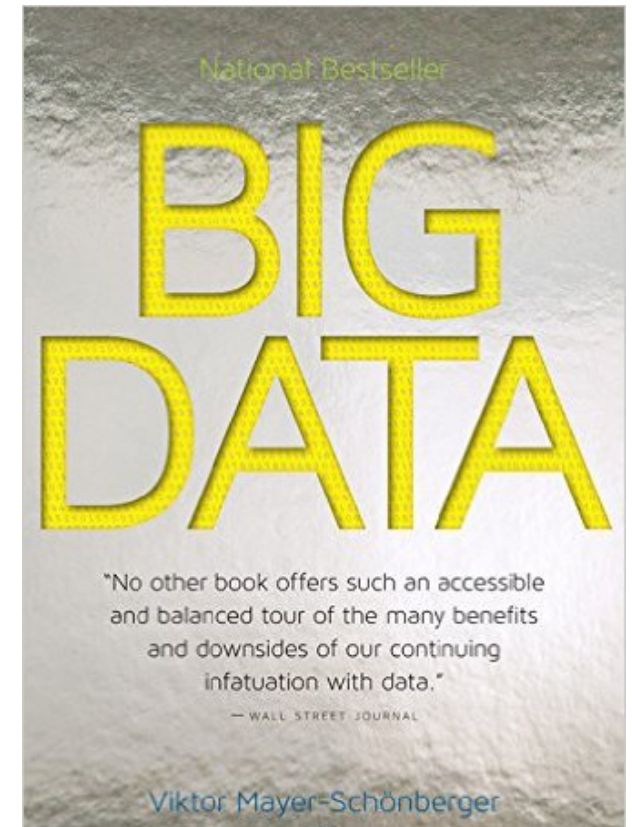# Lecture 5: Correlation & More SQL

Core 109S IDWT?, Spring 2017
Michael Hay

Discussion

- Book

  - The book describes three shifts in mindset in Ch. 2-4.

  - What is the shift described in Ch. 4 "Correlation"? (Find good quotes to support your answers!)

  - Example from chapter?

  - What did you find interesting?

  - What do you think of Chris Anderson's bold prediction?

National Bestseller

BIG DATA

"No other book offers such an accessible and balanced tour of the many benefits and downsides of our continuing infatuation with data."
— WALL STREET JOURNAL

Viktor Mayer-Schönberger

CHRIS ANDERSON   SCIENCE   06.23.08   12:00 PM

THE END OF THEORY: THE DATA DELUGE MAKES THE SCIENTIFIC METHOD OBSOLETE
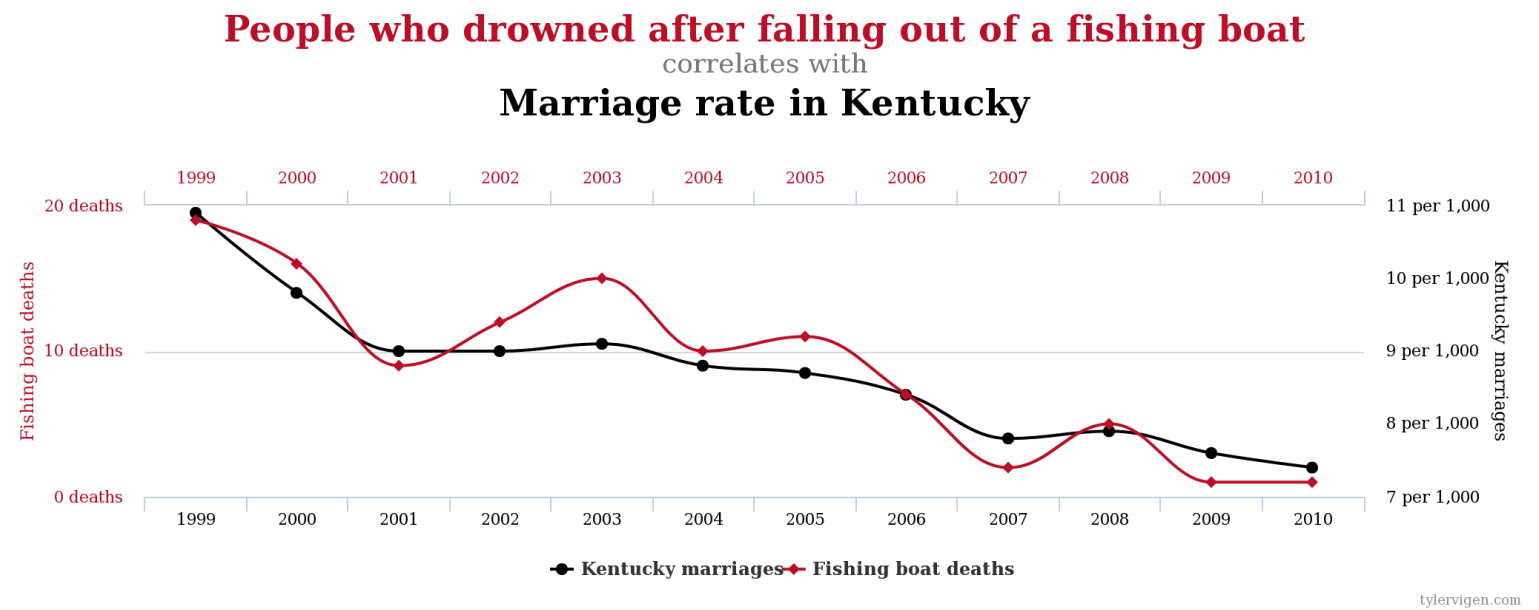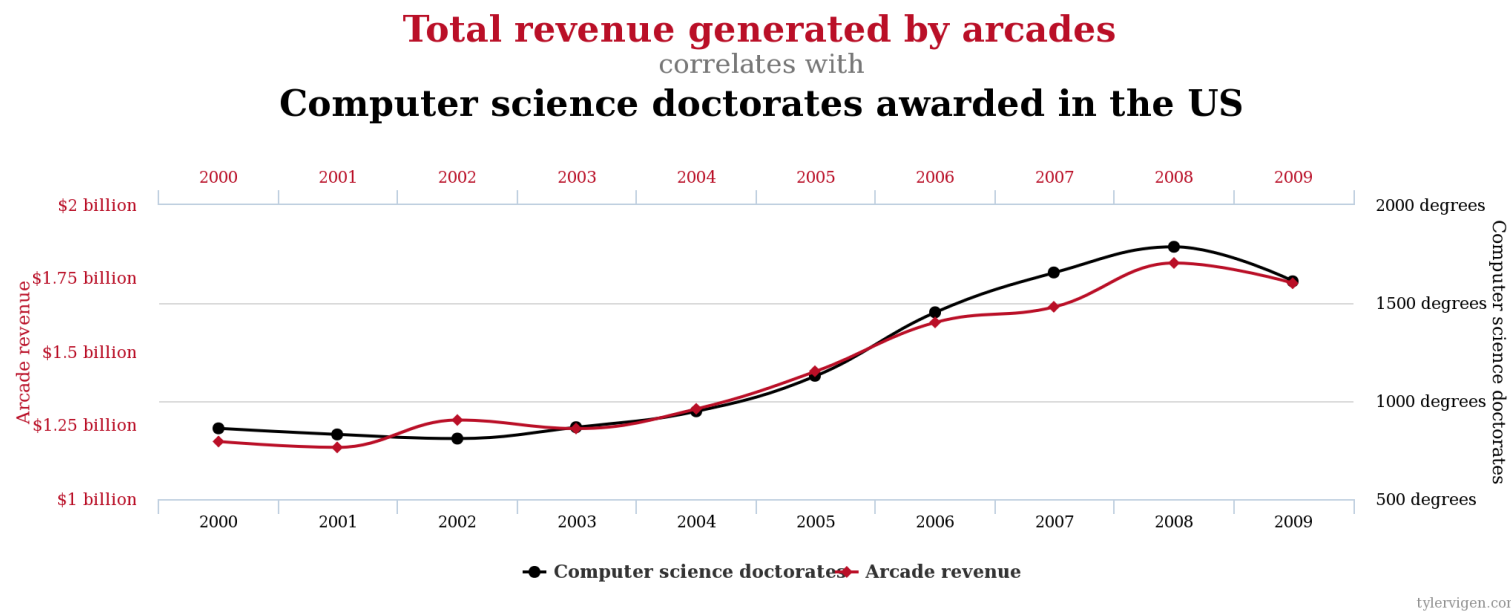

Illustration: Marian Bantjes

"**All models are wrong,** but some are useful."

This article has been reproduced in a new format and may be missing content or contain faulty links. Contact wiredlabs@wired.com to re-

https://www.wired.com/2008/06/pb-theory/

Correlation vs. causation

Total revenue generated by arcades
correlates with
Computer science doctorates awarded in the US

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009

Arcade revenue: $2 billion, $1.75 billion, $1.5 billion, $1.25 billion, $1 billion

Computer science doctorates: 2000 degrees, 1500 degrees, 1000 degrees, 500 degrees

2000 2001 2002 2003 2004 2005 2006 2007 2008 2009

— Computer science doctorates — Arcade revenue

tylervigen.com

People who drowned after falling out of a fishing boat
correlates with
Marriage rate in Kentucky

1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010

Fishing boat deaths: 20 deaths, 10 deaths, 0 deaths

Kentucky marriages: 11 per 1,000, 10 per 1,000, 9 per 1,000, 8 per 1,000, 7 per 1,000

1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010

— Kentucky marriages — Fishing boat deaths

tylervigen.com

http://www.tylervigen.com/spurious-correlations

3

SQL Introduction

- See lecture4handout.pdf and lecture5handout.pdf.

Exercise 1

- Write a query that returns college names and their enrollments for those colleges that begin with the letter "C" and end with the leter "e". Rename the "cname" attribute as "greatschool."

greatschool	enrollment
Colgate	2700

5

Exercise 2

- Write a query that returns the name, gpa, high school size of students with a GPA of more than 3.6 or from small high schools (those with fewer than 1000 students).

Exercise 3

- Names of colleges who have applications for either biology or CS majors.

  Could the same college name appear more than once in the result? Why? We can eliminate by replacing "SELECT" with "SELECT DISTINCT"