Data Thinking
Lecture 4

Wan Fang

Southern University of Science and Technology

Agenda

- Concepts of Data X
  - Data Eco-system & Lifecycle
  - Data Privacy & Ethics
  - Data Integrity
  - Data & Analytics Skills
  - How to Improve Your Skills

- Data Exploration and Analytics by Examples
  - Global Health Data by OurWorldInData

Concepts of Data X

Wan Fang

Southern University of Science and Technology

[Adapted from A Beginner's Guide to Data & Analytics by Harvard Business School]

Data Science vs. Data Analytics

- **Data science** is the process of building, cleaning, and structuring datasets to analyze and extract meaning.
- **Data analytics**, on the other hand, refers to the process and practice of analyzing data to answer questions, extract insights, and identify trends.

    - *You can think of data science as a precursor to data analysis. If your dataset isn't structured, cleaned, and wrangled, how will you be able to draw accurate, insightful conclusions?*

- Every analysis should be a feedback loop that deepens your learning.

    - *What can I learn from the results of this analysis about the underlying context, about competition, about customers, about suppliers?*
    - *How do the results of this analysis validate or reinforce hypotheses I had before I did the analysis?*

Data Science *in Business*

- To collect, organize, and maintain data—often to write algorithms that make large-scale analysis possible.
  - When designed correctly and tested thoroughly, algorithms can catch information or trends that humans miss.
  - They can also significantly speed up the processes of gathering and analyzing data.

Gain customer insights

- Data about your customers can reveal details about their habits, demographics, preferences, and aspirations.
- A foundational understanding of data science can help you make sense of and leverage it to improve user experiences and inform retargeting efforts.

Increase security

- You can also use data science to increase your business's security and protect sensitive information.
- For example, machine-learning algorithms can detect bank fraud faster and with greater accuracy than humans, simply because of the sheer volume of data generated every day.

Inform internal finances

- Your organization's financial team can utilize data science to create reports, generate forecasts, and analyze financial trends.
- Data on a company's cash flows, assets, and debts is constantly gathered, which financial analysts use to manually or algorithmically detect trends in financial growth or decline.

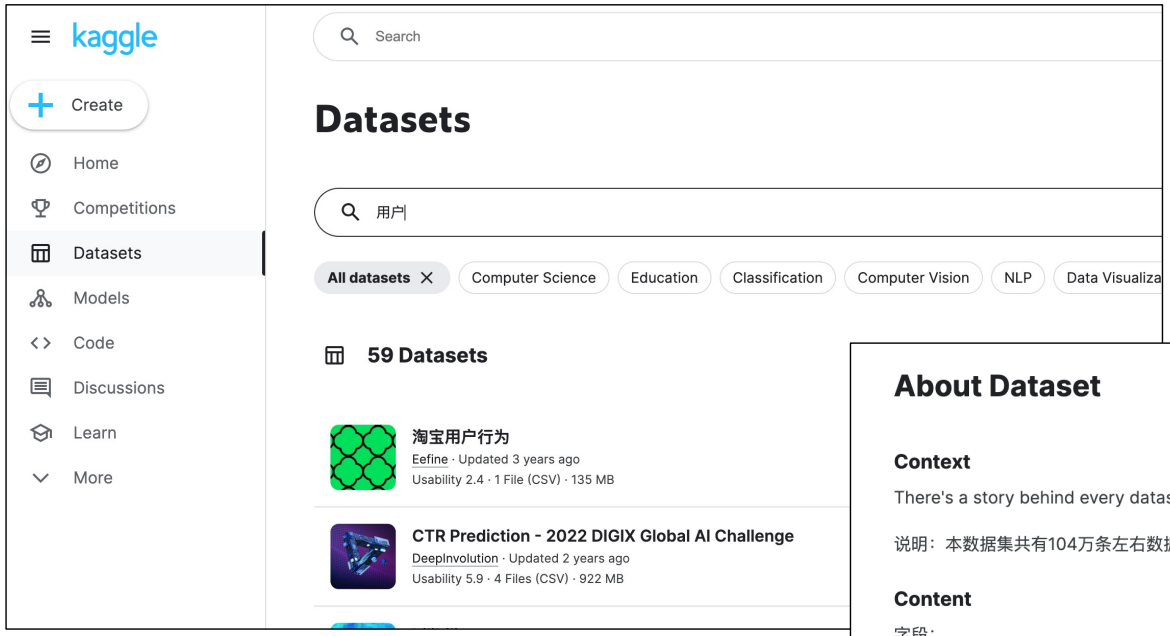
Streamline manufacturing

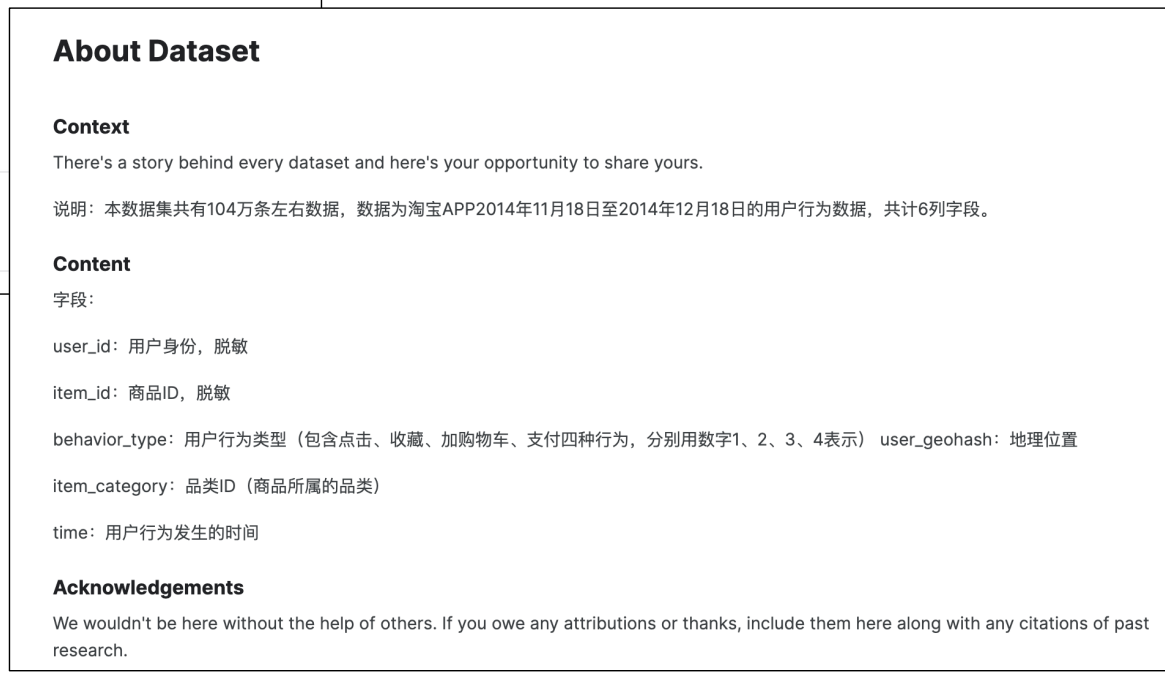
- Manufacturing machines gather data from production processes at high volumes.
- In cases where the volume of data collected is too high for a human to manually analyze it, an algorithm can be written to clean, sort, and interpret it quickly and accurately to gather insights that drive cost-saving improvements.

Predict future market trends

- Collecting and analyzing data on a larger scale can enable you to identify emerging trends in your market.
- By staying up to date on the behaviors of your target market, you can make business decisions that allow you to get ahead of the curve.

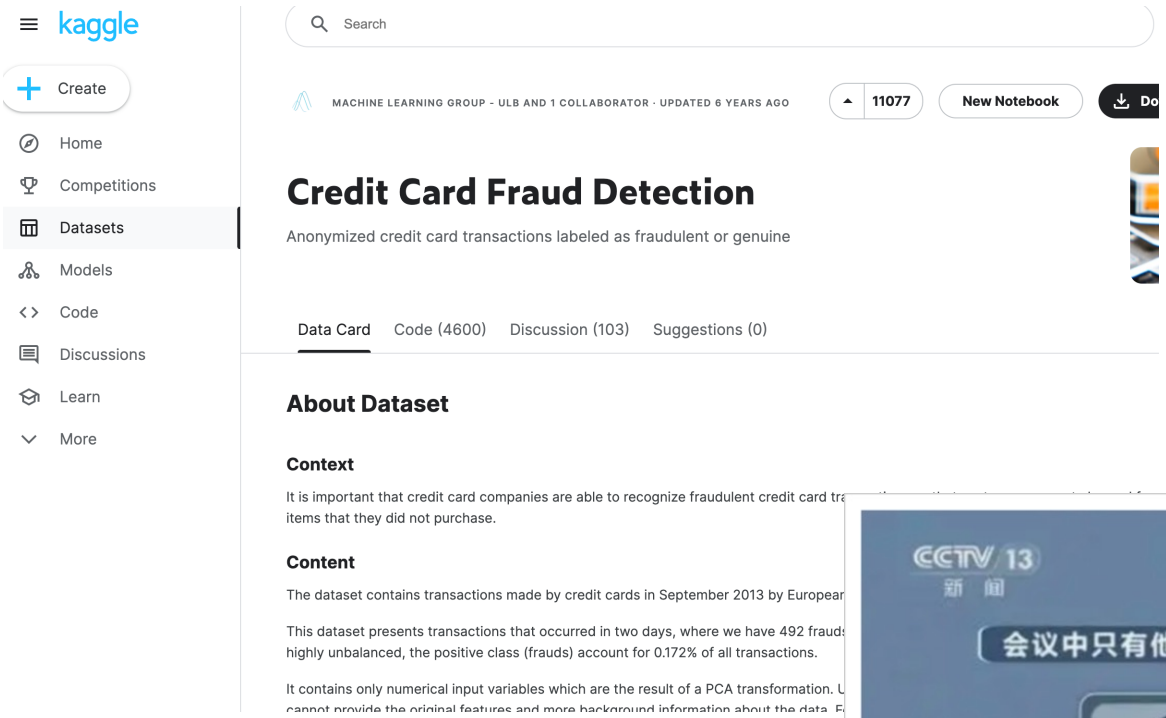
Gain customer insights





- Data about your customers can reveal details about their habits, demographics, preferences, and aspirations.
- A foundational understanding of data science can help you make sense of and leverage it to improve user experiences and inform retargeting efforts.

Increase security



Streamline manufacturing

Data Analytics *in Business*

- The main goal of business analytics is *to extract meaningful insights from data that an organization can use to inform its strategy and, ultimately, reach its objectives*.
- Business analytics can be used for:

Budgeting and forecasting

- By assessing a company's historical revenue, sales, and costs data alongside its goals for future growth, an analyst can identify the budget and investments required to make those goals a reality.

Risk management

- By understanding the likelihood of certain business risks occurring—and their associated expenses—an analyst can make cost-effective recommendations to help mitigate them.

Marketing and sales

- By understanding key metrics, such as lead- to-customer conversion rate, a marketing analyst can identify the number of leads their efforts must generate to fill the sales pipeline.

Product development (or research and development)

- By understanding how customers reacted to product features in the past, an analyst can help guide product development, design, and user experience in the future.

Four Types of Analytics

- Analytics is used to extract meaningful insights from data that can drive decision- making and strategy formulation.
  - There are **four types of analytics** you can leverage depending on the data you have and the type of knowledge you'd like to gain.

Descriptive analytics

- looks at data to examine, understand, and describe something that's already happened.

Diagnostic analytics

- goes deeper than descriptive analytics by seeking to understand the "why" behind what happened.

Predictive analytics

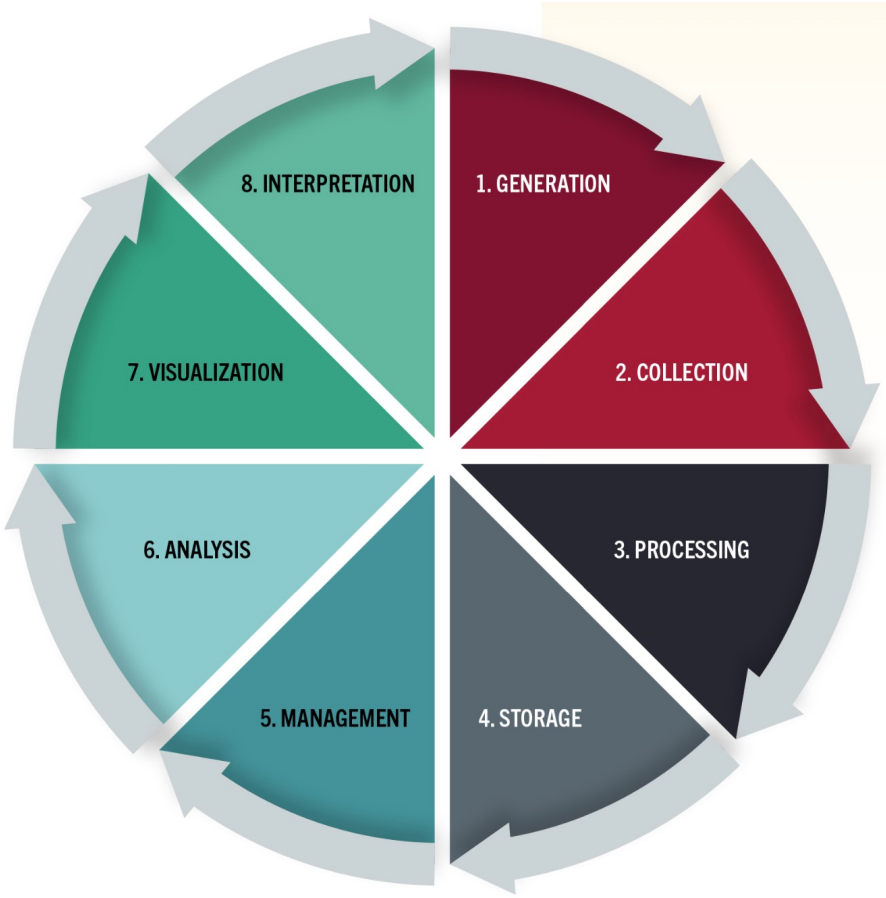
- relies on historical data, past trends, and assumptions to answer questions about what will happen in the future.

Prescriptive analytics

- identifies specific actions an individual or organization should take to reach future targets or goals.

Data Ecosystem & Lifecycle

- ***Data ecosystem*** refers to the programming languages, packages, algorithms, cloud-computing services, and general infrastructure an organization uses to collect, store, analyze, and leverage data.



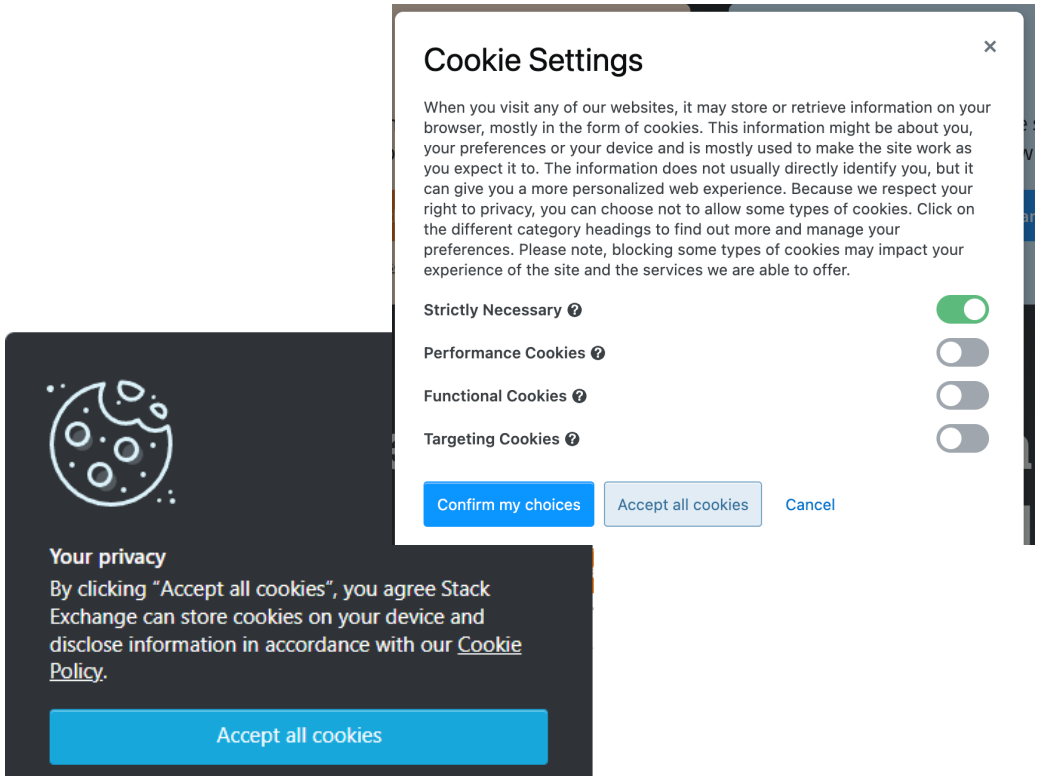
- ***Data life cycle*** describes the path data takes from when it's first generated to when it's interpreted into actionable insights.

- A data project's steps are often described as a cycle because the lessons learned, and insights gleaned from one project typically inform the next. In this way, the final step of the process feeds back into the first, enabling you to start again with new goals and learnings.

Data Privacy & Ethics

- ***Data privacy***, also known as *information privacy*, is a subcategory of data protection that encompasses the ethical and legal obligation to protect access to personally identifiable information (PII), which is any information that can be linked to a specific individual.
  - Some examples of PII include full name, address, ID number, and passport number.

- Data privacy is made up of three key questions:
  - 1. What data is collected?
  - 2. How is the data stored?
  - 3. Who can access the data?



Data Privacy & Ethics

- ***The ethics of data privacy*** can be boiled down to the fact that
  - an **individual's consent** is necessary to collect, store, and use their personal information.

- As a data handler, you have a responsibility to be transparent with your subjects about
  - your intentions,
  - what their data will be used for, and
  - who will have access to it.

- In addition, you need to ensure your use of data doesn't cause harm to an individual or group of people.
  - This is referred to as ***disparate impact*** and is unlawful.

Data Integrity

- Data integrity is the accuracy, completeness, and quality of data as it's maintained over time and across formats.
  - Preserving the integrity of your company's data is a constant process.

- Threats to a dataset's integrity include:
  - **Human error:**
    - For instance, accidentally deleting a row of data in a spreadsheet.
  - **Inconsistencies across formats:**
    - For instance, a dataset in Microsoft Excel that relies on cell referencing may not be accurate in a different format that doesn't allow those cells to be referenced.
  - **Collection error:**
    - For instance, data collected is inaccurate or lacking information, creating an incomplete picture of the subject.
  - **Cybersecurity or internal privacy breaches:**
    - For instance, someone hacks into your company's database with the intent to damage or steal information, or an internal employee damages data with malicious intent.

- To maintain your datasets' integrity,
  - diligently check for errors in the collection, formatting, and analysis phases,
  - monitor for potential breaches, and
  - educate your team about the importance of data integrity.

Data & Analytics Skills

- **1. Critical Thinking**
  - Data science is a discipline that's built on a foundation of critical thinking.
  - If you're interested in using data to solve business problems, you need to be adept at thinking critically about challenges and solutions.
    - While data can provide many answers, it's nothing without a human's discerning eye.
- **2. Hypothesis Formation and Testing**
  - At the heart of data and analytics is the desire to answer questions.
    - The proposed explanations for these leading questions are called hypotheses, which must be formed before analysis takes place.
  - An example of a hypothesis is, "I predict that a person's likelihood of recommending our product is directly proportional to their reported satisfaction with the product."
    - You predict the data will show this trend and must prove or disprove the hypothesis through analysis. Without a hypothesis, your analysis has no clear direction.

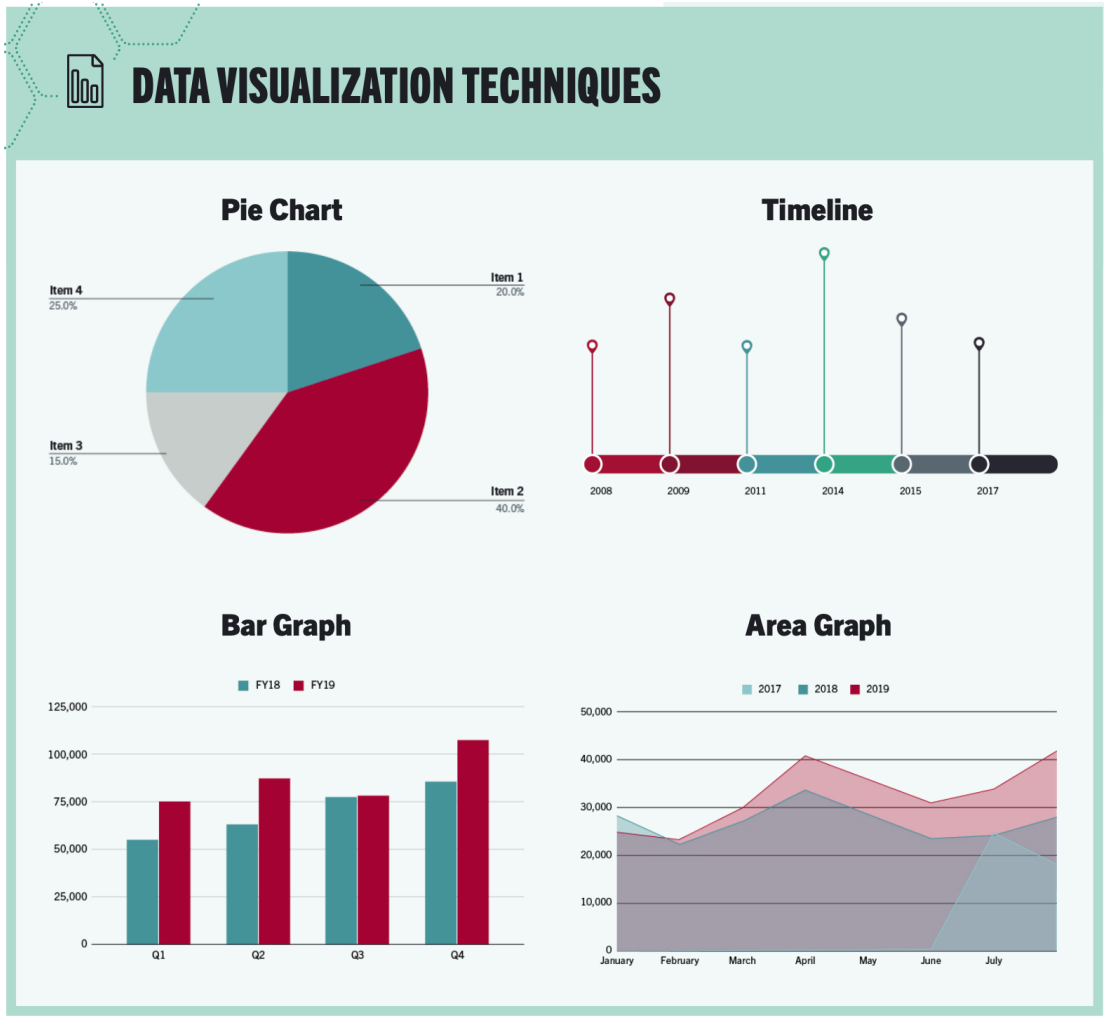
Data & Analytics Skills

- **3. Data Wrangling**
  - Data wrangling is the process of cleaning raw data in preparation for analysis.
    - It involves identifying and resolving mistakes, filling in missing data, and organizing and transferring it into an easily understandable format.
  - This is an important skill for anyone dealing with data to acquire because it leads to a more efficient and organized data analysis process.
    - You can extract valuable insights from data more quickly when it's cleaned and in its optimal viewing format.

- **4. Mathematical Ability**
  - You don't have to be a mathematician to become data literate, but strong math skills become increasingly important as you deal with more complex analyses.
    - If you're not a data scientist or analyst, your work may not require you to understand the more complex mathematical concepts, but having a basic understanding of statistics can go a long way.
  - A seasoned data professional needs a solid understanding of statistics, probability, linear algebra, and multivariable calculus.
    - Data scientists often call on statistical methods to find structure in data and make predictions, and linear algebra and calculus can make machine-learning algorithms easier to comprehend.

Data & Analytics Skills

- **5. Data Visualization**
  - It's crucial to know how to transform raw data into compelling visuals that tell a story.
  - Some popular data visualization techniques that all business professionals should know include pie charts, bar charts, and histograms.
  - To create these visualizations, use a data visualization tool.
  - Examples include Microsoft Excel and Power BI, Google Charts, Tableau.



Data & Analytics Skills

- **6. Programming**
  - Programming languages, like Python and R, are commonly used to solve complex statistical problems with data.
  - While programming skills are immensely valuable, they're not necessary for beginners dabbling in data.
    - It's more important to focus on effectively analyzing and visualizing data to draw conclusions.

- **7. Machine Learning**
  - As artificial intelligence grows in popularity, machine learning is a highly valuable skill for professionals working with big data.
  - Machine learning refers to the use of computer algorithms that automatically learn from and adapt in response to data.

How to Improve Your Skills

- **1. Embrace the Challenge**
  - The first step is to confront any mental barriers surrounding your ability to learn and develop data skills.
  - Although data science has a reputation for being code-based and complex, its concepts are accessible if you have the desire and drive to learn and put in the work.

- **2. Consider Opposing Viewpoints**
  - When analyzing data, it's crucial to consider all possible interpretations and avoid getting stuck in one way of thinking.
    - For instance, imagine you track users who click a button on your site to download an e-book. The data shows that the user's age is positively correlated with their likelihood to click the button; as age increases, downloads increase. At first glance, you may interpret this trend to mean that a user downloads the e-book because of their age.
  - This conclusion doesn't take into consideration the variables that change with age.
    - For instance, perhaps the real reason older users are more likely to download the e-book is their higher level of responsibility at work, average income, or likelihood of being parents.
    - This example illustrates the need to consider multiple interpretations of data, and it specifically shows the difference between **correlation** (the trending of two or more variables in the same direction) and **causation** (when a trend in one variable causes a trend to occur in one or more other variables).
  - To practice this skill, question your assumptions and ask others for their opinions. The more you actively engage with different viewpoints, the less likely you are to get stuck in a one-track mindset when analyzing data.

How to Improve Your Skills

- **4. Learn From Real-World Examples**
  - By exploring how other business professionals use data to solve problems, you can imagine what you'd do in their scenarios, evaluate the impact of their actions, and put that knowledge into practice.
  - You need to make it relevant and ask, 'Why do I care about this?' or 'Why do I want to look at a summary statistic?' or 'How is this going to be meaningful for a specific decision?'
  - By exposing yourself to cases from various industries.

- **4. Find a Community**
  - You can turn to online forums, social media, affinity groups.

Data Exploration and Analytics by Examples

Wan Fang

Southern University of Science and Technology

[Lecture Notes on Global Health by OurWorldInData]