WEEK 2: DATA SCIENCE / ANALYSIS

Dr. Kai Li
School of Information Sciences
University of Tennessee, Knoxville
Spring 2025

Review of Week 1

- Introduction
- Concepts about data and data analysis

- Questions / Comments?

Discussion of week 1

- The key take aways are:
  - *Every type of investigation / methods has its own pros and cons.*
    - Alignment of data, questions, and methods
  - *We often used them together (mixed-method approach) in research design, particularly using qualitative methods to identify potential patterns and then using quantitative methods to confirm these patterns.*

The video

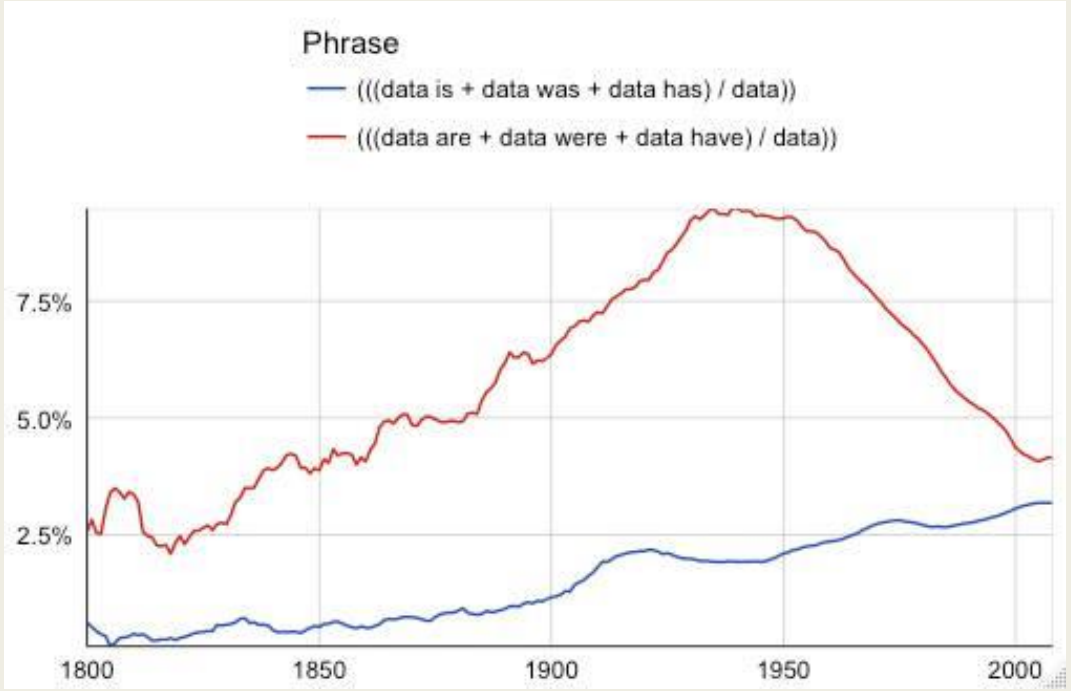
- Thank you for sharing your insights!
  - *Participatory, hand-on experience*
- The roles played by technologies can be exaggerated.

Overview of this lecture

- What is data science?
- What are the central data science skills?

Is data singular or plural?

- Created using the ngramr package of R.
- Will discuss the package later!



What is data science?

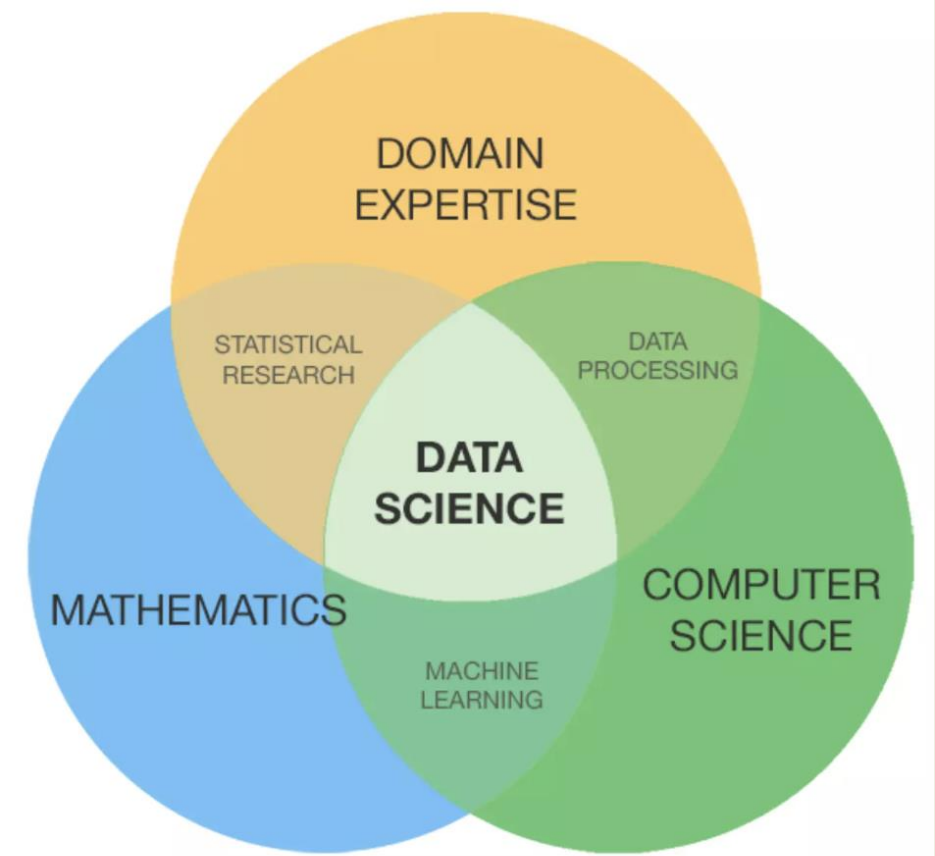
- Not surprisingly, there are many different definitions of data science, from some very different perspectives:
  - *A field*
  - *A set of methods*

What is data science?

- As a research field, data science is believed to be highly interdisciplinary.
- What are some of the central fields related to data science?
  - *Computer science*
  - *Statistics / Mathematics*
  - *But data science is also applied in many other fields!*

What is data science?

- But data science is more than just programming and statistics.
- Like information science, it is also about applying the methods to solve real-world problems in specific domains.



Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

What is data science?

- As a matter of fact, the composition of data science can be much more complex than the aforementioned statements.
- For example:

Accordingly, a *discipline-based data science formula* is given as follows:

$$\begin{aligned} data\ science\ = &\ statistics + informatics + computing + communication \\ &+ sociology + management \,|\, data + environment + thinking, \end{aligned} \quad (1)$$

Cao, L. (2017). Data science: a comprehensive overview. *ACM Computing Surveys (CSUR)*, *50*(3), 1-42.

What is NOT data science?

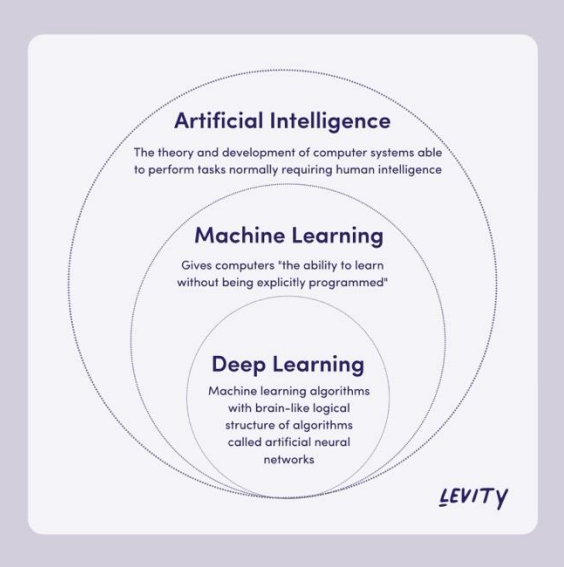
- Data science is not (just) computer science.
  - *Data science focuses on applying CS and computational methods in specific contexts but not developing these methods.*
- Data science is not (just) information science.
  - *Data science focuses on using computational methods to draw conclusions from the data, while information science focuses on the whole lifecycle of data.*
  - *But DS and IS are largely competing in the same space.*

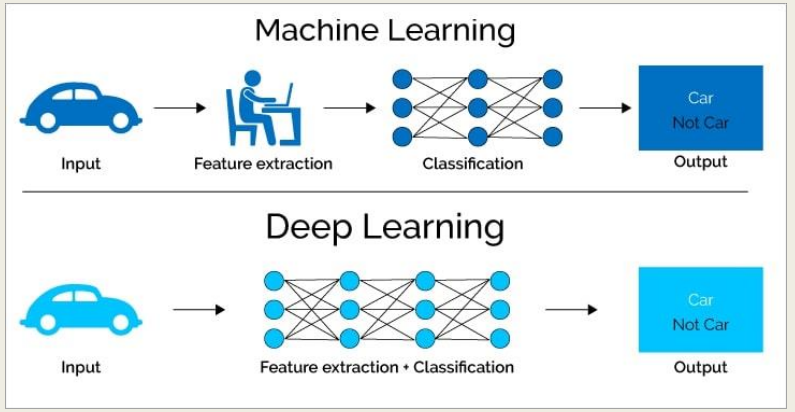
Data science vs. data analytics

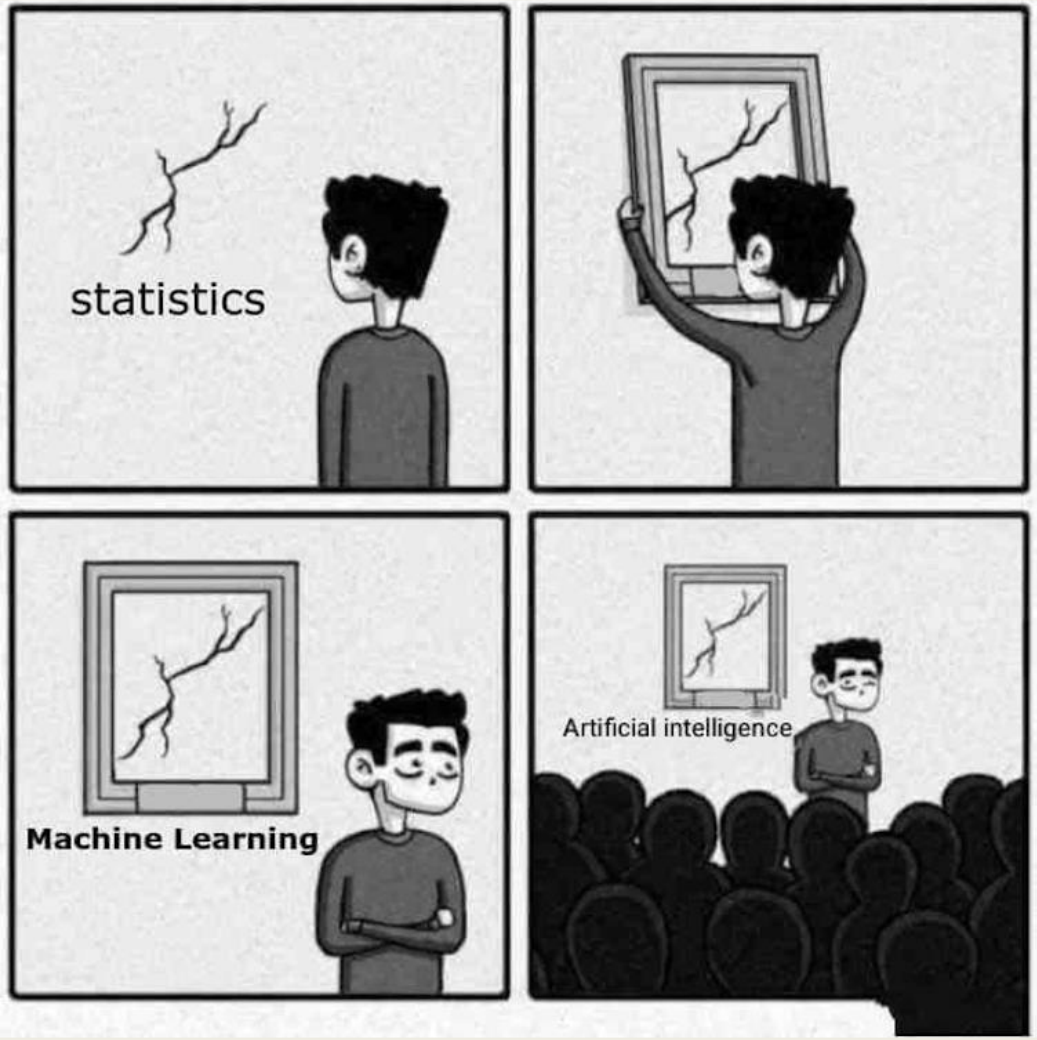
- A common feature of data science and data analytics is that they are both standardized methods that can be used to answer a wide spectrum of questions.
- And… they are both largely based on computer science and statistics.
- A lot of data analytics methods are rebranded as data science today, but it does not change the nature of these methods.
- Data science methods are more strongly connected to the bigness and complexity of data and rely on computational approaches.
  - *Such as deep learning*

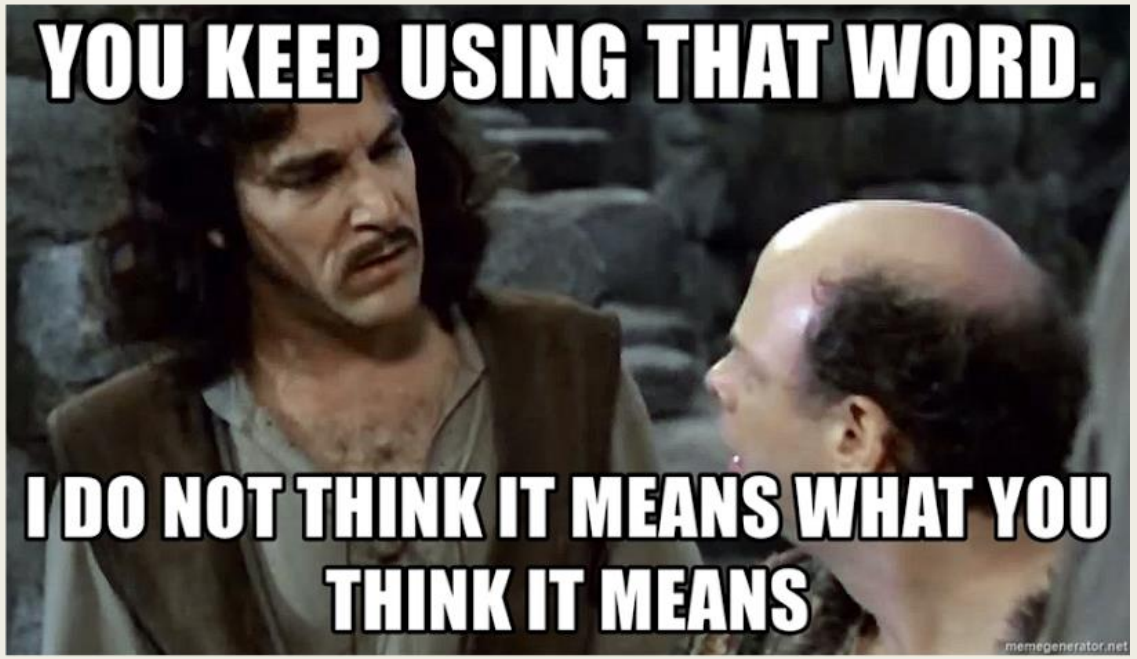
Deep learning vs. Machine learning

- A key difference between ML and DL is that DL is not statistical-driven!
  - *ML is based on statistical models, such as regression and classification.*
  - *But DL is much less driven by statistical methods.*
- DL is much more “theory-less” and much less transparent. → XAI (explainable ai)









https://windowsontheory.org/2022/06/20/the-uneasy-relationship-between-deep-learning-and-classical-statistics/

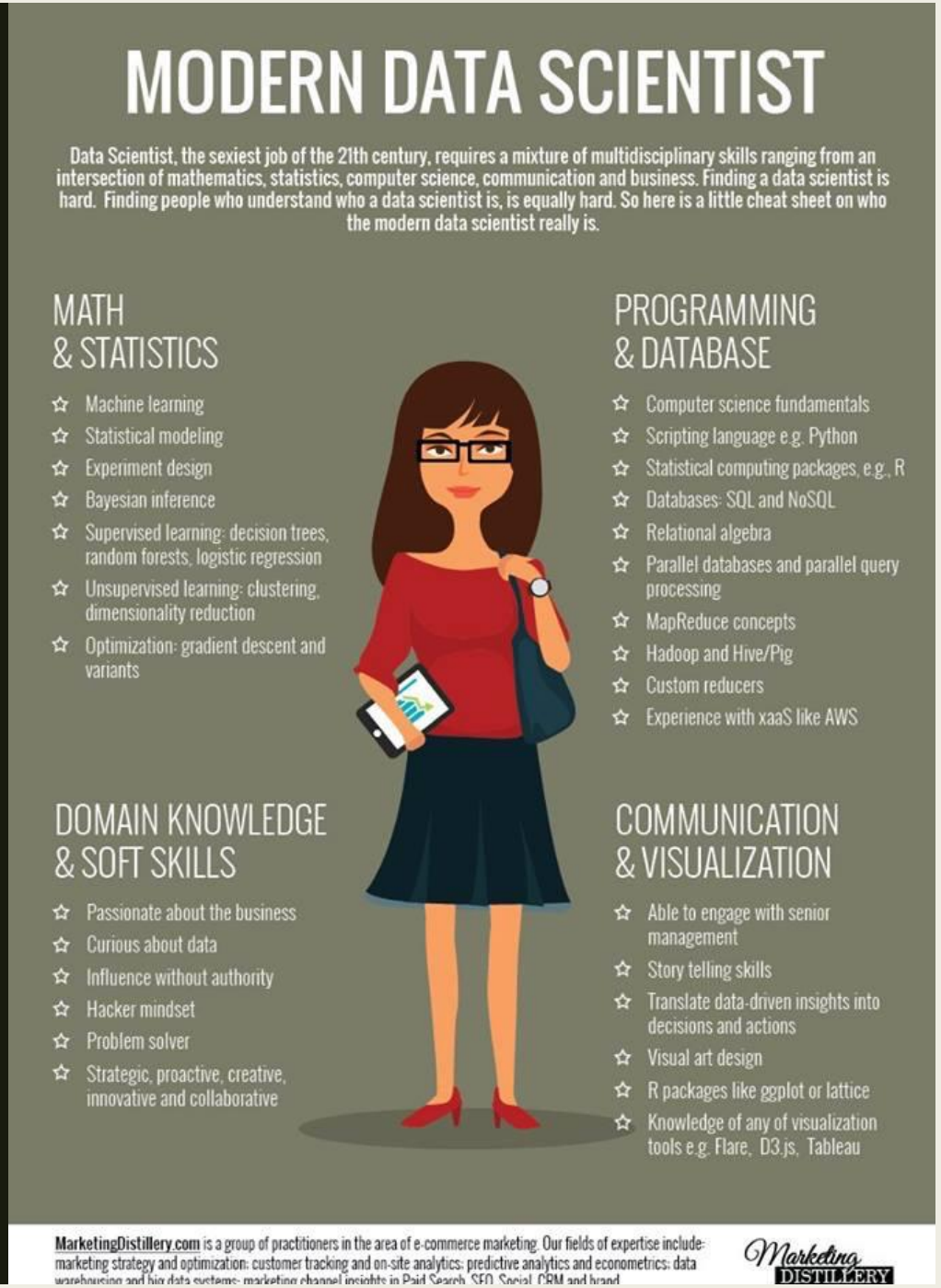
So what's the point so far?

- Being a buzzword, a lot of things are being re-branded as "data science," making this topic very broad and confusing.
- Even though we are talking about data science in the beginning of this course, I am not going to say data science is everything that we are covering in this class.
- HOWEVER, it is important to understand (1) what data science is, (2) what benefits it can bring to you (especially in the job market) and (3) how to brand yourself as a data scientists in the job market.

Skills of data scientists

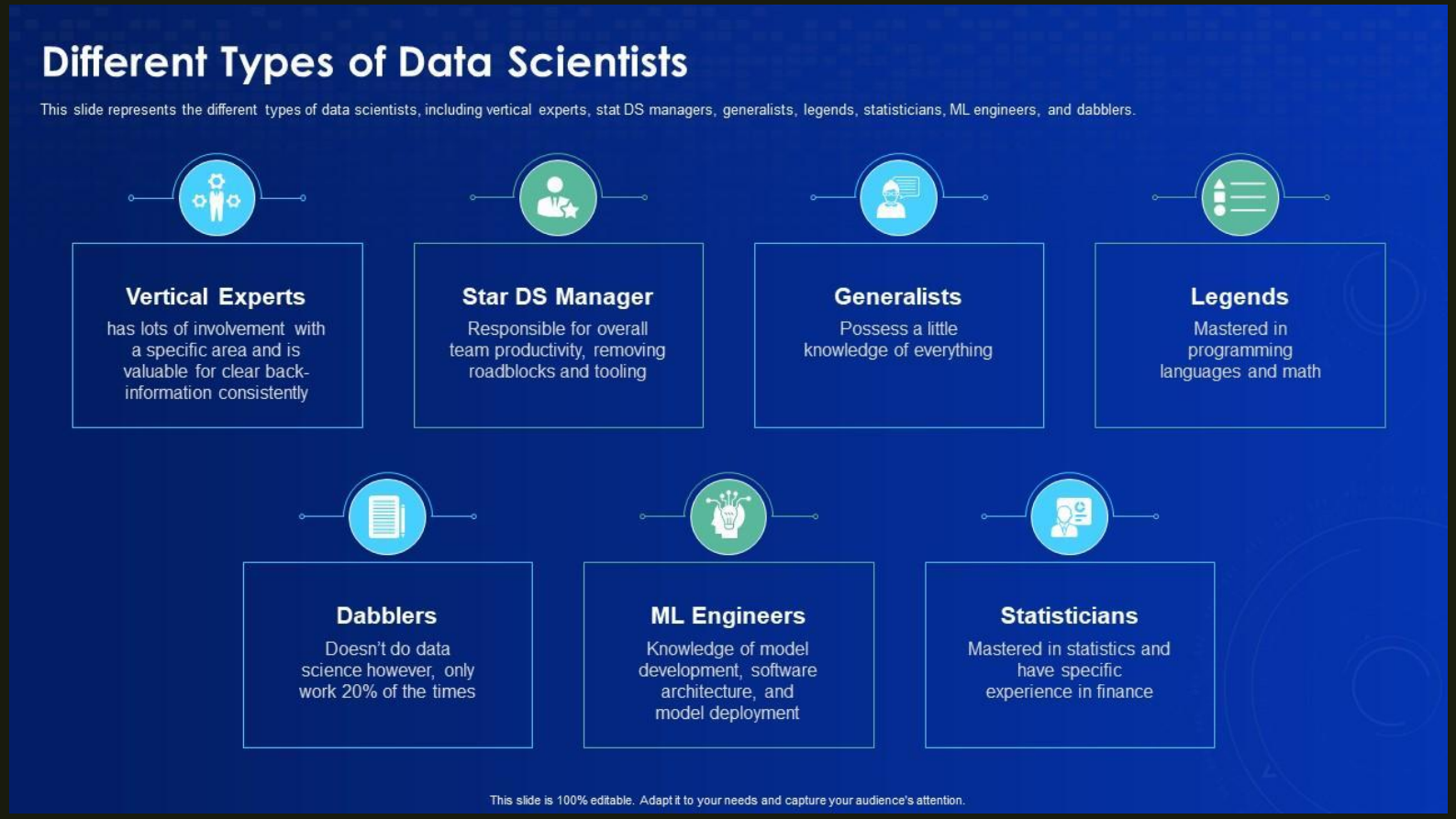
- While you are going to conduct a survey of this topic in your Assignment 1, I am going to make some general comments about this topic in this lecture, which I hope will make doing your assignment easier.

IT TAKES MANY DIFFERENT SKILLS TO BECOME A DATA SCIENTIST



Skills of data scientists

- Of course, you don’t need to have all skills to survive.
- There are different types of data science jobs on the market.
- I hope you can learn more about these “recipes” of data science jobs and think about what roles you want to play in your next job.
- Another source is the DS program at UTK: https://cecs.utk.edu/academics/undergraduate-programs/data-science-program/.



Types of data scientist jobs

- Another example is here.
- But of course, there can be many different classification schemes.
- I am NOT going to ask you to survey the different types of data scientist jobs this week. But I am hoping to see some discussion on this point in your Assignment 1 report.

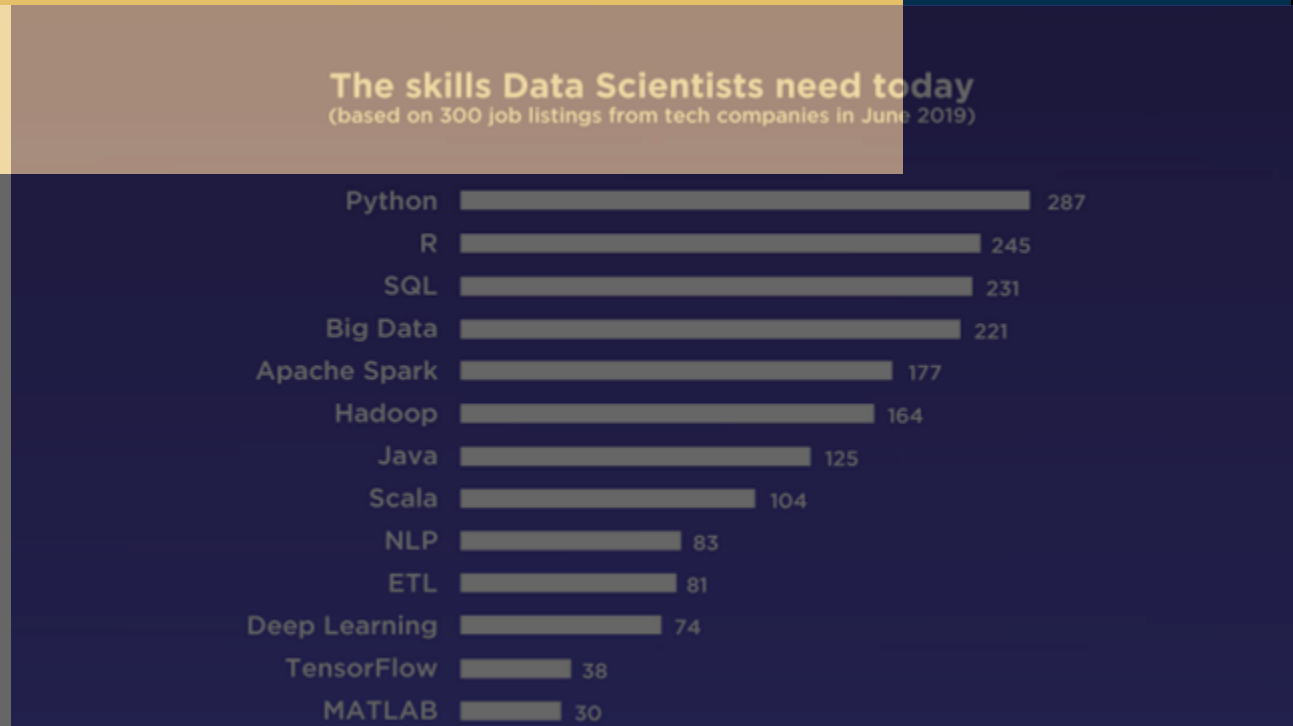
A personal reflection

- Based my skills:
  - *Statistics & Maths: working knowledge about most basic statistical methods*
  - *Programming language: I use R very frequently, and to a lesser extent, Python, but I won't see myself a programmer.*
  - *Visualization and communication: extensive experience of creating visualization*
  - *Domain knowledge: not much beyond my direct research field*
- While I am not going to ask you to do it, it would be helpful to reflect on what skills you have!

DATA SCIENCE / ANALYTICS TOOLS

Table 5. List of the top 50 most frequently used software from 9571 papers published in *PLoS ONE* in 2014.

Rank	Software	Mentions	Free or not	Rank	Software	Mentions	Free or not
1	SPSS	1868	No	26	ArcGIS	113	No
2	ImageJ	1065	Yes	27	PRIMER	110	Yes
3	SAS	611	No	28	MrBayes	106	Yes
4	Stata	578	No	29	BLASTP	106	Yes
5	MATLAB	452	No	30	BLASTX	106	Yes
6	BLAST	403	Yes	31	Bowtie	104	Yes
7	EXCEL	391	No	32	BEAST	100	Yes
8	MEGA	366	Yes	33	MetaMorph	99	No



All tools are useful but to different extents…

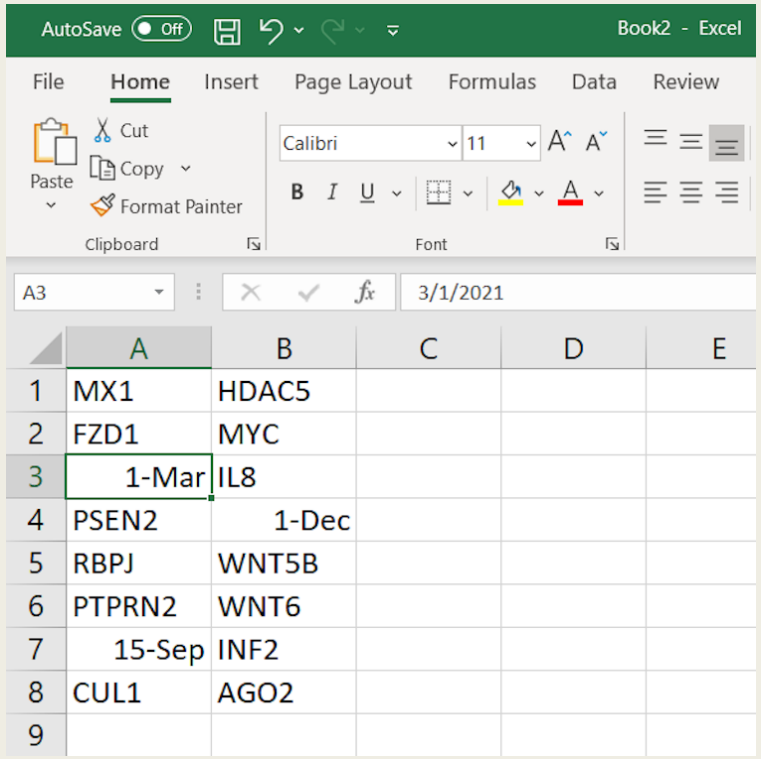
- There are many tools that we can use (open-source vs. commercial, programming language vs. packaged program).
- Pros and cons of programming languages:
  - *Extensibility: doing more tasks and being modified*
  - *Using just one language to do multiple tasks in the whole data lifecycle*
  - *But… steep learning curves.*
- A programming language is a new language to learn.

Programming languages

- Python and R are probably two most important programming languages for data science.
  - *Python is way more popular and more ”regular” programming language.*
  - *R is much more statistical and has better visualization packages.*
  - *In my opinion, R is much more friendly to beginners.*
    - Easy to install and configure
    - Much better documentation
    - R has only one version at the same time, unlike Python.

Microsoft Excel mistakes

- For example, it has been found that the autocorrect function of Microsoft Excel has caused a lot of mistakes in scientific research.
- https://theconversation.com/excel-autocorrect-errors-still-plague-genetic-research-raising-concerns-over-scientific-rigour-166554



Discussion question

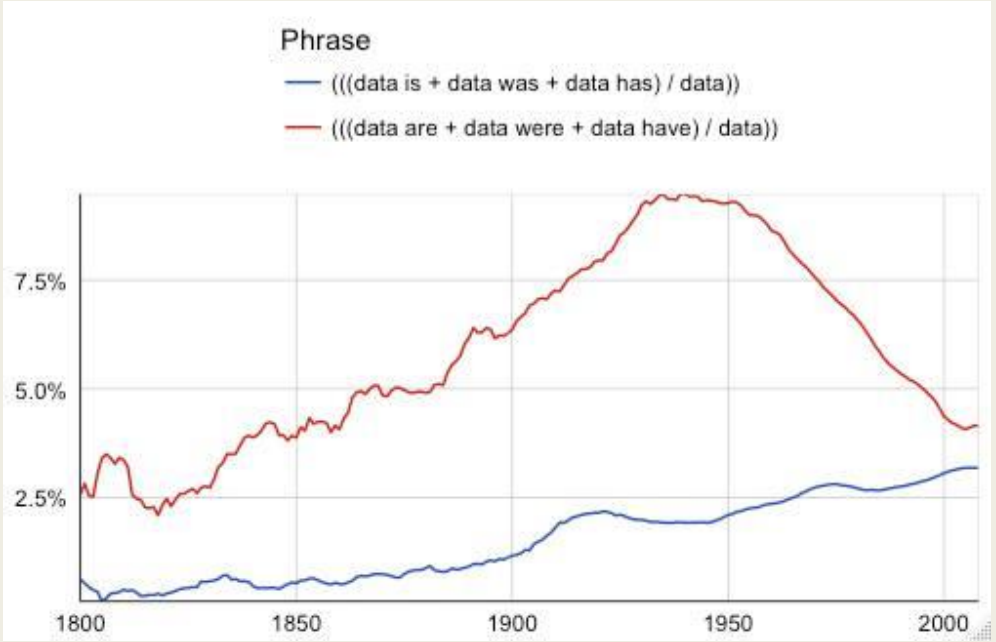
- We don’t have a discussion question per se, but the following task:
  - *I want you to install R and being able to load and run the example code file.*
  - *If you want, you can try Python.*
- You don’t need to post anything to the discussion board this week.
  - *But if you want, feel free to share any thought or questions about this process. We can keep talking about it in the next week!*

A demonstration of installing R

- R:
  - *Install R: https://www.r-project.org/*
  - *Install RStudio: https://posit.co/download/rstudio-desktop/*
    - RStudio is the most popular IDE for R language.
- Everything you can do in R is organized in packages (libraries), each of which is a list of functions that you can use.
  - *For example, the ggplot2 package contains ggplot() function (as well as many other functions) that we can use to draw graphs.*
  - *You need to install and load a package before using it, except for the base package.*

Is data singular or plural?

- Created using the ngramr package of R.
- We can find the package from the CRAN repository, the official R package repository.
  - https://cran.r-project.org/web/packages/ngramr/index.html
- We can also find the tutorial (vignette) of this package by searching it name on Google.
  - https://github.com/seancarmody/ngramr



A demonstration of installing Python

- R:
  - *Install R: https://www.r-project.org/*
  - *Install RStudio: https://posit.co/download/rstudio-desktop/*
    - RStudio is the most popular environment for R language.
- Python:
  - *You can install Python and use an IDE (integrated development environment), such as PyCharm.*
    - Configure Python: https://docs.python.org/3/using/mac.html
  - *Or, you can use an integrated environment like Anaconda. (My recommendation!)*
  - *You also need to use pip to install Python packages.*