Visual Data Analytics

A Short Tutorial

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)

Google "Polo Chau" if interested in my professional life.

Georgia Tech | College of Computing

Bio | CV | Students | Papers | Teaching | Funding | Press | Design

POLO CHAU

Legal name:
Duen Horng Chau

Associate Director, MS in Analytics
Associate Professor, School of Computational Science & Engineering
Machine Learning Area Leader, College of Computing
Georgia Tech

Linkedin | Twitter | Google Scholar | YouTube

Admin: Carolyn Young Financial Manager: Arlene Washington
polo@gatech.edu www.cc.gatech.edu/~dchau
Office: Klaus 1324 404-385-7682

POSITIONS

May 2014 -	Associate Director MS in Analytics, Georgia Tech
Aug 2018 -	Associate Professor School of Computational Science & Engineering, Georgia Tech
Aug 2012 - Aug 2018	Assistant Professor School of Computational Science & Engineering, Georgia Tech
Dec 2012 - Dec 2015	Adjunct Assistant Professor

My research group website:

Polo Club of DATA SCIENCE

Scalable. Interactive. Interpretable.

Students (see all)

- Haekyu Park, CS PhD
- Scott Freitas, ML PhD
- Nilaksh Das, CSE PhD
- Fred Hohman, CSE PhD
- Shang-Tse Chen, CS PhD
- Minsuk (Brian) Kahng, CS PhD
- Siwei (Bob) Li, CS UG
- Ángel (Alex) Cabrera, CS UG
- Joon Kim, CS UG
- Sudeep Agarwal, CS UG
- Kristina Marotta, CS OMS
- Matthew Keezer, MS CS

Recent Alumni (see all)

Polo Club
of
DATA SCIENCE

Scalable. Interactive. Interpretable.

At Georgia Tech, we innovate **scalable, interactive, and interpretable** tools that amplify human's ability to understand and interact with billion-scale data and machine learning models. Our current research thrusts: human-centered AI (interpretable, fair, safe AI; adversarial ML); large graph visualization and mining; cybersecurity; and social good (health, energy).

At Georgia Tech, I teach

Data & Visual Analytics

Year	Semester	Course Websites		Students
2019	Spring	Campus	Online	1000
2018	Fall	Campus	Online	677
2018	Spring	Campus	Online	287
2017	Fall	Campus		273
2017	Spring	Campus		214
2016	Fall	Campus		215
2016	Spring	Campus		187
2015	Fall	Campus		146
2015	Spring	Campus		113
2014	Fall	Campus		118
2014	Spring	Campus		95
2013	Spring	Campus		35

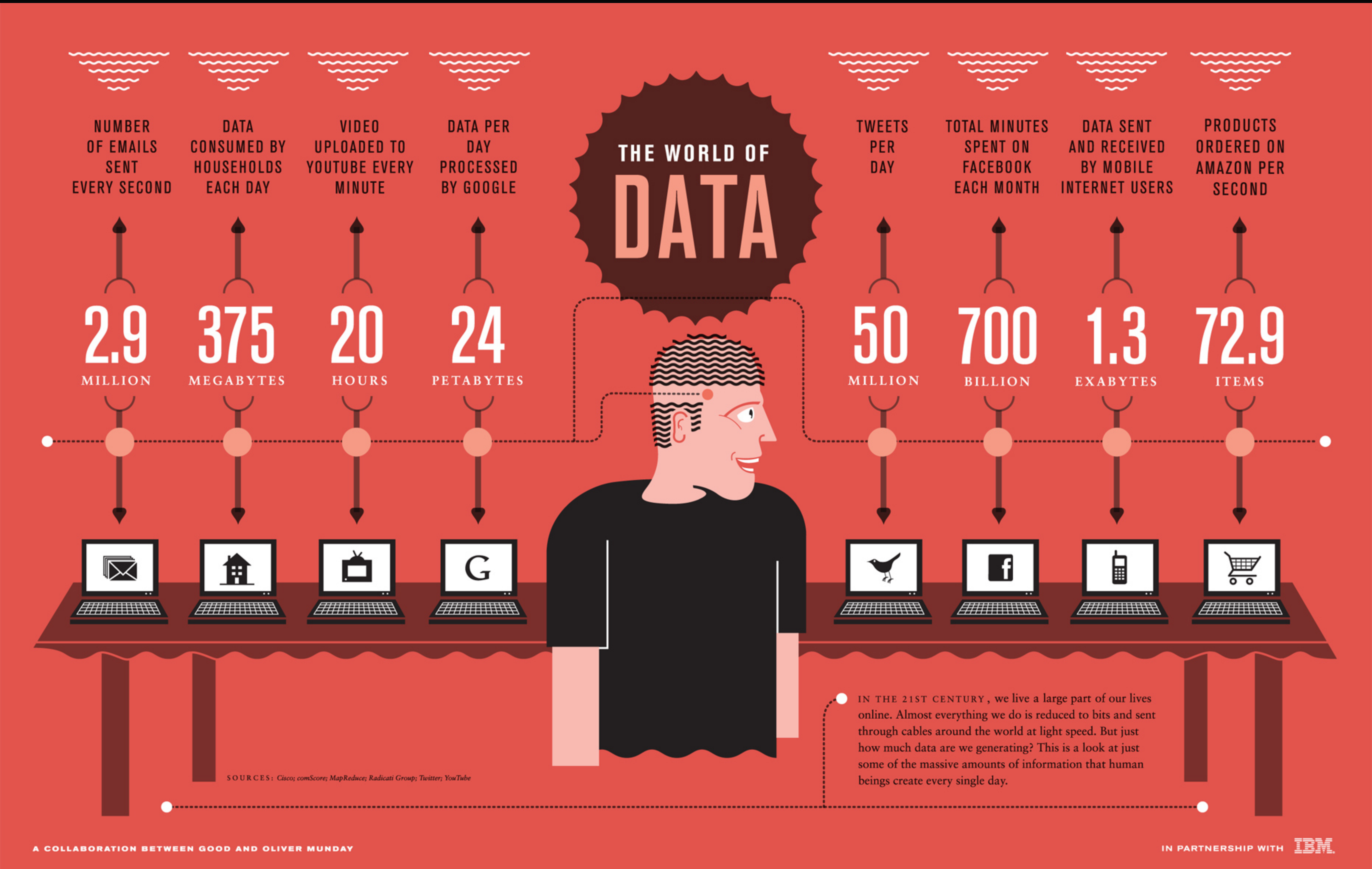
You need to learn **many things**.

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”?

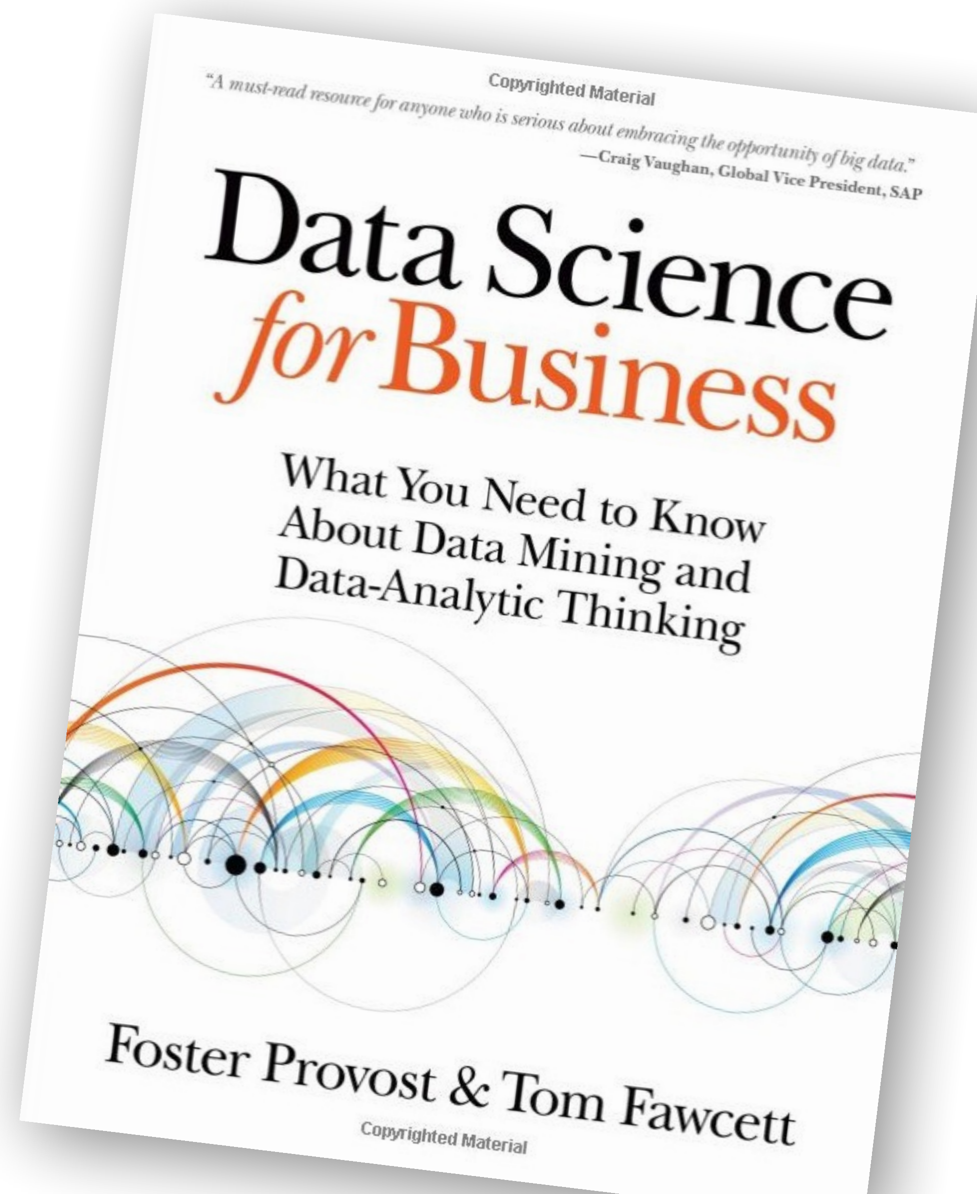
What are the "ingredients"?

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

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

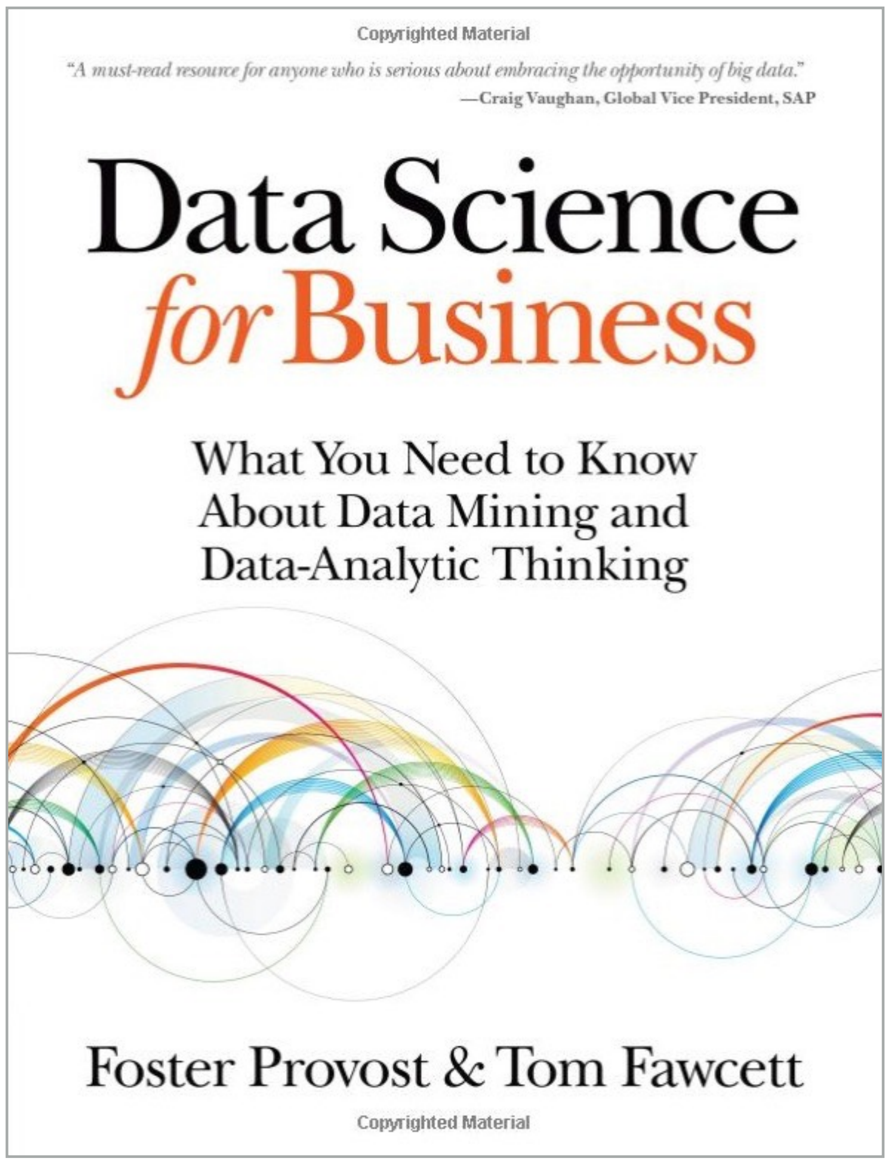
And here’s a good book.

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

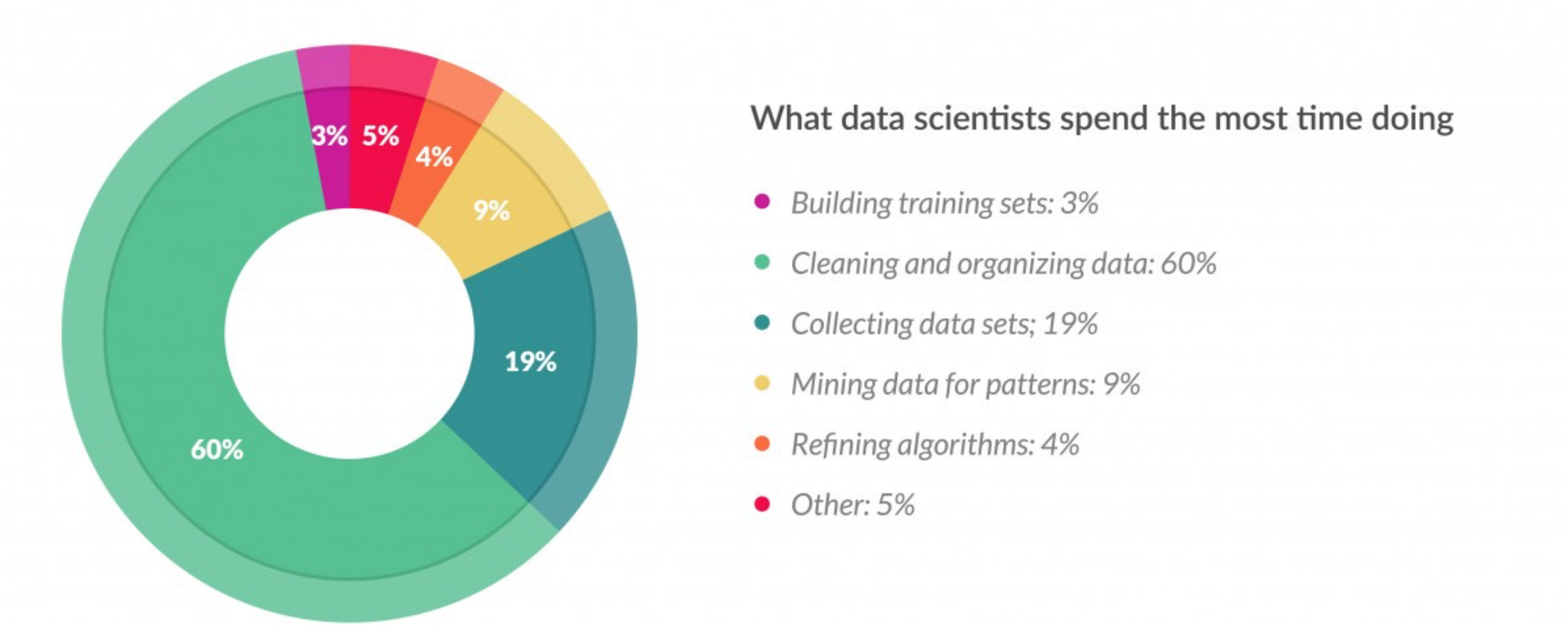
How dirty is real data?

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



We are all 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.”

OPEN Refine

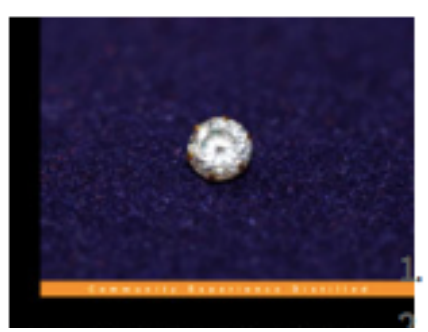
A free, open source, powerful tool for working with messy data

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)

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

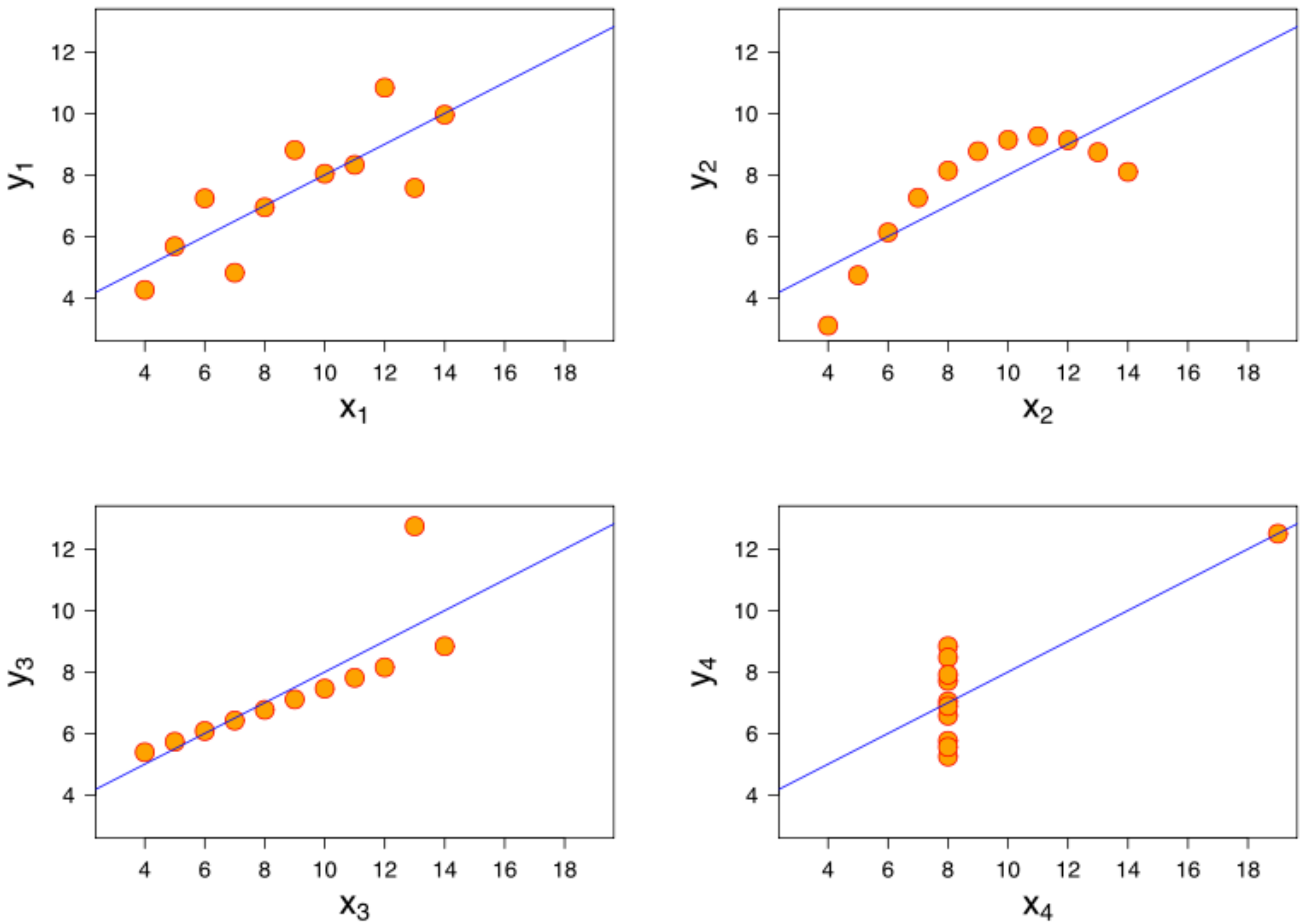
Visualization is **NOT** only about "making things look pretty"

(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

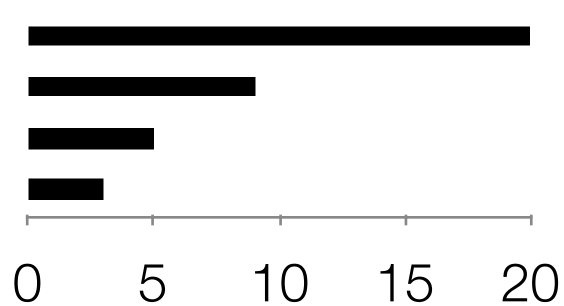


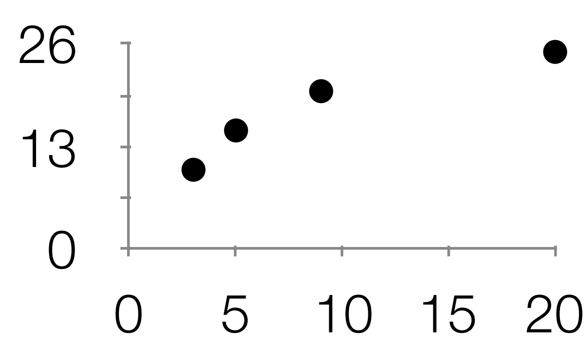
https://en.wikipedia.org/wiki/Anscombe%27s_quartet

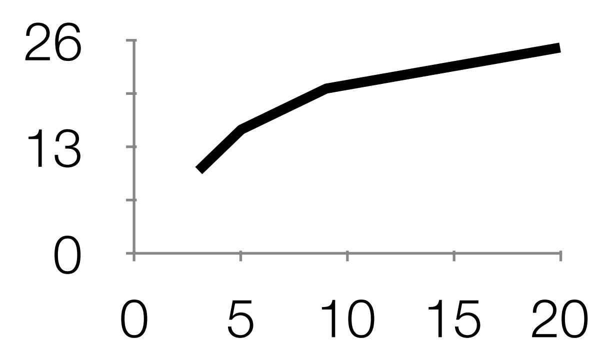
Designing **effective** visualization is **not hard if you learn the principles**.

Easy, because…

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

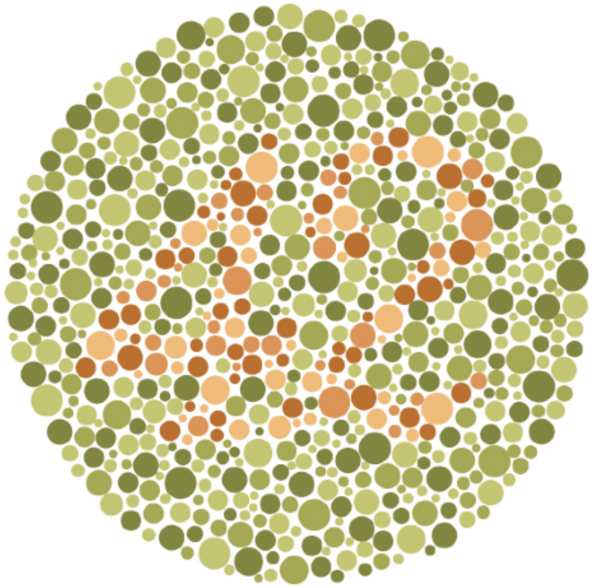


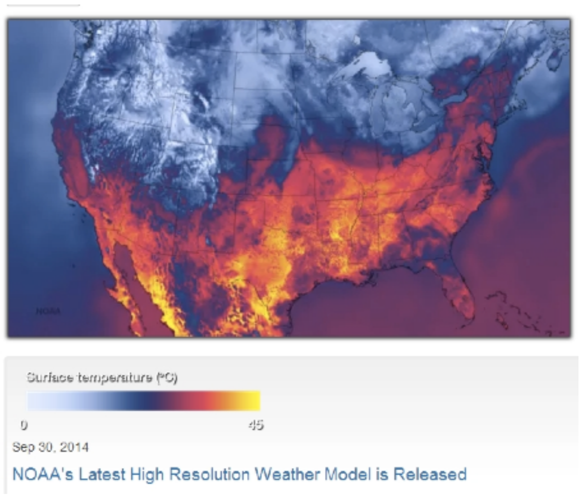




Designing **effective** visualization is **not hard if you learn the principles**.

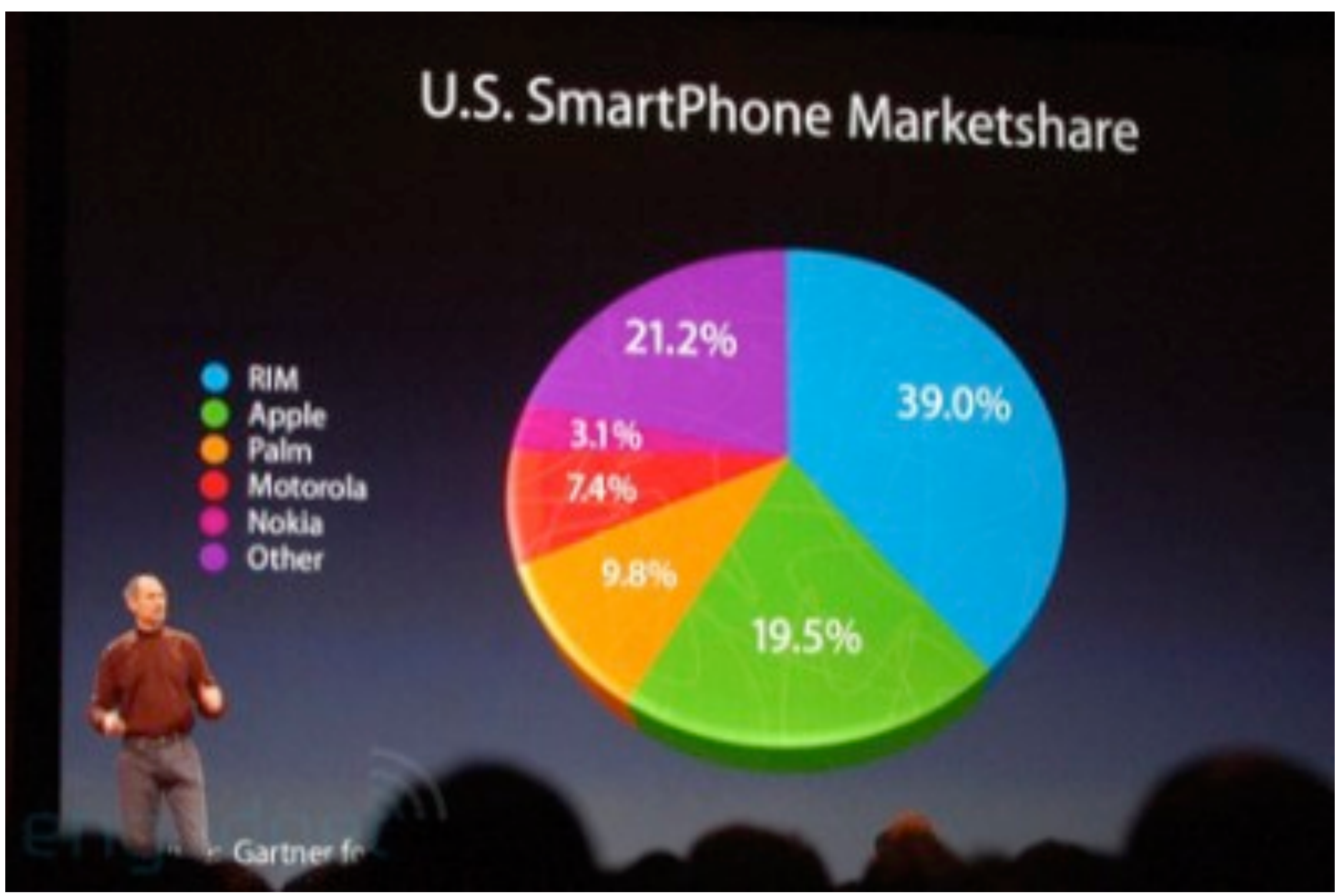
Colors (even grayscale) must be used carefully

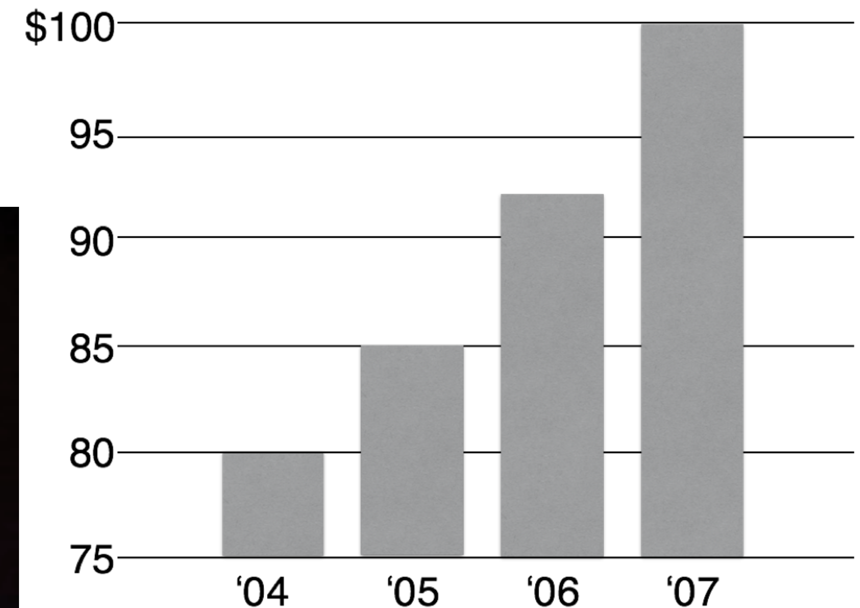


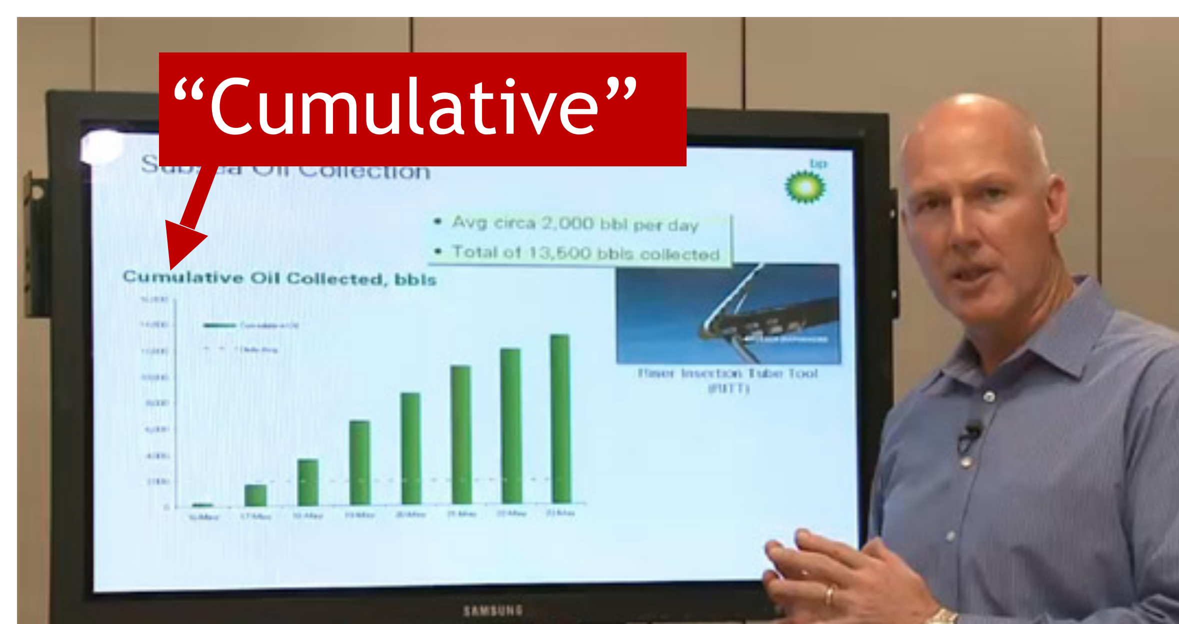


Designing **effective** visualization is **not hard if you learn the principles**.

Charts can mislead (sometimes intentionally)







Lesson 7

Learn D3 and visualization basics

Seeing is believing.
A huge competitive edge.

Lesson 8

Scalable interactive visualization

easier to deploy than ever before.

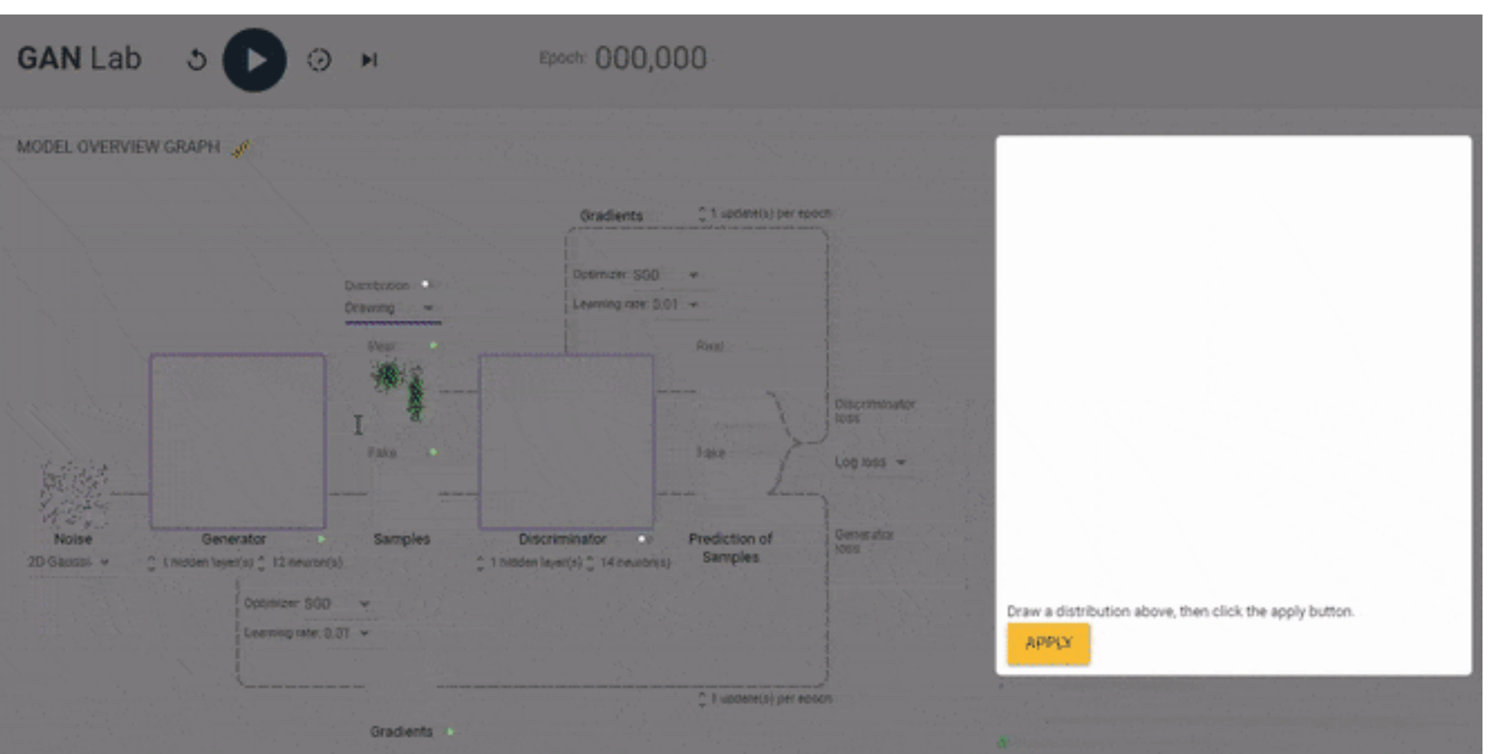
Many tools (internal + external) now run in browser.

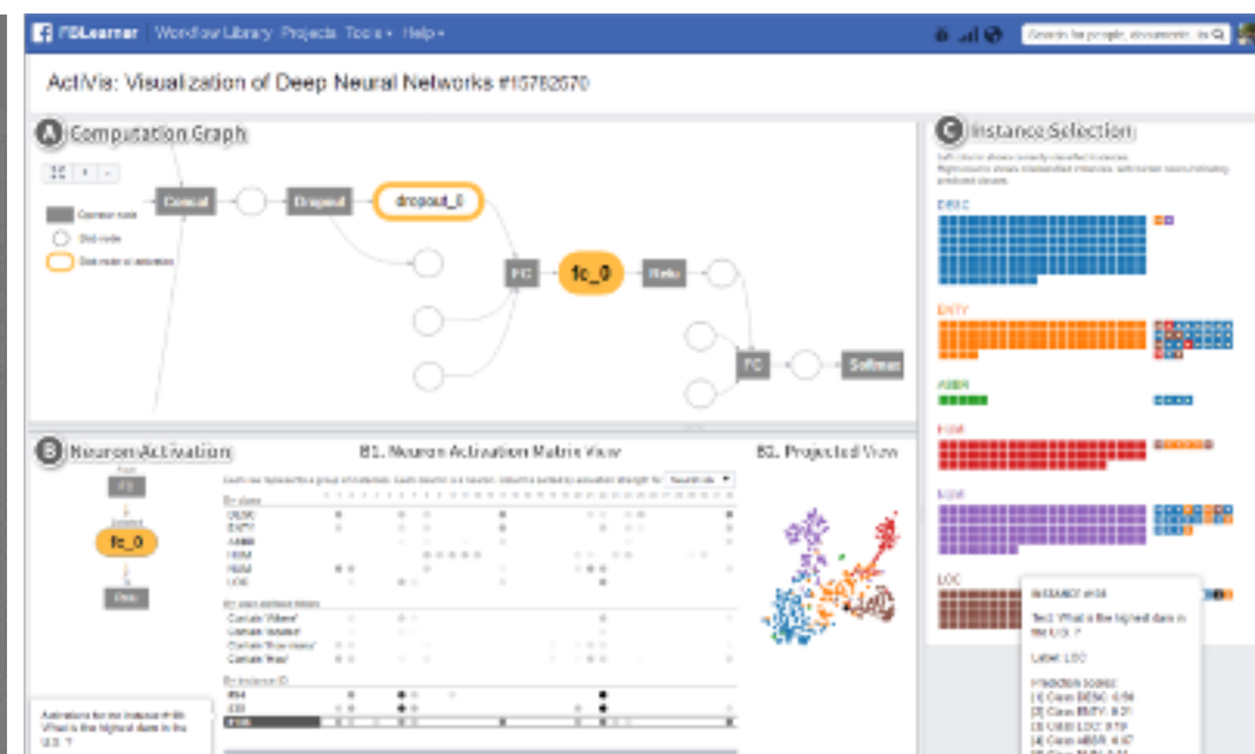
GAN Lab (with Google)

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

ActiVis (with Facebook)

Visual Exploration of Deep Neural Network Models





Lesson 8

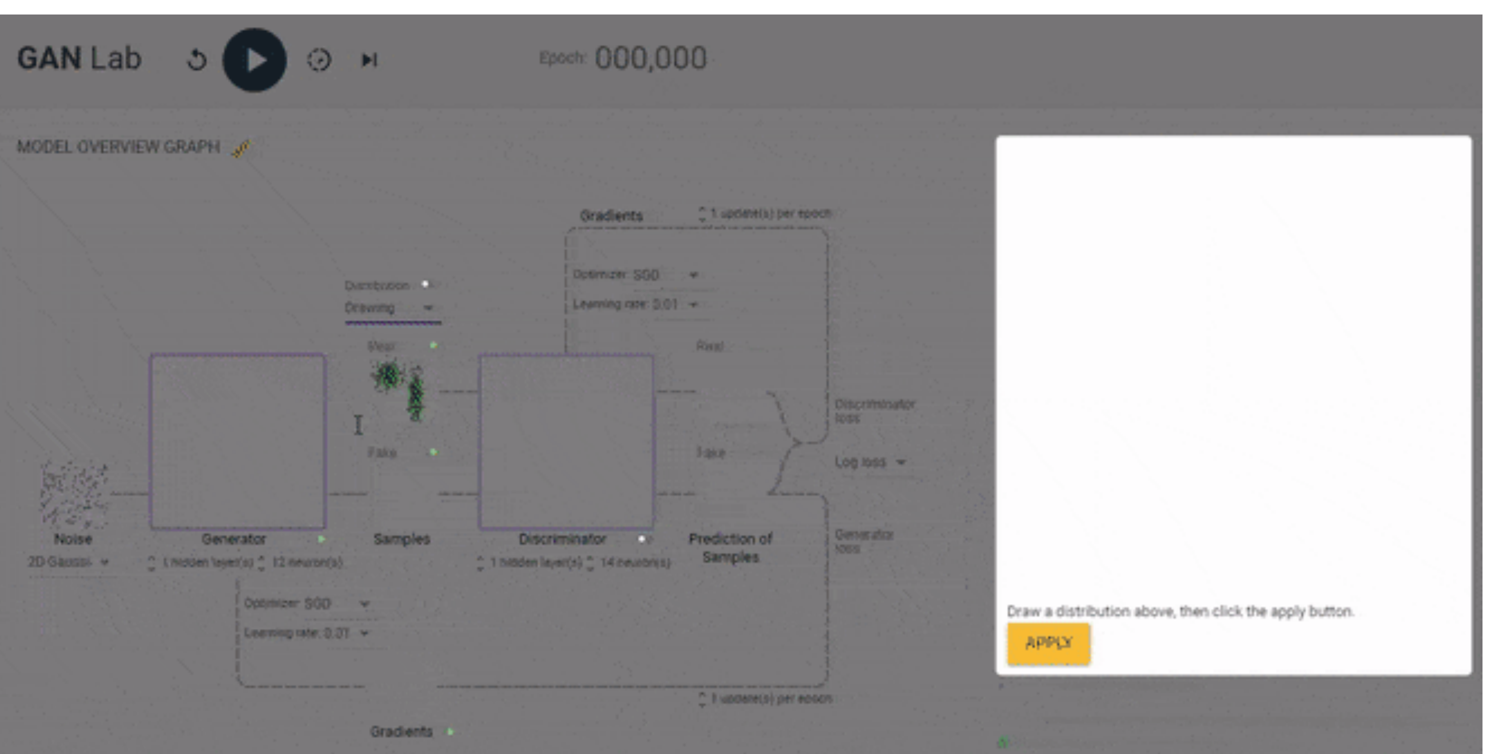
Scalable interactive visualization

easier to deploy than ever before.

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

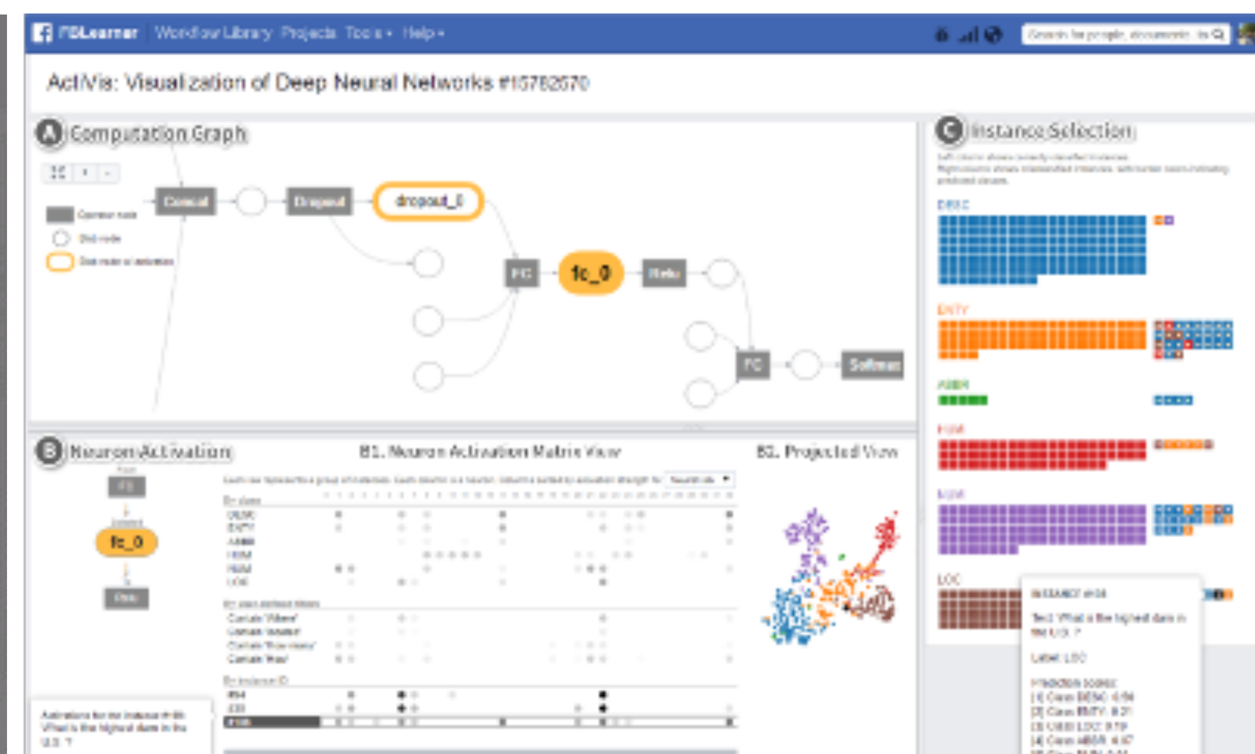
GAN Lab (with Google)

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



ActiVis (with Facebook)

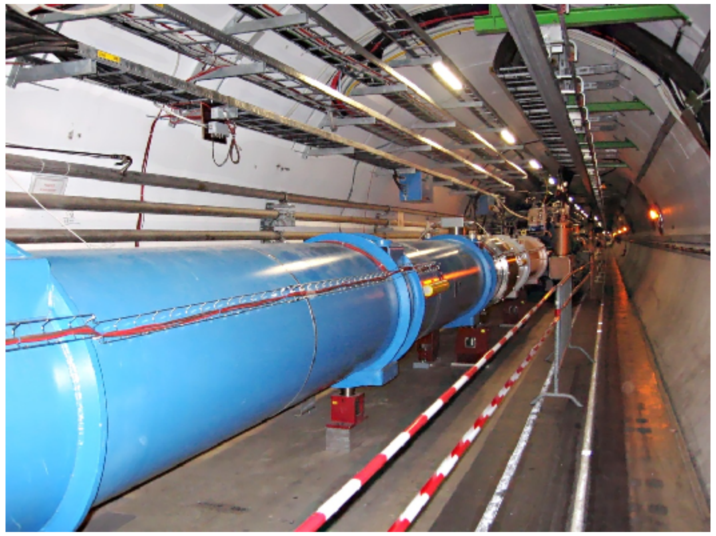
Visual Exploration of Deep Neural Network Models



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)

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

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?

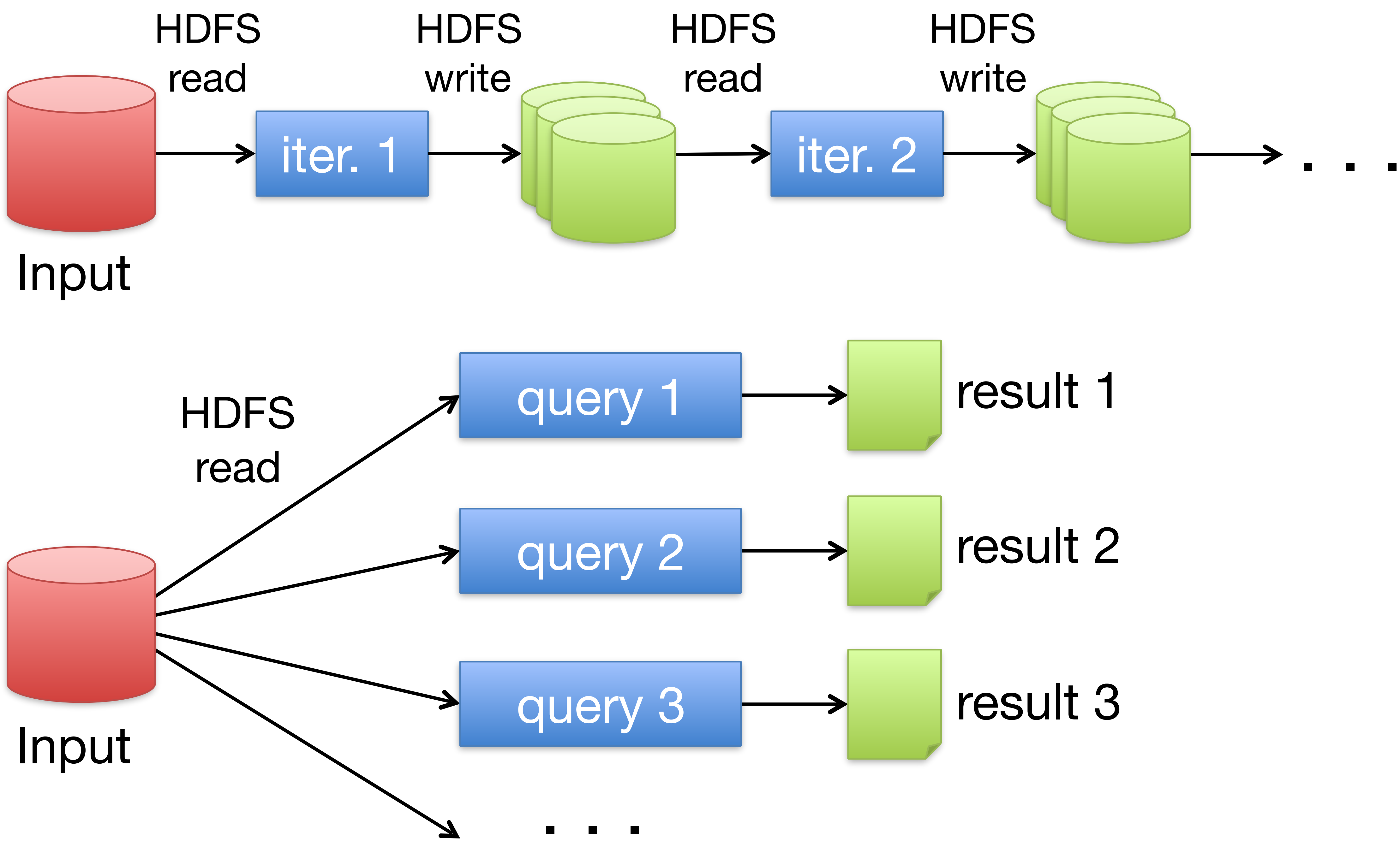
MapReduce greatly simplified big data analysis

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

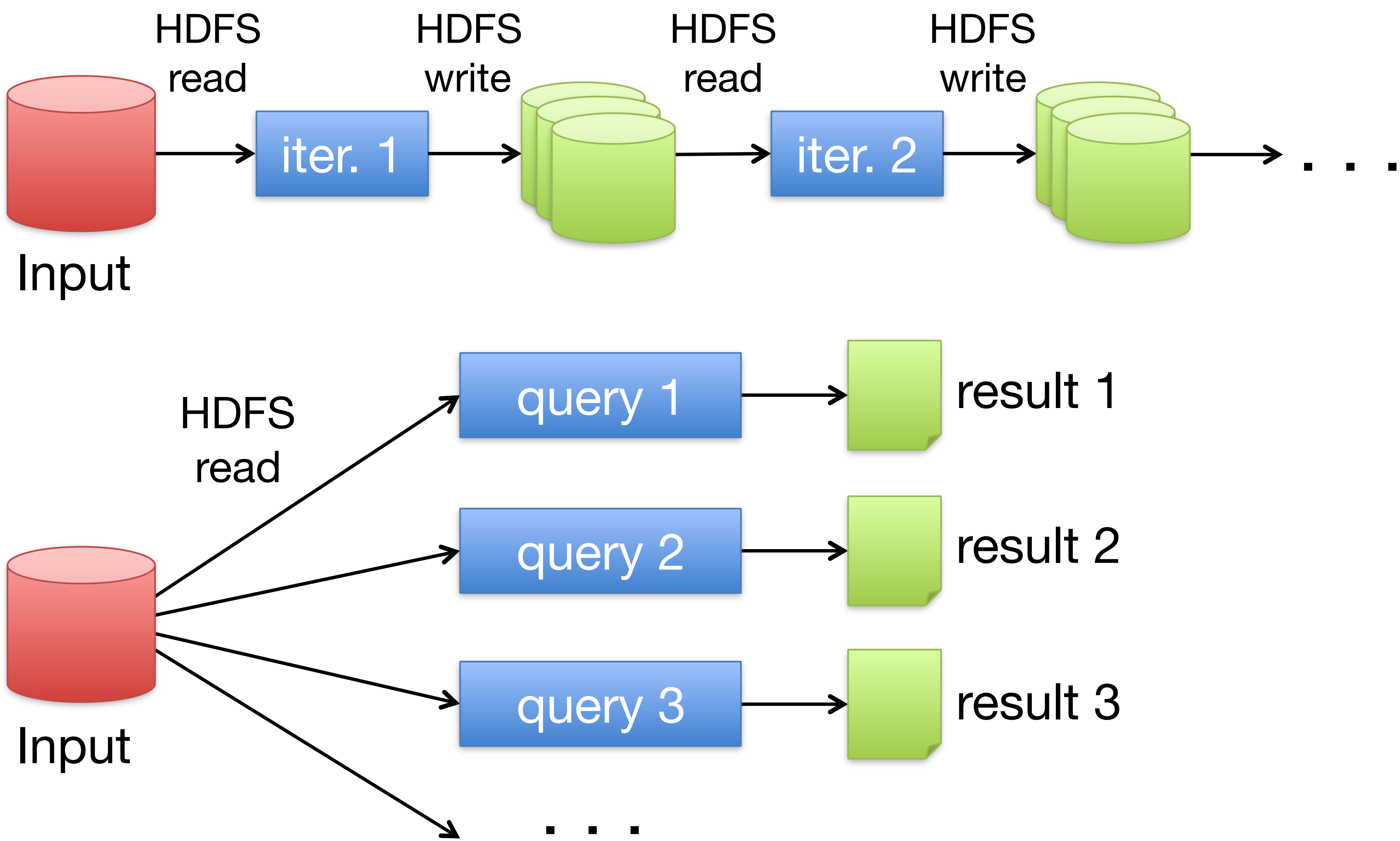
- More **complex**, multi-stage applications (e.g. iterative graph algorithms and machine learning)
- More **interactive** ad-hoc queries

Require faster **data sharing** across parallel jobs

Data Sharing in MapReduce

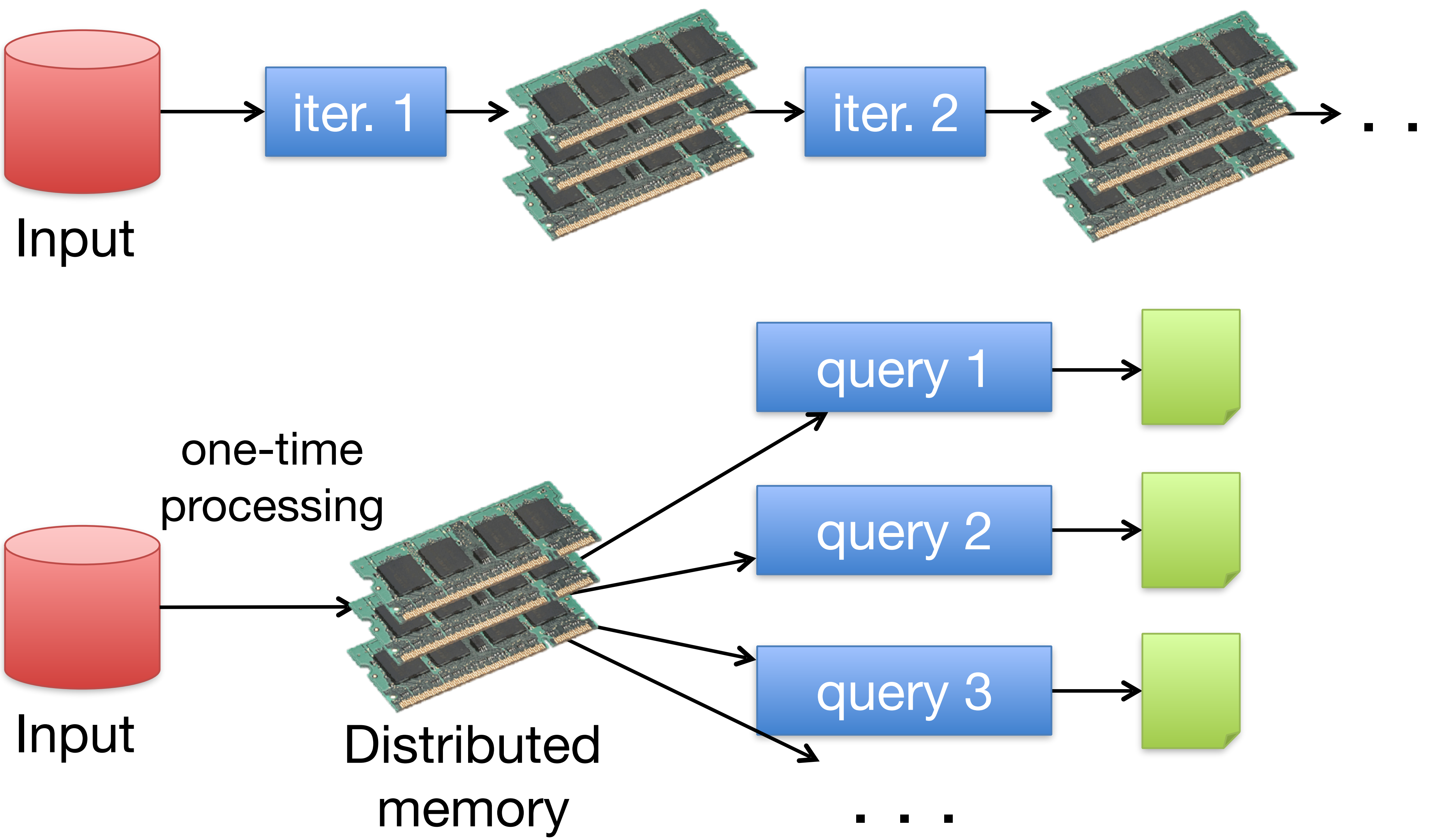


Data Sharing in MapReduce

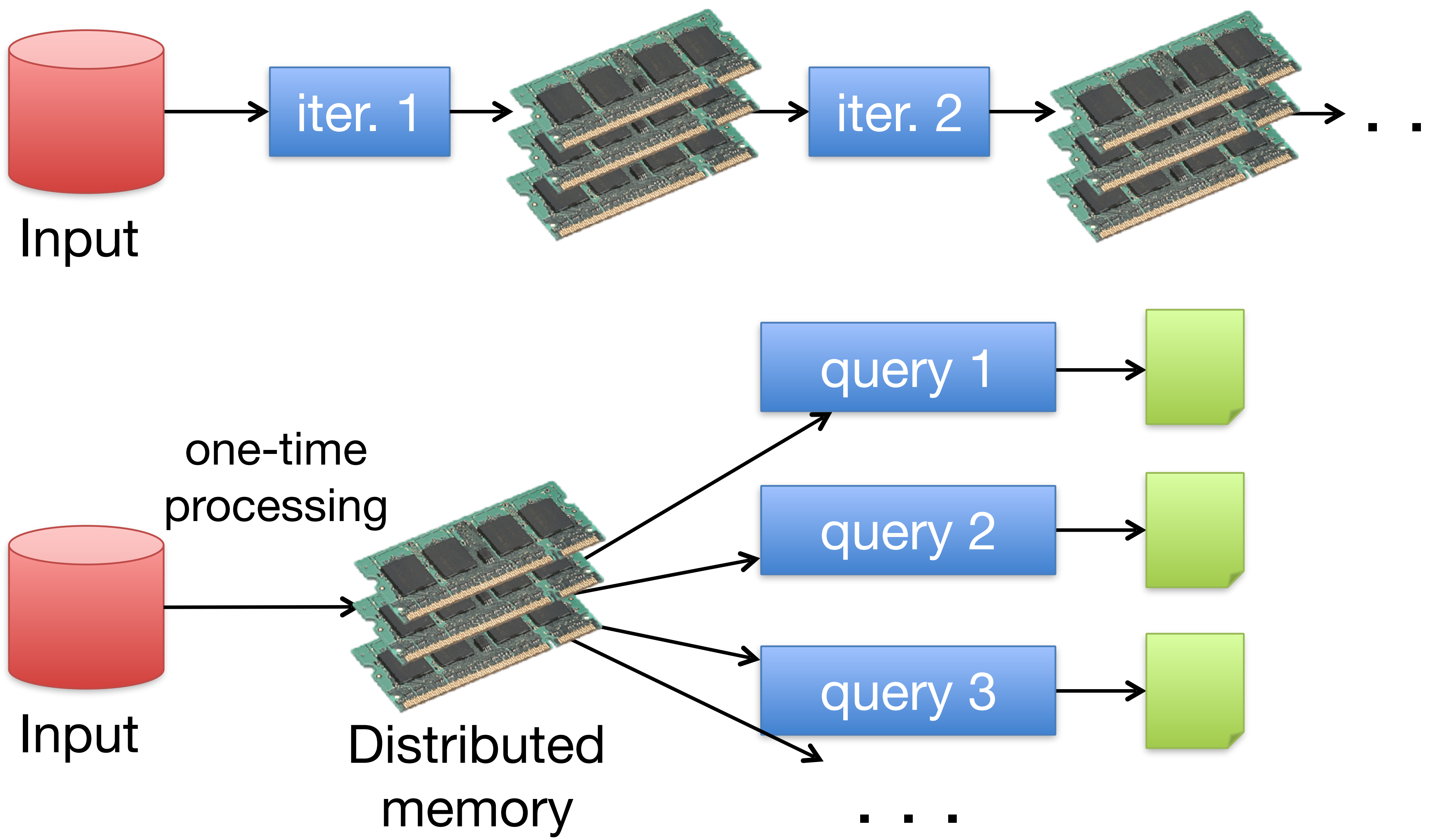


Slow due to replication, serialization, and disk IO

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.
Flume is a layer that runs on top of MapReduce that abstracts away the complexity into something that is much easier

Lesson 10

Industry moves fast.
So should you.

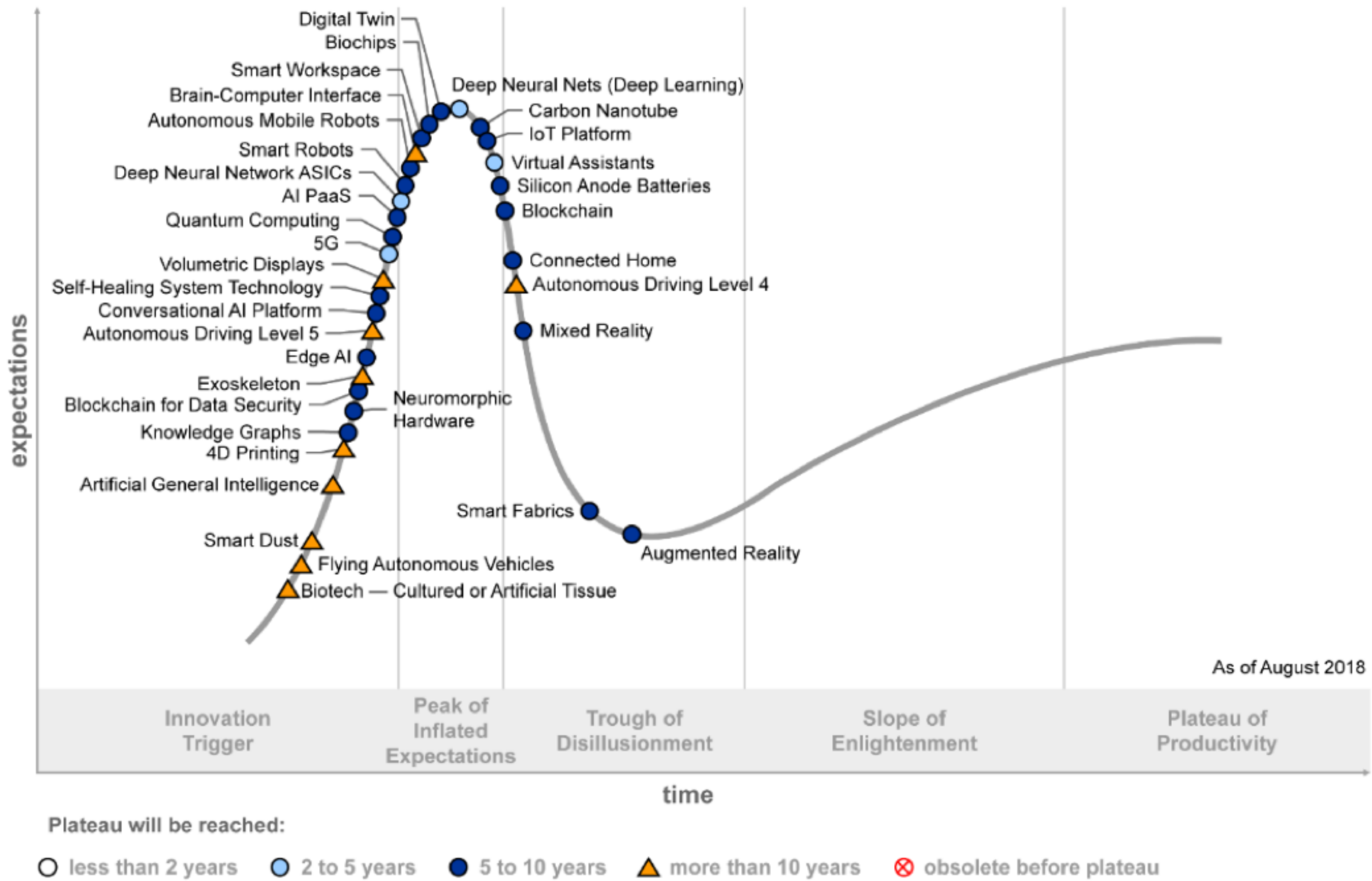
Be **cautiously optimistic**.
And be very careful of **hype**.

There were 2 AI winters.

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

Gartner Hype Cycle for Emerging Technologies

Debatable!



Source: https://www.gartner.com/en/newsroom/press-releases/2018-08-20-gartner-identifies-five-emerging-technology-trends-that-will-blur-the-lines-between-human-and-machine

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

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

Visual Data Analytics

A Short Tutorial

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

Twitter: @PoloChau