

# Introduction to Telling Stories with Data

Author: Nicholas G Reich

*This material is part of the **statsTeachR** project*

*Made available under the Creative Commons Attribution-ShareAlike 3.0 Unported  
License: [http://creativecommons.org/licenses/by-sa/3.0/deed.en\\_US](http://creativecommons.org/licenses/by-sa/3.0/deed.en_US)*

*Newton showed that the book of nature is written in the language of mathematics. Some chapters ... boil down to a clear-cut equation; but scholars who attempted to reduce biology, economics, and psychology to neat Newtonian equations have discovered that these fields have a level of complexity that makes such an aspiration futile.*

*This did not mean, however, that they gave up on mathematics.*

**A new branch of mathematics was developed over the last 200 years to deal with the more complex aspects of reality: statistics.**

- Yuval Noah Harari

*Sapiens: A Brief History of Humankind*

# Communicating ideas with evidence

## What is a narrative? [From the OED]

*An account of a series of events, facts, etc., given in order and with the establishing of connections between them; a narration, a story, an account.*

## What is data? [From Google: literally, “what is data”]

### da·tum

/ˈdātəm, ˈdætəm/ ⓘ

*noun*

plural noun: data

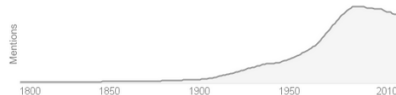
1. a piece of information.
  - an assumption or premise from which inferences may be drawn.
2. a fixed starting point of a scale or operation.

#### Origin



mid 18th century: from Latin, literally 'something given,' neuter past participle of *dare* 'give.'

#### Use over time for: data



# Uber Is Taking Millions Of Manhattan Rides Away From Taxis

The ride-share service probably isn't increasing congestion.

By REUBEN FISCHER-BAUM and CARL BIALIK

## Are Ubers Supplementing Or Replacing Cabs?

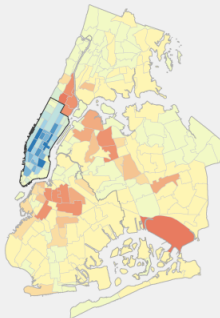
Change in number of Uber and taxi pickups by taxi zone, April-June 2014 versus April-June 2015



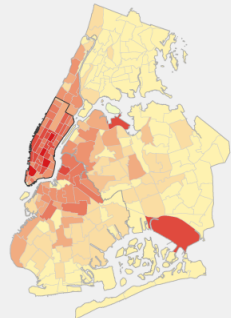
CABS + UBER



CABS ONLY



UBER ONLY





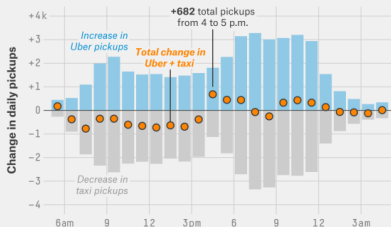
TRANSPORTATION | 11:19 AM | DEC 9, 2015

## Is Uber Making NYC Rush-Hour Traffic Worse?

By CARL BIALIK, REUBEN FISCHER-BAUM and DHRUMIL MEHTA

### Uber adds (a little) to Manhattan evening rush

Average change in Uber pickups, taxi pickups, and total Uber + taxi pickups by hour of day; April-June 2014 vs. April-June 2015



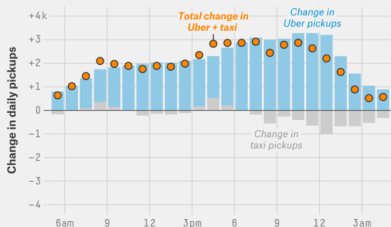
Non-holiday weekdays; Manhattan core taxi zones only

FIVETHIRTYEIGHT

SOURCE: NYC TAXI & LIMOUSINE COMMISSION

### Thousands of new pickups outside Manhattan core

Average change in Uber pickups, taxi pickups, and total Uber + taxi pickups by hour of day; April-June 2014 vs. April-June 2015



Non-holiday weekdays; includes Manhattan taxi zones outside the core

FIVETHIRTYEIGHT

SOURCE: NYC TAXI & LIMOUSINE COMMISSION

# When Teamwork Doesn't Work for Women

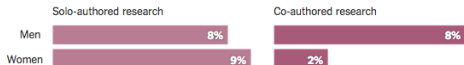
## Economic View

By **JUSTIN WOLFERS** JAN. 8, 2016

### Getting Credit Where Credit Is Due

Male and female economists are given roughly equal credit for work they perform alone, but in group work, women receive far less credit.

#### Consequences of writing one more paper on the probability of earning tenure

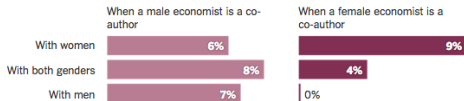


Source: Heather Sarsons, "Gender Differences in Recognition for Group Work"

### Who Gets the Credit for Collaboration?

Women get full credit, in terms of earning tenure, only when writing papers with other women. Writing one with a man has no impact on the female author, only the male.

#### Effect of writing an additional paper on the probability of earning tenure



Source: Heather Sarsons, "Gender Differences in Recognition for Group Work"

# Trump's tweets: from varianceexplained.org



**Todd Vaziri**

@tvaziri



Follow

Every non-hyperbolic tweet is from iPhone (his staff).

Every hyperbolic tweet is from Android (from him).



RETWEETS

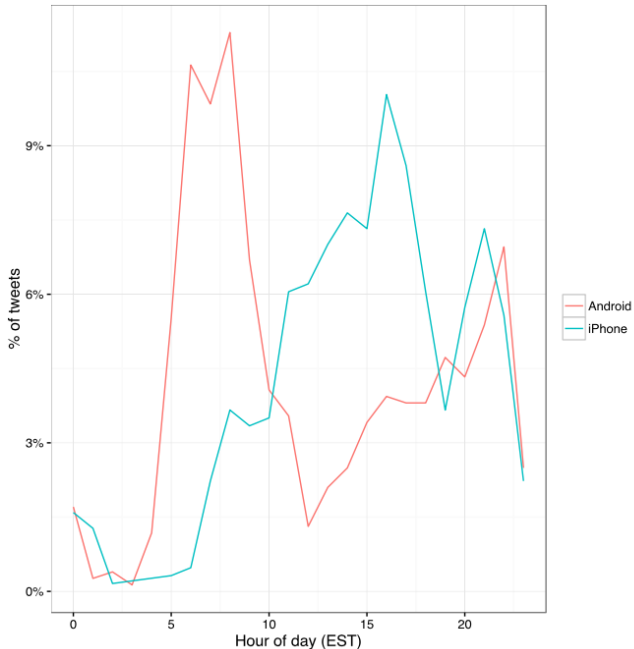
**10,739**

LIKES

**14,905**

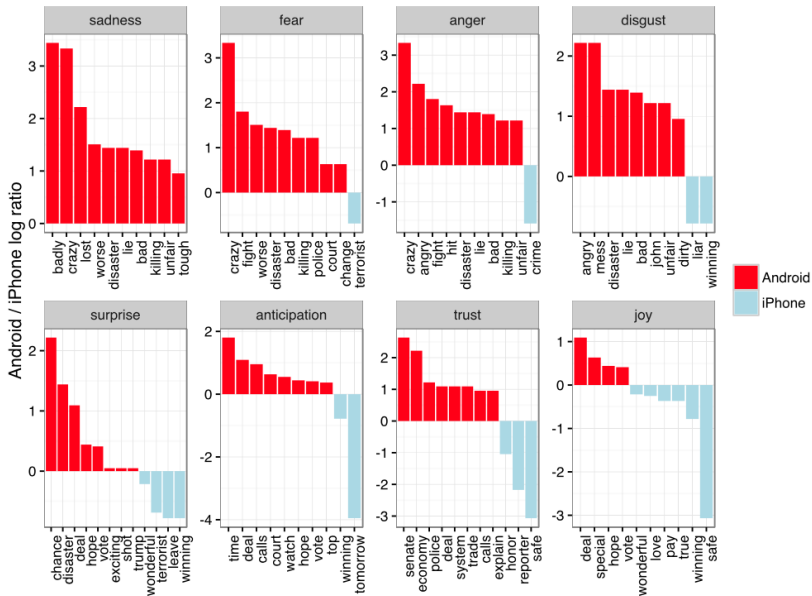


# Trump's tweets: from [varianceexplained.org](https://varianceexplained.org)





# Trump's tweets: from varianceexplained.org

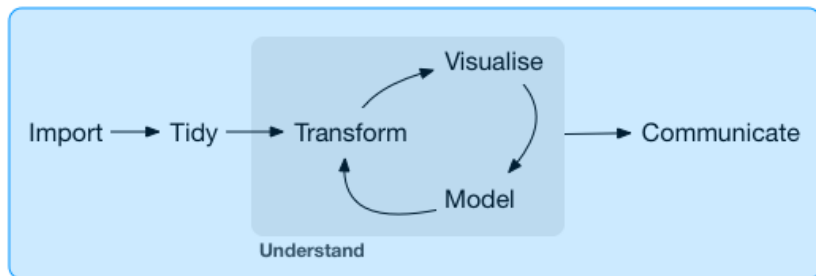


# How to tell a story using data

Telling stories with data requires

- ▶ detective work
- ▶ creativity, both scientific and artistic
- ▶ experimentation with different storylines
- ▶ good data, (good data does not necessarily equal “big data”)

# The tidy-verse: a process for data analysis

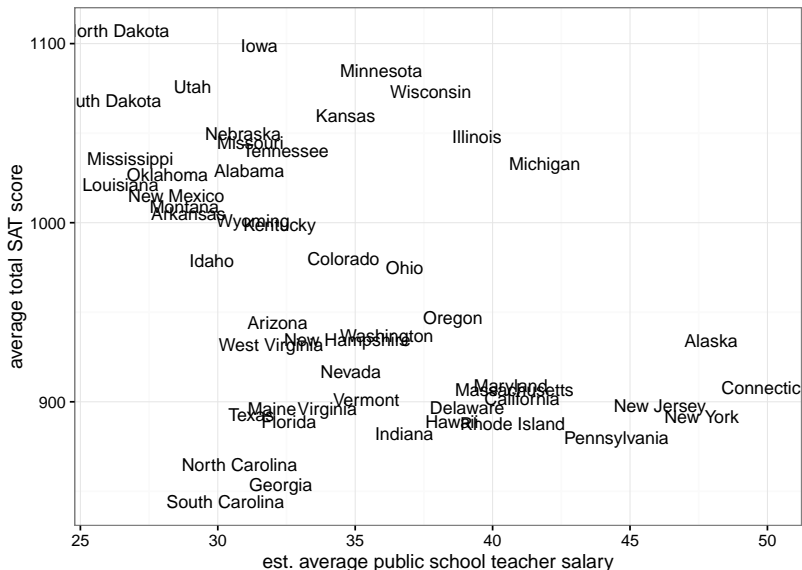


Program

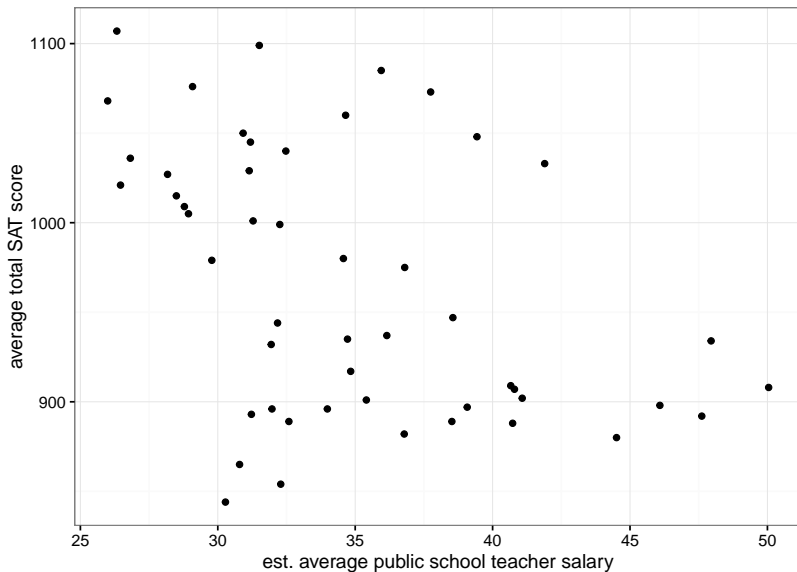
# A common modeling tool: regression

- The goal is to learn about the relationship between two variables: a “covariate” (or “predictor”) of interest and an “outcome” of interest.
  - Some models focus on prediction.
  - Other models focus on description.
- Regression is an exercise in inferential statistics: we are drawing evidence and conclusions from data about “complex aspects of reality”, i.e. “noisy” systems.

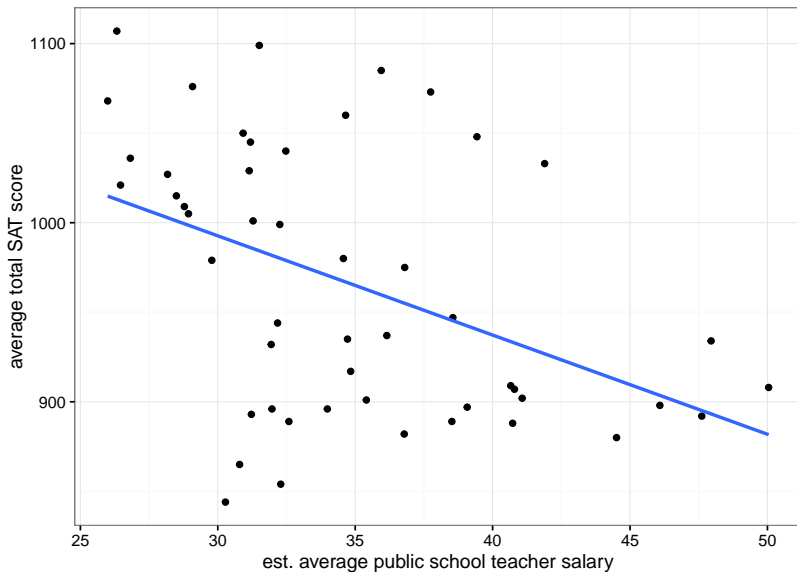
# State-level SAT score data (1994-95)



# State-level SAT score data (1994-95)



# State-level SAT score data (1994-95)



# The SAT example

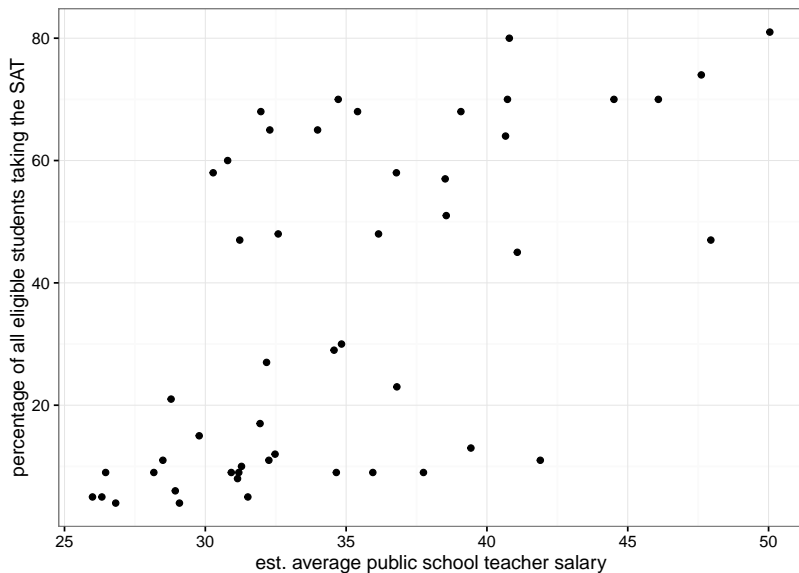
What is the outcome variable?

What is the covariate or predictor variable?

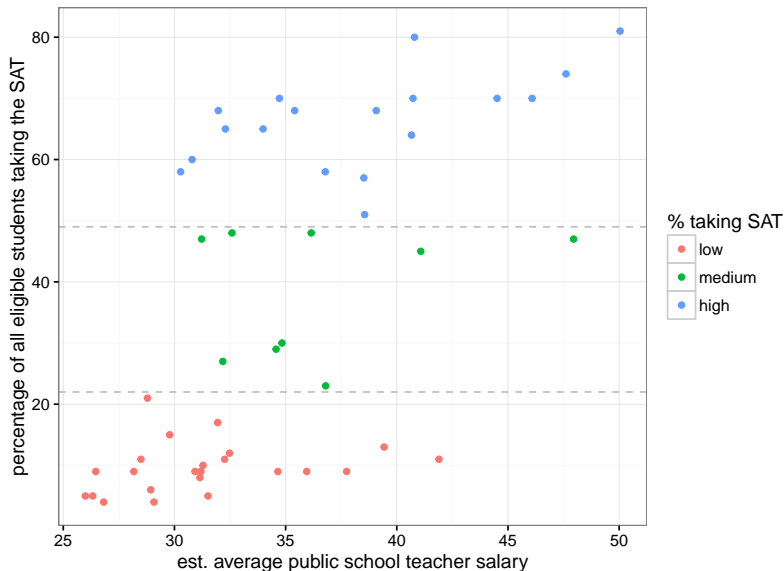
What other data might be part of this story?



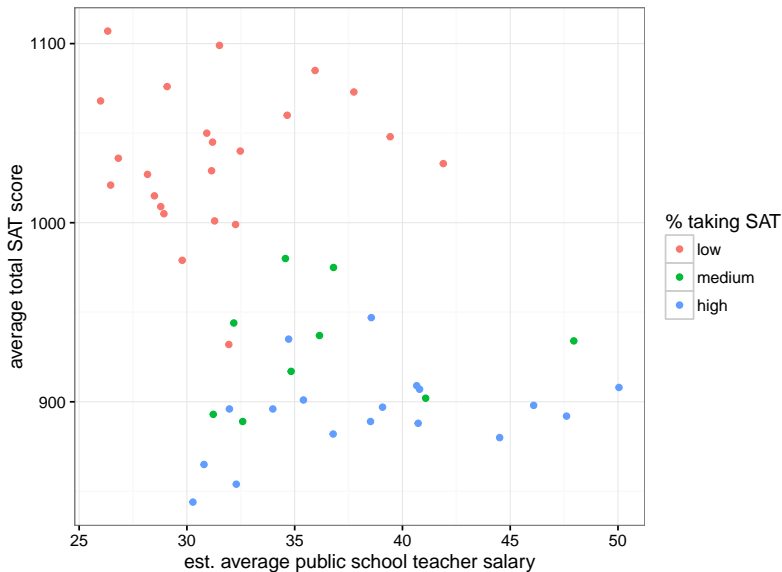
# State-level SAT score data (1994-95)



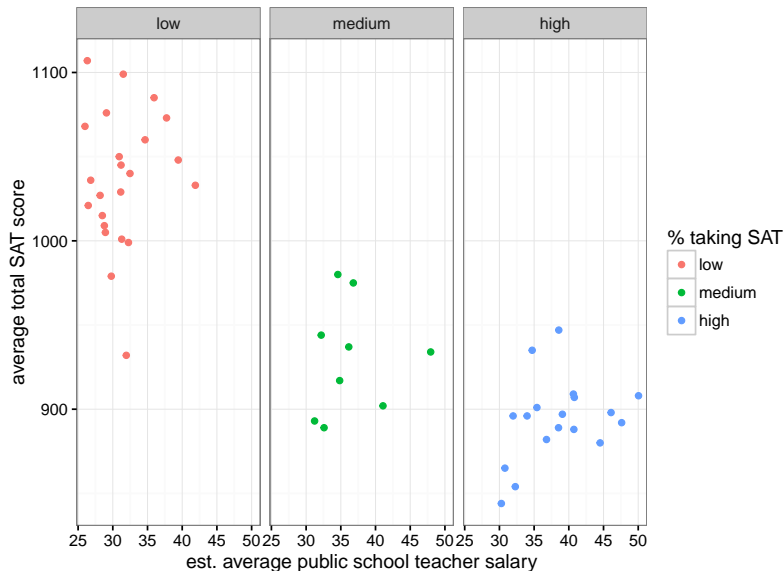
# State-level SAT score data (1994-95)



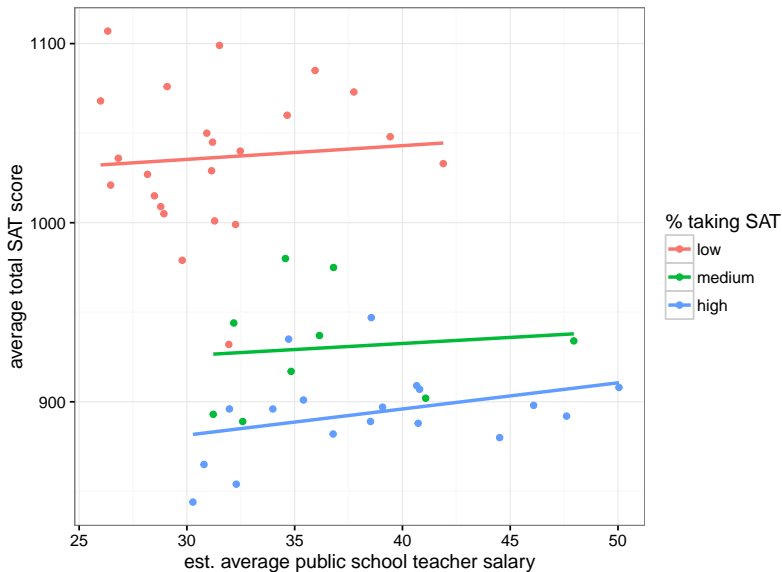
# State-level SAT score data (1994-95)



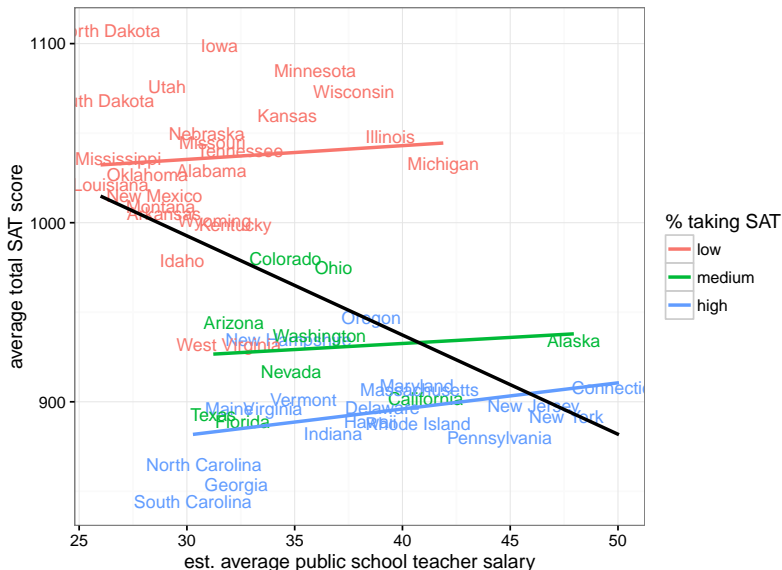
# State-level SAT score data (1994-95)



# State-level SAT score data (1994-95)



# State-level SAT score data (1994-95)



## State-level SAT score data (1994-95)

What can we conclude from all of this? (BTW, this is an example of "Simpson's Paradox".)

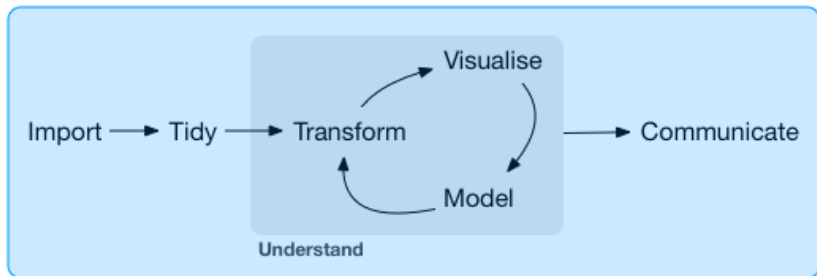
# Regression modeling

The process of using data to describe the relationship between outcomes and predictors is called modeling.

- Models are models, not reality.
- “All models are wrong, but some are useful.”
- Introduce structure to our model that balances realism with “goodness of fit” .

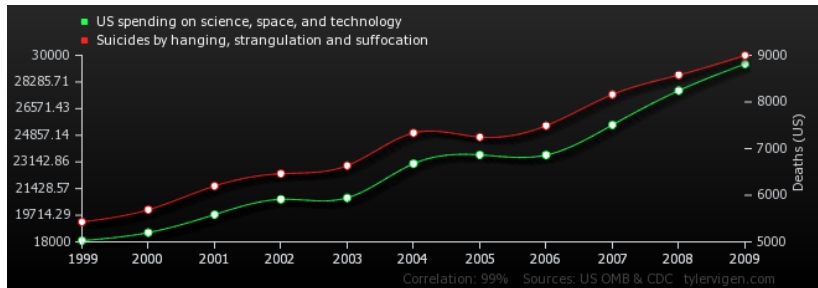


# Things to come

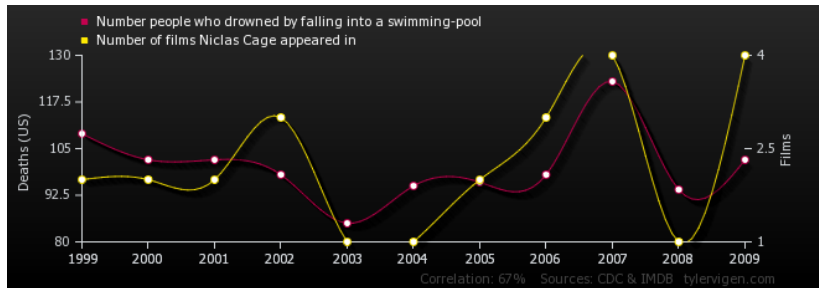


Program

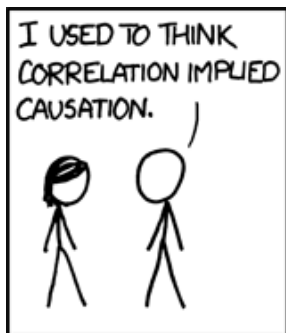
# Beware of correlation!



# Beware of correlation!



Hat tip to [www.tylervigen.com](http://www.tylervigen.com)



<https://xkcd.com/552/>