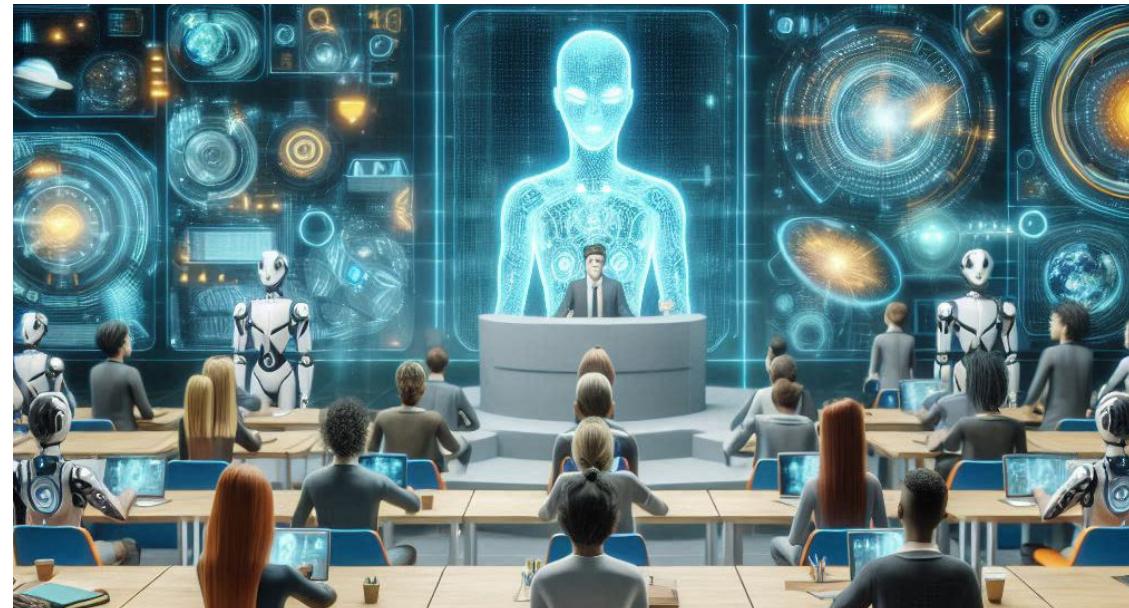




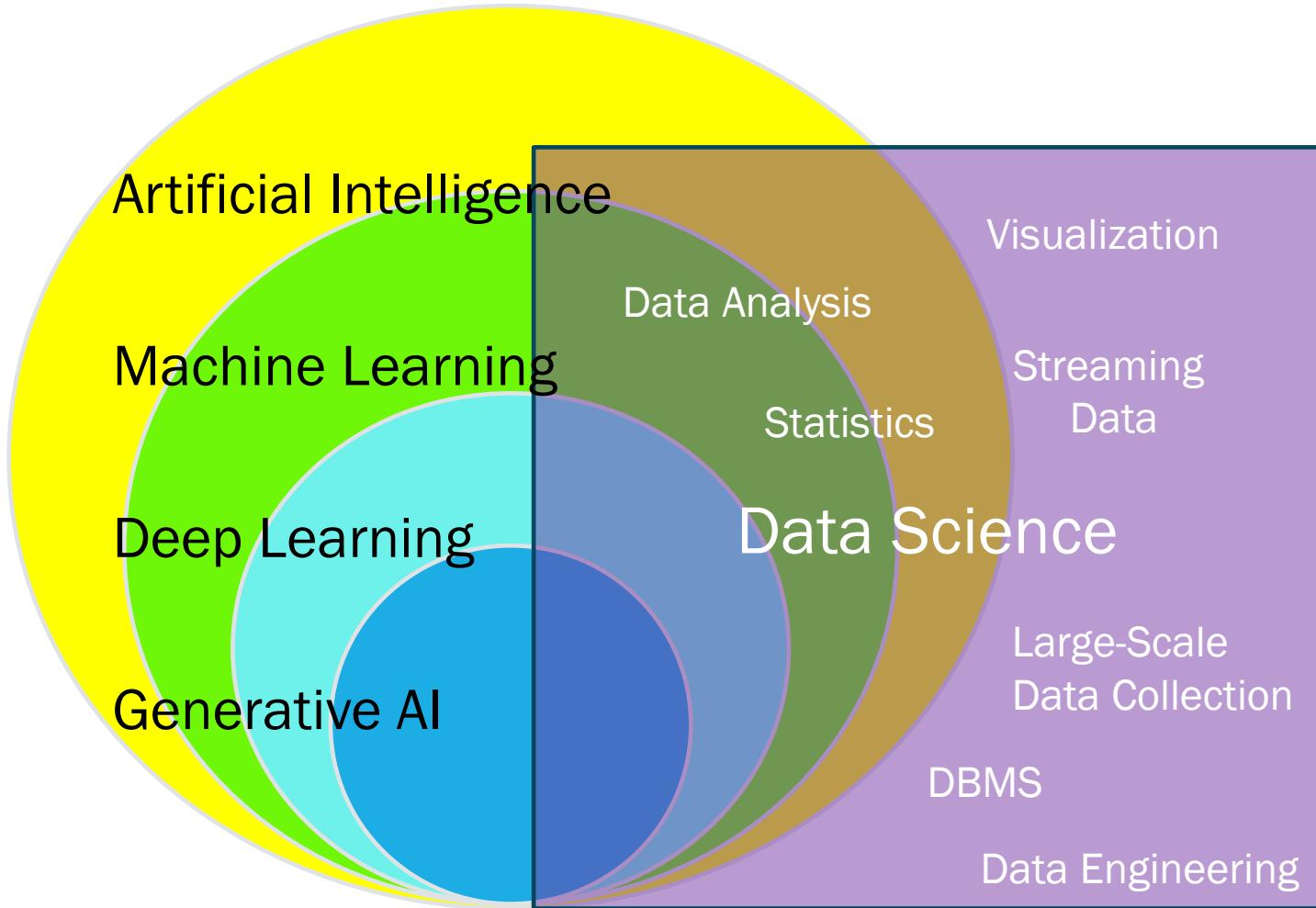
CS 180 INTRODUCTION TO DATA SCIENCE

INTRODUCTION, COURSE OVERVIEW, OBJECTIVES, HOW TO SUCCEED



Created by DALL-E Prompt: "Artificial Intelligence Classroom"

WHAT IS DATA SCIENCE?



Data Science:

- **Scope:** focuses on data as a whole, including data collection, processing, analysis, storage, and management.
- **End Goals:** Primarily concerned with extracting knowledge and **actionable insights** from data.
- **Techniques:** Uses data collection, data cleaning, data transformation, statistical analysis, data visualization, data management and data engineering tools.



WHY DATA SCIENCE?



Healthcare



Financial fraud detection



Internet search



Advertising



Logistics



E-commerce



Finance



Video game



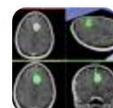
Airline route planning



Genomics



Retail



Medical image computing



Augmented reality



Data science in advertising



Drug development



Education



Manufacturing



Recommendation systems



Speech recognition



Transportation



Energy



Data cleansing



Sports

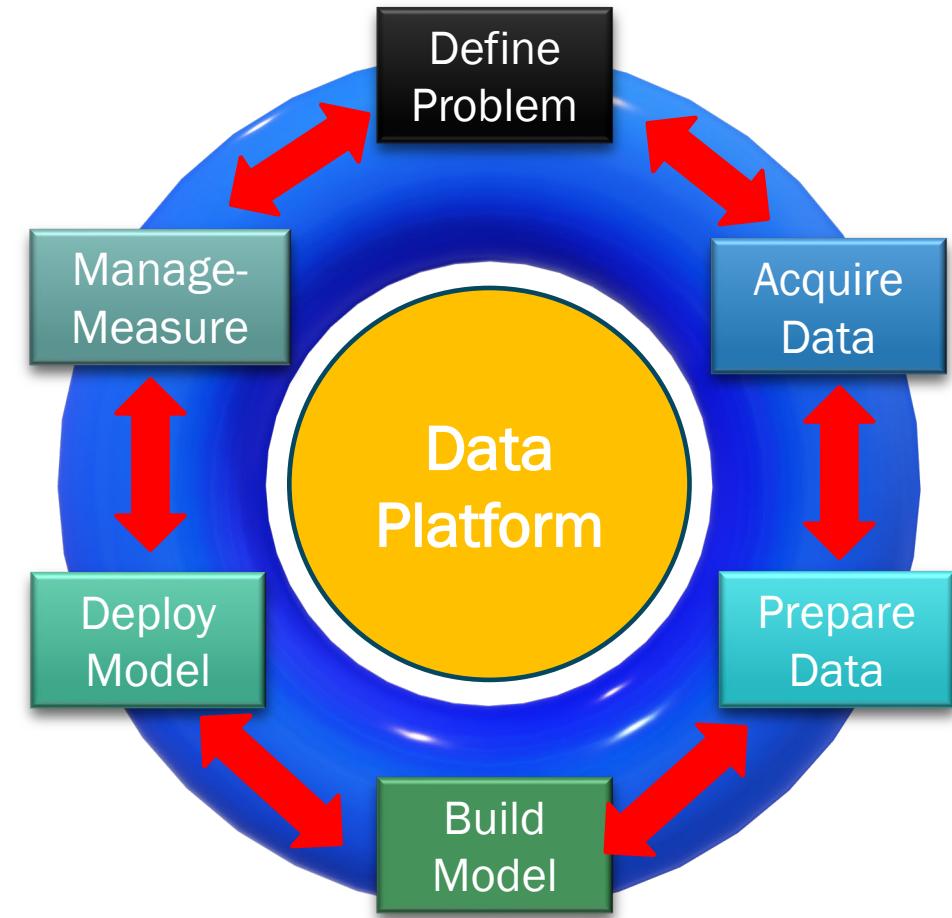


Virtual assistants

BYU

TECH STUFF WE'LL LEARN

- The Data Science Lifecycle process
- Using Python for data science (Numpy, Pandas, Matplotlib, Scikit-Learn, and more)
- How to prepare data for analysis
- How to explore data for insights
- Data Visualization (Python and Tableau)
- Data Management (SQL)
- Machine Learning basic algorithms
- Use of GenAI tools for language-based problems



COURSE GOALS

- Enhance your ability to quickly learn and master new concepts
- Utilize cutting-edge data science languages and tools to extract insights from data
- Effectively communicate insights through data visualization, dashboards, data storytelling
- Think critically about conclusions drawn from data analysis
- Recognize the potential to do good by applying Data Science to address critical global problems

WE ARE A LEARNING COMMUNITY

- Definition: A group of people who come together with the shared goal of learning, growing, and improving—often through collaboration, discussion, and mutual support. [ChatGPT]
- Get to know each other. How?
- At the beginning of every class, introduce yourself to those around you.
- Benefits: More effective learning, better performance, build a social network, better sense of belonging, and it's more fun!

AI “POLICY”

- Generative AI tools like ChatGPT, Gemini, Claude, Grok, etc., can be helpful to assistants in the learning process.
- If misused, GenAI can become a crutch, lead to sloppy thinking, and even a substitute for engagement with real people.
- We want you to have the best learning experience, resulting in foundational understanding and deep skill development.
- We recommend limiting GenAI use to answering questions about topics, looking up syntax for code, and understanding code examples, but not using GenAI to answer coding questions end-to-end.
- We trust you will follow this recommendation (aka policy). We will not police you, but if it is obvious, we will bring this to your attention.
- To help you better learn, we will have portions of the mid-term and final that you can't do with AI. As a result, it will benefit you to know the concepts to the point where you do not need AI.

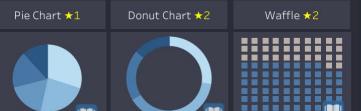
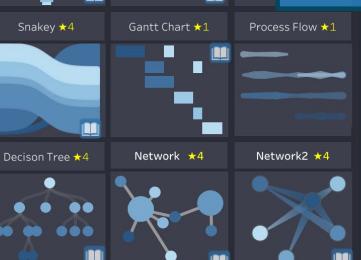
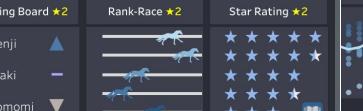
UPCOMING ASSIGNMENTS

1. Sign up for zyBook. How? Go to the first reading assignment in Canvas and click “Load Chapter 1 Reading...” This will initiate the process of buying the zyBook. You should not have to do this again.
2. Reading Assignment: Chapter 1. Points automatically accrue as you do the activities in the book.
3. Data Science Lab 1: Intro to Colab. Set up a Python Development Environment. Programming assignments will be turned in via Google Colab notebooks. However, if you are planning a career requiring programming, I recommend using a professional IDE, such as VS Code. If needed, we’ll go through the setup quickly next week.
4. Getting Set Up “Assignment”: This will help you (and me) hit the ground running!

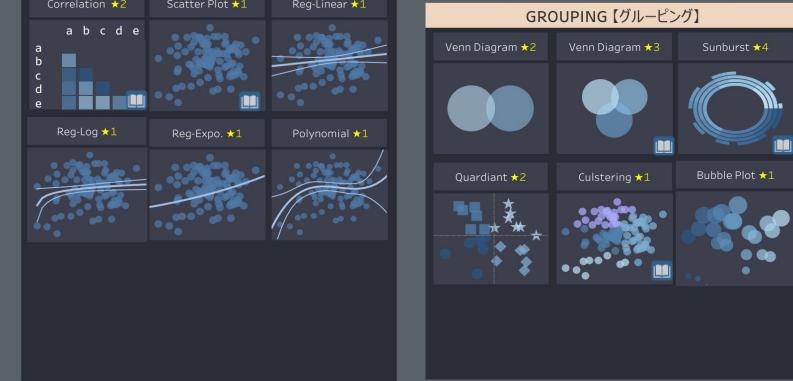




★1=SIMPLE ★2=MODERATE ★3=DIFFICULT ★4=ADVANCED ★5=MASTER

KPI VALUE - STATIC【KPI値 - 単一】**KPI VALUE - COMPARE【KPI値・比較】****KPI VALUE - TREND【KPI値・推移】****CONTRIBUTION - STATIC【貢献 - 単一】****CONTRIBUTION - COMPARE【貢献 - 比較】****CONTRIBUTION - TREND【貢献 - 推移】****FLOW【フロー】****RELATIONSHIP【関係性】****RANKING- COMPARE【ランクинг・比較】****DISTRIBUTION【分布】**

How-to-Build / Credits

GROUPING【グルーピング】**TEXT ANALYTICS【テキスト分析】****MAP【地図】**

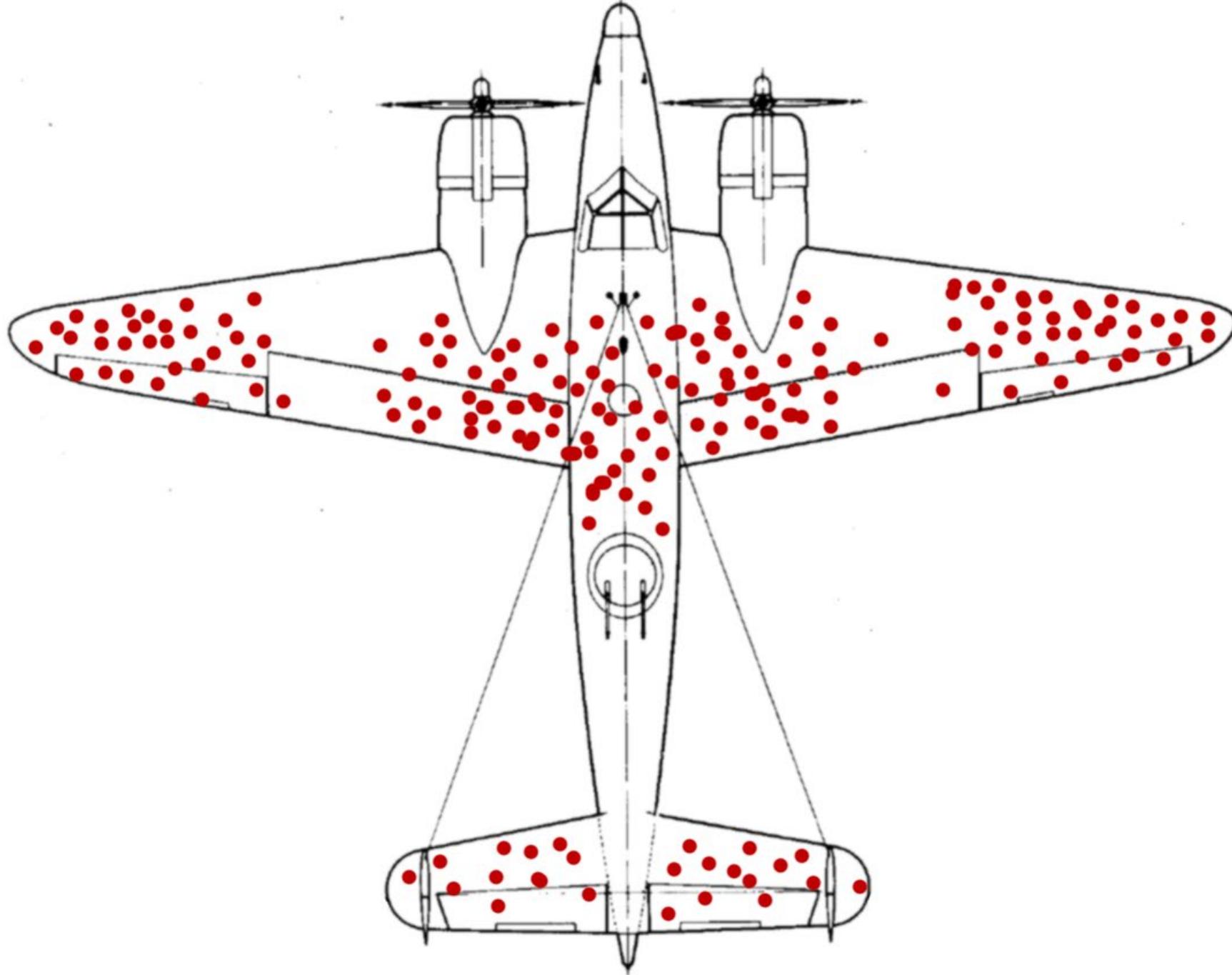
Let's Dive In!

TO BE CONTINUED つづく

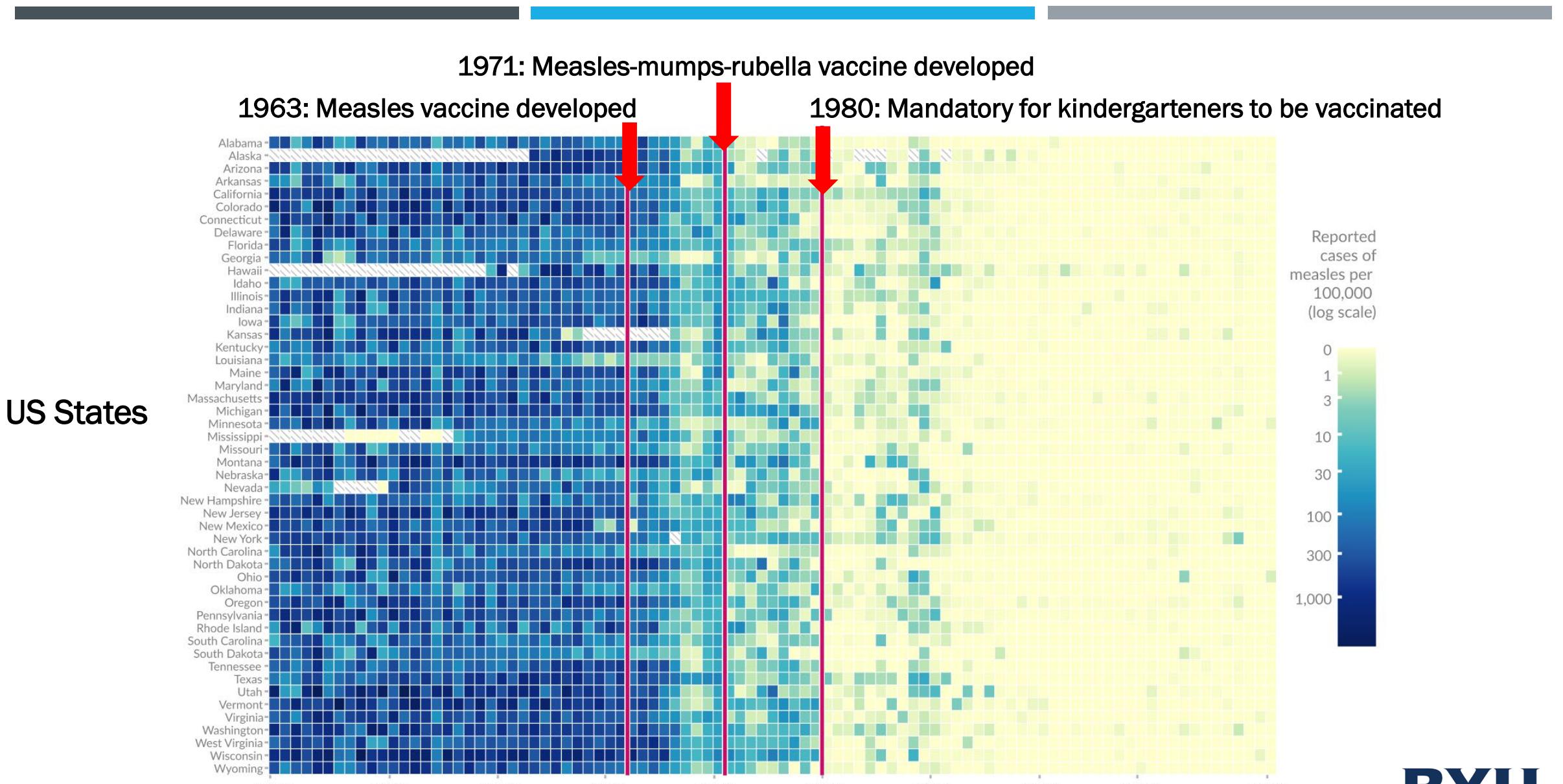


- What do you see?
- What should you do?
- Analyze or act?

BYU



BYU



Data source: Project Tycho (2018); Centers for Disease Control and Prevention (1959–2022)

BYU

GENDER BIAS AT BERKLEY (1973)

Are men applying to Berkeley more likely to get in than women?

	Men		Women	
	Applicants	Admitted	Applicants	Admitted
Total	8442	44%	4321	35%

GENDER BIAS AT BERKLEY (1973)

Are men applying to Berkeley more likely to get in than women?

	Men		Women	
	Applicants	Admitted	Applicants	Admitted
Total	8442	44%	4321	35%

- Departments have different acceptance rates,
- More women applied to departments with lower acceptance rates

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

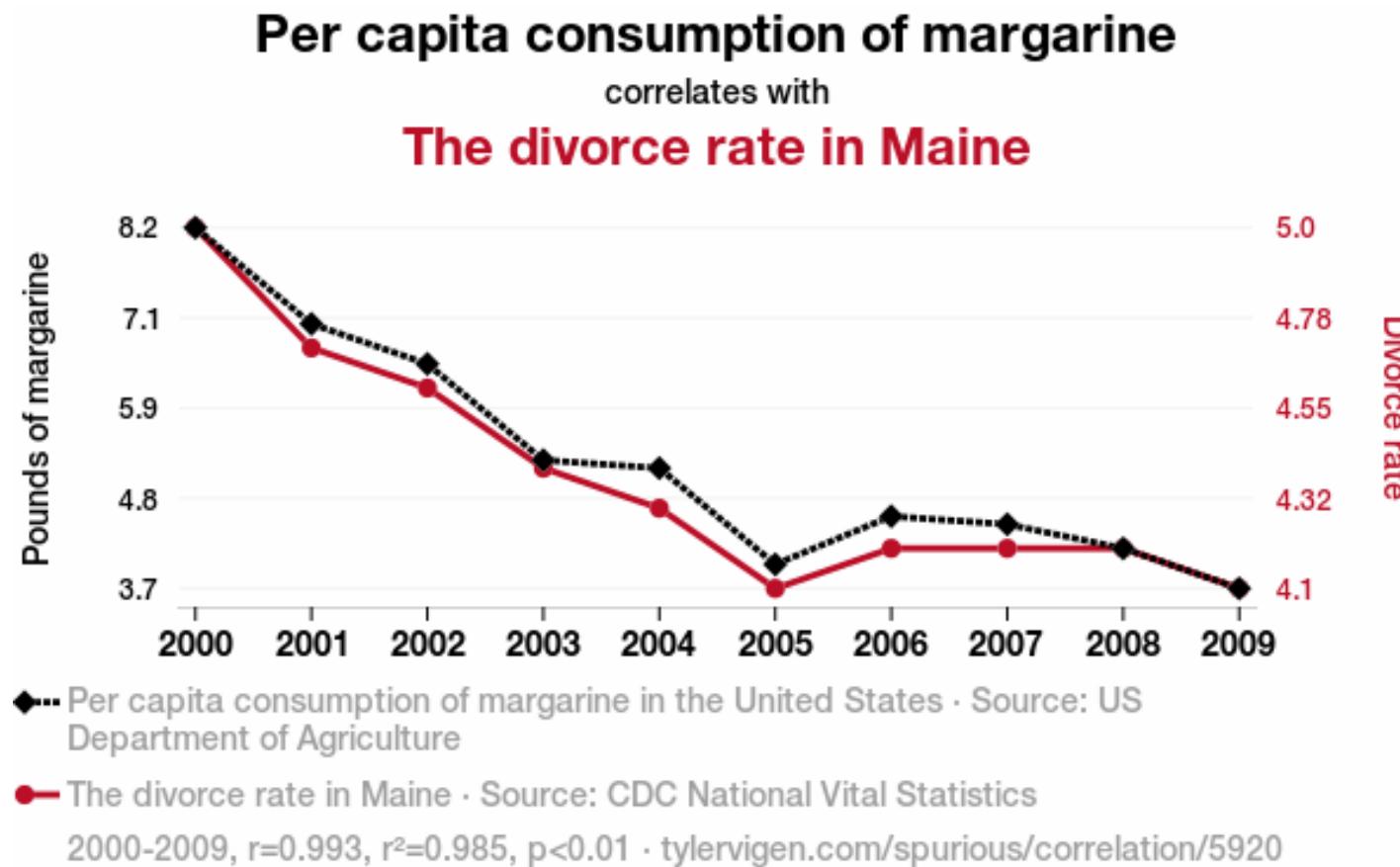
SPURIOUS CORRELATIONS



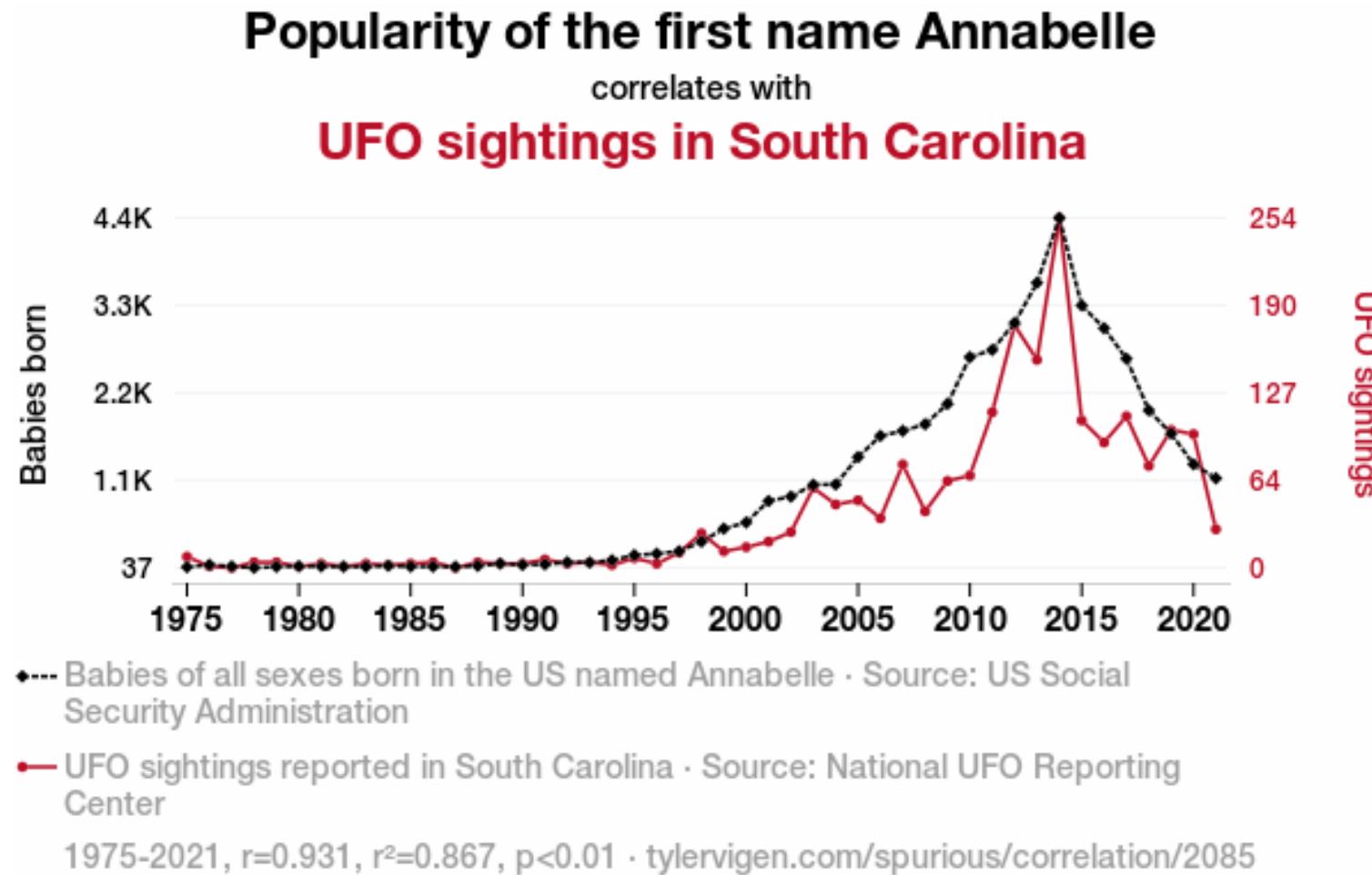
BYU



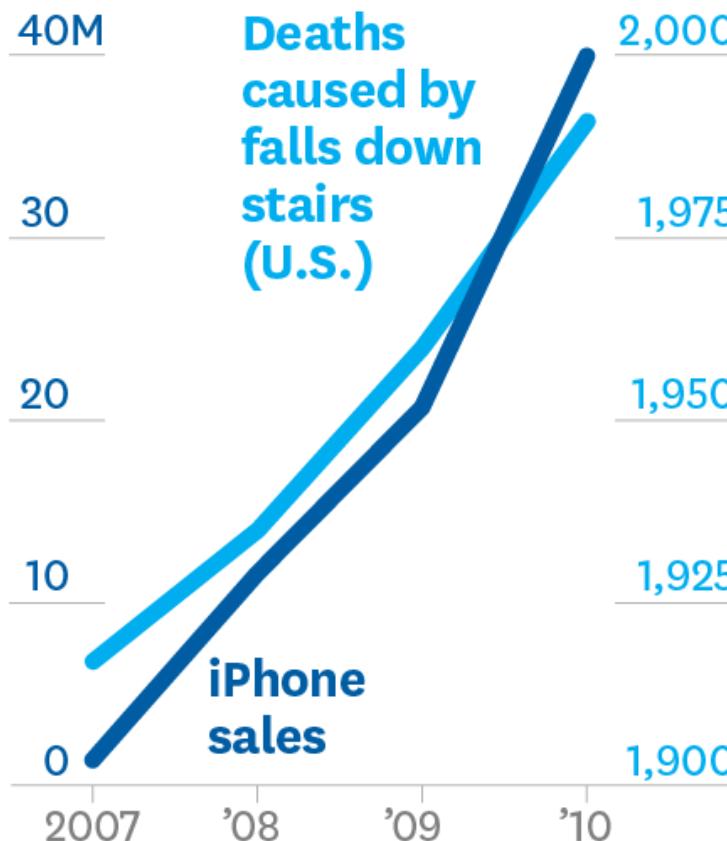
SPURIOUS CORRELATIONS



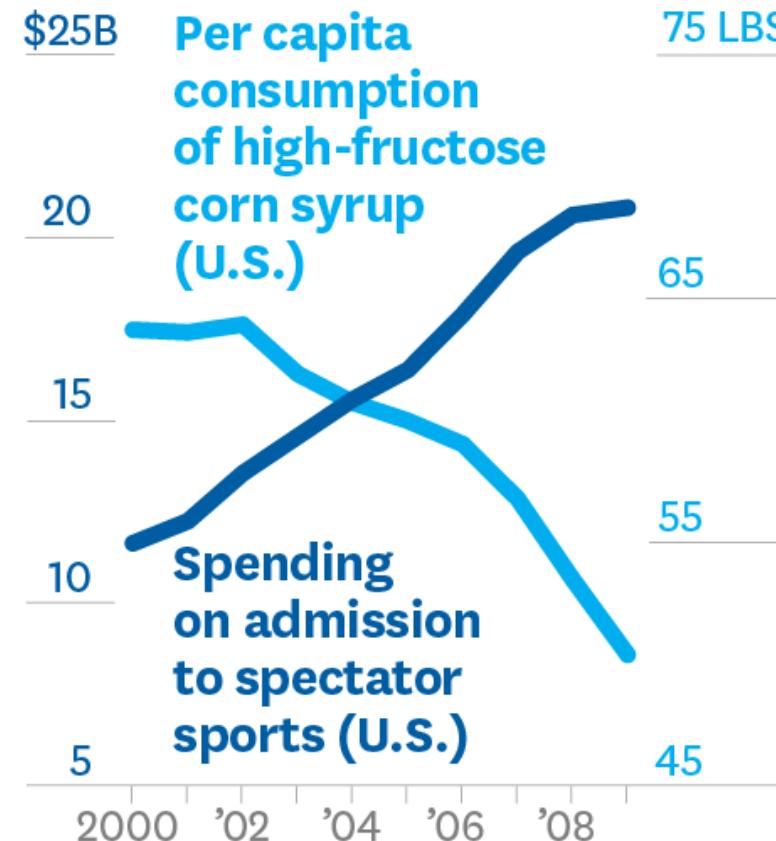
SPURIOUS CORRELATIONS



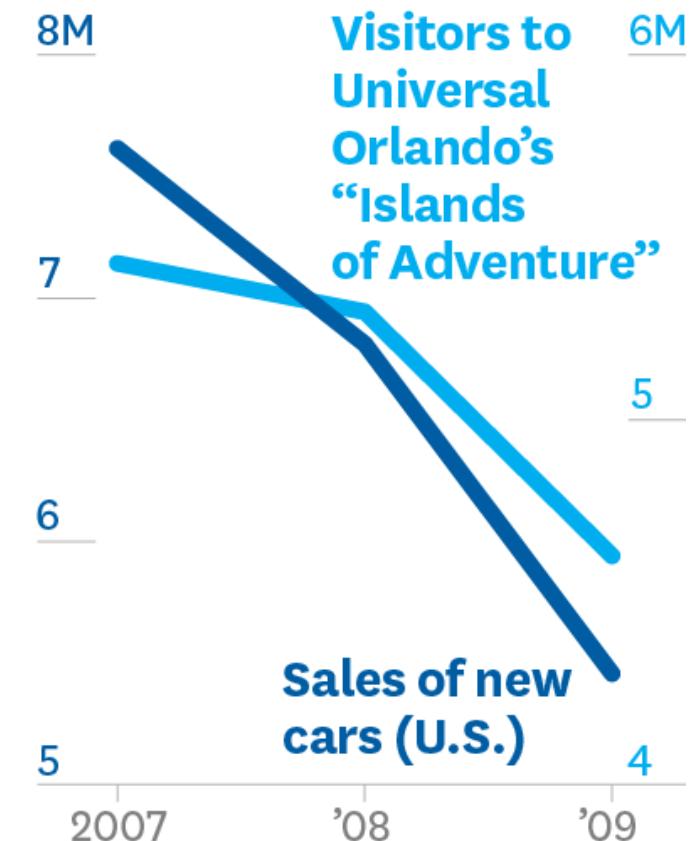
MORE IPHONES MEANS MORE PEOPLE DIE FROM FALLING DOWN STAIRS



LET'S CHEER ON THE TEAM, AND WE'LL LOSE WEIGHT



TO INCREASE AUTO SALES, MARKET TRIPS TO UNIVERSAL ORLANDO



SOURCE TYLERVIGEN.COM
FROM "BEWARE SPURIOUS CORRELATIONS," JUNE 2015

© HBR.ORG

ZYBOOK DATA SCIENCE LIFECYCLE FOR DATA ANALYSIS

Table 1.4.1: Data science lifecycle.

Step	Description
Step 1: Gathering data	Identify available and relevant data; gather new data if needed.
Step 2: Cleaning data	Reformat datasets, create new features, and address missing values.
Step 3: Exploring data	Create data visualizations and calculate summary statistics to explore potential relationships in the dataset.
Step 4: Modeling data	Use modeling skills and content knowledge to fit and evaluate models, measure relationships, and make predictions.
Step 5: Interpreting data	Describe and interpret conclusions from data through written reports and presentations.

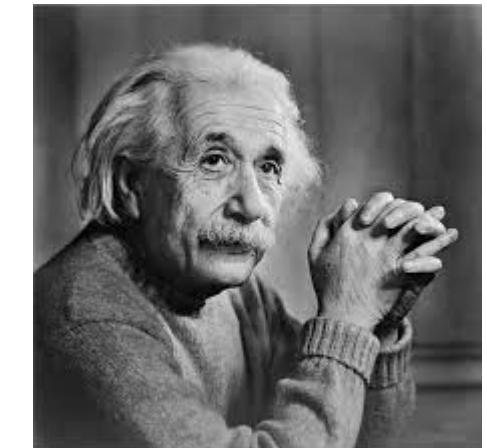
DEFINE THE PROBLEM

- What is the core problem?
- What processes, systems, orgs are affected?
- If solved, what is business value?
- How can problem be scoped?
- How is value measured?
- Characterize problem domain
- Is this a data-driven problem?
- What data is needed? (prelim)

“

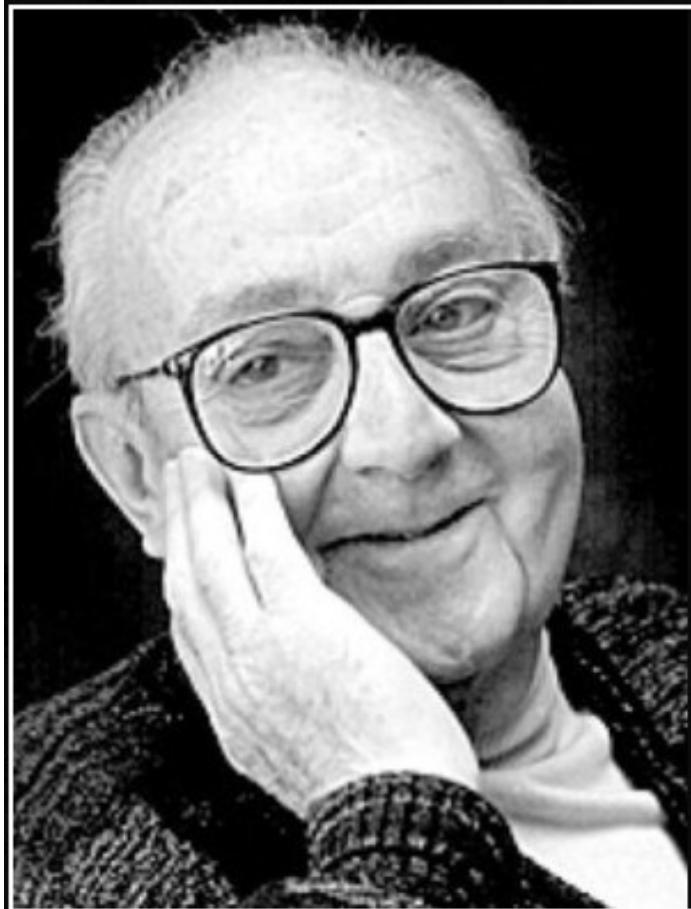
IF I HAD AN HOUR TO SOLVE A
PROBLEM I'D SPEND 55 MINUTES
THINKING ABOUT THE PROBLEM AND
5 MINUTES THINKING ABOUT
SOLUTIONS.

”



Albert Einstein

4. BUILD THE MODEL



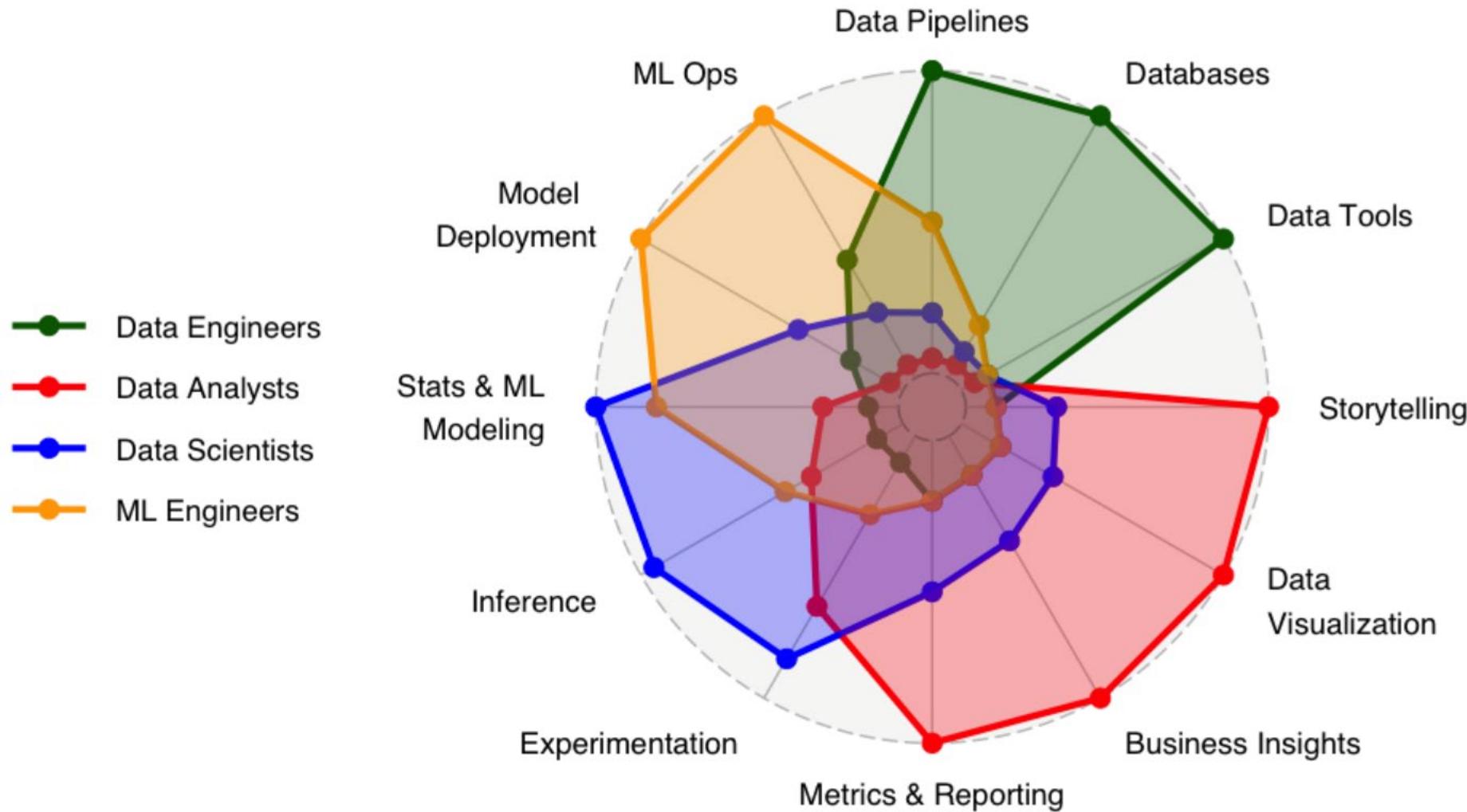
All models are approximations.
Essentially, all models are wrong, but
some are useful. However, the
approximate nature of the model
must always be borne in mind.

— *George E. P. Box* —

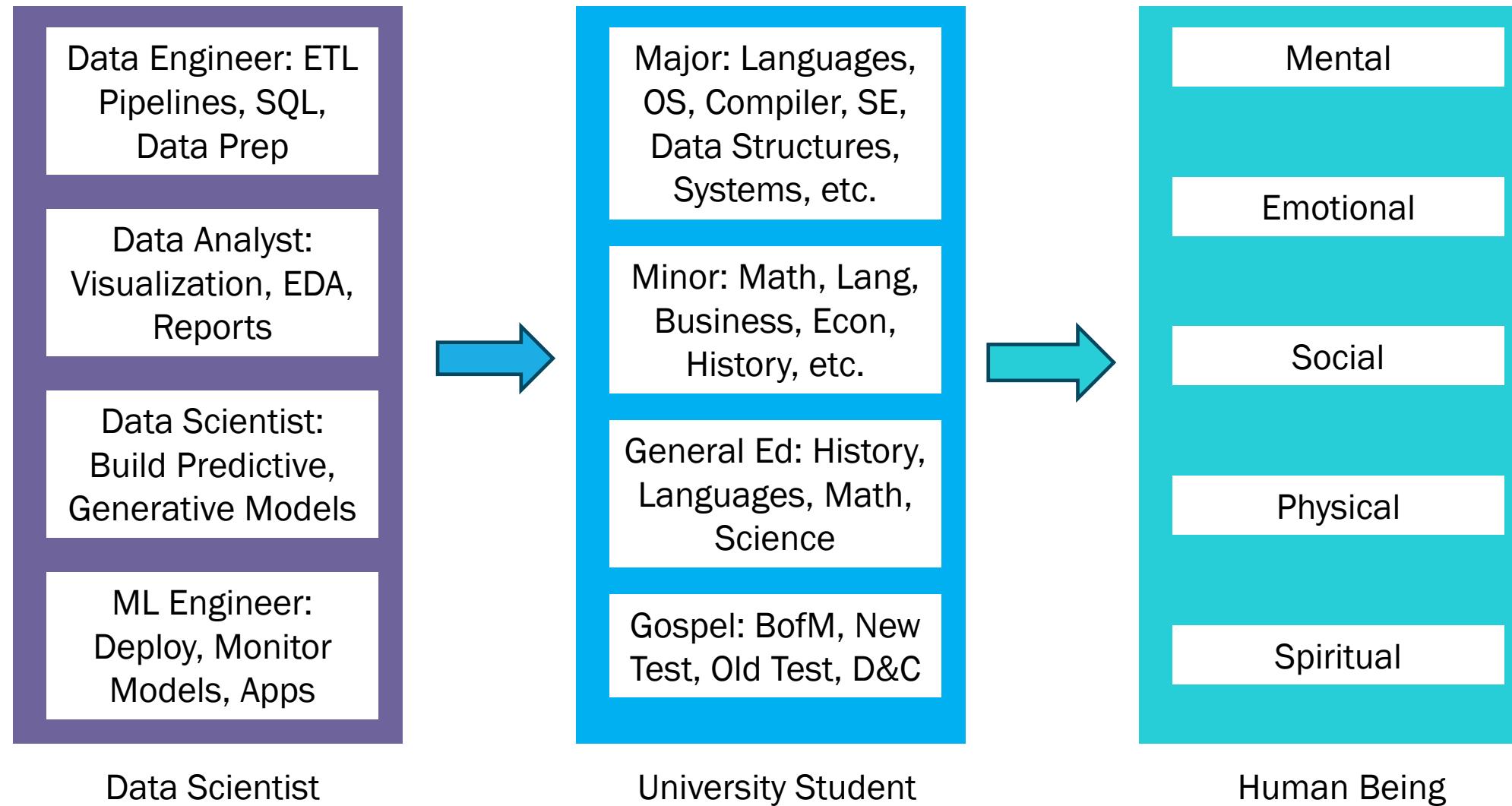
AZ QUOTES

BYU

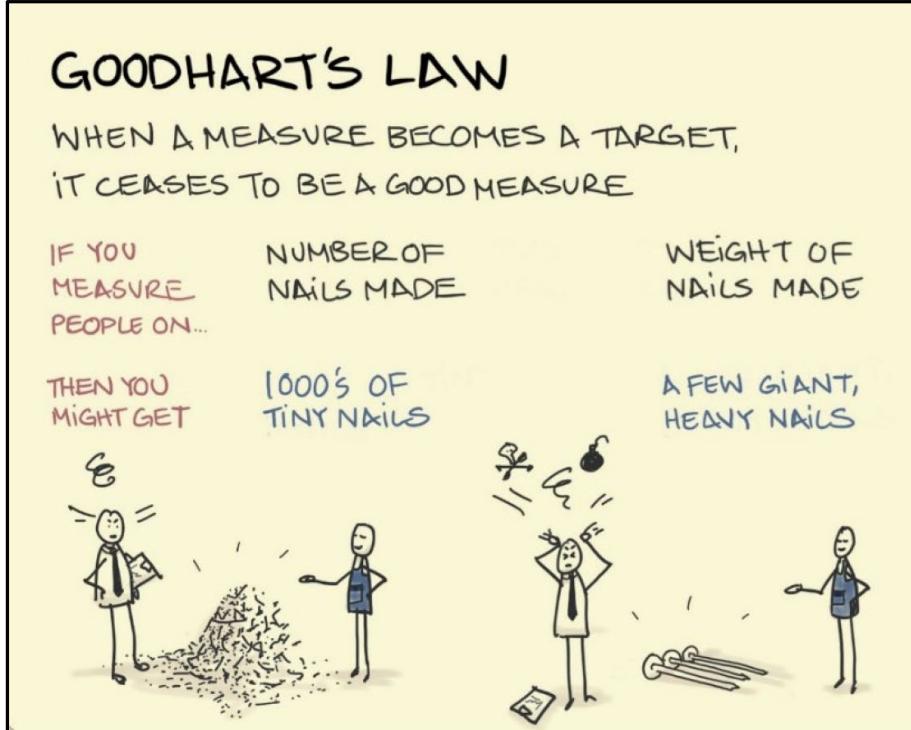
SPIDER CHART OF RELATIVE SKILLS FOR KEY DATA ROLES



THE GOAL: FULL-STACK DATA SCIENTIST AND BEYOND



GOOD TO KNOW...



Students understand this law very well. It is easy to get caught up focusing on getting the "A" instead of mastering the knowledge or skill.

**Don't let school...
get in the way of your
education ☺**