Data Analysis in Astronomy and Physics

Lecture 0: Syllabus

M. Röllig - SS 2022

Contents

The course introduces the basic aspects of data analysis and the application of statistical methods to data in astronomy and other physical sciences. The course covers the following topics (incomplete list):

- Descriptive statistics
- probability & probability distributions
- statistical inference
- hypothesis testing
- correlation analysis
- regression
- least-squares fitting and testing fits
- multivariate analysis

Contents

- data smoothing
- interpolation
- feature detection
- **machine learning concepts**

We will also cover practical aspects, such as:

- plotting and presenting data
- data formats, and
- work with real data
- image processing, etc.

The course will often use real astronomical data or applications from astronomy, but the contents of the course are of course applicable to all physical sciences.

Requirements

- Master level course
  - Mathematik für Studierende der Physik
  - Einführung in die Programmierung für Physiker
  - Grundlagen der Astronomie
- Some hands-on exercises require computer access and the basic understanding of a computational data analysis software of your choice (Origin, Excel, Matlab, Mathematica, R), or a programming language like python.
- A practical requirement is of course the technical possibilities to access, join and participate in the course.

  Contact me if you have any problems!

Course Organisation

- The course will take place as *'flipped classroom'*. This means:
  - I will publish pre-recorded lectures videos and share the videos (via YouTube) and the lecture notes one week before the scheduled lecture date.
  - ON the lecture date we will have an interactive session. Here we will:
    - Discuss open questions and problems you might have encountered while studying the notes/videos. Q&A style
    - Present supplementary material, e.g. extended worked examples, real-life cases, additional in-depth topics.
    - Do hands-on examples on your own laptops/machines.

Schedule

These are the original scheduled times. We currently expect the course to be fully in-presence with the option of having the exercise in hybrid mode (presence & online in parallel).

Lectures:	**Monday, 10:00-11:30 hrs**
Hand out of assignments:	**Monday, 10:00**
Submission deadline:	**Monday, 10:00 in the following week**
Exercise session ：	**Friday, 13:00-13:45 hrs**

Schedule First Week

Video(s) and lecture notes available on:	**Monday, 28.03.2022 10:00**
First lecture:	**Monday, 04.04.2022 10:00**
Hand out of first assignment:	**Monday, 04.04.2022 10:00**
Deadline of first assignments:	**Monday, 11.04.2022 10:00**
Correction of first assignment:	**Thursday, 14.04.2022**
1st exercise session:	**Friday, 15.04.2022 13:00**
Online Python introduction ：	**Friday, 08.04.2022 13:00**

Credit Points

This course is part of the **Astronomy/Astrophysics master module**.

Attendance of the course and successful completion the problem sheets (successful means **40% of all possible points**) of the course awards 4.5 CPs.

Exercise

- Most problems and maybe all projects will **require some computer work**, such as visualizing data, performing numerical experiments, realizing some algorithm in a programming language or software.
- If so, the choice of software/operation system/programming language is yours.
  - The results should always be presented in an independent manner, i.e. must not be presented in form of source code, excel files, etc. but in a report.
  - Source code, data files, scripts, etc should come as a supplement.
- Consider doing the exercise in Python. If you don’t know Python yet, use this course to learn it.
- Modern data analysis is predominantly done in Python. Python knowledge is a must for anybody looking for a job as data scientist.

Exercise

Exercises will be supervised by:

Craig Yanitski (yanitski@ph1.uni-koeln.de)
Dr. Christof Buchbender (buchbend@ph1.uni-koeln.de)

The exercises will be a *mix of standard problem sheets and projects*.

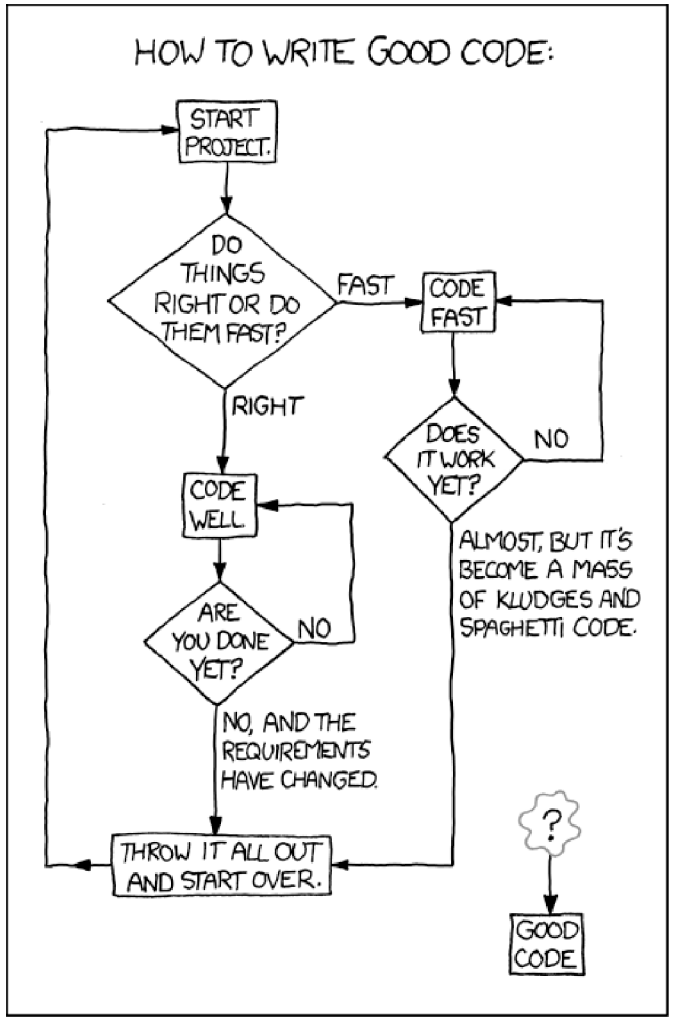
The problem sheets will be available for download on ILIAS every Monday together with the lecture notes for the week. Students have to hand in the completed problem sheets till Monday morning (10:00 am) the following week electronically! Please upload the completed problem sets on ILIAS. We will correct your submissions till Thursday and provide you a schematic template solution as a basis for discussion on Friday.

www.ilias.uni-koeln.de/ilias/goto_uk_crs _ 4522808.html

Out[ ]=

Exercise

- Provide a written solution in the form of a brief report as PDF. Do not just upload code! If you create figures, add labels, and figure captions to enable the reader to understand what is shown.
- Pack all your code, data and PDFs in a ZIP file and upload it to ILIAS.
- Add a minimum level of comments to your code.



Credits: https://xkcd.com/844/

Lecture notes

- Prepared lecture notes will be available online in form of PDF slides and *Mathematica* Notebook slides. Please check the course web-page @ ILIAS for any relevant material.
- The lecture (video) will be pre-recorded and published online.
- It is **important** that you read the respective lecture notes and watch the videos BEFORE coming to class.
- During class the contents of the notes will not necessarily be repeated (we can of course discuss related questions and clarify open issues). Instead selected topics will be covered by presenting worked examples and by demonstrating relevant principles.

Course web-page

www.ilias.uni-koeln.de/ilias/goto_uk_crs _ 4522808.html

Out[ ]=

Literature

- Wall and Jenkins, Practical Statistics for Astronomers (Cambridge University Press)
- Feigelson and Babu , Modern Statistical Methods for Astronomy (Cambridge University Press), 20% discount available
- Bevington and Robinson, Data Reduction and error analysis for the physical sciences (McGraw-Hill)
- Taylor, An Introduction to error analysis (Springer)
- Press et al., Numerical Recipes, (Cambridge University Press), available online
- Jean-Luc Starck and Fionn Murtagh, Handbook of Astronomical Data Analysis, free ebook
- Lecture notes of the The Summer School in Astrostatistics 2009.
- Coursera Online course : Data Analysis and Statistical Inference, Duke University
- Lecture notes on Astrostatistics, from John Peacock, Royal Observatory Edinburgh

Contact Information

DISCORD

In the previous (online) semesters we had a discord channel for course communications. Do we need this?

Open for other communication options.

TELEPHONE

+49 179 124 144 8 (M. Röllig)

EMAIL

roellig@ph1.uni-koeln.de (lecturer)
yanitski@ph1.uni-koeln.de (teaching assistant)
buchbend@ph1.uni-koeln.de (teaching assistant)

Introduction

Why do you need to know about data analysis?

http://dilbert.com/strip/2018-04-03

Age of data

CISCO Global Internet Traffic Forecast (Feb 2019) (https://www.cisco.com/c/dam/m/en_us/solutions/service-provider/vni-forecast-highlights/pdf/Global_2022_Forecast_Highlights.pdf)

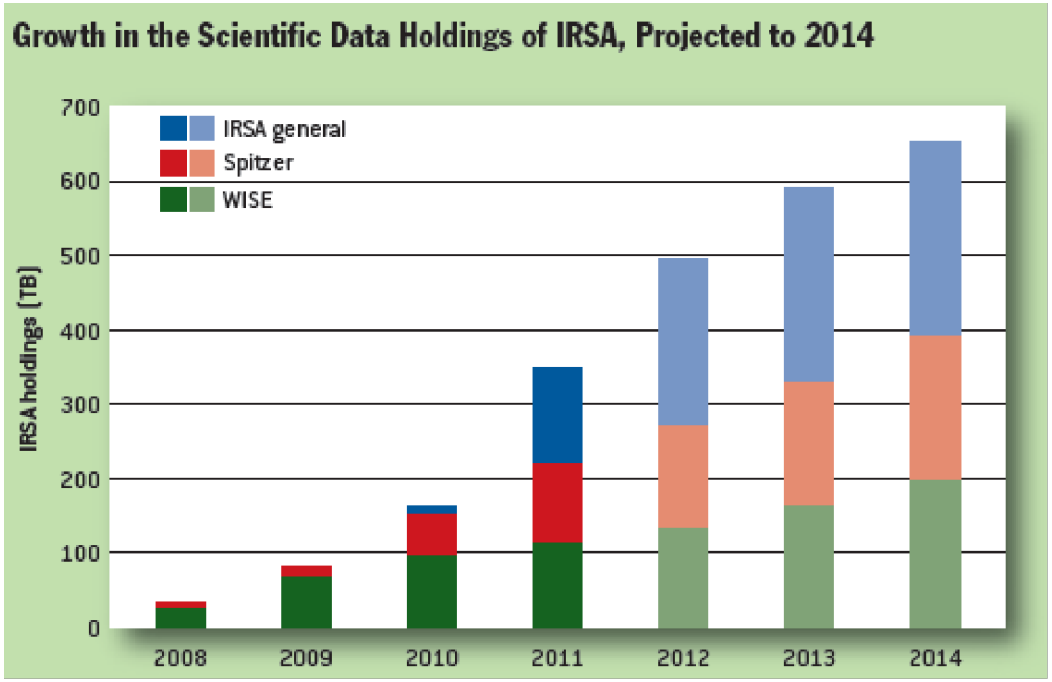
- Annual global IP traffic will reach 4.8 ZB per year by 2022, or 396 exabytes (EB) per month. In 2017, the annual run rate for global IP traffic was 1.5 ZB per year, or 122 EB per month.
- Global IP traffic will increase threefold over the next 5 years. Overall, IP traffic will grow at a Compound Annual Growth Rate (CAGR) of 26 percent from 2017 to 2022. Monthly IP traffic will reach 50 GB per capita by 2022, up from 16 GB per capita in 2017.
- The number of devices connected to IP networks will be more than three times the global population by 2022. There will be 3.6 networked devices per capita by 2022, up from 2.4 networked devices per capita in 2017. There will be 28.5 billion networked devices by 2022, up from 18 billion in 2017.
- Smartphone traffic will exceed PC traffic. In 2018, PCs accounted for 41 percent of total IP traffic, but by 2022 PCs will account for only 19 percent of IP traffic. Smartphones will account for 44 percent of total IP traffic by 2022, up from 18 percent in 2017.
- Globally, mobile data traffic will increase sevenfold between 2017 and 2022.

Age of data

- By 2022, the gigabyte equivalent of all movies ever made will cross the Internet every 1 minutes.
- Global Internet traffic by 2022 will be equivalent to 193x the volume of the entire Global Internet in 2005.
- Globally, gaming traffic will grow 9-fold from 2017 to 2022, a compound annual growth rate of 55%.
- Globally, the average Internet user will generate 84.6 gigabytes of Internet traffic per month by 2022, up 194%from 28.8 gigabytes per month in 2017, a CAGR of 24%.

Introduction

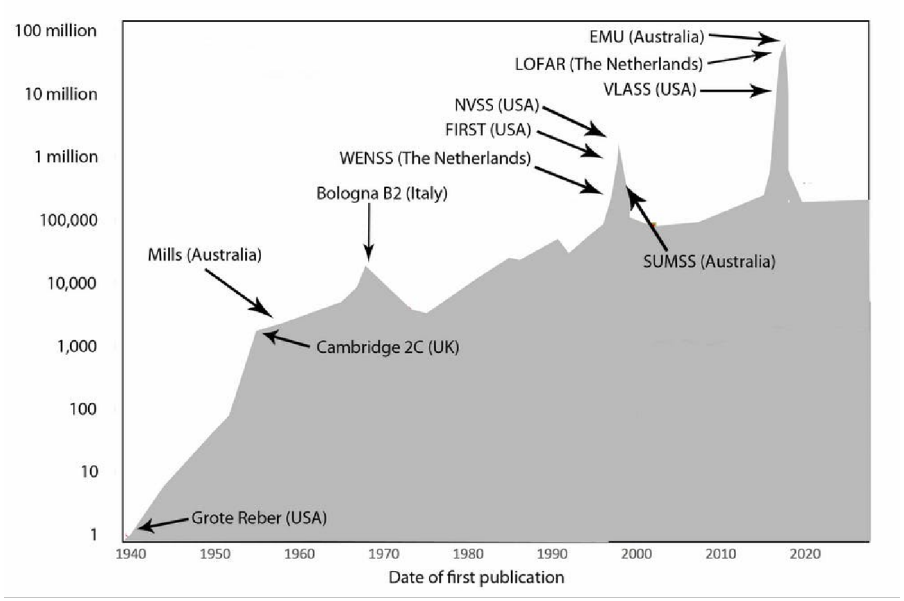
We are living in the age of information. The pure amount of data available today is staggering. The computational power at hand is doubling every other year.



Wide-Field Infrared Survey Explorer	(space telescope 2010-2011)
Spitzer	(space telescope 2004-2020)
IRSA - Infrared Science Archive	(NASA archive)

Introduction

The graph shows two spikes in number of radio sources detected in major surveys over the years, from the birth of radio astronomy to the next-generation surveys. Credit: Ray Norris, Author provided.



Read more at: https://phys.org/news/2017-09-unexpected-big-data-boom-radioastronomy.html#jCp

E.g.: The Gaia Data release 2 consists of data of 1.7 billion stars.

Introduction

If all sensor data were to be recorded in **LHC**, the data flow would be extremely hard to work with. The data flow would exceed 150 million petabytes annual rate, or nearly **500 exabytes** (10^{18} bytes **per day)**, before replication. To put the number in perspective, this is equivalent to 500 quintillion (5×10^{20}) bytes per day, almost 200 times more than all the other sources combined in the world.

On 29 June 2017, the CERN DC passed the milestone of 200 petabytes of data permanently archived in its tape libraries.

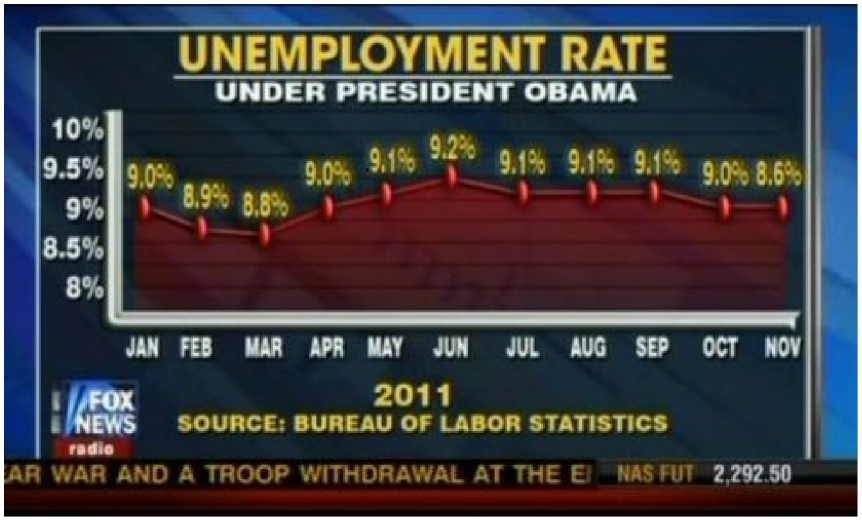
The Square Kilometre Array is a telescope which consists of millions of antennas and is expected to be operational by 2024. Collectively, these antennas are expected to gather 14 exabytes and store **one petabyte per day**.

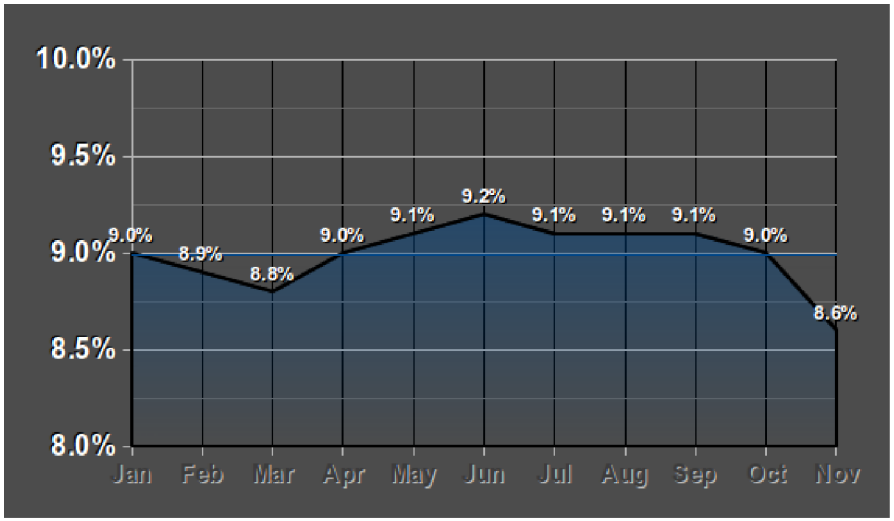
Source: Berimann & Groom 2011 http://en.wikipedia.org/wiki/Big_data

You will be working with data a lot. Δ(astronomical data)/yr$\approx$0.5 PB (10^{15} Byte)

Introduction

Data in the wrong hands can be dangerous and manipulative.

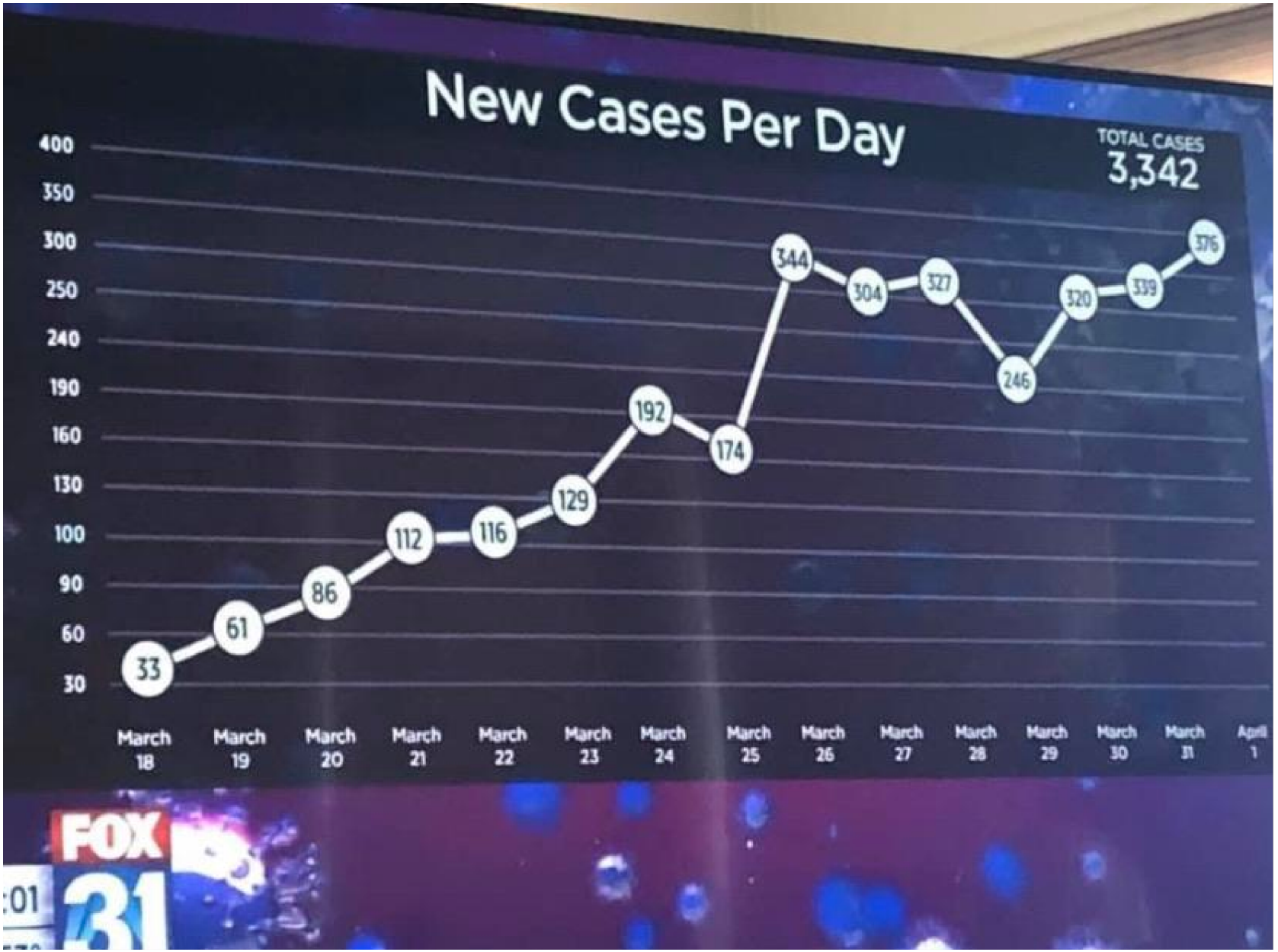




Source:http://freethoughtblogs.com/lousycanuck/2011/12/14/im-better-at-graphs-than-fox-news/

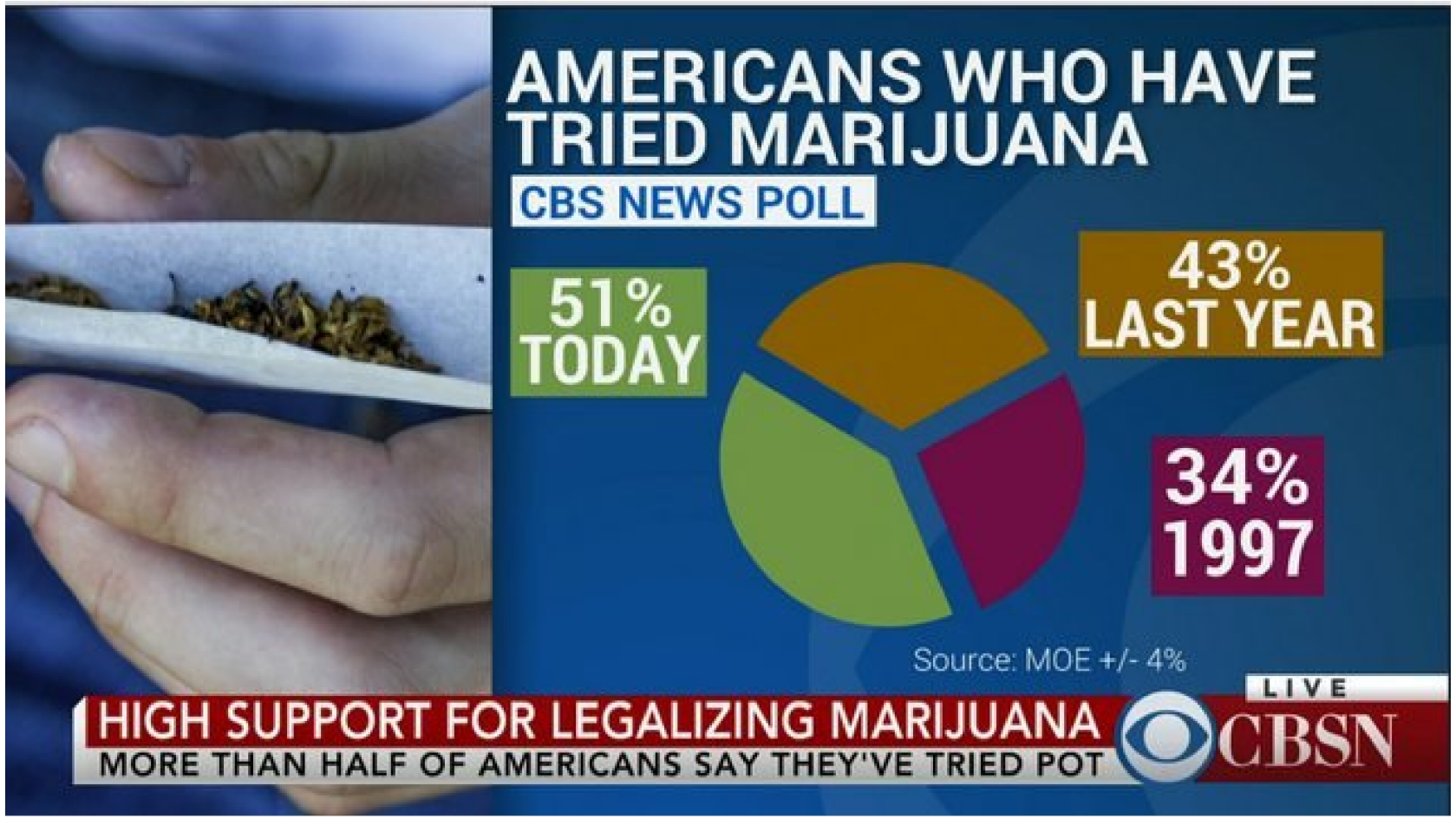
Introduction

Sometimes it is more subtle


New Cases Per Day
TOTAL CASES
3,342
400
350
300
250
240
190
160
130
100
90
60
30
33
61
86
112
116
129
192
174
344
304
327
246
320
339
376
March 18
March 19
March 20
March 21
March 22
March 23
March 24
March 25
March 26
March 27
March 28
March 29
March 30
March 31
April 1
FOX
31

Introduction

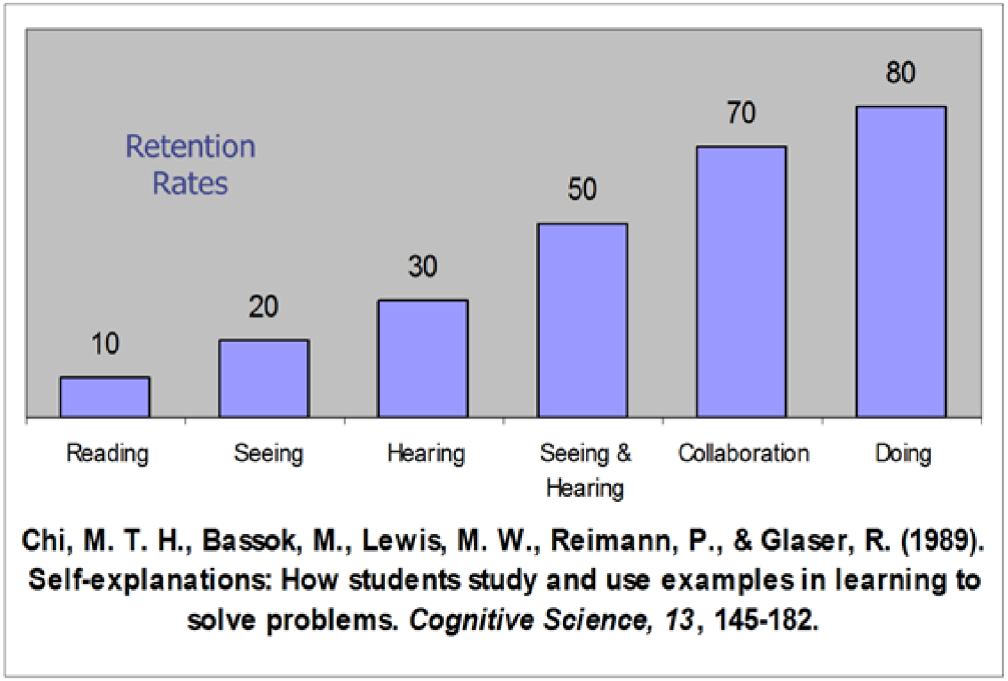
Sometimes it is just bad



Why?

Introduction

Data in the wrong hands can be dangerous and manipulative.



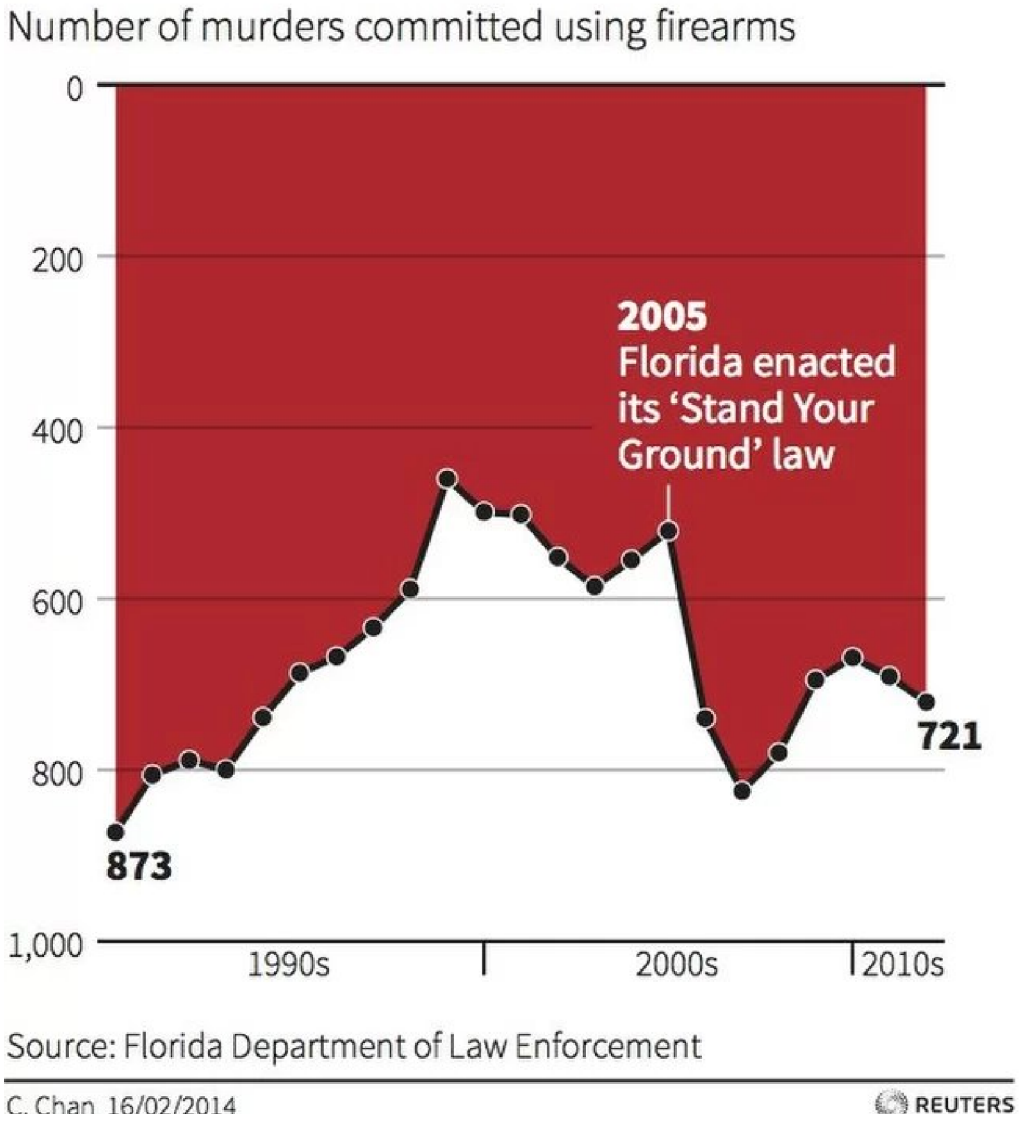
Bogus learning theory: "A quick glance is enough to be suspicious. Any study that produces a series of results with exact multiples of ten is highly suspicious"

Source:http://www.willatworklearning.com/2015/01/mythical-retention-data-the-corrupted-cone.html

Introduction

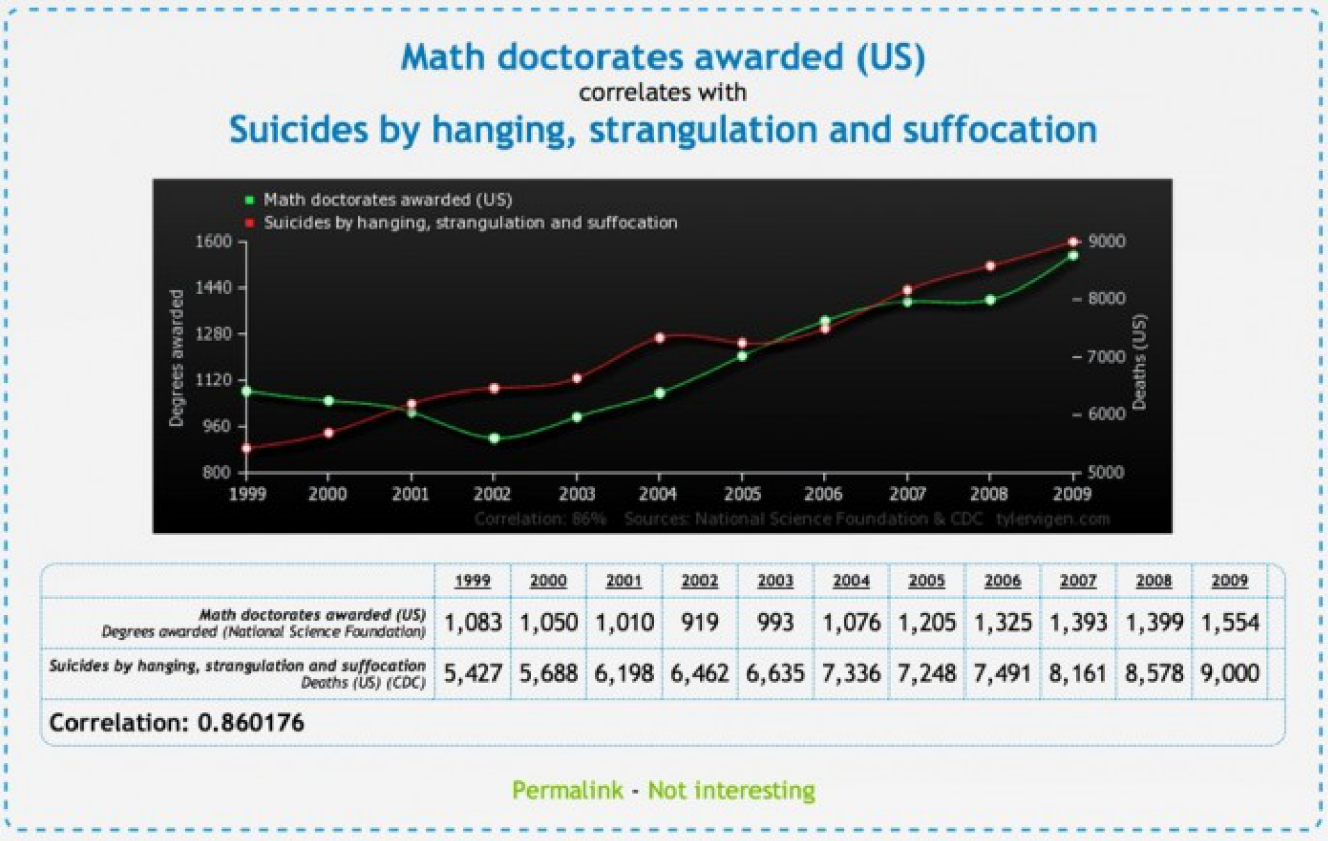
Data in the wrong hands can be dangerous and manipulative.

Gun deaths in Florida



Introduction

Correlation is not causality.



	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Math doctorates awarded (US) *Degrees awarded (National Science Foundation)*	1,083	1,050	1,010	919	993	1,076	1,205	1,325	1,393	1,399	1,554
Suicides by hanging, strangulation and suffocation *Deaths (US) (CDC)*	5,427	5,688	6,198	6,462	6,635	7,336	7,248	7,491	8,161	8,578	9,000

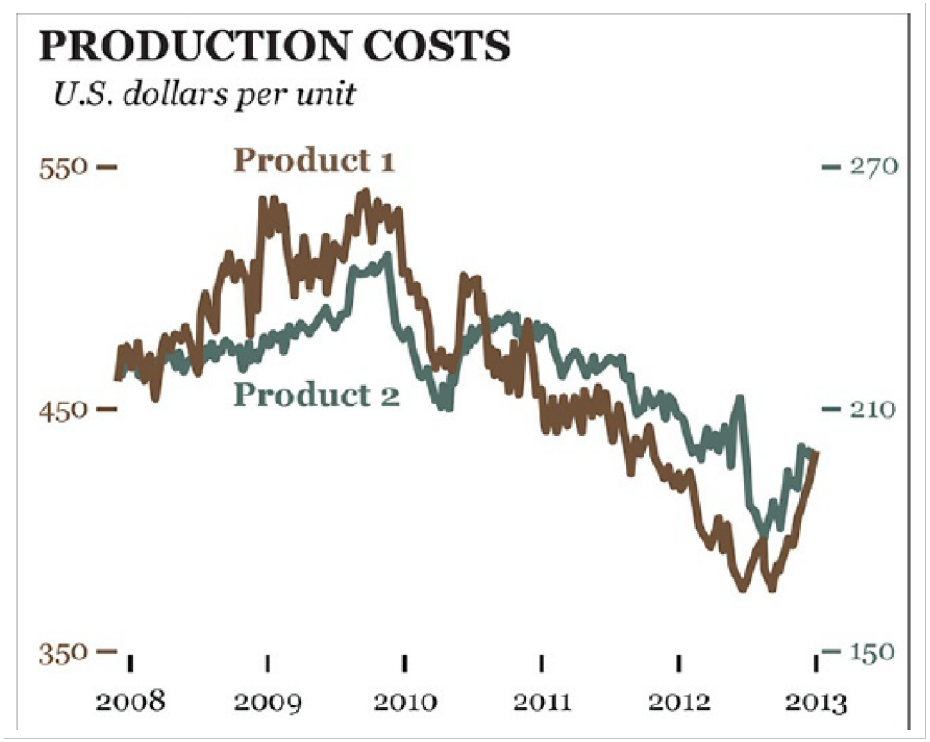
Correlation: 0.860176

Permalink - Not interesting

Source:http://twentytwowords.com/funny-graphs-show-correlation-between-completely-unrelated-stats-9-pictures/3/

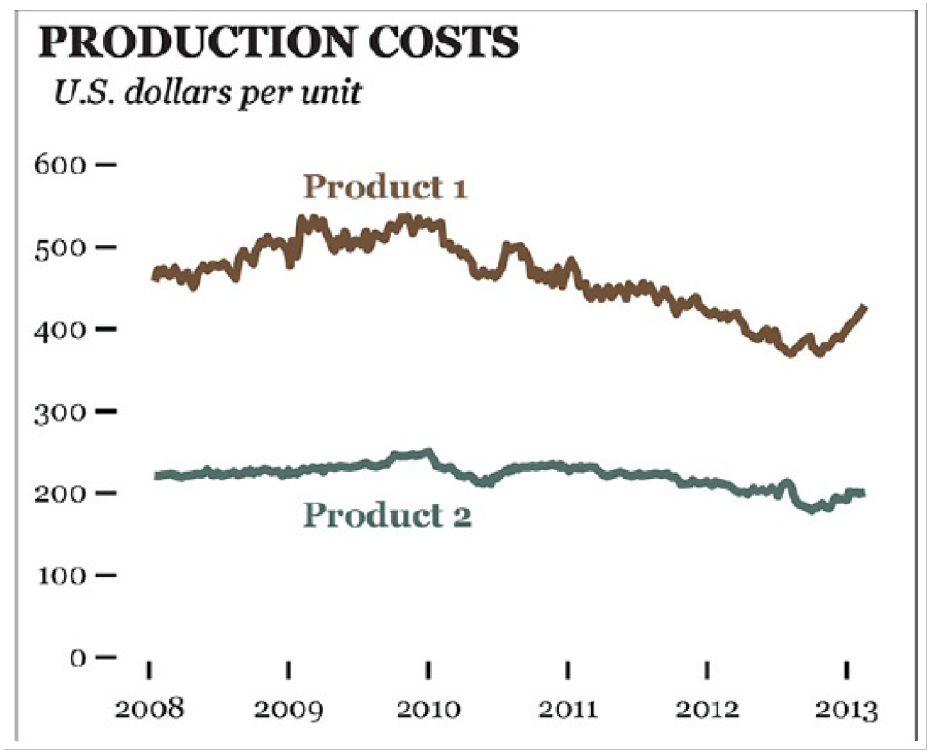
Introduction

We are desperate for correlations. From: The Truthful Art



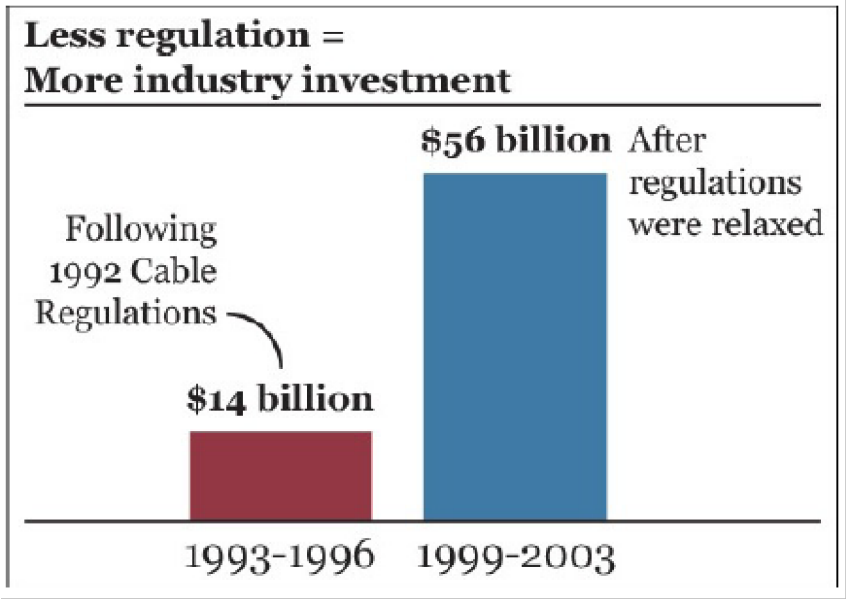
Introduction

We are desperate for correlations. From: The Truthful Art



Truthful

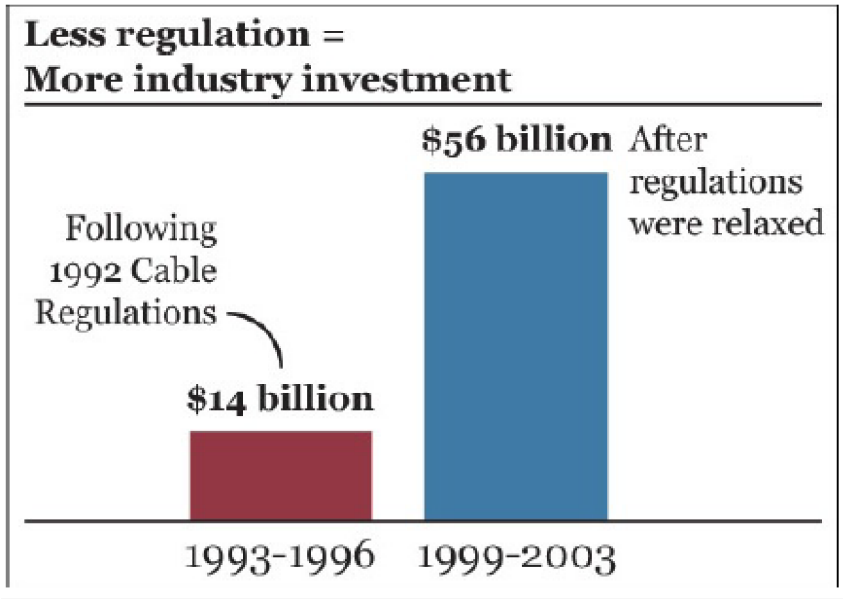
From: The Truthful Art

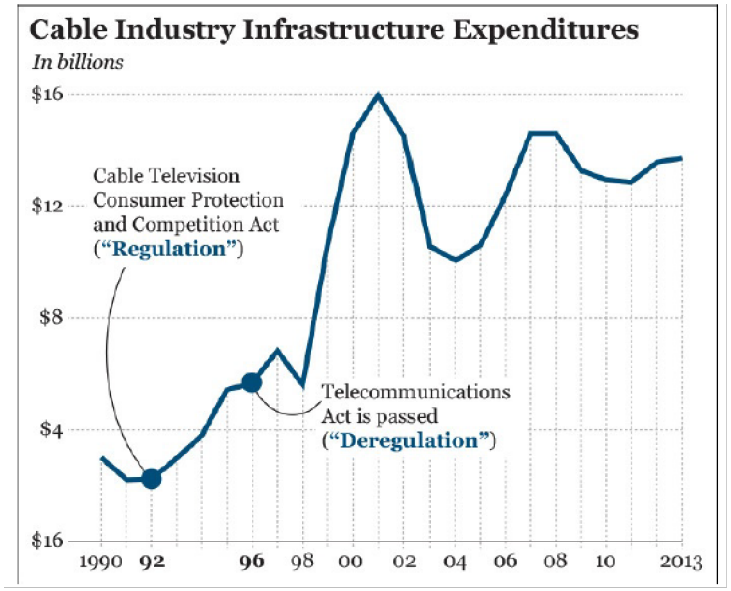


What is problematic with this plot?

Truthful

From: The Truthful Art





Is governmental regulation really the problem?

Or just ugly…

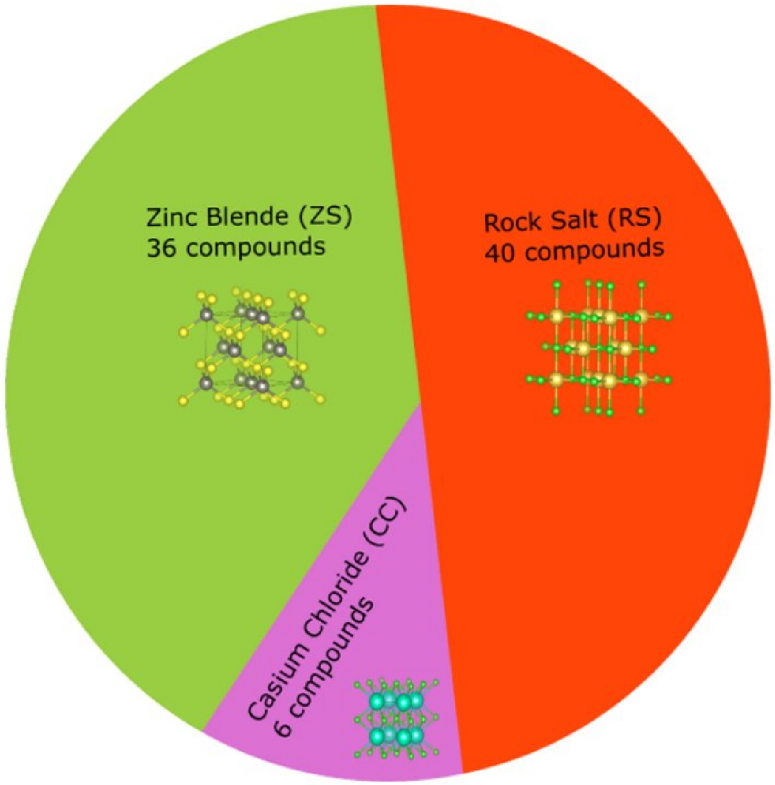
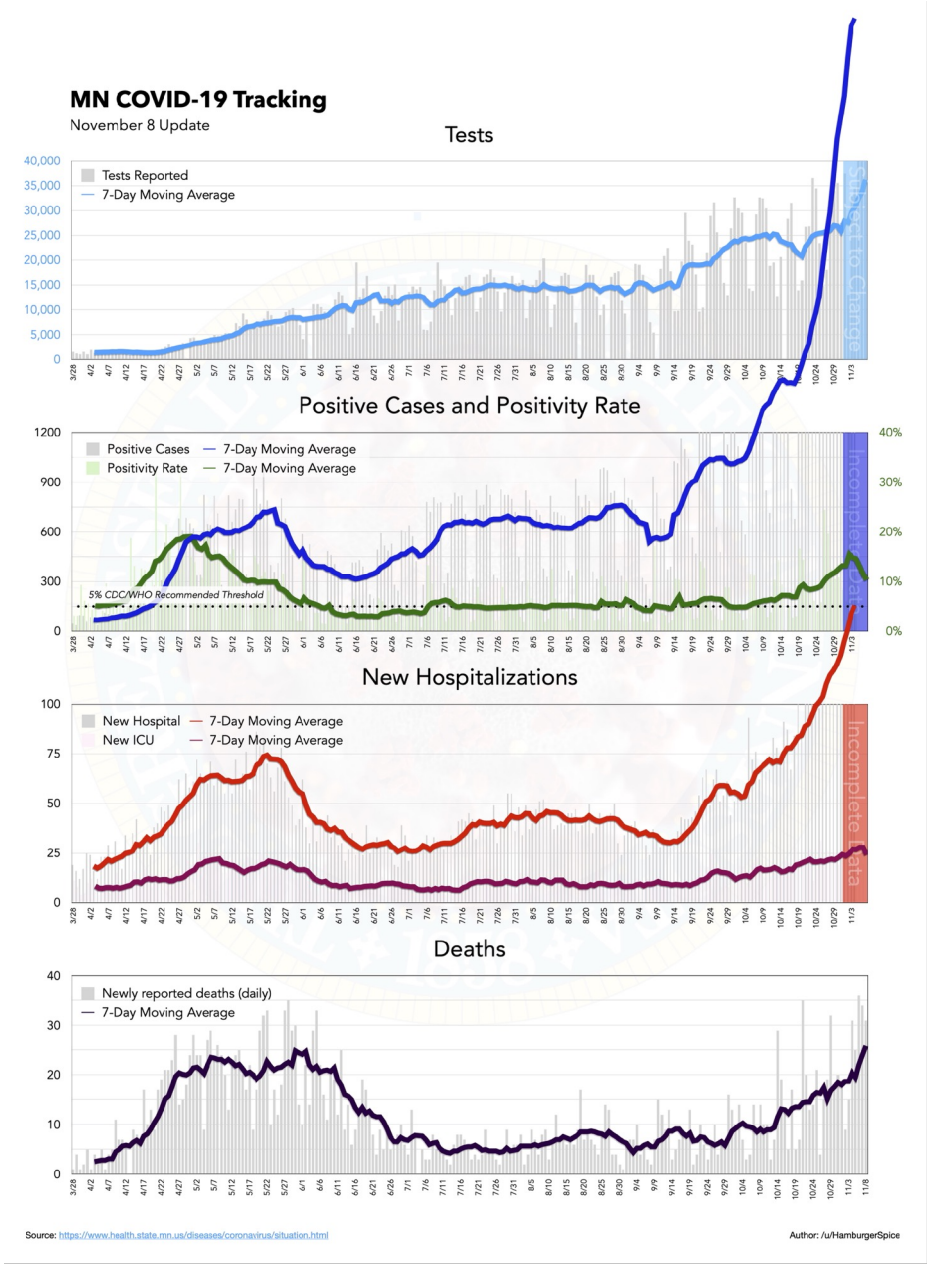


Figure 2. Distribution of crystal structures of the 82 dielectric materials[20] explored in this work. Each material in the dataset has one

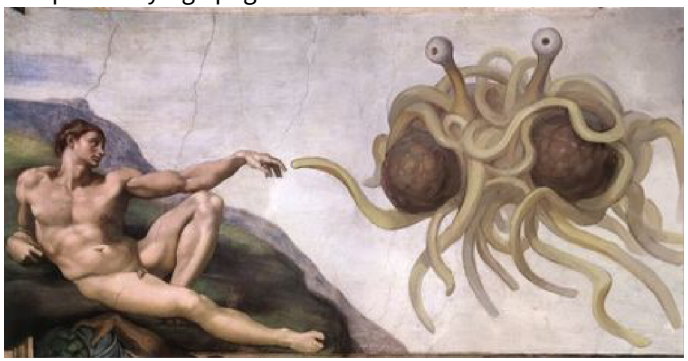
But sometimes breaking the rules might work...


MN COVID-19 Tracking
November 8 Update
Tests
Tests Reported
7-Day Moving Average
Subject to Change
Positive Cases and Positivity Rate
Positive Cases
7-Day Moving Average
Positivity Rate
7-Day Moving Average
5% CDC/WHO Recommended Threshold
Incomplete Data
New Hospitalizations
New Hospital
7-Day Moving Average
New ICU
7-Day Moving Average
Incomplete Data
Deaths
Newly reported deaths (daily)
7-Day Moving Average
Source: https://www.health.state.mn.us/diseases/coronavirus/situation.html
Author: /u/HamburgerSpice

Introduction

Additional Reading

- TED Talk – How to spot *a* misleading graph
- http://www.infoworld.com/article/2611729/big-data/big-data-without-good-analytics-can-lead-to-bad-decisions.html
- http://flowingdata.com/category/statistics/mistaken-data/
- CARGO CULT SCIENCE by Richard Feynman
- Open Letter To Kansas School Board
  Wikipedia: Flying Spaghetti Monster



- The Truthful Art by Alberto Cairo (ISBN 13: 9780321934079)
- https://www.callingbullshit.org/videos.html

Init