Course Review

CSE 6242 / CX 4242

Duen Horng (Polo) Chau
Associate Professor & ML Area Leader, College of Computing
Associate Director, MS Analytics
Georgia Tech

Twitter: @PoloChau

Alternative Title

11 Lessons Learned from Working with Tech Companies (Facebook, Google, Intel, eBay, Symantec)

You need to learn **many things**.

And I bet you agree.

- **HW1:** Data collection via API, SQLite, OpenRefine, Gephi
- **HW2:** Tableau, D3 (Javascript, CSS, HTML, SVG)
- **HW3:** AWS, Azure, Hadoop/Java, Spark/Scala, Pig, ML Studio
- **HW4:** PageRank, random forest, Scikit-learn

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.

THE WORLD OF DATA

NUMBER OF EMAILS SENT EVERY SECOND

2.9
MILLION

DATA CONSUMED BY HOUSEHOLDS EACH DAY

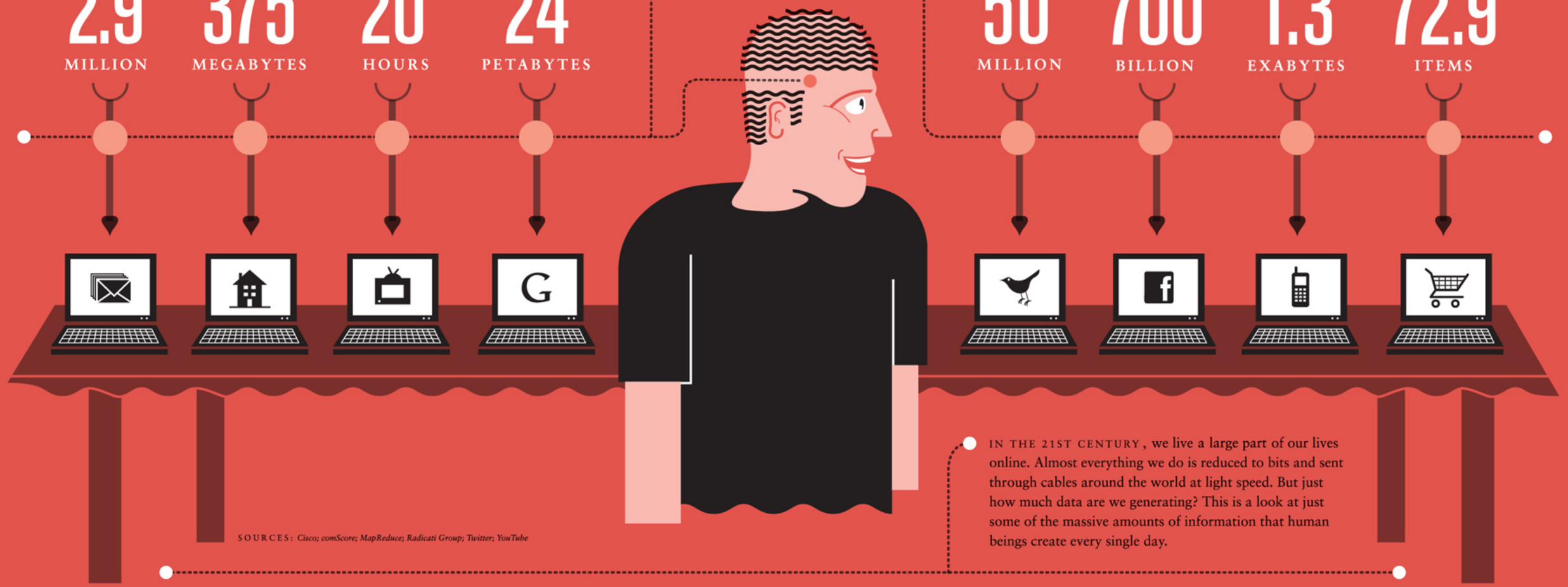
375
MEGABYTES

VIDEO UPLOADED TO YOUTUBE EVERY MINUTE

20
HOURS

DATA PER DAY PROCESSED BY GOOGLE

24
PETABYTES



TWEETS PER DAY

50
MILLION

TOTAL MINUTES SPENT ON FACEBOOK EACH MONTH

700
BILLION

DATA SENT AND RECEIVED BY MOBILE INTERNET USERS

1.3
EXABYTES

PRODUCTS ORDERED ON AMAZON PER SECOND

72.9
ITEMS

SOURCES: *Cisco; comScore; MapReduce; Radicati Group; Twitter; YouTube*

IN THE 21ST CENTURY, we live a large part of our lives online. Almost everything we do is reduced to bits and sent through cables around the world at light speed. But just how much data are we generating? This is a look at just some of the massive amounts of information that human beings create every single day.

A COLLABORATION BETWEEN GOOD AND OLIVER MUNDAY

IN PARTNERSHIP WITH IBM

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

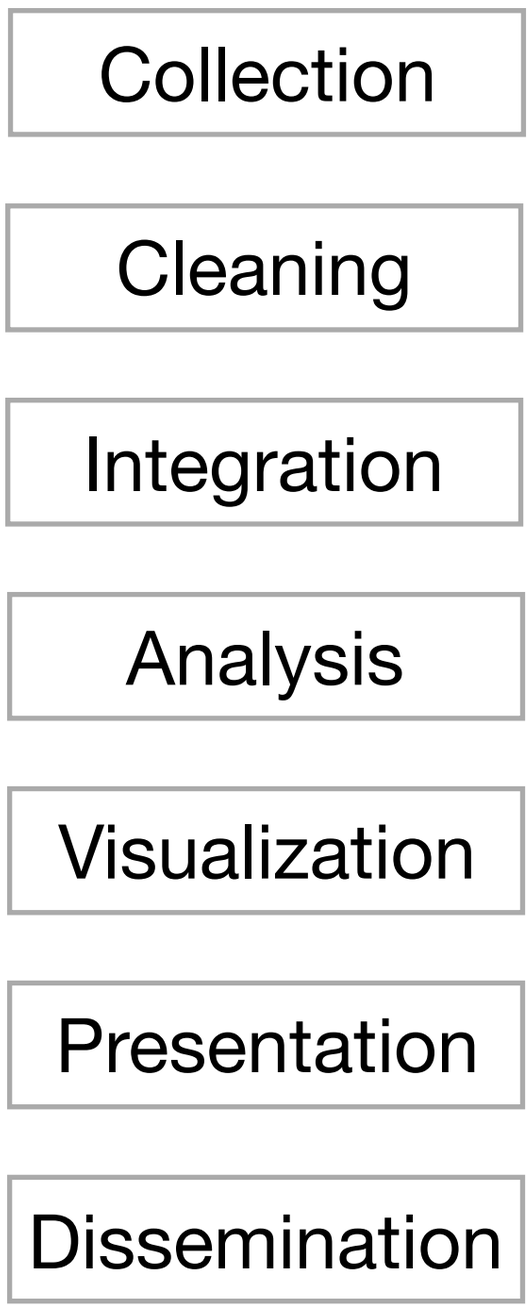
Analysis

Visualization

Presentation

Dissemination

Building blocks, not “steps”

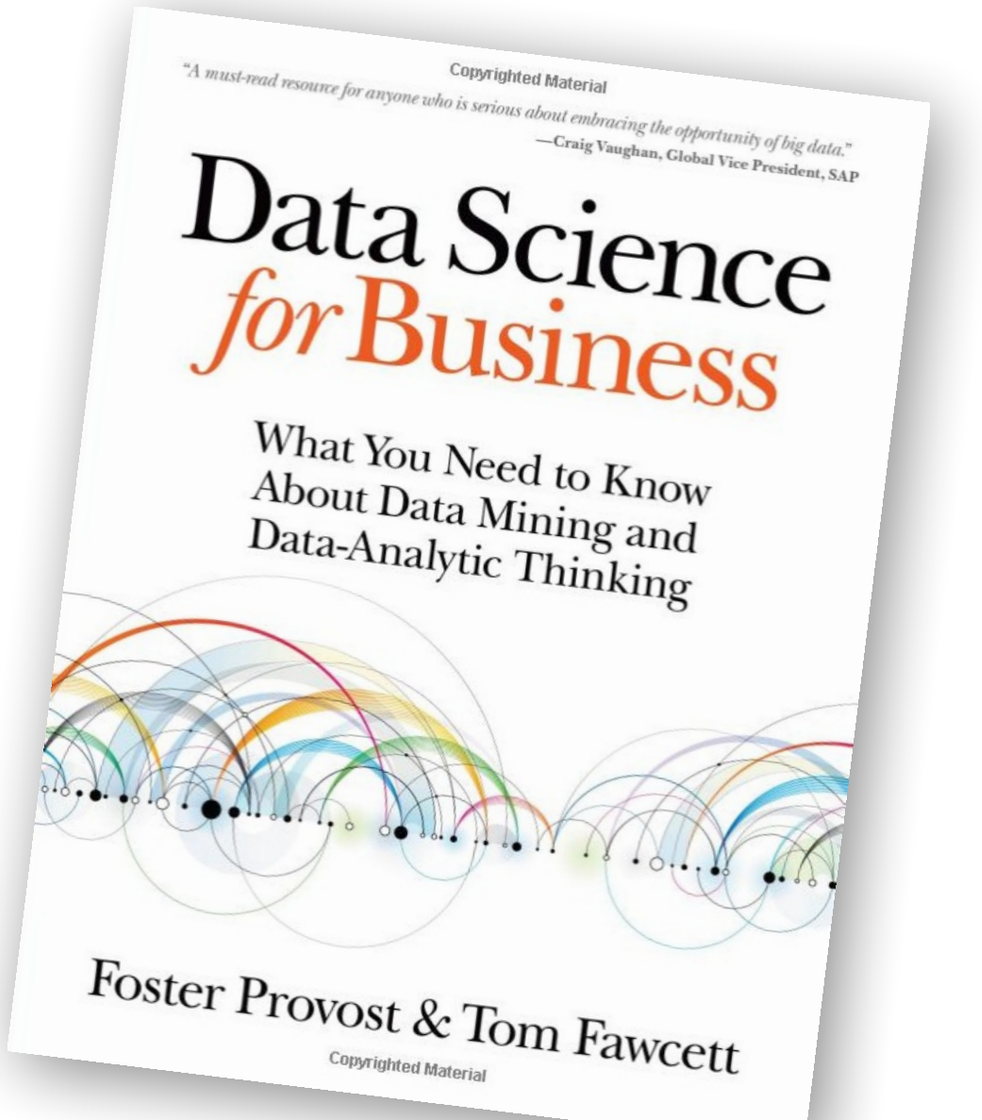


- Can skip some
- Can go back (two-way street)
- Examples
  - Data types inform visualization design
  - Data informs choice of algorithms
  - Visualization informs data cleaning (dirty data)
  - Visualization informs algorithm design (user finds that results don’t make sense)

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

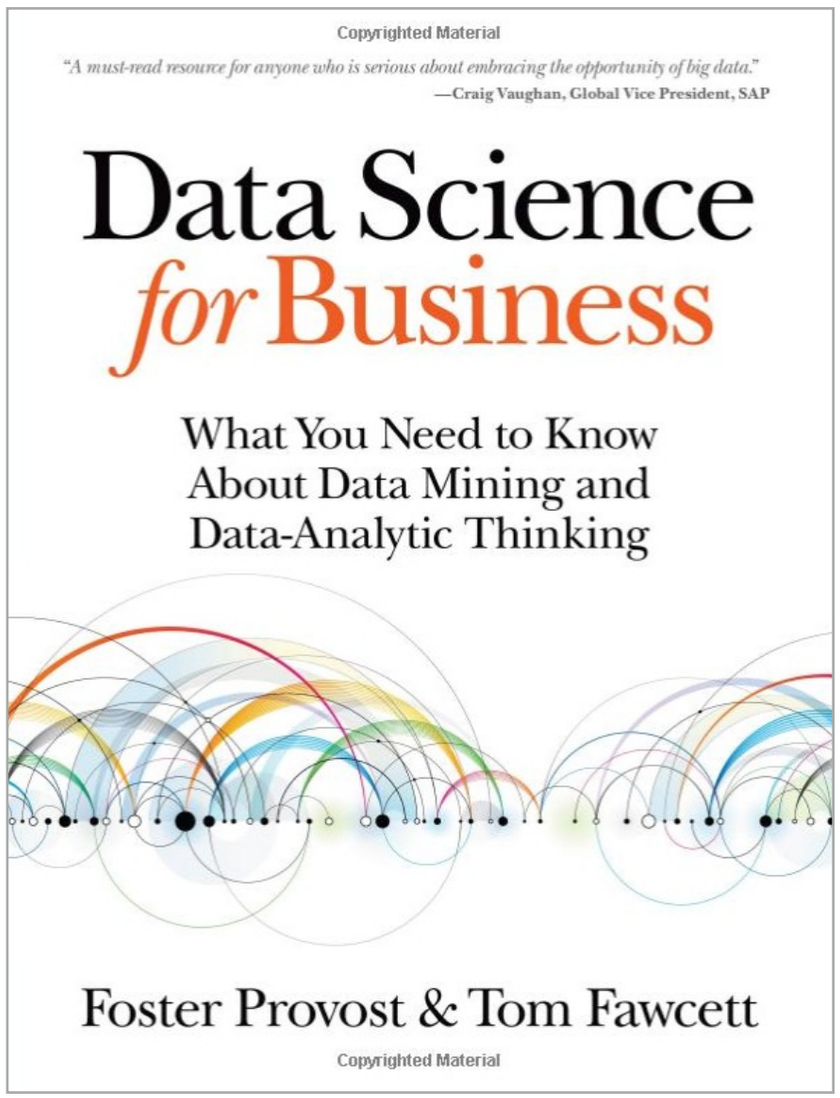
And here's a good book.

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

Great news!
Few principles!!



1. **Classification**
2. **Regression**
3. **Similarity Matching**
4. **Clustering**
5. **Co-occurrence grouping**
   (aka frequent items mining, association rule discovery, market-basket analysis)
6. **Profiling**
   (related to pattern mining, anomaly detection)
7. **Link prediction** / recommendation
8. **Data reduction**
   (aka dimensionality reduction)
9. **Causal modeling**

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

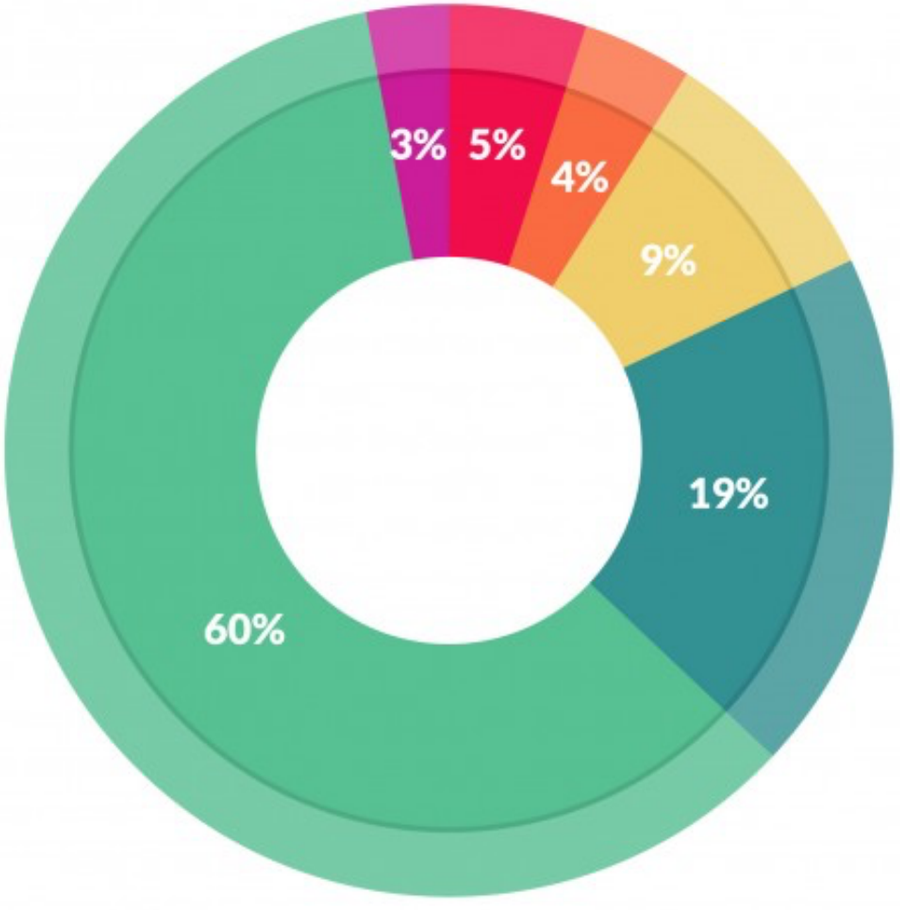
How dirty is real data?

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



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

We are all Data Janitors!

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.”

OpenRefine

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

Home

Download

Documentation

Community

Post archive

A Governance Model for OpenRefine

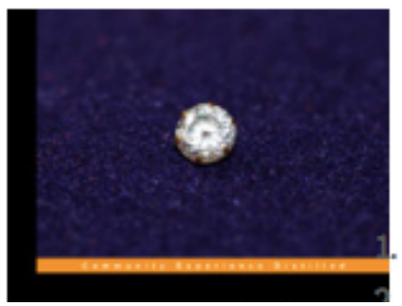
Using OpenRefine: a manual

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

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 5

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

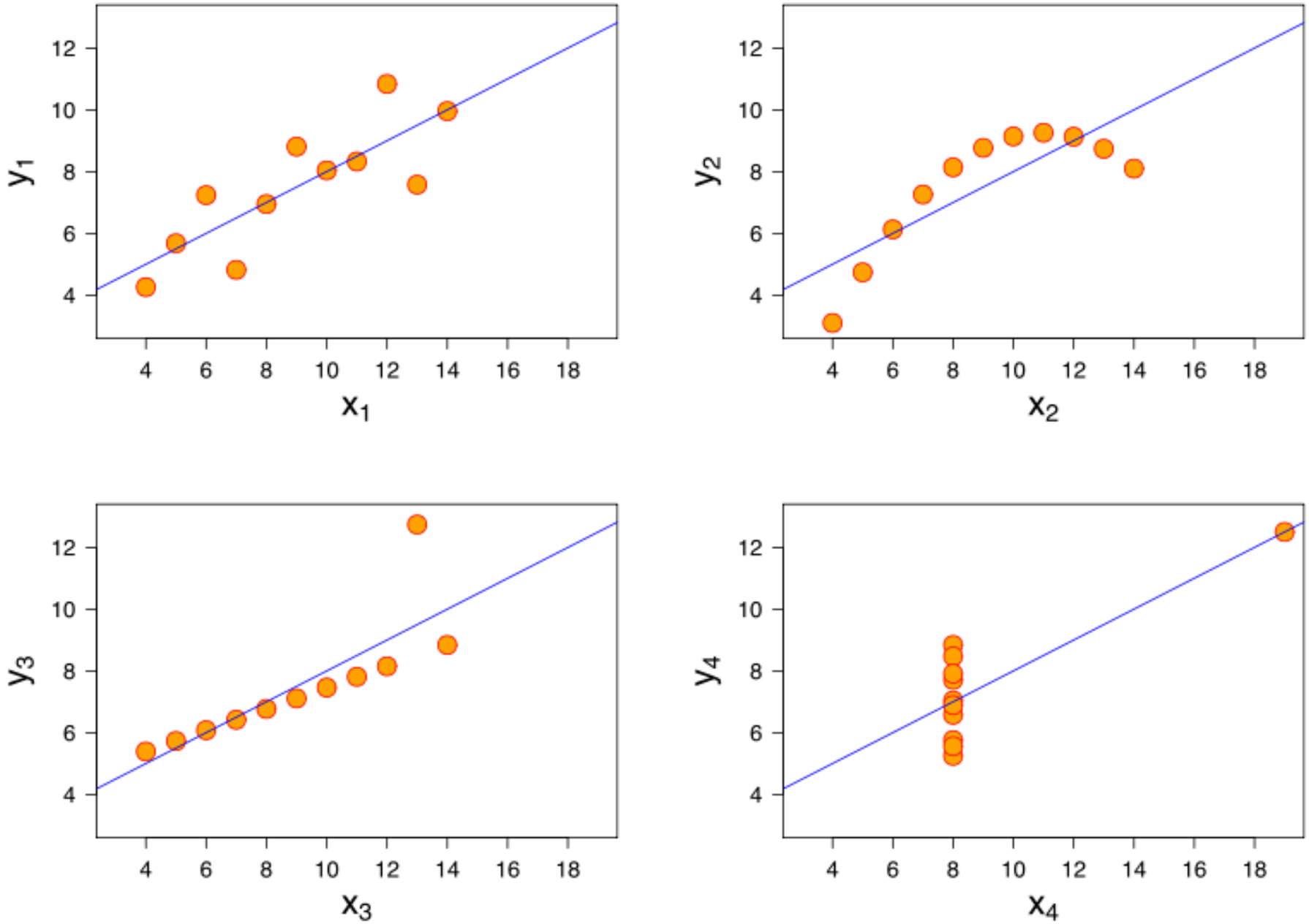
Visualization is **NOT** only about "making things look pretty"

(Aesthetics is important too)

Key is to design **effective** visualization to:
(1) **communicate** and
(2) help people **gain insights**

Why **visualize** data? Why not automate?

Anscombe's Quartet

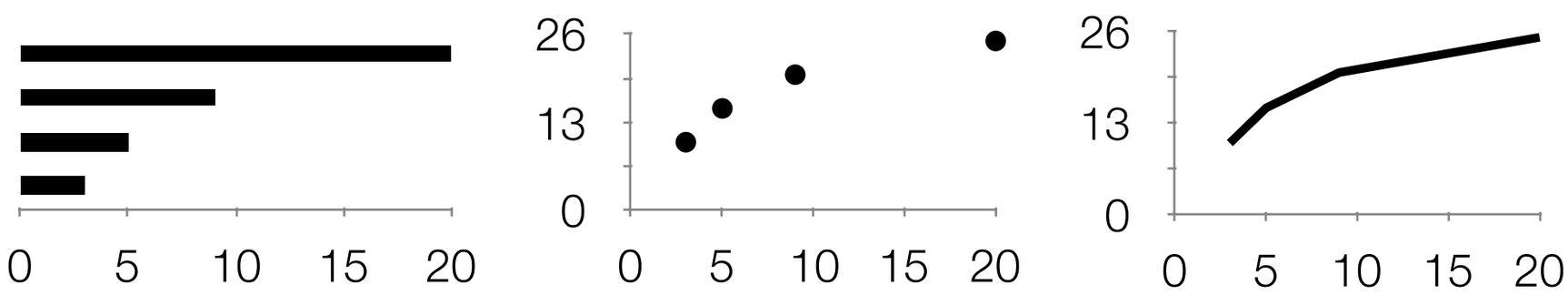


https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Designing **effective** visualization is **not hard if you learn the principles**.

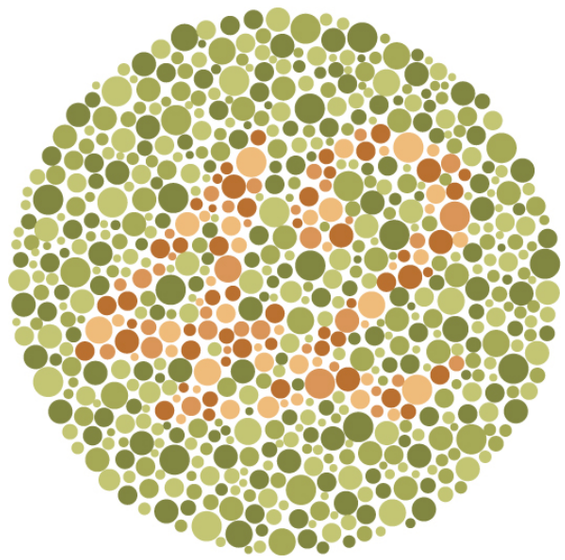
Easy, because…

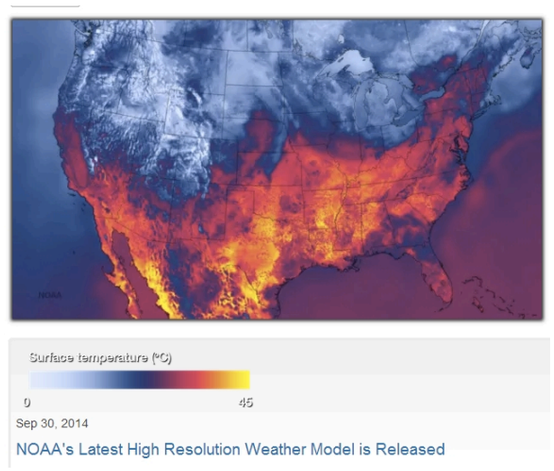
Simple charts (bar charts, line charts, scatterplots) are incredibly effective; handles most practical needs!



Designing **effective** visualization is **not hard if you learn the principles**.

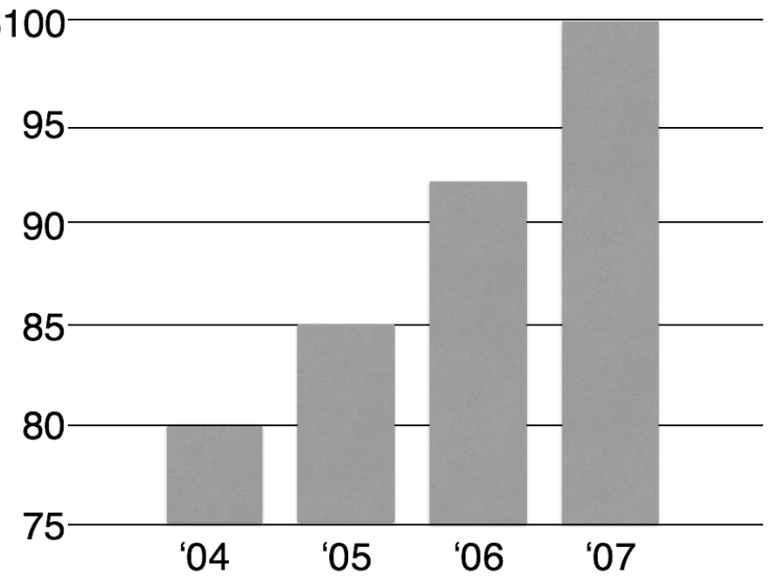
Colors (even grayscale) must be used carefully

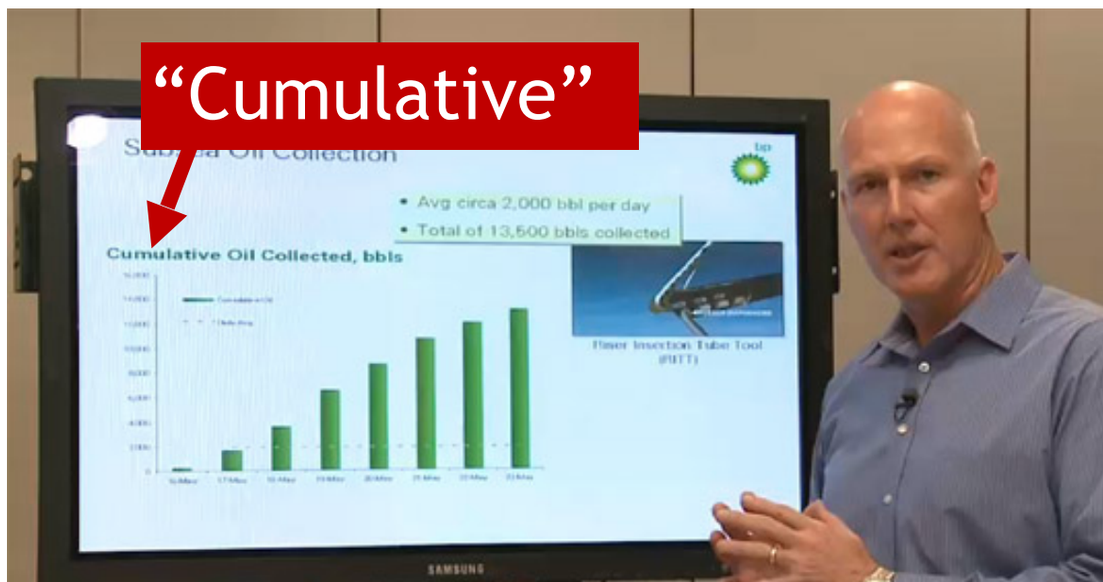




Designing **effective** visualization is **not hard if you learn the principles**.

Charts can mislead (sometimes intentionally)





Lesson 7

Learn **D3** and visualization basics

Seeing is believing.
A huge competitive edge.

Scalable interactive visualization

easier to deploy than ever before.

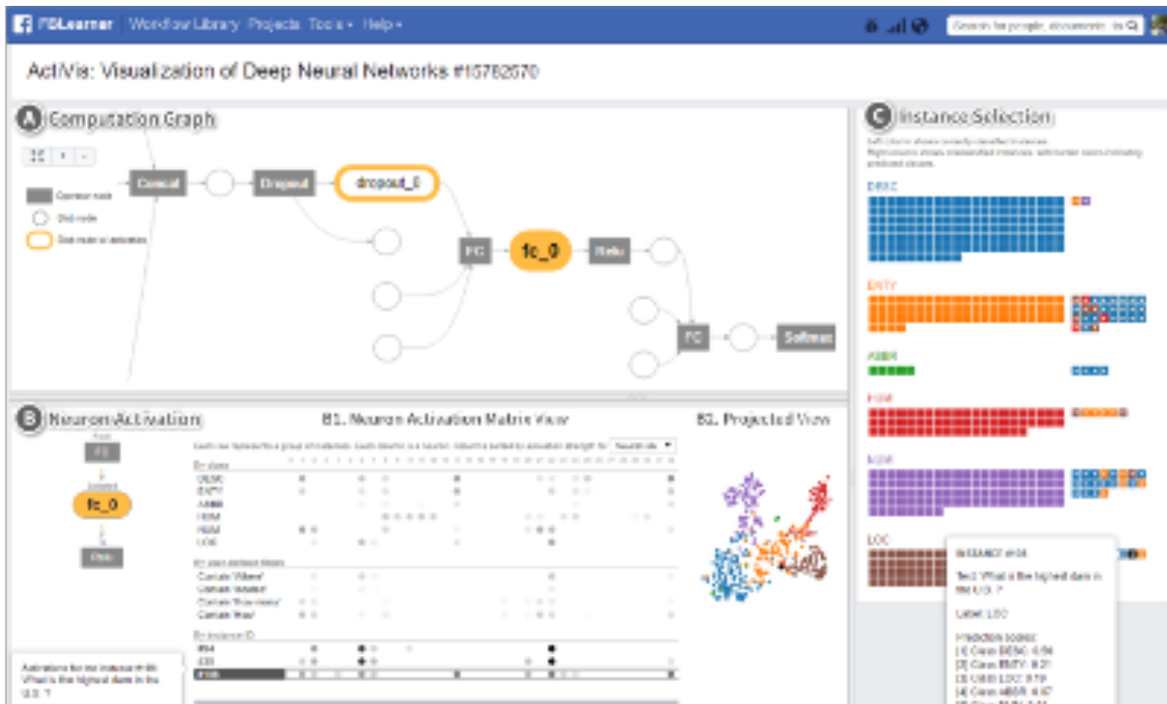
Many tools (internal + external) now run in browser.

GAN Lab (with Google)

Play with **Generated Adversarial Networks** (GAN) in browser

ActiVis (with Facebook)

Visual Exploration of Deep Neural Network Models



Scalable interactive visualization

easier to deploy than ever before.

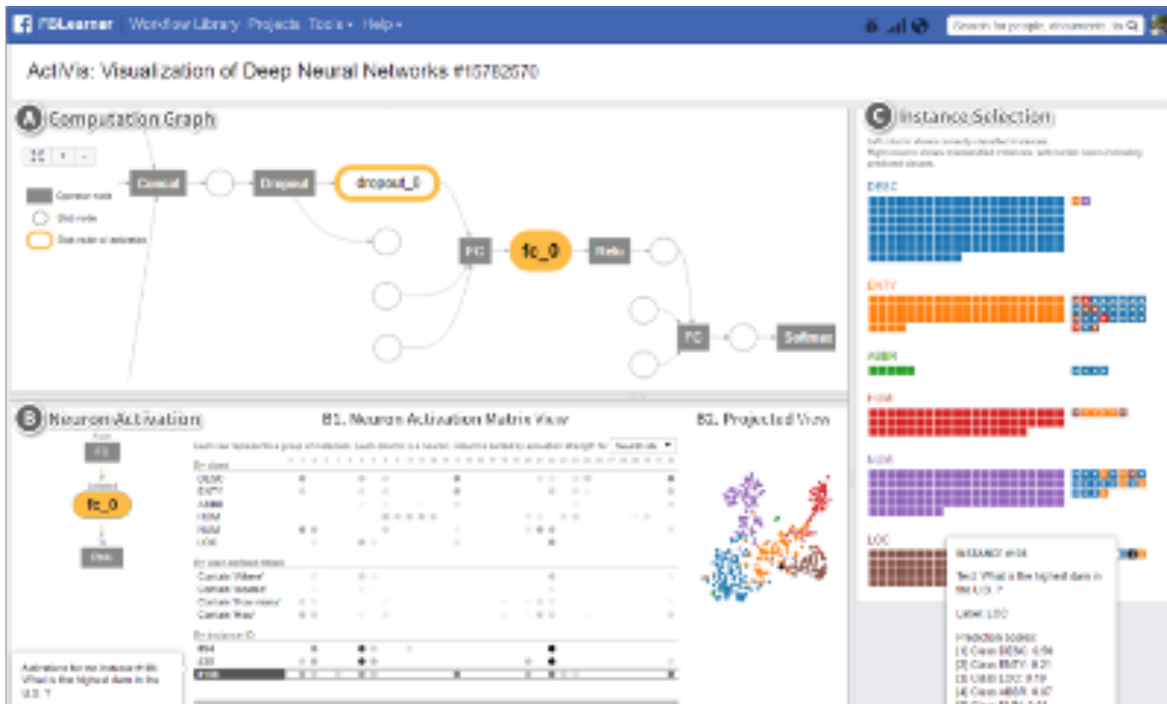
Many tools (internal + external) now run in browser.

GAN Lab (with Google)

Play with **Generated Adversarial Networks** (GAN) in browser

ActiVis (with Facebook)

Visual Exploration of Deep Neural Network Models

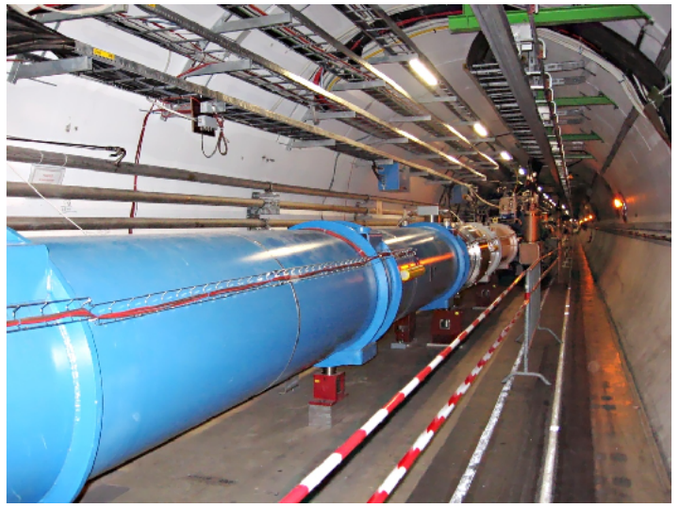


Companies expect you-all to know the “basic” **big data technologies**

(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

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)

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

Why learn Spark?

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?

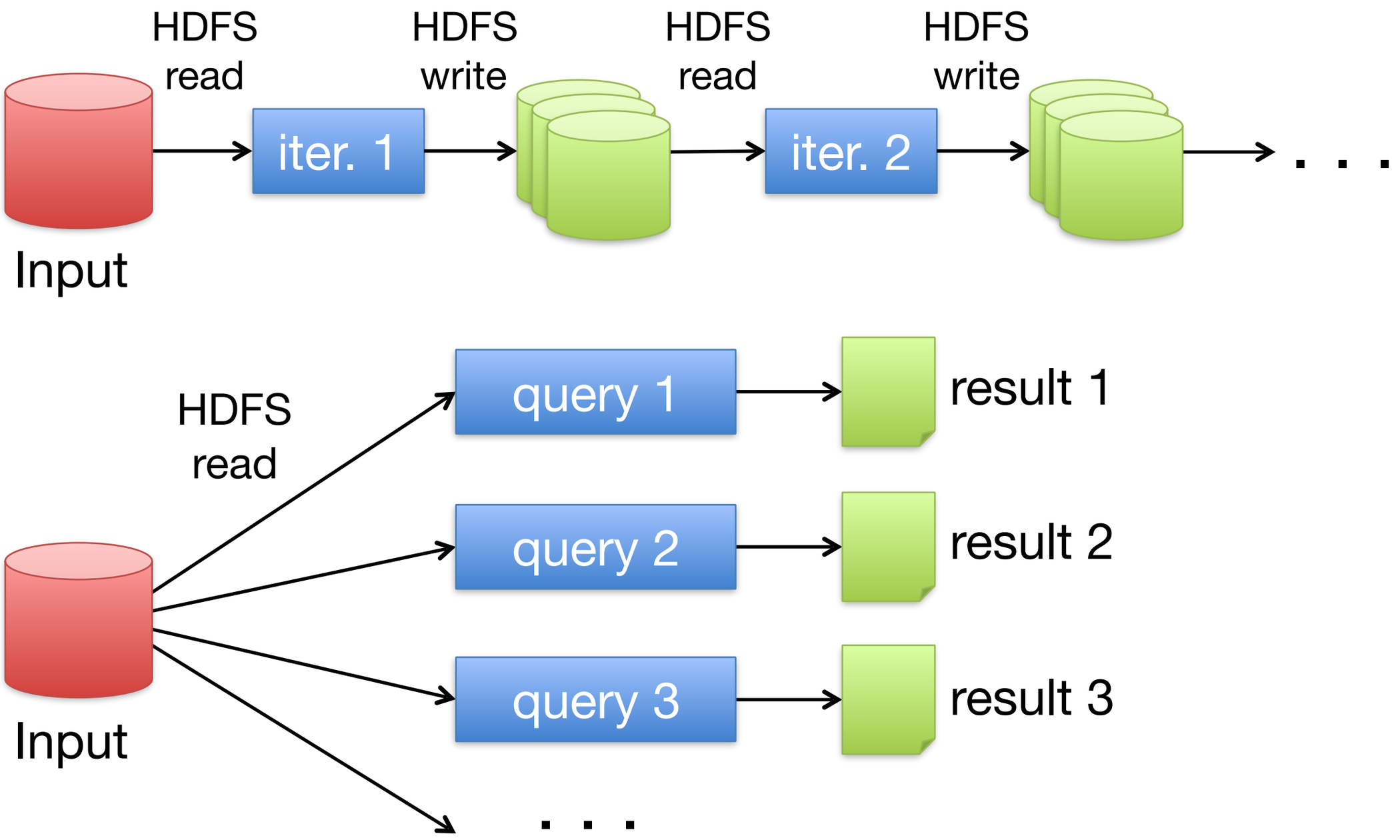
MapReduce greatly simplified big data analysis

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

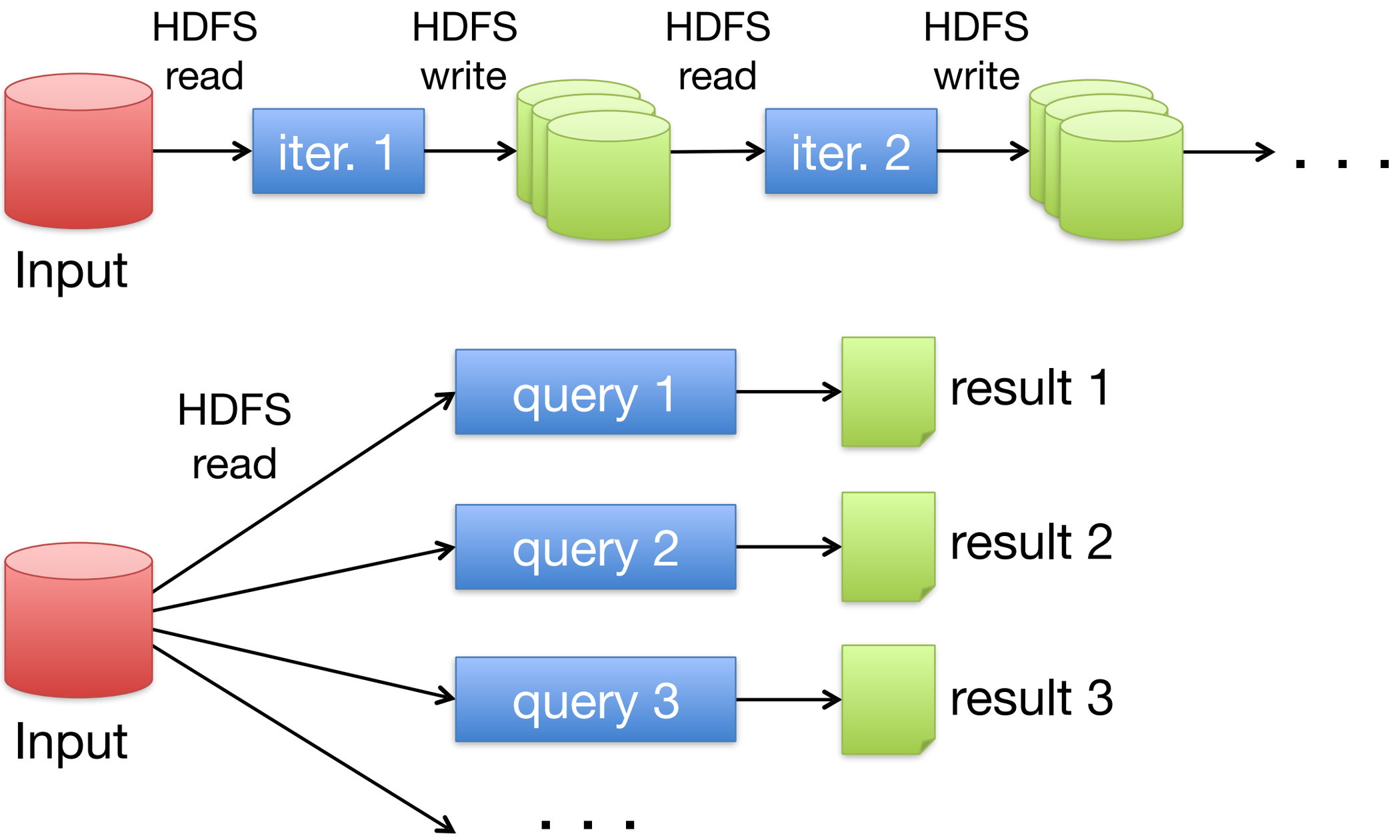
» More **complex**, multi-stage applications (e.g. iterative graph algorithms and machine learning)
» More **interactive** ad-hoc queries

Require faster **data sharing** across parallel jobs

Data Sharing in MapReduce

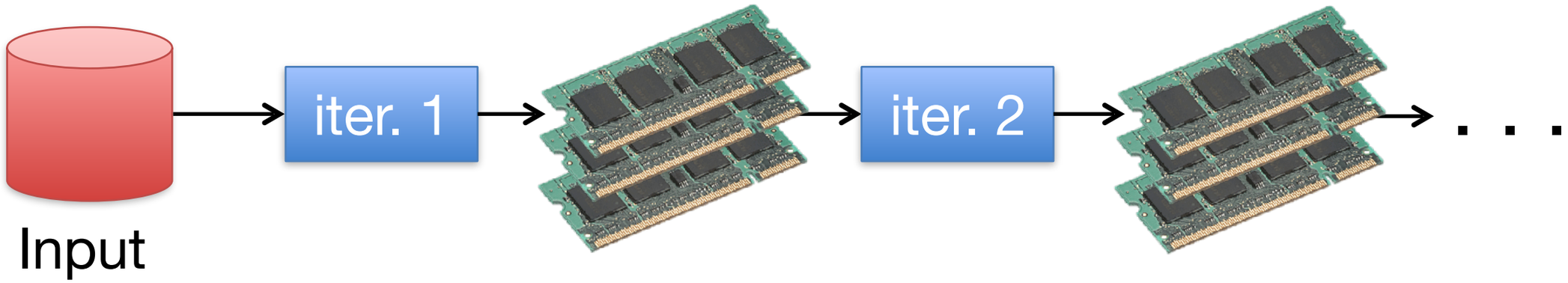


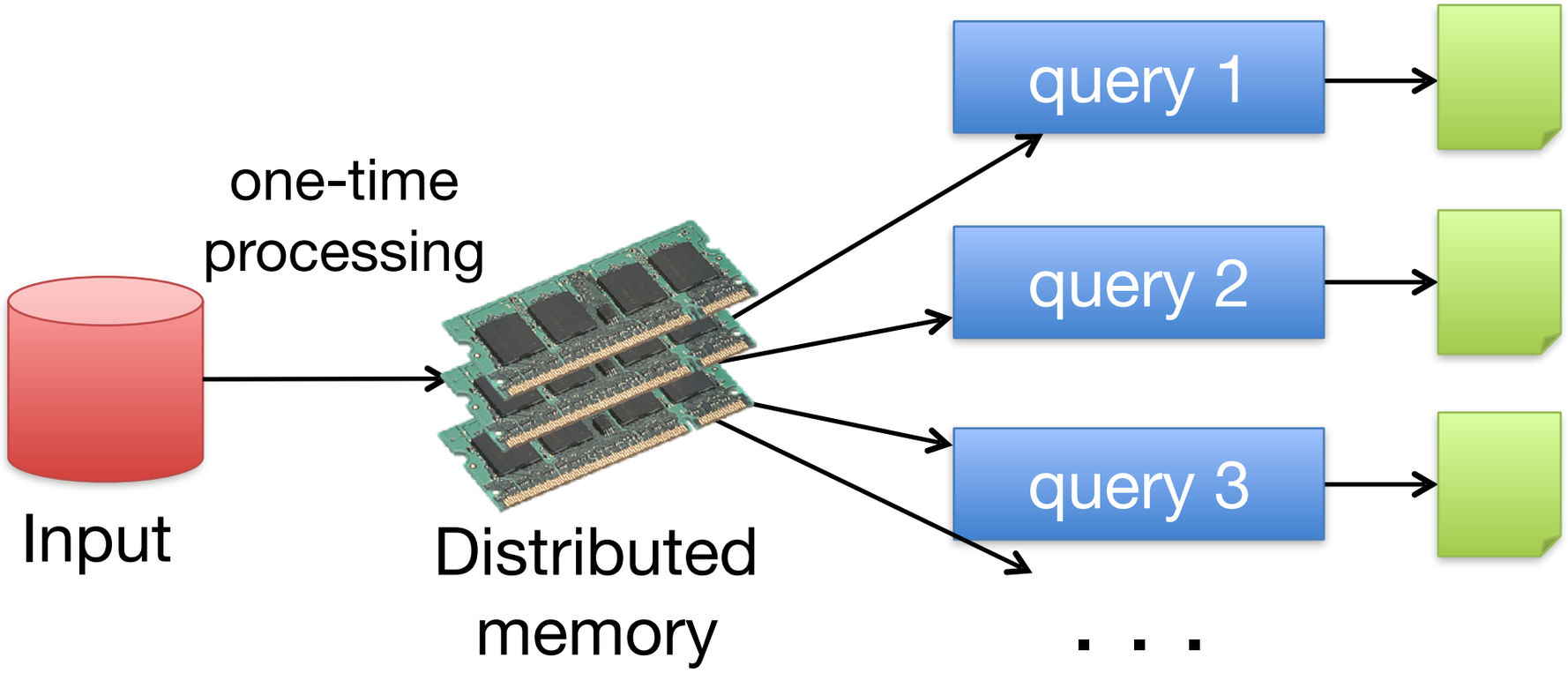
Data Sharing in MapReduce



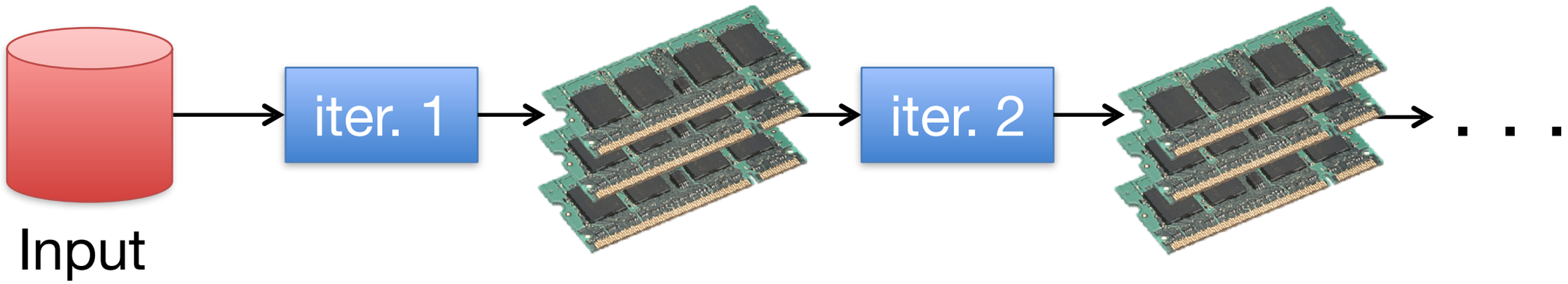
Slow due to replication, serialization, and disk IO

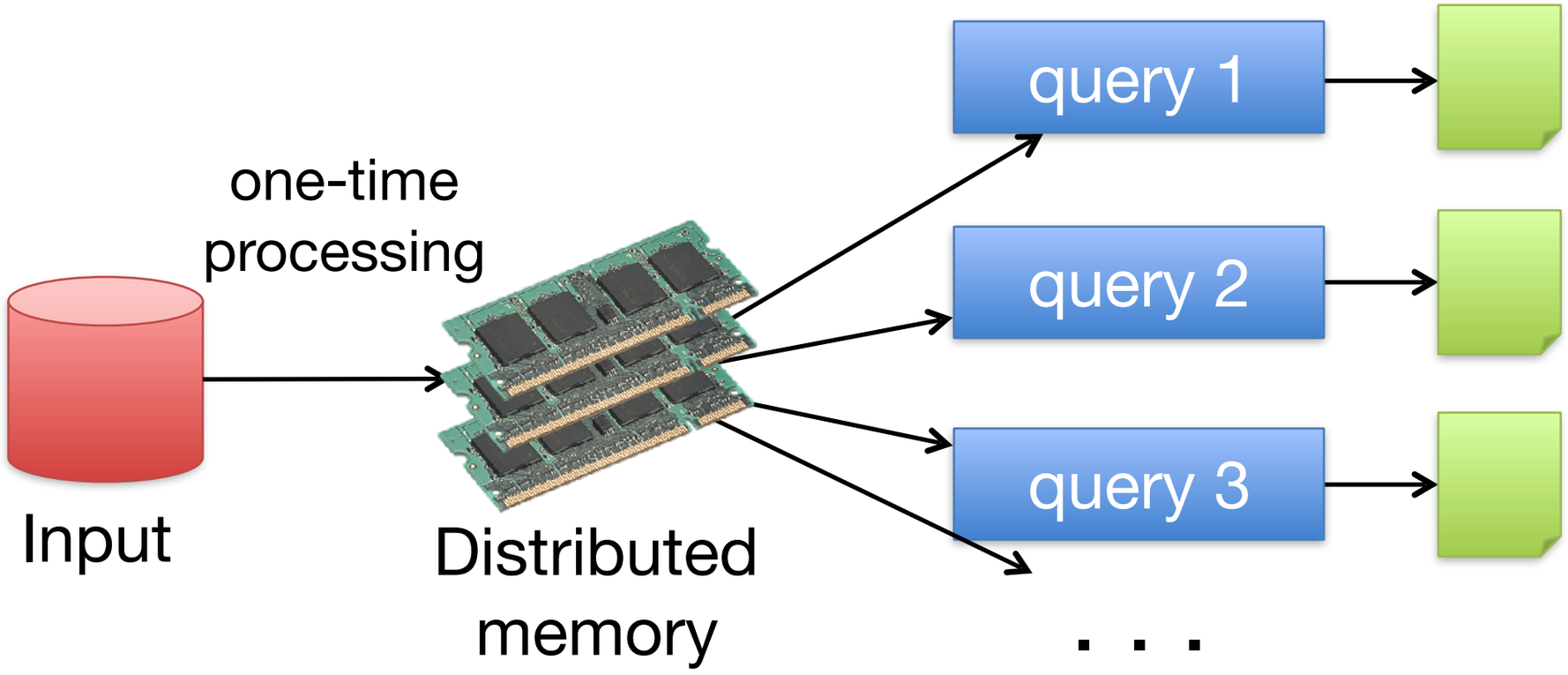
Data Sharing in Spark





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.

Industry moves fast.
So should you.

Be **cautiously optimistic**.

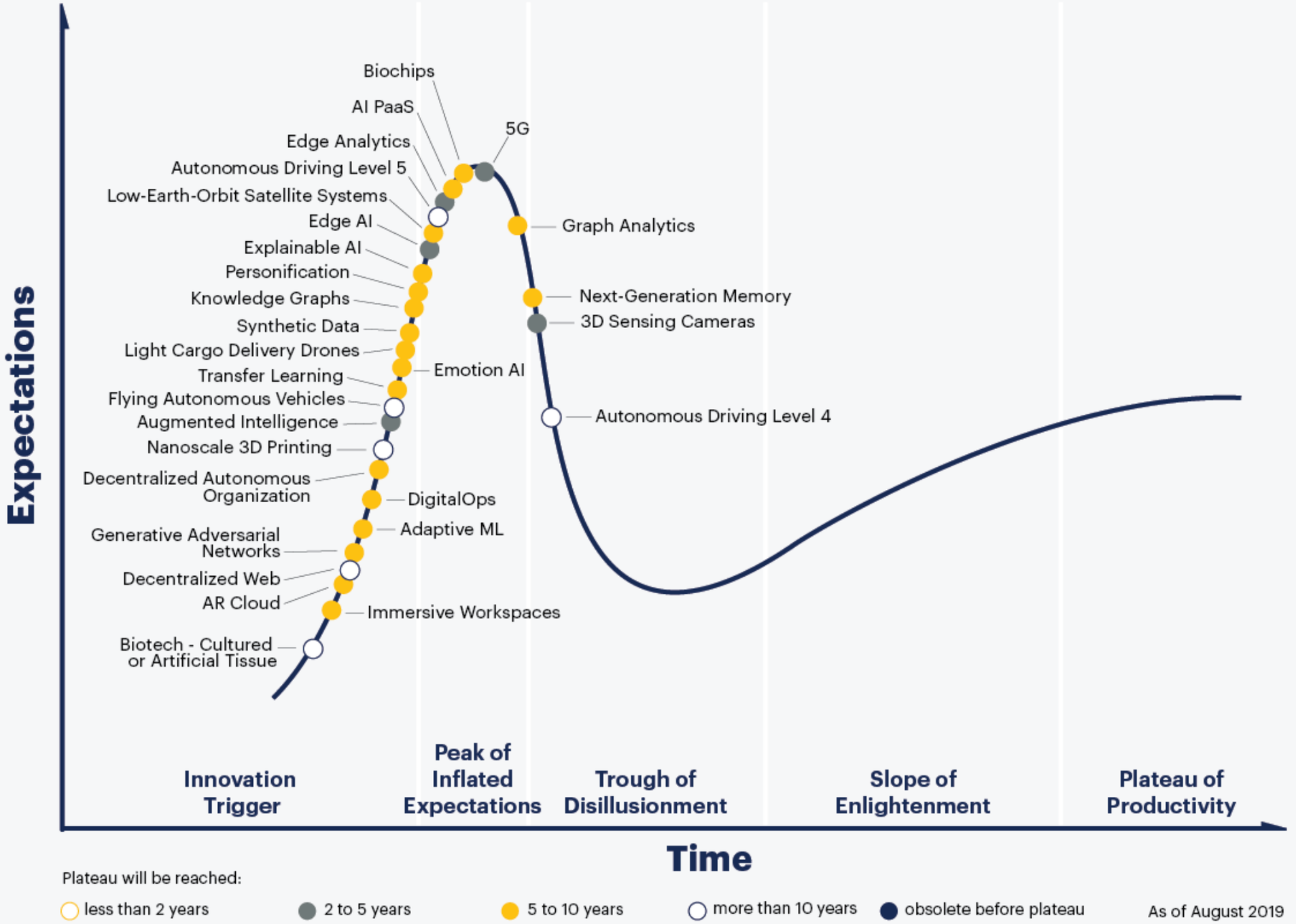
And be very careful of **hype**.

There were 2 AI winters.

https://en.wikipedia.org/wiki/History_of_artificial_intelligence

Gartner Hype Cycle for Emerging Technologies, 2019

Debatable!



gartner.com/SmarterWithGartner

Gartner

“Artificial Intelligence”

Self-Driving Taxis Hit the Streets of Singapore

by Kirsten Korosec @kirstenkorosec AUGUST 25, 2016, 4:09 AM EDT

Retrieved from: http://www.theaustralian.com.au/business/wall-street-journal/selfdriving-taxis-hit-the-road-in-singapore/news-story/73116ddc2e7c043578cb7b87d8264f5b

Google AI beats Go world champion again to complete historic 4-1 series victory

Posted Mar 15, 2016 by Jon Russell (@jonrussell)

Next Story

The battle between Google’s artificial intelligence and Go world champion Lee Sedol concluded today after the former (AlphaGo) triumphed to win the five-game series 4-1.

Retrieved from: https://techcrunch.com/2016/03/15/google-ai-beats-go-world-champion-again-to-complete-historic-4-1-series-victory/

Microsoft silences its new A.I. bot Tay, after Twitter users teach it racism [Updated]

Posted Mar 24, 2016 by Sarah Perez (@sarahintampa)

Next Story

How a Self-Driving Uber Killed a Pedestrian in Arizona

By TROY GRIGGS and DAISUKE WAKABAYASHI UPDATED MARCH 21, 2018

A woman was struck and killed on Sunday night by an autonomous car operated by Uber in Tempe, Ariz. It was believed to be the first pedestrian death associated with self-driving technology.

What We Know About the Accident

https://www.nytimes.com/interactive/2018/03/20/us/self-driving-uber-pedestrian-killed.html

https://www.tesla.com/en_GB/blog/tragic-loss?redirect=no

"Neither Autopilot nor the driver noticed the white side of the tractor trailer against a brightly lit sky, so the brake was not applied"

TESLA MODEL S MODEL

A Tragic Loss

The Tesla Team • 30 June 2016

We learned yesterday evening that NHTSA is opening a prelim performance of Autopilot during a recent fatal crash that occu the first known fatality in just over 130 million miles where Aut Among all vehicles in the US, there is a fatality every 94 millio a fatality approximately every 60 million miles. It is important t NHTSA action is simply a preliminary evaluation to determine worked according to expectations.

Following our standard practice, Tesla informed NHTSA abou

Good Read about AI:
White House Report

Preparing for The Future of Artificial Intelligence

https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

The Current State of AI

Remarkable progress has been made on what is known as **Narrow AI**, which addresses specific application areas such as playing strategic games, language translation, self-driving vehicles, and image recognition.

Narrow AI underpins many commercial services such as trip planning, shopper recommendation systems, and ad targeting, and is finding important applications in medical diagnosis, education, and scientific research. These have all had significant societal benefits and have contributed to the economic vitality of the Nation.

The Current State of AI

General AI (sometimes called Artificial General Intelligence, or AGI) refers to a notional future AI system that exhibits apparently intelligent behavior at least as advanced as a person across the full range of cognitive tasks.

A broad chasm seems to separate today's Narrow AI from the much more difficult challenge of General AI. Attempts to reach General AI by expanding Narrow AI solutions have made little headway over many decades of research. The current consensus of the private-sector expert community, with which the NSTC Committee on Technology concurs, is that **General AI will not be achieved for at least decades.**"

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

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

Course Review

CSE 6242 / CX 4242

Duen Horng (Polo) Chau
Associate Professor & ML Area Leader, College of Computing
Associate Director, MS Analytics
Georgia Tech

Twitter: @PoloChau