CSE 6242 / CX 4242
Course Review

Duen Horng (Polo) Chau
Associate Director, MS Analytics
Associate Professor, CSE, College of Computing
Georgia Tech

Alternate Title

10 Lessons Learned

from Working with Tech Companies

(e.g., Google, eBay, Symantec, Intel)

You need to learn **many things**.

And I bet you agree.

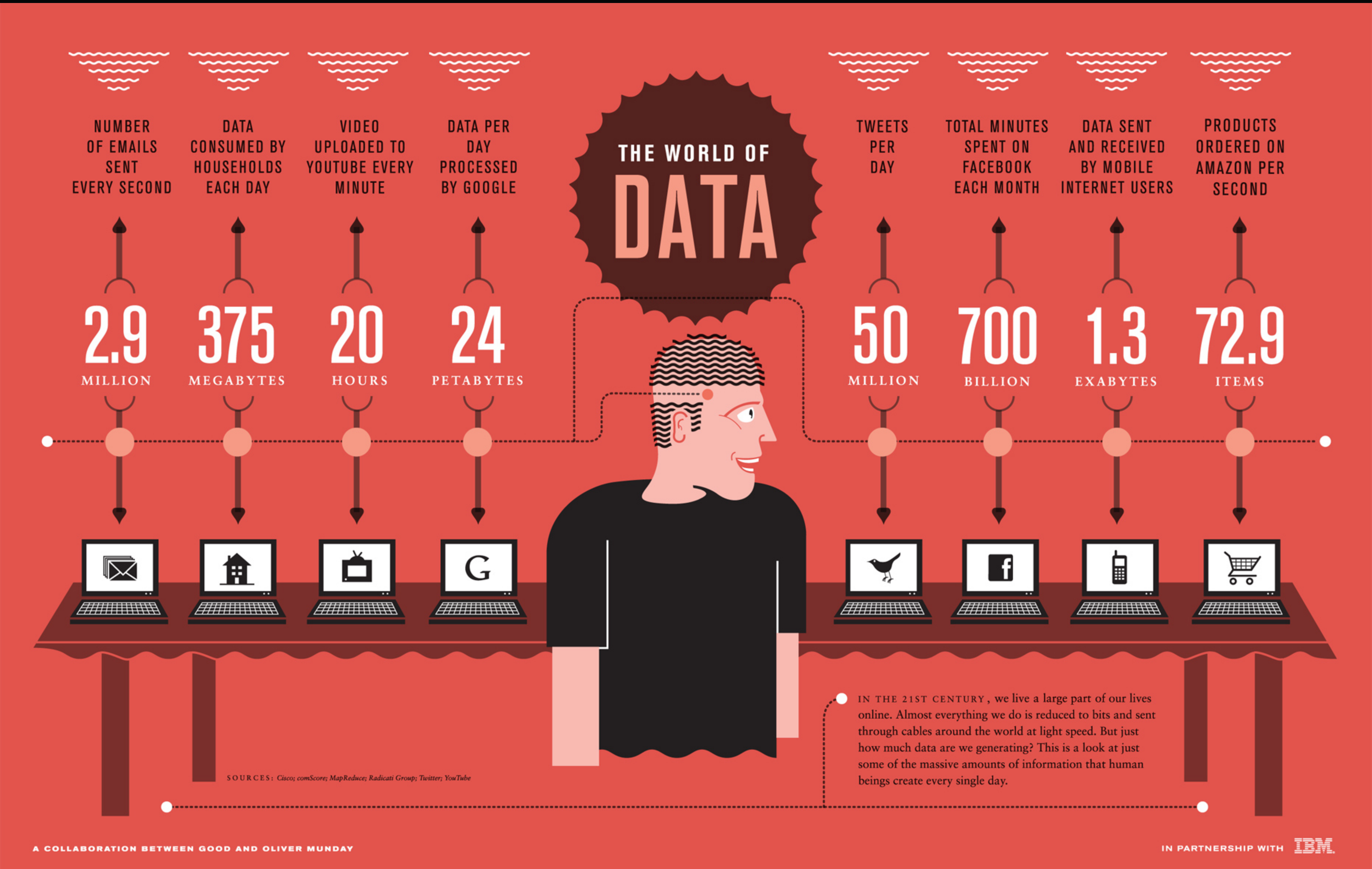
- **HW1:** Twitter API, Gephi, SQLite, OpenRefine, Gephi
- **HW2:** Tableau, D3 (Javascript, CSS, HTML, SVG)
  - Graph interaction/layout, scatter plots, heatmap/select box, sankey chart, interactive vis, Choropleth
- **HW3:** AWS, Azure, Hadoop/Java, Spark/Scala, Pig, ML Studio
- **HW4:** MMap, PageRank, random forest, Weka

Good news! Many jobs!

Most companies looking for “data scientists”

The data scientist role is critical for organizations looking to extract insight from information assets for ‘big data’ initiatives and requires a ***broad combination*** *of skills that may be fulfilled better as a team*
- Gartner (http://www.gartner.com/it-glossary/data-scientist)

Breadth of knowledge is important.



http://spanning.com/blog/choosing-between-storage-based-and-unlimited-storage-for-cloud-data-backup/

What are the “ingredients”?

What are the “ingredients”?

Need to think (a lot) about: storage, complex system design, scalability of algorithms, visualization techniques, interaction techniques, statistical tests, etc.

Analytics Building Blocks

Collection

Cleaning

Integration

Analysis

Visualization

Presentation

Dissemination

Building blocks, not "steps"

- Collection
- Cleaning
- Integration
- Analysis
- Visualization
- Presentation
- Dissemination

- Can skip some
- Can go back (two-way street)
- Examples
  - Data types inform visualization design
  - Data informs choice of algorithms
  - Visualization informs data cleaning (dirty data)
  - Visualization informs algorithm design (user finds that results don't make sense)

Python is a king.

Some say **R** is.

In practice, you may want to use the ones that have the widest community support.

Python

One of "**big-3**" programming languages at tech firms like Google.

- **Java** and **C++** are the other two.

Easy to write, read, run, and debug

- General programming language, tons of libraries
- Works well with others (a great "glue" language)

Lesson 3

You’ve got to know **SQL** and **algorithms** (and Big-O)

(Even though job descriptions may not mention them.)

Why?

(1) Many datasets stored in databases.

(2) You need to know if an algorithm can **scale** to large amount of data, and how to measure speed!

From on GT alum who are now **Googlers**:

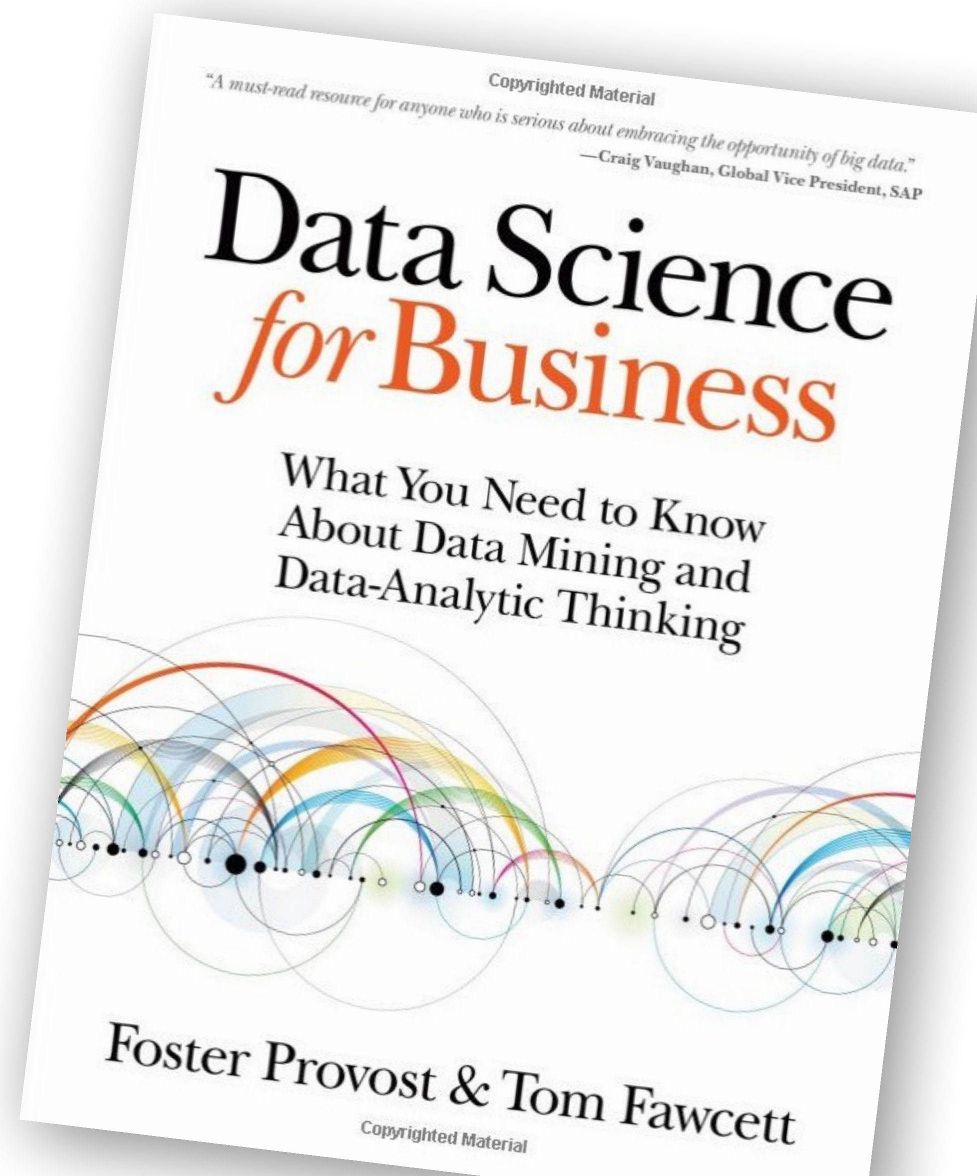
- Data structure and algorithm classes helped make them “Google ready”
- Course codes
  - CSE6140
  - CS1332, CS3510

Lesson 4

Learn **data science concepts** and key **generalizable techniques** to **future-proof** yourselves.

And here's a good book.

A critical skill in data science is the ability to decompose a data-analytics problem into pieces such that each piece matches a known task for which tools are available. Recognizing familiar problems and their solutions avoids wasting time and resources reinventing the wheel. It also allows people to focus attention on more interesting parts of the process that require human involvement—parts that have not been automated, so human creativity and intelligence must come into play.



http://www.amazon.com/Data-Science-Business-data-analytic-thinking/dp/1449361323

1. Classification

(or Probability Estimation)

Predict which of a (small) set of classes an entity belong to.

- email spam (y, n)
- sentiment analysis (+, -, neutral)
- news (politics, sports, …)
- medical diagnosis (cancer or not)
- face/cat detection
    - face detection (baby, middle-aged, etc)
- buy /not buy - commerce
- fraud detection

2. Regression ("value estimation")

Predict the **numerical value** of some variable for an entity.

- stock value
- real estate
- food/commodity
- sports betting
- movie ratings
- energy

3. Similarity Matching

Find similar entities (from a large dataset) based on what we know about them.

- price comparison (consumer, find similar priced)
- finding employees
- similar youtube videos (e.g., more cat videos)
- similar web pages (find near duplicates or representative sites) ~= clustering
- plagiarism detection

4. Clustering (unsupervised learning)

Group entities together by their similarity. (User provides # of clusters)

- groupings of similar bugs in code
- optical character recognition
    - unknown vocabulary
- topical analysis (tweets?)
- land cover: tree/road/…
- for advertising: grouping users for marketing purposes
- fireflies clustering
- speaker recognition (multiple people in same room)
- astronomical clustering

5. Co-occurrence grouping

(Many names: frequent itemset mining, association rule discovery, market-basket analysis)

Find associations between entities based on transactions that involve them

(e.g., bread and milk often bought together)

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/

6. Profiling / Pattern Mining / Anomaly Detection (unsupervised)

Characterize **typical** behaviors of an entity (person, computer router, etc.) so you can find **trends** and **outliers**.

Examples?
computer instruction prediction
removing noise from experiment (data cleaning)
detect anomalies in network traffic
moneyball
weather anomalies (e.g., big storm)
google sign-in (alert)
smart security camera
embezzlement
trending articles

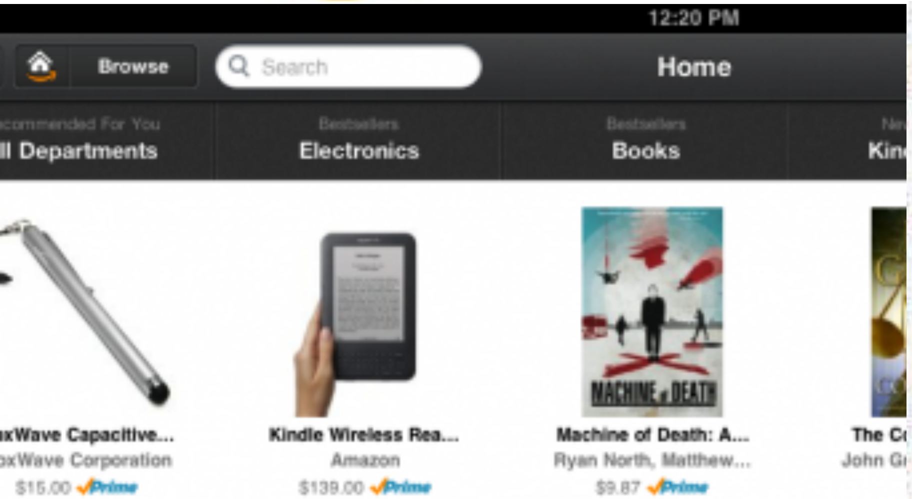
7. Link Prediction / Recommendation

Predict if two entities should be connected, and how strongly that link should be.

linkedin/facebook: people you may know

amazon/netflix: because you like terminator… suggest other movies you may also like

amazon.com



8. Data reduction ("dimensionality reduction")

Shrink a large dataset into smaller one, with as little loss of information as possible

1. if you want to visualize the data (in 2D/3D)
2. faster computation/less storage
3. reduce noise

More examples

- **Similarity functions**: central to clustering algorithms, and some classification algorithms (e.g., k-NN, DBSCAN)
- **SVD** (singular value decomposition), for NLP (LSI), and for recommendation
- **PageRank** (and its personalized version)
- **Lag plots** for auto regression, and non-linear time series foresting

Data are dirty.

Always have been.
And always will be.

You will likely spend majority of your time cleaning data. And that's important work! Otherwise, garbage in, garbage out.


Data Cleaning
Why data can be dirty?

How dirty is real data?

Examples

- Jan 19, 2016
- January 19, 16
- 1/19/16
- 2006-01-19
- 19/1/16

http://blogs.verdantis.com/wp-content/uploads/2015/02/Data-cleansing.jpg

How dirty is real data?

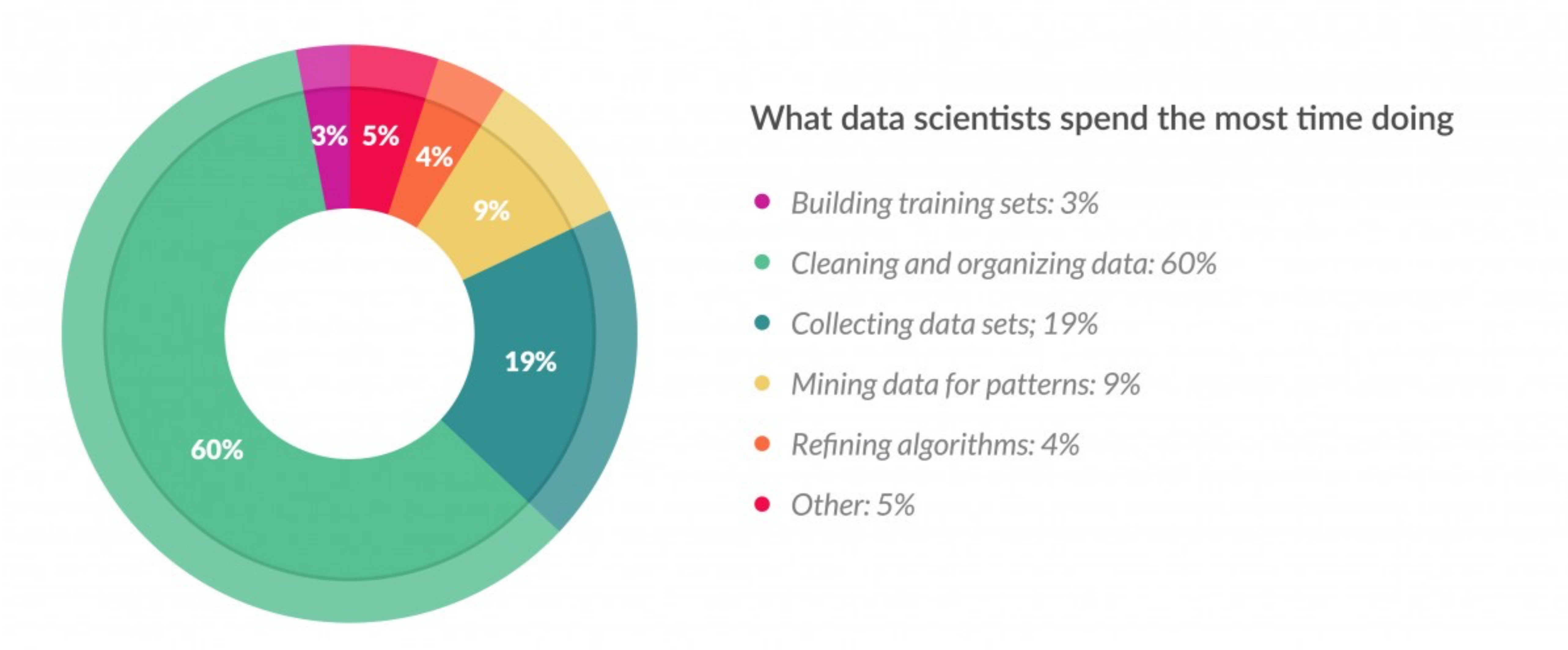
Examples

- duplicates
- empty rows
- abbreviations (different kinds)
- difference in scales / inconsistency in description/ sometimes include units
- typos
- missing values
- trailing spaces
- incomplete cells
- synonyms of the same thing
- skewed distribution (outliers)
- bad formatting / not in relational format (in a format not expected)

“80%” Time Spent on Data Preparation

Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says [Forbes]

http://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#73bf5b137f75



“80%” Time Spent on Data Cleaning

For Big-Data Scientists, ‘Janitor Work’ Is Key Hurdle to Insights [New York Times]

http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html?_r=0

Big Data's Dirty Problem [Fortune]
http://fortune.com/2014/06/30/big-data-dirty-problem/

Data Janitor

The Silver Lining

“Painful process of cleaning, parsing, and proofing one’s data”
— one of the three sexy skills of data geeks (the other two: statistics, visualization)

http://medriscoll.com/post/4740157098/the-three-sexy-skills-of-data-geeks

@BigDataBorat tweeted
**“Data Science is 99% preparation,
1% misinterpretation.”**

Follow us on: Github Twitter

Google™ Custom Search Search ×

OPEN Refine

A free, open source, powerful tool for working with messy data

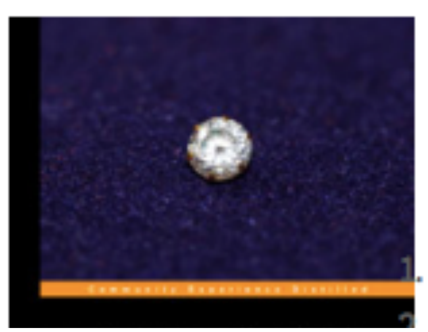
- Home
- Download
- Documentation
- Community
- Post archive
- A Governance Model for OpenRefine
- Using OpenRefine: a manual

Welcome!

OpenRefine (formerly Google Refine) is a powerful tool for working with messy data: cleaning it; transforming it from one format into another; extending it with web services; and linking it to databases like Freebase.

Please note that since October 2nd, 2012, Google is not actively supporting this project, which has now been rebranded to OpenRefine. Project development, documentation and promotion is now fully supported by volunteers. Find out more about the history of OpenRefine and how you can help the community.

Using OpenRefine - The Book



Using OpenRefine, by Ruben Verborgh and Max De Wilde, offers a great introduction to OpenRefine. Organized by recipes with hands on examples, the book covers the following topics:

1. Import data in various formats
2. Explore datasets in a matter of seconds

Lesson 6

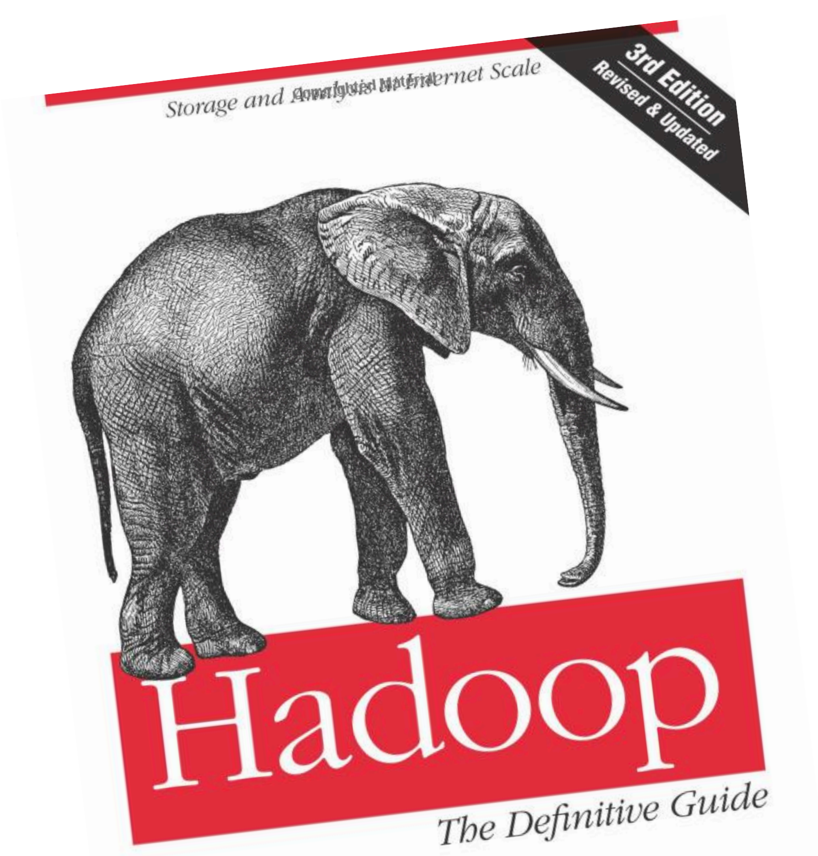
Learn D3 and visualization basics

Seeing is believing.

A huge competitive edge.

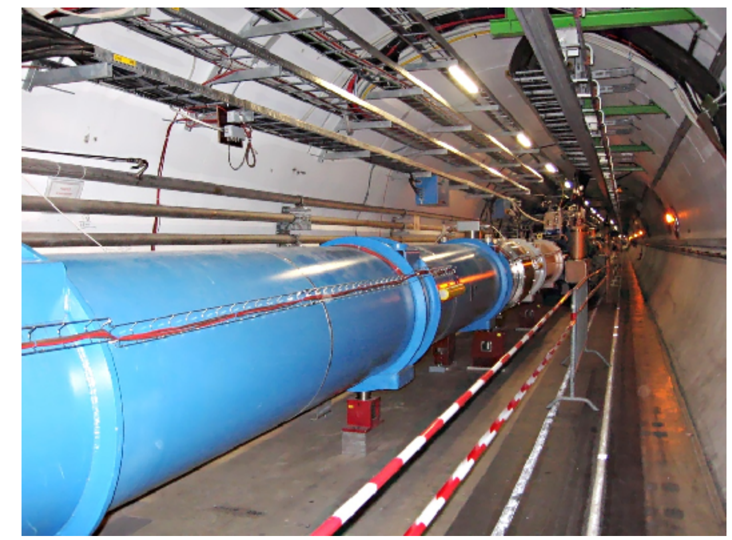
Companies expect you-all to know the “basic” **big data technologies**

(e.g., Hadoop, Spark)



“Big Data” is Common...

Google processed **24 PB / day** (2009)



Facebook’s add **0.5 PB / day** to its data warehouses

CERN generated **200 PB** of data from “Higgs boson” experiments

Avatar’s 3D effects took **1 PB** to store

http://www.theregister.co.uk/2012/11/09/facebook_open_sources_corona/
http://thenextweb.com/2010/01/01/avatar-takes-1-petabyte-storage-space-equivalent-32-year-long-mp3/
http://dl.acm.org/citation.cfm?doid=1327452.1327492

Machines and disks die

3% of 100,000 hard drives
fail within **first 3 months**

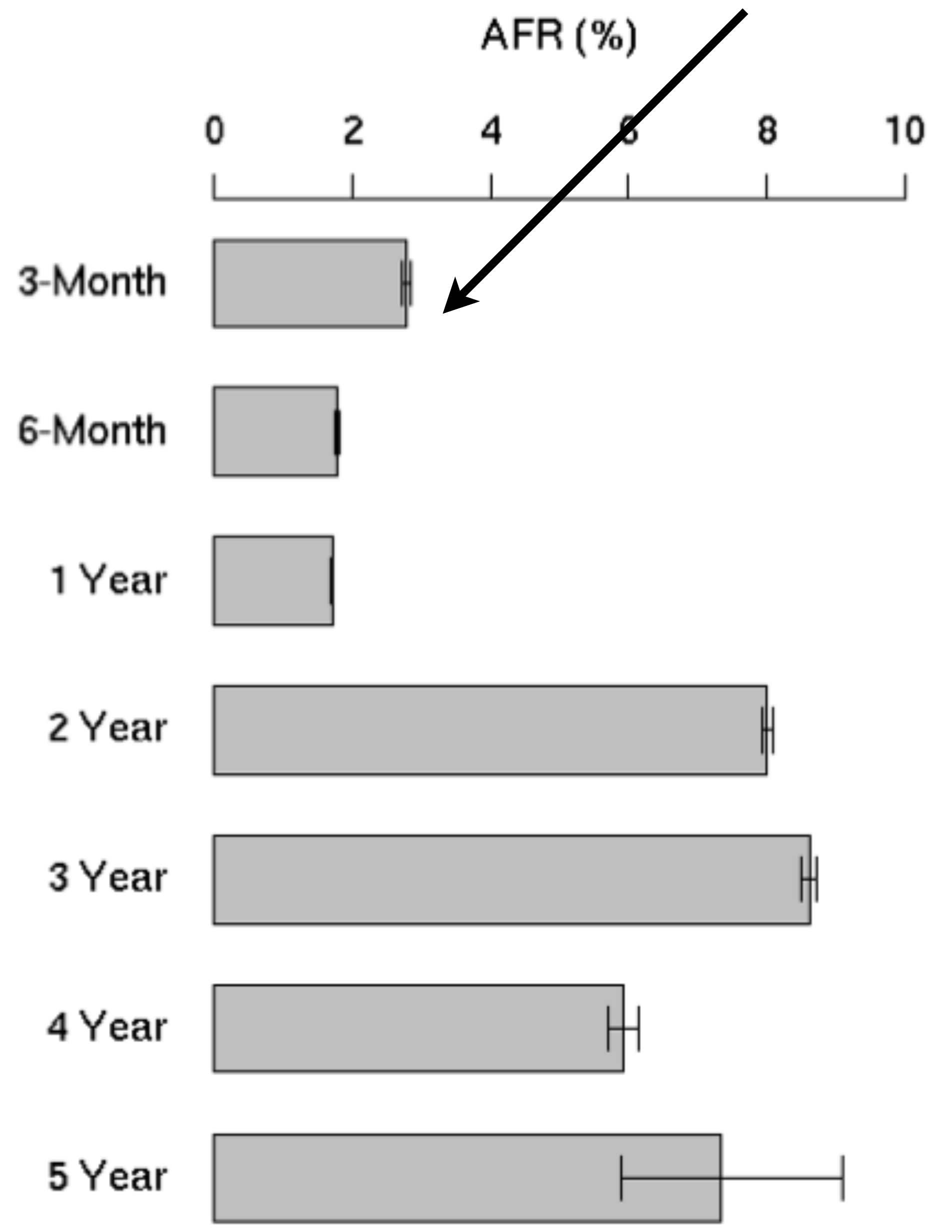


Figure 2: Annualized failure rates broken down by age groups

Failure Trends in a Large Disk Drive Population
http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en/us/archive/disk_failures.pdf

http://arstechnica.com/gadgets/2015/08/samsung-unveils-2-5-inch-16tb-ssd-the-worlds-largest-hard-drive/

Open-source software for reliable, scalable, distributed computing

Written in Java

Scale to thousands of machines

- Linear scalability (with good algorithm design): if you have 2 machines, your job runs twice as fast

Uses simple programming model (MapReduce)

Fault tolerant (HDFS)

- Can recover from machine/disk failure (no need to restart computation)

http://hadoop.apache.org

Why learn Hadoop?

Fortune 500 companies use it

Many research groups/projects use it

Strong community support, and favored/backed my major companies, e.g., IBM, Google, Yahoo, eBay, Microsoft, etc.

It's **free**, open-source

Low cost to set up (works on commodity machines)

Will be an "essential skill", like SQL

http://strataconf.com/strata2012/public/schedule/detail/22497

Spark is now pretty popular.

(Somewhat eclipsed by Tensorflow/deep learning etc.)

Project History

Spark project started in 2009 at UC Berkeley AMP lab, open sourced 2010

Became Apache Top-Level Project in Feb 2014

Shark/Spark SQL started summer 2011

Built by 250+ developers and people from 50 companies

Scale to 1000+ nodes in production

In use at Berkeley, Princeton, Klout, Foursquare, Conviva, Quantifind, Yahoo! Research, …

http://en.wikipedia.org/wiki/Apache_Spark

Why a New Programming Model?

MapReduce greatly simplified big data analysis

But as soon as it got popular, users wanted more:

- More **complex**, multi-stage applications (e.g. iterative graph algorithms and machine learning)
- More **interactive** ad-hoc queries

Why a New Programming Model?

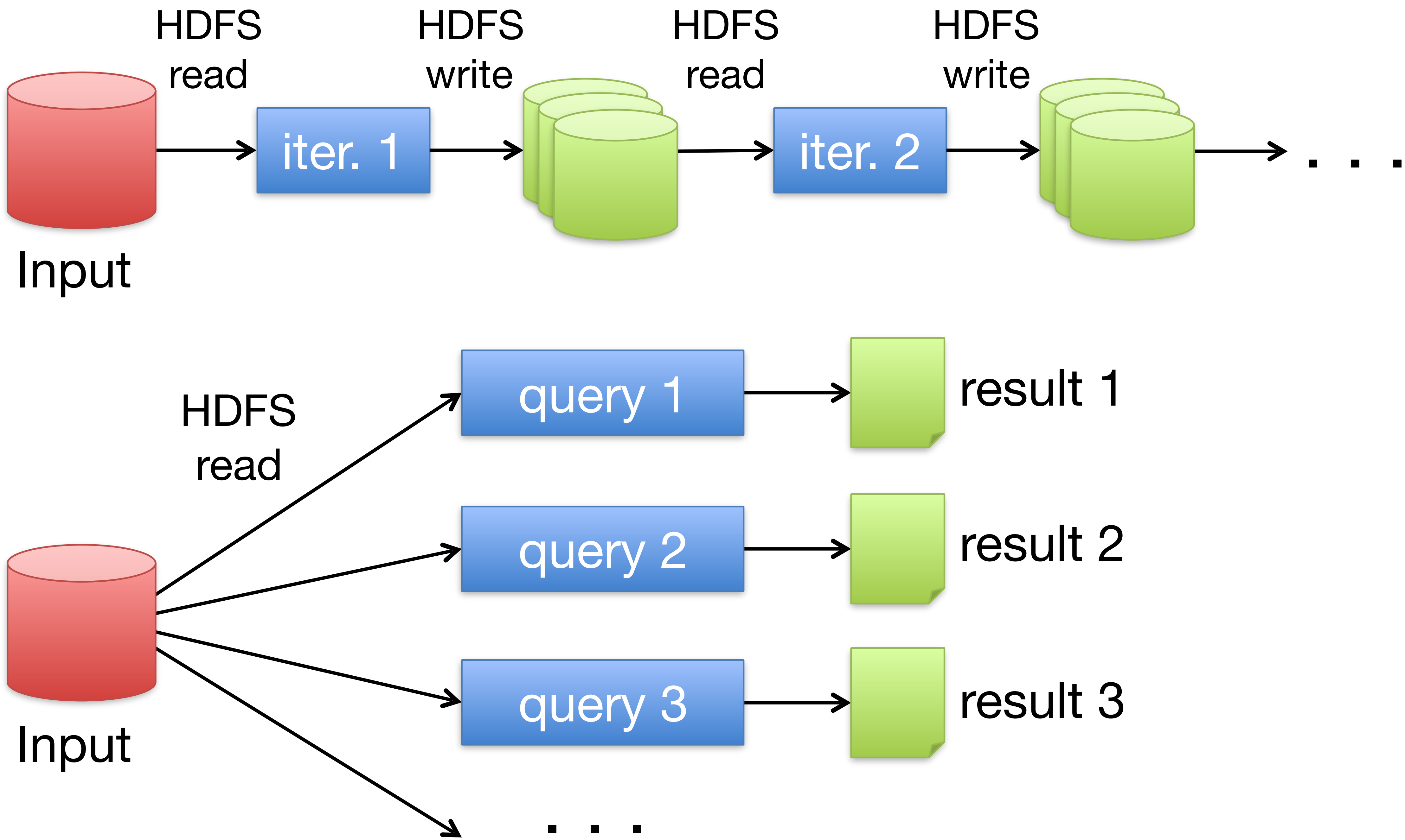
MapReduce greatly simplified big data analysis

But as soon as it got popular, users wanted more:

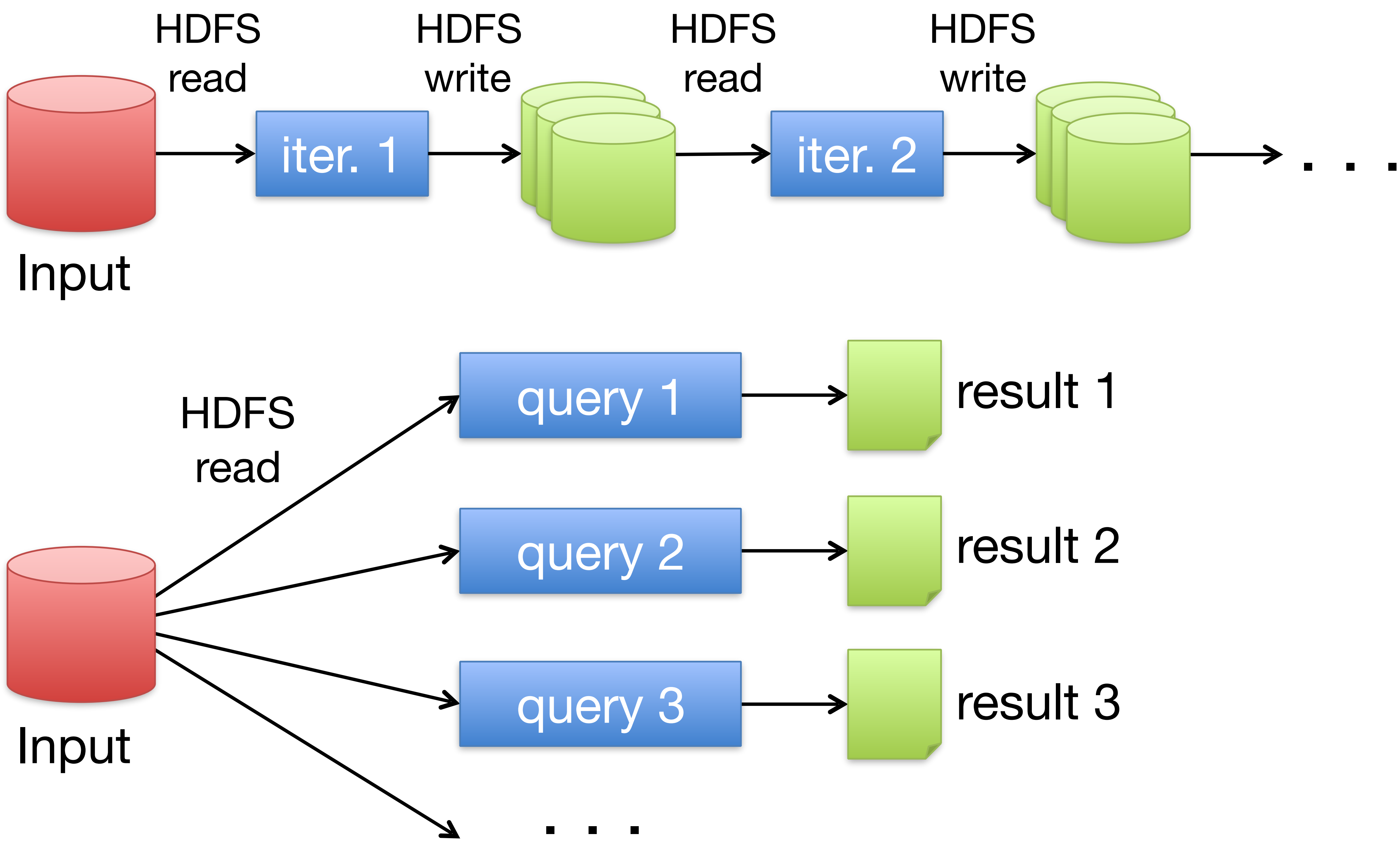
- More **complex**, multi-stage applications (e.g. iterative graph algorithms and machine learning)
- More **interactive** ad-hoc queries

Require faster **data sharing** across parallel jobs

Data Sharing in MapReduce

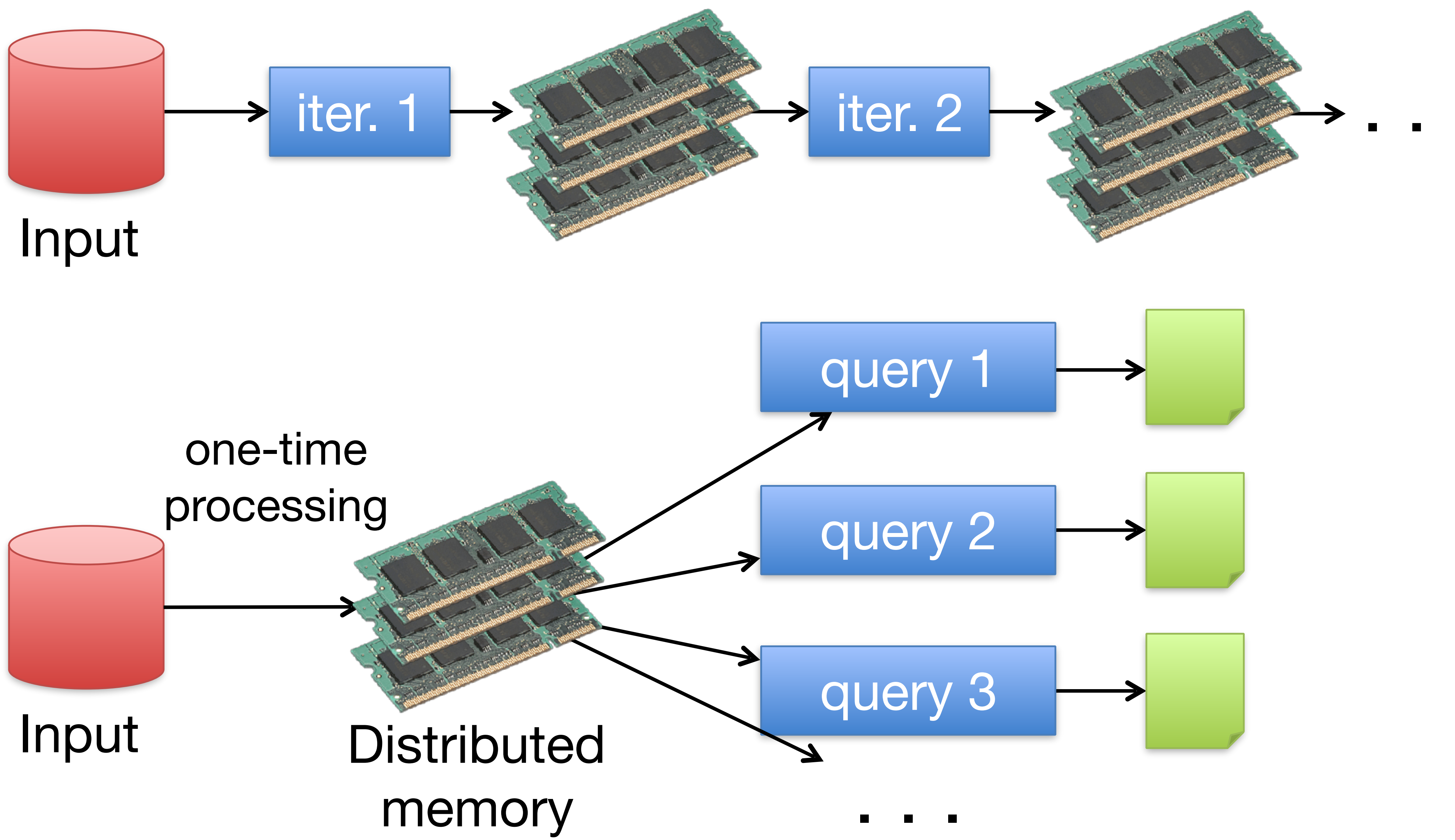


Data Sharing in MapReduce

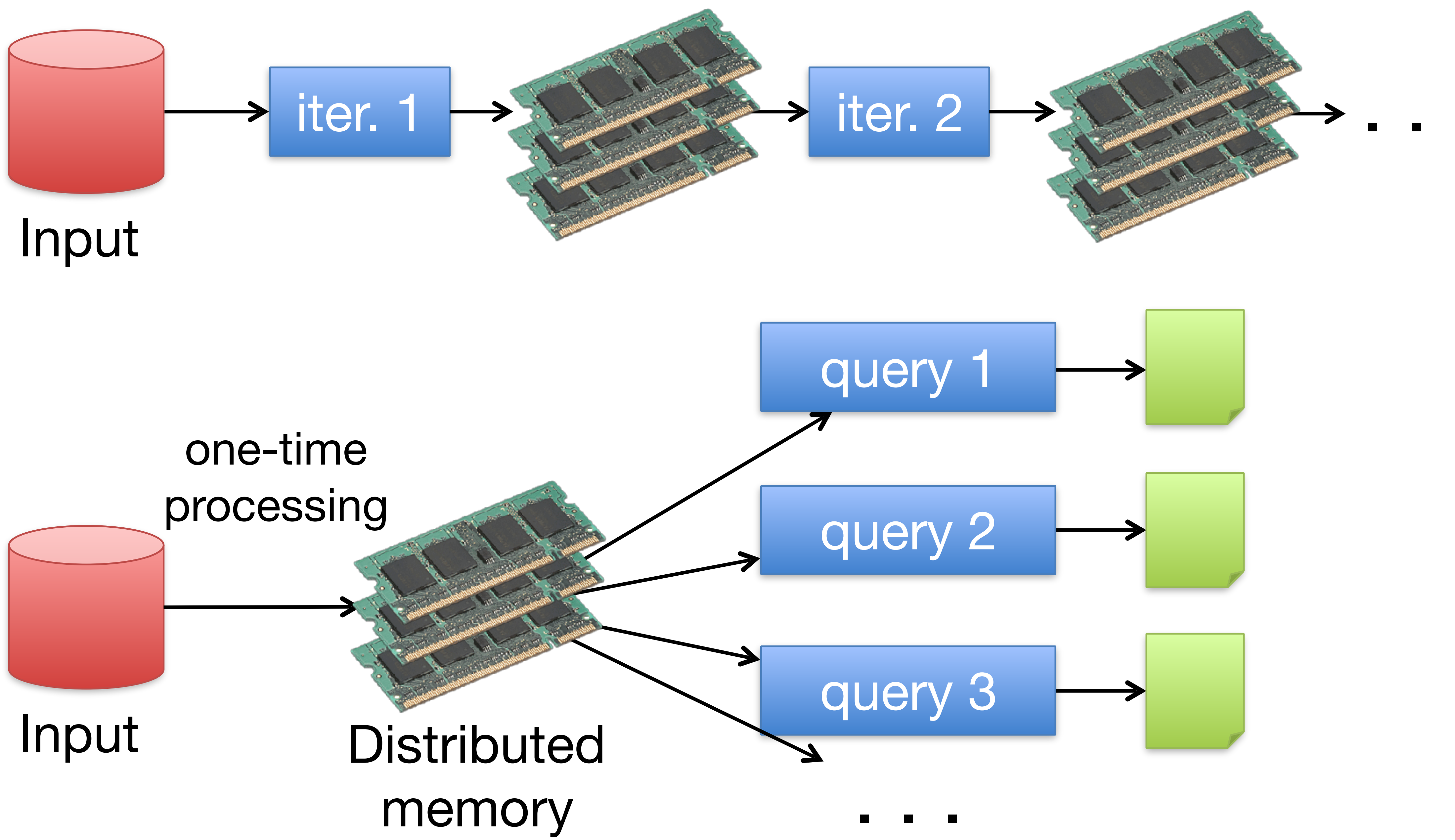


Slow due to replication, serialization, and disk IO

Data Sharing in Spark



Data Sharing in Spark



10-100× faster than network and disk

Is MapReduce dead? No!

Google Dumps MapReduce in Favor of New Hyper-Scale Analytics System

http://www.datacenterknowledge.com/archives/2014/06/25/google-dumps-mapreduce-favor-new-hyper-scale-analytics-system/

http://www.reddit.com/r/compsci/comments/296aqr/on_the_death_of_mapreduce_at_google/

P = NP reddit COMPSCI **comments** related other discussions (3)

87 On the Death of Map-Reduce at Google. (the-paper-trail.org)
submitted 3 months ago by qkdhfjdjdhd
20 comments share

all 20 comments
sorted by: best

[–] tazzy531 47 points 3 months ago
As an employee, I was surprised by this headline, considering I just ran some mapreduces this past week.
After digging further, this headline and article is rather inaccurate.
Cloud DataFlow is the external name for what is internally called Flume.
Flume is a layer that runs on top of MapReduce that abstracts away the complexity into something that is much easier

Industry moves fast.
So should you.

Be **cautiously optimistic**.
And be careful of **hype**.

There were 2 AI winters.
https://en.wikipedia.org/wiki/History_of_artificial_intelligence

Gartner's 2015 Hype Cycle

http://www.gartner.com/newsroom/id/3114217

Lesson 10

Your **soft skills** can be more important than your hard skills.

If people don't understand your approach, they won't appreciate it.