Lesson 1.3: Is Data Science Just Hype?

UMD DATA605: Big Data Systems

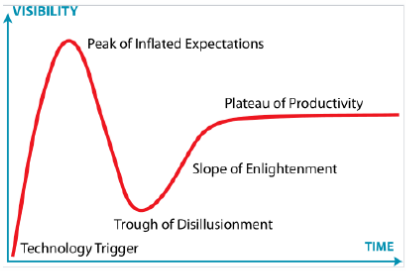
Lesson 1.3: Is Data Science Just Hype?

Instructor: Dr. GP Saggese, gsaggese@umd.edu

2 / 14: Is Data Science Just Hype?

Is Data Science Just Hype?

- **Big data (or data science) is everywhere**
  - *"Any process where interesting information is inferred from data"*
- Data scientist called the "sexiest job" of the 21st century
  - The term has becoming very muddled at this point
- **Is it all hype?**

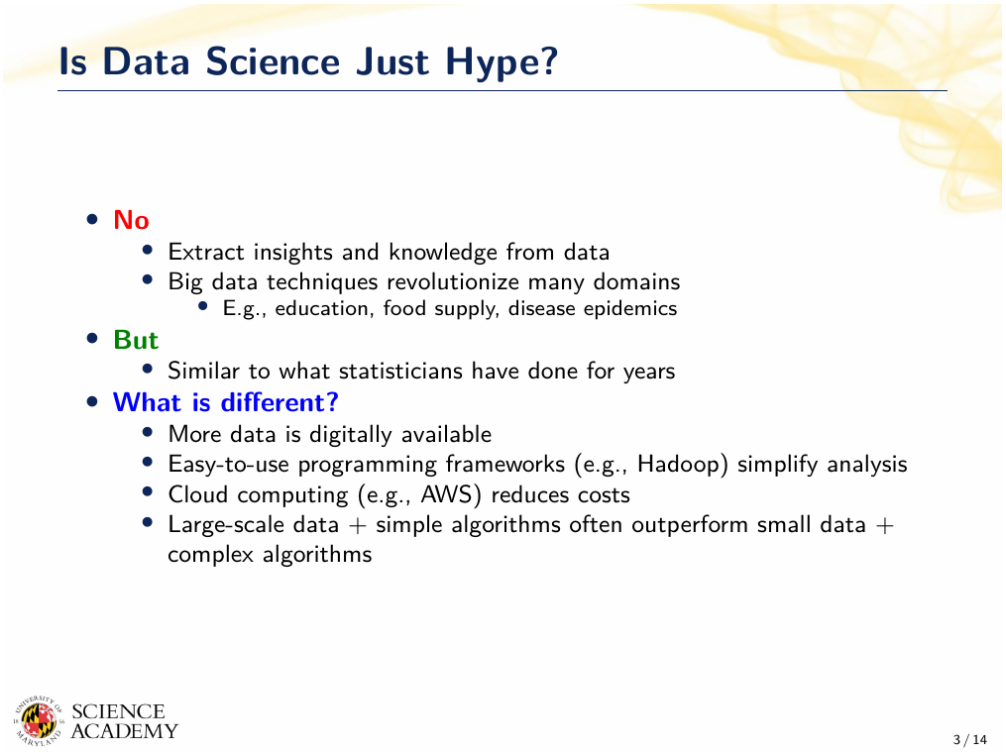


SCIENCE ACADEMY

2 / 14

- **Big data (or data science) is everywhere**
  - The phrase *"Any process where interesting information is inferred from data"* highlights the broad scope of data science. It means that data science involves analyzing data to extract meaningful insights, which can be applied across various fields like healthcare, finance, marketing, and more. This universality is why data science is so prevalent today.
- **Data scientist called the "sexiest job" of the 21st century**
  - This statement reflects the high demand and allure of the data scientist role. The job is considered "sexy" because it combines technical skills, creativity, and problem-solving to drive business decisions and innovations. However, the term "data scientist" has become muddled, meaning that as the field has grown, the role's definition has become less clear, encompassing a wide range of skills and responsibilities.
- **Is it all hype?**
  - This question prompts us to consider whether the excitement around data science is justified or if it's exaggerated. While data science has transformative potential, it's important to recognize that not all claims about its impact may be realistic. Understanding the true capabilities and limitations of data science is crucial to avoid overhyping its potential.

3 / 14: Is Data Science Just Hype?

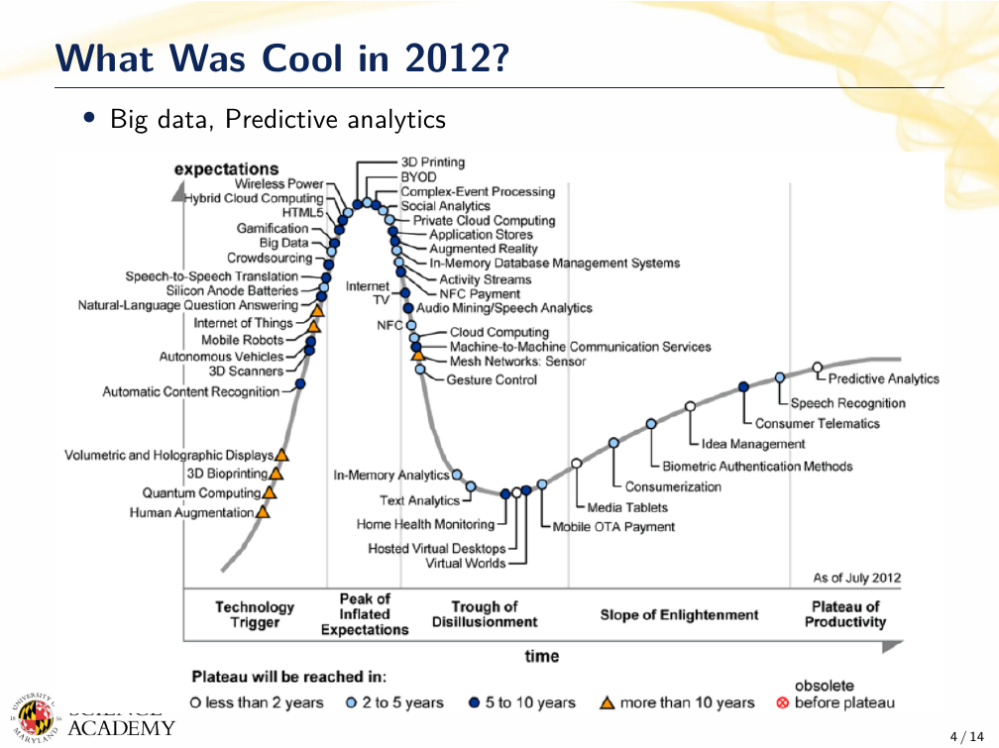


- **Is Data Science Just Hype?**
  - **No**
    - Data science is not just a buzzword; it plays a crucial role in extracting valuable insights and knowledge from vast amounts of data. This process helps organizations and researchers make informed decisions and predictions.
    - The advent of big data techniques has transformed various fields by providing new ways to analyze and interpret data. For example, in education, data science can help tailor learning experiences to individual students. In the food supply chain, it can optimize logistics and reduce waste. In managing disease epidemics, it can predict outbreaks and improve response strategies.
  - **But**
    - While data science is impactful, it's important to recognize that it builds on the foundation laid by statisticians. Statisticians have been analyzing data to draw conclusions for many years, and data science extends these principles with new tools and technologies.
  - **What is different?**
    - The key difference today is the sheer volume of data that is now digitally available, which was not the case in the past. This abundance of data allows for more comprehensive analysis.
    - Easy-to-use programming frameworks, such as *Hadoop*, have made it simpler for people to process and analyze large datasets without needing extensive programming skills.
    - Cloud computing services, like *AWS*, have significantly reduced the cost and complexity of storing and processing large datasets, making data science more accessible

to a wider audience.

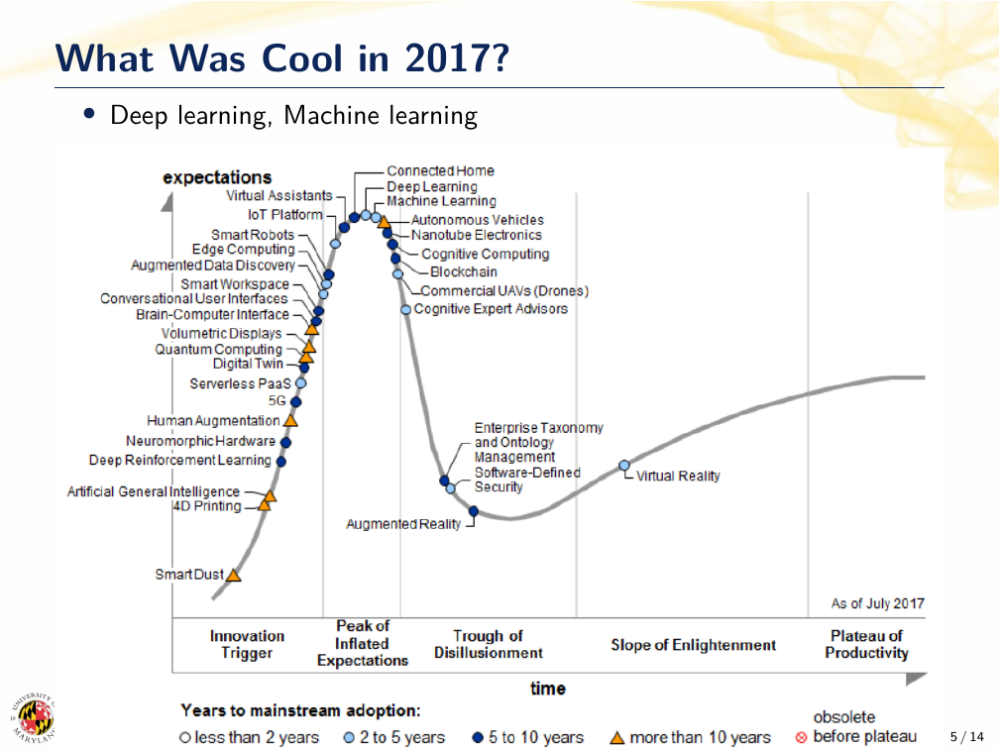
* Interestingly, having access to large-scale data allows for the use of simpler algorithms, which can often yield better results than using complex algorithms on smaller datasets. This is because more data can provide a clearer picture and reduce the noise in the analysis.

4 / 14: What Was Cool in 2012?



- **Big Data**
  - In 2012, the term *big data* was gaining significant traction. It refers to the vast volumes of data generated every second from various sources like social media, sensors, and transactions. The challenge was not just storing this data but also processing and analyzing it to extract meaningful insights. This was a time when companies started realizing the potential of data as a valuable asset, leading to investments in technologies and infrastructure to handle big data.
- **Predictive Analytics**
  - Predictive analytics involves using historical data to predict future outcomes. In 2012, this was becoming a hot topic as businesses sought to leverage data to forecast trends, customer behavior, and potential risks. The rise of machine learning algorithms played a crucial role in enhancing predictive analytics, allowing for more accurate and actionable predictions. This was particularly appealing to industries like finance, healthcare, and retail, where anticipating future trends could lead to significant competitive advantages.
- **Context**
  - The excitement around big data and predictive analytics in 2012 was driven by advancements in technology, such as improved data storage solutions and more powerful computing capabilities. This period marked the beginning of a data-driven approach in decision-making across various sectors, setting the stage for the data-centric world we live in today.

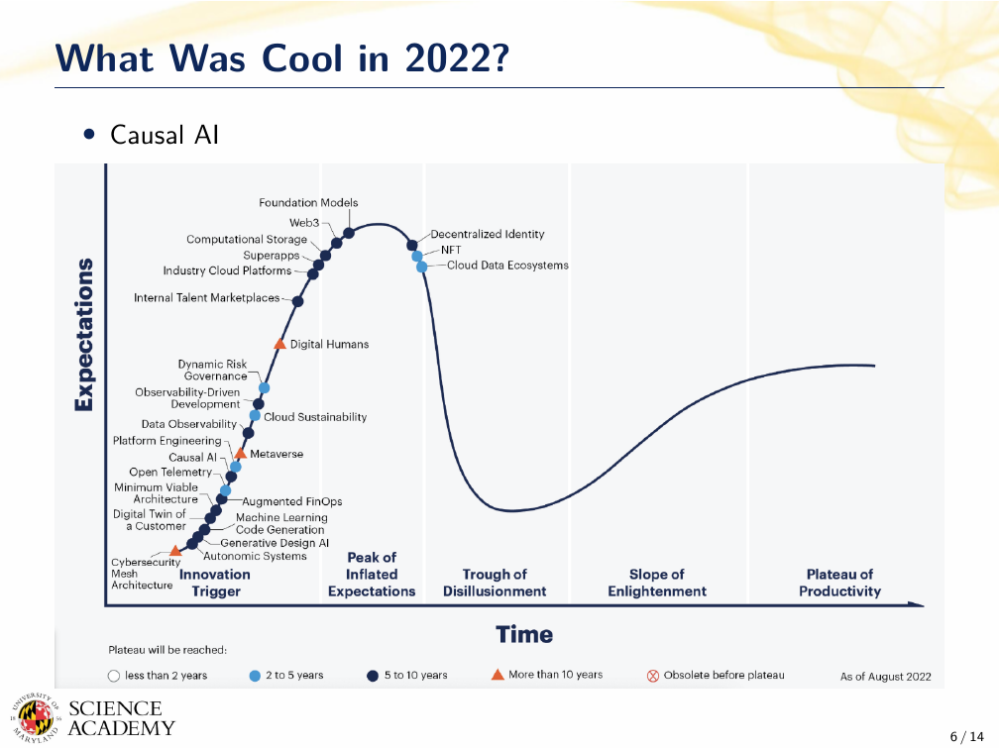
5 / 14: What Was Cool in 2017?



- **Deep Learning**: In 2017, deep learning was a major trend in the field of artificial intelligence. It involves using neural networks with many layers (hence "deep") to model complex patterns in data. This approach was particularly successful in areas like image and speech recognition, where it significantly improved accuracy and performance. The excitement around deep learning was due to its ability to automatically learn features from raw data, reducing the need for manual feature engineering.
- **Machine Learning**: Machine learning, a broader field that includes deep learning, was also a hot topic in 2017. It refers to the use of algorithms and statistical models to enable computers to improve their performance on a task through experience. Machine learning was being applied across various industries, from healthcare to finance, to automate processes and gain insights from large datasets. The rise of big data and increased computational power contributed to the growing interest and advancements in machine learning during this time.

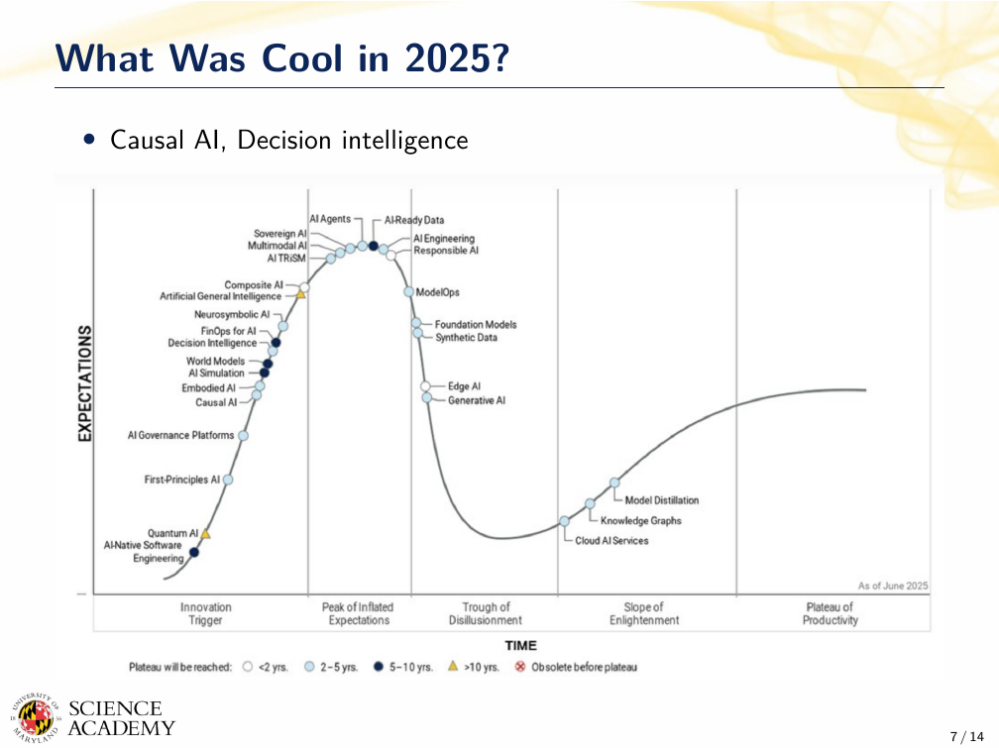
In summary, 2017 was a pivotal year for both deep learning and machine learning, as these technologies began to demonstrate their potential to transform industries and solve complex problems.

6 / 14: What Was Cool in 2022?



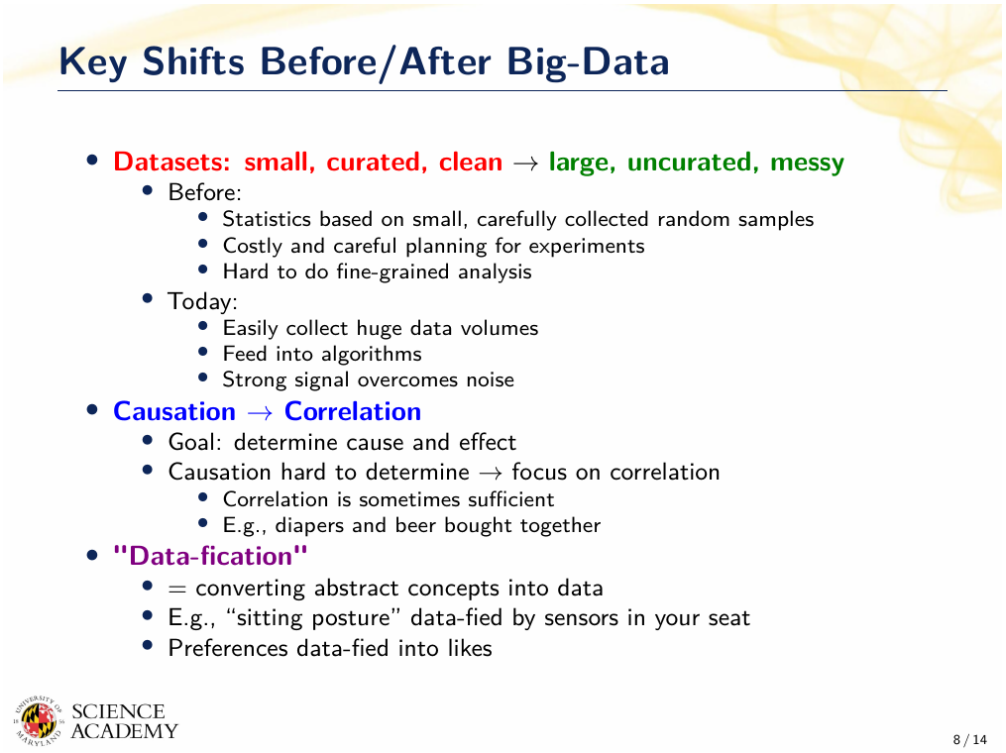
- **Causal AI**
  - Causal AI was a significant trend in 2022, gaining attention for its ability to go beyond traditional correlation-based machine learning models. Unlike standard AI models that often identify patterns and correlations, Causal AI focuses on understanding the *cause-and-effect* relationships within data. This is crucial because knowing what causes what can lead to more accurate predictions and better decision-making.
  - The importance of Causal AI lies in its potential to provide insights that are not just predictive but also prescriptive. For instance, in healthcare, understanding the causal factors of a disease can lead to more effective treatments. In business, it can help identify the true drivers of customer behavior, leading to more targeted marketing strategies.
  - Causal AI uses techniques like *causal inference* and *structural causal models* to identify these relationships. This approach can help in scenarios where interventions are possible, allowing organizations to simulate the effects of potential actions before implementing them.
  - The growing interest in Causal AI reflects a broader trend towards more interpretable and actionable AI systems, which are essential for building trust and ensuring ethical use of AI technologies.

7 / 14: What Was Cool in 2025?



- **Causal AI**:
  - Causal AI refers to a branch of artificial intelligence that focuses on understanding cause-and-effect relationships rather than just correlations. This is important because while traditional AI models can identify patterns and correlations in data, they often cannot determine if one thing causes another.
  - By understanding causality, AI systems can make better predictions and decisions. For example, in healthcare, causal AI can help determine whether a treatment actually improves patient outcomes or if the observed effects are due to other factors.
  - In 2025, causal AI was considered "cool" because it represented a significant advancement in AI's ability to provide insights that are more actionable and reliable.
- **Decision Intelligence**:
  - Decision intelligence is an emerging field that combines data science, social science, and managerial science to improve decision-making processes. It involves using AI and machine learning to analyze data and provide recommendations that help organizations make better decisions.
  - This approach is particularly valuable in complex environments where decisions need to be made quickly and with incomplete information. By leveraging decision intelligence, businesses can optimize operations, reduce risks, and improve outcomes.
  - In 2025, decision intelligence gained popularity as it empowered organizations to harness the power of AI not just for automation, but for strategic decision-making, making it a "cool" trend in the tech world.

8 / 14: Key Shifts Before/After Big-Data



- **Datasets: small, curated, clean → large, uncurated, messy**
  - *Before Big Data*:
    - In the past, data analysis relied on small datasets that were carefully selected and cleaned. This was because collecting data was expensive and required meticulous planning.
    - Researchers often used random samples to make statistical inferences, but this limited the ability to perform detailed analysis.
  - *Today with Big Data*:
    - We can now gather vast amounts of data quickly and at a lower cost. This data is often messy and uncurated, but the sheer volume allows us to extract meaningful insights.
    - Algorithms can process these large datasets, and the significant amount of data helps to highlight important patterns, even if there is some noise.
- **Causation → Correlation**
  - Traditionally, the aim was to understand cause and effect relationships. However, determining causation is complex and often not feasible with large datasets.
  - With big data, the focus has shifted to identifying correlations, which can be very useful. For example, noticing that diapers and beer are often purchased together can inform marketing strategies, even if the exact cause isn't clear.
- **"Data-fication"**
  - This term refers to the process of turning abstract concepts into quantifiable data.
  - For instance, sensors can capture data about your sitting posture, transforming it into something that can be analyzed. Similarly, online preferences are converted into data points like "likes," which can be used to understand user behavior and preferences.

9 / 14: Examples: Election Prediction

Examples: Election Prediction

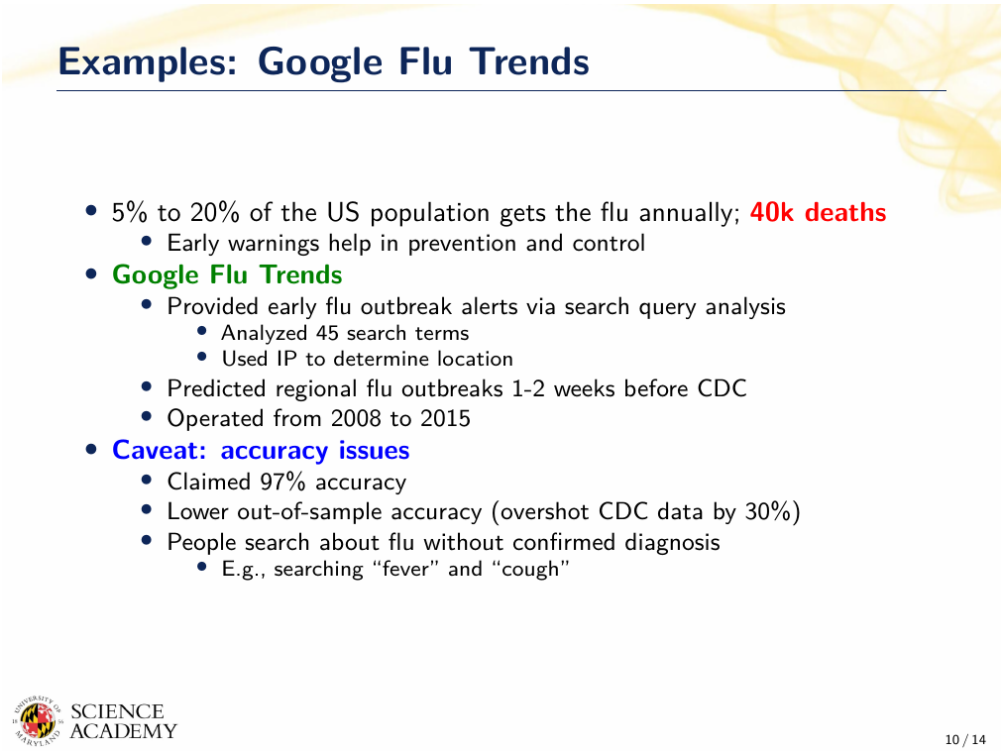
- Nate Silver and the 2012 Elections
  - Predicted 49/50 states in 2008 US elections
  - Predicted 50/50 states in 2012 US elections
- **Reasons for accuracy**
  - Multiple data sources
  - Historical accuracy incorporation
  - Statistical models
  - Understanding correlations
  - Monte-Carlo simulations for electoral probabilities
  - Focus on probabilities
  - Effective communication

SCIENCE ACADEMY

9 / 14

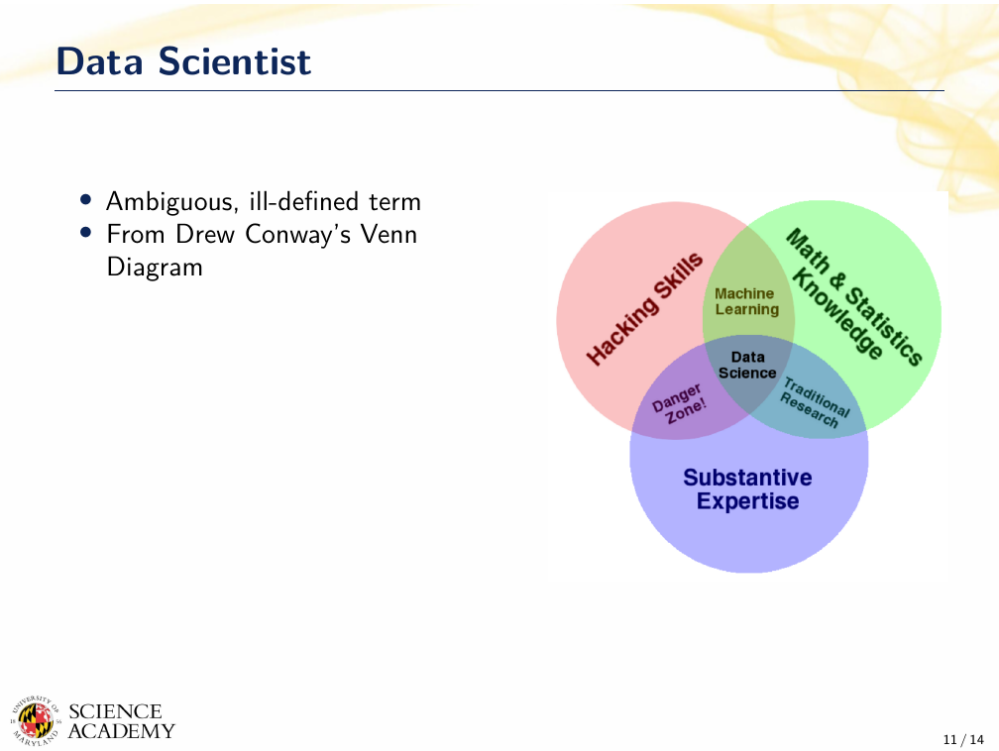
- **Nate Silver and the 2012 Elections**
  - Nate Silver is a well-known statistician and writer who gained fame for his accurate predictions in the US elections. In 2008, he correctly predicted the outcomes in 49 out of 50 states, and in 2012, he improved his accuracy by predicting all 50 states correctly. This achievement highlighted the power of data-driven approaches in making accurate predictions.
- **Reasons for accuracy**
  - **Multiple data sources**: Silver used a variety of data sources, including polls, economic indicators, and demographic data, to create a comprehensive view of the electoral landscape.
  - **Historical accuracy incorporation**: By considering historical data, Silver was able to identify patterns and trends that informed his predictions.
  - **Statistical models**: He employed sophisticated statistical models to analyze the data, which helped in understanding the dynamics of voter behavior.
  - **Understanding correlations**: Recognizing how different factors are related allowed Silver to refine his predictions and account for potential biases.
  - **Monte-Carlo simulations for electoral probabilities**: These simulations helped in estimating the likelihood of different electoral outcomes by running numerous scenarios.
  - **Focus on probabilities**: Instead of making absolute predictions, Silver focused on the probabilities of various outcomes, which provided a more nuanced understanding of the election.
  - **Effective communication**: Silver's ability to communicate complex statistical concepts in an accessible way helped the public understand the uncertainties and probabilities involved in election forecasting.

10 / 14: Examples: Google Flu Trends



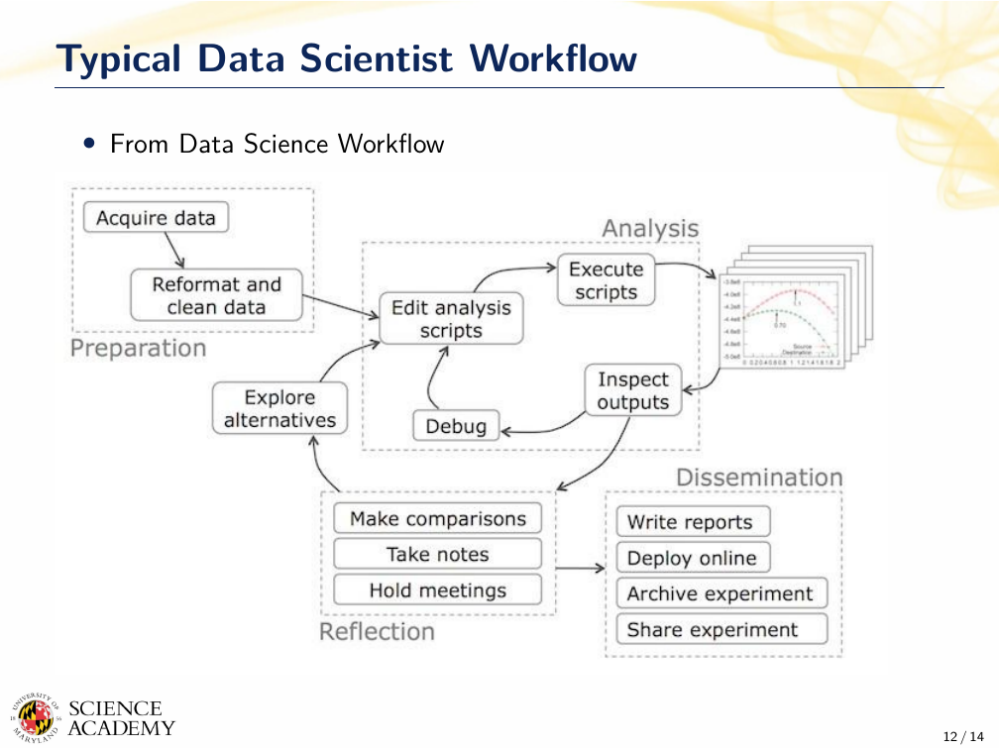
- **5% to 20% of the US population gets the flu annually; 40k deaths**
  - The flu is a significant health concern in the US, affecting a large portion of the population each year and resulting in approximately 40,000 deaths. This highlights the importance of early detection and prevention strategies to mitigate its impact.
- **Google Flu Trends**
  - Google Flu Trends was an innovative project that aimed to provide early warnings of flu outbreaks by analyzing search queries. By examining 45 specific search terms related to flu symptoms and using IP addresses to determine the location of searches, Google attempted to predict where flu outbreaks might occur.
  - The tool was able to predict regional flu outbreaks 1-2 weeks before the Centers for Disease Control and Prevention (CDC), offering potentially valuable lead time for public health responses.
  - The project ran from 2008 to 2015, showcasing the potential of big data and machine learning in public health surveillance.
- **Caveat: accuracy issues**
  - Although Google Flu Trends claimed a high accuracy rate of 97%, it faced challenges with accuracy, particularly when applied to new data (out-of-sample). It sometimes overestimated flu cases by as much as 30% compared to CDC data.
  - One reason for this discrepancy is that people often search for flu-related symptoms like "fever" and "cough" without having a confirmed flu diagnosis. This behavior can lead to overestimations in flu predictions, as not all searches correlate directly with actual flu cases.

11 / 14: Data Scientist



- **Ambiguous, ill-defined term**
  - The term *data scientist* is often used in various contexts, leading to some confusion about what it precisely means. It can encompass a wide range of skills and responsibilities, making it a bit tricky to pin down.
  - In general, a data scientist is someone who uses data to solve problems. This can involve analyzing data, building models, and communicating insights to help make decisions.
  - The role can vary significantly between organizations. Some might focus more on statistical analysis, while others might emphasize machine learning or data engineering.
- **From Drew Conway's Venn Diagram**
  - Drew Conway's Venn Diagram is a popular way to visualize the skill set of a data scientist. It highlights the intersection of three main areas: *hacking skills*, *math and statistics knowledge*, and *substantive expertise*.
  - *Hacking skills* refer to the ability to manipulate and analyze data using programming languages like Python or R.
  - *Math and statistics knowledge* is crucial for understanding data patterns and building predictive models.
  - *Substantive expertise* involves having domain knowledge to apply data science effectively in a specific field, such as finance, healthcare, or marketing.
  - The diagram helps illustrate why the role of a data scientist can be so varied and why it requires a diverse skill set.

12 / 14: Typical Data Scientist Workflow



- **Typical Data Scientist Workflow**
  - **From Data Science Workflow**

In this slide, we are looking at a visual representation of a *typical data scientist's workflow*. This workflow is a series of steps that data scientists follow to extract insights and value from data. Let's break down the key components:

- **Data Collection**: This is the first step where data scientists gather raw data from various sources. This could include databases, APIs, or even web scraping. The goal is to collect relevant data that will be used for analysis.
- **Data Cleaning**: Once the data is collected, it often needs cleaning. This involves removing duplicates, handling missing values, and correcting errors. Clean data is crucial for accurate analysis.
- **Data Exploration**: In this phase, data scientists explore the data to understand its structure and characteristics. They use statistical methods and visualization tools to identify patterns and trends.
- **Model Building**: After understanding the data, the next step is to build predictive models. This involves selecting appropriate algorithms and training them on the data to make predictions or classifications.
- **Model Evaluation**: Once a model is built, it needs to be evaluated to ensure its accuracy and reliability. This involves testing the model on new data and using metrics to assess its performance.
- **Deployment**: The final step is deploying the model into a production environment where it

can be used to make real-time decisions or predictions.

This workflow is iterative, meaning data scientists often revisit previous steps to refine their models and improve results. Understanding this workflow is essential for anyone looking to work in data science, as it provides a structured approach to solving complex data problems.

13 / 14: Where Data Scientist Spends Most Time

Where Data Scientist Spends Most Time

- 80-90% of the work is data cleaning and wrangling
- "Janitor Work" in Data Science

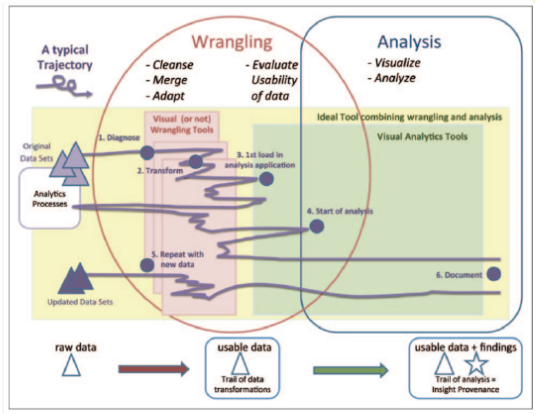


Figure 1. The iterative process of wrangling and analysis. One or more initial data sets may be used and new versions may come later. The wrangling and analysis phases overlap. While wrangling tools tend to be separated from the visual analysis tools, the ideal system would provide integrated tools (light yellow). The purple line illustrates a typical iterative process with multiple back and forth steps. Much wrangling may need to take place before the data can be loaded within visualization and analysis tools, which typically immediately reveals new problems with the data. Wrangling might take place at all the stages of analysis as users sort out interesting insights from dirty data, or new data become available or needed. At the bottom we illustrate how the data evolves from raw data to usable data that leads to new insights.

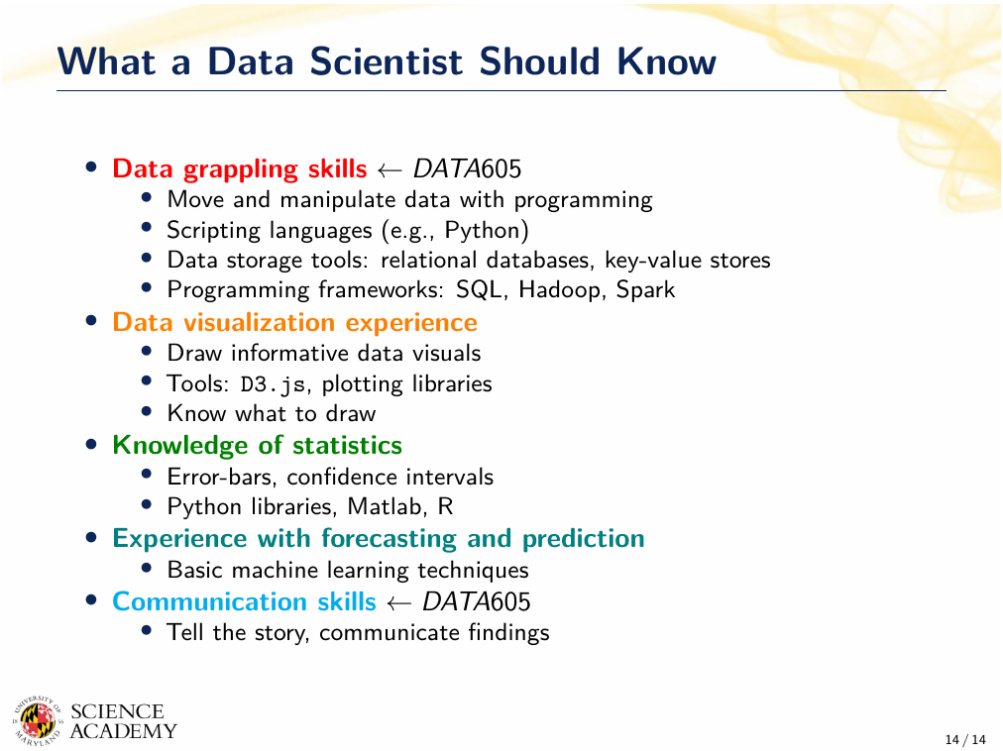
SCIENCE ACADEMY

13 / 14

- **80-90% of the work is data cleaning and wrangling**
  - Data scientists spend a significant portion of their time on tasks related to preparing data for analysis. This involves cleaning and organizing raw data to make it usable.
  - Data cleaning includes removing errors, handling missing values, and ensuring consistency in data formats.
  - Data wrangling involves transforming and mapping data from one "raw" form into another format to make it more appropriate and valuable for a variety of downstream purposes, such as analytics.
  - This step is crucial because the quality of the data directly impacts the quality of the insights derived from it.
- **"Janitor Work" in Data Science**
  - The term "janitor work" is often used to describe the less glamorous, yet essential, tasks of cleaning and organizing data.
  - While it might sound mundane, this work is critical because it lays the foundation for any successful data analysis or machine learning project.
  - Without proper data cleaning and wrangling, the results of data analysis can be misleading or incorrect.
  - This highlights the importance of attention to detail and thoroughness in the early stages of data science projects.

The image on the slide likely illustrates the proportion of time spent on these tasks, emphasizing their significance in the data science workflow.

14 / 14: What a Data Scientist Should Know



- **Data grappling skills** ← *DATA*605
  - *Move and manipulate data with programming*: This means being able to handle data efficiently using code. It's about cleaning, transforming, and preparing data for analysis.
  - *Scripting languages (e.g., Python)*: Python is a popular language for data science due to its simplicity and powerful libraries. Knowing how to write scripts in Python is crucial for automating data tasks.
  - *Data storage tools: relational databases, key-value stores*: Understanding how to store and retrieve data is essential. Relational databases like MySQL or PostgreSQL and key-value stores like Redis are common tools.
  - *Programming frameworks: SQL, Hadoop, Spark*: SQL is used for querying databases, while Hadoop and Spark are frameworks for processing large datasets. These tools help manage and analyze big data efficiently.
- **Data visualization experience**
  - *Draw informative data visuals*: Creating visuals that clearly communicate data insights is key. This involves choosing the right type of chart or graph to represent the data.
  - *Tools: `D3.js`, plotting libraries*: `D3.js` is a JavaScript library for creating dynamic, interactive data visualizations. Other plotting libraries like Matplotlib or Seaborn in Python are also widely used.
  - *Know what to draw*: It's important to understand which visual representation best suits the data and the message you want to convey.
- **Knowledge of statistics**
  - *Error-bars, confidence intervals*: These are statistical tools used to express the uncertainty in data. Understanding them helps in making informed decisions based on data analysis.

  - *Python libraries, Matlab, R*: These are tools and languages commonly used for statistical analysis. Each has its strengths, and knowing how to use them can enhance data analysis capabilities.
- **Experience with forecasting and prediction**
  - *Basic machine learning techniques*: This involves using algorithms to make predictions based on data. Understanding the basics of machine learning is crucial for building predictive models.
- **Communication skills** $\leftarrow DATA605$
  - *Tell the story, communicate findings*: Being able to explain data insights in a clear and compelling way is vital. This involves not just presenting data but also telling a story that highlights the key findings and their implications.