Getting Started With Data Science & Machine Learning

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

Twitter: @PoloChau

Alternative Title

(Some of the)

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

Polo Club of Data Science | Bio | CV | Students | Publications | Teaching | Funding | Press | Fun

Duen Horng (Polo) Chau

Associate Professor, School of Computational Science & Engineering
Associate Director, MS in Analytics
Director of Industry Relations, The Institute for Data Engineering and Science
Associate Director of Corporate Relations, The Center for Machine Learning
Machine Learning Area Leader, College of Computing
Georgia Tech

Linkedin   Twitter   Google Scholar   YouTube

Admin: Carolyn Young   Financial Manager: Arlene Washington
Please read our FAQ if you are interested joining my group.
polo@gatech.edu   www.cc.gatech.edu/~dchau
Office: CODA 1321   404-385-7682

POSITIONS

Oct 2019 - Director of Industry Relations
Institute for Data Engineering and Science, Georgia Tech

Oct 2019 - Associate Director of Corporate Relations for Machine Learning
The Center for Machine Learning , Georgia Tech

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

My research group website:

Polo Club of DATA SCIENCE

Scalable. Interactive. Interpretable.

Students (see all)

Rahul Duggal, CS PhD
Austin Wright, ML PhD
Zijie (Jay) Wang, ML PhD
Haekyu Park, CS PhD
Scott Freitas, ML PhD
Nilaksh Das, CSE PhD
Fred Hohman, CSE PhD
Jonathan Leo, CS UG
Rob Firstman, CS UG
Omar Shaikh, CS UG
Jon Saad-Falcon, CS UG
Robert Turko, CS UG
Zhiyan (Frank) Zhou, CS UG
Anish Upadhayay, CS UG
Megan Dass, CS UG
Alex Yang, CS UG
Kevin Li, CS UG

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
2020	Fall	Campus 6242 & 4242	Online 6242		1277
2020	Spring	Campus 6242	Campus 4242	Online 6242	966
2019	Fall	Campus 6242	Campus 4242	Online 6242	877
2019	Spring	Campus 6242 & 4242	Online 6242		1000
2018	Fall	Campus 6242 & 4242	Online 6242		677
2018	Spring	Campus 6242 & 4242	Online 6242		287
2017	Fall	Campus 6242 & 4242			273
2017	Spring	Campus 6242 & 4242			214
2016	Fall	Campus 6242 & 4242			215
2016	Spring	Campus 6242 & 4242			187
2015	Fall	Campus 6242 & 4242			146
2015	Spring	Campus 6242 & 4242			113
2014	Fall	Campus 6242 & 4242			118
2014	Spring	Campus 6242 & 4242			95
2013	Spring	Campus 6242 & 4242			35

You (likely) need to learn **many things**.

Why? Complexity of datasets and problems.

What are the “ingredients”?

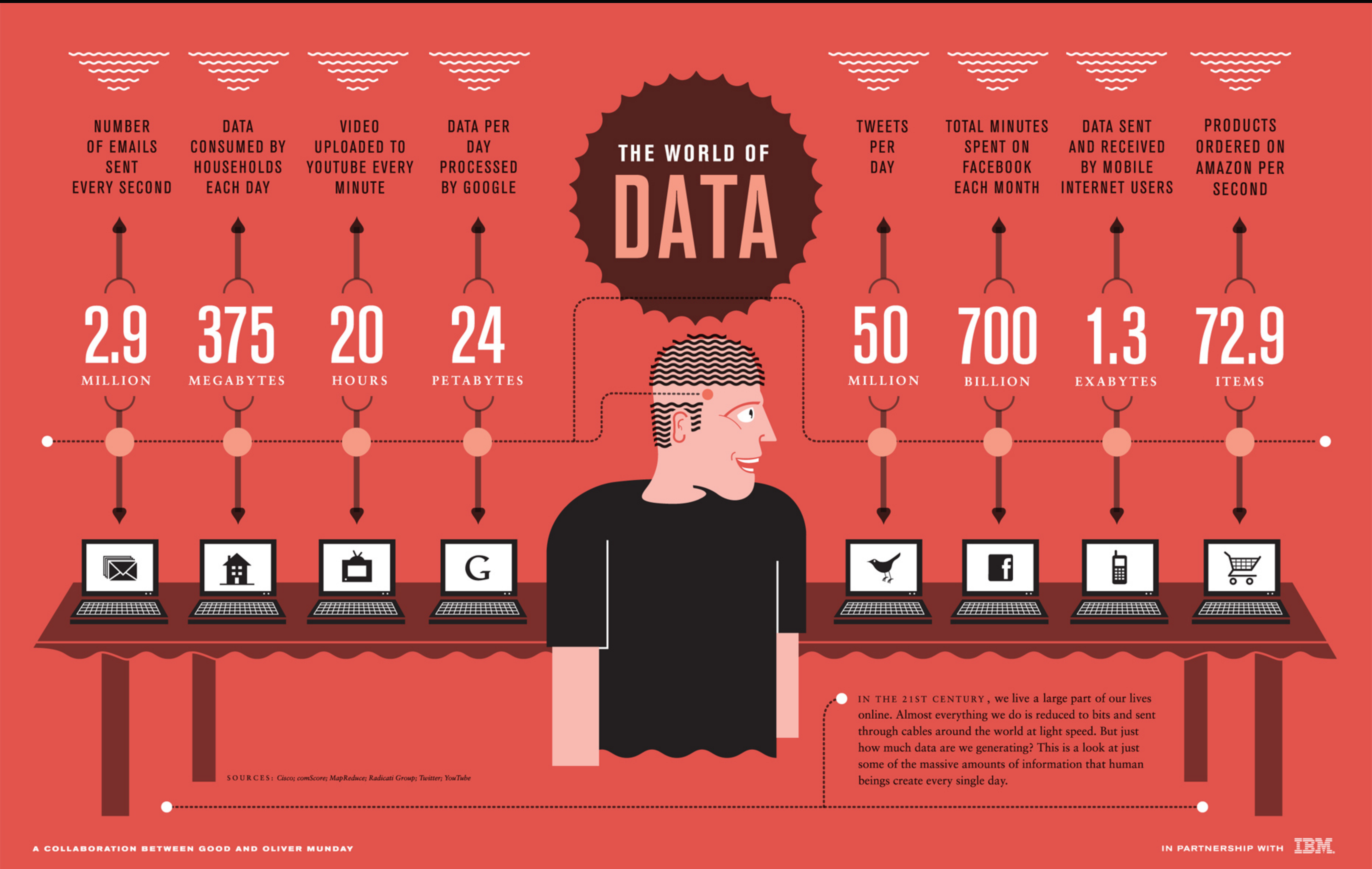
Need to think (a lot) about: storage, complex system design, scalability of algorithms, visualization techniques, interaction techniques, statistical tests, etc.

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.

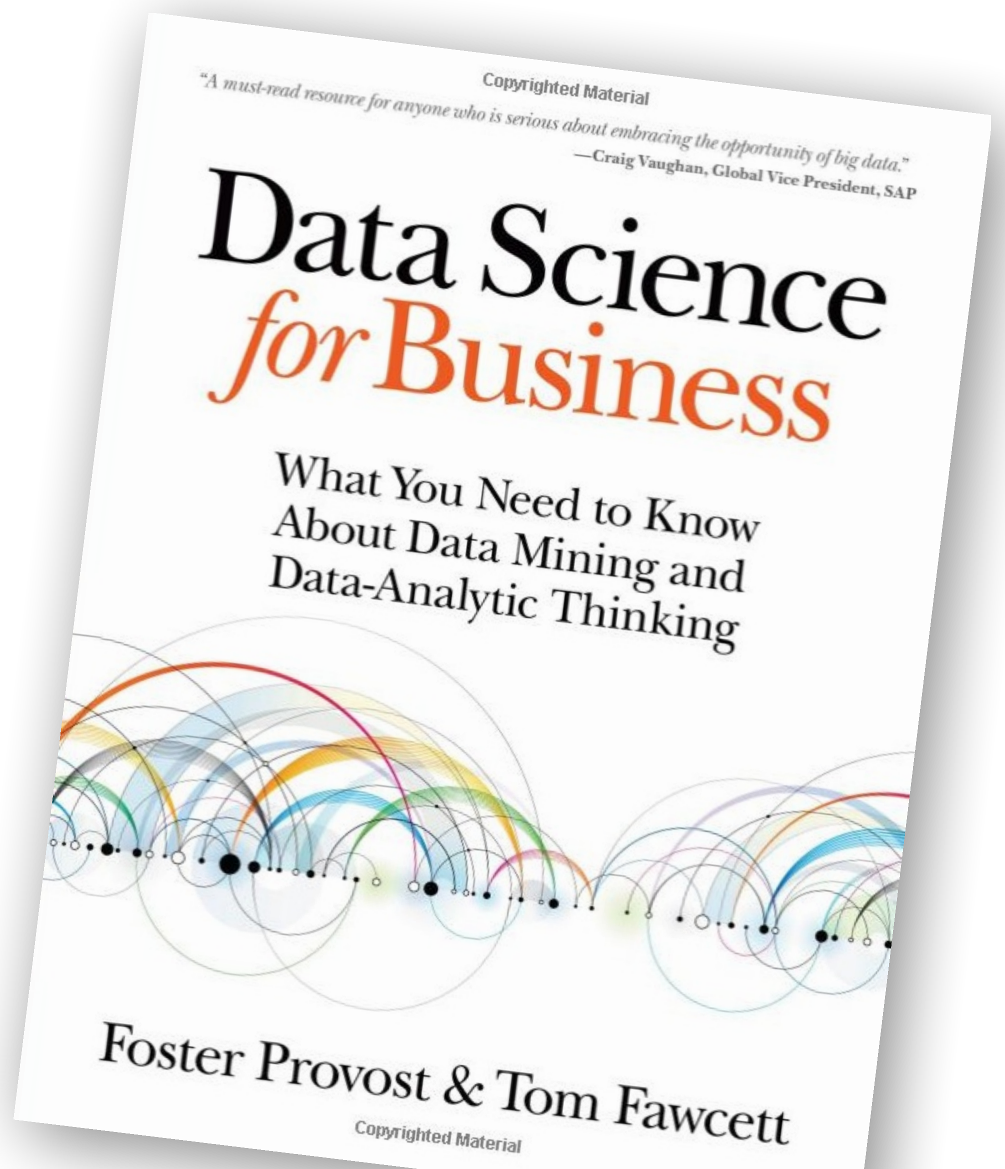


http://spanning.com/blog/choosing-between-storage-based-and-unlimited-storage-for-cloud-data-backup/

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

And here’s a good book.

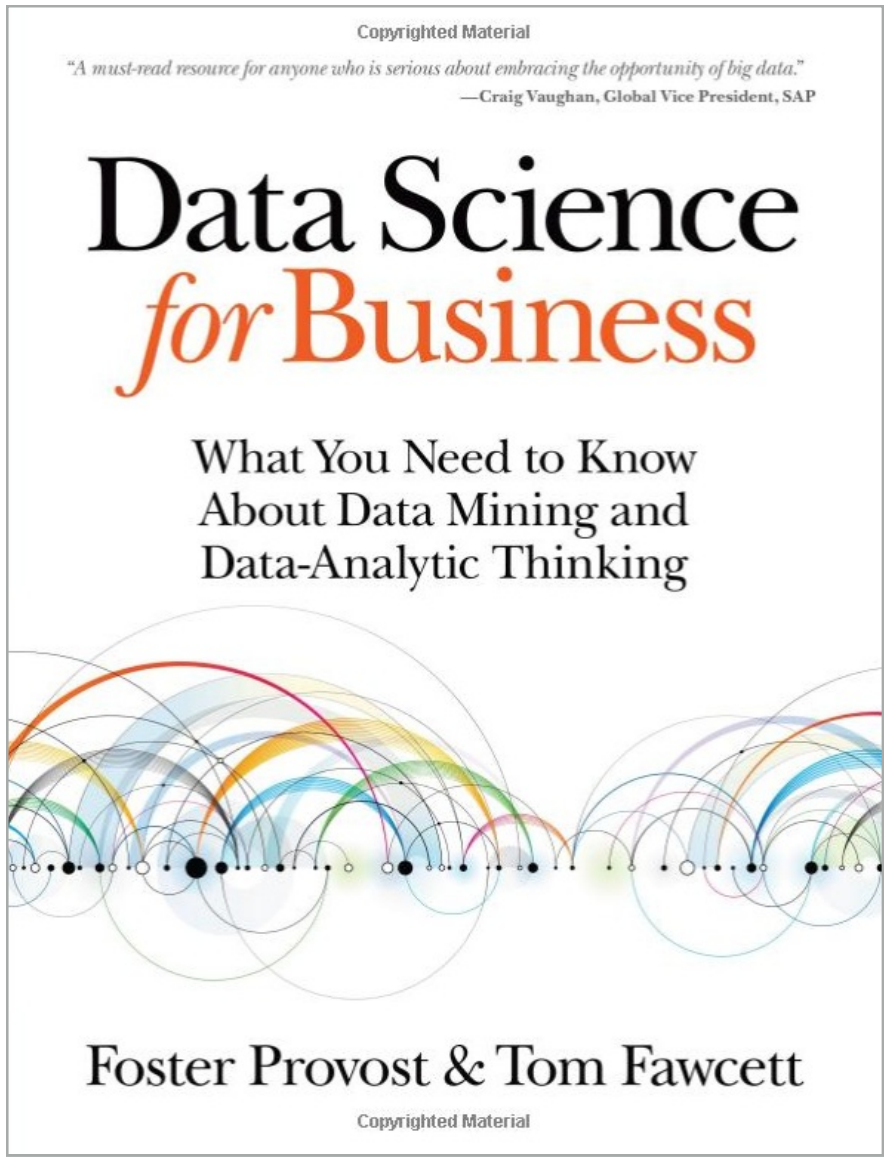
A critical skill in data science is the ability to decompose a data-analytics problem into pieces such that each piece matches a known task for which tools are available. Recognizing familiar problems and their solutions avoids wasting time and resources reinventing the wheel. It also allows people to 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.



FREE for all Georgia Tech users at O'Reilly's **Safari Books Online** (and also many other data science related books)

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**.

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

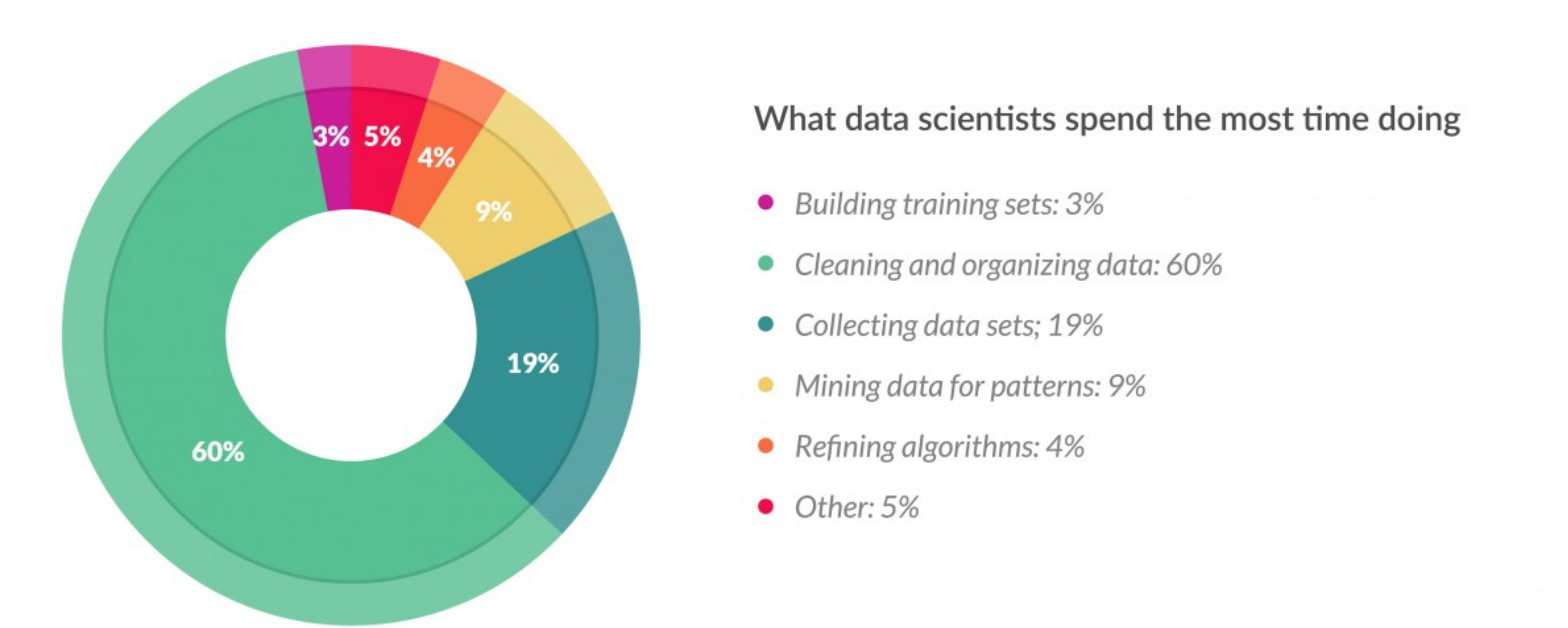
How dirty is real data?

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



Follow us on: Github Twitter

Google™ Custom Search Search ×

OPEN Refine

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

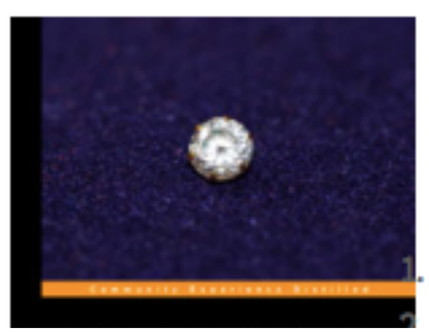
- Home
- Download
- Documentation
- Community
- Post archive
- A Governance Model for OpenRefine
- Using OpenRefine: a manual

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

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 (e.g., Scikit-learn, Pandas, NumPy, TensorFlow, PyTorch)
- 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

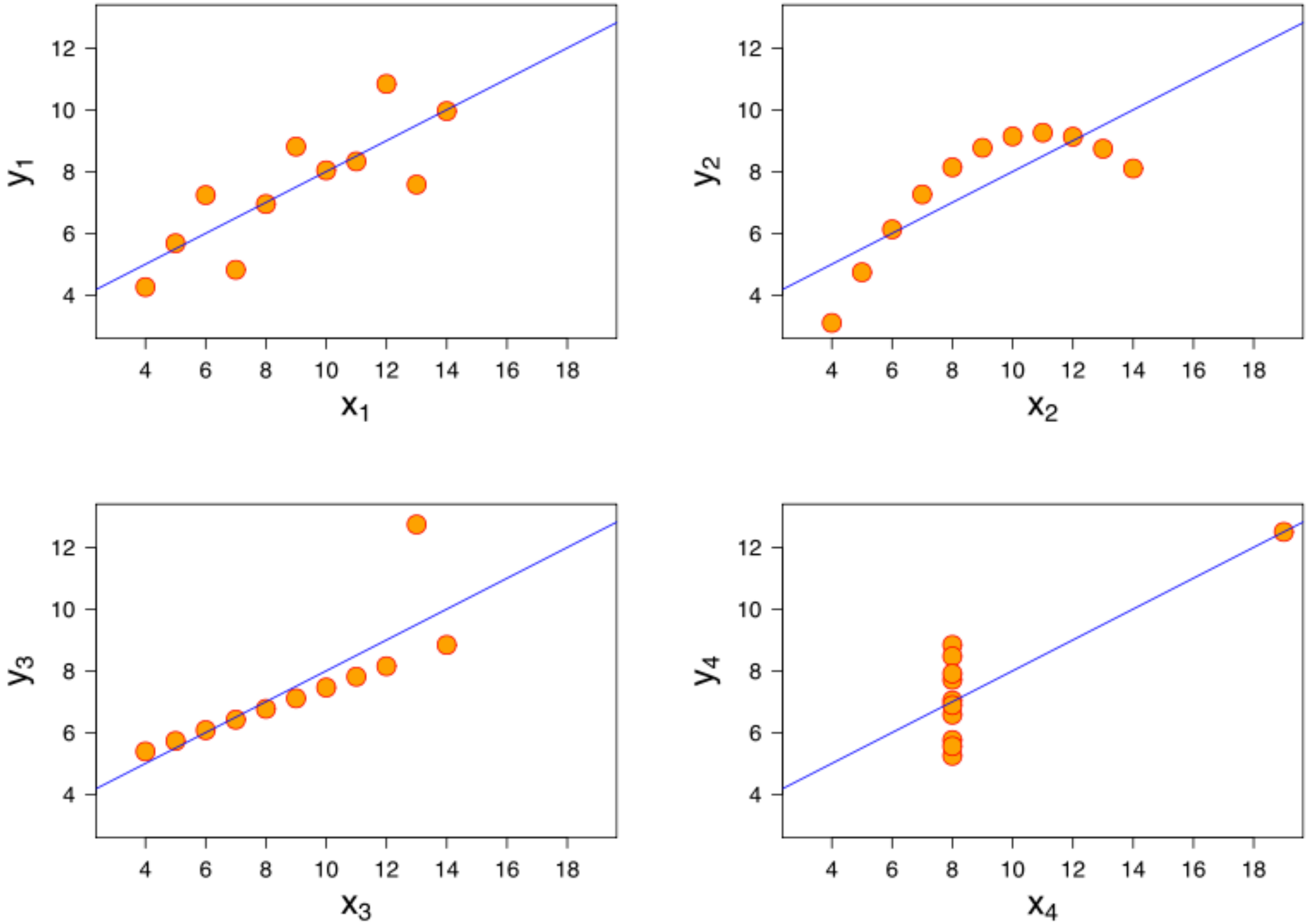
Visualization is **NOT** only about "making things look pretty"

(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

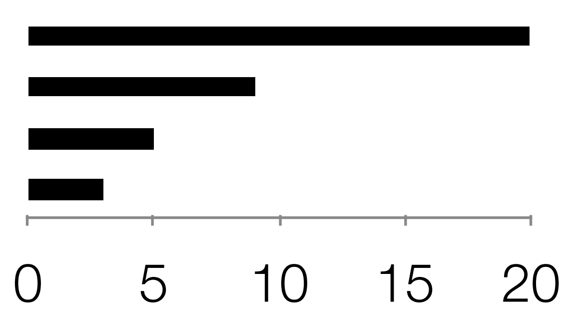


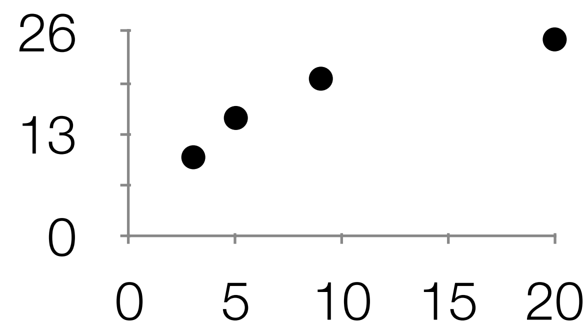
https://en.wikipedia.org/wiki/Anscombe%27s_quartet

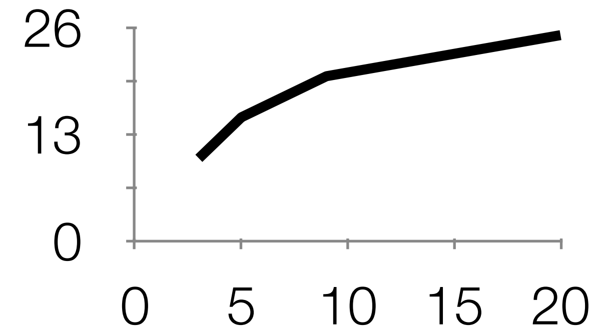
Designing **effective** visualization is **not hard if you learn the principles**.

Easy, because…

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

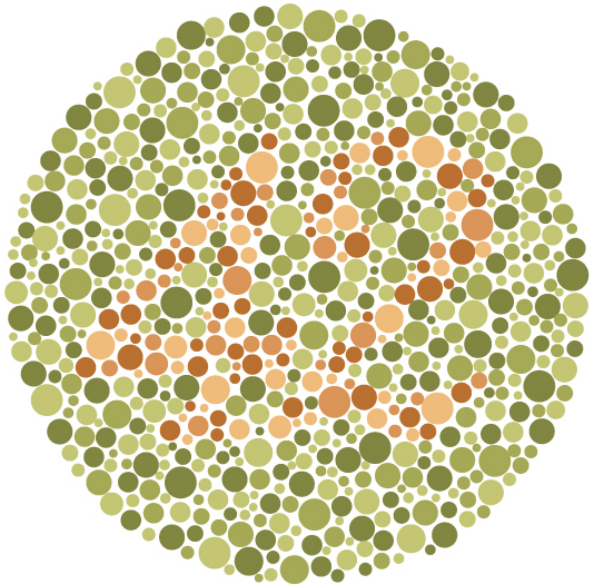


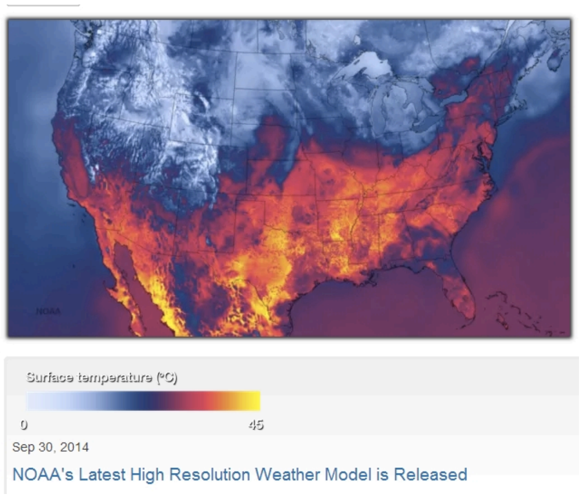




Designing **effective** visualization is **not hard if you learn the principles**.

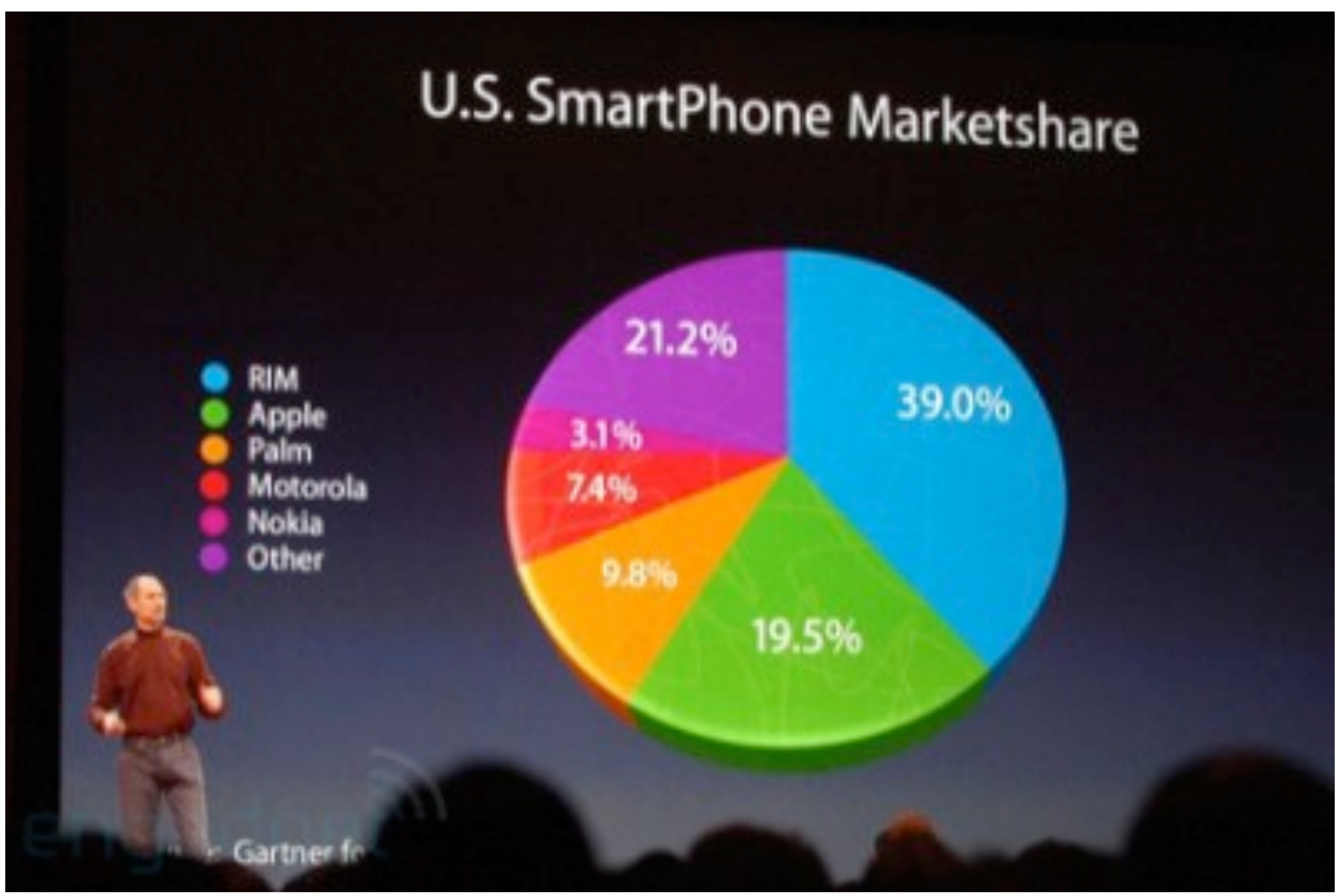
Colors (even grayscale) must be used carefully

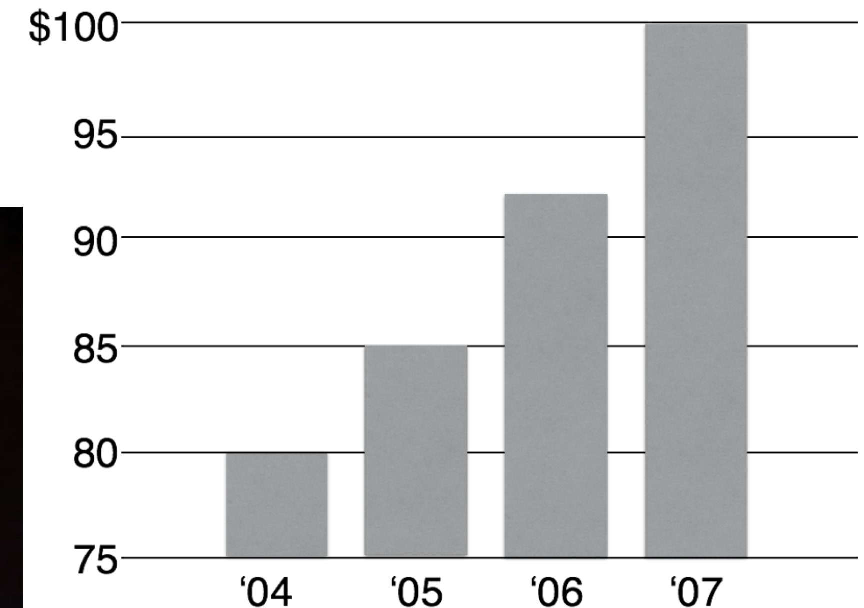


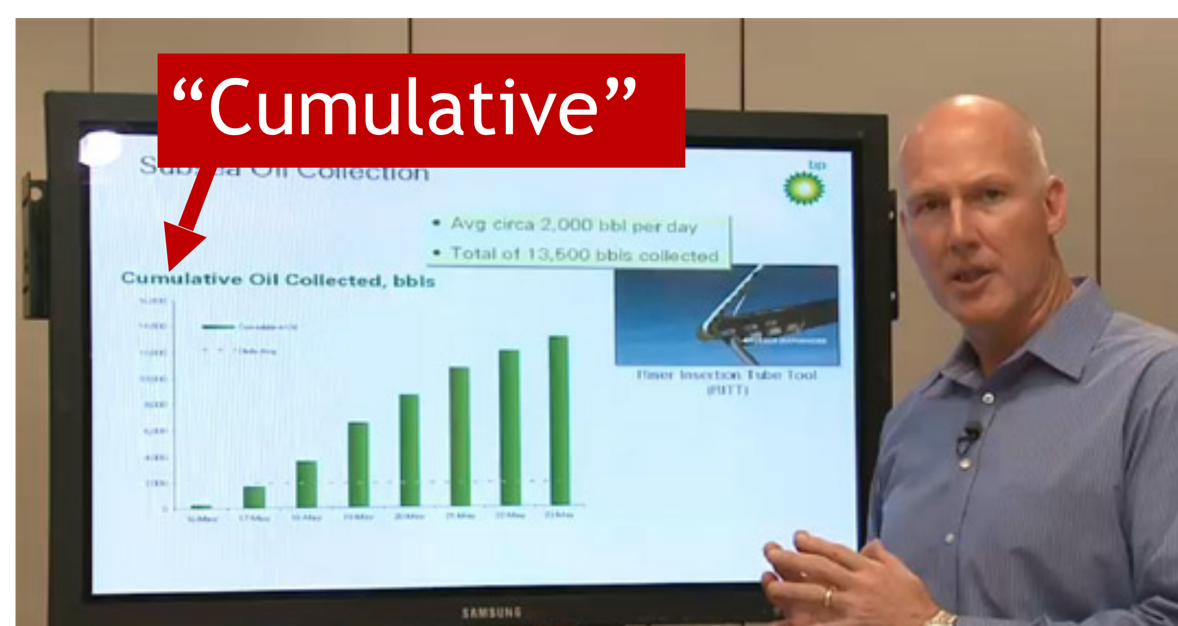


Designing **effective** visualization is **not hard if you learn the principles**.

Charts can mislead (sometimes intentionally)







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

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

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

Getting Started With Data Science and Machine Learning

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

Twitter: @PoloChau