UMD DATA605: Big Data Systems

Lesson 1.2: Introduction to Big Data

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

Data Science

- **Promises of data science**
  - Give a competitive advantages
  - Make better strategic and tactical business decisions
  - Optimize business processes
- **Data science is not new**, it was called:
  - Operation research (~1970-80s)
  - Decision support, business intelligence (~1990s)
  - Predictive analytics (Early 2010s)
  - ...
- **What has changed**
  - Now learning and applying data science is *easy*
    - No need for hiring a consulting company
  - Tools are *open-source*
    - E.g., Python + pydata stack (numpy, scipy, Pandas, sklearn)
  - *Large data sets* available
  - *Cheap computing*
    - E.g., cloud computing (AWS, Google Cloud), GPUs

Motivation: Data Overload

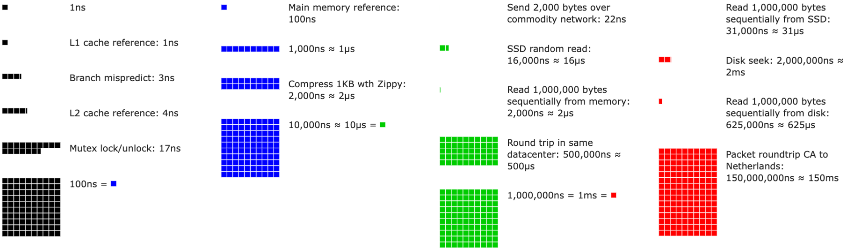
- *"Data science is the number one catalyst for economic growth"* (McKinsey, 2013)
- **Explosion of data in every domain**
  - Sensing devices/networks monitor processes 24/7
    - E.g., temperature of your room, your vital signs, pollution in the air
  - Sophisticated smart-phones
    - 80%+ of the world population has a smart-phone
  - Internet and social networks make it easy to publish data
  - Internet of Things (IoT): everything is connected to the internet
    - E.g., power supply, toasters
  - Datafication turns all aspects of life into data
    - E.g., what you like/enjoy turned into a stream of your "likes"
- **Challenges**
  - How to handle the increasing amount data?
  - How to extract actionable insights and scientific knowledge from data?

Scale of Data Size

- **Megabyte** $= 2^{20} \approx 10^{6}$ bytes
  - Typical English book
- **Gigabyte** $= 2^{30}$ bytes $=$ 1,000 MB
  - 1/2 hour of video
  - Wikipedia (compressed, no media) is 22GB
- **Terabyte** $=$ 1 million MB
  - Human genome: ~1 TB
  - 100,000 photos
  - \$50 for 1TB HDD, \$23/mo on AWS S3
- **Petabyte** $=$ 1000 TB
  - 13 years of HD video
  - \$250k/year on AWS S3
- **Exabyte** $=$ 1M TB
  - Global yearly Internet traffic in 2004
- **Zettabyte** $=$ 1B TB $= 10^{21}$ bytes
  - Global yearly Internet traffic in 2016
  - Fill 20% of Manhattan, New York with data centers
- **Yottabytes** $= 10^{24}$ bytes
  - Yottabyte costs \$100T
  - Fill Delaware and Rhode Island with a million data centers
- **Brontobytes** $= 10^{27}$ bytes

Constants Everybody Should Know

- CPU at 3GHz: 0.3 ns per instruction
- L1 cache reference/register: 1 ns
- L2 cache reference: 4 ns
- Main memory reference: 100 ns
- Read 1MB from memory: 20-100 us
- SSD random read: 16 us
- Send 1KB over network: 1 ms
- Disk seek: 2 ms
- Packet round-trip CA to Netherlands: 150 ms



1ns

L1 cache reference: 1ns

Branch mispredict: 3ns

L2 cache reference: 4ns

Mutex lock/unlock: 17ns

100ns = 🟦

Main memory reference: 100ns

1,000ns ≈ 1μs

Compress 1KB wth Zippy: 2,000ns ≈ 2μs

10,000ns ≈ 10μs = 🟩

Send 2,000 bytes over commodity network: 22ns

SSD random read: 16,000ns ≈ 16μs

Read 1,000,000 bytes sequentially from memory: 2,000ns ≈ 2μs

Round trip in same datacenter: 500,000ns ≈ 500μs

1,000,000ns = 1ms = 🟥

Read 1,000,000 bytes sequentially from SSD: 31,000ns ≈ 31μs

Disk seek: 2,000,000ns ≈ 2ms

Read 1,000,000 bytes sequentially from disk: 625,000ns ≈ 625μs

Packet roundtrip CA to Netherlands: 150,000,000ns ≈ 150ms

Big Data Applications: Marketing

- **Personalized marketing**
  - Target each consumer individually
  - E.g., Amazon personalizes suggestions using:
    - Shopping history
    - Search, click, browse activity
    - Other consumers and trends
    - Reviews (NLP and sentiment analysis)
- **Brands want to understand customer-product relationships**
  - Use sentiment analysis from:
    - Social media, online reviews, blogs, surveys
  - Positive, negative, neutral sentiment
- E.g.,
  - In 2022, $600B spent on digital marketing

Big Data Applications: Advertisement

- **Mobile advertisement**
  - Mobile phones are ubiquitous
  - 80% of world population has one
  - 6.5 billion smartphones
- **Integrate online and offline databases**, e.g.,
  - GPS location
  - Search history
  - Credit card transactions

- E.g.,
  - You've bought a new house
  - You google questions about house renovations
  - You watch shows about renovations
  - Your phone tracks where you are
  - Google sends you coupons for the closest Home Depot
  - *"I feel like Google is following me"*

Big Data Applications: Medicine

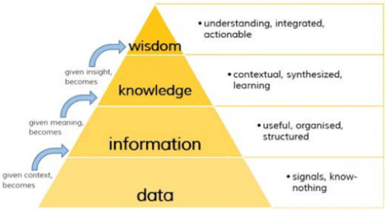
- **Personalized medicine**
  - Patients receive treatment tailored to them for efficacy
  - Genetics
  - Daily activities
  - Environment
  - Habits
- **Biomedical data**
- **Genome sequencing**
- **Health tech**
  - Personal health trackers (e.g., smart rings, phones)

Big Data Applications: Smart Cities

- **Smart cities**
    - Interconnected mesh of sensors
    - E.g., traffic sensors, camera networks, satellites
- **Goals**
    - Monitor air pollution
    - Minimize traffic congestion
    - Optimal urban services
    - Maximize energy savings

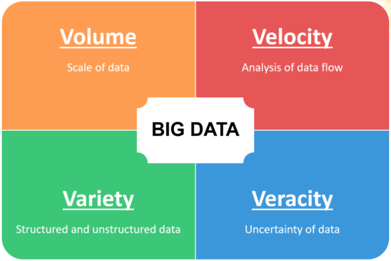
Goal of Data Science

- **Goal**: from data to wisdom
  - Data (raw bytes)
  - Information (organized, structured)
  - Knowledge (learning)
  - Wisdom (understanding)
- **Insights enable decisions and actions**
- Combine streams of big data to **generate new data**
  - New data can be “big data” itself



The Six V'S of Big Data

- What makes "Big Data" big?
- **Volume**
  - Vast amount of data is generated
- **Variety**
  - Different forms
- **Velocity**
  - Speed of data generation
- **Veracity**
  - Biases, noise, abnormality in data
  - Uncertainty, trustworthiness
- **Valence**
  - Connectedness of data in the form of graphs
- **Value**
  - Data must be valuable
  - Benefit an organization



The Six V's of Big Data

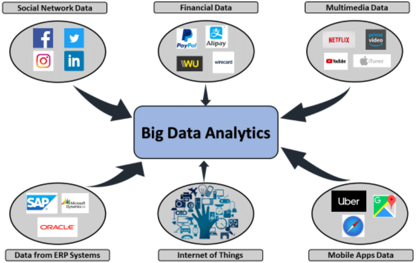
- **Volume**
  - Exponentially increasing data
  - 2.5 exabytes (1m TB) generated daily
    - 90% of data generated in last 2 years
    - Data doubles every 1.2 years
  - Twitter/X: 500M tweets/day (2022)
  - Google: 8.5B queries/day (2022)
  - Meta: 4PB data/day (2022)
  - Walmart: 2.5PB unstructured data/hour (2022)
- **Variety**
  - Different data forms
    - Structured (e.g., spreadsheets, relational data)
    - Semi-structured (e.g., text, sales receipts, class notes)
    - Unstructured (e.g., photos, videos)
  - Different formats (e.g., binary, CSV, XML, JSON)

The Six V's of Big Data

- **Velocity**
  - Speed of data generation
    - E.g., sensors generate data streams
  - Process data off-line or in real-time
  - Real-time analytics: consume data as fast as generated
- **Veracity**
  - Relates to data quality
  - How to remove noise and bad data?
  - How to fill in missing values?
  - What is an outlier?
  - How do you decide what data to trust?

Sources of Big Data

- Distinguish Big Data by source
  - Machines
  - People
  - Organizations



Sources of Big Data: Machines

- **Machines generate data**
    - Real-time sensors (e.g., sensors on Boeing 787)
    - Cars
    - Website tracking
    - Personal health trackers
    - Scientific experiments
- **Pros**
    - Highly structured
- **Cons**
    - Difficult to move, computed in-place or centralized
    - Streaming, not batch

Sources of Big Data: People

- People and their activities generate data
  - Social media (Instagram, Twitter, LinkedIn)
  - Video sharing (YouTube, TikTok)
  - Blogging, website comments
  - Internet searches
  - Text messages (SMS, Whatsapp, Signal, Telegram)
  - Personal documents (Google Docs, emails)
- Pros
  - Enable personalization
  - Valuable for business intelligence
- Cons
  - Semi-structured or unstructured data
    - Text, images, movies
  - Requires investment to extract value
    - Acquire → Store → Clean → Retrieve → Process → Insights
  - Surveillance capitalism

Sources of Big Data: Organizations

- Organizations generate data
    - Commercial transactions
    - Credit cards
    - E-commerce
    - Banking
    - Medical records
    - Website clicks
- Pros
    - Highly structured
- Cons
    - Store every event to predict future
        - Miss opportunities
    - Stored in "data silos" with different models
        - Each department has own system
        - Additional complexity
        - Data outdated/not visible
        - Cloud computing helps (e.g., data lakes, data warehouses)