

INTRO TO DATA SCIENCE

LECTURE 1: DATA EXPLORATION

Hari P Thadakamalla

Course Meeting → Mon/Wed: 7:00 – 10:00

Course Notes: https://github.com/hptdss/DAT_BOS_2

Course Emails : Schoology (Sign Up!)

Email:

Hari P Thadakamalla - hpt.dss@gmail.com

Mathew Mathew – mathewma@gmail.com

Stuart Powers – stuart.powers@gmail.com

I. WHAT IS DATA SCIENCE? II. THE DATA MINING WORKFLOW

LAB:

III. VISUALIZATION AS A MEDIUM IV. FIRST LINEAR MODEL

INTRO TO DATA SCIENCE

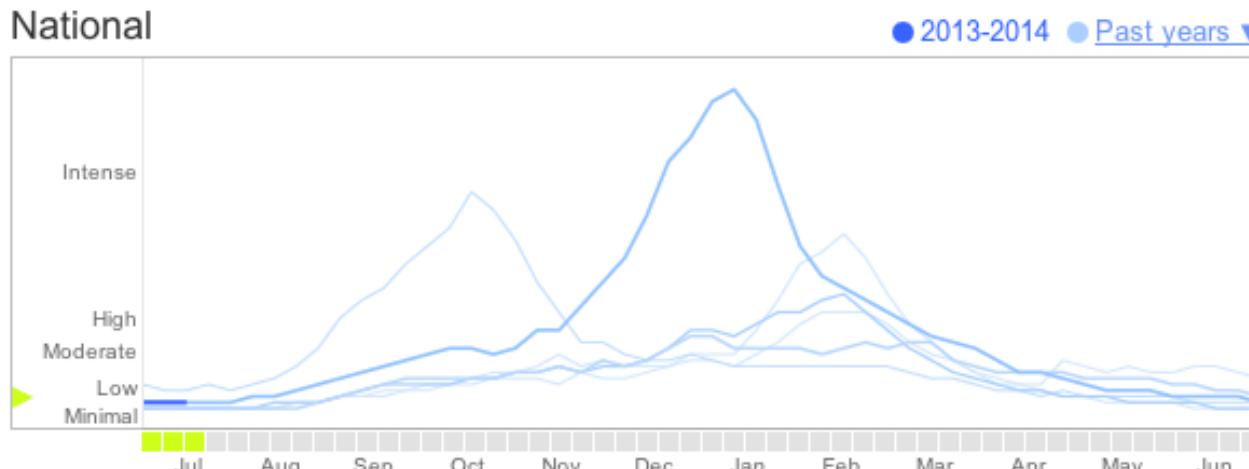
I. WHAT IS DATA SCIENCE?

- A set of tools and techniques used to extract useful information from data.

- A set of tools and techniques used to extract useful information from data.
- *“Data science, as it’s practiced, is a blend of Red-Bull-fueled hacking and espresso-inspired statistics”* – Michael Driscoll, Founder@metamarkets

Google FLU TRENDS*

Detecting flu trends and influenza epidemics using search engine query data



50 million search queries to
45 candidate queries

450 million models to identify
the candidate queries !!!!

Validation of the model

Complex !!!

2012-2013 ???

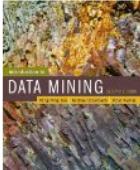
DATA SCIENCE IN ACTION

AMAZON.COM RECOMMENDATIONS*

Recommended Based on Your Browsing History



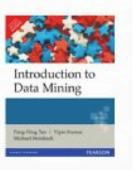
Introduction to Data Mining
Pang-Ning Tan, Michael Steinbach,
...
Hardcover
 (21)
\$132.40 **\$110.37**



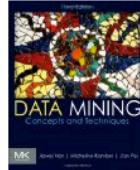
Introduction to Data Mining
► Pang-Ning Tan
Hardcover
\$124.00 \$117.80



Data Mining: Practical Machine...
► Ian H. Witten, Eibe Frank, Mark A. Hall
Paperback
 (31)
\$69.95 **\$41.30**



Introduction to Data Mining
Pang-ning Tan, Michael Steinbach,
...
Paperback



Data Mining: Concepts and Techniques...
Jiawei Han, Micheline Kamber, Jian Pei
Hardcover
 (18)
\$74.95 **\$54.63**

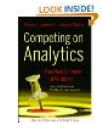
Customers Who Bought This Item Also Bought



Big Data: A Revolution That Will Transform ...
Viktor Mayer-Schonberger
 (115)
Kindle Edition
\$9.45



Predictive Analytics:
Microsoft Excel
► Conrad Carlberg
 (13)
Kindle Edition
\$14.40



Competing on Analytics:
The New Science of ...
► Thomas H. Davenport
 (97)
Kindle Edition
\$16.17



Keeping Up with the Quants: Your Guide to ...
► Thomas H. Davenport
 (8)
Kindle Edition
\$12.99



Naked Statistics: Stripping the Dread ...
Charles Wheelan
 (115)
Kindle Edition
\$12.99

- Several million catalog items and millions of customers
- Browsing history
- Purchase history

Item-based collaborating filtering

Business Goals: click-through, conversions, revenue impact

* <http://www.cs.umd.edu/~samir/498/Amazon-Recommendations.pdf>

DATA SCIENCE IN ACTION



INSIDE THE CAVE: Obama's Digital Campaign*



CREDIT: TIME

THE CAVE

The campaign's data & technology operations made up an estimated 30-40% of headquarters staff

Analytics to improve every aspect of the campaign - Web, Email, Field, and Communications

- A/B Testing for best Email subject
- Dynamic modeling and 'persuadability' score
- Mobile app optimization – Donate by single click like Amazon

Other Internet and e-commerce Companies

* <http://engagedc.com/download/Inside%20the%20Cave.pdf>

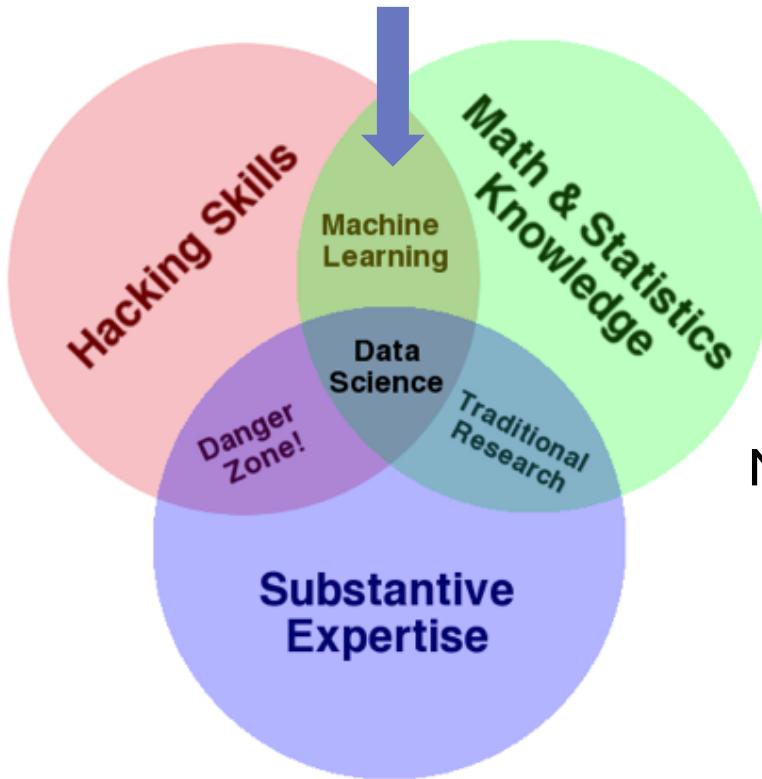
SKILLS NEEDED FOR A DATA SCIENTIST ?

10

- Manipulating text files
- Basic programming
- Thinking algorithmically
- Basic Statistics
- Modeling
- Intuition
- Problem solving skills
- Business knowledge
- Understanding vector operations (Linear Algebra !!!)

THE QUALITIES OF A DATA SCIENTIST

12



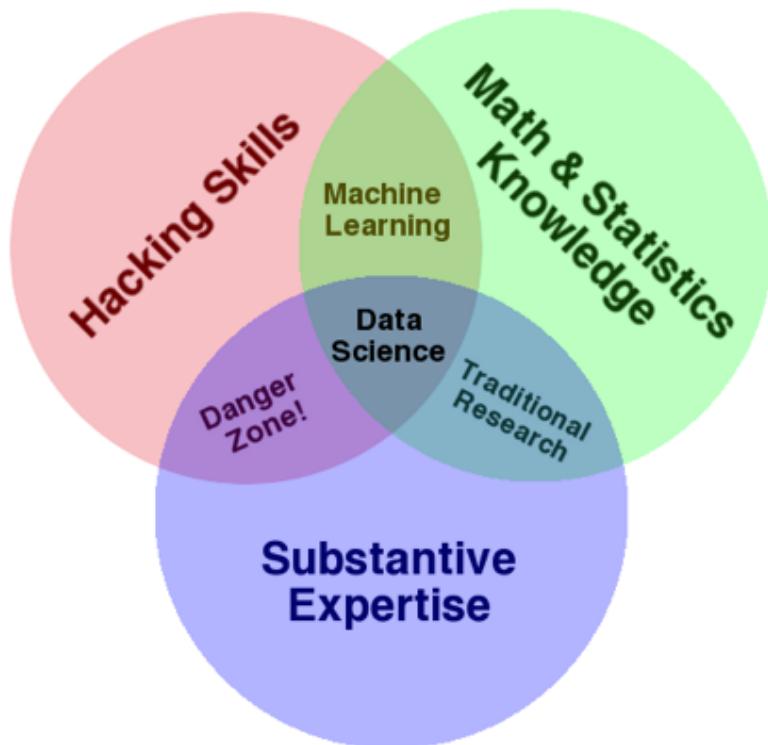
Machine Learning Examples

- Reading images of handwritten characters
- Face detection
- Spam filtering
- Voice recognition

By Drew Conway source: <http://www.dataists.com/2010/09/the-data-science-venn-diagram/>

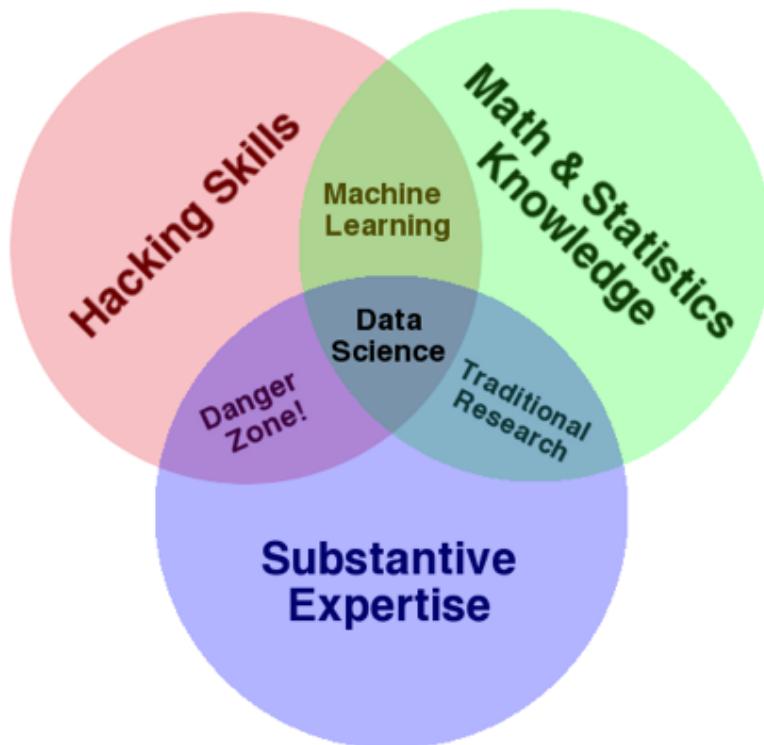
THE QUALITIES OF A DATA SCIENTIST

13



THE QUALITIES OF A DATA SCIENTIST

14



ONE MORE THING!

Communication skills

WHAT MAKES A GOOD DATA SCIENTIST?

15



Michael E. Driscoll

@medriscoll



Following

Data scientists: better statisticians than most programmers & better programmers than most statisticians bit.ly/NHmRqu
[@peteskomoroch](#)



More

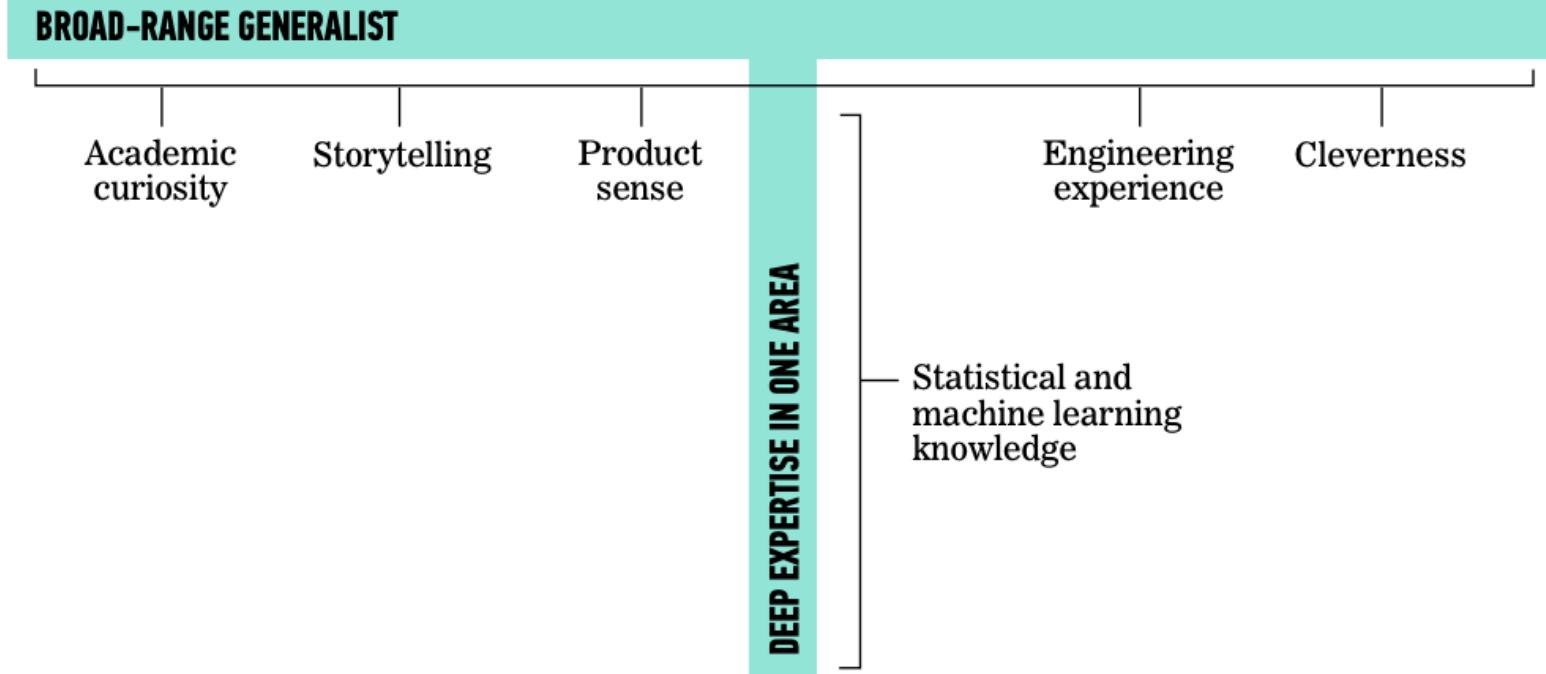


“A hybrid computer scientist software engineer statistician. The best tend to be really curious people, thinkers who ask good questions and are O.K. dealing with unstructured situations and trying to find structure in them.” – Rachel Schutt, News Corp

- Statistical and machine learning knowledge
- Engineering experience
- Academic curiosity
- Product sense
- Storytelling
- Cleverness

WHAT MAKES A GOOD DATA SCIENTIST?

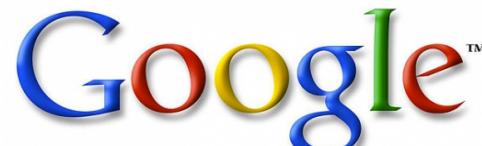
17



- A set of tools and techniques used to extract useful information from data.
- An interdisciplinary, problem-solving oriented subject.
- The application of scientific techniques to practical problems.
- A rapidly growing field.

WHO USES DATA SCIENCE?

19

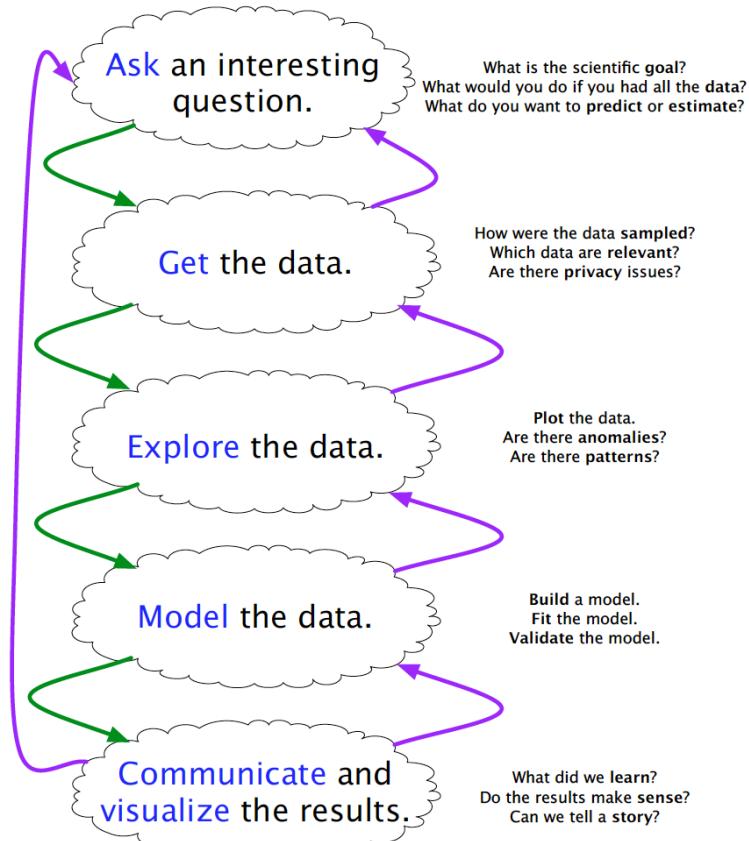


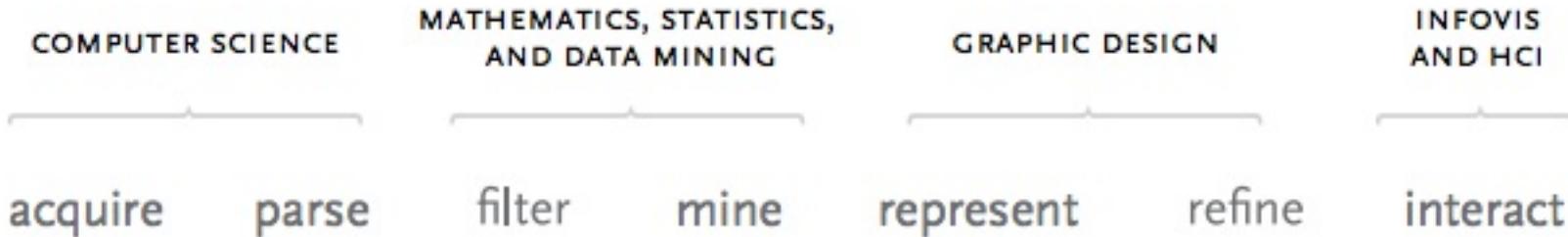
INTRO TO DATA SCIENCE

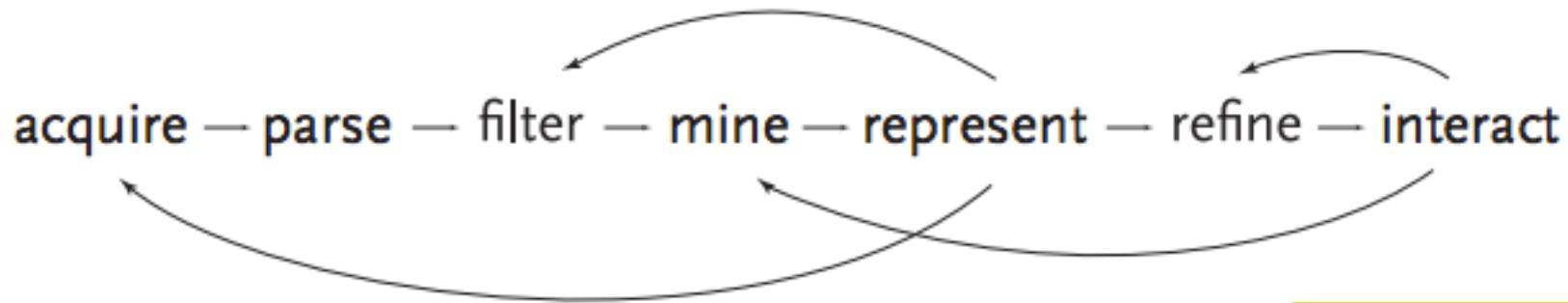
II. THE DATA SCIENCE WORKFLOW

THE DATA SCIENCE WORKFLOW

21







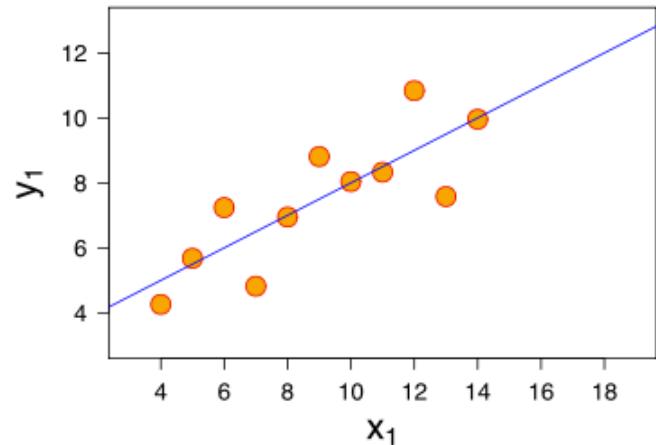
NOTE

This diagram illustrates
the *iterative* nature of
problem solving

III. VISUALIZATIONS AS A MEDIUM

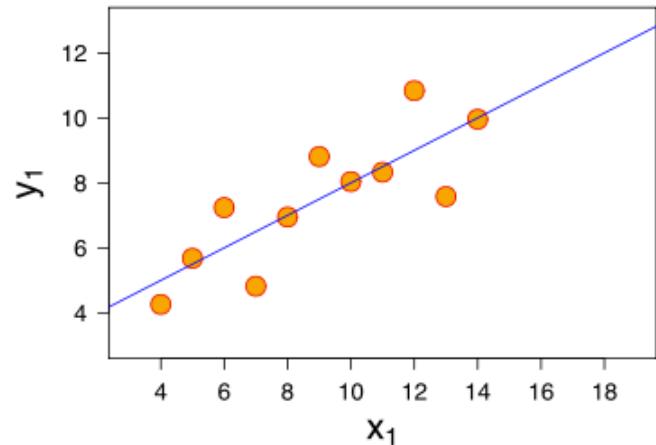
Consider the following dataset:

- eleven (x, y) points



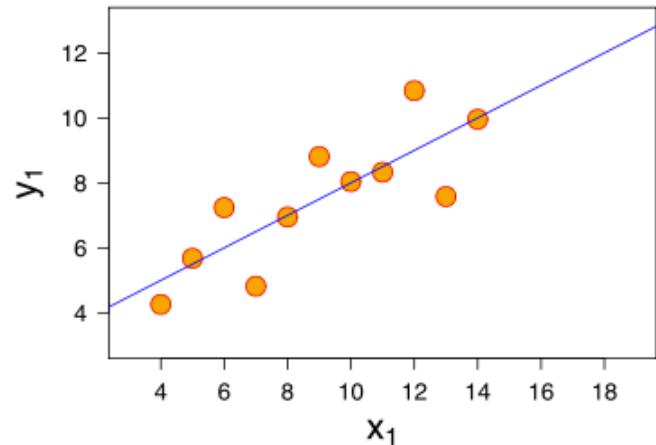
Consider the following dataset:

- eleven (x, y) points
- mean of x = 9, mean of y = 7.5



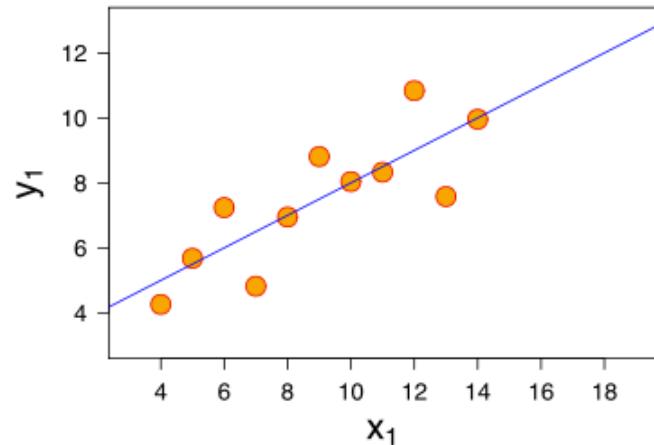
Consider the following dataset:

- eleven (x, y) points
- mean of $x = 9$, mean of $y = 7.5$
- variance of $x = 11$, variance of $y = 4.1$



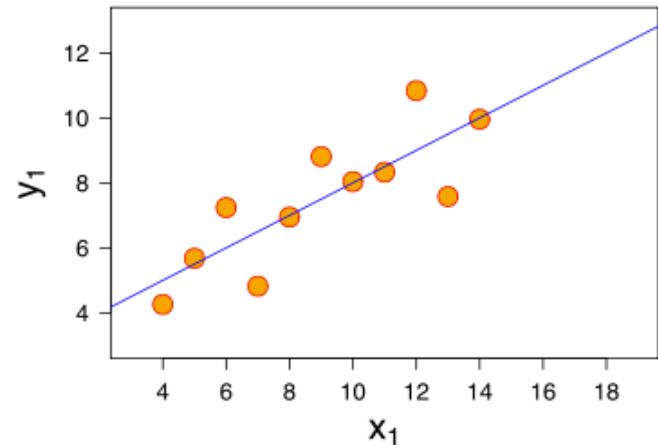
Consider the following dataset:

- eleven (x, y) points
- mean of $x = 9$, mean of $y = 7.5$
- variance of $x = 11$, variance of $y = 4.1$
- correlation of x and $y = 0.8$



Consider the following dataset:

- eleven (x, y) points
- mean of x = 9, mean of y = 7.5
- variance of x = 11, variance of y = 4.1
- correlation of x and y = 0.8
- line of best fit: $y = 3.00 + 0.500x$

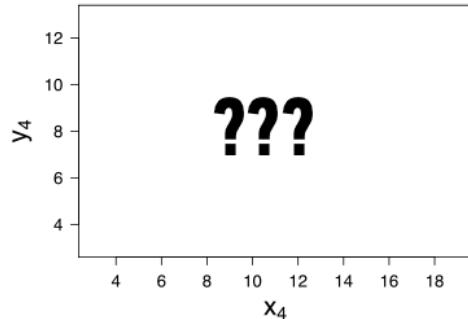
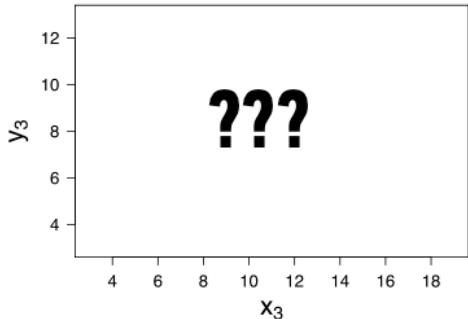
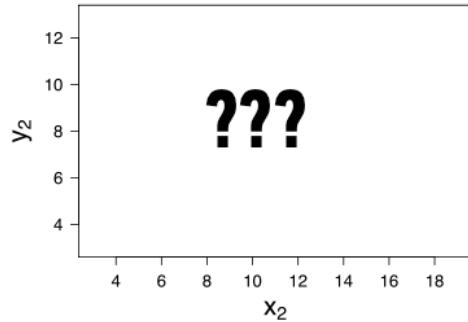
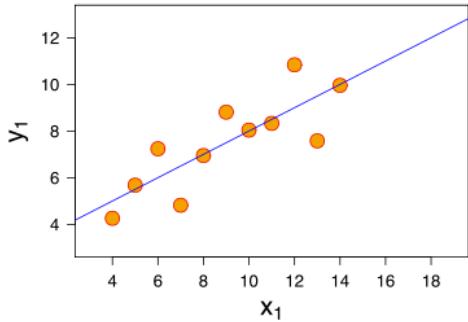


EXERCISE – WHY VISUALIZE DATA?

30

Now, suppose I give you
three more datasets
with exactly the same
characteristics...

Q: how similar are these
datasets?



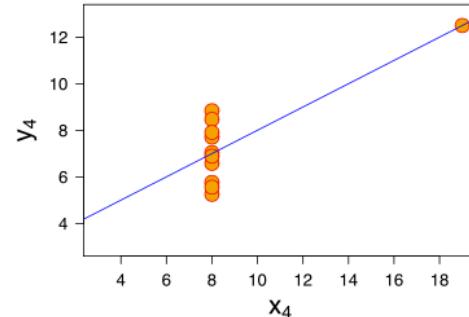
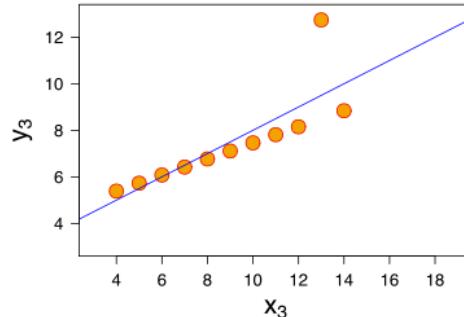
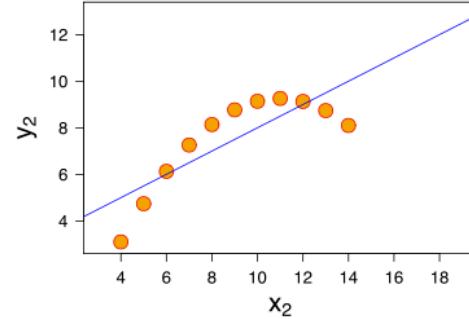
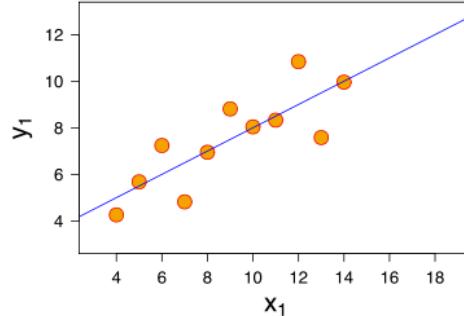
EXERCISE – WHY VISUALIZE DATA?

31

Now, suppose I give you
three more datasets
with exactly the same
characteristics.

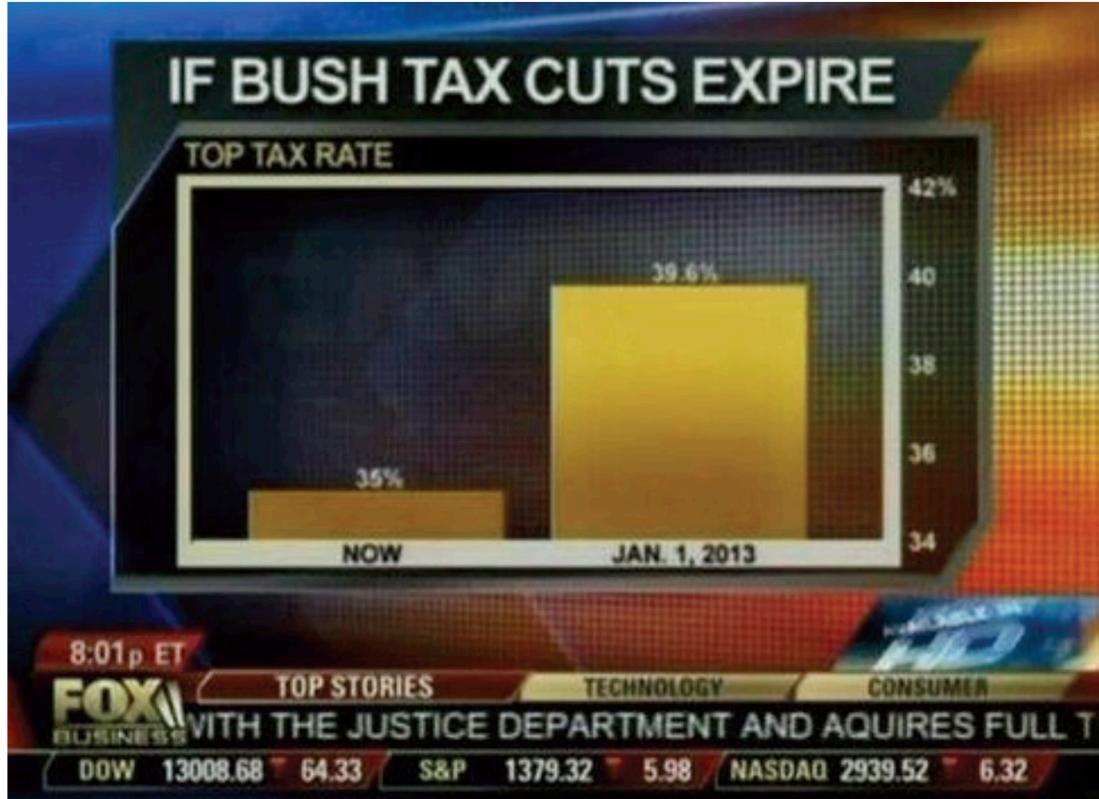
Q: how similar are these
datasets?

A: not very!



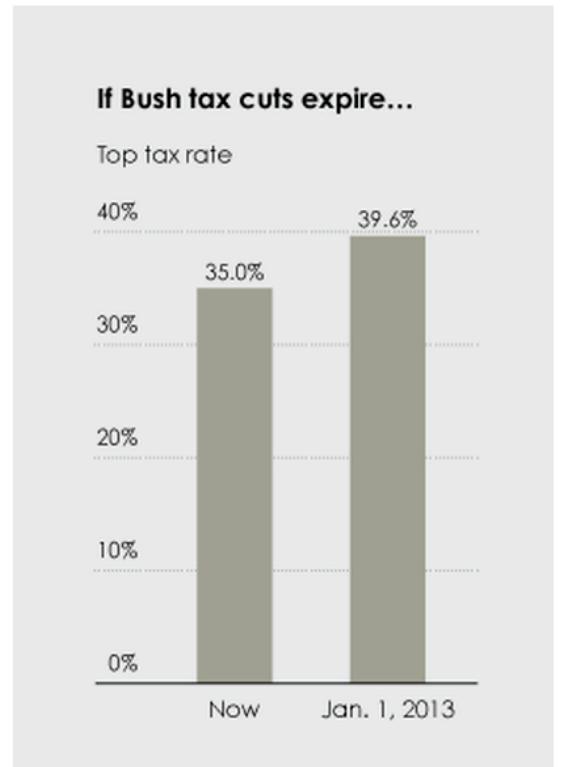
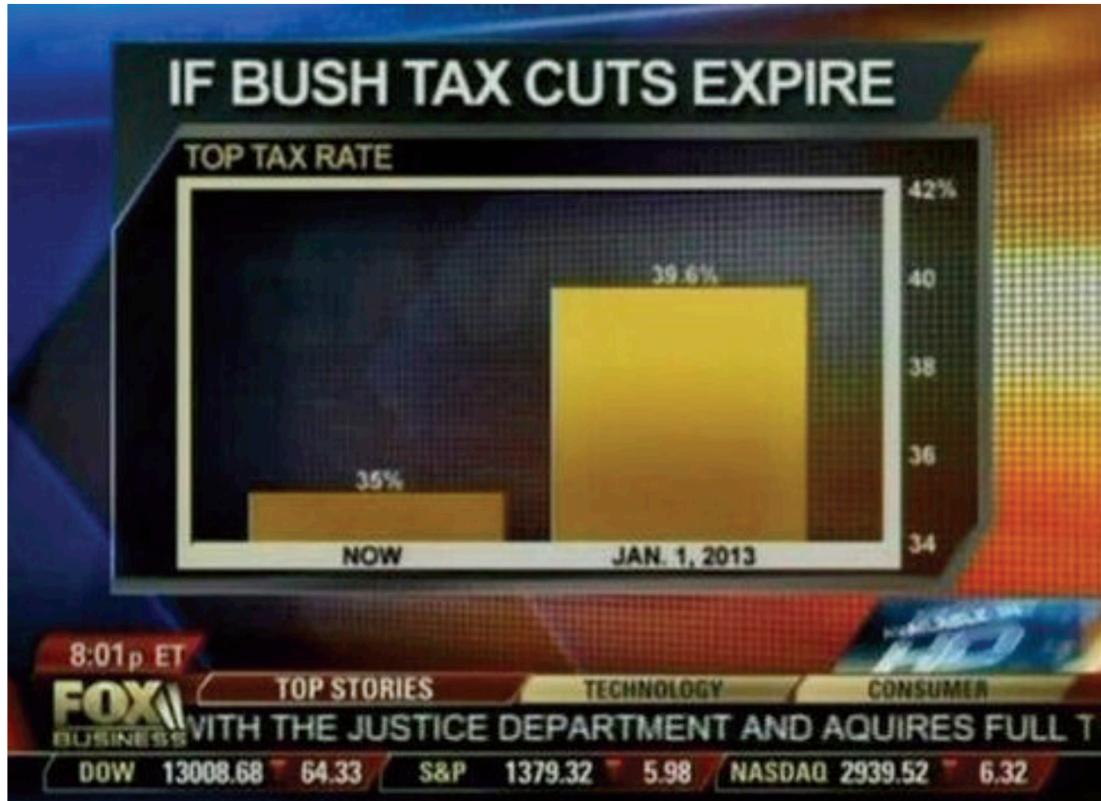
GRAPHICAL INTEGRITY

32



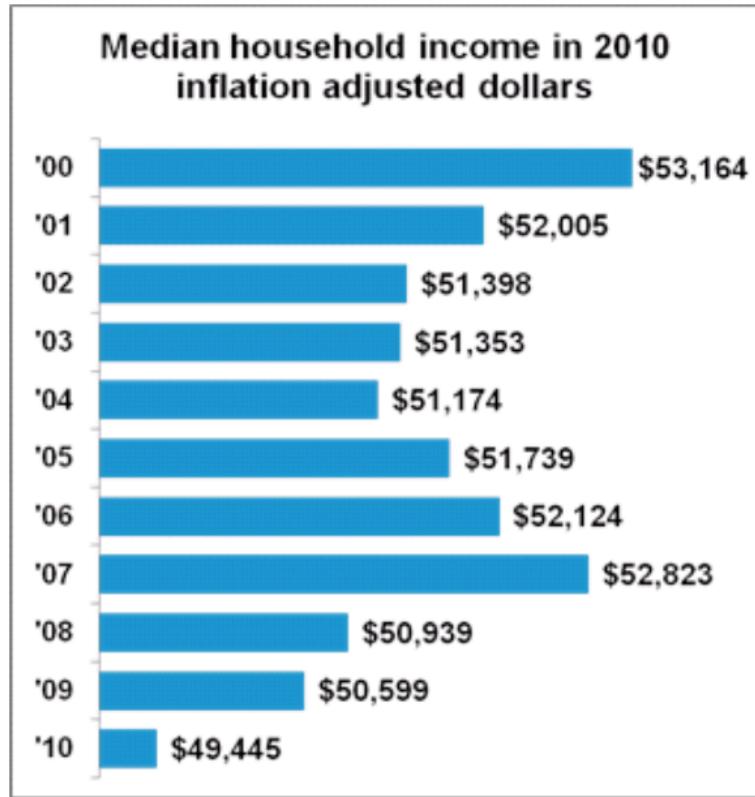
GRAPHICAL INTEGRITY

33



ANOTHER EXAMPLE

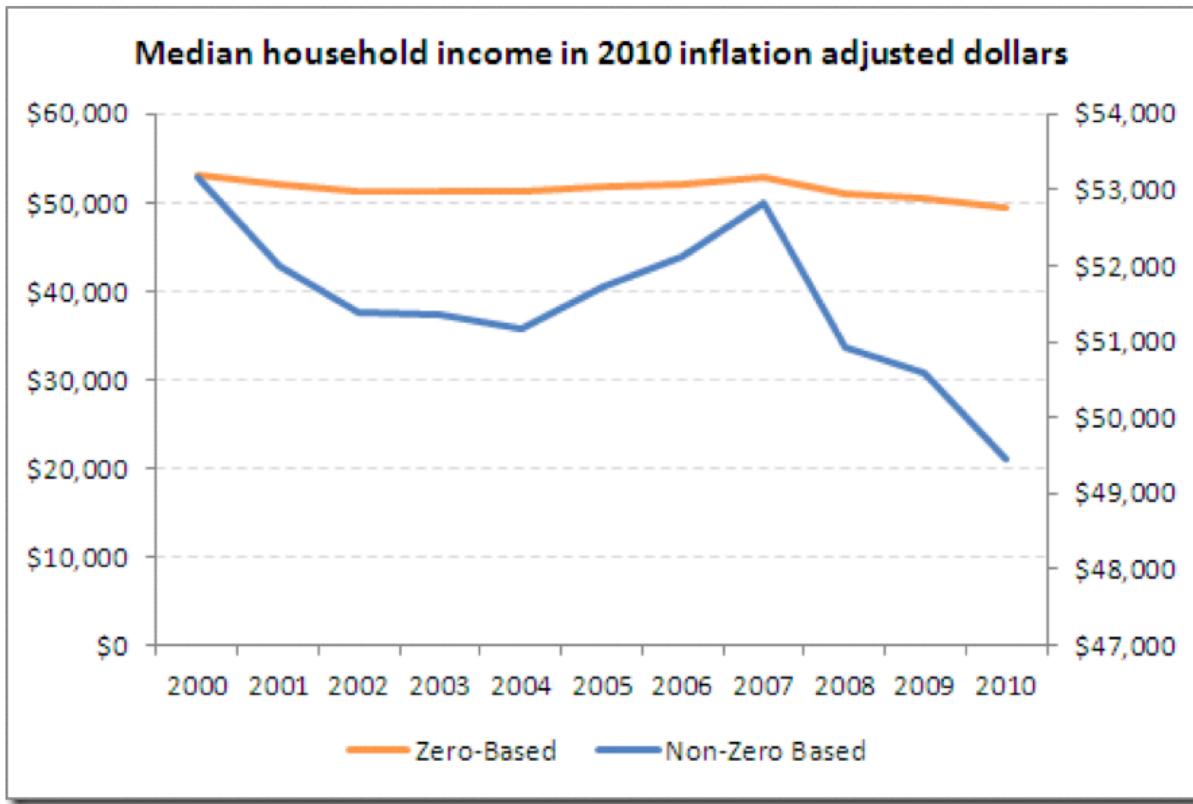
34

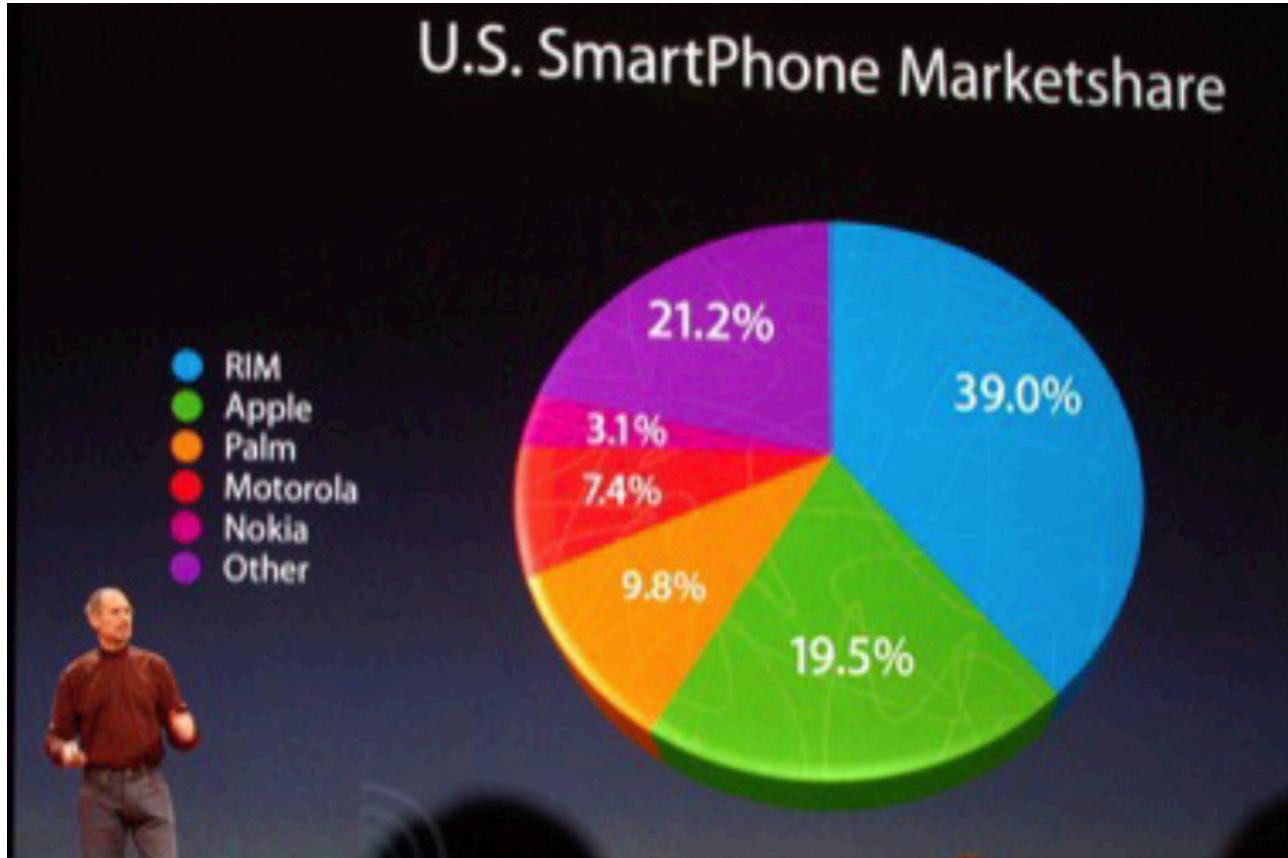


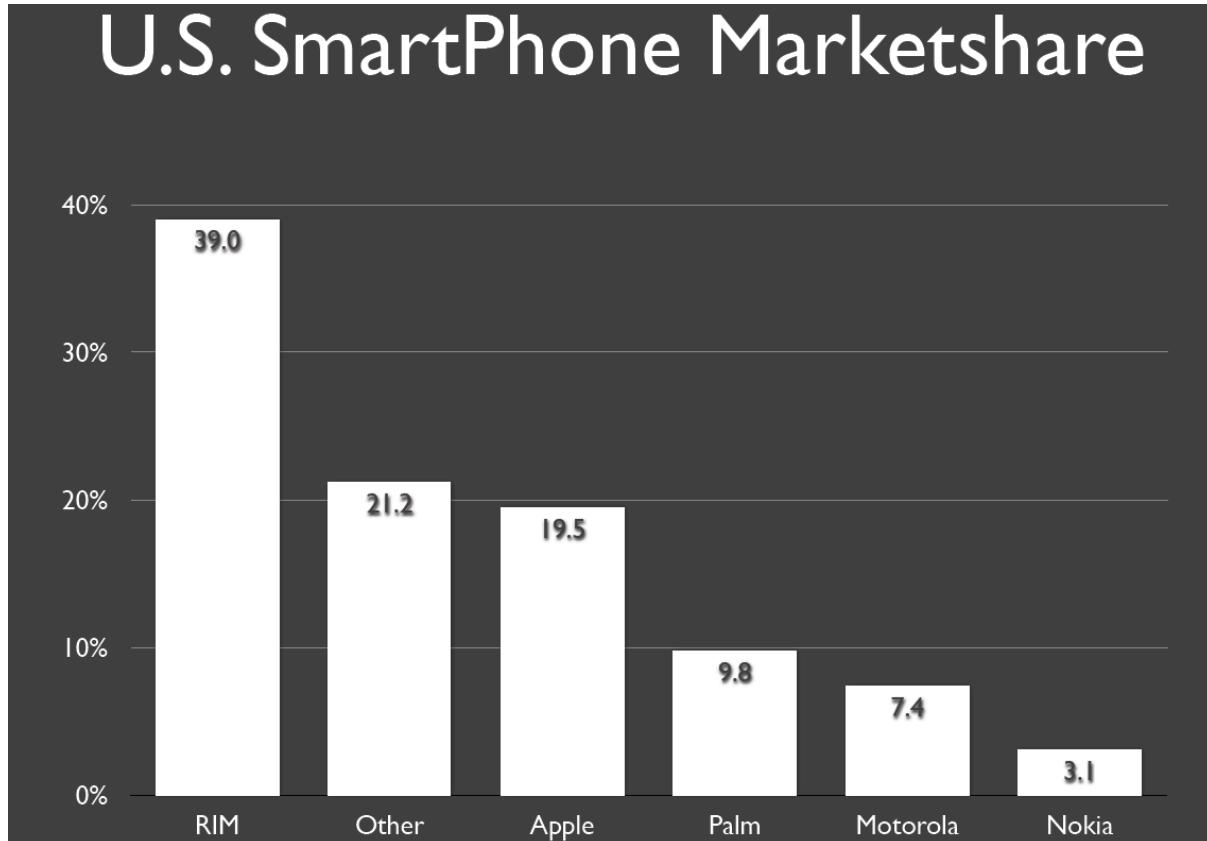
Always start your bar graphs
at Zero !

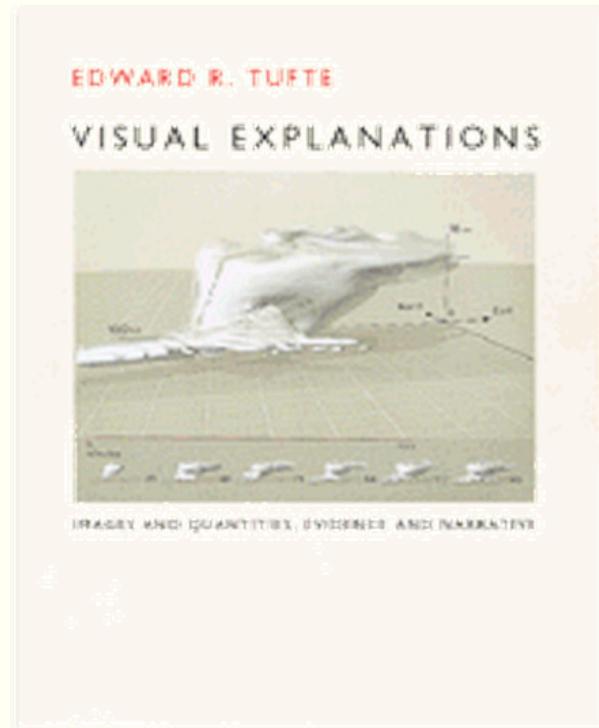
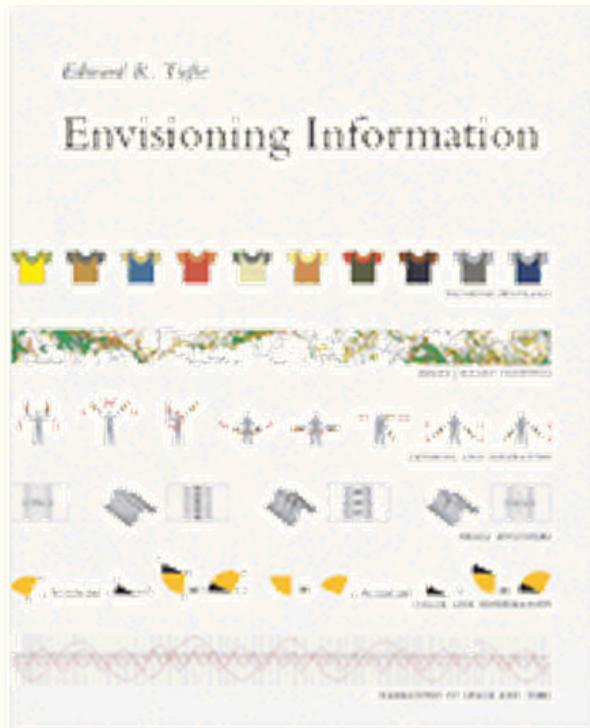
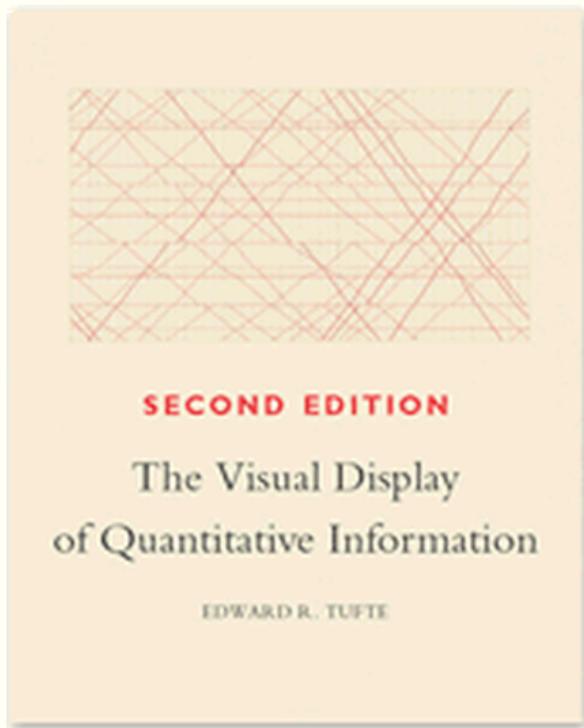
ANOTHER EXAMPLE

35









IV. FIRST LINEAR MODEL

KEY OBJECTIVES

- Become familiar with the R environment
- Explore data in R
- Visualize data using ggplot2
- Mathematical bonus: power laws

INTRO TO DATA SCIENCE

DISCUSSION