CS226
Big-Data Management

Instructor:

Ahmed Eldawy

Welcome (back) to UCR!
Happy New Year!

Class information

- Classes: M/W 2:00 – 3:20 PM
- Materials Science and Engineering 103
- Instructor: Ahmed Eldawy
- TAs: Bo Huang
- Office hours: T/Th 4:00 – 4:50 PM on Zoom
- Email: eldawy@ucr.edu
  - Subject: "[CS226]"
  - All caps, with brackets and no spaces
  - Email if you would like to meet in-person

Code of Conduct

- Instructor and TA provide:
  - Clear and detailed instruction
  - Help and support
  - Appropriate feedback
- Students should:
  - Follow the instructions responsibly
  - Ask for help when needed
  - Complete the work on their own
- Both should be:
  - Respectful, understanding, and honest
- Refer to the course syllabus for more details

Course work

Project (50%)

Course activities (15%)

Assignments (20%)

Exams (15%)

Project

- Groups of 4-5 students
- Milestones
  - Group Selection
  - Project proposal (6%)
  - Project proposal presentation (8%)
  - Literature survey (7%)
  - Report outline (5%)
  - Final report and deliverables (10%)
  - Final presentation (10%)
  - Presentation questions (4%)
- This year, we focus on Digital Twins
- PhD Students?

Use of Generative AI

- Use of generative AI, ChatGPT and the likes, is allowed
- You are required to disclose which parts were generated
- ChatGPT is not always right
- You are ultimately responsible for everything you deliver
- Exams will be on the test center

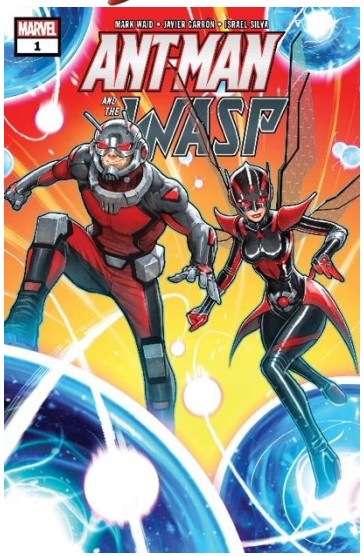
Learning Outcomes

- Understand the characteristics and aspects of big data.
- Learn how big data is stored in the distributed file system.
- Get familiar with big-data programming models and query execution engines.
- Build real applications using big-data systems, e.g., Spark and AsterixDB.
- Know the architectural differences between big-data systems.

Prerequisites

- Database Management Systems
- Relational data model
- Relational algebra
- SQL
- Indexing
- Query optimization
- Data normalization
- Transactions
- (Required task) Take the prerequisite test on Canvas

Super Hero


MARVEL
1
MARK WAID • JAVIER GARRÓN • ISRAEL SILVA
ANT-MAN
AND THE
WASP

Ant-Man/Wasp

Get smaller to understand how ants work and what they are capable of.

Use this knowledge to control thousands of ants and do amazing things!

Optional task: Watch Ant-Man and the Wasp this weekend ☺

Big-data Expert

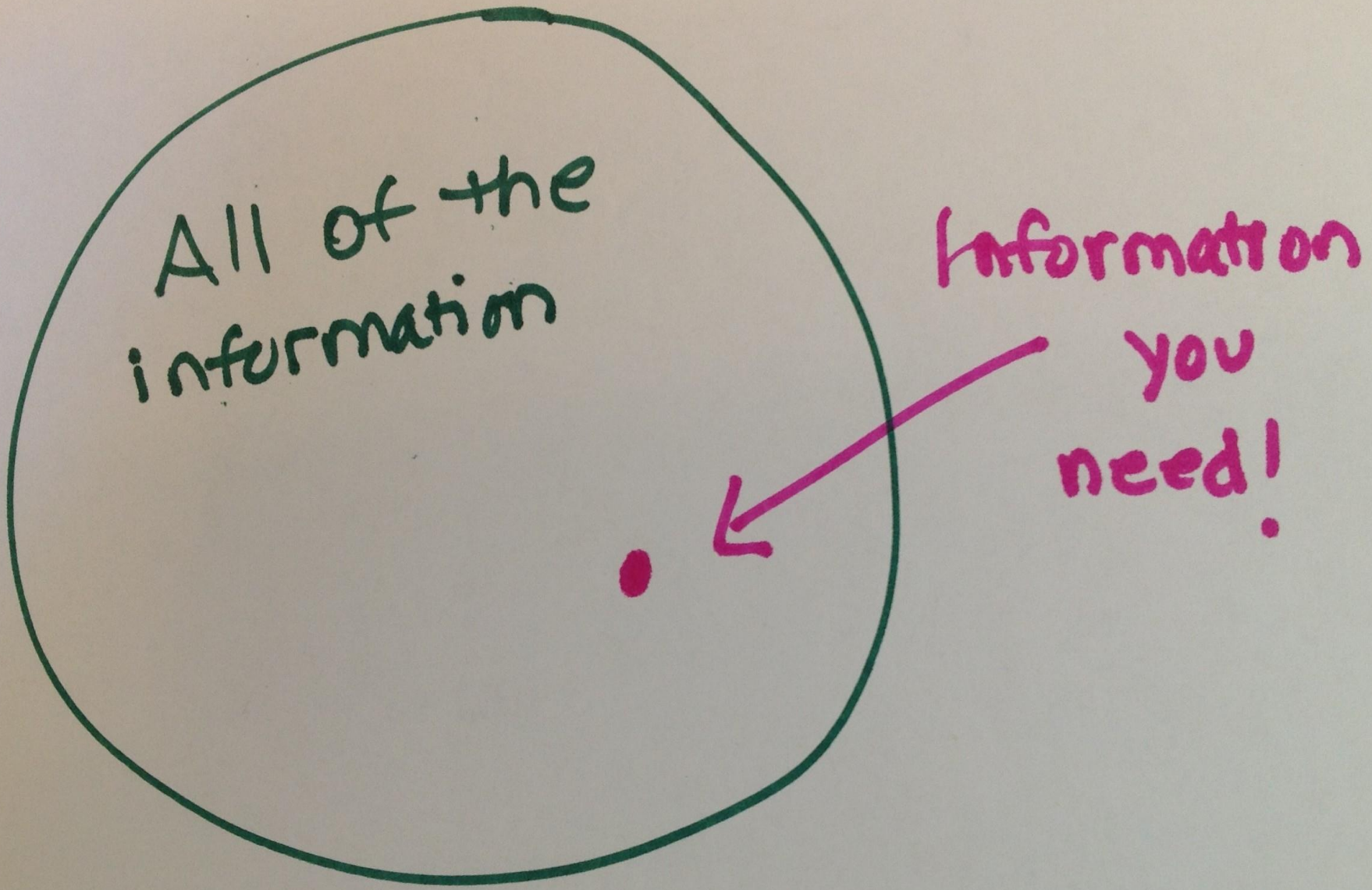
- Understand how the big-data platforms really work
- Control those thousands of processors efficiently to carry out your task

Syllabus

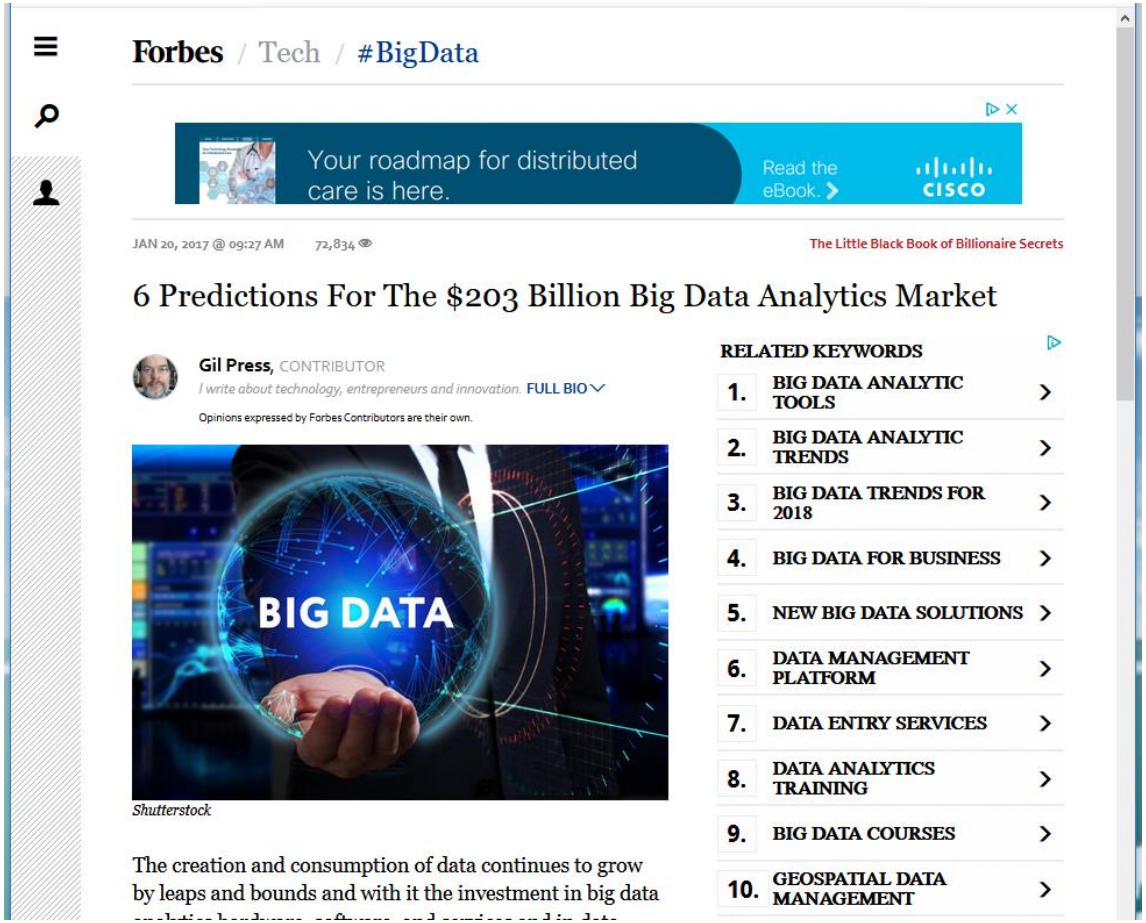
- Overview of big data
- Big-data storage
- Unstructured big-data processing
- (Semi-)Structured big-data processing
- Big-data layout and formats
- High-level applications on big-data

Introduction


Big Data
Straight Ahead


All of the information
Information you need!

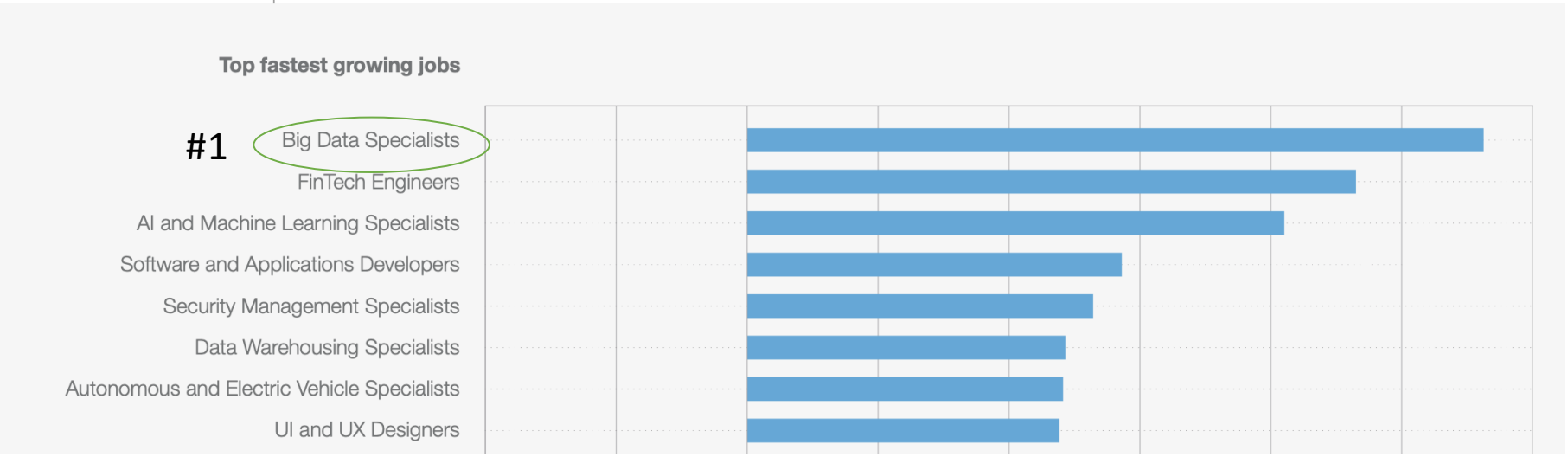
The Market of Big Data



Fastest Growing Jobs

FIGURE 2.2 **Fastest-growing and fastest-declining jobs, 2025-2030**

Top jobs by fastest net growth and net decline, projected by surveyed employers

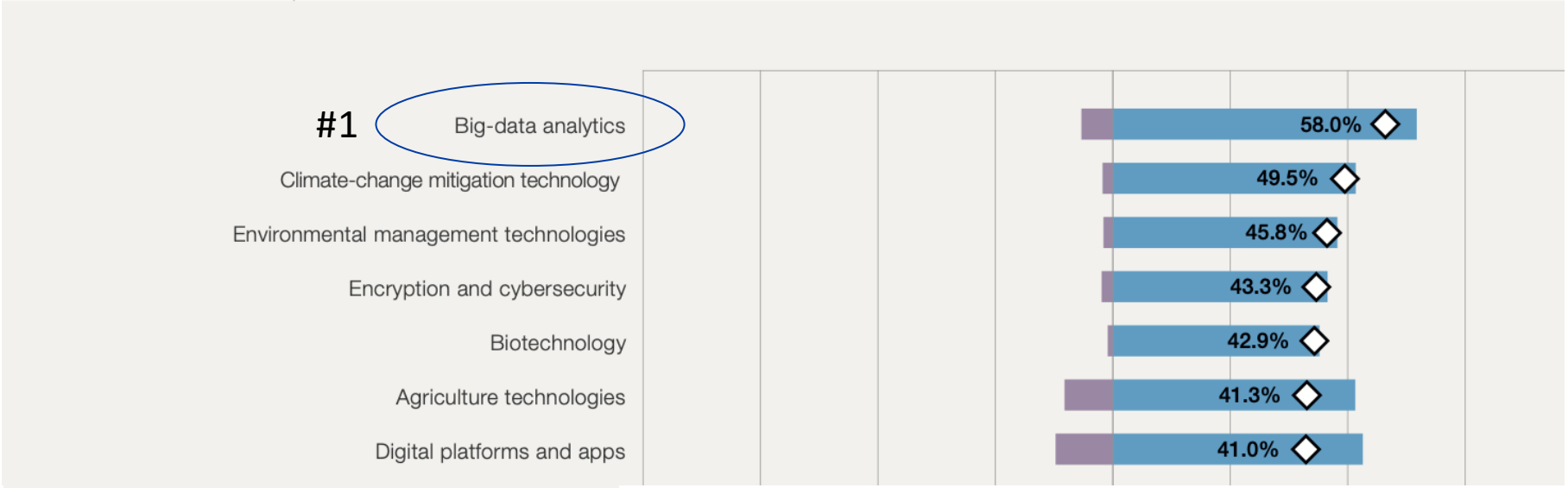


https://reports.weforum.org/docs/WEF_Future_of_Jobs_Report_2025.pdf

Technology Impact on Jobs

FIGURE 2.5 **Expected impact of technology adoption on jobs, 2023–2027**

Share of organizations surveyed that expect each technology to create or displace jobs, ordered by the job creation net effect. The shares of organizations which expect the impact of adopting these technologies to be neutral are not plotted.



Source

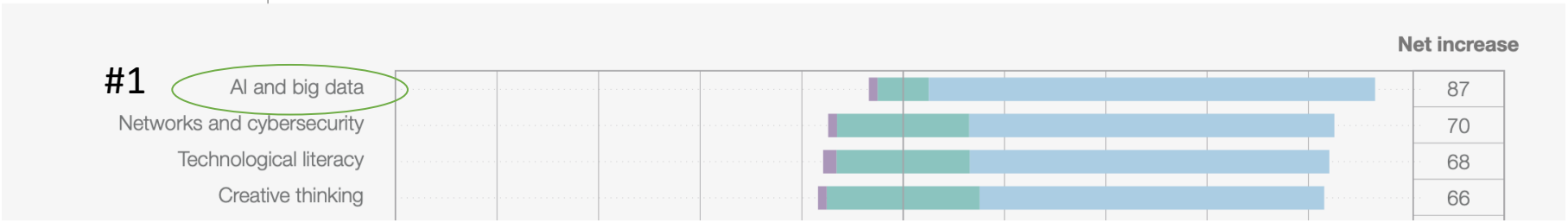
World Economic Forum, Future of Jobs Survey 2023.

https://www3.weforum.org/docs/WEF_Future_of_Jobs_2023.pdf

Skills on the Rise

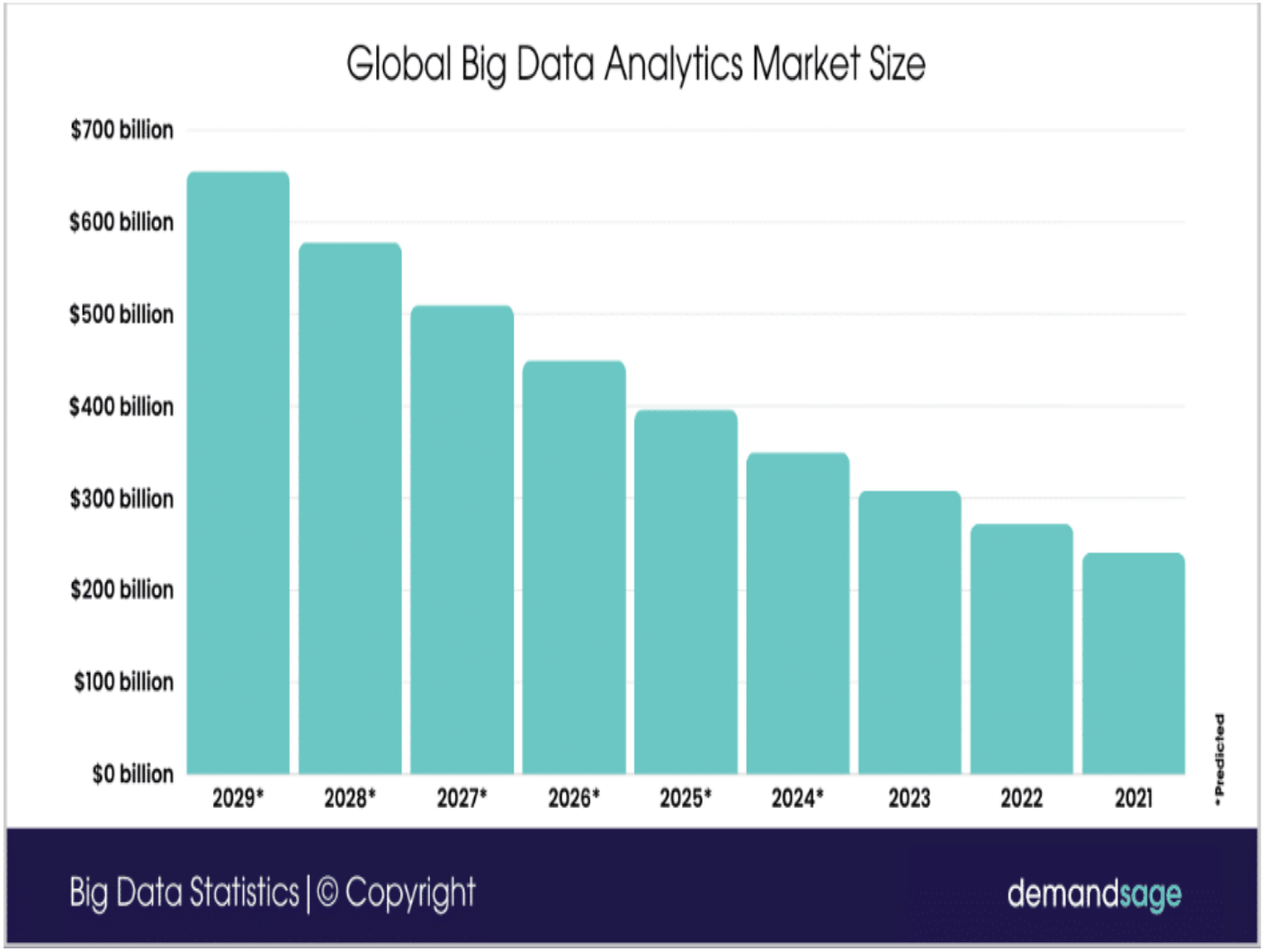
FIGURE 3.4 **Skills on the rise, 2025-2030**

Share of employers that consider skills to be increasing, decreasing, or remaining stable in importance. Skills are ranked based on net increase, which is the difference between the share of employers that consider a skill category to be increasing in use and those that consider it to be decreasing in use.



https://reports.weforum.org/docs/WEF_Future_of_Jobs_Report_2025.pdf

Big-data Market Size



https://www.demandsage.com/big-data-statistics/

OPEN Government Data Act

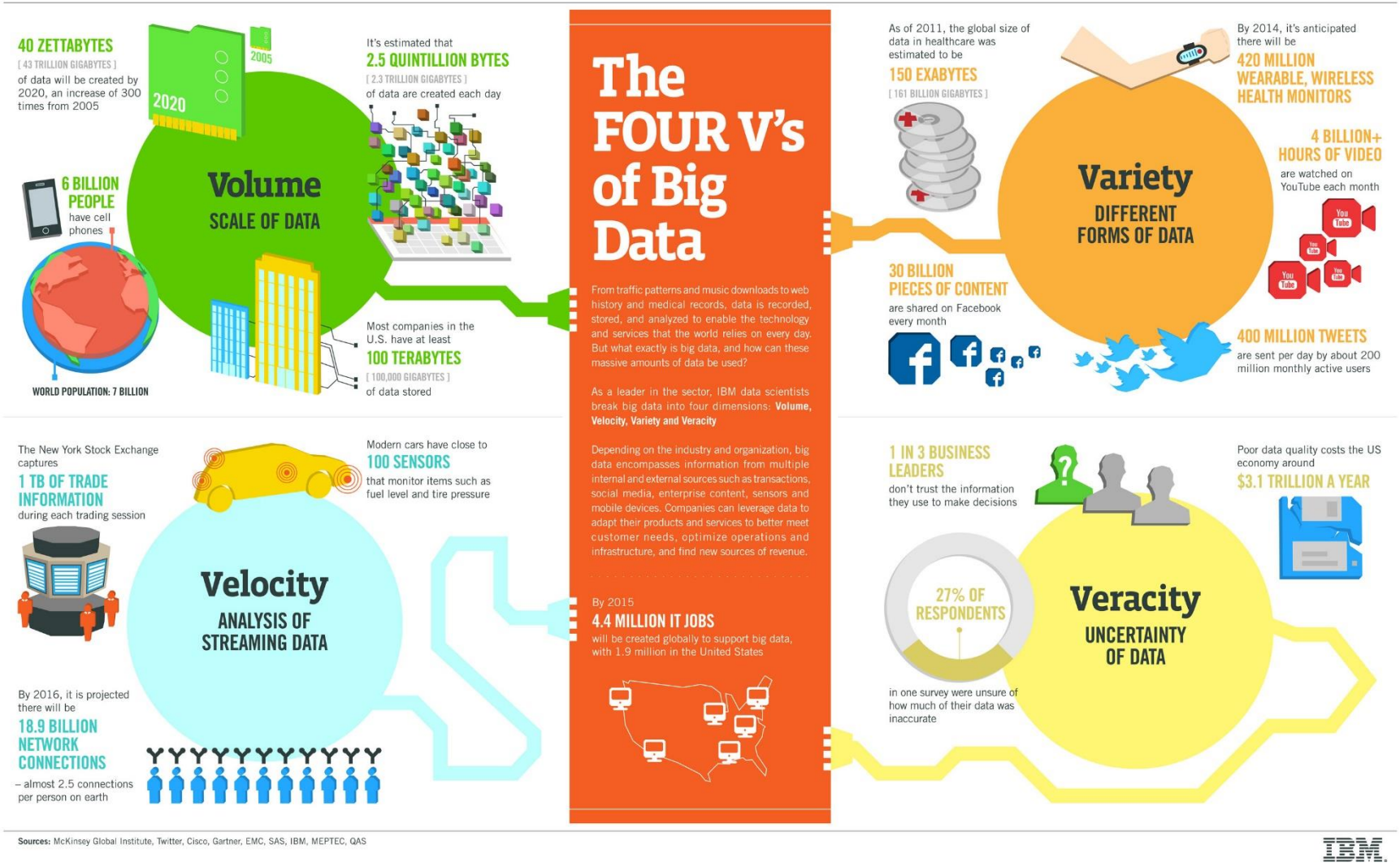
White House finalizes OPEN Government Data Act guidance, restarts CDO Council

The long-awaited guidance for the data and evidence statute comes six years after it became law.

BY MADISON ALDER • JANUARY 15, 2025

https://fedscoop.com/white-house-open-government-data-act-restarts-cdo-council/

V's of Big Data



Big Data Vs Big Computation

- Full scans (e.g., log processing)
- Range scans
- Point lookups
- Iterations
- Joins (self, binary, or multiway)
- Proximity queries
- Closures and graph traversals

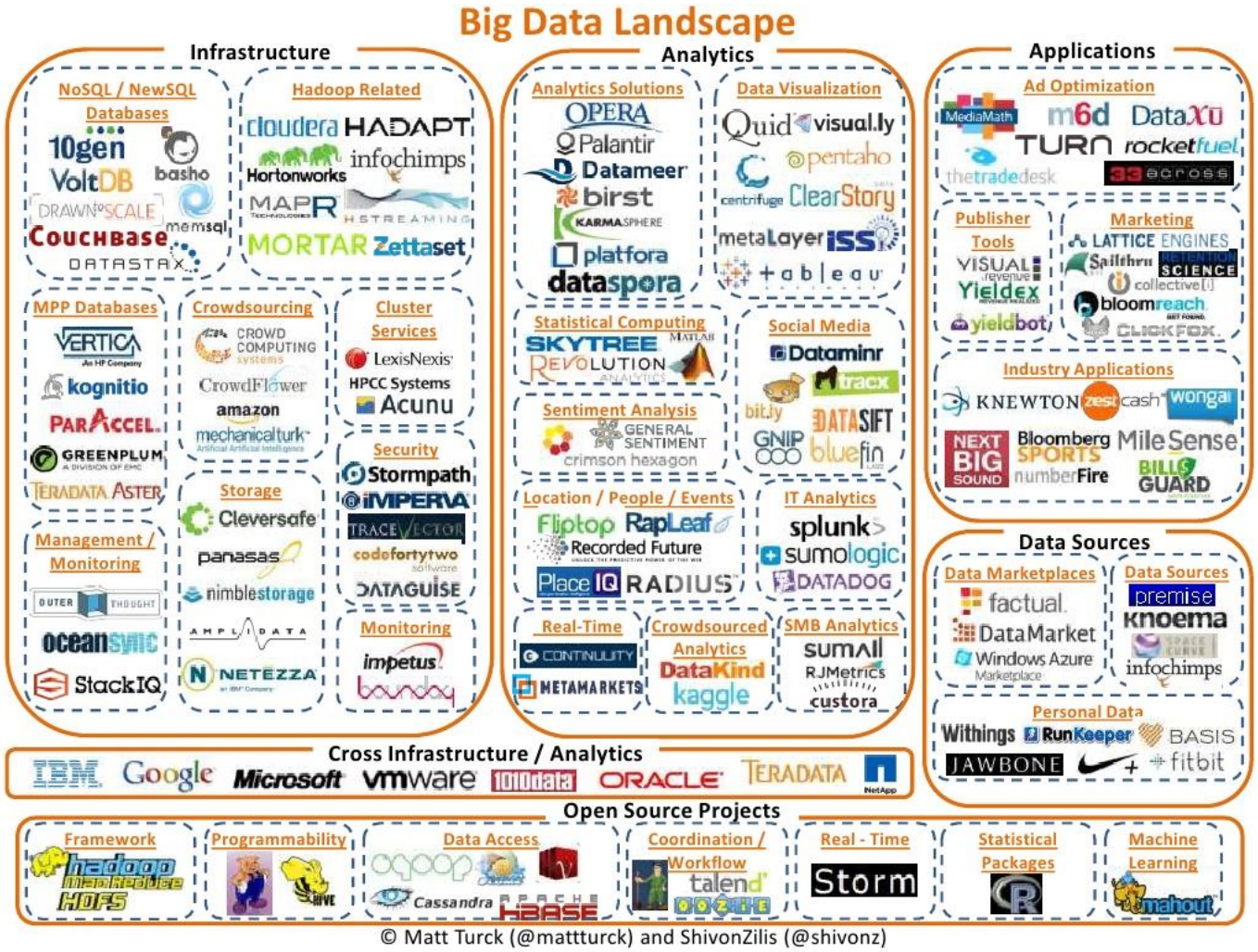
Big Data Applications

- Web search
- Marketing and advertising
- Data cleaning
- Knowledge base
- Information retrieval
- Internet of Things (IoT)
- Visualization
- Behavioral studies

Publicly Available Datasets

- Data.gov
- UCR Star [https://star.cs.ucr.edu]
- Flickr API
- Kaggle.com
- Industry: Yelp, Yahoo! Webscope, Zillow, …

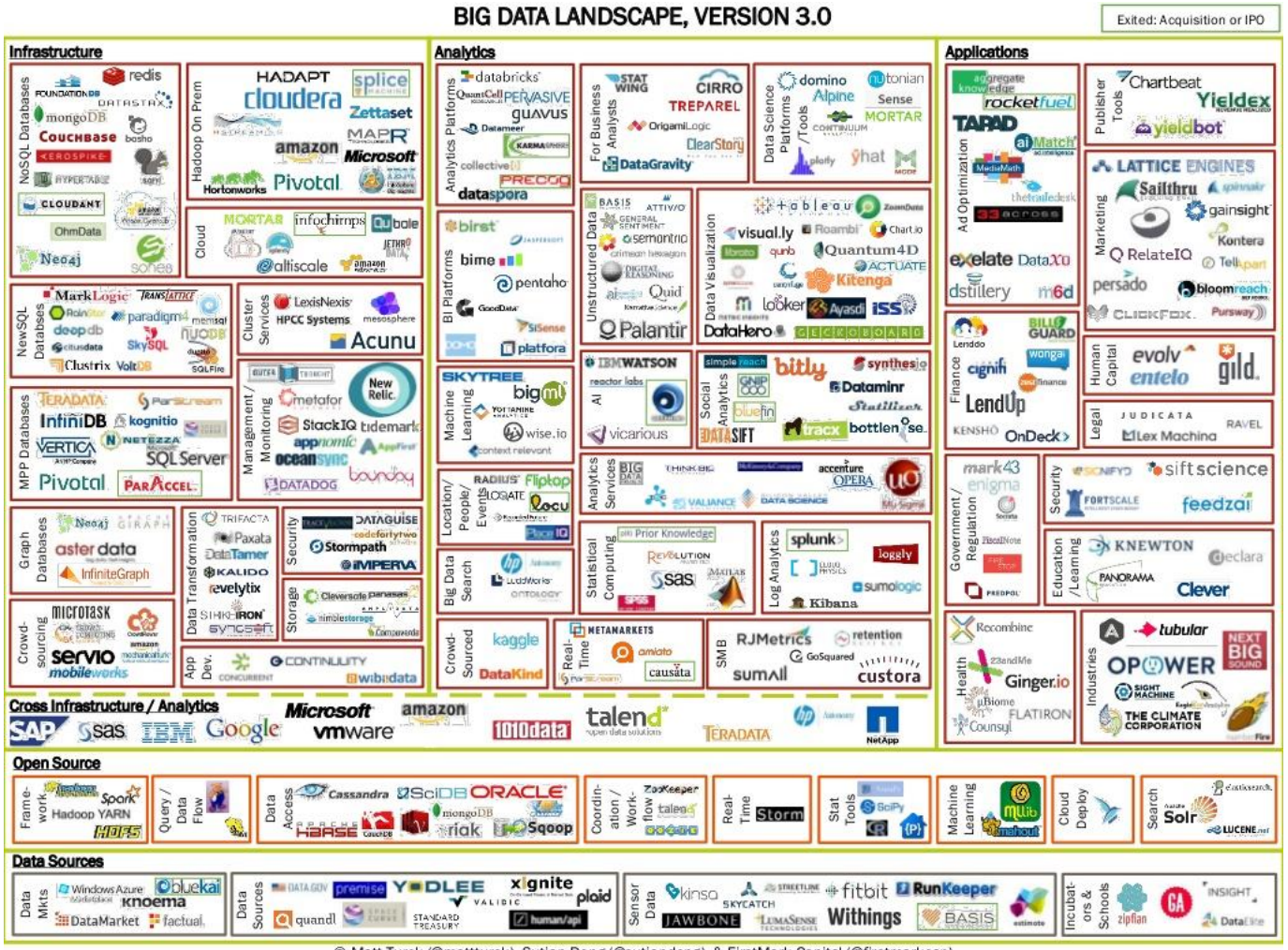
Big Data Landscape 2012



© Matt Turck (@mattturck) and ShivonZilis (@shivonz)

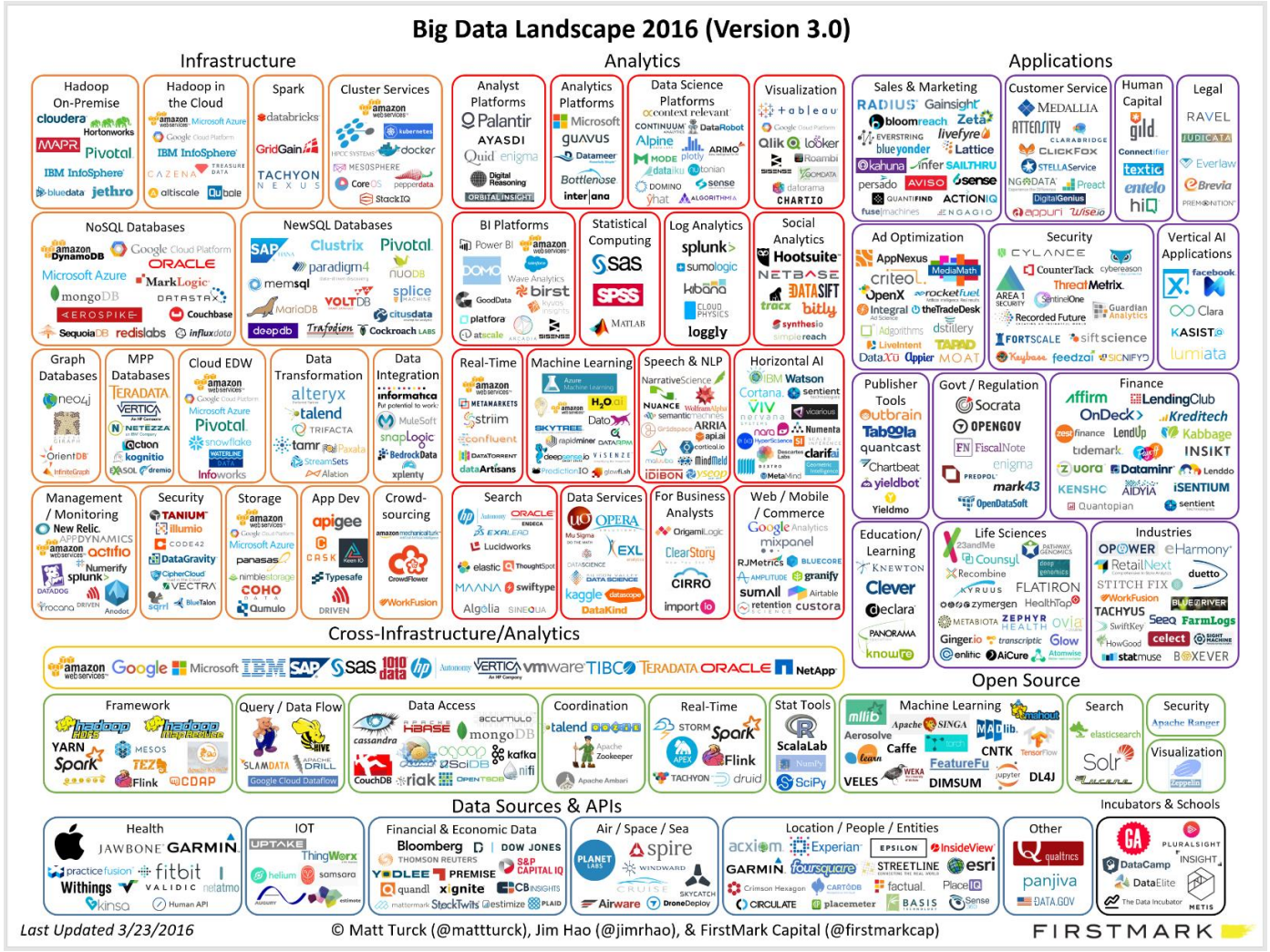
http://mattturck.com/2012/06/29/a-chart-of-the-big-data-ecosystem/

Big Data Landscape 2014



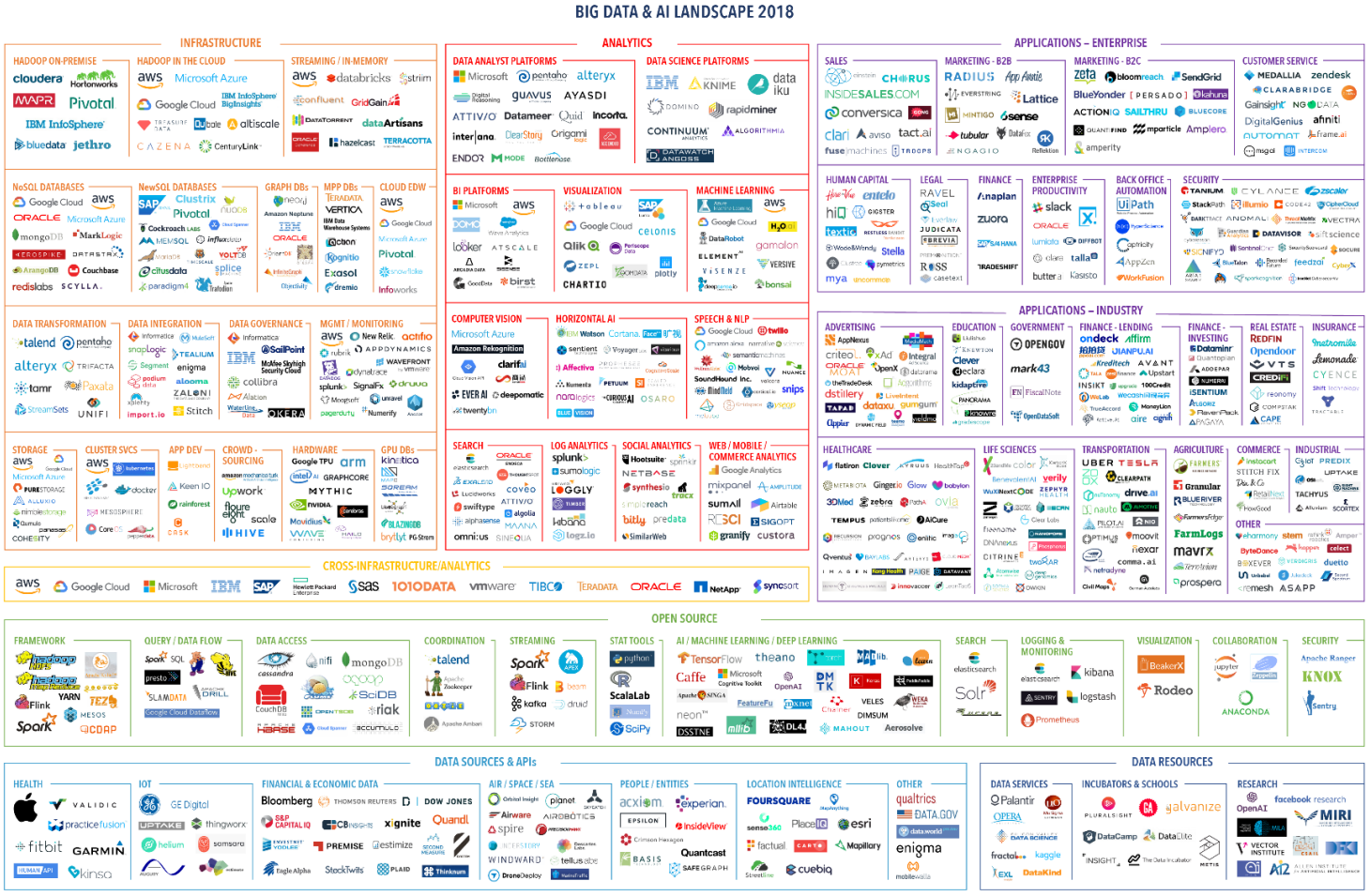
http://mattturck.com/2014/05/11/the-state-of-big-data-in-2014-a-chart/

Big Data Landscape 2016



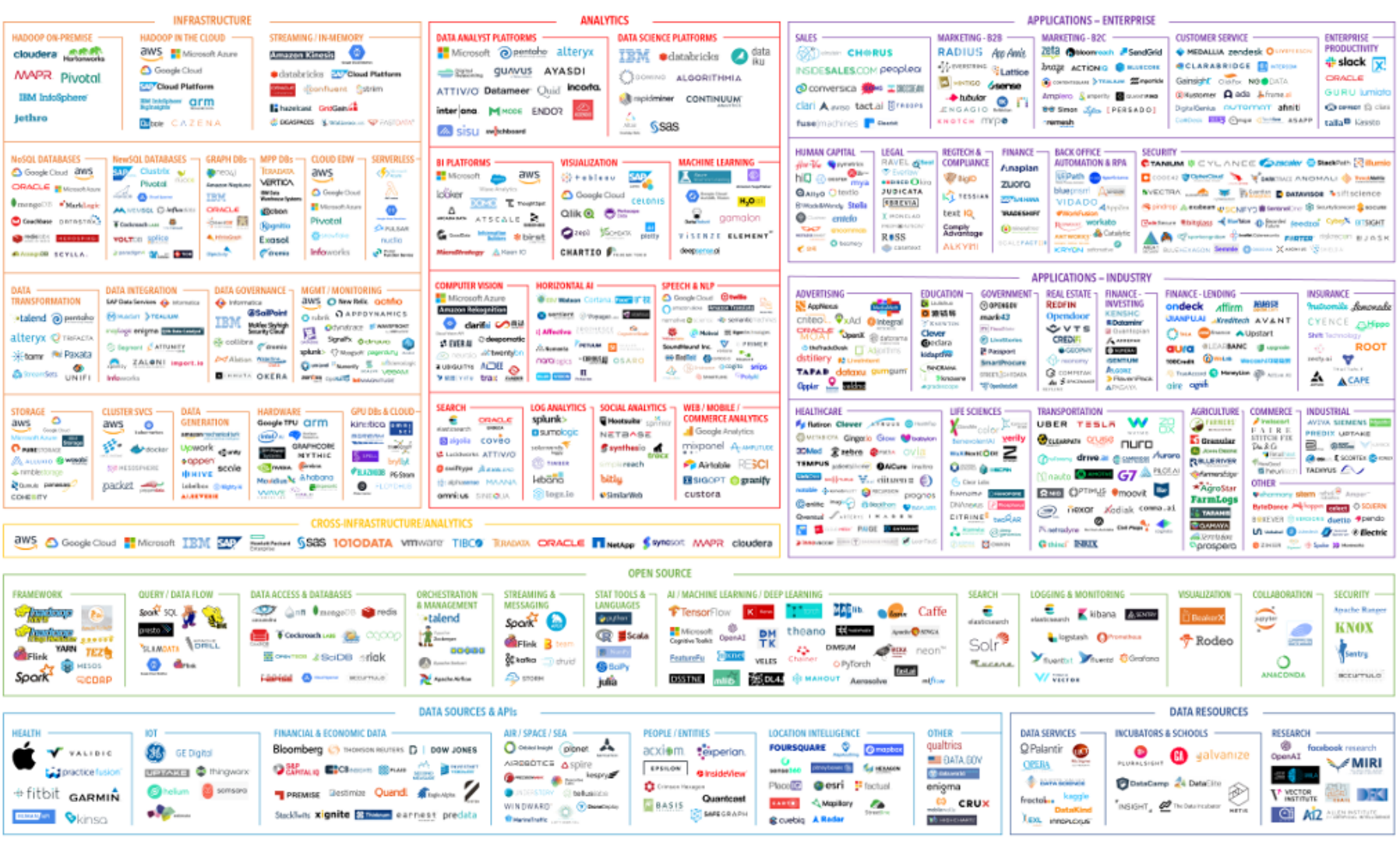
http://mattturck.com/2016/02/01/big-data-landscape/

Big Data Landscape 2018

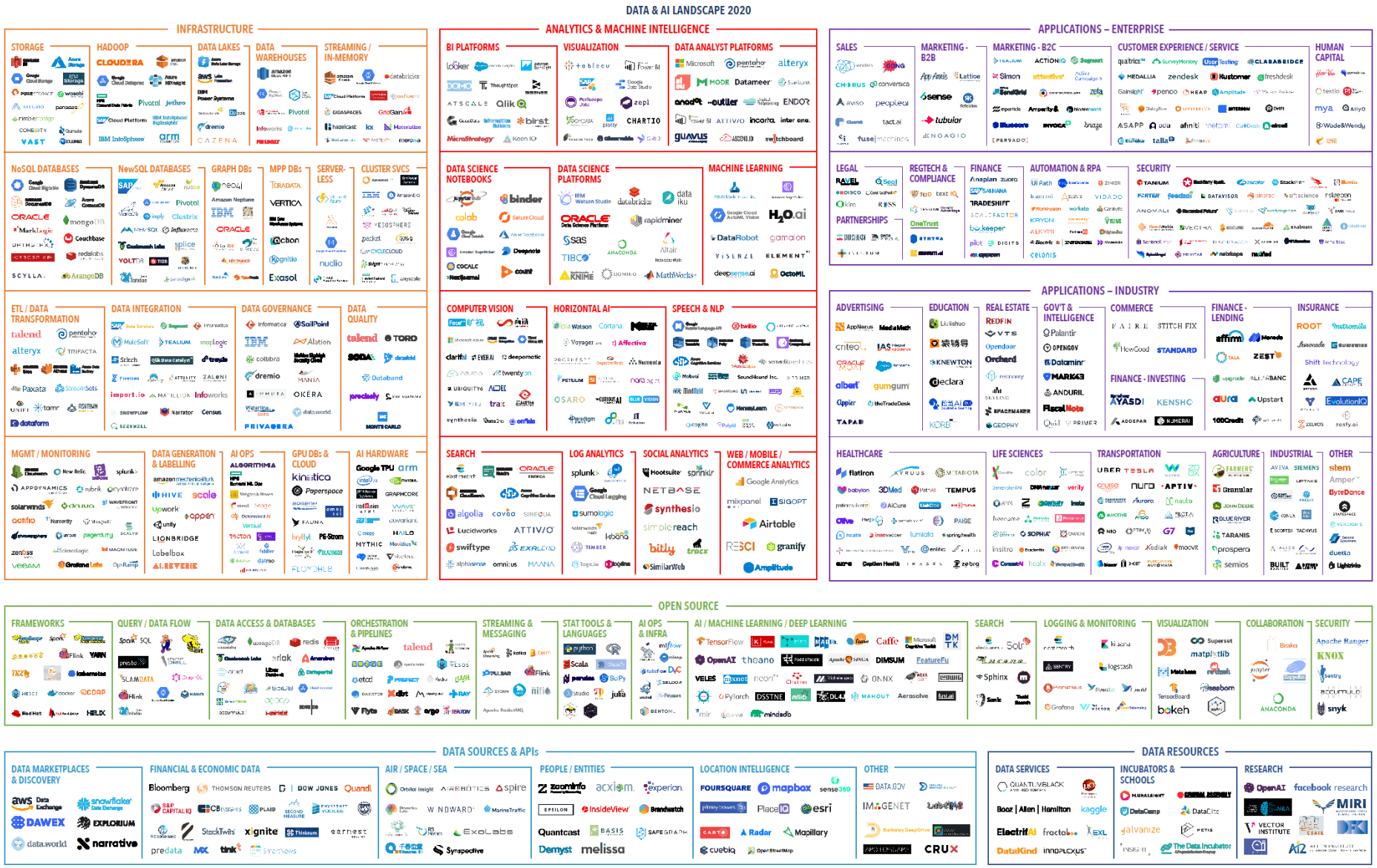


FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Data & AI Landscape 2019



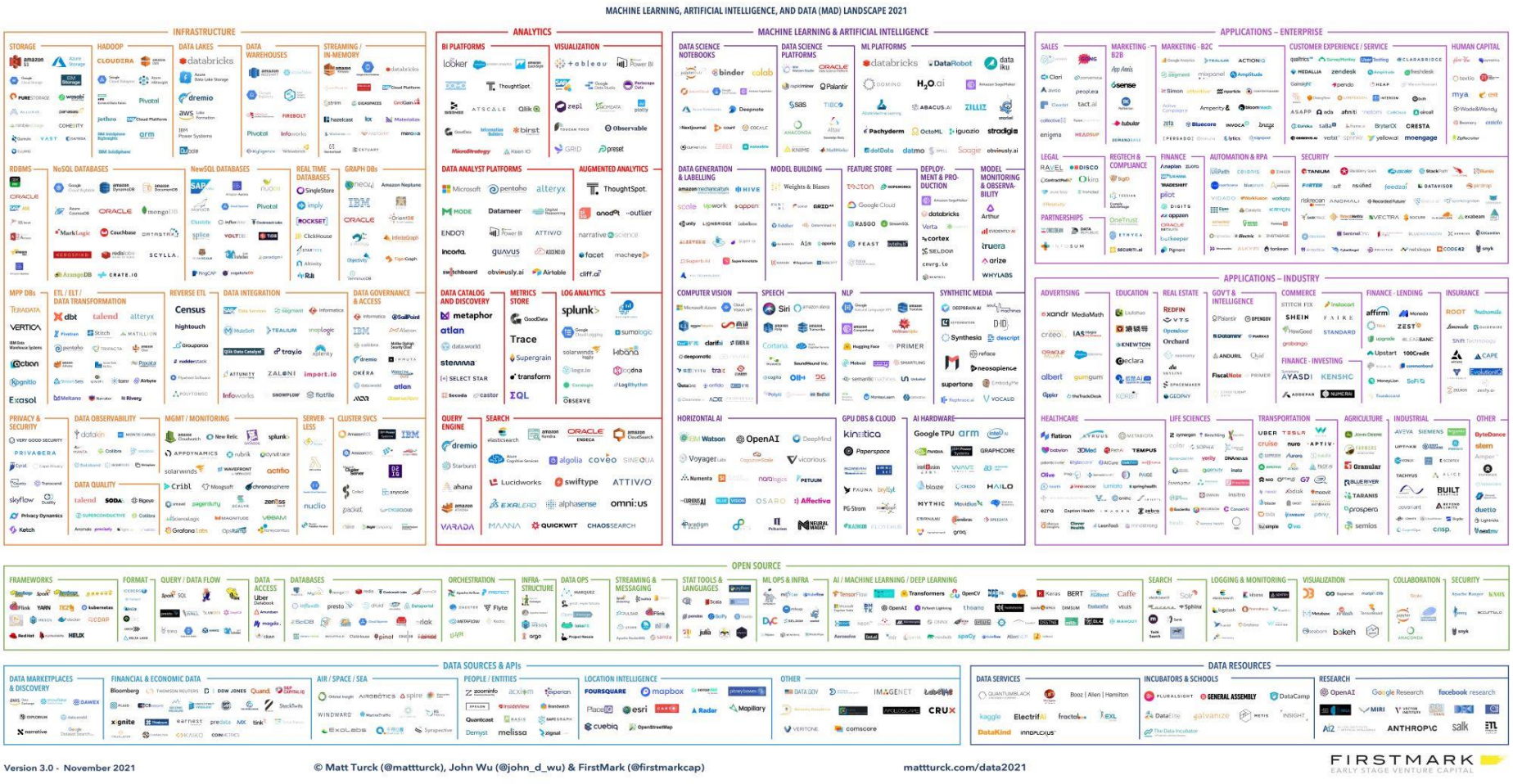
Data & AI Landscape 2020



mattturck.com/data2020

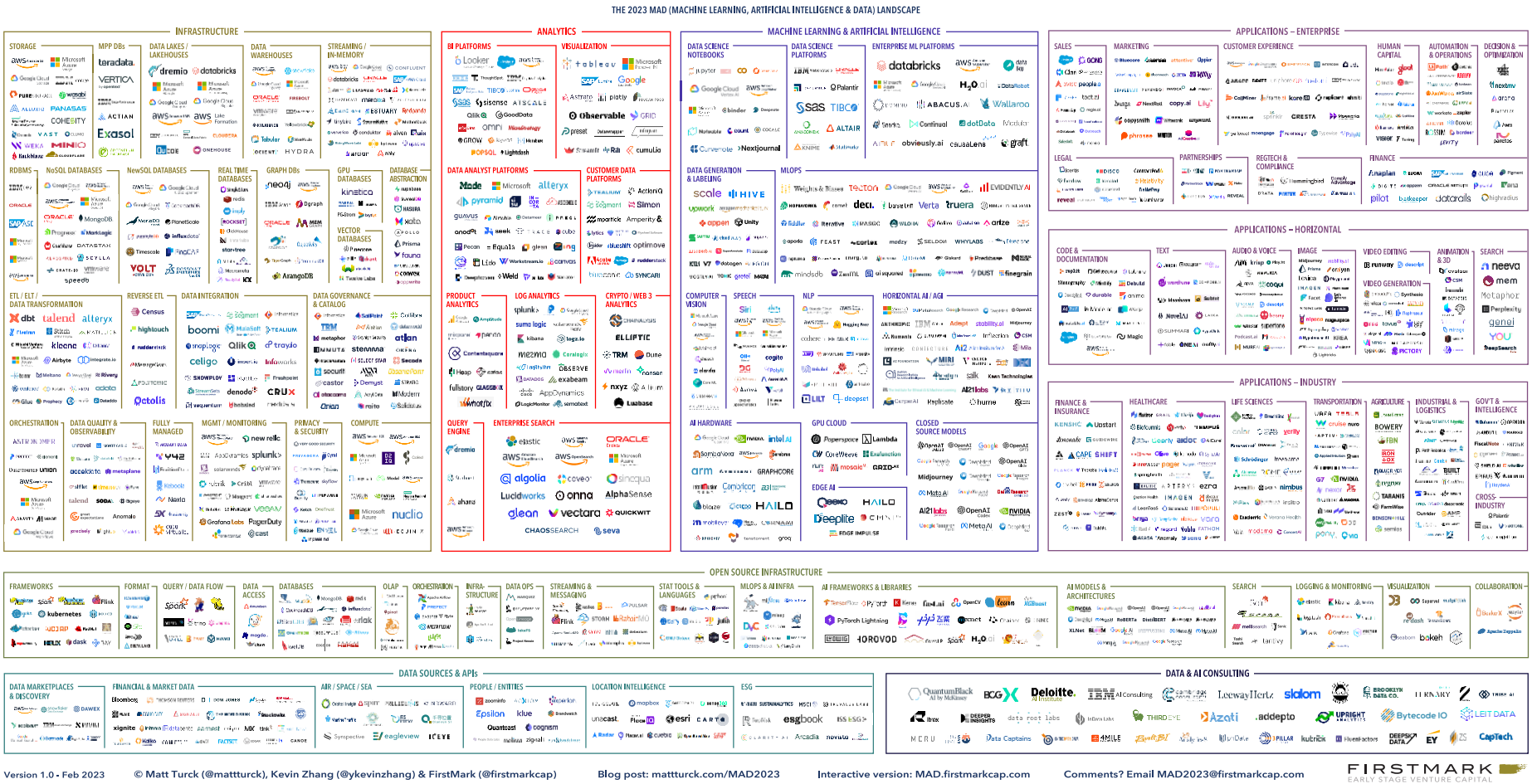
FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Machine Learning, AI, and Data (MAD) Landscape 2021



https://mattturck.com/data2021/

Machine Learning, AI, and Data (MAD) Landscape 2023



Components of Big Data

Components of Big Data

Big-data Libraries
MLlib (Machine Learning), GraphX, Visualization

Structured/ Semi-structured Data Processing
SparkSQL, Pig, SQL++, HiveQL

Distributed Computing
MapReduce (Hadoop and Google), Resilient Distributed Dataset (Spark), Hyracks (AsterixDB)

Big Data Distributed Storage
Hadoop Distributed File System, Cloud storage systems (Amazon S3 and Google File System), Key-value stores

Cloud Services
Amazon Web Services, Microsoft Azure, and Google Cloud Platform

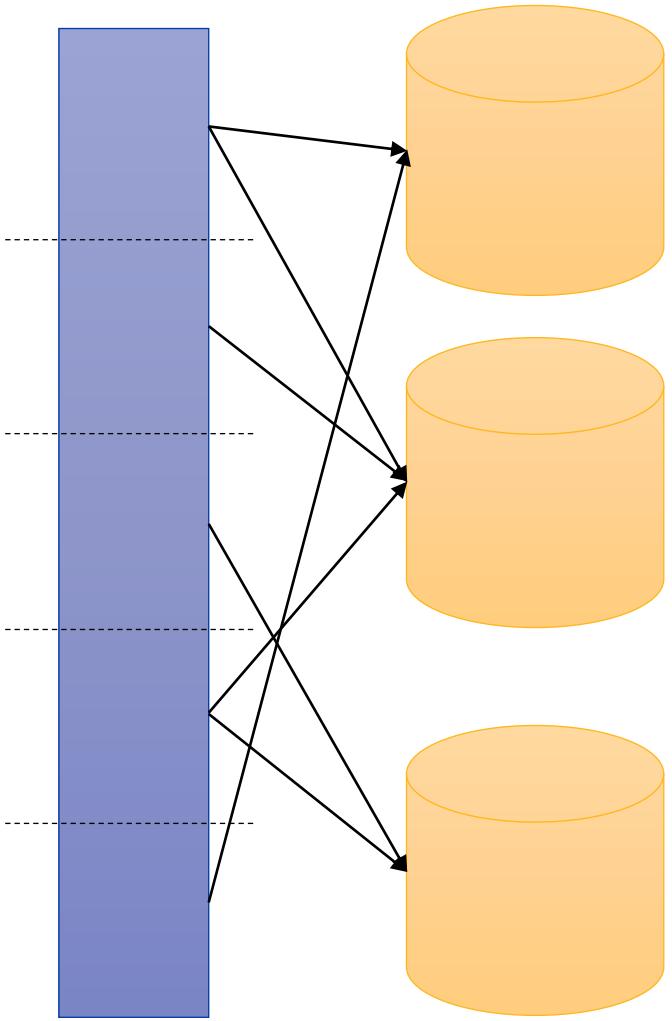
Coordination/Cluster Management
Oozie, Yarn, Kubernetes

Marlan and Rosemary Bourns
College of Engineering

Big Data Storage

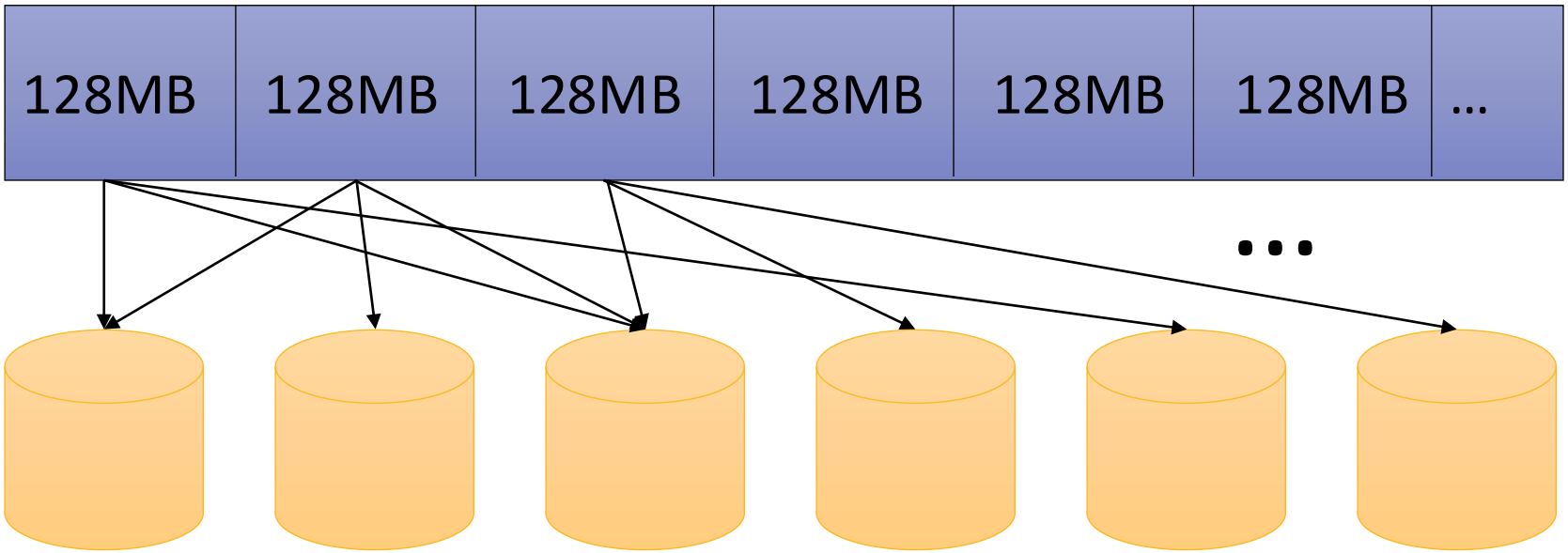
Storage of Big Data

- Data is growing faster than Moore's Law
- Too much data to fit on a single machine
- Partitioning
- Replication
- Fault-tolerance



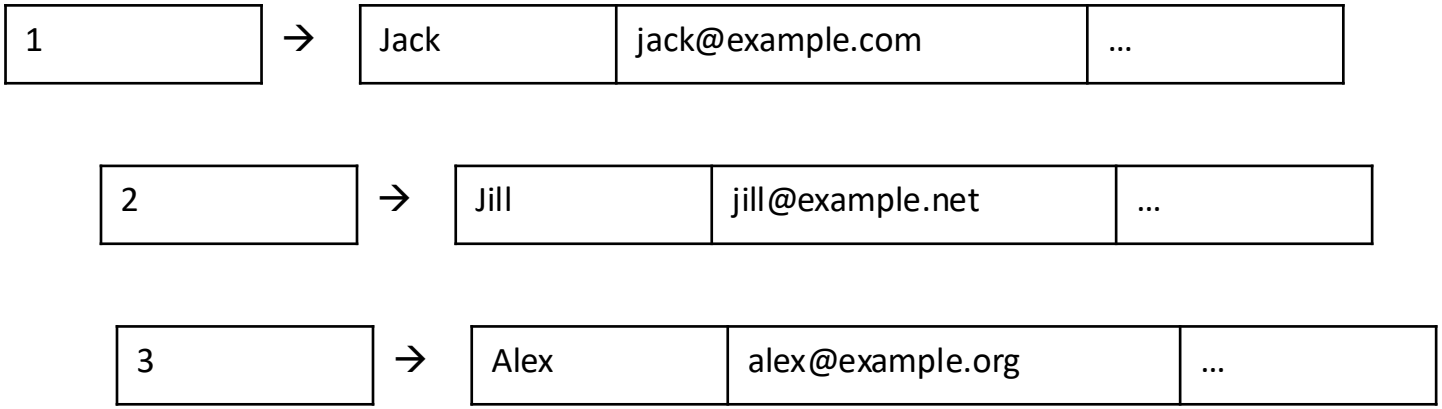
Hadoop Distributed File System (HDFS)

- The most widely used distributed file system
- Fixed-sized partitioning
- 3-way replication
- Write-once read-many
- See also: GFS, Amazon S3, Azure Blob Store

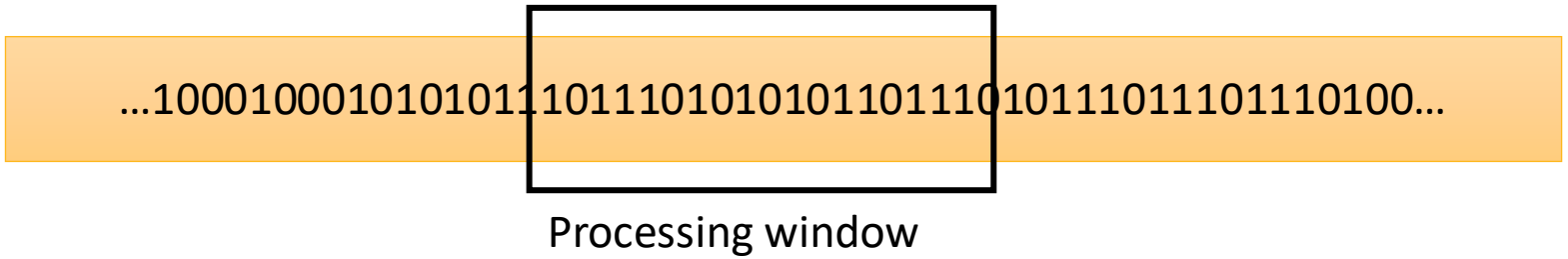


Key-value Stores

ID	Name	Email	…
1	Jack	jack@example.com	
2	Jill	jill@example.net	
3	Alex	alex@example.org	



Streaming

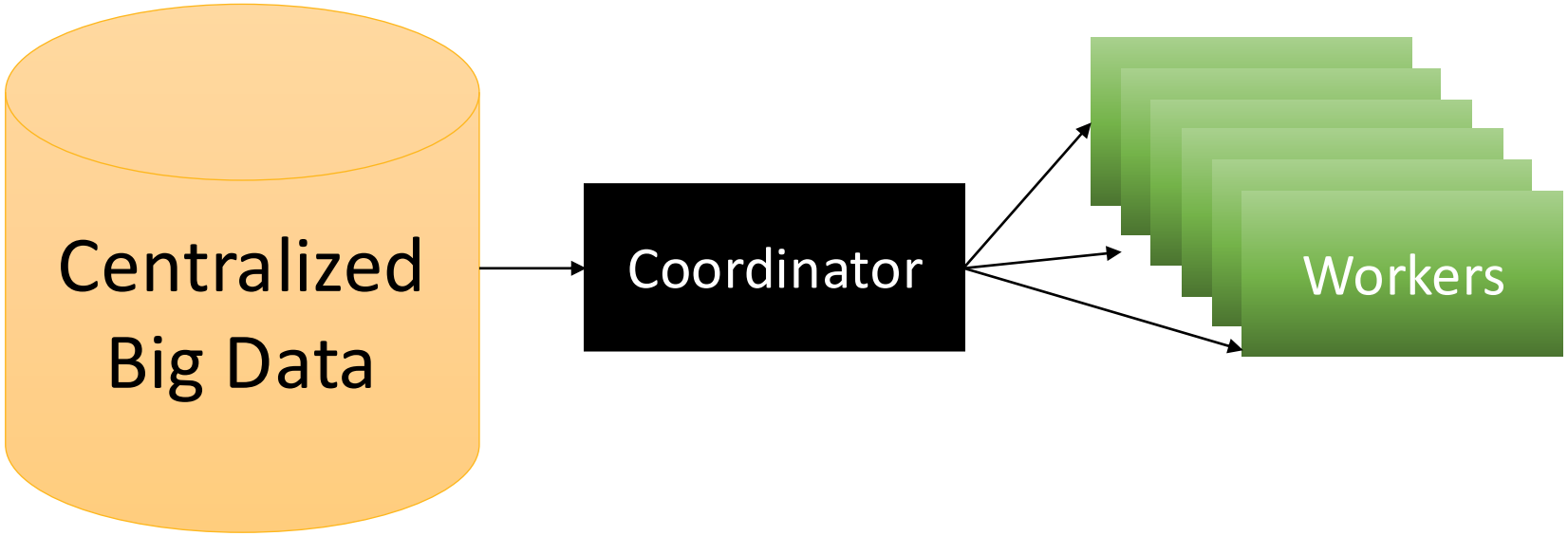


- Sub-second latency for queries
- One scan over the data
- (Partial) preprocessing
- Continuous queries
- Eviction strategies
- In-memory indexes

Marlan and Rosemary Bourns
College of Engineering

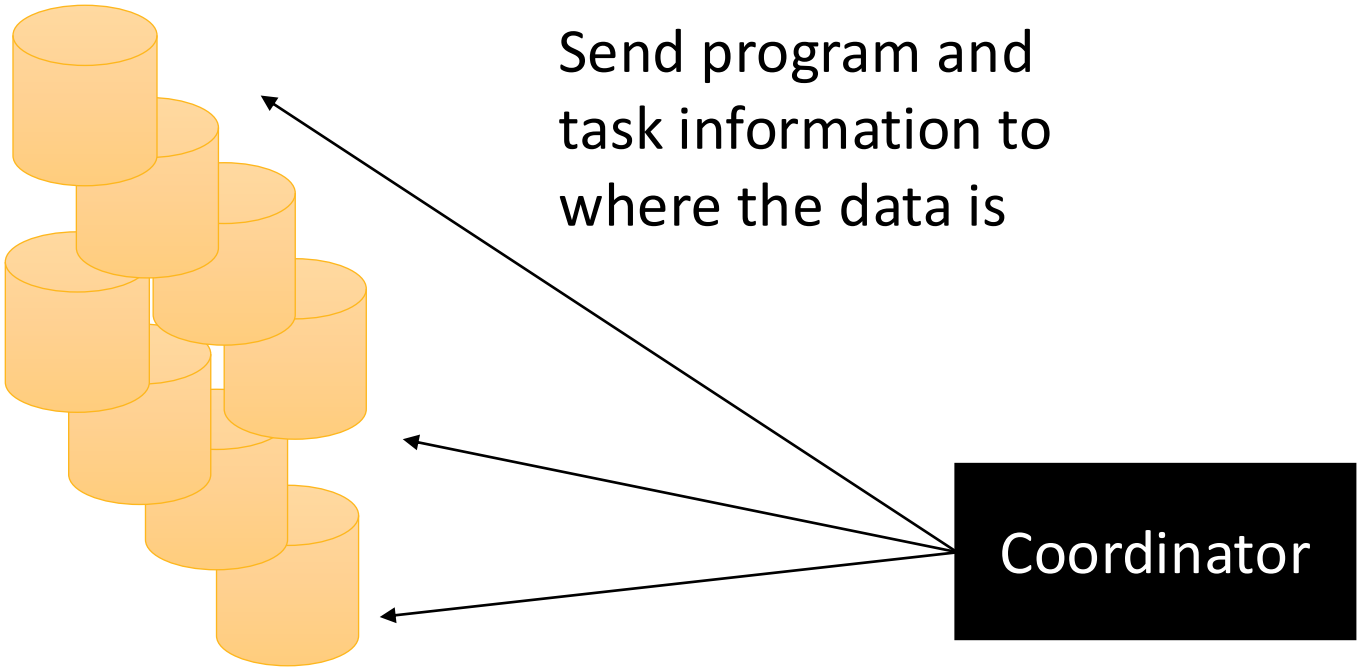
Big Data Computation Models

Traditional Distributed Computing



Ship data to computation paradigm
e.g., High performance computing (HPC)

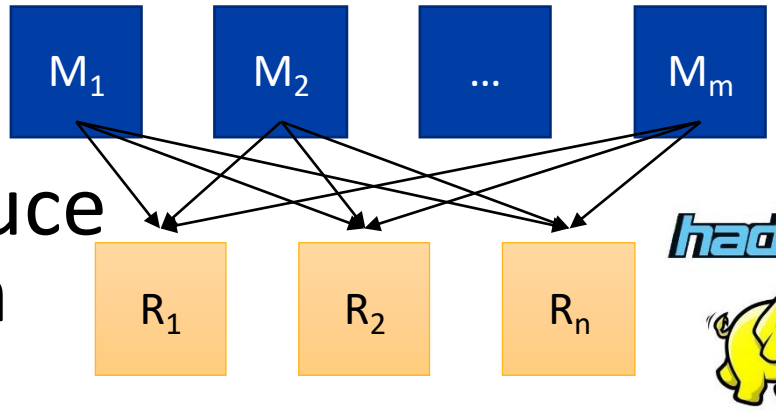
Big-data Computing

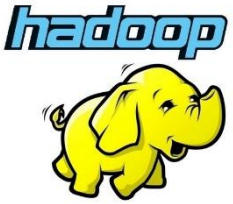


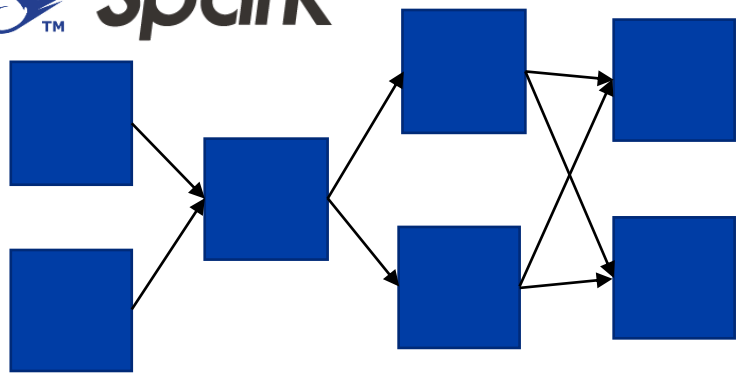
Ship compute to data paradigm

Task Execution

- MapReduce
  - Map-Shuffle- Reduce
  - Resiliency through materialization
- Resilient Distributed Datasets (RDD)
  - Directed-Acyclic-Graph (DAG)
  - In-memory processing
  - Resiliency through lineages
- Hyracks
- Stragglers
- Load balance

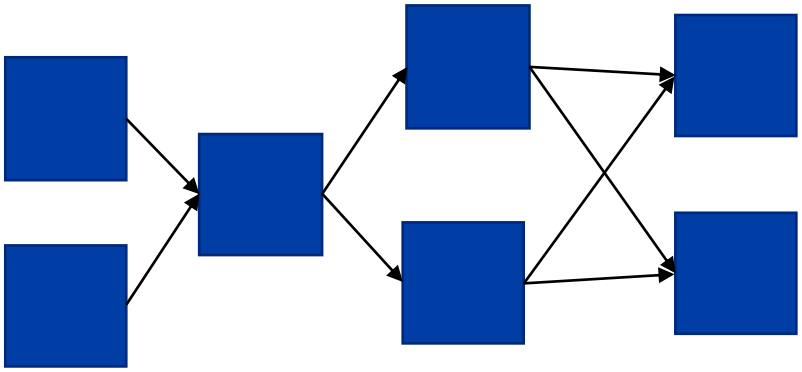






Provenance

- Debugging in distributed systems is painful



- We need to keep track of transformations on each record

Marlan and Rosemary Bourns
College of Engineering

(Semi-)Structured Big-data Processing

Structured Data Processing

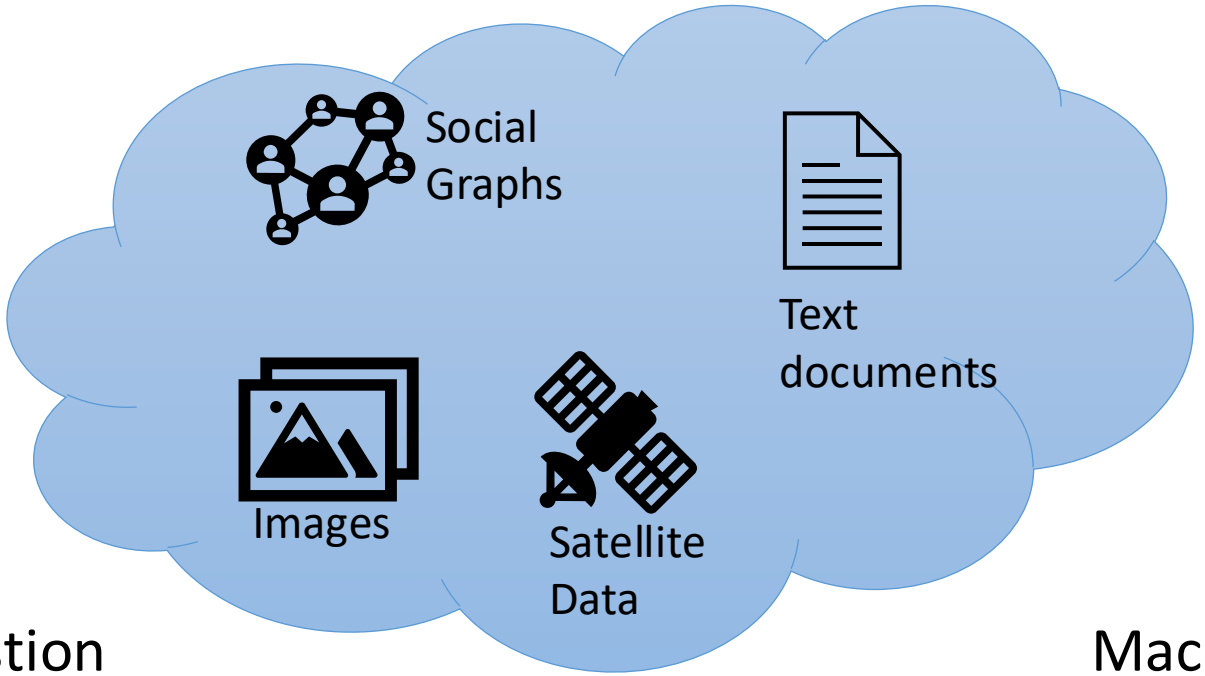
- There is a need for processing structured and semi-structured data
- The relational model and SQL are still popular
- Let the big-data system know about the structure of the data and processing
- Allow the system to optimize query processing
- Examples: Algebricks, SparkSQL, and Pig

Data Lakes

Data Transformation

Analytical Queries

Data Integration



Data Ingestion
and Retrieval

Machine Learning

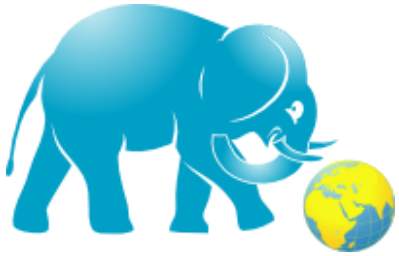
Data Lake

Marlan and Rosemary Bourns
College of Engineering

Big-data Applications

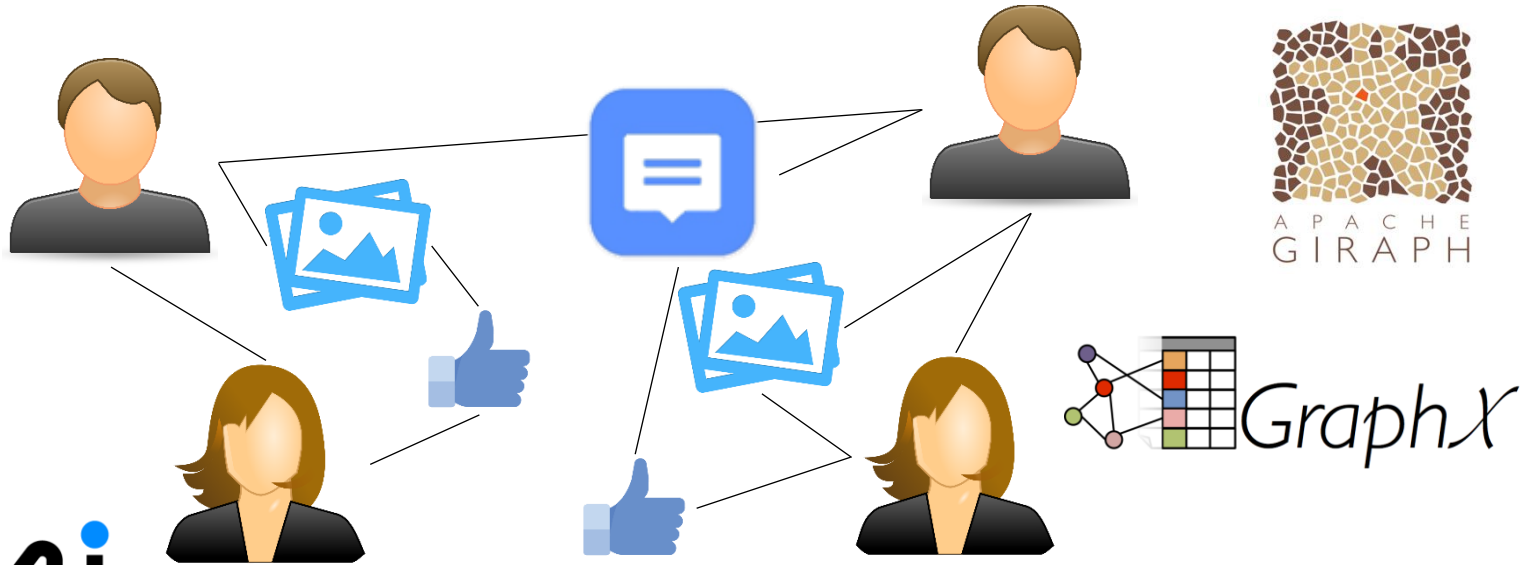
Big Spatial Data

- GPS-enabled devices
- Satellite images
- Autonomous vehicles
- Scientific applications



Big Graphs

- Motivated by social networks
- Billions of nodes and trillions of edges
- Tens of thousands of insertions per second
- Complex queries with graph traversals



neo4j

Machine Learning

- The rise of machine learning and data-driven models
- Modern generative AI models rely on big-data for training
- Data preparation and model training are costly for big data
- Use big-data processing to scale up the process

Streaming

- Internet-of-Things (IoT)
- GPS-enabled devices
- Surge of social media

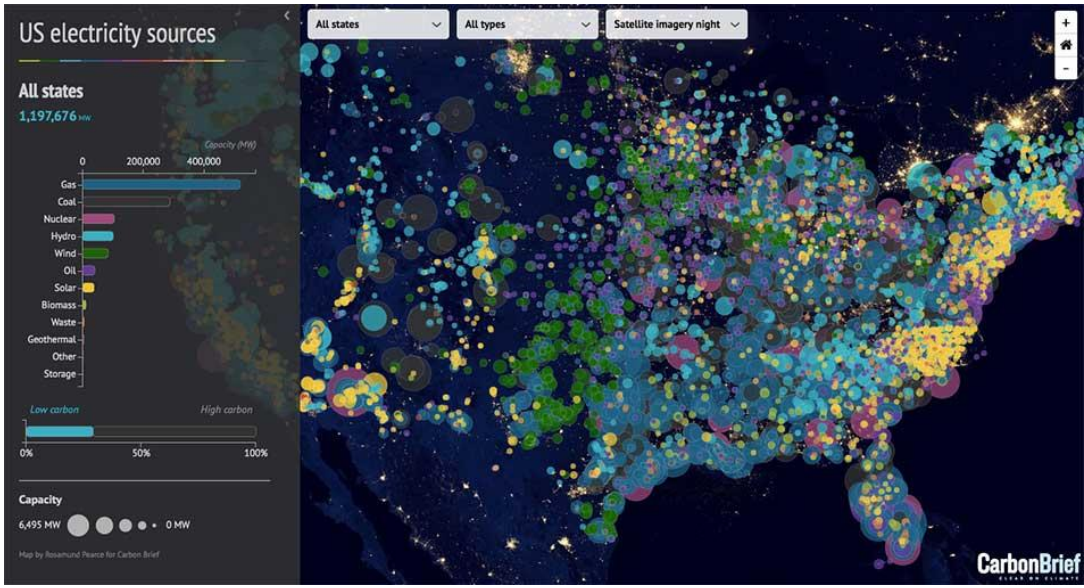
APACHE STORM™

Flink

Big-data Visualization

- Turn big-data into insights
- Effective way to communicate with non-technical users

Apache Superset™



Summary

- Data is getting big, so as computation
- Characteristics of big-data
- Big-data is a new technology for easy handling of large-scale data
- Big-data is a collection of systems, each focusing on a specific aspect of big-data, e.g., storage, processing, analysis, and visualization

Next Steps

- Think about project ideas and try to find interesting or relevant datasets
- (Required) Use the available discussion space on Canvas to initiate and discuss project ideas
  - This is your opportunity to find teammates
- (Required) Check your knowledge about the prerequisites using the short quiz on Canvas
- (Required) Check the syllabus and finish the course overview assignment