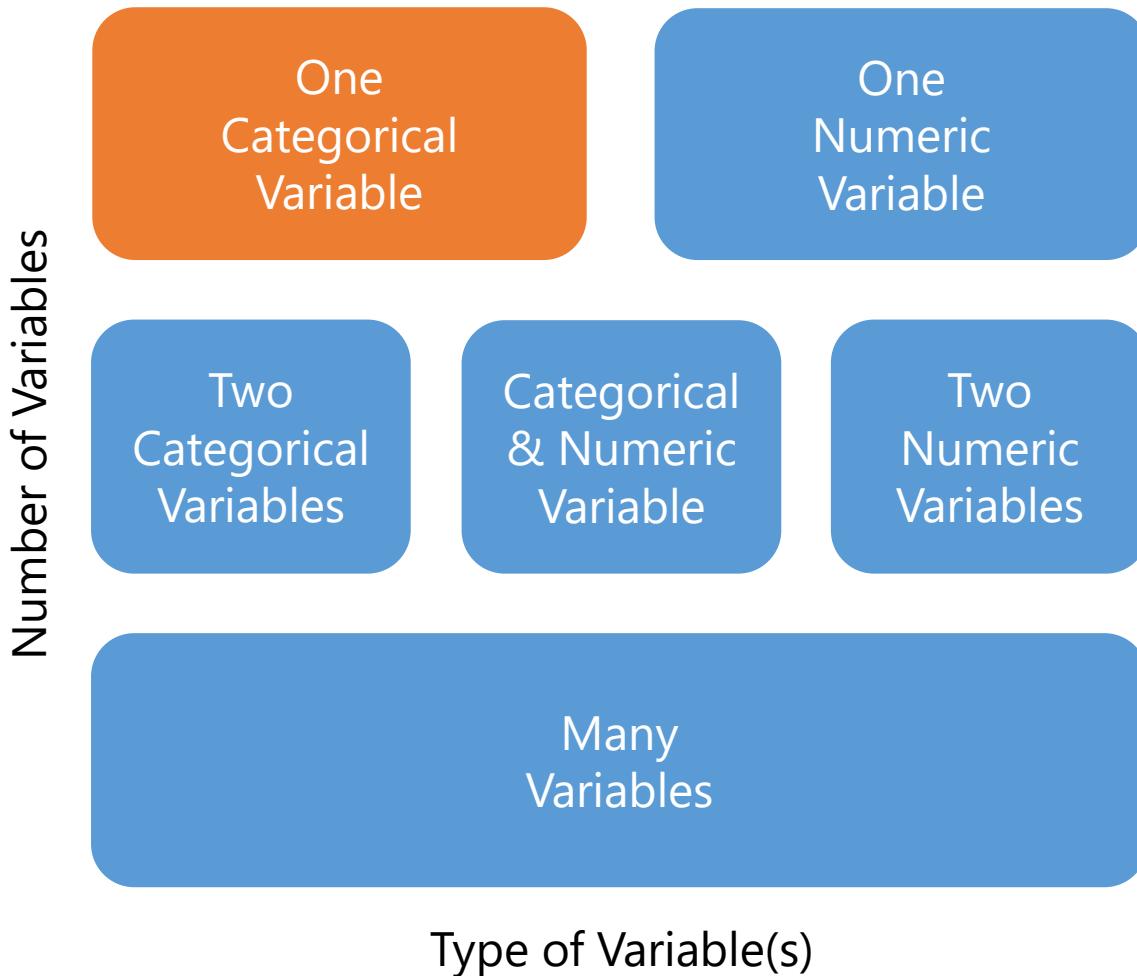
Numeric variables

| ID | Date | Customer | Product | Quantity |
|----|------------|----------|---------|----------|
| 1 | 2015-08-27 | John | Pizza | 2 |
| 2 | 2015-08-27 | John | Soda | 2 |
| 3 | 2015-08-27 | Jill | Salad | 1 |
| 4 | 2015-08-27 | Jill | Milk | 1 |
| 5 | 2015-08-28 | Miko | Pizza | 3 |
| 6 | 2015-08-28 | Miko | Soda | 2 |
| 7 | 2015-08-28 | Sam | Pizza | 1 |
| 8 | 2015-08-28 | Sam | Milk | 1 |

Types of Analysis



Analyzing One Categorical Variable



Analyzing One Categorical Variable

Frequency of observations

Frequency table

| Movies by Genre | | |
|-----------------|-----------|------------|
| Genre | Frequency | Percentage |
| Action | 612 | 9% |
| Adventure | 496 | 7% |
| Animation | 168 | 2% |
| Comedy | 1281 | 18% |
| Drama | 1570 | 22% |
| Horror | 269 | 4% |
| ... | ... | ... |

Analyzing One Categorical Variable

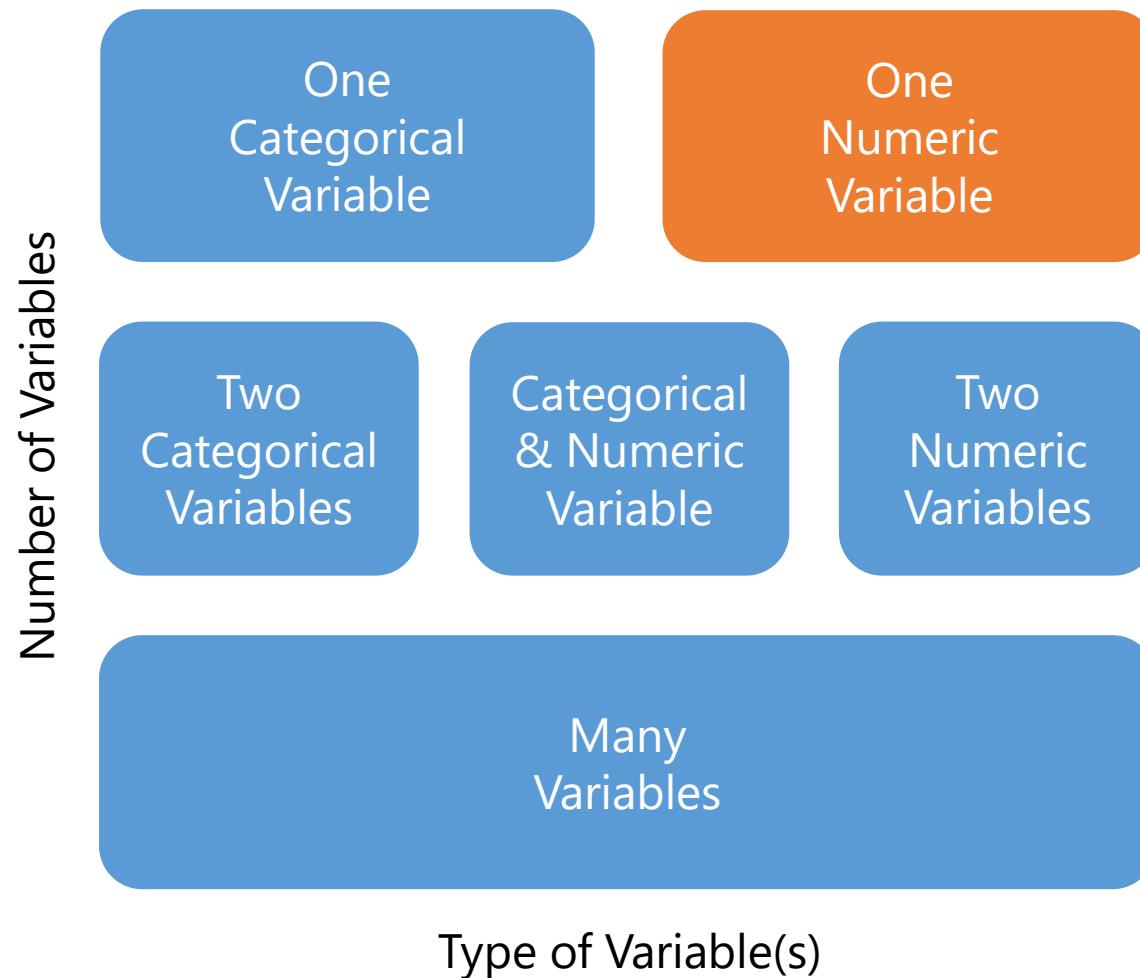
Frequency of observations

Frequency table

Proportions

| Movies by Genre | | |
|-----------------|-----------|------------|
| Genre | Frequency | Percentage |
| Action | 612 | 9% |
| Adventure | 496 | 7% |
| Animation | 168 | 2% |
| Comedy | 1281 | 18% |
| Drama | 1570 | 22% |
| Horror | 269 | 4% |
| ... | ... | ... |

Analyzing One Numeric Variable

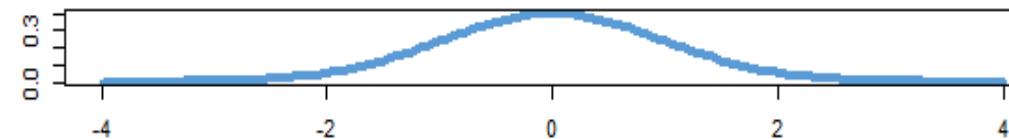
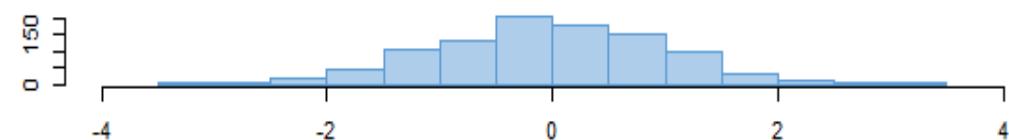
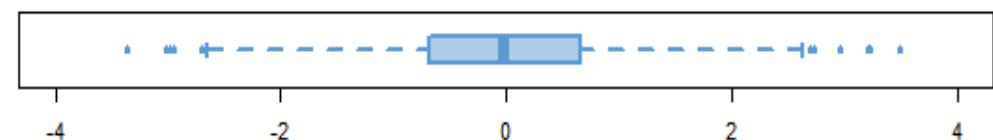


Analyzing One Numeric Variable

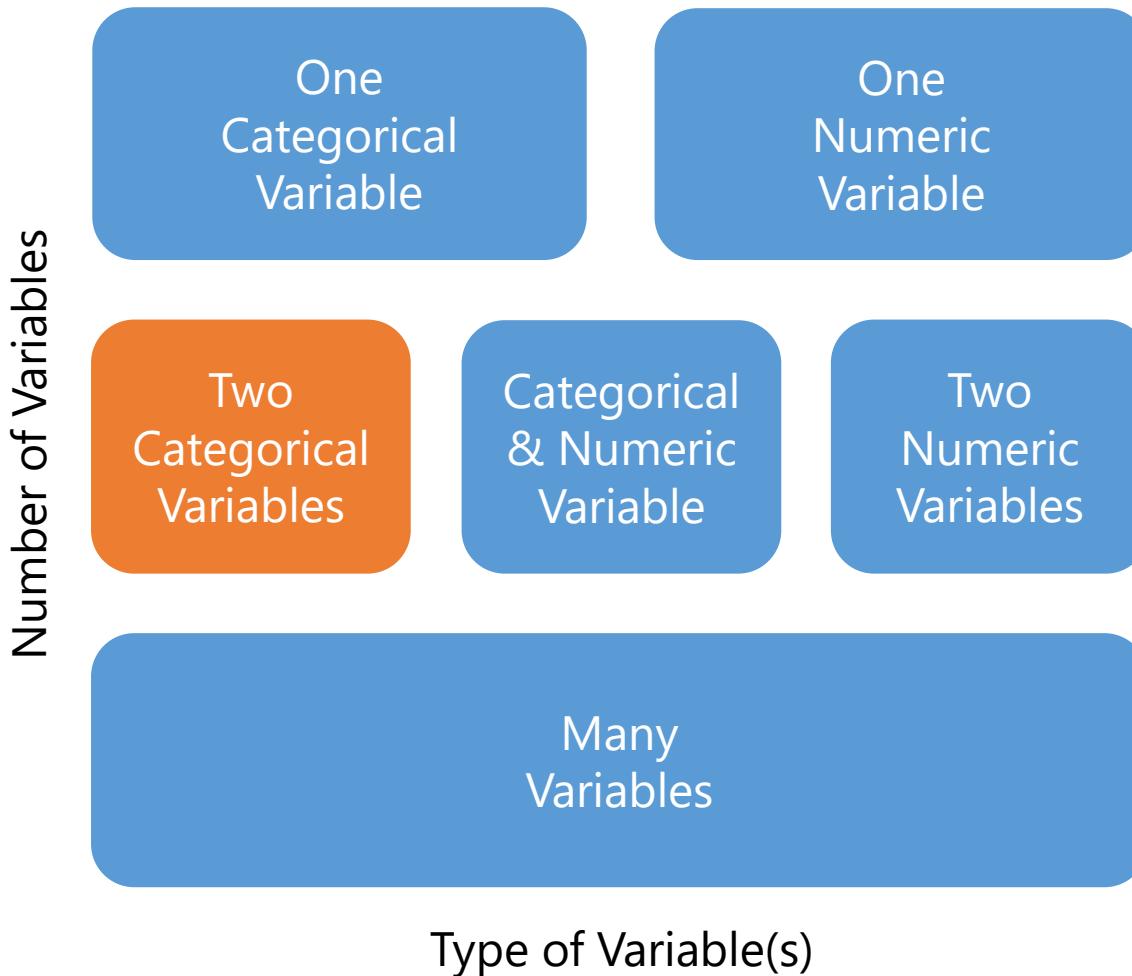
Central tendency

Dispersion

Shape



Analyzing Two Categorical Variables



Analyzing Two Categorical Variables

Joint frequency

| Movies by Genre and Rating | | | | | |
|----------------------------|-----|------|-------|------|-------|
| Genre | G | PG | PG-13 | R | Total |
| Action | 2 | 70 | 311 | 229 | 612 |
| Adventure | 44 | 179 | 209 | 64 | 496 |
| Animation | 43 | 111 | 8 | 6 | 168 |
| Comedy | 45 | 258 | 472 | 506 | 1218 |
| Drama | 12 | 136 | 586 | 836 | 1570 |
| Family | 38 | 181 | 10 | 1 | 230 |
| ... | ... | ... | ... | ... | ... |
| Total | 230 | 1207 | 2686 | 3058 | 7181 |

Analyzing Two Categorical Variables

Joint frequency
Contingency table

| Movies by Genre and Rating | | | | | |
|----------------------------|-----|------|-------|------|-------|
| Genre | G | PG | PG-13 | R | Total |
| Action | 2 | 70 | 311 | 229 | 612 |
| Adventure | 44 | 179 | 209 | 64 | 496 |
| Animation | 43 | 111 | 8 | 6 | 168 |
| Comedy | 45 | 258 | 472 | 506 | 1218 |
| Drama | 12 | 136 | 586 | 836 | 1570 |
| Family | 38 | 181 | 10 | 1 | 230 |
| ... | ... | ... | ... | ... | ... |
| Total | 230 | 1207 | 2686 | 3058 | 7181 |

Analyzing Two Categorical Variables

Joint frequency
Contingency table
Marginal frequency

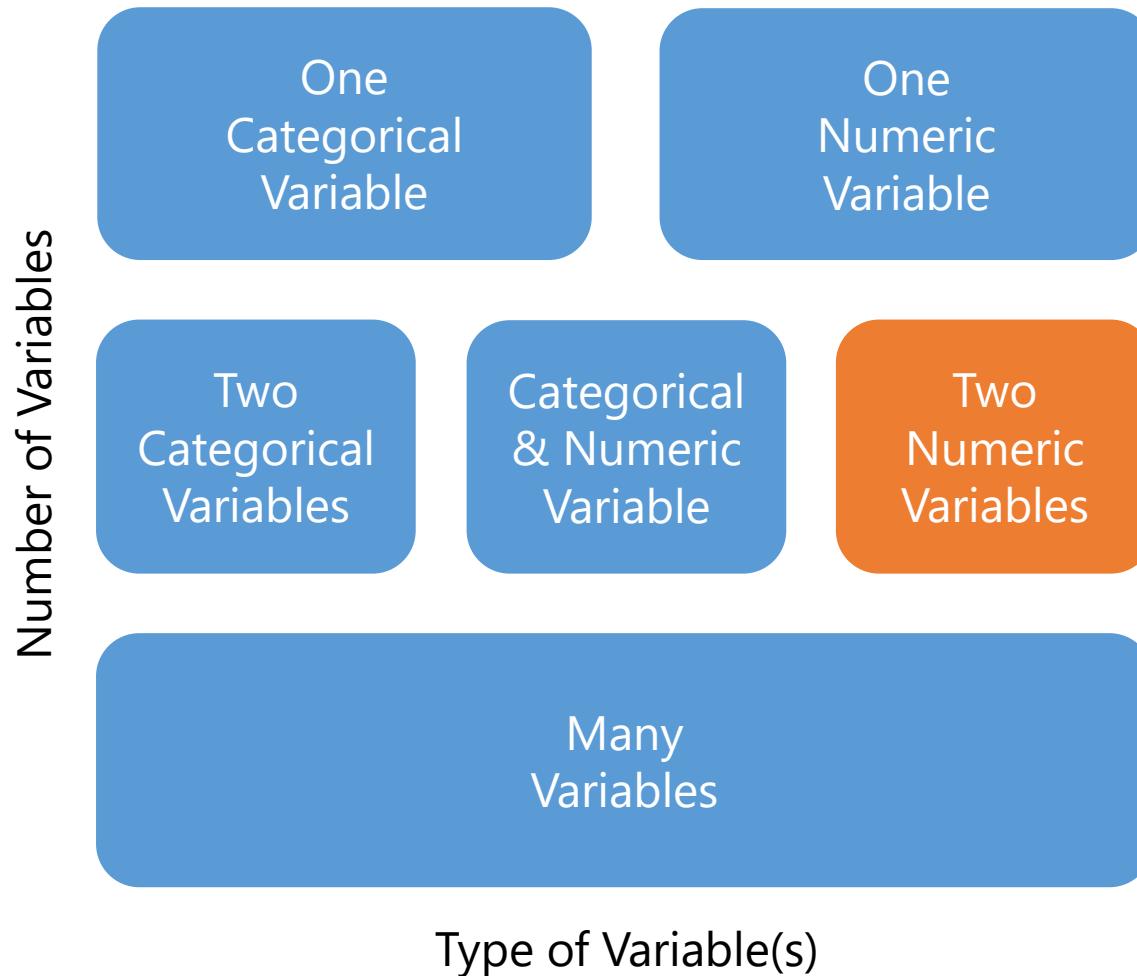
| Movies by Genre and Rating | | | | | |
|----------------------------|-----|------|-------|------|-------|
| Genre | G | PG | PG-13 | R | Total |
| Action | 2 | 70 | 311 | 229 | 612 |
| Adventure | 44 | 179 | 209 | 64 | 496 |
| Animation | 43 | 111 | 8 | 6 | 168 |
| Comedy | 45 | 258 | 472 | 506 | 1218 |
| Drama | 12 | 136 | 586 | 836 | 1570 |
| Family | 38 | 181 | 10 | 1 | 230 |
| ... | ... | ... | ... | ... | ... |
| Total | 230 | 1207 | 2686 | 3058 | 7181 |

Analyzing Two Categorical Variables

Joint frequency
Contingency table
Marginal frequency
Relative frequency

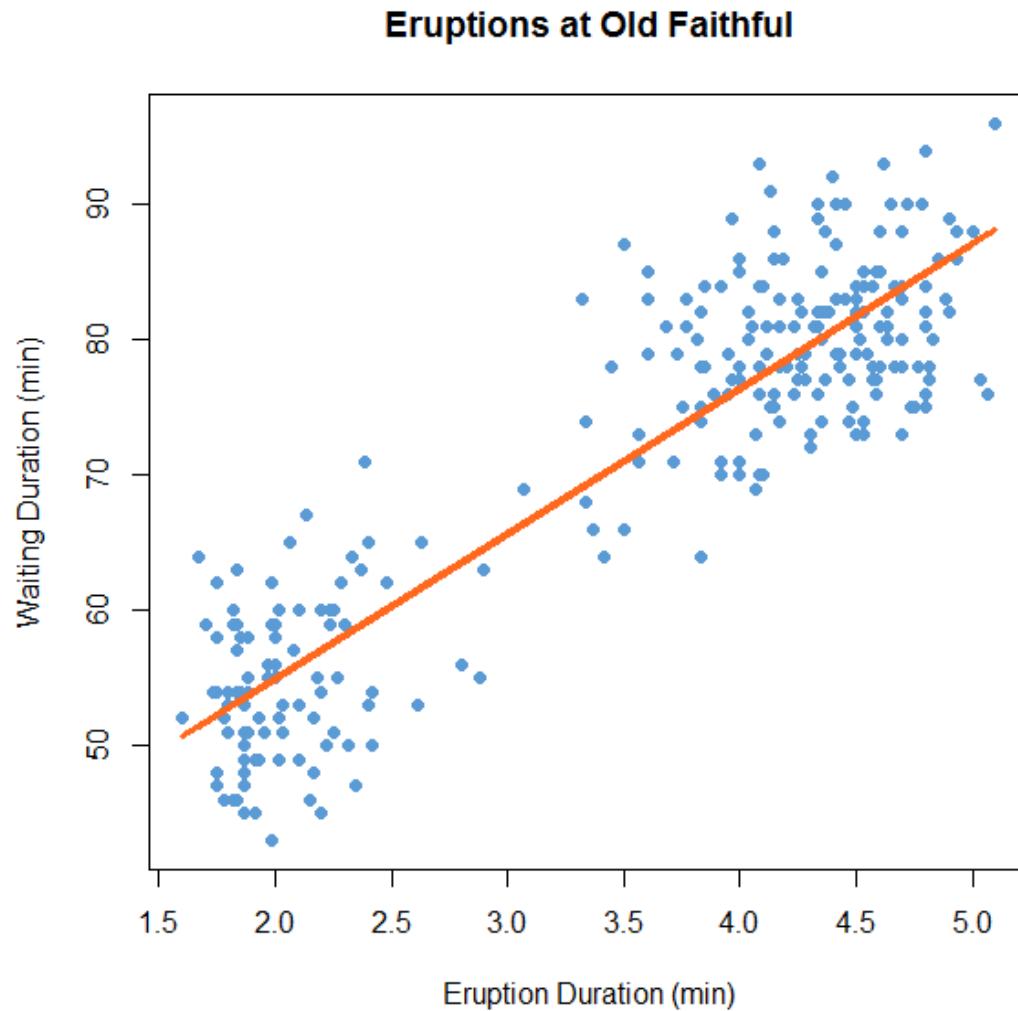
| Movies by Genre and Rating | | | | | |
|----------------------------|-------|-------|-------|-------|-------|
| Genre | G | PG | PG-13 | R | Total |
| Action | 0.001 | 0.010 | 0.043 | 0.032 | 0.086 |
| Adventure | 0.006 | 0.025 | 0.029 | 0.009 | 0.069 |
| Animation | 0.006 | 0.015 | 0.001 | 0.001 | 0.023 |
| Comedy | 0.006 | 0.036 | 0.066 | 0.070 | 0.170 |
| Drama | 0.002 | 0.019 | 0.082 | 0.116 | 0.219 |
| Family | 0.005 | 0.025 | 0.001 | 0.001 | 0.033 |
| ... | ... | ... | ... | ... | ... |
| Total | 0.032 | 0.168 | 0.374 | 0.426 | 1.000 |

Analyzing Two Numeric Variables

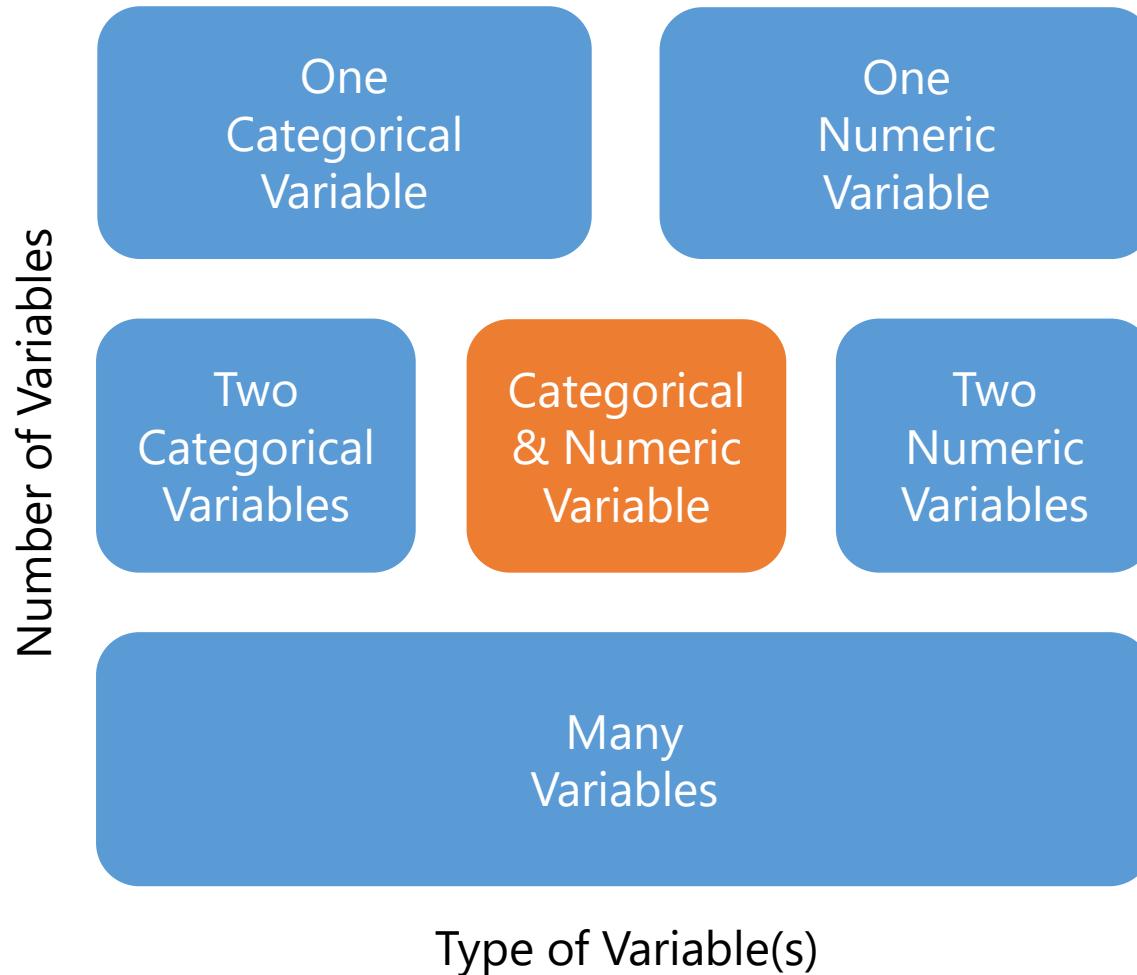


Analyzing Two Numeric Variables

Explanatory vs. outcome
Covariance
Correlation

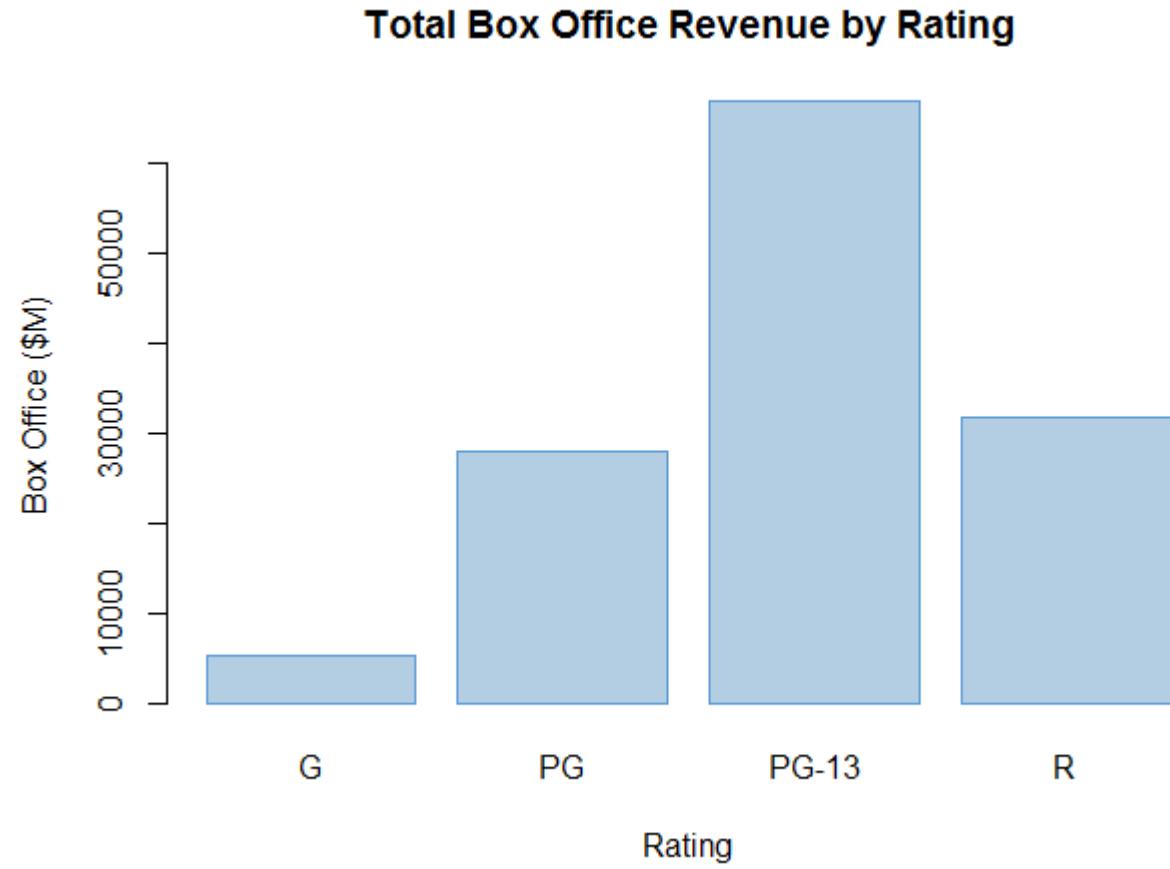


Analyzing a Numeric Variable Grouped by a Categorical Variable

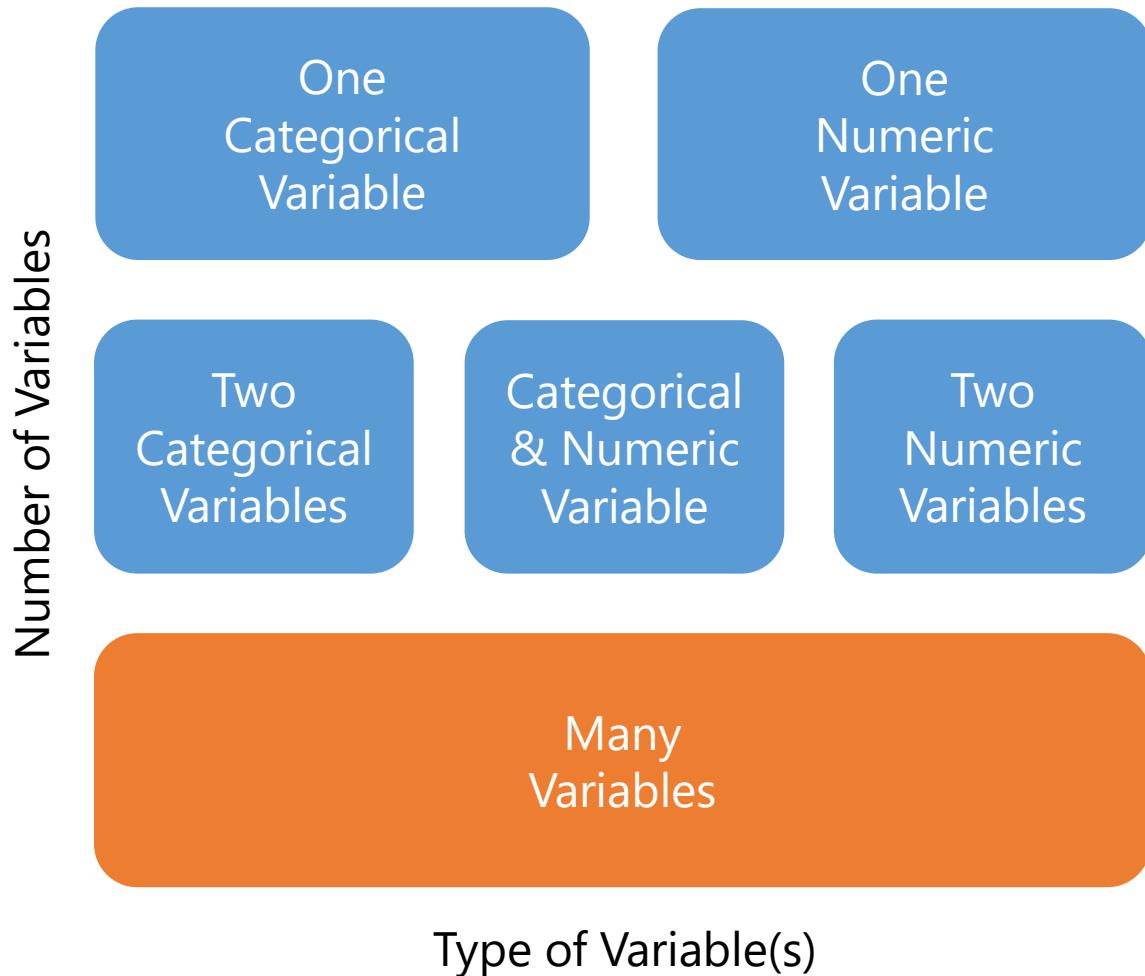


Analyzing a Numeric Variable Grouped by a Categorical Variable

One categorical variable
One numeric variable
Aggregate measures



Analyzing Many Variables







COWBOYS & Space Invaders: The Musical



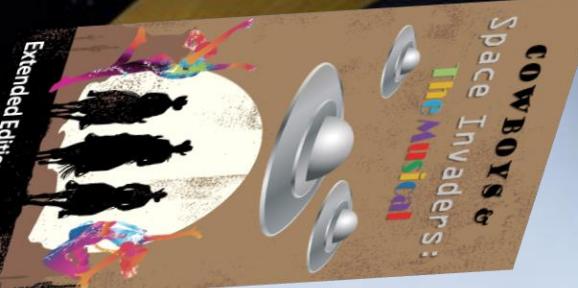
Extended Edition



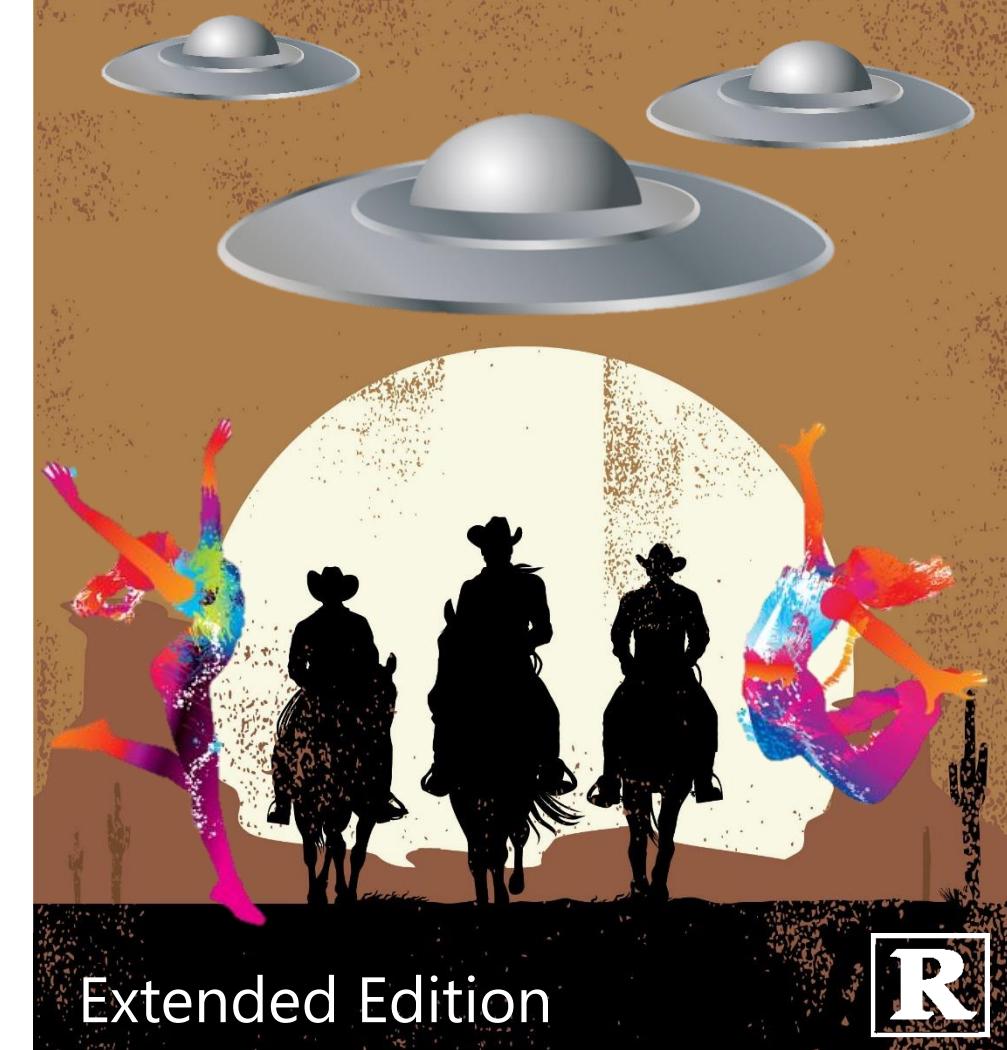
Code Demo

Lab 3

Descriptive Statistics



COWBOYS & Space Invaders: The Musical

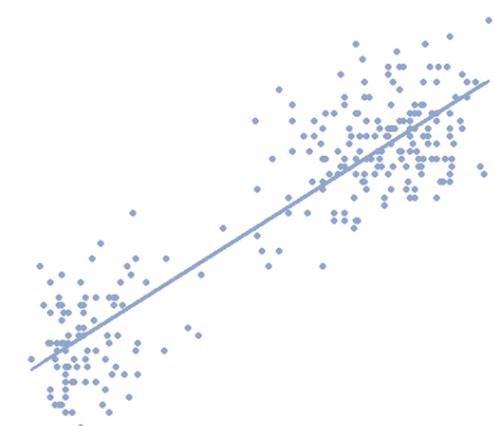
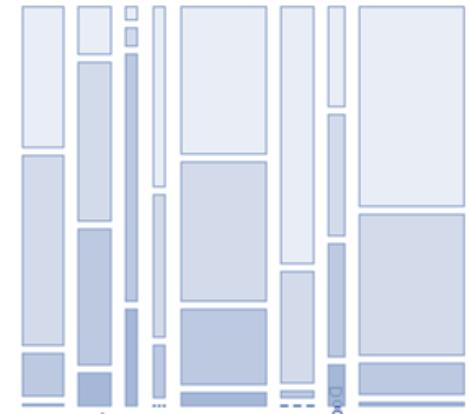
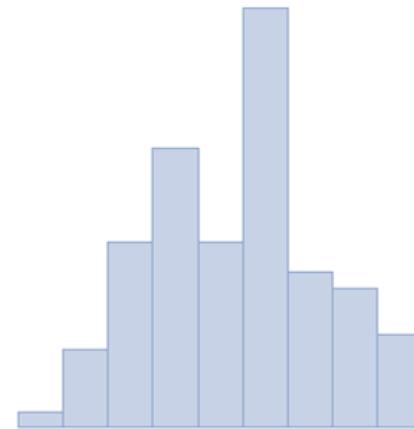




Data Visualization

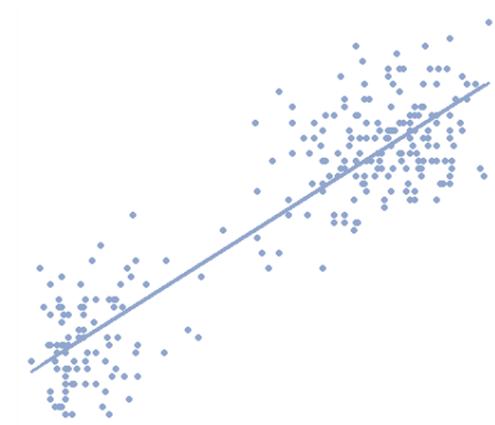
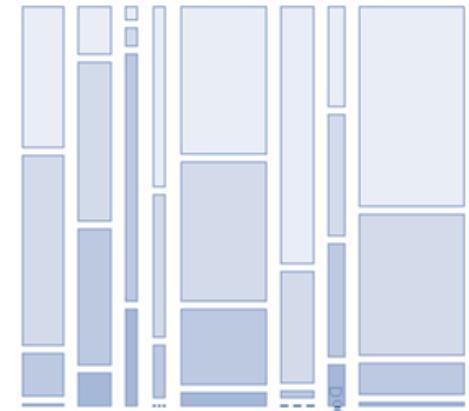
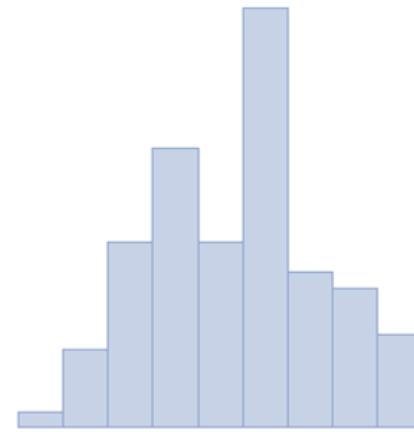
Data Visualization

Visual data representation



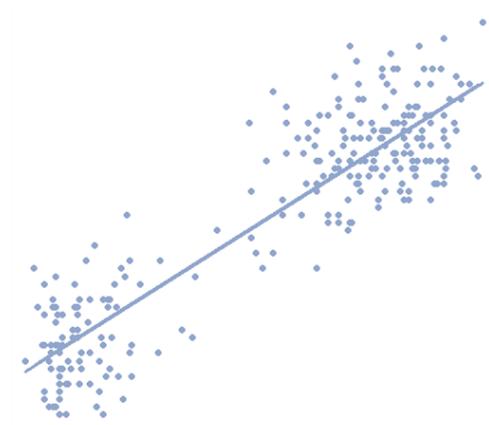
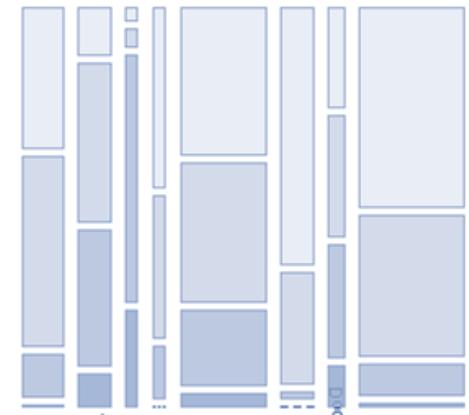
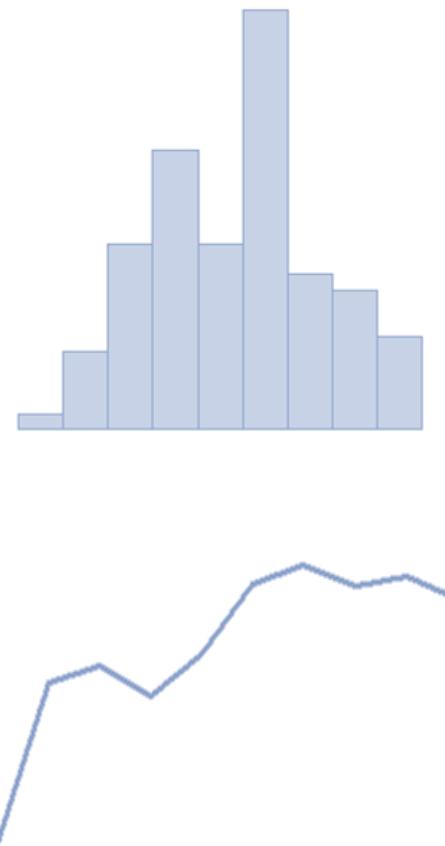
Data Visualization

Visual data representation
Human pattern recognition



Data Visualization

Visual data representation
Human pattern recognition
Map dimensions to visual

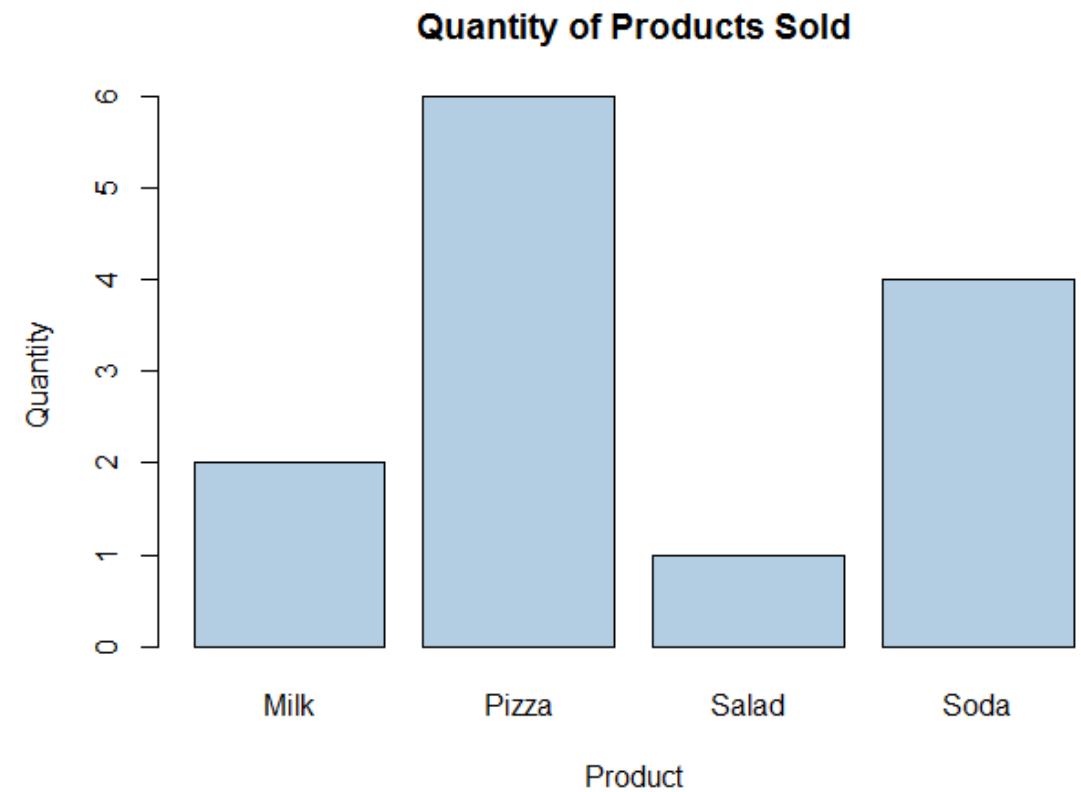


Data Visualization

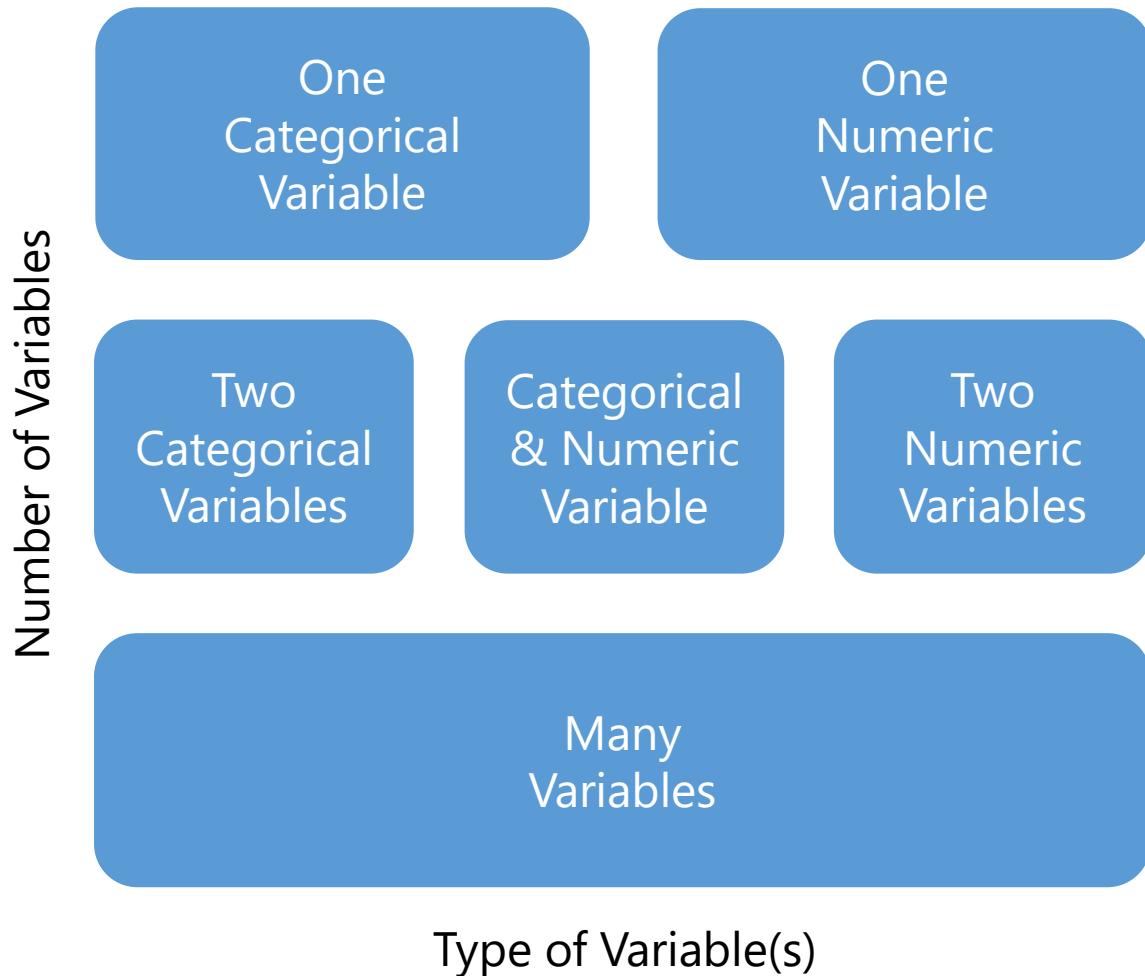
| ID | Date | Customer | Product | Quantity |
|----|------------|----------|---------|----------|
| 1 | 2015-08-27 | John | Pizza | 2 |
| 2 | 2015-08-27 | John | Soda | 2 |
| 3 | 2015-08-27 | Jill | Salad | 1 |
| 4 | 2015-08-27 | Jill | Milk | 1 |
| 5 | 2015-08-28 | Miko | Pizza | 3 |
| 6 | 2015-08-28 | Miko | Soda | 2 |
| 7 | 2015-08-28 | Sam | Pizza | 1 |
| 8 | 2015-08-28 | Sam | Milk | 1 |

Data Visualization

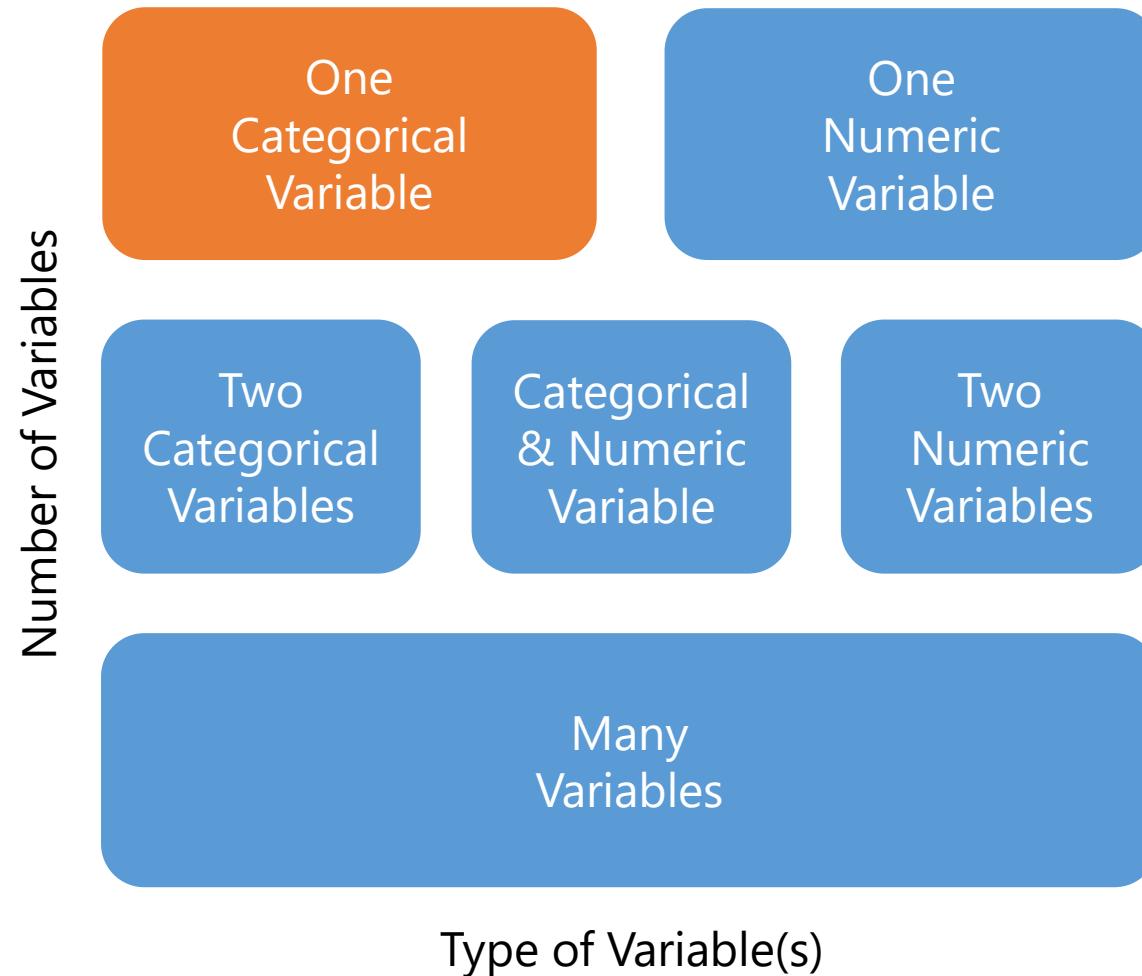
| ID | Date | Customer | Product | Quantity |
|----|------------|----------|---------|----------|
| 1 | 2015-08-27 | John | Pizza | 2 |
| 2 | 2015-08-27 | John | Soda | 2 |
| 3 | 2015-08-27 | Jill | Salad | 1 |
| 4 | 2015-08-27 | Jill | Milk | 1 |
| 5 | 2015-08-28 | Miko | Pizza | 3 |
| 6 | 2015-08-28 | Miko | Soda | 2 |
| 7 | 2015-08-28 | Sam | Pizza | 1 |
| 8 | 2015-08-28 | Sam | Milk | 1 |



Types of Analysis

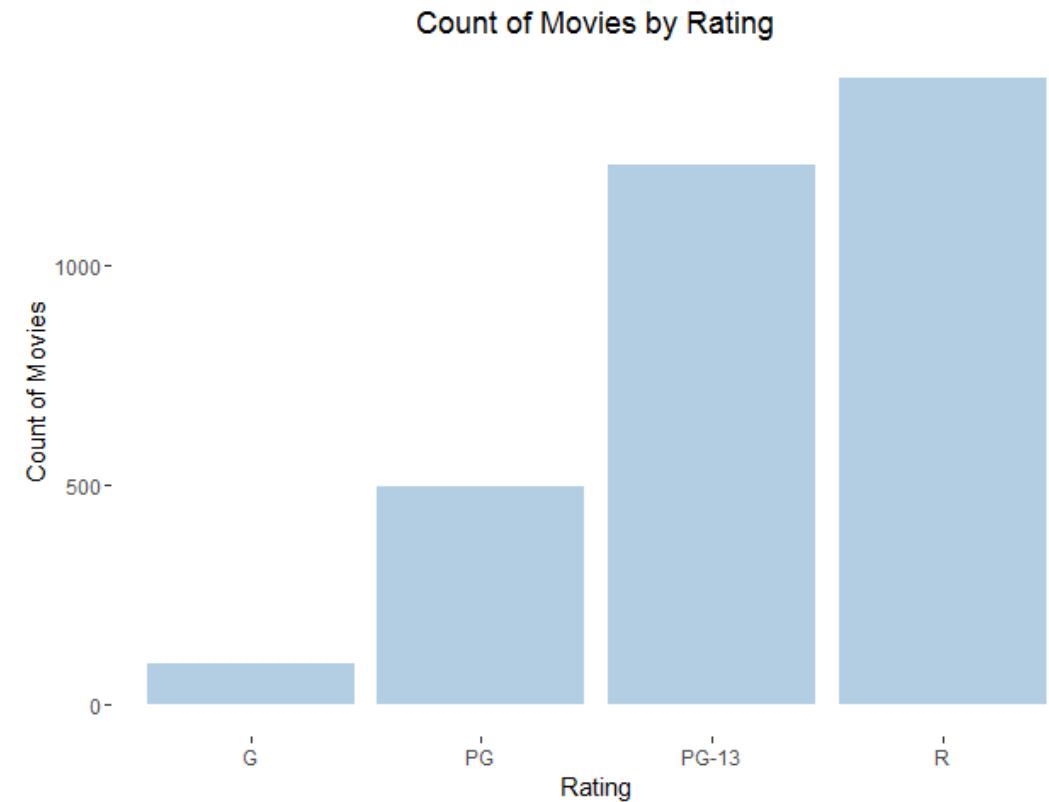


Visualizing One Categorical Variable

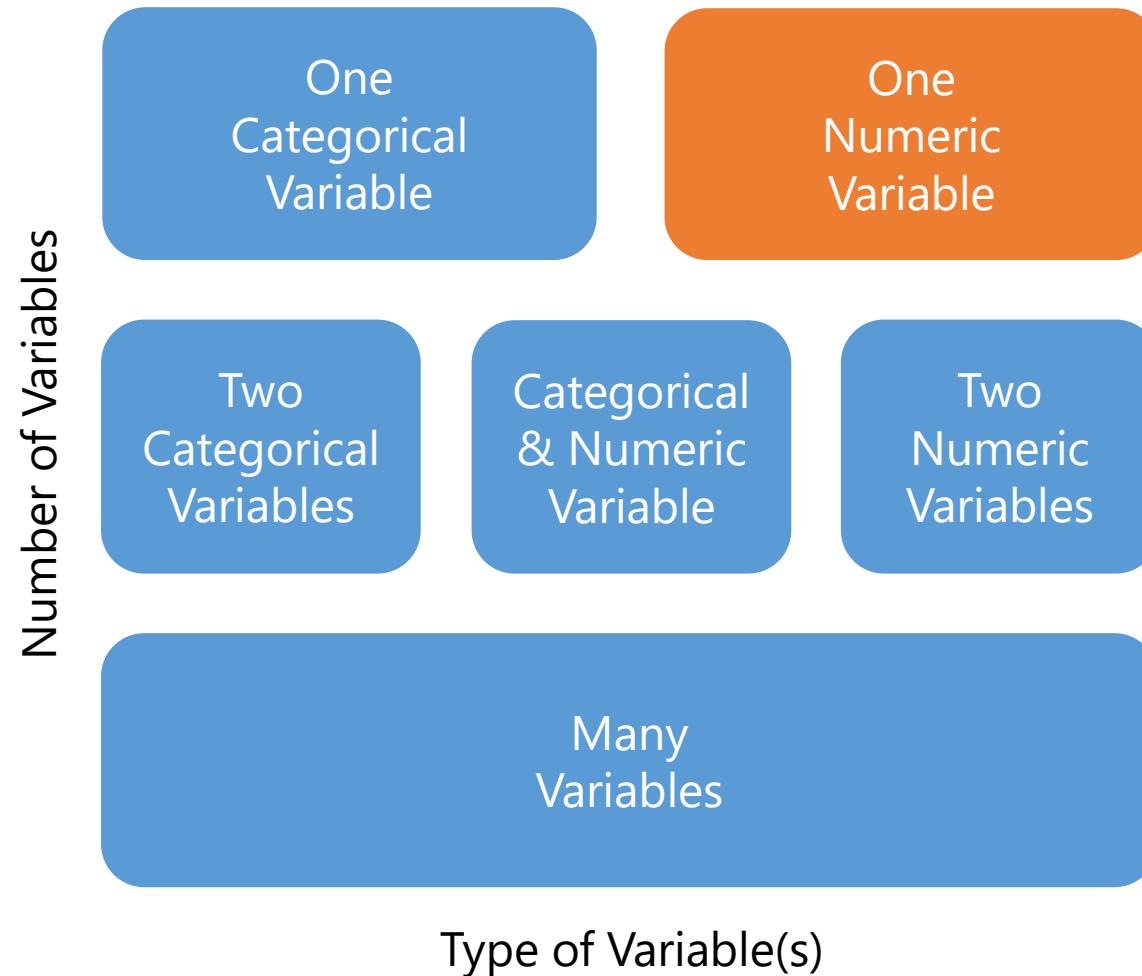


Visualizing One Categorical Variable

Frequency
Proportion

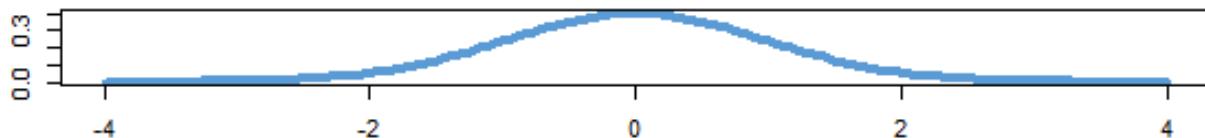
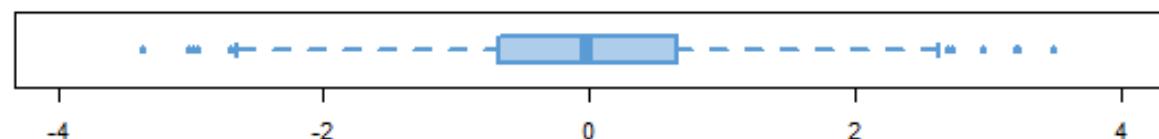


Visualizing One Numeric Variable

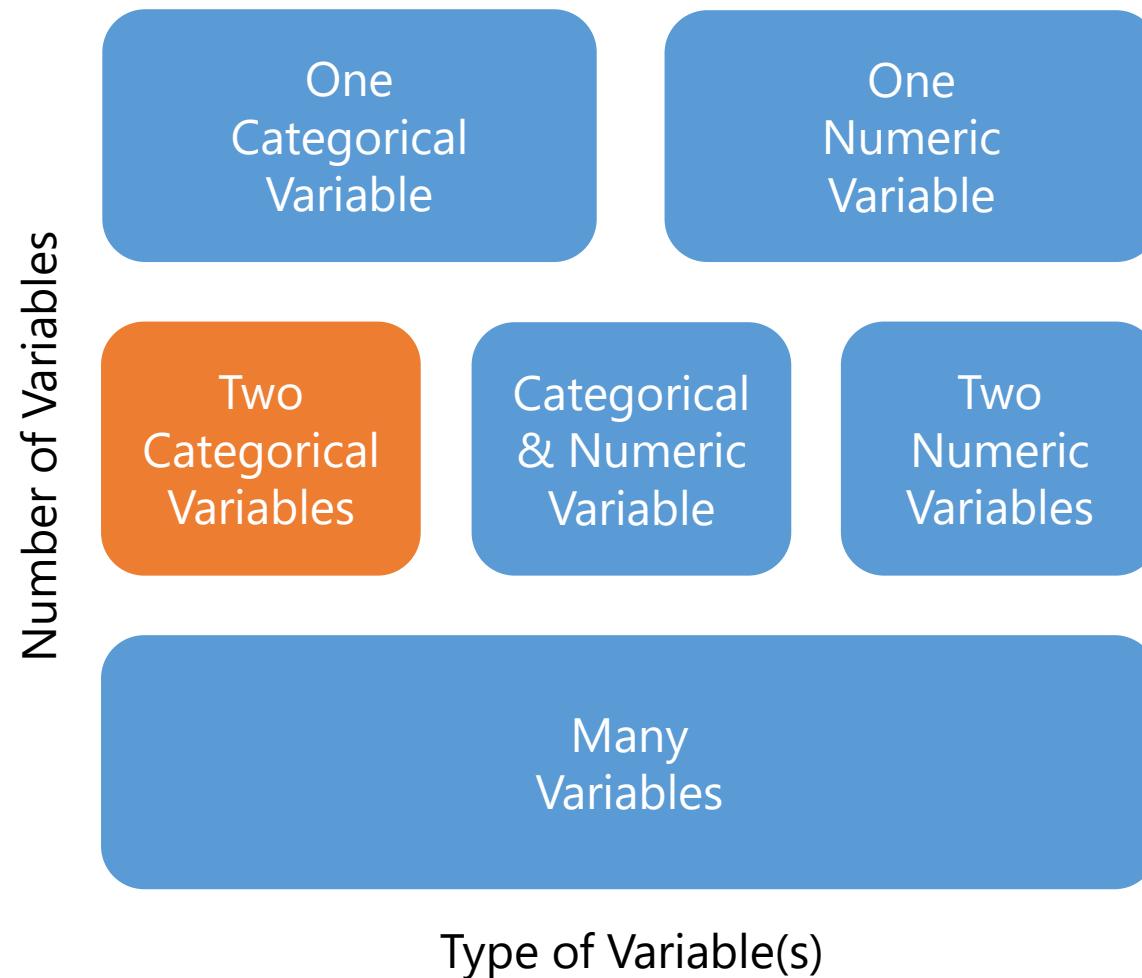


Visualizing One Numeric Variable

Location
Spread
Shape

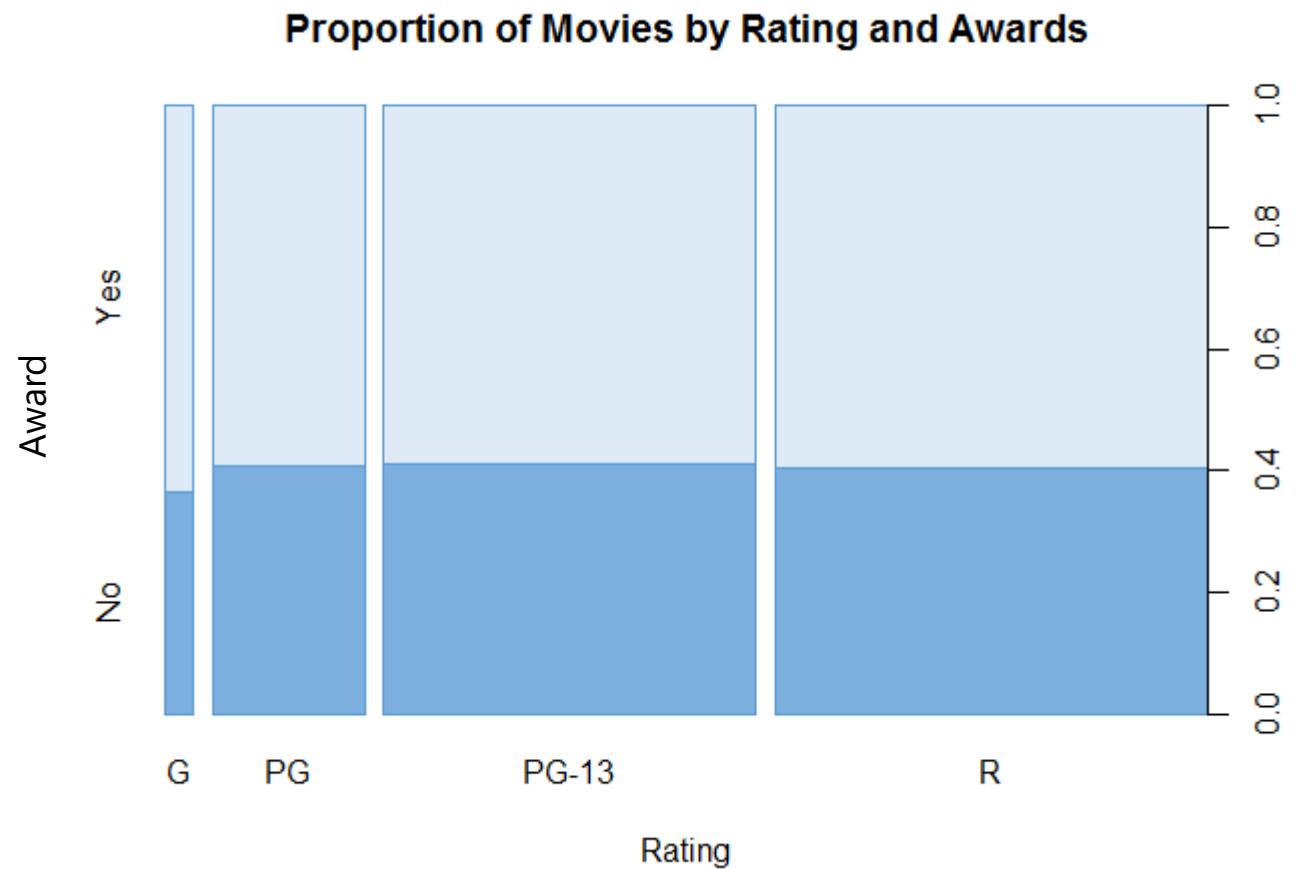


Visualizing Two Categorical Variables

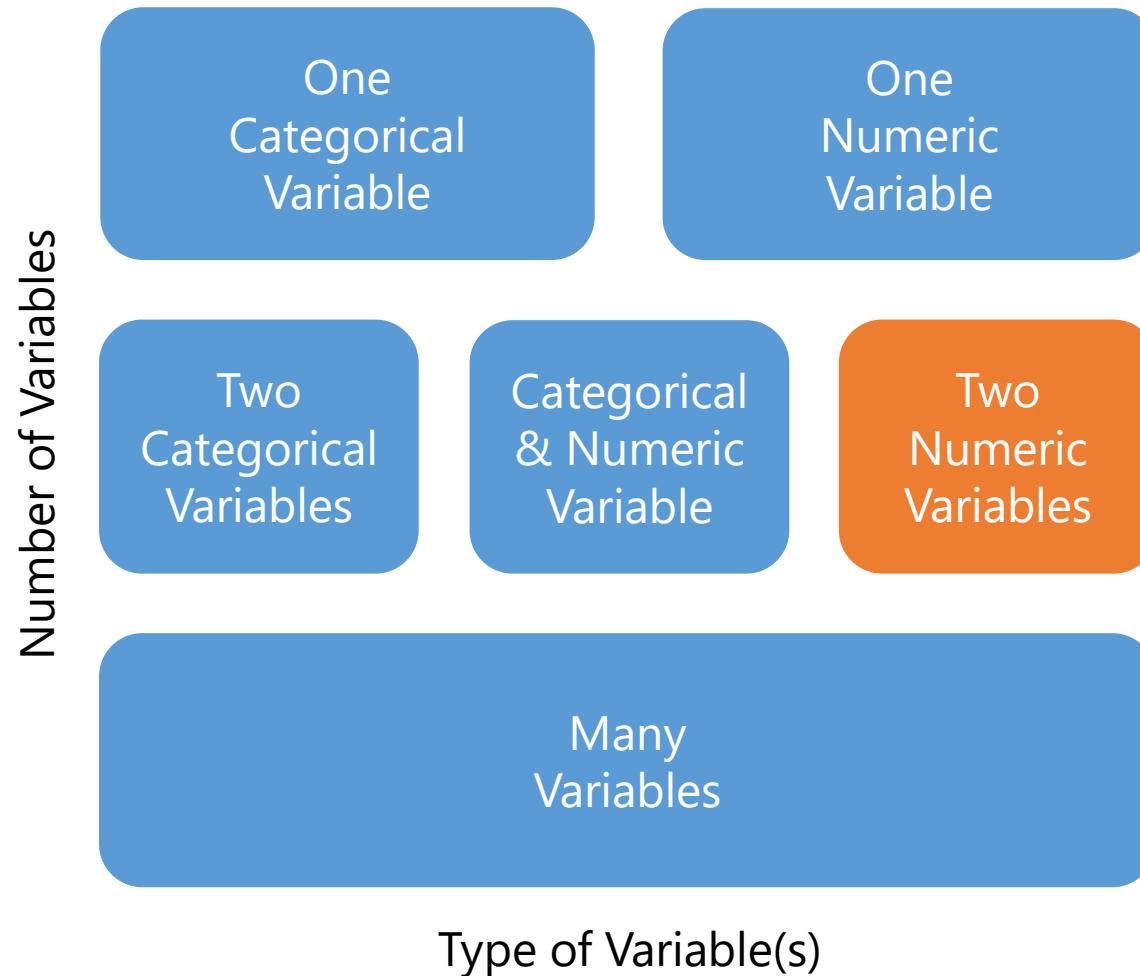


Visualizing Two Categorical Variables

Joint frequency
Marginal frequency
Relative frequency

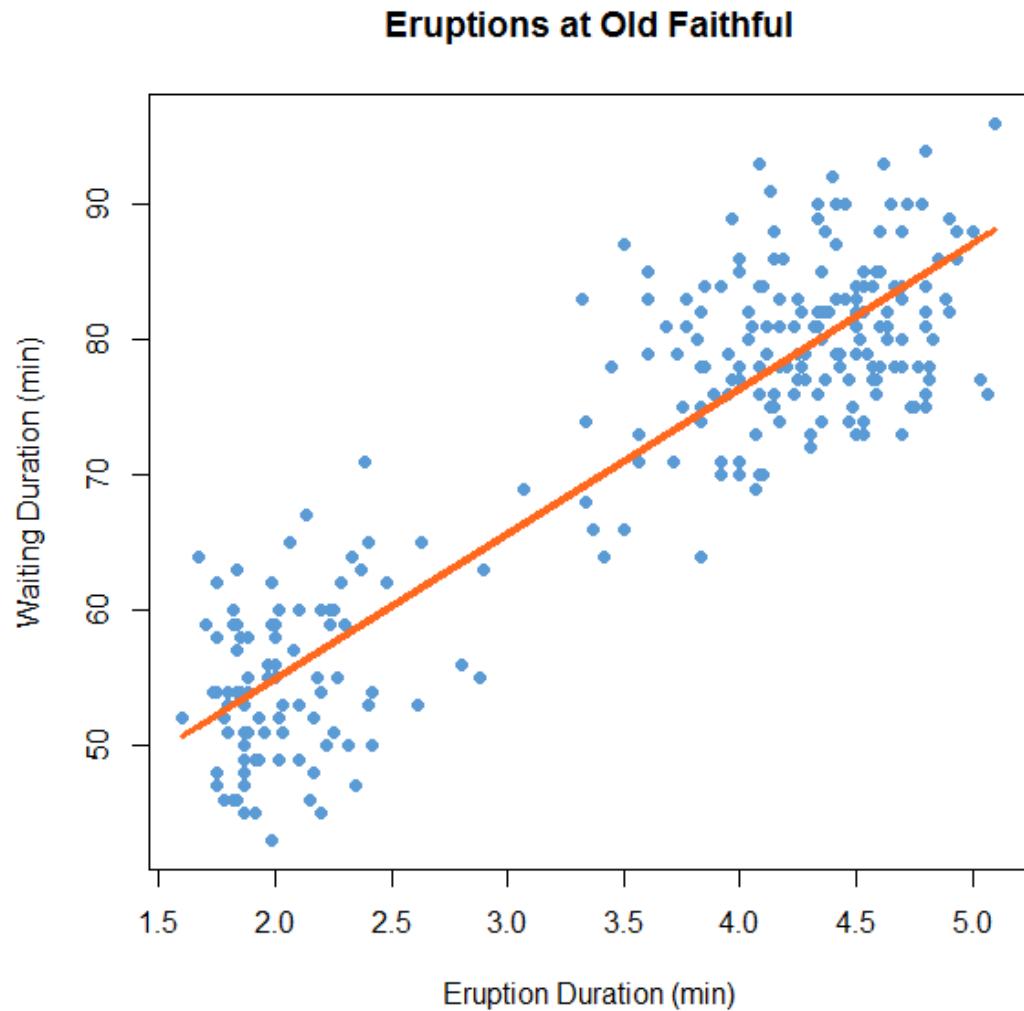


Visualizing Two Numeric Variables



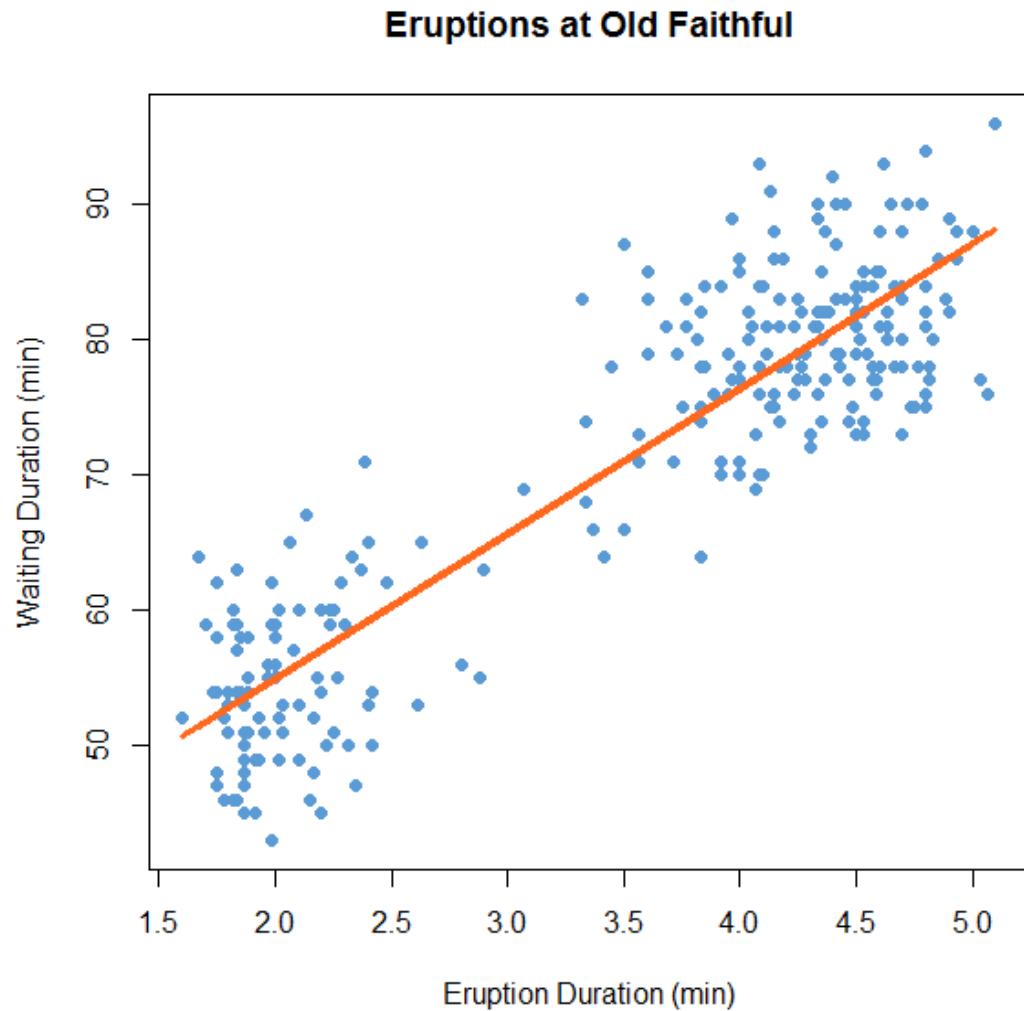
Visualizing Two Numeric Variables

Relationship



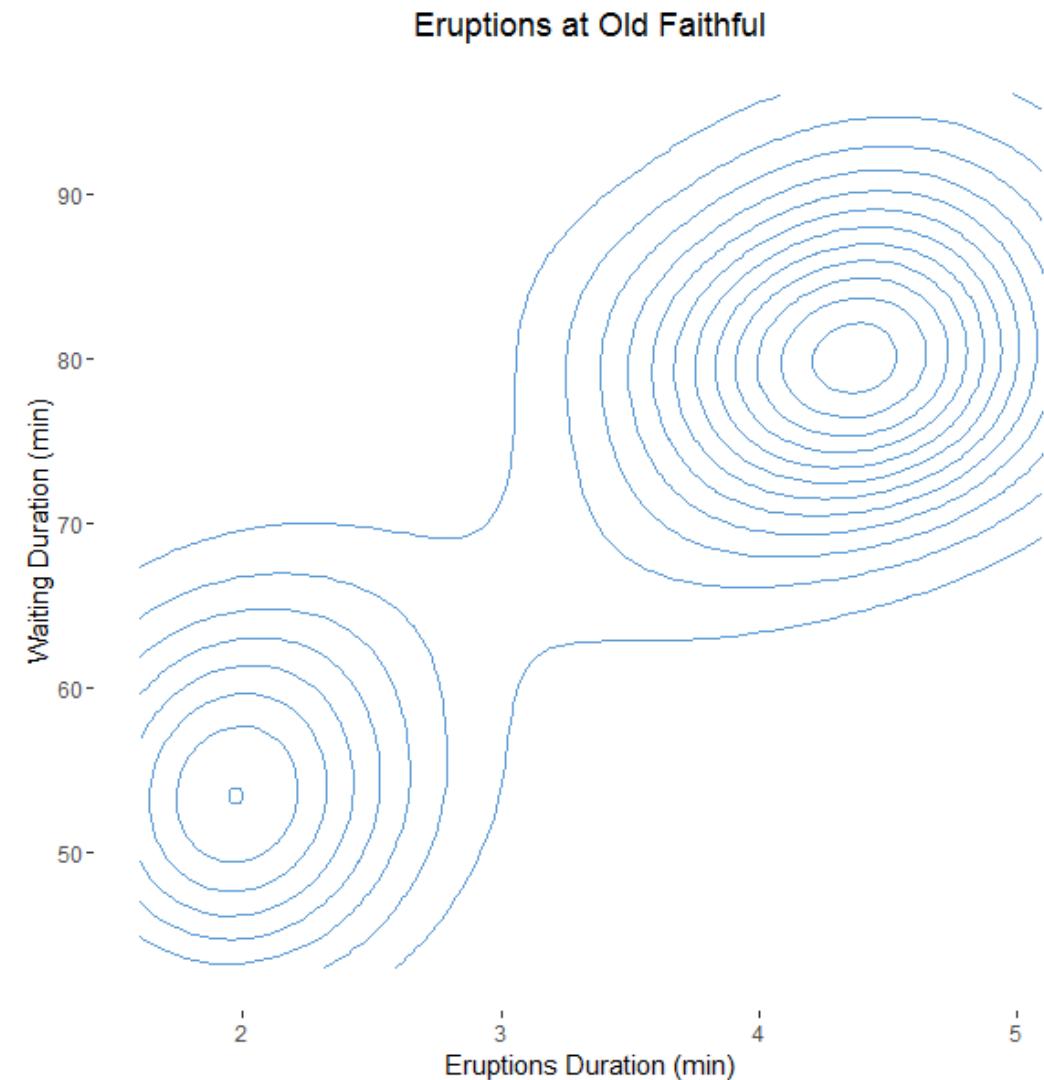
Visualizing Two Numeric Variables

Relationship
Correlation



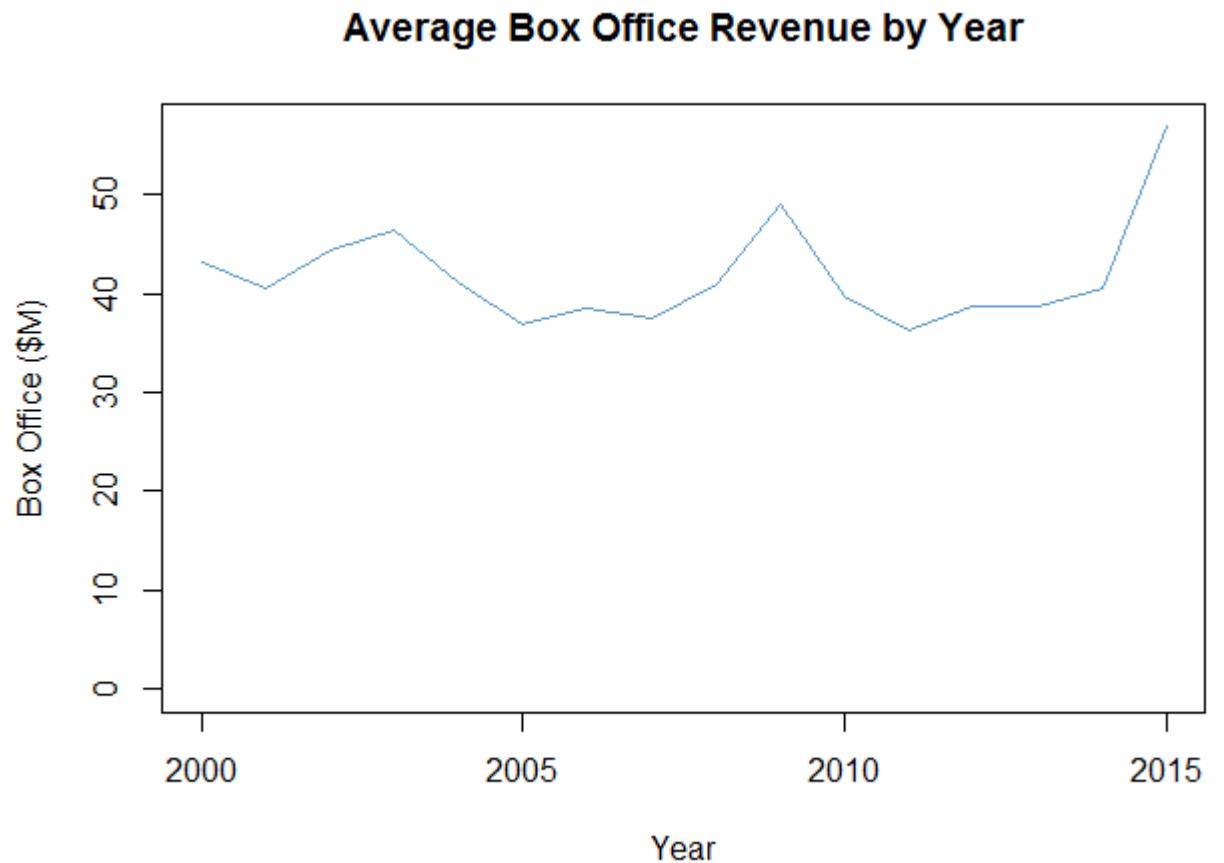
Visualizing Two Numeric Variables

Relationship
Correlation
Distribution



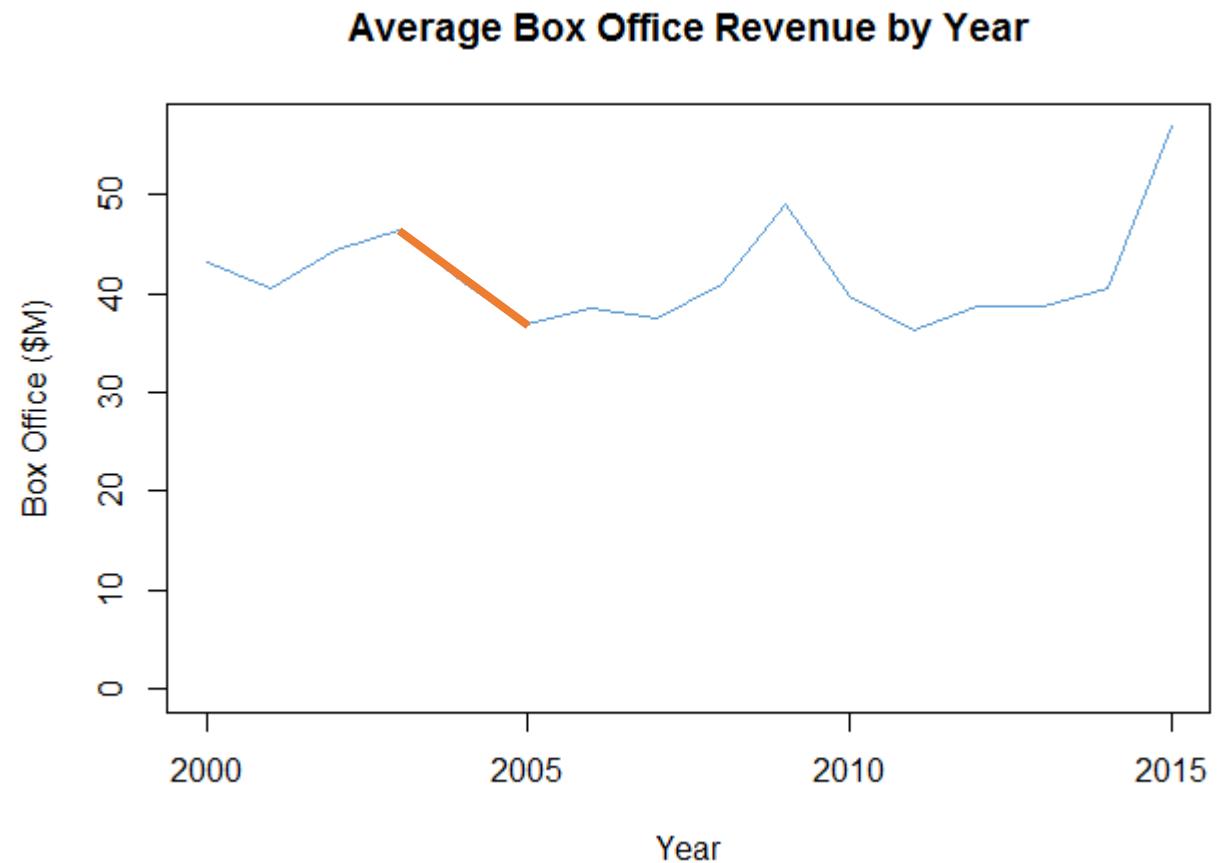
Visualizing Time Series Data

Values over time

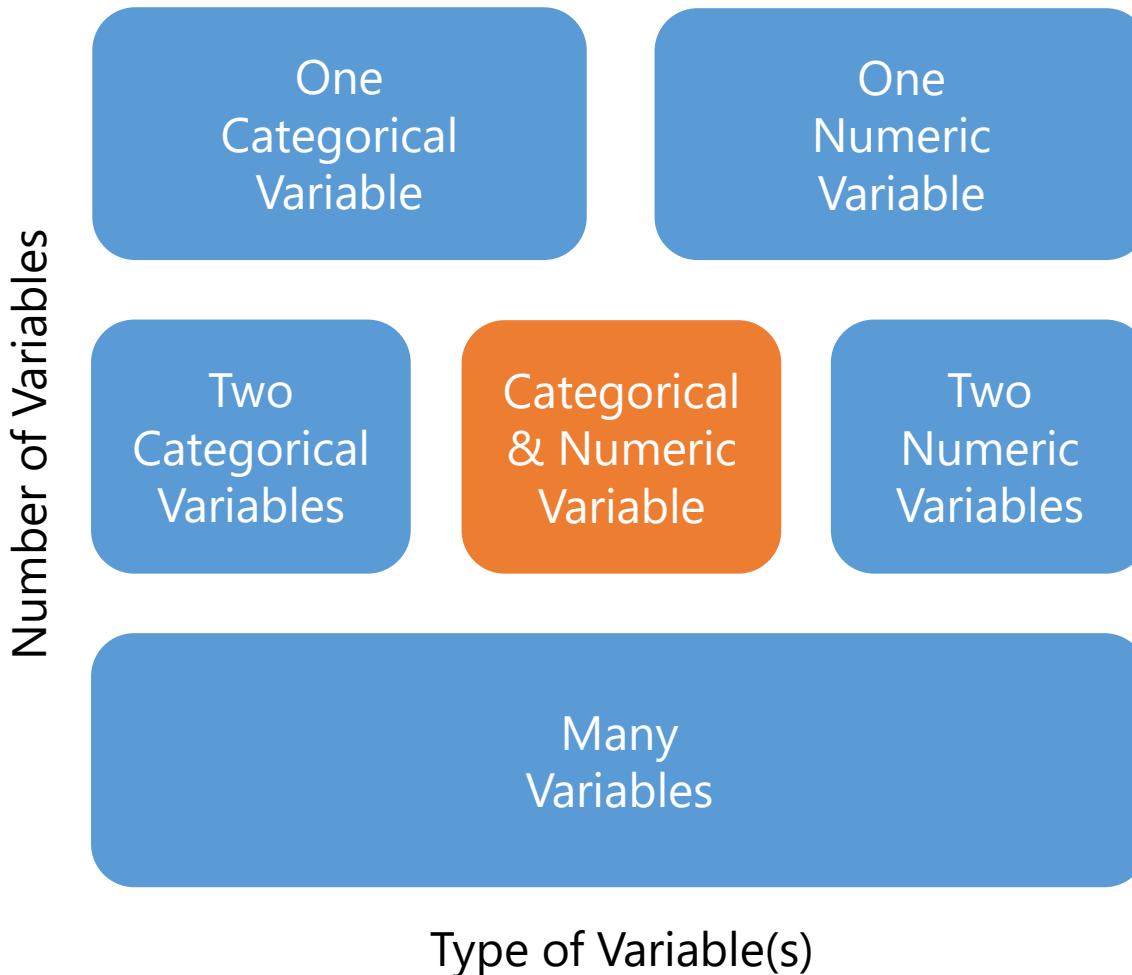


Visualizing Time Series Data

Values over time
Rate of change

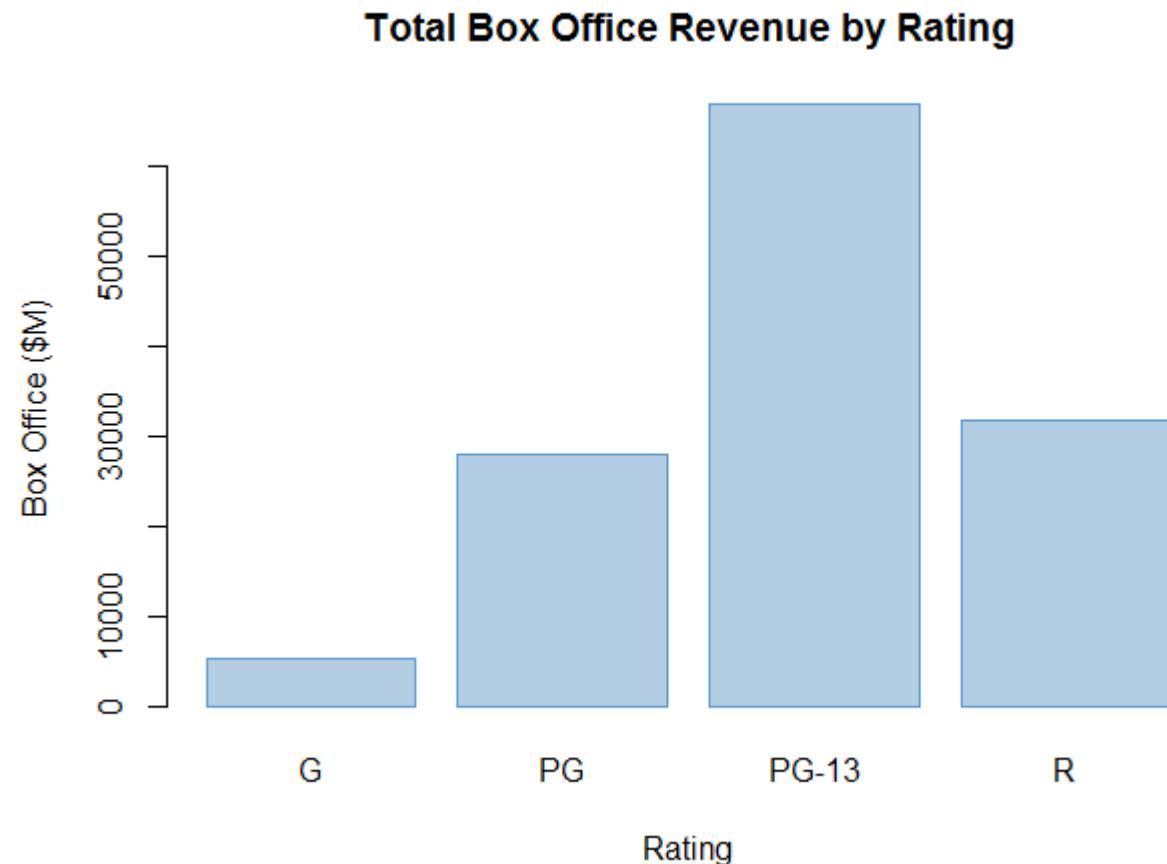


Visualizing a Numeric Variable Grouped by a Categorical Variable

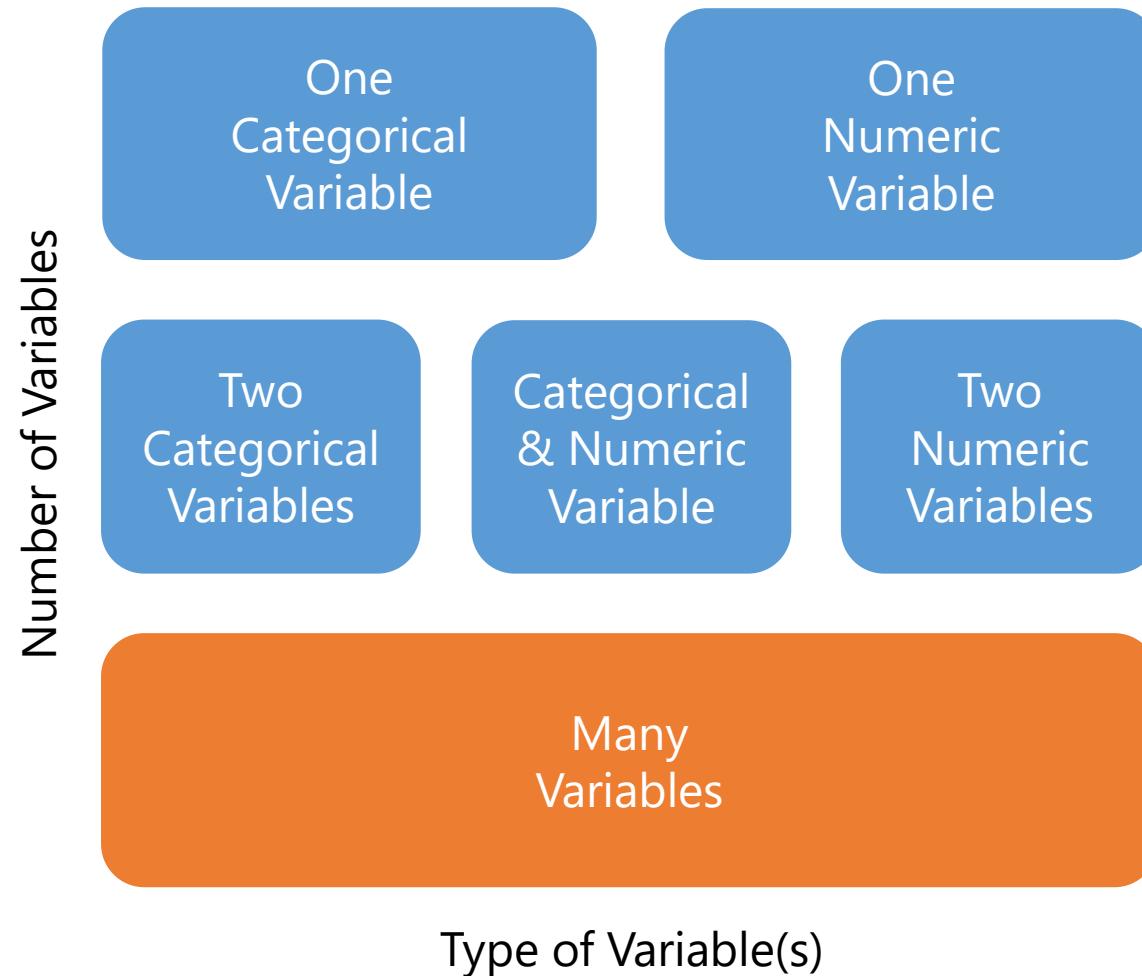


Visualizing a Numeric Variable Grouped by a Categorical Variable

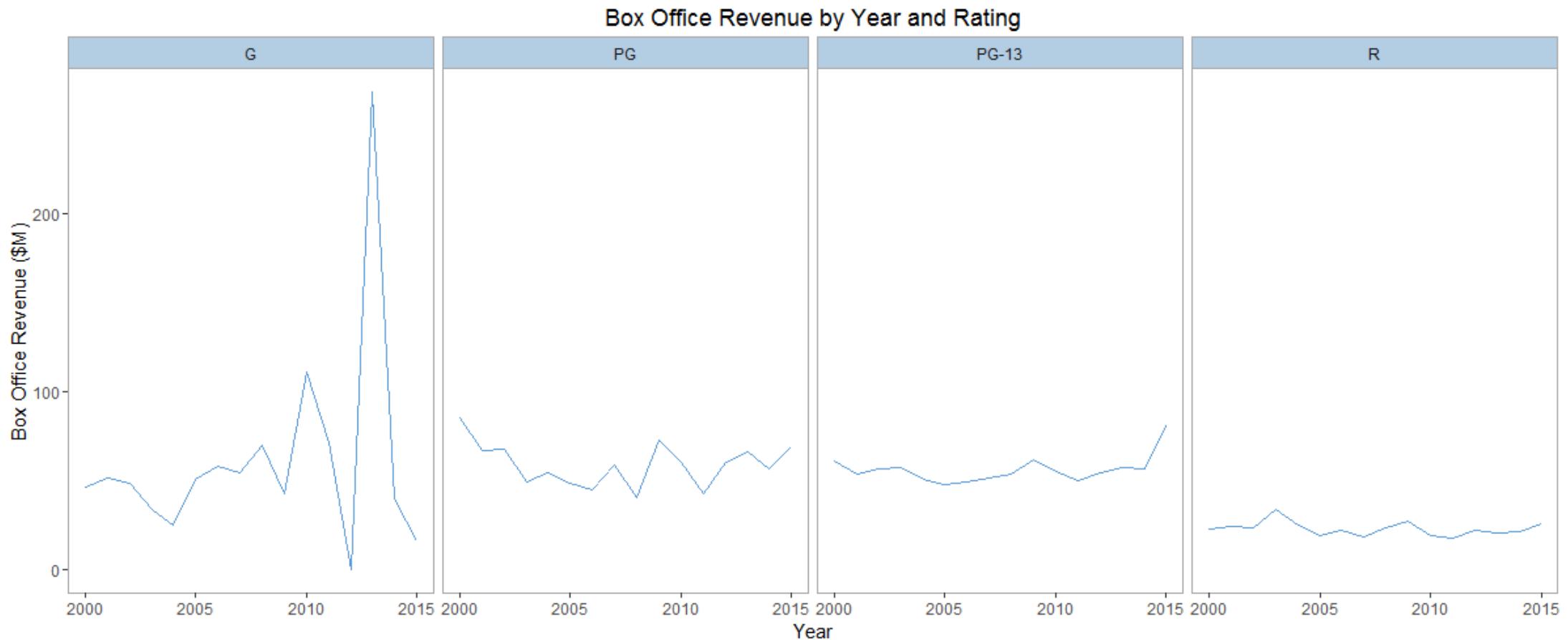
Aggregate
Grouped
Comparison



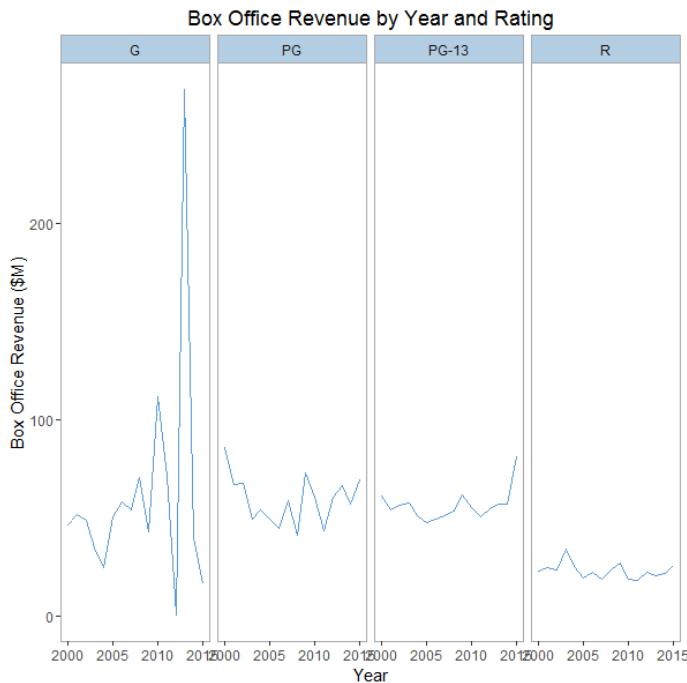
Visualizing Many Variables



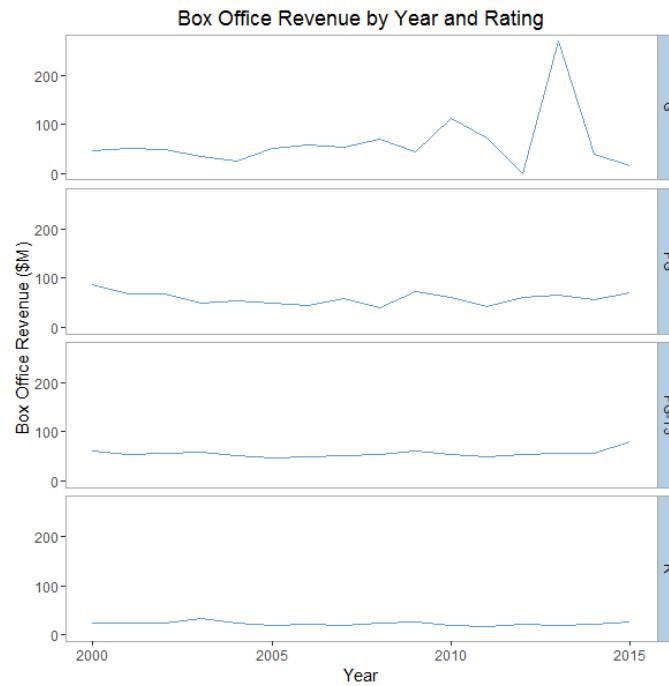
Visualizing Many Variables



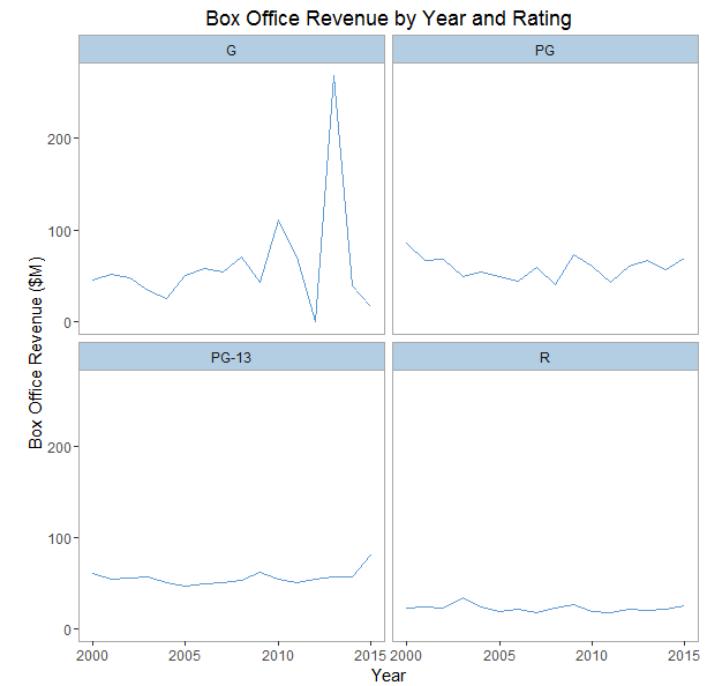
Visualizing Many Variables



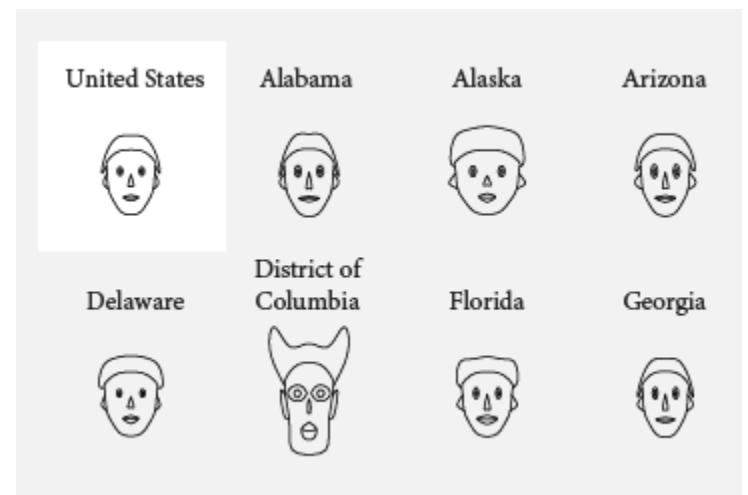
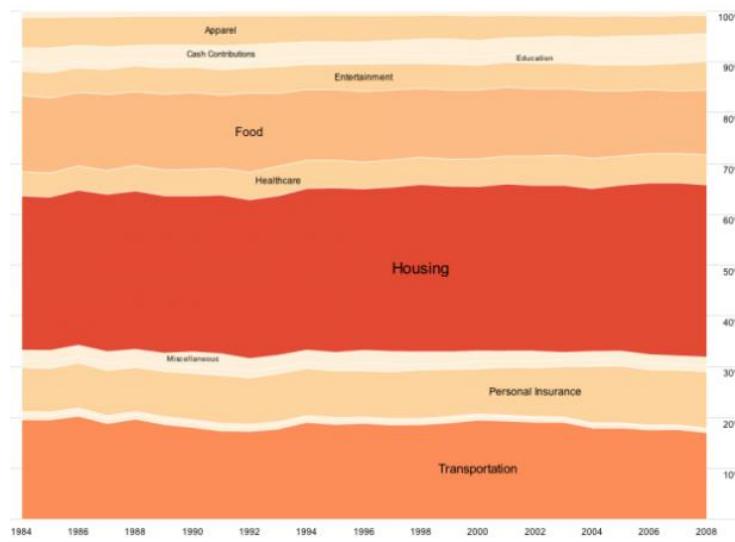
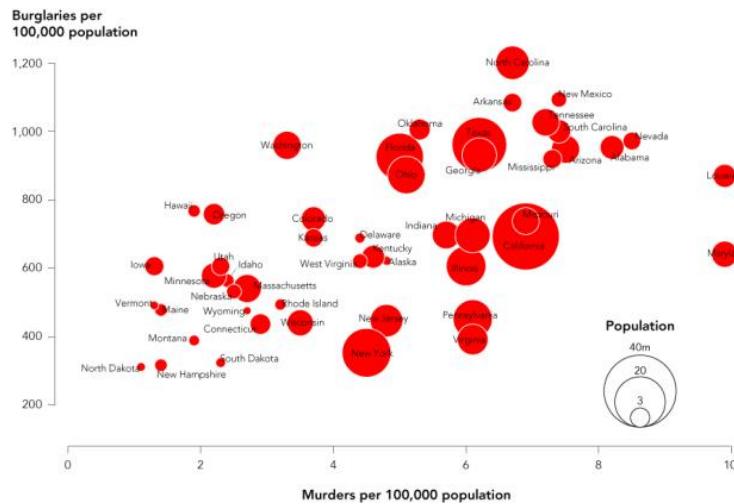
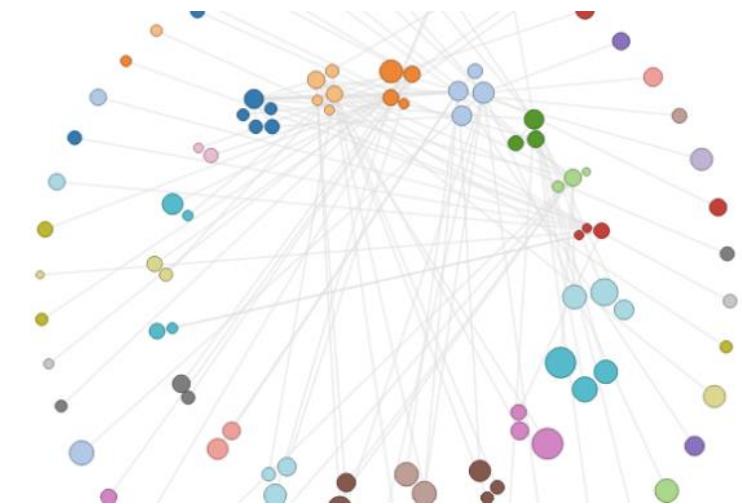
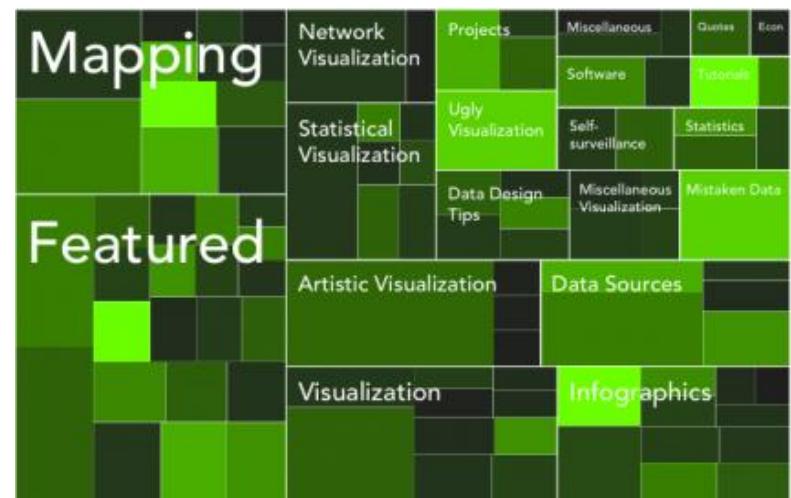
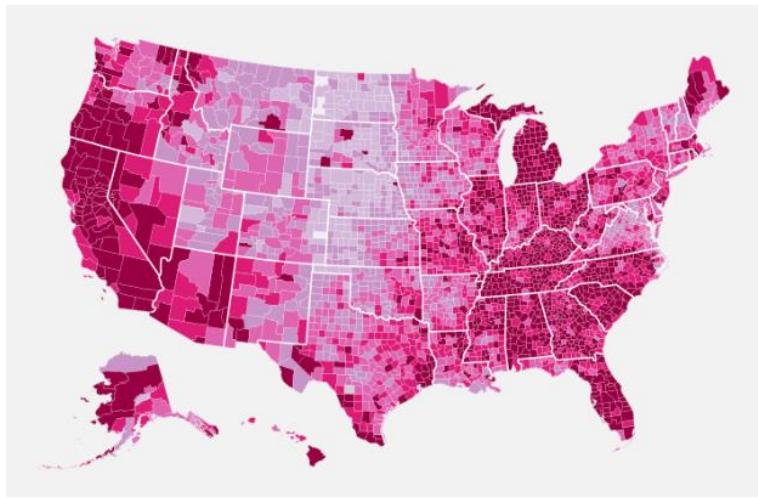
Horizontal



Vertical

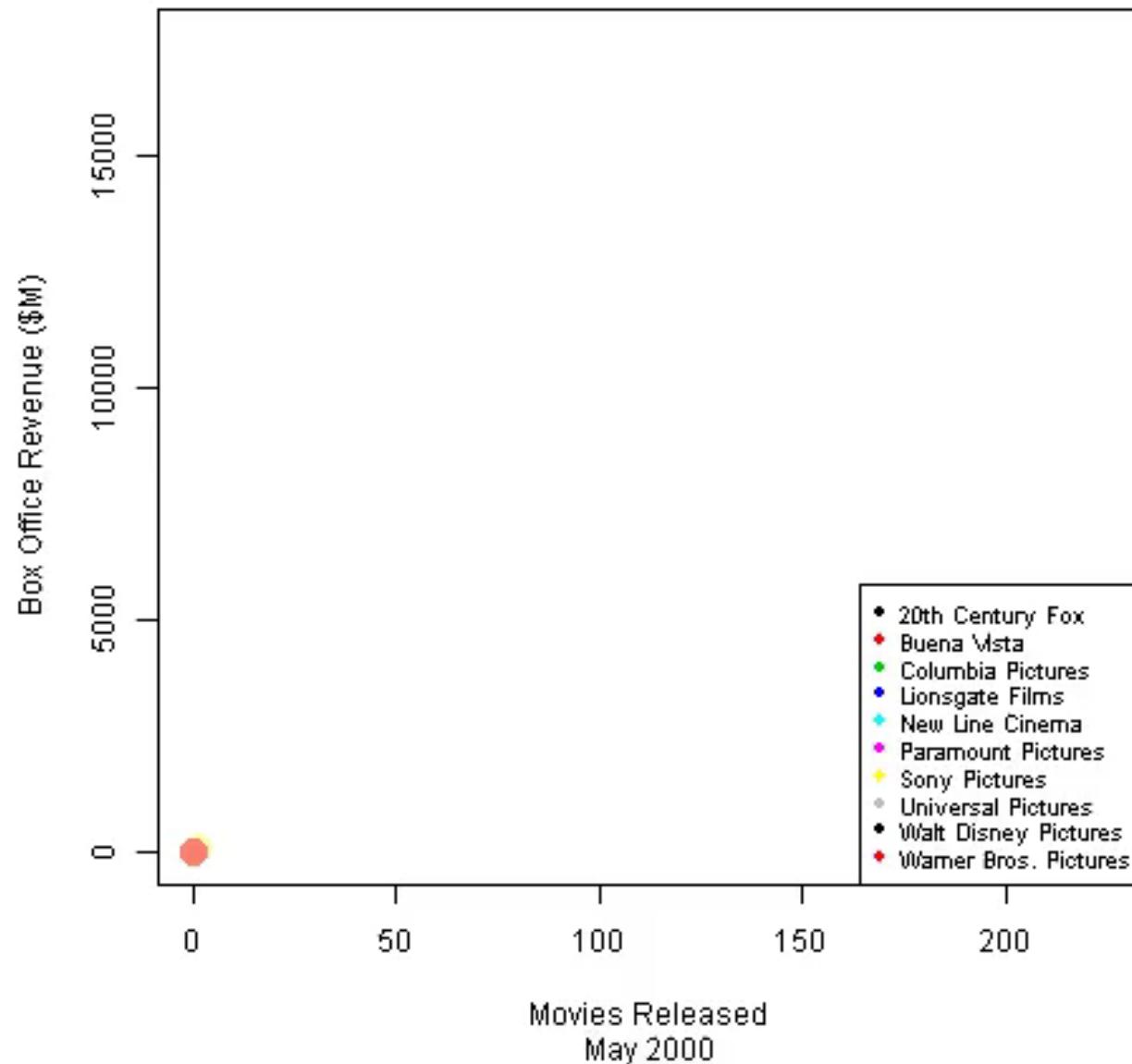


Wrapped



Source: Nathan Yau (www.flowingdata.com)

Top 10 Studios (2000-2015)



Interactive Movie Data

Year

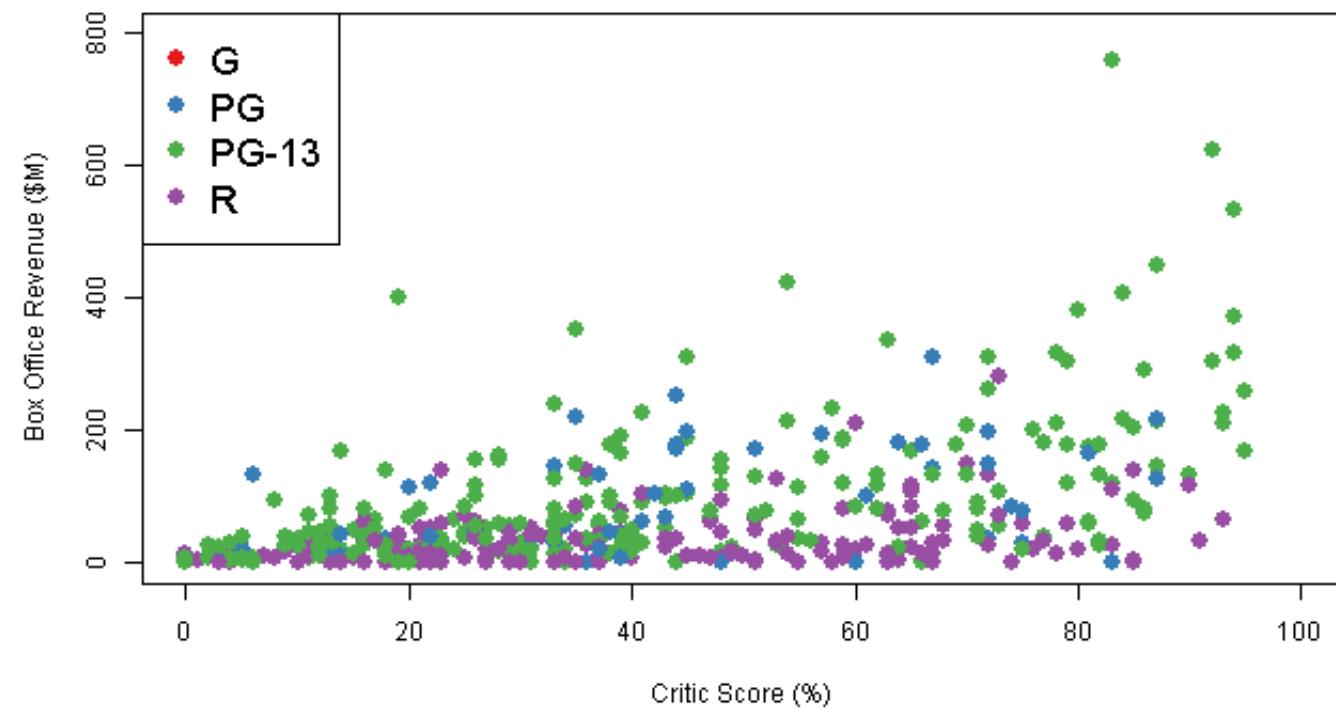
2002 2012 2015

2000 2002 2004 2006 2008 2010 2012 2014 2015

Rating

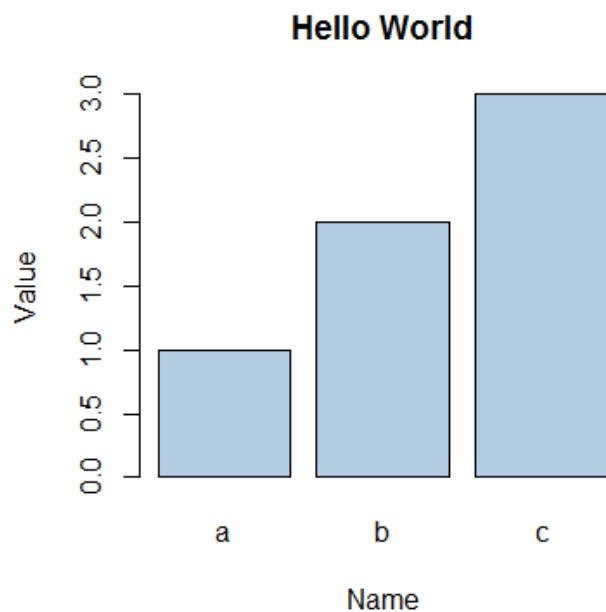
- G
- PG
- PG-13
- R

Genre



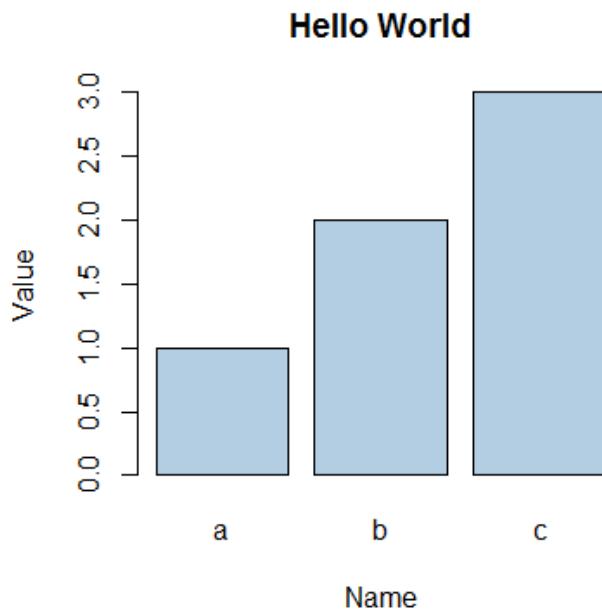
Plotting Systems in R

Plotting Systems in R

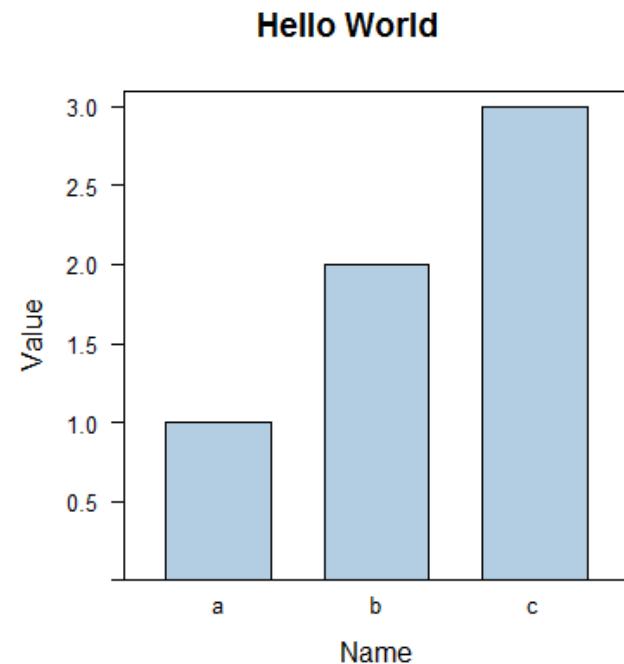


Base

Plotting Systems in R

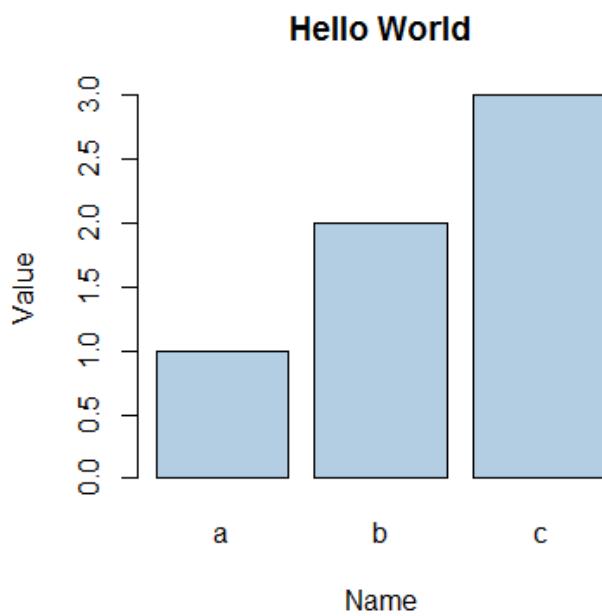


Base

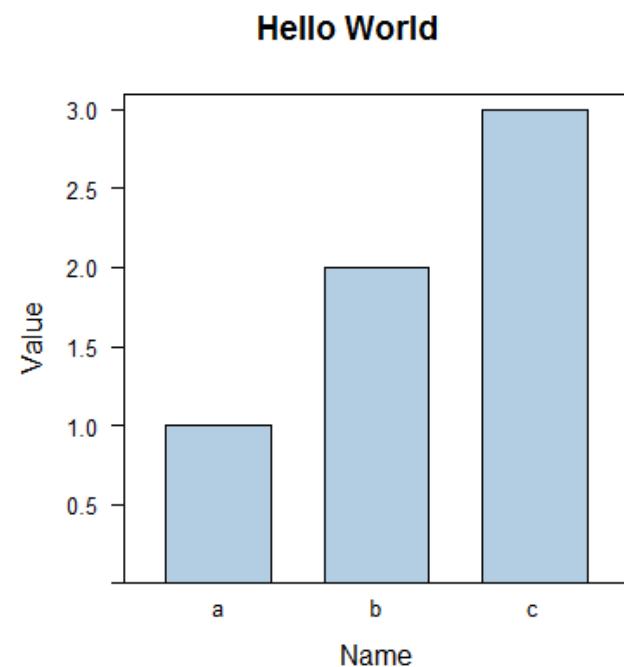


Lattice

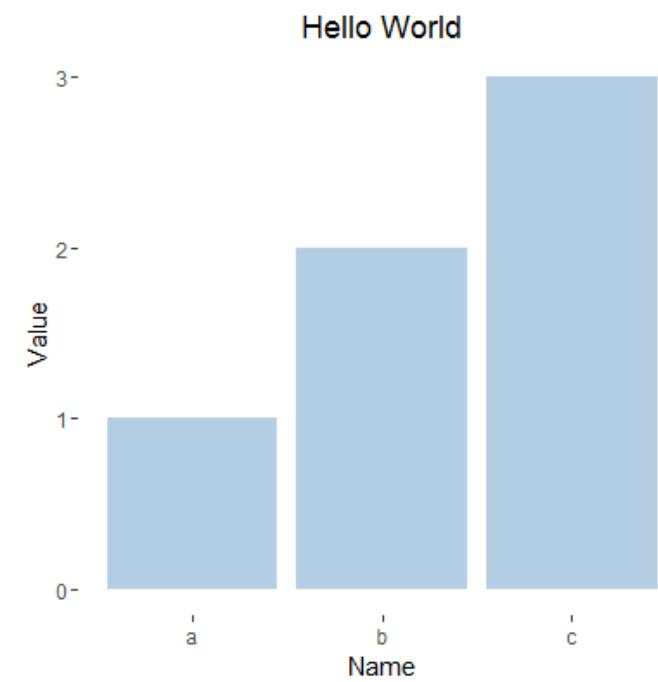
Plotting Systems in R



Base



Lattice



ggplot2



COWBOYS & Space Invaders:



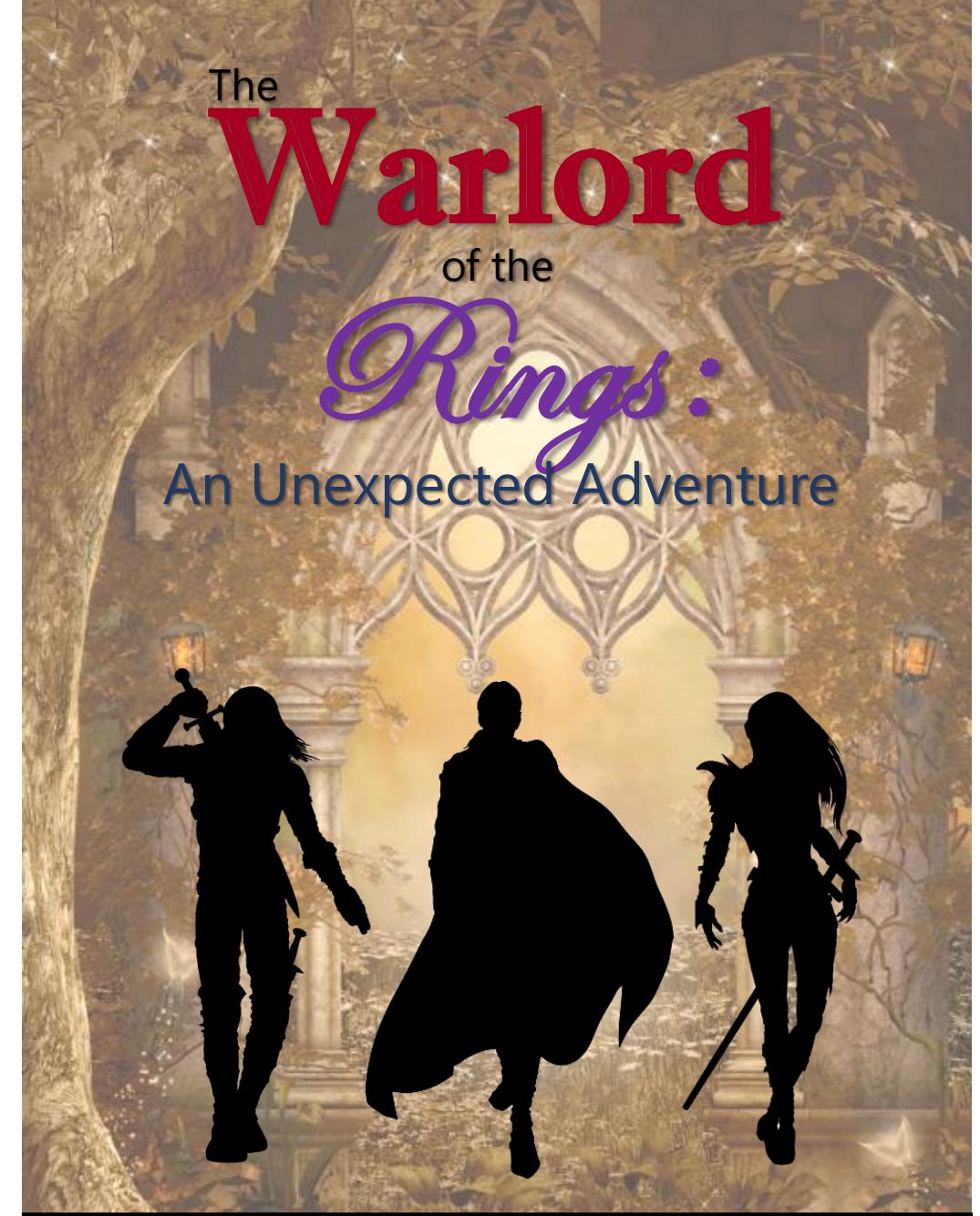
Extended Edition



Code Demo

Lab 4

Data Visualization



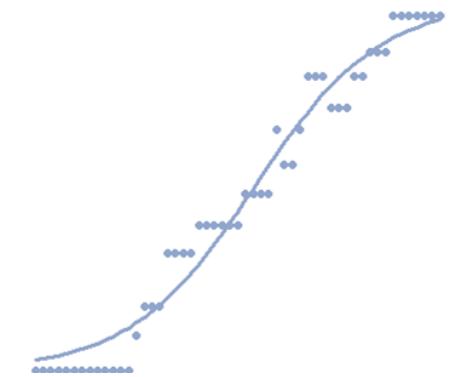
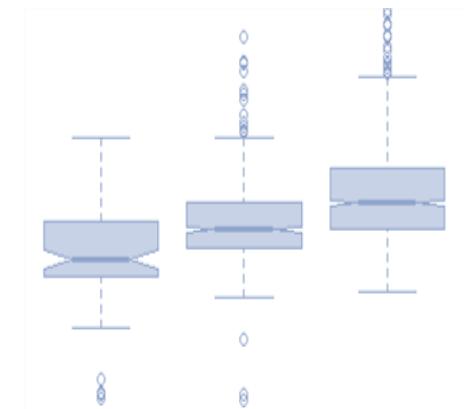
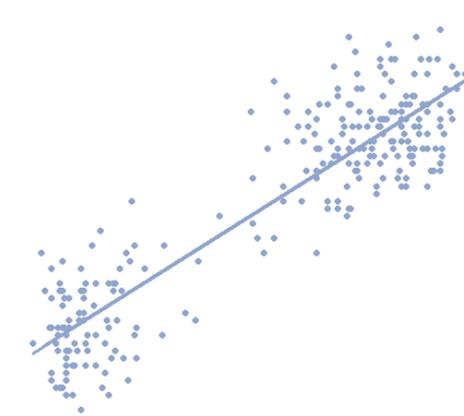
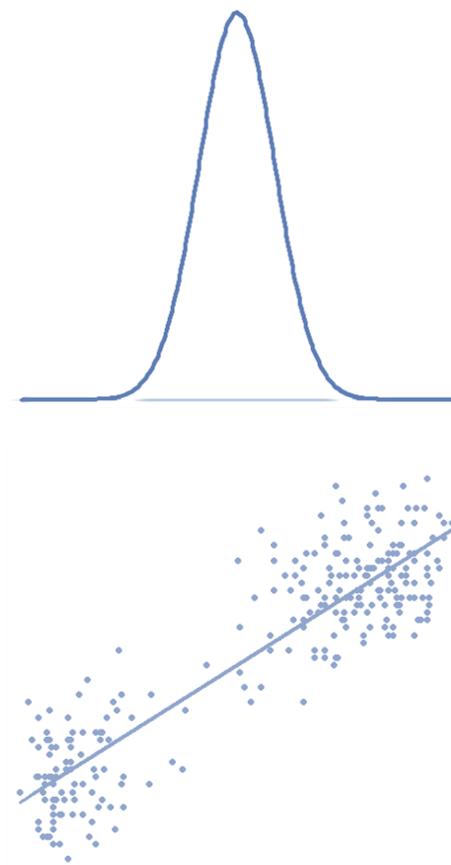
Feature Length

PG

Statistical Modeling

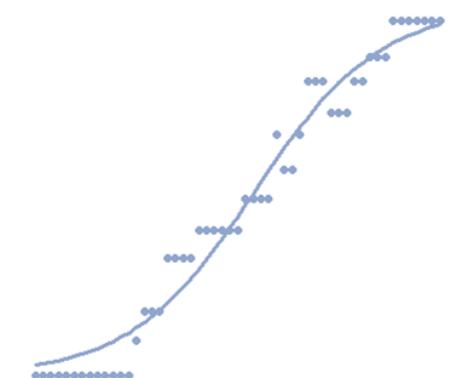
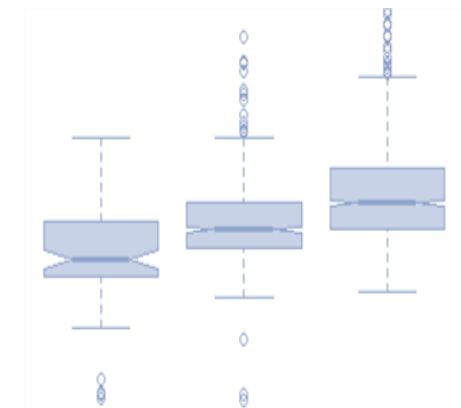
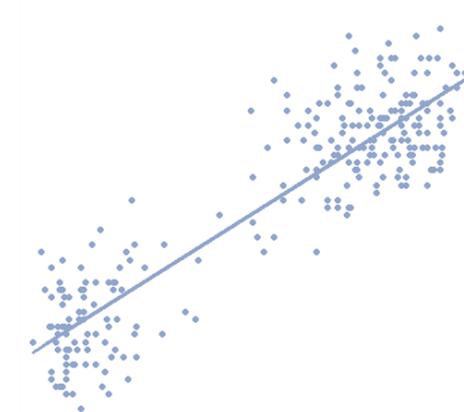
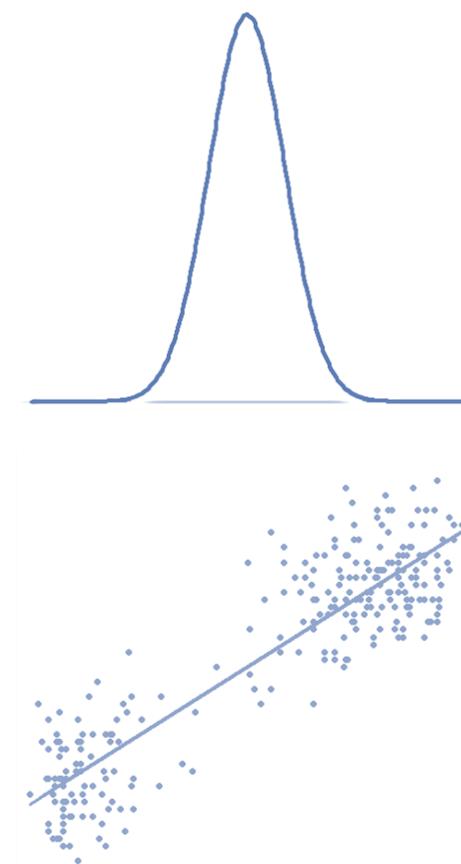
Statistical Model

Mathematical equations



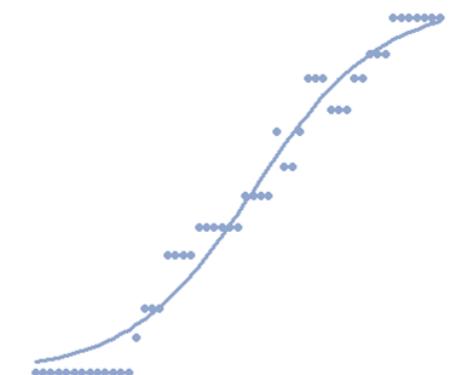
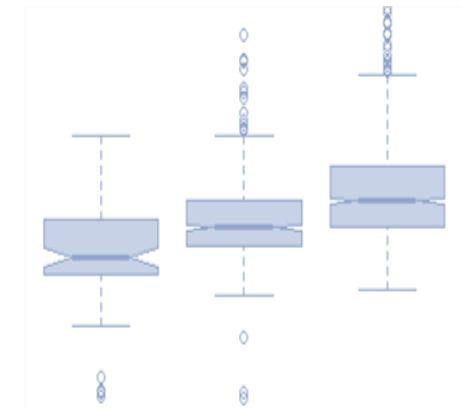
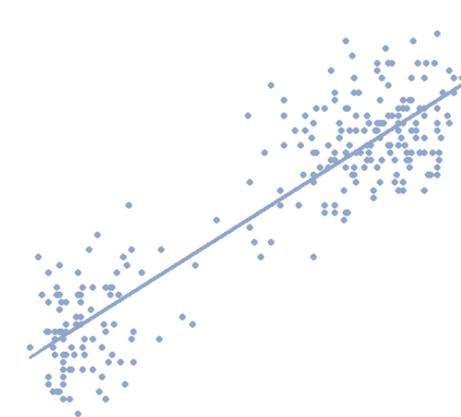
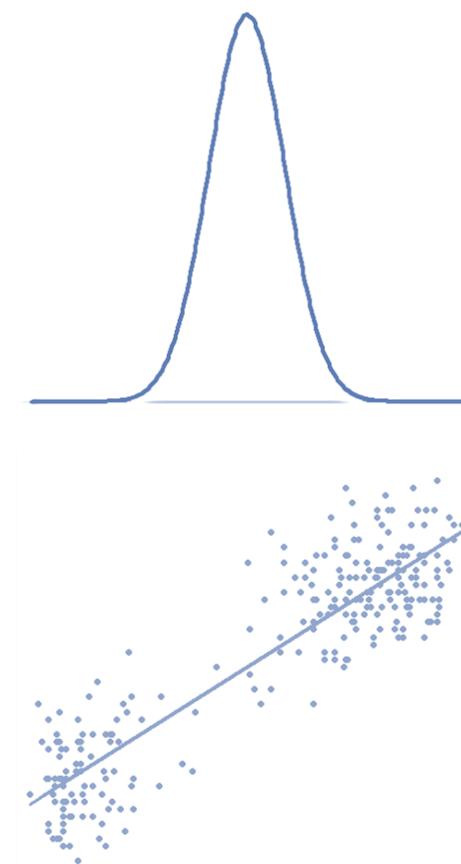
Statistical Model

Mathematical equations
Approximation of reality

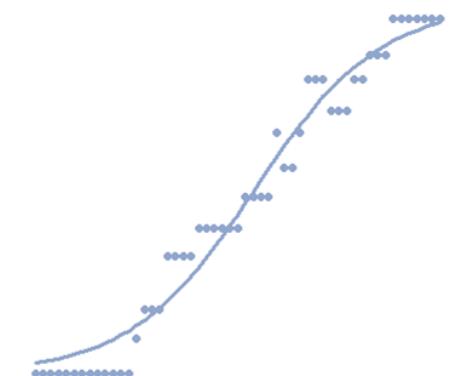
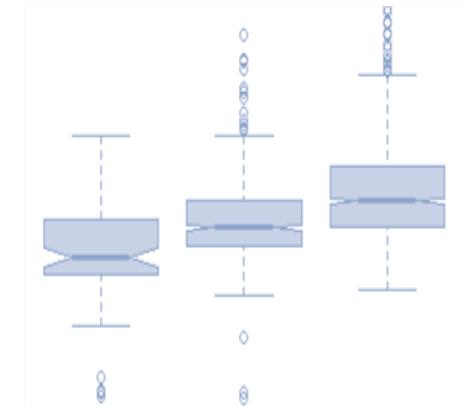
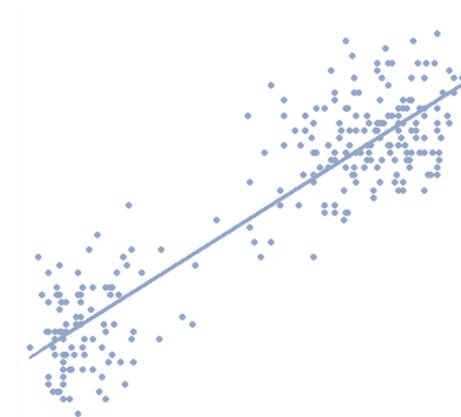
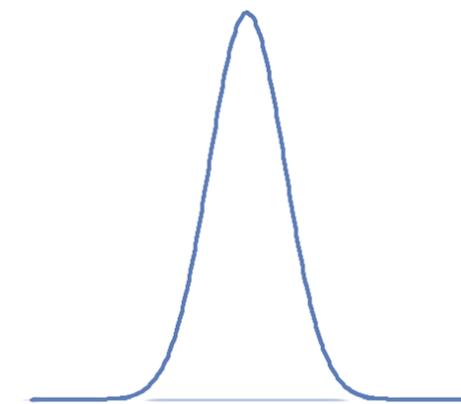


Statistical Model

Mathematical equations
Approximation of reality
Abstract representation

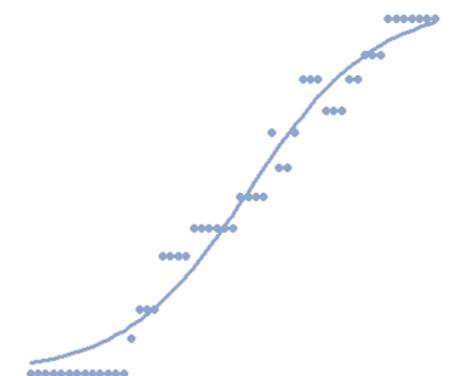
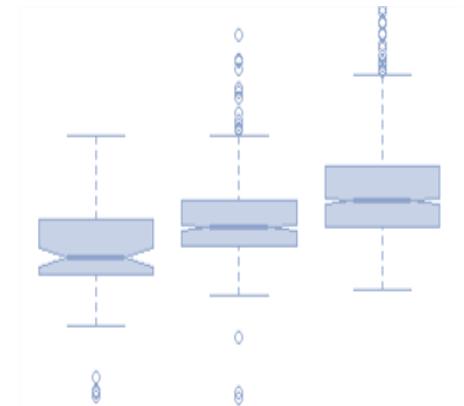
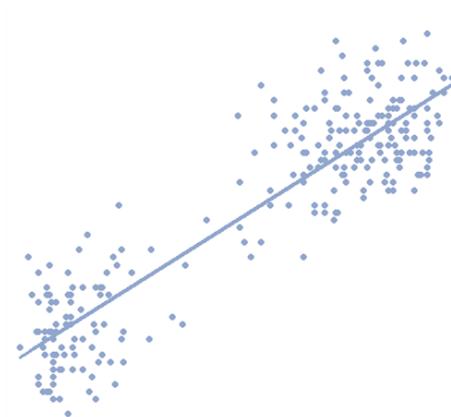
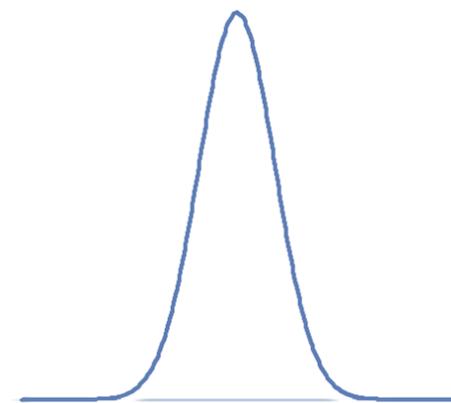


Purpose of Statistical Models



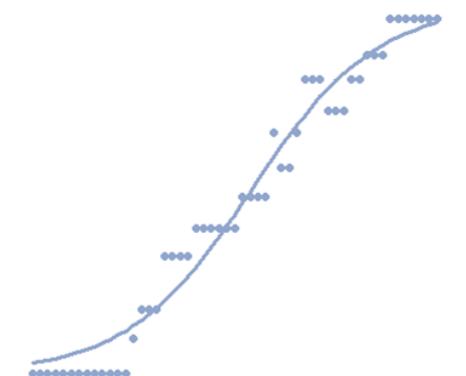
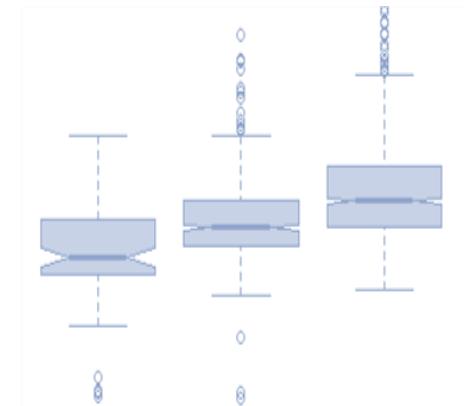
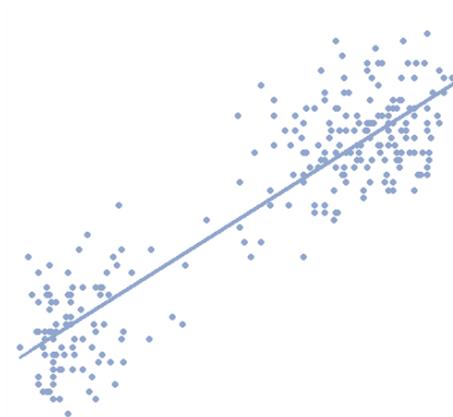
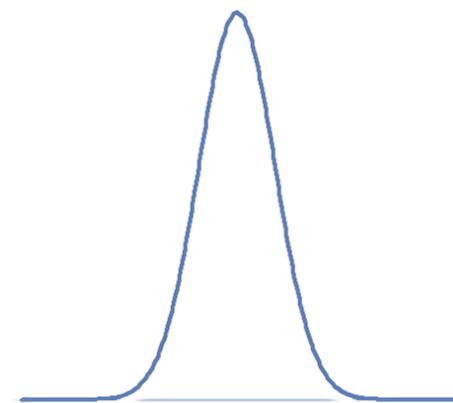
Purpose of Statistical Models

Description



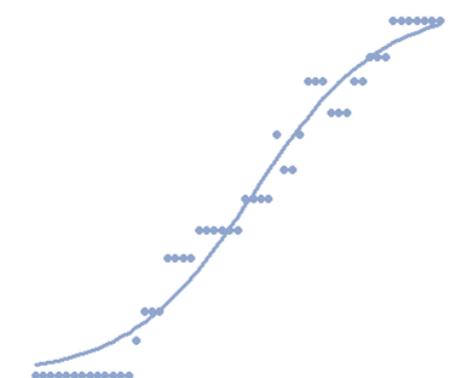
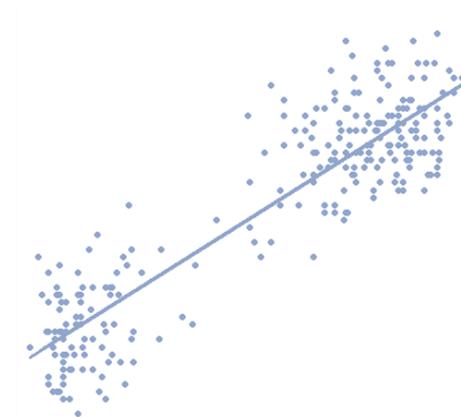
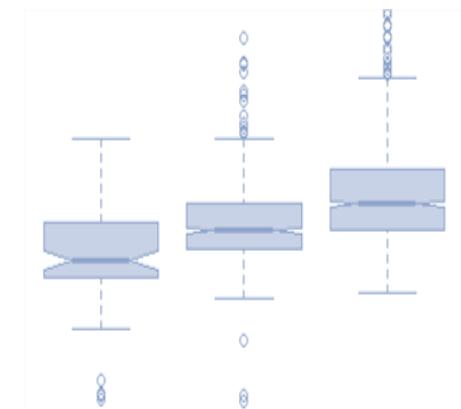
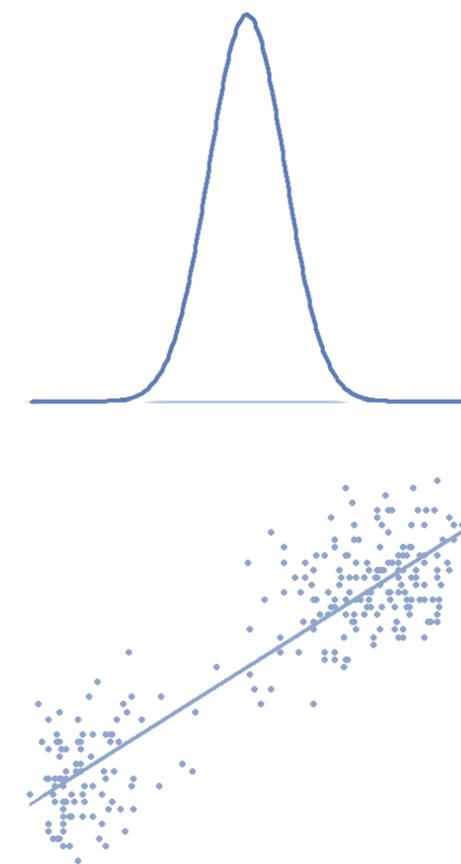
Purpose of Statistical Models

Description
Inference



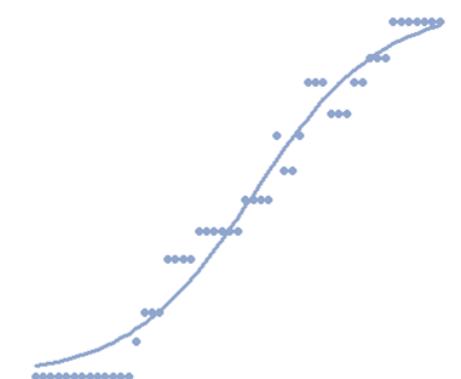
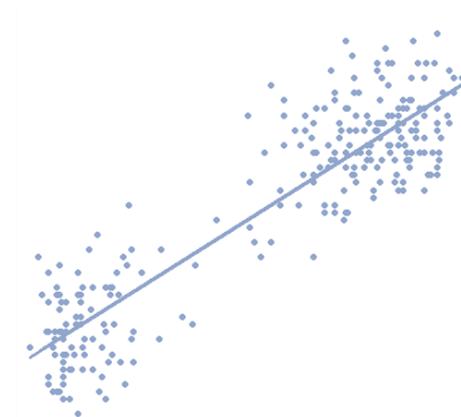
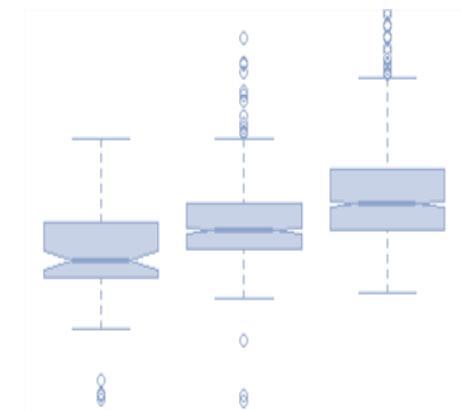
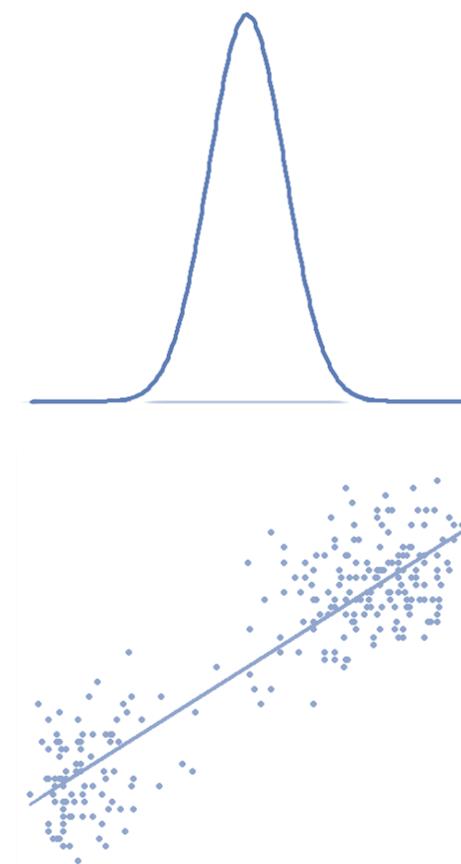
Purpose of Statistical Models

Description
Inference
Comparison



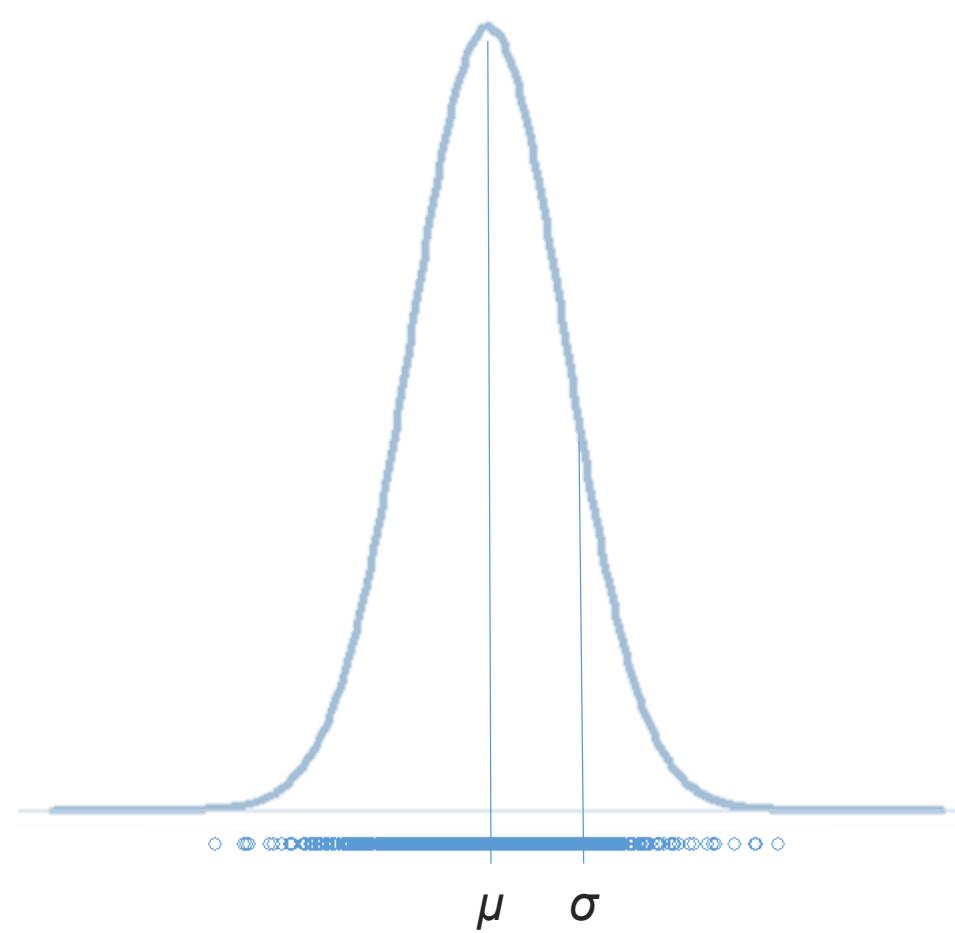
Purpose of Statistical Models

Description
Inference
Comparison
Prediction



Main Categories of Statistical Models

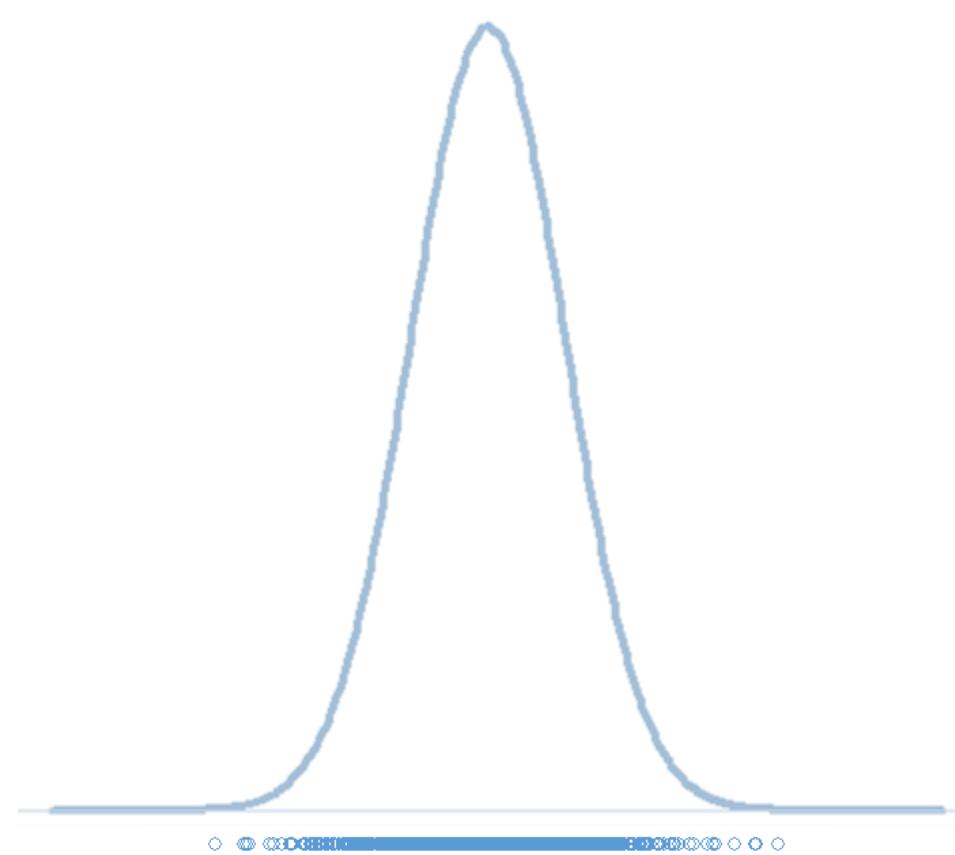
Parametric



Main Categories of Statistical Models

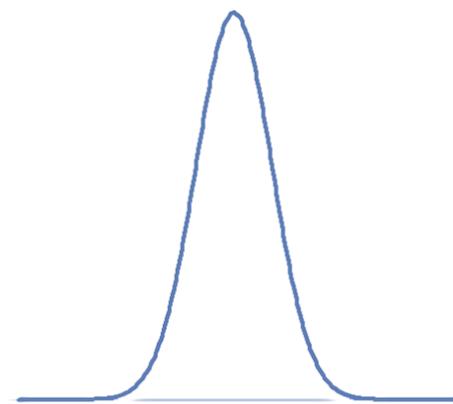
Parametric

Non-parameteric

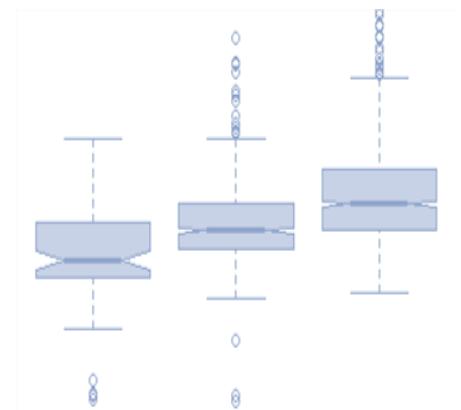


Types of Statistical Models

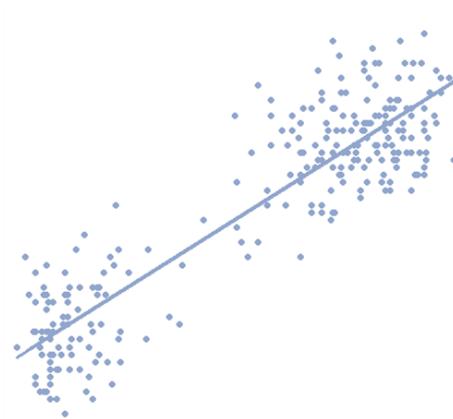
Probability distribution



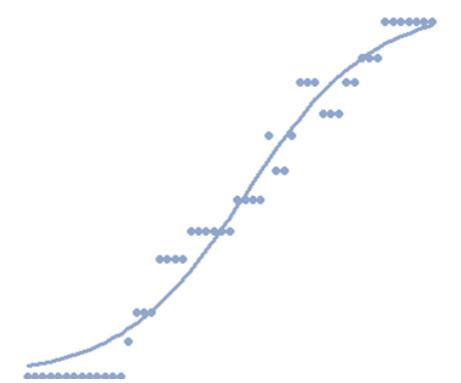
Analysis of variance (ANOVA)



Linear regression

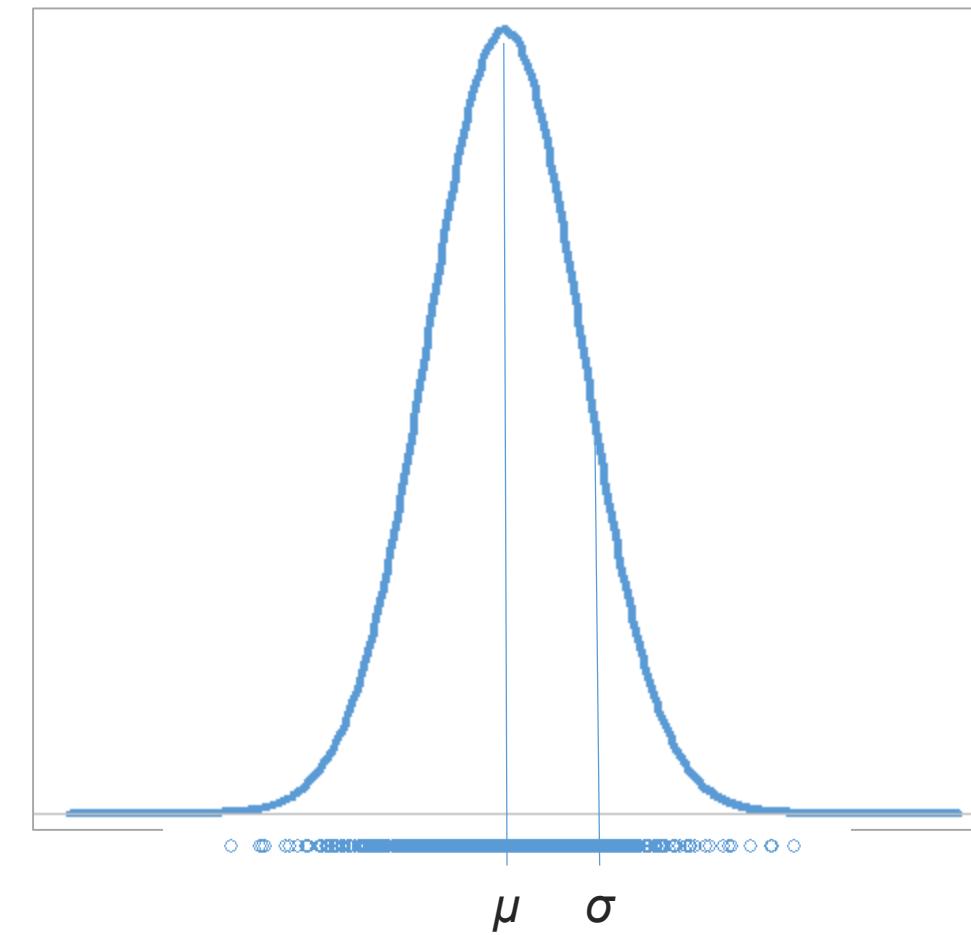


Non-linear regression



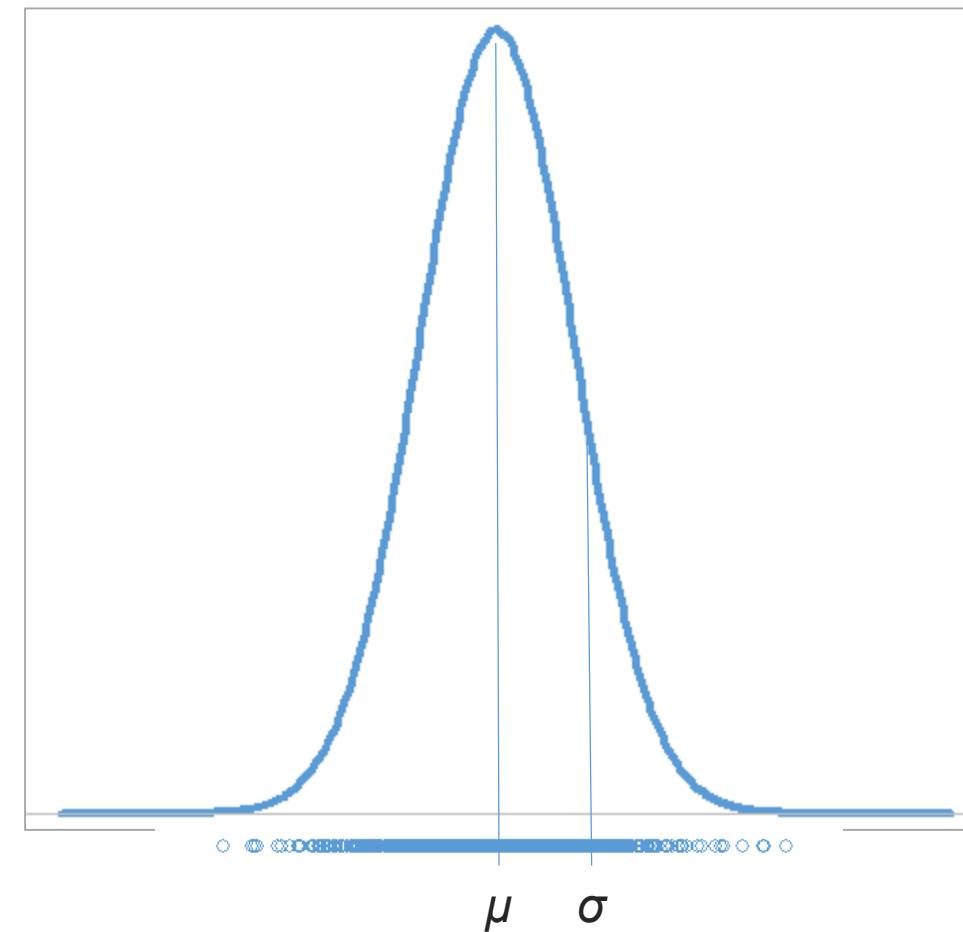
Bayesian network

Gaussian Distribution



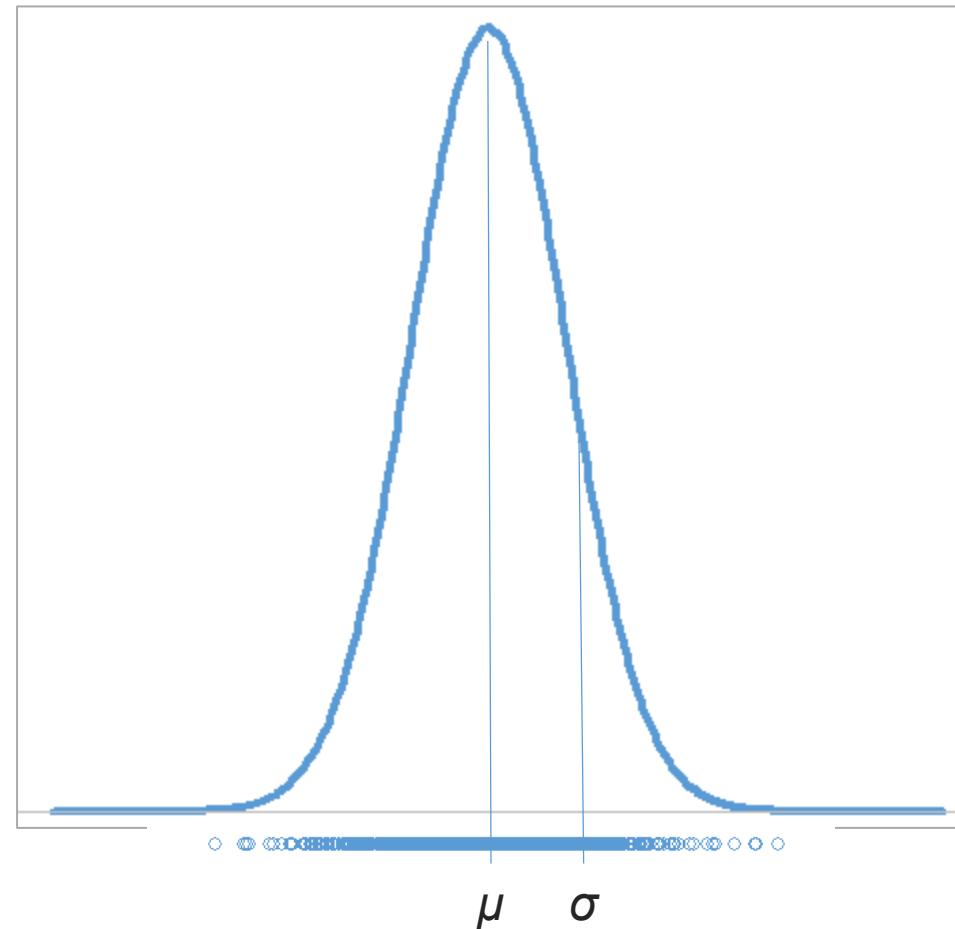
Gaussian Distribution

Probability distribution



Gaussian Distribution

Probability distribution
Parametric model
Mean (μ)
Standard deviation (σ)



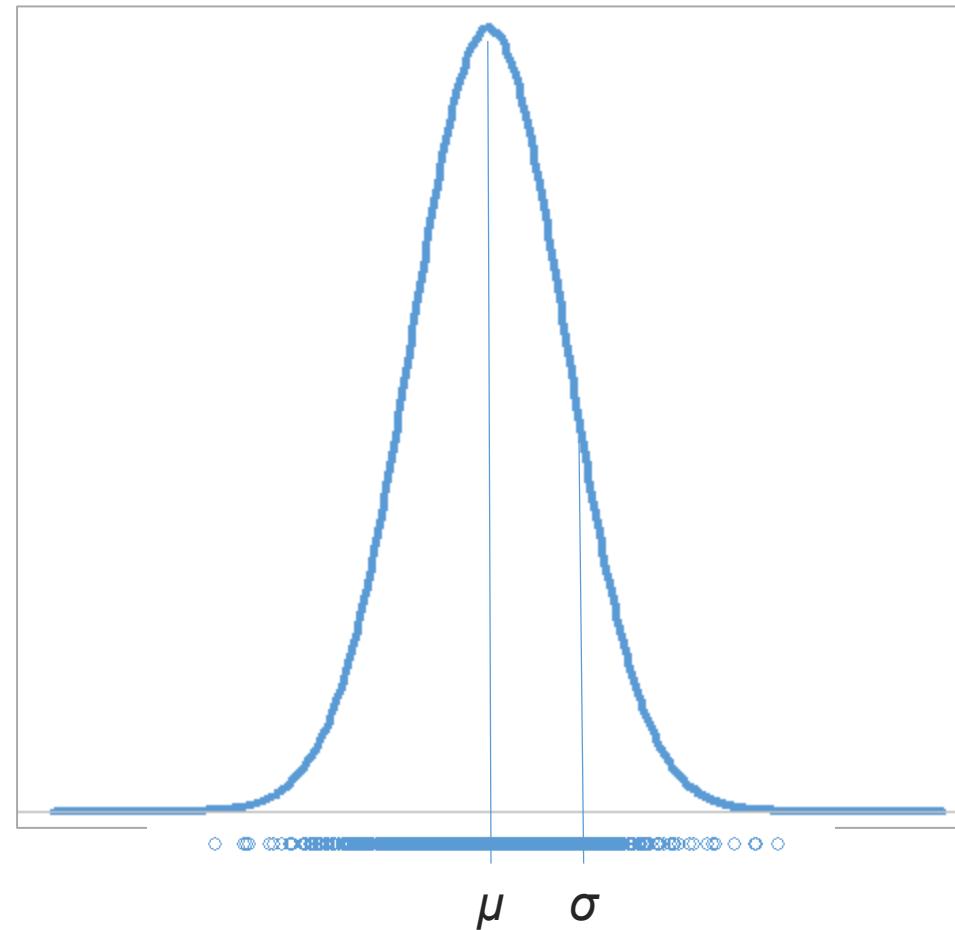
Gaussian Distribution

Probability distribution
Parametric model

Mean (μ)

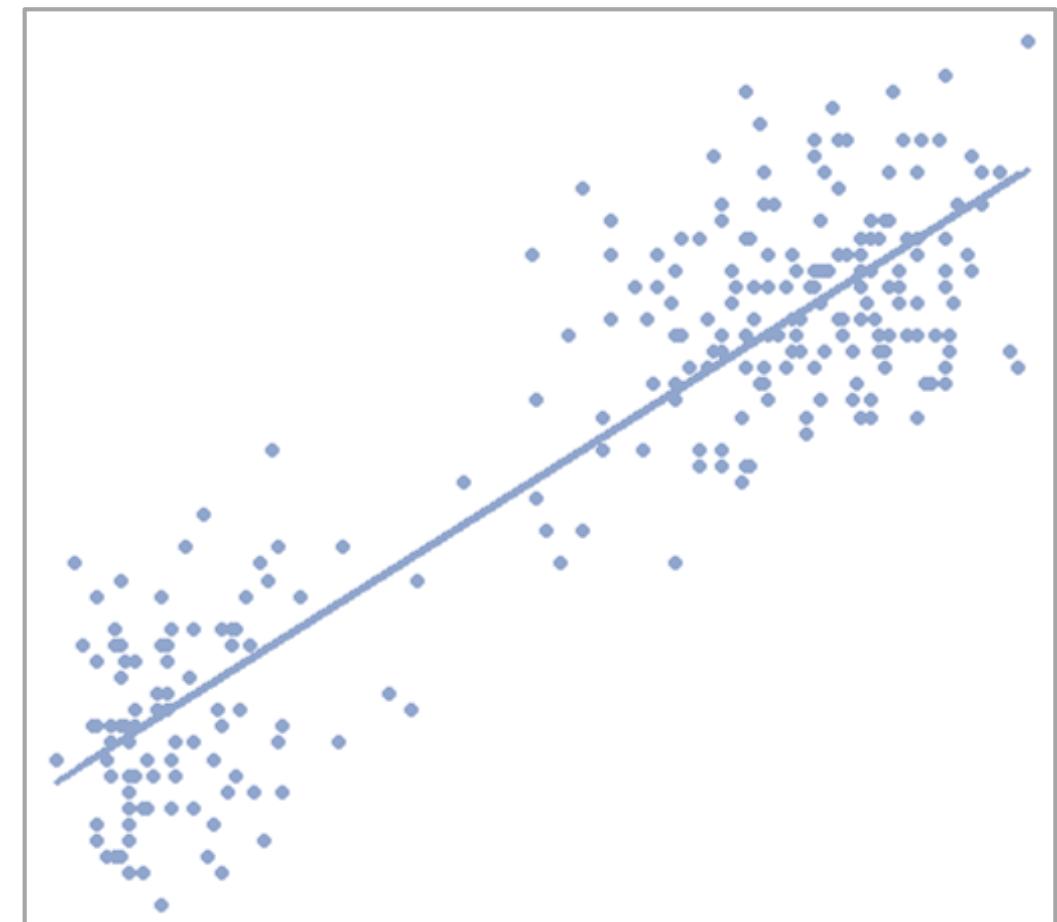
Standard deviation (σ)

Generative model



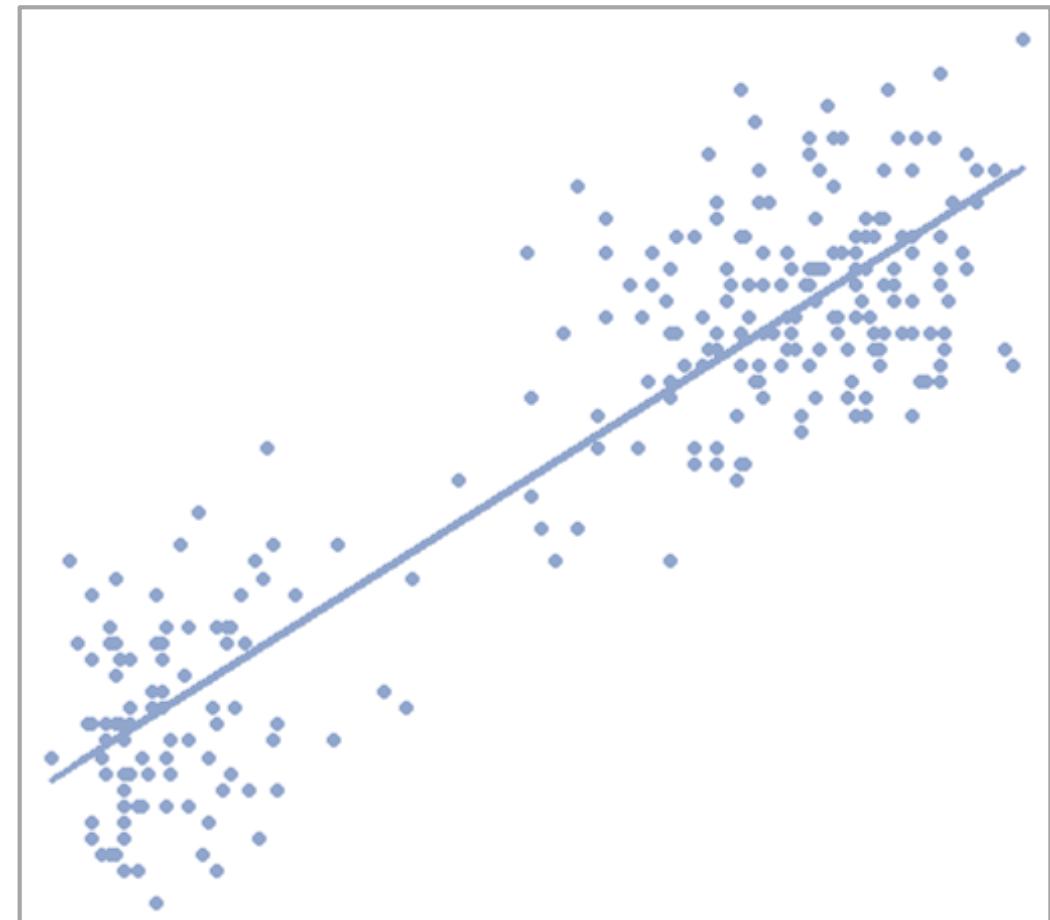
Simple Linear Regression

Relationship



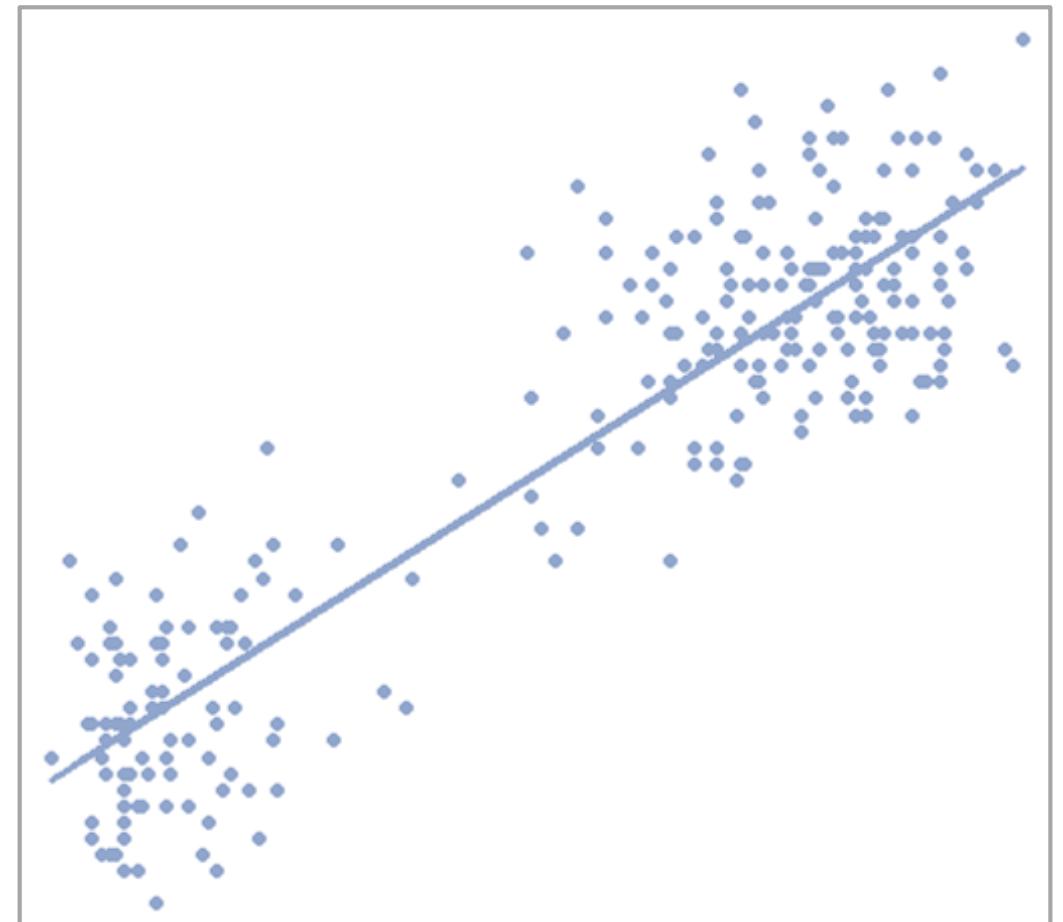
Simple Linear Regression

Relationship
Linear model



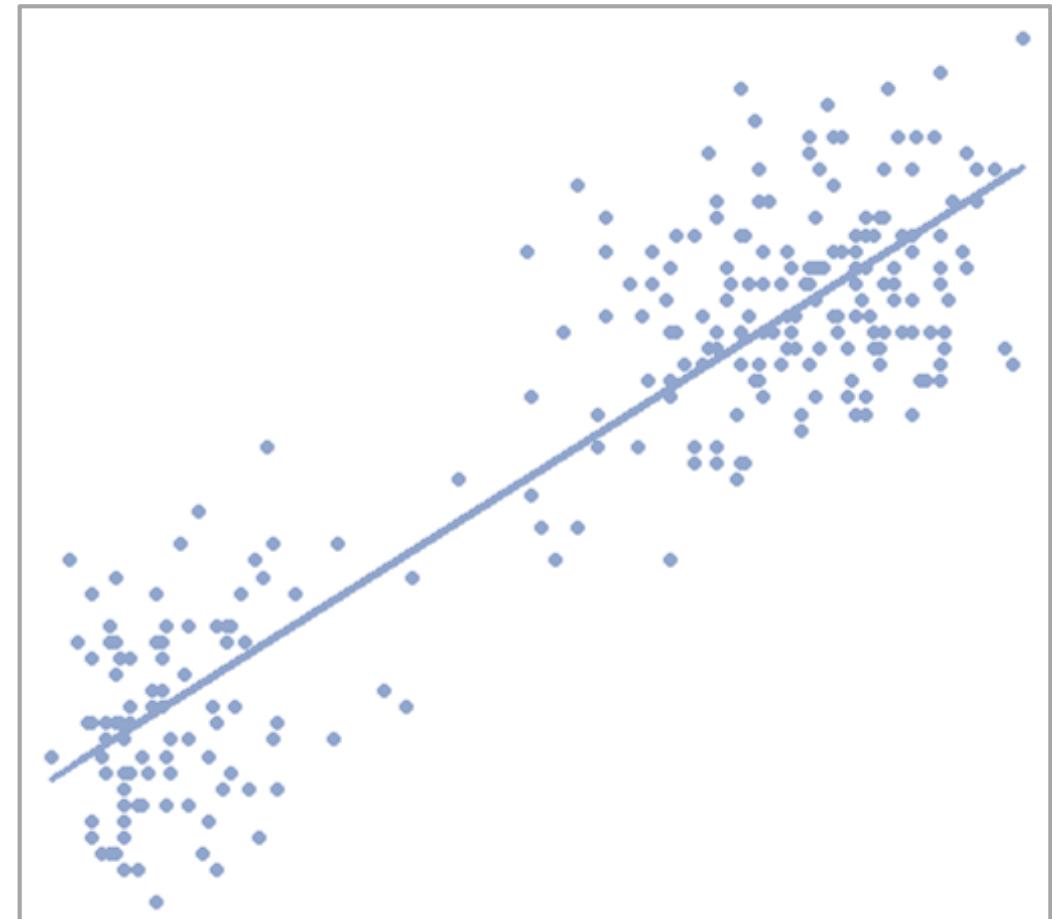
Simple Linear Regression

Relationship
Linear model
Explanatory variable



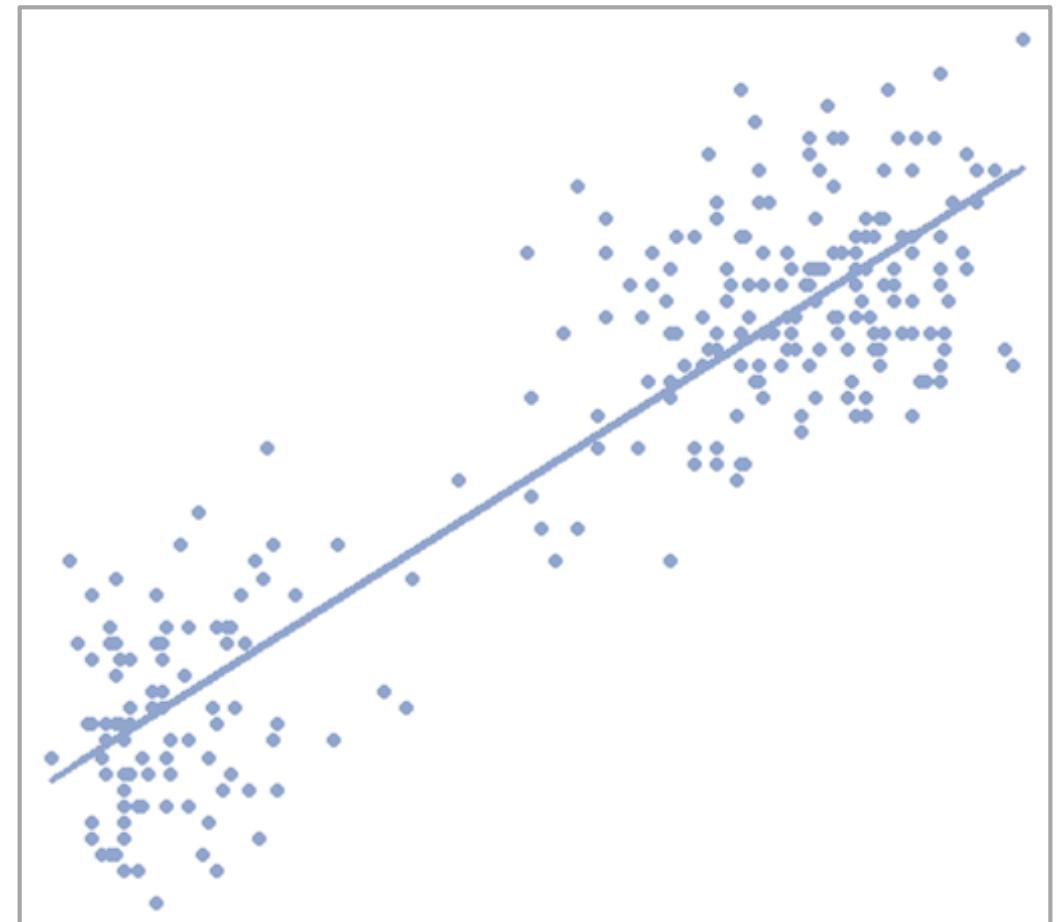
Simple Linear Regression

Relationship
Linear model
Explanatory variable
Outcome variable



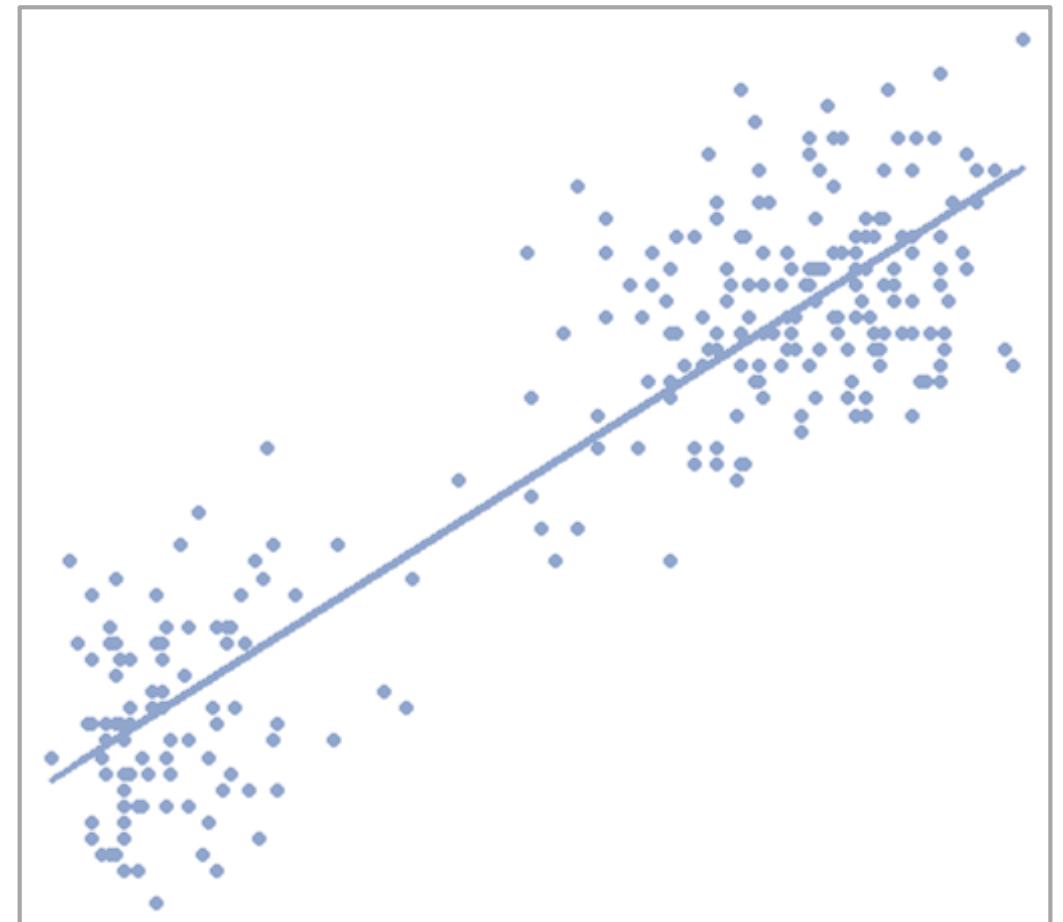
Simple Linear Regression

Linear predictor function



Simple Linear Regression

Linear predictor function
 $y = m \cdot x + b$

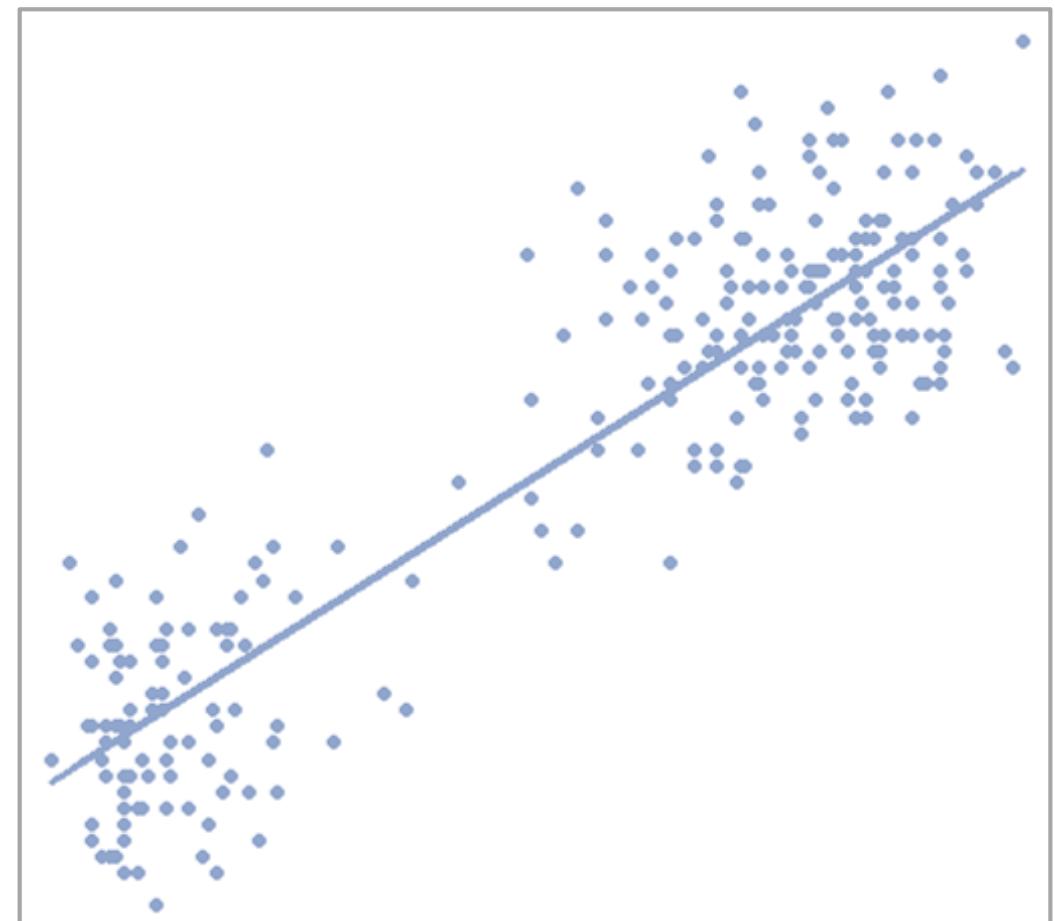


Simple Linear Regression

Linear predictor function

$$y = m \cdot x + b$$

Parameters estimated



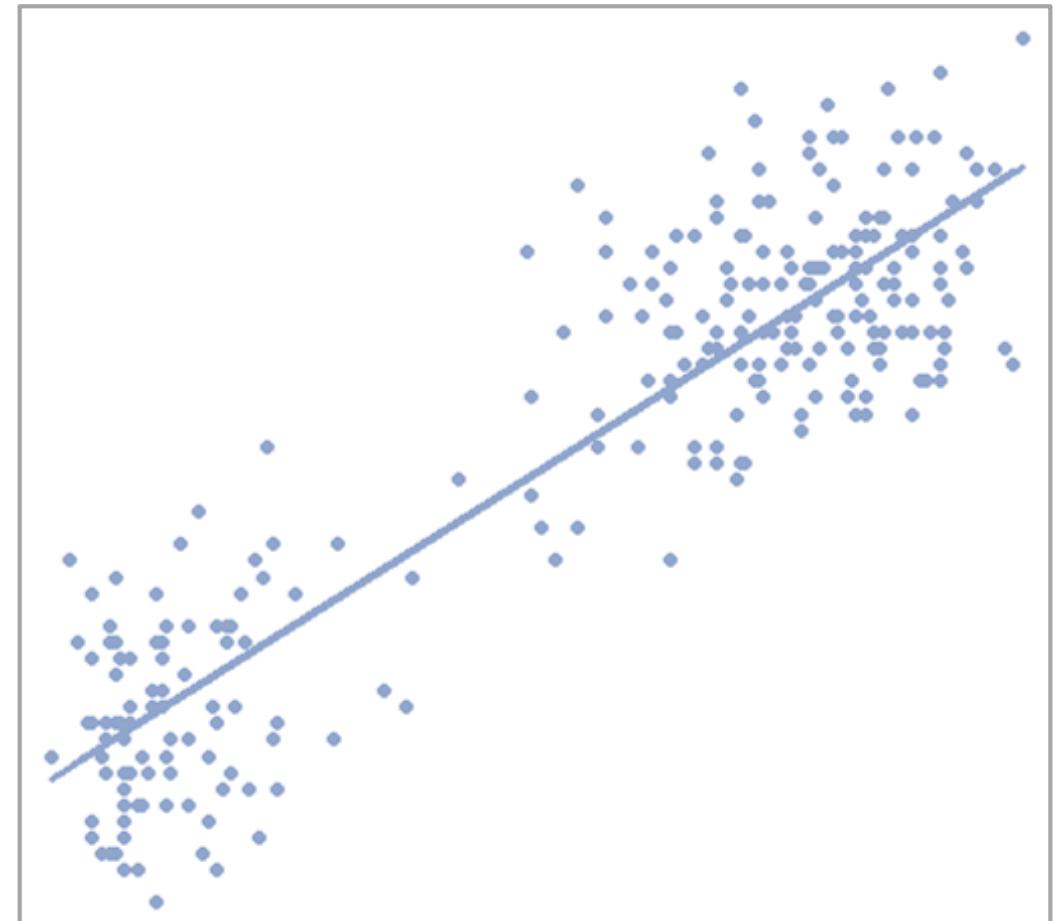
Simple Linear Regression

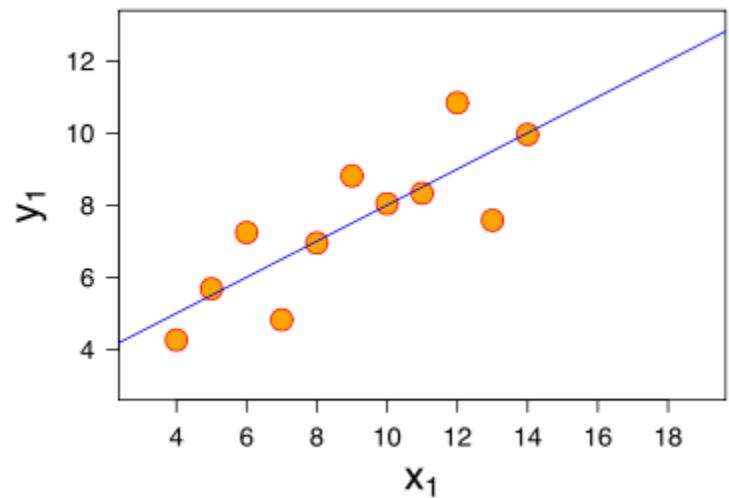
Linear predictor function

$$y = m \cdot x + b$$

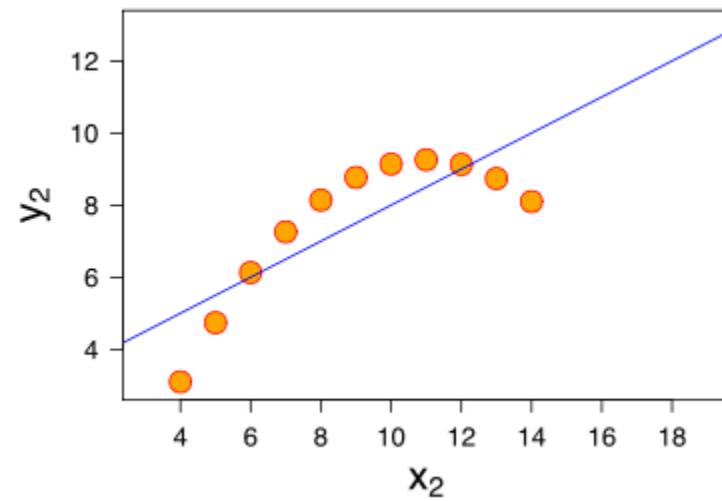
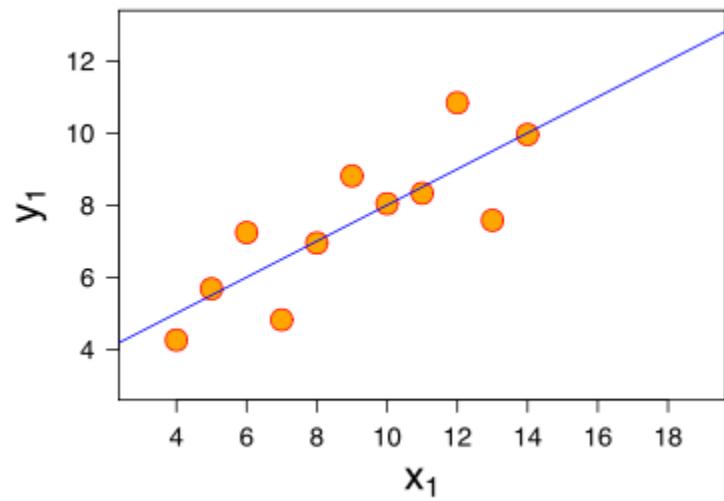
Parameters estimated

Relies on assumptions

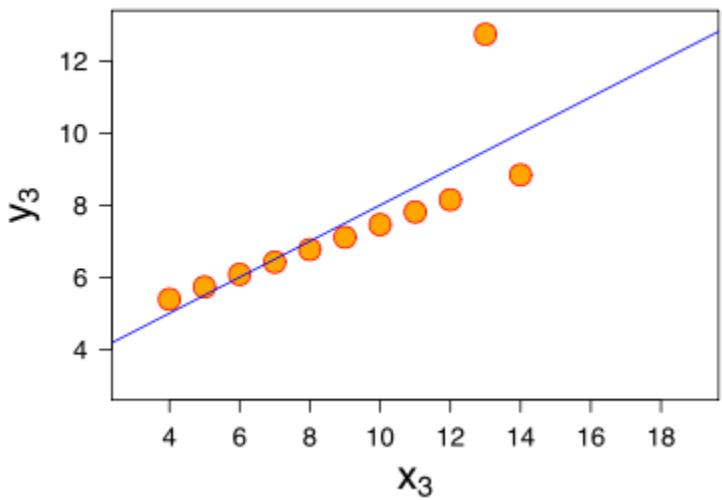
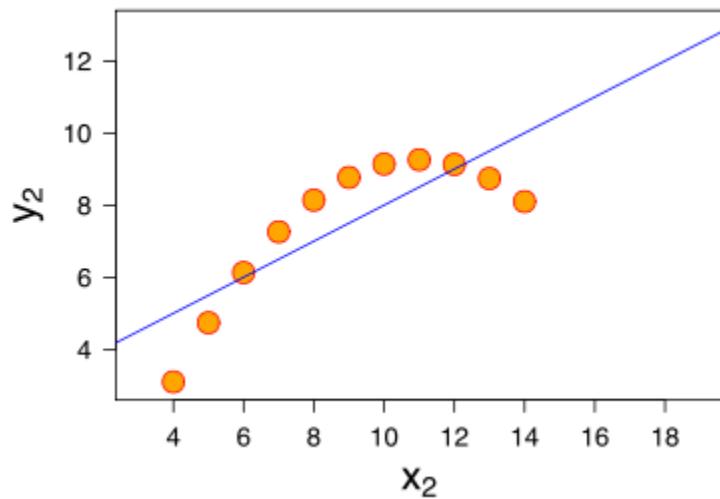
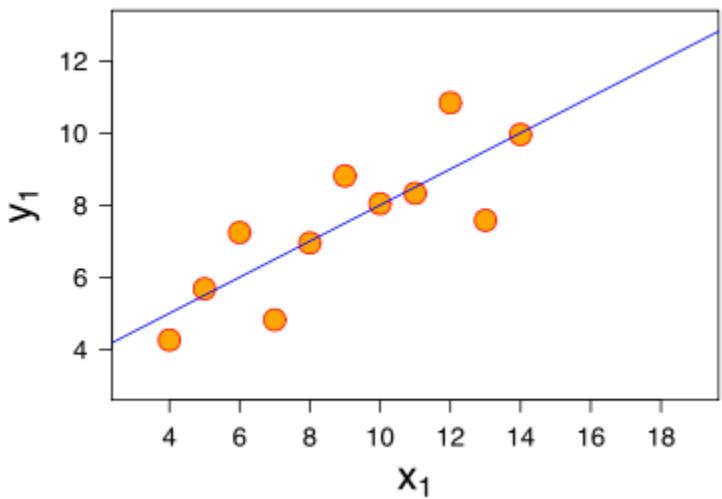




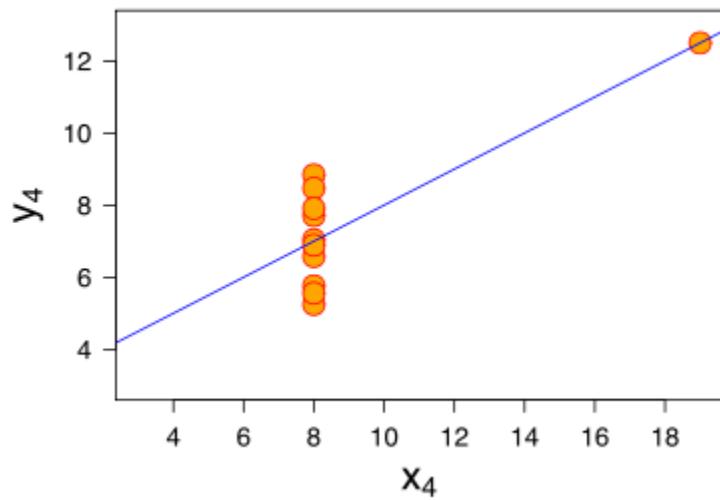
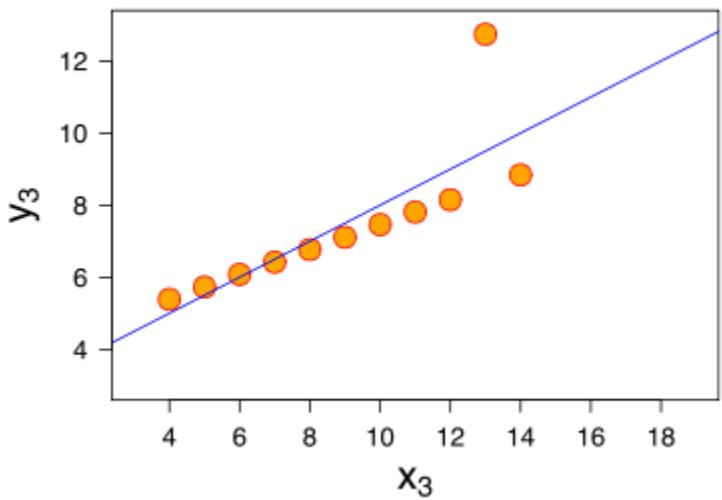
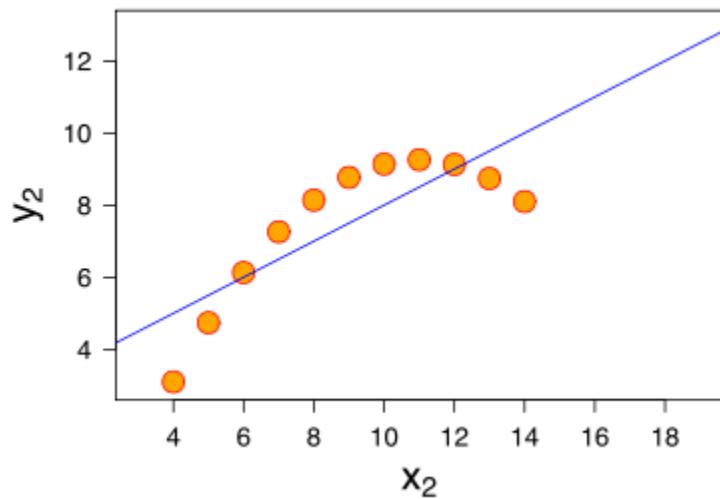
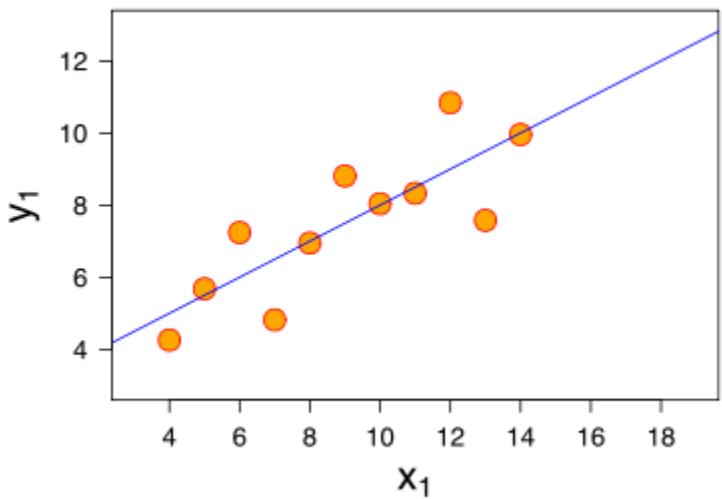
Source: https://en.wikipedia.org/wiki/Anscombe%27s_quartet



Source: https://en.wikipedia.org/wiki/Anscombe%27s_quartet



Source: https://en.wikipedia.org/wiki/Anscombe%27s_quartet



Source: https://en.wikipedia.org/wiki/Anscombe%27s_quartet





Photo by Danielle Langlois

Iris Data Set

| Fisher's Iris Data | | | | |
|--------------------|--------------|-------------|--------------|-------------|
| Species | Petal Length | Petal Width | Sepal Length | Sepal Width |
| setosa | 1.1 | 0.1 | 4.3 | 3 |
| setosa | 1.4 | 0.2 | 4.4 | 2.9 |
| setosa | 1.3 | 0.2 | 4.4 | 3 |
| setosa | 1.3 | 0.2 | 4.4 | 3.2 |
| setosa | 1.3 | 0.3 | 4.5 | 2.3 |
| ... | | ... | ... | ... |

Iris Data Set



Iris Setosa



Iris Versicolor



Iris Virginica

Code Demo

Lab 5

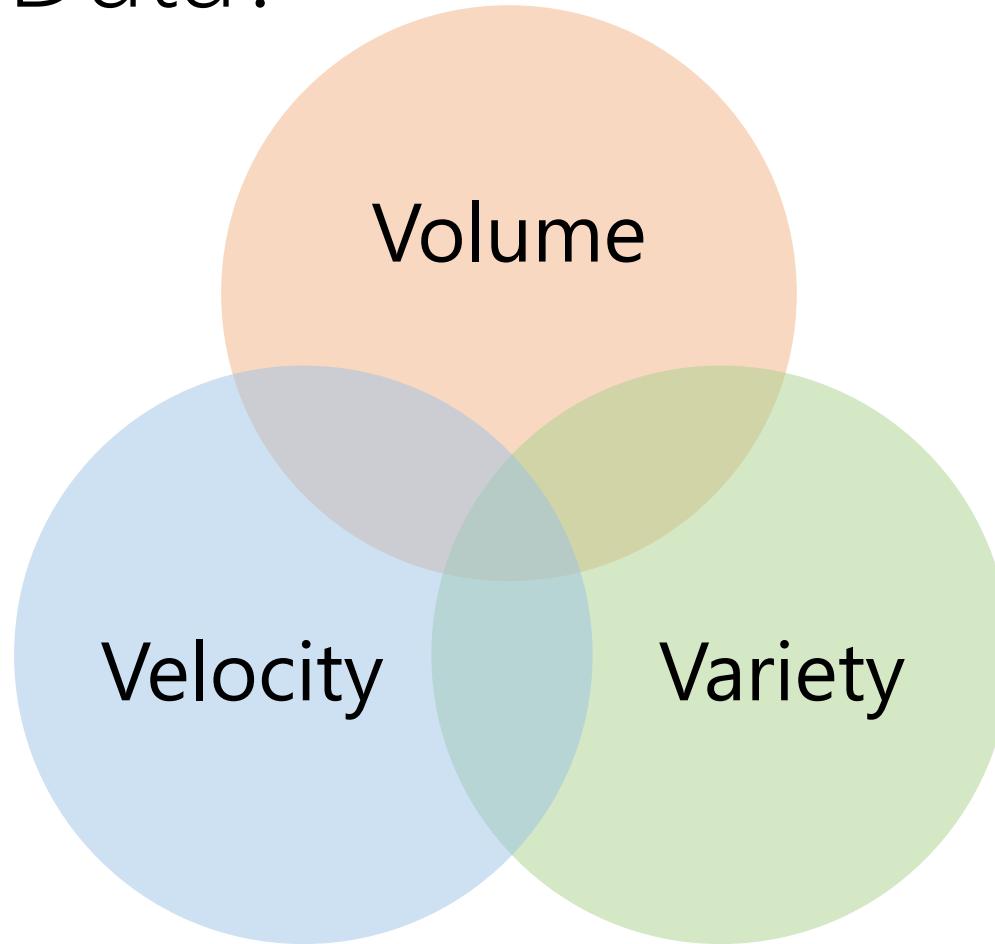
Statistical Models



Photos by Radomił Binek,
Danielle Langlois, and Frank Mayfield

Handling Big Data

What is Big Data?



Big Data is a Moving Target

What is Big Data today
Will not be Big Data tomorrow



Do I Have Big Data?

Can it fit in memory?

Can it fit on your hard drive?

Can you process it in a day?

Does it fit into tables?



Big Data Decision Table

| Big Data Decision Table | | | | |
|-------------------------|------|------|---------|------------|
| Class | Rows | Size | Storage | Management |
| | | | | |
| | | | | |
| | | | | |

Big Data Decision Table

| Big Data Decision Table | | | | |
|-------------------------|----------|-----------|---------|-------------------------|
| Class | Rows | Size | Storage | Management |
| Small | Millions | Gigabytes | Memory | R with desktop computer |
| | | | | |
| | | | | |

Big Data Decision Table

| Big Data Decision Table | | | | |
|-------------------------|----------|-----------|-----------|-------------------------------|
| Class | Rows | Size | Storage | Management |
| Small | Millions | Gigabytes | Memory | R with desktop computer |
| Medium | Billions | Terabytes | Hard disk | R with medium-data extensions |
| | | | | |

Big Data Decision Table

| Big Data Decision Table | | | | |
|-------------------------|-----------|-----------|-----------|-------------------------------|
| Class | Rows | Size | Storage | Management |
| Small | Millions | Gigabytes | Memory | R with desktop computer |
| Medium | Billions | Terabytes | Hard disk | R with medium-data extensions |
| Big | Trillions | Petabytes | Clusters | R with big-data extensions |

If you don't have a big data problem,
but you think you have a big data problem,
then you've just created a big data problem!

How to Handle Big Data in R

More hardware

Subsetting

Sampling

Microsoft R Open

Medium data packages

Big data packages

Microsoft R Server

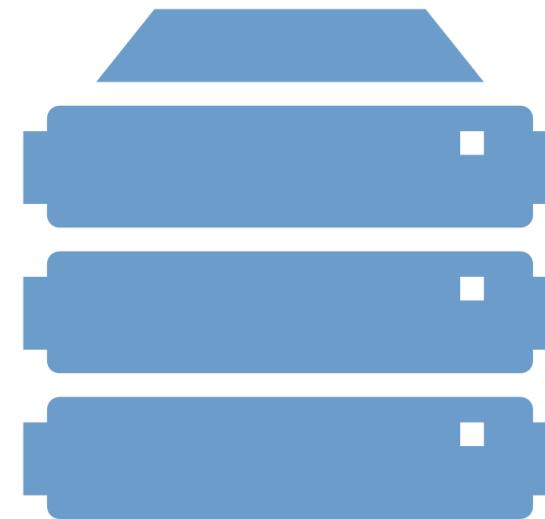


More Hardware

More CPU, memory, disk

Cost vs. benefit

Cloud Virtual Machine

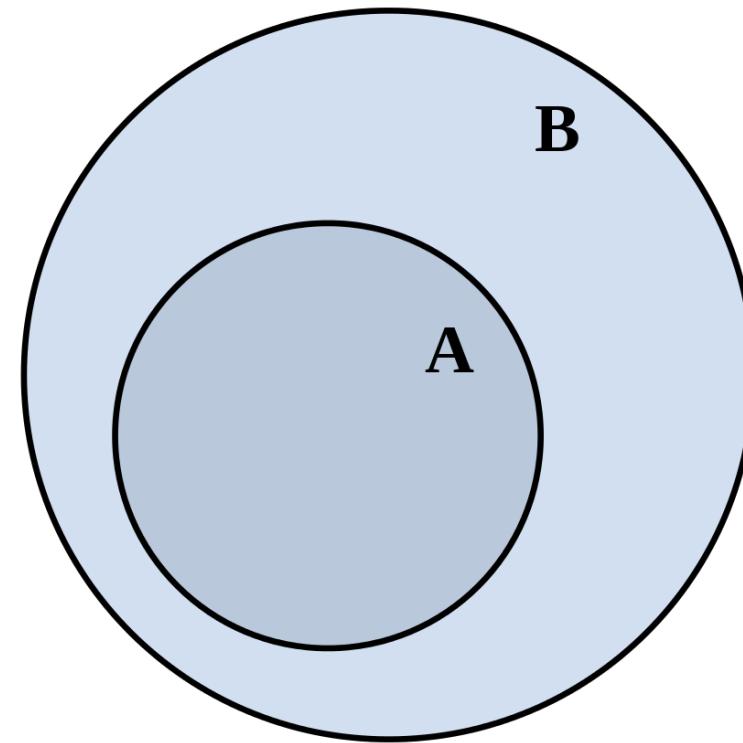


Subsetting

Select subset of data

Use dplyr with src

Various databases supported



Sampling

Randomly selected
Subset of original data
Low-cost solution



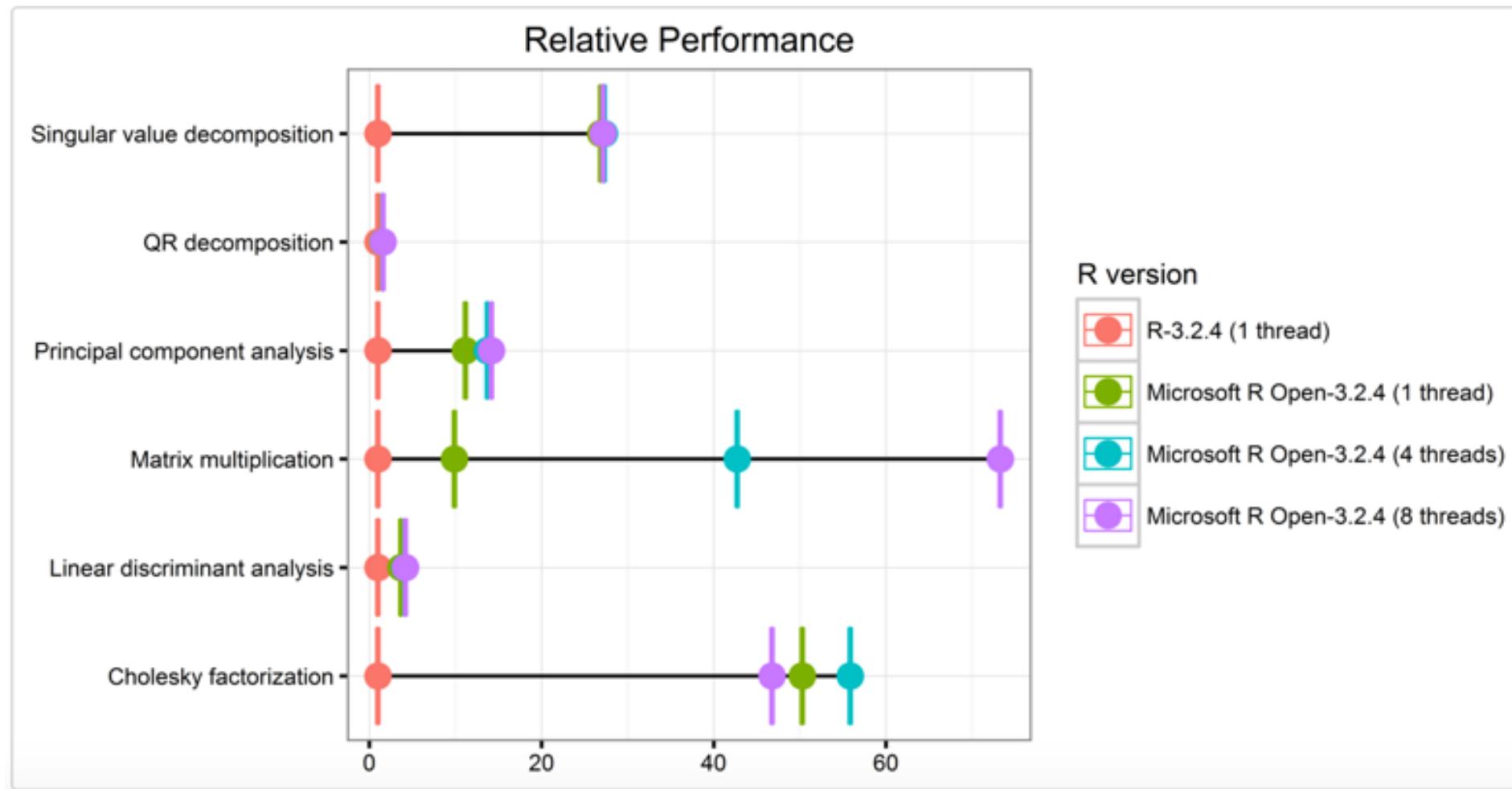
Microsoft R Open

- Enhance 64-bit R distribution
- Multithreaded performance
- Only helps with CPU constraint
- Package repository snapshots



Source: Microsoft R Open

Microsoft R Open Performance



Source: <https://mran.microsoft.com/documents/rro/multithread/#mt-bench>

3rd-Party Extension Packages

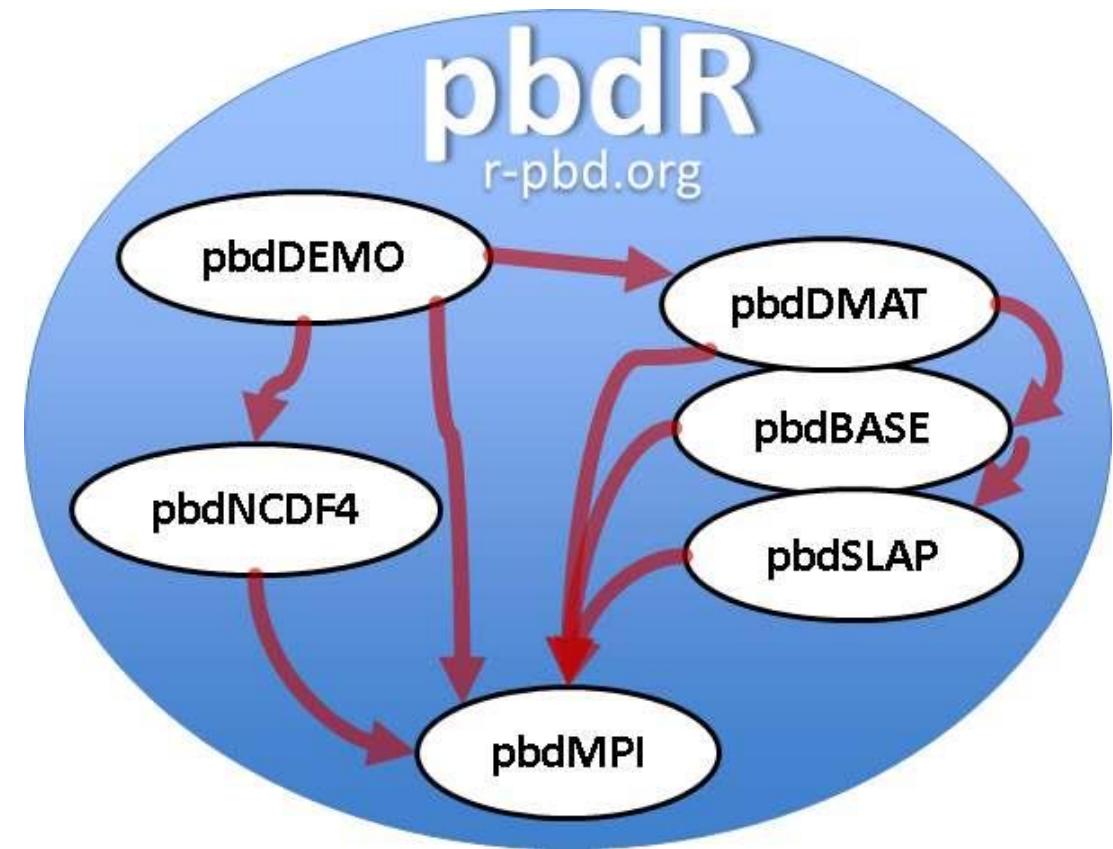
Big Memory

ff

Big LM

3rd-Party Extension Packages

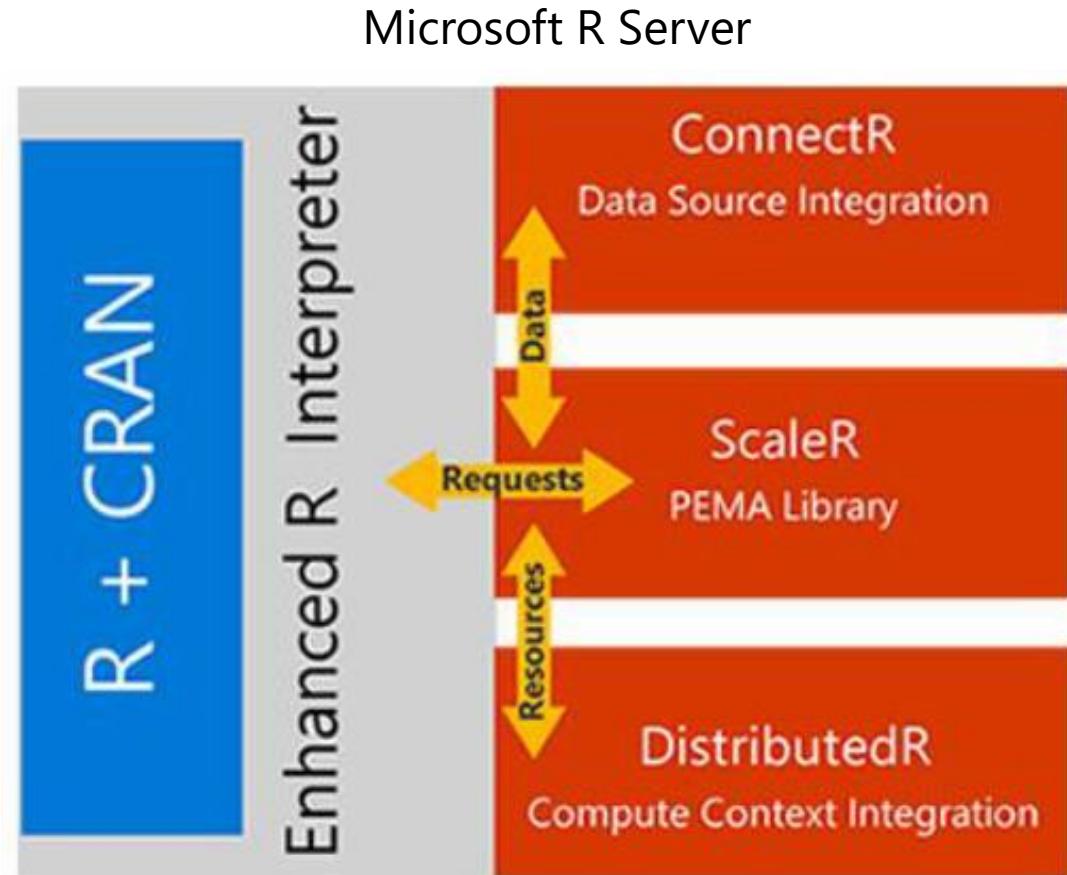
pbdR
Rhive
Rhive
Rbase
Rhdfs,
Rmr



Source: Wikipedia

Microsoft R Server

Windows
SQL Server 2016
SUSE Linux
Redhat Linux
Teradata
Hadoop
HD Insight



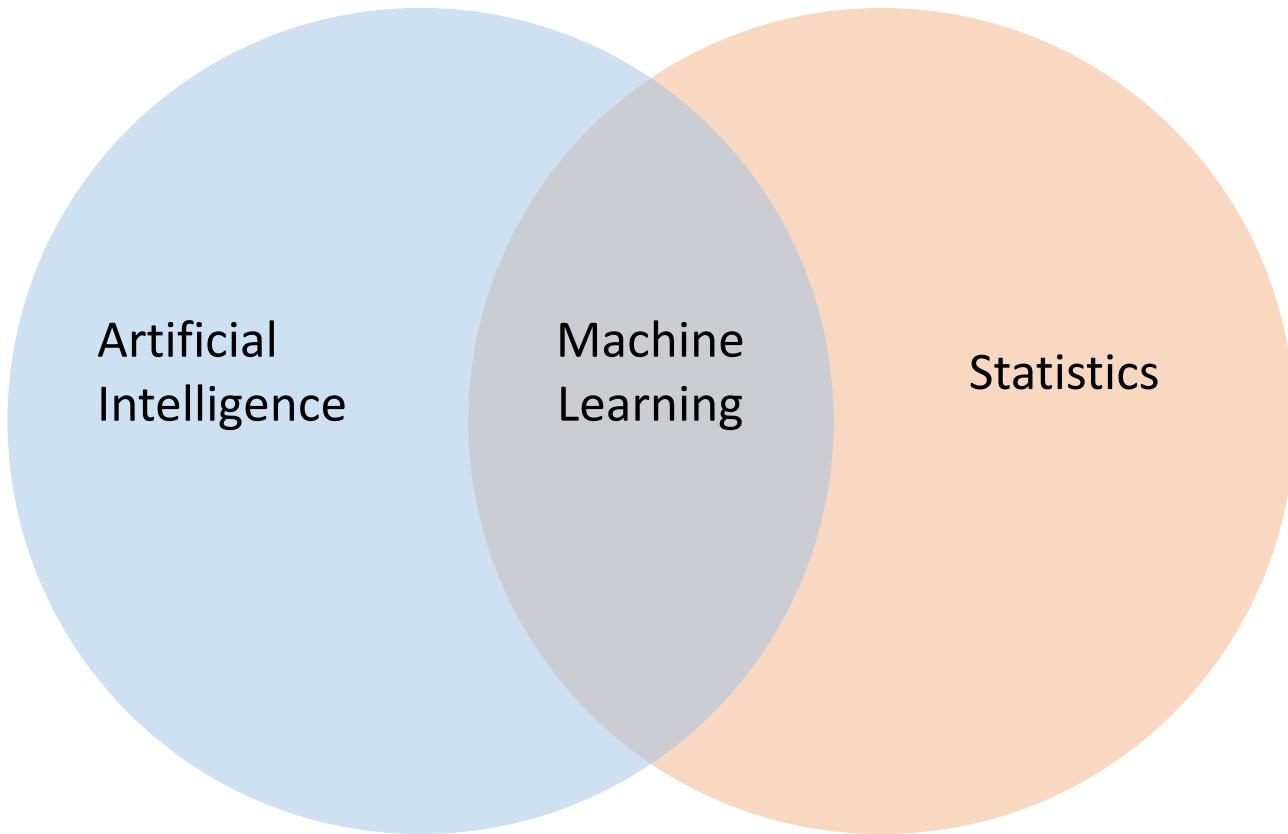
Source: Microsoft R Server

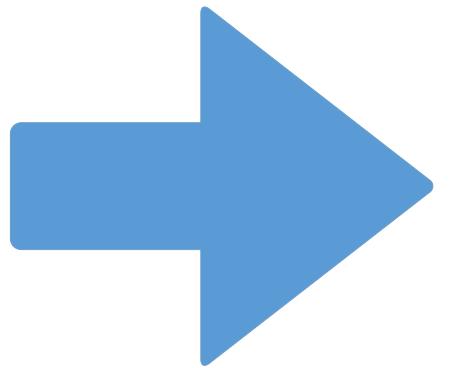
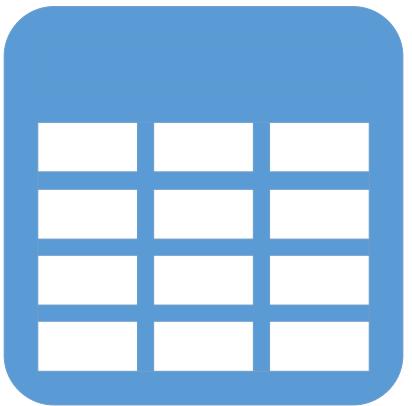
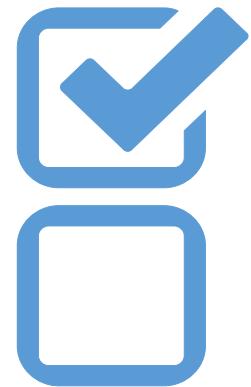
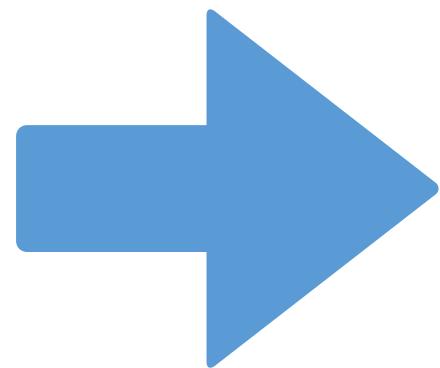
Code Demo

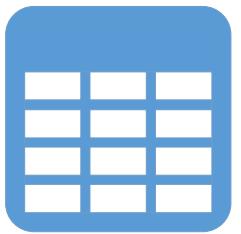
Lab 6

Handling Big Data

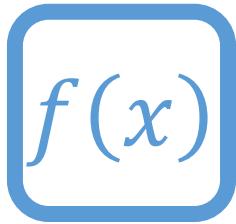
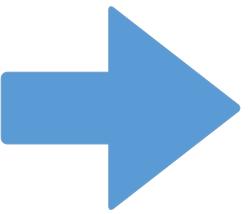
Machine Learning



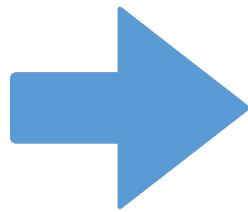
 $f(x)$ 



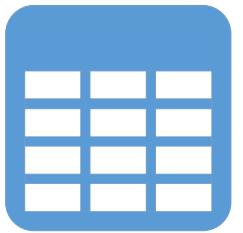
Data



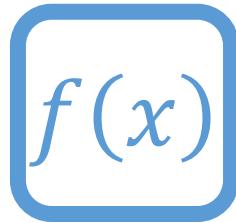
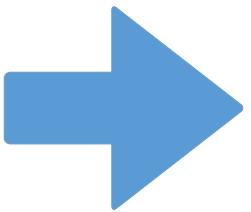
Function



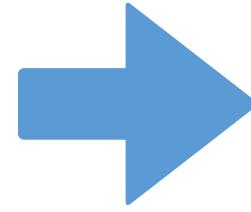
Prediction



Data

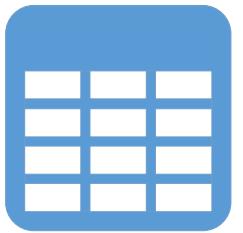


Function

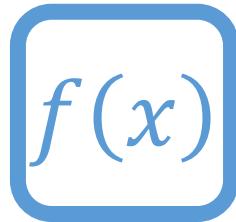
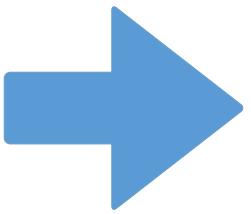


Prediction

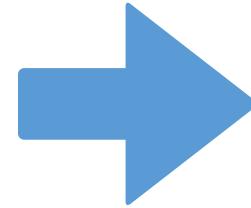




Data



Function



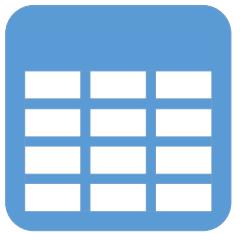
Prediction



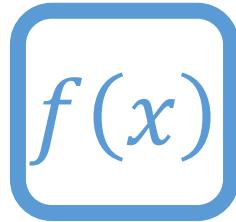
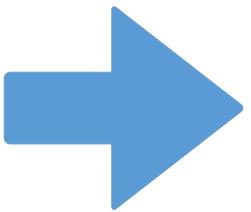
Cat



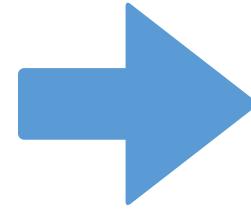
Dog



Data



Function



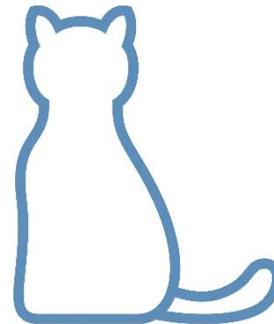
Prediction

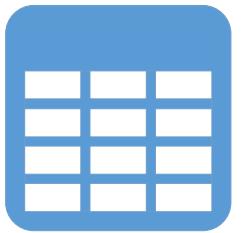


Cat

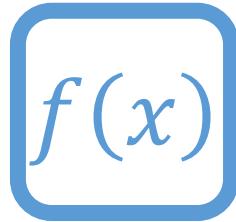
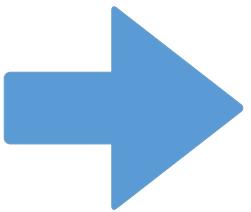


Dog

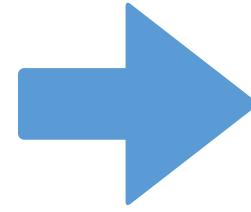




Data



Function



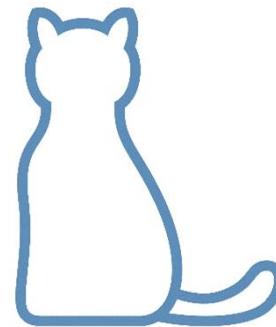
Prediction



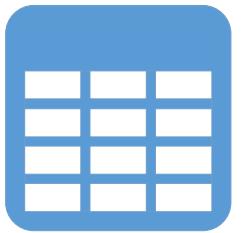
Cat



Dog



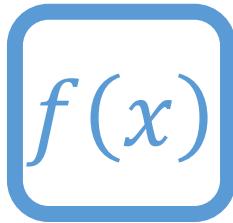
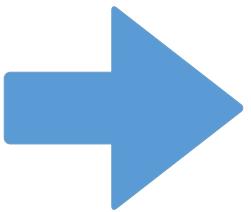
Is cat?



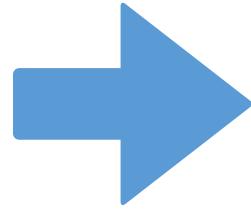
Data



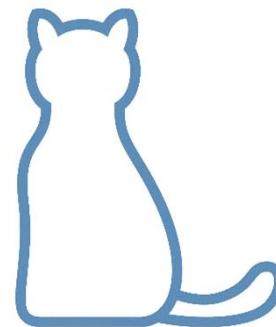
Cat



Function



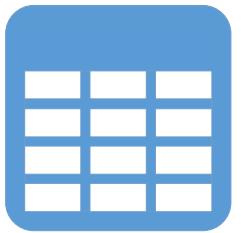
Prediction



Is cat?



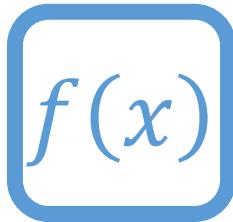
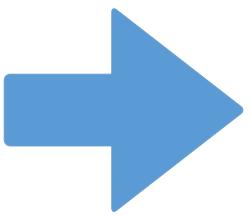
Dog



Data



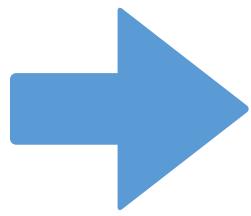
Cat



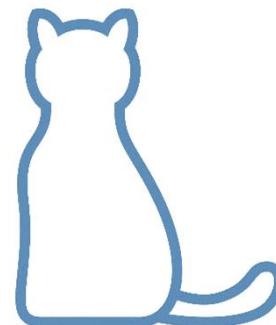
Function



Dog



Prediction



Is cat?



Yes

How does Machine Learning Work?

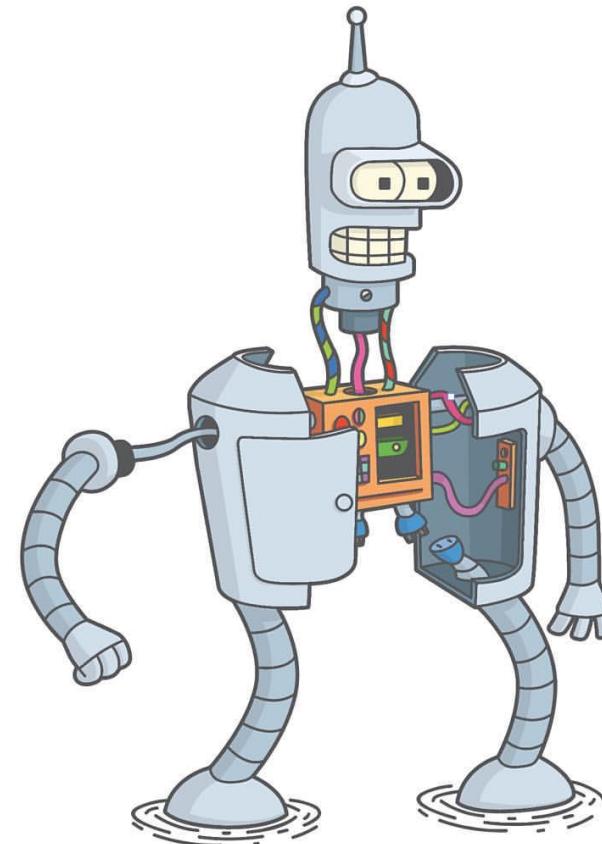
Uses statistical models

Model's parameters are trained

Model trained with algorithm

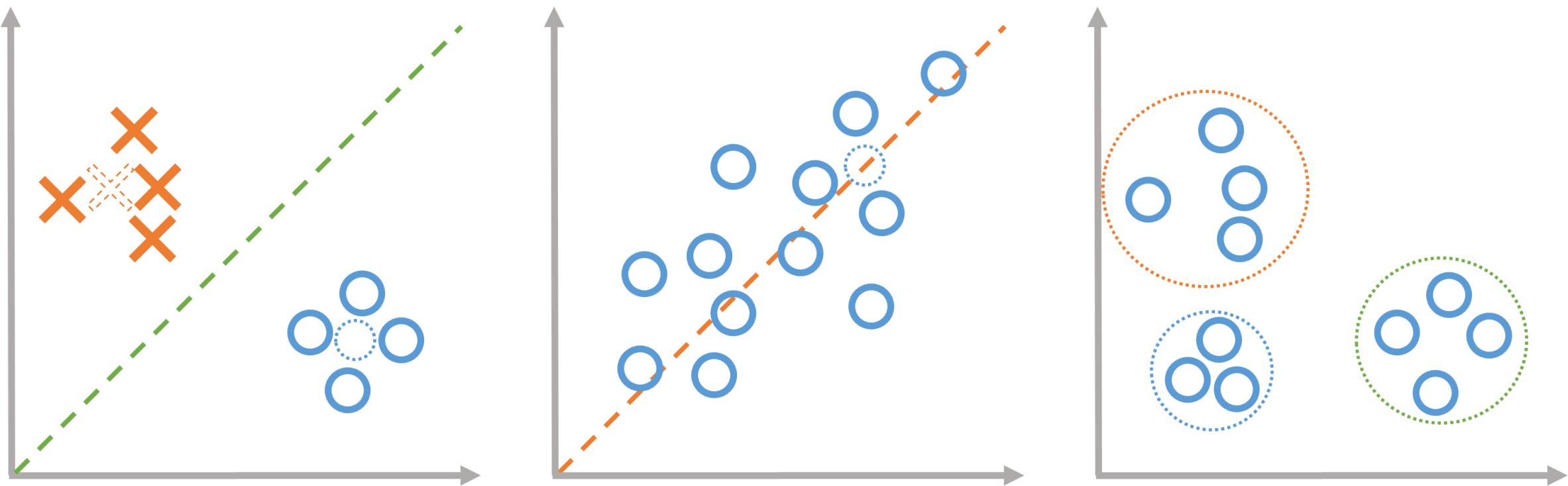
Model is used to predict output

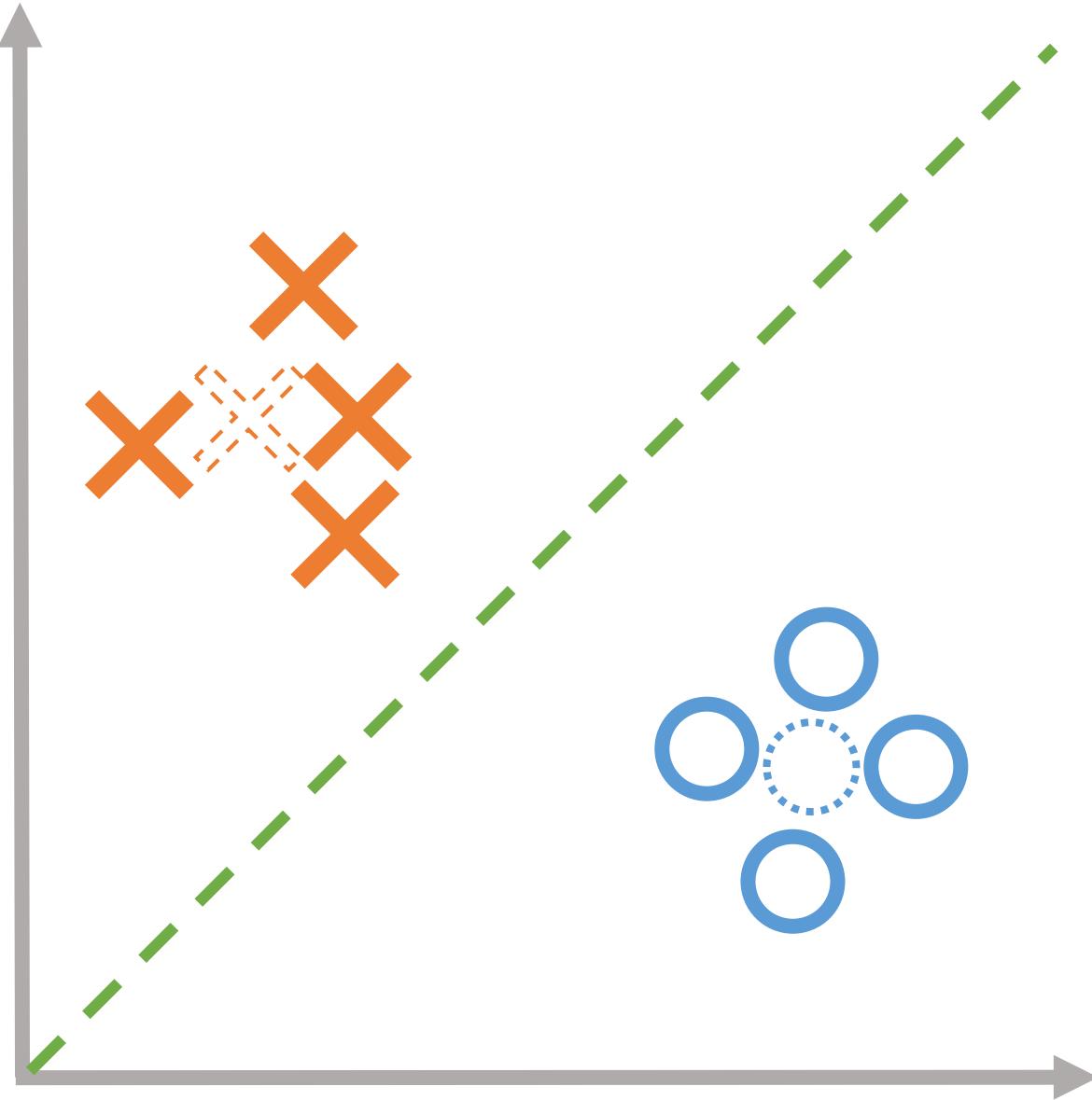
Prediction vs. explanation

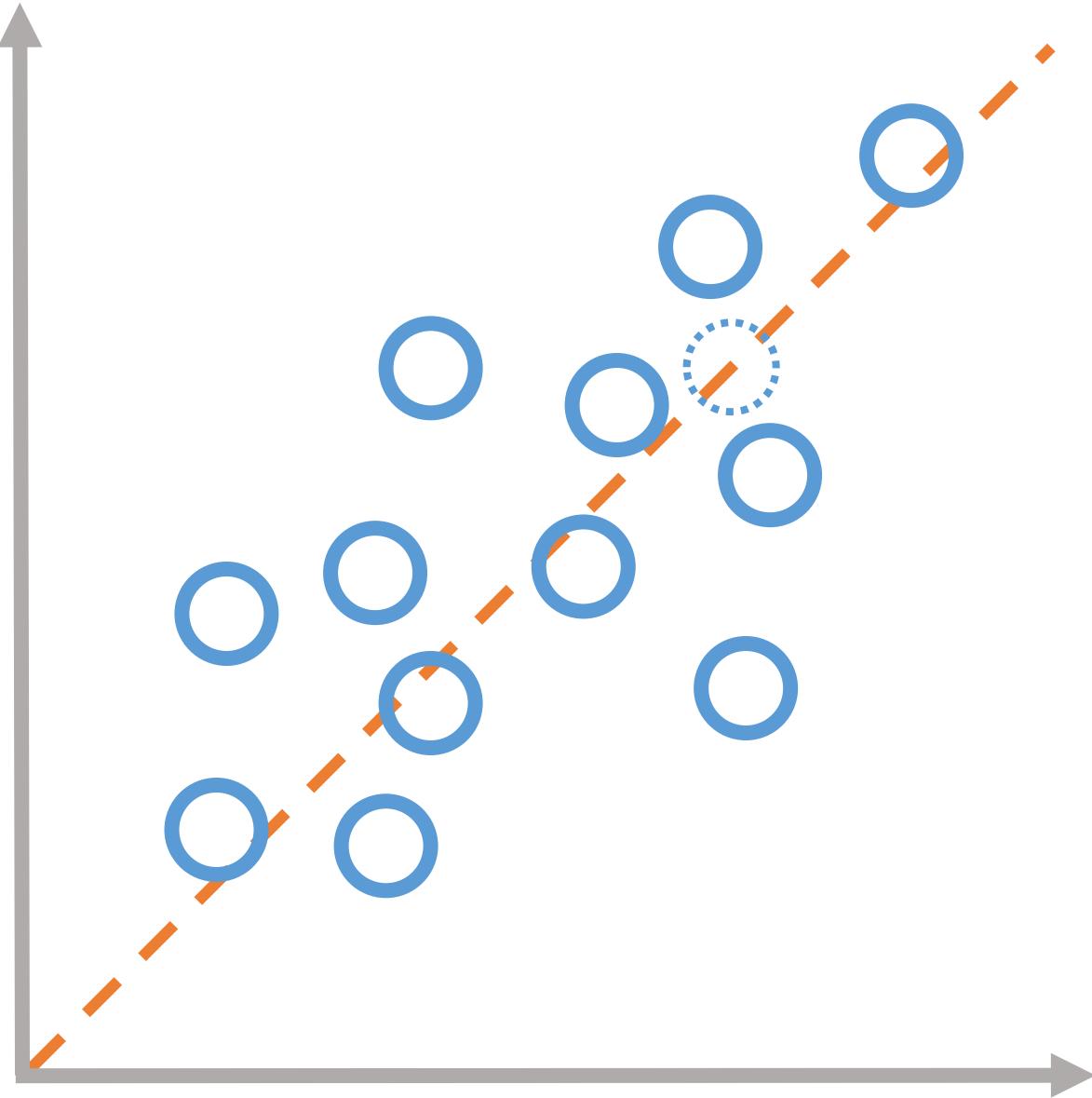


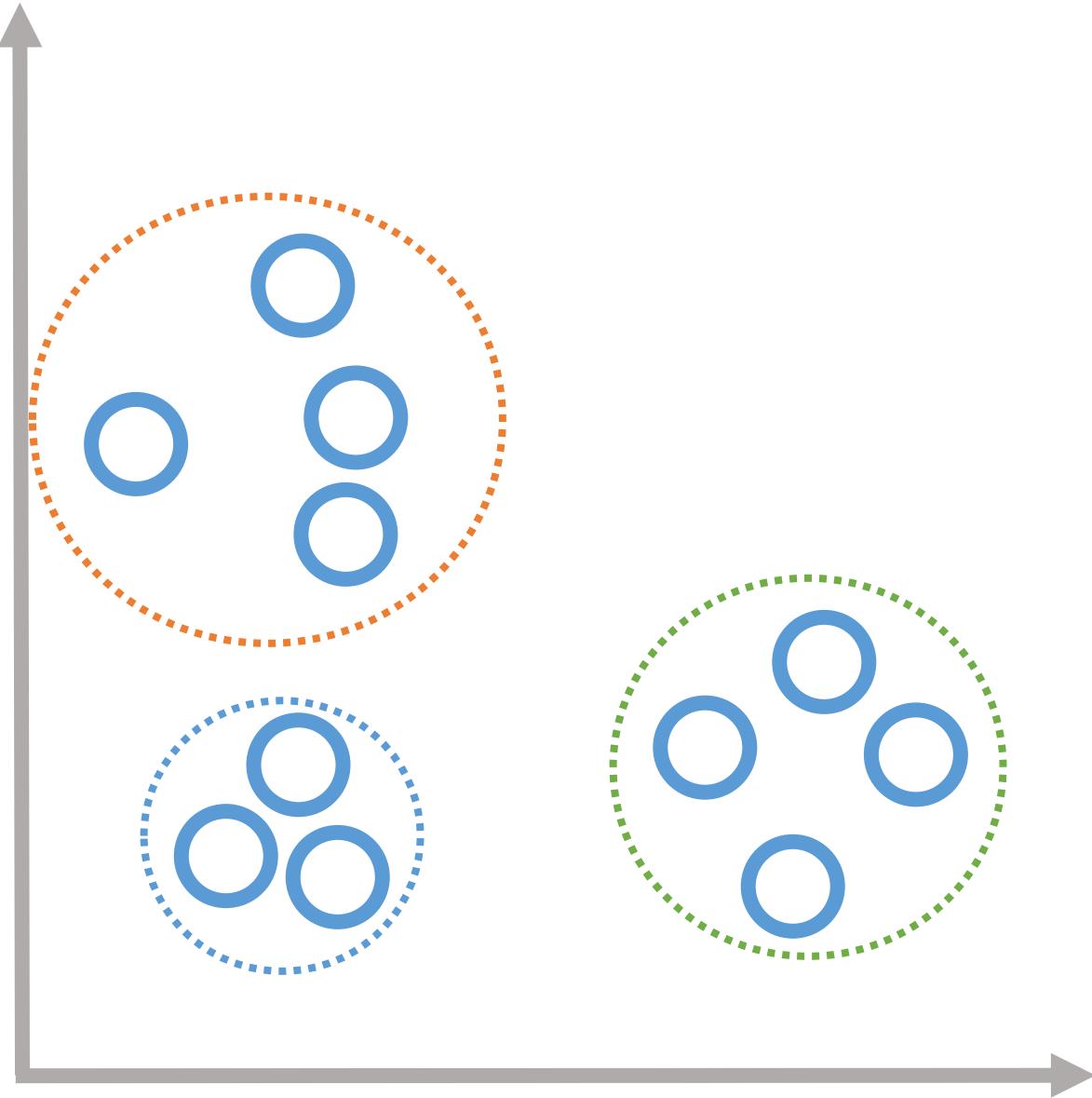
Source: Futurama

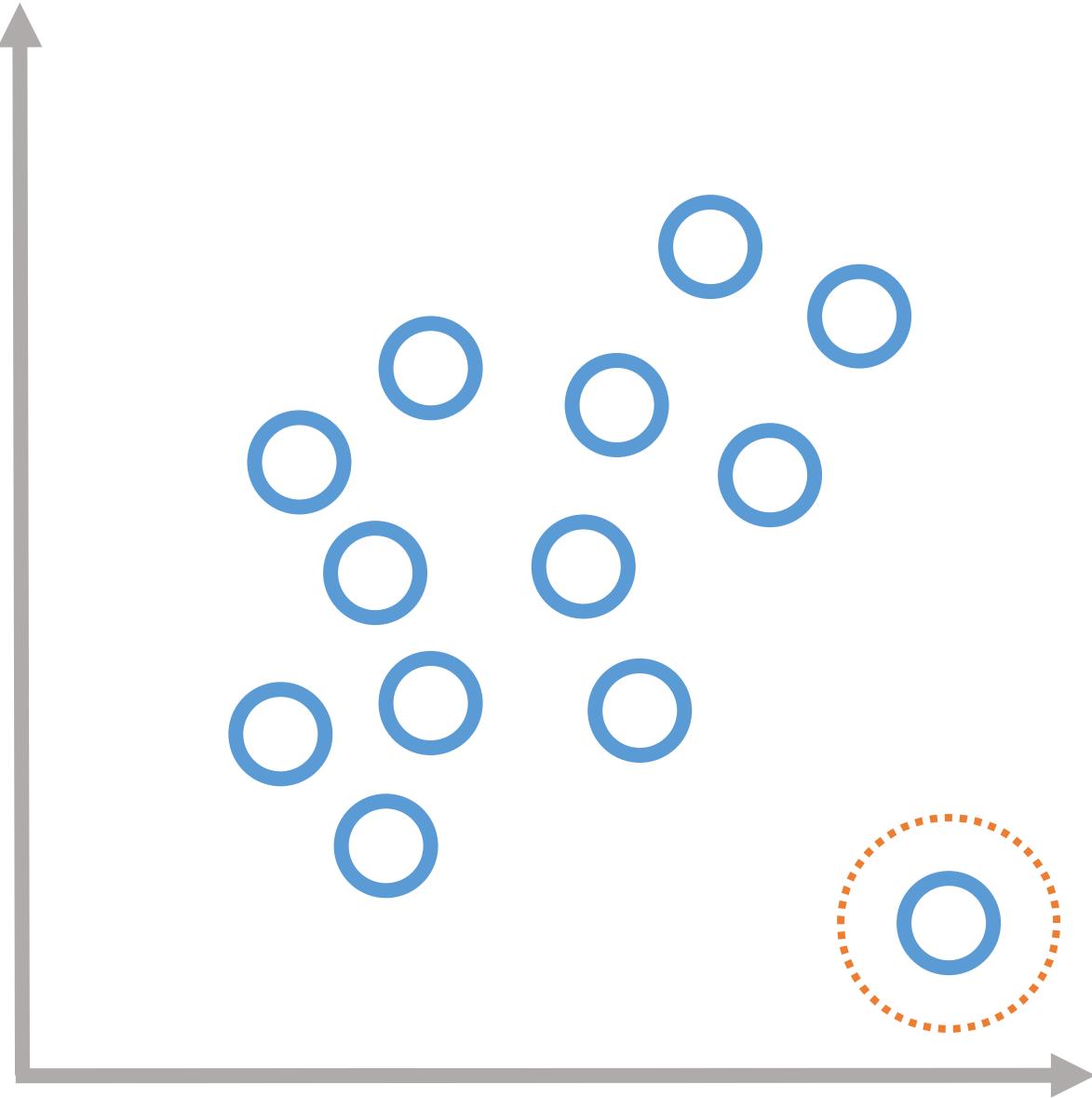
What Can Machine Learning Do?











Types of Machine Learning

Supervised
Unsupervised
Reinforcement



Source: Futurama

Types of ML Algorithms

Decision Trees

Naïve Bayes Classifier

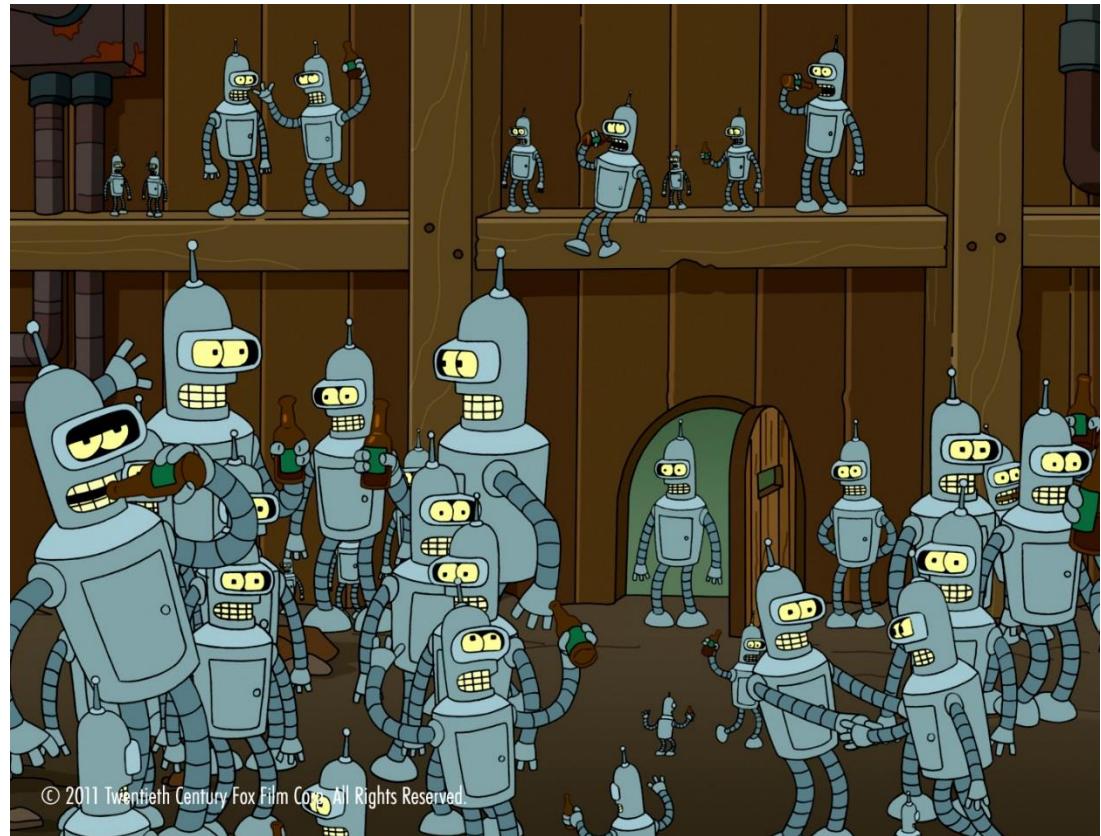
Linear Regression

Support Vector Machines

Neural Networks

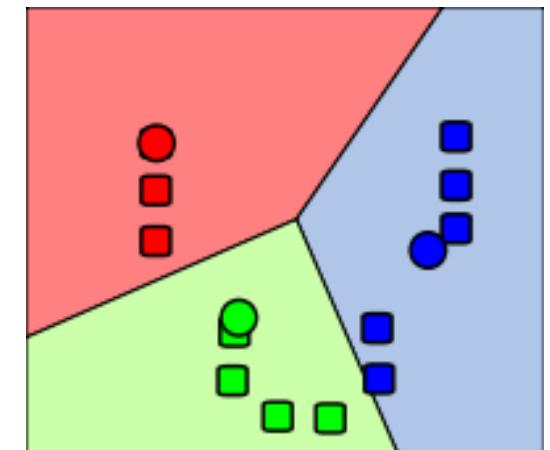
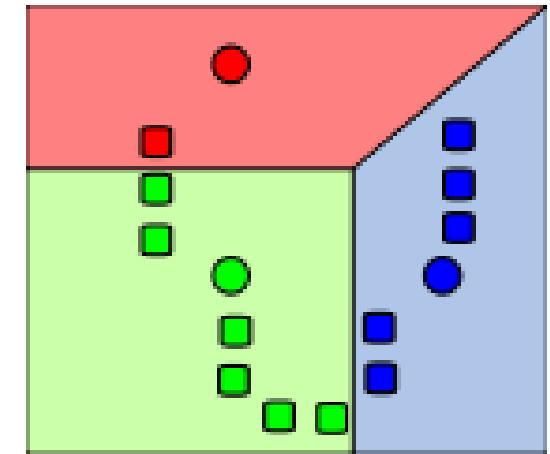
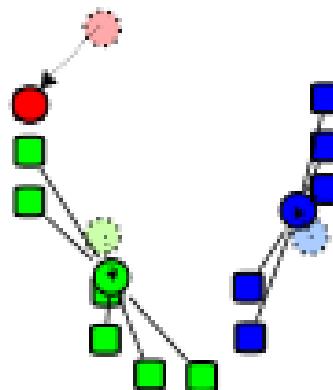
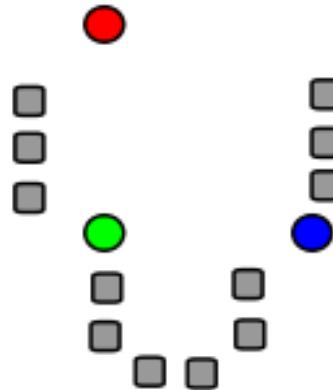
K-Means Clustering

Ensemble Learning



Source: Futurama

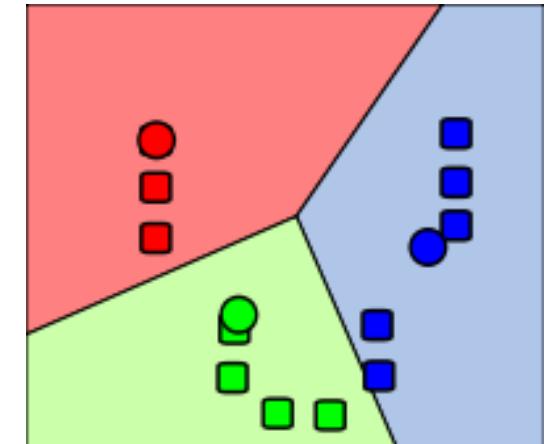
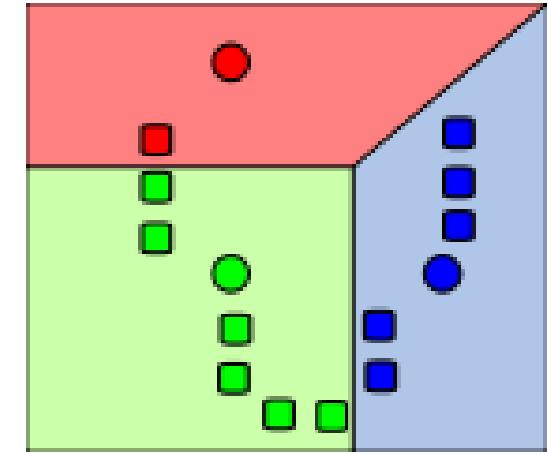
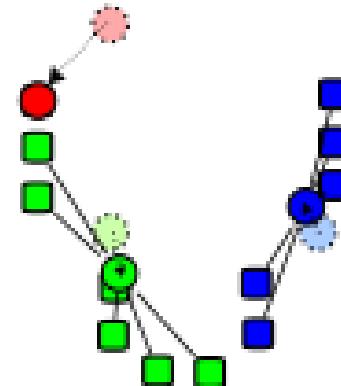
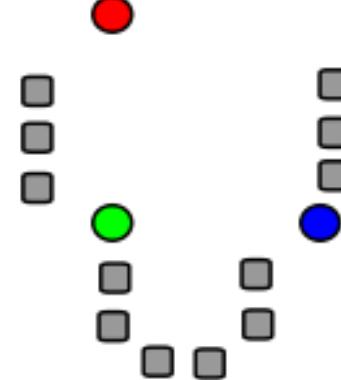
k-Means Clustering



Source: Wikipedia

k-Means Clustering

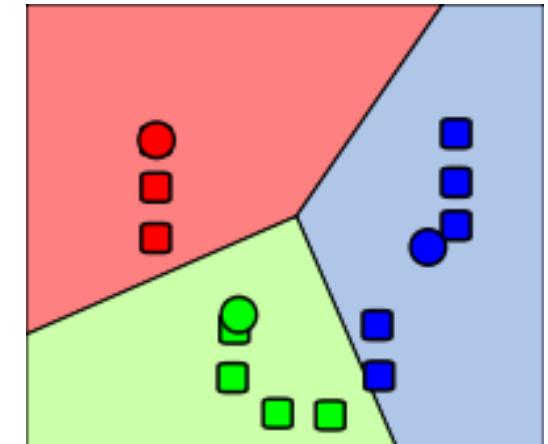
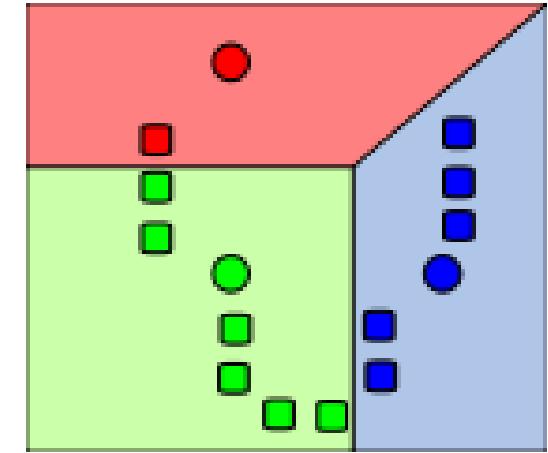
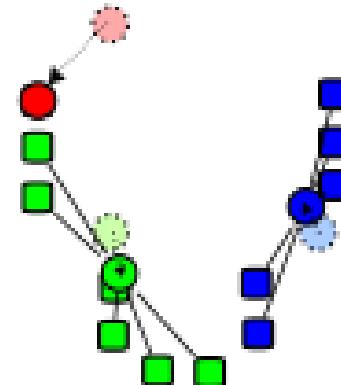
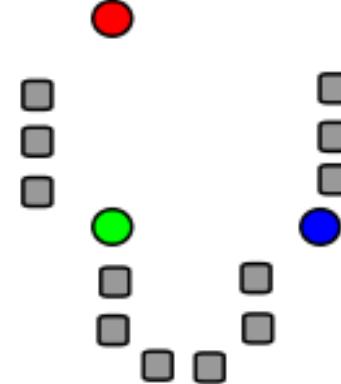
Unsupervised learning



Source: Wikipedia

k-Means Clustering

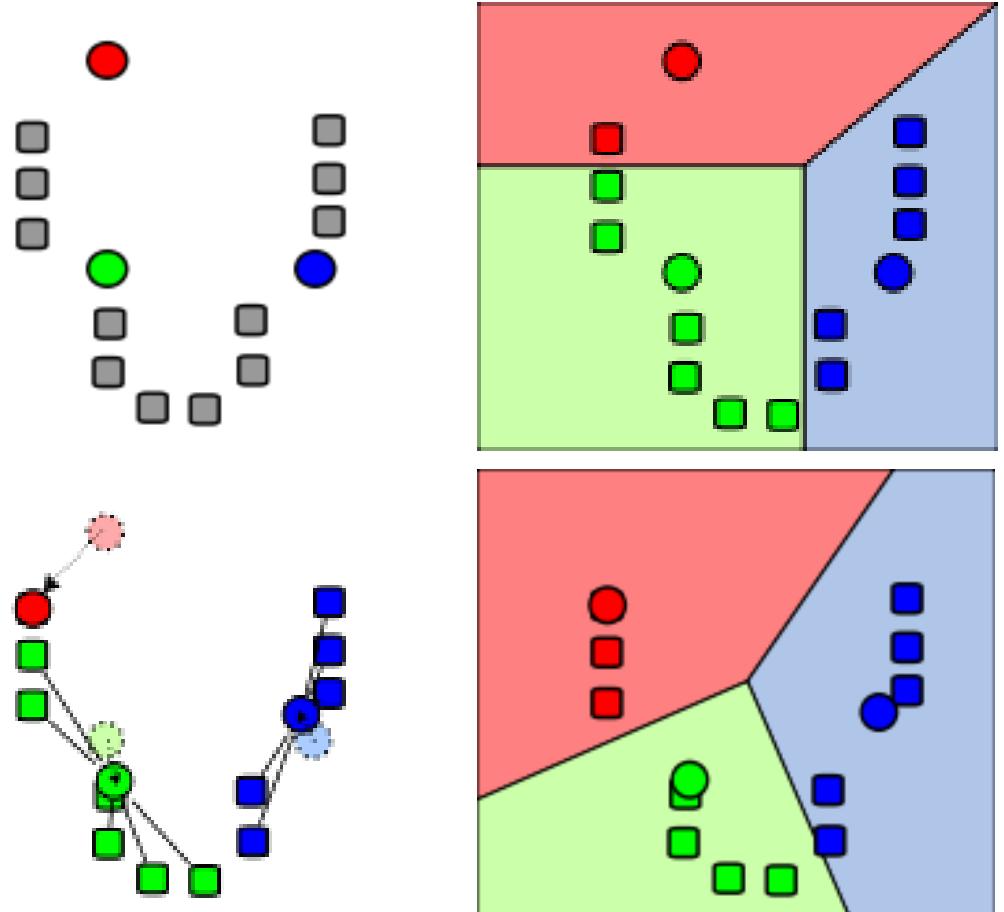
Unsupervised learning
Specify k (# of clusters)



Source: Wikipedia

k-Means Clustering

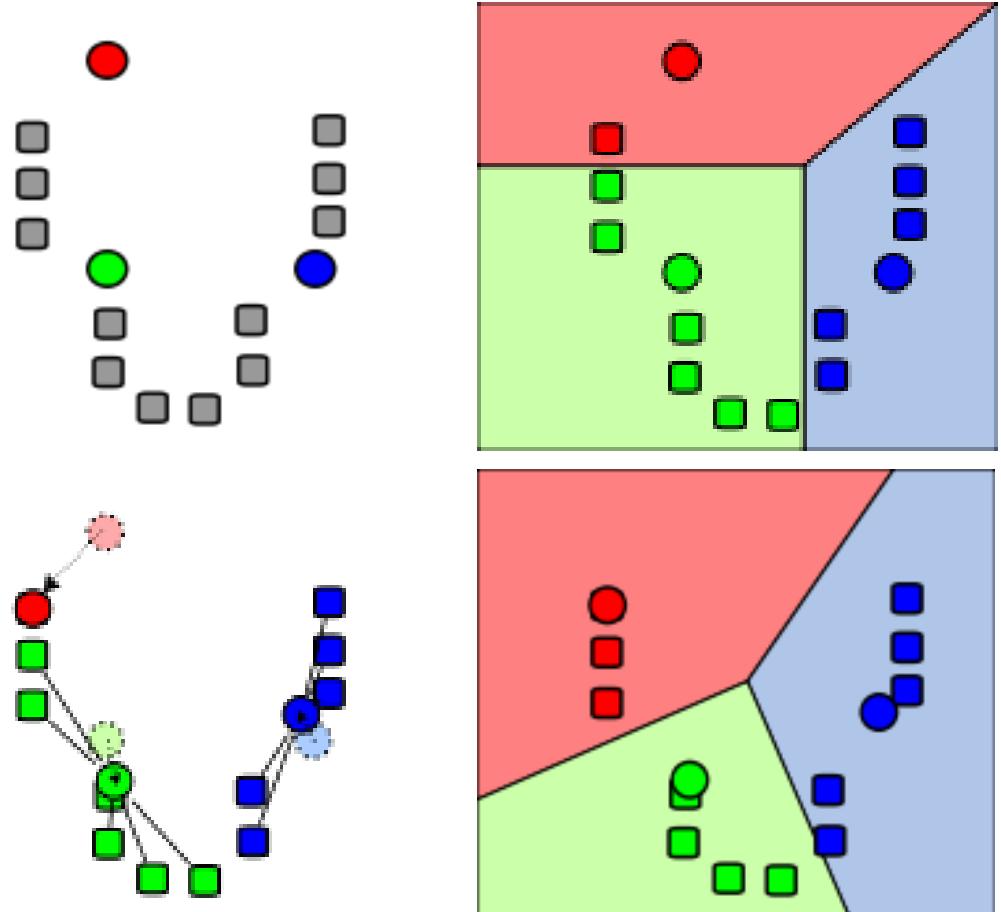
Unsupervised learning
Specify k (# of clusters)
Algorithm finds centers



Source: Wikipedia

k-Means Clustering

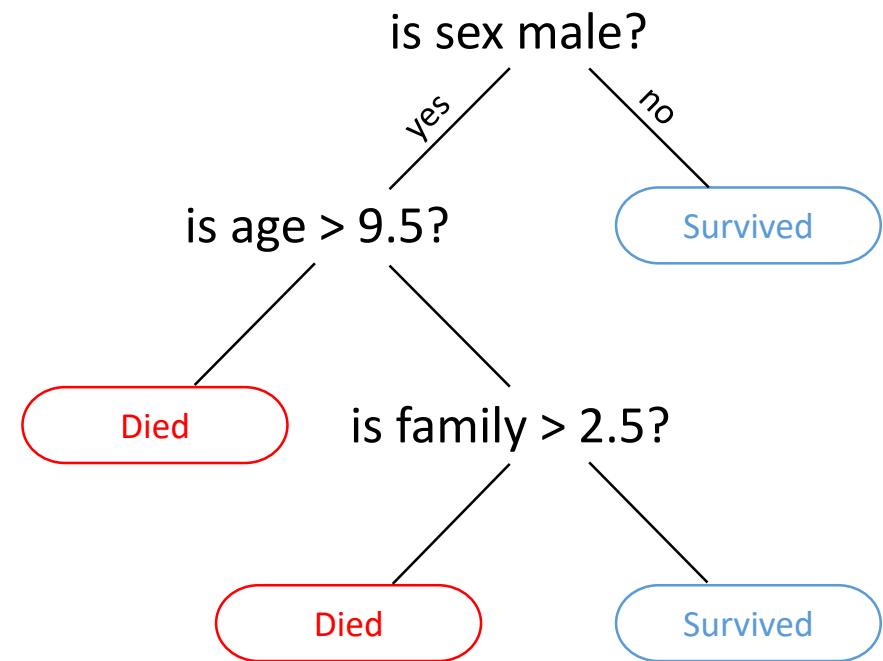
Unsupervised learning
Specify k (# of clusters)
Algorithm finds centers
Random restarts



Source: Wikipedia

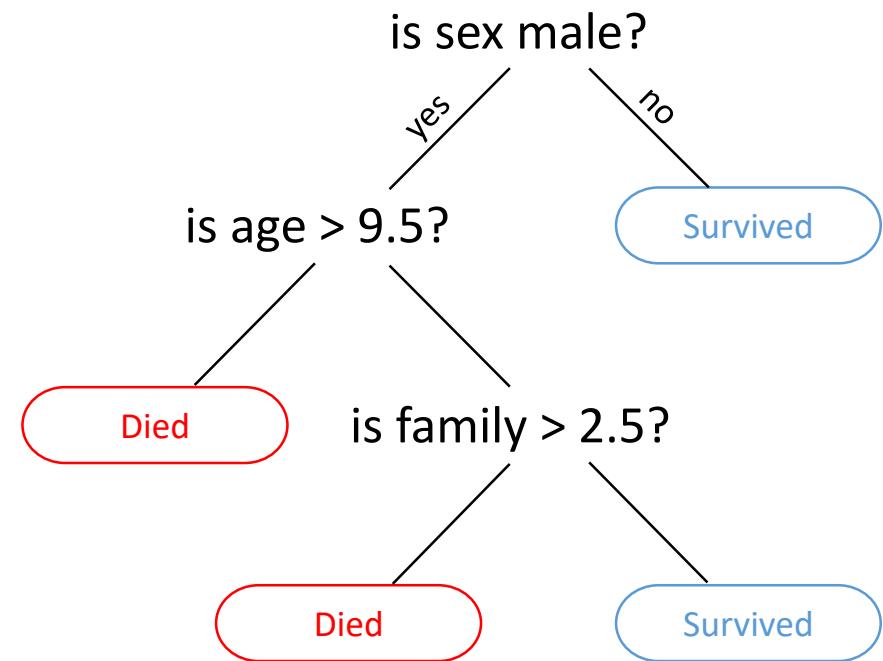
Decision Tree Classifier

Supervised learning



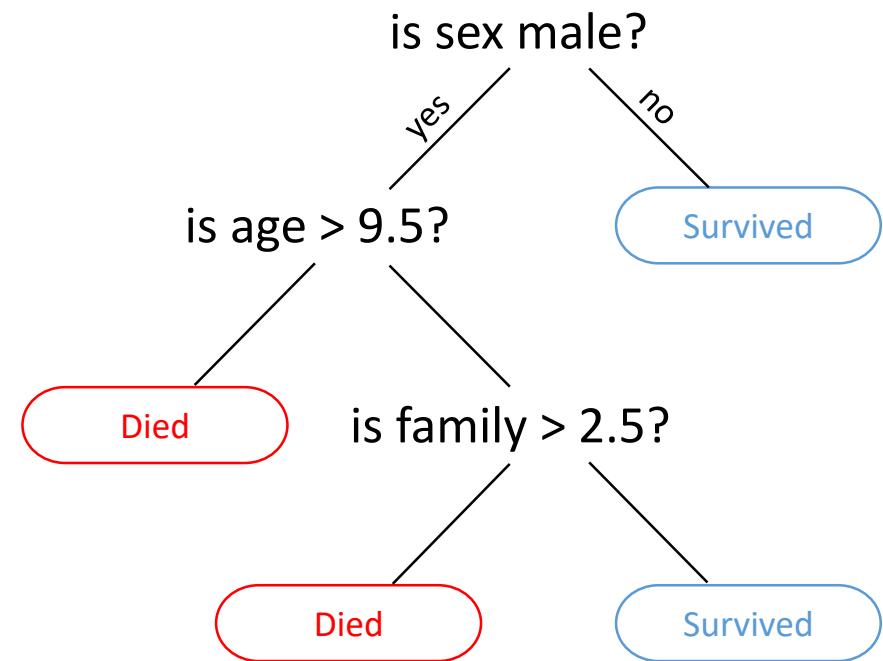
Decision Tree Classifier

Supervised learning
Tree of decisions



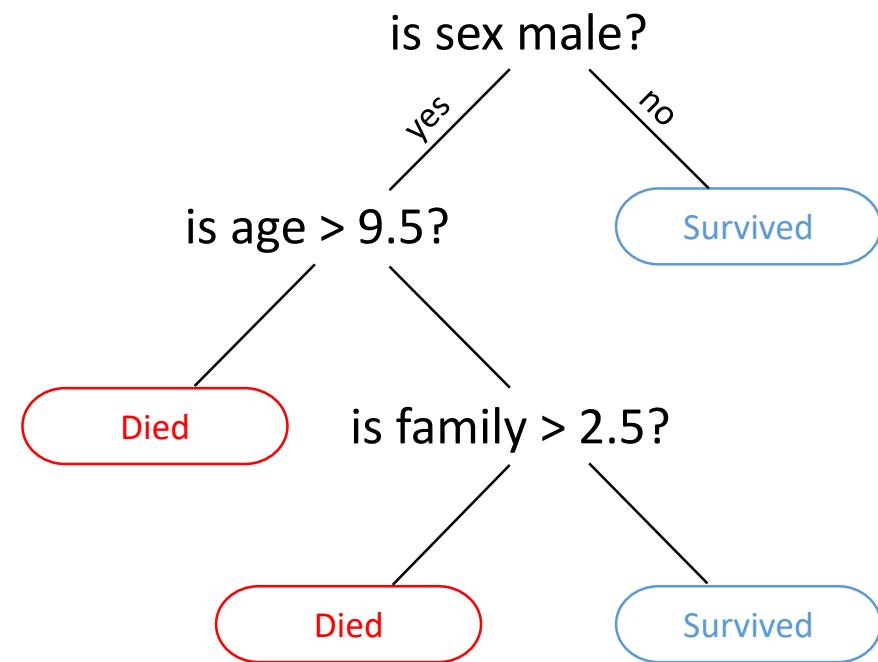
Decision Tree Classifier

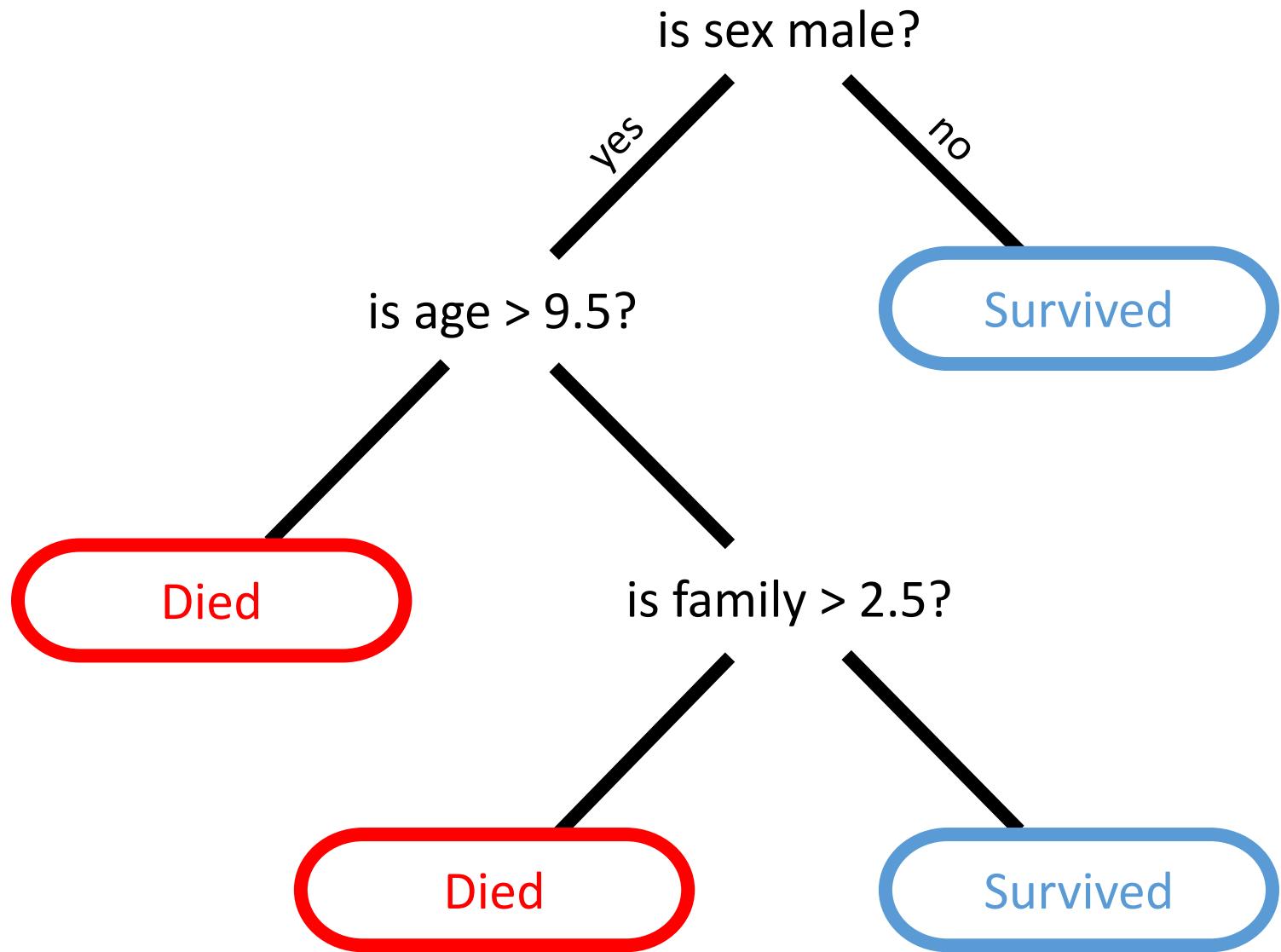
Supervised learning
Tree of decisions
Easy to understand



Decision Tree Classifier

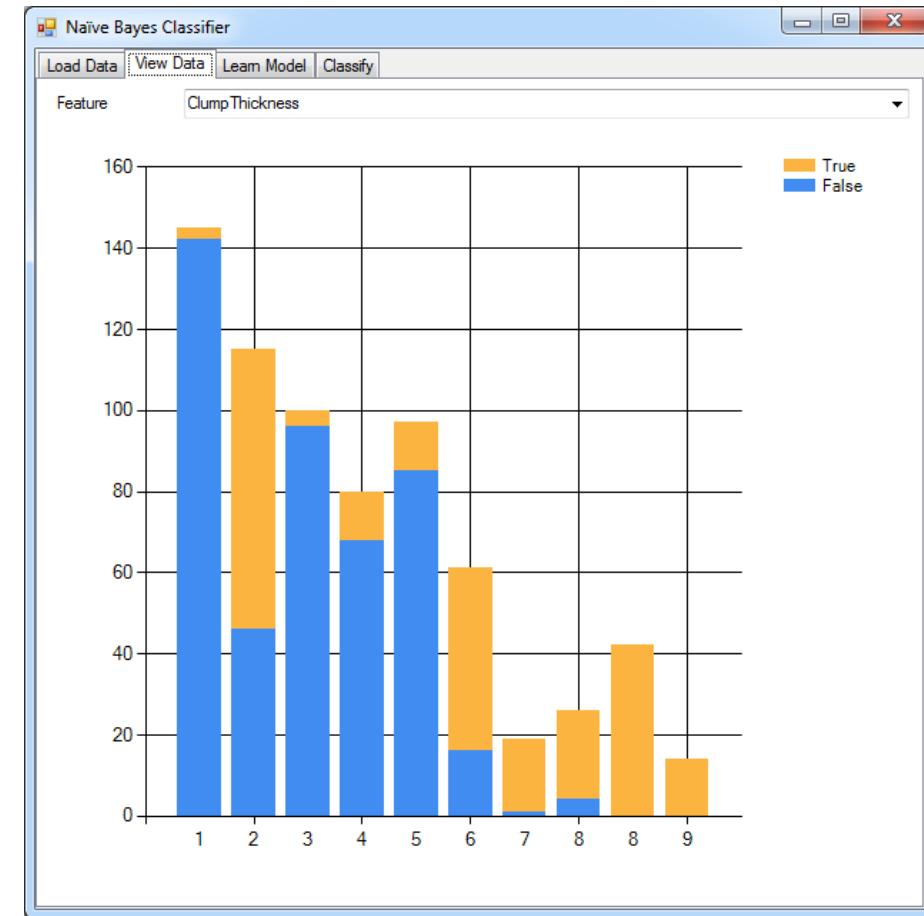
Supervised learning
Tree of decisions
Easy to understand
Transparent





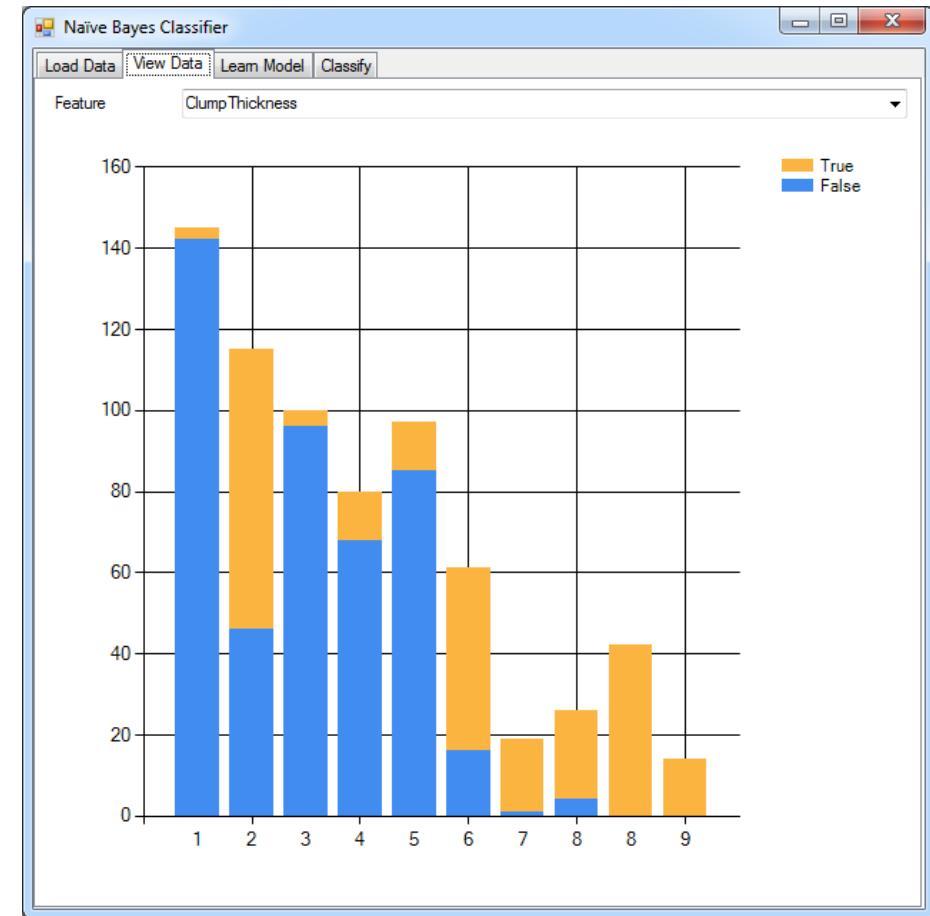
Naïve Bayes Classifier

Supervised learning



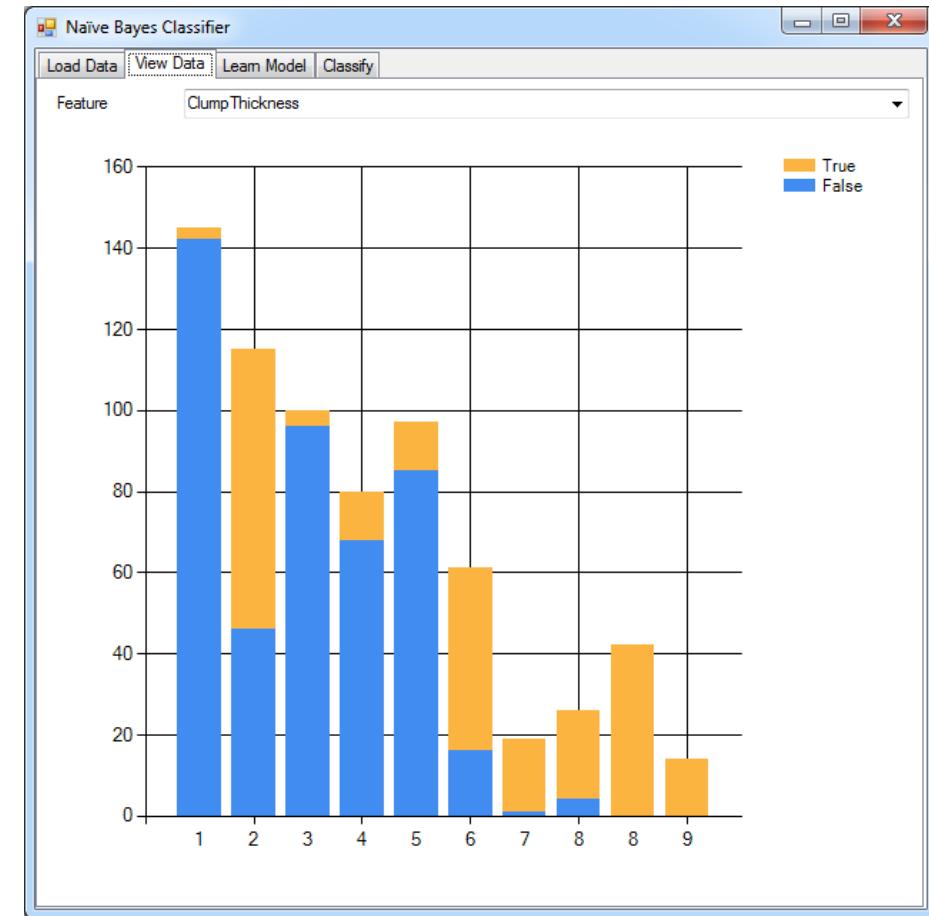
Naïve Bayes Classifier

Supervised learning
Bayesian statistics



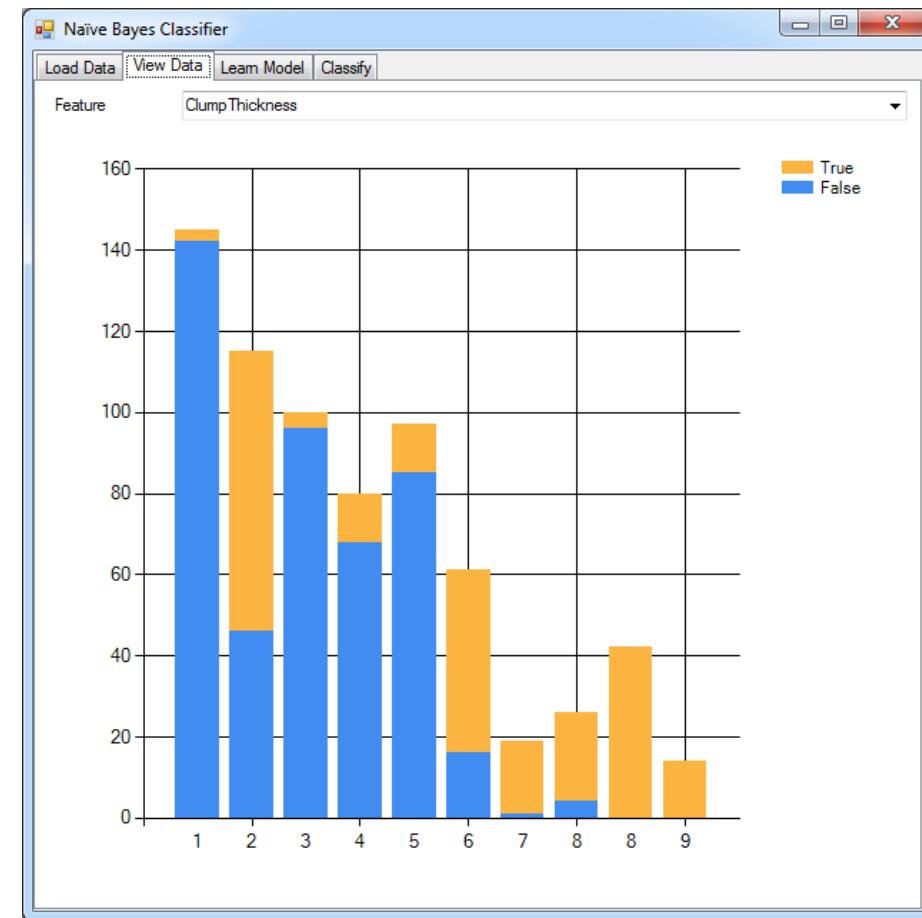
Naïve Bayes Classifier

Supervised learning
Bayesian statistics
Independence assumption



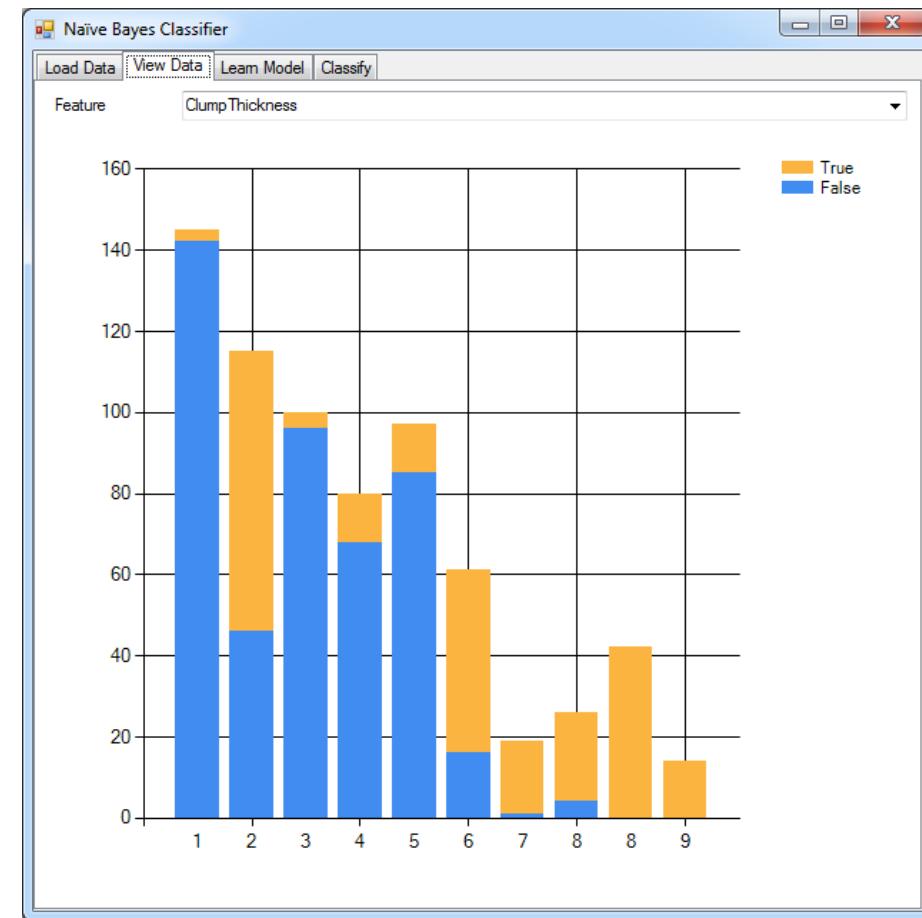
Naïve Bayes Classifier

Supervised learning
Bayesian statistics
Independence assumption
Relatively easy to understand



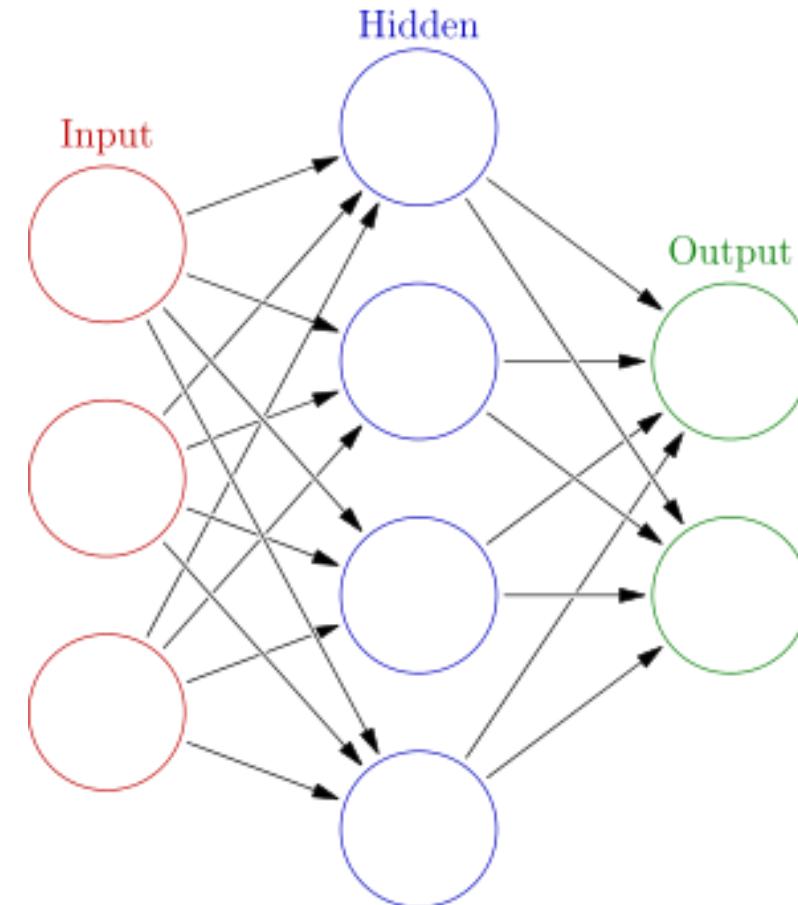
Naïve Bayes Classifier

Supervised learning
Bayesian statistics
Independence assumption
Relatively easy to understand
Transparent



Neural Network Classifier

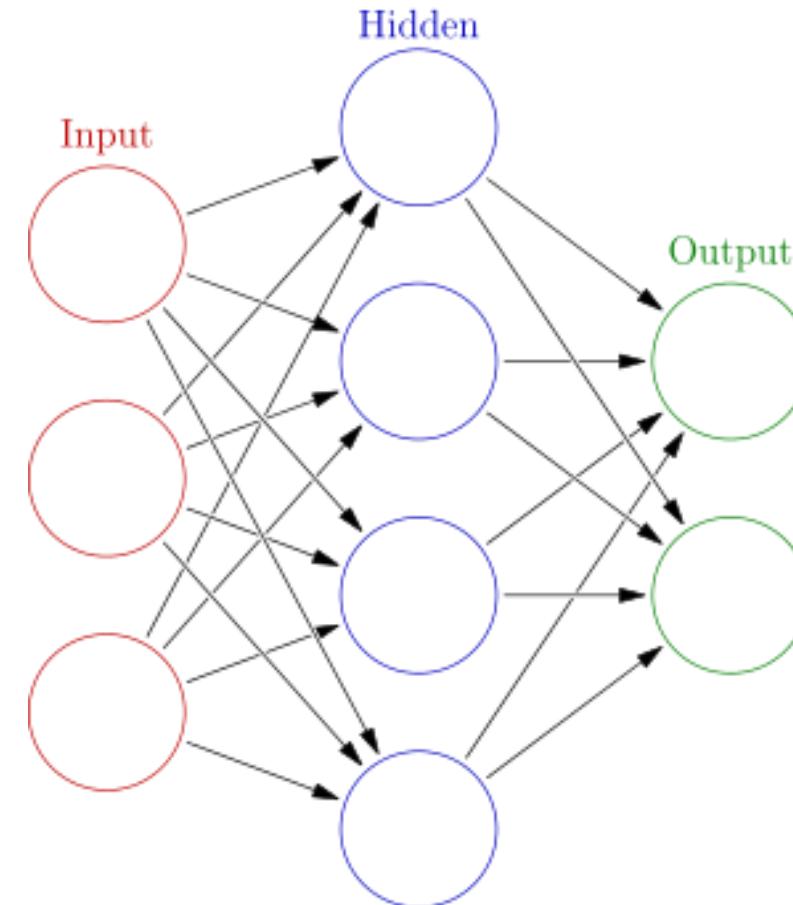
Supervised learning



Source: Wikipedia

Neural Network Classifier

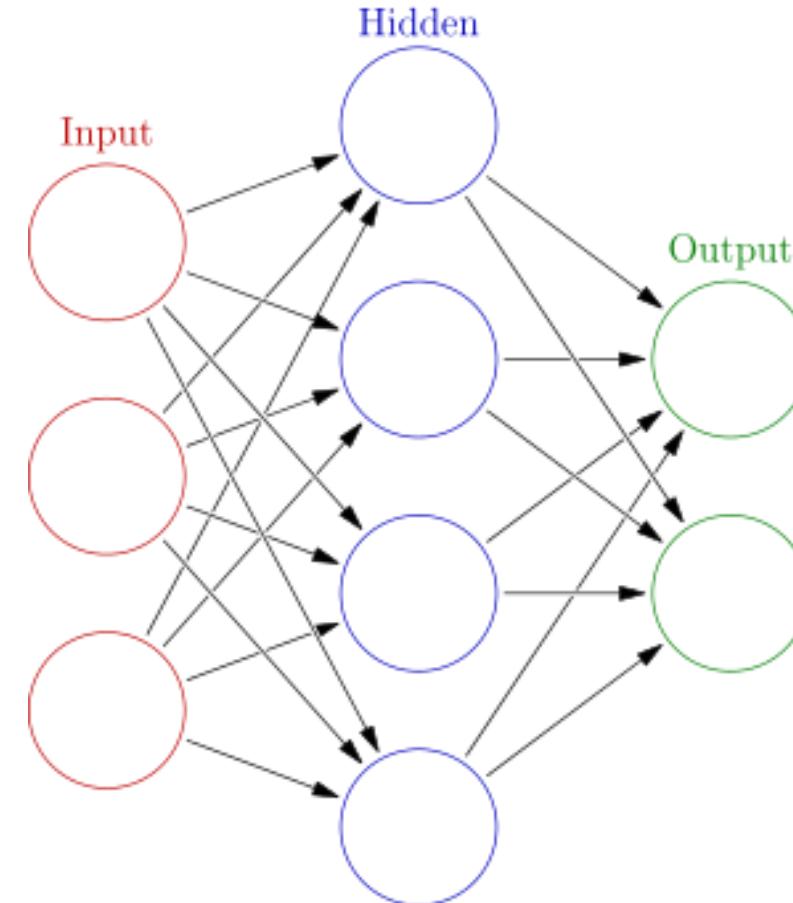
Supervised learning
Neurons in a brain



Source: Wikipedia

Neural Network Classifier

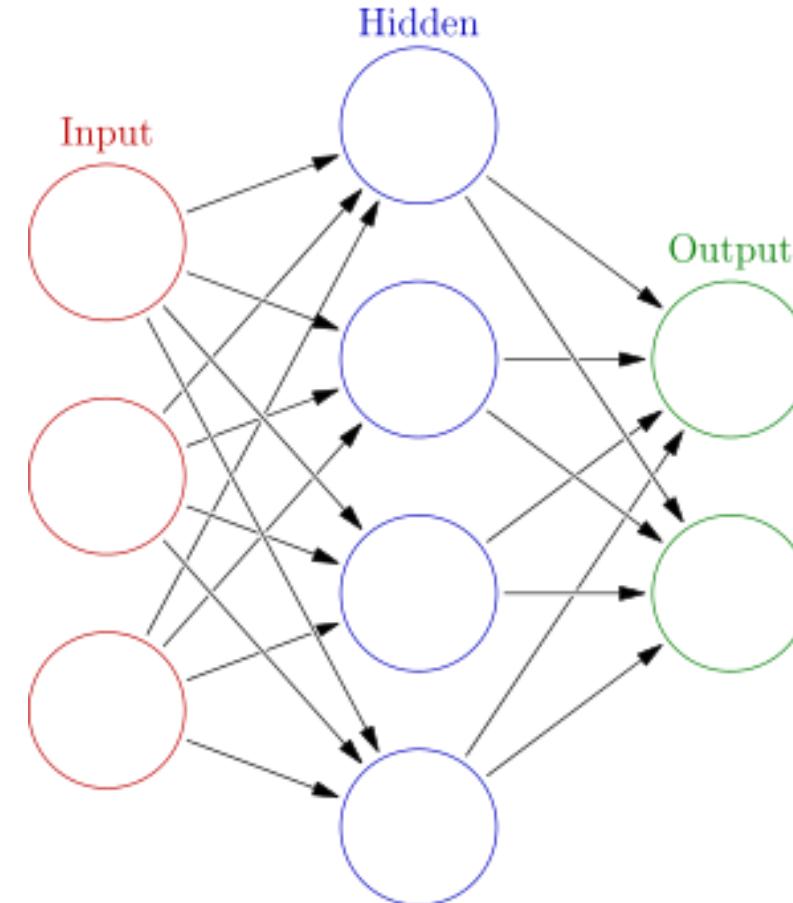
Supervised learning
Neurons in a brain
Complex



Source: Wikipedia

Neural Network Classifier

Supervised learning
Neurons in a brain
Complex
Not transparent



Source: Wikipedia

Find a
questio
n

Find a
questio
n

Prepare
the data

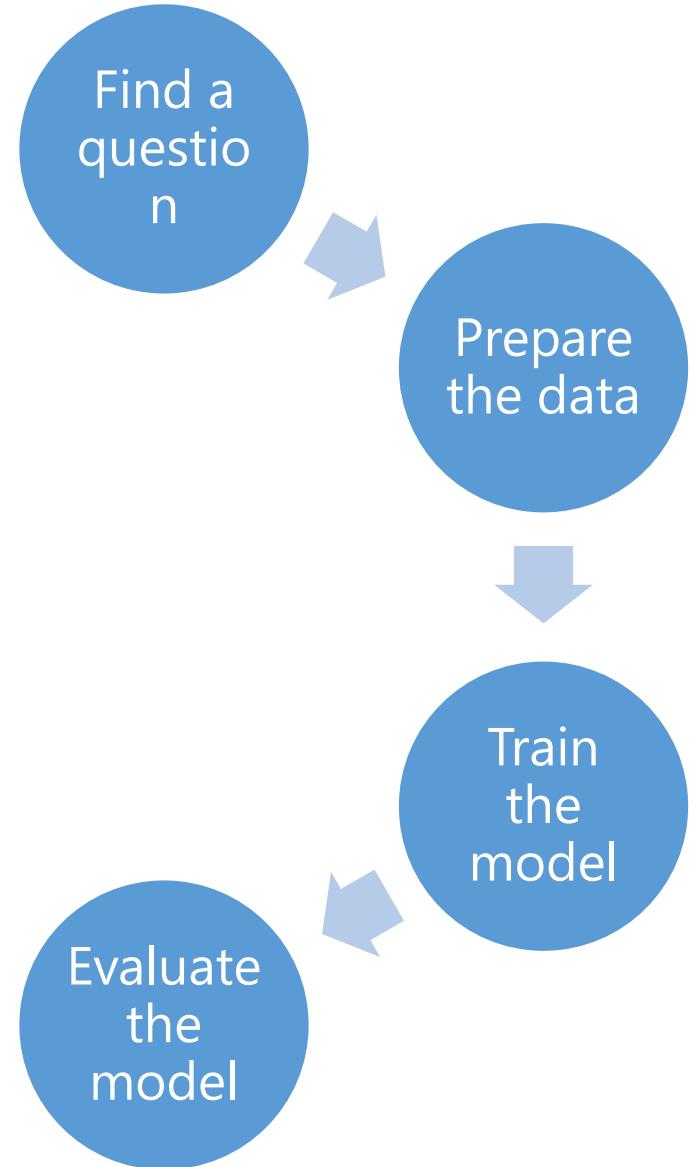


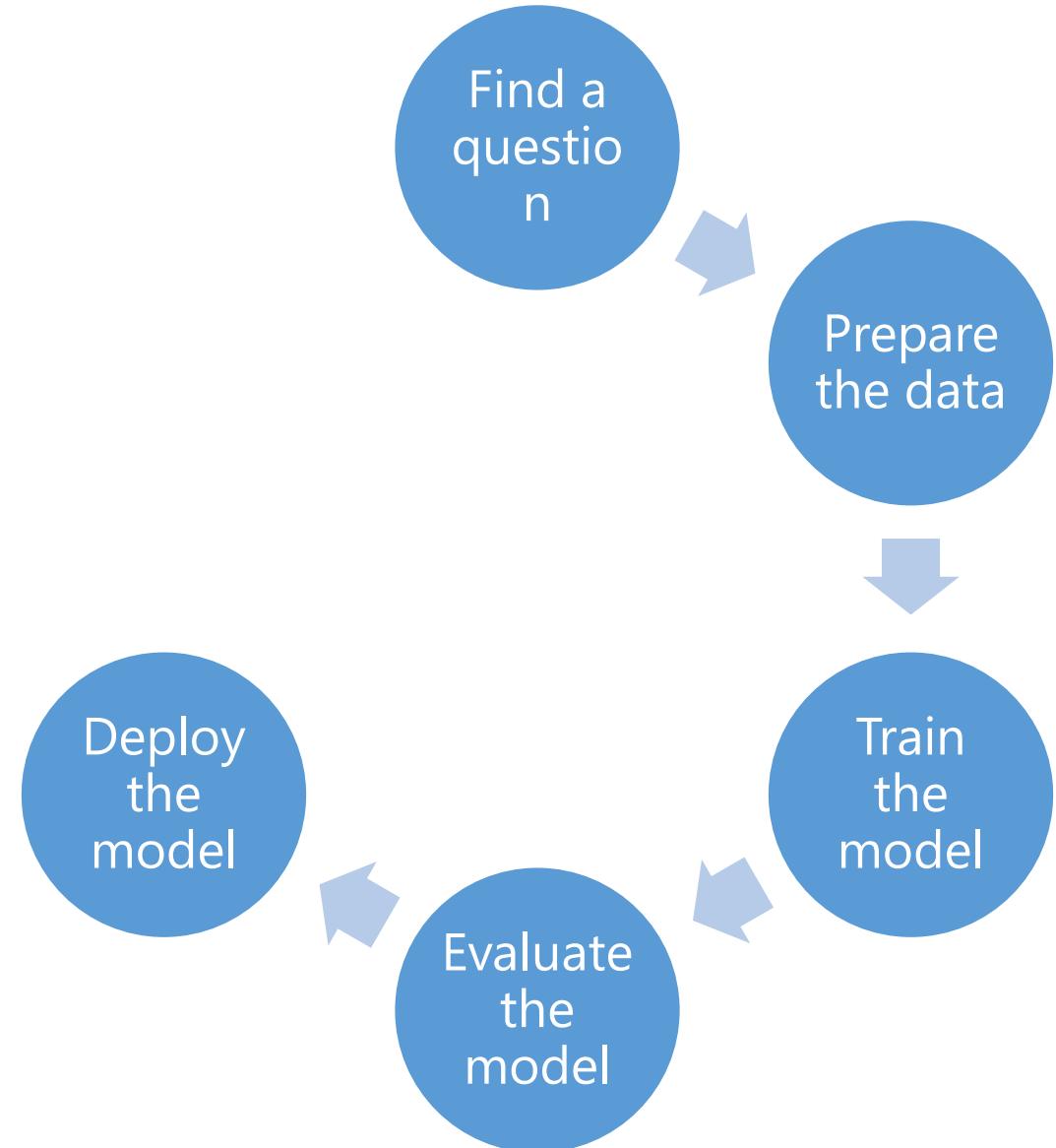
Find a
questio
n

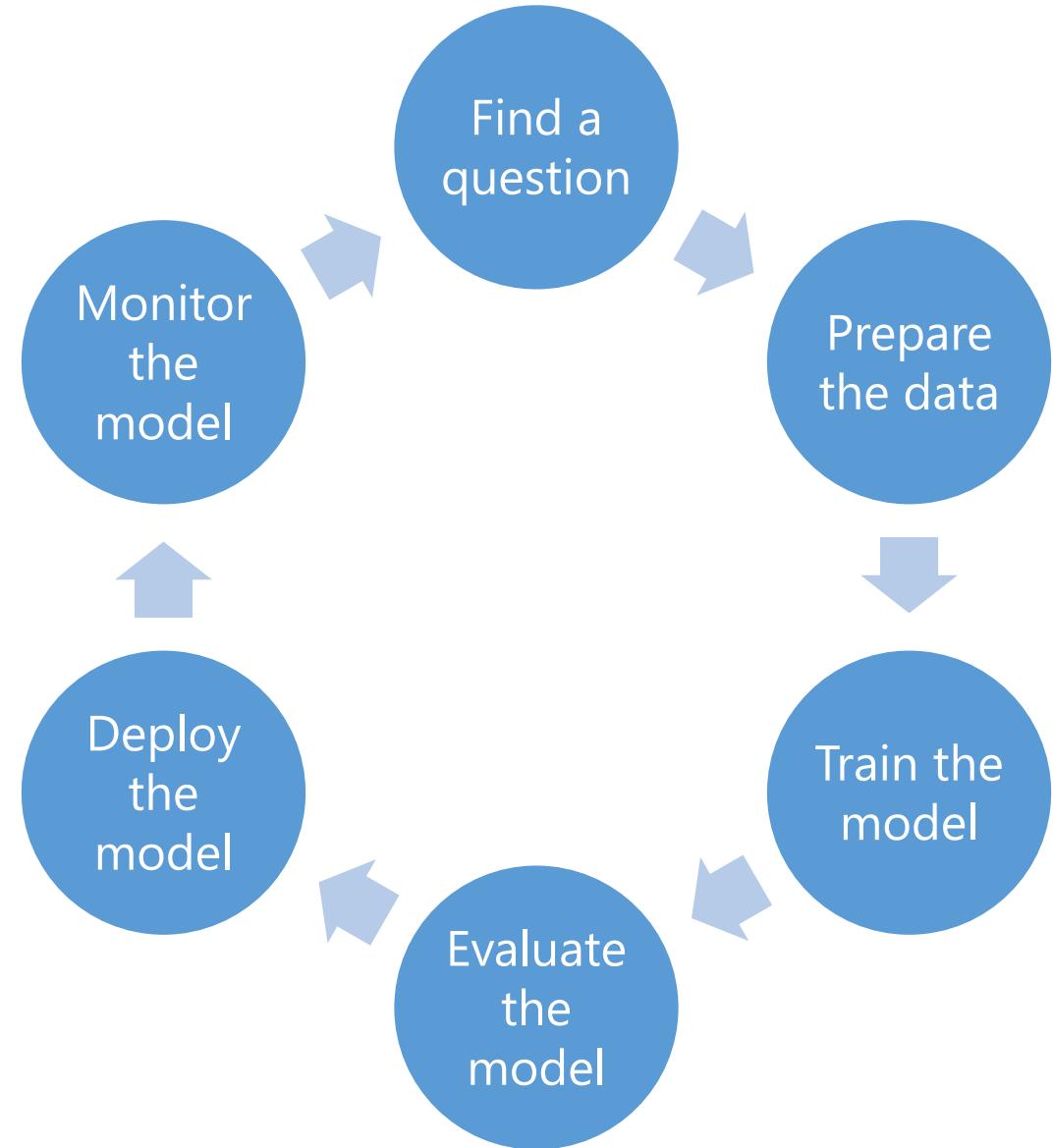
Prepare
the data

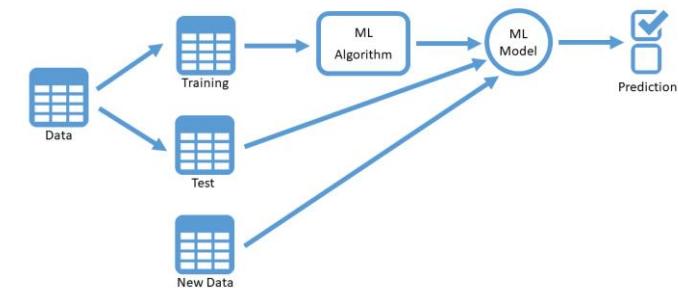
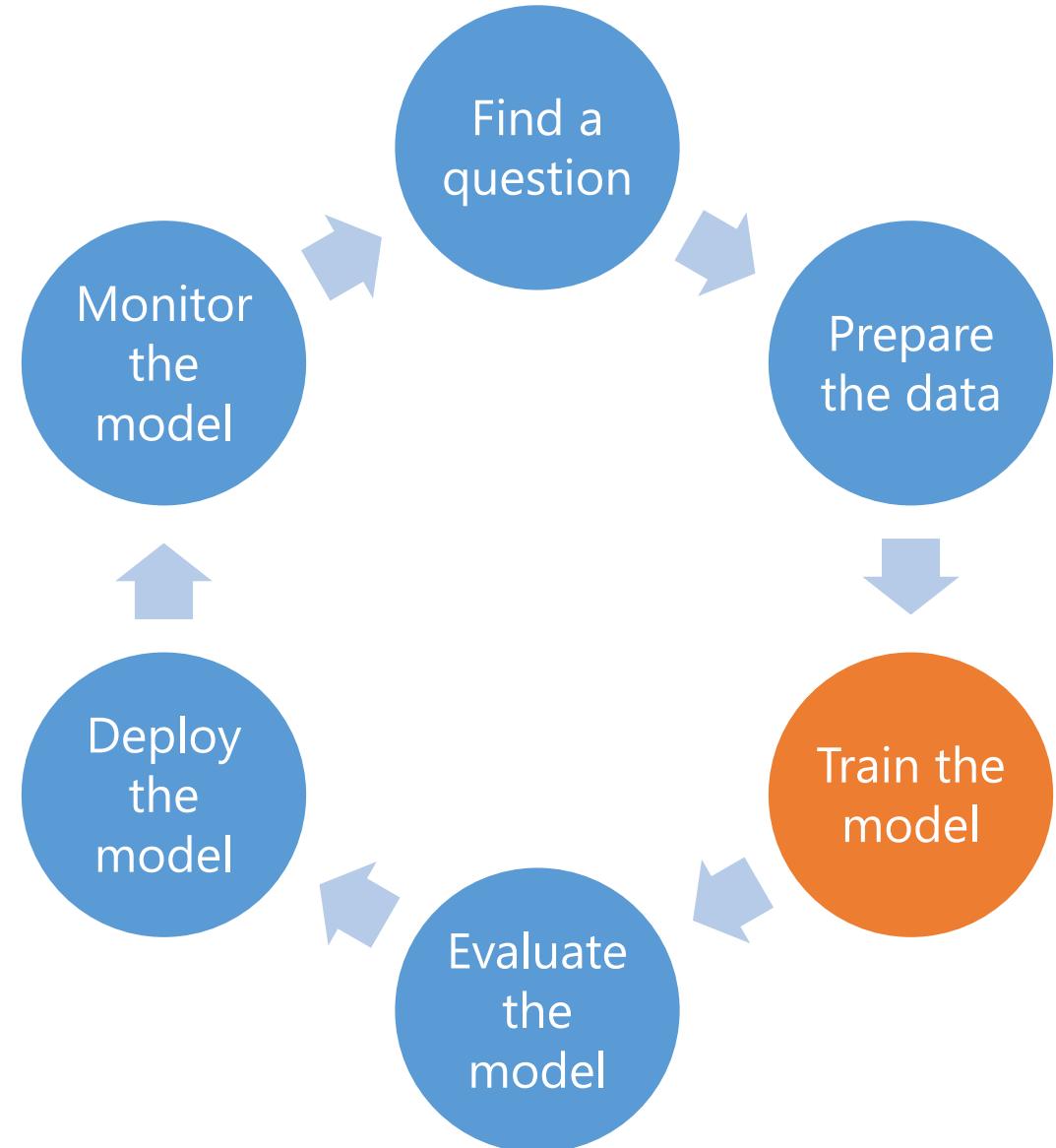
Train
the
model

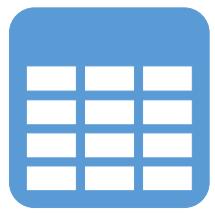




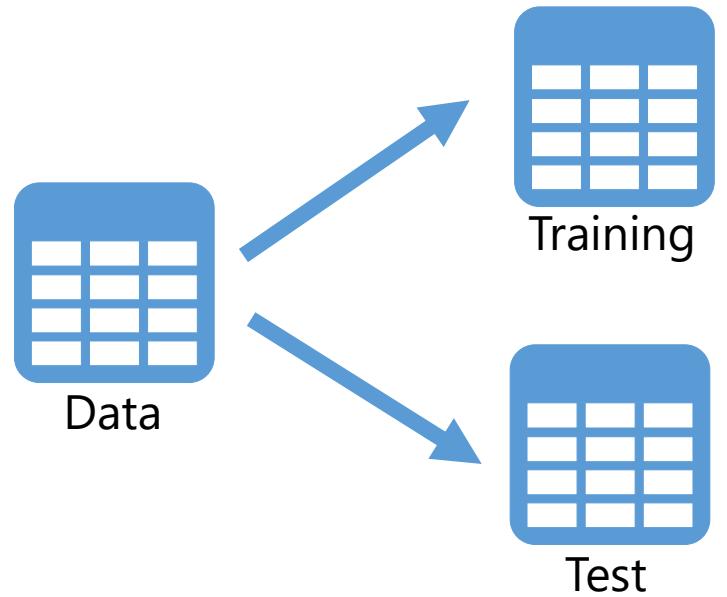


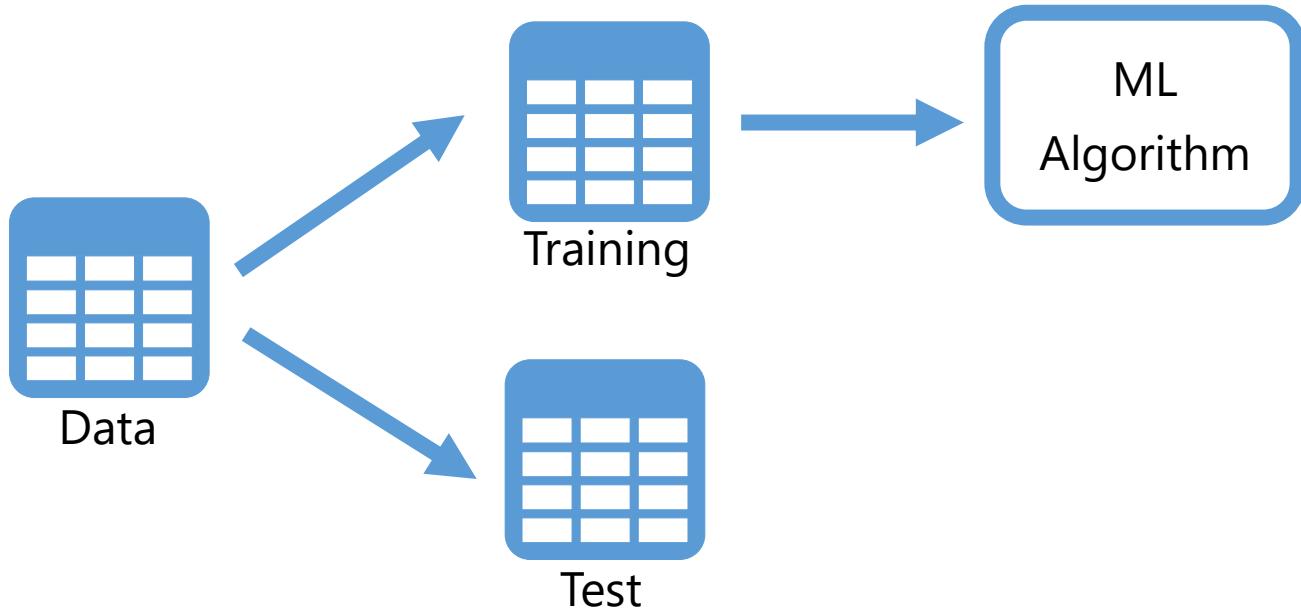


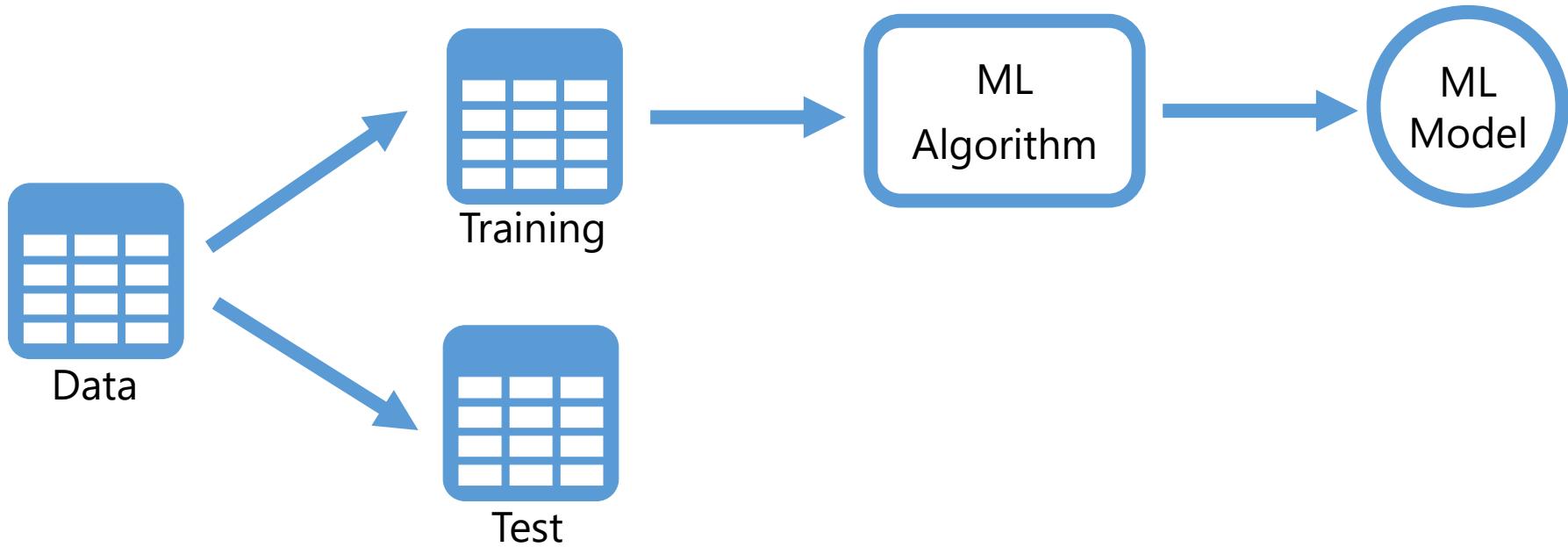


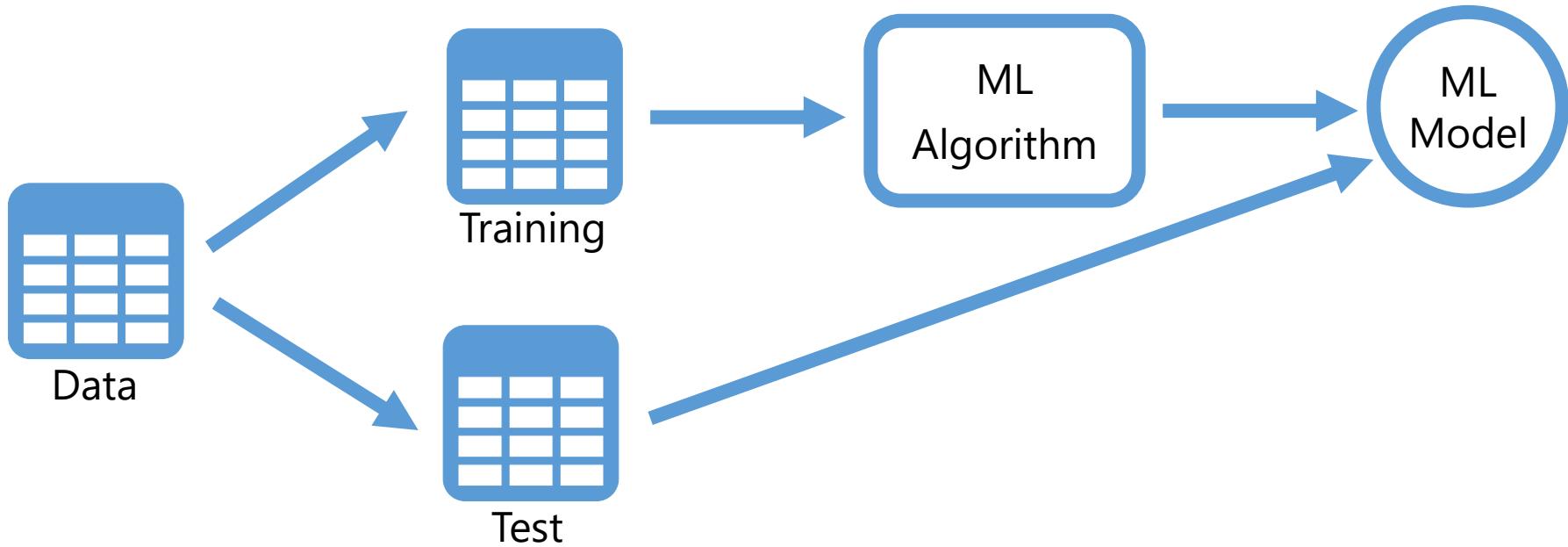


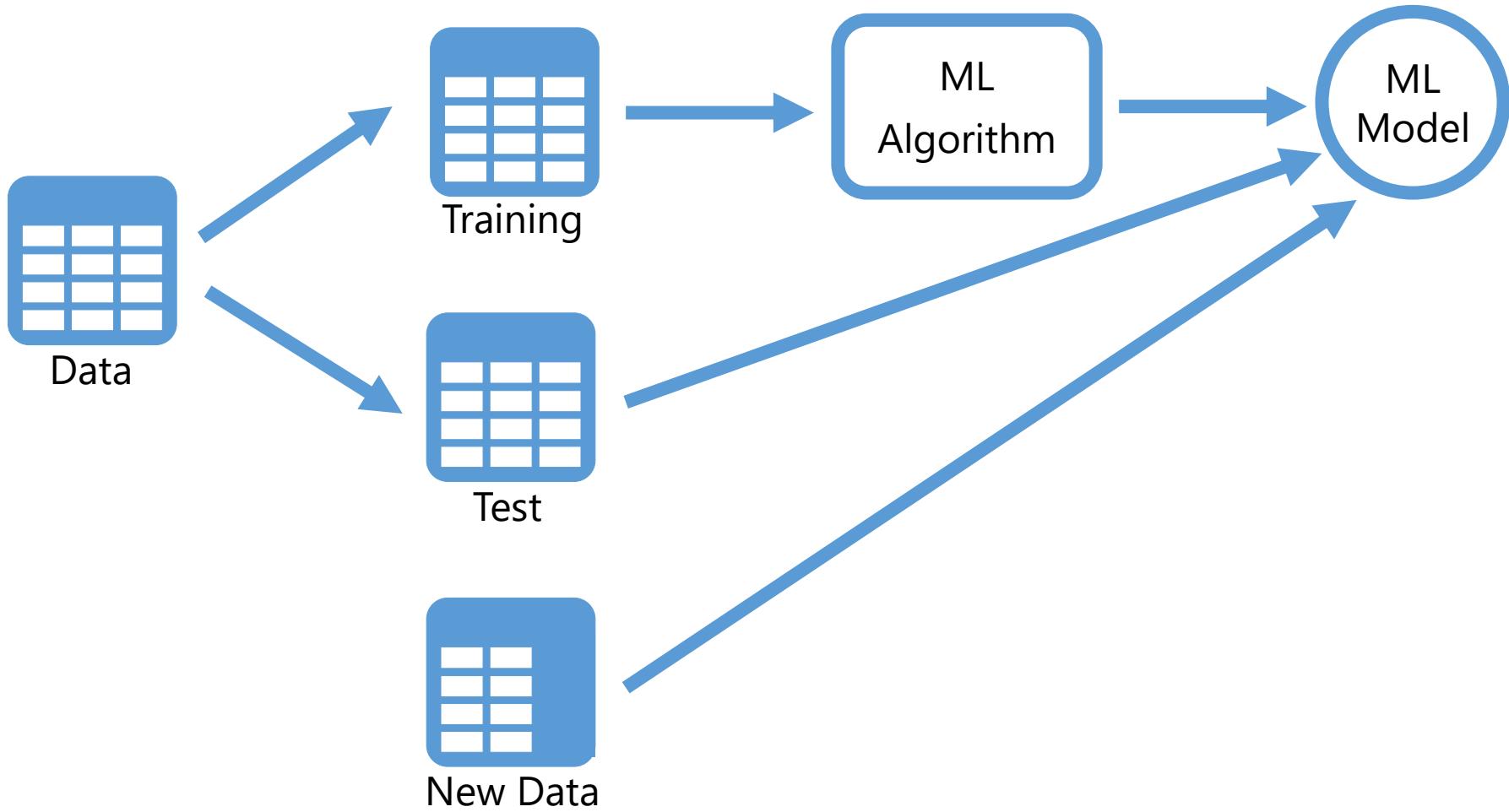
Data

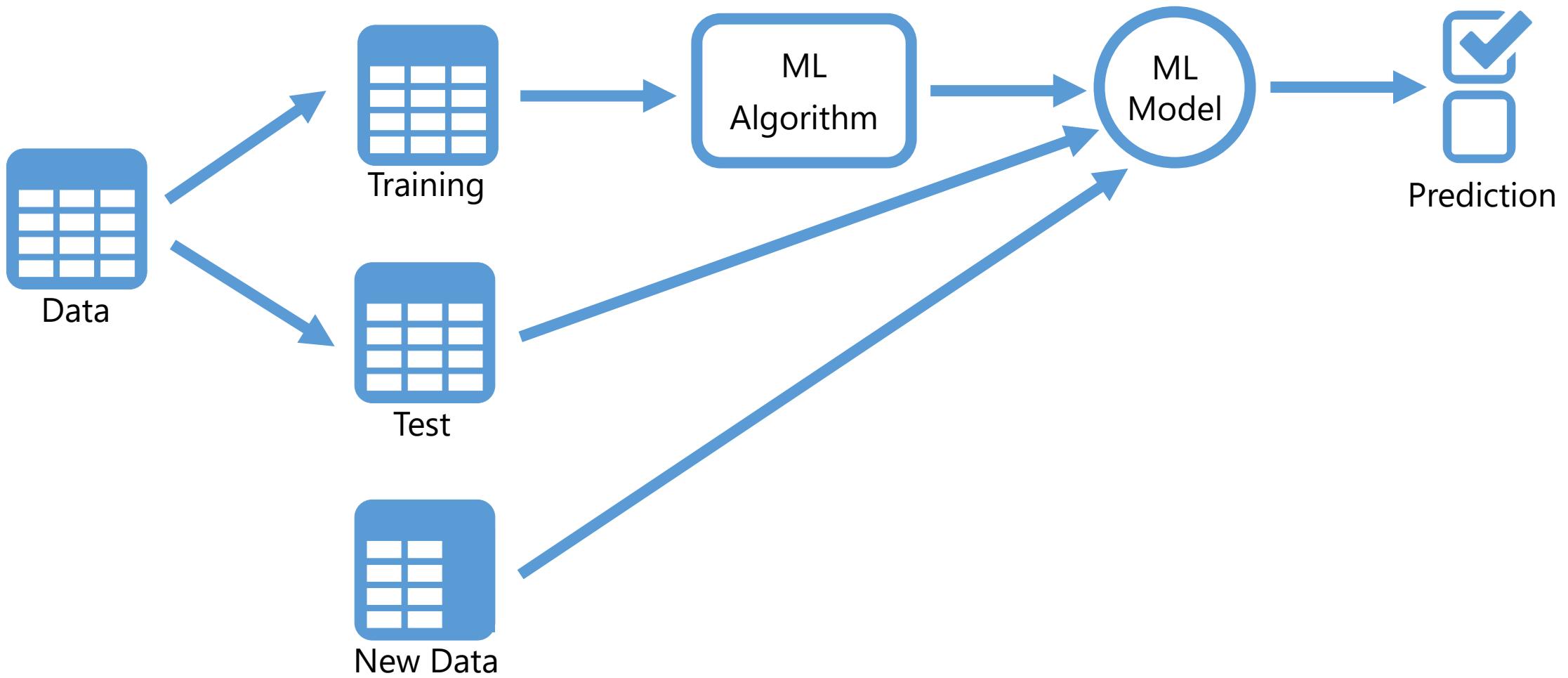


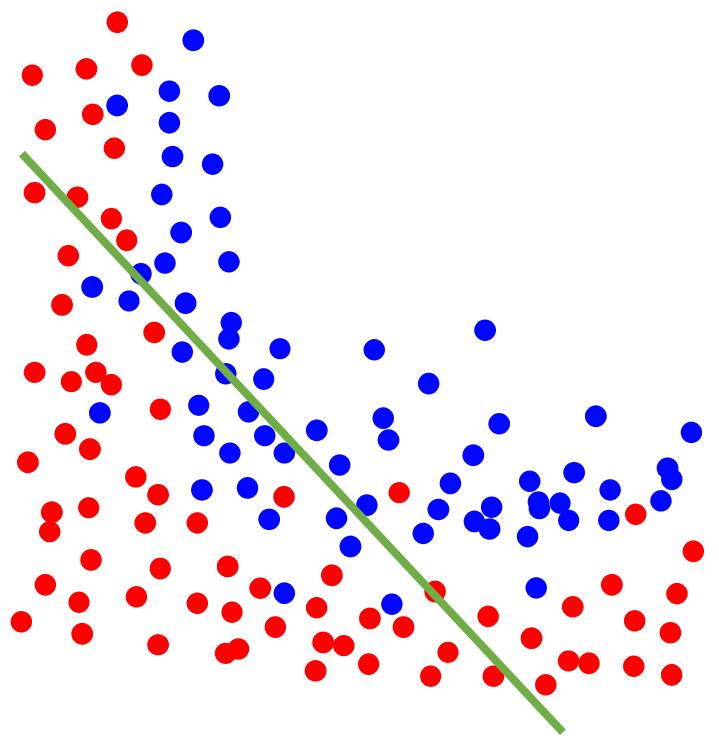




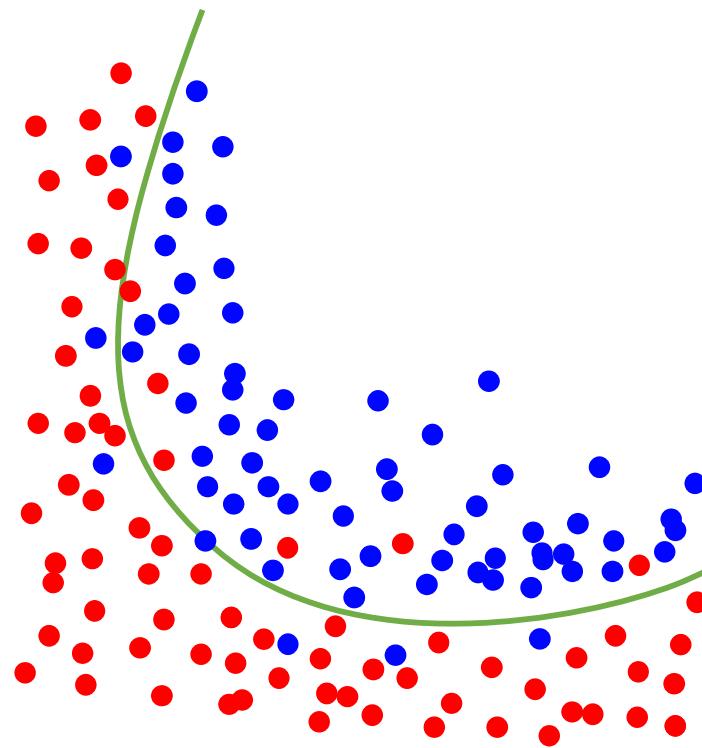




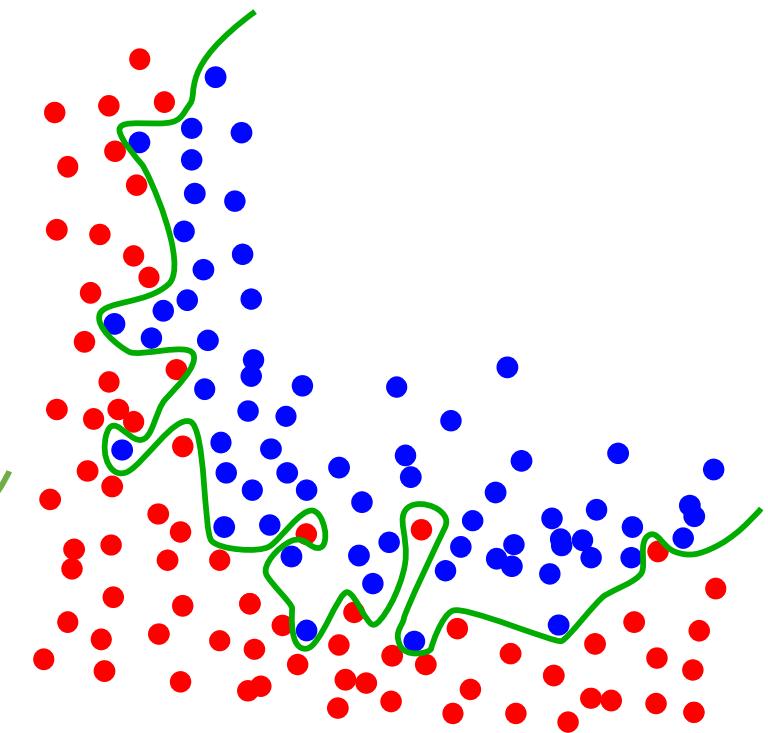




Underfit



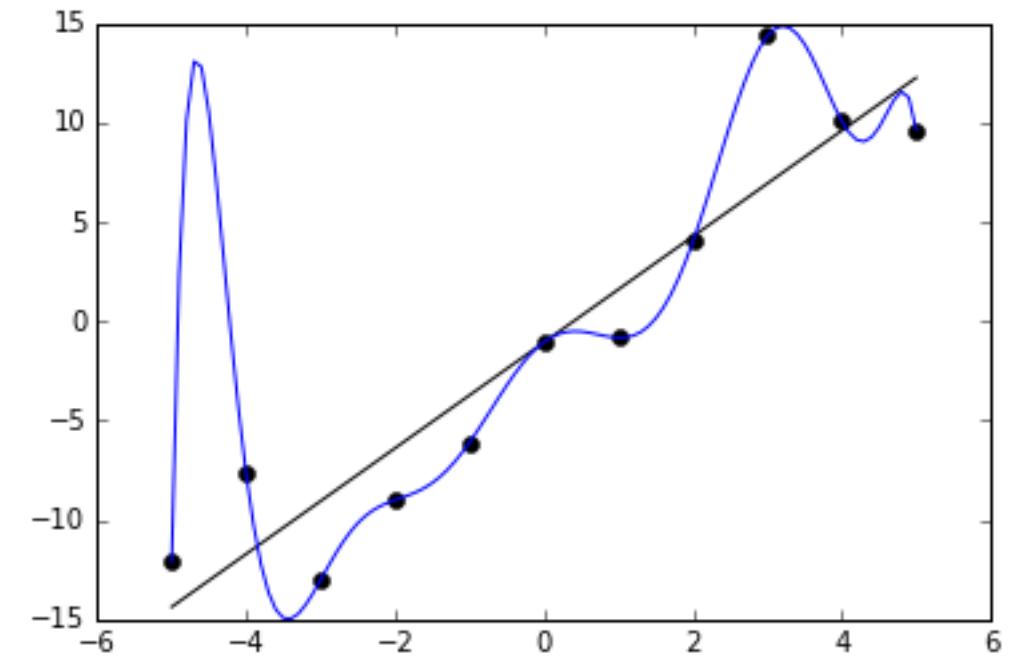
Good fit



Overfit

Regularization Techniques

- Early stopping
- Pruning (trees)
- Adding noise
- Parameter tuning



Source: Wikipedia



Photos by Radomił Binek,
Danielle Langlois, and Frank Mayfield

Iris Data Set



Iris Setosa



Iris Versicolor



Iris Virginica

Code Demo

Lab 7

Machine Learning



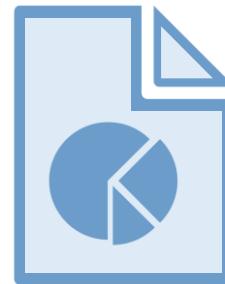
R in Practice

How to Use R in Practice?

1. Deploying R to production
2. Best practices
3. Creating reproducible research

How to Deploy to Production

Export charts (Rstudio)



Create documents (Markdown)



Create interactive reports (Shiny)

Deploy to Server (R Server)

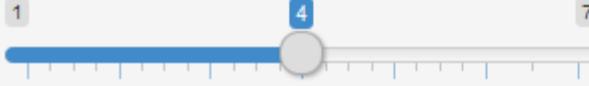


Deploy to Cloud (Azure ML)



Iris Species Predictor

Petal Length (cm)



A horizontal slider for Petal Length in cm, ranging from 1 to 7. The value is currently set at 4.

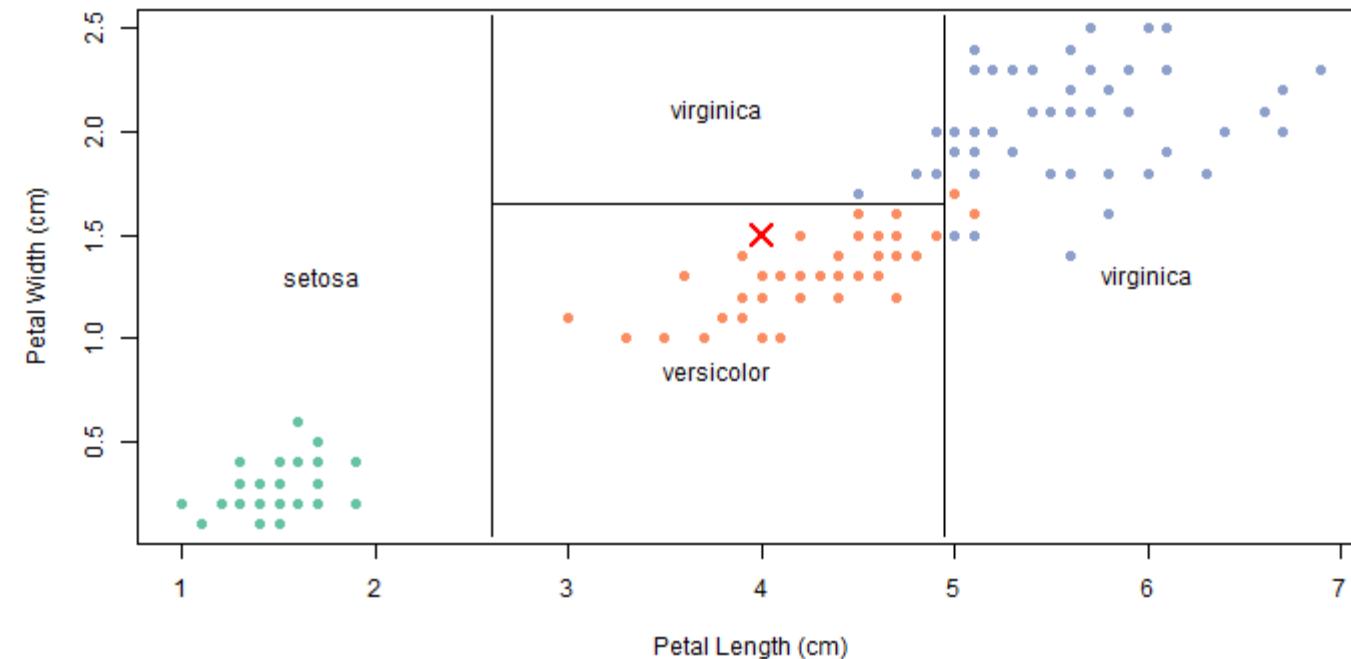
Petal Width (cm)



A horizontal slider for Petal Width in cm, ranging from 0 to 2.5. The value is currently set at 1.5.

The predicted species is versicolor

Iris Petal Length vs. Width



Code Demo



ADVICE

TIPS

GUIDANCE

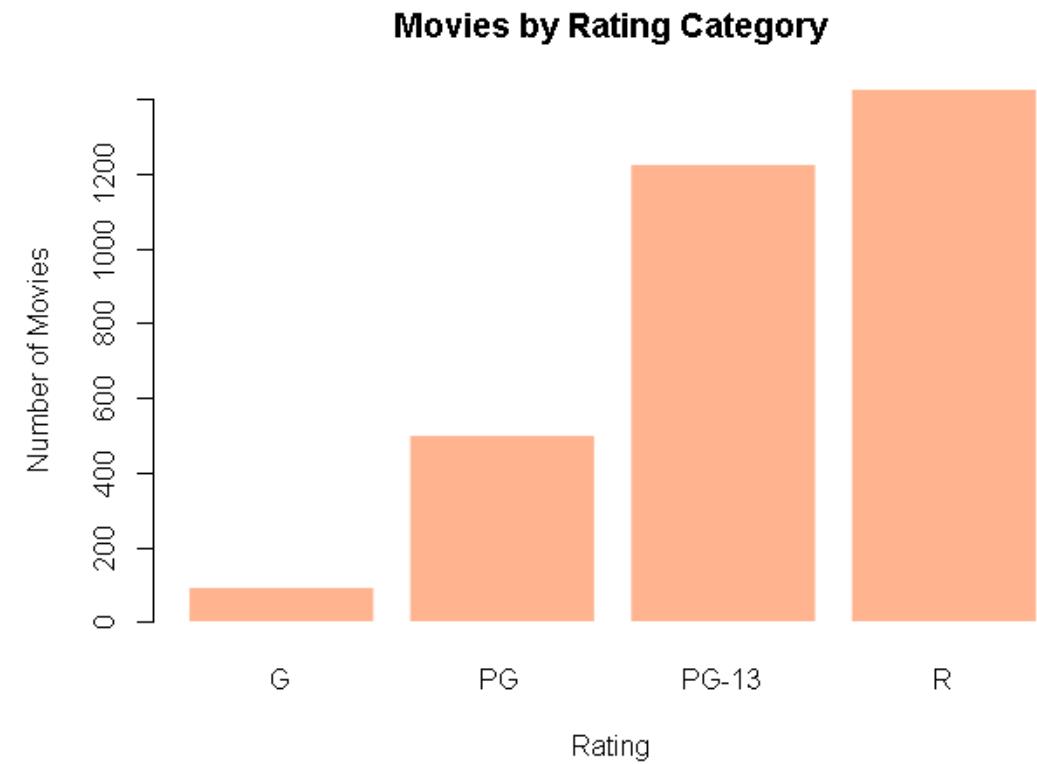
HELP

SUPPORT

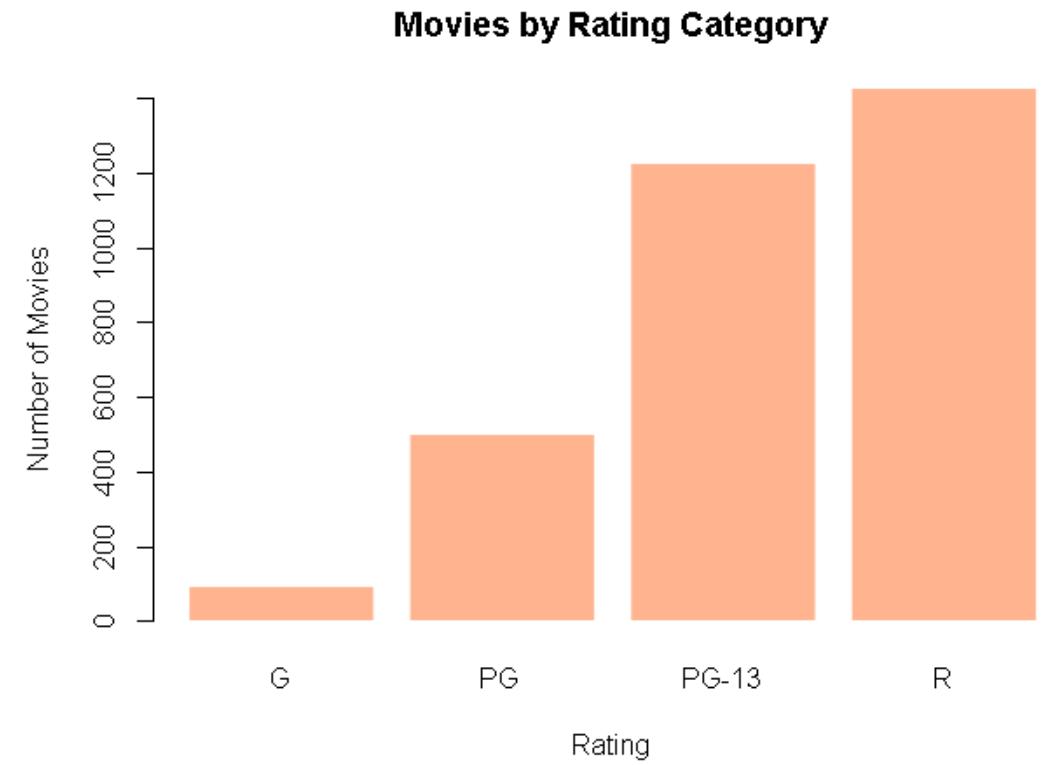
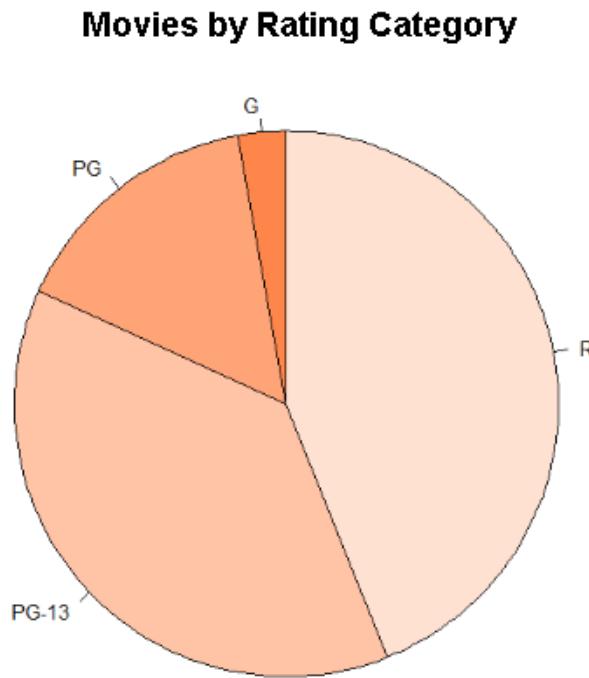
GUIDANCE

Start with a Question

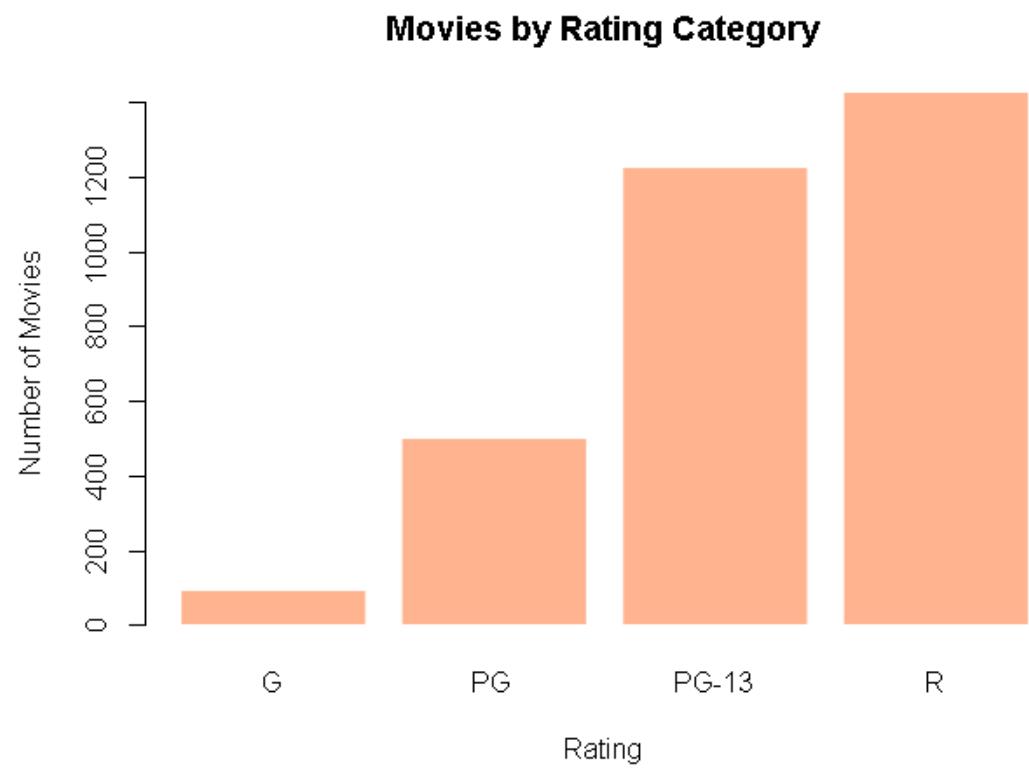
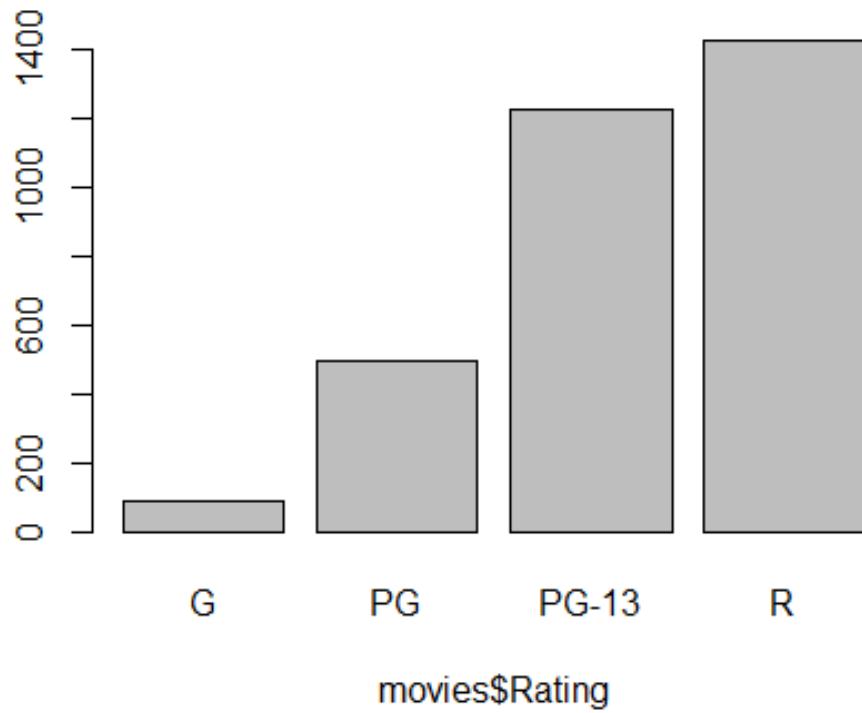
How many movies
were released in
each rating category
from 2000 to 2015?



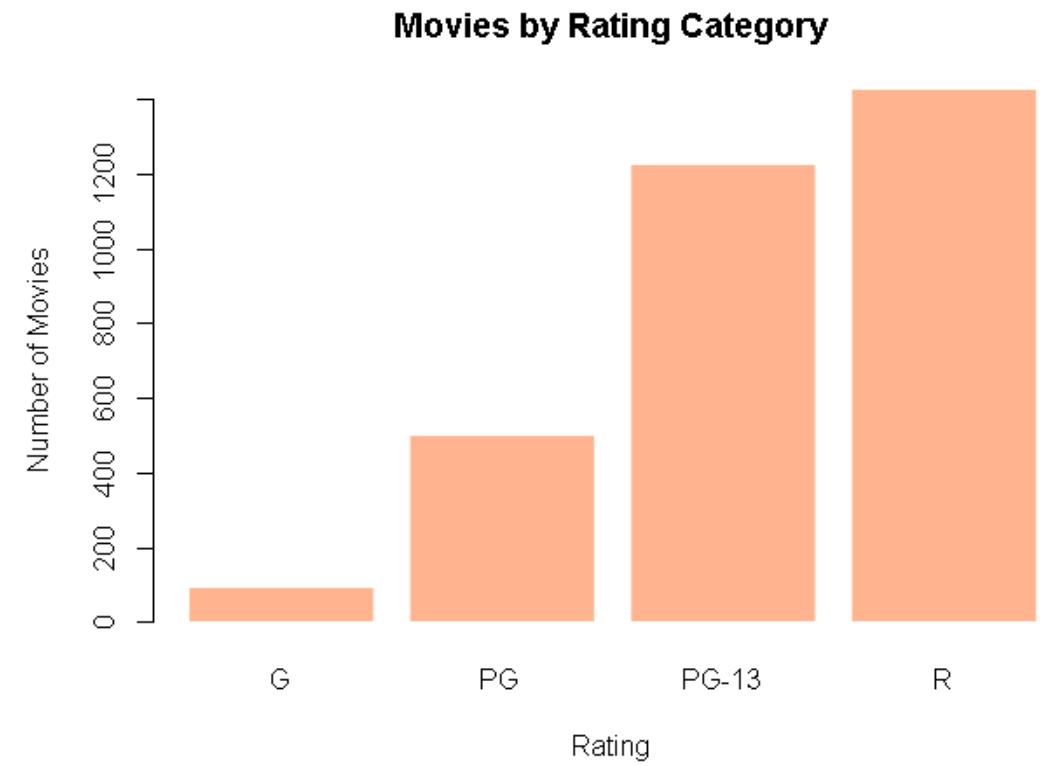
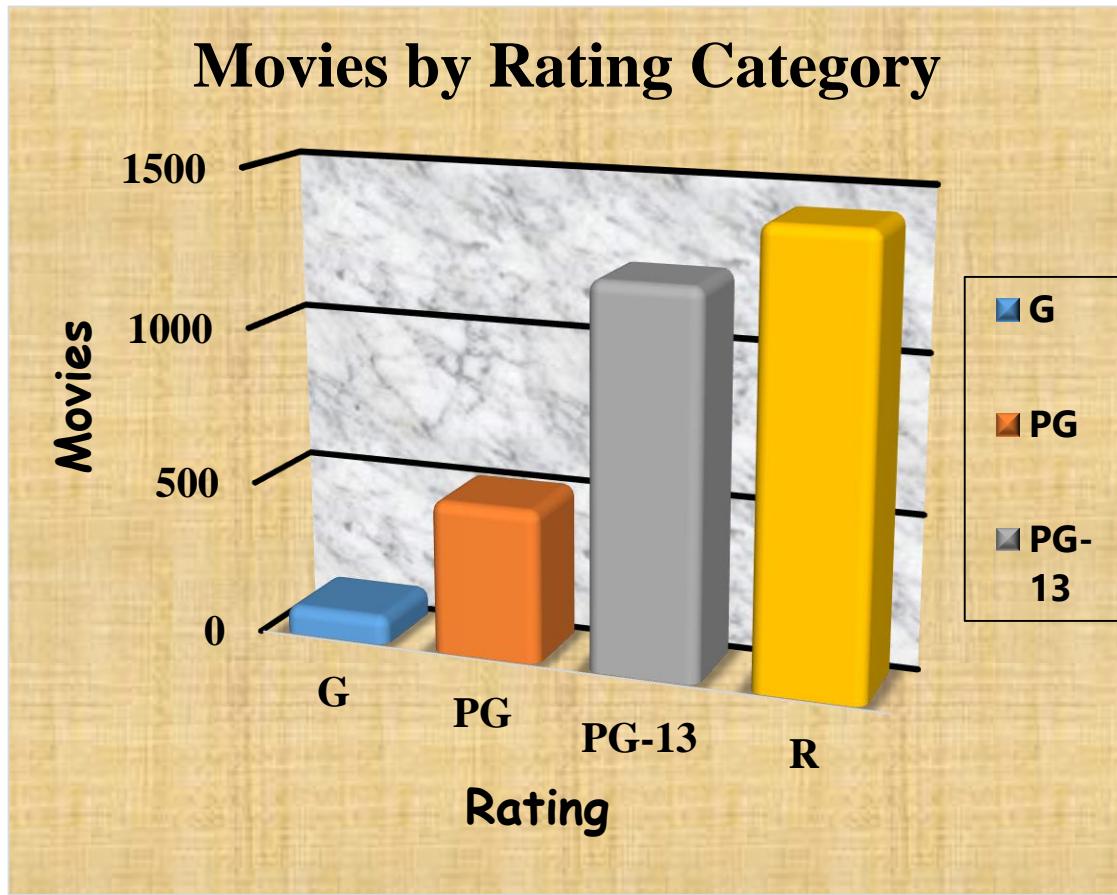
Use the Right Tool for the Job



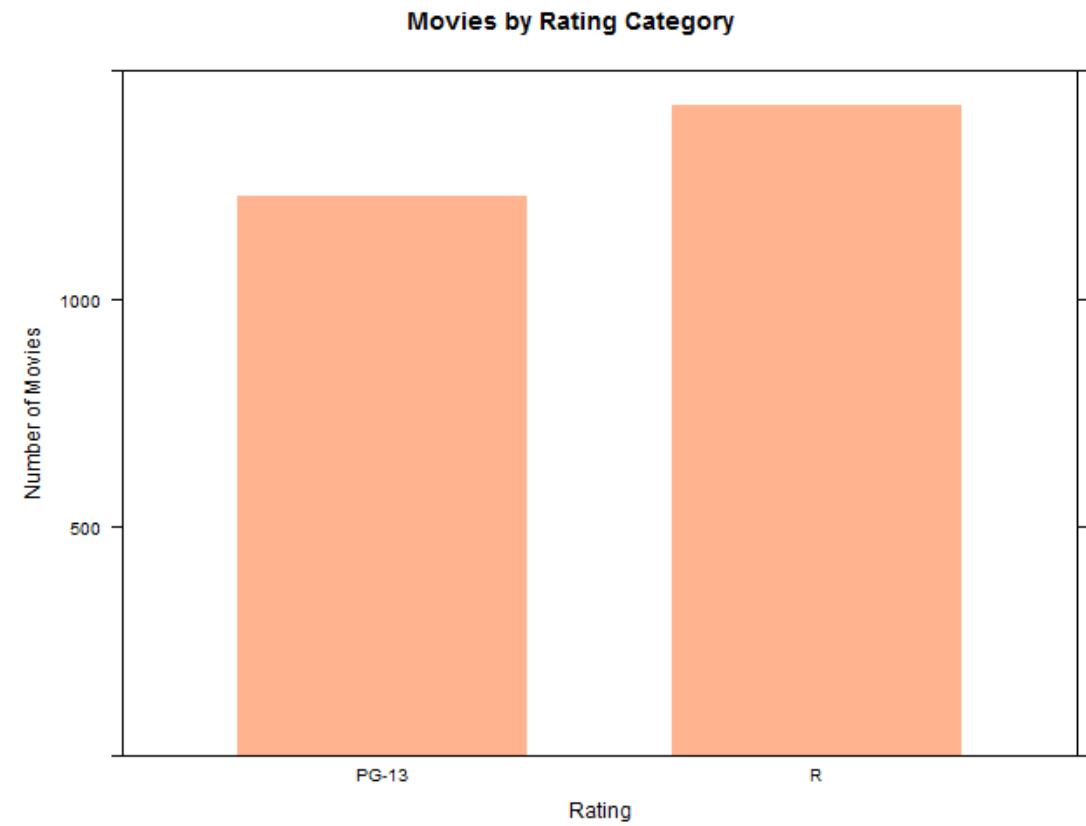
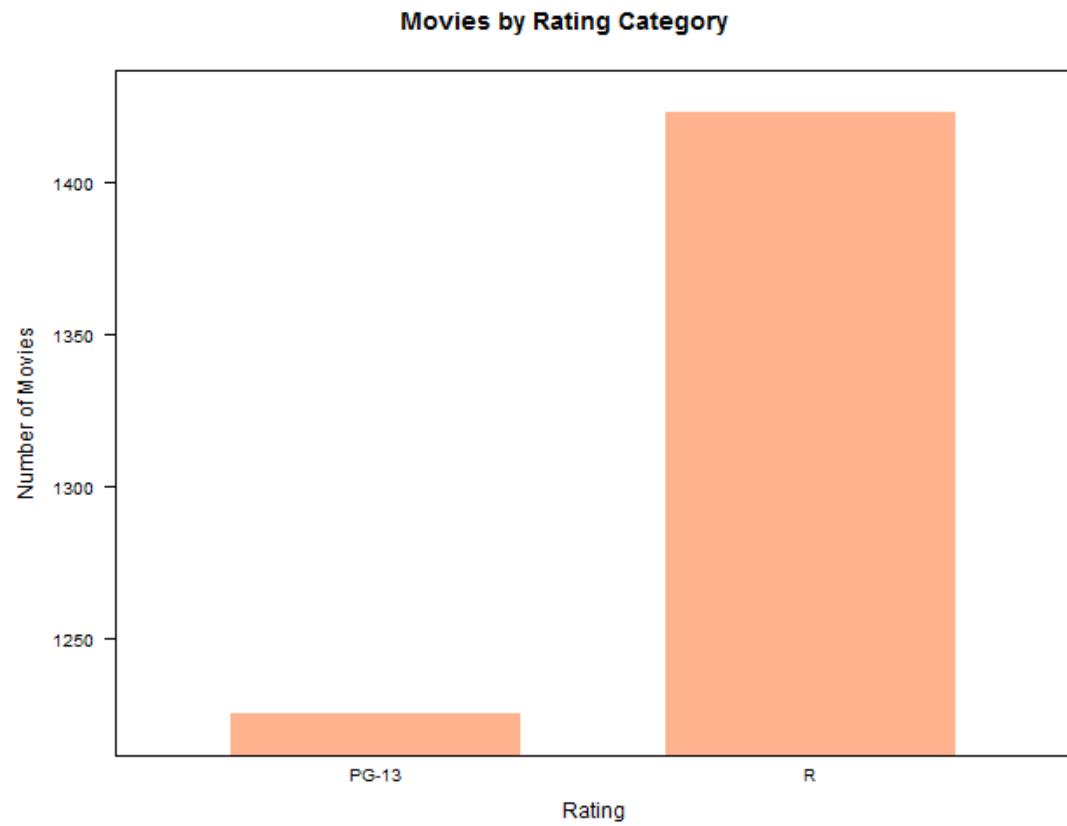
Know Your Audience



Create Clean Data Analyses



Avoid Biases and Information Distortions



Create Reproducible Research

Replication is hallmark of science
Big issue in science right now
Allows other to verify findings
Creates transparency
Allows other to build upon work



Source: <https://blog.mendeley.com>

How Do We Create Reproducible Research?

- Provide raw data and code
- Script all analysis steps
- Use source control
- State all assumptions
- Use markdown



Source: <https://blog.mendeley.com>

Where to Go Next...

Pluralsight: <https://www.pluralsight.com>

Coursera: <https://www.coursera.org/specializations/jhu-data-science>

Revolutions: <http://blog.revolutionanalytics.com>

Flowing Data: <http://flowingdata.com>

R-Blogger: <http://www.r-bloggers.com>

R-Seek: <http://rseek.org>

News

2017-08-25 - Invitation to Speak at Devoxx Morocco

Very excited to announce that I've been invited to give a keynote in Casablanca at [Devoxx Morocco](#) in November. My keynote presentation will be on [Artificial Intelligence](#).



2017-08-16 - Invitation to Speak at Microsoft Ignite

I've been invited to speak at [Microsoft Ignite](#) in Orlando, Florida in September. This will be my first time speaking at Ignite. Talks will include both Data Science and Machine Learning with R.



Matthew is a data science consultant, author for [Pluralsight](#), international public speaker, a [Microsoft MVP](#), [ASPIndier](#), and open-source software contributor.

2017-08-14 - Dev on Fire Interview



PLURALSIGHT

Data Science with R

Exploratory Data Analysis with R

Data Visualization with R (3-part)

Data Science: The Big Picture

Data Science with R



Matthew Renze
SOFTWARE CONSULTANT
@matthewrenze www.matthewrenze.com



www.pluralsight.com/authors/matthew-renze

Conclusion

Conclusion

1. Introduction
2. Working with Data
3. Descriptive Statistics
4. Data Visualization
5. Statistical Modeling
6. Handling Big Data
7. Machine Learning
8. R in Practice



Feedback

Very important to me!

What did you like?

What could I improve?



Thank You!

Matthew Renze
Data Science Consultant
Renze Consulting

Twitter: [@matthewrenze](https://twitter.com/matthewrenze)
Email: info@matthewrenze.com
Website: www.matthewrenze.com

