

## Lecture 1

# Overview and Values of Data Science

**Haoyu Yue** / [yohaoyu@u.washington.edu](mailto:yohaoyu@u.washington.edu)

Ph.D. Student, Interdisciplinary Urban Design and Planning  
University of Washington

RE 519 Real Estate Data Analytics and Visualization

Course Website: [www.yuehaoyu.com/data-analytics-visualization/](http://www.yuehaoyu.com/data-analytics-visualization/)

Autumn 2025



Welcome to start the journey in  
**data science**



**Haoyu Yue** ['haʊ.ju 'ju:.eɪ] | he/him | [yohaoyu@u.washington.edu](mailto:yohaoyu@u.washington.edu)

Ph.D. Student, Interdisciplinary Urban Design and Planning

Ph.D. Researcher, Urban Infrastructure Lab

Graduate Student, Statistics

**Research Area**

artificial intelligence, climate planning, renewable energy, quant methods



**Christian Phillips**

Ph.D. Student, Interdisciplinary Urban Design and Planning

Ph.D. Researcher, Urban Infrastructure Lab & Washington State Real Estate Research Center

2024



**Siman Ning**

Ph.D. Student, Interdisciplinary Urban Design and Planning

Ph.D. Researcher, Urban Infrastructure Lab

2022, 2023



**Feiyang Sun, Ph.D.**

Assistant Professor

University of California, San Diego

2020, 2021



# Brief Intro

## Class Survey Results

- **Did you take RE 506 Quantitative Methods, URBAN 520 Quantitative Methods, or at least 1 course about statistics in your undergrad or graduate study?**
  - I took at least one statistics course. **(17/18)**
  - I'm taking a statistics course this quarter.
  - I have no statistics background so far.
  - I know a lot about statistics without needing to take a class. **(1/18)**

# Brief Intro

## Class Survey Results

- **Did you take RE 597 Real Estate Data Modeling, URBAN 504 Introduction to GIS, or at least 1 course about GIS in your undergrad or graduate study?**
  - I took at least one GIS course. **(1/18)**
  - I'm taking a GIS course this quarter. **(1/18)**
  - I have no GIS background so far. **(15/18)**
  - I know a lot about GIS without needing to take a class.
  - I took R E 397 in my undergrad, but that was not about GIS. **(1/18)**

# Brief Intro

## Class Survey Results

- **Are you scared of coding (Python, R, C++, etc.)?**
  - Yes, I hate coding. **(1/18)**
  - Not much, but I can try some easy lines of coding. **(3/18)**
  - I don't know because I don't have any experience. **(8/18)**
  - No experience, but I'd love to learn with the help of AI. **(4/18)**
  - I code much in my work. **(2/18)**
- **Do you have any working experience in the real estate or other fields?**
  - Yes, but not related to any data science. **(15/18)**
  - Yes, and related to data science!
  - No working experience. **(3/18)**

# Brief Intro

## Class Survey Results

- Which programming language do you prefer to learn? Yes, I hate coding.
  - R (1/18)
  - Python (5/18)
  - I don't have any preference and am open to both. (9/18)
  - I don't know the difference, but I want to do data science in my future career. (1/18)
  - I don't have any interest in data science; I took this class because it was required. (2/18)

# Brief Intro

## Self Introduction

- Your name
- Your pronouns
- Your affiliation
- Year in the program
- Your experience/interests with data analysis and visualization
- Your expectation from this class
- Recent happiness (accomplishment, hobby, adventure, etc.)



# Course Overview

## Class Sessions

- **Two sessions per week** (MW 3:30-4:50 pm)
  - Typically, the first half of the class will be some sort of lecture, and the rest of the class will be dedicated to lab time, where you can work in groups
- **Location**
  - Mechanical Engineering Building 245 (sorry, the room is not so desirable!)
- **Office hours**
  - By appointment via [link](#) - online via Zoom
  - Anytime after each class session

# Course Overview

## Pre-requisites, Materials, Class Website, and Canvas

- **No prerequisites required for this class**
  - Basic knowledge of any programming language is appreciated.
- **Readings and slides**
  - Slides will be published on the website before the class
  - A few required readings, but lots of optional readings and resources
  - Can be accessed via the links on the website, although some may require a UW NetID login
- **Course website**
  - <https://www.yuehaoyu.com/data-analytics-visualization/>
  - Add to your bookmark for this quarter!
- **Canvas**
  - Only for survey, quiz, lab submission, and grading

# Course Overview

## Class Communication

- **Asynchronous Discussion Board**

- Using Ed Discussion for announcements, discussion, and technical questions
- linked from Canvas
- We'll try and respond to all questions by the end of each working day
- Please use the Ed Discussion as the first place to ask general questions. If you have a question about the course material or assignment, other students may have the same question. If you email me with a question like this, I will ask you to post it on the discussion board.
- I also encourage students to answer each other's questions on the discussion board

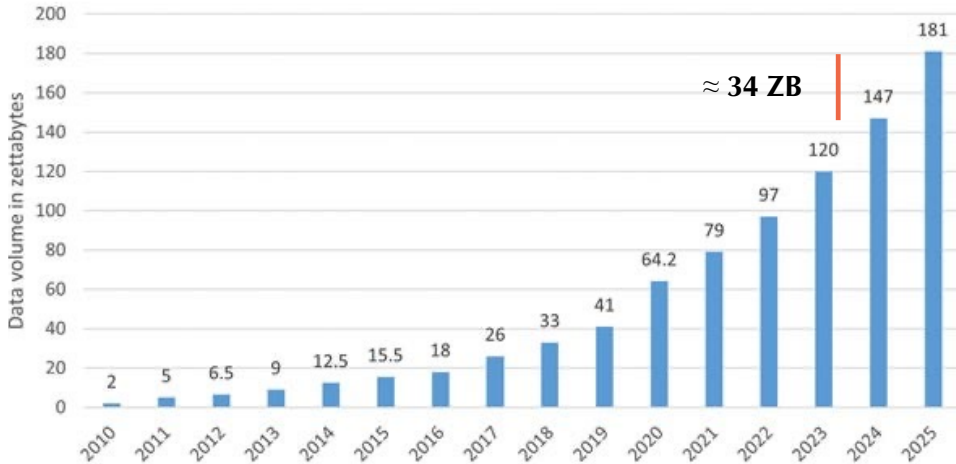
- **Email communication**

- Please use [yohaoyu@uw.edu](mailto:yohaoyu@uw.edu). Additionally, please put RE 519 in the subject line.

# The Value of Data Analysis and Visualization

## The Exponential Growth of Data

Volume of data created and replicated worldwide (source: IDC)



~2x every 3 years

1 ZB ≈

200 trillion photos taken by iPhone

Take 70 photos for each person on earth every day for a year

34 ZB ≈

Every 36 seconds, take one photo for each person on earth for a year

Source: Exponential Growth of Data  
<https://medium.com/@mwaliph/exponential-growth-of-data-2f53df89124>

# A DAY IN DATA

The exponential growth of data is undisputed, but the numbers behind this explosion - fuelled by internet of things and the use of connected devices - are hard to comprehend, particularly when looked at in the context of one day

**500m**

tweets are sent every day

Twitter

**294bn**

billion emails are sent

Radicati Group

**320bn**

emails to be sent each day by 2021

**306bn**

emails to be sent each day by 2020

**3.9bn**

people use email

**4PB**

of data created by Facebook, including

**350m** photos

**100m** hours of video watch time

Facebook Research

**4TB**

of data produced by a connected car

Intel

## ACCUMULATED DIGITAL UNIVERSE OF DATA

**4.4ZB**

2019

**44ZB**

2020

PwC

## DEMYSTIFYING DATA UNITS

From the more familiar 'bit' or 'megabyte', larger units of measurement are more frequently being used to explain the masses of data

Unit	Value	Size
<b>b</b> bit	0 or 1	1/8 of a byte
<b>B</b> byte	8 bits	1 byte
<b>KB</b> kilobyte	1,000 bytes	1,000 bytes
<b>MB</b> megabyte	1,000 <sup>2</sup> bytes	1,000,000 bytes
<b>GB</b> gigabyte	1,000 <sup>3</sup> bytes	1,000,000,000 bytes
<b>TB</b> terabyte	1,000 <sup>4</sup> bytes	1,000,000,000,000 bytes
<b>PB</b> petabyte	1,000 <sup>5</sup> bytes	1,000,000,000,000,000 bytes
<b>EB</b> exabyte	1,000 <sup>6</sup> bytes	1,000,000,000,000,000,000 bytes
<b>ZB</b> zettabyte	1,000 <sup>7</sup> bytes	1,000,000,000,000,000,000,000 bytes
<b>YB</b> yottabyte	1,000 <sup>8</sup> bytes	1,000,000,000,000,000,000,000,000 bytes

\*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.

**65bn**

messages sent over WhatsApp and two billion minutes of voice and video calls made

Facebook

Searches made a day

**5bn**

Searches made a day from Google

**3.5bn**

Smart Insights

**463EB**

of data will be created every day by 2025

IOE

**95m**

photos and videos are shared on Instagram

Instagram Business

**28PB**

to be generated from wearable devices by 2020

Statista

# The Value of Data Analysis and Visualization

How many 6 here? Which number is more likely to pair with A?

A	B	A	C	A	A	B	A	A	C	A	B
9	7	8	8	7	9	7	9	9	6	8	7

A	B	A	B	A	A	A	B	A	B	A	B
9	7	8	7	8	9	7	9	9	7	8	7

A	A	C	A	A	C	A	B	C	A	C	A
8	9	6	8	9	6	9	7	6	9	6	9

# The Value of Data Analysis and Visualization

How many 6 here? Which number is more likely to pair with A?

A	B	A	C	A	A	B	A	A	C	A	B
9	7	8	8	7	9	7	9	9	6	8	7

A	B	A	B	A	A	A	B	A	B	A	B
9	7	8	7	8	9	7	9	9	7	8	7

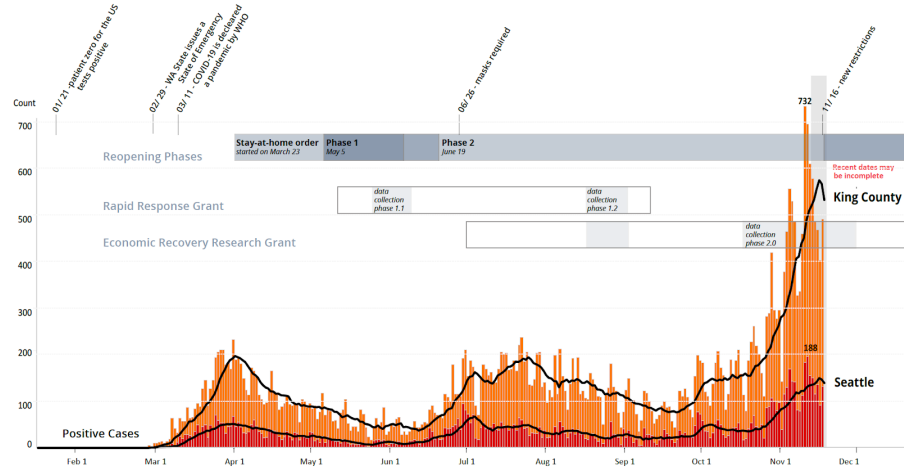
A	A	C	A	A	C	A	B	C	A	C	A
8	9	6	8	9	6	9	7	6	9	6	9

Pair	Frequency
A ~ 7	2
A ~ 8	6
A ~ 9	11
B ~ 7	7
B ~ 9	1
C ~ 6	5
C ~ 8	1

# The Value of Data Analysis and Visualization

## Why data visualization?

- Because our brains love graphics.
- **Analyze** data to support reasoning (exploratory visualization)
  - Develop and assess hypotheses
  - Find patterns / errors in data
  - Expand memory
- **Communicate** information to others (narrative/explanatory visualization)
  - Share and persuade
  - Collaborate and revise



Source: Whittington, J. and Proksch, G. "Design Determinants of COVID-19 Impacts to Food-Related Essential Business and Service" University of Washington, February 9, 2021

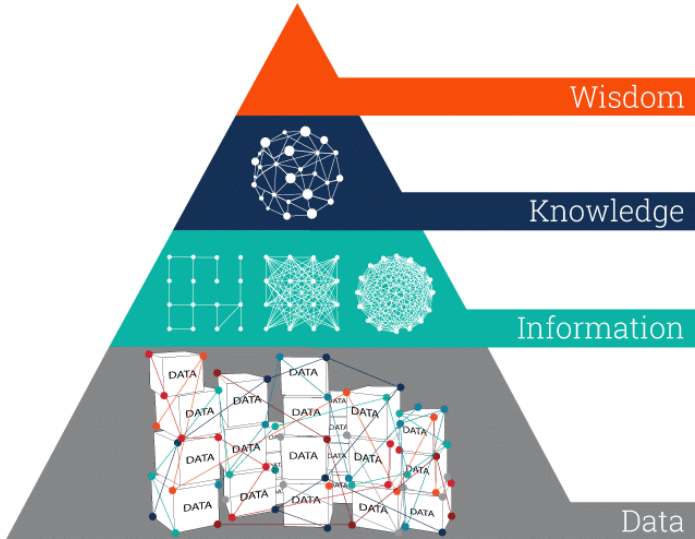


"The ability to take data—to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it—that's going to be a hugely important skill in the next decades, [...] because now we really do have **essentially free and ubiquitous data**. So, the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

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

# The Value of Data Analysis and Visualization

## The DIKW Pyramid

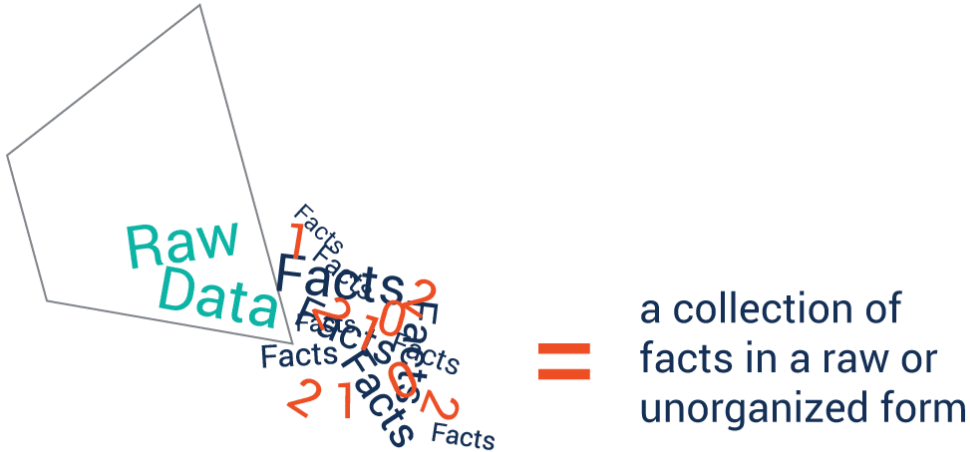


Each step up  
the pyramid  
answers  
questions  
about and  
adds value  
to the initial data.

Source: <https://www.ontotext.com/knowledgehub/fundamentals/dikw-pyramid/>

# The Value of Data Analysis and Visualization

## The DIKW Pyramid - Data



Base building block - Raw **Data**

Source: <https://www.ontotext.com/knowledgehub/fundamentals/dikw-pyramid/>

# The Value of Data Analysis and Visualization

## The DIKW Pyramid - Information



Second building block - Derived **Information**

Source: <https://www.ontotext.com/knowledgehub/fundamentals/dikw-pyramid/>

# The Value of Data Analysis and Visualization

## The DIKW Pyramid - Knowledge



Third building block - Relevant **Knowledge**

Source: <https://www.ontotext.com/knowledgehub/fundamentals/dikw-pyramid/>

# The Value of Data Analysis and Visualization

## The DIKW Pyramid - Wisdom



The top of the DIKW hierarchy - Guiding **Wisdom**

Source: <https://www.ontotext.com/knowledgehub/fundamentals/dikw-pyramid/>

# The Value of Data Analysis and Visualization

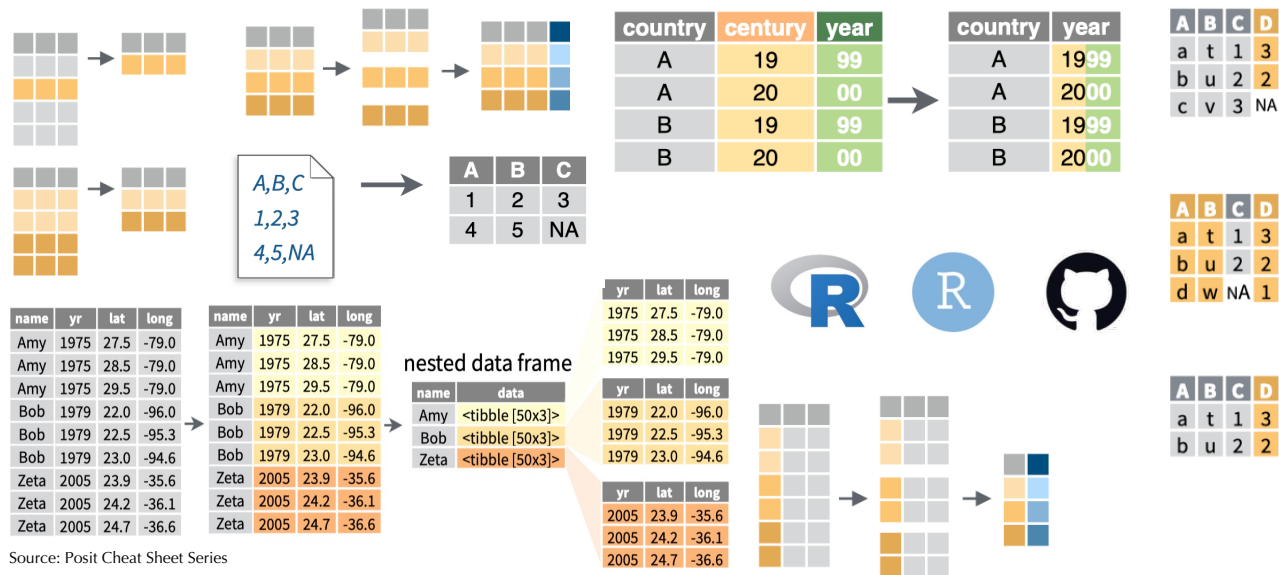
## Example of Job Posting

- Strong organizational and analytical skills
- Ability to provide efficient, timely, reliable and courteous service to customers
- **Ability to effectively present information**
- Requires knowledge of financial terms and principles
- Conducts basic financial analysis
- Ability to comprehend, analyze, and interpret documents
- Ability to solve problems involving several options in situations
- **Requires intermediate analytical and quantitative skills**
- **Experience in analyzing data to draw business-relevant conclusions and in data visualization techniques**
- **Technical expertise in techniques regarding data models and database design development**
- Sound knowledge of and experience with **reporting/dashboarding packages**, i.e., Power BI, **Tableau**, Datastudio, SSRS, Plx
- Strong knowledge of databases (MS SQL, BigQuery, etc)
- Adept at queries, report writing, and presenting findings
- Effective analytical skills with the ability to collect, organize, analyze, and disseminate significant amounts of information with attention to detail and accuracy

Source: CBRE

# Course Overview

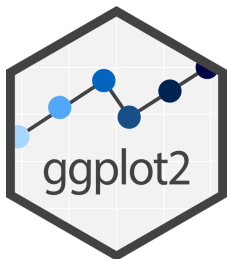
## Module 1 – Introduction and Data Processing



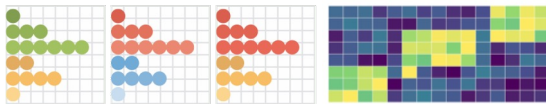


# Course Overview

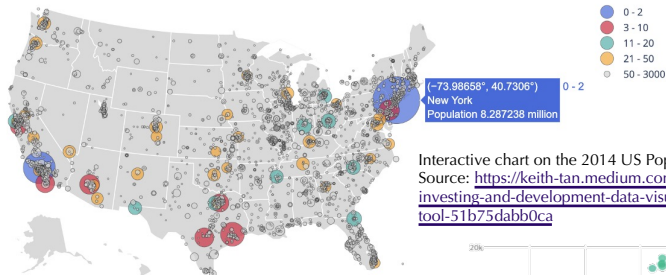
## Module 2 – Data Visualization



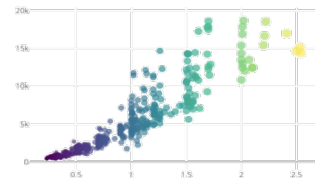
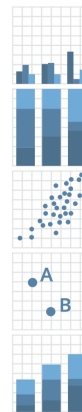
Source: Posit Cheat Sheet Series



Bubble Chart on real estate investment flow and GDP  
Source: <https://keith-tan.medium.com/real-estate-investing-and-development-data-visualization-as-a-tool-51b75dabb0ca>



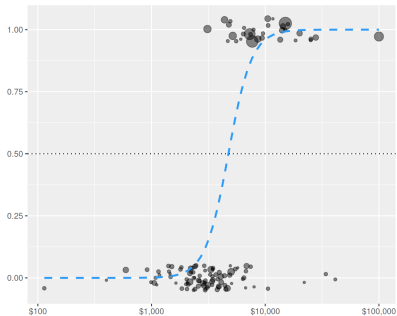
Interactive chart on the 2014 US Population, by city  
Source: <https://keith-tan.medium.com/real-estate-investing-and-development-data-visualization-as-a-tool-51b75dabb0ca>



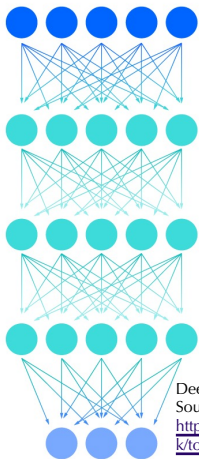
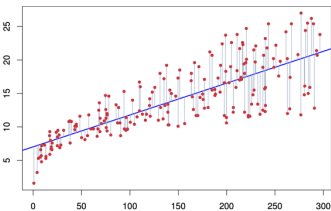
Infographic of SF Housing Price,  
Reddit: Original post by /u/surf2japan in r/bayarea

# Course Overview

## Module 3 – Data Modeling

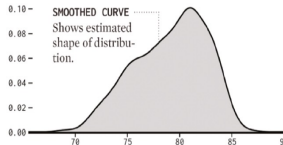


Source: Modern Data Science with R  
<https://mdsr-book.github.io/mdsr3e/>

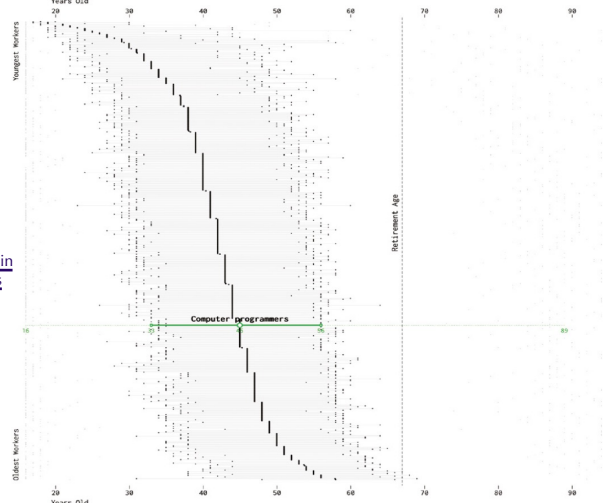
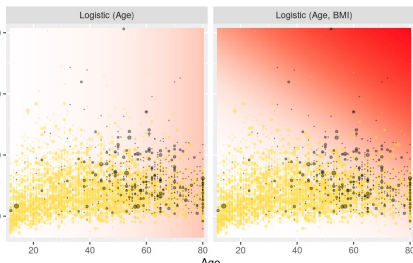


Deep neural network  
Source: IBM  
<https://www.ibm.com/think/topics/neural-networks>

DENSITY  
Represents proportion of population with given value.



Source: Visualize This - The Flowing Data Guide to Design, Visualization, and Statistics



Source: "Age and Occupation," Nathan Yau / 2007-Present FlowingData / <https://flowingdata.com/2021/09/30/age-and-occupation/> / last accessed February 08, 2024.

# Course Overview

## About this Course

### The class is about

- Data processing using R
- Exploratory data analysis
- Dashboard using Tableau and R
- Visual design for many forms of data
- Simple statistics methods

### The class is NOT about

- Analysis using Excel
- Advanced statistics methods
- Machine learning
- Artificial intelligence
- Databases and data management

### The class is

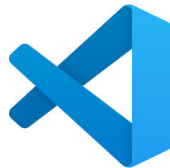
- Required for MSRE (Regular Standing)
- Selective for MSRE (Advanced Standing)
- Selective for the Master of Urban Planning
- An overview of data science and a gateway to future data science study

We will just introduce a little bit on those topics

# Course Overview

## Why are we using programming languages?

R vs. Python vs. Excel



**R** - A free, open-source software environment for statistical computing and graphics  
<http://www.r-project.org>

**RStudio** - An open-source integrated development environment (IDE)  
<https://posit.co/products/open-source/rstudio>

**GitHub Copilot** in RStudio  
<https://docs.posit.co/ide/user/ide/guide/tools/copilot.html>

# Course Overview

## Guest Speakers



### **Drew Dolan**

Principal, Fund Manager

DXD Capital

**Keywords:** Data-driven Real Estate, Self Storage Real Estate

October 27, 2025 (in-person at MEB 245)



### **Dr. Andy Krause**

Director of Applied Science in AI

Zillow

**Keywords:** Home Valuation, Machine Learning, Artificial Intelligence

Date & Location TBD

# Course Overview

## Statistics and Data Science @ UW

- **Coursework**

- [STAT 180/CSE 180/INFO 180 Introduction to Data Science](#)
- [CSE 583 Software Development for Data Scientists](#)
- [CSSS 508 Introduction to R for Social Scientists](#)
- [CSE 512 Data Visualization](#)
- [CSSS 569 Visualizing Data and Models](#)
- [CSE 416/STAT 416 Introduction to Machine Learning](#)
- [CEWA 567 Geospatial Data Analysis with Python](#)

- **Institutes and Centers (they usually offer a weekly seminar)**

- [eScience Institute \(Data Science\)](#)
- [Center for Studies in Demography and Ecology](#)
- [Center for Statistics and the Social Sciences](#)

# Course Overview

## Course Requirement – Labs (64%)

- In total, there will be 8 labs (8% each), and for each lab, there will be 2~3 parts.
- We will use class time to review certain parts of the labs, and you will have some tasks.
- In each lab, the expected finish time is ~1 to 2 hours after class.
- Each student is expected to submit their own lab, but study groups are allowed. But you're expected to acknowledge the names of collaborators along with a short description of the types of collaborations being done at the beginning of each lab submission.
- You may use generative AI tools, but please check the AI policy for each lab.
- You only need to submit once after each lab (due Monday at 11:59 PM PST), via Canvas.

# Course Overview

## Course Requirement - Data Analytics and Visualization Projects (30%)

- 1~3 students for each group.
- Start to think about the topic and data for the project now!
- The project could be, for example:
  - A modeling of interesting datasets to derive new insights
  - Pure visualization for some datasets
  - A replication of an interesting academic article
- The requirements include:
  - Project proposal (1 page; 5%)
  - Draft work presentation (in the last class, graded by peers and instructors; 10%)
  - Final delivery (could be any format, like report, website, poster; 15%)
- Details forthcoming.



# Course Overview

## Course Requirement – Participation and Extra Credits (6% + 3%)

- There will be several surveys and in-class quizzes (only graded on completion; 6%).
- There will be two ways to receive extra credits:
  - **Dataset sharing** (2%): From week 2 to 6, share high-quality online datasets directly related to real estate and housing on Ed Discussion.
    - The dataset cannot be repeated with the previous datasets shared by other students.
    - 10 unique datasets to receive 2%, 5 unique datasets to receive 1%.
  - **Course evaluation** (1%): We will leave some time in the last class to participate in the anonymous course evaluation.
- **Participating in the course evaluation is important to the course and me!**
  - Formal course evaluation occurs at the end of the quarter, university-widely. If you are experiencing a problem with the class, please let me know as soon as possible, as I might be able to make changes if needed within the course of the class.

# Course Overview

## Time Commitment, Final Grade, and Late Days

- According to the estimates for UW courses, it should take about **9** hours of work to complete a three-credit class each week. If you spend more than **6** hours beyond the classroom, please let me know as early as plan, and we will adjust the class content or specific study plans for you.
- The total scores will be curved and transformed into the UW numerical grading system for graduate courses, ranging from 4.0 to 1.7 in 0.1 increments as the final grade.
- Late days: You will have 6 penalty-free late days for assignments and projects (max 3 late days per assignment). Any delayed submission after the first 3 days will be penalized 10% per day for that specific assignment (but will not count towards your used late days).
  - Late days **cannot be used** for the project presentation and final delivery.

# Course Overview

## Generative AI Tools, GPT, Copilot, etc.

- **We encourage to use of generative AI tools beginning from Lab 4**
  - All sources, including AI tools, must be properly cited.
    - Example: “Describe the symbolism of the green light in the book The Great Gatsby by F. Scott Fitzgerald” prompt. ChatGPT, 13 Feb. version, OpenAI, 8 Mar. 2023.
- **Some caveats**
  - First, try finding the information you’re looking for yourself with Google, StackOverflow, etc.
    - It helps you learn how to format your questions
  - Don’t put sensitive information into ChatGPT...
  - ChatGPT will hallucinate and make up packages/functions that don’t exist
  - Sometimes, ChatGPT can make your code more unreadable or difficult to follow
- Note: Microsoft Copilot is the official AI tool for UW with commercial data protection.

# Course Overview

## Software and Computing

- R, R-studio: <http://www.r-project.org>, <http://www.rstudio.com>
- Tableau: <https://www.tableau.com/academic/students>
- Python: <https://www.python.org>
- Anaconda: <https://www.anaconda.com>
- GitHub: <https://github.com>
  
- Bring your computer (Windows, MacBook, or Linux is acceptable) to the class. If you have any trouble with having a computer, you may check the computing resources from the [college](#), [Student Technology Loan Program](#), or [UW libraries computer service](#).

# Lab Session 1-A

## Reminders

- Finish Lab 1-A (setting up R and RStudio)
- Talk to me if you are going to use Python, or try to waive this class, or have other questions or concerns.

**Thank you!**

**Haoyu Yue** / [yohaoyu@u.washington.edu](mailto:yohaoyu@u.washington.edu)

Ph.D. Student, Interdisciplinary Urban Design and Planning  
University of Washington

RE 519 Real Estate Data Analysis and Visualization

Course Website: [www.yuehaoyu.com/data-analytics-visualization/](http://www.yuehaoyu.com/data-analytics-visualization/)

Autumn 2025

The course was developed based on previous instructors: Christian Phillips, Siman Ning, Feiyang Sun  
Cover page credits: Visax