Welcome to DATA 151

I'm so glad you're here!

DATA 151: INTRODUCTION TO DATA SCIENCE (WITH R)

INTRODUCTION/SYLLABUS, TAXONOMY OF VARIABLES, DATA ORGANIZATION

NOTES PREPARED BY PROF. KITADA SMALLEY (FALL 2022)

FREE WRITE (THINK-PAIR-SHARE)

What is data science?

Describe an example

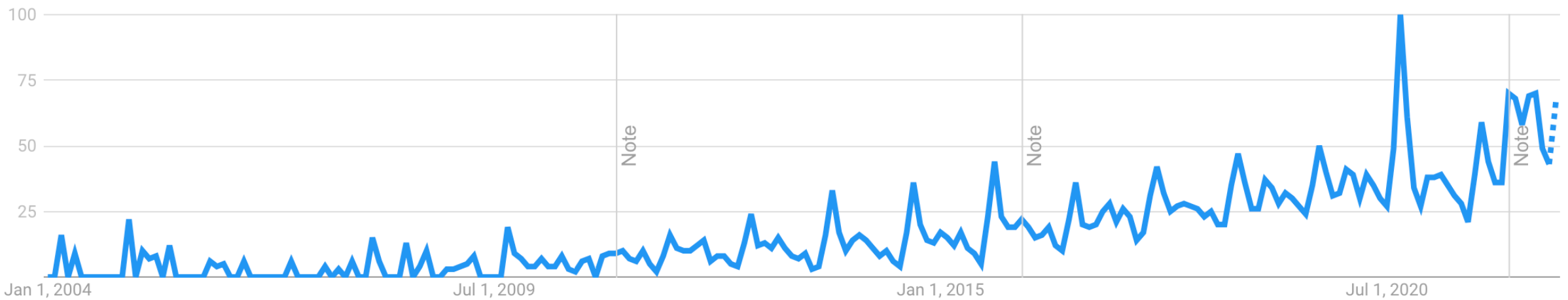
Google Trends | Explore

what is data science
Search term

+ Compare

United States | 2004 - present | All categories | Web Search

Interest over time



Google Trends | Explore

data analytics
Search term

+ Compare

United States | 2004 - present | All categories | Web Search

Interest over time

WHAT IS DATA SCIENCE?

Instagram

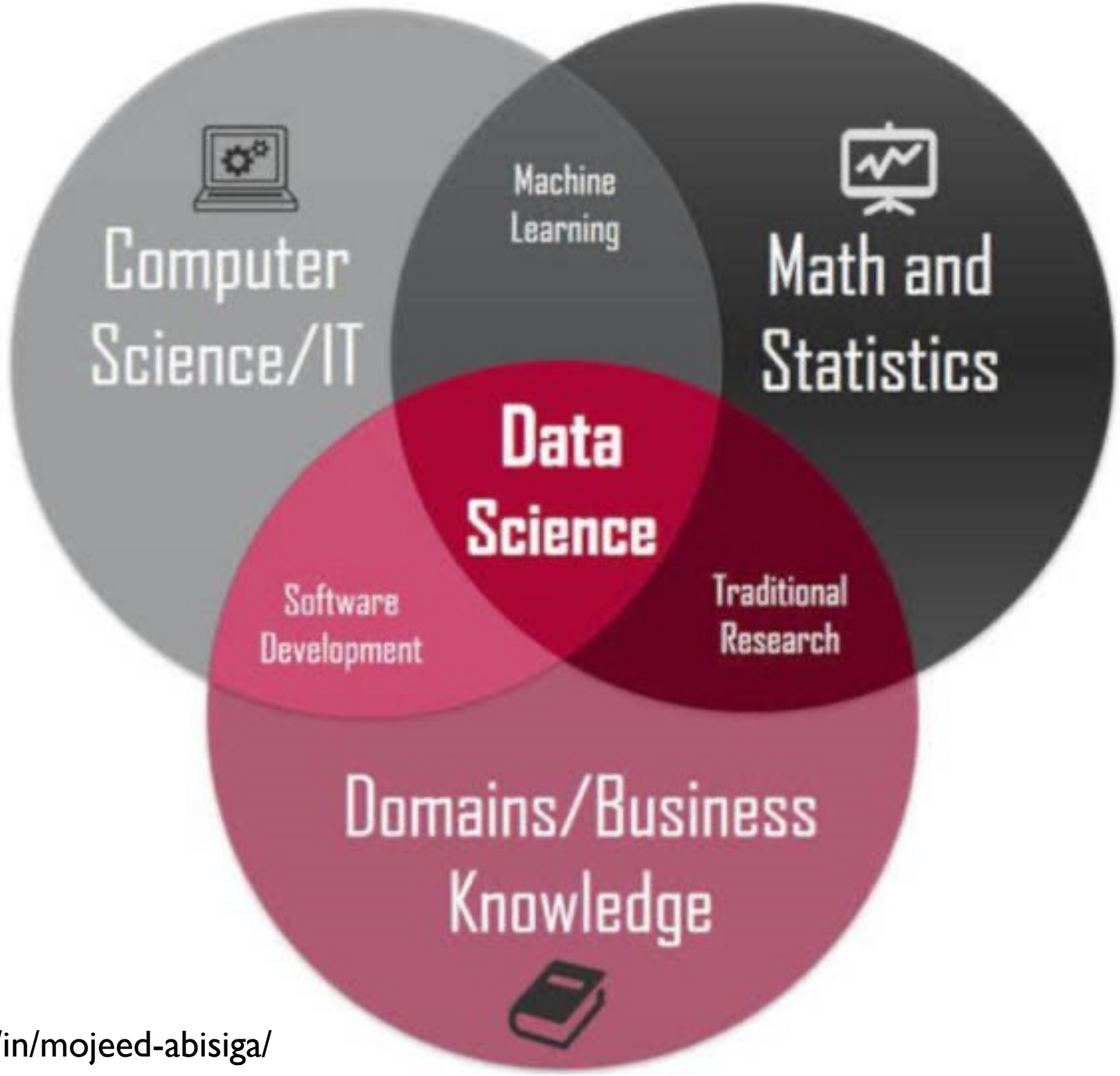
vs

REALITY

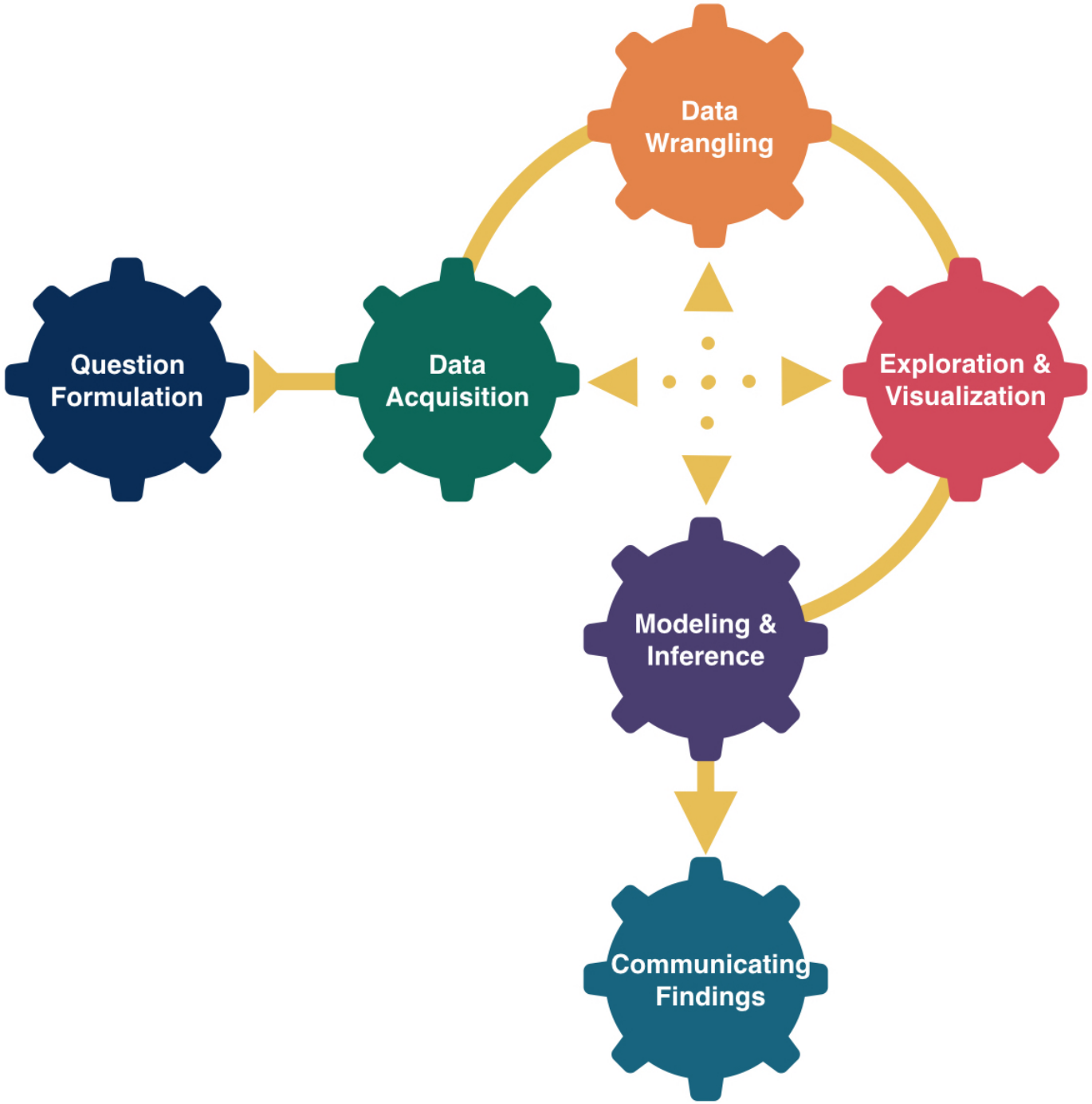
WHAT IS A DATA SCIENTIST?

"A data scientist is a rare hybrid, a computer scientist with the **programming** abilities to build software to scrape, combine, and manage **data** from a variety of sources and a **statistician** who knows how to **derive insights** from the information within"

– Jake Porway (founder DataKind)



https://www.linkedin.com/in/mojeed-abisiga/



Kelly McConville


DATA
#1
SORTED
#2
ARRANGED
#3
PRESENTED
VISUALLY
#4
EXPLAINED
WITH A STORY
#5
ACTIONABLE
(USEFUL)
#6

HOW DO YOU DATA SCIENCE?

Most requested skills according to LinkedIn

1. R
2. Python
3. Data Engineering (SQL)
4. Machine Learning

WELCOME TO DATA 151!

DATA 151: Introduction to Data Science (with R)

- Focuses on developing the foundational **skills of a modern data scientist including data cleaning, wrangling, visualization, and communication**
- Actively engage with **R and RStudio**, the most popular programming language and software environment for statistical computing
- Covers basic descriptive statistics (mean, standard deviation, quantiles, correlation) and introduces students to the tools they need to work with **large, real-world data sets.**

The course **does not assume** any previous background in statistics or programming.

DATA SCIENCE

- Data science is an exciting discipline that allows you to turn raw data into understanding, insight, and knowledge.
- We're going to learn to do this in a `tidy` way (more on this later)
- This is a course on introduction to data science, with an emphasis on statistical thinking.

Source: datasciencebox.org

COURSE FAQ

- Q – What data science background does this course assume?
- A – None! We designed this class to be accessible to a wide audience

Source: datasciencebox.org

COURSE FAQ

- Q – Is this an intro stat course?
- A - While statistics ≠ data science, they are very closely related and have tremendous of overlap. Hence, this course is a great way to get started with statistics. However this course is *not* your typical high school statistics course.

Source: datasciencebox.org

COURSE FAQ

- Q – Will we be doing computing?
- A – Yes. Please bring a laptop to each class that can access the Wi-Fi. This is essential to fully engaging and getting the most out of each class.

Source: datasciencebox.org

COURSE FAQ

- Q – Is this an intro CS course?
- A – No, but many themes are shared. We will be covering the basics of CS that pertain to manipulating data.

Source: datasciencebox.org

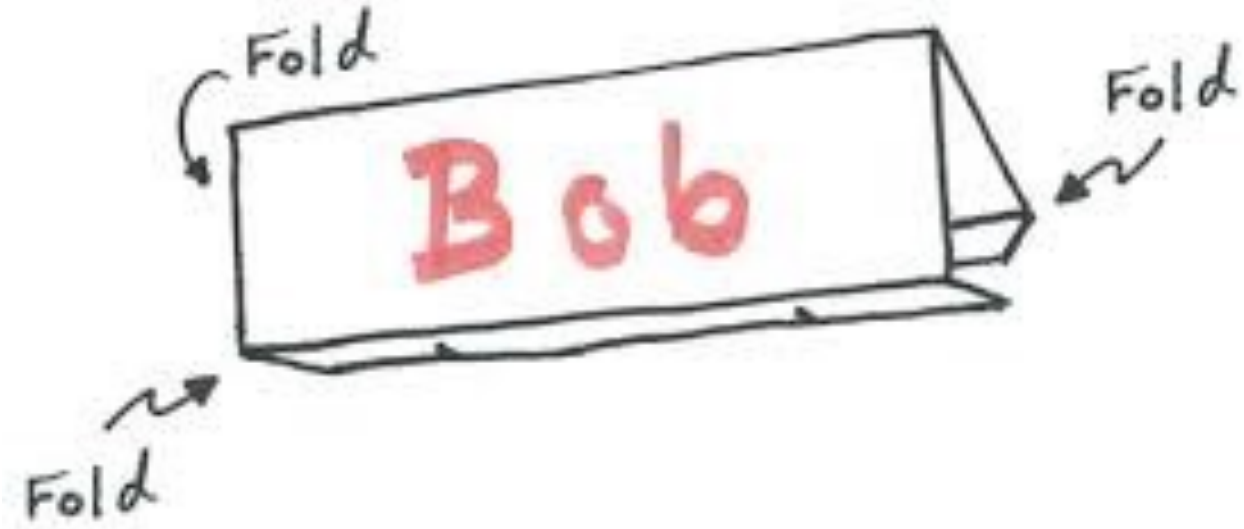
COURSE FAQ

- Q – What computing language will be learn?
- A – R (more in this later)

Source: datasciencebox.org

GETTING TO KNOW YOU

IT'S NICE TO MEET YOU!



Let's make table tents!

Everyone needs…

- A piece a paper
- A marker (not yellow)

Step 1: Folder you paper into <u>four</u> equal sections (long ways), the bottom will be two overlapping sections

Step 2: Write you name big enough on the on the middle sections so people can read it from more than 6ft away

INTRODUCTIONS

Please state your…

- Name
- Pronouns
- Hometown
- Major/Interests

Which emoji best describes how you feel about taking data science this term

GETTING TO KNOW ME

INTRODUCTIONS

Meet Your Professor…

Dr. Heather Kitada Smalley
(She/her/hers)
Albaugh Assistant Professor of Statistics

Email: hsmalley@willamette.edu

Office: FORD 216

INTRODUCTIONS

This is Hadley!

She's two!

MEET YOUR PROFESSOR

I'm from …

South Pasadena, CA

MEET YOUR PROFESSOR

…but I've been in Oregon for 14 years

I LOVE TO TEACH STATISTICS!

This term at Willamette I'm teaching…

- **DATA 151**: Introduction to Data Science with R
- **GSM 6019 / DATA 502**: Research Design, Visualization, and Presentation
  - If you're interested in the new Data Science program please feel free to reach out to me! I would be happy to talk to you!

MY TEACHING PHILOSOPHY GOALS

TRANSPARENT

ENGAGING

SUPPORTIVE

REAL WORLD SKILL BUILDING

COMMUNITY

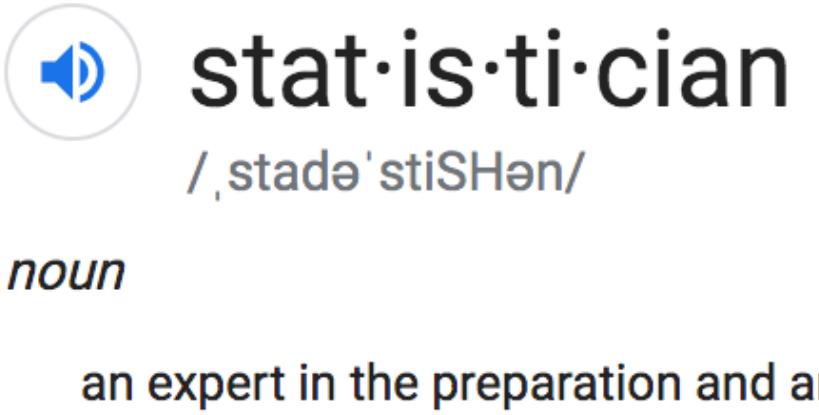
TRANSFORMATIVE

EMPOWERING

RESEARCH

I'M AN APPLIED STATISTICIAN

I'm a …



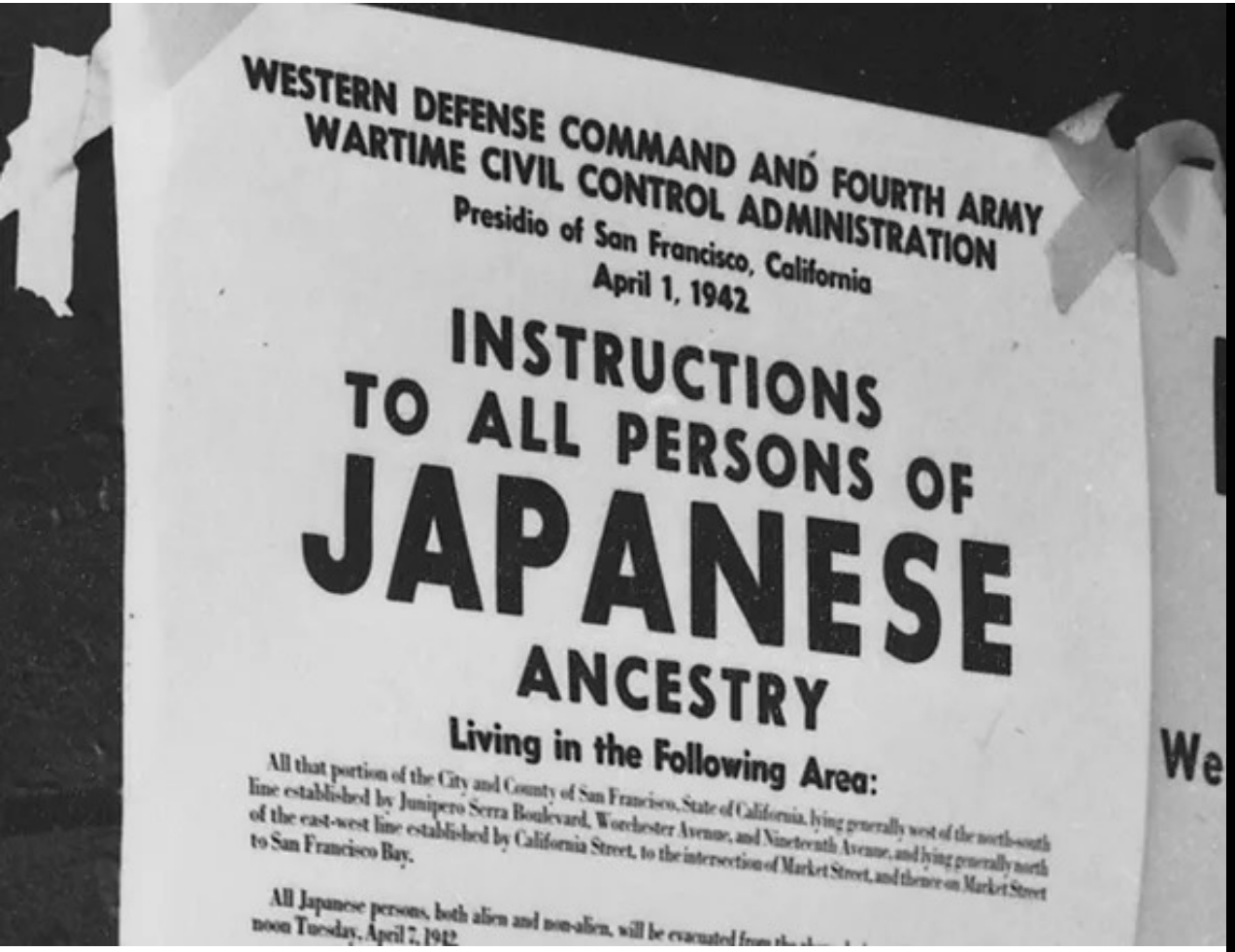
My research in public opinion and survey statistics has focused on developing and improving methodology for bias estimation and mathematical correction.

RESEARCH PHILOSOPHY

Be true to yourself!
Use data to help people

US CENSUS BUREAU

THE CENSUS AND DATA ETHICS
FEBRUARY 1942 - EXECUTIVE ORDER 9066



The Washington Post
Democracy Dies in Darkness

Retropolis

Secret use of census info helped send Japanese Americans to internment camps in WWII

Children at the Manzanar internment camp in California in 1943; photo taken by photographer Toyo Miyatake. (National Park Service/AP)

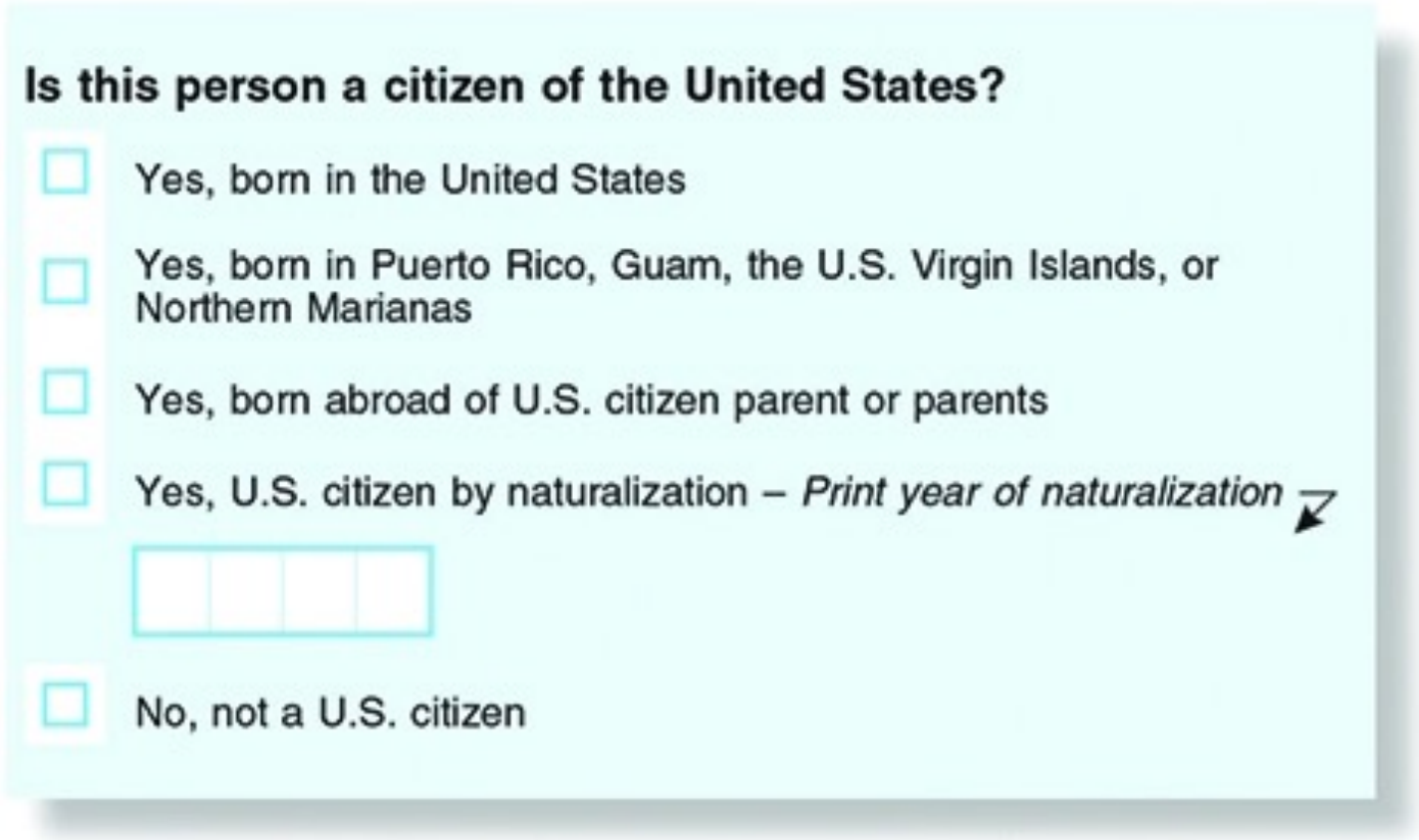
By Lori Aratani
April 6, 2018

US CENSUS BUREAU


DEFEND
OUR IMMIGRANT
COMMUNITIES
#DREAMActNow #SaveTPS
CASA
EVERY PERSON
COUNTS
AMERICA NEEDS A
FAIR AND ACCURATE
#2020CENSUS
#CountMeIn
IMMIGRATION
MARCH


AMERICA NECESITA UN
CENSO JUSTO PARA TODOS
#Censo2020
#CountMeIn
AMERICA NEEDS A
FAIR AND ACCURATE
#2020CENSUS
#CountMeIn
FIRM
CASA
nyic

DATA QUALITY AND DATA JUSTICE

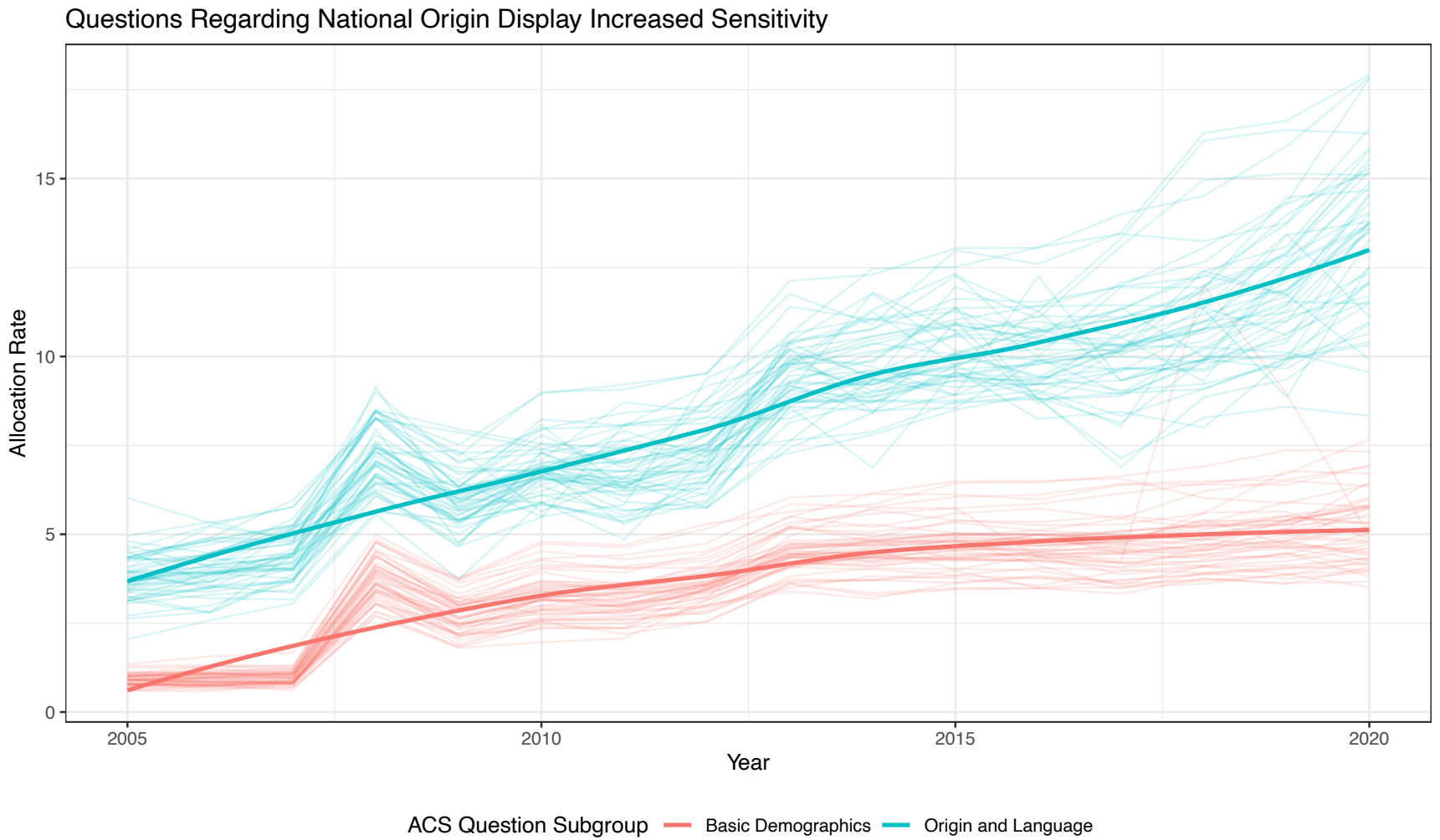


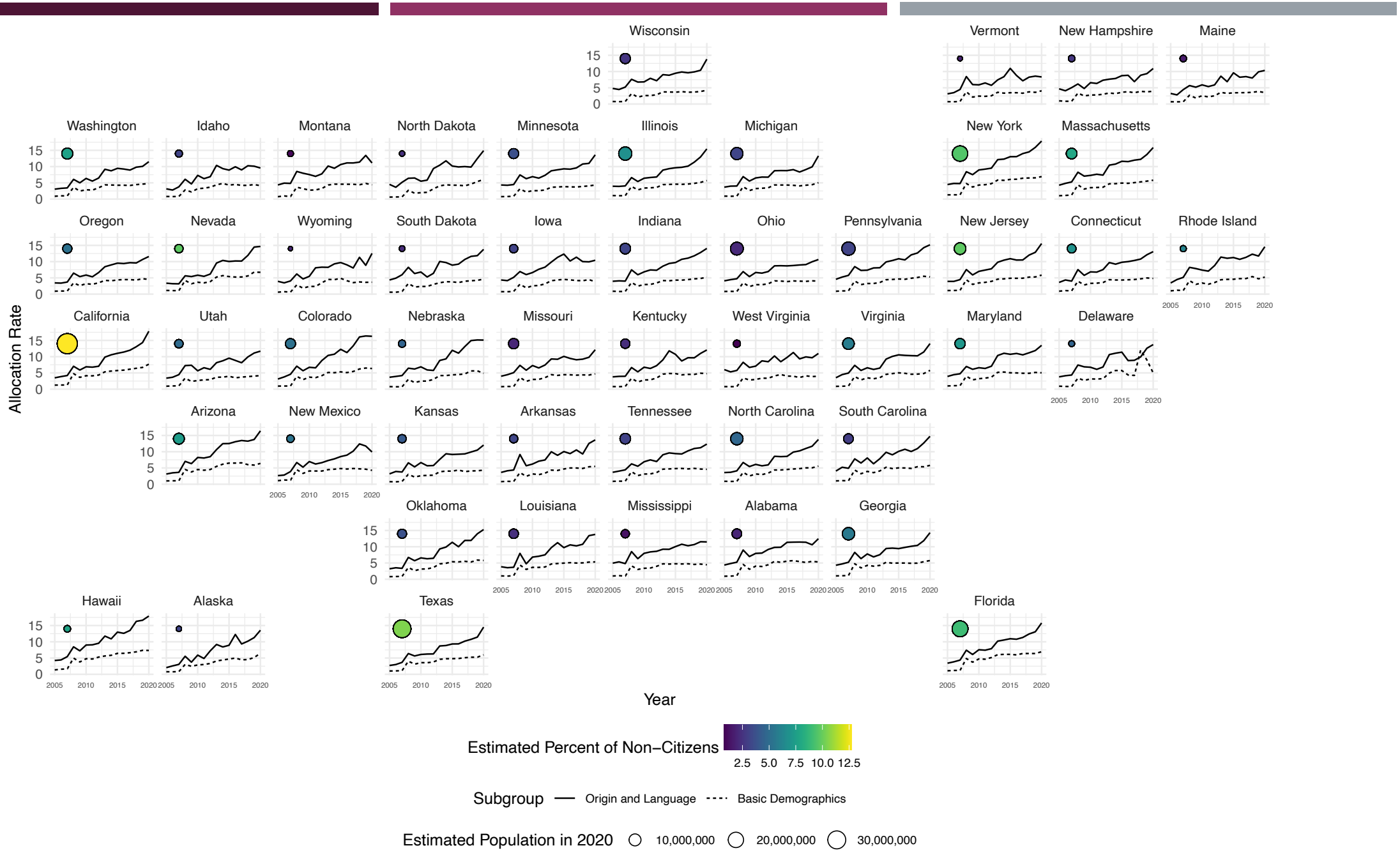
Is this person a citizen of the United States?

- ☐ Yes, born in the United States
- ☐ Yes, born in Puerto Rico, Guam, the U.S. Virgin Islands, or Northern Marianas
- ☐ Yes, born abroad of U.S. citizen parent or parents
- ☐ Yes, U.S. citizen by naturalization – *Print year of naturalization*

- ☐ No, not a U.S. citizen

DATA QUALITY AND DATA JUSTICE

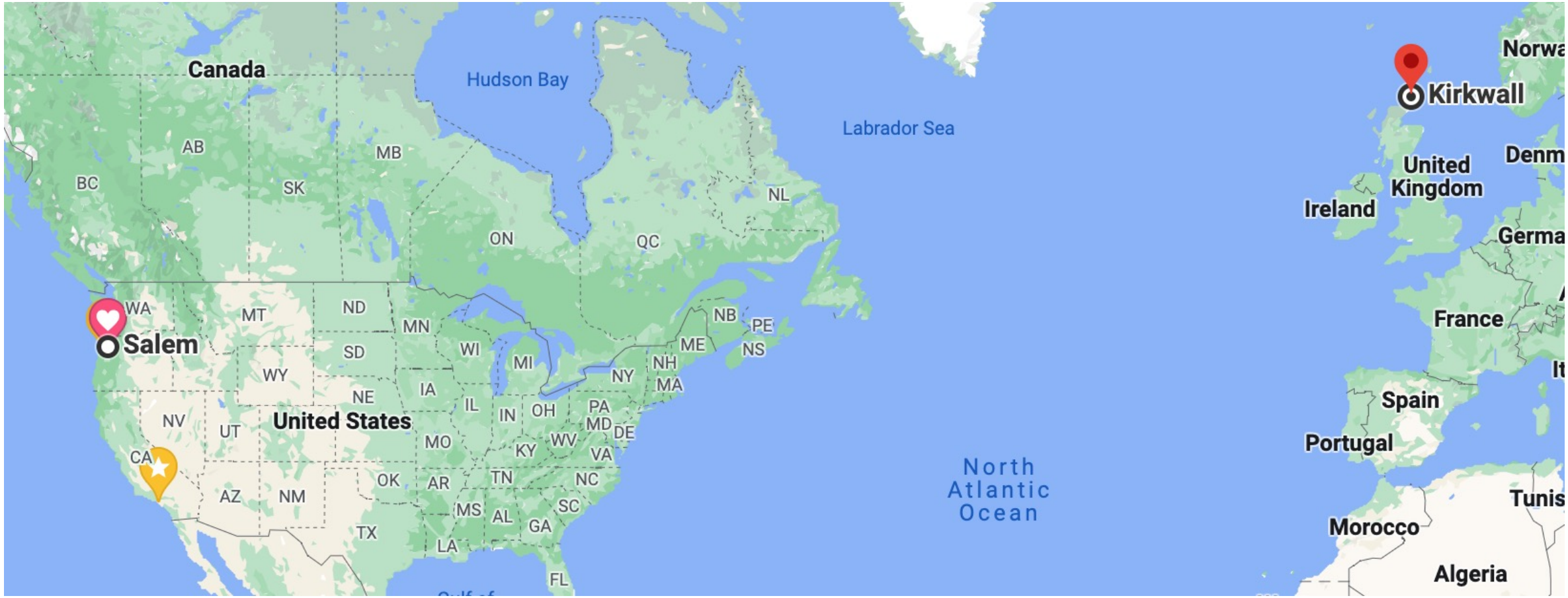



Wisconsin
Vermont
New Hampshire
Maine
Washington
Idaho
Montana
North Dakota
Minnesota
Illinois
Michigan
New York
Massachusetts
Oregon
Nevada
Wyoming
South Dakota
Iowa
Indiana
Ohio
Pennsylvania
New Jersey
Connecticut
Rhode Island
California
Utah
Colorado
Nebraska
Missouri
Kentucky
West Virginia
Virginia
Maryland
Delaware
Arizona
New Mexico
Kansas
Arkansas
Tennessee
North Carolina
South Carolina
Oklahoma
Louisiana
Mississippi
Alabama
Georgia
Hawaii
Alaska
Texas
Florida
Allocation Rate
Year
Estimated Percent of Non–Citizens
2.5 5.0 7.5 10.0 12.5
Subgroup
Origin and Language
Basic Demographics
Estimated Population in 2020
10,000,000
20,000,000
30,000,000

RESEARCH PHILOSOPHY

Look for inspiration everywhere!

DATA SCIENCE AND ARCHEOLOGY



DATA SCIENCE AND ARCHEOLOGY

SUMMER RESEARCH WITH STUDENTS

HELP WITH RESEARCH?!

Please let me know if you are interested in helping me out with some research

… or if you have your own research interests

FOUR OF A KIND! MAKE NEW FRIENDS

FOUR OF A KIND! MAKE NEW FRIENDS

Directions:

- Everyone will get a playing card
- Find people with the same rank as you
- Introduce yourselves and exchange information
  - Notecard: Please write names and contact information (email) for your group

THE SYLLABUS: (ON WISE)

ROADMAP/CONTRACT/OPERATOR'S MANUAL

COURSE MEETINGS AND STRUCTURE

Course meeting times (Pacific Time Zone)

- Tuesday and Thursday 2:30 – 4:00 PM
- in **FORD 102**

OFFICE HOURS

Office Hour Times:

- **In-person:** Tuesday 10:30 - 11:30 AM and Thursday 12:00 - 1:00 pm
  - In-person office hours will be in my Salem office.
- **Zoom <u>Only</u> Office Hours:** Monday and Wednesday 10:30 - 11:30 AM
- OR by appointment
  - If these times don't work for you please send me an email to schedule a meeting. I am happy to work with you.

OFFICE HOURS

Office Hour Sign-up Rules:

- I will create a Zoom meeting for each office hour interval.
- Office hours are considered open meetings, so people can drop-in (like in physical office hours).
- In order to organize this process I have created these tables so that students can "reserve" times. I would encourage people with similar topics to group together and have larger meetings.
- You can reserve more than 15 minutes by adding your name to multiple slots (but remember that there is a high demand)
- Please be polite and do not remove anyone's name.

OFFICE HOURS

Monday Aug 29, 2022: 10:30 am - 11:30 am

Time Block	Mode	Class	Name(s)	Topic
10:30 - 10:45	Zoom			
10:45 - 11:00	Zoom			
11:00 - 11:15	Zoom			
11:15 - 11:30	Zoom			

Tuesday Aug 30, 2022: 10:30 am - 11:30 am

Time Block	Mode	Class	Name(s)	Topic
10:30 - 10:45	SALEM			
10:45 - 11:00	SALEM			
11:00 - 11:15	SALEM			
11:15 - 11:30	SALEM			

COURSE FORMAT

Course Format:

- We will employ multiple pedagogical approaches to engage with the material. Classes will be composed of **interactive lectures, discussions, activities, and programming labs**. Please come to class ready to actively engage. I wish to promote a vibrant classroom community so that we may all learn from and teach each other.

LEARNING OUTCOMES

- **List** the major steps of the data science process and **apply** in practice.
- **Wrangle** different types of data using base R and Tidyverse functions in R.
- **Implement** exploratory data analysis techniques in R, including the following:
  - Data visualization using ggplot
  - Numerical data analysis (measures of central tendency, measures of precision & uncertainty, quantiles)
  - Categorical data analysis (proportions)
- **Perform** a simple linear regression analysis and **interpret** the results in context of the problem
- **Communicate** statistical results to audiences with varying degrees of statistical background knowledge.

COURSE CONTENT

- **Course Calendar** (see WISE page)
- **Units:**
  - Learning outcomes have been grouped into three units:
    - Unit #1: Working with data
      - Weeks 1 - 5
    - Unit #2: Exploratory Data Analysis
      - Weeks 6 - 10
    - Unit #3: Describing relationships between variables
      - Weeks 11 – 15

TIME COMMITMENT

- Willamette's Credit Hour Policy states that for every hour of class time there is an expectation of 2-3 hours of work outside of class.

- Since this class meets for three hours a week, you should expect 6-9 hours outside of class engaged in course-related activities.

LEARNING RESOURCES

TEXTBOOK

Introduction to Data Science: Data Analysis and Prediction with R

(iDatSci)

By Rafael A. Irizarry

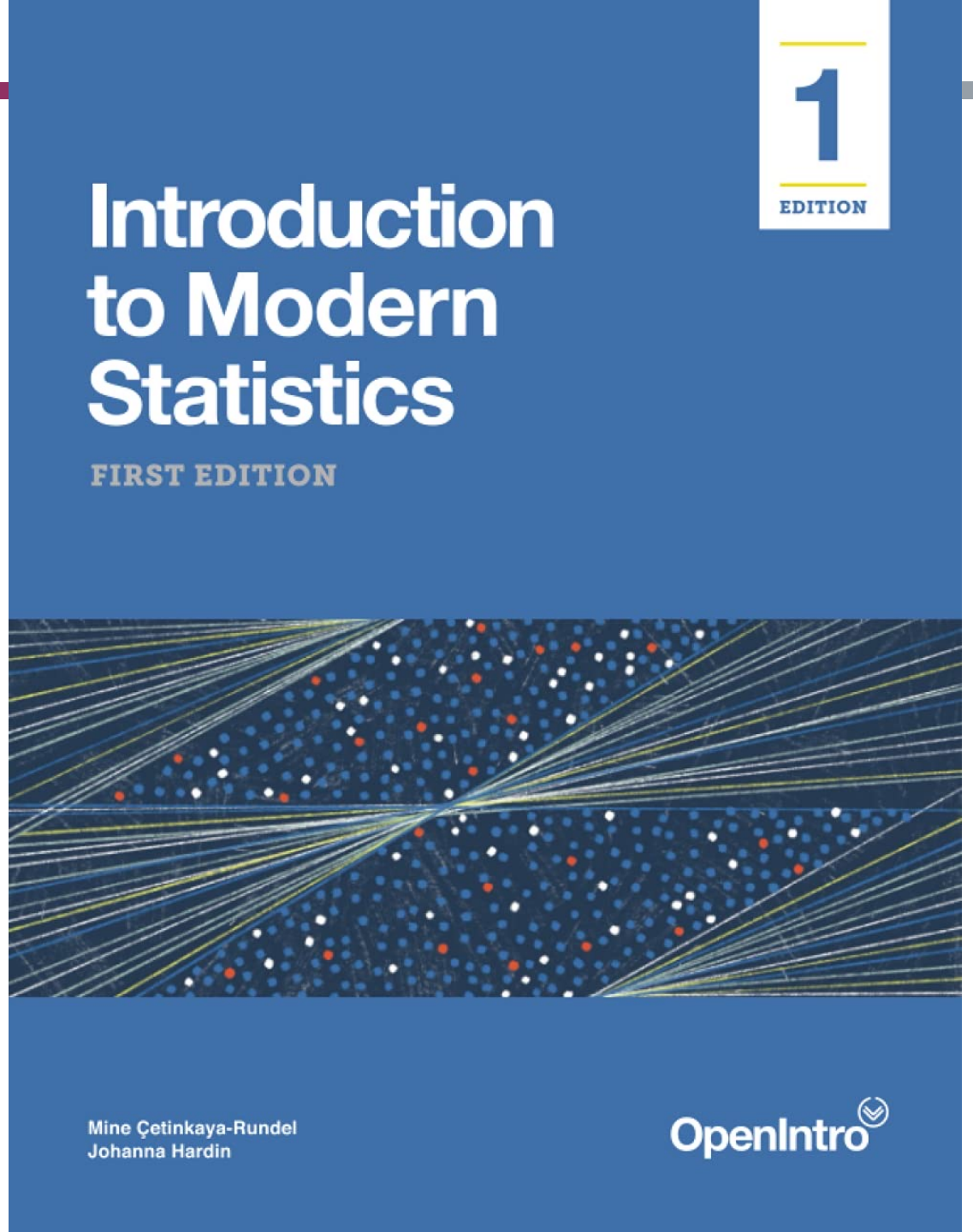
TEXTBOOK

Introduction to Modern Statistics

(iModStat)

By

Mine Çetinkaya-Rundel and Johanna Hardin

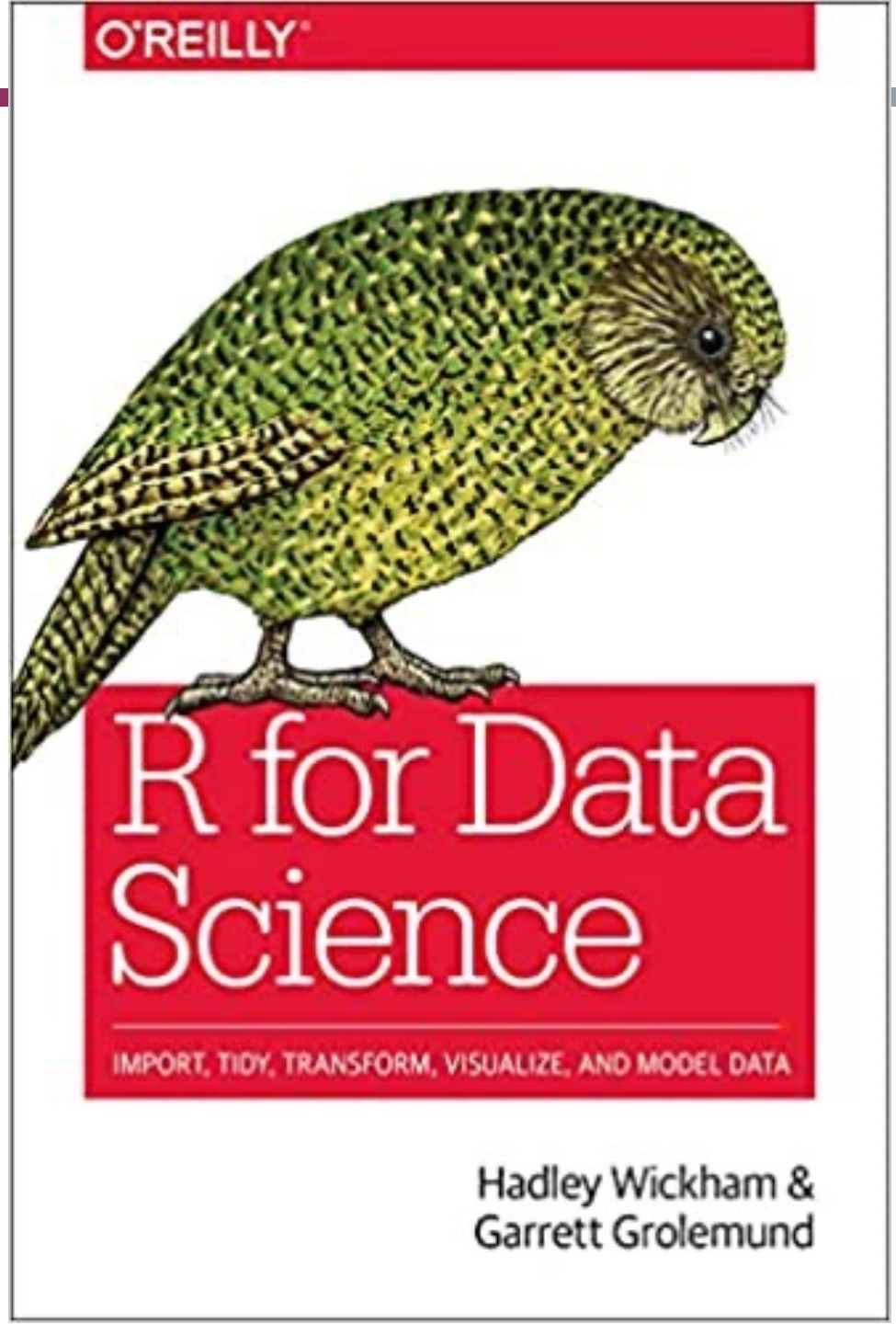


TEXTBOOK

R for Data Science

(R4DS)

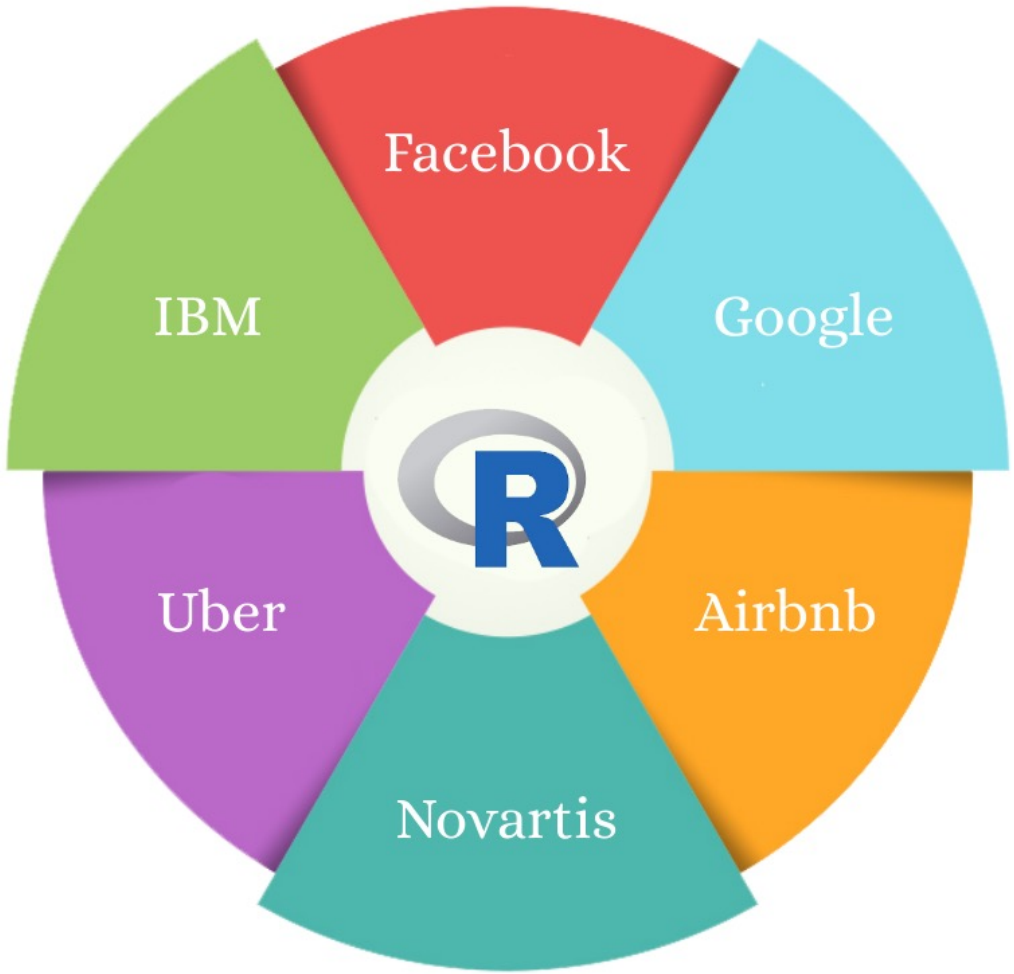
By Wickham and Grolemund

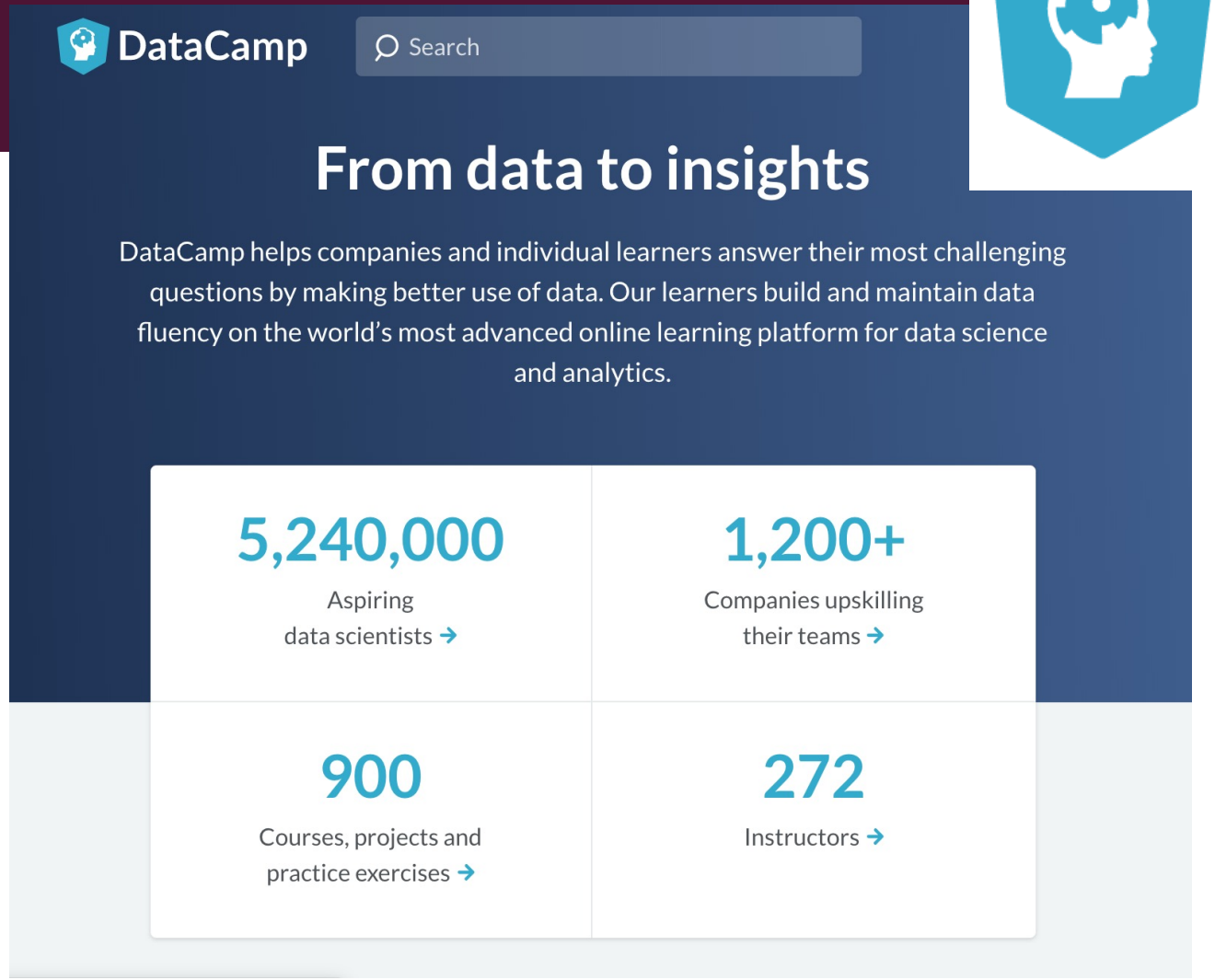


SOFTWARE

R and RStudio (free, opens source statistical computing software).

Data Science Companies that Use R




DataCamp
Search
From data to insights
DataCamp helps companies and individual learners answer their most challenging questions by making better use of data. Our learners build and maintain data fluency on the world's most advanced online learning platform for data science and analytics.
5,240,000
Aspiring data scientists →
1,200+
Companies upskilling their teams →
900
Courses, projects and practice exercises →
272
Instructors →

DataCamp

LEARN

Introduction to the Tidyverse ›

Data wrangling

4 hours to go

Keep Making Progress

PRACTICE

Introduction to R ›

DataCamp

HOW WILL I EARN MY GRADE?

HOMEWORK

- There will be approximately ***ten*** homeworks throughout the term (due on ***Thursdays at 11:59pm***) to reinforce the concepts from the reading as well as the material that we address in class.
- It is important that homework be done on-time because there will be opportunities to share your work with peers and get constructive feedback.
- Homeworks will be a combination for **practice problems, modules from Data Camp, and milestone check-ins for your projects**.

SUBMITTING WORK ONLINE

In order to cut down the use of paper and decrease passing papers back and forth ***all materials will be submitted online via WISE.***

Solutions should be written neatly or typeset and should use complete sentences.

- Please submit assignments as a single file.
- If you want to write assignments by hand you can scan them or use handy smartphone apps that stitch together photos of your assignments and turn them into good quality black and white text scans. I like to use Adobe Scan at home (it's free!)
- If an assignment requires programming please include your R markdown file that includes your solution, executable code, explanation, and any necessary output and/or plots and submit the R Markdown in WISE.

HOMEWORK

- If an assignment requires programming please also include your printed R markdown file which should include your solution, executable code, explanation, and any necessary output and/or plots.

ACADEMIC HONESTY ON HOMEWORK

While collaboration is valuable to the learning experience you must write your own solutions independently. Your written work should honestly represent your understanding of the material.

- You should acknowledge your collaborators and tutors by listing their names at the start of your solution
- The internet is a great source of information about mathematics; you are welcome to search information about the material of the course online, but you should not search for solutions to specific problems in the homework.
- The following are considered violations of the honor policy: cheating, fabrication, assisting, tampering, and plagiarism

LATE WORK POLICY

- Everyone will be allotted one free late assignment (must still be turned in within a week).
- If work is submitted after the specified due date the following table illustrates the deduction schedule from the total points earned:

Time late	Deduction
0 hr < x ≤ 24 hrs	20%
24 hr < x ≤ 48 hrs	40%
48 hr < x ≤ 72 hrs	60%
72 hr < x ≤ 96 hrs	80%
96 hr < x ≤ 129 hrs	100%

ASSESSMENTS

MIDTERMS

There will be two midterm exams, each worth 15%. Exams will cover theory as well as application. You can expect to be given a dataset and asked a series of questions to answer using the skills developed during the course. You may have one page of notes for the exam. No collaboration is allowed on exams.

- Midterm 1 covers content from Unit #1
- Midterm 2 covers content from Unit #2

ASSESSMENT #1

Skills Check: In order to make sure that students are on track for success and able to work on their final project, this midterm skills check will assess statistical programming skills in both base R and tidyverse.

ASSESSMENT #2

Exploratory Data Analysis: This assessment will cover the EDA process and construction of graphics in ggplot2, as well as the use of summary statistics.

Students will also be asked to dissect graphics to identify major components (geometry, aesthetics, scales, facets, etc). Finally, students will be asked to critically evaluate graphics for best practices and comment on elements that were particularly successful or poorly executed.

WEIGHTING FOR GRADED ITEMS

Item	Weight
Interactive Labs/Activities/Participation	10%
DataCamp Assignments	15%
Project Milestones	10%
Midterm #1: Skills Check	20%
Midterm #2: Exploratory Data Analysis	20%
Final Project and Presentation	25%
Total	100%

GRADES

	A = 92.5% or more	**A-** = 92.4% - 90%
B+ = 89.9% - 87.5%	**B** = 87.4% - 82.5%	**B-** = 82.4% - 80%
C+ = 79.9% - 77.5%	**C** = 77.4% - 72.5%	**C-** = 72.4% - 70%
D+ = 69.9% - 67.5%	**D** = 67.4% - 62.5%	**D-** = 62.4% - 60%
	F =59.9% or less	

CLASS PROJECT

PARTNER PROJECTS

The final project should be thought of as long-term projects with checkpoints throughout the term (please don't procrastinate). These are partner projects. Each team will write a statistical report on a dataset of their choosing by exploring the dataset, identifying scientific questions of interest, and applying the appropriate statistical learning (statistics or machine learning) methods.

More details on ***Project Milestones***

PARTNER PROJECTS

Public resources for datasets:

- SWD list of free public datasets: **LINK HERE**
- UC Irvine Machine Learning Repository http://archive.ics.uci.edu/ml/index.php
- Kaggle Datasets https://www.kaggle.com/datasets

Ultimately, this paper project is a great opportunity for you to showcase your data analysis skills and build your portfolio. Consider submitting your paper to a student paper competition.

- Undergraduate Class Project Competition (USCLAP)

UNDERGRADUATE CLASS PROJECT COMPETITION (USCLAP)

Website: https://www.causeweb.org/usproc/usclap

"Project Scope for USCLAP Competition:

The class project competition is for undergraduate students who conduct projects as part of an introductory or intermediate level statistics or data science course. Most projects submitted to the USCLAP competition involve analyzing real data using existing statistical techniques. Students may choose any topic on which to conduct a study and students may use existing data or collect their own."

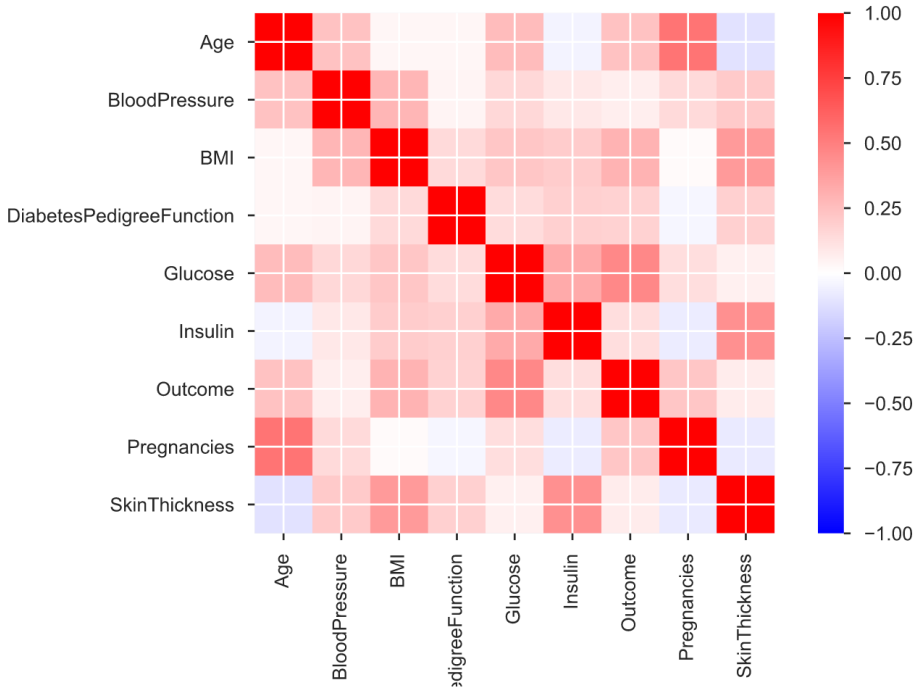
THE MOST IMPORTANT THING….

Find your passion!

Solve "wicked" problems!

EXAMPLES OF PROJECTS

Predicting diabetes diagnosis



EXAMPLES OF PROJECTS

Predicting whether a mushroom is poisonous or edible

EXAMPLES OF PROJECTS

Building a model for basketball (and baseball) performance (and salaries)

EXAMPLES OF PROJECTS

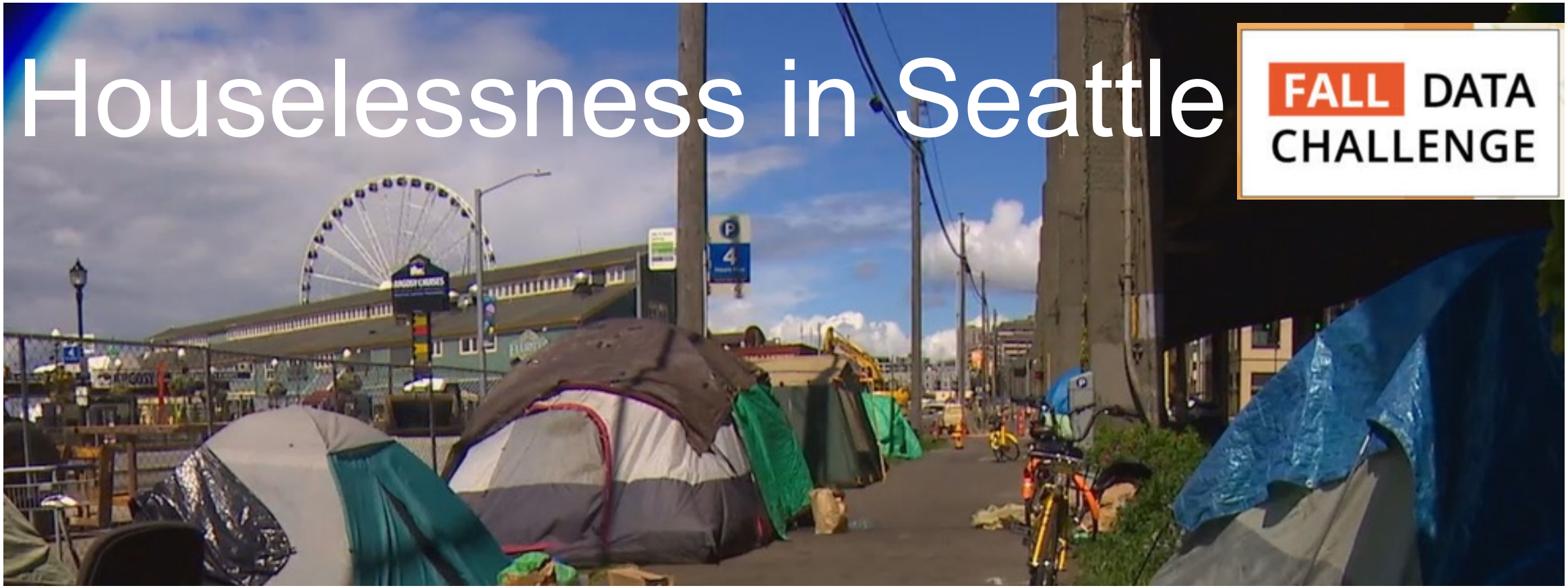
Testing the significance of demographics in sexual assault or violence against the LBTQ+ community

EXAMPLES OF PROJECTS

Testing the significance of chemical compounds in wine quality

OTHER OPPORTUNITIES

ASA FALL DATA CHALLENGE - 2019



ASA FALL DATA CHALLENGE - 2020

ASA FALL DATA CHALLENGE - 2021

ASA FALL DATA CHALLENGE - 2022

COURSE CALENDAR

COURSE CALENDAR

Course: Willamette University - DATA 151 - Introduction to Data Science		
Term: FALL 2022		**Assignments**
1A: Aug 30 Topics: ● Syllabus, Introductions, and Community Building ● Service available in accessible support *Motivating Questions: What is data science? What does data science mean to you?* **Related Reading:** ● *United Nations: Big Data for Sustainable Development* (link)	**1B: Sept 1** Topics: ● The Power of Data ● Types of variables *Motivating Questions: What is data? Where does data come from? How is data used for good?* **Related Reading:** ● iMStat - Ch 1 ○ 1: Hello Data	**Tasks: (Due 9/8)** ● Sign up to DataCamp ● Student Survey **HW #1: (Due 9/8)** ● Dear Data **Project: (Due 9/8)** ● Complete project partner survey
2A: Sept 6	**2B: Sept 8**	**HW #2: (Due 9/15)**

COURSE POLICIES

COMMUNICATION POLICY

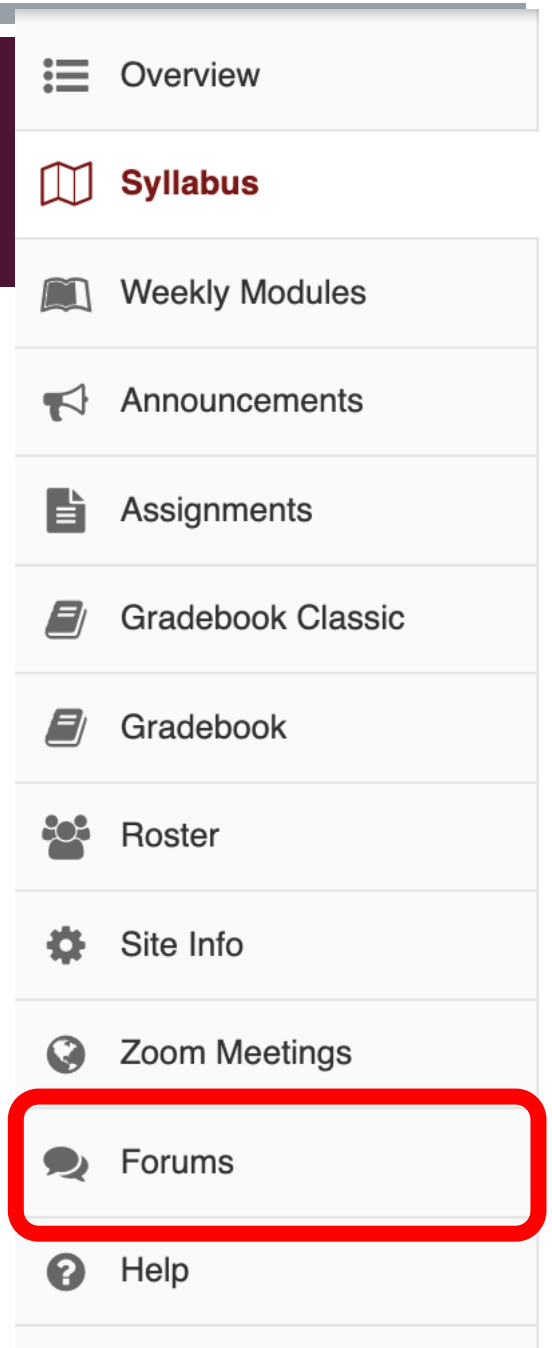
During Class

Please feel free to ask questions *during* class. Before and after class is not an ideal time to ask questions, please refrain until office hours or send me an email for out of class questions or concerns.

COMMUNICATION POLICY

FORUMS Discussion Board on WISE

- If, outside of class, you have a (non-personal) question please post it in the course forum. This allows other students to respond to the question, if they know the answer, and allows all students to view responses from the Instructor. This can be a great resource and is typically faster than email. All general questions asked through email will be redirected to be posted within the class forum.
  - Remember, if you have a question, many students probably have the same question! We are a classroom learning community.
- Extra credit points will be awarded to students who help out their peers by answering their questions.



COMMUNICATION POLICY

Email

- Put **DATA 151** and **YOUR NAME** in the title of your email.
- Only use email if you have a personal question, otherwise use the Discussion Board
- Exhaust all course materials and resources before emailing.
- Do not expect an instant response.
  - Your email will be responded to within 48 hours.

TECHNOLOGY IN THE CLASSROOM

- Computers will be used for a portion of the class to teach R statistical programming.
- While computers are being used for in class activities with R, students will be asked to stay on task and refrain from social media, browsing the internet, checking email, etc.
- In general, the use of electronic devices (computers, cell phone, tablets, etc) is prohibited during all other times because it is distracting for both the instructor and fellow students.
- If you have a specific reason for needing to use technology (for example, note taking) please let me know.

ZOOM ETIQUETTE

- Dress appropriately.
- Set up an intentional space where the class is going to happen.
- Be aware of your surroundings. Be mindful of what is in your camera's view. Make sure no personal information is visible and that you are in an area with good lighting.
- Let your family members know that you are on a "live call". It is important that you let everyone know so that they are aware that their voice/image might appear or be recorded.
- Make sure you are muted when not talking. Muting when you are not speaking allows others to hear more clearly without distraction.
- Think before you write a response.
- Be yourself. Be kind and respect others.
- Ask clarifying questions in the chat box.
- Be present and engaged in your learning.

WILLAMETTE POLICIES

- Academic Integrity
- Diversity and Disability Statement
- Commitment to Positive Sexual Ethics
- DACA/Undocumented Student Advocate
- Religious Practice
- SOAR Center Offerings
- Trans Inclusive Classroom Space
- WU Antiracism Statement

WHERE CAN I GET HELP?

ADDITIONAL RESOURCES

- Students are always welcome and encouraged to come to Prof. Smalley's office hours.
- Our class has embedded tutors, who are hosting office hours.
  - Please find the most up-to-date information on the hours on the WISE page.
- If additional help is needed, tutors can be arranged (each student is allotted 2 hours of tutoring a week) through the Student Success Hub.
- Help can also be obtained from the QUAD center

Link: https://willamette.edu/offices/quad/index.html

THE QUAD CENTER

Quantitative Understanding, Analysis and Design

FORD 224

HOURS:

Monday 6:00 - 9:00 PM
Tuesday 6:00 - 9:00 PM
Wednesday 3:00 - 9:00 PM
Thursday 3:00 - 9:00 PM
Friday 3:00 - 4:30 PM
Sunday 3:00 - 6:00 PM

HAYDEN

he/him/his

DATA 151 and QUAD Tutor

Major: 3+1 BS/MS Data Science

Minors: CS and Math

Can help with:

DATA 151/152, DATA 252, DATA 351, CS 151, CS 152, CS 351, MATH 138, MATH 280,

Programming and Software:

R Studio, Python, SQL, Excel

Shifts:

Sunday 3:00 – 4:30 PM

Monday 6:00 – 7:30 PM

MITCHELL

he/him/his

DATA 151 and QUAD Tutor

Majors: Computer Science and Data Science

Minor: Math

Can help with:

DATA 151/152, DATA 252, CS 151, CS 152, CS 351, MATH 249, MATH 251W, MATH 253, MATH 280,

Programming and Software:

R Studio, Python, C++

Shifts:

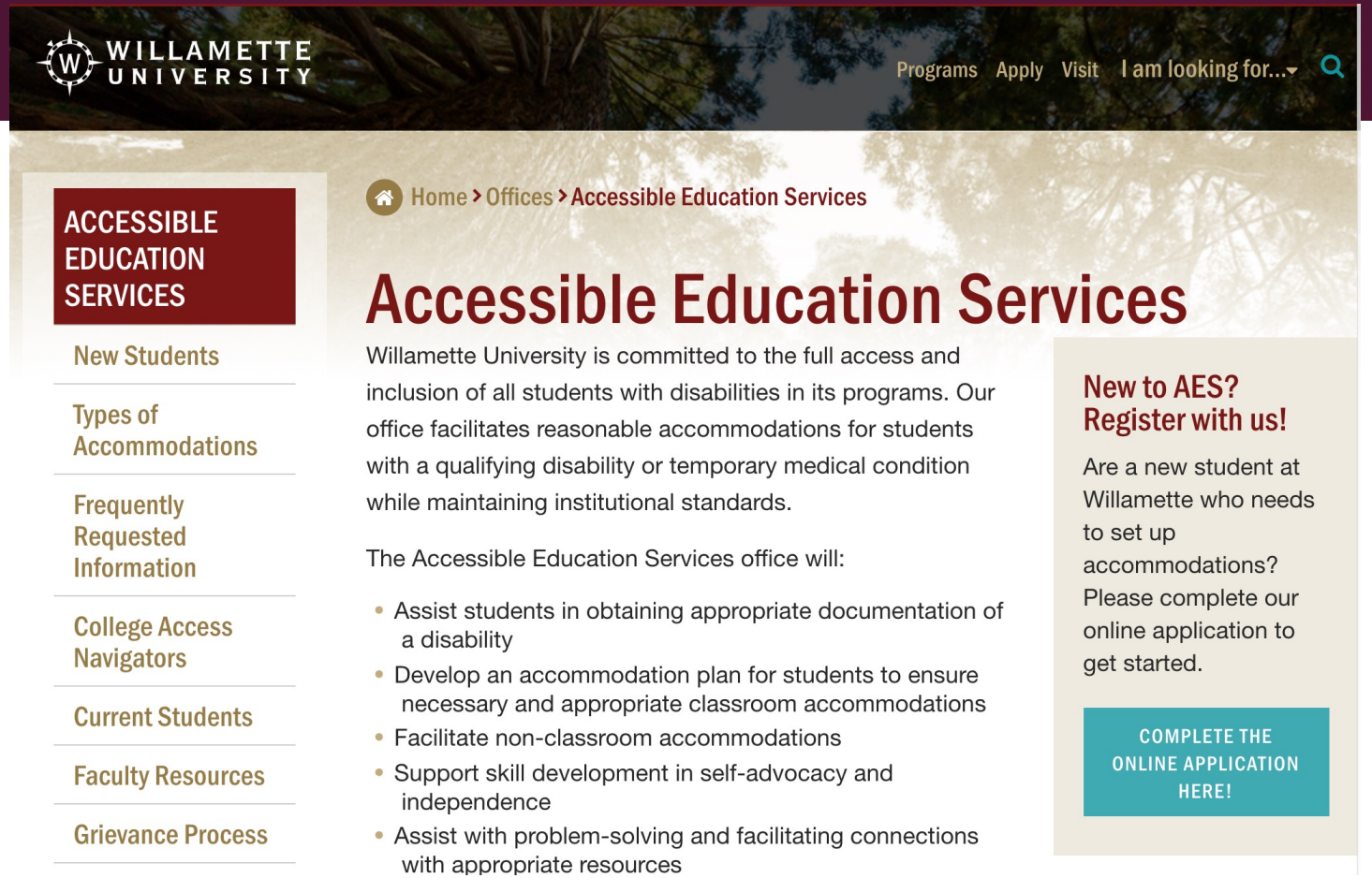
Monday 7:30 – 9:00 PM

Wednesday: 6:00 – 9:00 PM

Friday: 3:00 – 4:30 PM

DISABILITY SUPPORT SERVICES

- Students with disabilities are also encouraged to contact the Accessible Education Services office in Matthews 103 at 503-370-6737 or Accessible-info@willamette.edu to discuss a range of options to removing barriers in the course, including accommodations.
-



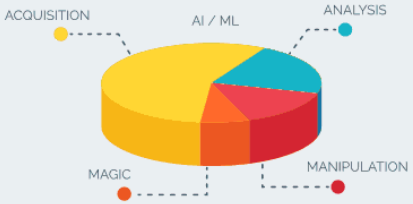
LET'S GET STARTED!

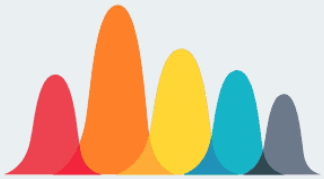
WHAT IS DATA?

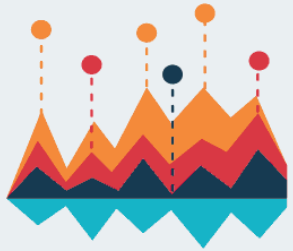
WHAT IS DATA?

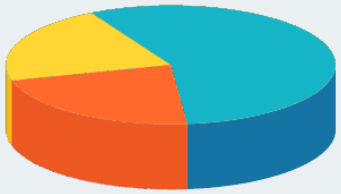
In the pursuit of knowledge, **data** is a collection of discrete states that convey information, describing quantity, quality, fact, statistics, other basic units of meaning, or simply sequences of symbols that may be further interpreted.

Data: Singular or Plural?









As a singular mass noun (like *information*)

All the data is available for download.

Our data shows that online businesses grew in 2020.

As the plural of *datum* (esp. in scientific and academic writing)

The collected data are then analyzed.

The data indicate that bias is pervasive across all fields of research.

#GOALS - DATA LITERACY

WHAT IS DATA LITERACY?

The ability to read, understand, create, and communicate data as information.

- Reading and writing/creating data
- Use analytics programs and interpret outputs

THREE FACETS

- Storing Data
- Handling Data
- Presenting Data

STORING DATA

- Storing Data
  - 1. Data is imperfect
  - 2. Know what the data quality is for
    - Missing data and incorrect data
  - 3. Data types

HANDLING DATA

- Handling Data
  - 4. Technical skills
  - 5. Choosing and using suitable tools
  - 6. Business skills / Domain knowledge

PRESENTING DATA

- Presenting Data
  - 7. Visualization skills
  - 8. Social skills
    - Talking to different audiences about data
  - 9. There's no one true answer
  - 10. Critical thinking
    - Decisions based on data

WHAT IS BIG DATA?

BIG DATA – FOUR V'S

- **Volume:** Scale of data
- **Variety:** Different forms
- **Velocity:** How fast sent
- **Veracity:** Trustworthiness
  - *Uncertainty, bias, or inaccuracies in the data make information less valuable for meaningful analysis and decision making*

THE 4 V'S OF BIG DATA

Volume

SCALE OF DATA

40 ZETTABYTES
of data will be created by 2020, an increase of 300 times from 2005

6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION

2.5 QUINTILLION BYTES
of data are created each day

Most companies in the U.S. have at least
100 TERABYTES
of data stored

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES

30 BILLION PIECES OF CONTENT
are shared on facebook every month

4 BILLION + HOURS OF VIDEO
are watched on You Tube each month

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

Velocity

ANALYSIS OF STREAMING DATA

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

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

Veracity

UNCERTAINITY OF DATA

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 data was inaccurate

Reference : http://www.ibmbigdatahub.com/infographic/four-vs-big-data

NEXT TIME:
READ AND DISCUSS "UNITED NATIONS: BIG DATA FOR SUSTAINABLE DEVELOPMENT"