Data 100: Principles and Techniques of Data Science

John DeNero[1] Sandrine Dudoit[2]

[1]Department of Electrical Engineering and Computer Sciences, UC Berkeley

[2]Department of Statistics and Division of Biostatistics, UC Berkeley

Spring 2019

Outline

1. Learning Objectives

2. Practical Matters

3. Data-Enabled Inquiries
   - 3.1 Examples
   - 3.2 General Aspects
   - 3.3 The Data Landscape

4. Defining Data Science
   - 4.1 Data Science Workflow
   - 4.2 A Science of Data Science

5. Computing Language for Data Science

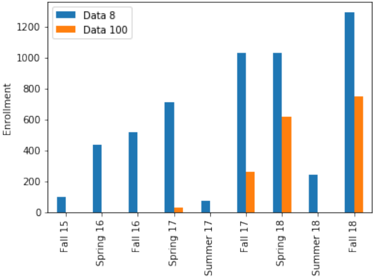
Version: 22/01/2019, 16:54

Acknowledgments

Sections 3–5 are based on discussions and a manuscript in preparation with Duncan Temple Lang, Professor, Department of Statistics, Director, Data Science Initiative, and Associate Dean, Graduate Studies, UC Davis.

Data Science Everywhere

- Data Science (DS) is ubiquitous, whether in academia, industry, or the media.
- Just at Berkeley:
  - Explosion in enrollment for Data 8 and Data 100.



Data Science Everywhere

  - New Data Science Major since Fall 2018 (`https://data.berkeley.edu/degrees/data-science-ba`): 294 declared majors so far.
  - New Division of Data Science and Information since Fall 2018 (`https://data.berkeley.edu/news/berkeley-announces-transformative-new-division`).
- Why so much interest and excitement?

Opportunities for Data-Enabled Discovery

- There are unprecedented and overwhelming opportunities to collect (e.g., drones, sensors, smart devices), access/disseminate (e.g., WWW, databases), and analyze (e.g., statistics/machine learning, high-performance computing, visualization) massive amounts of data.
- Most importantly, we are in a position to answer new types of questions and in new ways, with data collected from novel types of measurement processes.
- There are possibilities and promises to "learn from data" in every field of inquiry, from basic sciences to day-to-day activities, e.g., astronomy, environment, health, justice.
- Data Science is what enables us to seize these opportunities.

Opportunities for Data-Enabled Discovery

- As we will argue and experience in this course, DS is a new discipline, in its own right.
- In particular, it is fundamentally distinct from Computer Science (CS) and Statistics, but it leverages and complements them in essential ways.
- DS is a rapidly evolving field, pushed forward by novel research areas and technologies and building on the foundational disciplines of CS and Statistics.
- In pursuing a career in DS, you will be constantly learning!

Learning Objectives

The broad goal of the course is for you to develop, combine, and apply computational and inferential reasoning to address data-enabled questions.

This involves the ability to participate in the design and implementation of data-enabled workflows ▶ and includes the following.

- Framing and translating a possibly vague domain question into a data-enabled question and, if appropriate, a statistical inference question (i.e., estimation or testing).
- Identifying and acquiring the relevant data.
- Becoming a wise and effective "creator" / "maker" and "reader" / "viewer" of data visualization.
- Performing exploratory data analysis (EDA) for:
  - Data quality assessment/control (QA/QC).

Learning Objectives

- Data cleaning.
- Checking code.
- Understanding the main features of the data (good and bad).
- Revealing patterns.
- Suggesting new theories, models and further questions.

- Becoming a wise and effective "consumer" of statistical inference methods, i.e., appropriately selecting and applying these methods.
  - Making sure the methods actually answer the question vs. focusing on their mathematical and computational details (i.e., what's under the hood).
  - Understanding their scope (what they do), assumptions, and limitations.
  - Interpreting and validating their results.

Learning Objectives

- Presenting, translating, and interpreting the data analysis results back to the domain with the stakeholders.
- Implementing and applying the various steps of a DS workflow using appropriate and reliable software.
- Developing and adopting good scientific practice, including computational reproducibility.
- Developing the ability to evaluate data-driven analyses performed by others.

Learning Objectives

- Use common sense (not so common!) and critical thinking.
- Develop and refine your own judgment. Do not believe everything you read!
- In designing and implementing a Data Science workflow, focus on the question and data (vs. math, models, code).
- Don't forget to "look at data", "get a feel for the data".
- Don't lose the forest for the tree.
- Avoid having a hammer and looking for a nail.
- Avoid reinventing the wheel.
- Keep things simple and transparent. Sophisticated models and p-values are not answers to everything and can sometimes even be plain wrong!

Learning Objectives

Figure 1: *Shadok: Why do it the simple way when you can do it the hard way?* http://leocat.free.fr/shadok/generalites/.

Why Co-Teach?

- Reflect the multidisciplinary/interdisciplinary/transdisciplinary and collaborative nature of Data Science.
- Combine expertise in Computer Science and Statistics, two pillars of Data Science.
- Contribute different backgrounds and perspectives, e.g., perhaps different views of Data Science.
- Data Science is a novel discipline and very few, if any, faculty are trained in Data Science per se. You are the first generation being trained specifically as data scientists. We are in unchartered territory!

Staff

- Faculty Instructors.
  - John DeNero, Associate Teaching Professor, Department of Electrical Engineering and Computer Sciences.
    `http://denero.org`.
    E-mail: `denero@berkeley.edu`.
    Office hours: Monday, 9–10 am, and Wednesday, 10–11 am, 781 Soda Hall. Also by appointment (`denero.org/meet.html`).
  - Sandrine Dudoit, Professor, Department of Statistics and Division of Biostatistics.
    `http://www.stat.berkeley.edu/~sandrine`.
    E-mail: `sandrine@stat.berkeley.edu`.
    Office hours: Monday, 1–2 pm, and Thursday, 2–3 pm, 327 Evans Hall.
- Teaching Assistants.

Staff

- Ananth Agarwal.
- Philippe Boileau.
- Sasank Chaganty.
- Ashley Chien.
- Dan Crankshaw.
- Aman Dhar.
- Hatim Ezbakhe.
- Manana Hakobyan.
- Tony Hsu.
- Shrishti (Sona) Jeswani.
- Jinkyu Kim.
- Simon Mo.
- Maxwell Murphy.
- Samir Naqvi.
- Junseo Park.
- Suraj Rampure.
- Neil Shah.

Staff

  - Allen Shen.
  - Sumukh Shivakumar.
  - Janaki Vivrekar.
  - Daniel Zhu.
- See website for staff contact information: http://www.ds100.org/sp19/.

Resources

- Class website. All course materials, including syllabus, lecture notes, homeworks, labs, discussions, vitamins, projects, and references: `http://www.ds100.org/sp19/`.
- Piazza. All communication will be through Piazza: `http://piazza.com/berkeley/spring2019/data100`. Piazza will also be used to provide help with assignments and course concepts. If you have a private question, make a private post. Please do not post code publicly!
- Textbook. Lau et al. (2019), `http://www.textbook.ds100.org/`.

Prerequisites

- Official prerequisites.
  - Completion of Data 8 (Foundations of Data Science).
  - Completion of CS 61A (Structure and Interpretation of Computer Programs), CS 88 (Computational Structures in Data Science), or ENGIN 7 (Introduction to Computer Programming for Scientists and Engineers). We strongly recommend either CS 61A or CS 88.
  - Co-enrollment in linear algebra course EE 16A, MATH 54, or STAT 89A.
  - The official prerequisites will not be enforced in terms of enrollment. However, we expect knowledge from the official prerequisites for the course itself.
- Recommended: Python, Jupyter Notebook (`jupyter.org`).

Prerequisites

- If you have significant trouble with Discussion 1 or Homework 1, you should consider taking the course another semester.

Grading Policy

Data 100.

- 20% for weekly homework, two lowest scores automatically dropped.
- 10% for weekly labs, two lowest scores automatically dropped.
- 5% for weekly vitamins, which are short online quizzes about the week's lectures.
- 15% for projects (two or three).
- 20% for two midterm exams (10% each), February 28 and April 11, during class.
- 30% for final exam, Thursday, May 16, 11:30 am – 2:30 pm.

Grading Policy

Data 200. (Data 100 students may opt for the Data 200 grading scheme.)

- 20% for weekly homework, two lowest scores automatically dropped.
- 15% for projects (two or three).
- 15% for final project.
- 20% for two midterm exams (10% each), February 28 and April 11, during class.
- 30% for final exam, Thursday, May 16, 11:30 am – 2:30 pm.

Grading Policy

- If you miss a midterm for a pre-approved reason, such as a class conflict or illness, then you won't make it up, but instead your other midterm and the final exam will be used to compute your overall exam score (out of 50%).
- No late homework, lab, or vitamin, unless you are granted an extension by a staff member.
- Late projects will incur a score penalty of 20% per day late. More than a few minutes late means 1 day late.
- Collaboration. It is OK to discuss problems with friends. Please list their names at the top of your assignments (homework, lab, vitamin, or project). You must write your solutions individually; do not copy any other student's work!

Grading Policy

- If you need special accommodations for exams or any other aspects of the class, please make sure to contact the Disabled Students' Program (DSP; `https://dsp.berkeley.edu`) as soon as possible.

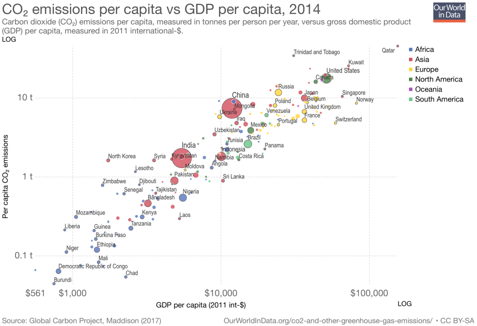
Waitlist

- Advice posted on Piazza: `https://piazza.com/class/jq38w3bffphe9?cid=13`.
- For Data 100, as of this morning: 850/850 enrolled, 170 waitlisted.
- For Data 200, as of this morning: 50/50 enrolled, 19 waitlisted.
- There are no plans to increase enrollment.
- There is typically a 10% drop rate for upper-division EECS courses. If your position on the waitlist is beyond 100, it's best to assume that you won't get into the class and make sure you have an alternative.
- You're welcome to come to class meetings and do the assignments if you're still on the waitlist.

Examples of Data-Enabled Inquiries

- Climatology and meteorology.
  - Is climate change “real”?
  - What are the causes of climate change?
  - What is the impact of climate change on agriculture, the economy, public health, violent conflicts?
  - When, where, and how strong the next floods, hurricanes?
  - What are relevant data for these questions, i.e., variables to measure?

Examples of Data-Enabled Inquiries



CO$_2$ emissions per capita vs GDP per capita, 2014

Carbon dioxide (CO$_2$) emissions per capita, measured in tonnes per person per year, versus gross domestic product (GDP) per capita, measured in 2011 international-$.

Source: Global Carbon Project, Maddison (2017) OurWorldInData.org/co2-and-other-greenhouse-gas-emissions/ • CC BY-SA

Figure 2: *Climatology: Our World in Data.* https://ourworldindata.org/grapher/co-emissions-per-capita-vs-gdp-per-capita-internation time=2014.

Examples of Data-Enabled Inquiries

- Computational linguistics.
  - Computational linguistics is broadly concerned with understanding how language is used and relies on computational and statistical methods to study linguistic phenomena. It involves natural language processing (NLP) and has wide-ranging applications.
  - Speech recognition (e.g., Apple's Siri), speech synthesis, spellchecking, machine translation (e.g., Google Translate).
  - Document retrieval. How to retrieve relevant documents in online searches.
    E.g. PubMed (`https://www.ncbi.nlm.nih.gov/pubmed/`).
  - Social media mining. How to mine user-generated content on social media to extract information about users, e.g., for marketing, security, public health/epidemiology, setting up filters.

Examples of Data-Enabled Inquiries

- What are the issues raised by the application of such algorithms, e.g., in terms of ethics, privacy, security, and governance?
- How do we develop/train and assess the accuracy of such algorithms?

- Crowd size.
  - How can we estimate the size of a crowd?
  - Using images, Twitter feeds, Facebook or Instagram check-ins?
  - E.g. Trump presidential inauguration, Womens' March, March for Science, March for Life, Gilets Jaunes demonstrations.

Examples of Data-Enabled Inquiries

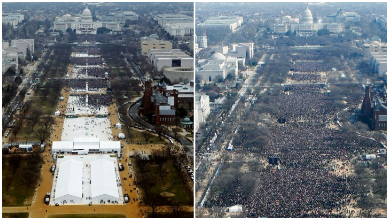


Figure 3: *Crowd size: Trump (left) and Obama (right) presidential inaugurations.*

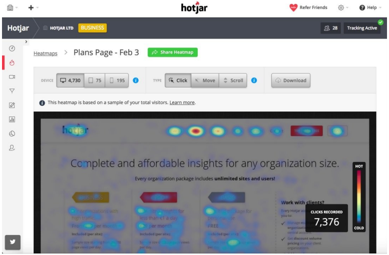
Examples of Data-Enabled Inquiries

- Death toll. How do we estimate death toll from natural disasters or conflicts?
  E.g. Hurricane Maria death toll controversy (https://en.wikipedia.org/wiki/Hurricane_Maria_death_toll_controversy), with a hundred-fold variation in estimates; Syrian Civil War death toll.
- Fake news. How do we assess the credibility of news articles and organizations?
  E.g. Public Editor (https://goodlylabs.org/pe.html).
- Immigration. What is the impact of immigration on the economy, on crime? How does one measure immigration, the economy, crime?
- Internet marketing.

Examples of Data-Enabled Inquiries

- In Internet marketing, conversion rate optimization (CRO), a.k.a., website optimization, is concerned with increasing the percentage of visitors to a website that take a desired action, e.g., convert into customers.
- A typical question in CRO is to determine which features of a website (e.g., landing page, headline, call to action, logo, search and navigation bar) increase the conversion rate.
- Two versions of the website ("A" and "B", respectively) are typically compared using A/B testing (e.g., two-sample z-test, two-sample t-test), where a subset of visitors are randomly split into two groups, one being directed to the A version and the other to the B version, and the conversion rates of the two groups are compared.

Examples of Data-Enabled Inquiries



This heatmap is based on a sample of your total visitors. Learn more.

Complete and affordable insights for any organization size.

Every organization package includes unlimited sites and users!

Work with clients?

Figure 4: *Internet marketing: Click heatmap for CRO.* https://www.hotjar.com/tour.

- Job market.
  - What is the distribution of salaries by field, job title, geography, experience, etc.?
  - What skills/technologies are required/optional in job descriptions?

Examples of Data-Enabled Inquiries

- What should somebody learn to get a better paying job?
- Should they get an extra degree or certification?
- What is the likely gain in pay relative to cost?
- How has the job market changed over time?

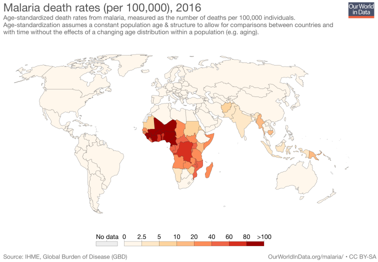
- Justice.
  - The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm, developed by the company Northpointe (now equivant), predicts recidivism risk based on variables related to criminal history, drug involvement, and juvenile delinquency.
  - It is used by US courts for case management.
  - What are the issues raised by the application of such algorithms, e.g., in terms of ethics, privacy, security, and governance?
  - Is the COMPAS algorithm accurate?
  - Is it racially biased?

Examples of Data-Enabled Inquiries

- https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm; https://www.nytimes.com/2017/10/26/opinion/algorithm-compas-sentencing-bias.html.

- Malaria research.
  - Much research is devoted to the development of new malaria prevention, diagnosis, and treatment strategies (e.g., genetically-engineered mosquitos, vaccine).
  - What are the issues raised by such interventions, e.g., in terms of ethics, privacy, security, and governance?
  - How do we assess the effectiveness of the new strategies? Their safety?
  - https://ourworldindata.org/malaria; https://www.gatesfoundation.org/What-We-Do/Global-Health/Malaria.

Examples of Data-Enabled Inquiries



Malaria death rates (per 100,000), 2016

Age-standardized death rates from malaria, measured as the number of deaths per 100,000 individuals. Age-standardization assumes a constant population age & structure to allow for comparisons between countries and with time without the effects of a changing age distribution within a population (e.g. aging).

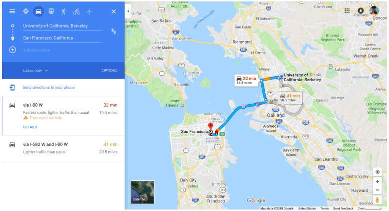
Source: IHME, Global Burden of Disease (GBD)

OurWorldInData.org/malaria/ • CC BY-SA

Figure 5: *Malaria research.*

- Maps and traffic. How to predict travel time between points A and B?

Examples of Data-Enabled Inquiries



University of California, Berkeley

San Francisco, California

Add destination

Leave now ▾ OPTIONS

Send directions to your phone

via I-80 W	32 min
Fastest route, lighter traffic than usual	14.4 miles
⚠ This route has tolls.	
DETAILS	
via I-580 W and I-80 W	41 min
Lighter traffic than usual	20.5 miles

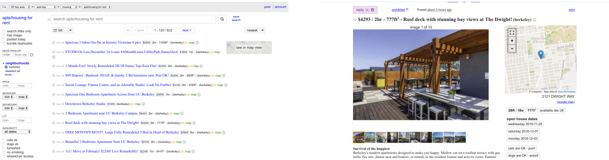
Figure 6: *Maps and traffic: Google Maps.*

- Personalized medicine. How can we guide and tailor disease prevention, diagnosis, and treatment based on individual variables, e.g., genome, exposome, family history?

Examples of Data-Enabled Inquiries

- Politics. How to target posts on social media? How to characterize the prevalence of political lobbying and its effects?
- Real estate.
  - How to find an affordable apartment while studying at Berkeley?
  - What is the relationship between rent/sale prices and mortgage rates, new construction, employment, commute times?
  - What is the influence of demographics, school district, crime, etc., on the real estate market?
  - How does the real estate market vary across time and geography?

Examples of Data-Enabled Inquiries



CL | SF bay area > east bay > housing > apts/housing for rent | post | account

apts/housing for rent

search apts/housing for rent | save search

list | << | < prev | 1 - 120 / 622 | next > | newest

□ search titles only
□ has image
□ posted today
□ bundle duplicates

MILES FROM ZIP: miles | from zip

neighborhoods
☑ berkeley
deselect all
more...

PRICE: min | max

BEDROOMS: min - max

BATHROOMS: min - max

FT²: min | max

AVAILABILITY: all dates

□ cats ok
□ dogs ok
□ furnished
□ no smoking
□ wheelchair access

see in map view

☆ Nov 28 Spacious 2 bdrm/1ba flat in historic Victorian 4 plex $2850 2br - 1100ft² - (berkeley) pic map

☆ Nov 28 STUDIO/In Law,December 1st Lease 4/6MonthLease,UtilityPark,SunnyDeck $1695 1br - (berkeley) pic map

☆ Nov 28 1 Month Free! Newly Remodeled 2B/1B Sunny Top-floor Flat! $2300 2br - (berkeley) pic map

☆ Nov 28 $99 Deposit~ Bushrod- HUGE & Quirky 2 Bd basement unit, Pets OK! $2295 2br - 900ft² - (berkeley) pic map

☆ Nov 28 Social Lounge, Fitness Center, and an Adorable Studio: Look No Further $3193 2br - 678ft² - (berkeley) pic map

☆ Nov 28 Spacious One Bedroom Apartment Across from UC Berkeley $3656 1br - (berkeley) pic map

☆ Nov 28 Downtown Berkeley Studio $1900 (berkeley) pic map

☆ Nov 28 2 Bedroom Apartment near UC Berkeley Campus $5424 2br - (berkeley) pic map

☆ Nov 28 Roof deck with stunning bay views at The Dwight! $4293 2br - 777ft² - (berkeley) pic map

☆ Nov 28 FREE MONTH'S RENT*: Large Fully Remodeled 5 Bed in Heart of Berkeley $5995 5br - (berkeley) pic map

☆ Nov 28 Beautiful 2 Bedroom Apartment Near UC Berkeley $5720 2br - (berkeley) pic map

☆ Nov 28 1x1! Move in February! $2248! Live Remarkably! $2248 1br - 541ft² - (berkeley) pic map

reply | □ prohibited | Posted about 3 hours ago | print

☆ $4293 / 2br - 777ft² - Roof deck with stunning bay views at The Dwight! (berkeley)

image 1 of 10

© craigslist - Map data © OpenStreetMap

2121 DWIGHT WAY
(google map)

2BR / 1Ba | 777ft² | available dec 28

open house dates
wednesday 2018-11-28
saturday 2018-12-01
monday 2018-12-03

cats are OK - purrr
dogs are OK - wooof

Survival of the happiest
Berkeley's modern apartments designed to make you happy. Mellow out on a rooftop terrace with gas grills, fire pits, dining area and heaters; or mingle in the resident lounge and activity room. Pamper

Figure 7: *Real estate: Craigslist listing.*

- Urban planning. Where should we put docking ports for bikes?

Examples of Data-Enabled Inquiries

- Not all Data Science projects start with a question.
- It can be quite fruitful to let the data suggest the question, e.g., by performing exploratory data analysis on publicly available data.
- Two examples of the data at your fingertips to learn about the World we live in:
  - Gapminder: `https://www.gapminder.org`.
  - Our World in Data: `https://ourworldindata.org`.
- Data Science can of course address a much broader range of questions, from the very specific to the very general.

Examples of Data-Enabled Inquiries

What questions are YOU interested in?

General Aspects of Data-Enabled Inquiries

- The above are interesting and important practical problems that clearly require or benefit from data to be answered.
- However, the questions are rather vague, some more than others, and it is not clear exactly what the relevant data are.
- This raises two essential and challenging aspects of data-enabled inquiries, that are typically not addressed by existing disciplines or curricula: Framing questions and identifying the data, i.e., what to measure, in order to answer the question.
- Addressing a data-enabled question is a highly exploratory, interactive, and iterative process, where the questions, data, analyses, and answers are gradually refined.

General Aspects of Data-Enabled Inquiries

- We do not necessarily know what we will do next until we see the results of the current step and we often return to one or more of the previous steps based on information we discover in the current step.
- There are trial-and-error and sleuthing aspects to DS.
- In addition to the computational and statistical challenges and potential domain insights, some of these studies have serious "real-life" implications (e.g., put patients at risk, jail innocent individuals), raise ethics and governance concerns, and are quite controversial.
- Although covering vastly different domains, there are commonalities between these problems, in terms of the types of questions, data, and analysis methods. Cf. Science of Data Science.

Getting the Question Right

“Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.”

John Tukey (1915–2000; `https://en.wikipedia.org/wiki/John_Tukey`): Data scientist (way before the term became mainstream!) at Bell Labs. Coined the terms “bit” and “software”, developed the boxplot.

Getting the Question Right

Now let's go back to some of the questions presented earlier to see if we can formulate them more precisely and sketch approaches for addressing them.

The Data Landscape

- The expression "big data" is pervasive, but not particularly well-suited and too narrow to describe the data landscape.
- The data can be "big" in terms of dimensionality, of course, but, more importantly, they are often complex, as detailed below.
- Data are often collected without a purpose, i.e., a precisely formulated driving research question. It can thus be unclear what to do with these data.
- Data are produced at a rate outpacing analysis capabilities as well as ethics and governance standards.

The Data Landscape

- Data relevance, provenance, and reliability. Whenever using data, we should be concerned about
  - whether they are actually useful for answering our question, e.g., when one cannot measure what we need and uses proxies or found data instead);
  - where they come from;
  - whether they can be trusted.

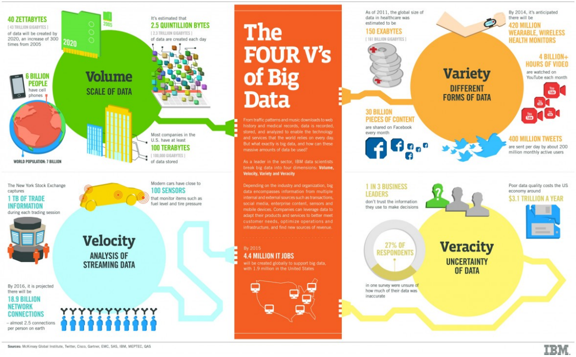
The Data Landscape

- High-dimensional data. Large sample sizes (n) and/or numbers of variables (p).
- Different types of data. Quantitative (continuous, discrete), qualitative, text, graph (i.e., edges and vertices), image, sound.
- Streaming data. Data continuously generated by different sources, e.g., sensors.
- Censored, missing, or sparse (cf. zero inflation) data.
- Various levels of data processing and formats. E.g. High-throughput genome sequencing data: Images (TIFF files), base calls (FASTQ files), read counts (CSV files).
- Multiple data sources and locations. In-house, WWW.

The Data Landscape

- Evolving data. E.g. DNA sequence (GenBank), ads on Craigslist, literature (Google Scholar, PubMed).
- Unreliable or erroneous data. Cf. Bad data.
- Four V's of Big Data. Volume, velocity, variety, veracity.

The Data Landscape



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015 **4.4 MILLION IT JOBS** will be created globally to support big data, with 1.9 million in the United States

40 ZETTABYTES [43 TRILLION GIGABYTES] of data will be created by 2020, an increase of 300 times from 2005

6 BILLION PEOPLE have cell phones

It's estimated that **2.5 QUINTILLION BYTES** [2.3 TRILLION GIGABYTES] of data are created each day

Most companies in the U.S. have at least **100 TERABYTES** [100,000 GIGABYTES] of data stored

The New York Stock Exchange captures **1 TB OF TRADE INFORMATION** during each trading session

Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be **18.9 BILLION NETWORK CONNECTIONS** – almost 2.5 connections per person on earth

As of 2011, the global size of data in healthcare was estimated to be **150 EXABYTES** [161 BILLION GIGABYTES]

By 2014, it's anticipated there will be **420 MILLION WEARABLE, WIRELESS HEALTH MONITORS**

4 BILLION+ HOURS OF VIDEO are watched on YouTube each month

30 BILLION PIECES OF CONTENT are shared on Facebook every month

400 MILLION TWEETS are sent per day by about 200 million monthly active users

1 IN 3 BUSINESS LEADERS don't trust the information they use to make decisions

27% OF RESPONDENTS in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around **$3.1 TRILLION A YEAR**

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

Figure 8: *Four V's of Big Data.* `https://www.ibmbigdatahub.com/infographic/four-vs-big-data`.

The Data Landscape

No longer just numerical data, $X_{n \times p}$.

Big bad data $\Longrightarrow$ Garbage in, garbage out (GIGO).

Defining Data Science

- So, what is Data Science, exactly? We've mentioned various aspects of DS, but we haven't defined it precisely yet.
- This is not an easy task, as with other subjects that generate a lot of buzz, such as, "artificial intelligence", "machine learning", and "bioinformatics".
- Definitions of and views on Data Science abound. Although related and sharing common themes, they also greatly vary in emphasis and precision depending on the background of the author and target audience.

Defining Data Science

- Data Science has been broadly defined as "the science of learning from data" (Donoho, 2017)[Section 2.5], with aspects of such a discipline already being discussed more than 50 years ago by Tukey (1962) in the "The Future of Data Analysis".

  Great reads for anyone interested in DS!
- Data Science is also often described as an interdisciplinary and integrative field, in terms of its
  - foundation/pillar disciplines, i.e., Computer Science and Statistics;
  - domain applications, i.e., disciplines to which a DS problem pertains (e.g., Environmental Science, History);
  - societal implications (e.g., ethics, policy).
- Rarely is Data Science presented as a discipline in its own right or its essence and uniqueness precisely defined.

Defining Data Science

- As we will argue and experience in this course, Data Science actually is a new discipline, i.e., it is fundamentally distinct from Computer Science and Statistics, but it leverages and complements them in essential ways.

Defining Data Science

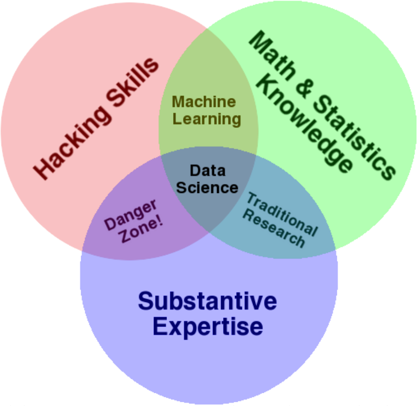


Figure 9: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram.

Defining Data Science

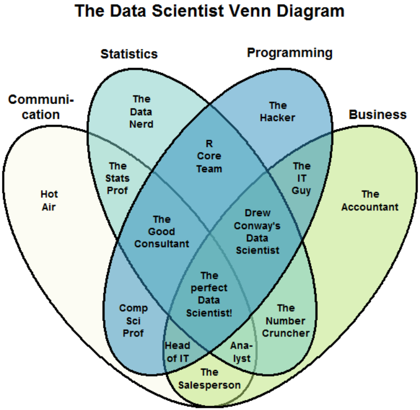


Figure 10: `https://whatsthebigdata.com/2016/07/08/the-new-data-scientist-venn-diagram/`.

Defining Data Science

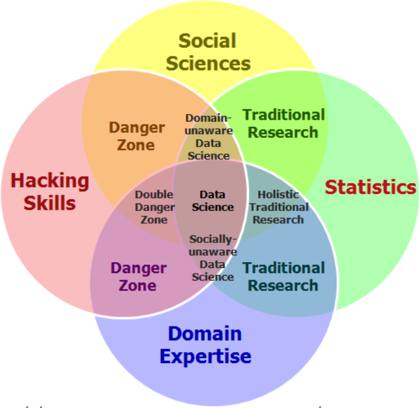


Figure 11: `http://datascienceassn.org/content/fourth-bubble-data-science-venn-diagram-social-sciences`.

Defining Data Science

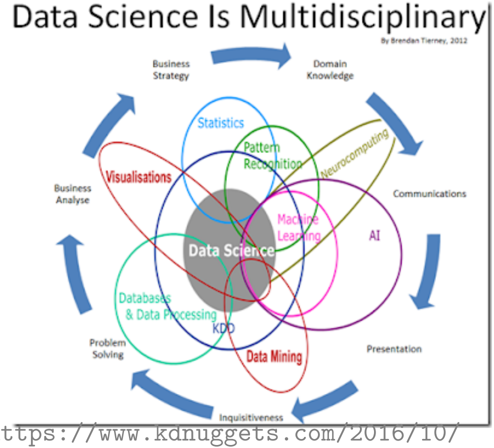
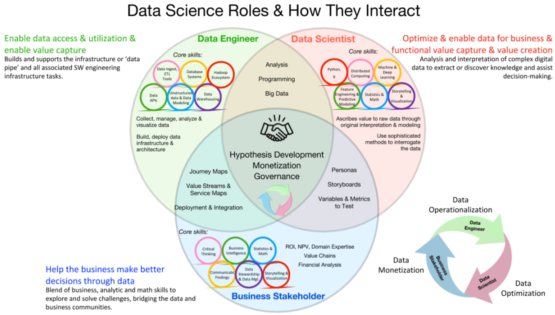


Figure 12: https://www.kdnuggets.com/2016/10/battle-data-science-venn-diagrams.html.

Defining Data Science



Data Science Roles & How They Interact

Enable data access & utilization & enable value capture

Builds and supports the infrastructure or 'data pipe' and all associated SW engineering infrastructure tasks.

Optimize & enable data for business & functional value capture & value creation

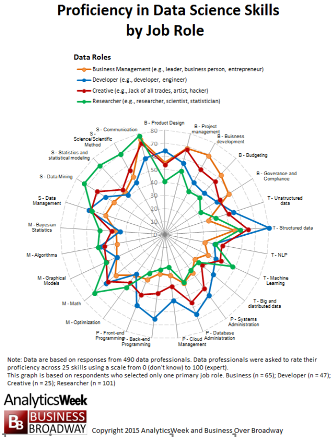
Analysis and interpretation of complex digital data to extract or discover knowledge and assist decision-making.

Help the business make better decisions through data

Blend of business, analytic and math skills to explore and solve challenges, bridging the data and business communities.

Figure 13: https://www.kdnuggets.com/2018/09/winning-game-plan-building-data-science-team.html.

Defining Data Science



Proficiency in Data Science Skills
by Job Role

Note: Data are based on responses from 490 data professionals. Data professionals were asked to rate their proficiency across 25 skills using a scale from 0 (don't know) to 100 (expert).
This graph is based on respondents who selected only one primary job role. Business (n = 65); Developer (n = 47); Creative (n = 25); Researcher (n = 101)

Figure 14: `http://businessoverbroadway.com/2015/09/23/investigating-data-scientists-their-skills-and-team-makeu`

Defining Data Science

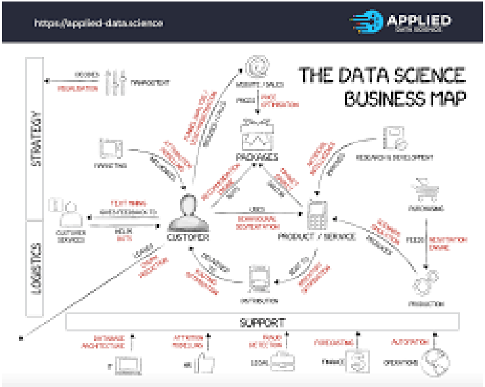


Figure 15: https://medium.com/applied-data-science/every-arrow-on-this-diagram-is-a-data-science-project-775

Defining Data Science

Confused by all these definitions and Venn diagrams? I sure am!

Still want to be data scientist after seeing all of this? ☺

Defining Data Science

Let's step back and think.

Instead of defining Data Science in terms of other disciplines, let's start from scratch and define it in terms of how it is actually practiced.

As a start, perhaps name some data scientists or describe what a data scientist does.

Data Science Workflow

- Data Science is broadly concerned with "using data" (i.e., envisioning, collecting, accessing, and analyzing data) to derive knowledge about a particular domain.
- It involves mapping a domain problem/question to a data-driven workflow that leads to an answer, i.e., insight/meaning/decisions/actions.
- A Data Science workflow includes both sequential and iterative aspects as well as transversal aspects that permeate the workflow.
- In particular, computing with data is a transversal aspect that is crucial throughout any data-enabled inquiry. We will discuss general aspects of computing with data later in this lecture and devote a good portion of the course to this topic.

Data Science Workflow

- Likewise for data visualization.

Data Science Workflow: Sequential and Iterative Aspects

Sequential and iterative aspects of a data-enabled discovery workflow. Depending on the question and data, different versions of the following steps are performed in a sequential and iterative manner to address a data-enabled question.

1. Identifying and framing data-enabled questions.
   - What do we want to know, exactly?
   - What type of answer is required, i.e., how accurate, precise, polished?
   - What are metrics for success, costs for errors?
   - How do we iteratively discuss and refine questions from domain stakeholders and then map these to a data-driven problem?

Data Science Workflow: Sequential and Iterative Aspects

- Very narrowly, in terms of statistical inference (i.e., estimation and testing), this involves defining parameters of interest and performance measures such as risk, Type I and II error rates (i.e., false positive and false negative error rates).
  But Data Science is much broader than this and does not always involve statistical inference.
- Also, recall that not all Data Science projects start with a question. It can be quite fruitful to let the data suggest the question, e.g., by performing exploratory data analysis on publicly available data.

2. Imagining and identifying the nature of the data[1] needed in order to answer the question of interest.
   - This involves mapping a concept to potential variables (quantitative or qualitative), i.e., determining
     - what to measure,

Data Science Workflow: Sequential and Iterative Aspects

- the sources of available data, and
- the technologies for capturing new data.

- We may need to generate data of a novel nature through new types of measurement processes, e.g., by deploying sensors.
- We may leverage existing available data, e.g., through Webscraping.

3. Designing the data collection procedure, e.g., survey/questionnaire, sampling scheme, or randomized controlled experiment.
4. Collecting the data. Acquiring available data/found data, generating new data, and fusing/merging data sources (e.g., record linkage).
5. Exploratory data analysis (EDA) for:
    - Data quality assessment/control (QA/QC).

Data Science Workflow: Sequential and Iterative Aspects

- Data cleaning.
- Checking code.
- Understanding the main features of the data (good and bad).
- Revealing patterns.
- Suggesting new theories, models, and further questions.

6. Data pre-processing.
   - Data cleaning.
   - Data filtering, i.e., removing observations and/or variables.
   - Data transformation, e.g., taking logs, centering and scaling.
   - Data calibration/normalization, i.e., adjusting for nuisance effects (e.g., different runs of an instrument) to allow comparisons across observations and/or variables.
   - Data imputation.
   - Dimensionality reduction.

Data Science Workflow: Sequential and Iterative Aspects

Pre-processing should not be taken lightly. It involves non-trivial decisions that can have a large impact on the final results. In particular, it is not uncommon for ad hoc pre-processing decisions to have a greater impact on the results than the choice of formal statistical inference method (only conditionally optimal given the data that are provided to it).

7. Answering the question.
    - Numerical and graphical summaries of the data, i.e., statistics. – Can often be sufficient or even the only appropriate type of answer.
    - Optimal statistical inference. – Not always applicable or needed.
        - This is the main step where statistical inference/machine learning (ML) come into play [2].

Data Science Workflow: Sequential and Iterative Aspects

        - Given the statistical formulation of the question (i.e., parameter of interest) and the data, provide optimal estimator/predictor/test, i.e., get the most information out of the data given available resources.
        - This involves reporting performance measures such as standard errors, risk, Type I and II error rates, and model diagnostics and comparison.
    - Assessing the accuracy and robustness of the answer. E.g. Using resampling (bootstrap, cross-validation), simulation, control observations and/or variables.

8. Translating the results back to the domain.
    - Presenting, translating, and interpreting the data analysis results back to the domain with the stakeholders.
    - Validating the results (cf. relevance, robustness).
    - Implementing the results into decisions and actions.

Data Science Workflow: Sequential and Iterative Aspects

9. Deploying and disseminating the Data Science scholarship.
   - Providing data-driven evidence justifying the insights and integrating the supporting reproducible and verifiable aspects of the workflow.
   - Implementing the DS workflow on a website for ongoing evolution and updating.
   - Allowing continual evaluation by stakeholders, e.g., with dashboards.

[1] We chose not to use the expression "type of data" here, due to possible confusion about its meaning. Indeed, depending on the context, "type of data" could refer to different properties of data. Here, it would mean"what the data measure", e.g., temperature, crime, blood pressure. Whereas in Statistics it could refer to whether the data are continuous or discrete and in CS to whether the data are integer/double, tree, or hash.

[2] This step includes Tukey's confirmatory data analysis (CDA).

Data Science Workflow: Sequential and Iterative Aspects

- The above sequential steps range from conceptual to analytical.
- Importantly, the workflow is highly interactive and iterative, i.e., each of the previous steps can be updated and refined using insights from the current step. Expect to revisit each step several times and be curious if this doesn't happen.
- Researchers often perform these steps intuitively and simultaneously, without clearly separating them from each other or the techniques involved in each.
- This makes DS in some sense an art or at least very hard to convey to new researchers.
- Furthermore, there are essential aspects and technologies that span all steps.

Data Science Workflow: Transversal Aspects

Transversal aspects of a data-enabled discovery workflow. The following related aspects are integral and essential to the workflow, i.e., they apply throughout a data-enabled inquiry (with varying emphasis depending on question and data).

- Team science, project management, and communication. Build the right team and integrate the right skills to design and implement the workflow. Ensure proper communication and feedback throughout.
- Domain context, knowledge, and considerations.
- Computing with data.
- Data visualization.
- Data technologies. E.g. For data acquisition, management, curation, visualization, dashboards.

Data Science Workflow: Transversal Aspects

- Computational reproducibility and verifiability. ▸
- Research responsible conduct and integrity, ethics, privacy, security, and governance. ▸

DS Workflow: Computing with Data

- Addressing a data-enabled question requires a different type of computing than is typical in programming and software engineering courses.
- Computing with data is exploratory and adaptive high-level computing.
- Given the exploratory and iterative nature of Data Science, we do not typically write full-blown programs or applications right away.
- Instead, we tend to implement small tasks to go from one step to the next and examine current data and results (e.g., using visualization) to dictate our next move.
- Also, we often return to one or more of the previous steps based on information we discover in the current step.

DS Workflow: Computing with Data

- This calls for an interactive computing language.
- We also may use different languages at different points in the workflow and for different purposes.
- Once we have refined the workflow, we may of course need to implement the analysis sequence into a polished and efficient software product.
- We will get back to computing later in this ▶ and upcoming lectures.

DS Workflow: Data Visualization

- Data visualization is a fundamental aspect of Data Science.
- It is essential to "look at data" throughout the workflow, from EDA to model diagnosis and reporting the results of the inquiry.
- An effective plot can be good enough to answer the question on its own. In some cases, it may even be the only appropriate type of answer.
- An effective plot can also be sufficient to convince stakeholders of the findings from a full-blown statistical inference procedure.

DS Workflow: Data Visualization

- Graphs are by definition functions of the data, i.e., statistics. Although not typically viewed this way, visualization can therefore be used as part of statistical inference.
- One can produce the same types of plots for a sample and for a population, in that sense, the plot for the sample can be viewed as an estimator of the plot for the population, i.e., the parameter.
- A pattern that we detect from plotting data for a sample can be used to infer properties of the population from which the sample was drawn. A formalized special case of such an approach is given by linear regression.
- We will discuss visualization in greater detail in upcoming lectures.

DS Workflow: Computationally Reproducible and Verifiable Research

- Computational reproducibility refers to the ability to regenerate, given the same input data, all of the computational output/results reported in a study/publication, e.g., tables, figures.
- Succinctly put, computational reproducibility is good research practice, in line with the scientific method which calls for direct evidence to support scientific claims: *Nullius in verba*, i.e., "take nobody's word for it".
- Computational reproducibility allows us and our collaborators to verify that our results are correct and get help if they aren't. It also allows independent verification by other investigators.

DS Workflow: Computationally Reproducible and Verifiable Research

- Computational reproducibility involves collecting into a script the actual commands we used for each step in the workflow leading to our final results. We typically omit the trial-and-error or the paths we started but didn't pursue to an end, as these are not of consequence to our final results [3].
- Additionally, we should ideally programatically generate the document which reports our findings, including tables and plots, so that we don't have to manually recreate these when the data or analyses change.

DS Workflow: Computationally Reproducible and Verifiable Research

- Compendia are integrated and executable input documents, with text, data, code, and software, that allow the generation of dynamic and reproducible output documents, intermixing text and code output (textual and graphical) (Gentleman and Temple Lang, 2004). Different views of the output document (e.g., PDF, HTML) can be automatically generated and updated whenever the text, data, or code change.
- Compendia are modern versions of the lab book.
- With compendia, we avoid the extra labor and error-prone manual intervention involved in recreating the plots and tables and keeping the numbers in the text up-to-date when either the data or code change.

DS Workflow: Computationally Reproducible and Verifiable Research

- Compendia allow us to link all of the evidence for the scholarship (i.e., data, text, and code) into one document.
- Examples of systems for generating dynamic and reproducible documents are: Jupyter Notebook for Python and R (`jupyter.org`); Sweave and knitr for R; DocBook-based XML.
- Note that this very narrow and "light" definition of reproducibility is often confused with the much more ambitious notion of scientific reproducibility, e.g., whether the biological findings from one study hold in another (that collected different data).

DS Workflow: Computationally Reproducible and Verifiable Research

- Notions of reproducibility actually go all the way back to the Middle Ages! Roger Bacon (c. 1214–1294) calls for a repeating cycle of observation, hypothesis, and experimentation, and the need for independent verification. The manner in which experiments are conducted should be recorded in precise detail so that others can reproduce and independently test results.

[3] This could be problematic if the results were the reason we didn't pursue a path. With a different dataset, we would get different results and maybe follow that path and not the one we did follow.

DS Workflow: Research Responsible Conduct and Integrity, Ethics, Privacy, Security, and Governance

- Although distinct concepts, research responsible conduct and integrity, ethics, privacy, security, and governance are closely intertwined in the DS workflow.
- Depending on the question and data, one should take appropriate measures to ensure data privacy and security.
  - FERPA (Family Educational Rights and Privacy Act, 1974) is a federal law that governs the access of educational information and records.
  - HIPAA (Health Insurance Portability and Accountability Act,) is a federal law providing data privacy and security provisions for safeguarding medical information and records.

DS Workflow: Research Responsible Conduct and Integrity, Ethics, Privacy, Security, and Governance

- One should constantly be vigilant about and able to recognize when ethical issues arise, that is, whether a course of action is “right” or “wrong” (“good” or “bad”) according to “norm”, i.e., the set of values or standards for conduct or practice held by society, members of a group, or individuals.
  - Should we be releasing genetically-engineered mosquitos in the wild in attempt to control malaria?
  - When a variable is very expensive to measure but very effective at predicting response to treatment, is it fine not to use it when training a predictor?
  - Should we be discarding outliers when fitting a model to make inference about a population?

DS Workflow: Research Responsible Conduct and Integrity, Ethics, Privacy, Security, and Governance

- Ethical concerns come into play in how we frame a question and how we define the costs for errors, i.e., loss function. In some cases, one may trade off good and bad, in others one may have to clearly rule out doing something bad.
- Privacy, security, and governance issues arise when we use others scholarship (i.e., data, code, methods), as well as when we consider making our own scholarship publicly available.

  E.g. When Webscraping, there are terms of service constraining our use of data.

DS Workflow: Research Responsible Conduct and Integrity, Ethics, Privacy, Security, and Governance

- Research responsible conduct and integrity involves ethics, privacy, security, and governance. Different professions and organizations have varying, but related definitions and guidelines.
  E.g. `https://ori.hhs.gov/education/products/ucla/default.htm`.

Data Science Workflow

Again, let's go back to some of the questions presented earlier to see what a sensible workflow might look like.

In Homework #1, we ask you to describe, for a question of your choice, your approach and thought process in addressing this question.
This is more about ***what*** *to do, than* ***how*** *to do it exactly.*

Defining Data Science

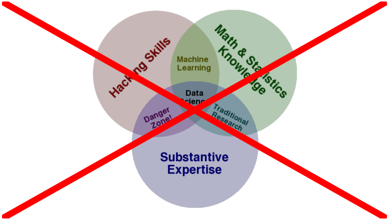
- Data Science involves an entire translational and transdisciplinary data-enabled problem-solving/discovery workflow.
- Data Science is translational in that it starts from a domain question and “uses data” (i.e., envisioning, collecting, accessing, and analyzing data) to derive knowledge about that particular domain. It also has potential implications/impact beyond the domain, on society and individuals.
  E.g. Malaria research.
- Data Science is transdisciplinary/transversal (vs. cross/inter/multidisciplinary), in the sense that it goes beyond the limits of existing disciplines.

Defining Data Science

- Computational, mathematical, and statistical foundations, as well as domain knowledge, are necessary, but not sufficient for Data Science.
- CS and Statistics are means for DS, not goals.
- Data Science emphasizes very different aspects of existing disciplines, such as Statistics, CS, and domain sciences.
- These disciplines may claim to do or be Data Science, but in fact they each only concern very specific aspects of DS.
- There are cultural differences between CS and Statistics in terms of the emphasis they place on different aspects of the workflow.
- Attempts to define DS in terms of (set operations of) other disciplines are limiting.

Defining Data Science

- In particular, DS is not the intersection of Statistics, Computer Science, and a domain science, as in Drew Conway's popular Data Science Venn diagram (http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram).



In fact, what would this intersection be? If anything, DS would be more like a union of these disciplines; we want to

Defining Data Science

be able to use all of CS and Statistics to address a DS question.

Defining Data Science

What is a data scientist?

- The label data scientist is used to refer to individuals with vastly different backgrounds, expertise, and objectives, e.g., a programmer in industry, an academic biologist generating and analyzing their own data.
- Data scientists, self-proclaimed or not, come in many flavors; it can be unclear what you are getting.
- Cf. The variety of descriptions (and diagrams!) you get when you search for “data scientist” on Google.
- There is a depth-breadth trade-off and a risk of being a jack-of-all-trades-master-of-none.

Defining Data Science

- Many Data Science problems are best addressed by having breadth of depth, i.e., teams of experts in relevant disciplines, speaking a common language and in constant feedback with each other.
- It is important to be clear about your strengths and limitations (not a bad thing, that's why we have collaborators and learn from them!), for your own benefit and that of others you work with or who are looking to hire you.

A Science of Data Science

- As currently presented and performed, Data Science consists of a collection of foci of activity (Donoho, 2017)/technical areas of work (Cleveland, 2001). It does not qualify as a science with general and unifying foundations and principles.
- A science of Data Science would unify these activities by laying down general principles and theories for the workflow.

A Science of Data Science

- This would involve abstracting from the specifics of a particular data-enabled inquiry and finding common themes across inquiries to derive general approaches and methods to improve the efficiency and accuracy of the workflow.
  E.g. Making code faster without knowing what it does. Applying general principles (across domains) for framing data-enabled questions and for identify relevant data.
- Envisioning a science of DS is in line with Tukey (1962)'s call for a "science of data analysis", more than 50 years ago!
- It is also hinted at in Donoho (2017).

Computing Language for Data Science

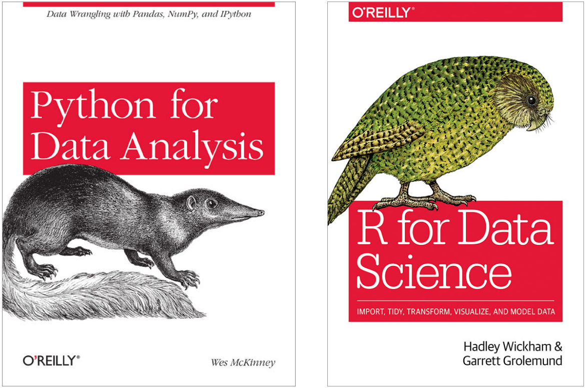

Data Wrangling with Pandas, NumPy, and IPython
Python for Data Analysis
O'REILLY®
Wes McKinney
O'REILLY®
R for Data Science
IMPORT, TIDY, TRANSFORM, VISUALIZE, AND MODEL DATA
Hadley Wickham & Garrett Grolemund

Figure 16: *Computing languages for Data Science.* Python (http://shop.oreilly.com/product/0636920023784.do) and R (https://r4ds.had.co.nz).

Computing Language for Data Science

- The main computing language for this course is Python (`https://www.python.org`). However, this is only one of several languages that are pertinent for Data Science.
- An obvious and equally appropriate alternative is R (`https://www.r-project.org`).
- The same functionality is often implemented in several languages (for efficiency or to take advantage of existing implementations).
- Furthermore, many real-world projects call for a combination of programming languages.
- In the course, we'll also use a reasonable amount of Structured Query Language or SQL, a domain-specific language used for managing data held in a relational database management system.

Computing Language for Data Science

- Although computing is an essential aspect of DS, programming is not the main point of DS. Likewise for statistical methodology and proving theorems.
- The focus of DS should be the question and data; computing and statistical inference are means for answering the question. In particular, good software should be invisible, i.e., a transparent means to an end.
- This is neither a CS nor a Statistics course, so we'll have to set aside some of our programming practices/habits.

Computing Language for Data Science

The criteria for choosing a language for a DS project differ from those in CS and Statistics. It is a matter of context as to which features of a language are favored.

- Properties of the language itself, e.g., general-purpose vs. domain-specific, functional vs. reference-based.
- Available functionality, i.e., packages/modules.
- Trustworthiness/reliability/quality of software, not just of the language itself, but also of the packages/modules. Actually, how can we evaluate the trustworthiness of individual packages/modules? What are good proxies for measuring this? This is a good Data Science question!
- Community testing of the language and packages/modules.

Computing Language for Data Science

- Speed of running code (and number of times it will be run) vs. speed of writing code.
- Extensibility vs. getting the job done as a one-off.
- Open-source, licensing, and cost.
  Open-source is key for learning by studying and modifying existing code, for assessing code, and for trusting code. E.g. Matlab, SAS, SPSS, and Stata are not open-source.
- Availability of knowledgeable programmers in the language.
- Ability to get help.
- Documentation.

Computing Language for Data Science

The following are important considerations when choosing a computing language for a Data Science project.

- Exploratory/Evolutionary aspect of computing.
  - In a DS project, the sequence of data analysis steps is rarely set in stone. Rather, there are exploratory, trial-and-error, and iterative aspects to DS.
  - The analysis, and hence the methods and code, evolve as we iterate through the workflow and collect and look at data.

    E.g. After EDA, we may decide to collect more data, discard some observations, log-transform some of the variables, use regularized regression (instead of standard unpenalized regression), or use dimensionality reduction before clustering.
  - We are context-switching focus between programming and interpreting.

Computing Language for Data Science

- DS therefore calls for a flexible interactive programming environment that allows us to adapt and revise the analysis [4].
- A general-purpose programming language with a read-eval-print loop (REPL) model as in Python and R is invaluable for this purpose.

- Social aspect of computing.
  - Ease of communication with collaborators is an important aspect of DS.
  - If a community works in a particular language, it may not be wise to use another language.
  - Not "speaking the same language" (computing, as well as domain-specific) can lead to miscommunication and setbacks on the project.

Computing Language for Data Science

  - Important practical questions to ask: Who am I working with on the code base now? Who will maintain the code? Who will take over the code?

- Trade-off between development speed and runtime speed. In many DS projects, the speed of running the code is not the primary concern. Instead, it may be more fruitful to focus on the ease and speed of writing pertinent and accurate code in order to complete the project in a timely manner.
- Not reinventing the wheel.

Computing Language for Data Science

- If a language already provides reliable and fast enough implementations of the required methodology, you are just reinventing the wheel by reimplementing the methodology in another language. And bound to make mistakes in the process.
  "Every non trivial program has at least one bug."
- It is perfectly fine to reuse code, as long, of course, as it is reliable and performs the required task.
- The main reasons for rewriting code are the learning process or making it faster.

- Computational reproducibility. ▸
  - The language should be amenable to computational reproducibility, as it is an essential aspect of Data Science.
  - Computation in both Python and R is reproducible.

Computing Language for Data Science

- Computation in Excel is not reproducible. With Excel, it is hard to read what the code is doing, because the code is spread across many cells and often repeated. There is no ordered sequence of computation. Therefore, one cannot trace back the steps leading to results, i.e., the logic cannot be audited or tested.

[4] This is not incompatible with computational reproducibility.

Computing Language: R

- R is an open-source implementation of the S language.
- S was initially developed at Bell Labs by John Chambers (1998 Association for Computing Machinery (ACM) Software System Award) and colleagues, with the first lines of code going back to 1967.
- Aside: Recall that Bell Labs is where Tukey was. It is also where C (Dennis Ritchie) and Unix (Ken Thompson, Dennis Ritchie, and colleagues) were developed.
- R was created in 1992 by Ross Ihaka and Robert Gentleman and has been developed by a core team of about twenty.

Computing Language: R

- R now dominates S+ (commercial implementation of S), not the least because it is open-source and under active development.
- Like Python, R is a general-purpose programming language.
  - Python is used for broader purposes and its community is not as specialized as R's.
  - R was designed more specifically for data analysis and its community is focused on high-quality statistical package.
- R is the *lingua franca* of statistical computing.
  - Whenever new statistical/machine learning methodology is developed, it is typically first implemented as an R package.

Computing Language: R

  - As a result, there is a very rich repository of packages for cutting-edge statistics/ML and visualization (The Comprehensive R Archive Network (CRAN); `https://cran.r-project.org`).

- Like Python, R is an interactive/REPL-based programming environment.
  - A fundamental aspect of R (and of the S language on which it is based) is to allow thought and computation to evolve.
  - Python is more intended to be a programming vs. an interactive tool.
  - Python's use of indentation for scoping makes it more cumbersome to cut-and-paste commands out of context (easy way around this is to remove white space, but this has limitations).
- Functional programming.

Computing Language: R

- R is a functional programming language, with no side effects, i.e., when we pass data to a function, the function will not change the data held by the caller of the function.
- Python is reference-based, i.e., data objects can be modified by the function to which the data were passed.
- A programming language for DS should support an iterative analysis process and ensure data fidelity and integrity.
- Functional programming is a very important design principle in a REPL environment, where we could otherwise continue on after previous commands have perturbed input.

- Computational reproducibility. As with Python, there are tools for computational reproducibility with R. E.g. Creation of dynamic documents using Sweave, knitr, Jupyter Notebook.

Computing Language: R

- Open-source. Like Python, R is open-source.
- Python vs. (and) R for Data Science: `https://blog.usejournal.com/python-vs-and-r-for-data-science-833b48ccc91d`. Can use both in Jupyter (`https://jupyter.org`).

Computing Language: R vs. R tidyverse and Python 3 vs. Python 2

Even within languages it's not simple!

- There are incompatibilities between code from the same language.
- There are difficulties to understand code from the same language.
- Both Python and R have challenges with adoptions/conventions.

Computing Language: Other

Other useful languages for some aspects of Data Science.

- C++. (https://isocpp.org)
  - Fast to run, not to write. Useful if code will be run many times.
  - Not amenable to the exploratory and iterative aspect of computing in DS (application/executable so need to know ahead of time the entire sequence of computations).
- JavaScript. JavaScript is the main language for creating interactive data visualizations, e.g., in data journalism.
- Julia. (https://julialang.org)
  - Good computational model: Interactive, compiled (just-in-time compilation), pretty fast, can work with code from other languages which is also compiled.
  - Currently lacking extensive packages, but this may change over time depending on sustained adoption.

Computing Language: Other

  - Current community of users focused on high-performance numerical analysis (e.g., optimization), not so much data analysis, although some visualization and machine learning packages are available.

- Matlab. Very good for some purposes and communities, e.g., image processing, neuroscience. Not open-source.
- SAS. Unlike Python and R, not a general-purpose programming language. Doesn't allow you to adapt what it does or experiment. Not open-source. Idem for SPSS (Statistical Package for the Social Sciences; https://www.ibm.com/analytics/spss-statistics-software) and Stata (https://www.stata.com).

Computing Language for Data Science

- There are trade-offs in every language.
- There are intersystem interfaces between different languages (e.g., Omega Project, `http://www.omegahat.net`).
- Data scientists should be proficient in several languages to switch between communities, use the right language for a particular task, and not reimplement from scratch but rather reuse and extend existing quality code (cf. reinventing the wheel).
- This requires familiarity with general programming concepts and reasoning, but also actually "speaking the language" to be able to understand and evaluate the code base and build on top of it.

Computing Language for Data Science

- At some points, you will need to learn about other languages, including domain-specific languages (DSL) (e.g., regex, SQL, xpath). Such DSL are used in general-purpose languages like Python and R.
- Take-home messages.
  - Be open-minded and agnostic.
  - Choose the right language for the matter at hand.
  - Leverage the strengths of different languages.
  - There is not much to be gained by being dogmatic an opposing different languages, often out of ignorance or laziness.
  - The same could be said for statistical methodology, e.g., applying hidden Markov model or neural networks to address every problem.

References

W. S. Cleveland. Data science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review*, 69(1): 21–26, 2001.

D. Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, 2017. doi: 10.1080/10618600.2017.1384734. URL https://www.tandfonline.com/doi/abs/10.1080/10618600.2017.1384734.

R. Gentleman and D. Temple Lang. Statistical analyses and reproducible research. Technical Report 2, Bioconductor Project Working Papers, 2004.

S. Lau, J. Gonzalez, and D. Nolan. *Principles and Techniques of Data Science*. 2019. URL https://www.textbook.ds100.org.

J. Tukey. The future of data analysis. *The Annals of Mathematical Statistics*, 33(1):1–67, 1962. doi: 10.1214/aoms/1177704711. URL https://projecteuclid.org/euclid.aoms/1177704711.