CS4225/CS5425 Big Data Systems for Data Science

Logistics + Introduction

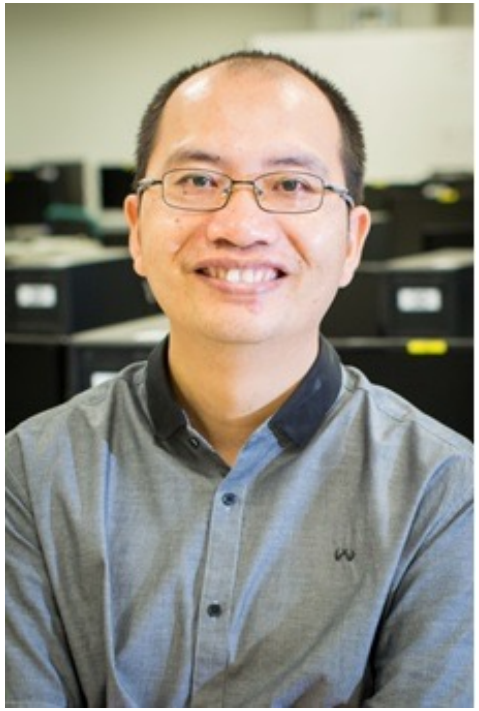
Bingsheng He
School of Computing
National University of Singapore
hebs@comp.nus.edu.sg

School of Computing

Learning Objectives

- Course logistics
- Introduction
  - What is (big) data science?
  - Why (big) data science?
  - *Infrastructure* for big data

About Bingsheng



- Bingsheng HE
- Office: COM3 #02-12
- Phone: 6516-7998
- Email: dcsheb@nus.edu.sg
- Consultation:
  - By appointment through email
- My research interests: big data systems, cloud computing, parallel and distributed computing
- Industrial experiences and consultation

About Ai Xin

- Ai XIN
- Office: COM3 #B1-24
- Phone: 660 16657
- Email: aixin@nus.edu.sg
- Consultation:
  - By appointment through email
- My research interests: Machine Learning and Deep Learning, Data Analytics and Big Data
- Industrial experiences: BHP Marketing Asia

Teaching Assistants

- Responsibility
  - Tutorials
  - Assist you in matters pertaining to the coding assignments
- We are fortunate to have many great TAs.
- Each assignment will indicate the TAs in charge of them, so you can contact the relevant TAs (or lecturers) for assistance

Schedule

Week	Date	Topics	Tutorial	Due Dates
1	15-Jan	Course Overview and Introduction		
2	22-Jan	Principles of Big Data Systems		
3	29-Jan	MapReduce/Hadoop- Intro		
4	5-Feb	MapReduce- Performance Analysis	Assignment 1 Briefing	Assignment 1 released
5	12-Feb	MapReduce- Database and Data Mining	Tutorial: MapReduce	
6	19-Feb	NoSQL Overview and Revision		
Recess				
7	5 Mar	Apache Spark 1		Assignment 1 due (8 Mar 11.59pm), Assignment 2 released
8	12 Mar	Apache Spark 2	Assignment 2 Briefing	Interview of Assignment 1
9	19 Mar	Stream Processing 1	Tutorial: Spark	
10	26 Mar	Stream Processing 2	Tutorial: Stream Processing	
11	2 Apr	NUS Well-Being Day		
12	9 Apr	Large Graph Processing 1		Assignment 2 due (12 Apr 11.59pm)
13	16 Apr	Large Graph Processing 2	Tutorial: Graph Processing	Interview of Assignment 2
	29 Apr	**Final Exam**		

Lecture

- In-person lecture
- Bingsheng for the first half, Ai Xin for the 2nd half
- Video recording will be given by CSIT
- Format
  - Session 1 about 45 minutes
  - 10 minutes for some short quiz (ungraded) and a break
  - Session 2 about 45 minutes

Lectures

- Reference textbooks
  - Jimmy Lin and Chris Dyer. 2010. Data-Intensive Text Processing with Mapreduce. Morgan and Claypool Publishers. https://lintool.github.io/MapReduceAlgorithms/MapReduce-book-final.pdf
  - Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. 2020. Mining of Massive Datasets (3rd ed.). Cambridge University Press. http://www.mmds.org/
- Study materials
  - Related chapters in the reference textbooks +
  - The related technical articles (for the state of the art)
- Only content discussed in class can appear on tests. Content marked “optional”, or only appearing in the powerpoint notes, will not appear on tests.

Tutorials

- Start from Week 4
- Not counted for final grade
- Tutorial questions will be available on Canvas before the tutorial: we recommend attempting questions before tutorial
- Some questions are samples for tests
- On tutorial weeks, tutorials will start about 10 minutes after lecture ends. That means they may start earlier than 8.30pm.

Assessment

- 2 assignments on Hadoop and Spark (25% each)
  - **Declare the usage of AI tools, and attach your prompts**
  - **Face-to-face interview**
- Final exam (50%) – held in-person.
- (Note: no weightage for in-lecture-quizzes)

Final Exam

- Date: 29 APR 2026 09:00-11:00
  - Held in person; open book + notes, but no electronics usage (other than calculators)
- Focus is on understanding and application, not facts / memorization
- Scope:
  - Material is out of scope for exams if it is not in the lecture slides, or is indicated as “optional”
- Examples of questions
  - Integrative: Combine knowledge from different chapters
  - Application: Apply your conceptual understanding to practical scenarios.
  - “Why not”: Example: Tommy proposed a solution A to solve problem B. Explain the problem with solution A.

Course Policies

- Zero-tolerance for plagiarism
- Plagiarism resources
  - http://www.cdtl.nus.edu.sg/ug/resources/plagiarism.htm
- Plagiarism prevention
  - http://cit.nus.edu.sg/plagiarism-prevention/
- Policy for Use of AI in Teaching and Learning.
  - https://ctlt.nus.edu.sg/wp-content/uploads/2024/08/Policy-for-Use-of-AI-in-Teaching-and-Learning.pdf

What will we learn?

- We will learn to process different types of data:
  - Large data volume
  - Graph data
  - Stream data (infinite/never-ending)
- We will learn to use different models of computation:
  - MapReduce/Spark
  - Large graph processing engines
  - Streams and online algorithms

What is Data Science?

- Standard definition:

"Data science is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms."

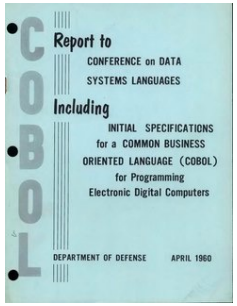
What is Data Science?

- Standard definition:

"Data science is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms."

- Historical view: in 1962, statistician John Tukey described a field called "data analysis", which emerged out of statistics, but with broader scope including the computing aspects of data analysis: including collecting, storing and analyzing large datasets, presenting data, etc.

Early transistor computers (TX-0, IBM 608, 7000 series)



Early DBMS, statistical software

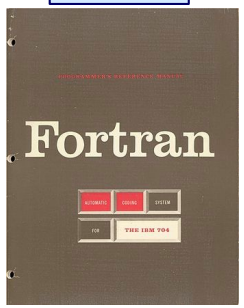
COBOL

1950 UNIVAC I

1957

1960

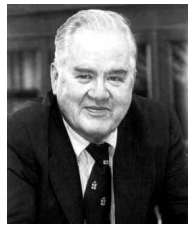
1962



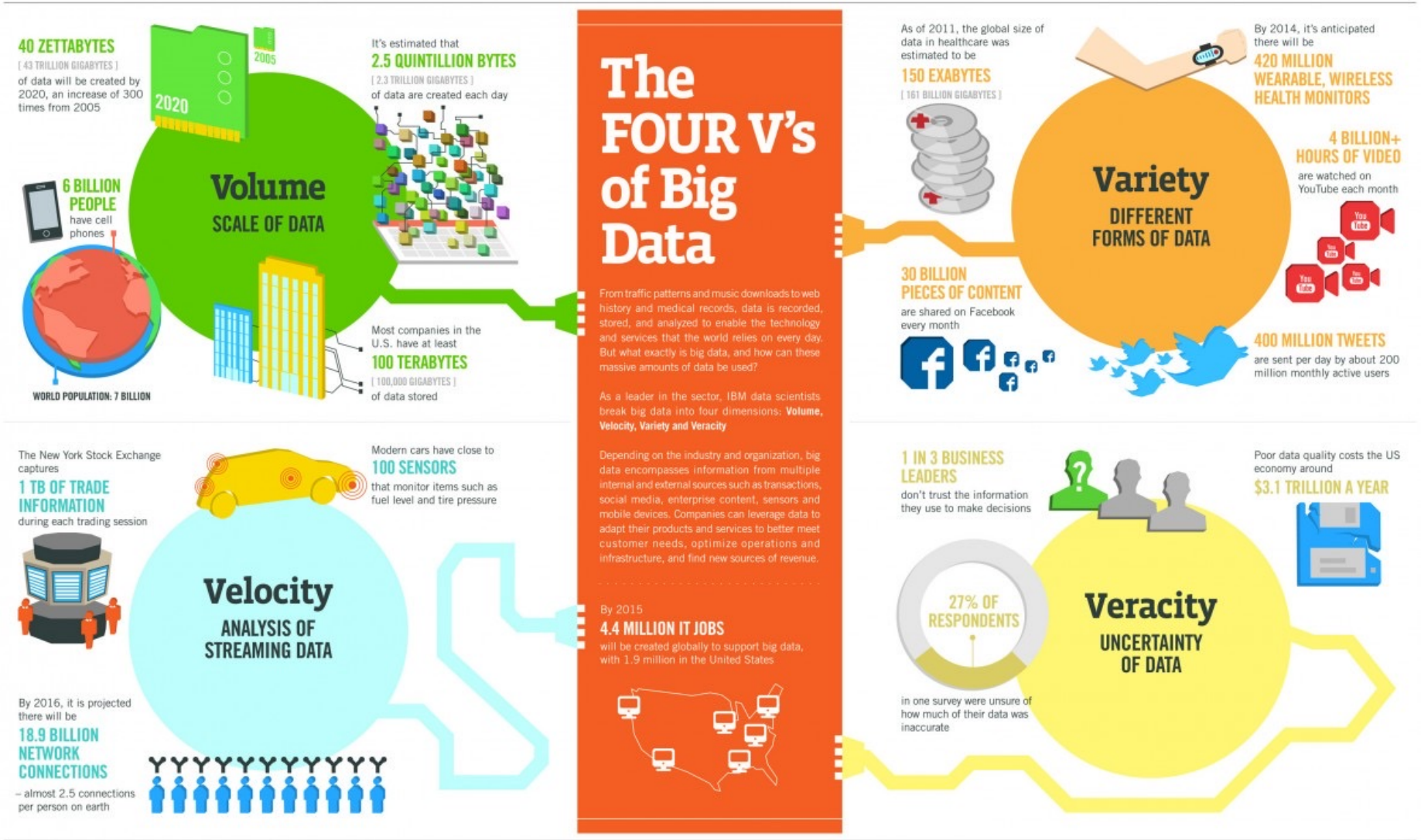
THE FUTURE OF DATA ANALYSIS[1]

By John W. Tukey

Princeton University and Bell Telephone Laboratories

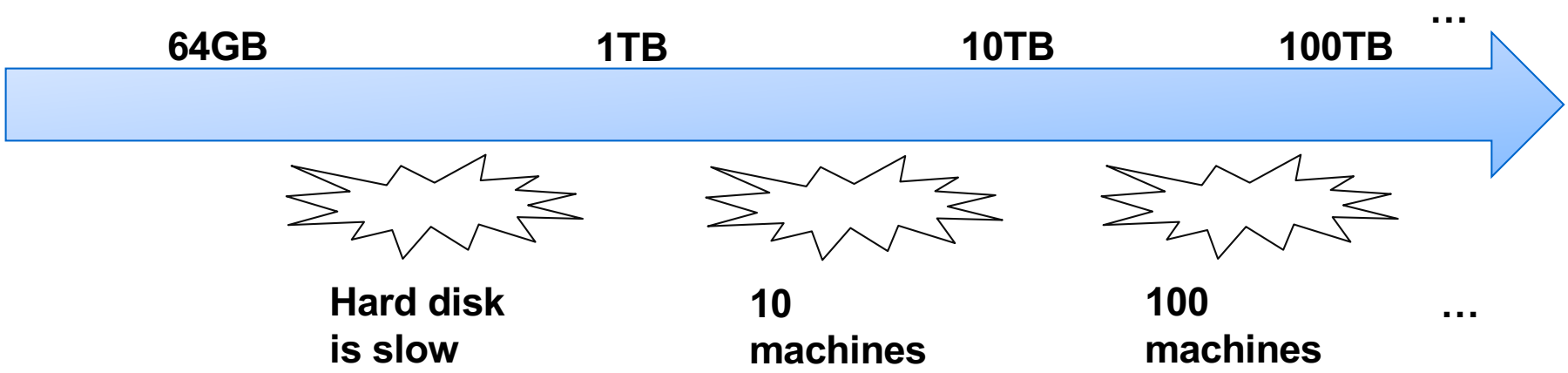


Challenges of Big Data: the 4 ‘V’s



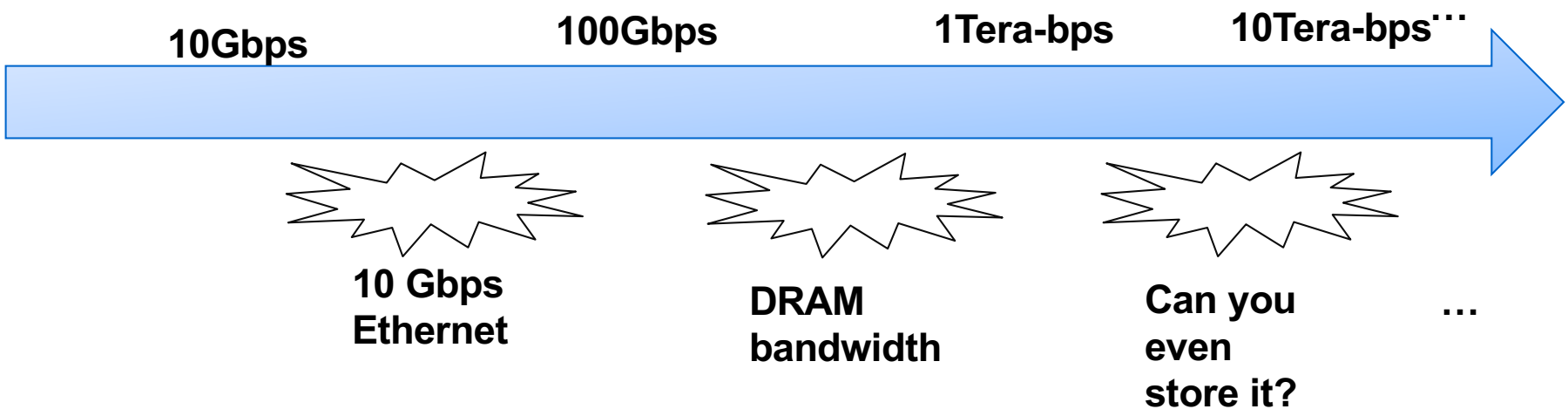
Volume

- Volume: The scale of data
- The challenges of large volume
  - Performance
  - Cost
  - Reliability
  - Algorithm design complexity



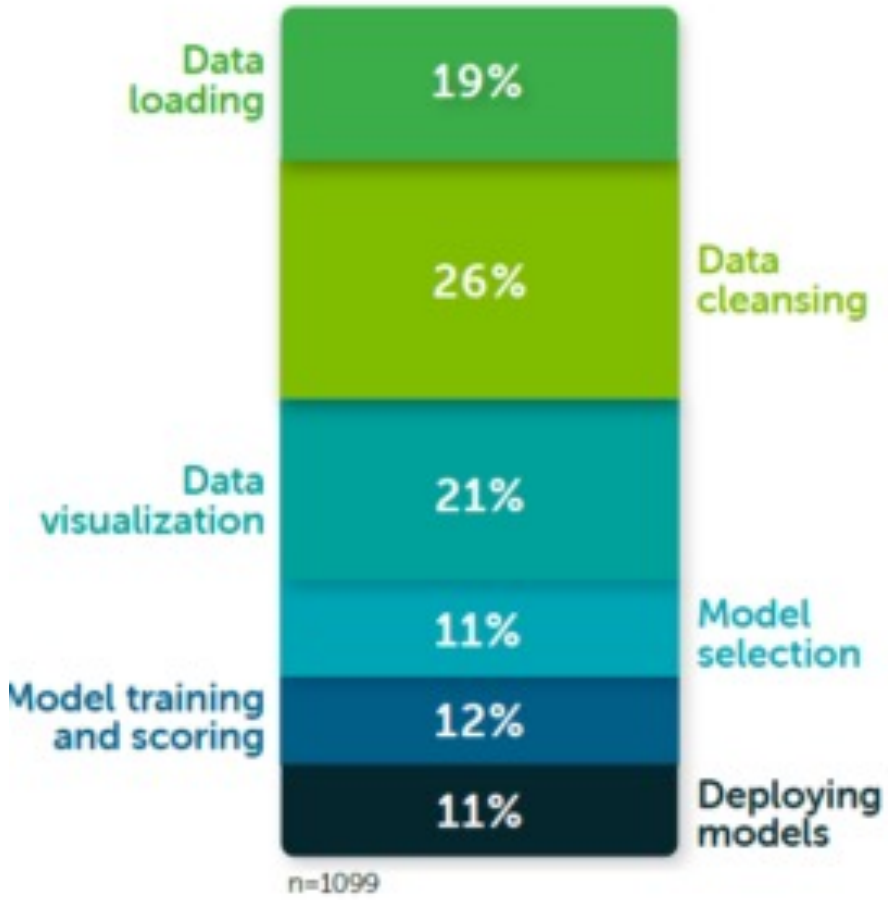
Velocity

- Velocity: the speed of new data (streaming data)
- The challenges of high velocity
  - Performance
  - Cost
  - Reliability
  - Algorithm design complexity



Variety and Veracity

- Why Variety matters?
  - “One size does not fit all”
  - Data integration
  - Multi-modal learning
- Why Veracity matters?
  - Dirty and noisy data
  - Data provenance
  - Data uncertainty



https://www.datanami.com/2020/07/06/data-prep-still-dominates-data-scientists-time-survey-finds/

Veracity: ~3.3% of data in some of the most popular datasets are mislabelled

Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks

Curtis G. Northcutt* ChipBrain, MIT
Anish Athalye MIT
Jonas Mueller Amazon


given: cat
corrected: frog


given: lobster
corrected: crab


given: ewer
corrected: teapot


given: white stork
corrected: black stork

[1] Northcutt, Curtis G., Anish Athalye, and Jonas Mueller. "Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks." NeurIPS 2021

Veracity: Importance of Data Quality

Forbes

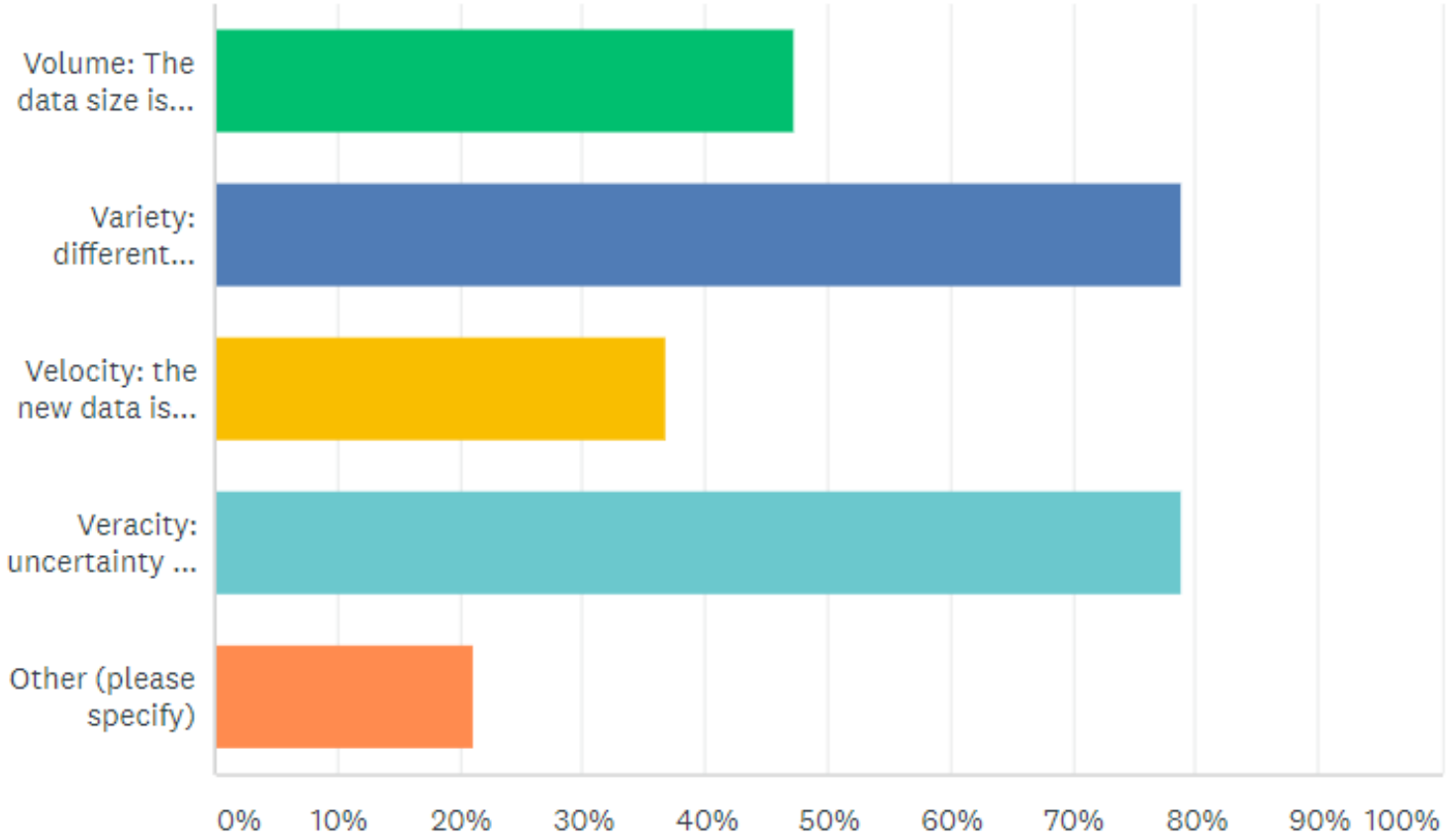
ENTERPRISE & CLOUD

Andrew Ng Launches A Campaign For Data-Centric AI

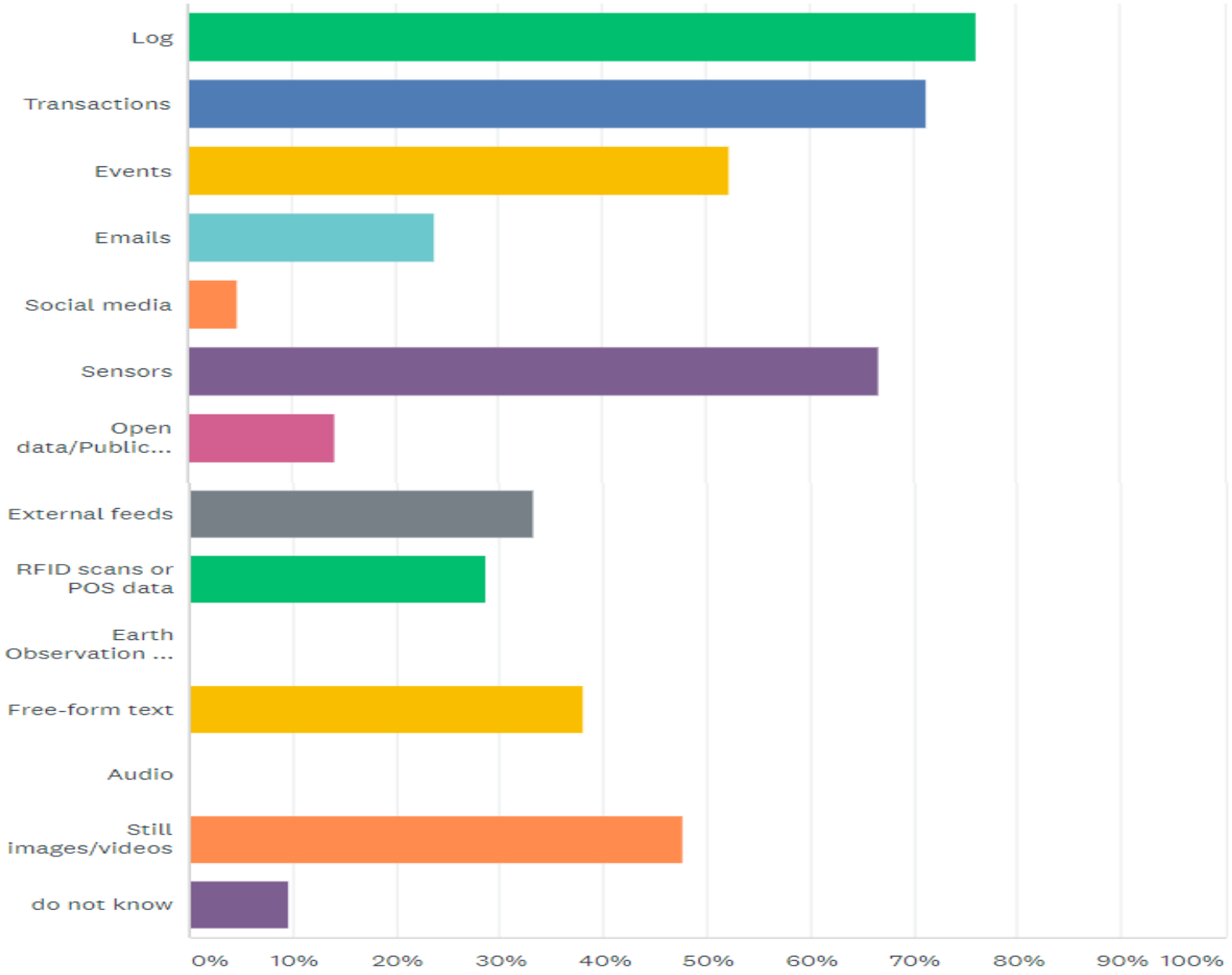
Data is eating the world so Andrew Ng wants to make sure we radically improve its quality. "Data is food for AI," says Ng, and he is launching a campaign to shift the focus of AI practitioners from model/algorithm development to the quality of the data they use to train the models.

Our Survey

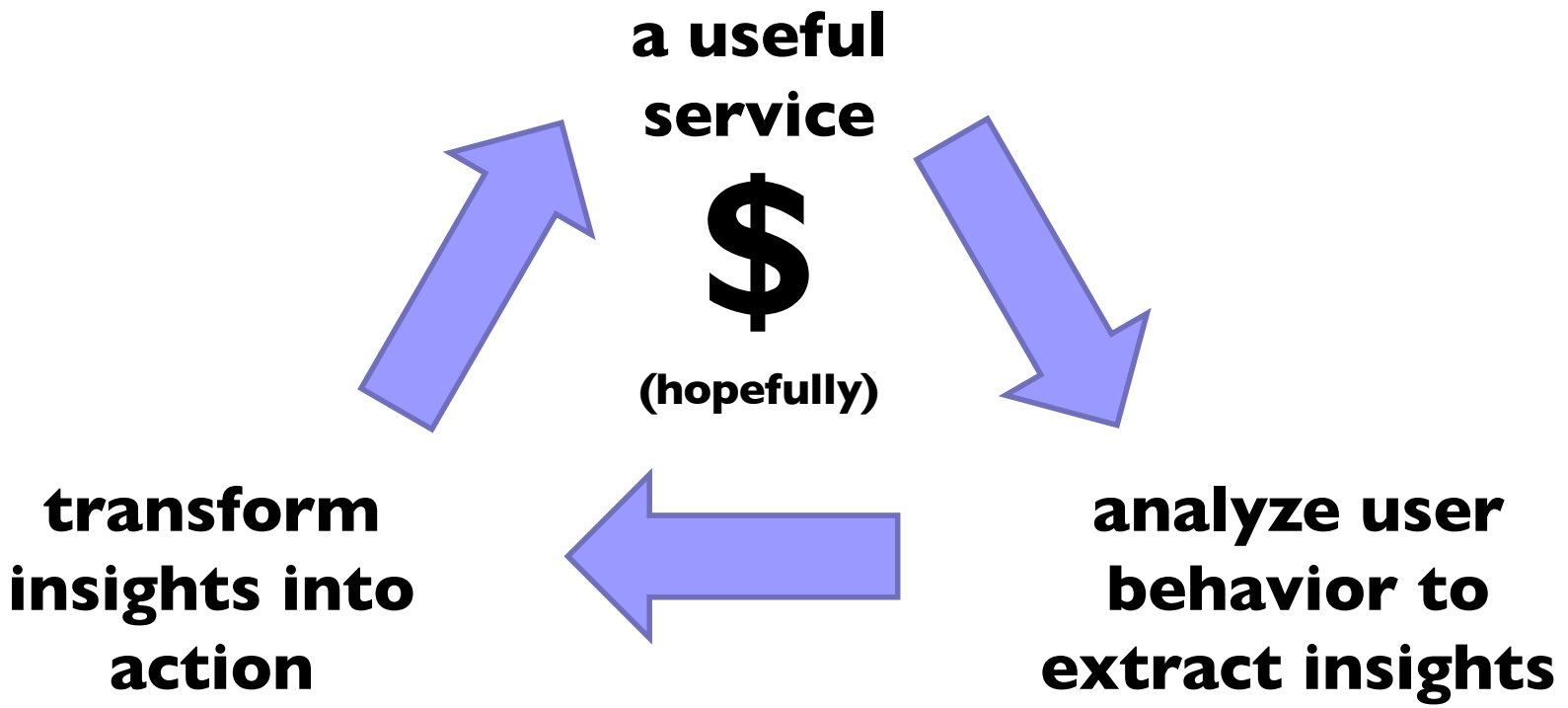
What are the big data challenges in your company?



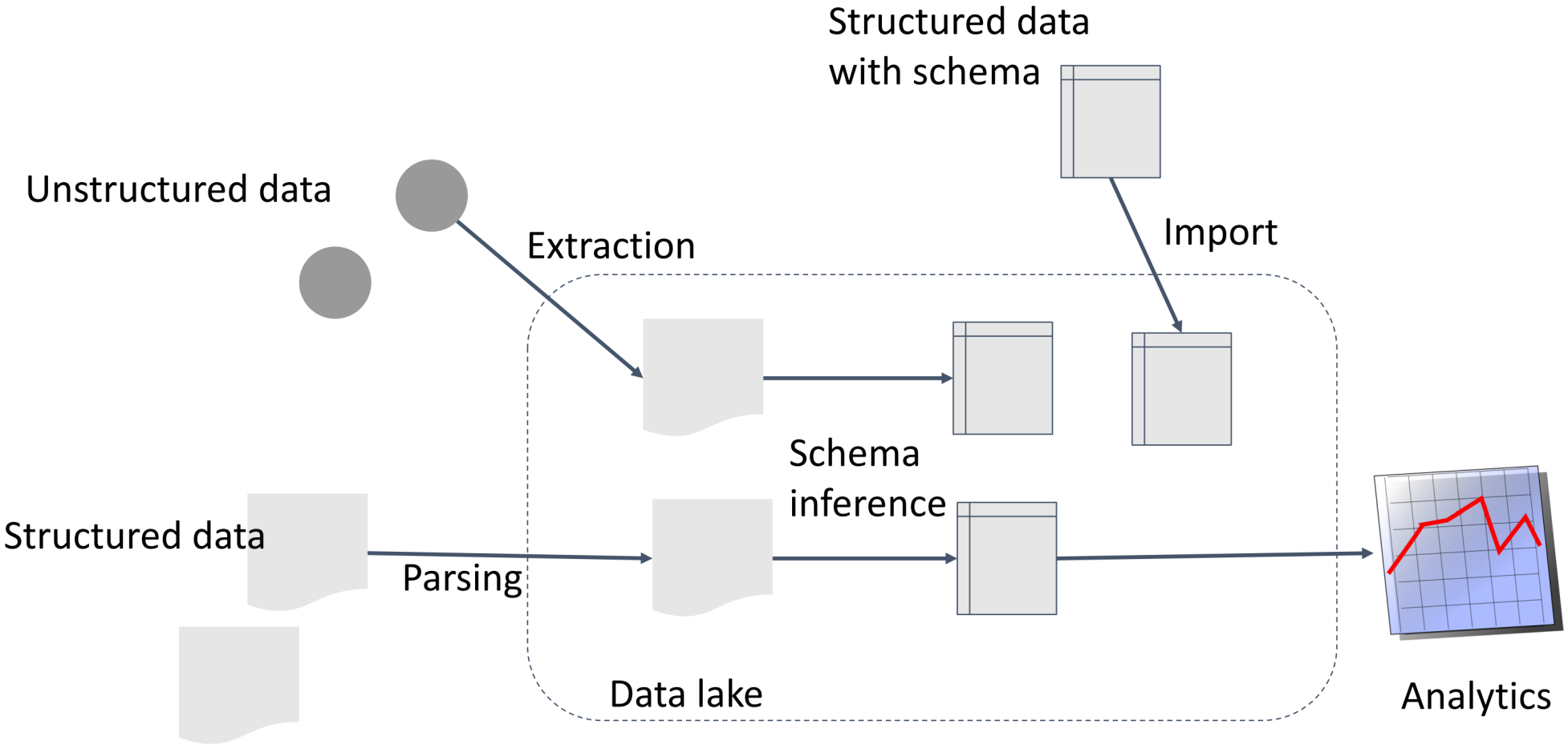
From what sources does your organisation collect, or expects to collect, data?



Virtuous Product Cycle



Data Lake: The Next Gen of Big Data?



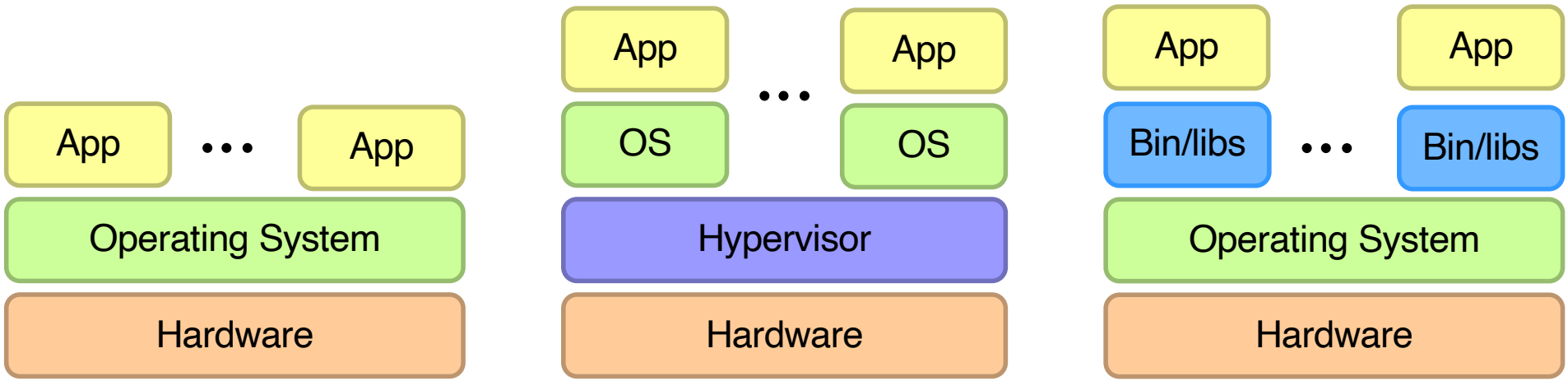

Infrastructure: Cloud Computing
Source: Wikipedia (Clouds)

Utility Computing

- What?
  - Computing resources as a metered service (“pay as you go”)
  - Ability to dynamically provision virtual machines
- Why?
  - Scalability: “infinite” capacity
  - Elasticity: scale up or down on demand

Thomas J. Watson (attributed?)

Enabling Technology: Virtualization and Containers



- **Virtual Machines**: enable sharing of hardware resources by running each application in an isolated virtual machine.
  - High overhead as each VM has its own OS.
- **Containers**: enable lightweight sharing of resources, as applications run in an isolated way, but still share the same OS.
  - A container is a lightweight software package that encapsulates an application and its environment.

Everything as a Service

- **Infrastructure as a Service (IaaS)**: Utility Computing
  - User rents a virtual machine and makes all the decisions on what to run on it
  - Examples: Amazon's EC2, Rackspace, Google Compute Engine
- **Platform as a Service (PaaS)**
  - Provides hosting for web applications and takes care of the hardware maintenance, upgrades, …
  - Example: Google App Engine. User provides their web application (e.g. in Python / Java) and the system takes care of all the details for hosting it.
- **Software as a Service (SaaS)**
  - User typically doesn't write code, and is just using an existing app
  - Example: Gmail, Dropbox, Zoom

Cloud Computing & Big Data Systems in AI?

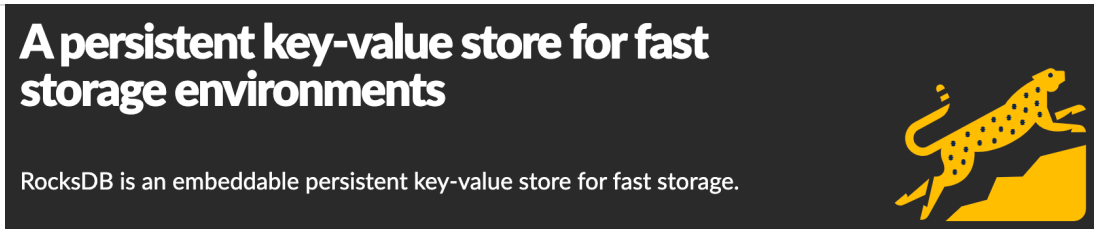
Reuters World Business Markets Sustainability Legal Breakingviews Technology Investiga

Artificial Intelligence

OpenAI buys database analytics firm Rockset in nine-figure stock deal, sources say

By **Krystal Hu** and **Akash Sriram**

June 22, 2024 7:33 AM GMT+8 · Updated 2 months ago



Rockset, founded by former engineers at Meta, builds real-time search and analytics databases…

…the technology will power the retrieval infrastructure of the ChatGPT maker's enterprise products.

Rockset's expertise in real-time data processing and vector search will enhance OpenAI's ability to quickly access and analyze vast amounts of information, likely leading to faster and more accurate responses from AI models…

https://www.reuters.com/markets/deals/openai-acquires-database-analytics-firm-rockset-2024-06-21/

MCQ Quiz

Sensor readings have precision issues or errors. Such scenario shows _____ of big data.

- (A) Volume
- (B) Velocity
- (C) Veracity
- (D) Variety
- (E) None of the above answers

Answer: C, Veracity means the accuracy of the data.

Take-away

- Data contains value and knowledge.
- Data science is a cross-disciplinary and emerging research area with interesting applications in science, engineering and commerce etc.
- Clouds are natural infrastructures for data science.
- Further readings:
  - Chapter 1. Jimmy Lin and Chris Dyer. 2020. Data-Intensive Text Processing with Mapreduce. Morgan and Claypool Publishers. https://lintool.github.io/MapReduceAlgorithms/MapReduce-book-final.pdf
  - Vasant Dhar. 2013. Data science and prediction. Commun. ACM 56, 12 (December 2013), 64–73. https://dsmilab.github.io/assets/file/reading_list/data_science_and_prediction.pdf
  - Introduction to Data Lakes. https://www.databricks.com/discover/data-lakes

Take-away in the AI Era

- (Why learning system principles is still essential, even with AI coding tools)
- 🤖 What AI coding tools are very good at
  - Writing correct-looking code
  - Using existing APIs and frameworks
  - Following known design patterns
  - Optimizing **within** a given design
- 🧠 What AI cannot do for you (yet)
  - Decide **what the right design is**
  - Decide **where data should live**
  - Decide **what trade-offs are acceptable**

AI writes code.
Humans decide systems.

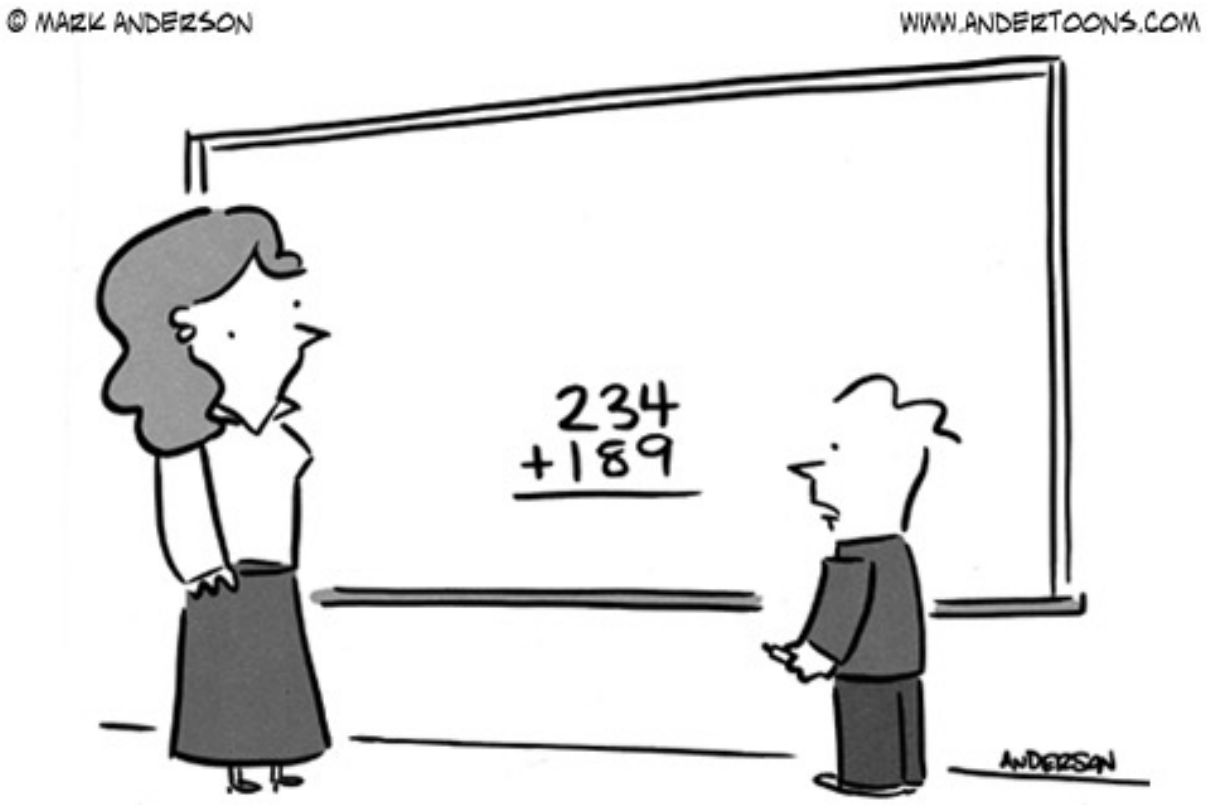
Acknowledgement

- Slides adopted/revised from
  - Jimmy Lin, http://lintool.github.io/UMD-courses/bigdata-2015-Spring/
  - Bryan Hooi
- Some slides are also adopted/revised from
  - Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. 2014. Mining of Massive Datasets (2nd ed.). Cambridge University Press. http://www.mmds.org/

Video Lectures

- Video recordings for lecture and tutorials
  - Canvas: Go to our course, and then look for "Videos/Panopto" > "Web Lectures".
- FYI. Crash course videos.
  - Memory & Storage: Crash Course Computer Science #19 https://www.youtube.com/watch?v=TQCr9RV7twk
  - Operating Systems: Crash Course Computer Science #18 https://www.youtube.com/watch?v=26QPDBe-NB8

Questions?



"Does this count as big data?"