ENSF 612: Fall 2021
Lecture 2. Data Analysis and Big Data

Dr. Gias Uddin, Assistant Professor
Electrical and Software Engineering
Schulich School of Engineering
University of Calgary
https://giasuddin.ca/

Topics

- Course objectives - revisited
- Data analysis - definition
- Big data - definition
- Sources of big data
- Big data case studies

Course Objectives Revisited

What will we learn in this course?

- What is data analysis (data science)?
- What is “big” data and big data analysis?
- What are sources of big data?
- Why is there so much excitement about it?
- What are the benefits of analyzing big data?
- What are some platforms available to develop algorithms to analyze big data?
- How to write big data programs on these platforms?
- How to develop data-driven models for prediction?

Course Objectives Revisited

- What will we learn in this course?
  - **What is data analysis (data science)?**
  - What is “big” data and big data analysis?
  - What are sources of big data?
  - Why is there so much excitement about it?
  - What are the benefits of analyzing big data?
  - What are some platforms available to develop algorithms to analyze big data?
  - How to write big data programs on these platforms?
  - How to develop data-driven models for prediction?

Data Analysis

- Data analysis or data science - been around for a long time
- Data science definition
  - Exploit data to find useful patterns and information
  - Involves methods to
    - **Collect** data
    - **Inspect quality** of data
    - **Extract subsets** of data
    - **Transform** data
    - Do **Exploratory Data Analysis**
    - **Build models** from data

Data Analysis - Example

- Illustrate data analysis process with an example
- Deerfoot Trail traffic analysis
- City planners interested in following questions
  - What is commute time from point A to B on the highway?
  - How does this commute time change over the day?
  - How is the time impacted by factors weather and accidents?
  - What are the most congested sections of the highway?
  - How do these sections change with time and season?
  - Can we predict future commute times?
- **What are the various steps in this analysis?**

Data Analysis - Example

- Figure out data sources – collect data
  - Commute time – road sensors, Google maps
  - Weather – Environment Canada
  - Accidents – police reports, tweets from users
- Inspect data quality
  - Check for duplicate or missing or incorrect data
  - E.g., no +30 C days in January ☺
  - Does sensor data agree with Google maps?

Data Analysis - Example

- Extract subset of data
  - Which year do we want to study?
  - Which roads do we want to study?
  - Which seasons do we want to study?
- Transform data
  - Merge data – transform it to format good for analysis tools
  - E.g., Excel likes data in comma-separated format
  - *Date, Time, Highway section, commute time, # of accidents, temperature, snow on ground,…….*

Data Analysis - Example

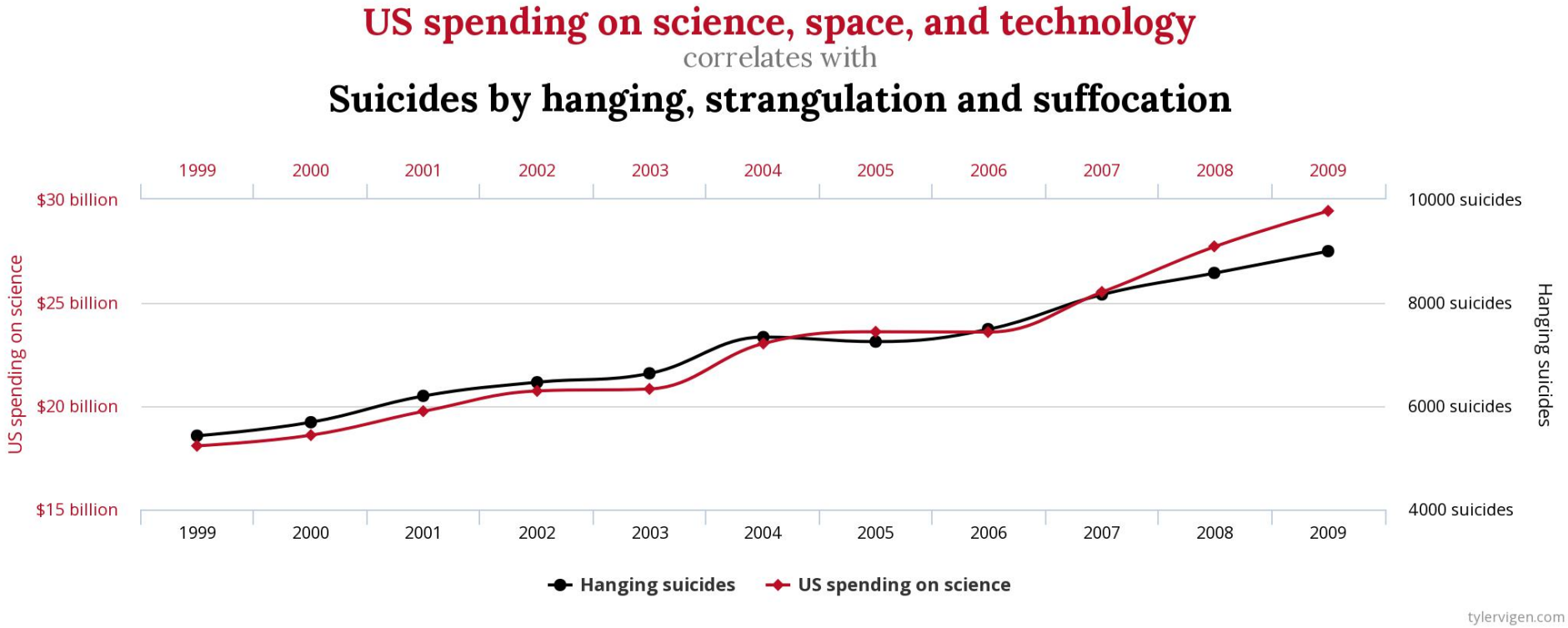
- Do preliminary or exploratory data analysis (EDA)
  - Compute statistics such as mean and median
  - Visualize data for weekends, weekdays, winter, etc.
  - EDA can help understand how to answer our questions
  - EDA can help us build **models**
- Build models
  - Predict output var as a function of input variables or *features*
  - Example
    - Output variable – commute time
    - Features – day, time, accidents, weather
  - Models need to be **trained** and **validated**
- **Data analysis process is typically iterative**

Data Analysis Pitfalls

- Trap of falling in to "correlation is causation"
- Say variable *a* and *b* are positively correlated
- Temptation might be to conclude
  - "as a increases b increases"
  - "as a decreases b decreases"
- This might not be necessarily true!
- Let's look at a few examples

Data Analysis Pitfalls – fun graphs

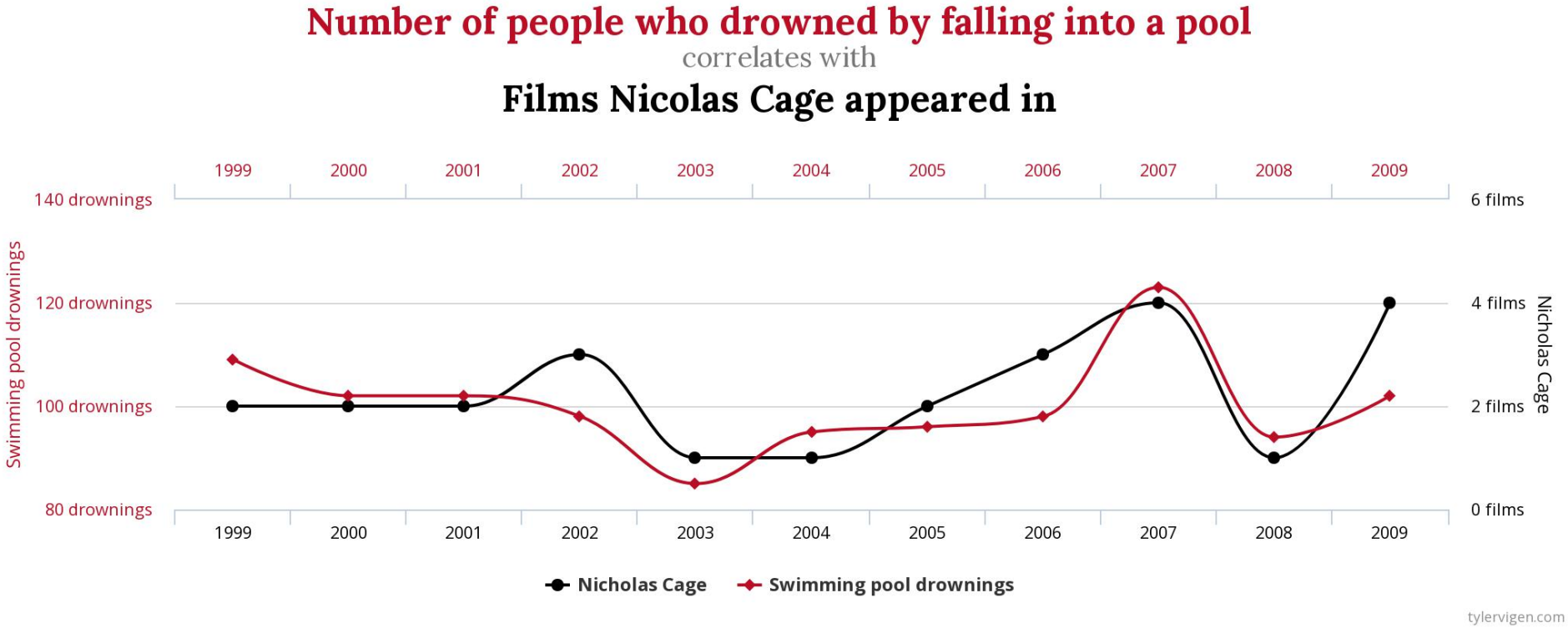
Correlation is not causation



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

Data Analysis Pitfalls – fun graphs

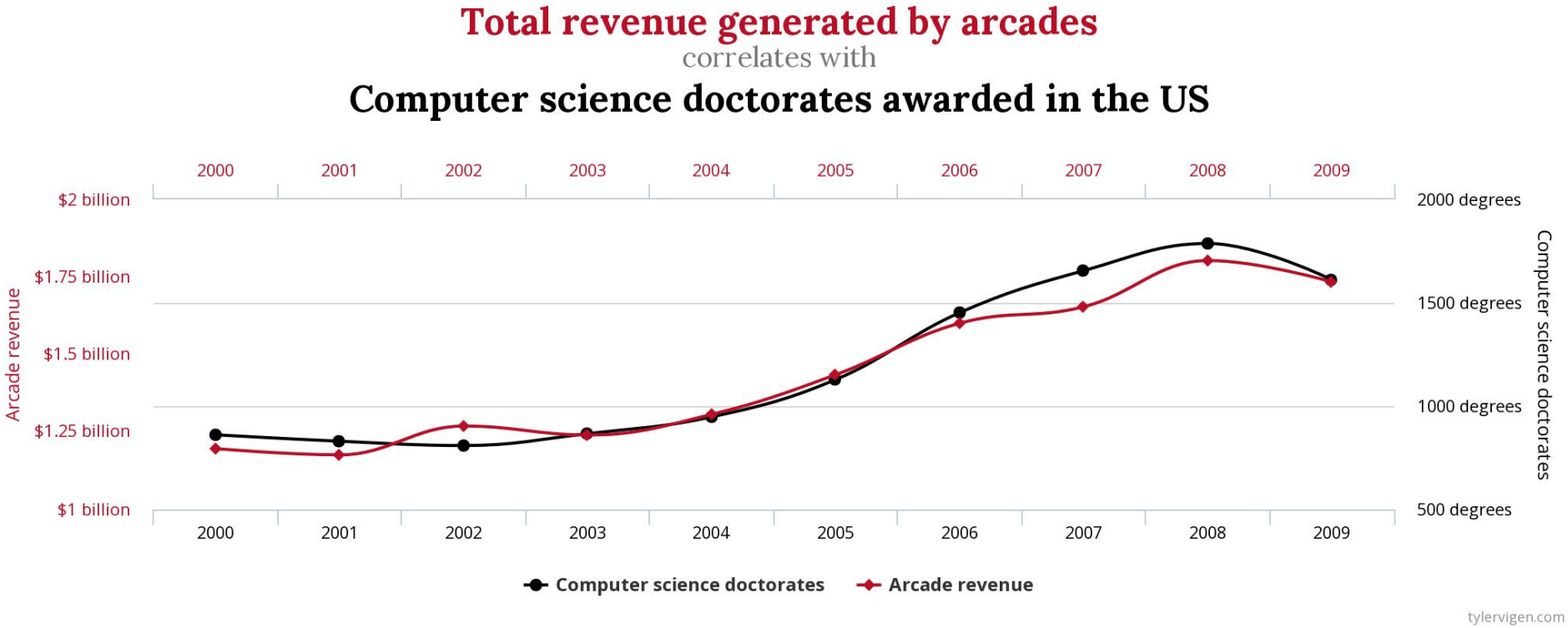
Correlation is not causation



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

Data Analysis Pitfalls – fun graphs

Correlation is not causation



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

Data Analysis Pitfalls – when it becomes serious

Correlation is not causation - Seven Countries Study by Ancel Keys

"Risk and rates of heart attack and stroke (CVR), both at the population level and at the individual level, correlated directly and independently to the level of total serum cholesterol"

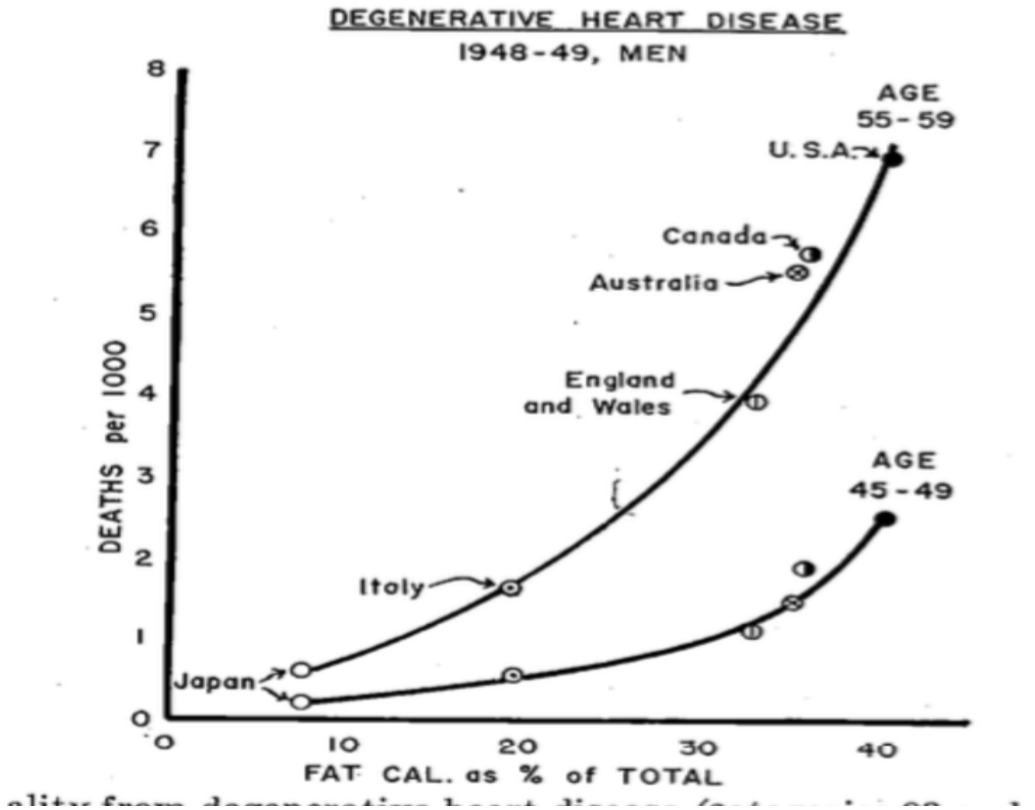
- Started in 1958 – followed

  13,000 subjects over 5-40 years

- Significant controversy

- Failed to consider other factors –

  e.g., sugar consumption



https://en.wikipedia.org/wiki/Seven_Countries_Study

Data Analysis Pitfalls – when it becomes serious

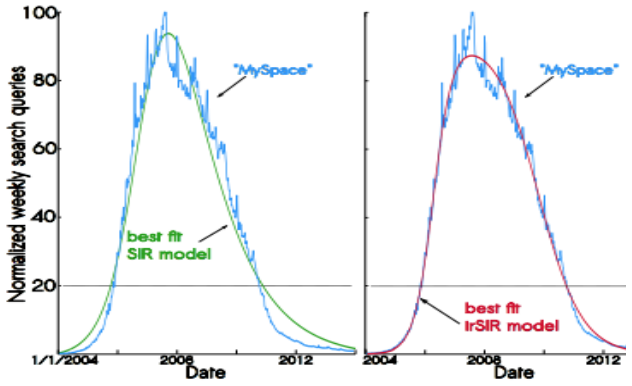
Correlation is not causation – Facebook Doomsday Prediction

- "Epidemiological modeling of online social network dynamics" by Cannarella and Spechler from Princeton
- From the abstract of the paper: "*Extrapolating the best fit model into the future predicts a rapid decline in Facebook activity in the next few years*"
- **How did they arrive at this conclusion?**
- Paper available at http://arxiv.org/abs/1401.4208

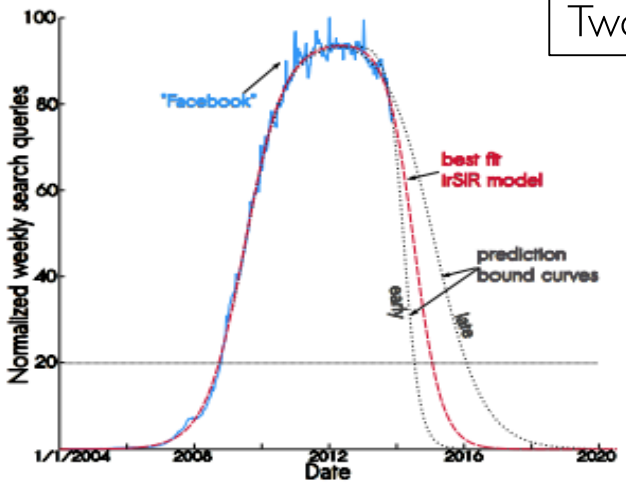
Data Analysis Pitfalls – when it becomes serious

Correlation is not causation – Facebook Doomsday Prediction

Google Trends searches for "MySpace"



ta for search query "Myspace" with best fit (a) SIR an… …
search query data are normalized such that the maximum data poi…t corresponds to a



Two Figures from the paper

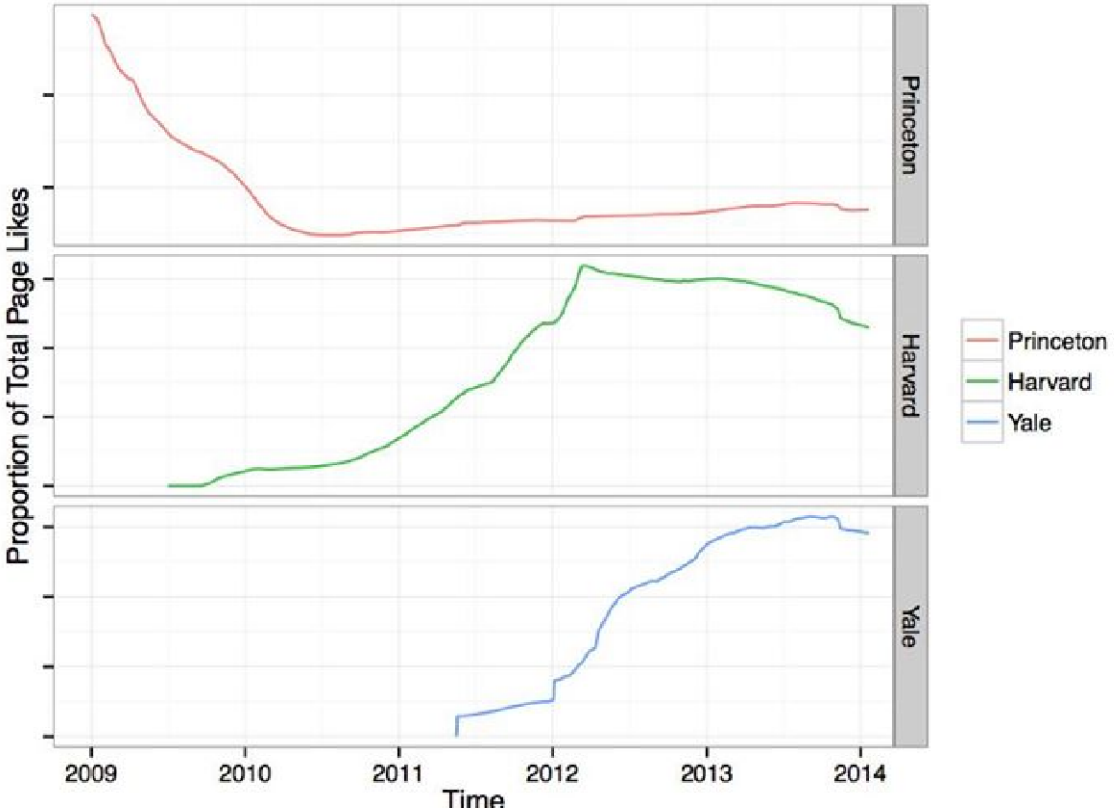
Searches for "Facebook"

- Declining search correlates to declining popularity
- Falling prey to correlation is causation!

Data Analysis Pitfalls – when it becomes serious

Correlation is not causation – Facebook Doomsday Prediction

Facebook issued a tongue in cheek rejoinder to study



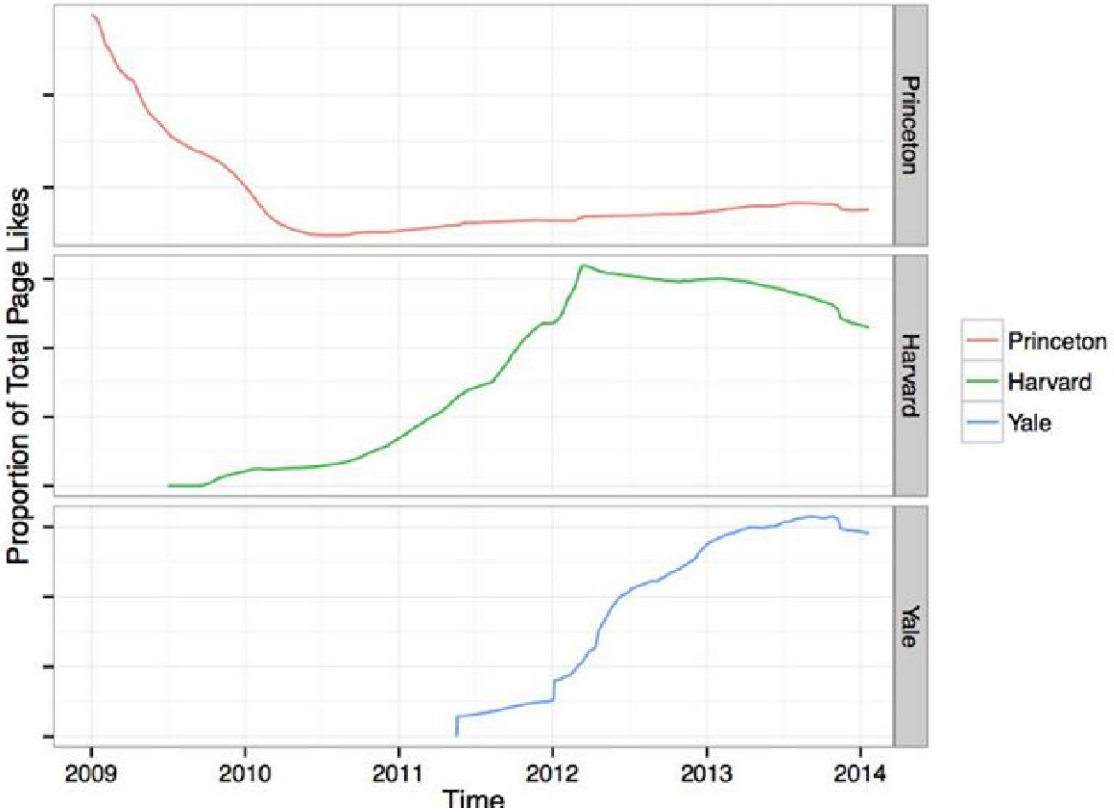
In keeping with the scientific principle "correlation equals causation," our research unequivocally demonstrated that Princeton may be in danger of disappearing entirely."

Study at https://www.facebook.com/notes/mike-develin/debunking-princeton/10151947421191849/

Data Analysis Pitfalls – when it becomes serious

Correlation is not causation – Facebook Doomsday Prediction

Facebook issued a tongue in cheek rejoinder to study



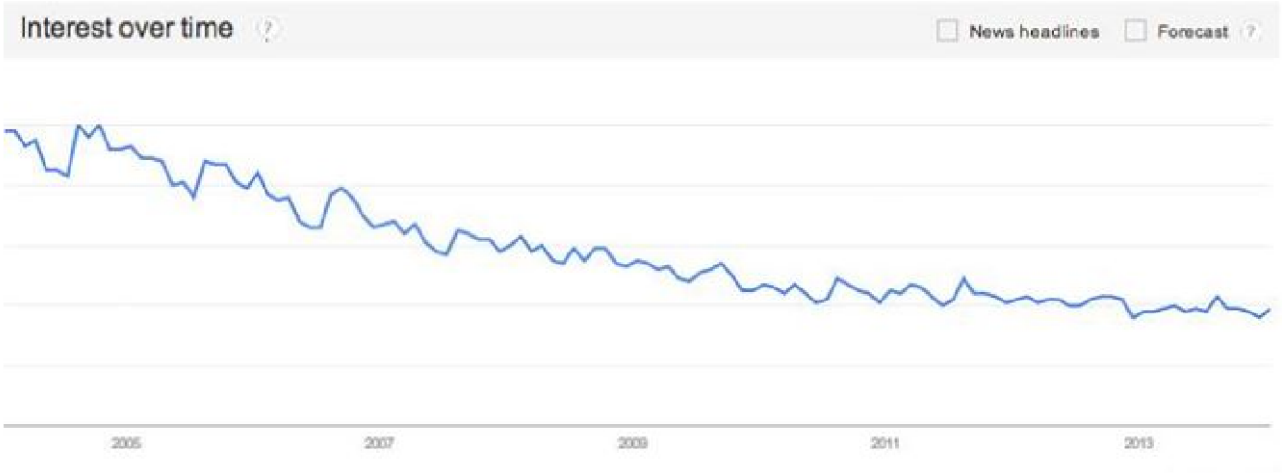
In keeping with the scientific principle "correlation equals causation," our research unequivocally demonstrated that Princeton may be in danger of disappearing entirely."

Study at https://www.facebook.com/notes/mike-develin/debunking-princeton/10151947421191849/

Data Analysis Pitfalls – when it becomes serious

Correlation is not causation – Facebook Doomsday Prediction

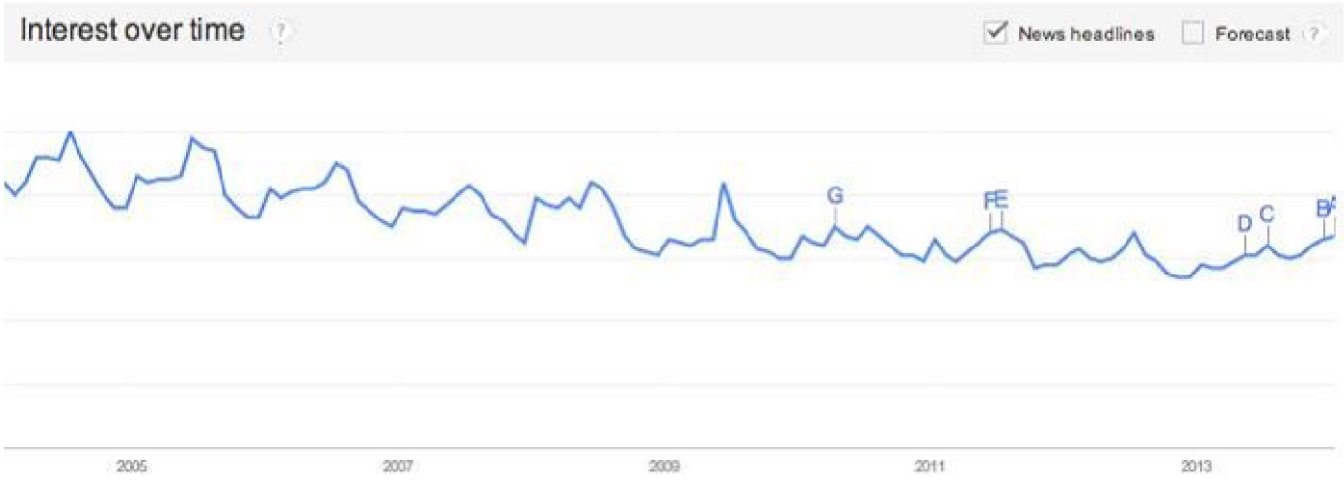
… and based on Princeton search trends:"
"
"This trend suggests that Princeton will have only half its current enrollment by 2018, and by 2021 it will have no students at all,…""



Data Analysis Pitfalls – when it becomes serious

Correlation is not causation – Facebook Doomsday Prediction

> While we are concerned for Princeton University, we are even more concerned about the fate of the planet — Google Trends for "air" have also been declining steadily, and our projections show that by the year 2060 there will be no air left:"



Course Objectives Revisited

- What will we learn in this course?
  - What is data analysis (data science)?
  - **What is "big" data and big data analysis?**
  - What are sources of big data?
  - Why is there so much excitement about it?
  - What are the benefits of analyzing big data?
  - What are some platforms available to develop algorithms to analyze big data?
  - How to write big data programs on these platforms?
  - How to develop data-driven models for prediction?

Big data refers to exponential growth in data. Some stat during COVID - 19

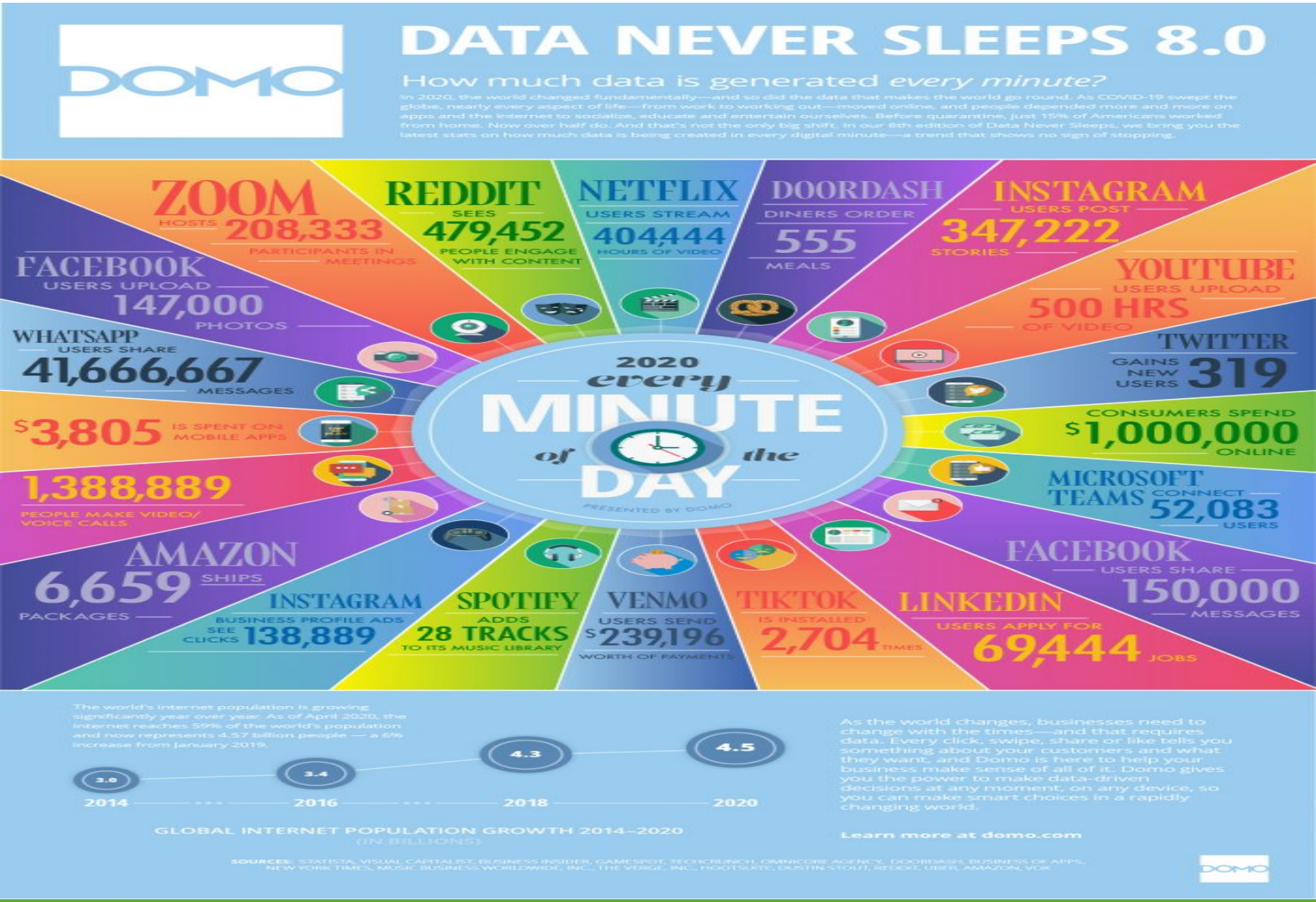
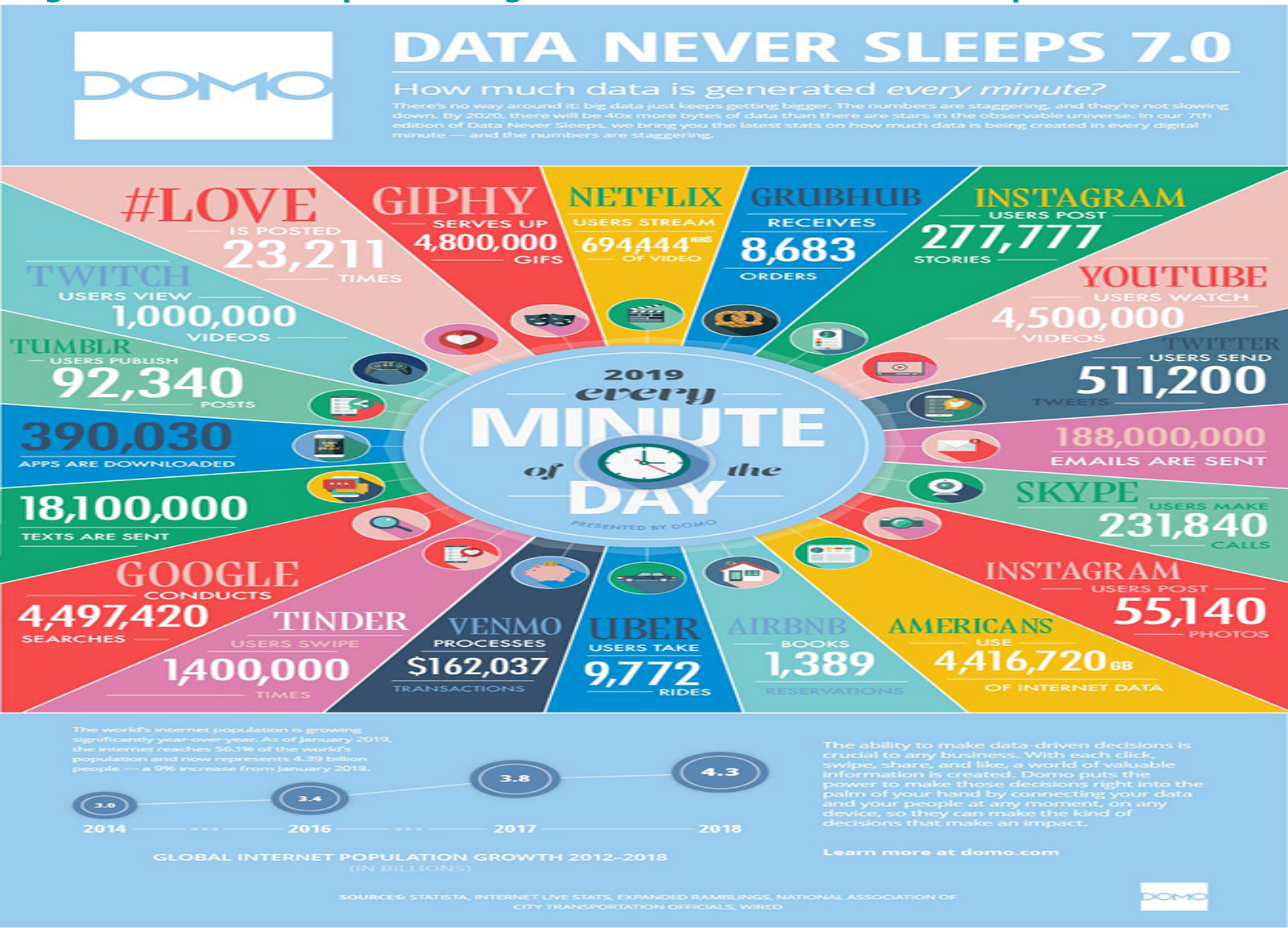
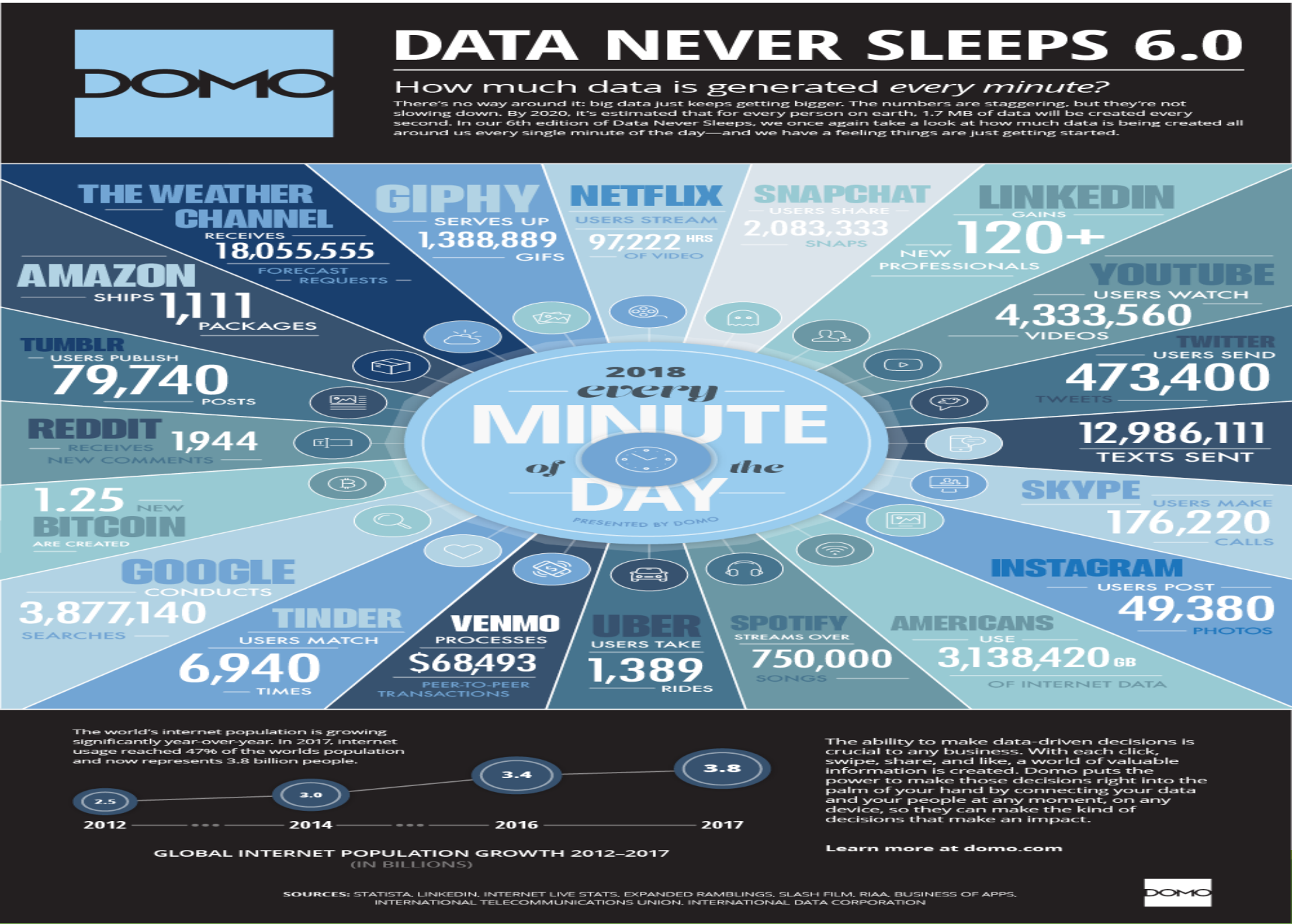


Image: https://www.domo.com/learn/data-never-sleeps-8

Big data refers to exponential growth in data. Data never sleeps 2019 Edition



Big data refers to exponential growth in data. Data never sleeps 2018 Edition



The world's internet population is growing significantly year-over-year. In 2017, internet usage reached 47% of the worlds population and now represents 3.8 billion people.

GLOBAL INTERNET POPULATION GROWTH 2012–2017
(IN BILLIONS)

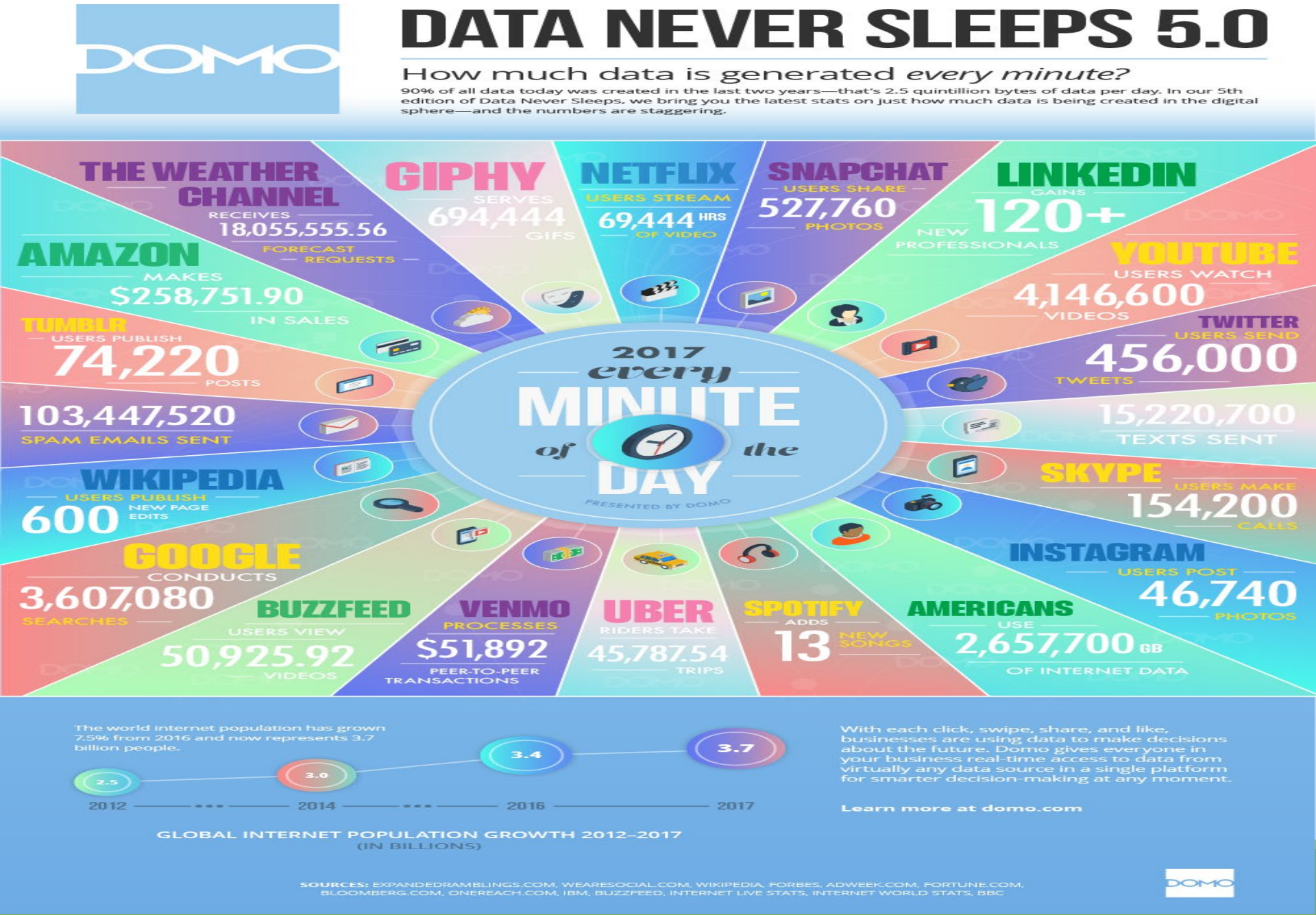
The ability to make data-driven decisions is crucial to any business. With each click, swipe, share, and like, a world of valuable information is created. Domo puts the power to make those decisions right into the palm of your hand by connecting your data and your people at any moment, on any device, so they can make the kind of decisions that make an impact.

Learn more at domo.com

SOURCES: STATISTA, LINKEDIN, INTERNET LIVE STATS, EXPANDED RAMBLINGS, SLASH FILM, RIAA, BUSINESS OF APPS, INTERNATIONAL TELECOMMUNICATIONS UNION, INTERNATIONAL DATA CORPORATION

DOMO

Big data refers to exponential growth in data. Data never sleeps 2017 Edition



Big data refers to exponential growth in data

- How big is big data?
  - Social networks
    - Twitter – 316 million users; 500 million tweets/ day
    - Facebook – 968 million users; 55 million status updates/day
  - Search
    - Google – 30 trillion pages indexed – 10^8 gigs of index data
    - Google uses 1 million compute hours to build index
  - Science
    - The square kilometer array – 700 TB/sec data to be persisted
  - More examples will follow later…..
- **How does one define big data?**

Big data refers to exponential growth in data

- Definition of big data?
  - Many definitions exist
  - Coarse definition for this course
    - "*Any dataset that doesn't fit reasonably in a single computer*"
  - Problematic definition since computer specs keep changing
    - Sample high end computer
    - *HP Superdome – 16 Xeon processors – 3 TB RAM*
  - Data will expand to make single computer inadequate

Course Objectives Revisited

- What will we learn in this course?
  - What is data analysis (data science)?
  - What is "big" data and big data analysis?
  - **What are sources of big data?**
  - **Why is there so much excitement about it?**
  - **What are the benefits of analyzing big data?**
  - What are some platforms available to develop algorithms to analyze big data?
  - How to write big data programs on these platforms?
  - How to develop data-driven models for prediction?

Sources of Scientific big data

CERN Large Hadron Collider

Greater than Wikipedia content, Tweets/day, library of congress

CERN Data Centre passes the 200-petabyte milestone

The CERN Data Centre passed a major milestone on 29 June 2017 with more than 200 petabytes of data now archived on tape

6 JULY, 2017 | By Mélissa Gaillard

CERN's Data Centre (Image: Robert Hradil, Monika Majer/ProStudio22.ch)

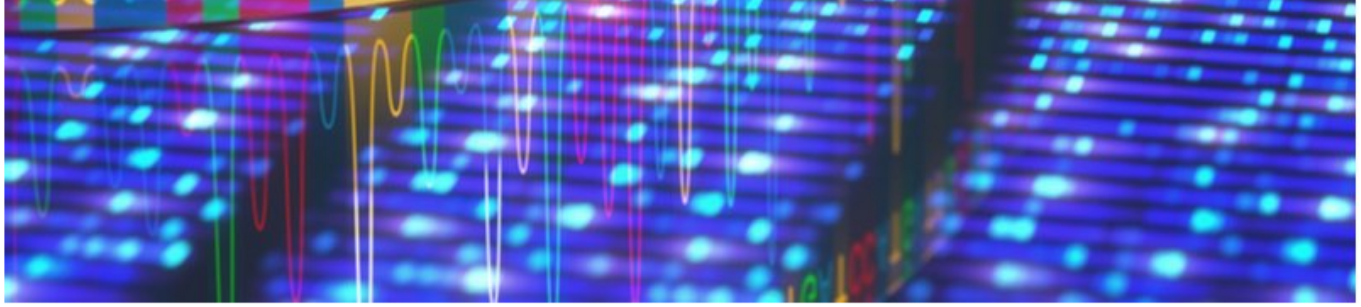
On 29 June 2017, the CERN DC passed the milestone of 200 petabytes of data permanently archived in its tape libraries. Where do these data come from? Particles collide in the Large Hadron Collider (LHC) detectors approximately 1 billion times per second, generating about one petabyte of collision data per second. However,

https://home.cern/news/news/computing/cern-data-centre-passes-200-petabyte-milestone

Sources of Scientific big data

Genome Sequencing Data

large genome databases can be used to detect diseases



According to the Global Alliance for Genomics and Health, more than 100 million genomes will have been sequenced in a healthcare setting by 2025. Most of these genomes will be sequenced as part of large-scale genomic projects stemming from both big pharma and national population genomics initiatives. These efforts are already garnering immense quantities of data that are only likely to increase over time. With the right analysis and interpretation, this information could push precision medicine into a new golden age.

Are we ready to deal with enormous quantities of data?

Genomics is now considered a legitimate big data field – just one whole human genome sequence produces approximately 200 gigabytes of raw data. If we manage to sequence 100M genomes by 2025 – we will have accumulated over 20B gigabytes of raw data. The massive amount of data can partially be managed through data compression technologies, with companies such as Petagene, but that doesn't solve the whole problem.

What's more, sequencing is futile unless each genome is thoroughly analyzed to achieve meaningful scientific insights. Genomics data analysis normally generates an additional 100 gigabytes of data per genome for downstream analysis, and requires massive computing power supported by large computer clusters – a feat that is economically unfeasible for the majority of companies and institutions.

https://www.labiotech.eu/genomics/cloud-genomics-big-data-problem/

Sources of Scientific big data

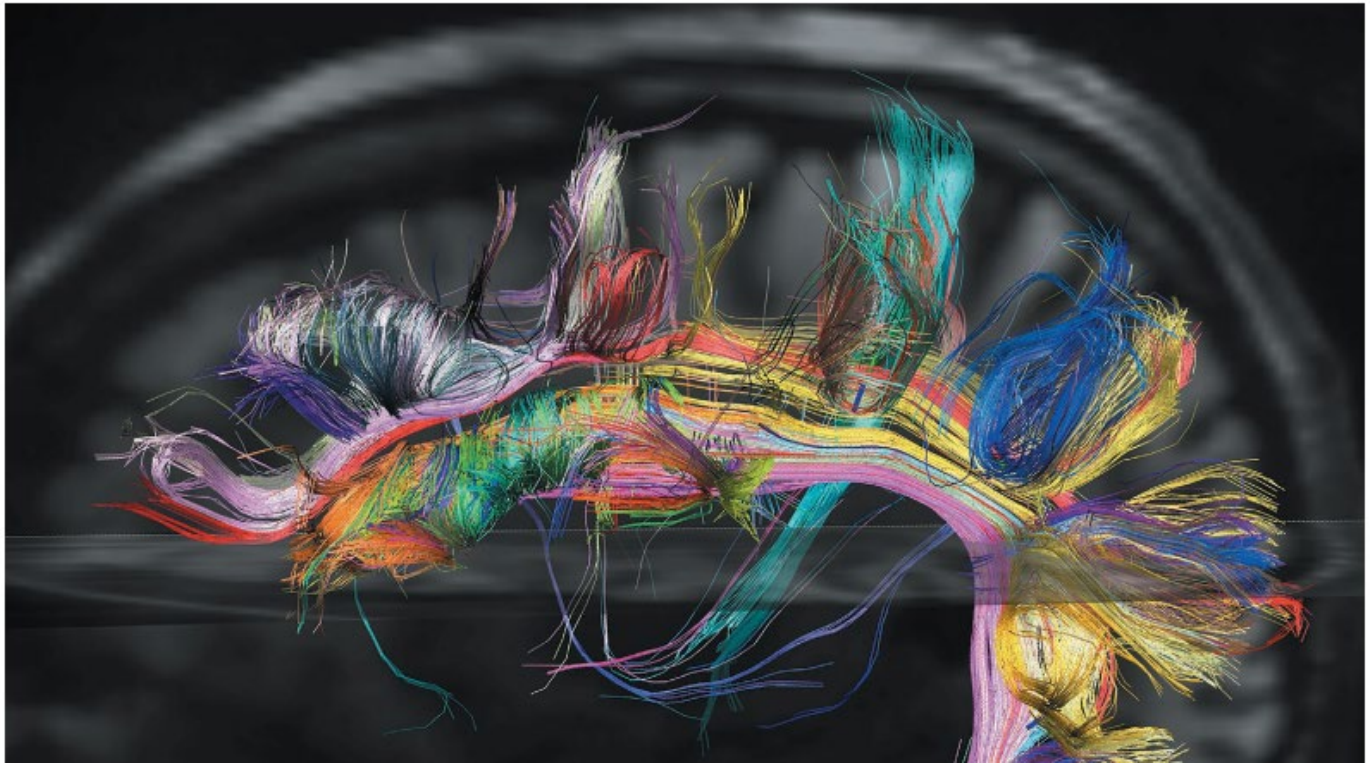
Brain Data

Brains have neurons

Each neuron cell needs around 1GB for imaging

BIG BRAIN, BIG DATA

Neuroscientists are starting to share and integrate data — but shifting to a team approach isn't easy.



Fruit fly: 60,000 neurons = 60,000 GB = 60 TB (Terabyte)
Mouse: 80M neurons = 80 Thousand TB = 80 PB (Petabyte)
Human: 86B neurons = 86,000M = 86,000 PB = 86 EB (Exabyte)

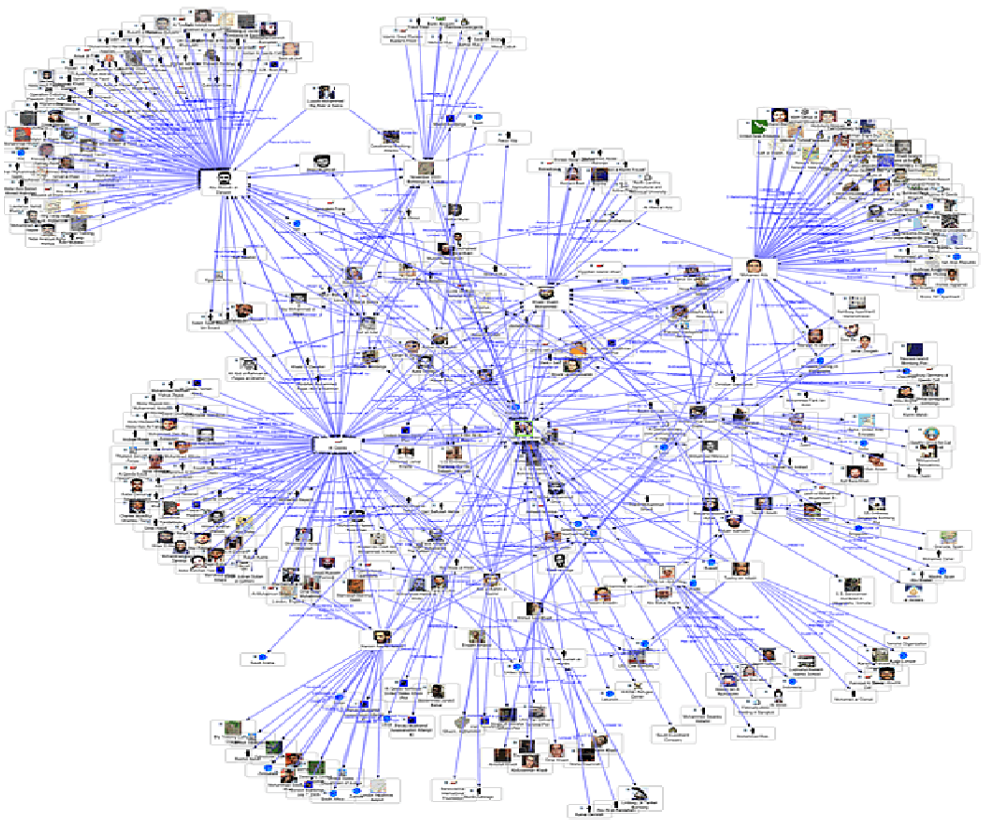
https://www.nature.com/articles/541559a

Everyday Web Usage

- Actions of Web users
  - Looked at Twitter and Facebook earlier
  - Every click, pause/stop/play, ad click stored
  - Data can be analyzed for profit
    - Targeting ads to users (Google)
    - Recommender systems (Netflix, Amazon)
- Content generated by Web users
  - Tweets, wall posts, pictures posted
  - Individually not much – but together can mean a lot
    - Facebook sells mined info on users to others
    - Graph analytics – identify "communities" in social networks

Data Center Log Files, Social Networks

- Log files from computers in data centers
  - Detection of security and performance problems
- Graph analytics
  - Mining graphs for patterns
  - E.g., friend recommendations on Facebook
  - E.g., commute time prediction in road networks



Sources of Big Data

Stack Exchange Online Technical Forums

- Across all of Stack Overflow and the Stack Exchange network, we saw **9+ billion pageviews** from **100+ million users** over the course of the year.

company JANUARY 18, 2019

State of the Stack 2019: A Year in Review

A loooong time ago, we used to post an annual "State of the Stack" update on the company and community. Then at some point it became an infographic which was… listen, everyone was doing infographics in 2011. Now it's 2019, the company has grown and changed in so many ways, so we're bringing this tradition…

David Fullerton
President and Chief Technology Officer (former)

StackExchange

All Sites Top Users Digests

All Technology Culture / Recreation Life / Arts Science Professional Business

Sort by: Traffic

Stack Overflow Q&A for professional and enthusiast programmers	20m questions	30m answers	70% answered	13m users	9.8m visits/day	6.7k questions/day	12y2m site age
Super User Q&A for computer enthusiasts and power users	449k questions	646k answers	66% answered	932k users	654k visits/day	119 questions/day	11y2m site age

twitter

Sources of Big Data

100+ millions GitHub software repositories

40 m+

developers on GitHub, including 10M new users in 2019.*

44 m+

repositories created in the last year—and 44% more developers created their first repository in 2019 than in 2018.*

87 m+

pull requests merged in the last year—and 28% more developers opened their first pull request in 2019 than in 2018.*

20 m+

issues closed in the last year. That's a lot of decisions made, bugs fixed, and boxes checked.*

https://octoverse.github.com/

Success Stories Using Big Data Analytics

- Google's flu trends
  - Traditional methods of predicting outbreak are slow
    - Manually collect reports from clinics/hospitals
  - By the time an alert is issued, outbreak has happened
  - Google predicted outbreak 2 weeks before government
  - How did they do it?
    - Start with 50 million search queries from 2003-2008
    - Identify 45 terms related to flu used by people with flu
    - Use these terms as features in a model
    - Model predicts flu cases on a week by week basis
    - Model found to be 97% accurate

Link: https://www.youtube.com/watch?v=6111nS66Dpk

Success Stories Using Big Data Analytics

- President Obama's reelection campaign
  - Big data analytics credited for successful campaign
  - Data models for targeting ads and fundraising
  - Algorithms to identify electors on the fence
- Interview with the data science team:

  http://poy.time.com/2012/12/19/obamas-data-team/

Success Stories Using Big Data Analytics

- Mining Twitter for customer service
- Restaurant used big data analysis to satisfy a customer
- More on the story in these links
- http://searchcio.techtarget.com/opinion/Ten-big-data-case-studies-in-a-nutshell
- http://shankman.com/the-best-customer-service-story-ever-told-starring-mortons-steakhouse/

Topics

- Course objectives - revisited
- Data analysis - definition
- Big data - definition
- Sources of big data
- Big data case studies

Acknowledgements

Portions of these slides were adapted from external material available under creative commons license CC-BY-NC-SA 4.0. This license grants the ability to share and adapt the material for non-commercial purposes.

Name of the creator: Dr. Anthony Joseph, University of Berkeley and team

License notice: https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode

Copyright notice: CC-BY-NC-SA 4.0

Link to material: https://courses.edx.org/courses/BerkeleyX/CS100.1x/1T2015/info