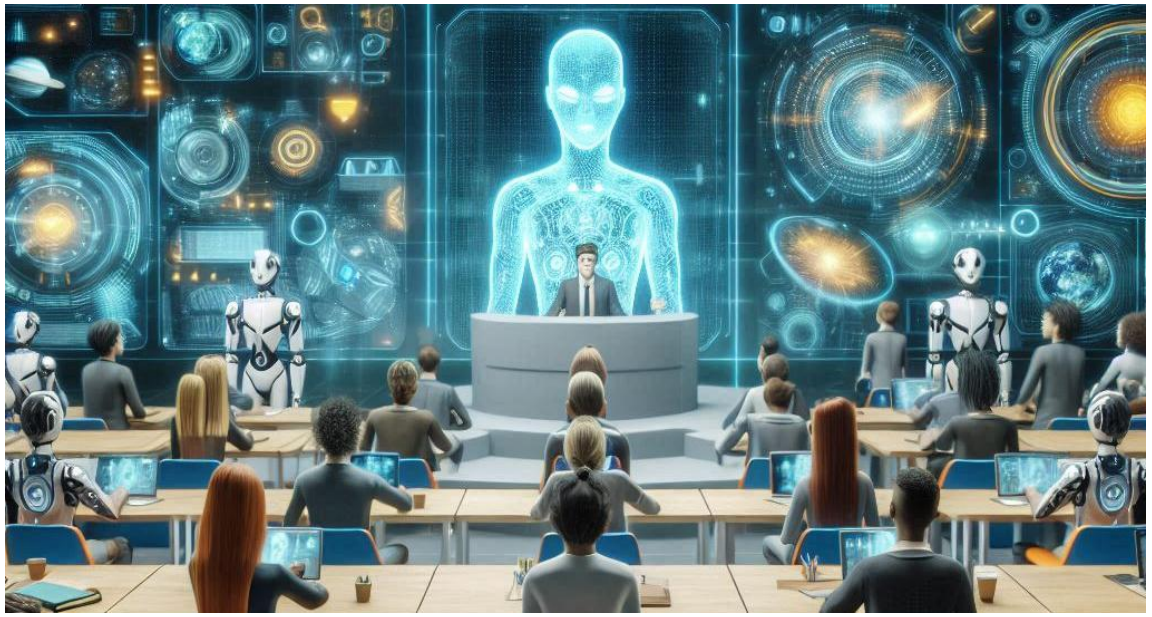
CS 180 INTRODUCTION TO DATA SCIENCE

INTRODUCTION, COURSE OVERVIEW, OBJECTIVES, HOW TO SUCCEED



Created by DALL-E Prompt: “Artificial Intelligence Classroom”

ANNOUNCEMENTS

Dr. Jake Rhodes

rhodes@stat.byu.edu

WVB 2177

Office Hours: TBD

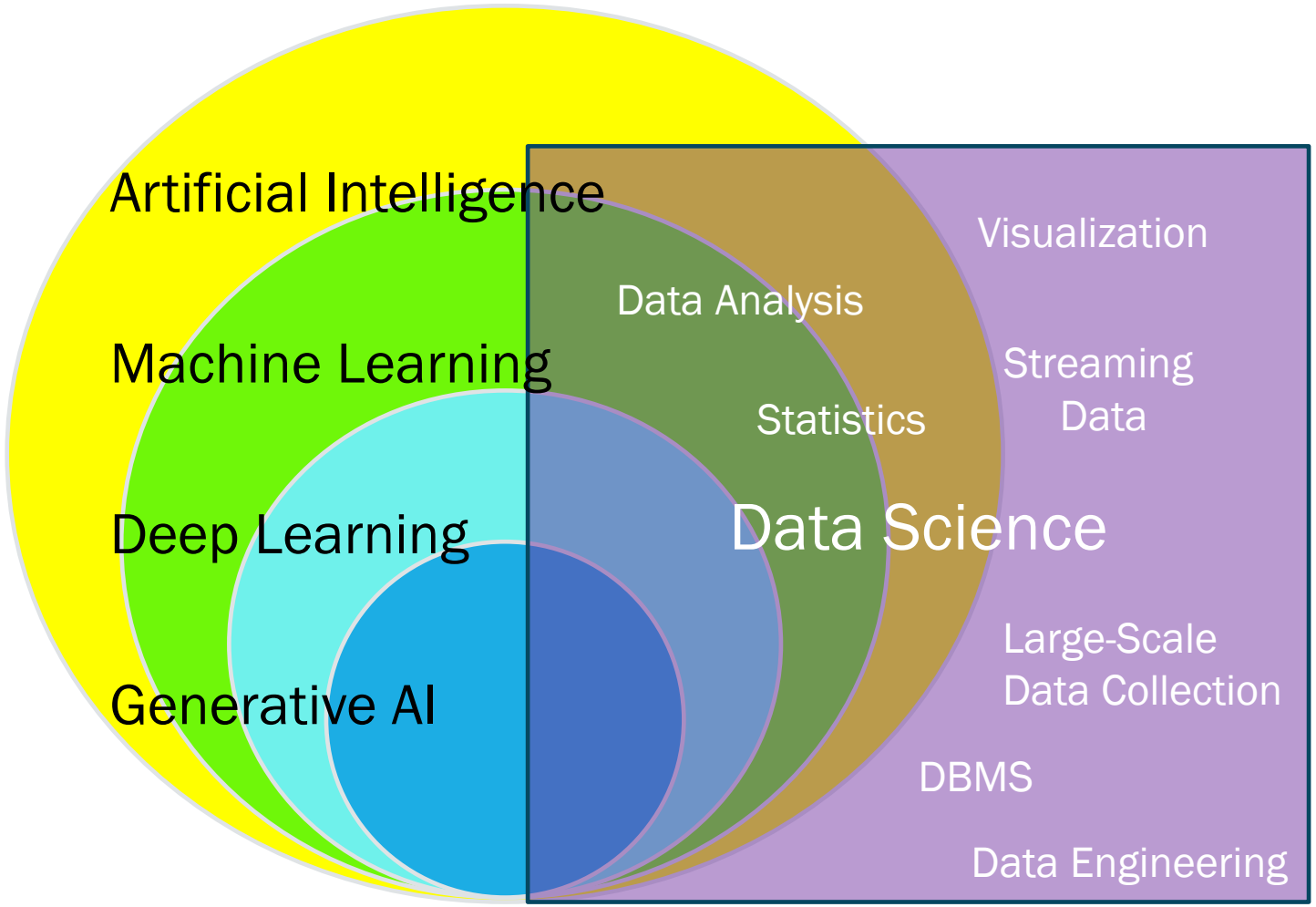
Teaching Assistants (WVB 1151):

1. Toby Allen
2. Corbin Christiansen
3. Michael Jensen
4. Spencer Marshall
5. Ian Villanueva

Office Hours: Posted on Syllabus

BYU

WHAT IS DATA SCIENCE?



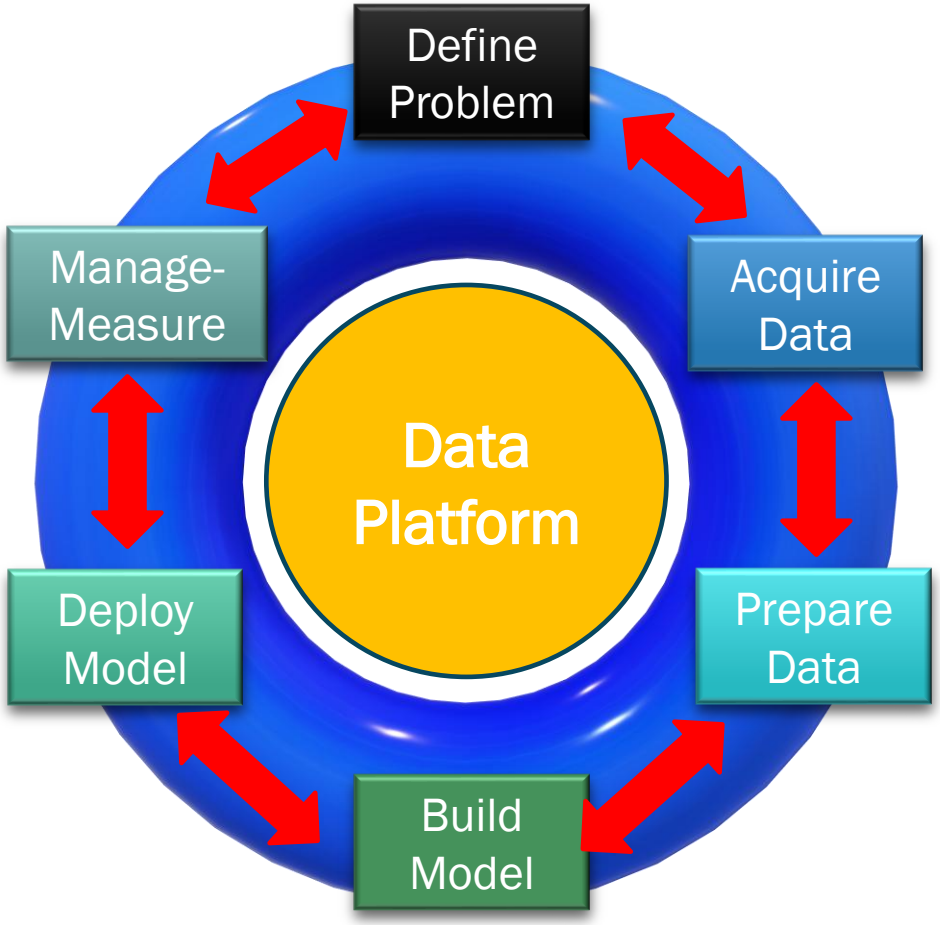
Data Science:

- **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.

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



BYU

WE'LL ALSO TACKLE *DATA LITERACY*

- Critical thinking about potential conclusions of an analysis
- Potential pitfalls of overreliance on unreliable data
- When is someone lying to you with statistics?

BYU

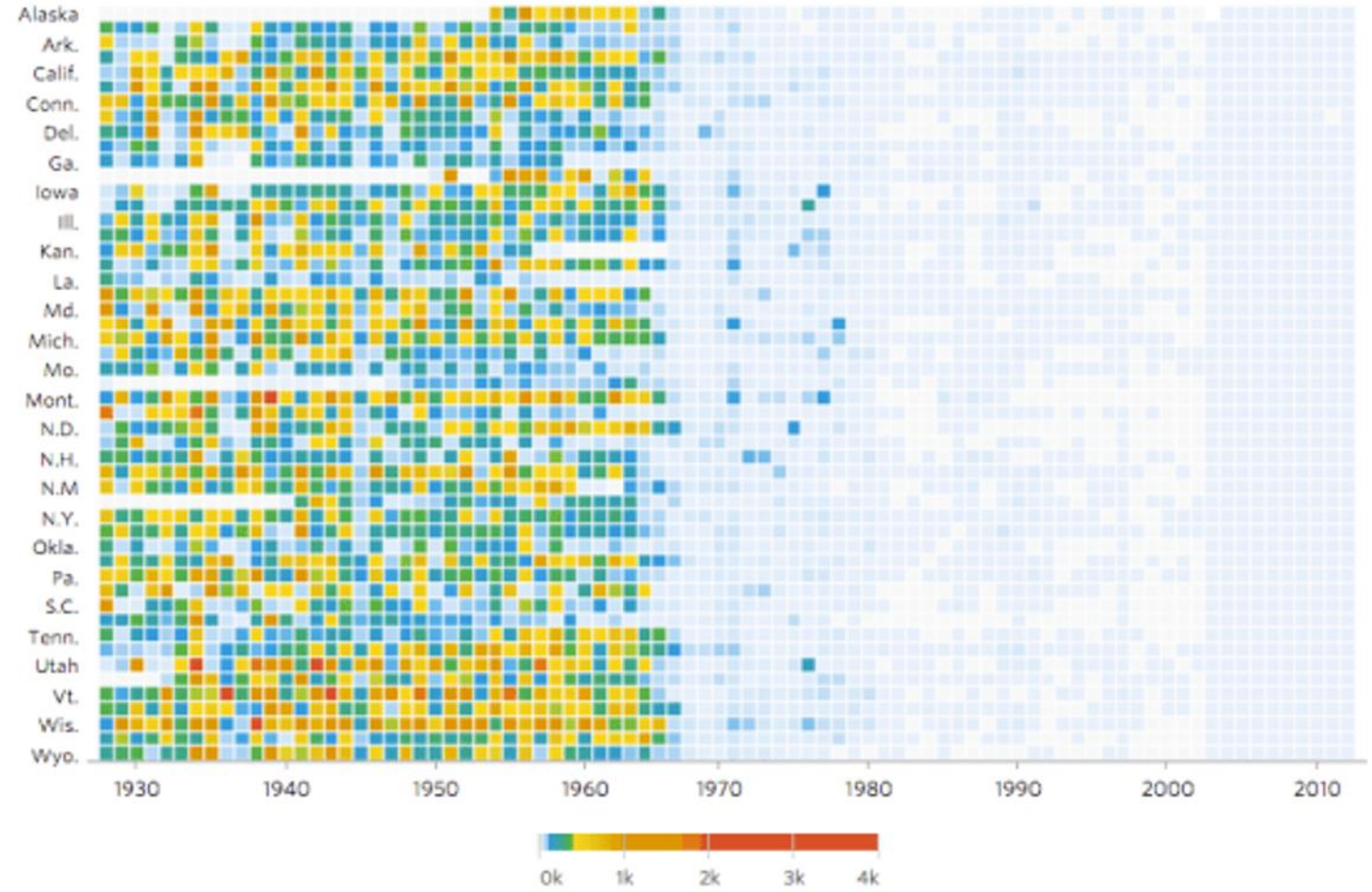
COURSE GOALS

- Begin the journey to become a “Full-Stack” Data Scientist by:
  - Using advanced tools to extract insights from data
  - Solving novel problems with core data science principles
  - Communicating insights effectively through visualization
  - Critically evaluating data-driven conclusions

BYU

+tableau
102 VISUZLIATIONS
VIZ LIBRARY
THE COLLECTION OF TABLEAU VISUALIZATIONS & ANALYSIS
V.2018.8.3
Kritidikoon Woraitthinan
KPI
STATIC (3)
COMPARISON (8)
TREND (14)
CONTRIBUTION
STATIC(3)
COMPARE(4)
TREND(3)
RANKING
STATIC (3)
COMPARISON (3)
TREND (2)
AMOUNT (5)
FLOW (9)
TIME (7)
RELATIONSHIP (10)
DISTRIBUTION (6)
GROUPING (5)
TEXT (2)
MAP (8)
EXTRA (2)
★1=SIMPLE ★2=MODERATE ★3=DIFFICULT ★4=ADVANCED ★5=MASTER
How-to-Build / Credits
KPI VALUE - STATIC【KPI値 - 単一】
Gauge1 ★2
Gauge2 ★3
Number ★1
95%
475
95%
+75
KPI VALUE - COMPARE【KPI値-比較】
Radial Gauge ★3
Circular Radial ★4
Radial Bar ★5
Likert Scale ★3
Scale ★2
Radar Chart ★3
Double Bar ★1
Bullet ★1
KPI VALUE - TREND【KPI値-推移】
Bar Chart ★1
Bar & Line ★2
Bar & Shape ★2
Bold Line ★1
Gap Trend ★3
Waterfall ★3
Lollipop ★2
Spiral ★3
Line Chart ★1
Line Cumu ★1
Align Start ★3
Sparklines ★1
Multiple Chart ★1
Forecast ★1
CONTRIBUTION - STATIC【貢献 - 単一】
Pie Chart ★1
Donut Chart ★2
Waffle ★2
CONTRIBUTION - COMPARE【貢献 - 比較】
Tree ★1
Bubble ★1
Growth Ring ★2
Contri. Bar ★1
Butterfly ★2
CONTRIBUTION - TREND【貢献 - 推移】
Stacked Bar ★1
Area ★1
FLOW【フロー】
Funnal1 ★2
Funnal2 ★3
Snakey ★4
Gantt Chart ★1
Process Flow ★1
Decison Tree ★4
Network ★4
Network2 ★4
RELATIONSHIP【関係性】
Heat Map1 ★1
Heat Map2 ★1
Pereto ★2
RANKING- COMPARE【ランキング-比較】
Ranking Board ★2
1 Kenji
2 Maki
3 Tomomi
Rank-Race ★2
Star Rating ★2
Arrow ★2
Slopegraph ★2
Slopegraph 2 ★3
RANKING- TREND【ランキング-推移】
Bump Chart1 ★2
Bump Chart2 ★2
TIME【時間】
Timeline ★3
Multi-timeline ★3
Calendar ★2
Clock ★2
Jump Plot ★4
Jump Plot2 ★4
Circular Dot ★3
DISTRIBUTION【分布】
Histrogram ★1
Box Plot ★1
Scatter Plot ★2
Scatter Plot2 ★2
Jitters ★1
Barcode Plot ★1
Correlation ★2
Scatter Plot ★1
Reg-Linear ★1
Reg-Log ★1
Reg-Expo. ★1
Polynomial ★1
GROUPING【グルーピング】
Venn Diagram ★2
Venn Diagram ★3
Sunburst ★4
Quardiant ★2
Culstering ★1
Bubble Plot ★1
TEXT ANALYTICS【テキスト分析】
Word Cloud ★2
Word Count ★2
Sentiment Analysis ★4
MAP【地図】
Map - Pie ★2
Map-Custom ★3
Path - Curve ★4
Path - Curve Multi-line ★4
Position ★3
EXTRA【特別】
3D Model ★5
Geometric Art - Drawing ★5
Shape- Heart ★3
Let’s Dive In!

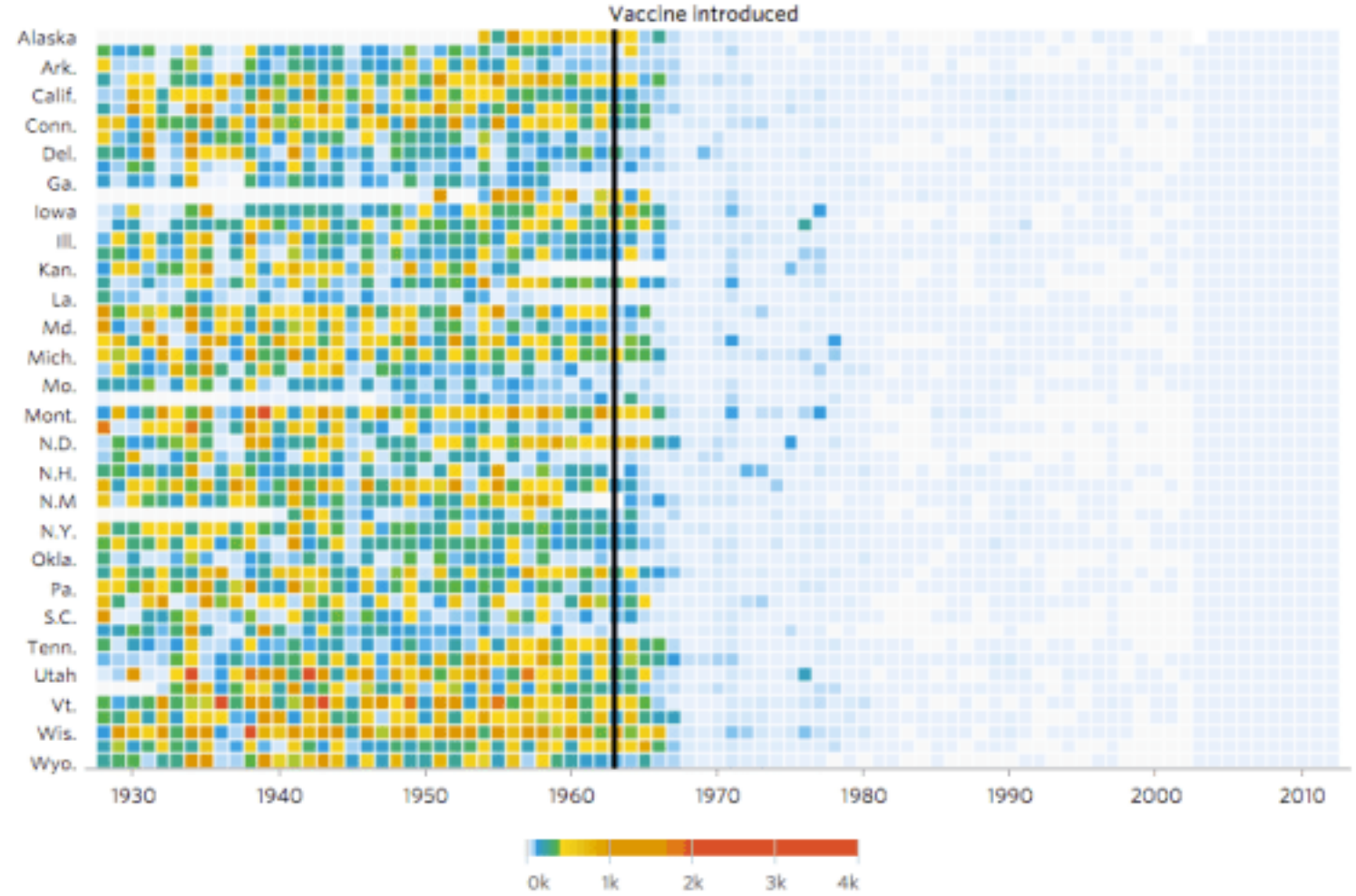
Measles



Note: CDC data from 2003-2012 comes from its Summary of Notifiable Diseases, which publishes yearly rather than weekly and counts confirmed cases as opposed to provisional ones.

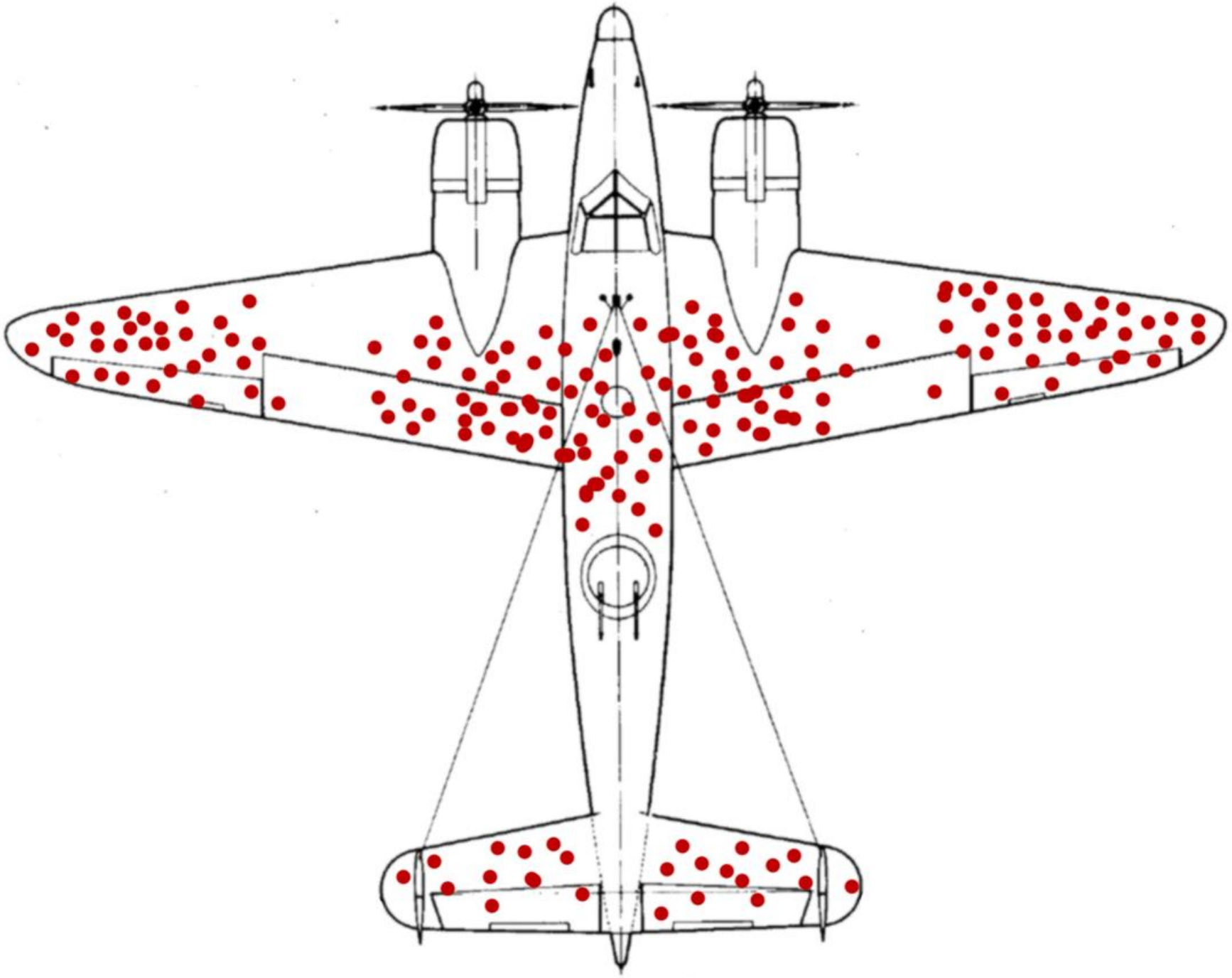
BYU

Measles

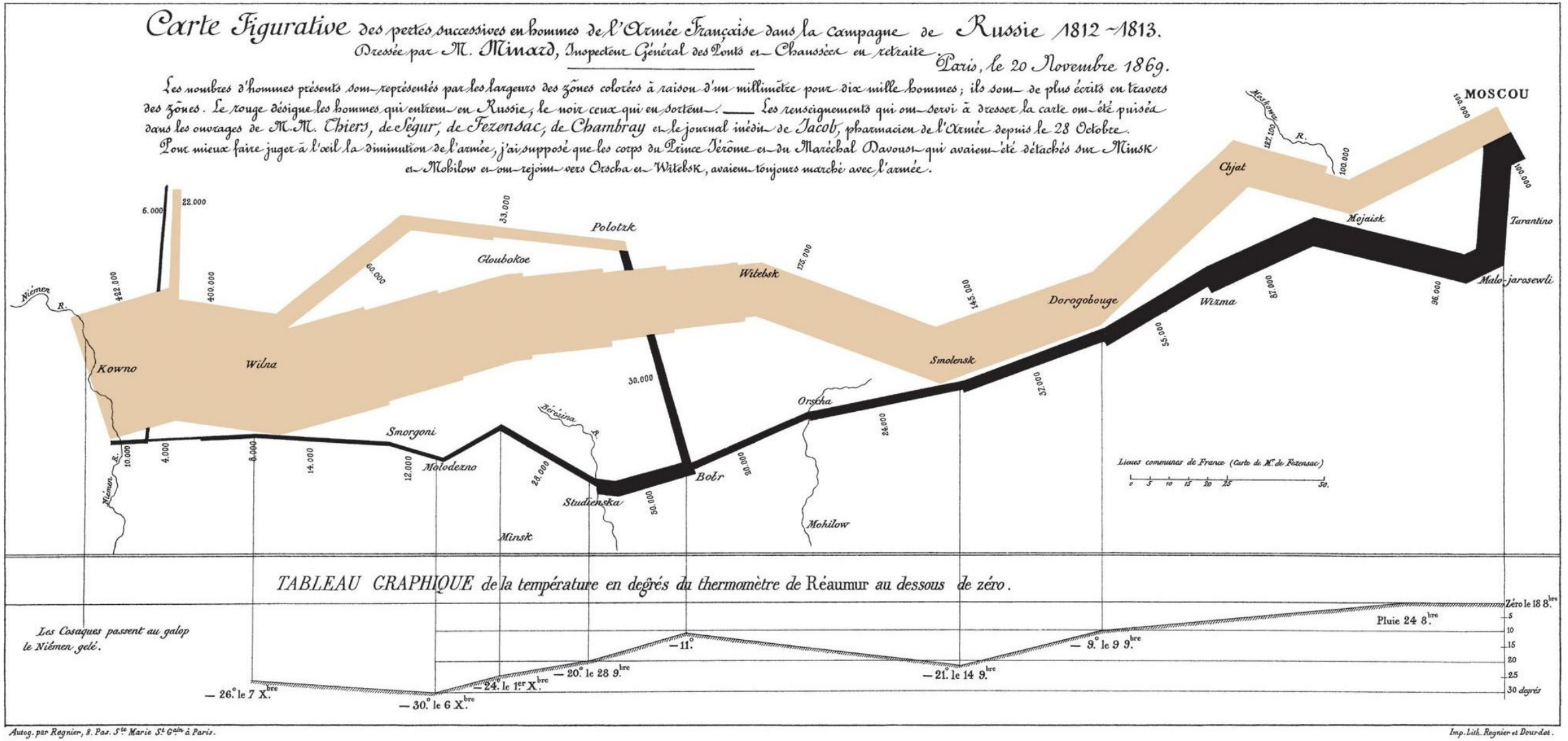


Note: CDC data from 2003-2012 comes from its Summary of Notifiable Diseases, which publishes yearly rather than weekly and counts confirmed cases as opposed to provisional ones.

BYU



NAPOLEON'S DISASTROUS INVASION OF RUSSIA IN 1812



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%

BYU

GENDER BIAS AT BERKLEY (1973)

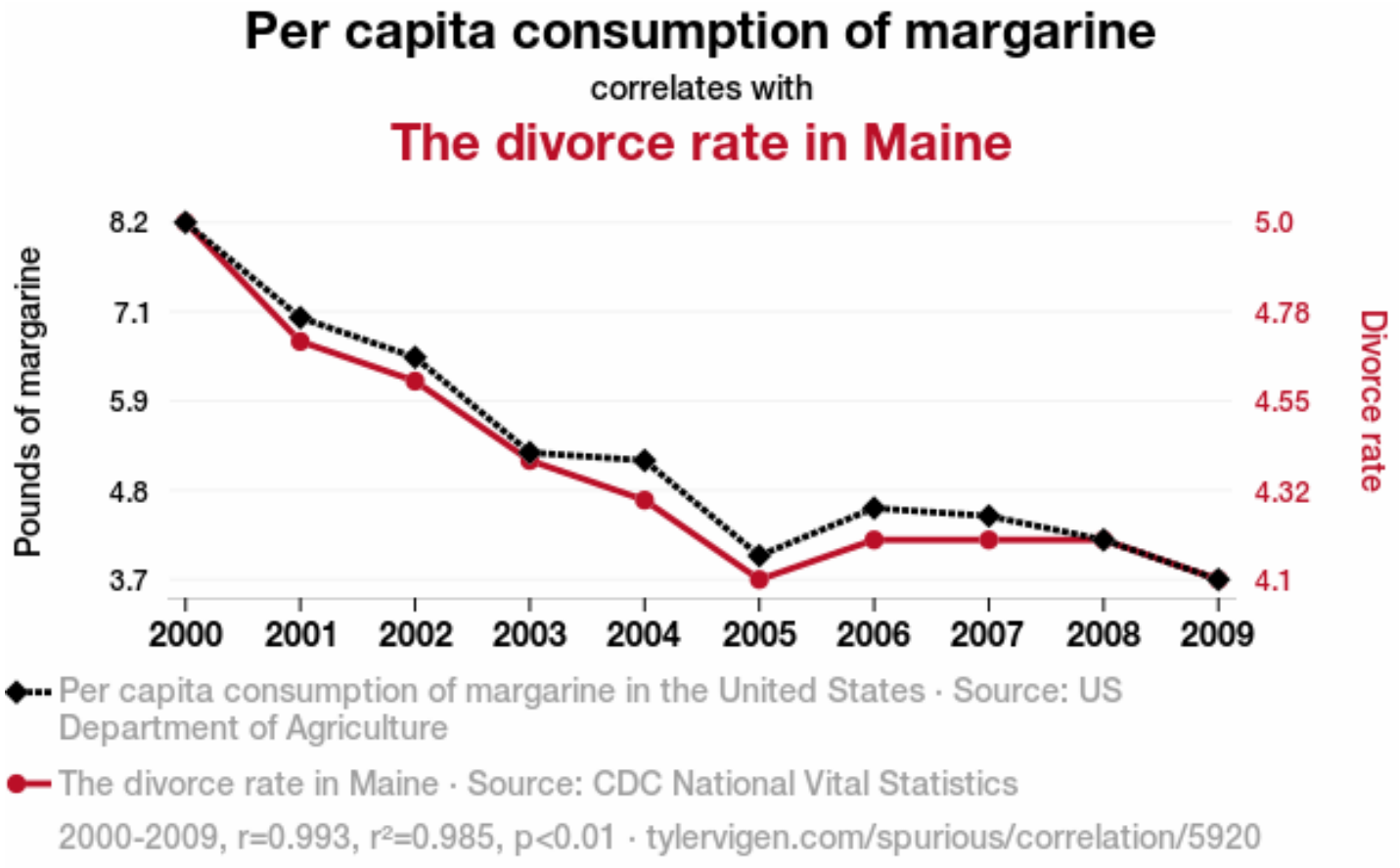
Are men applying to Berkeley more likely to get in than women?

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

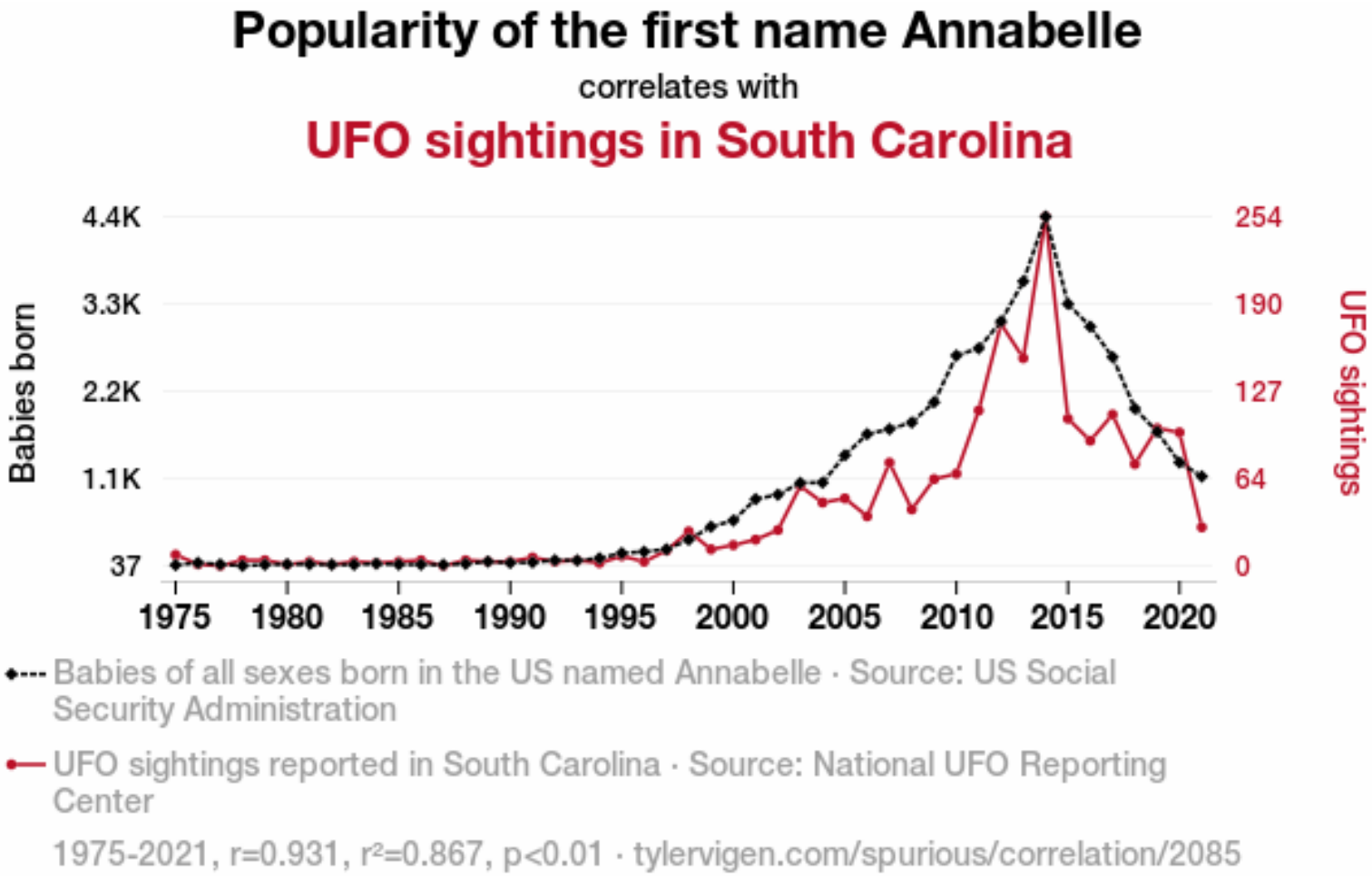
- *Departments have different acceptance rates,*
- *More women applied to departments with lower acceptance rates*

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



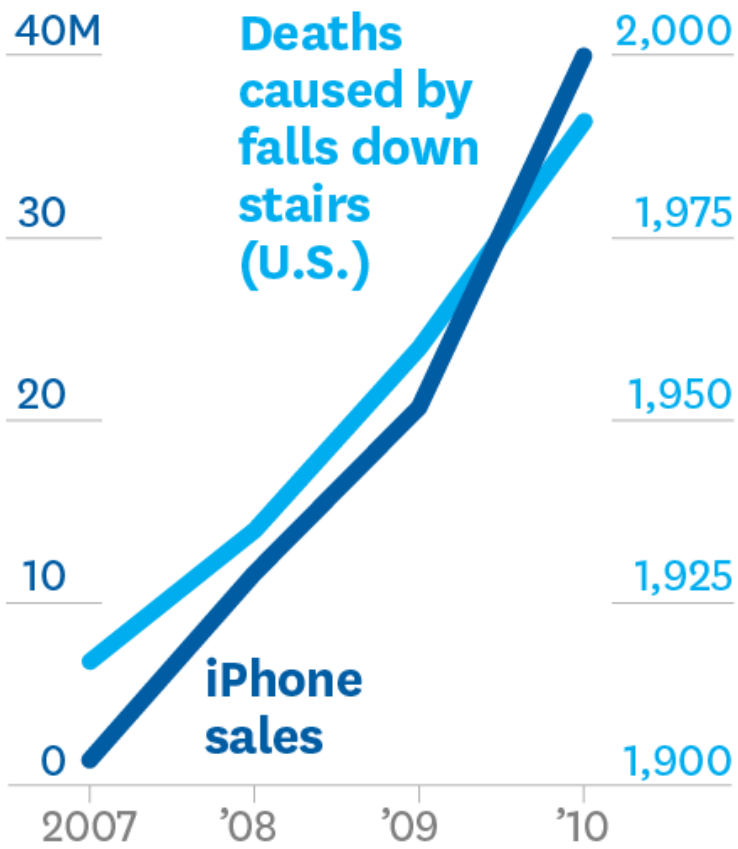
SPURIOUS CORRELATIONS



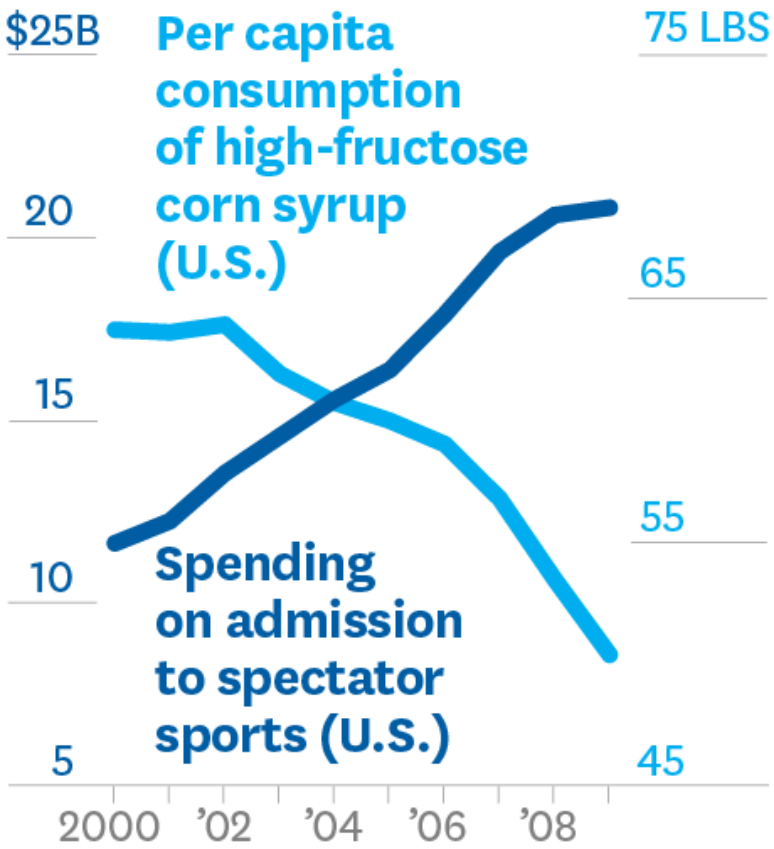
https://www.tylervigen.com/spurious-correlations

BYU

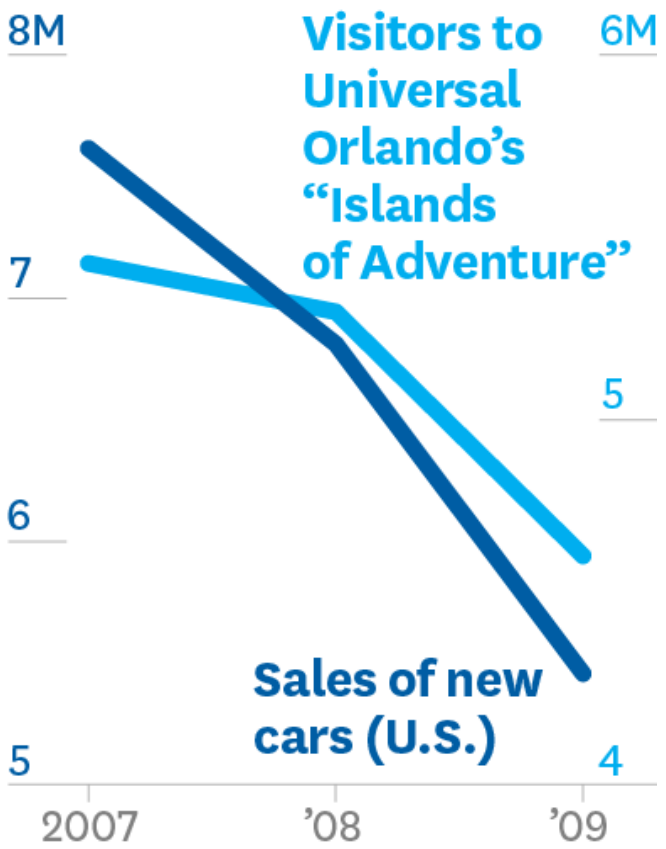
MORE IPHONES MEANS MORE PEOPLE DIE FROM FALLING DOWN STAIRS



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.

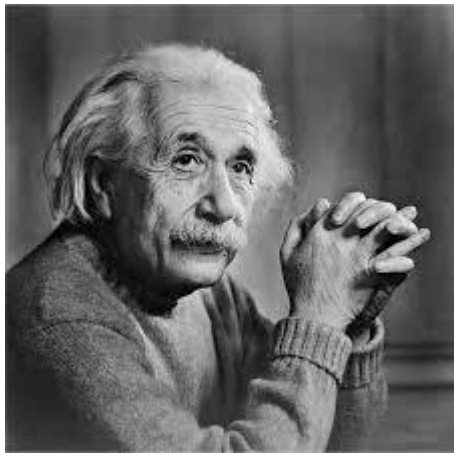
Step	Description
Step 1: Gathering data	Identify available and relevant data; gather new data if needed.
Step 2: Cleaning data	Reformat datasets, create new features, and address missing values.
Step 3: Exploring data	Create data visualizations and calculate summary statistics to explore potential relationships 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.

BYU

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)

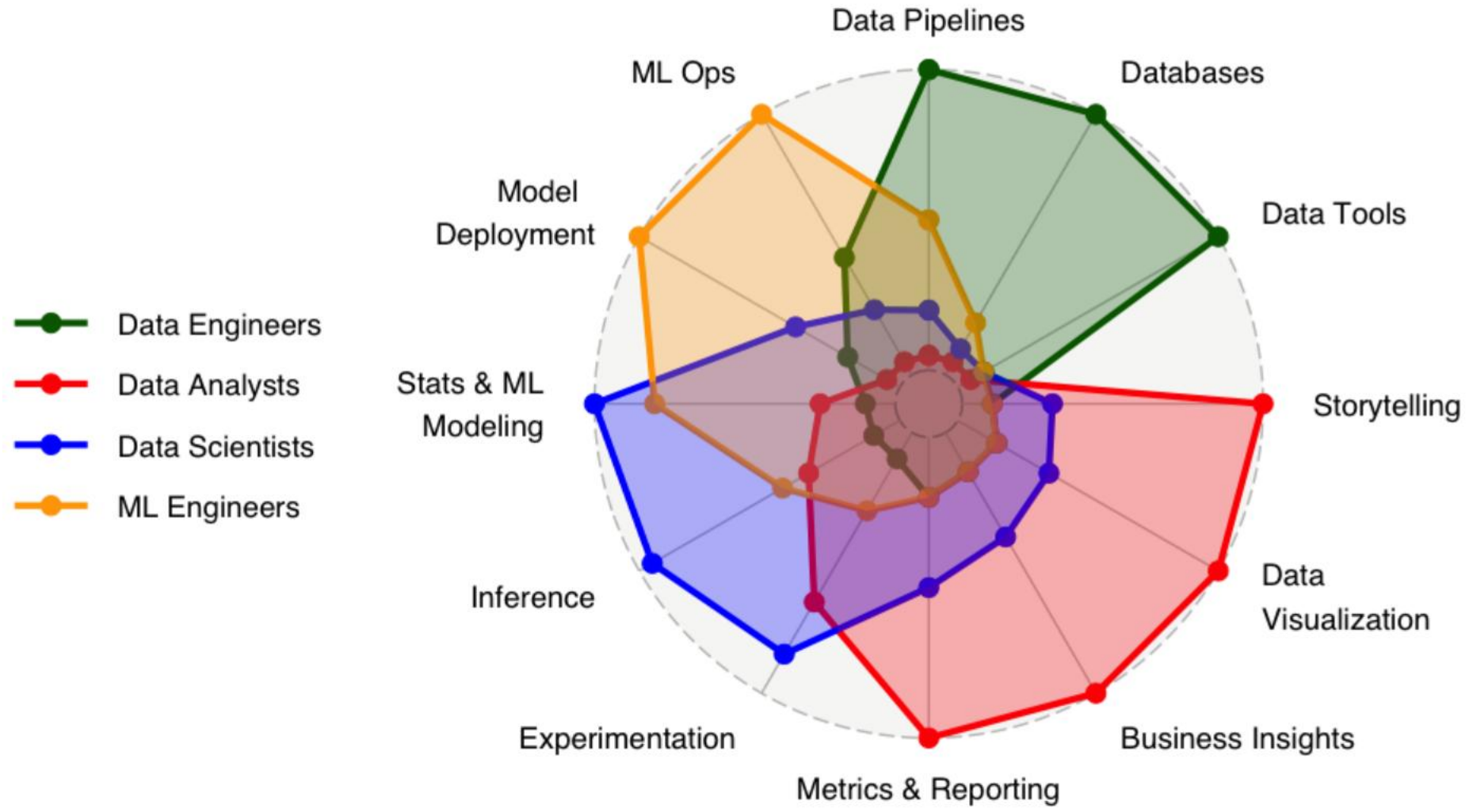
> IF I HAD AN HOUR TO SOLVE A PROBLEM I'D SPEND 55 MINUTES THINKING ABOUT THE PROBLEM AND 5 MINUTES THINKING ABOUT SOLUTIONS.



Albert Einstein

BYU

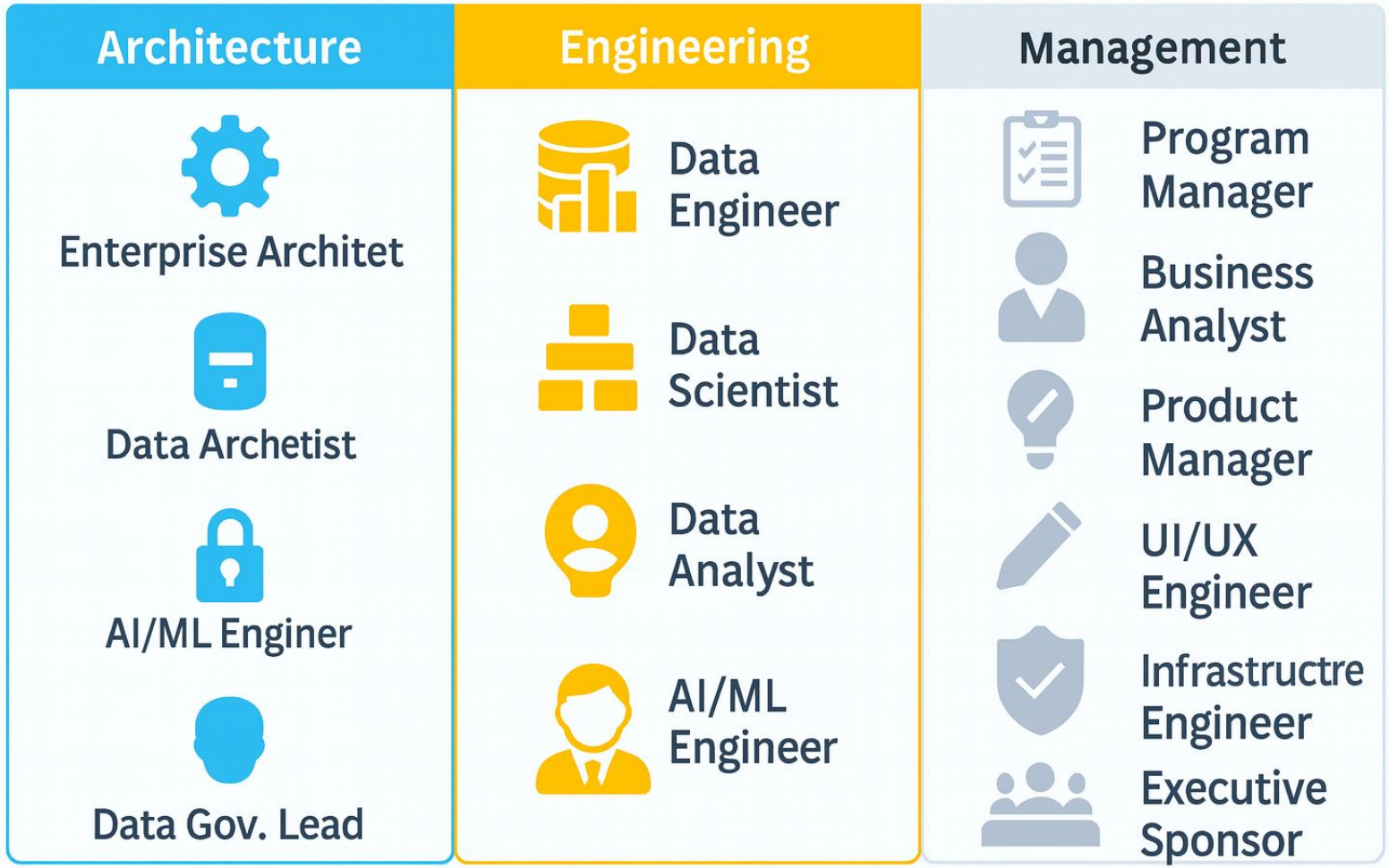
SPIDER CHART OF RELATIVE SKILLS FOR KEY DATA ROLES



From https://www.datacaptains.com/blog/guide-to-data-roles

BYU

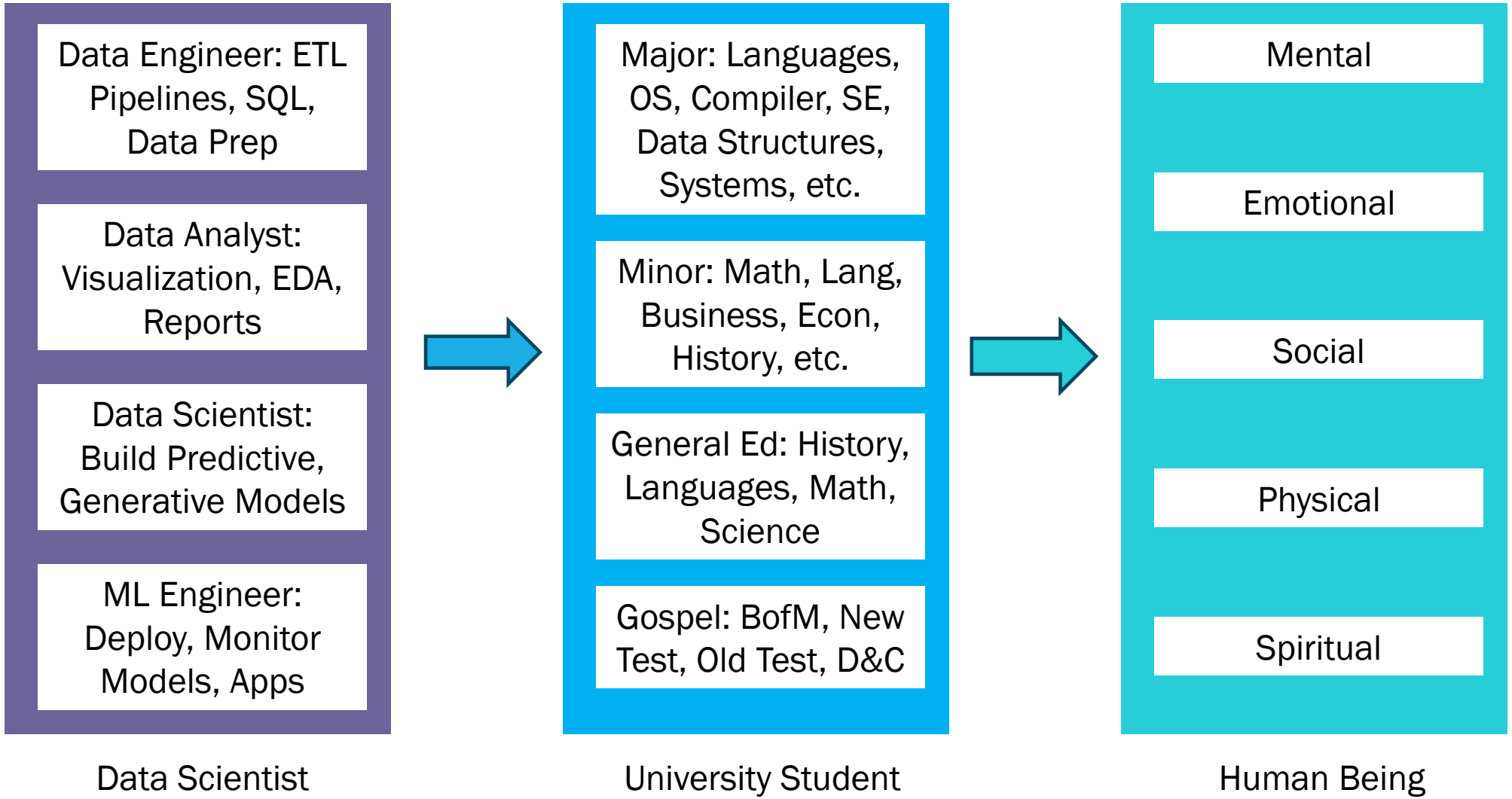
ROLES ON A DATA SCIENCE APPLICATION PROJECT



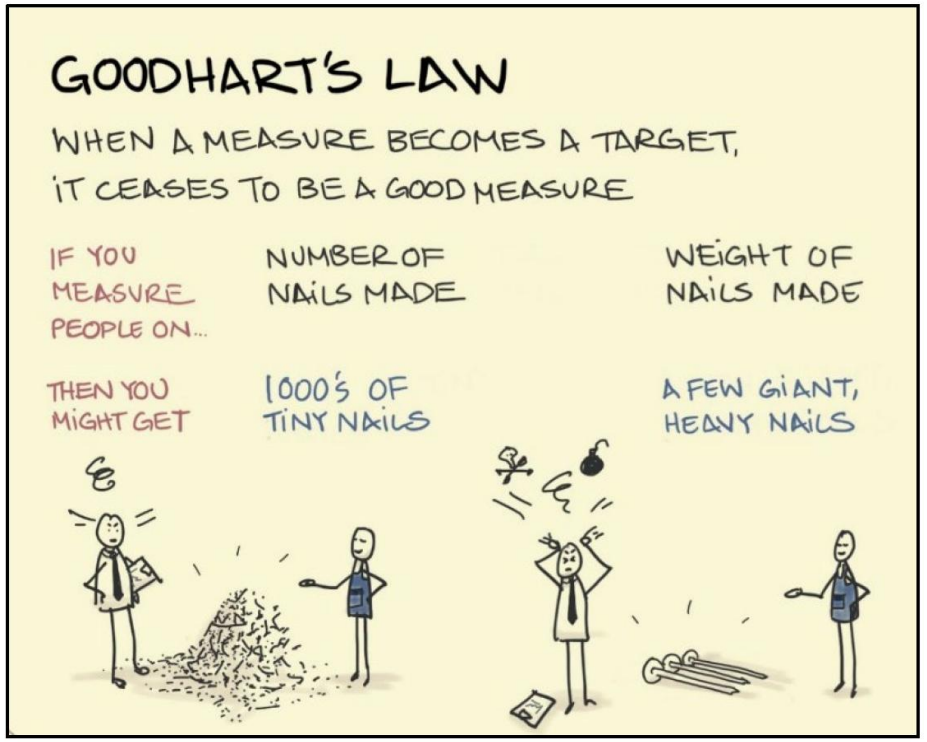
EXTENDED ROLES ON A DATA SCIENCE APPLICATION PROJECT

Role	Description
Data Engineer	Builds data pipelines, joins tables, converts data formats, prepares data for use by Data Scientists.
Data Scientist	Prepares data for modeling, extracts features, builds models
Data Analyst	Expert in SQL, BI, Excel, Analyzing data but not necessarily a domain expert (Tableau, PowerBI most popular tools)
AI/ML Engineer	Builds ML pipeline, integrate enterprise systems, monitor & manage models, skilled software engineer & data scientist
Enterprise Architect	Integrates DS applications into enterprise system (e.g., microservices, API gateway, event brokers, etc.)
Data Architect	Defines data management system architecture, data model
Data Governance Lead	Responsible for meta data, data catalog, data access, change management policies
Business Analyst	Expert in a particular domain (e.g., Finance), can use BI tools and Excel if set-up by the Data Analyst.
Program Manager	Keeps track of projects, personnel, budgets, identifies conflicts, dependencies, resource constraints
Application Engineer	Expert in technology for a particular domain or problem (e.g., Finance, Marketing, Sales, Manufacturing, etc.)
UI/UX Engineer	Designs, prototypes, and builds the user interface (mobile and web)
Product Manager	Responsible for overall product design, prioritization, deployment and assuring business value
Infrastructure Engineer	Expert in cloud and data management infrastructure
Security/Privacy Engineer	Assures application architecture is compliant with Enterprise security and privacy standards
Executive Sponsor	Oversees application development. Responsible for resourcing. Communicates with Executive Leadership.

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 ☺

BYU