Georgia Data
Innovation Hub
Carl Vinson Institute of Government
UNIVERSITY OF GEORGIA

GET SMART
About Data Science

James Byars
Senior Data Scientist

No Code Zone

Learning Objectives

After this session, participants should be able to:

1. Discuss core analytical concepts from data science.
2. Explain the data science life cycle.
3. Evaluate their data science strategy to inform the recruitment and retention of data scientists.

Join @
slido.com

#1627779

slido

What industry do you come from?

ⓘ Start presenting to display the poll results on this slide.

slido

What are you hoping to get out of this session?

ⓘ Start presenting to display the poll results on this slide.

Session’s Outline

1. The Data Science Landscape
2. Formulating the Right Question
3. Knowing the Data Science Process
4. Building Effective Data Science Team

#1: Know The Landscape

Decoding Technobabble

artificial intelligence (AI).

According to John McCarthy, Stanford University Professor of Computer Science, it is "the **science and engineering of making intelligent machines**, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable."

machine learning (ML).

A discipline of AI focusing on computers (machines) using algorithms and data for pattern discovery and prediction **without being explicitly programmed**. Increasing available data and model training will improve the machine learning models’ accuracy — like the **“human” adage of “practice makes perfect”**.

Introducing ChatGPT

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.

Try ChatGPT ↗ Read about ChatGPT Plus

User: this code is not working like i expect — how do i fix it?

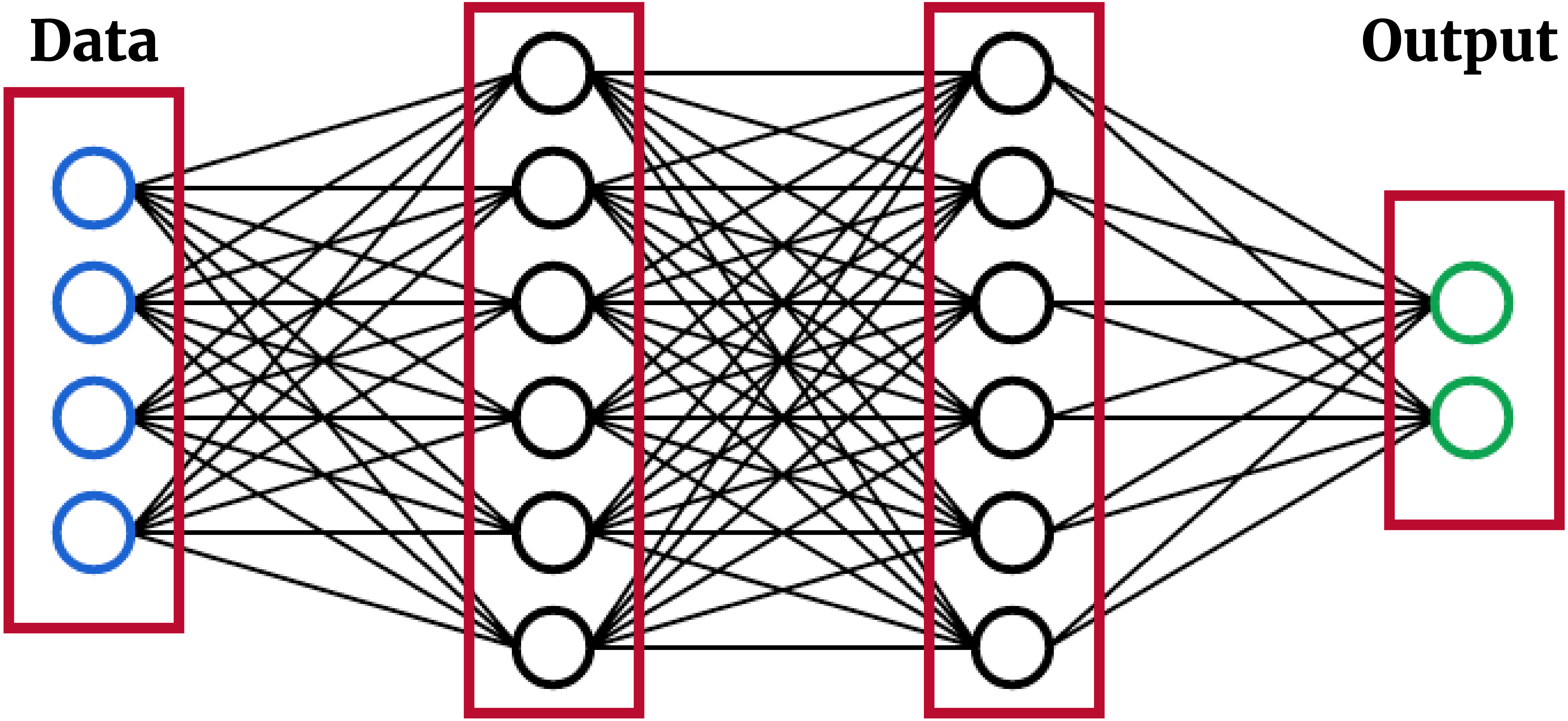
```
resultWorkerErr := make(chan error)
defer close(resultWorkerErr)
go func() {
    defer cancel()
    resultWorkerErr <- b.resultWorker(ctx)
}()

err := b.worker(ctx)
cancel()
if err == nil {
    return <-resultWorkerErr
}
return multierror.Append(err, <-resultWorkerErr)
```

ChatGPT: It's difficult to say what's wrong with the code without more context. Can you provide more information about what the code is supposed to do and what isn't working as expected? Also, is this the entire code or just a part of it?

User: just part of it — the error never surfaces. I think there is something wrong with the channel

neural network.

Also known as an artificial or simulated neural network is a popular artificial intelligence algorithm and the backbone of deep learning. Their name and structure are inspired by the **human brain** in that it mimics the way that **biological neurons send signals (information)** to one another.



Hidden Layers (“Brain Nodes”)

deep learning (DL).

Deep learning refers to multilayered “deep” neural networks (i.e., **three or more layers**). DL algorithms perform tasks repeatedly to promote model refinement for improvement in accuracy. DL algorithms depend on **vast amounts of data to drive “learning”**.

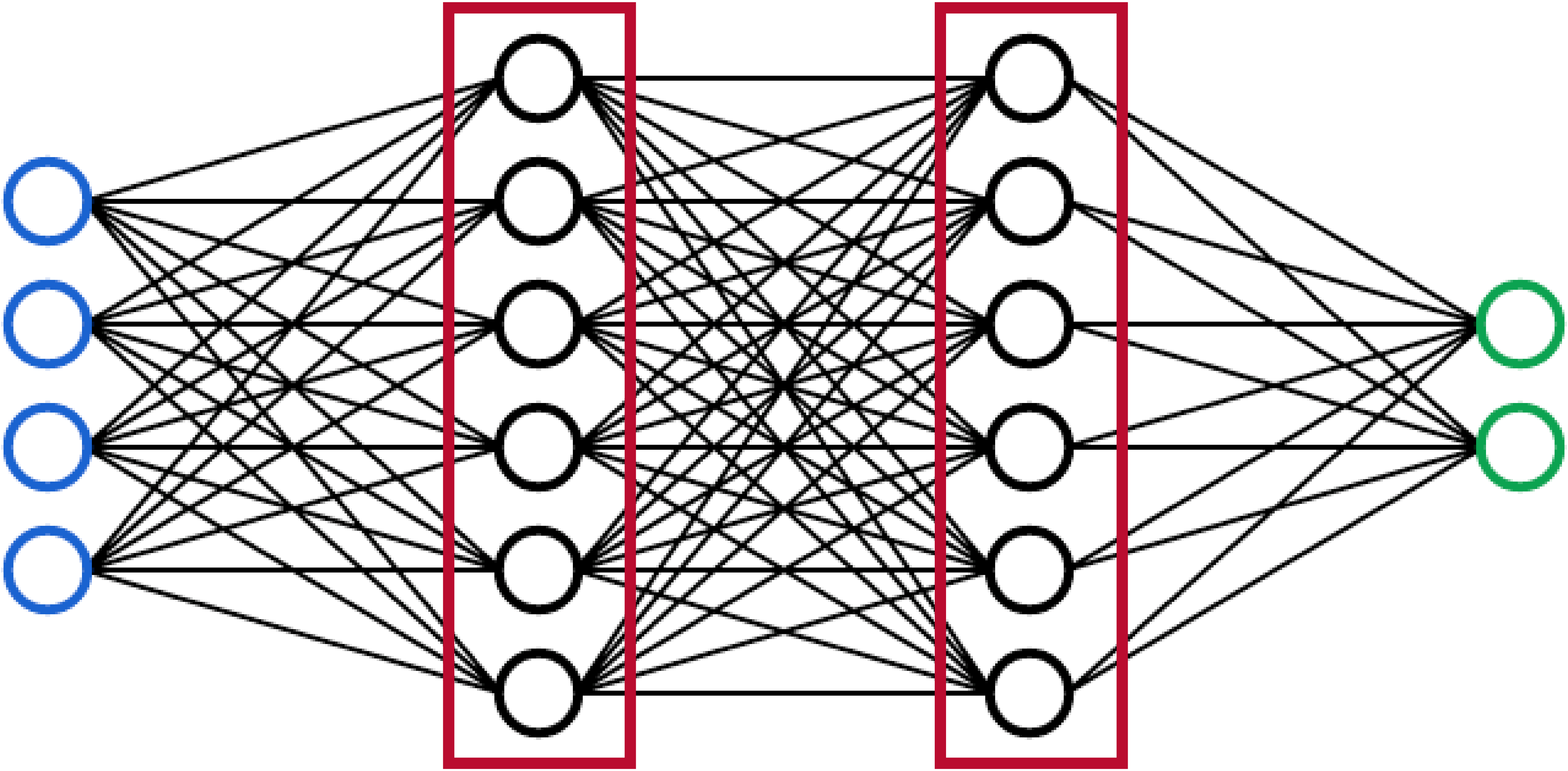
AlphaGo

Making history

AlphaGo is the first computer program to defeat a professional human Go player, the first to defeat a Go world champion, and is arguably the strongest Go player in history.

black box algorithm.

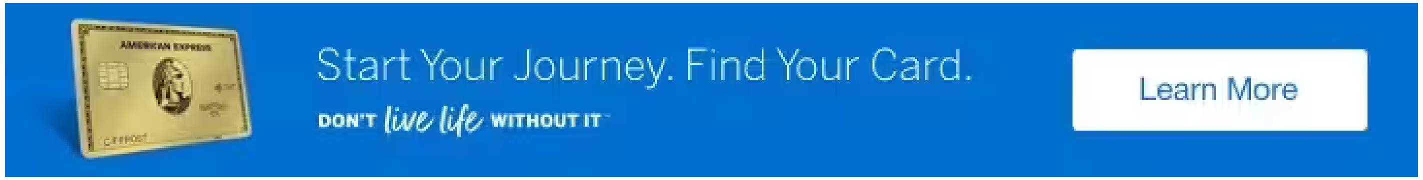
Algorithms that does not model the problem in a way which allows humans to directly state what happens for any given input. While these algorithms may be **successful for prediction, they lack explainability** (i.e., why the model made the specific prediction?).



Hidden Layers ("Brain Nodes")

predictive modeling.

A branch of machine learning that makes predictions about **future outcomes using historical data** by using algorithms from data science, artificial intelligence, machine learning, statistics, and computational processes.



How Amex Protects You Against Credit Card Fraud

American Express helps protect you from credit card fraud with its advanced fraud protection methods. Here are a few simple steps to keep you safe and alert.

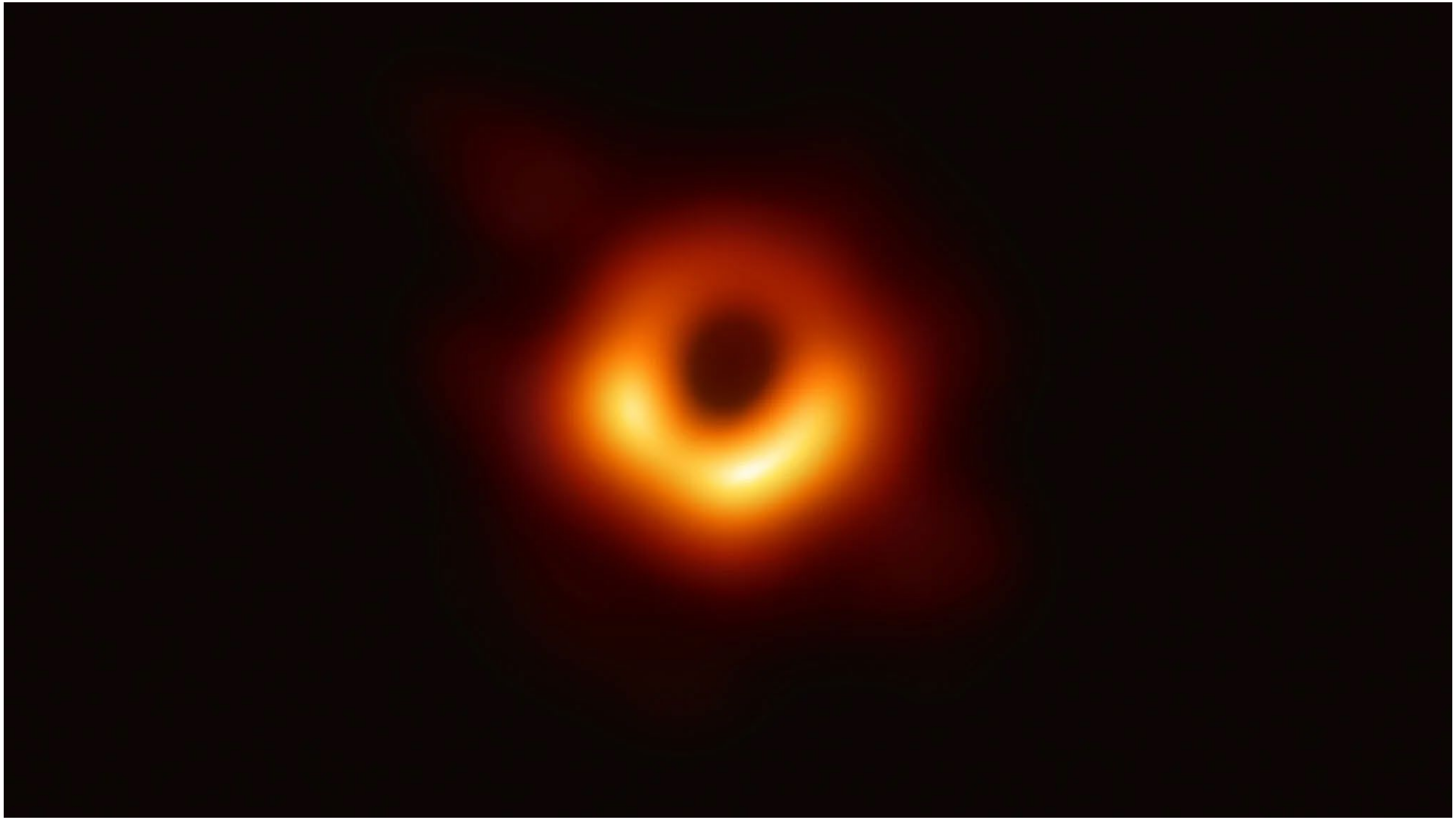
By Megan Doyle | American Express *Credit Intel* Freelance Contributor
5 Min Read | October 28, 2020 in Cards

Credit card fraud is increasingly common. Americans pay by credit card more and more, and fraudsters constantly evolve their schemes to steal personal information. But credit card fraud protection techniques are always improving too. American Express was an early developer of artificial intelligence expert systems for fraud detection,[1] and has remained on the cutting-edge of fraud protection methods ever since.

From basic account security to sophisticated machine learning models that analyze thousands of data points in real time, here's a close look at how American Express fraud protection techniques work behind the scenes, 24/7/365, to help keep you safe and secure.

big data.

Data that is **too large and complex** to fit on a single computer. Conventional strategies like traditional database storage and management approaches are computationally unproductive and ill-equipped.



data science (DS).

A discipline that works with and analyzes large volumes of data to provide actionable intelligence for data-informed decision making. DS is a **substantively and computationally interdisciplinary field** solve problems. Data science integrates concepts and approaches from **traditional statistics, business analytics, artificial intelligence, data architectures, storytelling and journalism, and others**.

#2: Know Your Question

Questions Drive The Algorithms

What is an algorithm?

algorithm.

Broadly: a step-by-step procedure for solving a problem or accomplishing an objective. **Repeatable sets of instructions which people or machines can use to process data into insight.**

Descriptive Questions

1. Uses algorithms such as visual and statistical approaches (e.g., frequencies, central tendencies, and variability).

2. Use data to provide a **quantitative summary of certain features or phenomenon**.

3. The limitation with descriptive analytics is that it is focuses only on historical events and **cannot be used to draw inferences or predictions**.

Cost of Attendance

Student Loan Outcomes

Time to Degree Cost Calculator

Use this page to compare costs per year between two USG colleges/universities. Choose the residency status below and then select two colleges/universities from the dropdown menus on the right and the table will adapt.

Which schools should I compare?

Select Residency Status

- (●) In-State
- ( ) Out-of-State

How much is the annual cost for students?

	College/University 1	College/University 2
	Albany State University	Clayton State University
Tuition & Mandatory Fees	$6,500	$6,084
Total Cost of Attendance (Tuition and fees, room and board, books and supplies)	$20,525	$17,456
Average Scholarships and Grants	$6,166	$7,368
Net Price (Total Cost of Attendance - Average Scholarships and Grants)	$14,359	$10,088
Net Price Calculator	Click Here	Click Here

74%

Microsoft Power BI

1 of 3

Georgia Data Innovation Hub
Carl Vinson Institute of Government
UNIVERSITY OF GEORGIA

Inferential Questions

1. Uses algorithms to estimate conclusions (inferences) using a small sample of data in order to **generalize to a larger population**.

2. Useful for testing hypothesis about how your data will reflect larger data collections.

3. Examines the **relationships between variables** in a sampled dataset larger population.

High School GPAs and ACT Scores as Predictors of College Completion: Examining Assumptions About Consistency Across High Schools

Elaine M. Allensworth ✉ and Kallie Clark View all authors and affiliations

Volume 49, Issue 3 | https://doi.org/10.3102/0013189X20902110

Abstract

High school GPAs (HSGPAs) are often perceived to represent inconsistent levels of readiness for college across high schools, whereas test scores (e.g., ACT scores) are seen as comparable. This study tests those assumptions, examining variation across high schools of both HSGPAs and ACT scores as measures of academic readiness for college. We found students with the same HSGPA or the same ACT score graduate at very different rates based on which high school they attended. Yet, the relationship of HSGPAs with college graduation is strong and consistent and larger than school effects. In contrast, the relationship of ACT scores with college graduation is weak and smaller than high school effects, and the slope of the relationship varies by high school.

Predictive Questions

1. Uses algorithms, like statistical modeling, artificial intelligence, and machine learning, to **generate educated forecasts** about future actions **based upon historical data**.

2. Predictive questions often focus on forecasting numerical outcomes (e.g., future housing prices) and categories (e.g., will customer purchase a ticket [yes/no]?)

AI, MACHINE LEARNING & RESEARCH

Building the Neural Zestimate

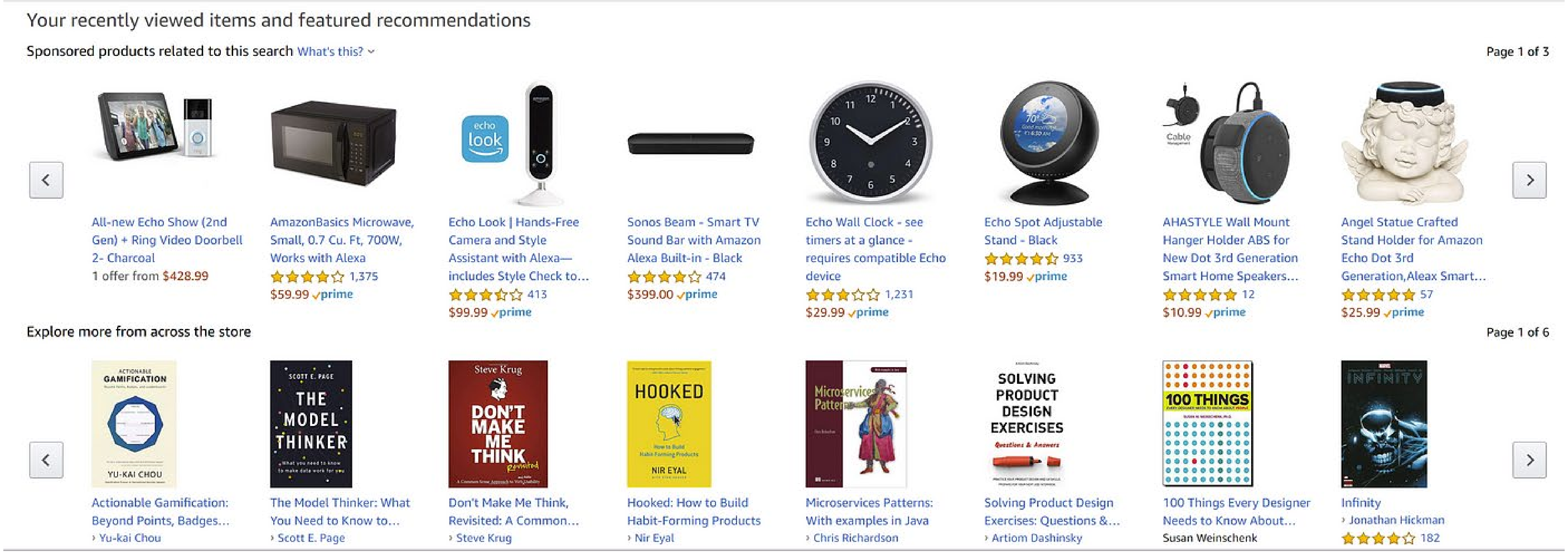
Reid Johnson • FEB 23 2023

Share

Prescriptive Questions

1. Uses algorithms to determine an **optimal course of action (i.e., recommendation) to maximize an intended outcome**.

2. Combines predictive, inferential, and descriptive methods with optimization, operational methods, and business logic to provide data-informed guidance on **evaluating the rewards and consequences of specific decisions**.


Your recently viewed items and featured recommendations
Sponsored products related to this search What's this?
Page 1 of 3
All-new Echo Show (2nd Gen) + Ring Video Doorbell 2- Charcoal
1 offer from $428.99
AmazonBasics Microwave, Small, 0.7 Cu. Ft, 700W, Works with Alexa
1,375
$59.99 prime
Echo Look | Hands-Free Camera and Style Assistant with Alexa—includes Style Check to...
413
$99.99 prime
Sonos Beam - Smart TV Sound Bar with Amazon Alexa Built-in - Black
474
$399.00 prime
Echo Wall Clock - see timers at a glance - requires compatible Echo device
1,231
$29.99 prime
Echo Spot Adjustable Stand - Black
933
$19.99 prime
AHASTYLE Wall Mount Hanger Holder ABS for New Dot 3rd Generation Smart Home Speakers...
12
$10.99 prime
Angel Statue Crafted Stand Holder for Amazon Echo Dot 3rd Generation,Alexa Smart...
57
$25.99 prime
Explore more from across the store
Page 1 of 6
Actionable Gamification: Beyond Points, Badges...
Yu-kai Chou
The Model Thinker: What You Need to Know to...
Scott E. Page
Don't Make Me Think, Revisited: A Common...
Steve Krug
Hooked: How to Build Habit-Forming Products
Nir Eyal
Microservices Patterns: With examples in Java
Chris Richardson
Solving Product Design Exercises: Questions &...
Artiom Dashinsky
100 Things Every Designer Needs to Know About...
Susan Weinschenk
Infinity
Jonathan Hickman
182

Casual Questions

1. Casual analysis, also known as explanatory analysis, uses algorithms to explore the **cause-and-effect relationship** between a well-defined treatment (e.g., event, phenomenon, or circumstances.

2. To achieve non-biased results, data scientists must be aware and control confounding variables which are often hidden (e.g., not collected or considered).

Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program

Alberto ABADIE, Alexis DIAMOND, and Jens HAINMUELLER

Building on an idea in Abadie and Gardeazabal (2003), this article investigates the application of synthetic control methods to comparative case studies. We discuss the advantages of these methods and apply them to study the effects of Proposition 99, a large-scale tobacco control program that California implemented in 1988. We demonstrate that, following Proposition 99, tobacco consumption fell markedly in California relative to a comparable synthetic control region. We estimate that by the year 2000 annual per-capita cigarette sales in California were about 26 packs lower than what they would have been in the absence of Proposition 99. Using new inferential methods proposed in this article, we demonstrate the significance of our estimates. Given that many policy interventions and events of interest in social sciences take place at an aggregate level (countries, regions, cities, etc.) and affect a small number of aggregate units, the potential applicability of synthetic control methods to comparative case studies is very large, especially in situations where traditional regression methods are not appropriate.

KEY WORDS: Observational studies; Proposition 99; Tobacco control legislation; Treatment effects.

Descriptive	**What happened?**
Inferential	**Why could it have happened?**
Predictive	**What will happen next?**
Prescriptive	**How can we make it happen going forward?**
Causal	**What is the impact of what we made happen?**

Descriptive	How many undergraduates have graduated with a Bachelor's in Data Science or related field since 2010?
Inferential	Is their statistical evidence to conclude a relationship between class size and students' retention of course content?
Predictive	Does Georgia have enough graduates in teaching and nursing over the next five years to fulfill current and future openings?
Prescriptive	What majors should our institution recommend to prospective students based on their aptitude and interest?
Causal	What is the impact on increased peer-to-peer tutoring on our institution's four-year graduation rate?

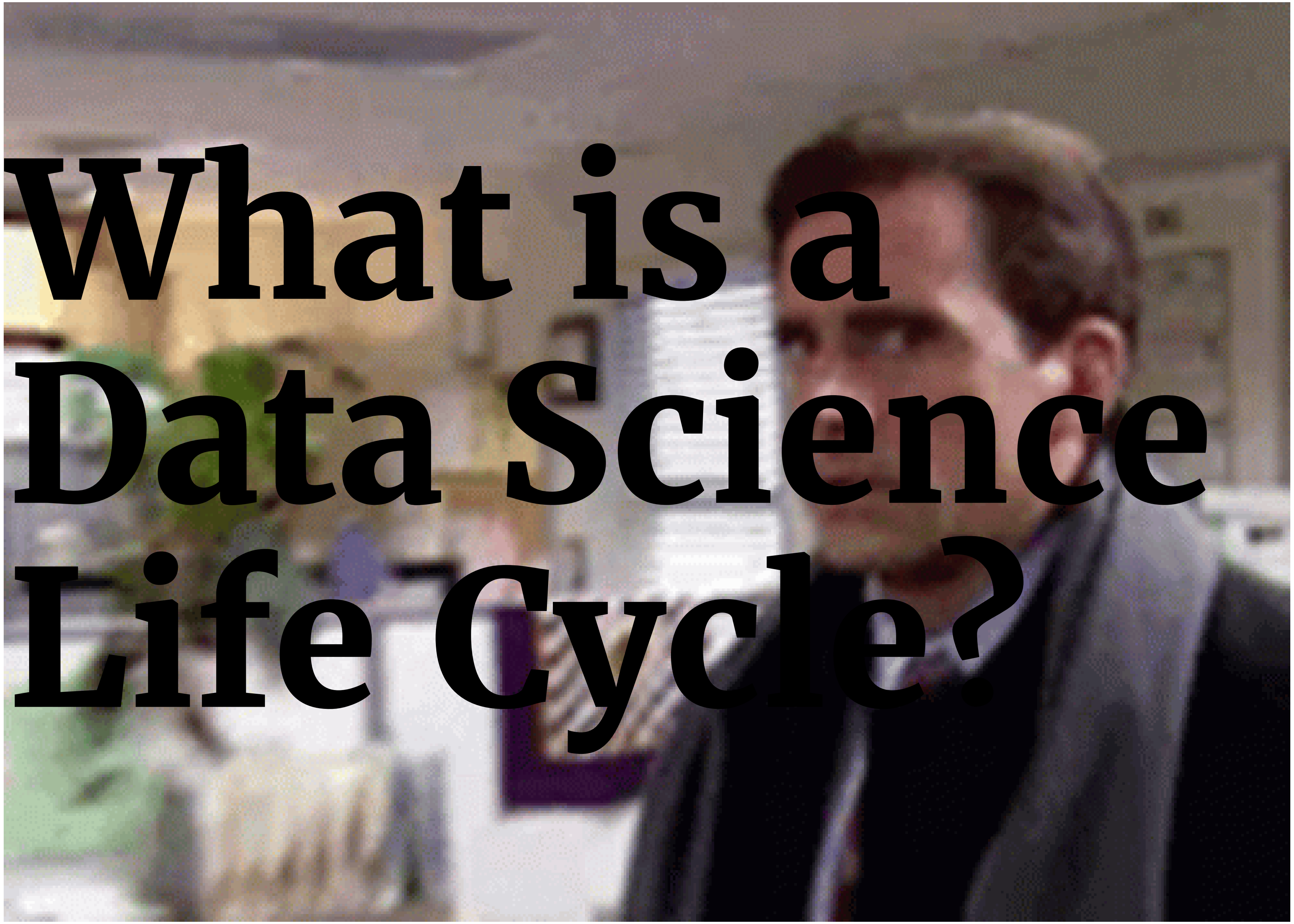
Descriptive	Knowledge of basic statistical measures (e.g., mean, medians, modes, variance, and rolling averages) and data management
Inferential	Knowledge of basic statistical procedures (e.g., linear regression), basic probability and mathematics, and data management
Predictive	Machine learning and computer/data science techniques, advanced probability and mathematics, and linear algebra
Prescriptive	Machine learning and computer/data science techniques, advanced probability and mathematics, and linear algebra
Causal	Knowledge of quasi-experimental methods and randomized controlled trial, advanced probability and mathematics, and graph theory

Descriptive	1) Basic statistical and computational programming language knowledge and training. Alternatively, 2) spreadsheets, database management, and business intelligence software
Inferential	1) Intermediate statistical and computational programming language knowledge and training. Alternatively, 2) spreadsheets and business intelligence software
Predictive	Advanced statistical and computational programming language knowledge and training
Prescriptive	Advanced statistical and computational programming language knowledge and training
Causal	Advanced statistical and computational programming language knowledge and training

#3: Know Your Process

Data Science Life Cycles

What is a Data Science Life Cycle?



Data Science Life Cycles

1. General set of core technical and substantive steps vital for completing a data science project.

2. Often discussed as a linear process, but it should, however, be iterative. It is healthy and normal to move back and forth between steps.

3. Require a comprehensive set of statistical, programming, and illustrative skills, software, and methods.

4. Time and resources spent on specific steps will differ from project to project.

O • S • E • M • N

Pronounced AWESOME

O S E M N

O	S	E	M	N
Obtain Data	Scrub Data	Explore Data	Model Data	Interpret Results

O

Obtain Data

<u>Objectives</u>

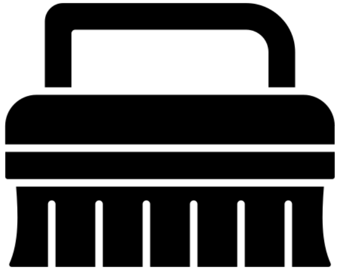
Extract data from relevant sources such as:

1. Relational and non-relational databases.
2. Single data tables (i.e., CSV and XLSX flat files).
3. Application programming interfaces (APIs).
4. The scraping of web pages.
5. Data gathered through surveys and experiments.

<u>Skills Required</u>

1. Database Management: PostgreSQL and Microsoft SQL Server.
2. Querying: Query Languages (e.g., SQL and GraphQL).
3. Data Retrieval: Scripting Languages (e.g., Java, Python, and R).
4. Distributed Storage: Hadoop and Spark.

S



Scrub Data

Objectives

Examining your data by:

1. Identifying outliers and errors.
2. Examining missing values.
3. Understanding your data's metadata.

Tidying up your data by:

1. Removing corrupted and duplicated records.
2. Accounting for anomalies and outliers.
3. Parsing categorical and string variables.
4. Replace (impute) missing records.
5. Accurately joining data sources through valid keys.

Skills Required

1. Programming: Scripting Languages (e.g., Python, R, and SAS).
2. Methods: Statistical imputation and advanced data management through programming and their statistical/computing packages.

E

Explore Data

Objectives

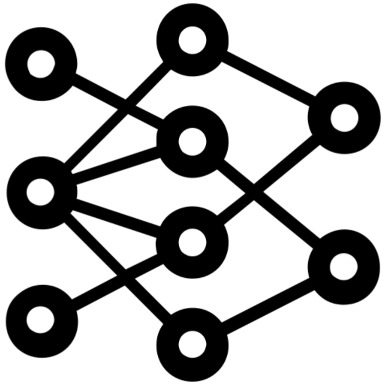
Understanding patterns with your data and your data's value by:

1. Exploring data through statistics and visuals.
2. Testing and extracting relevant features (variables).
3. Performing feature (variable) engineering.
4. Consider dimensionality reduction of dataset.

Skills Required

1. Dynamic Visualization: PowerBI, Qlik Sense, and Tableau.
2. Static Visualization: Scripting Languages (e.g., Python, R, and SAS).
3. Methods: Descriptive statistics (i.e., correlation and central tendency) and inferential statistics (e.g., significance tests).
4. Approaches: Dimensionality reduction through methods like cluster analysis and principal components analysis.
5. Tools: Scripting Languages (e.g., Python, R, and SAS).

M



Model Data

Objectives

Translating data into insights by:

1. Identifying the applicable modeling strategy.
2. Identifying the appropriate algorithm(s).
3. Building a scalable and stable statistical model(s).
4. Tuning your statistical model(s).
5. Appropriately evaluating said statistical model(s) and repeat steps.

Skills Required

1. Methods: Conditional on organizational question
2. Requirements: Strong understanding of linear algebra, calculus, statistics, and probability.
3. Tools: Scripting Languages (e.g., Python, R, and SAS).
4. Bonus: Distributed, cloud, and GPU computing.

N

Interpret Results

Objectives

Communicating your findings by :

1. Knowing your audience.
2. Knowing your purpose.
3. Knowing the right medium.
4. Relating findings to actionable organizational decisions.
5. Monitoring and evaluating the impact of those decisions.

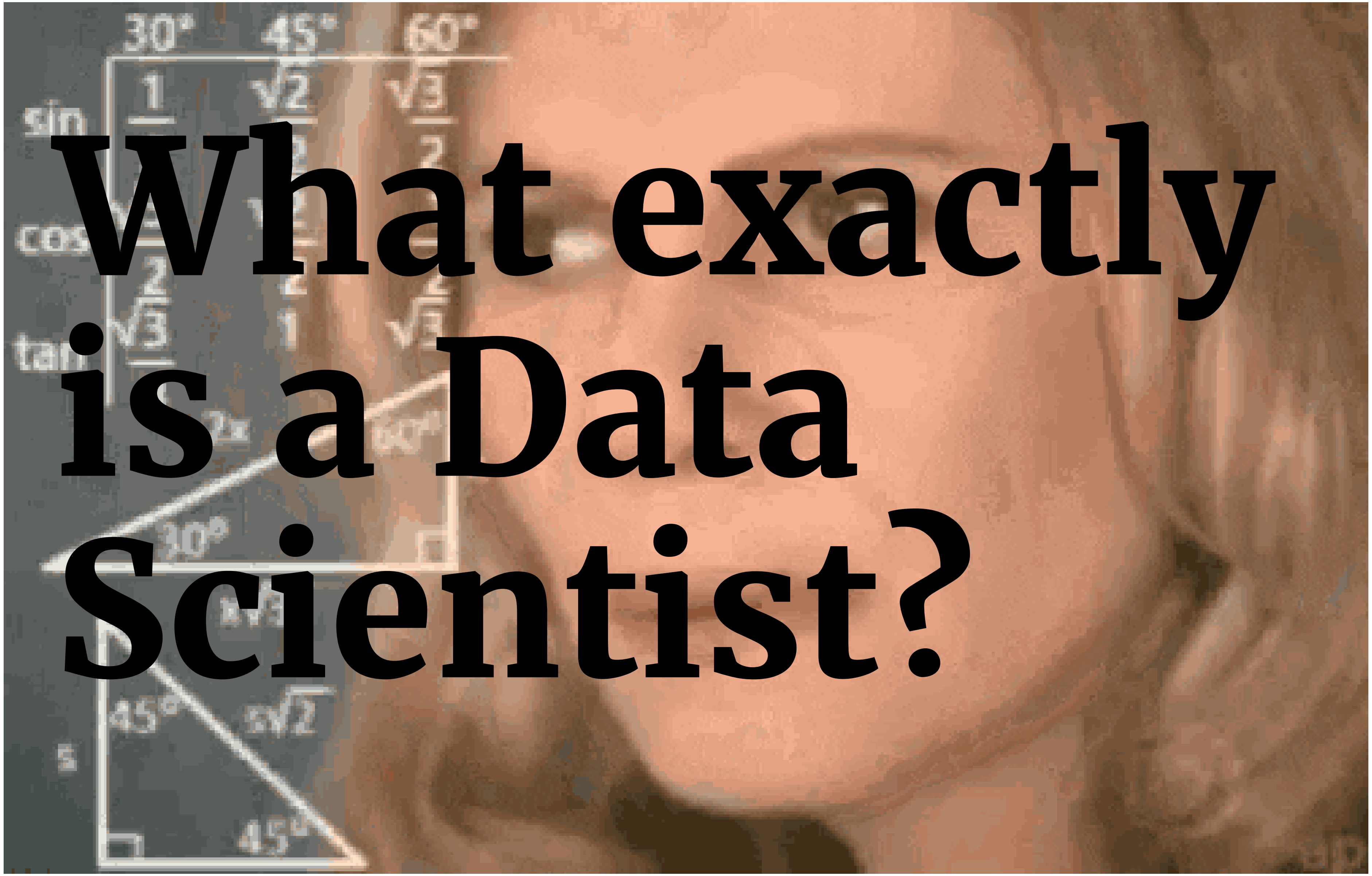
Skills Required

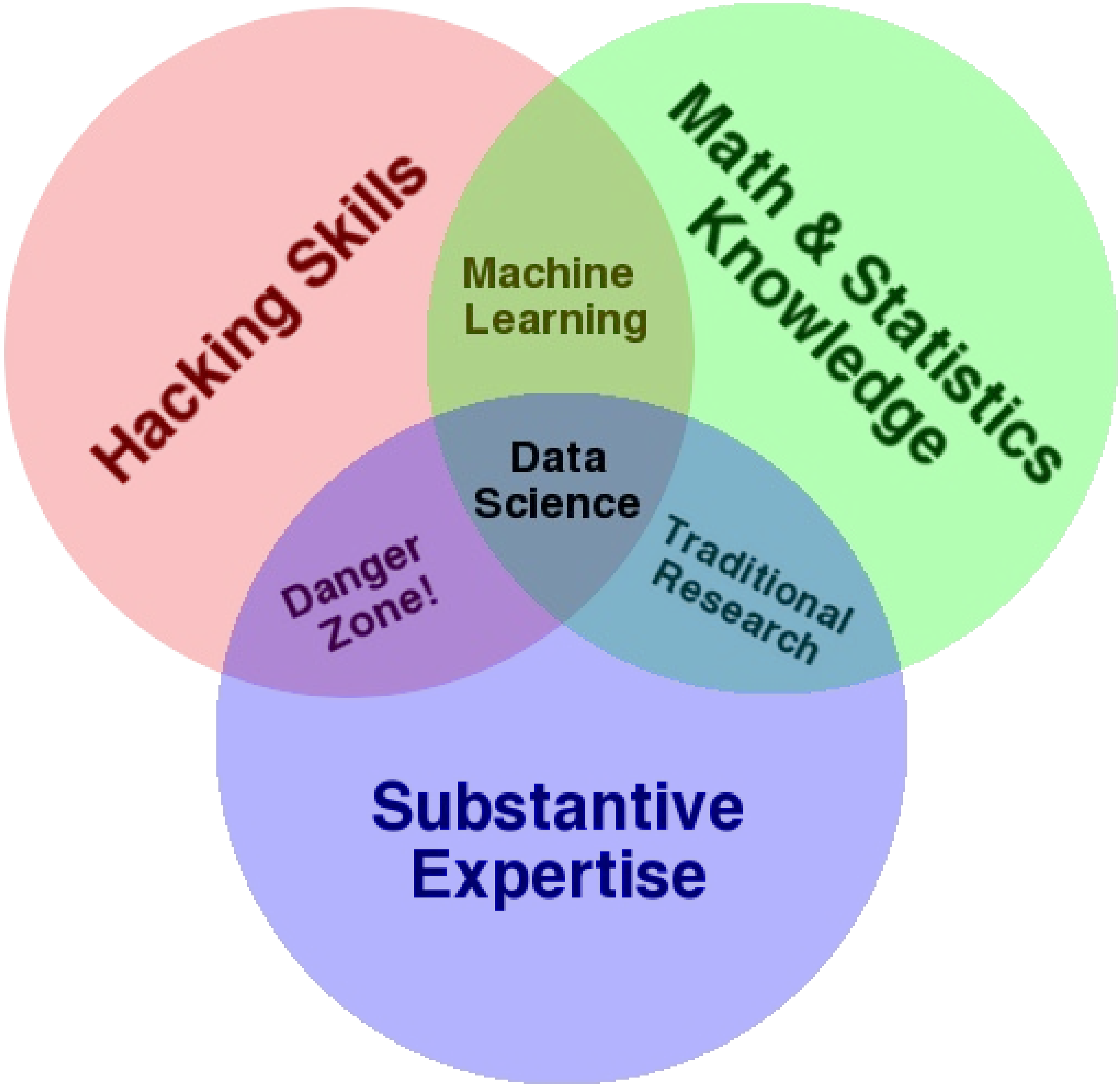
1. Dynamic Visualization: PowerBI, Qlik Sense, and Tableau.
2. Static Visualization: Scripting Languages (e.g., Python, R, and SAS).
3. Clean and concise business writing skills
4. Strong public speaking skills

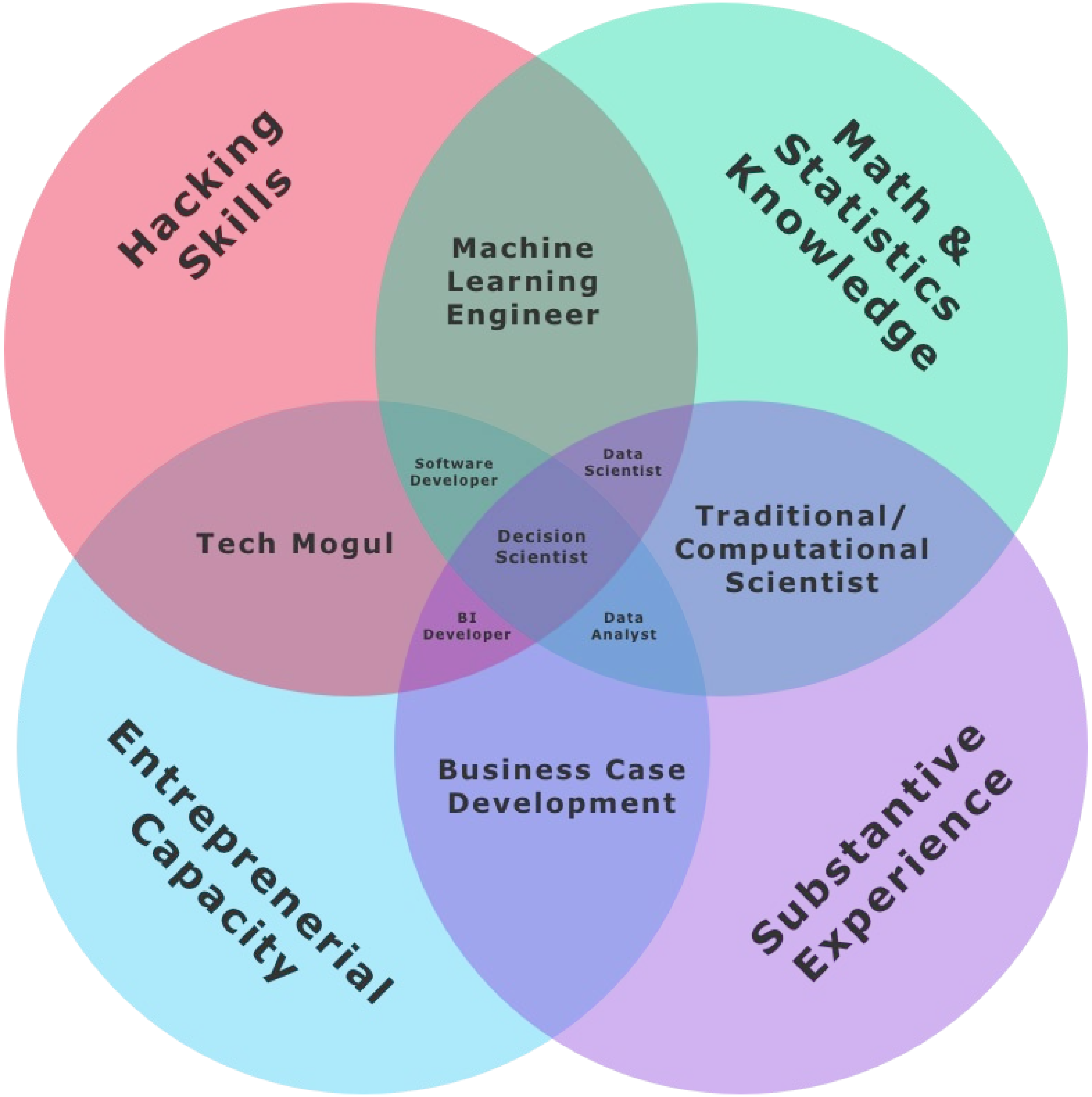
#4: Know Your Team

Data Scientists

What exactly is a Data Scientist?



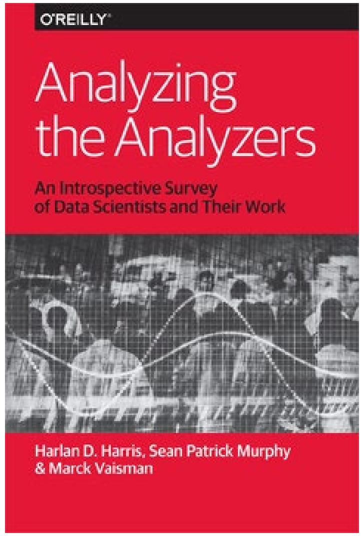

Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise


Hacking Skills
Math & Statistics Knowledge
Machine Learning Engineer
Software Developer
Data Scientist
Tech Mogul
Decision Scientist
Traditional/ Computational Scientist
BI Developer
Data Analyst
Business Case Development
Entrepreneurial Capacity
Substantive Experience

Data Science Archetypes

According to Harris, Murphy, and Vaisman (2013) there are four main archtypes of data scientists:

- Data Businesspeople
- Data Developers
- Data Researchers
- Data Creatives



Harris, H., Murphy, S., & Vaisman, M. (2013). Analyzing the analyzers: An introspective survey of data scientists and their work. O'Reilly Media, Inc. https://www.oreilly.com/content/analyzing-the-analyzers/

data businesspeople.

Data scientists who focus on project management, leading and managing data scientists, assessing the scalability and sustainability of data science projects, and **emphasizing a project's return on investment**.

data developers.

Data scientists who focus primarily on the **back-end, technical requirements** of the data science life cycle such as the data infrastructure, the extraction, transformation, and loading (ETL) process, scrubbing data, and feature engineering.

data researchers.

Data scientists who have **extensive and intensive academic training and experience** in the physical (e.g., statistics and computer science) or social (e.g., econometrics) sciences heavy in mathematics and statistics. They are often **designing or integrating cutting-edge and complex methods** to model complex social and business processes.

data creatives.

Data scientists that can tackle all aspects of the data science life cycle. They **have a breadth of the needed technical and substantive skills**, but not necessarily the expertise for deeper analysis. They work with a broad range of statistical and visual platforms and consider themselves as a **"jack of all trades"**, **artists**, or **hackers**.

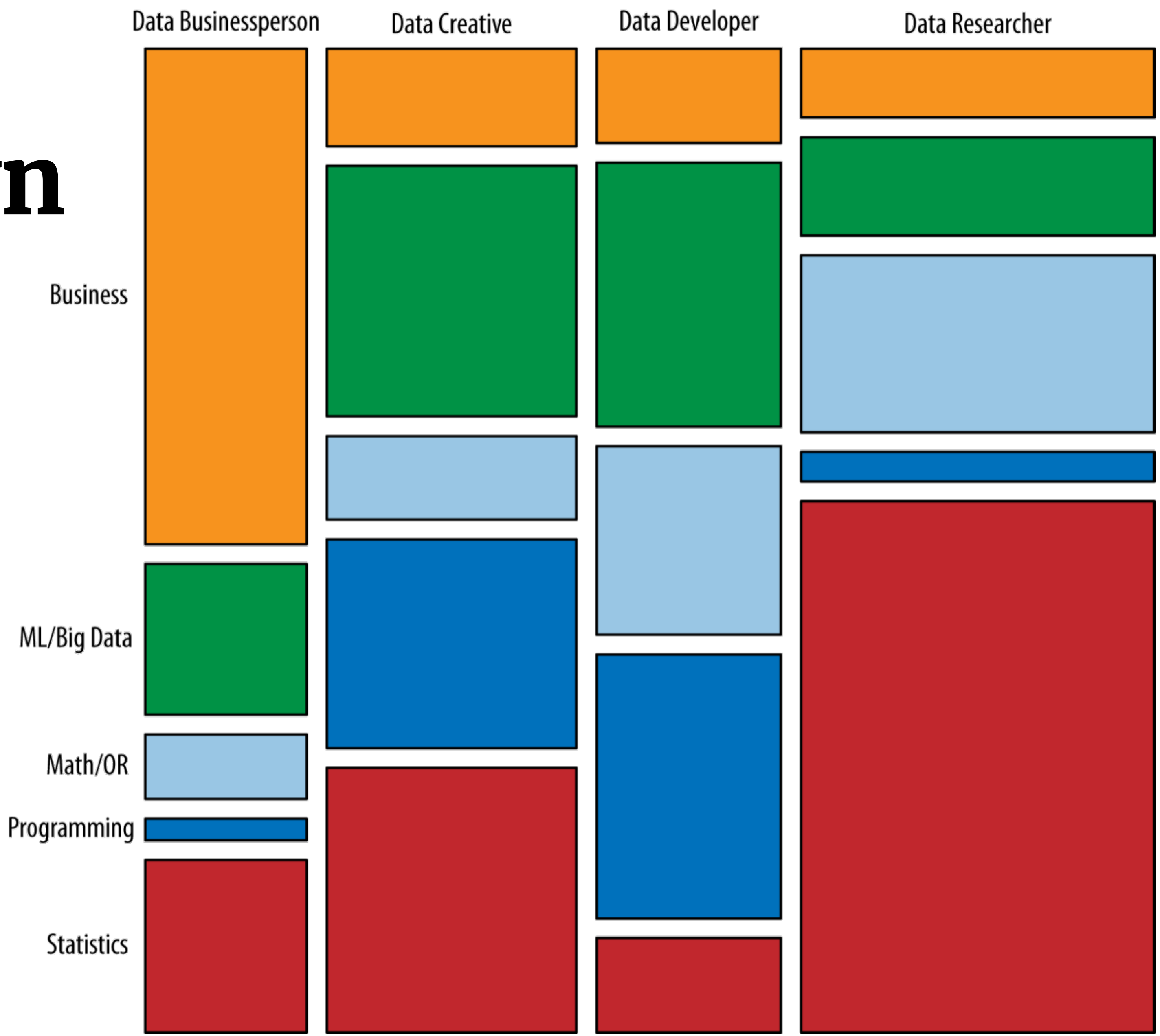
Aliases of Data Scientists

Data Developer	Developer	Engineer	
Data Researcher	Researcher	Scientist	Statistician
Data Creative	Jack of All Trades	Artist	Hacker
Data Businessperson	Leader	Businessperson	Entrepeneur

Skills of Data Scientists

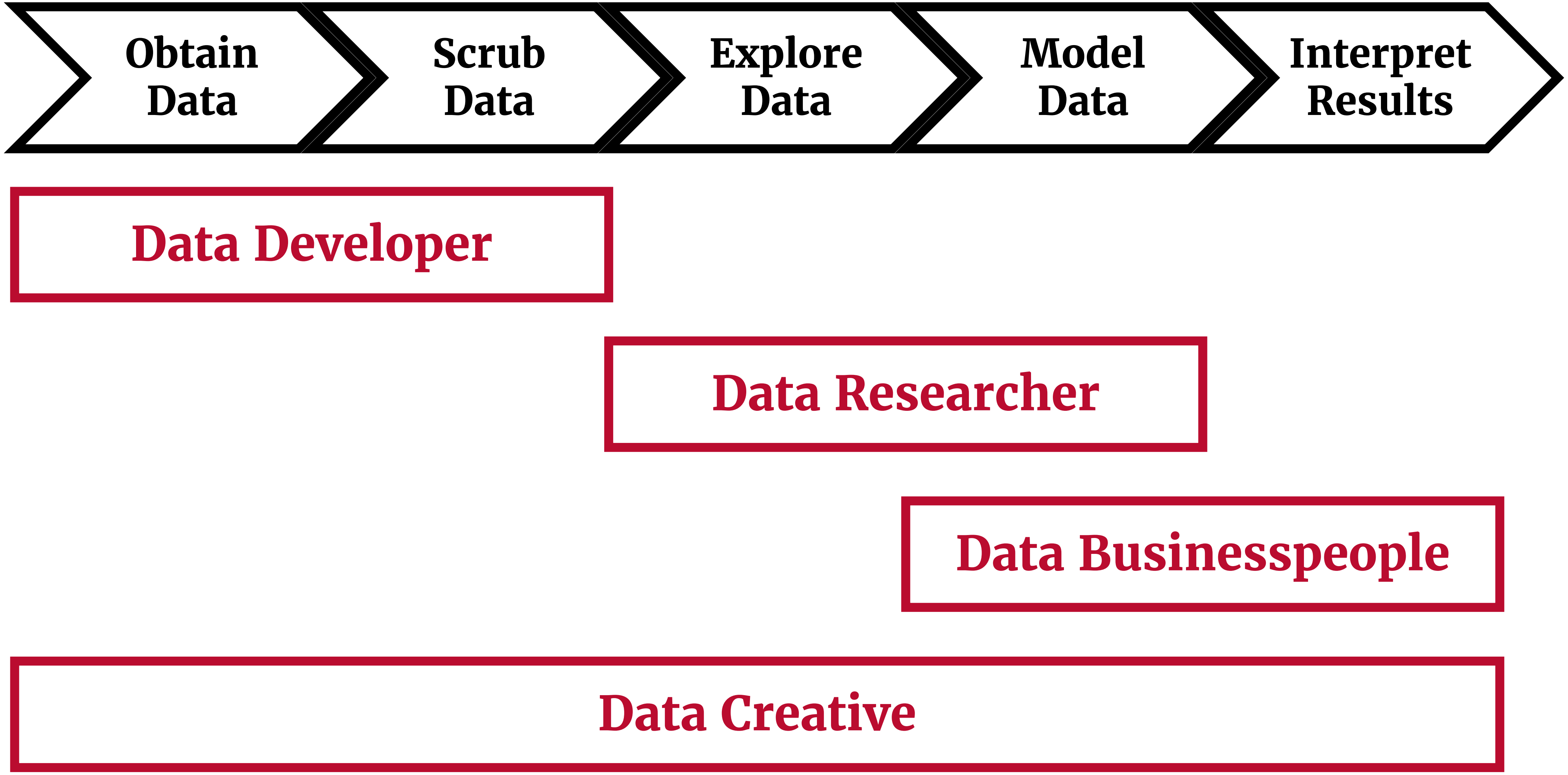
Business	ML / Big Data	Math / OR	Programming	Statistics
Product Developement	Unstructured Data	Optimization	Systems Administration	Visualization
Business	Structured Data	Math	Back End Programming	Temporal Statistics
	Machine Learning	Graphical Models	Front End Programming	Surveys and Marketing
	Big and Distributed Data	Bayesian / Monte Carlo Statistics		Spatial Statistics
		Algorithms		Science
		Simulation		Data Manipulation
				Classical Statistics

Data Science Skill Breakdown



Differences

Data Developer	**Question:**	What data architecture should the project utilize and what about the data's integrity, quality, and frequency?
	Output:	Back-end infrastructure, feature engineering, synthetic data, data ingestion pipelines, data architectures, descriptive dashboards
Data Researcher	**Question:**	Given that I want to improve X, how can I build or improve my model?
	Output:	Quantitative models (e.g., inferential, casual, predictive, prescriptive)
Data Creative	**Question:**	Given all the data and model findings, how can I improve profitability?
	Output:	Data-informed recommendations and technical solutions
Data Businesspeople	**Question:**	How can I deploy and scale our data science solutions?
	Output:	Monitoring dashboards, line-of-business products and services, and client/customer (end-user) applications


Obtain Data
Scrub Data
Explore Data
Model Data
Interpret Results
Data Developer
Data Researcher
Data Businesspeople
Data Creative

Q & A

Statistics are like bikinis. What they reveal is suggestive, but what they conceal is vital.

Aaron Levenstein

To Learn More…

1. Data Science
2. Data & Statistical Modeling
3. Static & Dynamic Visualizations
4. Predictive Analytics
5. Machine Learning
6. Casual Modeling

Connect With Us!

facebook.com
/VinsonInstitute

Carl Vinson
Institute of Government

@CVIOG_UGA

www.cviog.uga.edu

UNIVERSITY OF GEORGIA

Carl Vinson Institute of Government

Since 1927, the Carl Vinson Institute of Government has been an integral part of the University of Georgia. A public service and outreach unit of the university, the Institute of Government is the largest and most comprehensive university based organization serving governments in the United States through research services, customized assistance, training and development, and the application of technology.