DS 420
Big Data

Instructor:
Dr. Peilong Li

Lecture 1:

Course overview; Intro to Big Data

Hello!

Welcome to DS 420 Big Data

You are about to embark on an awesome journey of data analytics at large scale

Lecture outline

Announcements/notes

- Finish Reading 1 this week.
- HW 1 out, due by next Monday

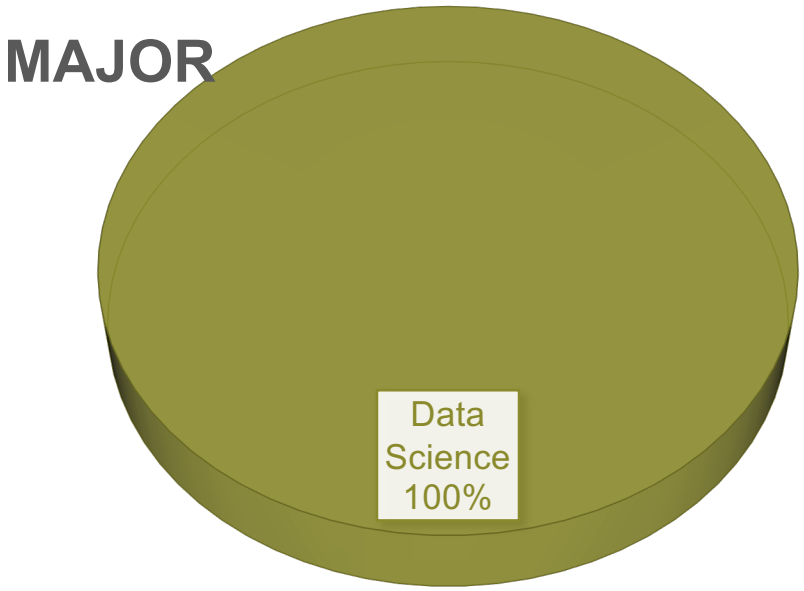
Today's lecture

- Course overview
  - Instructor information
  - Course materials, policies and resources
  - Course outline
- Intro to Big Data

Who am I?

- **Dr. Peilong Li**
- **E-mail:** lip@etown.edu
- **Office:** Esbenshade 284B
- **Phone:** (717) 361-4761
- **Office hours:**
  - MWF 2-3:30 PM and by appointment
  - https://calendly.com/dr_li/
  - Student questions are top priority during these hours
  - Also in office during my lunch break.
- **Personal/Course Website:**
  - https://Peilong.github.io/

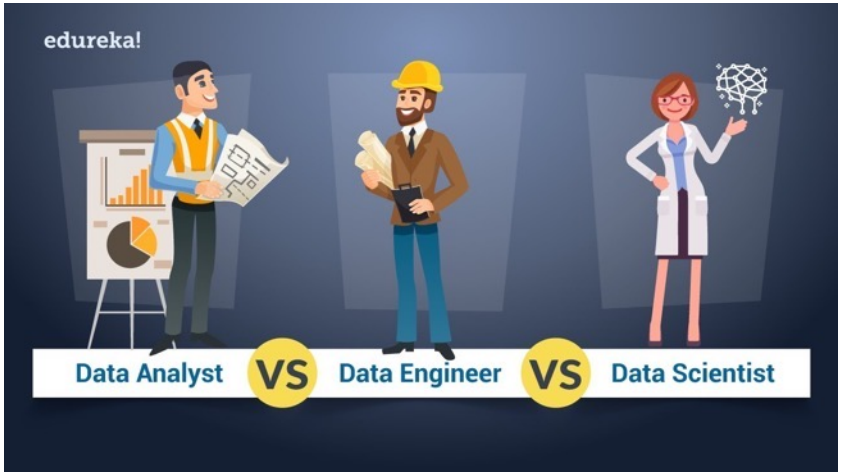
Who are you?



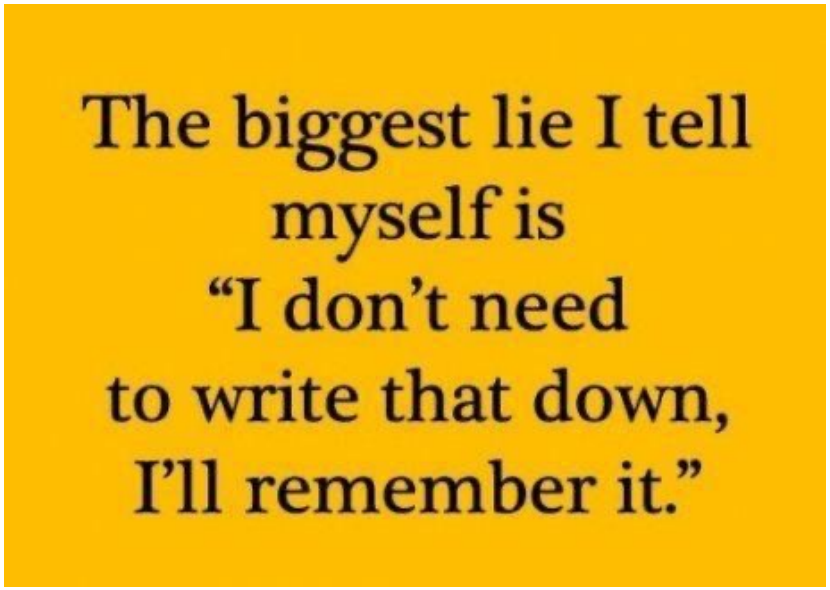
- Survivors of DS 200 and/or DS 300
- Emerging data analysts/engineers
- Your turn to share your holiday fun, or your plan this summer.

Why are we here?

- What's the purpose of this class?
  - Tame humongous of data around us.
  - Machine learning at scale.
  - Data engineering
  - Enrich your resume with buzzwords: Hadoop, Spark, Kafka, etc.



What I Encourage



- Prepare ahead and take notes during
- Code along with me
- Program independently

COURSE OVERVIEW

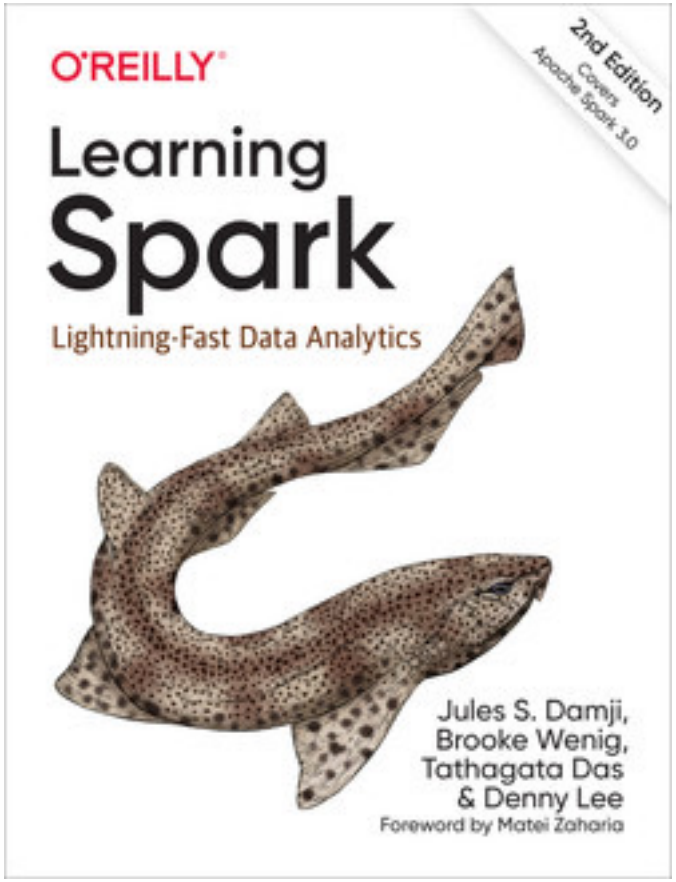
Course focuses

- Variety
  - A wide spectrum of software components
  - Different use case scenarios
- Volume
  - Store, extract and analyze data at large scale
- Programmability
  - Tame the big data with various programmable weapons
- Velocity
  - Handle and analyze streaming data in real-time

Tentative course outline

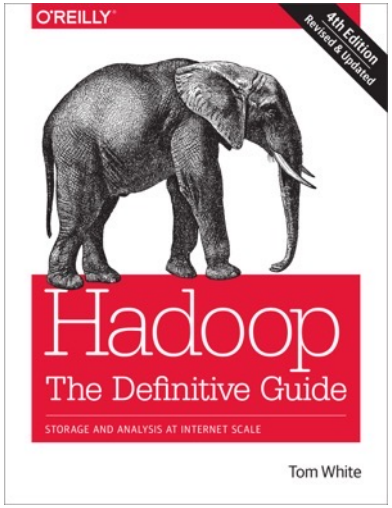
- Hadoop ecosystem
- HDFS and MapReduce
- Scripting and SQL – Pig
- NoSQL databases – Hbase, MongoDB
- Apache Spark – RDD, DataFrame
- Spark MLlib – Linear regression, Logistic regression, trees, k-means
- Uses cases – NLP, etc.
- Stream processing – Flume, Kafka, Spark Streaming

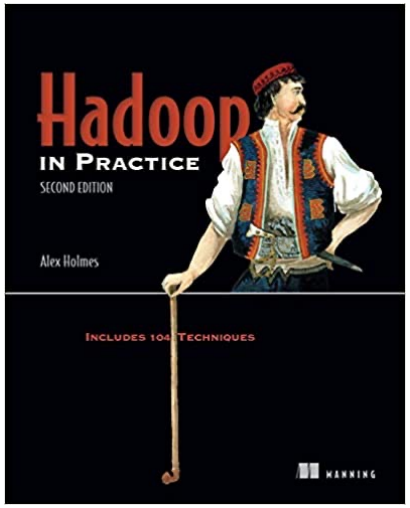
Required Textbook

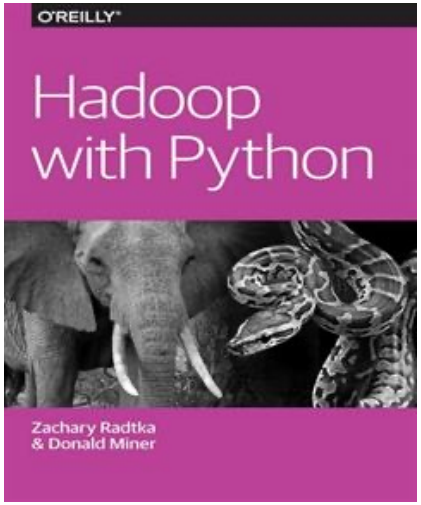


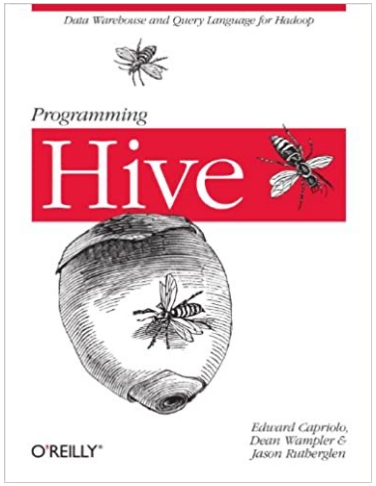
- Damji et al.
- ***Learning Spark, 2nd Edition***
- 2020, O'Reilly Media
- **Required textbook**
- A legitimate e-copy of the book is posted on the course website.

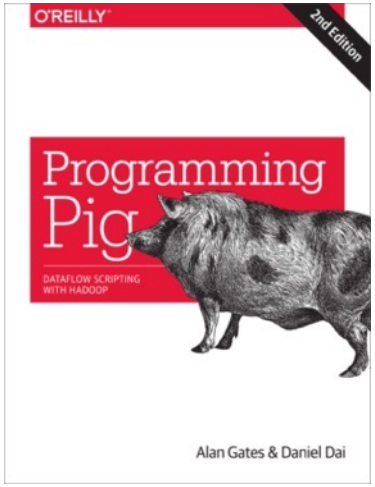
Optional Textbooks

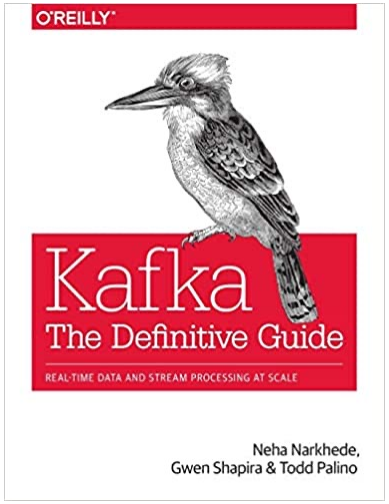


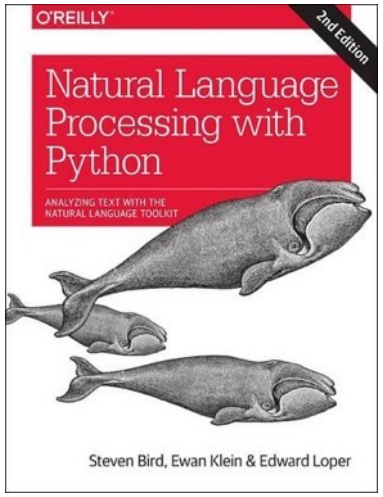












Reading Assignments

- Thank goodness you don't need to buy them all – at least for now!
- Chapter excerpts from these textbooks as reading assignments.
- Take a look at the schedule page …
  - **LS** refers to the required textbook.
  - Reading1-5
  - Paper1-5

Reading Assignments

Reading will be assigned for each week.

Need to finish before the start of next week.

E.g., week 1's reading is the preparation for week 2.

Reading helps your homework and exams.

Course tools

- Campus server:
  - Available to you at: http://adal.etown.edu
- PySpark on Jupyter Notebook:
  - Your familiar environment
  - Available to you at: http://adal.etown.edu:12345
  - Install by yourself? Dependencies:
    - Java JDK and JRE
    - Apache Spark version 3.2.1
    - Anaconda 3
    - PySpark

Optional Software

- Hortonworks Data Platform (HDP)
  - Sandbox version 2.6.5 (3.xx is available)
  - Installs with either virtual machine or Docker
  - Want to install on your own computer?
    - Recommend Docker installation
    - Requires 4 cores and 8GB of **FREE** RAM
    - Takes hundred GBs of disk space

Additional course materials

- Course website (bookmark it now):
  - https://peilong.github.io/courses/ds420/
  - The **up-to-date** schedule
  - Detailed course policies and guidelines
- Canvas will contain:
  - Course announcements
  - Lecture slides, handouts, assignments.
  - Reminders about projects and exams.
  - Syllabus (you can request a printed copy if you like).

Course Announcements & Discussion

- All course announcements will be posted via the Canvas "**Announcements**" page.
  - You are responsible for checking the email updates (or app notifications) from Canvas regularly.
- Discussions are encouraged to be posted via Microsoft Teams. Invitation link is on Syllabus.
- Private instant messages can be sent directly to your recipient on the channel by searching on the "search" bar.

Grading and important due dates

- Grading breakdown
  - X pop quizzes: 4%
  - Lecture presentation: 7%
  - 6 take-home quizzes: 24% - 4% each
  - 5 homework: 20% - 4% each
  - 2 midterm exams: 25%
  - Team final project: 20%

Grading scale

A	93-100	B-	80-82	D+	67-69
A-	90-92	C+	77-79	D	63-66
B+	87-89	C	73-76	D-	60-62
B	83-86	C-	70-72	F	0-59

TAKE-HOME QUIZZES

Programming exercises

- Note on course schedule: several days marked as “PE #”
  - Those classes will contain supervised, in-class programming exercises
    - We’ll write/complete short data analytics tasks to illustrate previously covered concepts
  - If you have a laptop, feel free to bring it
  - Extremely helpful to prepare for the take-home quizzes

Take-home quizzes

- Labs are re-branded as "take-home quizzes"
- Mini assignments that can be completed in 80 minutes.
- Runs week-long.
- PE during the same week to get you ready.
- No late submission.
- Work should be done individually, but the goal is to learn, and I will help everyone

Homework assignments

HOMEWORK

Homework

- Will submit all code via Canvas
  - https://etown.instructure.com/
  - I will start grading on the due date
  - Grade period is given
- Penalty after due date: $-(2^{n-1})$ points per day
  - i.e., -1 after 1 day, -2 after 2 days, -4 after 3 days …
  - Assignments that are 8+ days late receive 0

Homework

- Must "run all" without exceptions
  - If your program throws an exception somewhere, you gain no points starting from the first exception
- Must be done individually
  - If I can ascertain that code from one student's project appears in another student's project, both projects will score zero points
  - Both students will also have a full letter grade reduction at the end of the semester

EXAMS

Exams format

- Two in-class midterm exams totaling 25%
- Open book, open Internet
- Somewhat similar to labs and homework
- More comprehensive than a single lab
- Focus on solving big data problems with programming and analytical skills
- There will be open-ended questions that test your innovation.

LECTURE PRESENTATION

Paper Reading and Presentation

- Why do I have to read and present a paper?
  - Data science tech is changing day by day. You need to keep up with the world.
  - Learn how to use the library database and online resources.
  - Public speaking skill!!! You want to articulate your findings to your boss/team.
  - Understand the structure of a paper for your final project report.

Individual Lecture Presentation

- Note a few days on the schedule denoted as "Paper X Presentation"
- Claim your paper and read it.
- Present as an instructor. 30 minutes
  - Need to be interactive
  - Need to have a live demonstration with the presented technology
  - Followed by a 10-minute Q&A
  - Will be evaluated by peers and the instructor.

Teamwork!!!!

FINAL PROJECT

Teamwork is the key!

- Why do I need a team?
  - Mimics real-world data positions.
- Do I need to find the partner by myself?
  - Yes, unless you explicitly ask me for help.
- What's the team size?
  - Two. Soloists are not permitted unless the class size is odd.

Teamwork is the key!

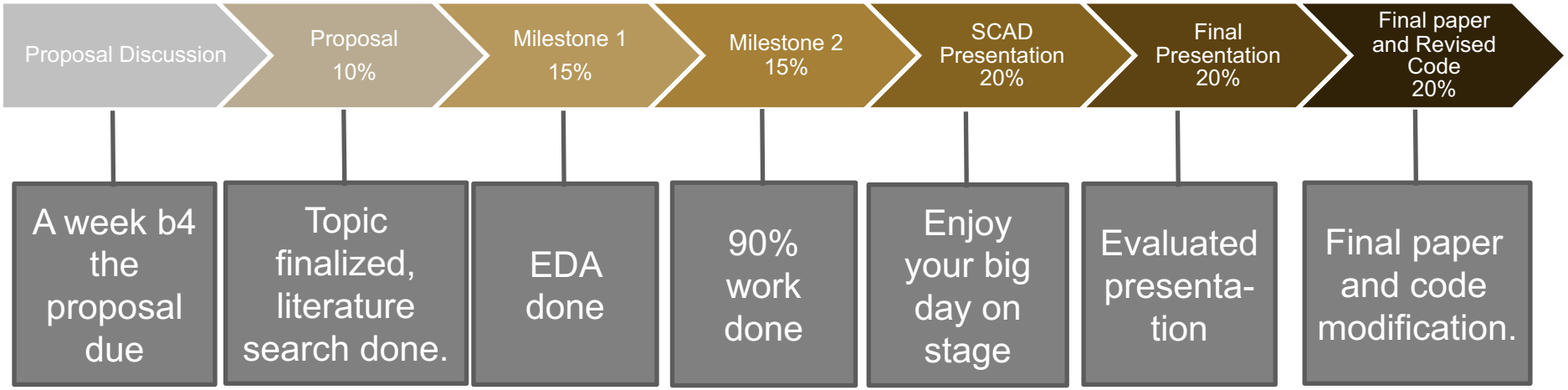
- How to show our teamwork?
  - Members cannot "work on the same piece of code" - must have distinct tasks.
  - Your individual notebook for your part of work needs to be explicit. Combining them as a team for group submission.
  - Will be graded differently if I can discern the workloads are not even.

Final project

- Find your own project topics
  - Discuss your topic with the instructor a week before the proposal due date
  - Proposal writing requirements will be posted on Canvas.
- Potential project directions
  - Develop a data-powered tool
  - Extend your previous projects from DS200 and DS300
  - Find an open dataset, solve the data problem and render the knowledge
    - Dataset size should be over 100 MB.
    - Should avoid using Pandas

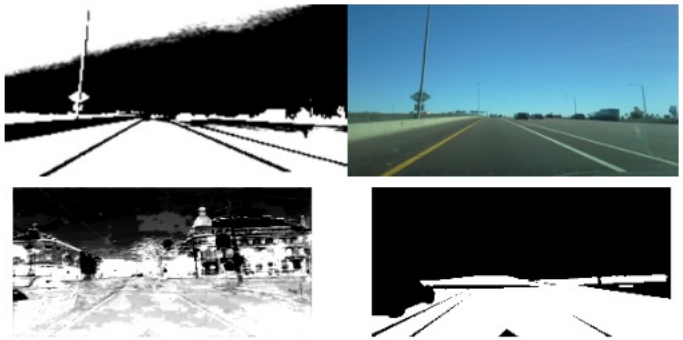
Final project

- Timeline:
  - Check on the course schedule page for the exact date: https://peilong.github.io/courses/ds420/schedule/

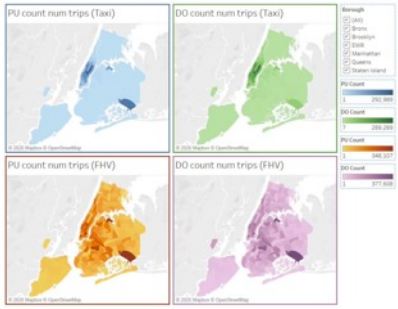


Food for Thoughts

Semantic segmentation for autonomous driving



NYC taxis vs. for-hire vehicles



POLICIES

Attendance

- You are expected to attend classes in person unless I explicitly mention "no class today".
  - Check your emails often.
- You are expected to have read the material we are going to cover **before** class
- Missed exams cannot be made up
- Exams and labs must be made up **before** the scheduled time, for excused absences

Academic honesty

- Don't cheat (theme of all classes this week)
- Some misunderstandings:
  - Harm your fellow students
    - True if graded on a curve
  - Harm the reputation of E-Town College
- Real reason:
  - Constructing the person you are going to be
  - Behavior ←→habits
  - Hurting yourself

Disability

Elizabethtown College welcomes otherwise qualified students with disabilities and is committed to providing access for all students to courses, programs, services, and activities. If you have a documented disability such as a learning disability or chronic illness or a new circumstance such as a concussion and would like to request accommodations please contact the Director of Disability Services by phone (717-361-1227) or e-mail (daviesl@etown.edu). The Office of Disability Services can provide resources to you and facilitate communication with faculty about reasonable accommodations. After meeting with the Office of Disability Services, please set up an appointment to meet with me, the instructor, to discuss the accommodations as they pertain to my class.

Diversity and Inclusivity Statement

- I consider this classroom to be a place where you will be treated with respect, and I welcome individuals of all ages, backgrounds, beliefs, ethnicities, genders, gender identities, gender expressions, national origins, religious affiliations, sexual orientations, ability - and other visible and non-visible differences. All members of this class are expected to contribute to a respectful, welcoming and inclusive environment for every other member of the class.

WHAT'S BIG DATA


American's Favorite Pie is …

See new and see better …

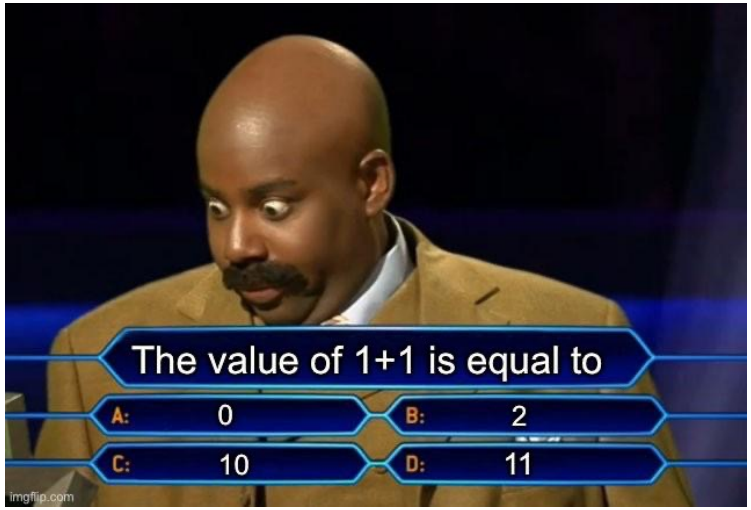
The Answer to Global Challenges

Where Was "Big Data" Originated?

- Early in the cloud era, research at companies like **Google** and **Amazon** made it clear that people respond well to social networking tools and smarter advertising placement and recommendations.

The Million-dollar Question

- The idea is simple:
  - “People with Ken’s interest find this store fantastic.”
  - “Anne really like Eileen Fisher and might want to know about this 15% off sale on spring clothing.”
  - “Sarah had a flat tire and needs new ones.”
- Seemingly simple, but there is a key obstacle.



They had lots of customers and data

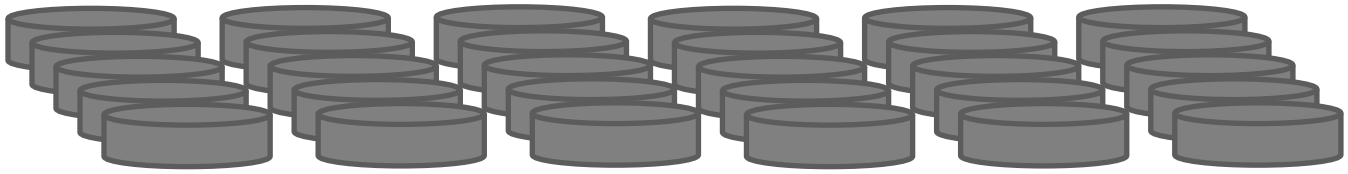
- Web search and product search tools needed to deal with **billions** of web pages and **hundreds of millions** of products.
- **Billions** of people use these platforms.
- So simple ideas still involve enormous data objects that simply can't fit in memory on modern machines.
- And yet in-memory computing is far faster than any form of disk-based storage and computing!

What are the big data files?

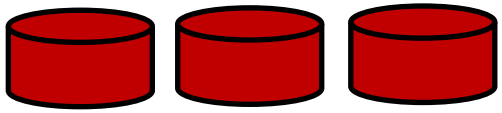
- Snapshot of all the web pages in the world, updated daily.
- Current product data & price for every product Amazon knows about.
- Social networking graph for all of Facebook

The solution: Parallel Computing

Data starts out sharded over servers



Early pipeline stages are extremely parallel: they ***extract***, ***transform***, ***summarize***



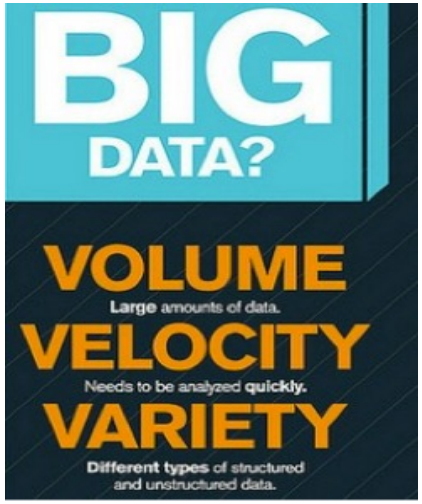
Eventually we squeeze our results into a more useful form, like a trained machine-learning model. The first stages can run for a long time before this converges

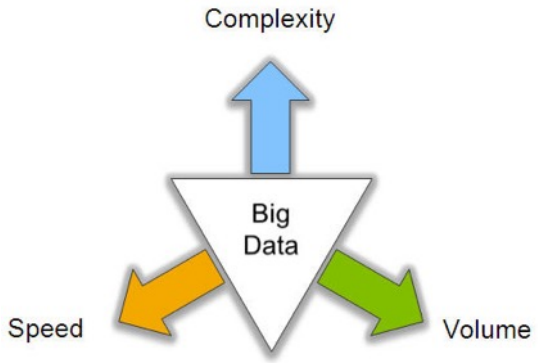
Copy the model to wherever we plan to use it.

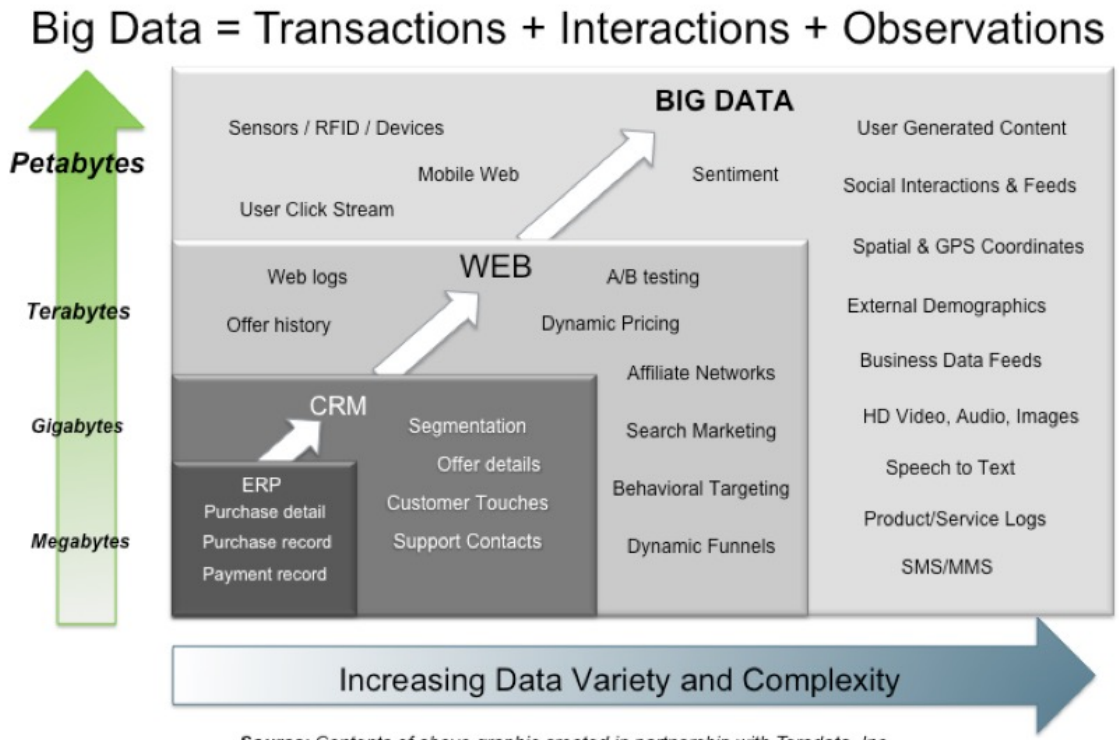
Big Data's Characteristics

BIG DATA 3 V'S

Big Data: 3V's

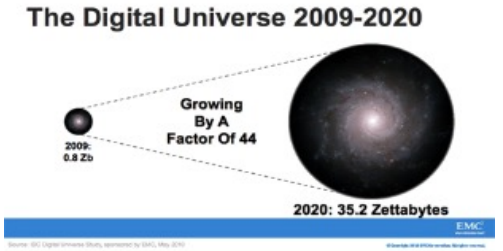




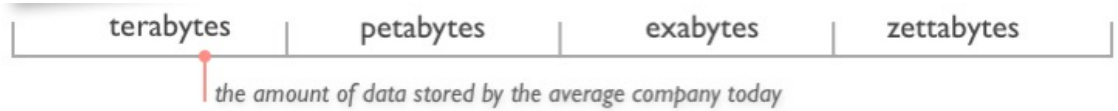


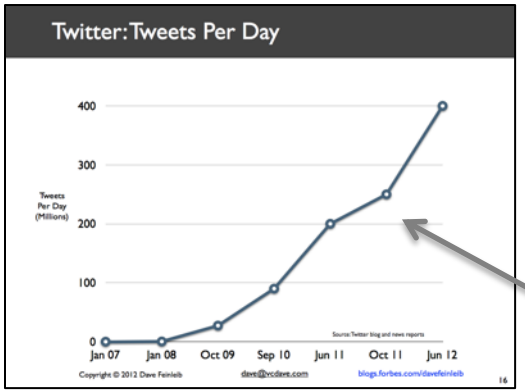
Source: *Contents of above graphic created in partnership with Teradata, Inc.*

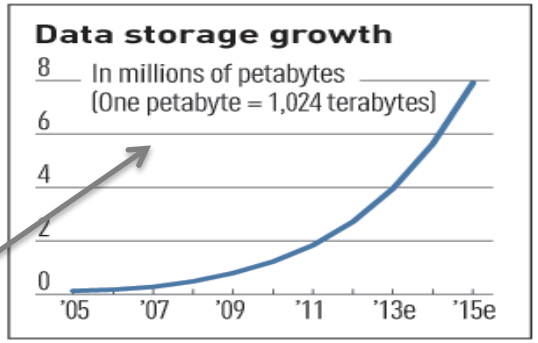
Volume (Scale)



- Data Volume
  - 44x increase from 2009 2020
  - From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially







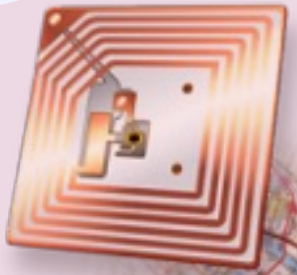
Exponential increase in collected/generated data

12+ TBs
of tweet data every day

? TBs of data every day

25+ TBs of
log data every day

30 billion RFID
tags today
(1.3B in 2005)



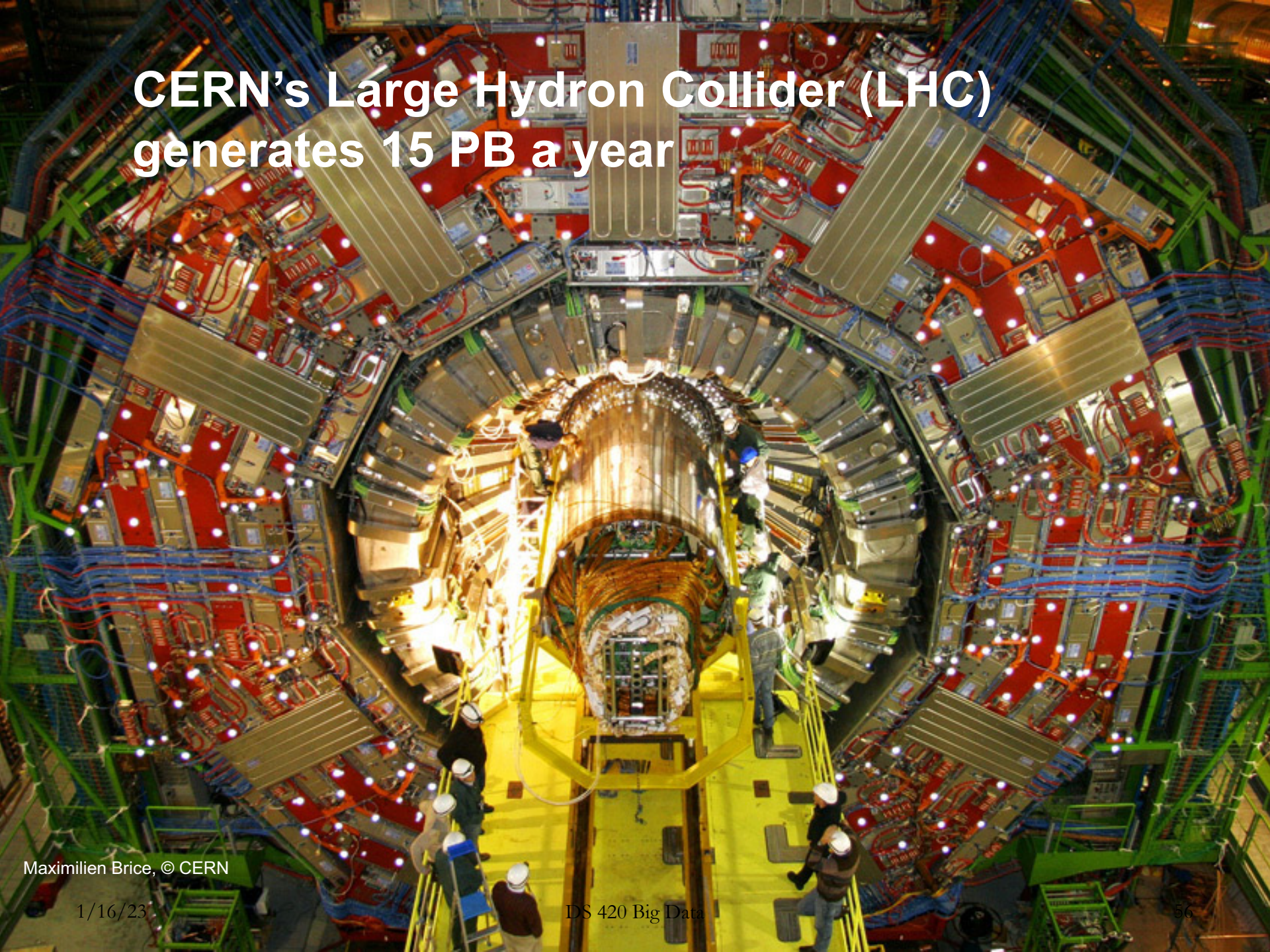
4.6 billion
camera phones world wide

100s of millions of GPS enabled
devices sold annually

76 million smart meters in 2009…
200M by 2014

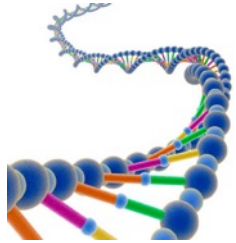
2+ billion
people on the Web by end 2011

CERN's Large Hydron Collider (LHC) generates 15 PB *a* year

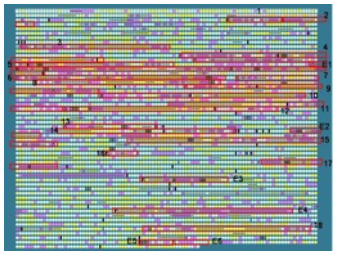


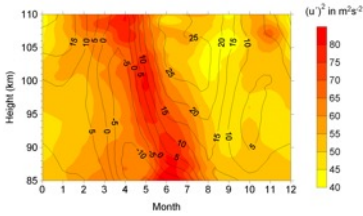
Maximilien Brice, © CERN

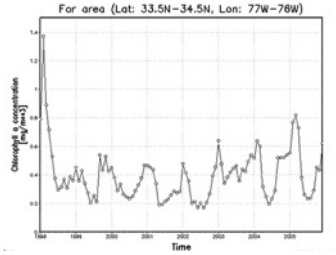
Variety (Complexity)



- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
  - Social Network, Semantic Web (RDF), …
- Streaming Data
  - You can only scan the data once
- A single application can be generating/collecting many types of data
- Big Public Data (online, weather, finance, etc)







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

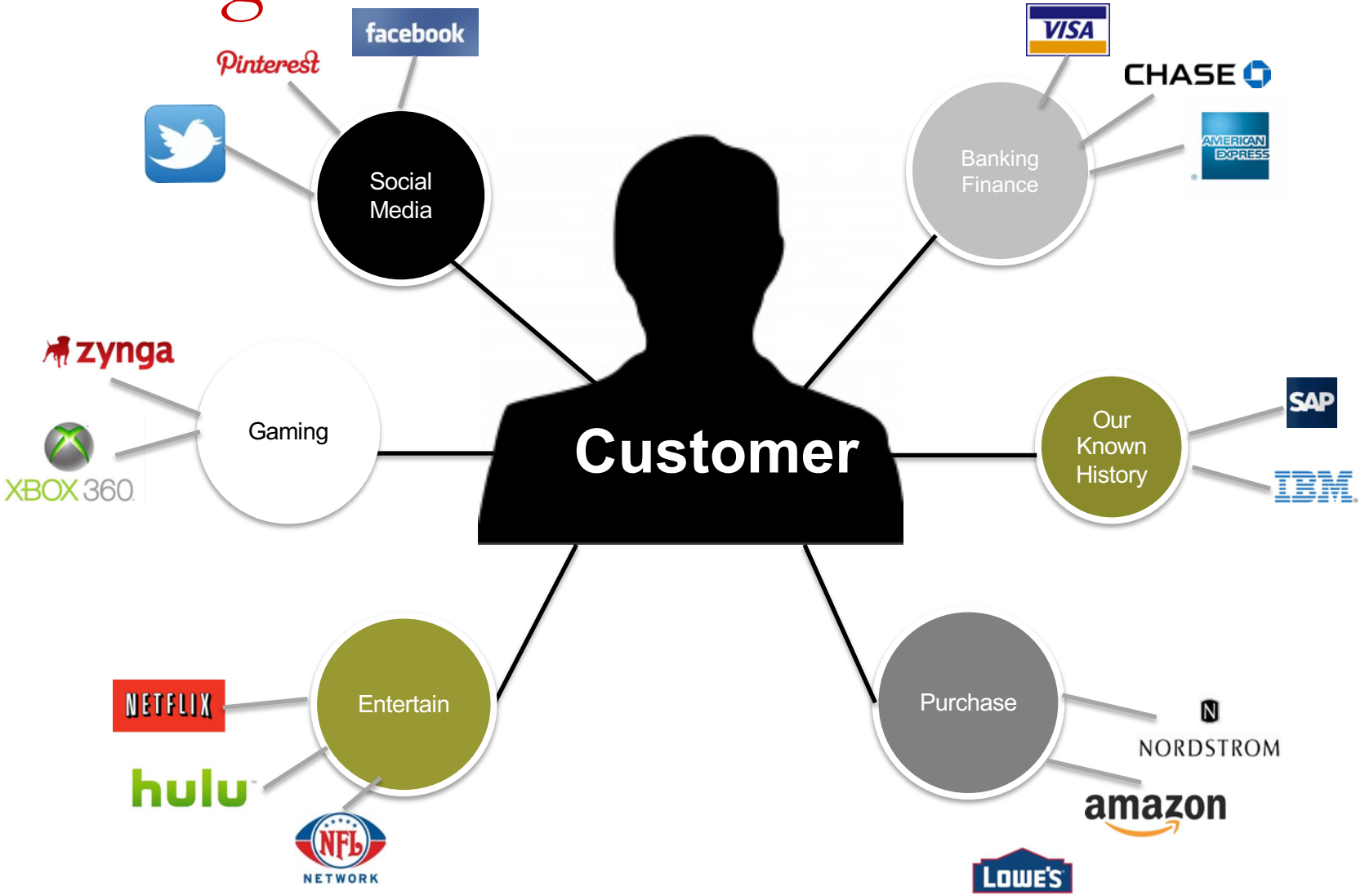
The Model Has Changed…

- The Model of Generating/Consuming Data has Changed

Old Model: Few companies are generating data, all others are consuming data

New Model: all of us are generating data, and all of us are consuming data

A Single View to the Customer

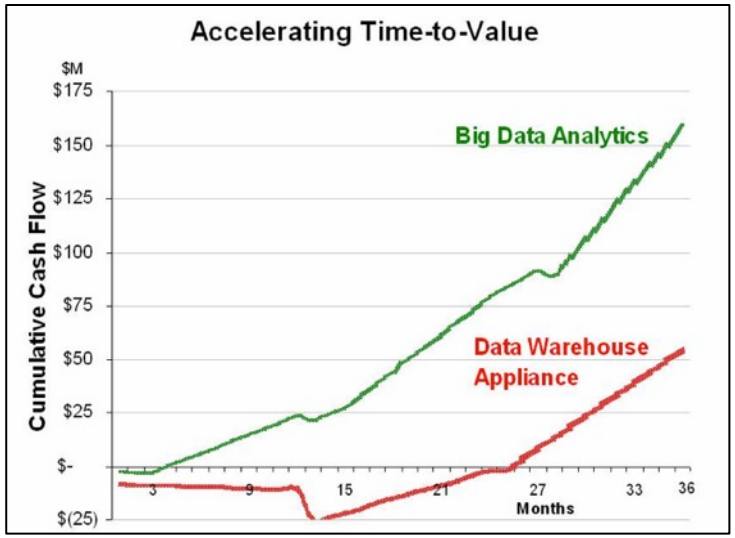


Velocity (Speed)

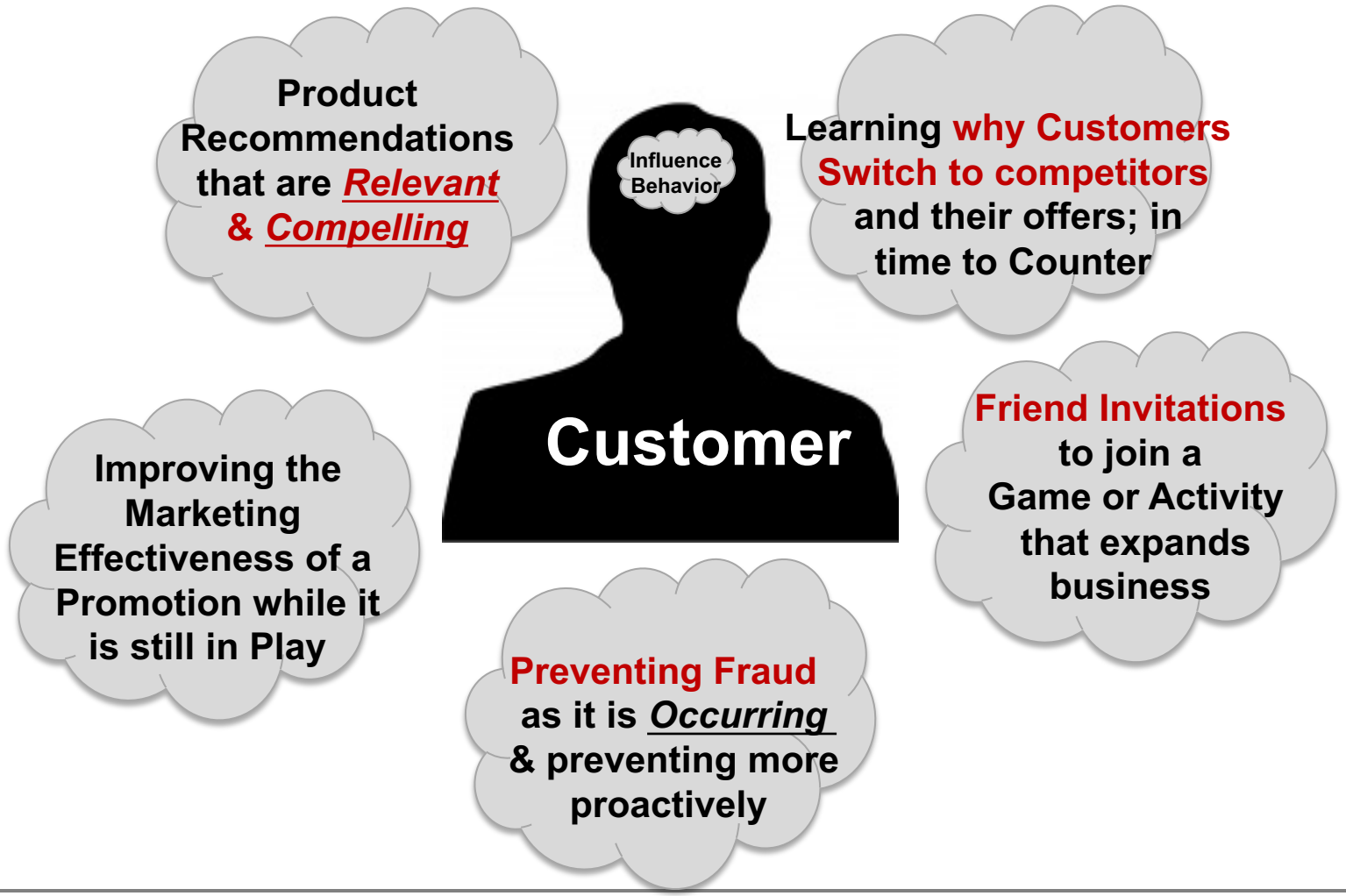
- Data is being generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities
- Examples
  - E-Promotions: Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - Healthcare monitoring: sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

Big Data Analytics

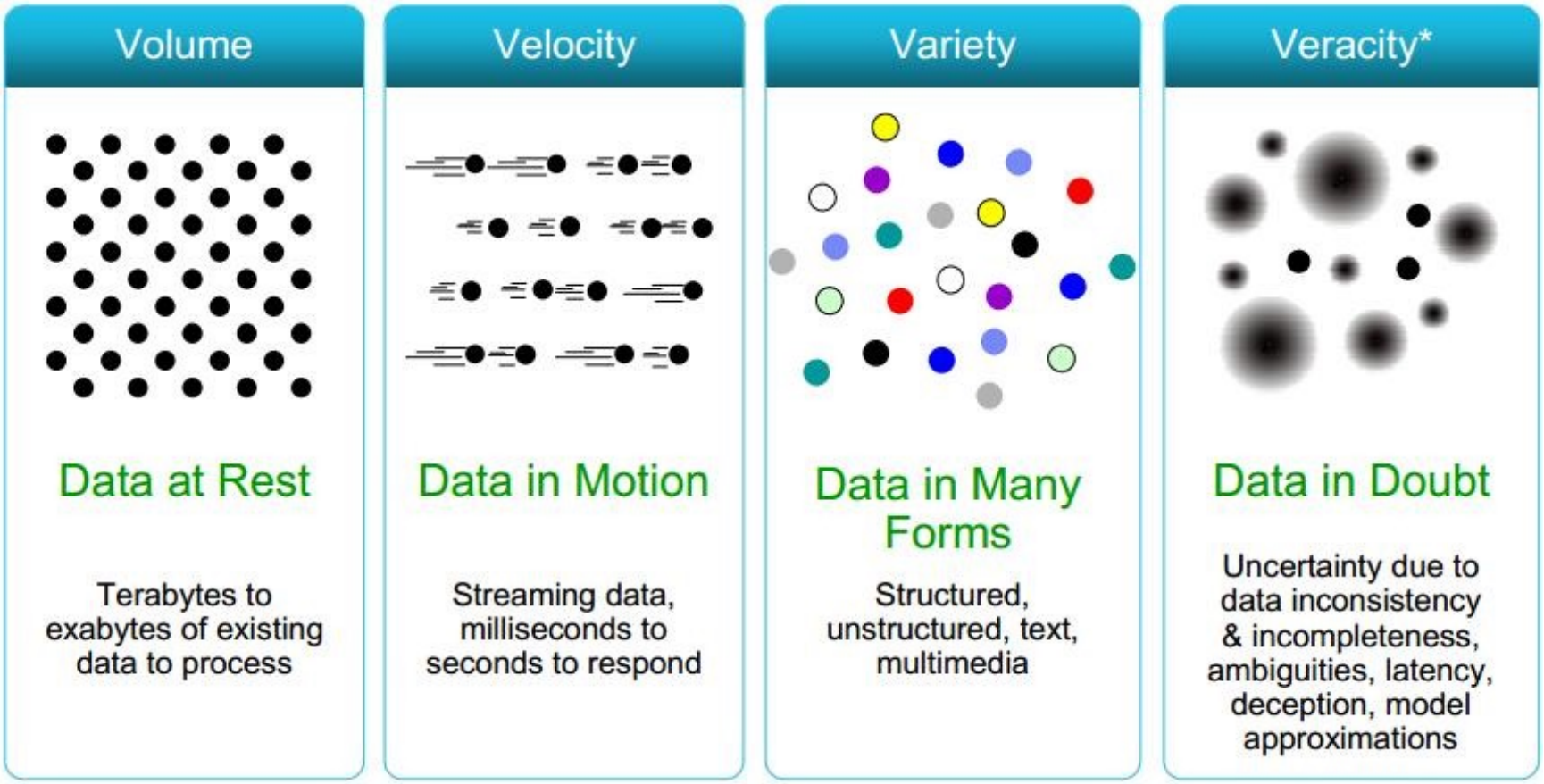
- Big data is more real-time in nature than traditional DW applications
- Traditional DW architectures (e.g. Exadata, Teradata) are not well-suited for big data apps
- Shared nothing, massively parallel processing, scale out architectures are well-suited for big data apps



Real-Time Analytics/Decision Requirement



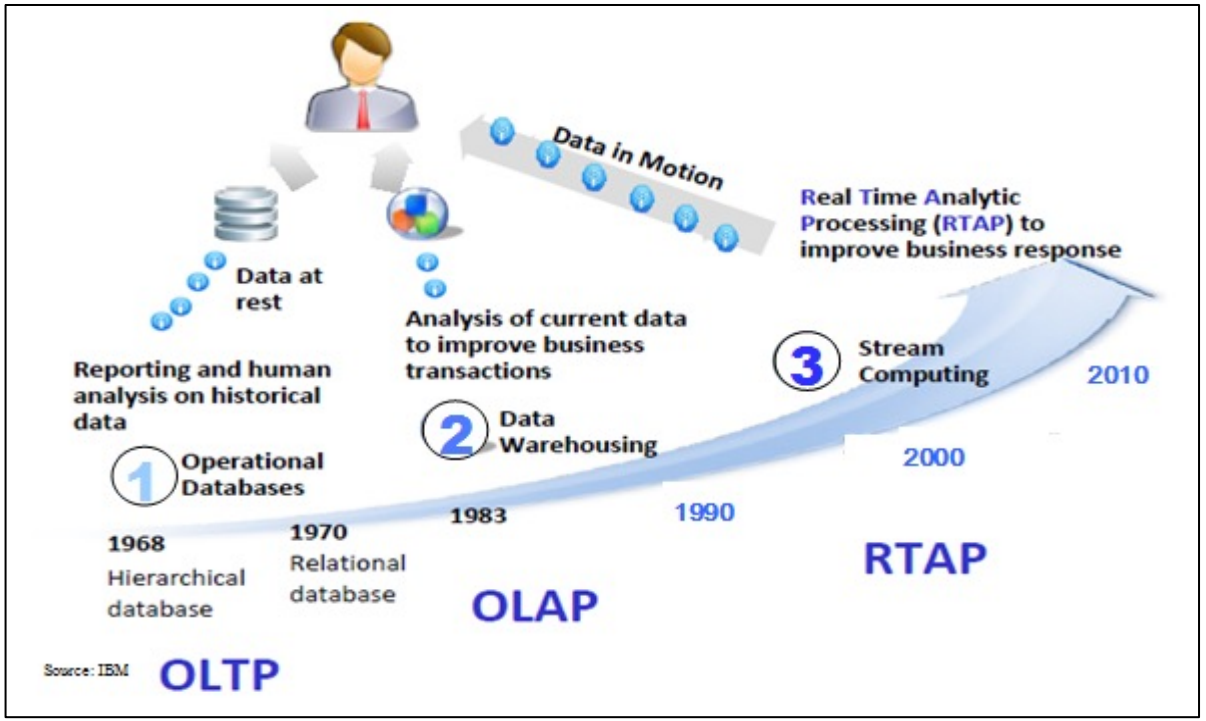
Some Make it 4V's



Then Out of Control …

EVOLUTION OF DATA PROCESSING

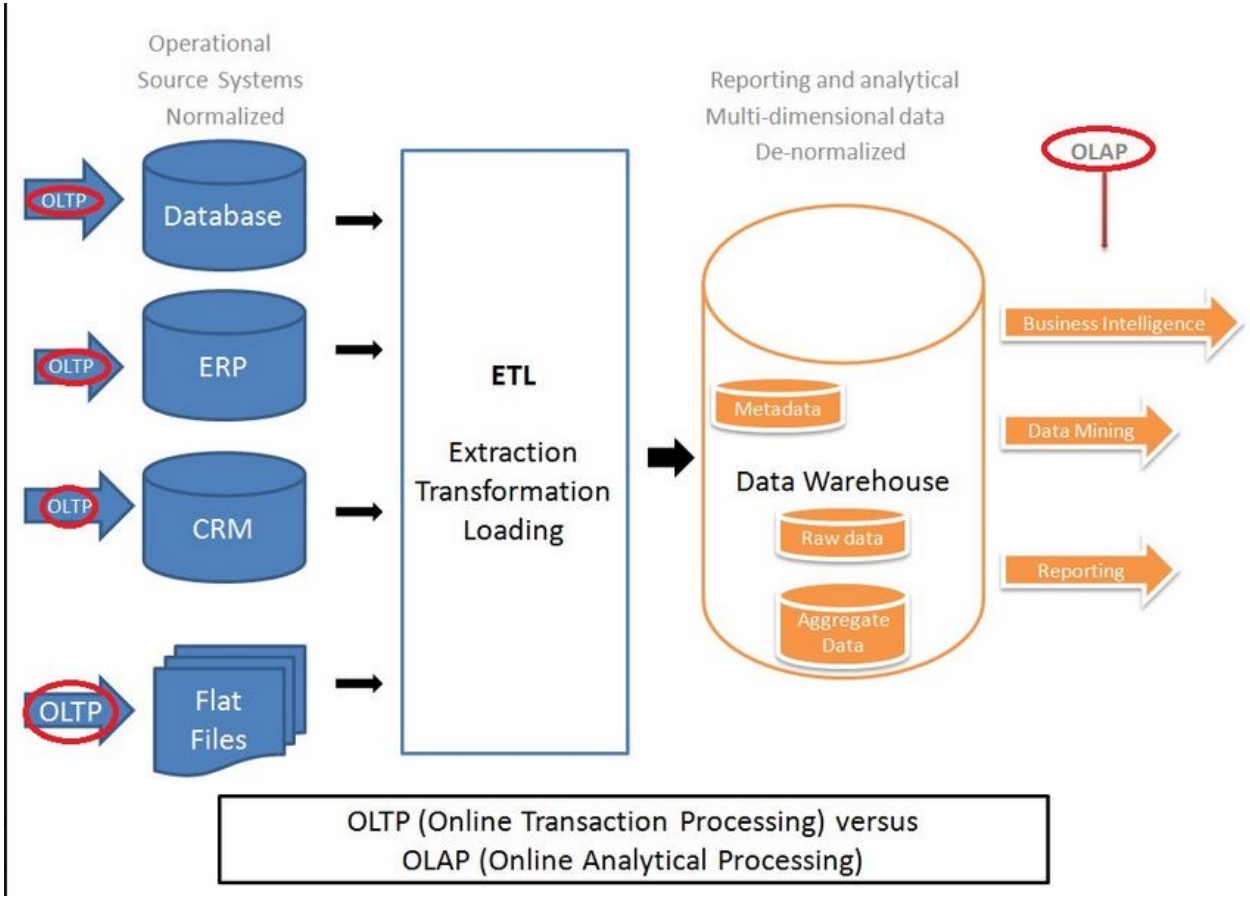
Evolution of Data Processing



- **OLTP:** Online Transaction Processing – DBMSs, business/operation logic
- **OLAP:** Online Analytical Processing -- Data Warehousing, finding value
- **RTAP:** Real-Time Analytics Processing -- Big Data Architecture & technology

OLTP vs. OLAP vs. ETL

- ETL is the enabler

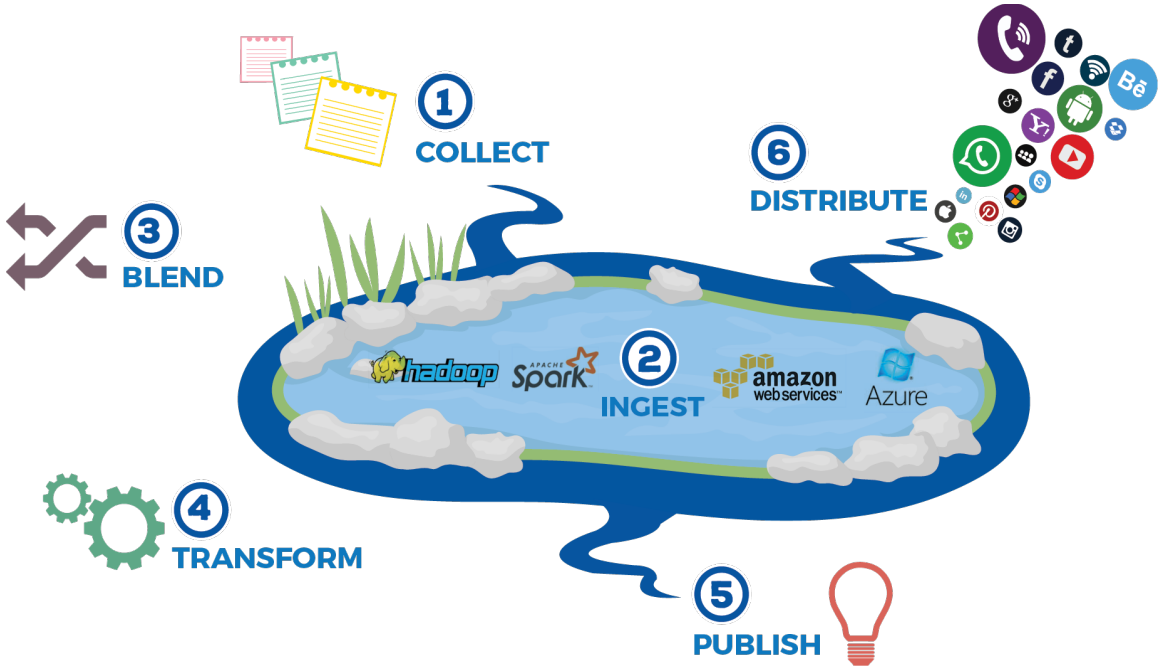


Traditional ETL

- ETL – Extract, Transform, Load
  - Extract – Find gold nuggets on the river bank
  - Transform – melt the gold and remove impurities
  - Load – make the gold a wedding ring

Data Lake "ELT"

- Unstructured data, raw data
  - Hard to interpret, can be useful or useless
  - Business does not want to lost a single data



Deploying a Big Data Solution

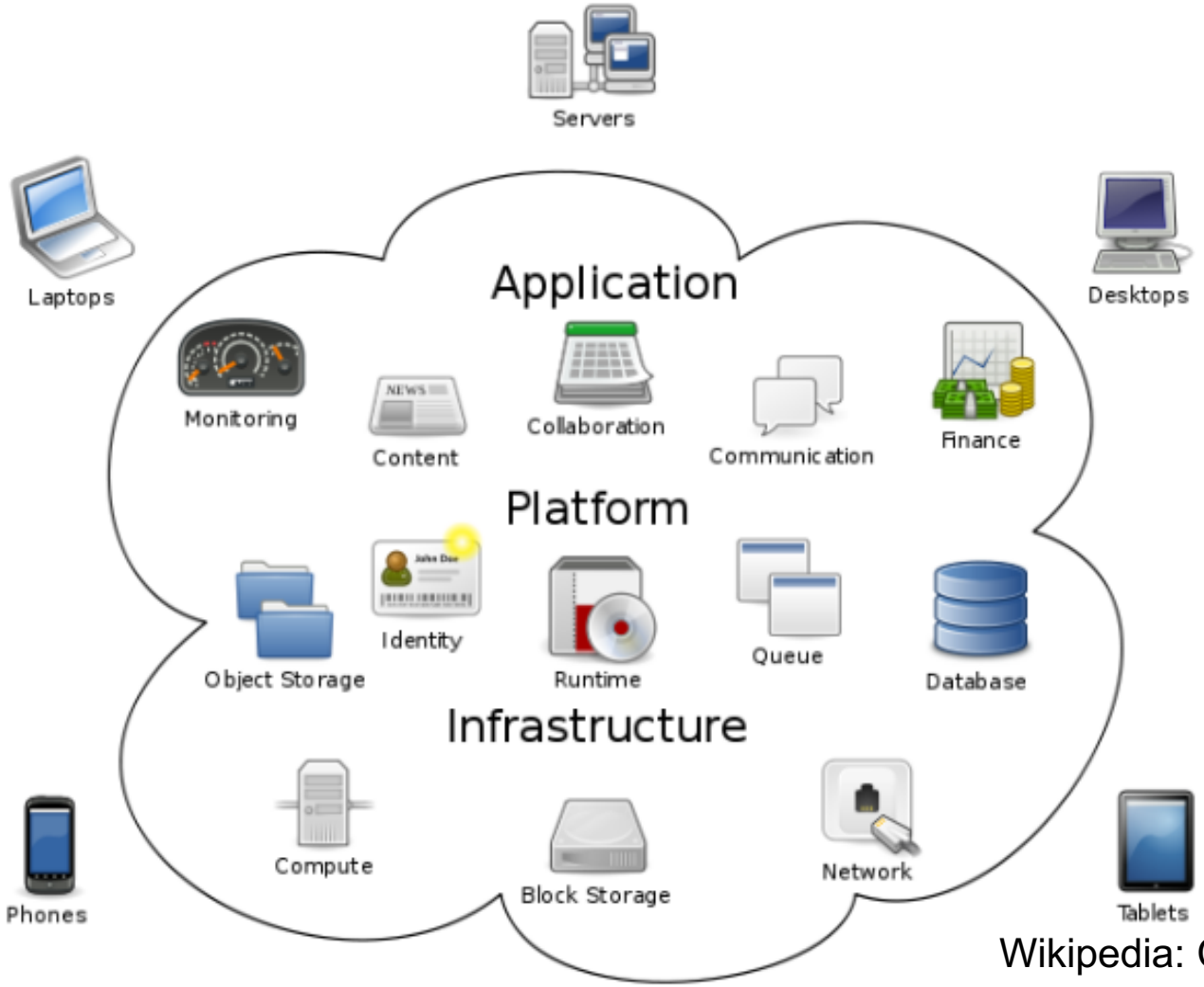
- Data Ingestion
  - Extraction of data from various sources
- Data Storage
  - HDFS (sequential access) or NoSQL databases (random R/W)
- Data Processing
  - MapReduce, Pig, Spark

BIG DATA IN PRODUCTION

Cloud Computing

- IT resources provided as a service
  - Compute, storage, databases, queues
- Clouds leverage economies of scale of commodity hardware
  - Cheap storage, high bandwidth networks & multicore processors
  - Geographically distributed data centers
- Offerings from Microsoft, Amazon, Google, …

Cloud Computing



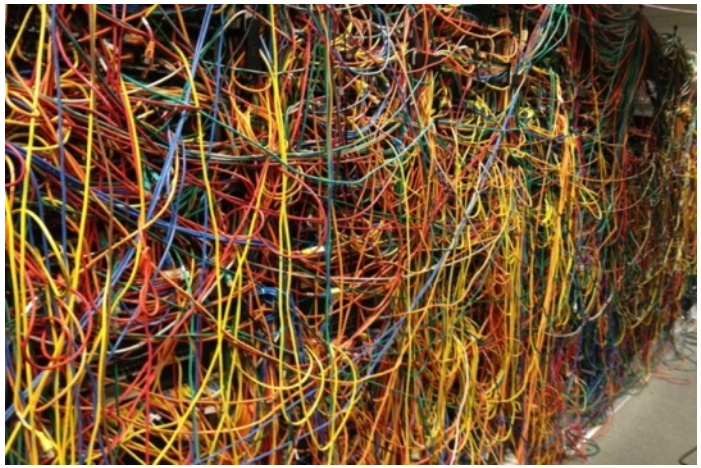
Wikipedia: Cloud Computing

Benefits

- Cost & management
  - Economies of scale, “out-sourced” resource management
- Reduced Time to deployment
  - Ease of assembly, works “out of the box”
- Scaling
  - On demand provisioning, co-locate data and compute
- Reliability
  - Massive, redundant, shared resources
- Sustainability
  - Hardware not owned

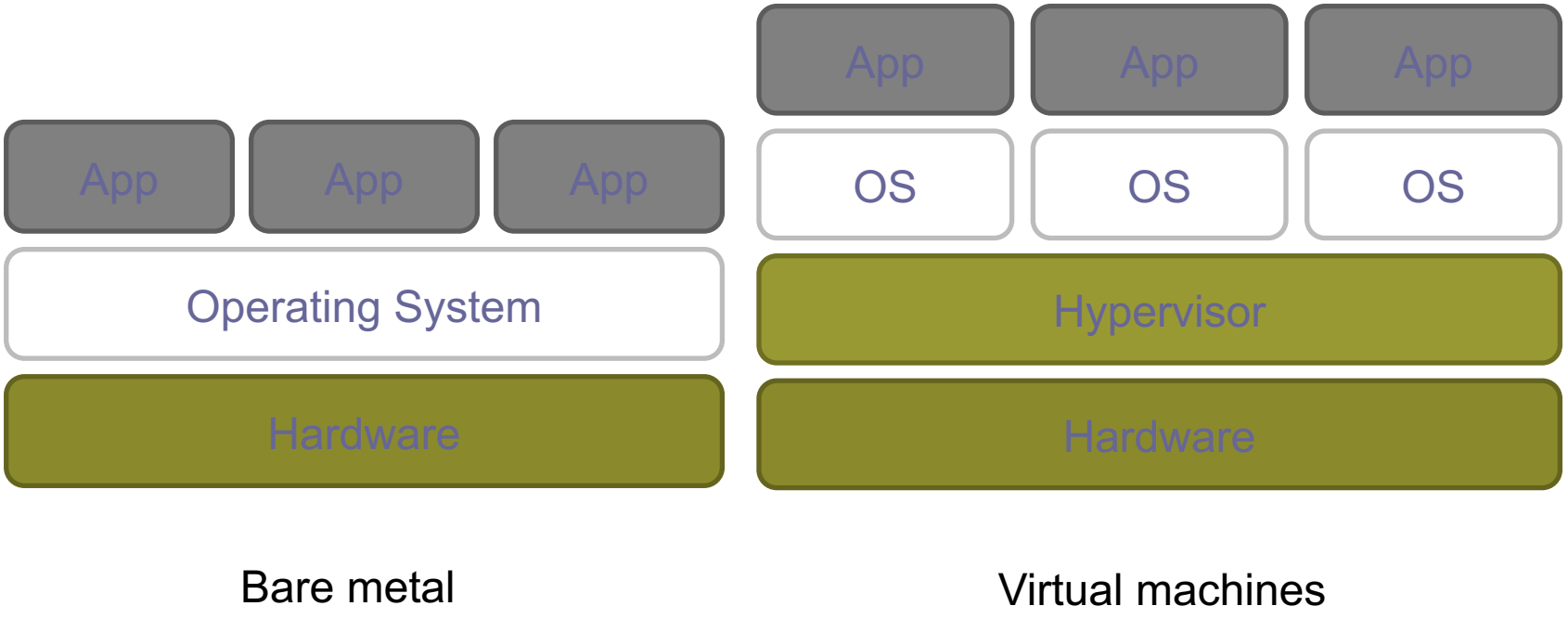
Types of Cloud Computing

- Public Cloud: Computing infrastructure is hosted at the vendor's premises.
  - E.g. Amazon EC2, Microsoft Azure, Google Cloud
- Private Cloud: Self-organized and maintained and is not shared with other organizations.



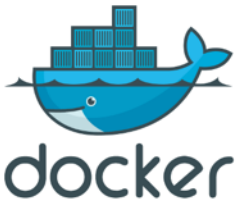
Enabling Technology: Virtualization

- Hardware abstraction level:
  - Example: virtual machines

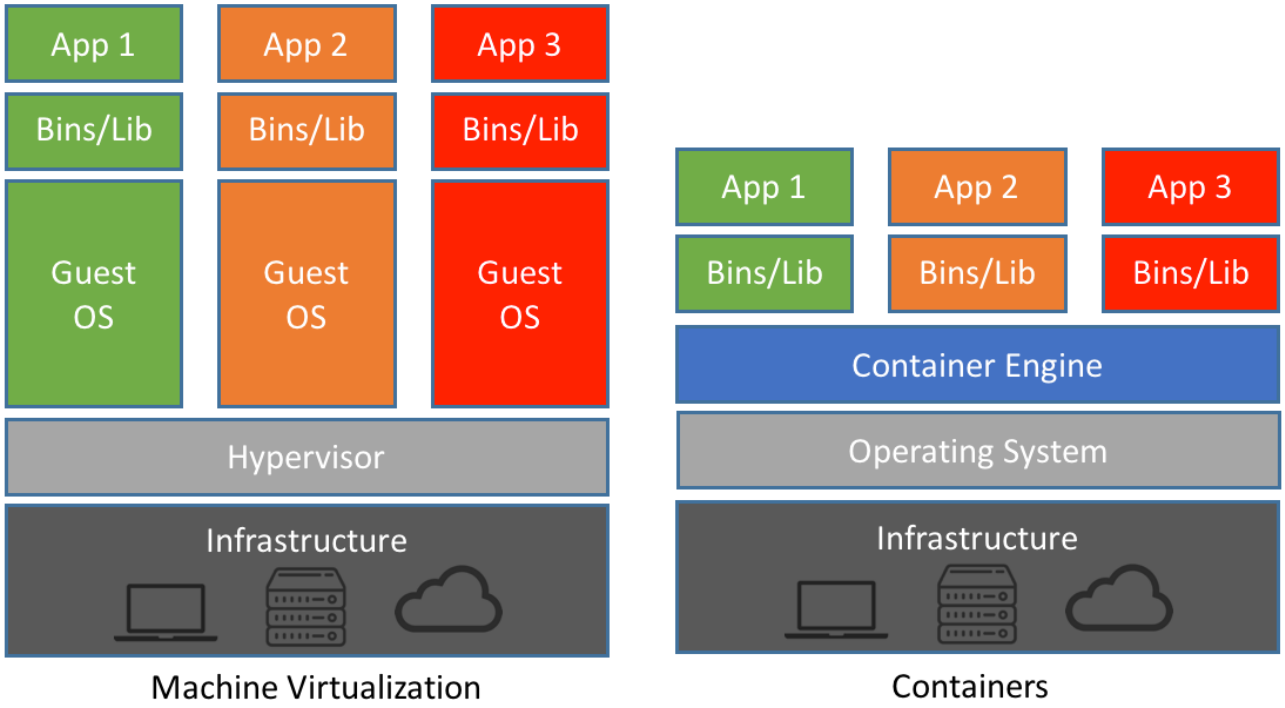


Enabling Technology: Virtualization

- Operation system level:
  - Example: containers



LXC



Benefits of Containerization

- Portability between platforms and clouds
- Resource efficiency over VM and bare metal
- DevOps agility: microservices
- Improved security thanks to isolation
- Faster app startup and easier scaling
- Easier management (install, upgrade …)

Summary

- What: Big data is NOT more data, it is better data.
- Where: Webpages, online products, social network …
- Why: The 3 V's
- When:
  - OLTP → OLAP → RTAP
  - ETL vs ELT
- How:
  - Cloud computing, virtualization, containerization.

Final notes

Next time:

Lec2: Hadoop Ecosystem

Reminders:

Finish Reading 1 for this week.

HW 1 out, due next Monday.