Human Centered Data Science

DATA 512 — Jonathan T. Morgan & Os Keyes

Course overview | Week 1 | September 27, 2018

Introductions

Your instructors

Jonathan Morgan

- Pronouns: he/him
- Senior Design Researcher @ Wikimedia Research (research.wikimedia.org)
- PhD in Human Centered Design & Engineering (hcde.uw.edu)
- Member of Partnership on AI (partnershiponai.org)
- Member of Community Data Science Collective (communitydata.cc)
- Resident of Whidbey Island (en.wikipedia.org/wiki/Whidbey_Island)

Your instructors

Os Keyes

- Pronouns: They/Them
- PhD student in Human-Centred Design & Engineering
- Professional ~~cynic~~ ethicist
- Science & Technology Studies (STS) geek
- Recovering data scientist

Overview of the day

- Syllabus review
- Pre-course survey results
- What do we mean by ‘data science’?
- What do we mean by ‘human centered’?
- How does HCD relate to DS?
- For week 2

Class breaks

Syllabus review

https://wiki.communitydata.cc/HCDS_(Fall_2018)#Schedule

Class policies

https://wiki.communitydata.cc/HCDS_(Fall_2018)#Policies

Accomodations

- If you need an accommodation due to a disability, you will need to file a disability accommodation with UW.
- Accommodation = any change to class policies: due dates, attendance, grading, recording lectures, etc.
- Jonathan can *only* accommodate your needs if you have applied for a disability accommodation beforehand.

For information on accommodations and how to apply for one, see:
http://depts.washington.edu/uwdrs/current-students/accommodations/

Plagiarism

- Plagiarism is using someone else’s words, images, code, or data in a class assignment without attributing it appropriately.
- If you plagiarize:
    - You probably won’t get away with it
    - You will almost certainly fail the assignment
    - You may fail the course

When in doubt… ask Jonathan and Os *before* you turn in the assignment.

See: https://wiki.communitydata.cc/HCDS_(Fall_2018)#Academic_integrity_and_plagiarism

> *“The key to avoiding plagiarism is that you show clearly where your own thinking ends and someone else's begins.”*
>
> - UW HCDE Plagiarism and academic conduct policy

Assignments

https://wiki.communitydata.cc/HCDS_(Fall_2018)#Assignments

In-class activities

- 2 points each, not graded on content
- Due every week, unless otherwise specified
- Turn in by 11:59pm the day *after* class (via Canvas)
- If group assignment
    - Choose 1 group member to turn in (via Canvas)
    - Include all group members' names in the Canvas post

https://wiki.communitydata.cc/Human_Centered_Data_Science_(Fall_2018)/Assignments#Weekly_in-class_activities

Reading reflections

- 2 points each, not graded on content
- Due every week, unless otherwise specified
- Turn in before next class (via Canvas)
- If multiple readings are assigned
    - Read both, reflect on one

Format

- Answer the question "How does this reading inform your understanding of human centered data science?" (2-3 full sentences)

- +1 question that this reading made you think of, and say why.

https://wiki.communitydata.cc/Human_Centered_Data_Science_(Fall_2018)/Assignments#Weekly_reading_reflections

Major assignments

- **A1 - 5 points** (due 10/18): Data curation (programming/analysis)
- **A2 - 10 points** (due 11/1): Sources of bias in data (programming/analysis)
- **A3 - 10 points** (due 11/8): Crowdwork Ethnography (written)
- **A4 - 10 points** (due 11/22): Final project plan (written)
- **A5 - 10 points** (due 12/6): Final project presentation (oral, slides)
- **A6 - 15 points** (due 12/9): Final project report (programming/analysis, written)

Breakdown of assignments

- 20% in-class work
- 20% reading reflections
- 60% assignments
    - 25% Stand-alone assignments
    - 35% Final project assignments

Late assignments

- Assignment due dates are FIRM
  - Late assignments can *only* be accepted if you have already filed a disability accommodation with Jonathan beforehand.
- For more information about disability accommodations, see:
  - https://wiki.communitydata.cc/HCDS_(Fall_2018)#Disability_and_accommodations

Missed classes

- If you are unable to attend class, you may request a make-up assignment to receive credit for that day's in-class activity, AS LONG AS…
  - You have a disability accommodation on file with Jonathan

    OR

  - You contact both Jonathan and Os by email before 5pm on the day of class
- You are responsible for reviewing course slides, which will be posted to Canvas within 24 hours of each class
- Follow up with fellow students or instructors (use Slack!) about announcements
- Showing up at office hours if you need to

Extra credit

- One extra credit assignment will be made available in late November
- This assignment will be worth up to 6 points

Communication

Office hours

- **Jonathan Morgan:** via Google Meet, as needed.
  - Schedule via email (jmo25@uw.edu)

- **Os Keyes:** Monday 5-7pm and Wednesday 5-7pm, in Sieg Hall 431.
  - Scheduling is appreciated (okeyes@uw.edu)
  - Drop-in’s are okay

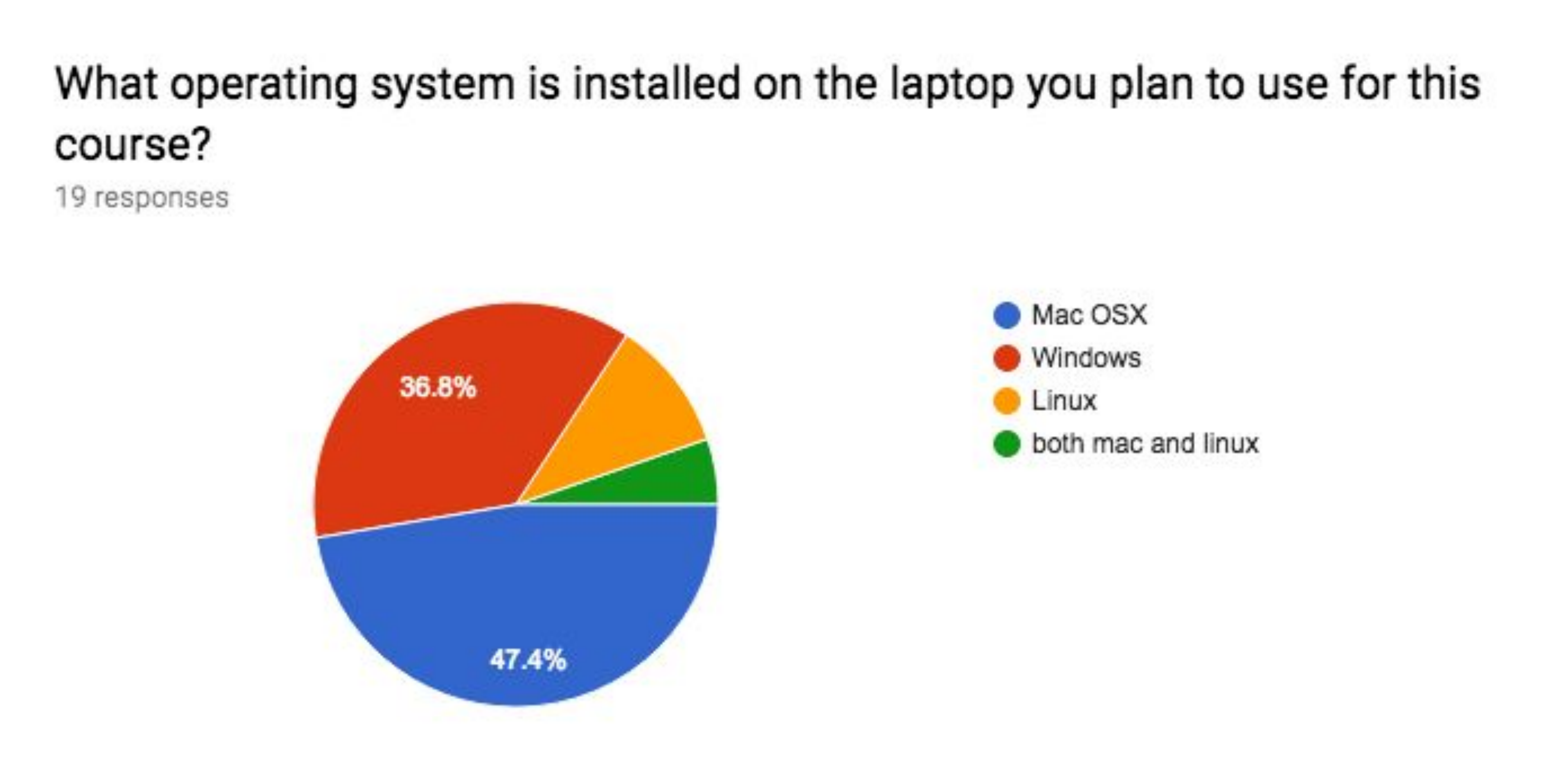
Slack and email

- Use the course Slack channel for general questions, discussions, posting interesting links
    - You can also use the class email list for these things

- Email Jonathan *and* Os (both of us!) directly if you have private questions, if you need to miss class, or if you have any concerns.

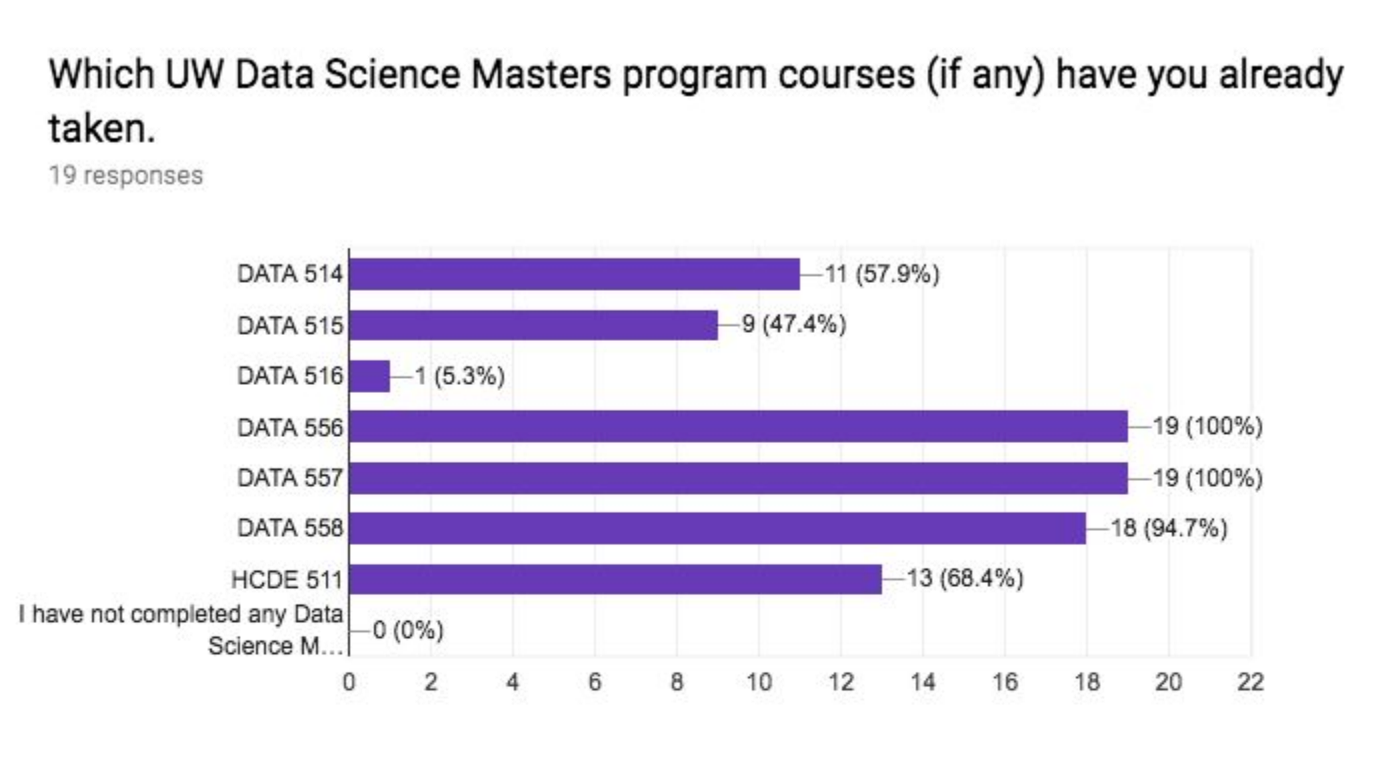
Questions?

Pre-course survey



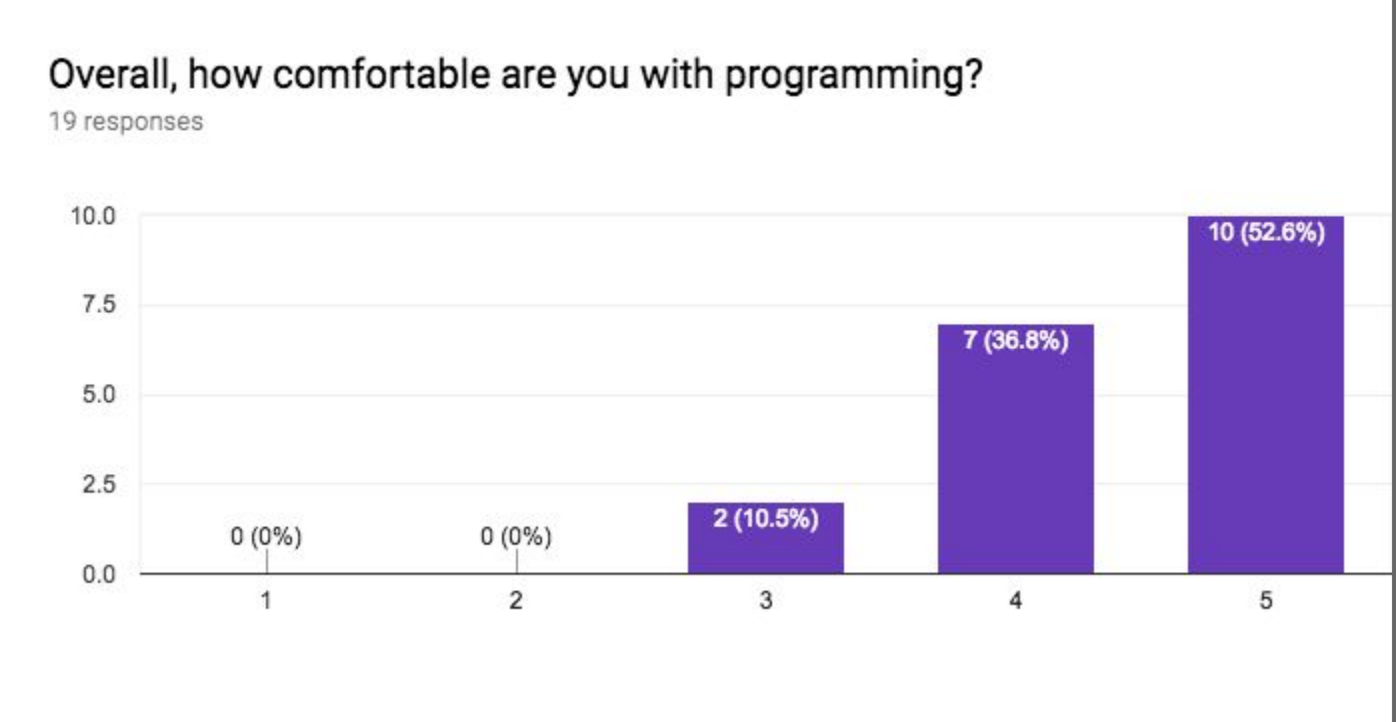
What operating system is installed on the laptop you plan to use for this course?

19 responses



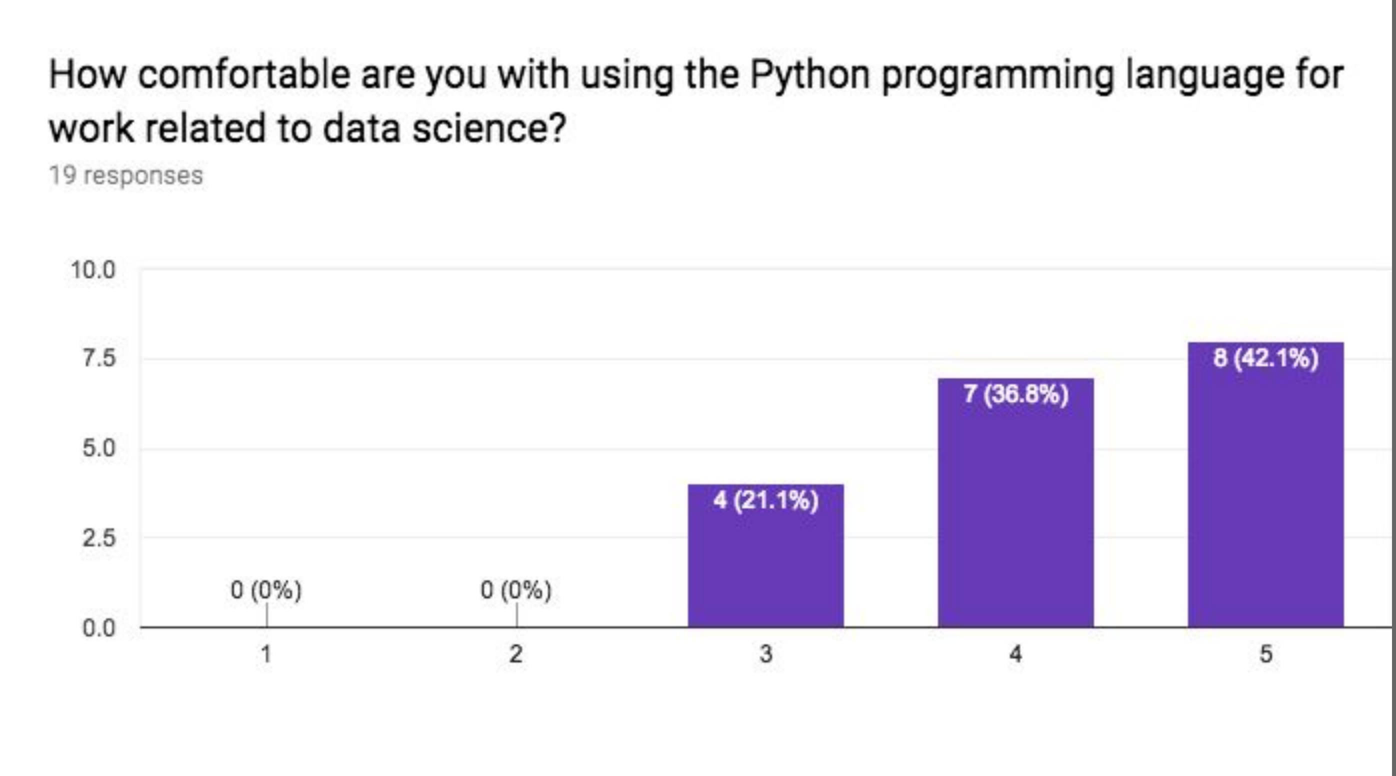
Which UW Data Science Masters program courses (if any) have you already taken.

19 responses



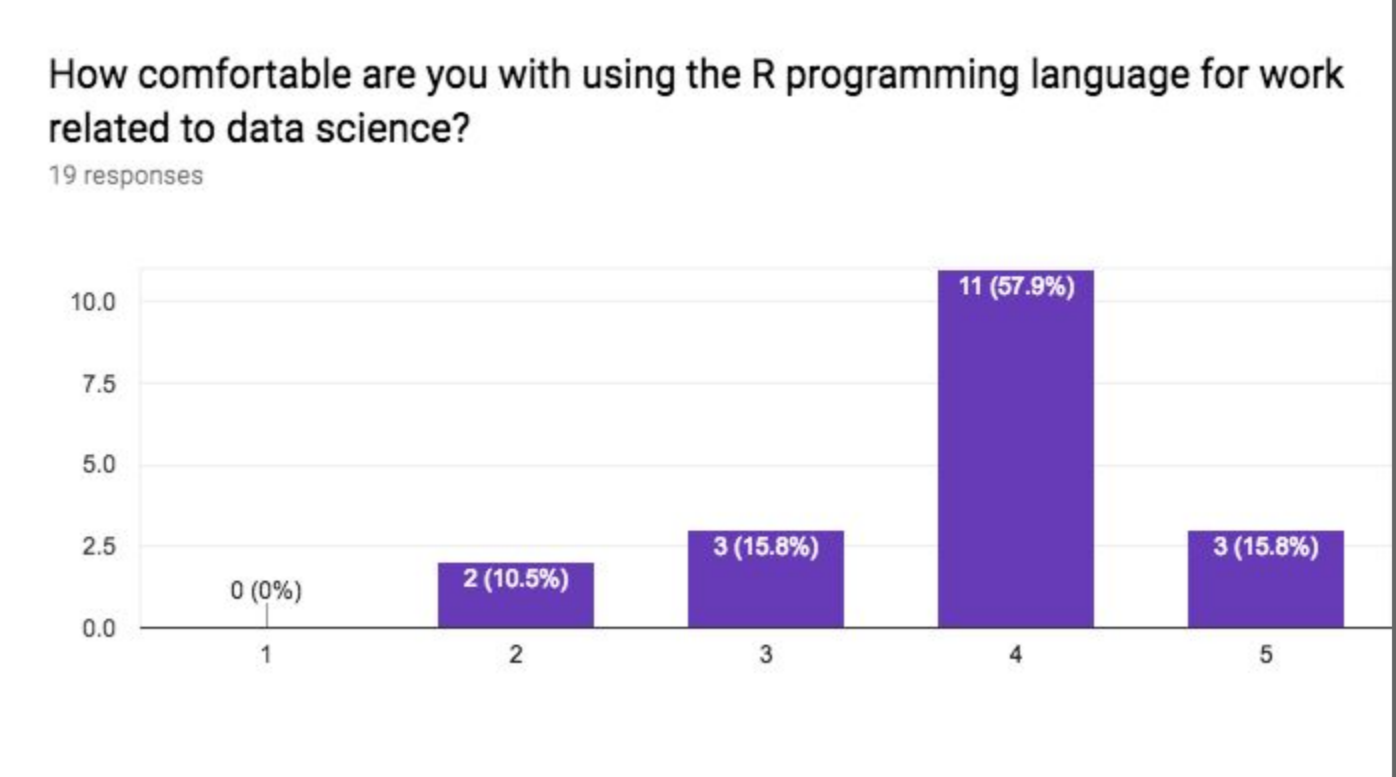
Overall, how comfortable are you with programming?

19 responses



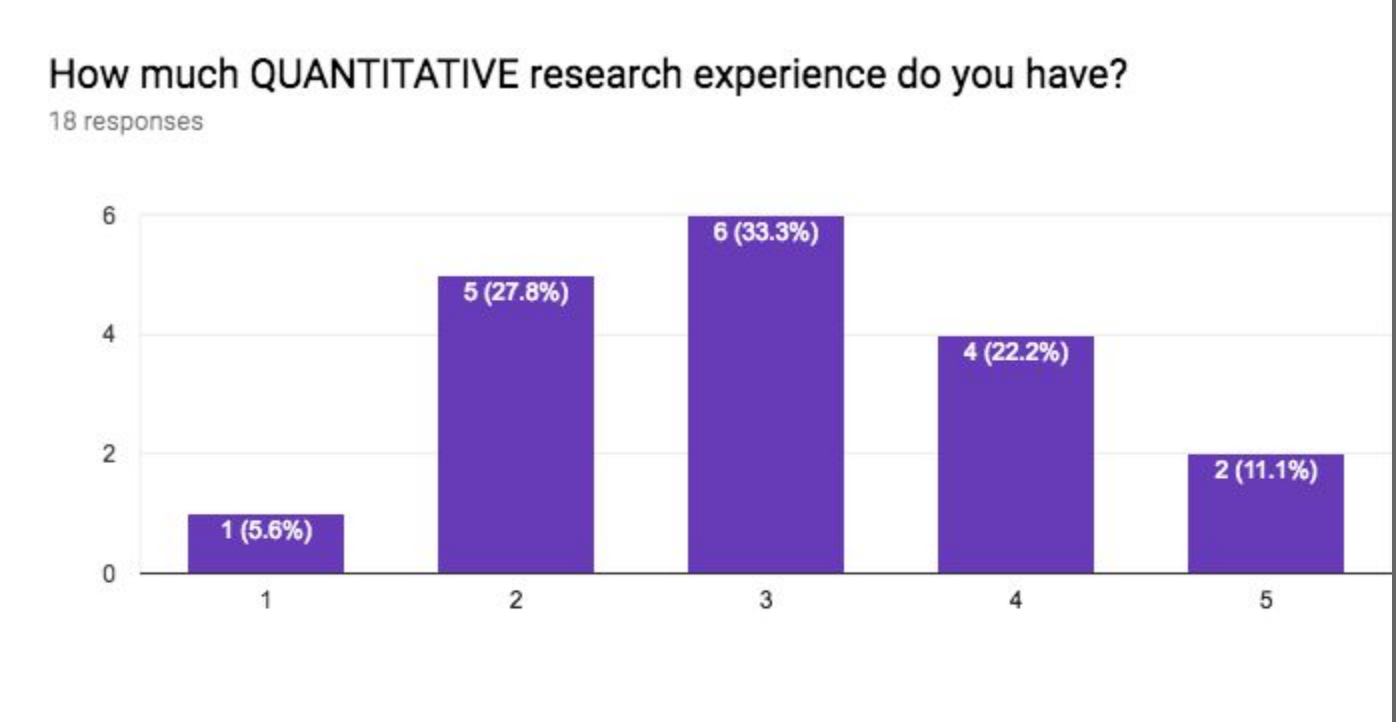
How comfortable are you with using the Python programming language for work related to data science?

19 responses



How comfortable are you with using the R programming language for work related to data science?

19 responses

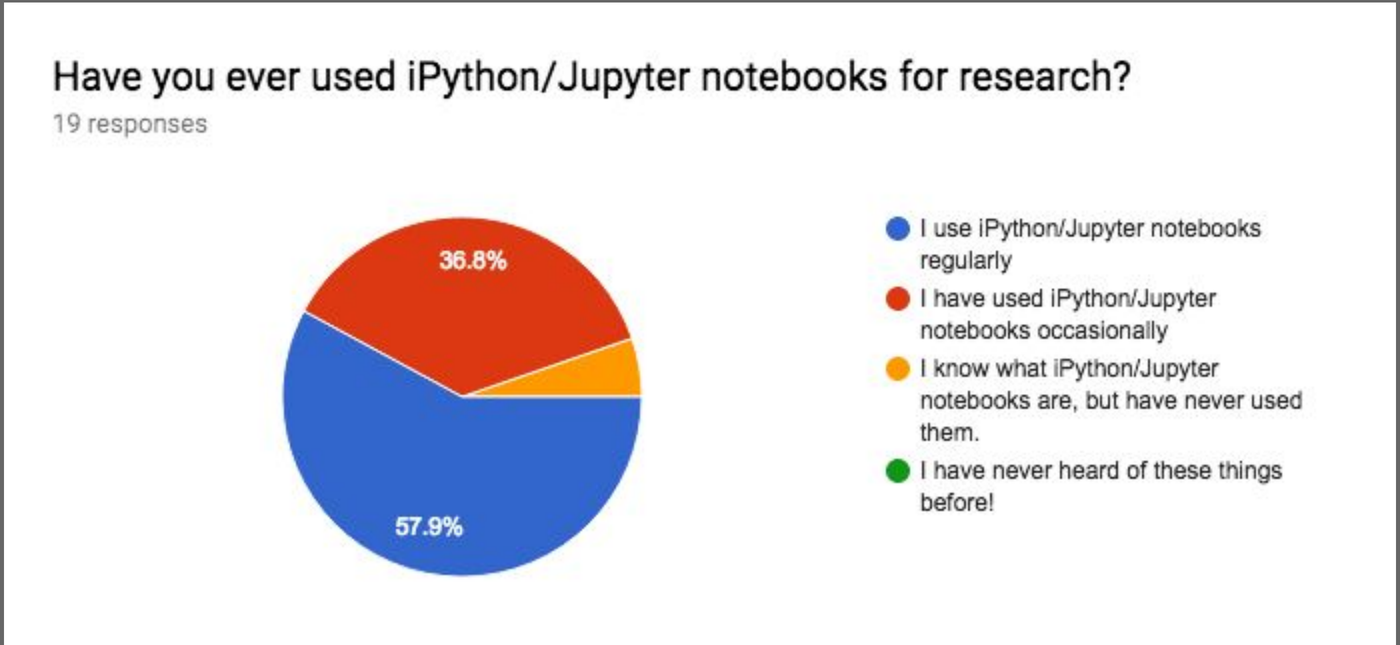


How much QUANTITATIVE research experience do you have?

18 responses

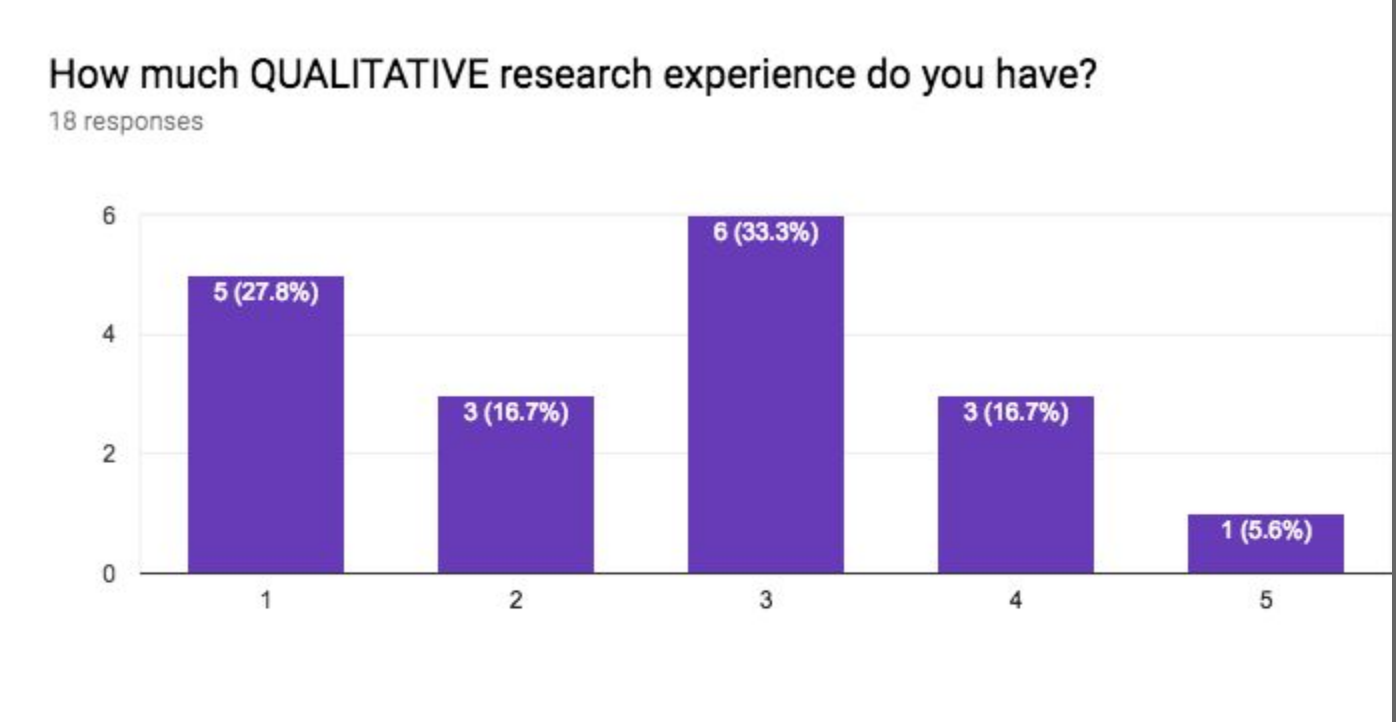
What kind of quant research?

- I've built data pipelines and used logistic regression, clustering, and PCA to obtain interpretable models. Also built news article recommender system using LDA.
- During my undergrad degree I used C++ to analyze particle collisions from the Large Hadron Collider, and constructed histograms of the results.
- 3 years of developing atmospheric air pollution chemistry models
- Built models to predict customer spend. Built models to build customer frequency of spend. Build model to identify facial emotion from video. Built visualizations using Tableau and D3 to visualize nutrition facts of fast food menu items.
- I use rudimentary statistical methods to make my picks in my NFL Pick 'Em league.
- None outside courses



Have you ever used iPython/Jupyter notebooks for research?

19 responses



How much QUALITATIVE research experience do you have?

18 responses

What kind of qual research?

- I previously organized focus groups at my previous job
- I have helped plan and conduct usability tests, surveys, and feedback gathering programs, primarily on documentation, as my work
- User testing completed in HCDE 511
- I've taken courses on how to conduct user research and ethnographic studies for visualization design and development.
- I've participated in usability studies and briefly covered them in INFX 562.

What kind of data are you most interested in working with?

- Business analytics, product analytics
- Linguistic data, because I think data science will be able to produce the "oracle" of documentation someday, given the right semantic modeling of all our content.
- Financial data, Sports data.
- Social media data
- government data, non-profit data. Prior to this program I am interested in applying what I learn in the program to something related to climate change, environmental protection, etc.
- I'm most interested in using ecommerce and web data. I really love predicting user behavior in the future based on past data.
- As long as the data is legally available, and the problem is interesting, I'll be keen to work on it.

What do you hope to get from DATA 512?

- User Interface, how to communicate analysis in a way that can be easily understood.
- How to better understand your end user needs.
- What the challenges and paths to resolution look to be for all the socially interesting areas of data science, from privacy to the control problem of AI.
- Understanding the ethical aspects in the data science industry and impact on society.
- Data privacy and security and the rules/regulations (new and existing) surrounding how we companies manage and use data. I'd also be interested in learning about methods to ensure equality (gender, racial, etc.) in data collection, studies, etc.
- Role of ethics/philosophy in data science; how pervasive data science is going to affect individual privacy; reasons for the success of open source systems; innovative ways to tackle UX issues.

What concerns you about the course?

- The relevance to the industry or to obtain a job because I am good at it.
- I would really like the course to be as practical as possible, and help me analyze the concepts being taught through practical examples.
- Based on discussions with students who took the course last year, it sounded like the material was light on covering current discussions around data security, privacy, and regulation.
- I've heard complaints about this course from 1st year students that it was not technical or organized enough.

In your own words, what is ‘data science’?

- The study of data and data analysis to gain an understanding of a problem.
- Process of getting, aggregating, analyzing, modeling, visualizing and interpreting data.
- Applied use of statistics, ML and visualization.
- Application of scientific and mathematical methods to study and extract information from data.
- Asking the right questions and figuring out a way to get the best answers from given data, or finding the proper data if required.
- Using Data to answer business questions
- Statistics + Computer Science + Graphics Design
- The science of extracting useful information from data by combining statistics with programming.

What do we mean by *data science*?

Data Science and its relationship to data-driven decision making

Foster Provost & Tom Fawcett, *Big Data,* 2013

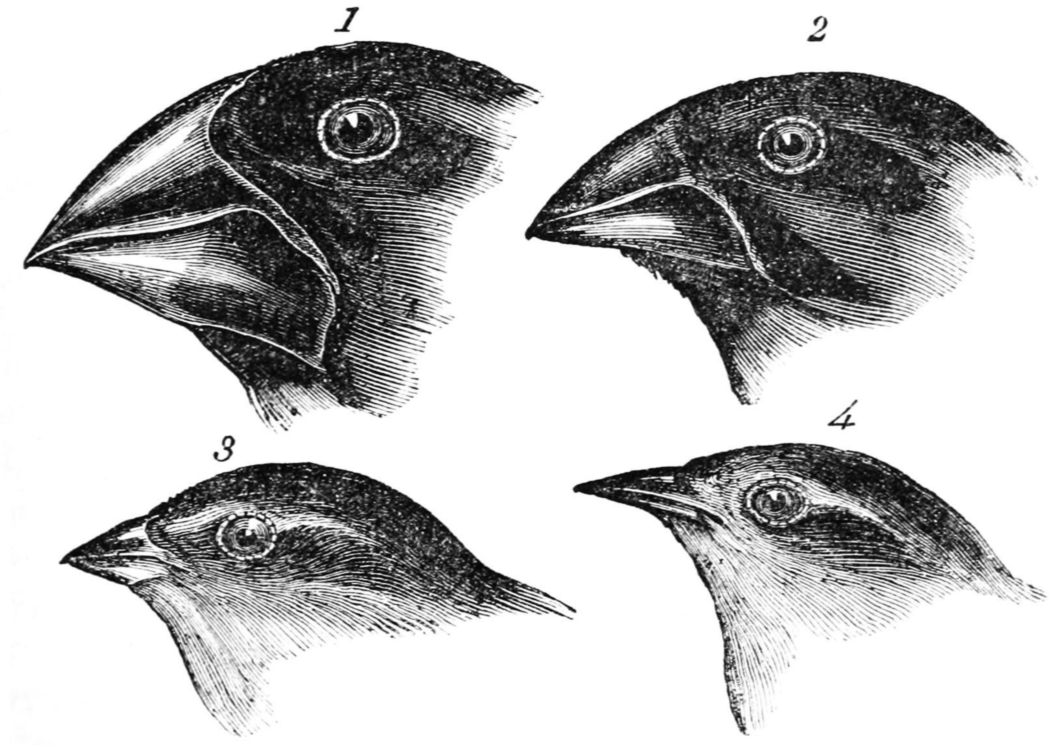
Definitions

- Data science
- Big data
- Data mining
- Data engineering
- Data-driven decision making
- Data assets

“Data science is a set of fundamental principles that support and guide the principled extraction of information and knowledge from data.”

Fundamental concepts of data science

- Look for correlations
- Avoid overgeneralization
- Identify assumptions and confounding factors



1. Geospiza magnirostris.
3. Geospiza parvula.

2. Geospiza fortis.
4. Certhidea olivacea.

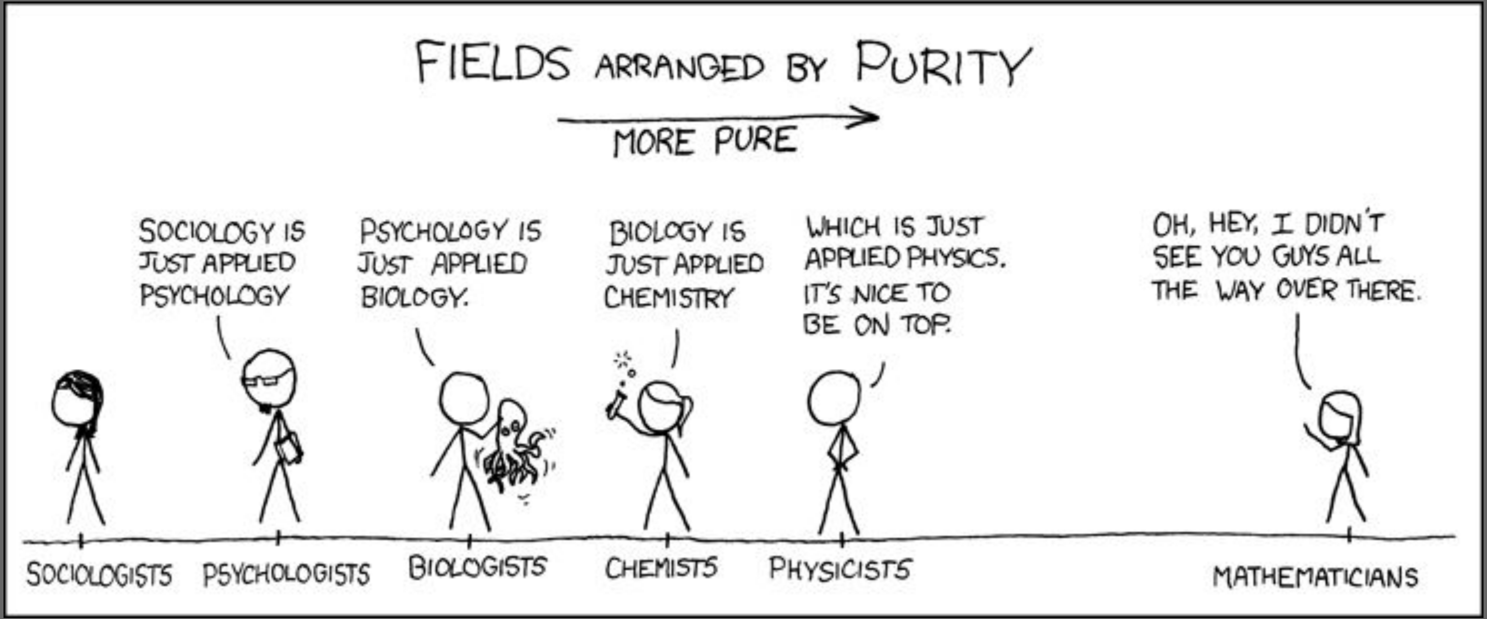
Source: https://en.wikipedia.org/wiki/File:Darwin%27s_finches_by_Gould.jpg

“I think data-scientist is a sexed up term for a statistician. Statistics is a branch of science. Data scientist is slightly redundant in some way and people shouldn’t berate the term statistician.”

~ Nate Silver, Founder of fivethirtyeight.com

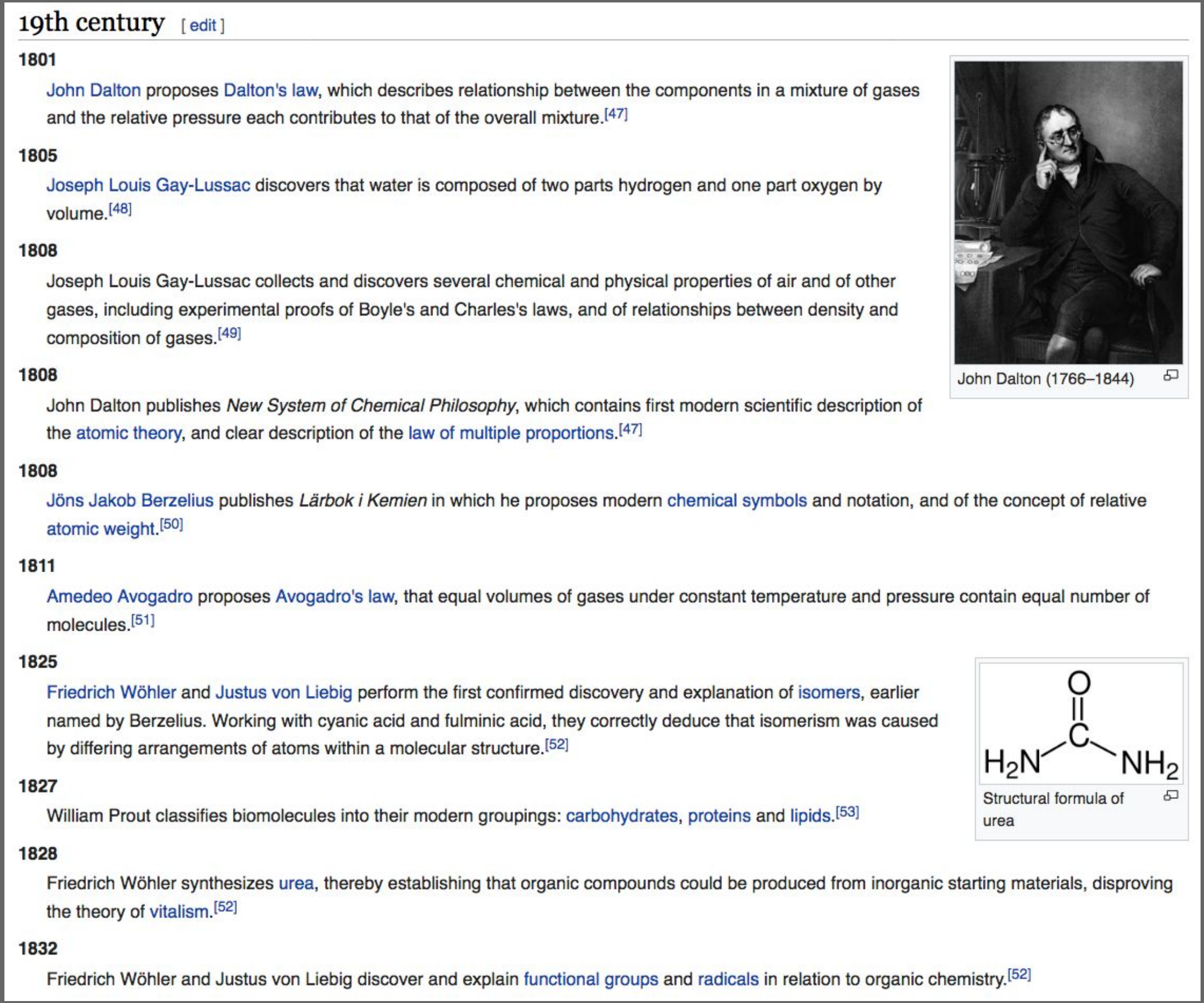
Source:
http://www.statisticsviews.com/details/feature/5133141/Nate-Silver-What-I-need-from-statisticians.html

Key concepts


FIELDS ARRANGED BY PURITY
MORE PURE
SOCIOLOGY IS JUST APPLIED PSYCHOLOGY
PSYCHOLOGY IS JUST APPLIED BIOLOGY.
BIOLOGY IS JUST APPLIED CHEMISTRY
WHICH IS JUST APPLIED PHYSICS. IT'S NICE TO BE ON TOP.
OH, HEY, I DIDN'T SEE YOU GUYS ALL THE WAY OVER THERE.
SOCIOLOGISTS
PSYCHOLOGISTS
BIOLOGISTS
CHEMISTS
PHYSICISTS
MATHEMATICIANS

Source: https://xkcd.com/435/

“The particular concerns of data science in business are fairly new, and businesses are still learning how best to address them. The state of data science may be likened to that of chemistry in the mid-19th century, when theories and general principles were being formulated and the field was largely experimental.”



19th century [edit]

1801

John Dalton proposes Dalton's law, which describes relationship between the components in a mixture of gases and the relative pressure each contributes to that of the overall mixture.[47]

1805

Joseph Louis Gay-Lussac discovers that water is composed of two parts hydrogen and one part oxygen by volume.[48]

1808

Joseph Louis Gay-Lussac collects and discovers several chemical and physical properties of air and of other gases, including experimental proofs of Boyle's and Charles's laws, and of relationships between density and composition of gases.[49]

1808

John Dalton publishes *New System of Chemical Philosophy*, which contains first modern scientific description of the atomic theory, and clear description of the law of multiple proportions.[47]

John Dalton (1766–1844)

1808

Jöns Jakob Berzelius publishes *Lärbok i Kemien* in which he proposes modern chemical symbols and notation, and of the concept of relative atomic weight.[50]

1811

Amedeo Avogadro proposes Avogadro's law, that equal volumes of gases under constant temperature and pressure contain equal number of molecules.[51]

1825

Friedrich Wöhler and Justus von Liebig perform the first confirmed discovery and explanation of isomers, earlier named by Berzelius. Working with cyanic acid and fulminic acid, they correctly deduce that isomerism was caused by differing arrangements of atoms within a molecular structure.[52]

Structural formula of urea

1827

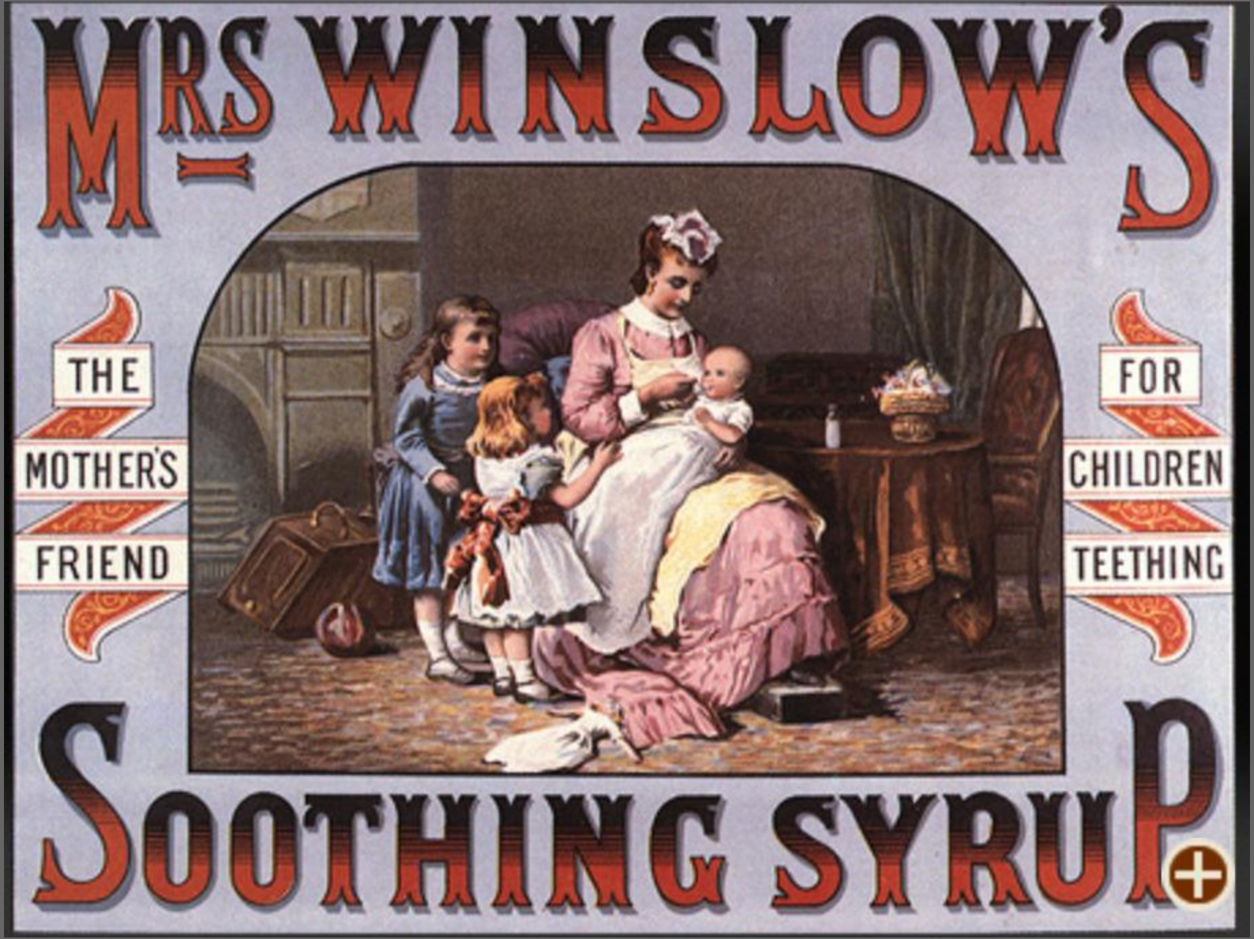
William Prout classifies biomolecules into their modern groupings: carbohydrates, proteins and lipids.[53]

1828

Friedrich Wöhler synthesizes urea, thereby establishing that organic compounds could be produced from inorganic starting materials, disproving the theory of vitalism.[52]

1832

Friedrich Wöhler and Justus von Liebig discover and explain functional groups and radicals in relation to organic chemistry.[52]



Source: https://www.nlm.nih.gov/exhibition/pickyourpoison/exhibition-opium.html

“Many new companies are being developed with data mining as a key strategic component.

Facebook and Twitter, along with many other Digital 100 companies, have high valuations due primarily to **data assets they are committed to capturing and creating.**”

This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology.

Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. **With enough data, the numbers speak for themselves.**

Chris Anderson *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*
(WIRED, June 23 2008)

Five years ago, a team of researchers from Google announced a remarkable achievement in one of the world’s top scientific journals, *Nature*.

Without needing the results of a single medical check-up, they were nevertheless able to track the spread of influenza across the US. What’s more, they could do it more quickly than the Centers for Disease Control and Prevention (CDC).

Harford, T. (2014). Big data: A big mistake? *Significance*, *11*(5), 14–19.

Not only was “Google Flu Trends” quick, accurate and cheap, it was theory-free.

Google’s engineers didn’t bother to develop a hypothesis about what search terms – “flu symptoms” or “pharmacies near me” – might be correlated with the spread of the disease itself. The Google team just took their top 50 million search terms and let the algorithms do the work.

Harford, T. (2014). Big data: A big mistake? *Significance*, *11*(5), 14–19.

Course goals

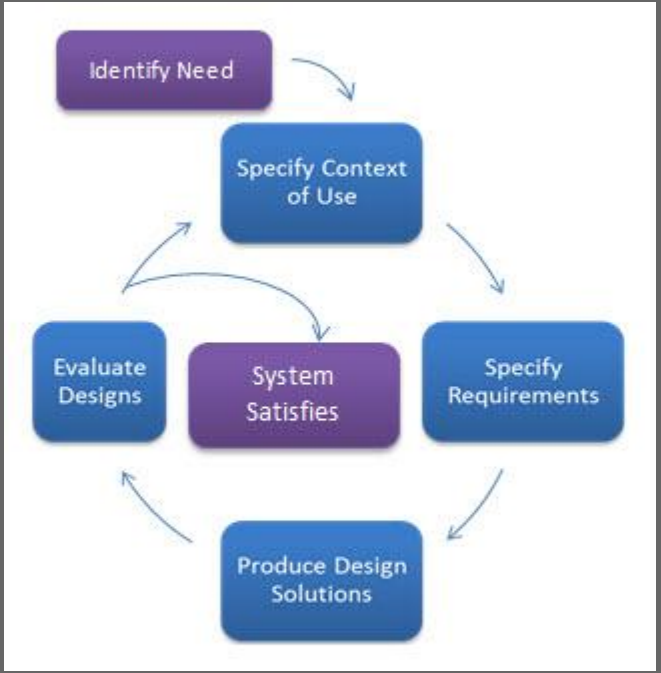
The goal of this course is to help you, the data scientists of the future, think critically about the *scientific* and *societal* implications of your work.

And how understanding and performing your work in a *human centered* way helps you do better data science.

What does ‘human centered’ mean?

Data science as design

User-centered design process



Source: https://www.usability.gov/what-and-why/user-centered-design.html

User-centered design principles

- **Audience:** Who is the product for?
- **Purpose:** What problem is the product intended to solve?
- **Context:** What aspects of the audience’s mental, physical, emotional, temporal, social, economic, etc. circumstances might affect how, when, and why they use (or don’t use!) the product?

ISO 9241-210: Human Centered Design for interactive systems

- The design is based upon an explicit understanding of users, tasks and environments.
- Users are involved throughout design and development.
- The design is driven and refined by user-centred evaluation.
- The process is iterative.
- The design addresses the whole user experience.
- The design team includes multidisciplinary skills and perspectives.

Source: International Organization for Standardization:
https://www.iso.org/standard/52075.html

ISO 9241-210: Human Centered Design for interactive systems

- The design is based upon an explicit understanding of users, tasks and environments.
- **Users are involved throughout design and development.**
- The design is driven and refined by user-centred evaluation.
- The process is iterative.
- **The design addresses the whole user experience.**
- **The design team includes multidisciplinary skills and perspectives.**

Source: International Organization for Standardization:
https://www.iso.org/standard/52075.html

Human Centered Systems in the Perspective of Organizational and Social Informatics

Rob Kling & Leigh Star, 1997
(in Week 1 ‘Resources’)

Based in an analysis of human tasks

“[A human-centered system] must be analysis which encompasses the complexity of social organization.

The analysis cannot be based upon a vague idea of what a generic individual would like, in a stereotypic situation.”

Built to take account of human skills

“The basic architecture of the system must reflect a realistic relationship between people and machines

As with the architecture of buildings, the architecture of machines embody questions of livability, usability and sustainability.”

Designed to address human needs

“The question of whose purposes are served in the development of a system [must] be an explicit part of design, evaluation and use.

The question of whose ideas get put into the design process [and] the question of whose problems are being solved [are] important for human centered systems.”

Monitored for performance in terms of human benefits

“Human-centered is not a one-off or timeless attribute of a system at a given point in time… it is a process.”

“It include the participation of stakeholder groups -- such as involving patient groups in the development of specialist medical technologies, or teachers in the development of instructional technology.”

Implications of HCD for data science

Audience: Who are you designing for? Who else may be affected by your work?

Purpose: Whose purposes are served by your analysis, or your algorithm? Whose purposes are not?

Context: How will your work impact people’s lives? How will they understand and interact with the results of your work? What unintended consequences might result from performing this analysis or deploying this algorithm?

Implications of HCD for data science

- Privacy, ethics, and consent
- Data provenance, preparation, and reproducibility
- Sources and consequences of bias in data
- Interdisciplinary science and mixed-methods research
- Algorithmic fairness, transparency, and accountability
- Societal implications of data science
- Human centered design methods for algorithm design & evaluation
- User interface design for algorithmically-driven software systems
- Effective communication of methods, outcomes, and implications

How does human centered design relate to data science?

In-class activity

2 points, 20 minutes, 5 people per group

Analyzing HCDS scenarios

- Read through the scenario with your group
- Flesh out the scenario with additional details of your choice (document these)
- Describe at least 3 human-centered design considerations in your scenario
- Follow submission instructions on Canvas (discussion “Week 1 in-class activity: Real-life Human Centered Data Science scenarios”)
- Be prepared to discuss your choices

This activity is graded. There are no right or wrong answers; what matters is that you show that you’re thinking.

See: https://wiki.communitydata.cc/HCDS_(Fall_2018)/Assignments#Weekly_in-class_activities

Example

Scenario description

You work for a health insurance company in Mexico. The company uses machine learning to mine data about current and potential customers, such as shopping history, and identify patterns associated with high-risk individuals so that they can charge those individuals higher prices for insurance.

Additional details (I made these up)

- The company plans to mine public social media feeds and match that data with individual shopper profiles
- Mexico has no state-sponsored health care plans
- Mexico has recently adopted a GDPR-style “right to explanation” law

Example

Scenario description

You are working for a health insurance company in Mexico. The company uses machine learning to mine data about current and potential customers, such as shopping history, and identify patterns associated with high-risk individuals so that they can charge those individuals higher prices for insurance.

HCD considerations

1. The people who are poorest and sickest will be least likely to have access to health insurance
2. Social media data could be mis-identified, leading to people being charged higher prices unfairly
3. How will the insurance company explain to customers why they are being charged a particular rate?

For Week 2

Homework

“Big Data’s end-run around Anonymity and Consent”
Solan Barocas & Helen Nissenbaum, 2014

https://wiki.communitydata.cc/HCDS_(Fall_2018)/Assignments#Weekly_reading_reflections

Questions?

Unused slides

HCDE Mission Statement

Human Centered Design & Engineering (HCDE) faculty and students are advancing the research and design of technologies by using innovative techniques to study human activity and develop meaningful information and sociotechnical systems.

HCDE is designing the future by:

- Considering the role of technology in human activity.
- Prioritizing the needs, desires, and behaviors of people and communities who interact through sociotechnical systems.
- Addressing the specifics of design by working with interdisciplinary communities of researchers to build the technologies of tomorrow.

Source: https://www.hcde.washington.edu/mission

“Human-Centered Design (HCD) is **a process and a set of techniques** used to create new solutions for the world. Solutions include products, services, environments, organizations, and modes of interaction. The reason this process is called “human-centered” is because **it starts with the people we are designing for.**

The HCD process begins by examining the needs, dreams, and behaviors of the people we want to affect with our solutions. We seek to listen to and understand what they want. We view the world through this lens **throughout the design process.**”

Source: https://ideo.org

Group activity

20 minutes, 4-5 people per group

Sketching the boundaries of human centered design

- Discuss the question *“What do we mean by human centered-ness?”*
- Come up with AT LEAST three bullet points about what ‘human centered’ means
- Be prepared to discuss your decisions

This activity is not graded. There is no right or wrong answer.

Group activity

20 minutes, 4-5 people per group

Sketching the boundaries of data science

- Divide your paper in half: “data science” and “not data science”
- Discuss the question *“What do we mean by data science?”*
- Consider key terms, concepts, activities, examples, data, domains etc…
  - that you associate with data science.
  - that you DO NOT associate with data science.
- Add these to sticky notes, place them on either side of the paper
- Be prepared to discuss your decisions

This activity is not graded. There is no right or wrong answer.