



UMD DATA605: Big Data Systems

Lesson 1.2: Introduction to Big Data

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

- ***Big Data***

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

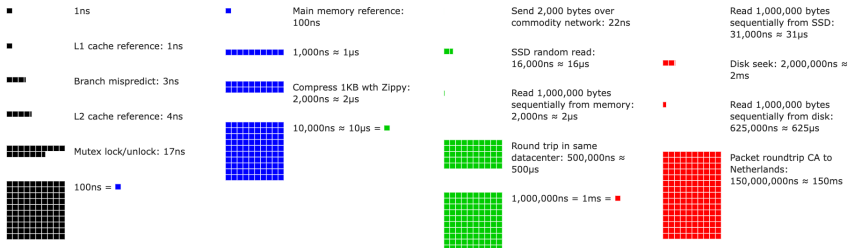
- *“Data science is the number one catalyst for economic growth”* (McKinsey, 2013)
- **Explosion of data in every domain**
 - Sensing devices/networks monitor processes 24/7
 - E.g., temperature of your room, your vital signs, pollution in the air
 - Sophisticated smart-phones
 - 80% of the world population 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** = 10^9 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
- **Zetabyte** = 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 μ s
- SSD random read: 16 μ s
- Send 1KB over network: 1 ms
- Disk seek: 2 ms
- Packet round-trip CA to Netherlands: 150 ms



Big Data Applications

- **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 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

- **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

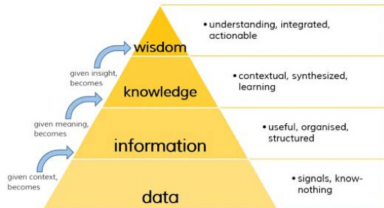
- **Biomedical data**
- Personalized medicine
 - Patients receive treatment tailored to them for efficacy
 - Genetics
 - Daily activities
 - Environment
 - Habits
- Genome sequencing
- Health tech
 - Personal health trackers (e.g., smart rings, phones)

Big Data Applications

- **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

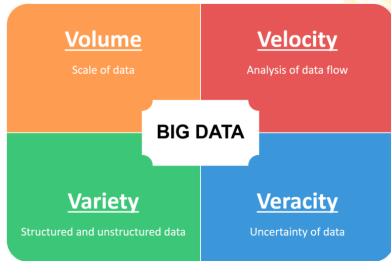
Goal of Data Science

- **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

- **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**

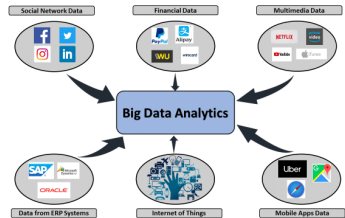
- Different data forms
 - Structured (e.g., spreadsheets, relational data)
 - Semi-structured (e.g., text, sales receipts, class notes)
 - Unstructured (e.g., photos, videos)
- Different formats (e.g., binary, CSV, XML, JSON)

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

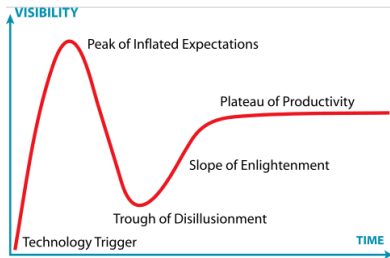


Sources of Big Data: Orgs

- **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)

Is Data Science Just Hype?

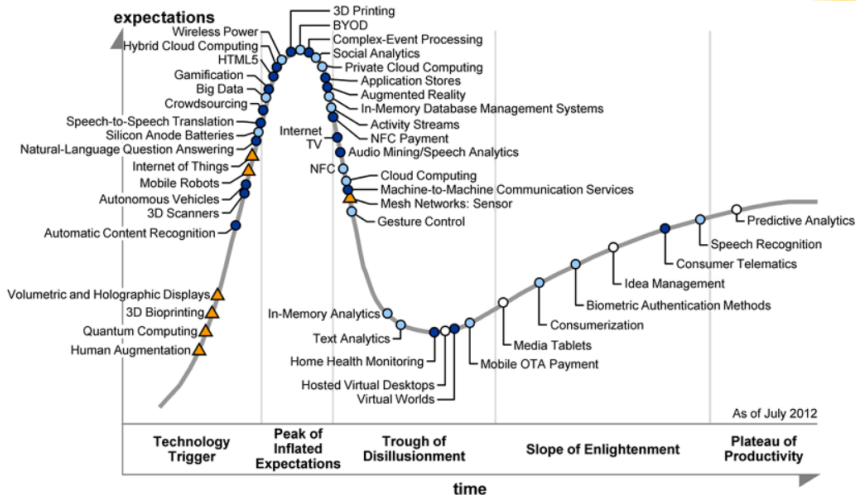
- Big data (or data science)
 - “Any process where interesting information is inferred from data”
- Data scientist called the “sexiest job” of the 21st century
 - The term has becoming very muddled at this point
- **Is it all hype?**



Is Data Science Just Hype?

- **No**
 - Extract insights and knowledge from data
 - Big data techniques revolutionize many domains
 - E.g., education, food supply, disease epidemics
- **But**
 - Similar to what statisticians have done for years
- **What is different?**
 - More data is digitally available
 - Easy-to-use programming frameworks (e.g., Hadoop) simplify analysis
 - Cloud computing (e.g., AWS) reduces costs
 - Large-scale data + simple algorithms often outperform small data + complex algorithms

What Was Cool in 2012?



Plateau will be reached in:

○ less than 2 years

● 2 to 5 years

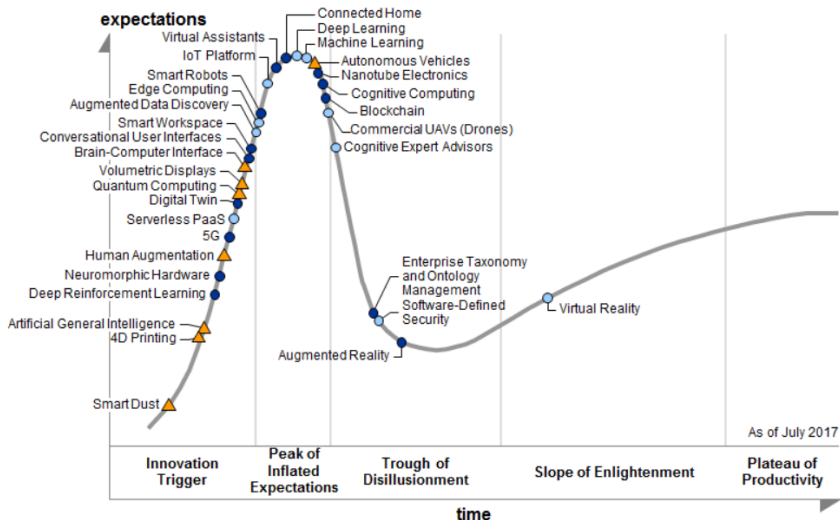
● 5 to 10 years

▲ more than 10 years

○ obsolete

⊗ before plateau

What Was Cool in 2017?



Years to mainstream adoption:

○ less than 2 years

● 2 to 5 years

● 5 to 10 years

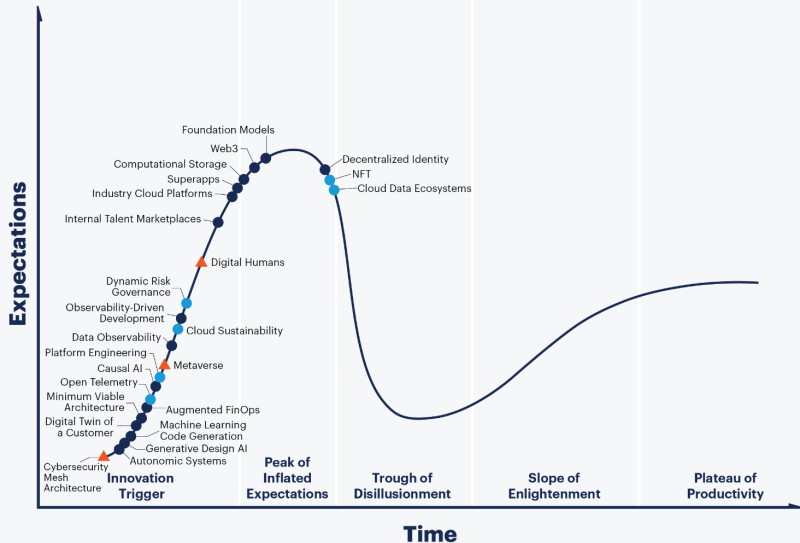
▲ more than 10 years

○ obsolete

⊗ before plateau

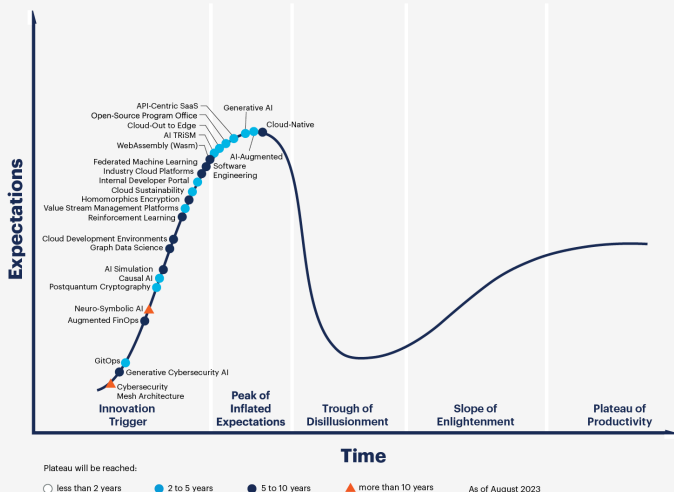
What Was Cool in 2022?

Hype Cycle for Emerging Tech, 2022



What Was Cool in 2023?

Hype Cycle for Emerging Technologies, 2023



Key Shifts Before/After Big-Data

- **Datasets: small, curated, clean → large, uncurated, messy**
 - Before:
 - Statistics based on small, carefully collected random samples
 - Costly and careful planning for experiments
 - Hard to do fine-grained analysis
 - Today:
 - Easily collect huge data volumes
 - Feed into algorithms
 - Strong signal overcomes noise
- **Causation → Correlation**
 - Goal: determine cause and effect
 - Causation hard to determine → focus on correlation
 - Correlation is often sufficient
 - E.g., diapers and beer bought together
- **"Data-fication"**
 - = converting abstract concepts into data
 - E.g., "sitting posture" data-fied by sensors in your seat
 - Preferences data-fied into likes
- From: Rise of Big Data, 2013

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

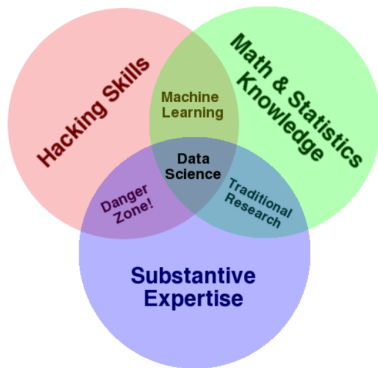


Examples: Google Flu Trends

- 5% to 20% of US population contracts flu yearly; 40k deaths
- Early warnings enable prevention and control
- Google Flu Trends
 - Early flu outbreak warnings via search query analysis
 - 45 search terms analyzed
 - IP used to determine location
 - Predict regional flu outbreaks 1-2 weeks before CDC
 - Active from 2008 to 2015
- Caveat: accuracy declined
 - Claimed 97% accuracy
 - Out of sample accuracy lower (overshot CDC data by 30%)
 - People search about flu without knowing diagnosis
 - E.g., searching for “fever” and “cough”
 - Google Flu Trends: The Limits of Big Data

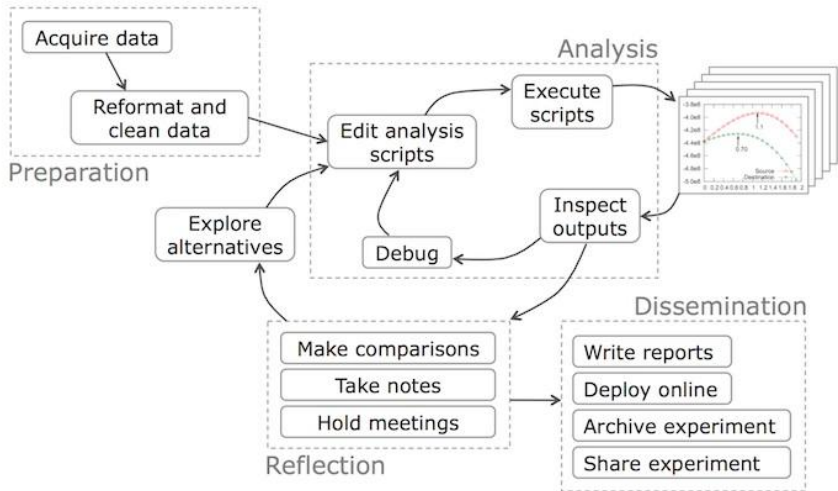
Data Scientist

- Ambiguous, ill-defined term
- From Drew Conway's Venn Diagram



Typical Data Scientist Workflow

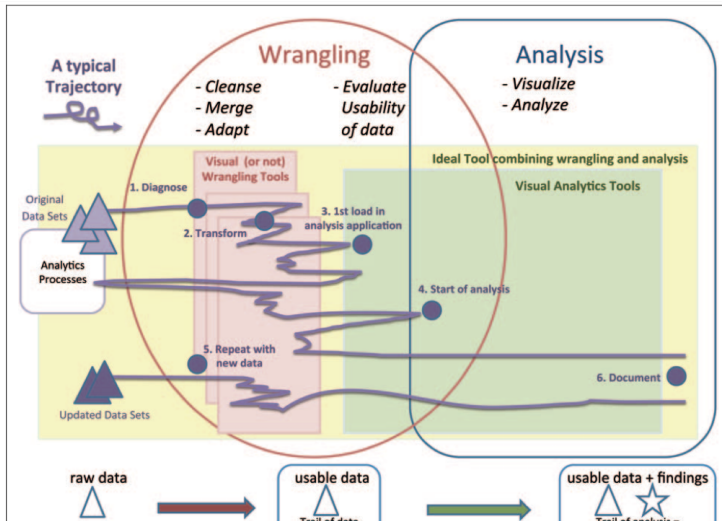
- From Data Science Workflow



[

Where Data Scientist Spends Most Time

- 80-90% of the work is data cleaning and wrangling
- “Janitor Work” in Data Science
- Research Directions in Data Wrangling



What a Data Scientist Should Know

- From: How to hire a data scientist
- **Data grappling skills**
 - Move and manipulate data with programming
 - Scripting languages (e.g., Python)
 - Data storage tools: relational databases, key-value stores
 - Programming frameworks: SQL, Hadoop, Spark
- **Data visualization experience**
 - Draw informative data visuals
 - Tools: D3.js, plotting libraries
 - Know what to draw
- **Knowledge of statistics**
 - Error-bars, confidence intervals
 - Python libraries, Matlab, R
- **Experience with forecasting and prediction**
 - Basic machine learning techniques
- **Communication skills**
 - Tell the story, communicate findings