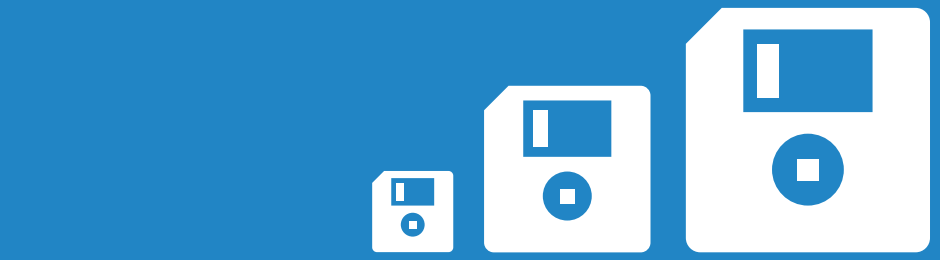
Indian Institute of Science
Bangalore, India
भारतीय विज्ञान संस्थान
बंगलौर, भारत

DS221

Data Structures, Algorithms & Data Science Platforms

Instructor: Chirag Jain
(slides from Prof. Simmhan)

CDS
The Department of Computational and Data Science



Big Data Concepts

What, Where, Why?

What is Big Data?

Image credits: http://www.seekbig.in/1128-tnpsc-economics-questions/

The term *is* fuzzy … *Handle with care!*

Wordle of "Thought Leaders'" definition of Big Data, © Jennifer Dutcher, 2014
https://datascience.berkeley.edu/what-is-big-data/

Data Generation View

"Big data refers to the approach to data of 'collect now, sort out later'…The low cost of storage and better methods of analysis mean that you generally don't need to have a specific purpose for the data in mind before you collect it."

Rohan Deuskar, CEO and Co-Founder, Stylitics

Wordle of "Thought Leaders'" definition of Big Data, © Jennifer Dutcher, 2014
https://datascience.berkeley.edu/what-is-big-data/

Image Credits: https://community.uservoice.com/wp-content/uploads/benefits-of-effective-questions-800x448-300x168.jpg

Data Systems View

"Big data is when your business wants to use data to solve a problem, answer a question, produce a product, etc., but the standard, simple methods break down on the size of the data set, causing time, effort, creativity, and money to be spent crafting a solution to the problem that leverages the data without simply sampling or tossing out records."

John Foreman, Chief Data Scientist, MailChimp

Wordle of "Thought Leaders'" definition of Big Data, © Jennifer Dutcher, 2014
https://datascience.berkeley.edu/what-is-big-data/

Image Credits: http://heae.tk/question/

Data Analysis View

"While the use of the term is quite nebulous … I've understood "big data" to be about analysis for data that's really messy or where you don't know the right questions or queries to make — analysis that can help you find patterns, anomalies, or new structures amidst otherwise chaotic or complex data points."

Philip Ashlock, Chief Architect, Data.gov

Wordle of "Thought Leaders'" definition of Big Data, © Jennifer Dutcher, 2014
https://datascience.berkeley.edu/what-is-big-data/

Image Credits:http://www.clipartpanda.com/categories/question-and-answer-clipart

So…What is Big Data?

Data whose characteristics exceeds the capabilities of conventional *algorithms, **systems** and techniques* to derive useful value.

https://www.oreilly.com/ideas/what-is-big-data

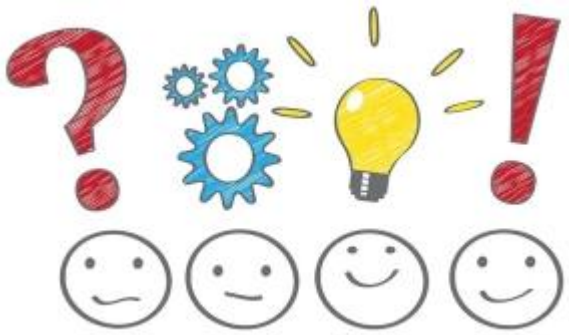
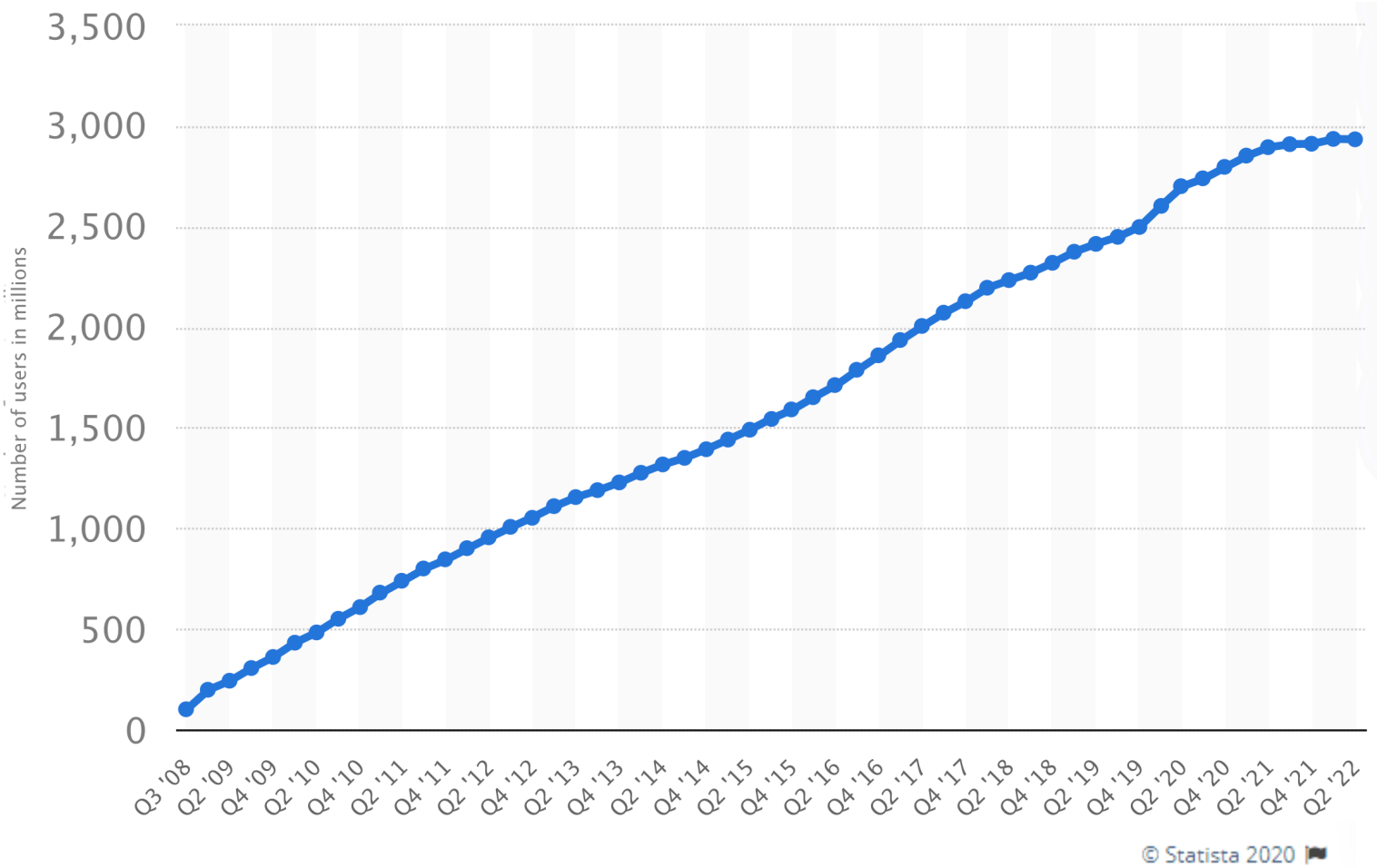


Image Credits: https://community.uservoice.com/wp-content/uploads/benefits-of-effective-questions-800x448-300x168.jpg

Facebook Active Users

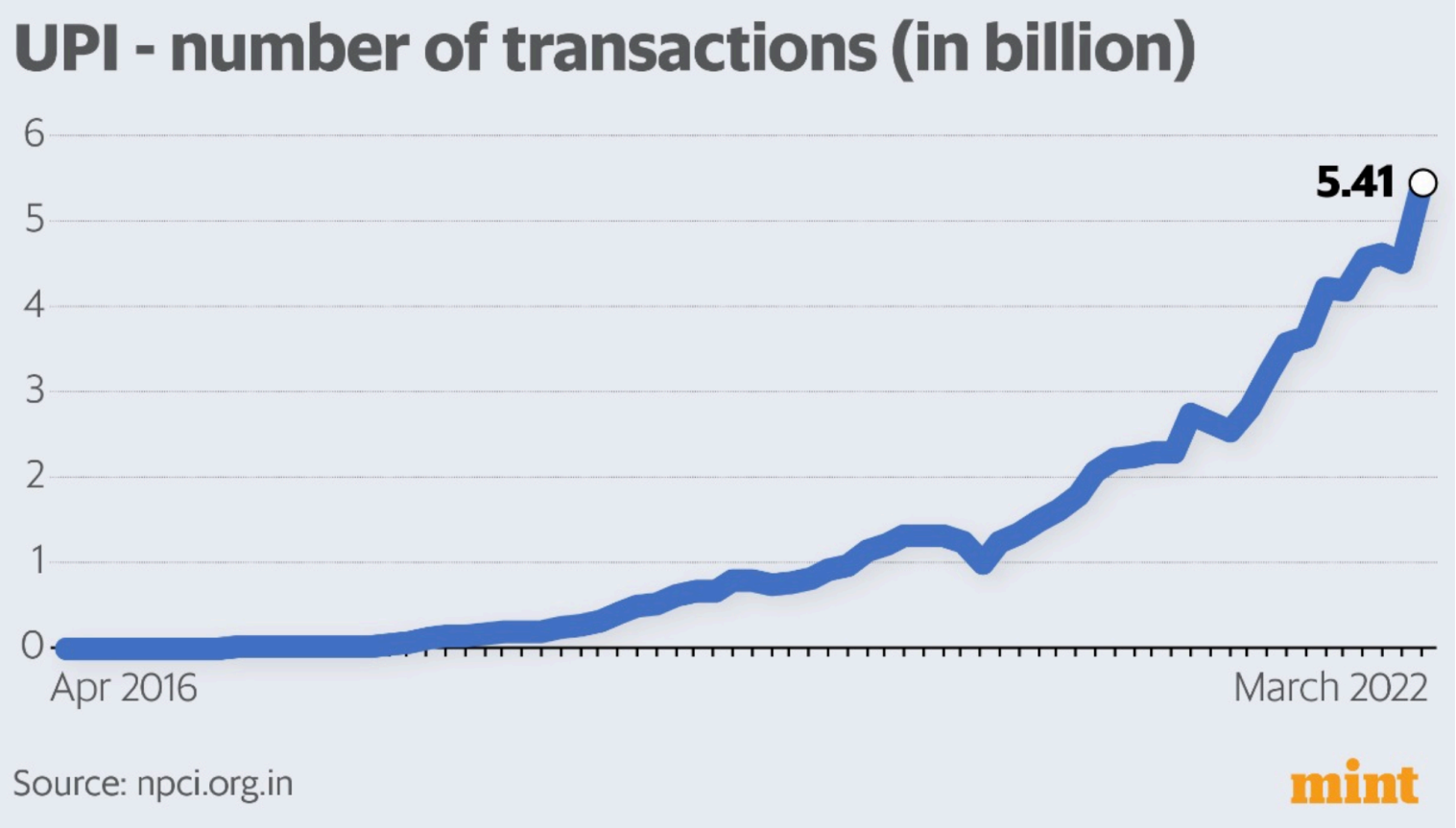


Over 2 billion users each month

https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

FinTech: UPI Transactions

Monthly UPI Transactions



https://www.npci.org.in/product-statistics/upi-product-statistics

Big Data and Science

AAAS Become a Member

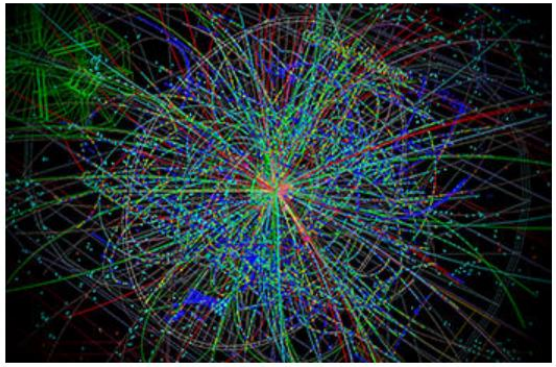
Science Contents News Careers Journals

AI's early proving ground: the hunt for new particles

Particle physicists began fiddling with artificial intelligence (AI) in the late 1980s, just as the term "neural network" captured the public's imagination. Their field lends itself to AI and machine-learning algorithms because nearly every experiment centers on finding subtle spatial patterns in the countless, similar readouts of complex particle detectors—just the sort of thing at which AI excels. "It took us several years to convince people that this is not just some magic, hocus-pocus, black box stuff," says Boaz Klima, of Fermi National Accelerator Laboratory (Fermilab) in Batavia, Illinois, one of the first physicists to embrace the techniques. Now, AI techniques number among physicists' standard tools.

Particle physicists strive to understand the inner workings of the universe by smashing subatomic particles together with enormous energies to blast out exotic new bits of matter. In 2012, for example, teams working with the world's largest proton collider, the Large Hadron Collider (LHC) in Switzerland, discovered the long-predicted Higgs boson, the fleeting particle that is the linchpin to physicists' explanation of how all other fundamental particles get their mass.

Such exotic particles don't come with labels,

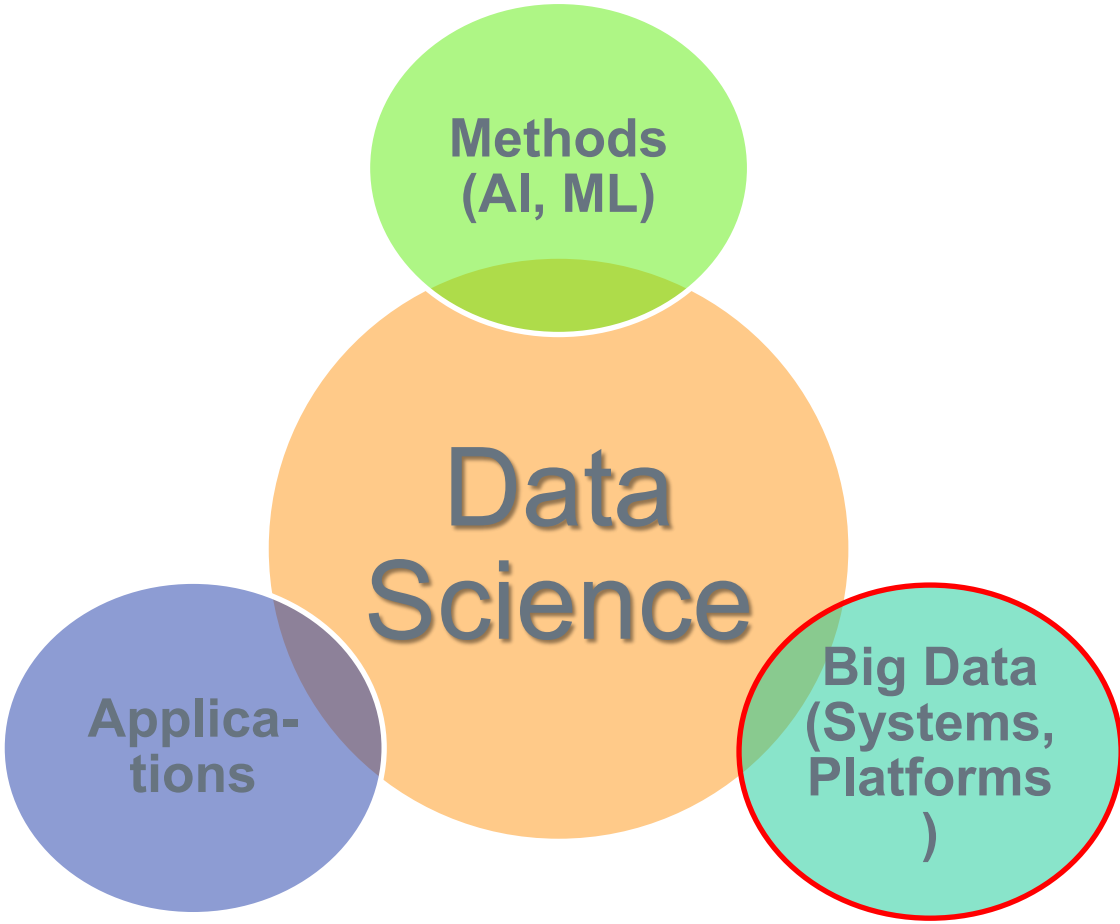


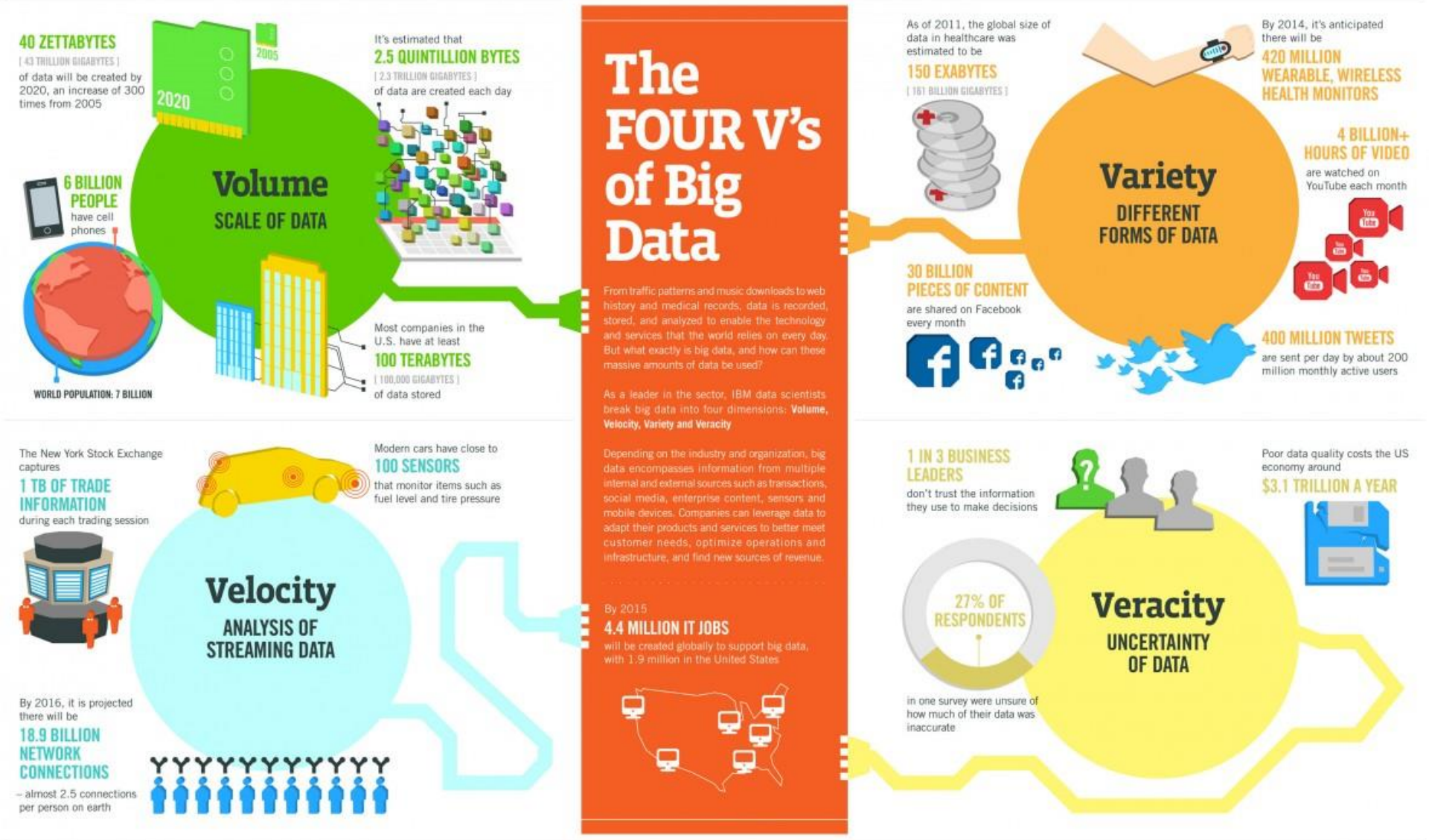
Neural networks search for fingerprints of new particles in the debris of collisions at the LHC. © 2012 CERN, FOR THE BENEFIT OF THE ALICE COLLABORATION

https://www.sciencemag.org/news/2017/07/ai-changing-how-we-do-science-get-glimpse

Data Science

Inter-disciplinary domain at the intersection of data analysis methods, Big Data Systems and data-driven applications.





Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Data Analysis Lifecycle

Acquire

- Acquire Data
  - Sensors, Web logs & crawls, Transactions

Goal

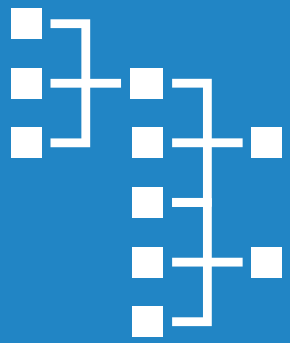
- Define Analysis & Analytics
  - Trends, Clusters, Outliers, Classification

Process

- Translate to Scalable Applications
  - Develop algorithms, Map to abstractions, Implement on Platforms

Additional Reading

- A Survey of Big Data Research, H Fang, et al., *IEEE Network,* September/October 2015, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4617656/
- Beyond the hype: Big data concepts, methods, and analytics, A. Gandomi and M. Haider, *International Journal of Information Management*, Volume 35, Issue 2, 2015, https://doi.org/10.1016/j.ijinfomgt.2014.10.007
- Uncertainty in big data analytics: survey, opportunities, and challenges. R.H. Hariri, et al. *J Big Data* **6,** 44 (2019). https://doi.org/10.1186/s40537-019-0206-3

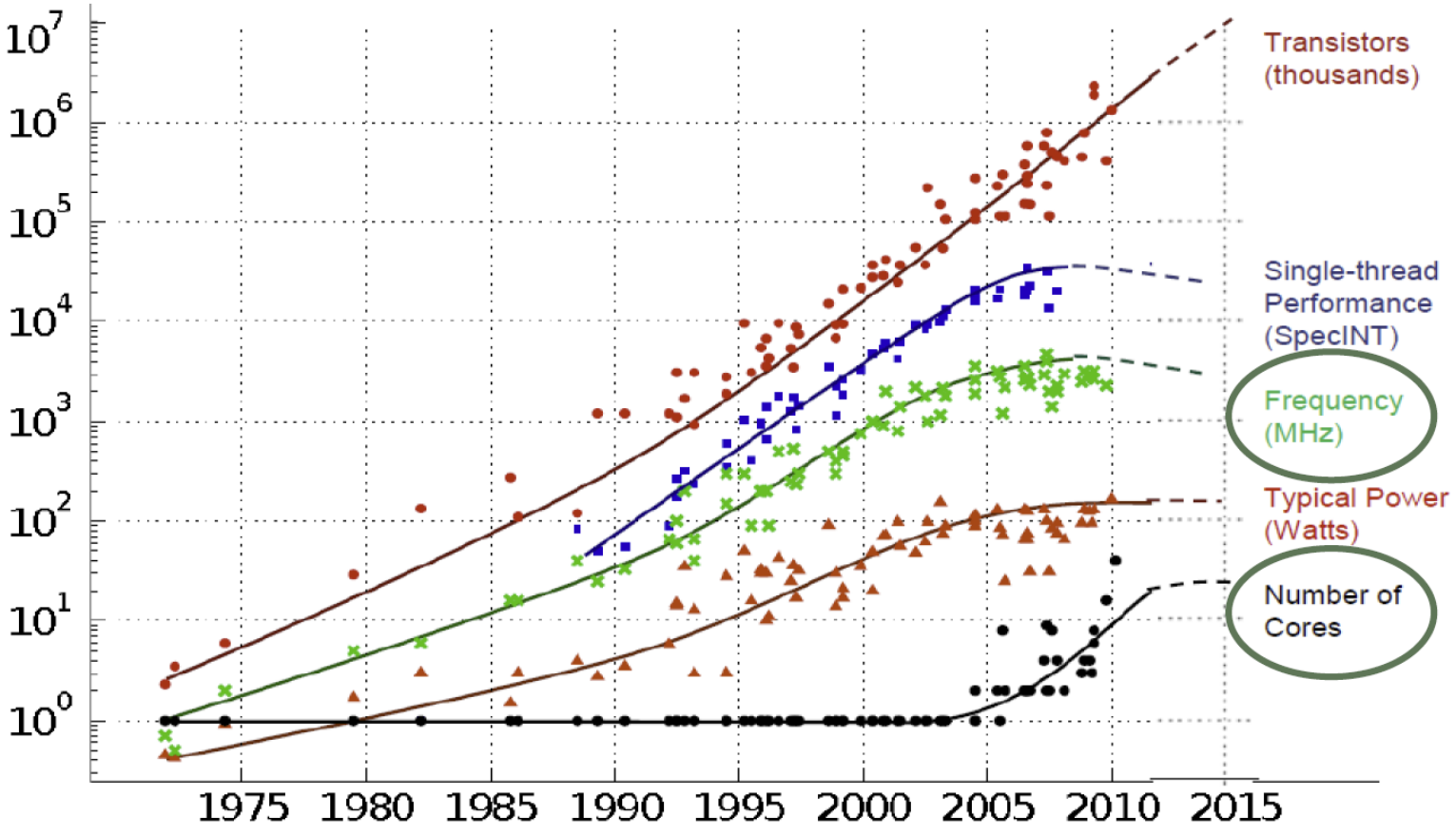


Distributed Systems

Helping Data Science Scale

Resources: Scale Up

Power and Heat Problems Led to Multiple Cores and Prevent Further Improvements in Speed



Jeffrey Funk, NUS

Original data collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond and C. Batten
Dotted line extrapolations by C. Moore

Source: Chuck Moore, Data Processing in Exascale-Class Systems, April 27, 2011. Salishan Conference on High Speed Computing

Resources: Scale Out

- More servers, working collaboratively
  - Connected over the network
  - High-speed LAN
  - Ethernet LAN
  - WAN
- Software is critical to enable this

Distributed Systems

- Distributed Computing
  - Clusters of machines
  - Connected over network
- Distributed Storage
  - Disks attached to clusters of machines
  - Network Attached Storage
- *How can we make effective use of multiple machines?*

- **Commodity** clusters vs. **HPC** clusters
  - Commodity: Available off the shelf at large volumes
  - Lower Cost of Acquisition
  - Cost vs. Performance
    - Low disk bandwidth, and high network latency
    - CPU typically comparable (Xeon vs. i3/5/7)
    - Virtualization overhead on Cloud
- *How can we use many machines of modest capability?*

Cloud Computing

- Large Data Centers
- 1000's of racks, 100k servers
- Virtualized infrastructure
  - On-demand: Rent VMs by the minute
  - 100 machines for 10mins costs same as 1 machine for 1000 minutes
- Makes use of economies of scale. Much cheaper than on-premises servers.
  - Reduce operations costs (energy, personnel)
  - Ensures capital costs (servers) are fully used
- *Have you used the Cloud?*

Scalability

- **System Size**: Higher performance when adding more machines
- **Software**: Can framework and middleware work with larger systems?
- **Application**: As problem size grows (compute, data), can the system keep up?
- **Vertical vs Horizontal:** ?

Scalability Metric

- If the *problem size* is x and the number of *processors available* is p, **Speedup** is:

$$Speedup(p, x) = \frac{time(1, x)}{time(p, x)}$$

- **Speedup Efficiency**: *How good is the speedup relative to perfect linear speedup?*

$$Speedup\ Efficiency = \frac{Speedup(p, x)}{p}$$

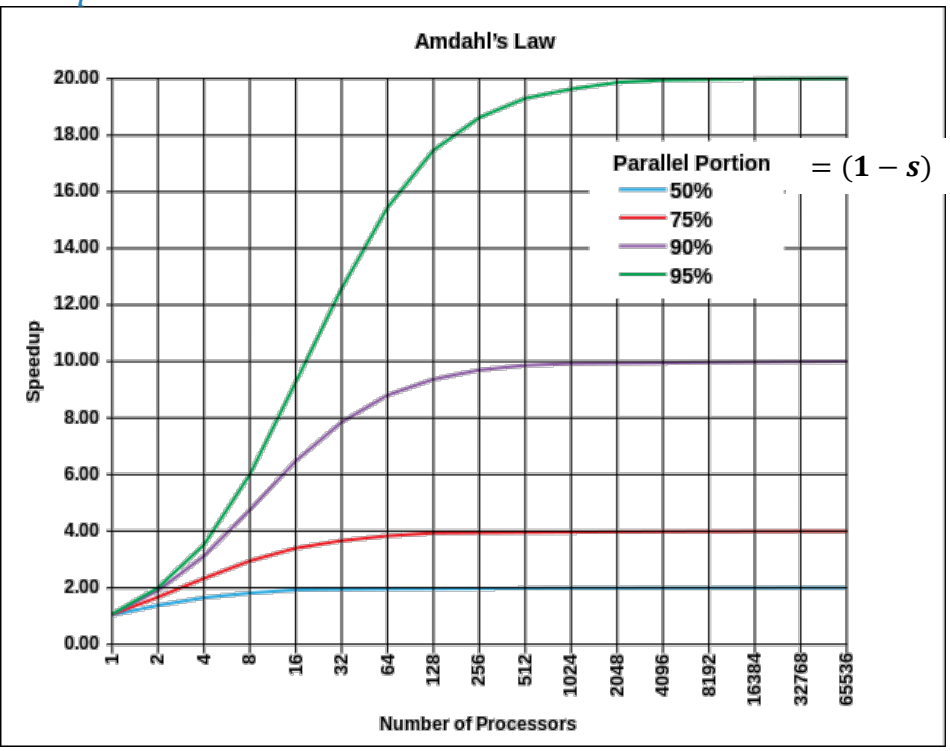
Amdahl's Law of Strong Scaling

- Amdahl's Law for Application Scalability
  - Total problem size is fixed
  - *Speedup limited by sequential bottleneck*
- s is *serial* fraction of application, cannot be parallelized
- $(1 - s)$ is fraction of application that can be *parallelized*
- p is number of processors
- t is time taken to process input of size x on 1 processor
- Then the speedup is limited by the serial fraction

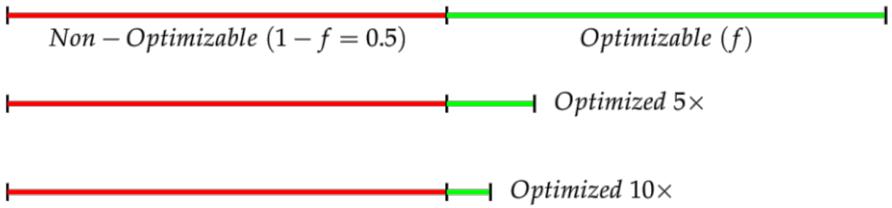
$$Speedup(p, x) = \frac{time(1,x)}{time(p,x)} = \frac{t}{s.t + \frac{(1-s)}{p}.t} = \frac{\mathbf{1}}{\boldsymbol{s + \frac{(1-s)}{p}}}$$

Amdahl's Law

$$\frac{1}{s + \frac{s}{p}}$$



$= (1 - s)$



Meaningful only for small values of ***p***, small fractions of ***s***

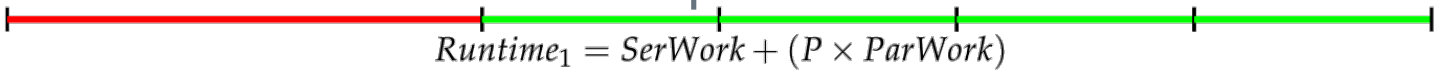
Gustafson's Law of weak scaling

- Gustafson's Law of Application Scalability
  - Problem size (***p.x***) increases with number of processors (***p***)
  - "Scaled speedup"
- $Speedup(p, x.p) = \frac{time(1,p.x)}{time(p,p.x)} = \frac{s.p.t+(1-s).p.t}{s.p.t+\frac{(1-s).p.t}{p}}$
- If $s.p.t \approx s.t$ …i.e., only parallelizable fraction of input work increases with number of processors
- $Speedup(p, x.p) = \frac{s.t+(1-s).p.t}{s.t+\frac{(1-s).p.t}{p}} = \frac{s+(1-s).p}{s+(1-s)}$

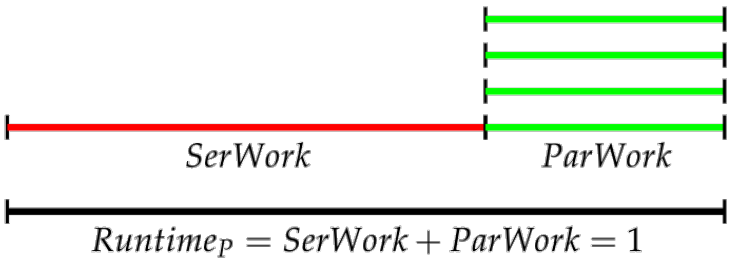
$$= \boldsymbol{s + (1 - s).p}$$

Gustafson's Law of weak scaling

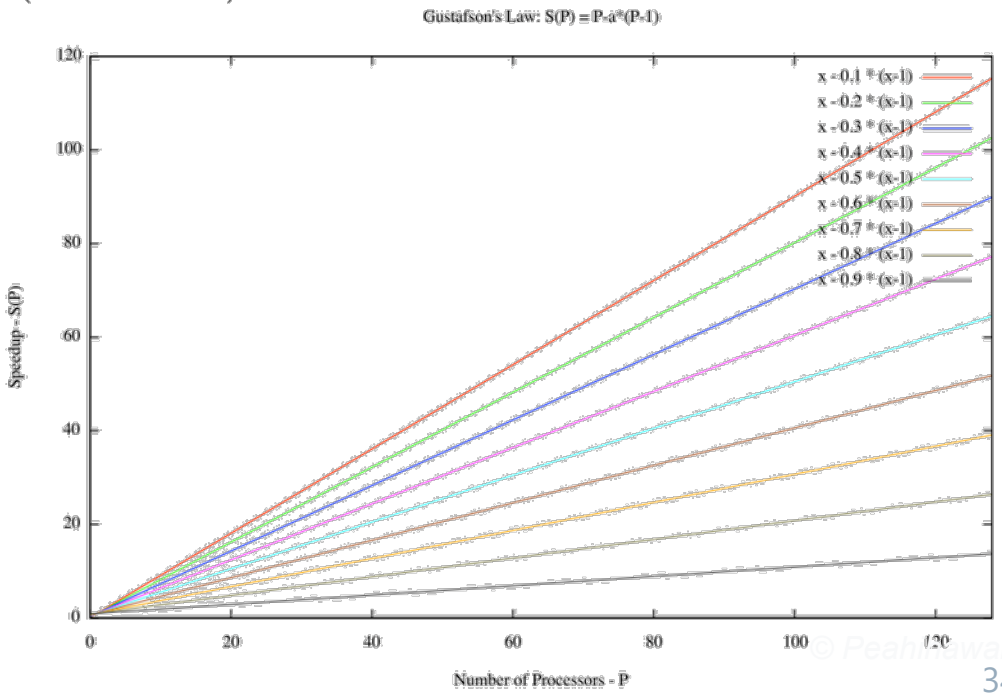
- As input work increases, all incremental work can be parallelized. Total time remains constant as we increase number of processors.



$$Runtime_1 = SerWork + (P \times ParWork)$$



$$Runtime_P = SerWork + ParWork = 1$$



Scalability

- Strong vs. Weak Scaling
- **Strong Scaling**: How the performance varies with the # of processors for a *fixed total problem size*
- **Weak Scaling**: How the performance varies with the # of processors for a *fixed problem size per processor*
  - Big Data platforms are intended for "Weak Scaling"

Additional Reading

- The evolution of distributed computing systems: from fundamental to new frontiers. Lindsay, D., Gill, S.S., Smirnova, D. et al. *Computing* (2021). https://doi.org/10.1007/s00607-020-00900-y
- Big Data computing and clouds: Trends and future directions, M. D. Assunção, et al., Journal of Parallel and Distributed Computing, Volumes 79–80, 2015, Pages 3-15, https://doi.org/10.1016/j.jpdc.2014.08.003

- TRY GOOGLE COLAB BEFORE NEXT CLASS
  - https://colab.research.google.com/
-
- Will be sharing Spark Notebook