Introduction: What is Data Science

Materials Science and Engineering

CASE WESTERN RESERVE UNIVERSITY

What is data science?

1
- Coding, math, and statistics in applied settings

2
- The analysis of diverse data (data that did not fit into analytic approaches)

3
- Inclusive analysis. (all of the data/information you have in order to get the most insightful and compelling answer to your research questions

What is data science?

Problem Formulation	• Identify an outcome of interest and the type of task: classification / regression / clustering • Identify the potential predictor variables • Identify the independent sampling units
Collect & Process Data	• Conduct research experiment (e.g. Clinical Trial) • Collect examples / randomly sample the population • Transform, clean, impute, filter, aggregate data • Prepare the data for machine learning — X, Y
Machine Learning	• Modeling using a machine learning algorithm (training) • Model evaluation and comparison • Sensitivity & Cost Analysis
Insights & Action	• Translate results into action items • Feed results into research pipeline

Data science cycle

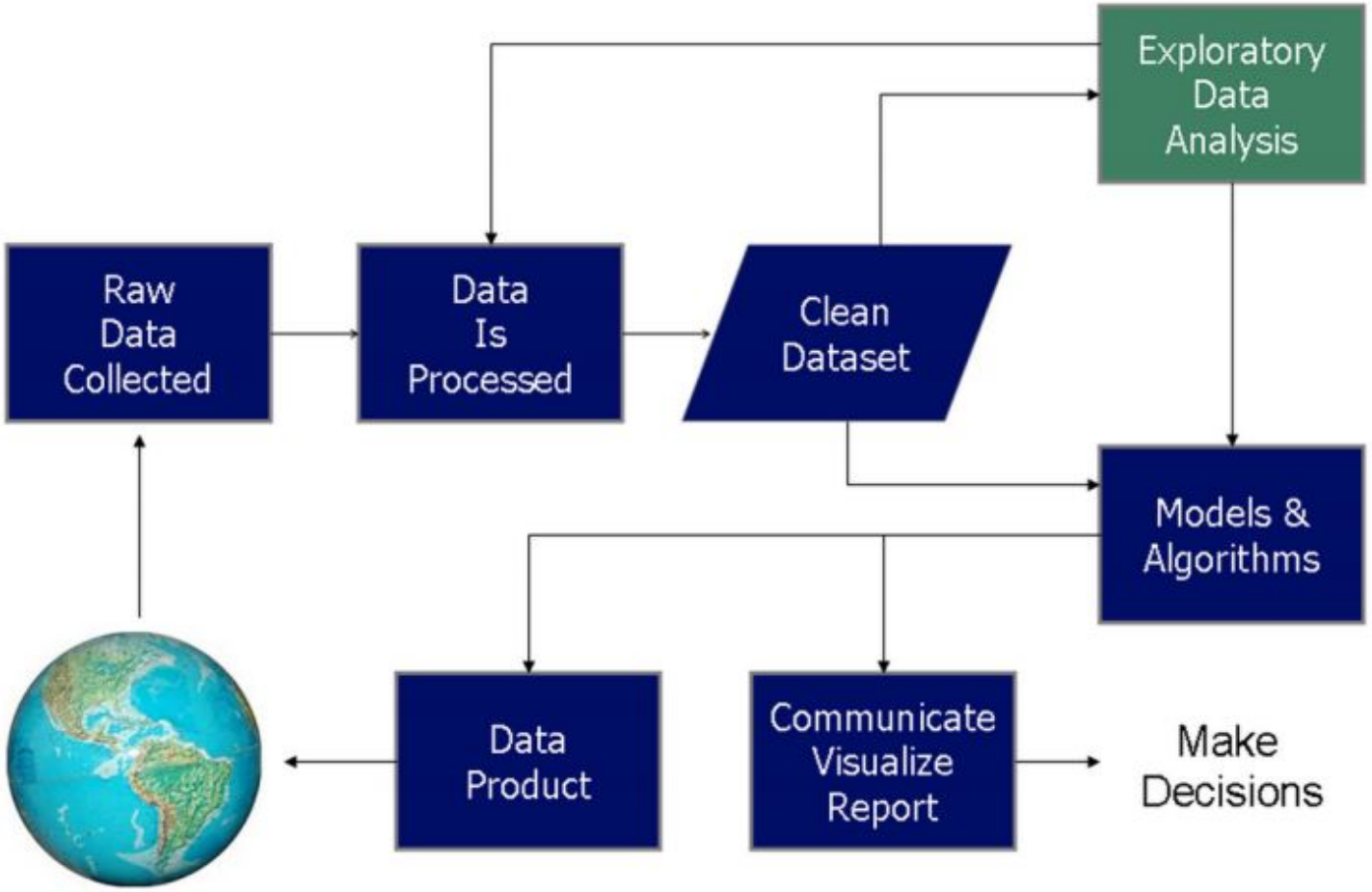
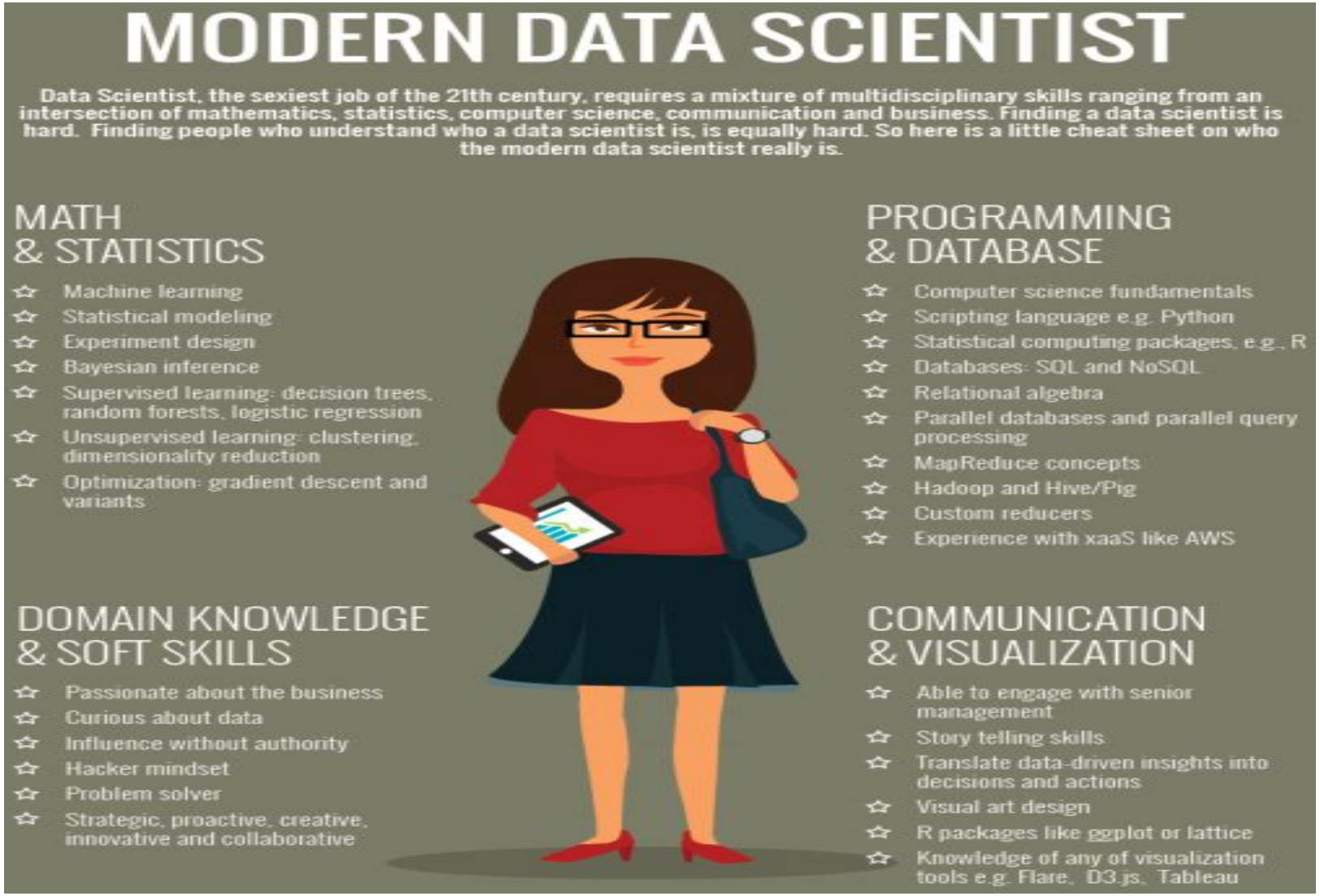


Image: https://en.wikipedia.org/wiki/Data_science

What does a Data scientist do?

- Data curiosity, explore data, discover unknowns
- Understand data relationships
- Understand the business, has domain knowledge
- Can tell relevant stories with data
- Holistic view of the business
- Knows machine learning, statistics, probability
- Can hack and code
- Define and test an hypothesis, run experiences
- Asks good questions

What does a Data scientist do?



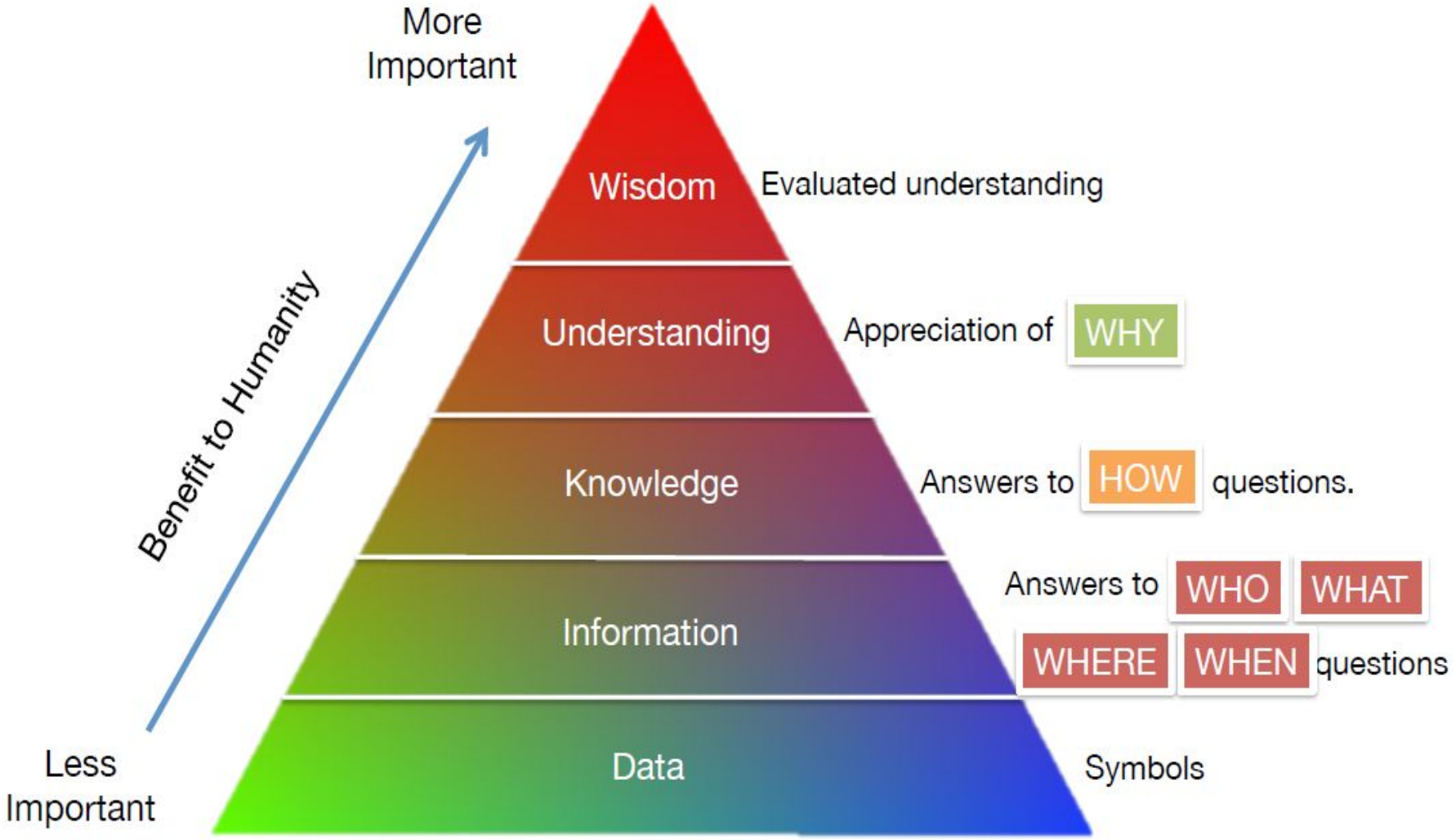
Data scientist tools

- We recommend learning Data Science using the R language.
- Pro & Con:
  - Pro: User-friendly data analytics, statistics and graphical models. Good for research and academic purpose. Reproducible research
  - Con: “Steep learning curve at the beginning.”
  - Disagree with this. I think Python learning curve is more diffcult
- Quatro: Integrated R, Python, Julia, VS code
  - Bilingual especially because the philosophy is similar between R and Python (except those 0 and 1

Data scientist tools

- Python is another high level, scripting language.
  - https://www.learnpython.org/
- Pro & Cons:
  - Pro: Powerful libraries, general-purpose language, good readability, easy to integrate into cloud-related software.
  - Con: You need to read a lot of documents when learning different libraries
  - More like the “Wild Wild West”; anyone can make a package
  - Less focus on statistics and visualization
- Google “Python VS R for Data Science”
  - to find out the comparison of these two,
  - or best is learn both.

Data scientist tools



Data science: Example 1

Better understand and target customers

- Expand traditional data sets
  - with social media data, browser, text analytics, etc.
- Get a more complete picture of their customers
- Create predictive models
  - Retailers can predict what products will sell well
  - Car insurance companies can understand how well the drivers actually drive.

Data science: Example 2

Understand and optimize business processes

- Retailers can optimize their stock based on the predictive models
- Supply chain or delivery route optimization based on big data (geographic positioning data and RFID)

Data science: Example 3

Improve health:

- Find new cures and
  - better understand and predict disease patterns
- Using data from smart wearable devices
  - can better understand connections
  - between lifestyles and diseases
- Monitor and predict epidemics and disease outbreaks

Data science: Example 4

Improve sports performance:

- Sensor technology is built into sport equipment
- Video analytics to track the performance of every player in a football or baseball game
- Track nutrition and sleep
- Monitor emotional wellbeing

Is data science in need?

Rare Qualities

Data science takes unstructured data, then finds order, meaning and value

High Demand

Data science provides insight and competitive advantage (huge thing in business setting

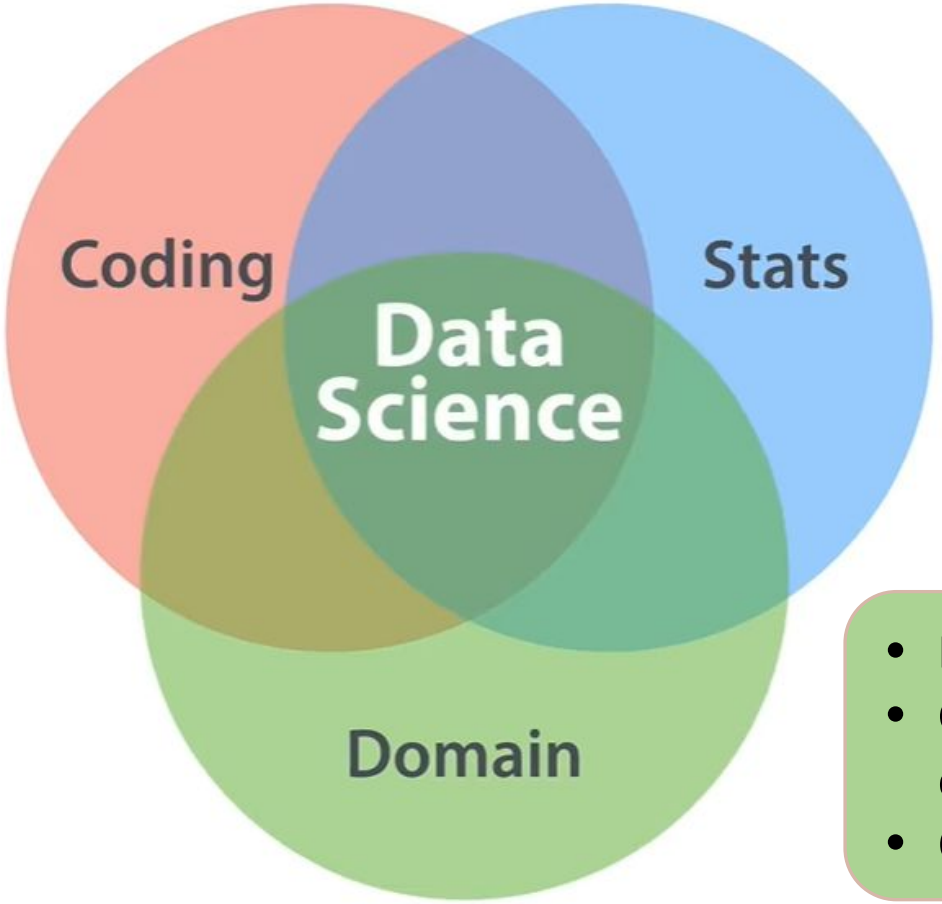
McKinsey & Company: Executive summary Projecting a need

- **140-190K deep analytical talent position**
- **1.5M data-savvy managers (no analysis/understands/speaks data**

Data science Venn Diagram

Gather, prepare data and requires creativity

- **Statistical (R & Python)**
- **Database (SQL)**
- **Command line (Bash)**
- **Regex**

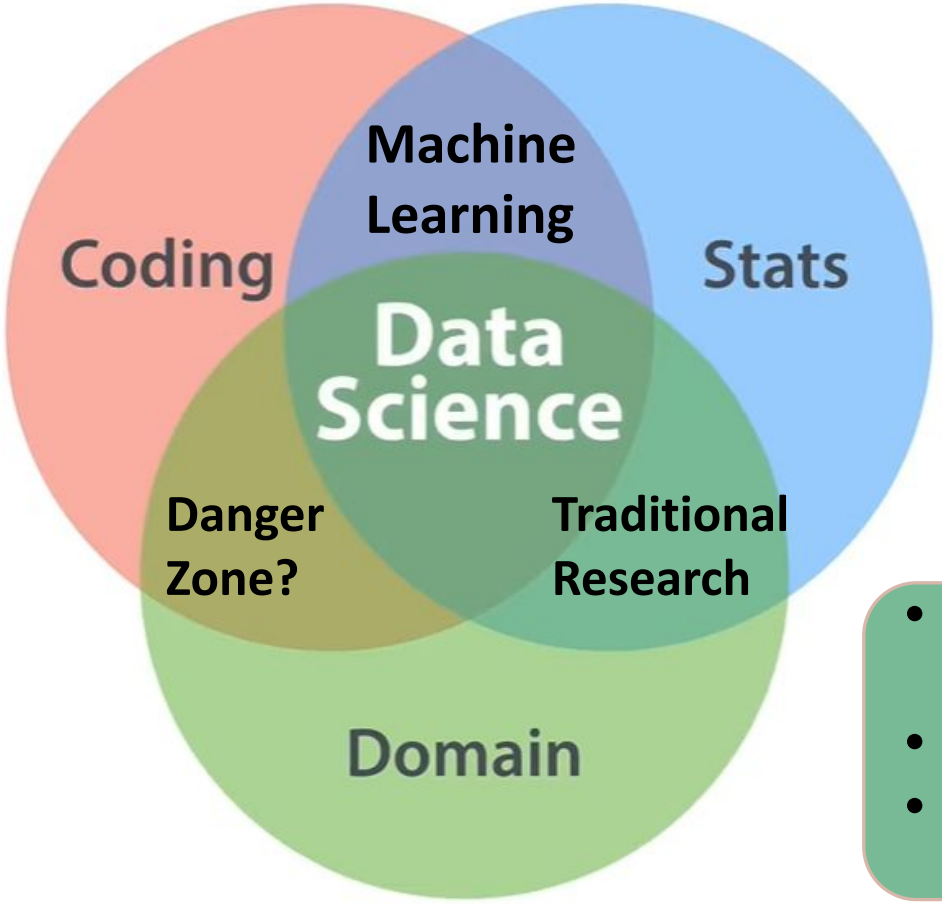


- **Probability, algebra, regression, etc.**
- **Choose procedure**
- **Diagnose problems**
- **Understand the mechanics of the problem**

- **Expertise in field.**
- **Goals, methods, and constraints**
- **Can implement well**

Diagram is from Drew Conway j.mp/ds-venn

Data science Venn Diagram



- ML: Machine Learning
- Coding and Math without domain
- “Black box” models

- Coding and domain without math
- Unlikely to happen
- Word counts, maps

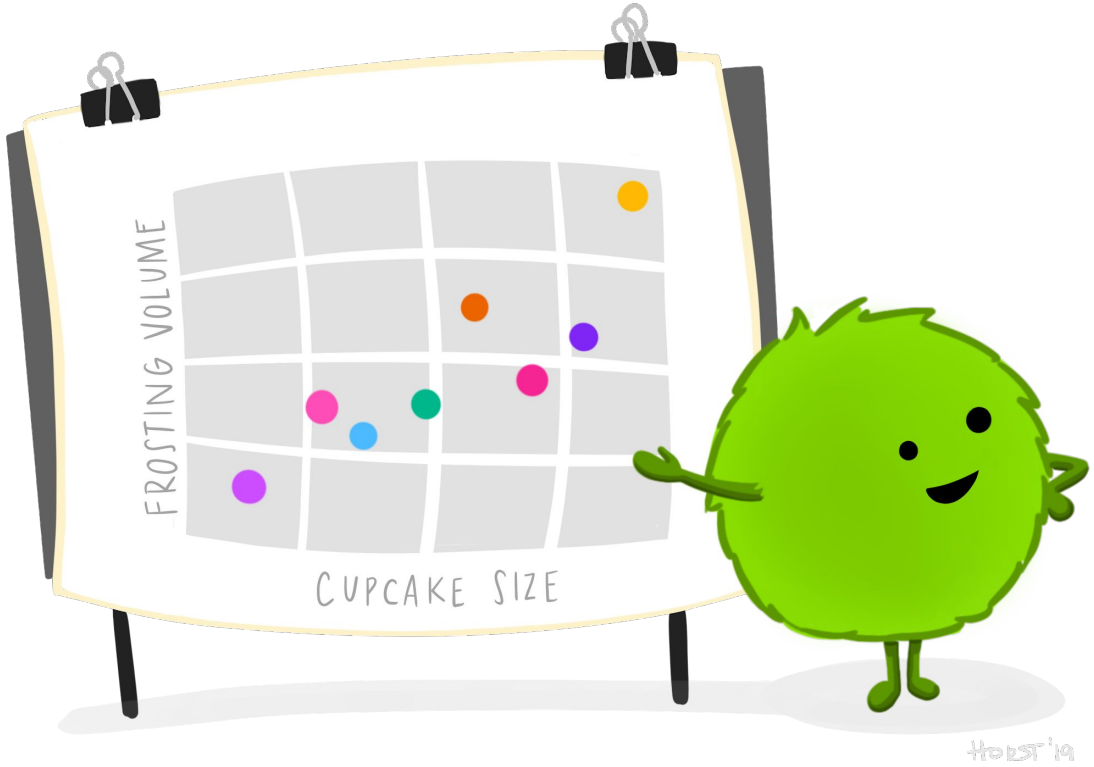
- Math and Domain without coding
- Data is structured
- Effort is in method and interpretation

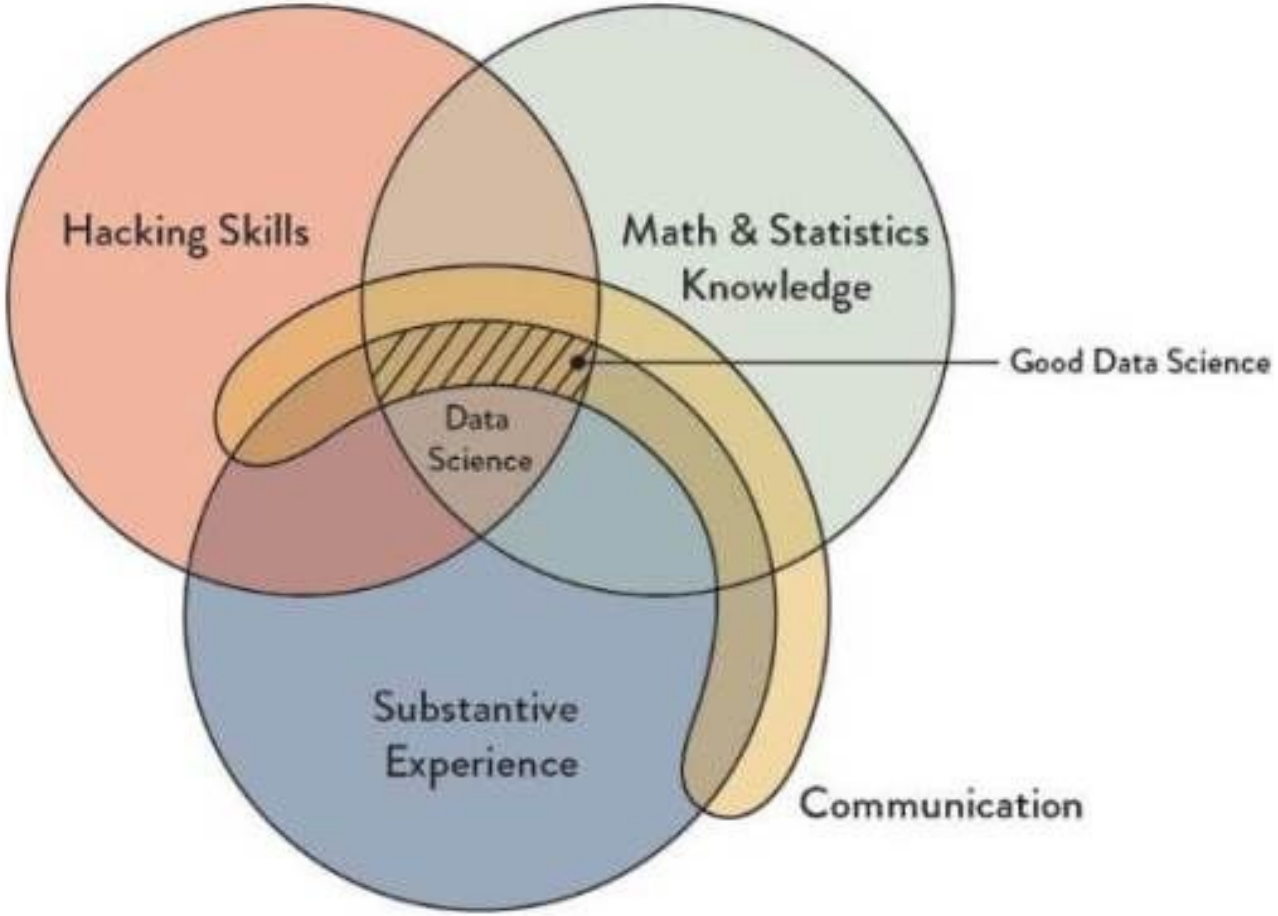
Diagram is from Drew Conway j.mp/ds-venn

Data Science Skills

- Communication

  Need to communicate results!!





Data science demographics

Code	Statistics	Domain
Coders who can do math, stats, and business	Statisticians who can code and do business	Business people who can code and do numbers
Most Common	Less Common	Least Common

All of them are important to data science

- Several fields make Data Science
  - Diverse skills needed
  - Many roles involved

Data science pathway

Planning

- Define goals
- Organize resources
- Coordinate people
- Schedule project

Data Prep

- Get data
- Clean data
- Explore data
- Refine data

Modeling

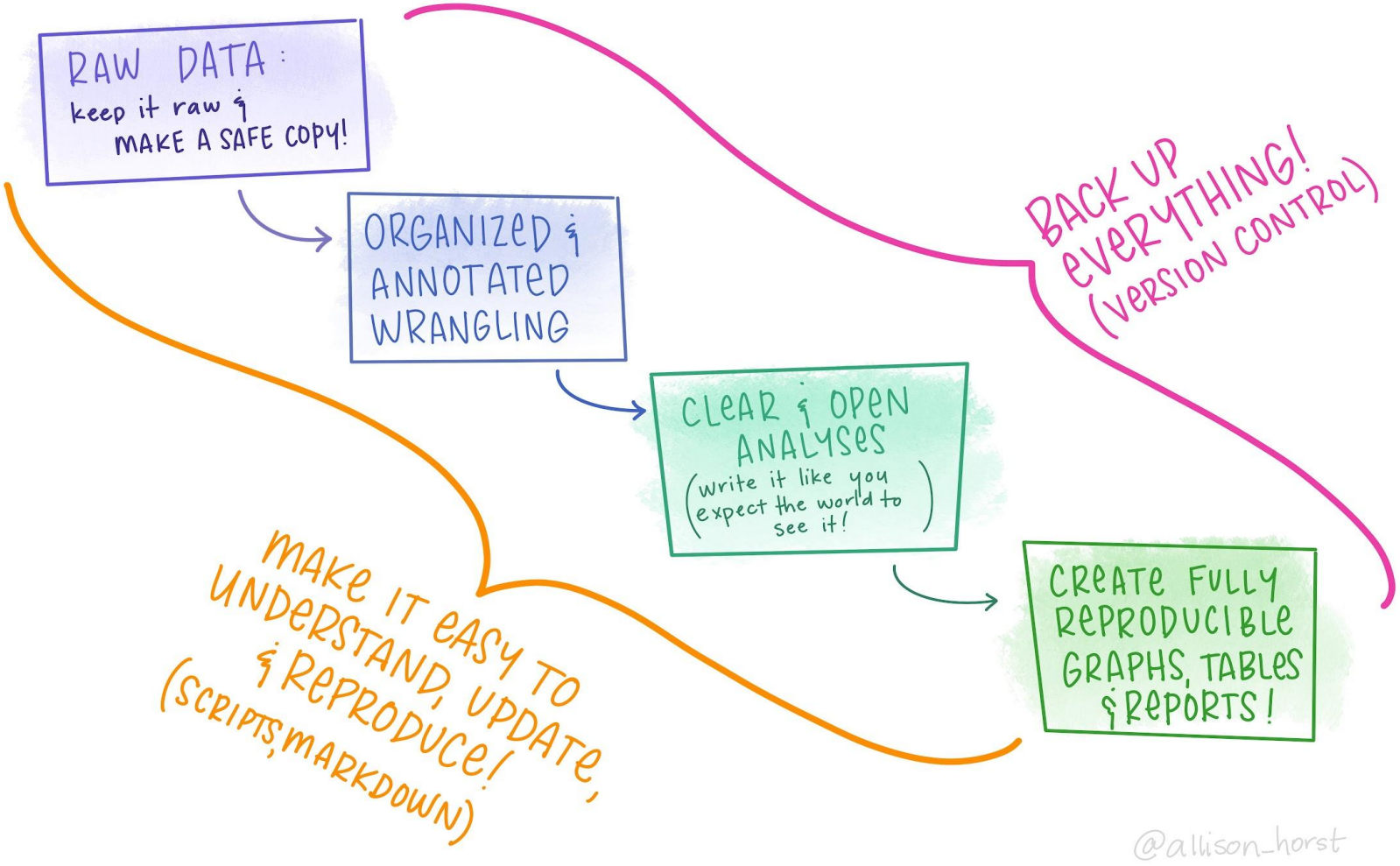
- Create model
- Validate model
- Evaluate model
- Refine model

Follow up

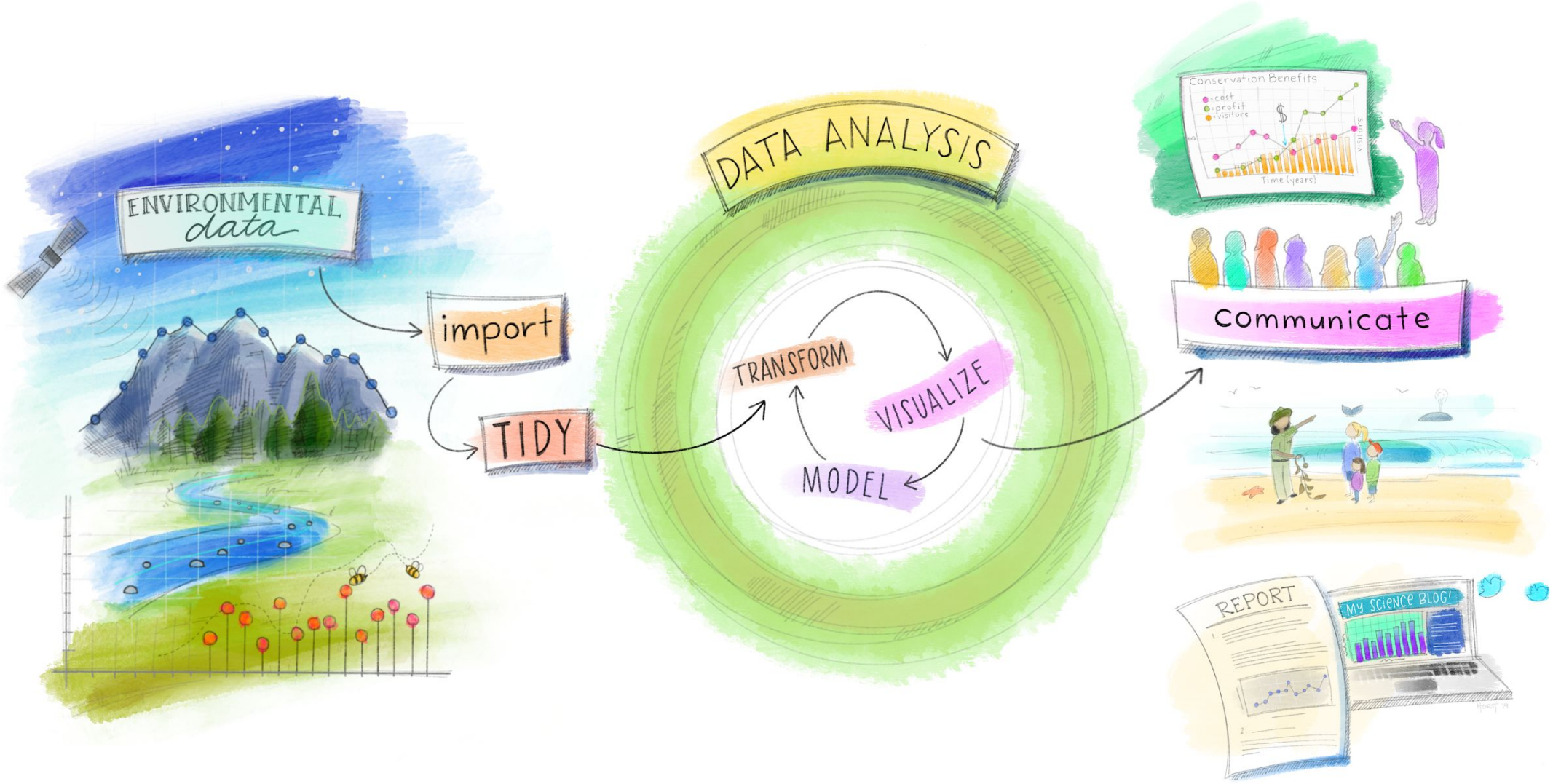
- Present model
- Deploy model
- Revisit model
- Archive assets

- **Data science is not just technical**
  - **Contextual skills matter**
    - **One step at a time**

Data Science Pathway



Total Data Science Pathway



Roles in Data science

Engineers

- Focus on back end hardware, software.
- Makes DS possible
- Developer, Data base administrators

Big Data

- Focus on computer science and math
- Machine learning
- Data products

Researchers

- Focus on domain specific research
- Physics, genetics, material science
- Strong statistics

Roles in Data science

Analyst

- Day-to-day tasks
- Web analytics, SQL
- Good for business
- **Not exactly DS?**

Business

- Frame business relevant questions
- Manages projects
- Must “speak data”

Entrepreneur

- Data startups
- Needs data and business skills
- Creative throughout

Roles in Data science

Full-Stack

- **Knows all of the other professions**
- **Very rare (unicorns)**

A mythical creatures with magical abilities

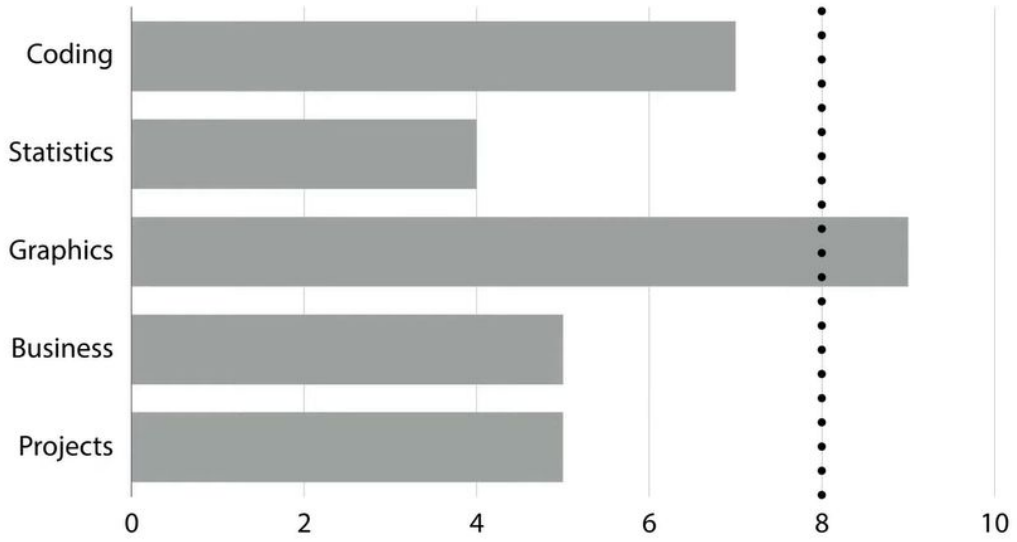
A mythical data scientist with universal abilities

- **Data science is diverse**
- **Different goals and skills**
- **Different contexts**

Teams in Data science

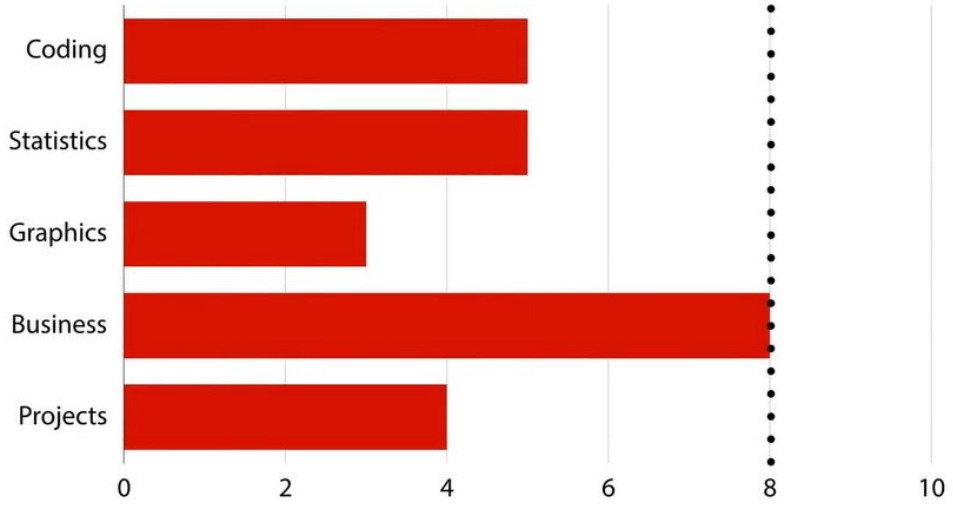
John

- **Strong visualization**
- **Good coding**
- **Limited analytics**



Sara

- **Strong business**
- **Good tech skills**
- **Limited graphics**



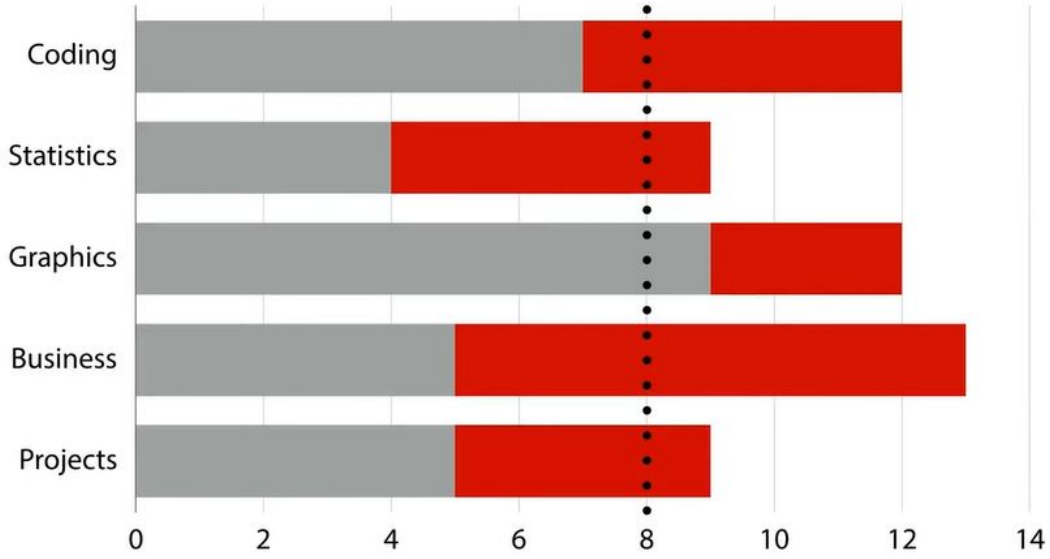
Teams in Data science

John

- Strong visualization
- Good coding
- Limited analytics

Sara

- Strong business
- Good tech skills
- Limited graphics

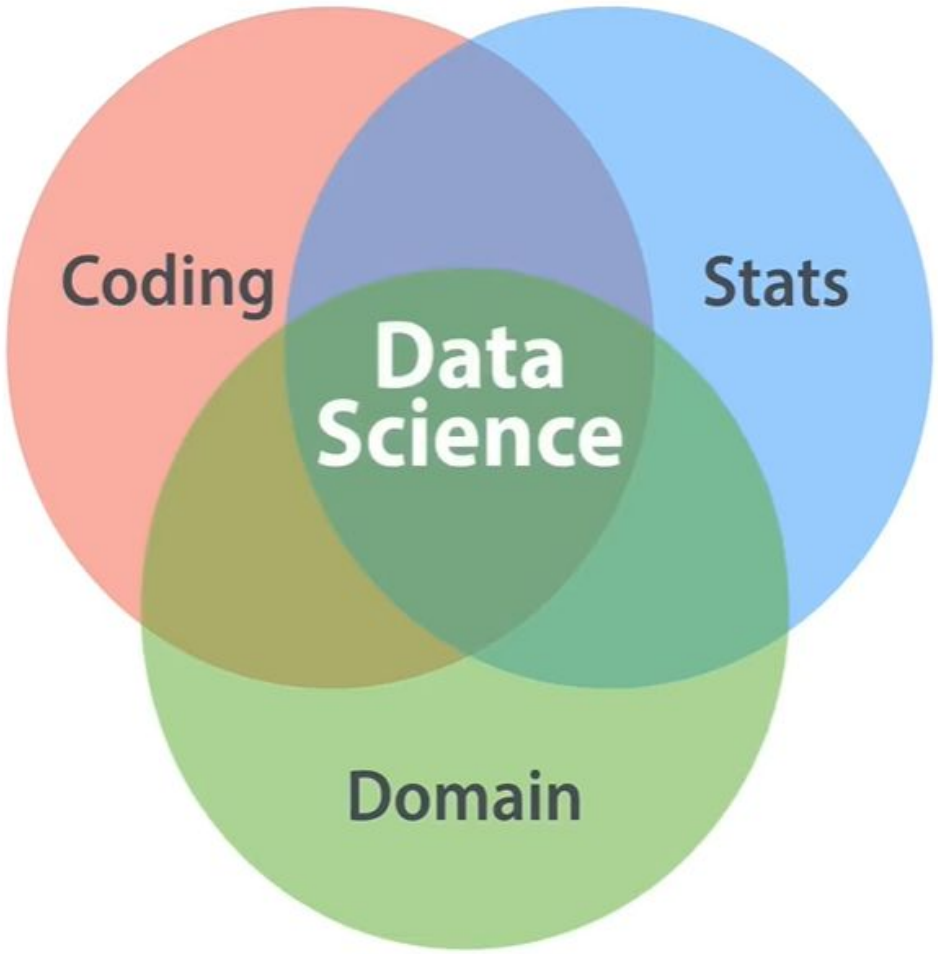


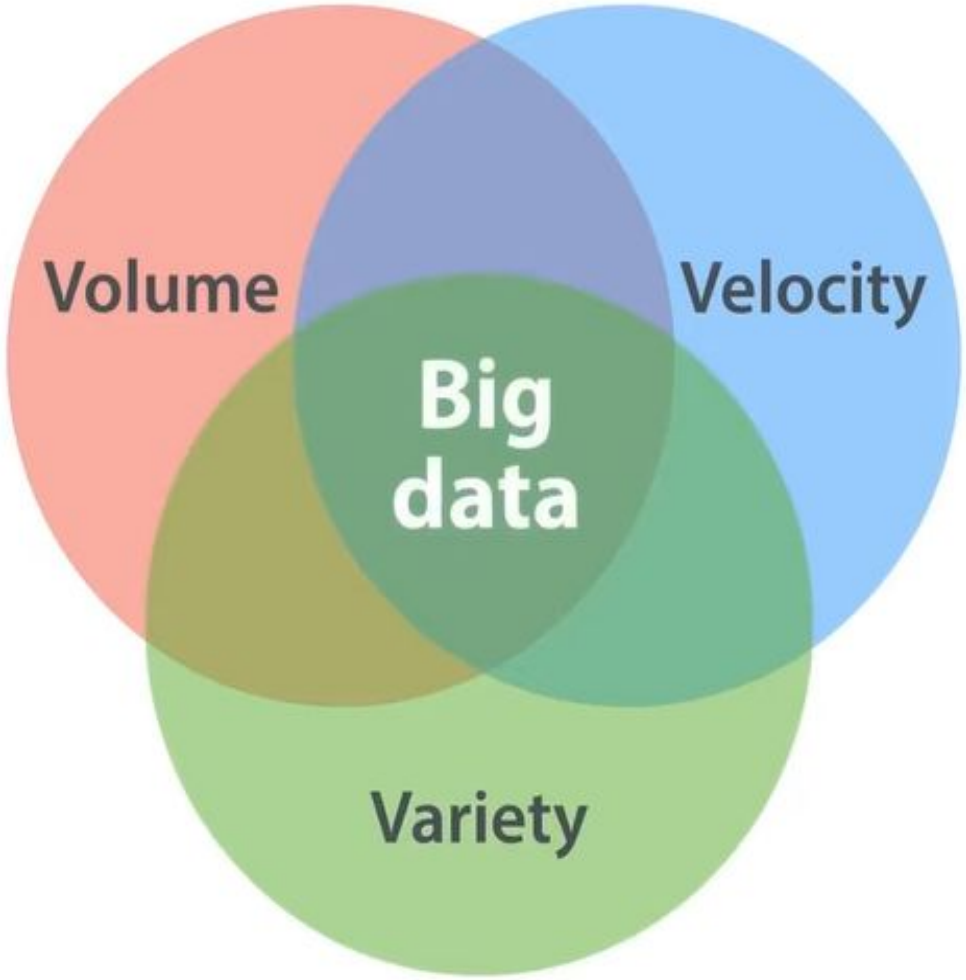
Full-Stack-Team

- Can’t do DS on your own
- People need people
- Make unicorns collectively

Data science vs. Big data

Data science vs. Big data

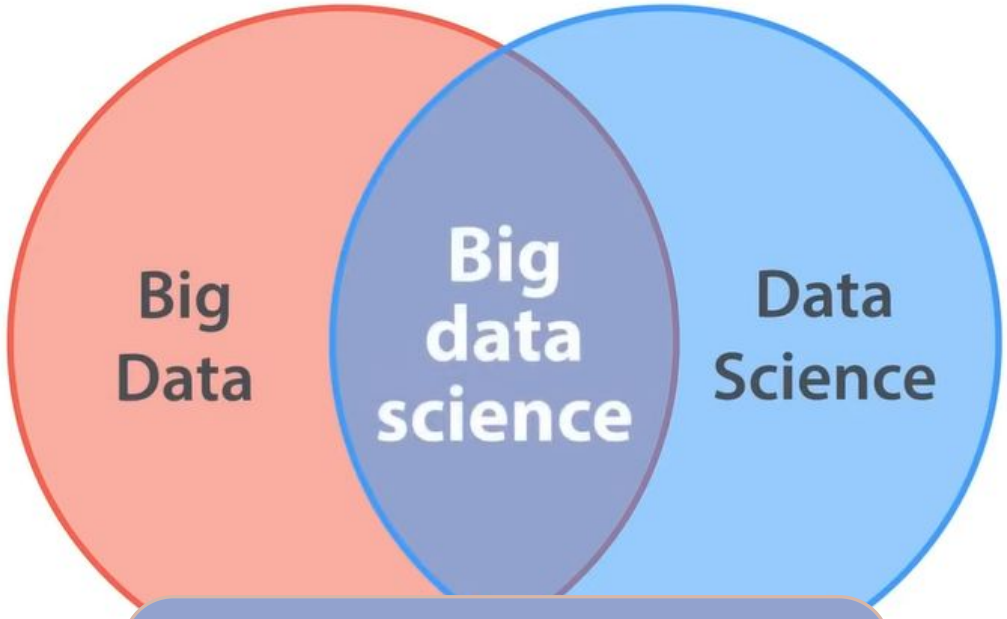




Data science vs. Big data

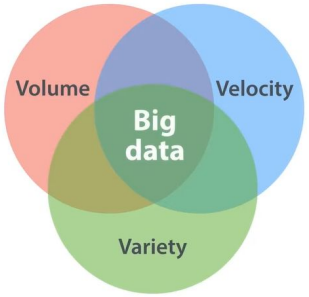
Big data without all V's?

- **Machine learning and word counts**
- **Need at least two skills**



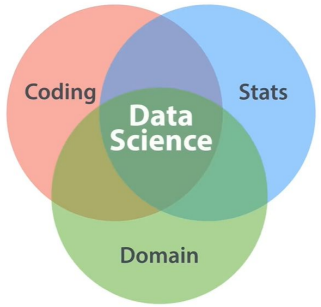
Data with 1 V?

- **Genetics data**
- **Streaming sensor data**
- **Facial recognition**

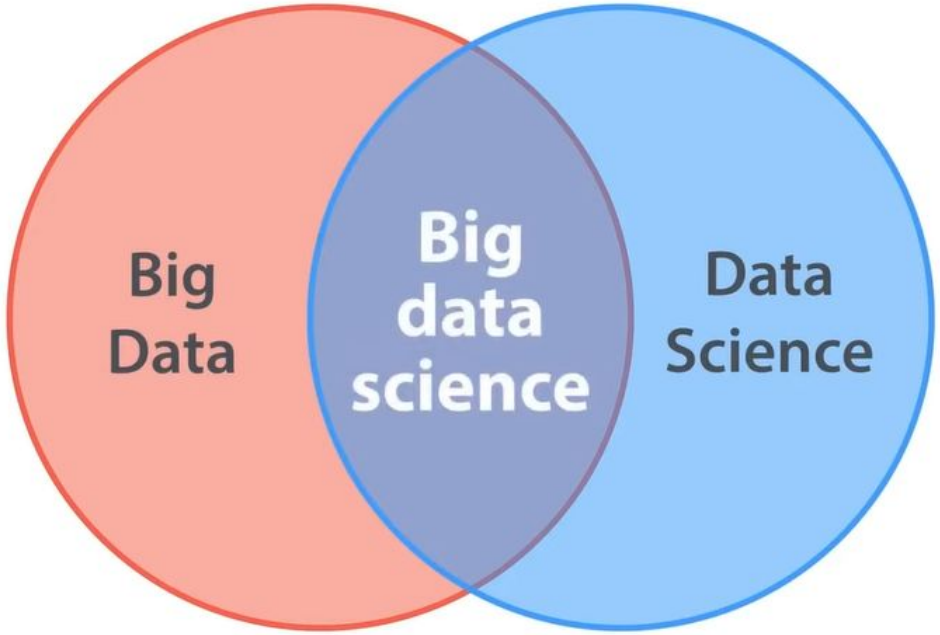


Volume, velocity, and variety

- **Need the full skill set**
- **Coding, statistics, and domain expertise**



Data science vs. Big data



- **Big data ≠ Data science**
- **Some common ground**
- **Big data science unifies**

Data science vs. Coding

Coding

- Task instructions.
- Like a recipe.
- User input
- If, for, while
- Print “Hello, world”

Coding and data

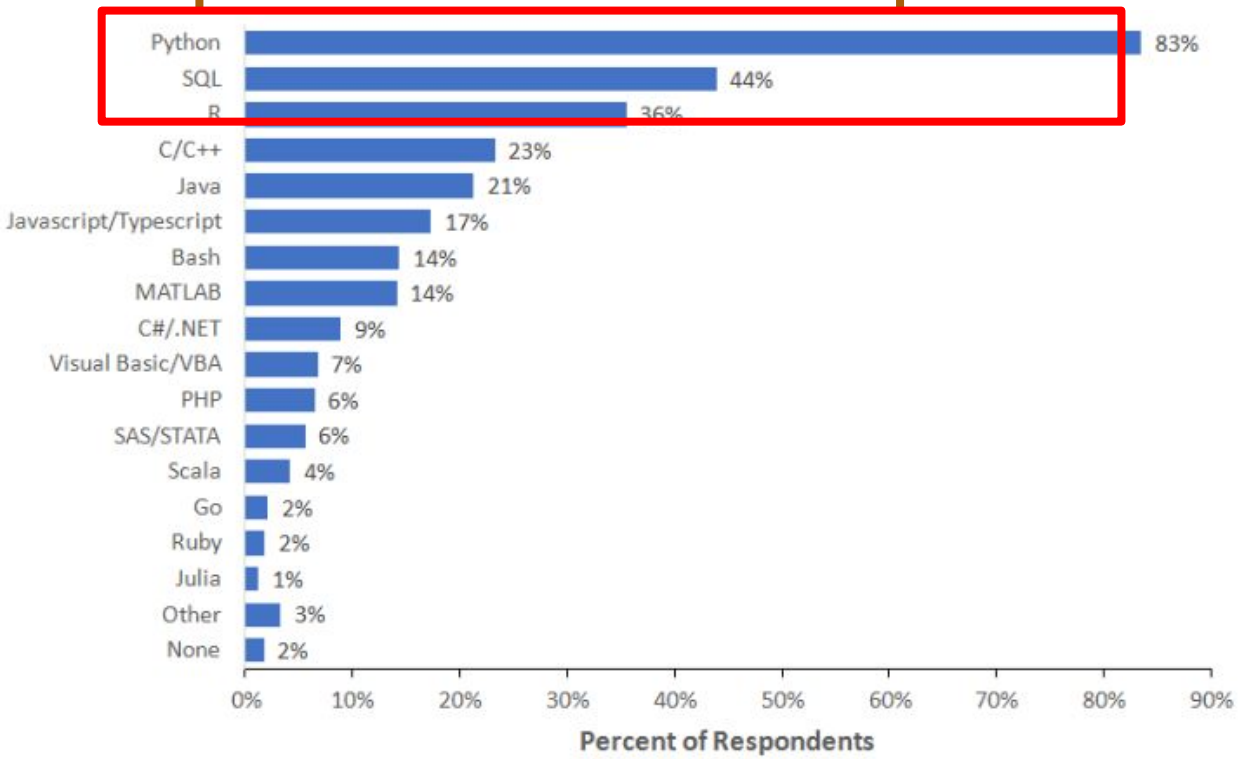
- Conceptually simple
- Domain expertise and math/stats not vital

Data science vs. Coding

Tools for Coding

Rank	Language	Type	Score
1	Python		100.0
2	Java		96.3
3	C		94.4
4	C++		87.5
5	R		81.5
6	JavaScript		79.4
7	C#		74.5
8	Matlab		70.6
9	Swift		69.1
10	Go		68.0

Tools for data science



https://spectrum.ieee.org/computing/software/the-top-programming-languages-2019

Data science vs. Coding

Tools for Coding

- Task instructions.
- Like a recipe.
- User input
- If, for, while
- Print “Hello, world”

Tools for data science

- Word counts
- Conceptually simple
- Domain expertise and math/stats not vital

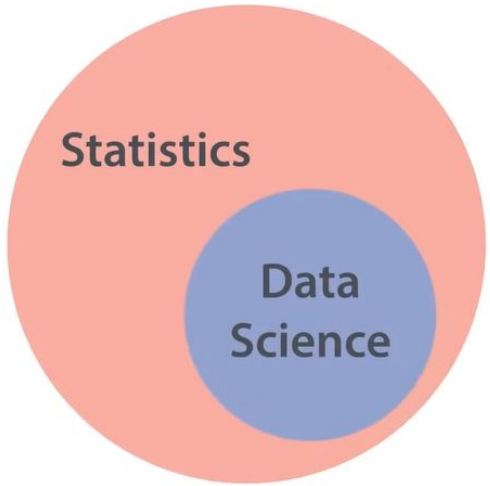
- To make valid inference and generalizations
  - in the face of variability and uncertainty,
    - you need statistics

Need DATA SCIENCE

Data science vs. Statistics

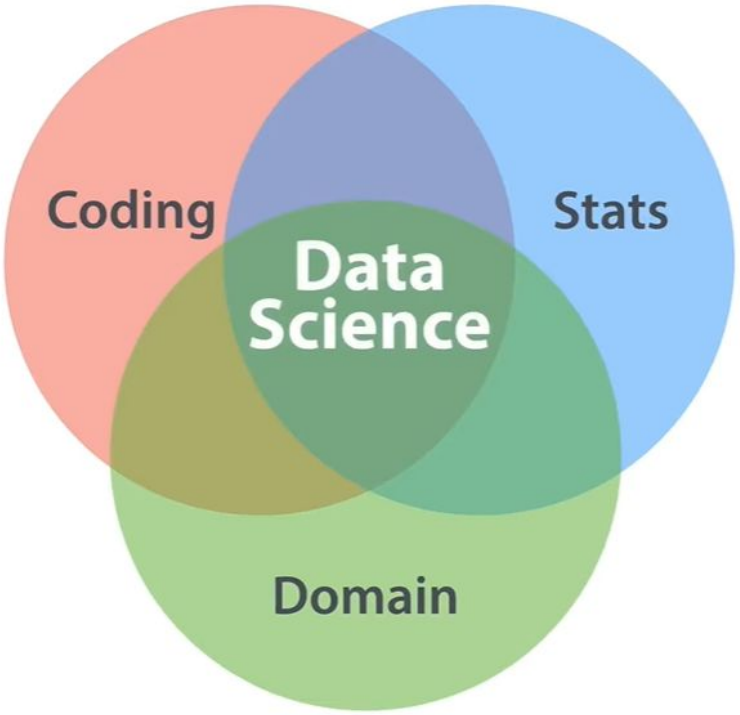
All data scientists do not have stats knowledge

- **NOT TRUE**



All data scientists would first be statisticians

NOT TRUE



Both share similarities but different

Data science vs. Statistics (Differences)

Training

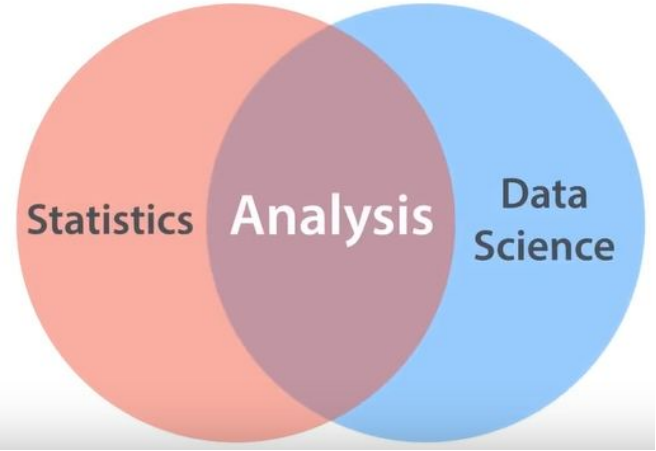
- **Most data scientists are not trained as statisticians.**

Practice

- **Machine learning and big data are not shared with statistics.**

Context

- **Data scientists work in different settings.**



- **DS and Stats both use data**
  - **Different backgrounds**
- **Different goals and contexts**

Data science ethics

Privacy

- Confidentiality
- Should not share
- Sources not intended for sharing

Anonymity

- Not hard to identify
- HIPAA
- Proprietary data may have identifiers

Copyright

- Scraping data is common and useful
- Webpages, PDFs, images, audio, etc.
- Check copyright

Data Security

- Keep data safe
- Make sure data remains anonymous

Data science ethics

Potential Bias

- **Algorithms are only as neutral as the rules and data that they get**

Overconfidence

- **Analyses are limited simplifications; still need humans in the loop**

- **DS has potential and risks**
- **Analyses can't be neutral**
- **Good judgement is vital**