# What is Data Science?

CAC 350

From Ch. 1: A Hands-On Introduction to Data Science by Chirag Shah

# What is Data Science?


Photo: (C)Robert Viglasky/Hartswood Films and BBC Wales for BBC One and MASTERPIECE

- Data science is "a field of study and practice that involves the collection, storage, and processing of data in order to derive important insights into a problem or phenomenon." - Chirag Shah
- "Data Science is a multidisciplinary blend of data inference, algorithm development, and technology in order to solve analytically complex problems." Frank Lo, Director of Data Science at Wayfair

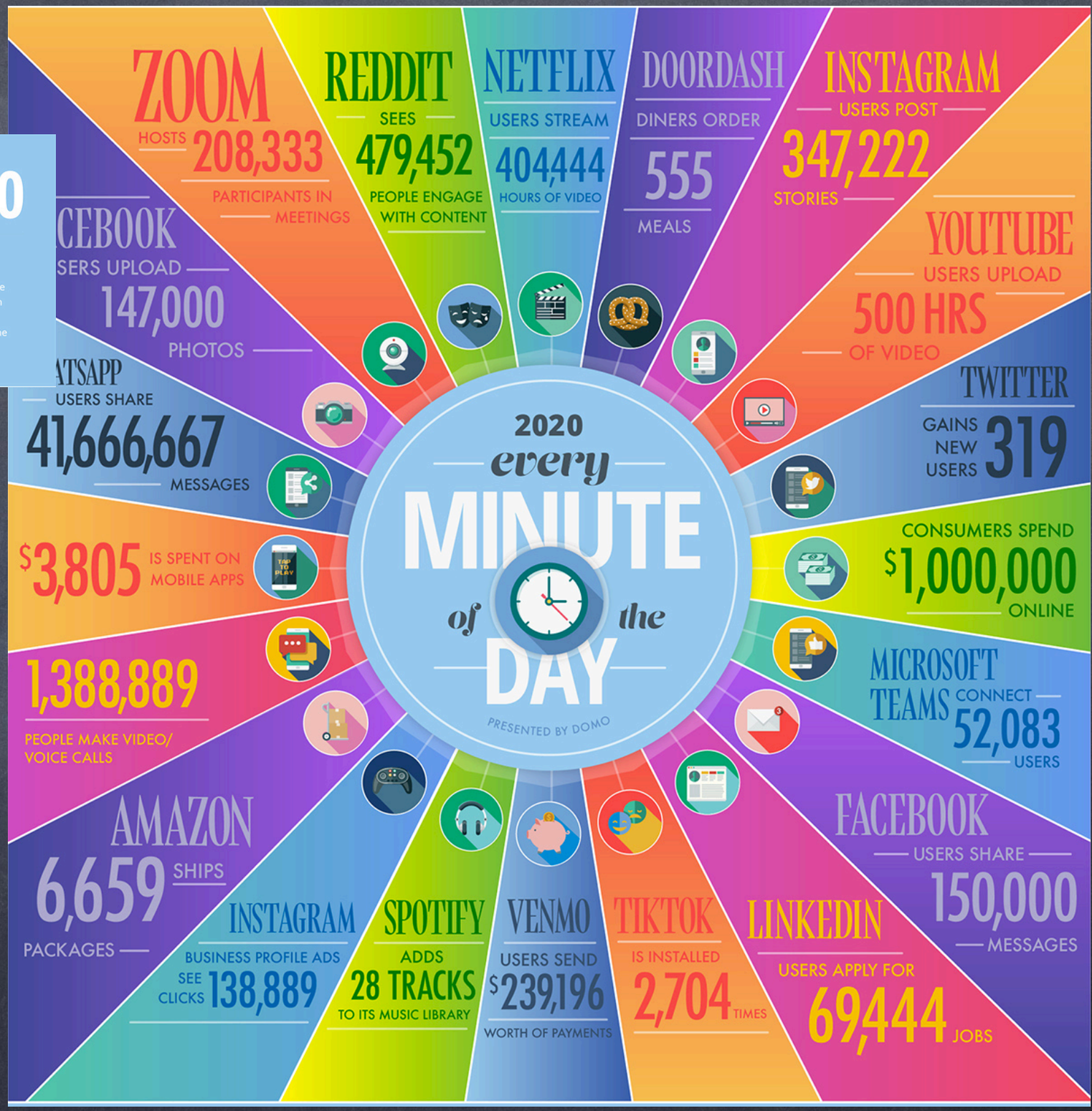
# Data...

DOMO

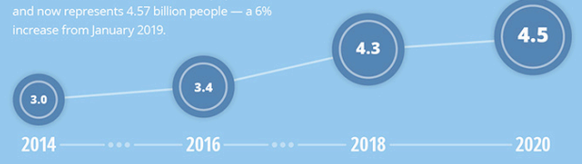
## DATA NEVER SLEEPS 8.0

How much data is generated *every minute?*

In 2020, the world changed fundamentally—and so did the data that makes the world go round. As COVID-19 swept the globe, nearly every aspect of life—from work to working out—moved online, and people depended more and more on apps and the Internet to socialize, educate and entertain ourselves. Before quarantine, just 15% of Americans worked from home. Now over half do. And that's not the only big shift. In our 8th edition of Data Never Sleeps, we bring you the latest stats on how much data is being created in every digital minute—a trend that shows no sign of stopping.




The world's internet population is growing significantly year over year. As of April 2020, the internet reaches 59% of the world's population and now represents 4.57 billion people — a 6% increase from January 2019.




GLOBAL INTERNET POPULATION GROWTH 2014–2020
(IN BILLIONS)

As the world changes, businesses need to change with the times—and that requires data. Every click, swipe, share or like tells you something about your customers and what they want, and Domo is here to help your business make sense of all of it. Domo gives you the power to make data-driven decisions at any moment, on any device, so you can make smart choices in a rapidly changing world.

**Learn more at domo.com**

**SOURCES:** STATISTA, VISUAL CAPITALIST, BUSINESS INSIDER, GAMESPOT, TECHCRUNCH, OMNICORE AGENCY, DOORDASH, BUSINESS OF APPS, NEW YORK TIMES, MUSIC BUSINESS WORLDWIDE, INC., THE VERGE, INC., HOOTSUITE, DUSTIN STOUT, REDDIT, UBER, AMAZON, VOX

DOMO

https://www.visualcapitalist.com/every-minute-internet-2020/

# 3V Model

- Velocity: The speed at which data is accumulated
- Volume: The size and scope of the data

Expected to reach 40 ZB by end of 2020, 40 billion times 1 TB
8 billion people = 5 TB of data/person

- Variety: The massive array of data and types (structured and unstructured)

Massive increase here: text, images, video

# Where do we see data science?

# Finance

- Sources: social media, mobile interactions, server logs, real-time market feeds, customer service records, transaction details, existing databases

- Build predictive models, Run real-time simulations of market events, Fraud detection, risk reduction

- Analyze customer's purchasing power to customize bank products

# Public Policy

- "The application of policies, regulations, and laws to the problems of society through the actions of government and agencies for the good of a citizenry." (Econ, PS, Sociology, etc.)
- Open repositories:
  - US government (https://www.data.gov)
  - Chicago (https://data.cityofchicago.org)
  - NYC (https://nycopendata.socrata.com)
- Data Science for Social Good project: 25 data analytics experts work to find clues to solve relevant problems impacting society

# Politics

- 2016: Trump's campaign was an example of using data science in social media to tailor individual messages to individual people
- Twitter analysis during 2016 campaign:
  - Clinton emphasized masculine traits and feminine issues
  - Trump focused on masculine issues paying no attention to traits
  - Trump used user-generated content as sources of tweets more than Clinton
  - 3/4 of Clinton's tweets were original content compared to 1/2 of Trump's (retweets and replies to citizens)

# Dark Side of DS

- Cambridge Analytica data scandal (March 2018)
- Obtained data on approximately 87 million FB users from academic researcher in order to target political ads during the 2016 US presidential campaign
- Not the first incident...advertisers, spammers, and cybercriminals use data (obtained legally or illegally) for feeding us information

# Healthcare

- COVID-19...data galore
- Data: Clinical studies, insurance information, hospital records, biological data (gene expression, next-generation DNA sequencing, proteomics, and metabolomics)
- Personal wearable health trackers (fitbit): heart rate, blood glucose, sleep patterns, stress levels, brain activity
- Apple partnered with Stanford Medicine to collect and analyze data from Apple Watch to identify irregular heart rhythms - insurance companies will provide Apple Watches to help keep patients healthier

# Urban Planning

- The Urban Center for Computation and Data (UrbanCCD), at the University of Chicago, is working to change the field of urban planning based on data

- "The consequences are seen in inefficient transportation networks belching greenhouse gases and unplanned city-scale slums with crippling poverty and health challenges. There is an urgent need to apply advanced computational methods and resources to both explore and anticipate the impact of urban expansion and find effective policies and interventions." - Charlie Catlett, UrbanCCD Director

- chicagoshovels.org - plow tracker, snow corps, bus tracker

# Education

It's called iReady, and imNotAFan

- Reading
  - Traditional method: short stories, test every other week, graded papers
  - Future: computerized software program constantly measuring and collecting data, linking to websites providing further assistance, and giving the student instant feedback and teacher

# Libraries

"Imagine that Alice, a scientist conducting research on diabetes, asks Mark, a research librarian, to help her understand the research gap in previous literature. Armed with the digital technologies, Mark can automate literature reviews for any discipline by reducing ideas and results from thousands of articles into a cohesive bulleted list and then apply data science algorithms, such as network analysis, to visualize trends in emerging lines of research on similar topics. This will make Alice's job far easier than if she had to painstakingly read all the articles."

Is DS it's own field or is it a subset of another?

Data Science and...

# Statistics

- "data-scientist is a sexed up term for a statistician." - Nate Silver, FiveThirtyEight
  - Stats: developed to help people deal with pre-computer "data problems"
  - DS: emphasizes the data problems of the 21st century (large datasets, computer code, visualizing data)
- Nathan Yau of Flowing Data suggests data scientists have at least three basic skills:
  1. A strong knowledge of basic stats and machine learning - or at least enough to avoid misinterpreting correlation for causation or extrapolating too much from a small sample
  2. CS skills to take an unruly dataset and use a programming language to make it easy to analyze
  3. Ability to visualize and express data and analysis in a meaningful way

# Computer Science

- Computer scientists have developed techniques and methods that assist in the field of DS:
  - Database systems (structured and unstructured)
  - Visualization techniques (and associated algorithms)
  - Algorithms making it possible to compute complex and heterogeneous data in less time (e.g., machine learning, pattern recognition)
- The two subjects overlap and are "mutually supportive"

# Engineering

- Engineers (chemical, civil, computer, mechanical, etc.) need data to solve problems
- Mutually supportive: DS develops method and techniques, Engineering has developed hardware (e.g., CPU, GPU)
- "Smart" building techniques:
  - Predictive algorithms - better cost estimates (location, time of year, total value, relevant cost indices, etc. )
  - 3D printing to predict weak spots
  - Drones for monitoring

# Business Analytics

- "refers to the skills, technologies, and practices for continuous iterative exploration and investigation of past and current business performance to gain insight and be strategic"
- Need explanative and predictive modeling, fact-based management for decision-making
- Four types of analytics:

1. Decision analytics: supports decision-making with visual analytics that reflect reasoning
2. Descriptive analytics: provides insight from historical data with reporting, score cards, clustering, etc.
3. Predictive analytics: employs predictive modeling using statistical and machine learning techniques
4. Prescriptive analytics: recommends decisions using optimization, simulations, etc.

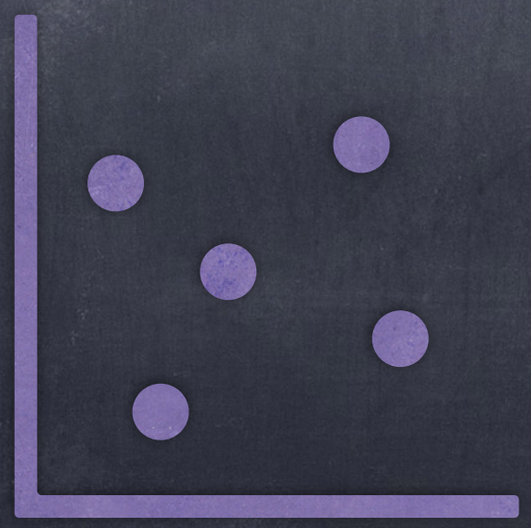
# Social & Computational Social Science

- anthropology, archaeology, economics, linguistics, political science, psychology, public health, and sociology
- Computational social science helps connect theories or results from one discipline to another
  - "How will the information revolution in this digital age transform society?"
- Government policies, people's mandates in elections, and hiring strategies are examples of applications of computational social science

# Data Science = Information Science?

# Data vs. Information

- How do you define these two concepts?
- I have a number: 523...data? information?



# Data Scientists

- Interested in investigating the characteristics of data
- Look for patterns that reveal how people and society can benefit from the data
- What's missing? The processes and people behind the data (more system side, phenomena focused). Where is the user's perspective?

# Information Scientists

- Look at data in the context they are generated and used
- Bridges the gap between quantitative analysis and an examination of data that tells a story
- Focus on the human side of data and information, in addition to the system perspective
- Scholars in IS tend to combine the user and system sides to understand how and why data is generated and the info they convey within a given context

In summary:
Data Science: Technical
Information Science: Practical + Human

Alternative definition of Information Science - the storage and retrieval of data...how you manage the information

# Computational Thinking

- "thinking like a computer scientist"
- "Computational thinking is using abstraction and decomposition when attacking a large complex task or designing a large complex system" - Jeannette Wing, CMU


- Iterative process:
  1. Problem formulation (abstraction)
  2. Solution expression (automation)
  3. Solution execution and evaluation (analyses)

# Example

Find the largest number:

7, 24, 62, 11, 4, 39, 42, 5, 97, 54

Let's do it systematically....

# Example

- First, we used decomposition
- The process we used could be applied to 100 numbers or more, abstraction and generalization
- Abstraction: treating the actual object of interest (10 numbers) as a series of numbers
- Generalization: ability to devise a process that is applicable to the abstracted quantity (series of numbers) and not the specific objects (10 numbers)
- Sorting is another excellent example - lots of strategies there as you may have seen in 210

# Skills for Data Science

- From Twitter: "Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician."

1. Willing to experiment
2. Proficiency in mathematical reasoning
3. Data literacy (ability to extract meaningful information from a dataset)

# Example

- Let's work on a data-driven problem:
- start with a problem, identify a data source, collect data, clean the data, analyze the data, and present our findings

# Example

- What can you tell me about this data?
- Any easy stats?
- Relationships?
- On average, how much increase in weight with 1" increase in height?
- Is it uniform?
- What about someone 57"? 73"?

| Observation | Height (in) | Weight (lbs) |
|---|---|---|
| 1 | 58 | 115 |
| 2 | 59 | 117 |
| 3 | 60 | 120 |
| 4 | 61 | 123 |
| 5 | 62 | 126 |
| 6 | 63 | 129 |
| 7 | 64 | 132 |
| 8 | 65 | 135 |
| 9 | 66 | 139 |
| 10 | 67 | 142 |
| 11 | 68 | 146 |
| 12 | 69 | 150 |
| 13 | 70 | 154 |
| 14 | 71 | 159 |
| 15 | 72 | 164 |

# Tools

- No special tools but definitely some more suitable than others
- Any programming language would do, but you can plot a graph in Python in one line...it takes more in C or Java
- R is very popular as well
- Datasets can be stored in CSV and load in into Python/R...what's the issue?
- SQL Database - can access through Python/R
- UNIX - allows one to solve many data problems and data processing needs without writing any code

# Ethics, Bias, and Privacy

- Anytime you're dealing with data, you have to consider these topics.
- Politics and media - excellent example
- How, where, and why was the data collected?
- Who collected it?
- What did they intend to use it for?
- If the data was collected from people, did these people know that:
  - such data was being collected?
  - how the data would be used?

# Origin of Data

- Availability of data != Right to use

> "if you are not paying for it, you are the product"

- Google blog: https://ai.google/responsibilities/responsible-ai-practices/
- https://www.wired.com/story/prominent-ai-ethics-researcher-says-google-fired-her/

# Issues: Ethics, Bias, and Privacy

- Anytime you're dealing with data, you have to consider these topics.
- Politics and media - excellent example

# What are you worth?

| Company | Value/User |
| --- | --- |
| Facebook | $158 |
| Google | $182 |
| Amazon | $733 |

# Ethical Data

- If the data is ethical, that means, we're good, right?

- If a data scientist is not careful, inherent bias in the data can show up in the analysis and the insights developed

# Next Class

- Participation activity on Moodle - individual or a partner
- Download Jupyter Notebook, can be with Anaconda - make sure you get the Python interpreter
- Get the books, we'll go over Ch. 1 in the Python book on Thursday - reading ahead is highly encouraged