COMP9313: Big Data Management

Lecturer: Xin Cao

Course web site: http://www.cse.unsw.edu.au/~cs9313/

Chapter 1: Course Information and Introduction to Big Data Management

Part 1: Course Information

Course Info

- Lectures: 14:00 – 16:00 (Tuesday) and 14:00 – 16:00 (Wednesday)
  - Hybrid (online access through Moodle using Collaborate)
- Labs: Weeks 1-10 (except the recess week)
- Consultation (Weeks 1-10): Questions *regarding lectures, course materials, assignements, exam, etc.*
  - Time: 13:00 – 14:00 (Thursday)
  - Place: K17-201D and online
- Course Admin:
  - Siqing Li, *siqing.li@unsw.edu.au*
- Discussion and QA:
  - WebCMS3
  - Ed: https://edstem.org/au/join/uB64ta

Lecturer in Charge

- Lecturer: Xin Cao
  - Office: 201D K17 (outside the lift turn left)
  - Email: *xin.cao@unsw.edu.au*

- Research interests
  - Database
  - Data Mining
  - Big Data Technologies
  - Machine Learning Applications
  - My homepage: https://xincao-unsw.github.io/
  - My publications list at google scholar: https://scholar.google.com.au/citations?user=kJIkUagAAAAJ&hl=en

Course Aims

- This course aims to introduce you to the concepts behind Big Data, the core technologies used in managing large-scale data sets, and a range of technologies for developing solutions to large-scale data analytics problems.

- This course is intended for students who want to understand modern large-scale data analytics systems. It covers a wide range of topics and technologies and will prepare students to be able to build such systems as well as use them efficiently and effectively to address challenges in big data management.

- *Not possible to cover every aspect of big data management.*

Lectures

- Lectures focusing on the frontier technologies on big data management and the typical applications
- Try to run in more interactive mode and provide more examples
- A few lectures may run in more practical manner (e.g., like a lab/demo) to cover the applied aspects
- Lecture length varies slightly depending on the progress (of that lecture)
- Note: attendance to every lecture is assumed

Resources

- Textbooks
  - Hadoop: The Definitive Guide. Tom White. 4th Edition - O'Reilly Media
  - Mining of Massive Datasets. Jure Leskovec, Anand Rajaraman, Jeff Ullman. 2nd edition - Cambridge University Press
  - Data-Intensive Text Processing with MapReduce. Jimmy Lin and Chris Dyer. University of Maryland, College Park.
  - Learning Spark . Matei Zaharia, Holden Karau, Andy Konwinski, Patrick Wendell. O'Reilly Media
- Reference Books and other readings
  - Apache MapReduce Tutorial
  - Apache Spark Quick Start
  - Many other online tutorials … …

- Big Data is a relatively new topic (so no fixed syllabus)

Prerequisite

- Official prerequisite of this course is COMP9024/2521 (Data Structures and Algorithms) and COMP9311/3311 (Database Systems).
- Before commencing this course, you should:
  - have experiences and good knowledge of algorithm design (equivalent to COMP9024/2521)
  - have a solid background in database systems (equivalent to COMP9311/3311)
  - **have solid programming skills in Java/Python**
  - **be familiar with working on a Unix-style operating systems**
  - have basic knowledge of linear algebra (e.g., vector spaces, matrix multiplication), probability theory and statistics , and graph theory
- No previous experience necessary in
  - MapReduce/Spark
  - Parallel and distributed programming

Please do not enrol if you

- Don’t have COMP9024/9311 knowledge
- Cannot produce correct Python program on your own
- Never worked on Unix-style operating systems
- Have poor time management
- Are too busy to attend lectures/labs

- *Otherwise, you are likely to perform badly in this subject*

Learning outcomes

- After completing this course, you are expected to:
  - describe the important characteristics of Big Data
  - develop an appropriate storage structure for a Big Data repository
  - utilise the map/reduce paradigm and the Spark platform to manipulate Big Data
  - use a high-level query language to manipulate Big Data
  - develop efficient solutions for analytical problems involving Big Data

Assessment

Number	Name	Full Mark
1**	Coding Project 1	12
2**	Coding Project 2	16
3**	Coding Project 3	22
4	Final Exam	50

Later Submission Penalties:
** : 5% reduction of your marks for up to 5 days

The final mark is calculated by:
Final Mark= proj1 + proj2 + proj3 + FinalExam

You also need to achieve at least 20 marks in the final exam to pass the course.

Projects

- Projects:
  - 1 project on MapReduce
  - 1 project on Spark
  - 1 project on a real cloud platform (Dataproc)

- Both results and source codes will be checked.
  - If not able to run your codes due to some bugs, you will not lose all marks.

Final exam

- Final online written exam (50 pts)
- If you are ill on the day of the exam, do not attend the exam – I will not accept any medical special consideration claims from people who already attempted the exam
- **You need to achieve at least 20 marks in the final exam**
- No supplementary exam will be given

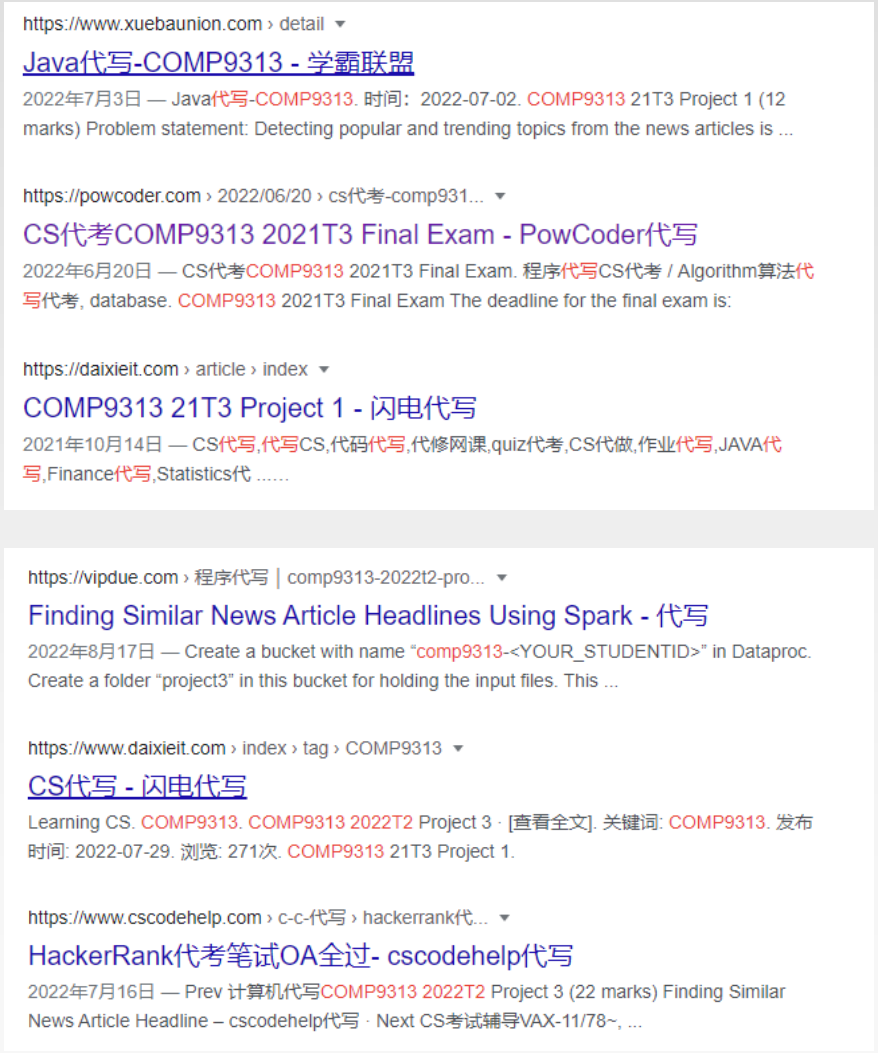
You May Fail Because …

- *Plagiarism*
- Code failed to compile due to some mistakes
- Late submission
  - 1 sec late = 1 day late
  - submit wrong files
- Program did not follow the spec

- I am unlikely to accept the following excuses:
  - “Too busy”
  - “It took longer than I thought it would take”
  - “It was harder than I initially thought”
  - … …

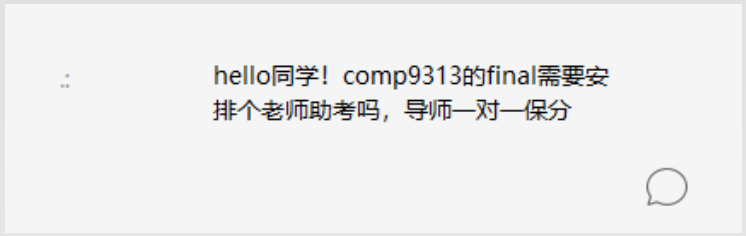
Plagiarism Detection

- Plagiarism detection will be performed on all projects



Plagiarism Detection

- Plagiarism detection will be performed on the final exam



Computing Environment

- Use Linux/command line (a virtual machine image will be provided)
  - Projects marked on Linux servers
  - You need to be able to upload, run, and test your program under Linux
- Assignment submission
  - Use Moodle to submit and check your marks

Tentative Course Schedule

Week	Topic	Assignment
1	Course info and introduction to big data	
	Hadoop, HDFS, and YARN	
2	Hadoop MapReduce 1	
3	Hadoop MapReduce 2	
4	Spark 1	Proj1 due
5	Spark 2	
6	Recess Week	
7	Finding Similar Items	Proj2 due
8	Streaming data mining	
9	NoSQL, HBase, and Hive	
10	Link Analysis/Revisions	Proj3 due

Labs

- 1 lab on Hadoop setup HDFS practice
- 3 labs on MapReduce
- 3 labs on Spark
- 1 lab on high level MapReduce tools
- 1 lab on a real cloud platform (AWS/Dataproc)

Virtual Machine

- Software: Virtualbox/VMWare
- VM image:
  - Xubuntu 22.04
    - Download the VM image at: https://drive.google.com/file/d/1ymUkS422jiNnEKU2witPb2fIL8wf6eME/view?usp=sharing
    - Open VirtualBox, File->Import Applicance
    - Browse the image folder, select the "*.ova" file
    - The image will be imported to your computer, which may take 10 minutes
    - comp9313 is used as both username and password.
- You can also install Hadoop and Spark in your own computer (Linux and Mac OS)

Your Feedbacks Are Important

- Big data is a new topic, and thus the course is tentative
- The technologies keep evolving, and the course materials need to be updated correspondingly
- Please advise where I can improve after each lecturer, at the discussion and QA website
- myExperience system
  - Please focus on the teaching…

Reasonably allocate task progress. I support 3 assignments but hope to allow time for review at the end of the semester.

more interactive questions during lectures would be good

Maybe give us more confidence when the course just started. He described this course as an extremely hard one but it's actually achievable

too much homework and too difficult

All the lab should be in person, online lab is useless

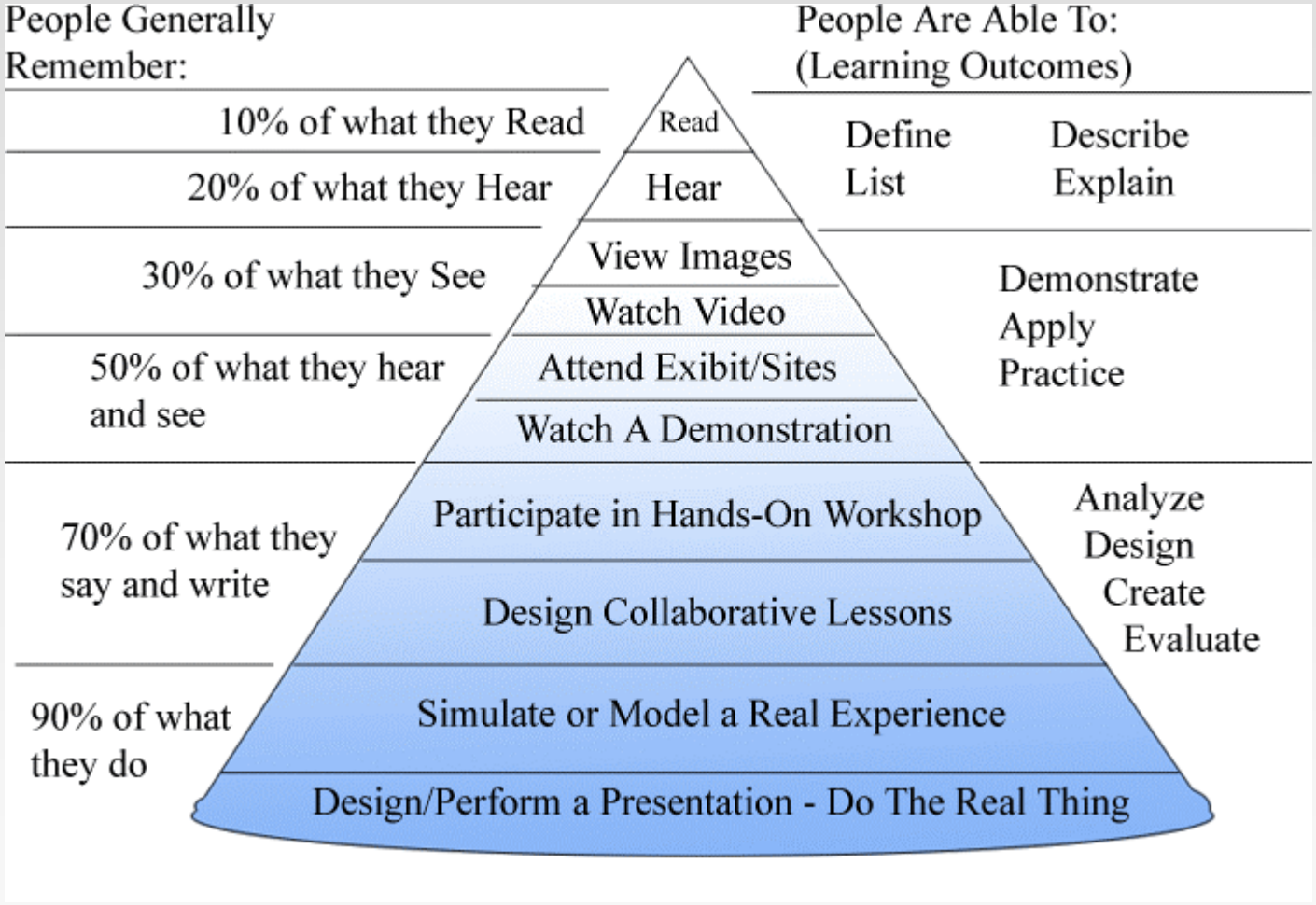
lab instruction can be formatted better – try to make it a markdown document and render it better
can use ed forum rather than webcms3 forum

It would be better if Mr Cao can speak more concisely and make sure the pronunciations of words are correct (it is not about accent as he mentioned in the first lecture).

The course was poorly taught and it was very difficult to understand the lecturer. I personally had to rely on online resources to learn the content myself.

sometime I feel like that I am a examiner of IELTS speaking and the lecturer is a nervous examinee

Why Attend the Lectures?



Part 2: Introduction to Big Data

What is Big Data?

- No standard definition! here is from Wikipedia:
  - Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software
  - Challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, information privacy, and data source..
  - The term "big data" often refers simply to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from big data, and seldom to a particular size of data set.

Instead of Talking about "Big Data"…

- Let's talk about a crowded application ecosystem:
  - Hadoop MapReduce
  - Spark
  - High-level query languages (e.g., Hive)
  - NoSQL (e.g., HBase, MongoDB, Neo4j)
  - Pregel
  - Flink
  - … …
- Let's talk about data science and data management:
  - Finding similar items
  - Graph data processing
  - Streaming data processing
  - … …

Who is generating Big Data?

Social

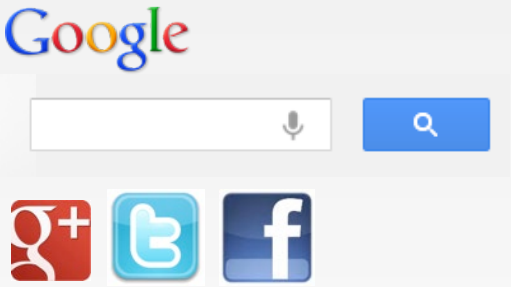
User Tracking & Engagement

Homeland Security

eCommerce

Financial Services

Real Time Search

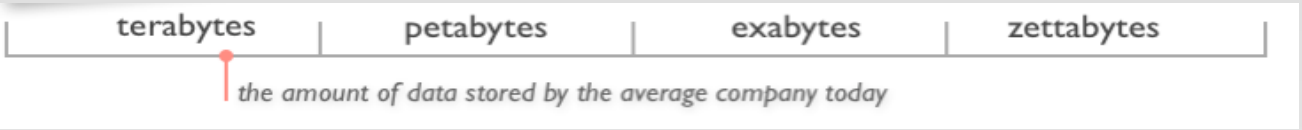


Big Data Characteristics: 3V

- The Vs of big data were often referred to as the "three Vs"
  - **Volume**: In a big data environment, the amounts of data collected and processed are much larger than those stored in typical relational databases.
  - **Variety**: Big data consists of a rich variety of data types.
  - **Velocity**: Big data arrives to the organization at high speeds and from multiple sources simultaneously.

Volume (Scale)

- In the big data era, huge amount of data is being generated every day



Recent Twitter statistics

Quick Twitter Statistics

Total Number of Monetizable Daily Active Users:

217 million (source)

Last updated: 21/02/22

Total Number of Tweets Sent per Day:

500 million (source)

Last updated: 21/02/22

Q4 2022 Total Twitter Revenue:

$1.57 billion (source)

Last updated: 21/02/22

The number of US Adults Who Use Twitter:

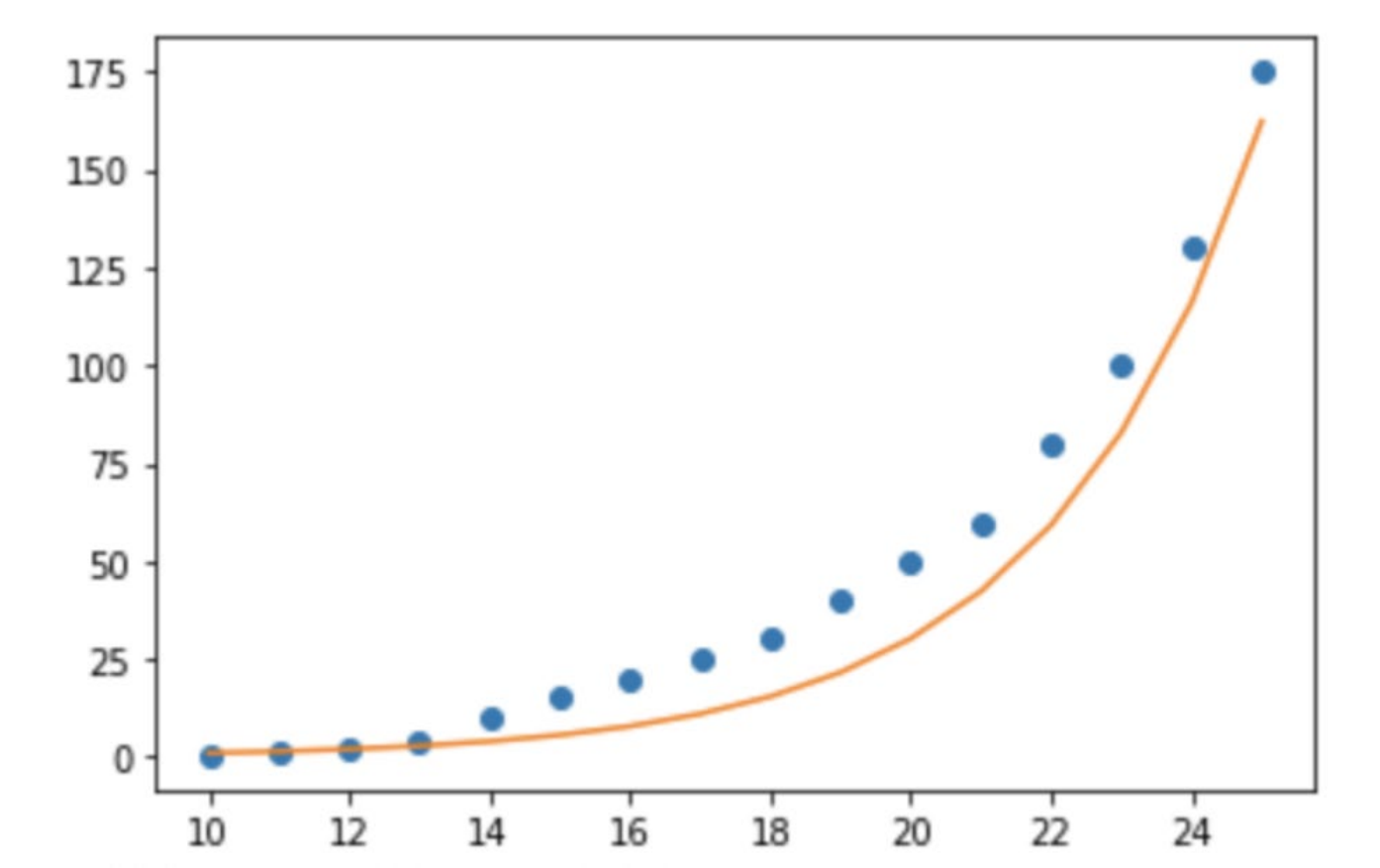
23% (source)

Last updated: 21/02/22

https://www.omnicoreagency.com/twitter-statistics/

Volume (Scale)

- Data volume is increasing exponentially (40% increase per year)



Data amount in Zetabytes from 2010 to 2025

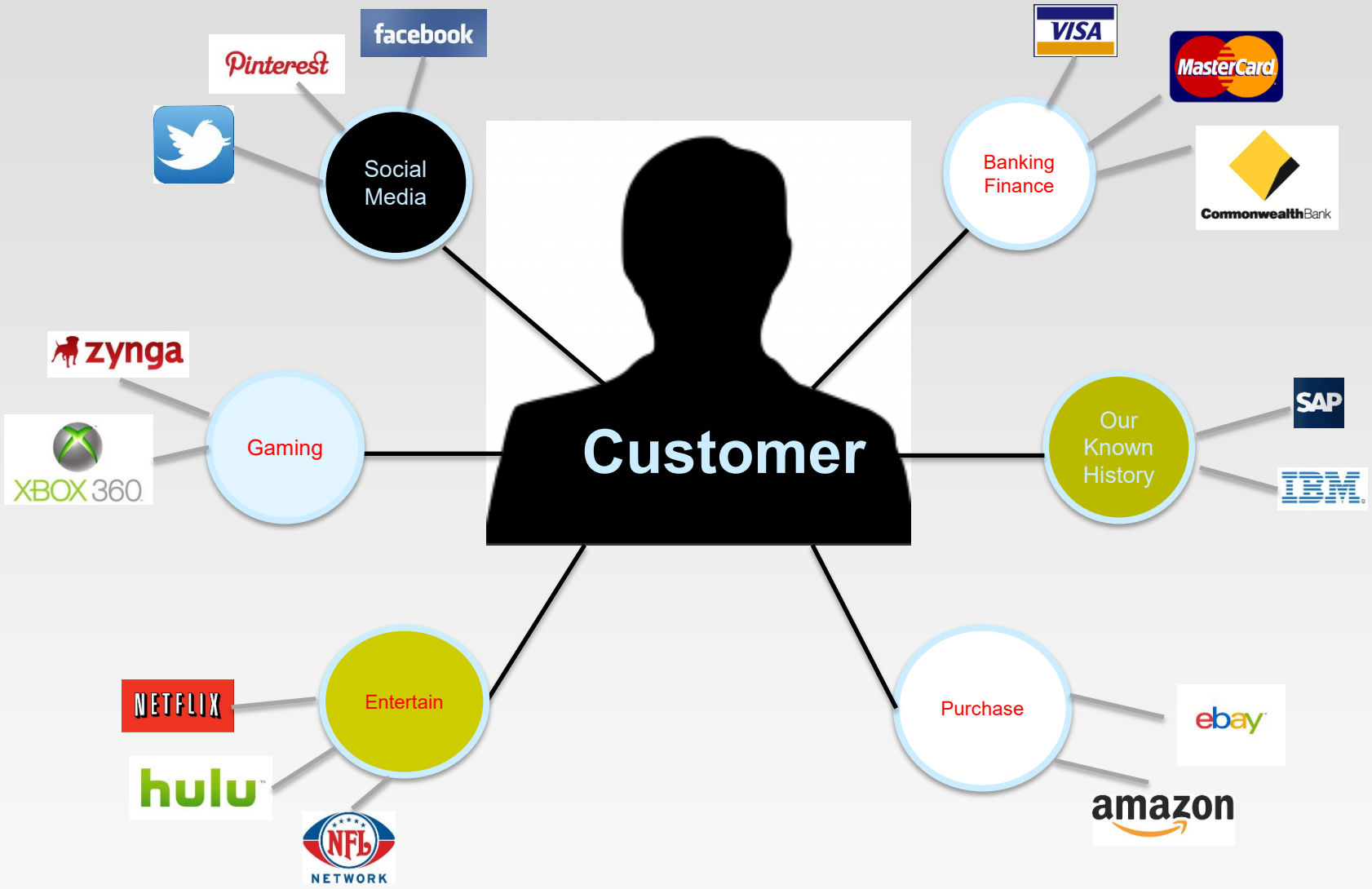
A forecast by IDC & SeaGate. Image by Sven Balnojan.

Variety (Complexity)

- Different Types:
  - Relational Data (Tables/Transaction/Legacy Data)
  - Text Data (Web)
  - Semi-structured Data (XML)
  - Spatial Data
  - Temporal Data
  - Graph Data
    - Social Network, Semantic Web (RDF), …
  - One application can be generating/collecting many types of data
- Different Sources:
  - Movie reviews from IMDB and Rotten Tomatoes
  - Product reviews from different provider websites

To extract knowledge➔ all these types of data need to linked together

A Single View to the Customer



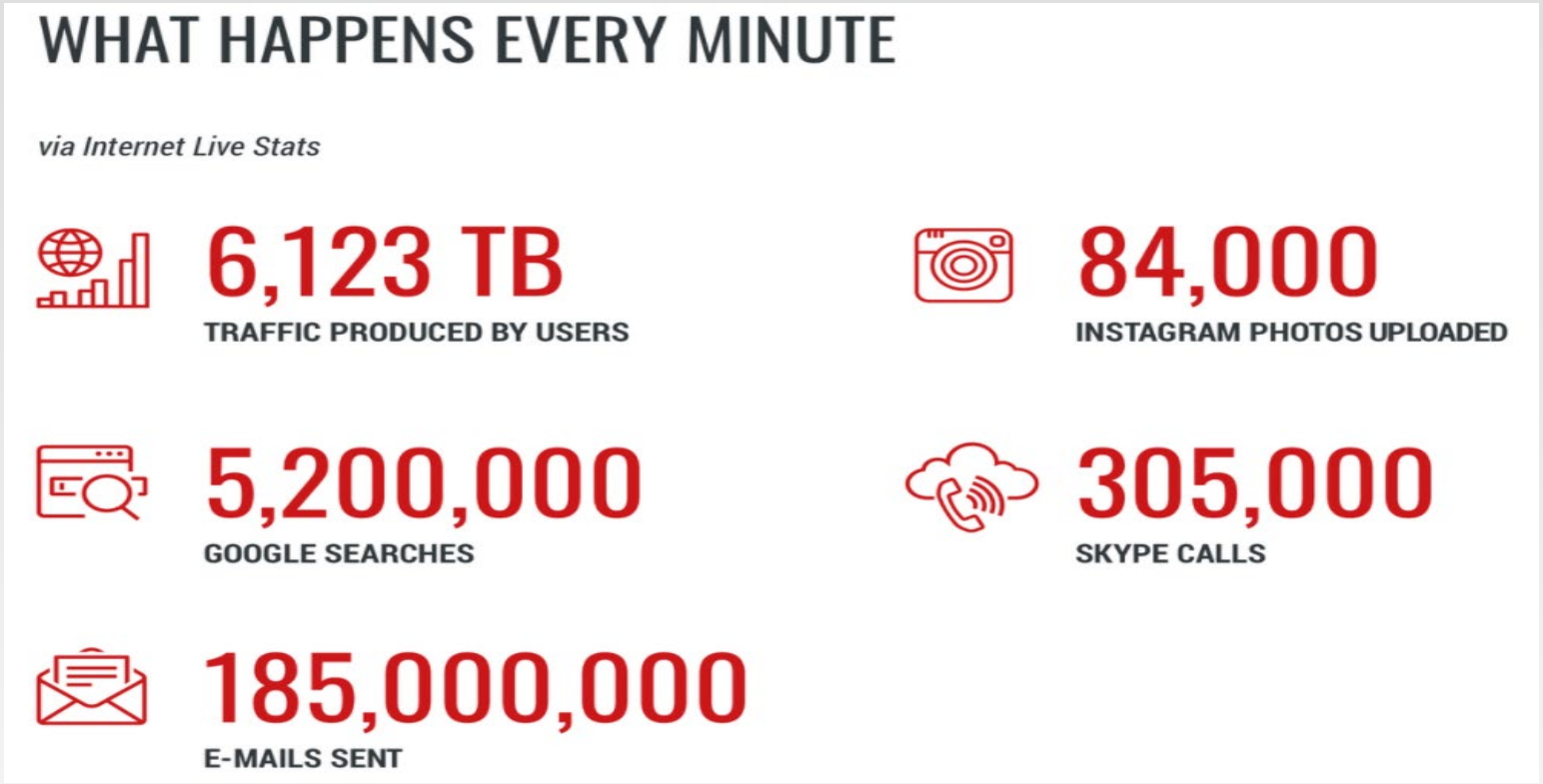
A Global View of Linked Big Data

Diversified social network

Velocity (Speed)

- Velocity essentially measures how fast the data is coming in.
- Data is being generated fast and need to be processed fast
  - Late decisions -> missing opportunities
- It is usually met in online data analytics, for example
  - **E-Promotions**: based on your current location, your purchase history, what you like -> send promotions right now for store next to you
  - **Healthcare monitoring**: sensors monitoring your activities and body -> any abnormal measurements require immediate reaction
  - **Pandemic management and response**: contact tracing for new infected COIVD-19 cases and future hotspots prediction to slow down the spread of infectious diseases

Velocity in Real-world



- The statistics for 1 second in many applications.
  http://www.internetlivestats.com/one-second/

Extended Big Data Characteristics: 6V

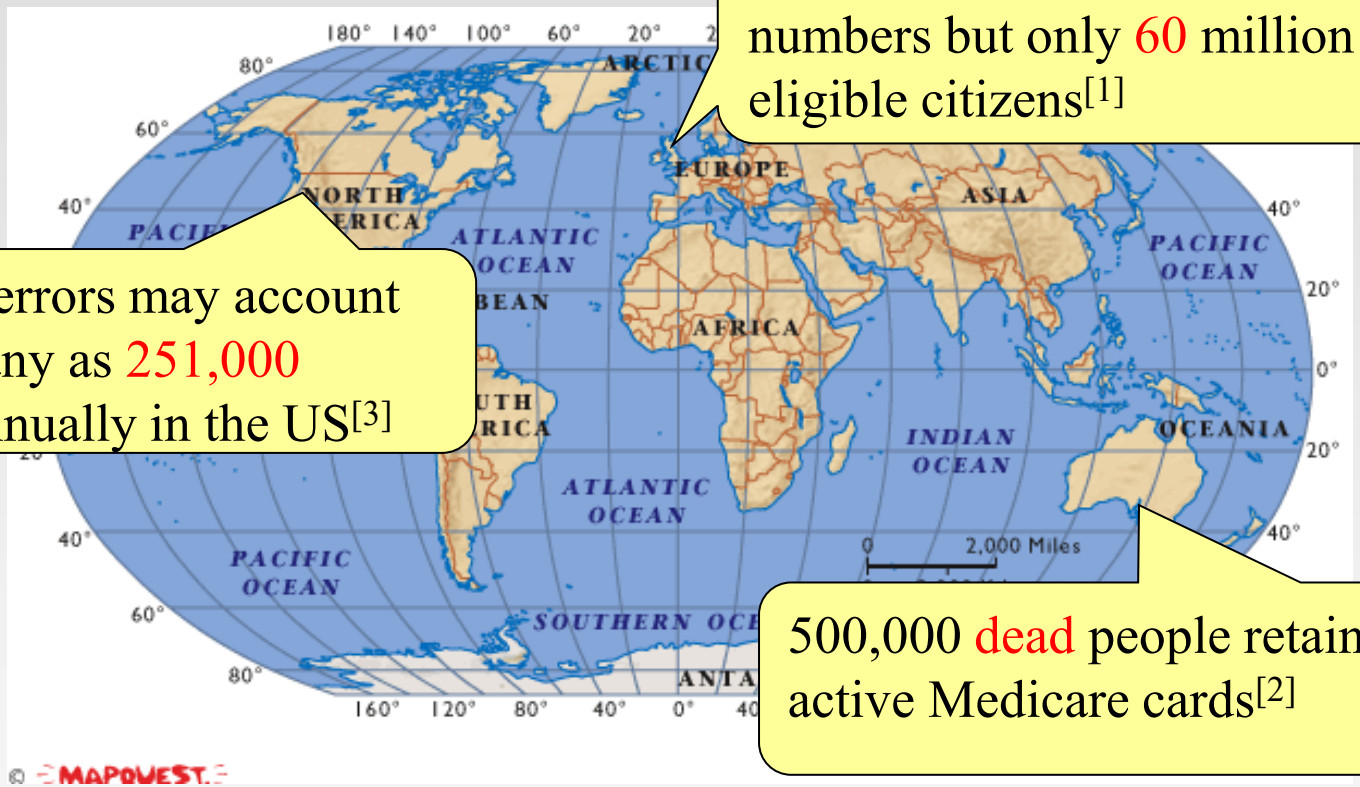
- Volume: In a big data environment, the amounts of data collected and processed are much larger than those stored in typical relational databases.
- Variety: Big data consists of a rich variety of data types.
- Velocity: Big data arrives to the organization at high speeds and from multiple sources simultaneously.

- Veracity: Data quality issues are particularly challenging in a big data context.
- Value: Ultimately, big data is meaningless if it does not provide value toward some meaningful goal.
- Visibility/Visualization/Variability/Validity/….

Veracity (Quality & Trust)

- Data = quantity + quality
- When we talk about big data, we typically mean its quantity:
  - What capacity of a system provides to cope with the sheer size of the data?
  - Is a query feasible on big data within our available resources?
  - How can we make our queries tractable on big data?
  - . . .
- Can we trust the answers to our queries?
  - Dirty data routinely lead to misleading financial reports, strategic business planning decision -> loss of revenue, credibility and customers, disastrous consequences
- *The study of data quality is as important as data quantity*

Data in real-life is often dirty



[1] https://publications.parliament.uk/pa/cm200001/cmhansrd/vo010327/debtext/10327-21.htm
[2] https://www.privacy.org.au/Campaigns/ID_cards/MedicareSmartcard.html
[3] Your Health Care May Kill You: Medical Errors. https://pubmed.ncbi.nlm.nih.gov/28186008/

Value

- Big data is meaningless if it does not provide value toward some meaningful goal

Other V's

- Visibility: the state of being able to see or be seen is implied.
  - Big Data – visibility = Black Hole?

- Visualization: Making all that vast amount of data comprehensible in a manner that is easy to understand and read.

A visualization of Divvy bike rides across Chicago

- Big data visualization tools:

BI/Visualization/ Analytics

pentaho COGNOS tableau sas

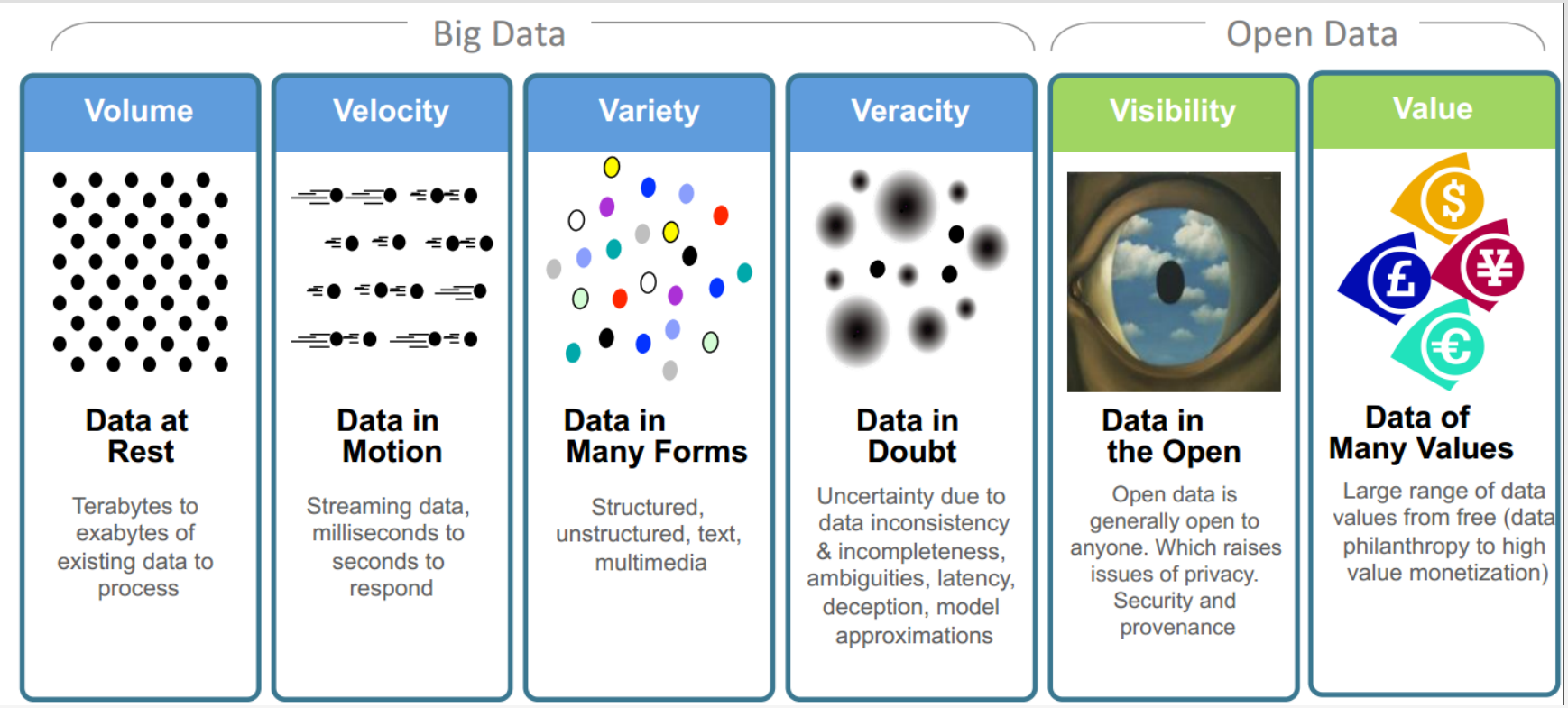
looker

DOMO ClearStory QlikView MicroStrategy Business Objects

Other V's

- Variability
  - Variability refers to data whose meaning is constantly changing. This is particularly the case when gathering data relies on language processing.
  - It defines the need to get meaningful data considering all possible circumstances.
- Viscosity
  - This term is sometimes used to describe the latency or lag time in the data relative to the event being described. We found that this is just as easily understood as an element of Velocity.
- Volatility
  - Big data volatility refers to how long is data valid and how long should it be stored. You need to determine at what point is data no longer relevant to the current analysis.
- More V's in the future …
  - How many v's are there in big data? http://www.clc-ent.com/TBDE/Docs/vs.pdf

Big Data: 6V in Summary



Transforming Energy and Utilities through Big Data & Analytics. By Anders Quitzau@IBM

Tag Clouds of Big Data

Why Study Big Data Technologies?

- The hottest topic in both research and industry
- Highly demanded in real world
- A promising future career
  - Research and development of big data systems:

    distributed systems (eg, Hadoop), visualization tools, data warehouse, OLAP, data integration, data quality control, …
  - Big data applications:

    social marketing, healthcare, …
  - Data analysis: to get values out of big data

    discovering and applying patterns, predicative analysis, business intelligence, privacy and security, …

- Get enough credits

Big Data Open-Source Tools

https://datafloq.com/big-data-open-source-tools/os-home/

What Will the Course Cover

- Topic 1. Big data management tools
  - Apache Hadoop
    - YARN/HDFS/Hive (briefly introduced)
    - MapReduce
    - Spark
    - NoSQL (HBase)
    - Google Dataproc
- Topic 2. Big data typical applications
  - Finding similar items
  - Graph data processing
  - Data stream mining
  - Link Analysis

End of Chapter 1.1