CS 180 Intro to Data Science

Introduction & Course Overview

Names to Know

- Tim Kapp

  “Tim” | “TJ” | “Professor” |
  “Oh Captain My Captain”

- tkapp@byu.edu
- TMCB 2254
- Office Hours: TBD

- Teaching Assistants (WVB 1151):
  - Kayla Ou
  - Toby Alley
  - Spencer Hales
  - Spencer Marshall
  - Michael Jensen
  - Patrick Wilmot
- Office Hours: Posted on Syllabus

Tim Kapp

- **Cinco.ai** – CEO & AI Architect
- **University of Utah** – Adjunct Professor, AI & ML
- **U.N. AI for Good** – Chair, Silicon Valley Innovation Factory
- **U.S. National AI & Cybersecurity ISAO** – Founding Member
- **Tortora Brayda Institute for AI & Cybersecurity** – Fellow
- **Gerald Huff Fund for Humanity** – Board Member
- **Keynote Speaker** – AI Strategy, Post AI-Economics, The Future of Work
- **Sample Clients:** SEACEN, GoodRx, Adobe, Disney, Vatican, M1, GoDaddy, Blizzard Entertainment, Visa, ZEO Energy, InterContinental Hotels Group, Macmillan

My Team - France, Madagascar, Romania, US







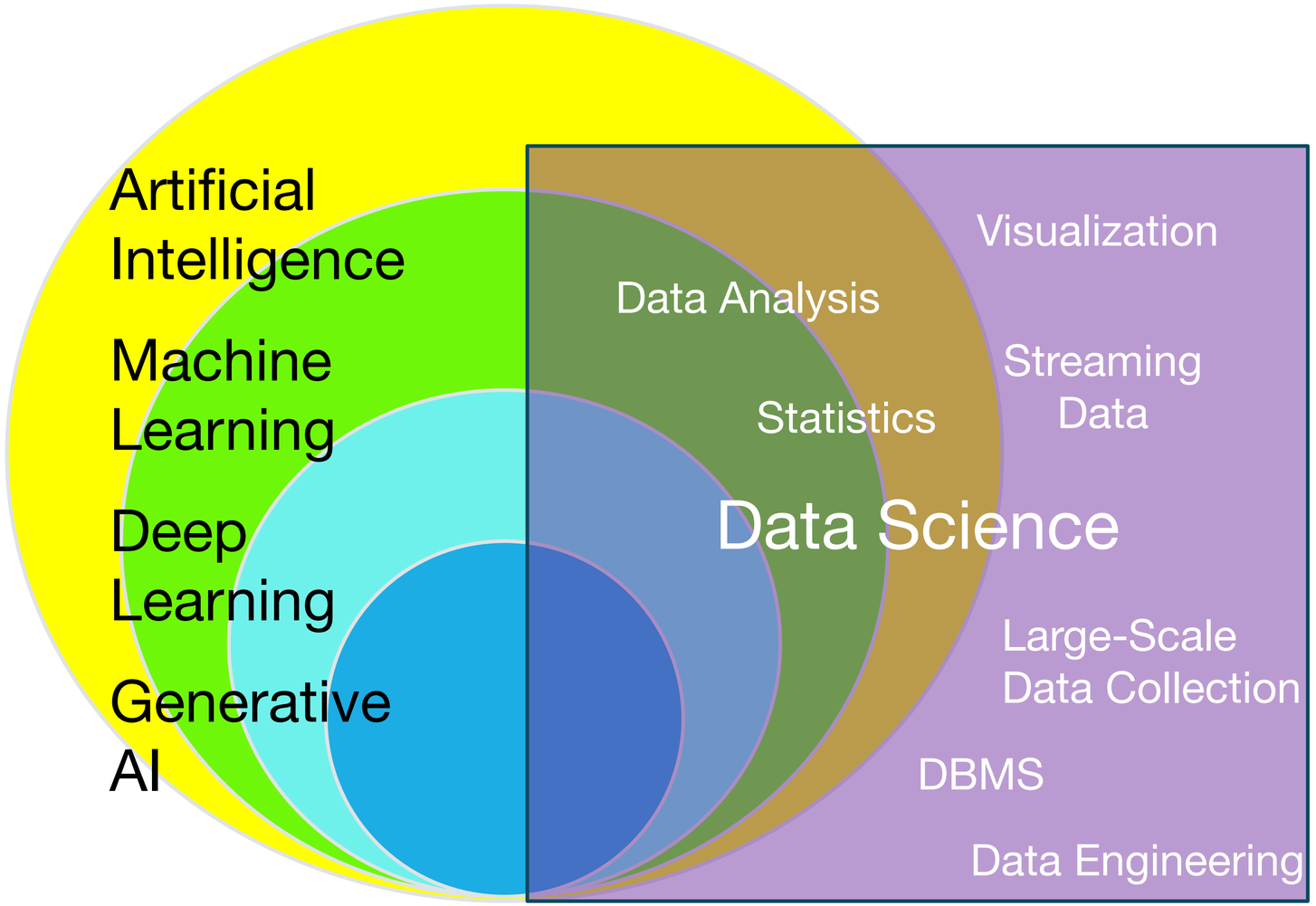









What is Data Science?

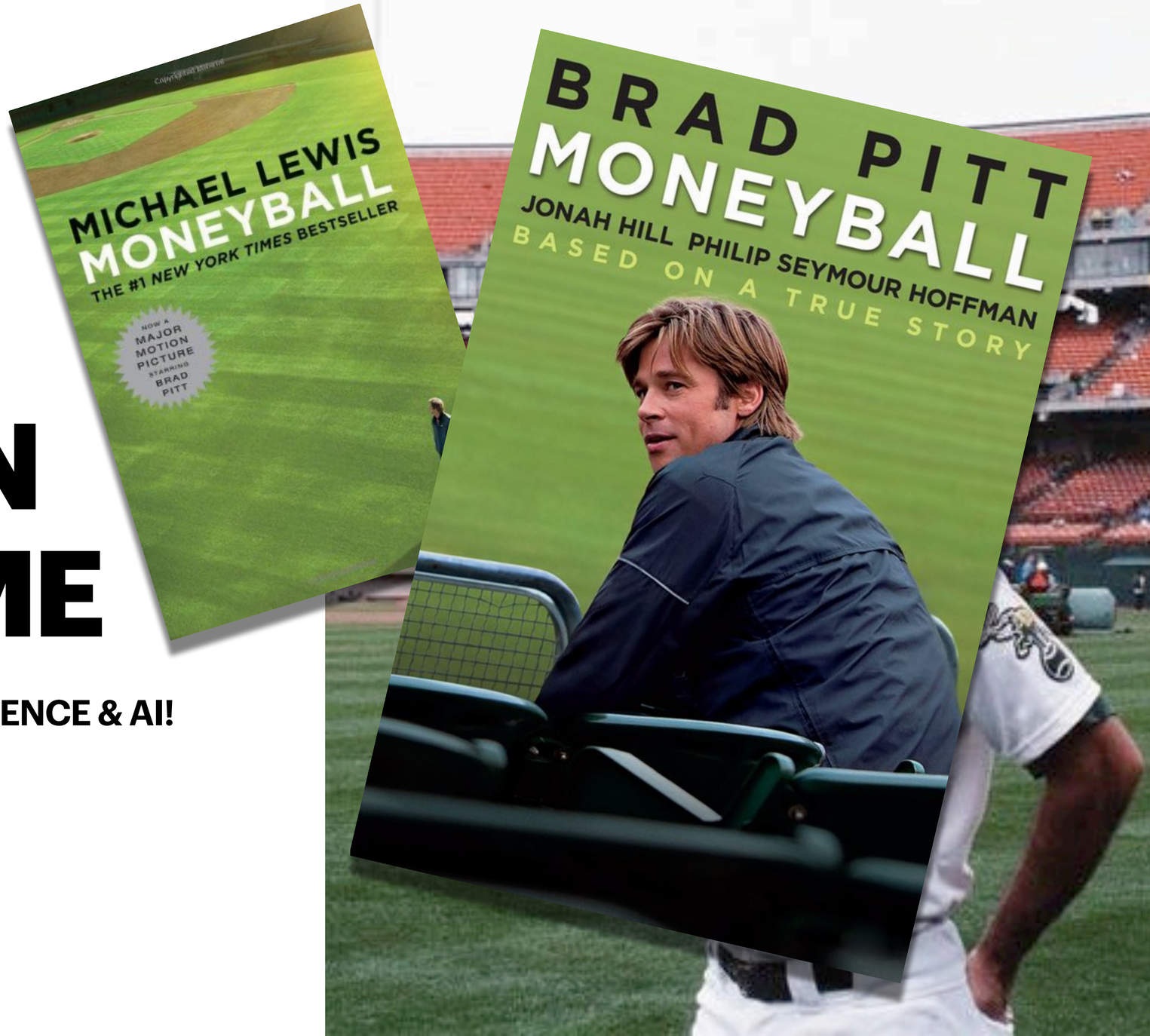


Data Science:

- **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.

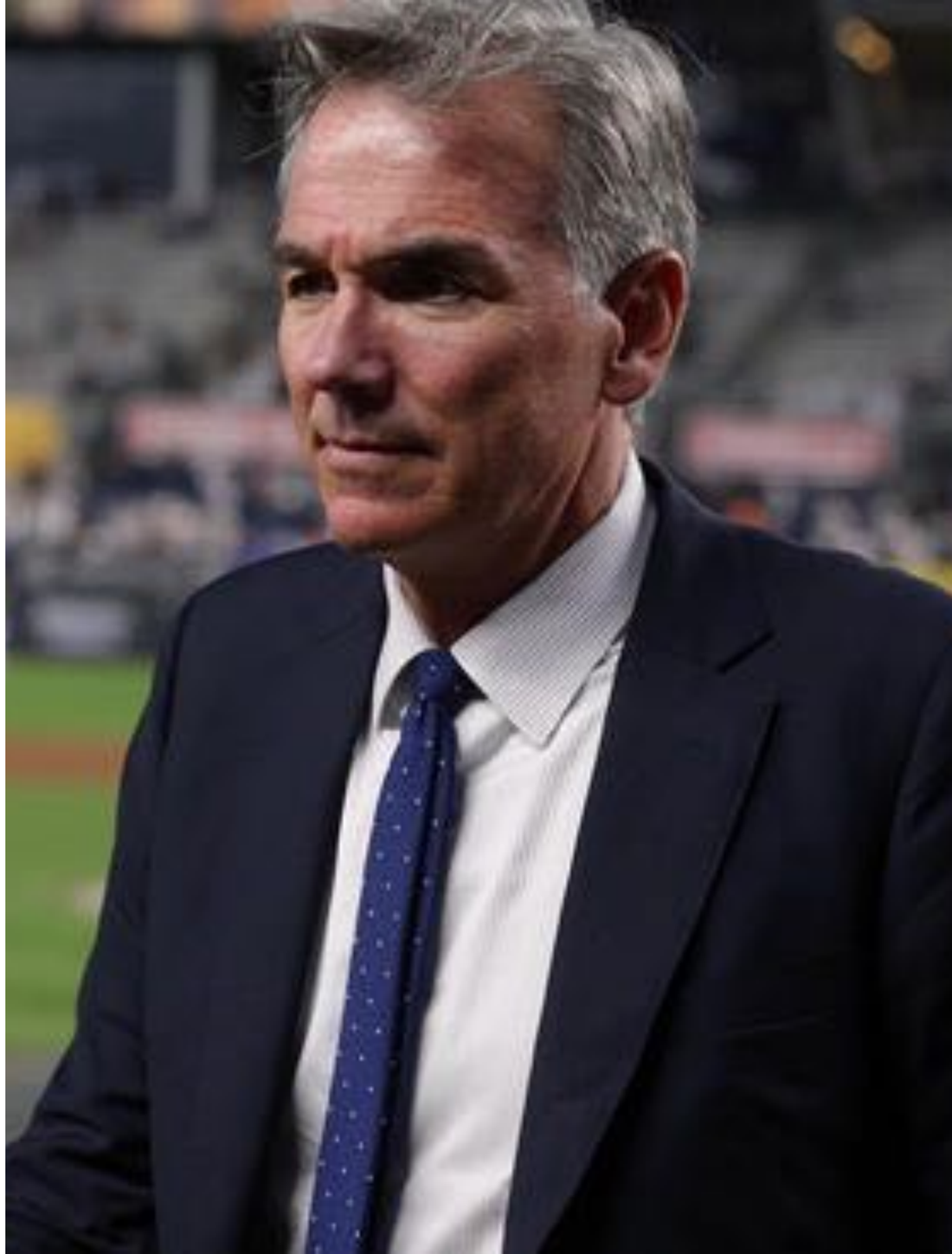
THE ART OF WINNING AN UNFAIR GAME

A MUCH BETTER DEFINITION OF DATA SCIENCE & AI!



BILLY BEANE

- A first-round pick in the MLB draft by the Mets
- 1984 to 1989 he played in MLB as an outfielder for the New York Mets, Minnesota Twins, Detroit Tigers, and Oakland Athletics but failed to ever meet the expectations of scouts.
- He joined the Athletics' as a scout in 1990 just as budgets were being slashed.
- Bean applied statistical analysis (sabermetrics) to baseball in order to build a team of amazing players that everyone else had overlooked.
- In the 2006 MLB season, the Athletics ranked 24th of 30 major league teams in player salaries but had the 5th-best regular-season record.
- The A’s reach the playoffs in 4 consecutive years (2000 - 2003)
- First team in 100 years to win 20 consecutive games.



HOW I BOUGHT MY FIRST HOUSE AND WON A TRIP TO HAWAII WITH PREDICTIVE ANALYTICS

“The Rent Distance Tradeoff for Student Housing”

1. Rent does correlate to distance from important “hot spots”
2. Small things can make a big difference - i.e. garbage disposal
3. You often discover the unexpected - i.e. pools have negative value

THE RENT-DISTANCE TRADEOFF FOR STUDENT HOUSING: AN EMPIRICAL ANALYSIS

*W. Cris Lewis and Tim J. Kapp**

Introduction

The purpose of this paper is to estimate bid-rent or rent-offer functions for single-student apartments surrounding two major universities in Utah—Brigham Young University (BYU) in Provo and Utah State University (USU) in Logan. The statistical results provide an empirical test of the monocentric land-use model and have implications for evaluating proposed investments in student housing projects, for loan appraisals of apartment buildings, and for property tax assessments. The analysis also provides implied estimates of the value or opportunity cost that the average student assigns to time.

The Monocentric Model of Urban Structure and Land Rent

The conventional monocentric household location model, as articulated by Muth (1969 and 1985), Mills and Hamilton (1989), and Mosca and Williamson (1967), analyzes the decision-making process of households and resulting pattern of land use and rents in an area characterized by one employment center where one or more members of the household commute to each work day. The model is based on a set of fairly restrictive assumptions:

- All households have identical utility functions and income levels;
- The households (or individuals forming these households) maximize the utility of consuming housing, nonhousing goods and services, and commuting to work (or school) subject to a budget constraint;
- All land surrounding the center is identical; and
- All employment (or education) opportunities and consumer services for the households are provided only at the center.

While these assumptions do not hold exactly in any real world setting, the model still has generated consistent and useful results in a myriad of applications. For example, the model's predictions of declining land rent and rent per unit of housing services with increasing distance from the center has been confirmed empirically by Muth (1984), Wiend and Muth (1972), and others. The prediction that suburban residents will

* W. Cris Lewis is with the Department of Economics at Utah State University. Tim J. Kapp was a student at Brigham Young University at the time this research was done.

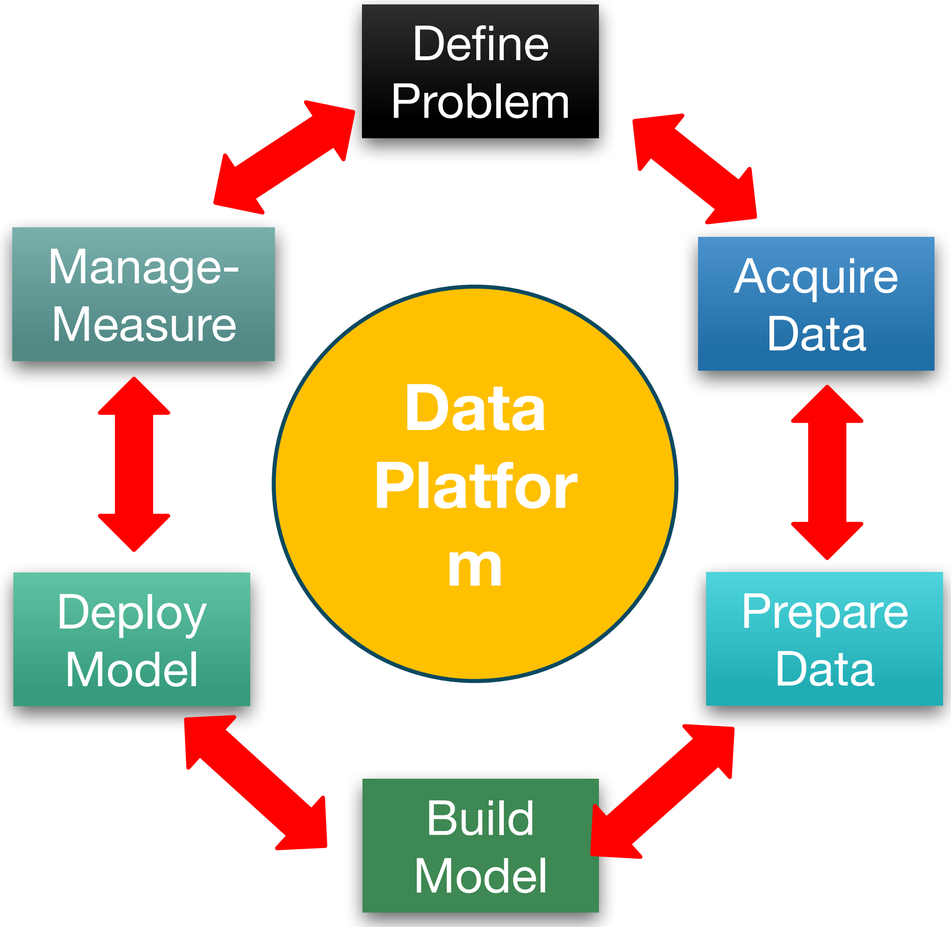
42

Why Data Science?

- Healthcare
- Advertising
- Finance
- Genomics
- Augmented reality
- Education
- Speech recognition
- Data cleansing
- Financial fraud detection
- Logistics
- Video game
- Retail
- Data science in advertising
- Manufacturing
- Transportation
- Sports
- Internet search
- E-commerce
- Airline route planning
- Medical image computing
- Drug development
- Recommendation systems
- Energy
- Virtual assistants

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



Gen AI Applications

Code

Video

Text

Image

Audio/Music

We will cover a Lot this Semester!

Class #	Week #	Month	Date	Topic	Reading	Labs
1	1	Jan	8	Welcome, Introduction, Course Objectives, DS Lifecycle	Chapter 1 Intro DS	Lab 1: Colab Set Up, GitHub
2	2	Jan	13	Python setup, Google colab, Github		
3	2	Jan	15	NumPy, Vectorization	Chapter 2.1-2.5 Python	Lab 2: Vectorization
4	3	Jan	20	Pandas, Matplotlib, Seaborn		
5	3	Jan	22	Data Cleaning and Preparation	Chapter 2.6-2.8 Python	Lab 3: NumPy, Pandas
6	4	Jan	27	Data Acquisition, ETL, Populations, Sampling	Chapter 3 Data Prep	
7	4	Jan	29	Descriptive Statistics	Chapter 4 Prob & Stat	Lab 4: Data Preparation
8	5	Feb	3	Exploratory Data Analysis (EDA)		
9	5	Feb	5	Principles of Data Visualization	Chapter 5 EDA	Lab 5: Data Visualization
10	6	Feb	10	Data management - databases, SQL queries		
11	6	Feb	12	More SQL Features, Joins	Chapter 7 DB and SQL	Lab 7: Data Engineering with SQL
12	7	Feb	17	**MONDAY SCHEDULE, NO CLASS**		
13	7	Feb	19	SQLite		
14	8	Feb	24	MIDTERM REVIEW		
15	8	Feb	26	MIDTERM		Midterm
16	9	Mar	3	Overview of ML		
17	9	Mar	5	Unsupervised Learning- Kmeans	Chapter 8 Unsupervised Learning	
18	10	Mar	10	Unsupervised Learning- Hierarchical, DBSCAN		
19	10	Mar	12	Supervised Learning: Part 1	Chapter 9 Supervised Learn	Lab 8: Cluster Analysis
20	11	Mar	17	Supervised Learning: Part 2		
21	11	Mar	19	Evaluation of models, comparing performance	Chapter 10 Decision Trees	Lab 9: ML Classification/Regression
22	12	Mar	24	Feature Importance with RF and Logistic Regression		
23	12	Mar	26	ANN, Multi-Layer Perceptron, Backpropagation	Chapter 12 Eval	**New Lab?**
24	13	Mar	31	Deep Learning		
25	13	Apr	2	GenAI - Introduction	Chapter 13 ANN	Lab 10: MLP and Backpropagation
26	14	Apr	7	GenAI - Applications		
27	14	Apr	9	Ethics / Data Privacy / Business and Government Policy+	Chapter 14 GenAI	Lab 11: GenAI Applications
28	15	Apr	14	Review and Wrap Up		
29	16	Apr	16	EXAM PREP DAY (NO CLASS)	Chapter 15 AI Ethics	Final
		Apr	22	LAST DAY OF FINALS		

Course Goals

- Learn new data science and AI tools quickly in a fast-changing technical world
- Use data, machine learning, and AI to extract real insight from messy, real-world data
- Work effectively with AI systems, knowing when to trust them—and when not to
- Communicate insights clearly through visualization, dashboards, and data storytelling
- Think critically about results, assumptions, bias, and uncertainty
- **DO GOOD THINGS…Apply data science and AI for real-world impact and social good**

We are a Learning Community

What this means

- We learn faster and deeper together than alone
- Everyone contributes: questions, ideas, mistakes, and insights

How we do it

- Say your name for the first few classes at least
- Introduce yourself to the people around you—every class
- Talk, compare approaches, challenge ideas, help each other think

Why it matters

- Better understanding and performance
- Stronger professional and social connections
- A real sense of belonging
- And honestly… it’s more fun

AI “Policy”

Generative AI in This Course

- GenAI tools can be powerful learning assistants
- Misuse turns them into a crutch and weakens understanding
- Our goal is deep skills, not just correct answers

Use AI to: clarify concepts, look up syntax, understand examples

Don’t use AI to: solve assignments end-to-end or replace your thinking

We trust you to follow this policy. We won’t actively police AI use—but if misuse is obvious, we’ll address it.

Some parts of the midterm and final will be AI-free by design.

Upcoming Assignments

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.

Reading Assignment: Chapter 1

- Points automatically accrue as you do the activities in the book.

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.

Getting Set Up “Assignment”

- This will help you (and me) hit the ground running!

Why data science and AI?

What do you see?

A metaphor for data science

The Snow Leopard Represents:

- Signal hidden in noise
- Rare but important patterns
- Outliers that matter
- Ground truth you miss if you don’t look carefully

The Cliff Represents:

- Raw data — messy, unstructured, overwhelming
- Where most people stop looking

Why It’s Hard to See:

- Your brain optimizes for speed, not correctness

- **“The hardest part of data science isn’t computing.It’s knowing what you’re looking for — and what you might be missing.**
- **“Most failures in data science aren’t wrong answers. They’re unseen answers.”**
- **And because…**

Parts of the World are too Complex for Humans!

Airways. @PythonMaps
This map shows the world's flight paths and airports. It maps 10,000 airports and 67,663 routes linking those airports.
Data source - https://openflights.org/data.html

Interesting applications of data science and things to watch out for

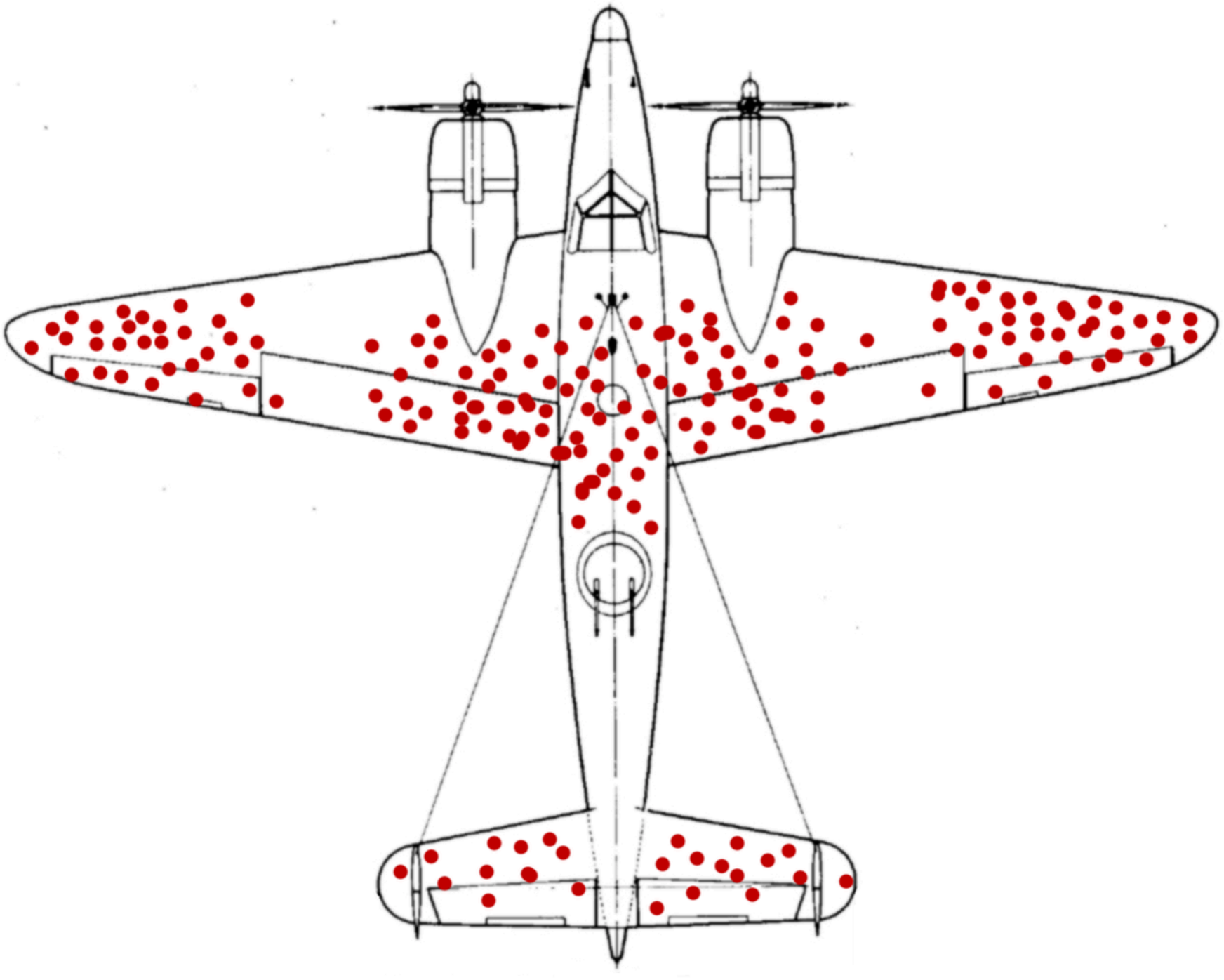
1971: Measles-mumps-rubella vaccine developed

1963: Measles vaccine developed

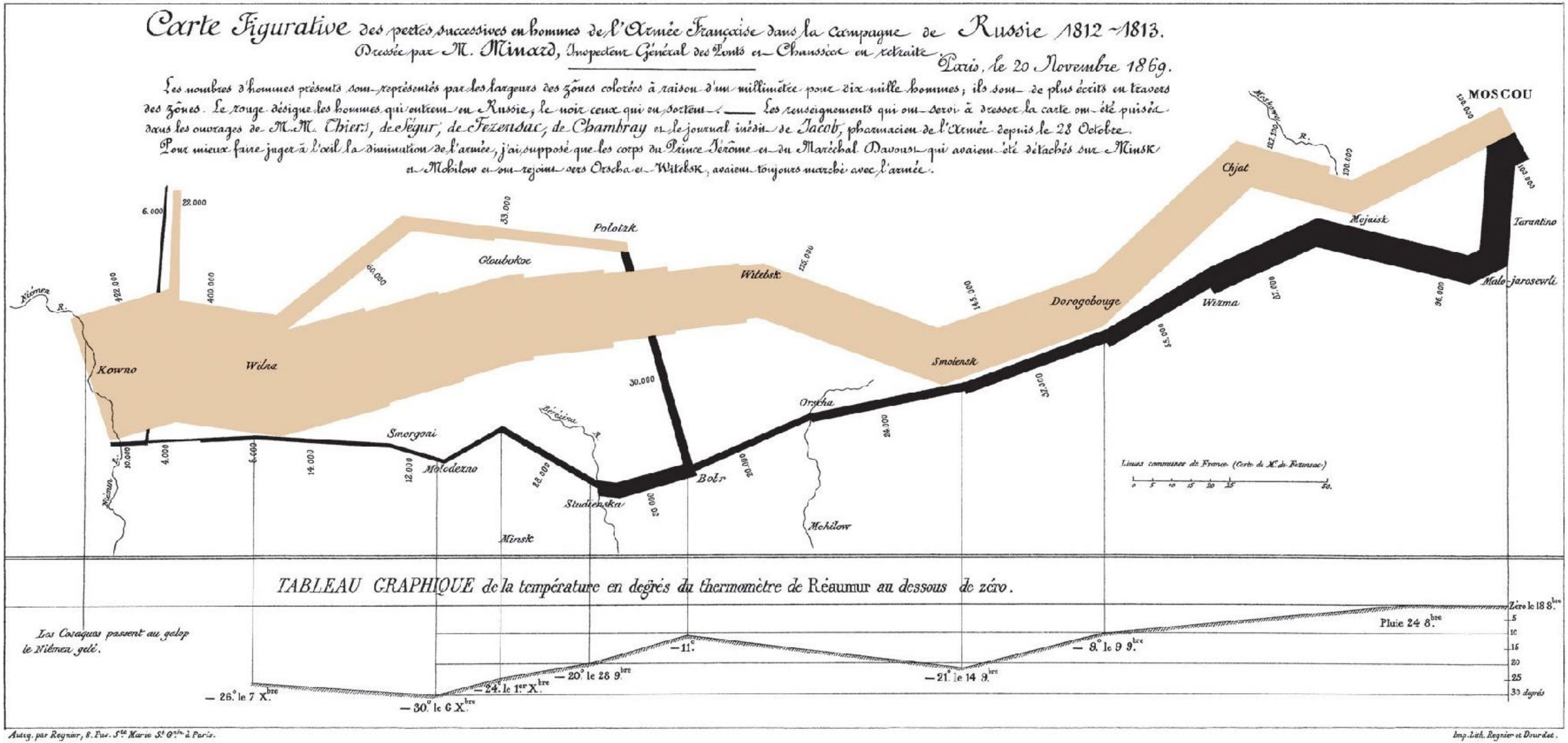
1980: Mandatory for kindergarteners to be vaccinated

US States

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



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?

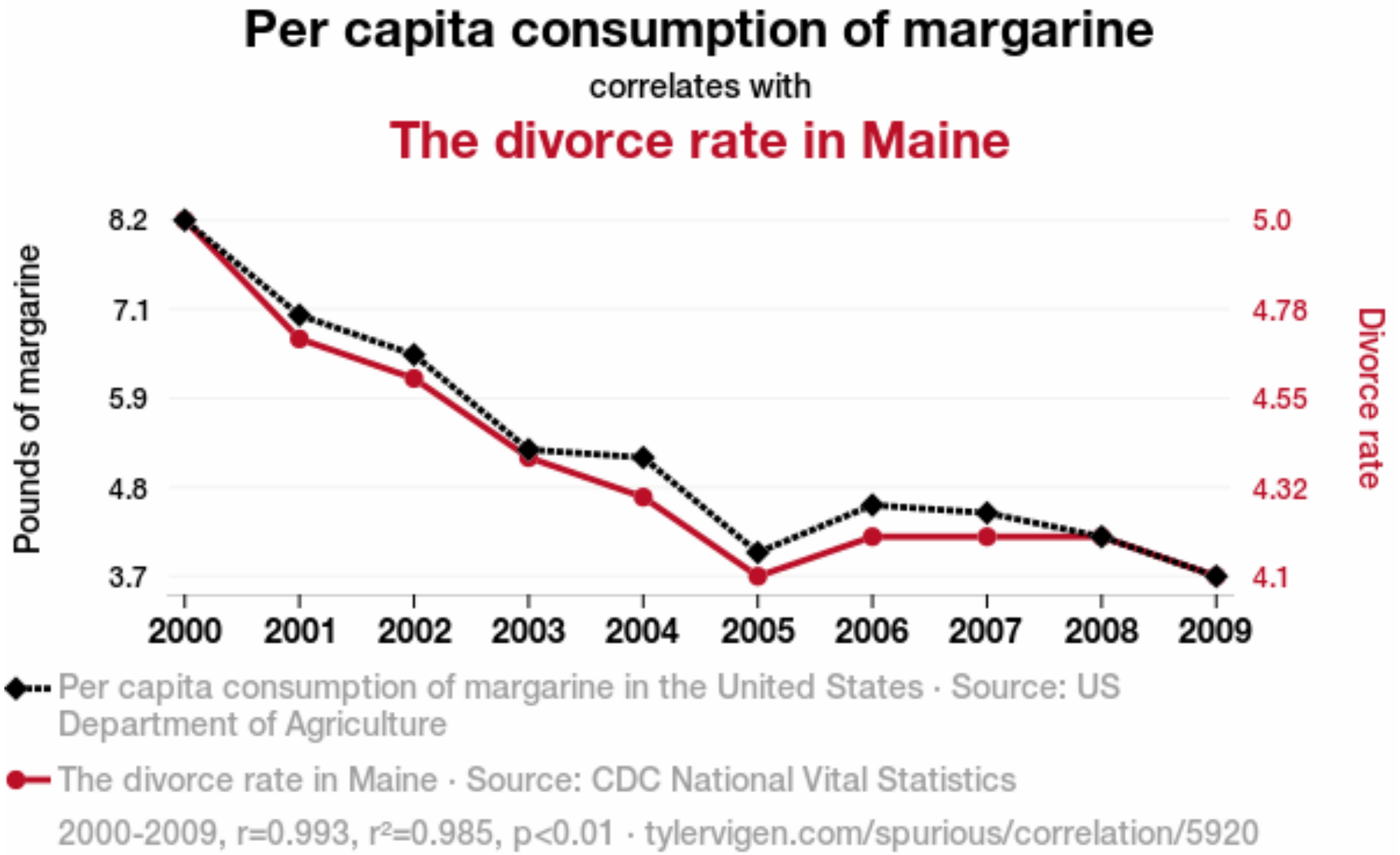
	Men		Women	
	Applicants	Admitted	Applicants	Admitted
Total	8442	**44%**	4321	

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

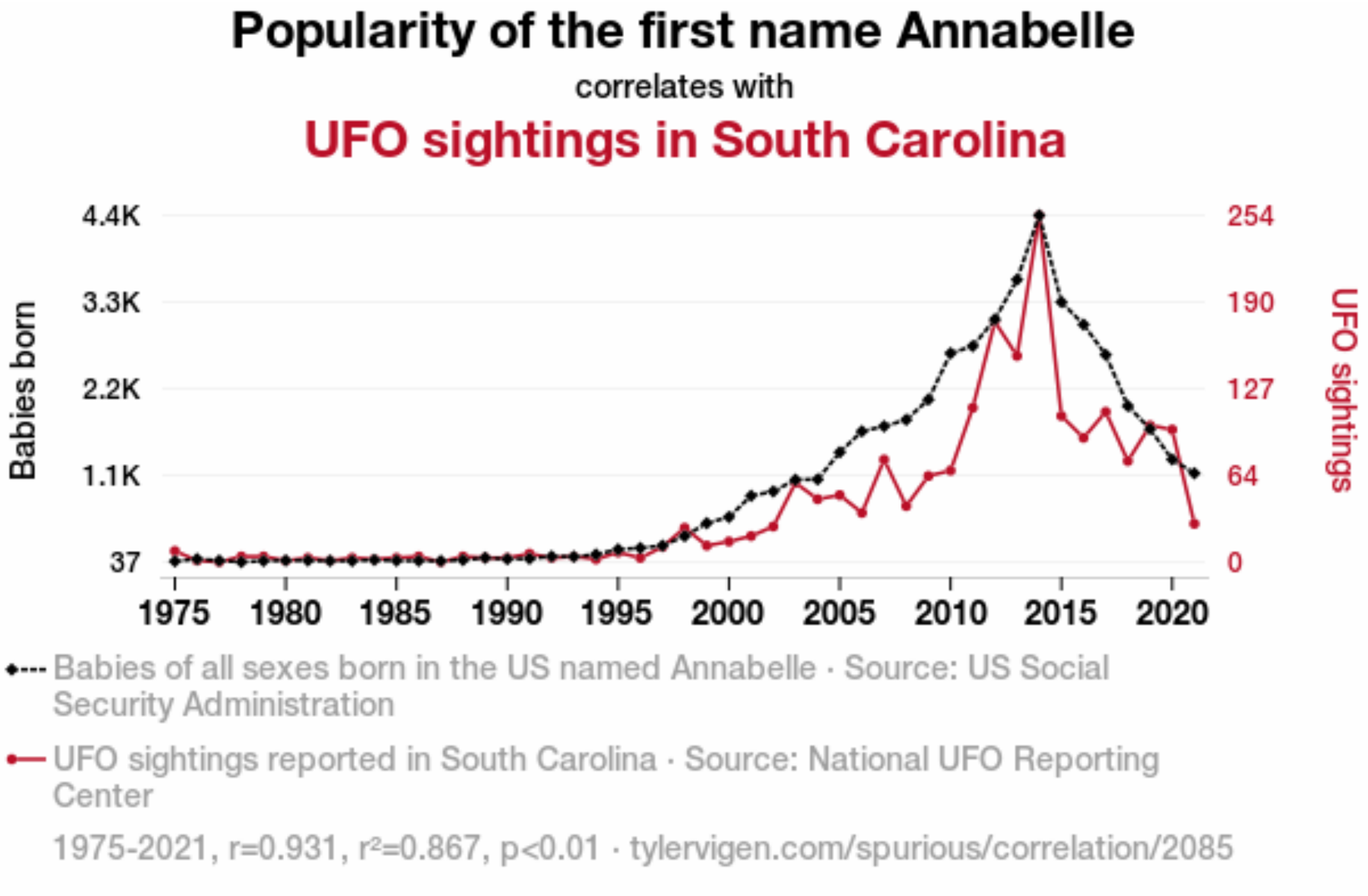
Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	***825***	62%	108	**82%**
B	***560***	63%	25	**68%**
C	325	**37%**	***593***	34%
D	417	33%	375	**35%**
E	191	**28%**	***393***	24%
F	373	6%	341	**7%**

Spurious Correlations

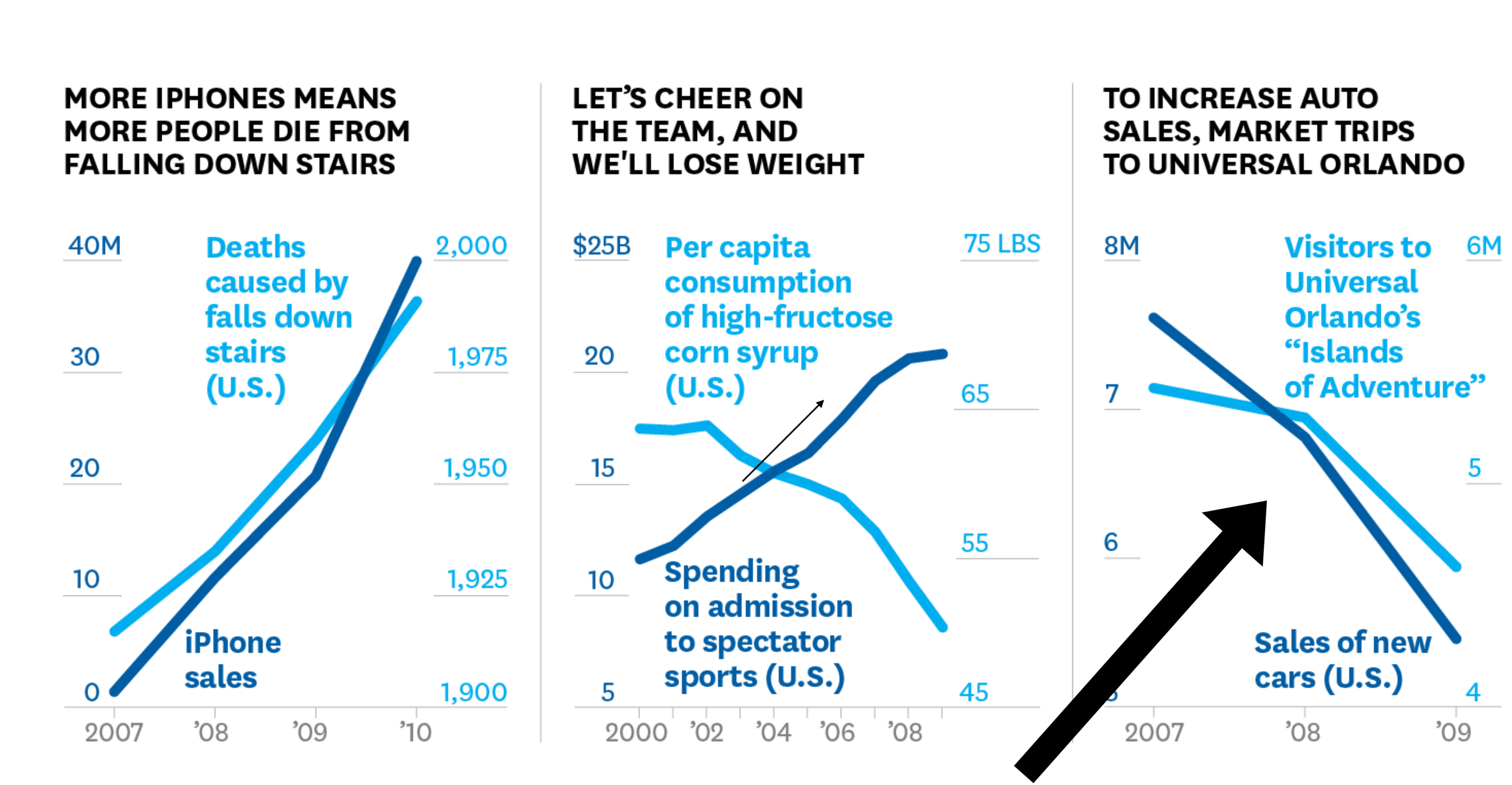
Spurious Correlations



Spurious Correlations



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



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

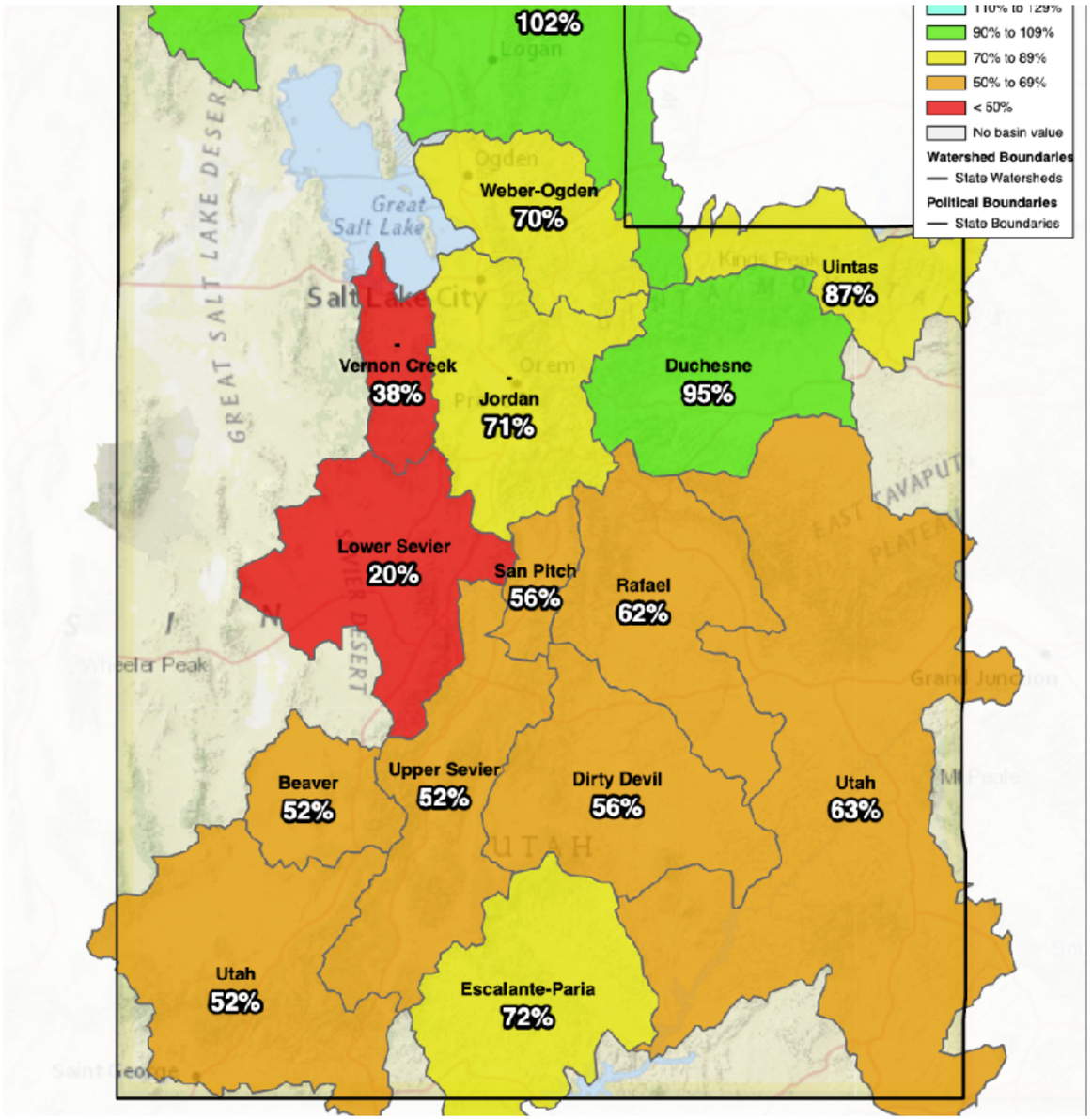
Confounding Variable Economy

What’s the cost of bad data science?

Crying “Wolf”

Utah Snow Pack

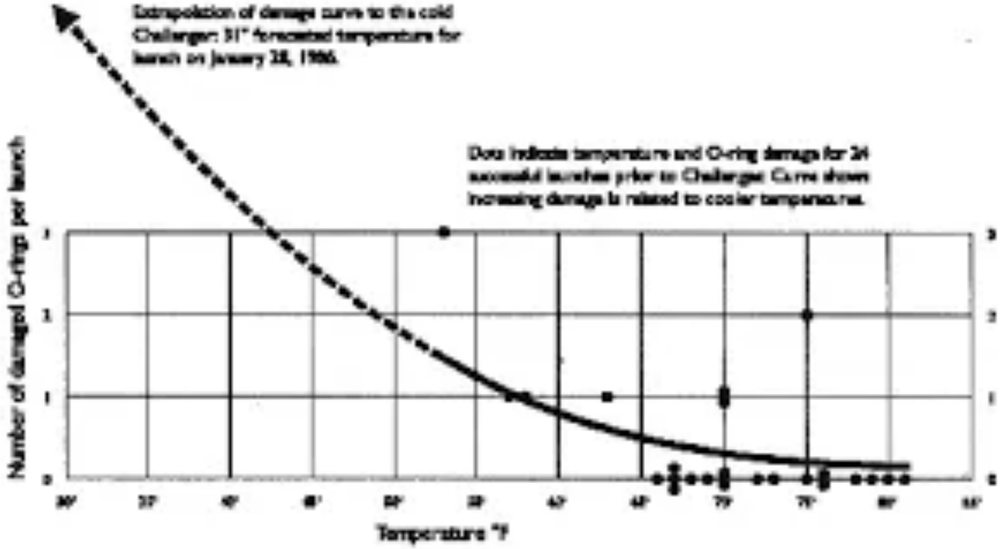
- Even when the data is correct, you have to worry about interpretation.
- At the end of a snow season, this chart would be disastrous. At the beginning it’s “meh”. Why?
- There are many systems that act like expiring “options” that behave this way.

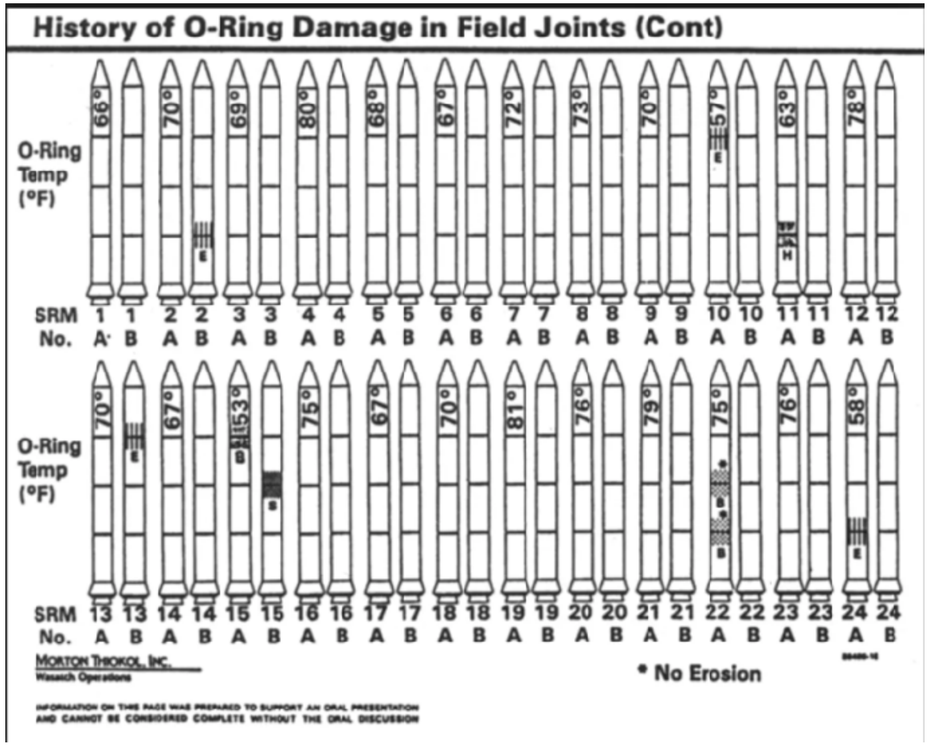


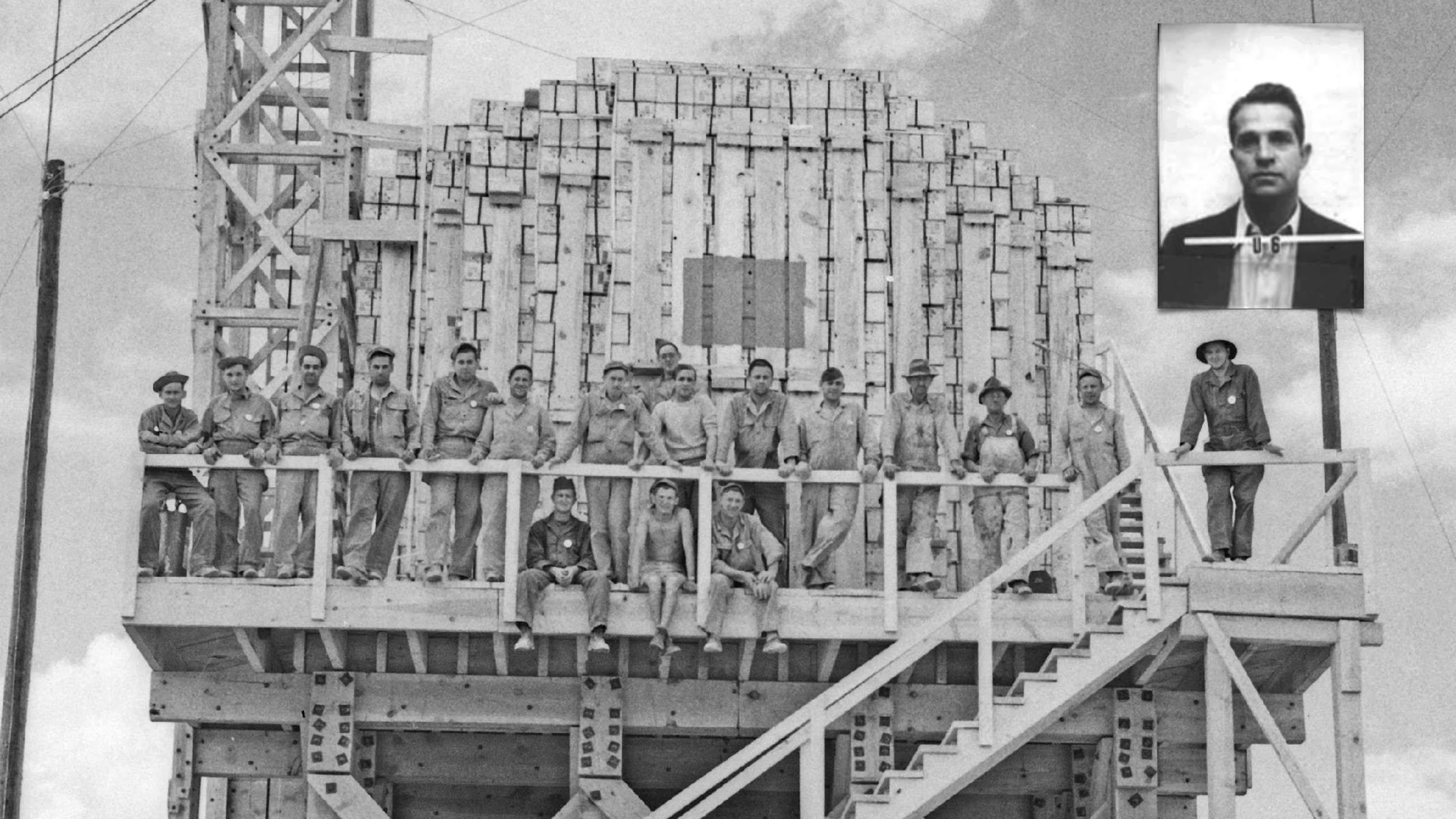
Bad data costs lives

The Space Shuttle

- Rubber becomes less flexible as temperature drops. O-rings might not seal properly during ignition,
- At launch time, the temperature at the Kennedy Space Center was 36 degrees Fahrenheit. This was 15 degrees colder than any prior Space Shuttle launch.
- Before the launch, engineers attempted to persuade NASA management to delay the mission. They brought data from earlier shuttle flights showing instances of O-ring erosion.
- Charts shown to decision-makers listed launches in tabular form, documenting which O-rings had erosion and which did not. What the charts failed to do was plot O-ring damage against temperature in a clear, visual manner.
- Launches with no O-ring damage were largely excluded from the analysis. This omission distorted the picture.

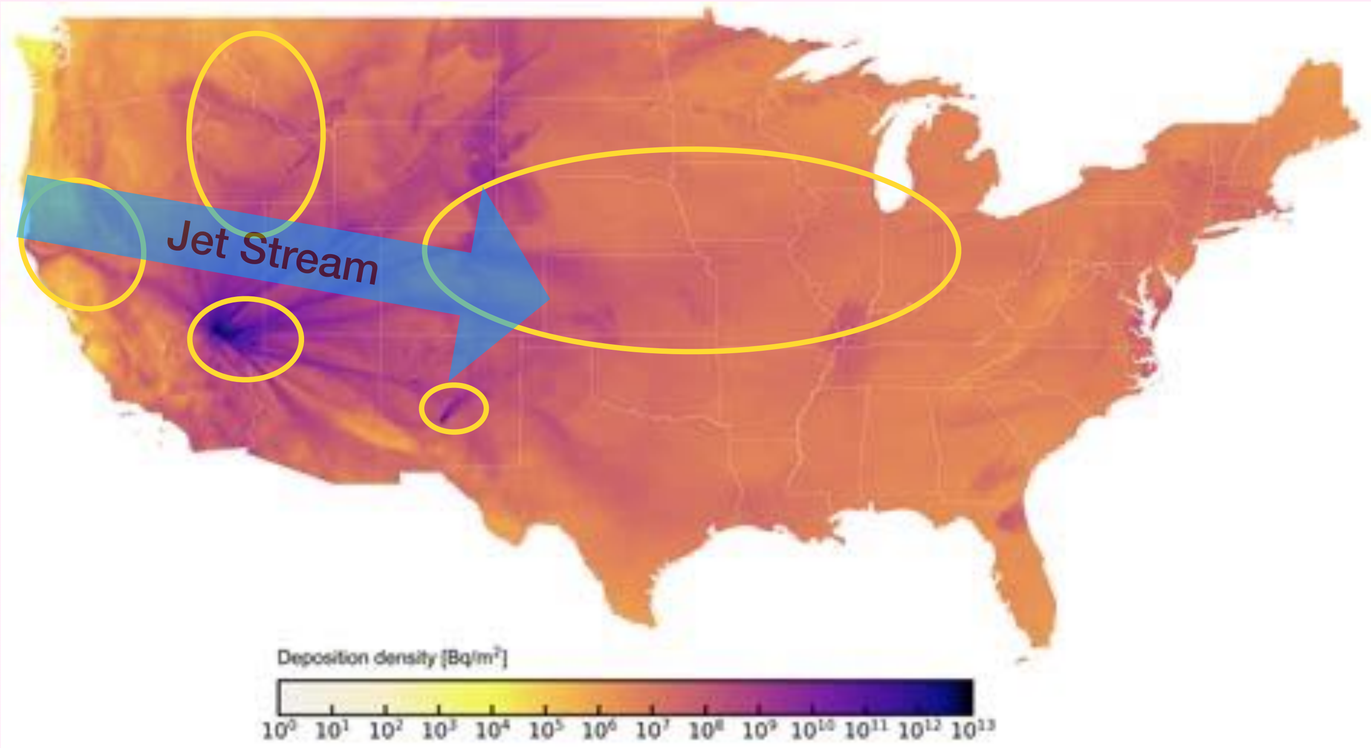





U. G

The Fallout

- Trinity bomb was 200x larger than the TNT test.
- 4-10x larger than consensus estimate (~21 kilotons)
- Land based testing: cancer-related deaths near blast zones ~ 10-24
- Bomb tests pushed radio active material into the stratosphere and jet stream
- Between 17k - 340k excess deaths (US only) from cancer nationwide



Nuclear Test → Fallout on Grass → Cows Eat Grass → Milk → People → Thyroid & Cancer Risk

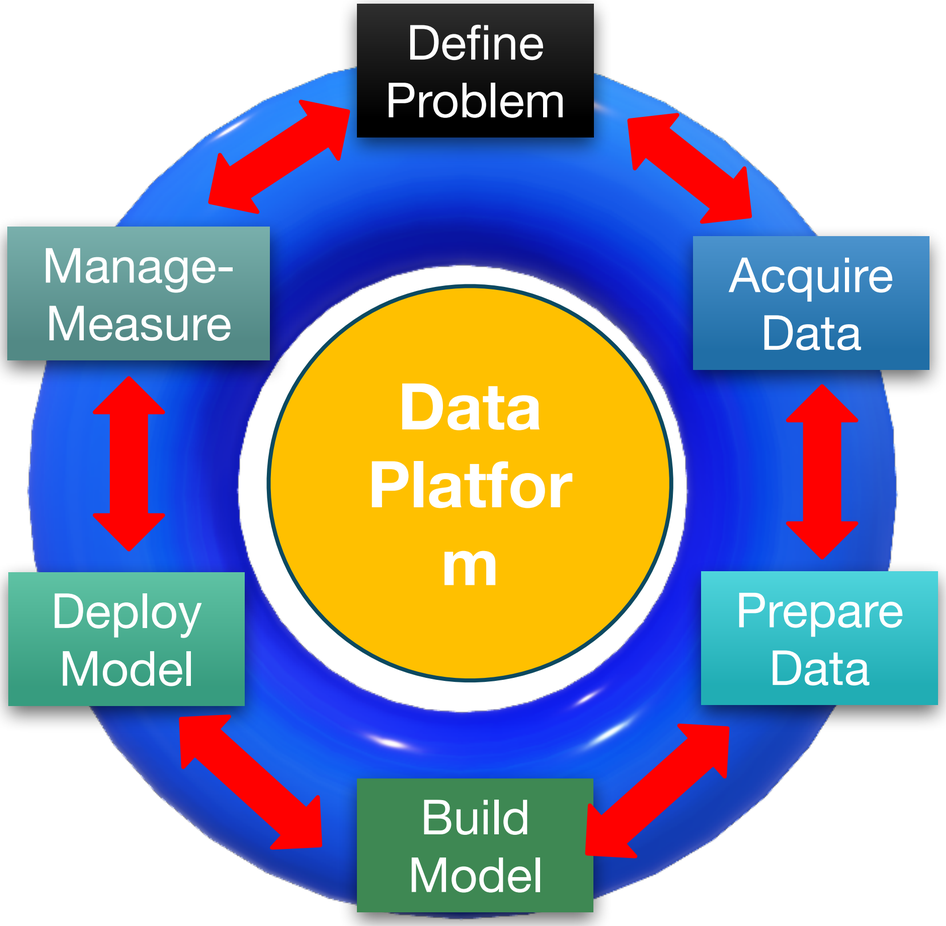
zyBook Data Science Lifecycle for Data Analysis

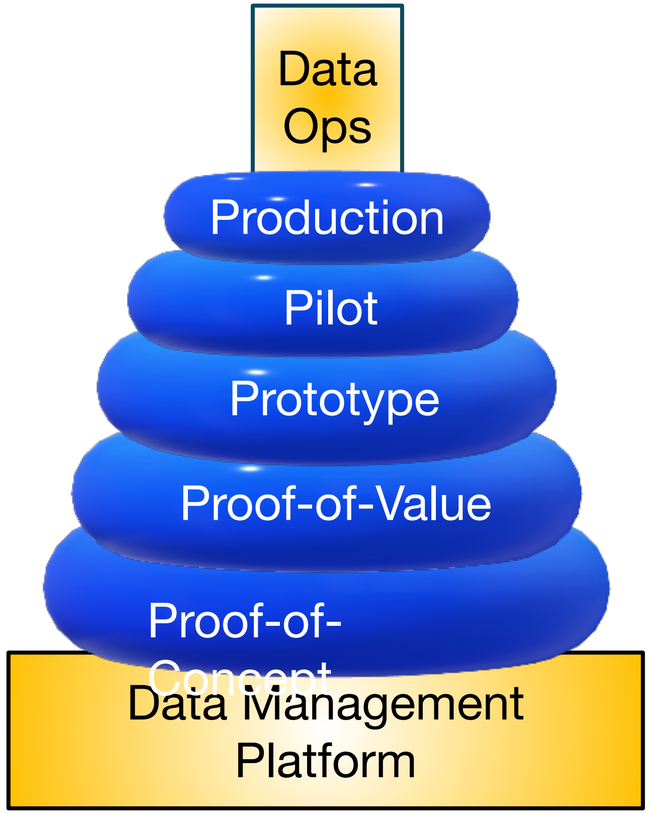
Table 1.4.1: Data science lifecycle.

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.

How would you approach a data-driven problem?

THE 6S5P DATA SCIENCE LIFECYCLE FOR TAKING ACTION WITH DATA





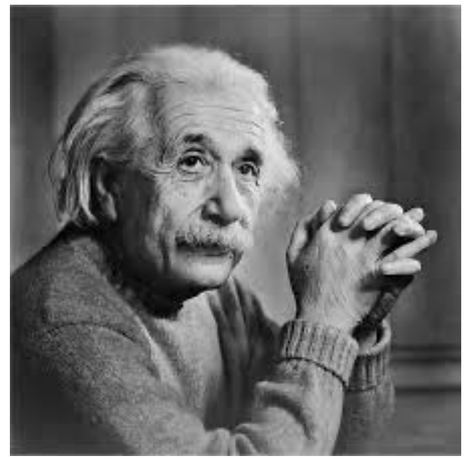
Inspiration for 6S5P process

BYU

1. DEFINE THE (REAL) 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."



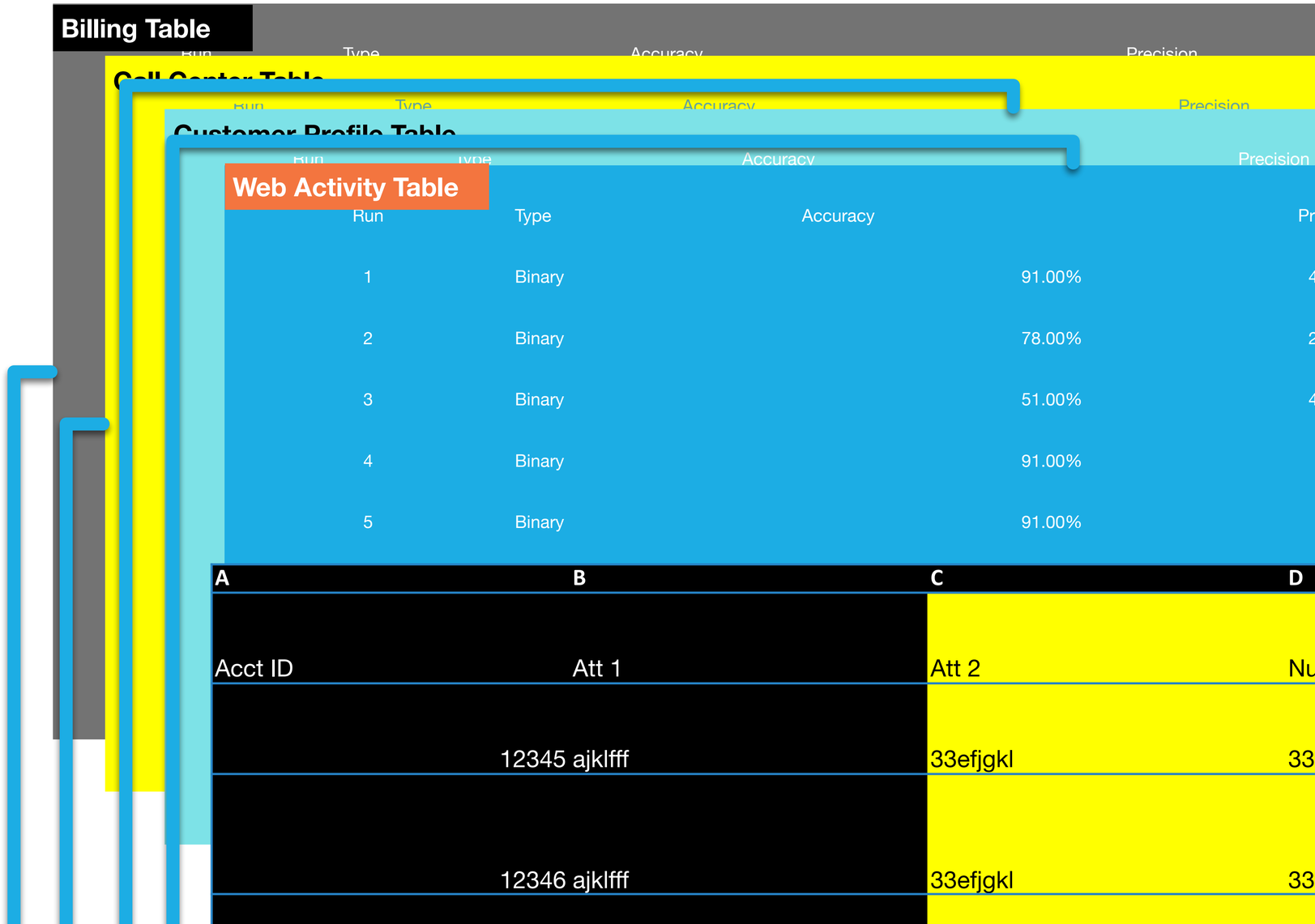
Albert Einstein

BYU

2. IDENTIFY AND ACQUIRE THE DATA

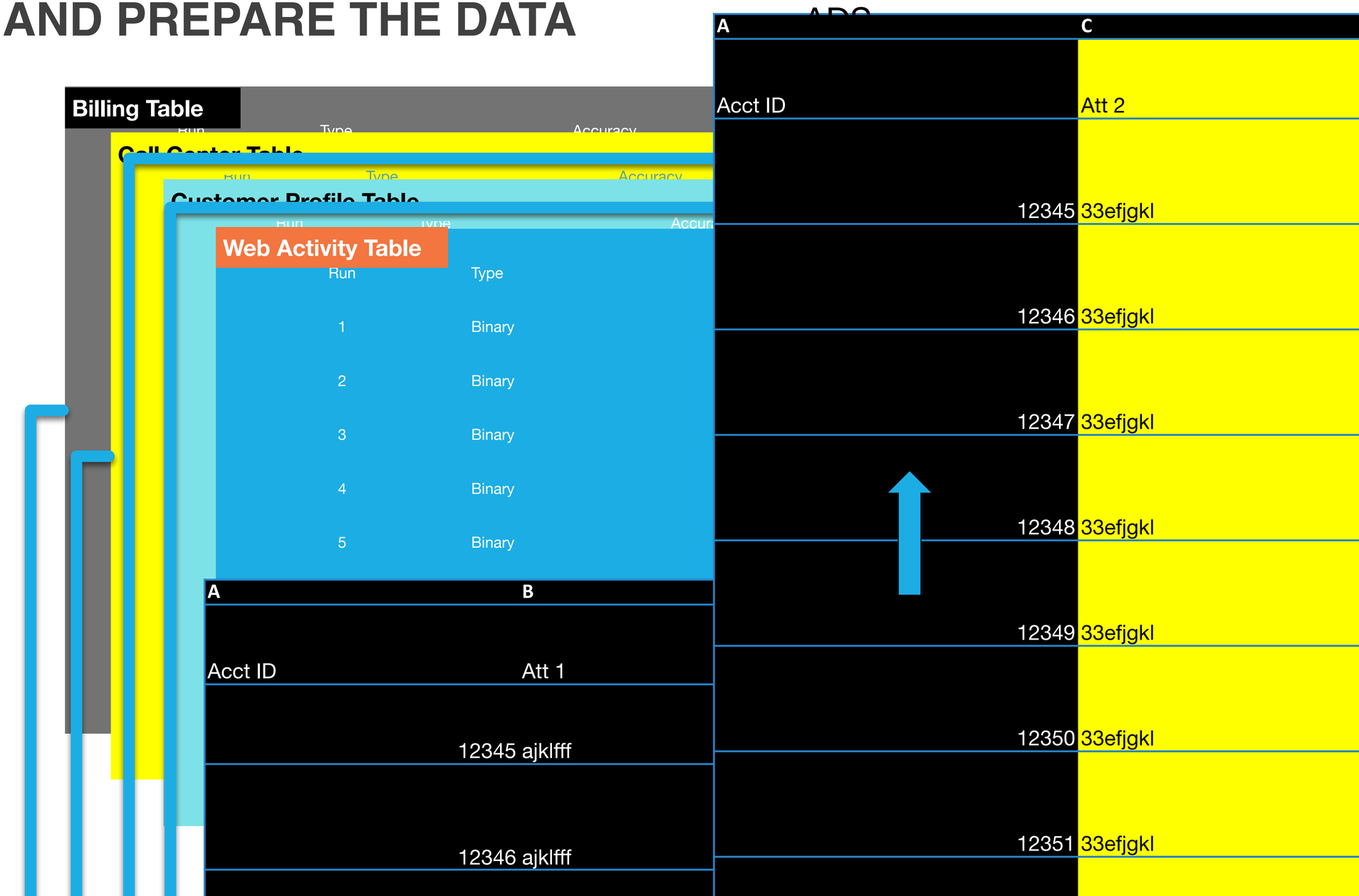
- Detailed data needed to address problem
- Where is data?
- Minimum data needed for MVP*
- Strategy to collect all data
- Join data sources
- ETL# tools, processes, execution
- Result: Raw Analytic Data Set or View

Source Data

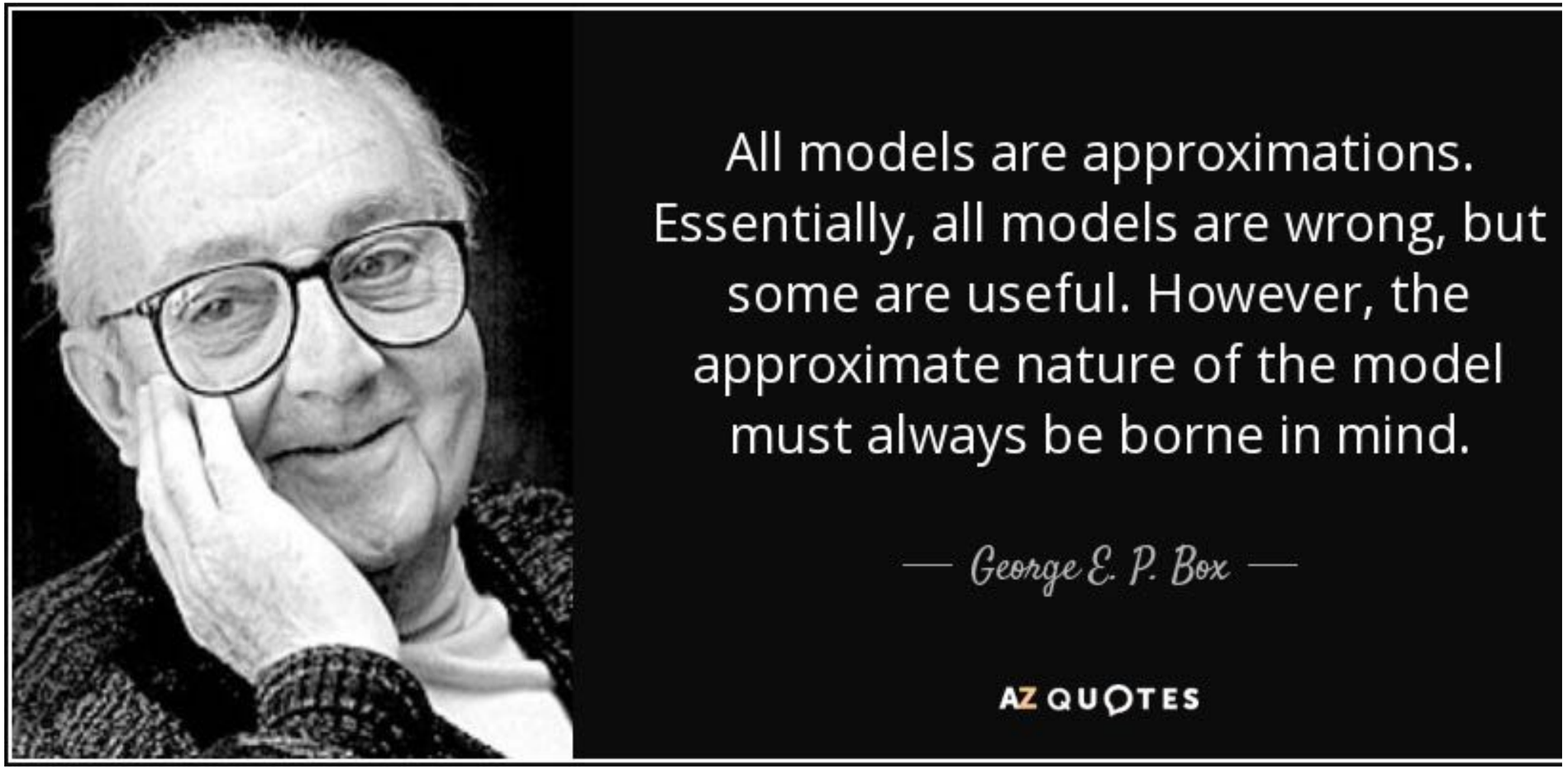


3. UNDERSTAND AND PREPARE THE DATA

- Clean Data (missing / corrupted values)
- Outlier analysis
- Explore Data (EDA*)
- Visualize Data
- Transform Data (e.g., Normalize)
- Generate Features
- Create Analytic Data Set (ADS)

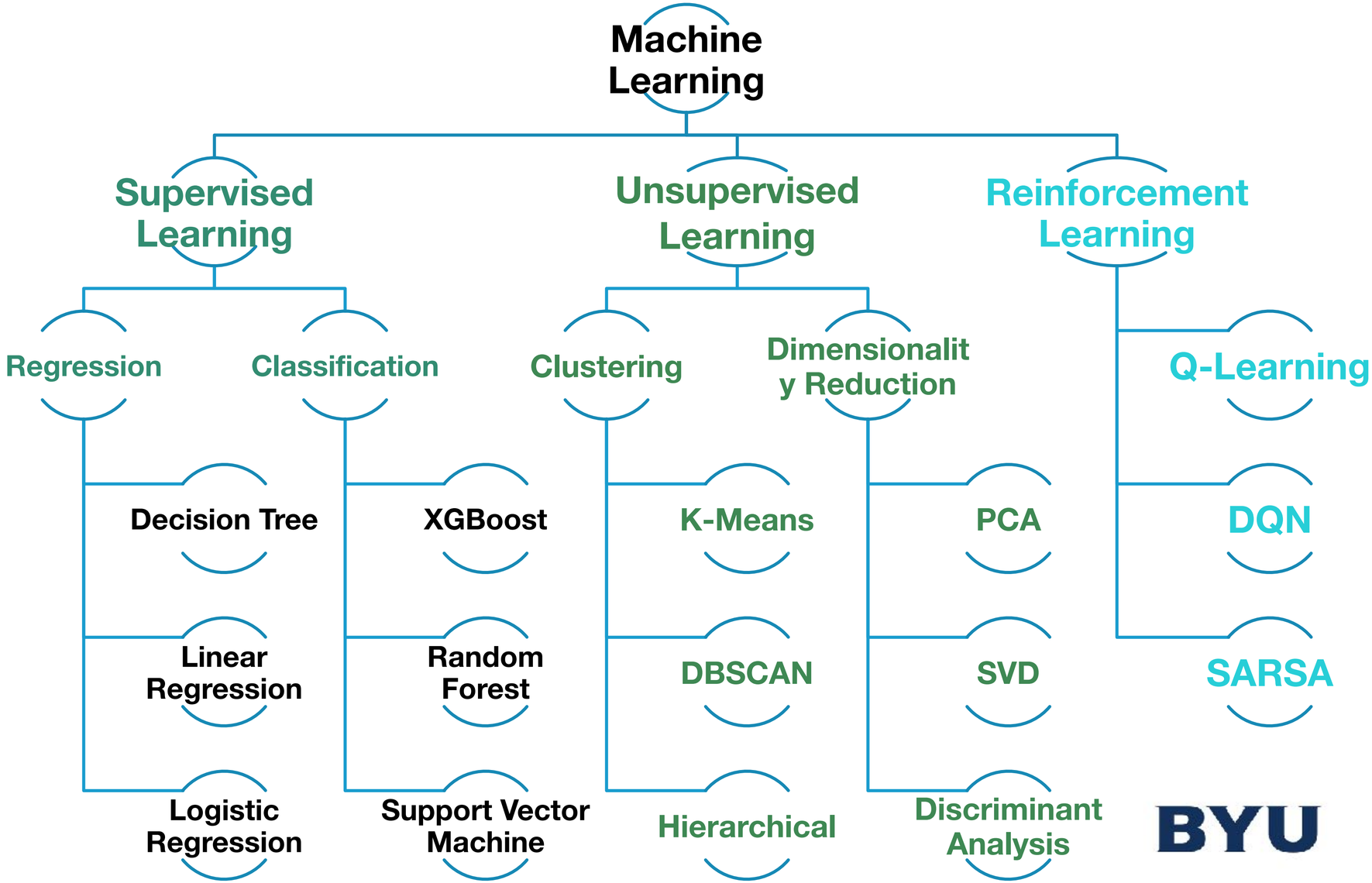


4. Build the Model



4. BUILD THE MODEL

- Select algorithms
- Define model metrics
- Optimize hyper-parameters
- Create cross-validation training & testing sets
- Run & Validate models
- Select the best-performing model set
- Interpret Model output
- Iterate to improve performance



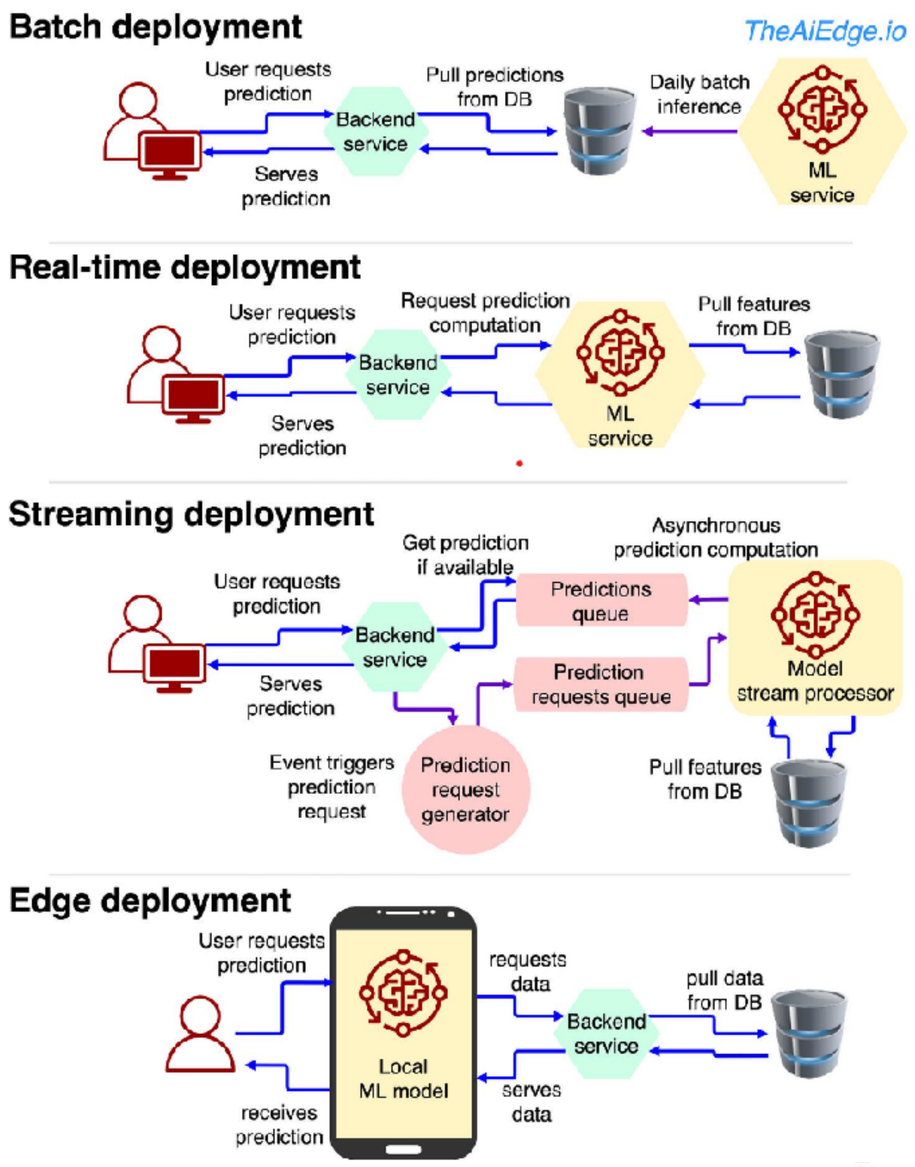
BYU

5. DEPLOY INTO PRODUCTION

Deploy

- Define requirements
- ID business process, tools, & org changes
- Determine deployment approach:
  - Batch Table
  - REST API Endpoint
  - Container
  - Pickle File, PMML, ONYX
  - Web Service
  - Stand alone application
- Develop app S/W
- Integrate with other systems
- Prototype
- Pilot
- Full scale implementation

Deployment Patterns



6. MANAGE AND MEASURE MODELS IN PRODUCTION

Manage

- Version Control
- Collect data about all model performance
- Store model code
- Track training and testing sets
- Track all model experiments
- Trigger retraining when models drift
- Run A/B testing
- Evaluate model outputs against business goals

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

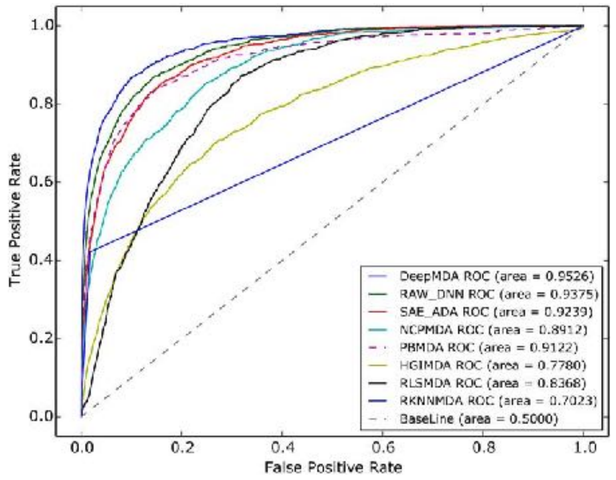
$$\text{Precision} = \frac{TP}{TP + FP}$$

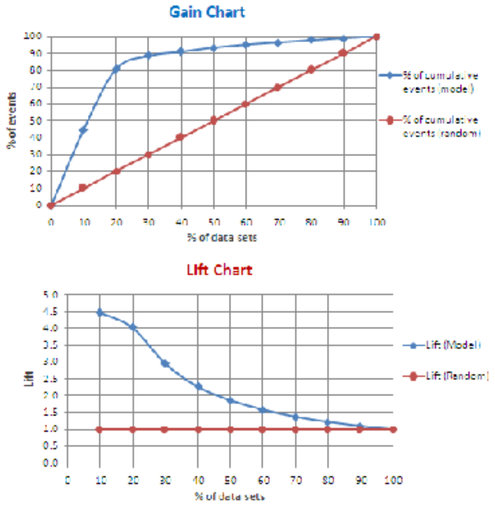
$$\text{Recall} = \frac{TP}{TP + FN}$$

Model Repository
- Model Metadata
- Model Scoring Metadata
- Model Store

Model Management
- Version Control
- Automated Testing
- Release Management

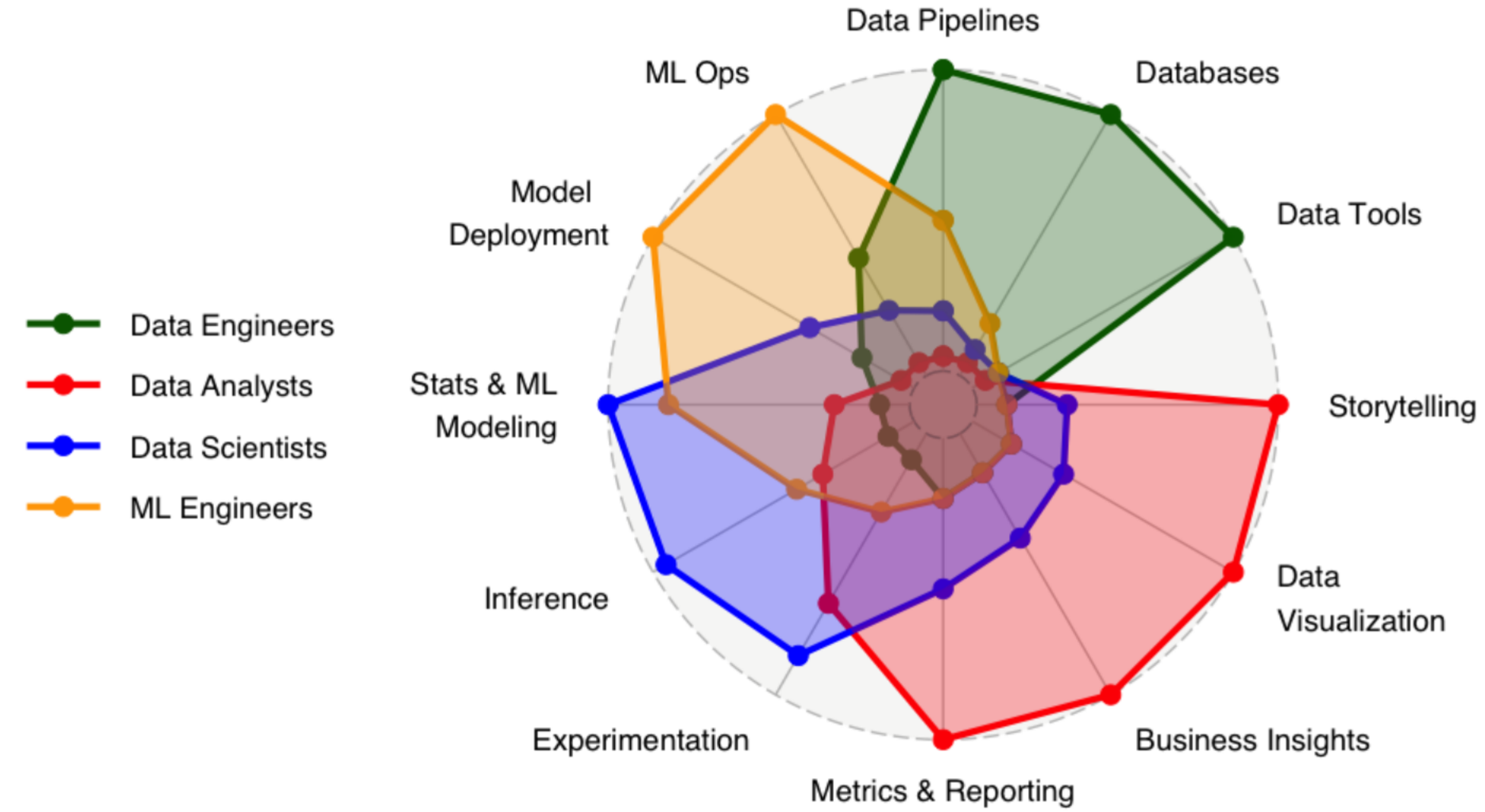
Model Execution
- Training
- Evaluation
- Scoring





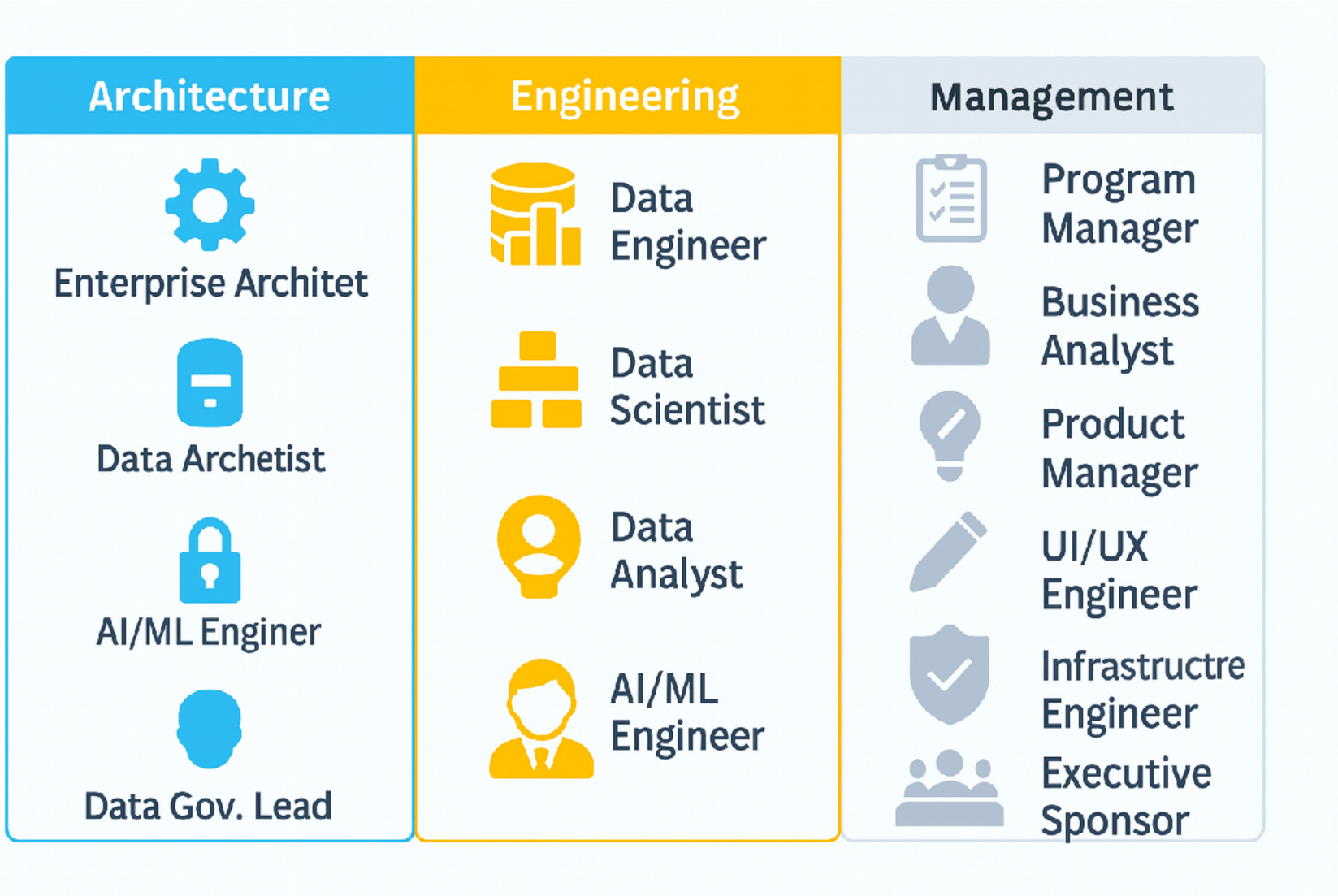
Lift Charts

Spider chart of relative skills for key data roles



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

Roles on a Data Science Application Project

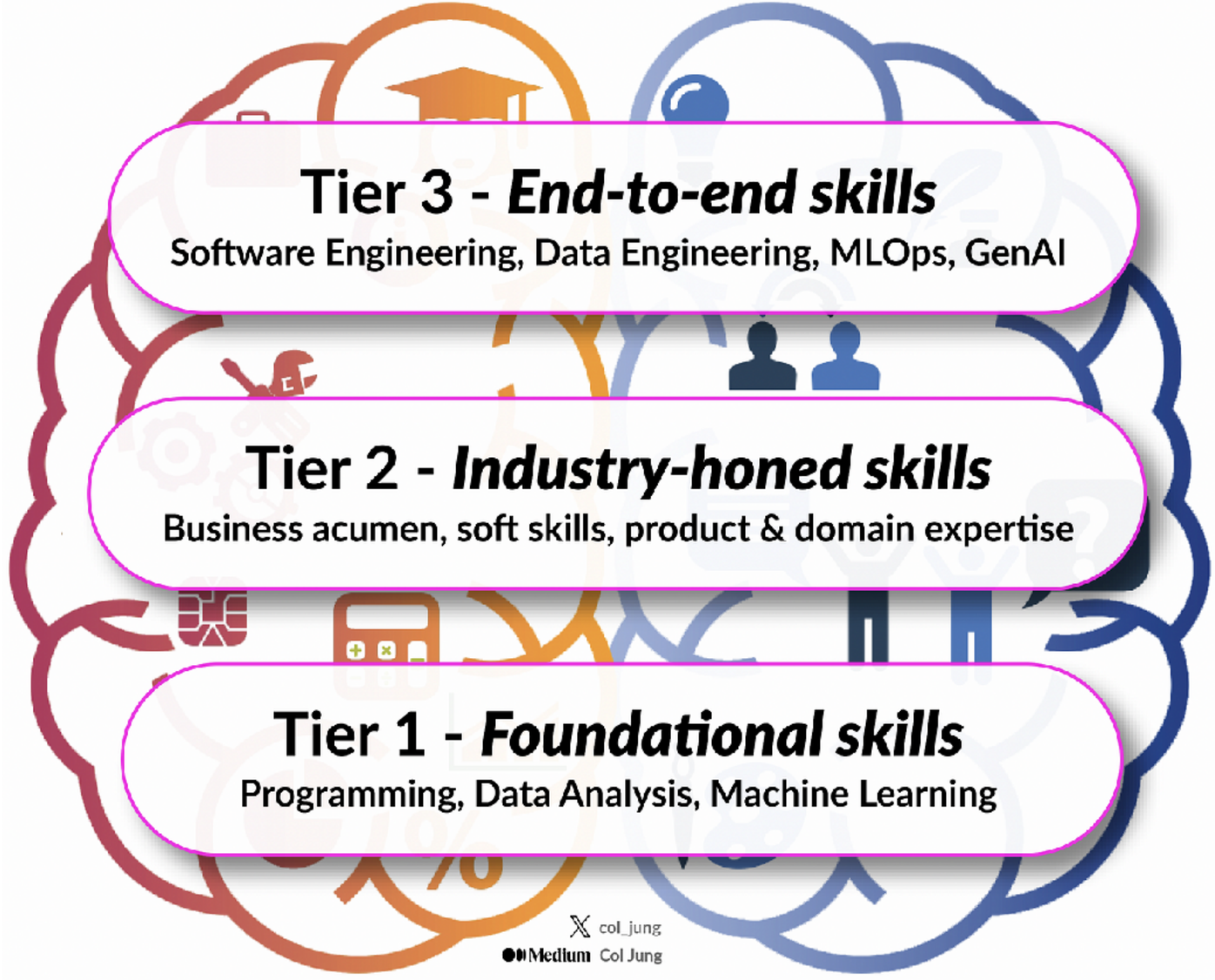


Extended Roles on a Data Science Application Project

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.

Skill Sets for Full Stack Data Scientists

"Evolution of Data Science: New Age Skills for the Modern End-to-End Data Scientist," by Col Jung, Medium, July 23, 2024.



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.

We have just one job…

BYU’s role is to build Christ-like leaders

What did Christ do?

- **Heal the Sick** → Improve Health Outcomes
- **Feed the Hungry** → Reduce Scarcity and Waste
- **Raise the Marginalized** → Expand Access and Inclusion
- **Defend the Vulnerable** → Protect Against Exploitation
- **Calm Chaos** → Reduce Risk and Instability
- **Heal Without Judgment** → Deliver Help Without Stigma
- **Cross Boundaries** → Integrate Disconnected Systems
- **Expose Hypocrisy** → Reveal Truth in Systems
- **Empower Others** → Democratize Insight and Tools
- **Reject Coercive Power** → Build Non-Manipulative Systems
- **Judge by Fruit** → Measure Real-World Impact
- **Serve Quietly** → Prioritize Outcomes Over Recognition

Marissa Fayer, Deeplook Medical

5% of breast cancer tests are still false negatives

Taylor Sheed, Stemuli

Improving inner-city education outcomes with AI-driven video games


Health Transformer
EXPERIENCE
STARTUP
+HEALTH
CHRISTIAN DANSEREAU, PHD
Perceiv AI
Solving Alzheimers through early prediction

And you…