STAT 5600 & STAT 6600
Fall 2023
Probability and Statistics for Data Science

Lecture 0: Preliminaries and Logistics

Instructor: Elvan Ceyhan

Office: Extension Hall 203

Emails: ezc0066@auburn.edu

ceyhan@auburn.edu

Outline

1. Logistics
    - Course info
    - Lectures
    - Course webpage
    - Office hours
    - Grading

2. Syllabus

Course Info

- New course (almost)
- Looks at prob and stat topics from a DS perspective
  - Probability theory
  - Statistical inference
  - DS techniques
- MS level course
  - Probability Theory
  - Random Variables
  - Stochastic Processes
  - Statistical Inference
  - Hypothesis Testing
  - Regression and Time Series Analysis

Course Info

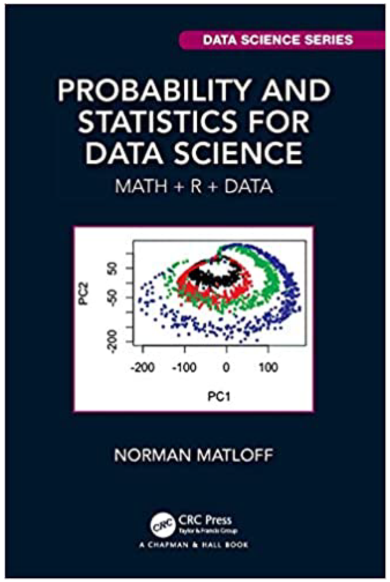
- Prerequisites:
  - Probability and Statistics at STAT 3600 and 3610 levels
    - Will greatly help (not absolutely necessary)
    - Students are presumed to know calculus and some linear algebra
    - No advanced knowledge of prob & stat assumed or required (though it helps)
  - Basic CS background
    - We will use R (and possibly Python)

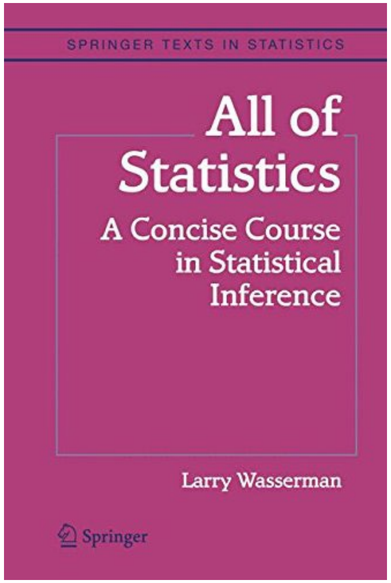
Course Info

- This is NOT a computational course or a computer science course
- More of a theory + methods course
  - ➢ Class is fast-paced & demanding
  - ➢ Many results without proof
- Course website: Under Canvas:
  - ➢ This is your best resource
  - ➢ Will be regularly updated

Course Info

RECOMMENDED




SPRINGER TEXTS IN STATISTICS
All of Statistics
A Concise Course in Statistical Inference
Larry Wasserman
Springer

RECOMMENDED

Prob & Stat for
Data Science (PSDS)
by
Carlos Fernandez-Granda
(freely available
online)

- Software: R or RStudio (or Python)
- Available from CRAN or RStudio websites

Lectures

- Tu-Th 3:30 pm - 4:45 pm
- ACLC (Academic Classroom/Lab Complex) 010
  - ➢ Slides + Annotations
  - ➢ Occasionally some programming (*R* and maybe *Python*)
    - ➢ Posted on Canvas before or after class
  - ➢ May have cancellations due to weather or unavailability or …
    - ➢ Will be emailed and updated on Canvas and/or email.
    - ➢ Weather- or emergency-related class cancellations decided by AU.

In-class

- Interactive (please)
- Have a book (or soft copy of it)
- Please mute your phones
- No audio/video recording allowed (unless required)

- Attendance is not mandatory but strongly encouraged.
  - If you are present and active in class, I will notice.
  - If you are absent, I will notice.
  - If you are present and inactive, I will notice that too! ☺
  - May help to bump your grade if you are on the border.

Office Hours

- Tuesday 12:00 pm (noon) – 1:00 pm or by appointment
  - Tentative
  - May revise - visit after add/drop date
- TA and TA Office hours:
  - TBD

Grading

- 20% assignments
- 80% exams (in-class mid-terms and a final)
- 10% project - bonus (in groups, if time allows, more on this later)
- *Some parts are tentative!*

Grading – HW Assignments

- 20% Assignments
  - ~5-7 assignments (~ once every 2 weeks)
    - ~5-8 problems per assignment
    - Later assignments may have more programming
  - Collaboration is allowed
    - Acknowledge help received or collaboration (i.e., joint work).
    - DO NOT DIRECTLY COPY FROM OTHERS or INTERNET!

Grading – HW Assignments

- Assignment questions will be based on lectures
  - But harder than examples done in class
  - Will require some effort, helps to discuss with classmates
  - Okay to discuss with your classmates, but need to acknowledge

- Assignments due **at the beginning of** class
  - NO LATE SUBMISSIONS
  - Hard-copies only (typed/hand-written, scanned versions for the D01 section)

Grading - Exams

- 80% exams
  - Mid-terms 1 and 2
    - 25% mid-term 1 (prob & stat), Sept end/early Oct
    - 25% mid-term 2 (inference), mid-Nov
    - Non-cumulative (i.e., non-overlapping)
    - 30% Final Exam (cumulative)
  - In-class exams
    - Somewhat easier than assignments
    - Based on material/examples covered in lectures (attend!)
    - No collaborations, obviously
    - Closed-book, closed-notes
    - 75 mins

Grading – Mini Group-projects

- 10% Group mini-project (bonus, if time permits)
- Basically, last assignment, due at end of semester

- Data analysis project
  - Programming involved
  - Same as assignment group
  - Can start early, in the 2nd half of the semester
  - Will discuss details as we go along

Grading - Recap

- 20% assignments
- 80% exams (in-class mid-terms)
- 10% mini group-projects (bonus)

- Will try to provide mid-semester grades (after MT1)
  - ➢ For self-evaluation purposes only

Syllabus

Probability Theory (8-10 lectures)

- Probability review (events, computing probability, conditional prob., Bayes' Thm.)
- Random variables (Geometric, Exponential, Normal, expectation, moments, etc.)
- Probability inequalities (Markov's, Chebyshev's, Central Limit Thm., etc.)
- Markov chains (stochastic processes, balance equations, etc.)

MID-TERM 1 (Late September - Early October)

Statistical Inference (8-10 lectures)

- Non-parametric inference (empirical PDF, bias, kernel density, plug-in estimator)
- Confidence intervals (percentiles, Normal-based CIs)
- Parametric inference (method of moments, max likelihood estimator)
- Hypothesis testing (Wald's test, t-test, KS test, p-values, permutation test)
- Bayesian inference (Bayesian reasoning, inference, etc.)

MID-TERM 2 (Mid-November)

Data Science Models (3-5 lectures)

- Regression (simple LR, multiple LR, non-linear regression)
- Time series analysis (moving average, EWMA, AR, ARMA, ARIMA)

MINI-PROJECT (Due late November - early December)

STAT 5600 & STAT 6600, Fall 2023
Probability and Statistics for Data Science

What **is Data Science (DS)?**

Analysis of data (using several tools/techniques)

Statistics/Data Analysis + Computer Science (CS) + ???

Who **is a Data Scientist?**

Someone who is better at stats than the average CS person
and
someone who is better at CS than an average statistician.

What is Data Science?

- **Definition:** Using data to answer questions
- Statistics, Data Mining, and Machine Learning (ML) are all concerned with collecting and analyzing data
- Data Science is a combination of software programming, statistics, and storytelling
- Broad field involving statistics, computer science, mathematics

- Includes data cleaning, formatting, and visualization
- Previously, stat research was in stat departments and data mining and ML were in Comp Sci departments

- An Economist Special Report sums up this mélange of skills well - they state that a data scientist is broadly defined as someone:
- "who combines the skills of software programmer, statistician, and storyteller slash artist to extract the nuggets of gold hidden under mountains of data"

- And by the end of courses in the DS program, hopefully you will feel equipped to do just that!

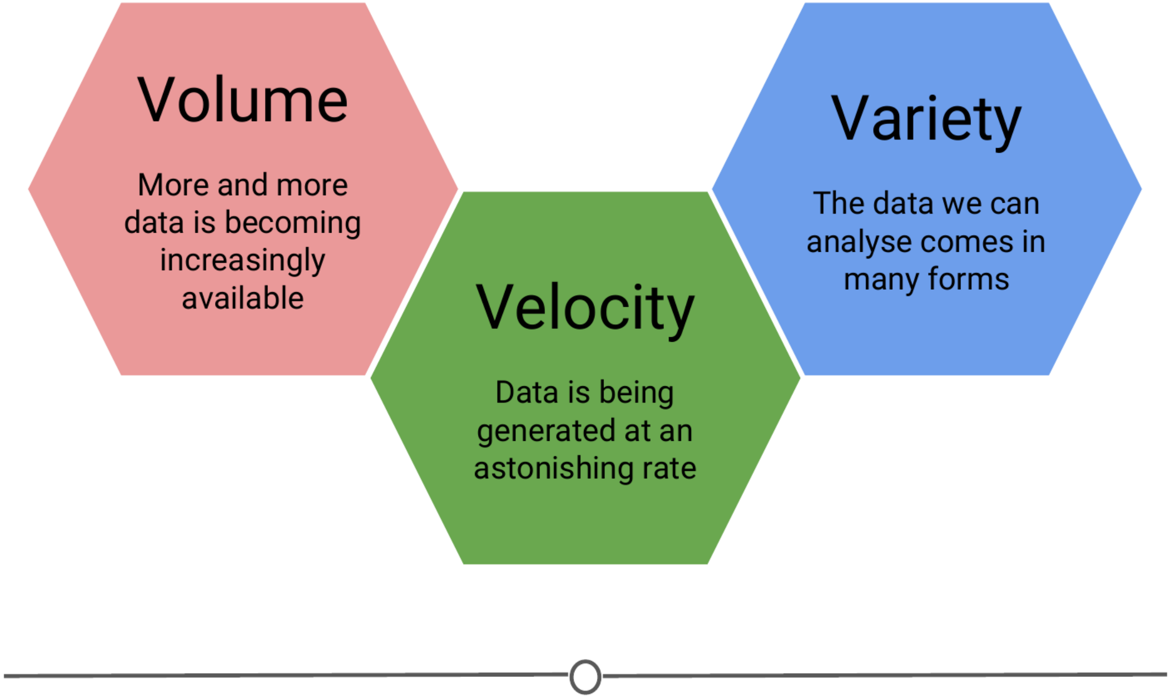
Why Data Science?

- Rise of data science due to vast available and generated data
- Availability of inexpensive computing
- Perfect storm: Rich data + Analytical tools
- Exponential growth of data generation;
  more data than ever before

- There is a little example that illustrates the remarkably rapid expansion of data creation that we are presently undergoing. Around the 3rd century BC, it was thought that the Library of Alexandria contained all the human knowledge. In today's context, the world contains such a vast amount of information that each living person could possess 320 times more information than that (historians believe) was kept in the entire collection of Alexandria.
- And that is still growing.

Big Data Characteristics

1. **Volume:** Large datasets (e.g., YouTube's 300 hours of video/minute)
2. **Velocity:** Rapid data generation and collection (real-time GPS data)
3. **Variety:** Different types of data (structured and unstructured)



Volume

More and more data is becoming increasingly available

Velocity

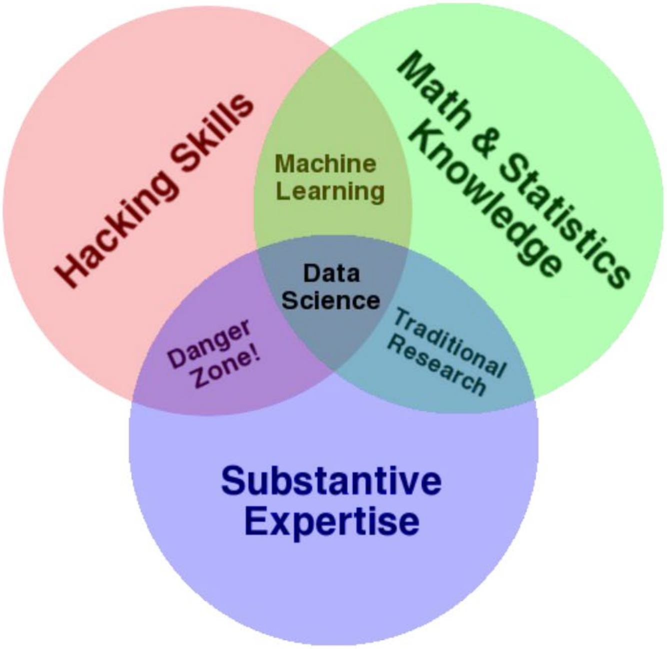
Data is being generated at an astonishing rate

Variety

The data we can analyse comes in many forms

Who is a Data Scientist?

- **Definition:** Someone who uses data to answer questions
- Venn diagram by Drew Conway: Substantive expertise, hacking skills, math & statistics
- *Substantive expertise:* Formulate questions, select relevant data
- *Hacking skills:* Data cleaning, formatting, programming
- *Math & statistics:* Data analysis, deriving insights



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Why Choose Data Science?

- High demand for data science skills
- Machine learning engineers, data scientists, big data engineers among top emerging jobs
- Data scientist roles have grown over 650% since 2012
- Demand exceeds supply; Opportunities in diverse sectors

- Additionally, according to Glassdoor, in which they ranked the top 50 best jobs in America, Data Scientist is **THE top job** in the US in 2017, based on job satisfaction, salary, and demand.

Examples of Famous Data Scientists

1.Daryl Morey - Sports (Houston Rockets GM)

2. Hilary Mason - Social media analysis

3. Nate Silver - FiveThirtyEight (election predictions)

He uses large amounts of totally free public data to make predictions about a variety of topics;
most notably he makes predictions about who will win elections in the United States and has a remarkable track record for accuracy doing so.

Data Science in Action

- Google's flu outbreak analysis: Analyzed search terms vs. CDC flu data
- Identified correlations using big data
- Predicted flu outbreaks based on search terms

Conclusion

- Data science is about using data to answer questions
- **Key skills:** Substantive expertise, hacking skills, math & statistics
- High demand and diverse opportunities in data science
- Exciting journey ahead; Let's get started!

Data & Statistics

- Definition and exploration of "data" in the context of data science.
- Review of previous discussions on data science.
- Definitions of "Data“
- *Cambridge English Dictionary:* Data is information, including facts or numbers, collected for decision-making.
- *Wikipedia:* Data is a set of qualitative or quantitative values.
- Notice the differences in Definitions!
- Cambridge definition emphasizes actions around data (collection, examination, decision-making).
- Wikipedia definition focuses on data as a set of values.

Variables

- To have data, a set of values is required.
- This set, often called a **population** in statistics, forms the basis of measurement.
- **Variables:** Variables are measurements or characteristics of items.
- Examples include measuring height, time on a website, or qualitative traits.

population = whole set of item

Variable : measurements on items.

Qualitative and Quantitative Variables

- **Qualitative variables:** Information about qualities (e.g., (country of) origin, sex).
- **Quantitative variables:** Information about quantities (e.g., height, weight).
- Qualitative described by words, quantitative by numbers on a scale.

- Summary of Wikipedia Definition:
- Measurements (qualitative/quantitative) on a set of items constitute data.

Data Presentation

"tidy data" movement

- Structured dataset example: tidy data, rectangular, spreadsheet with variables (e.g., origin, sex, height, weight).

Hadley Wickam

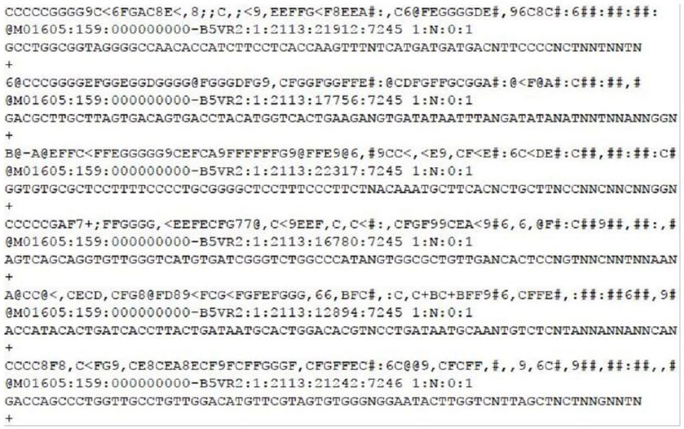
Name	Country of origin	Sex	Weight (kg)	Height (cm)
A. Bee	Canada	M	75	163
C. Dee	UAE	M	80	180
E. Eff	China	F	72	175
G. Haitch	South Africa	F	68	172
I. Jay	Poland	M	77	168
K. Elle	Japan	N/A	76	173
M. Enn	Chile	M	80	190

Messy Data

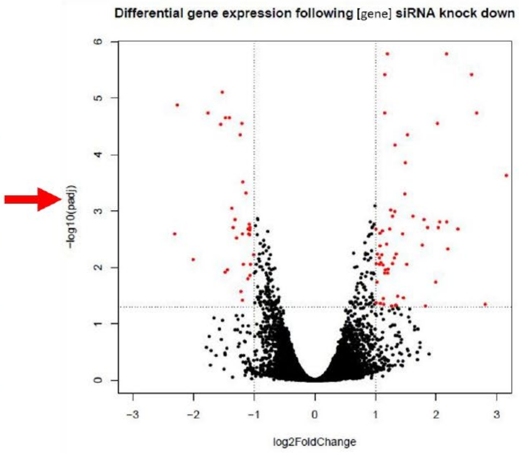
- Importance of extracting and organizing information from messy data.
- Commonly encountered messy data sources in data science:
- (Genome) sequencing data
- Population census data
- Electronic medical records (EMR), other large databases
- Geographic information system (GIS) data (mapping)
- Image analysis and image extrapolation
- Language and translations
- Website traffic
- Personal/Ad data (e.g.: Facebook, Netflix predictions, etc.)

Messy Data Example - Sequencing

- One type of messy data is (genome) sequencing data. This data is generally first encountered in the FASTQ format, the raw file format produced by sequencing machines. These files are often hundreds of millions of lines long, and it is DS's job to parse this into an understandable and interpretable format and infer something about that individual's genome. Below, this data was interpreted into expression data, and produced a plot called a "volcano plot".

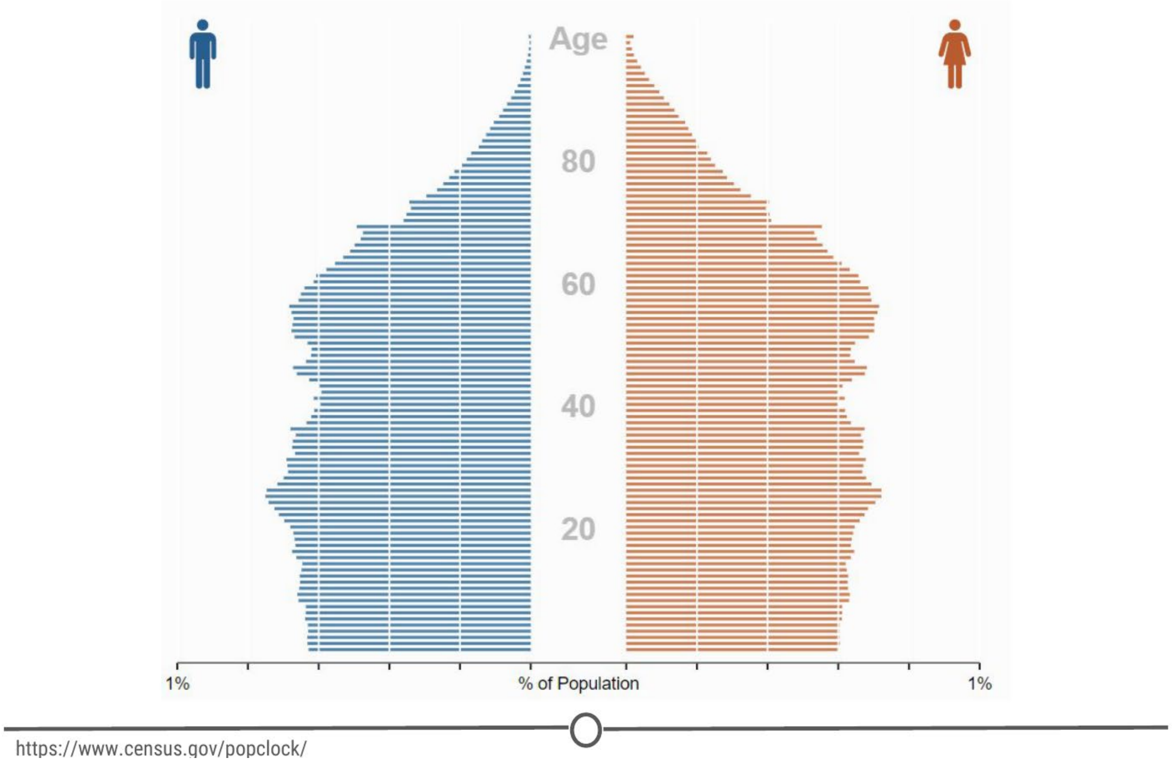


```
CCCCCGGGG9C<6FGAC8E<,8;;C,;<9,EEFFG<F8EEA#:,C6@FEGGGGDE#,96C8C#:6##:##:##:
@M01605:159:000000000-B5VR2:1:2113:21912:7245 1:N:0:1
GCCTGGCGGTAGGGGCCAACACCATCTTCCTCACCAAGTTTNTCATGATGATGACNTTCCCCNCTNNTNNTN
+
6@CCCGGGGEFGGEGGDGGGG@FGGGDFG9,CFGGFGGFFE#:@CDFGFFGCGGA#:@<F@A#:C##:##,#
@M01605:159:000000000-B5VR2:1:2113:17756:7245 1:N:0:1
GACGCTTGCTTAGTGACAGTGACCTACATGGTCACTGAAGANGTGATATAATTTANGATATANATNNTNNANNGGN
+
B@-A@EFFC<FFEGGGGG9CEFCA9FFFFFFG9@FFE9@6,#9CC<,<E9,CF<E#:6C<DE#:C##,##:##:C#
@M01605:159:000000000-B5VR2:1:2113:22317:7245 1:N:0:1
GGTGTGCGCTCCTTTTCCCCTGCGGGGCTCCTTTCCCTTCTNACAAATGCTTCACNCTGCTTNCCNNCNNCNNGGN
+
CCCCCGAF7+;FFGGGG,<EEFECFG77@,C<9EEF,C,C<#:,CFGF99CEA<9#6,6,@F#:C##9##,##:,#
@M01605:159:000000000-B5VR2:1:2113:16780:7245 1:N:0:1
AGTCAGCAGGTGTTGGGTCATGTGATCGGGTCTGGCCCATANGTGGCGCTGTTGANCACTCCNGTNNCNNTNNAAN
+
A@CC@<,CECD,CFG8@FD89<FCG<FGFEFGGG,66,BFC#,:C,C+BC+BFF9#6,CFFE#,:##:##6##,9#
@M01605:159:000000000-B5VR2:1:2113:12894:7245 1:N:0:1
ACCATACACTGATCACCTTACTGATAATGCACTGGACACGTNCCTGATAATGCAANTGTCTCNTANNANNANNCAN
+
CCCC8F8,C<FG9,CE8CEA8ECF9FCFFGGGF,CFGFFEC#:6C@@9,CFCFF,#,,9,6C#,9##,##:##,,#
@M01605:159:000000000-B5VR2:1:2113:21242:7246 1:N:0:1
GACCAGCCCTGGTTGCCTGTTGGACATGTTCGTAGTGTGGGNGGAATACTTGGTCNTTAGCTNCTNNGNNTN
+
```



Messy Data Example 2: Census data

- Census data as a rich information source.
- Challenges due to data size, plotting distributions (population pyramid).



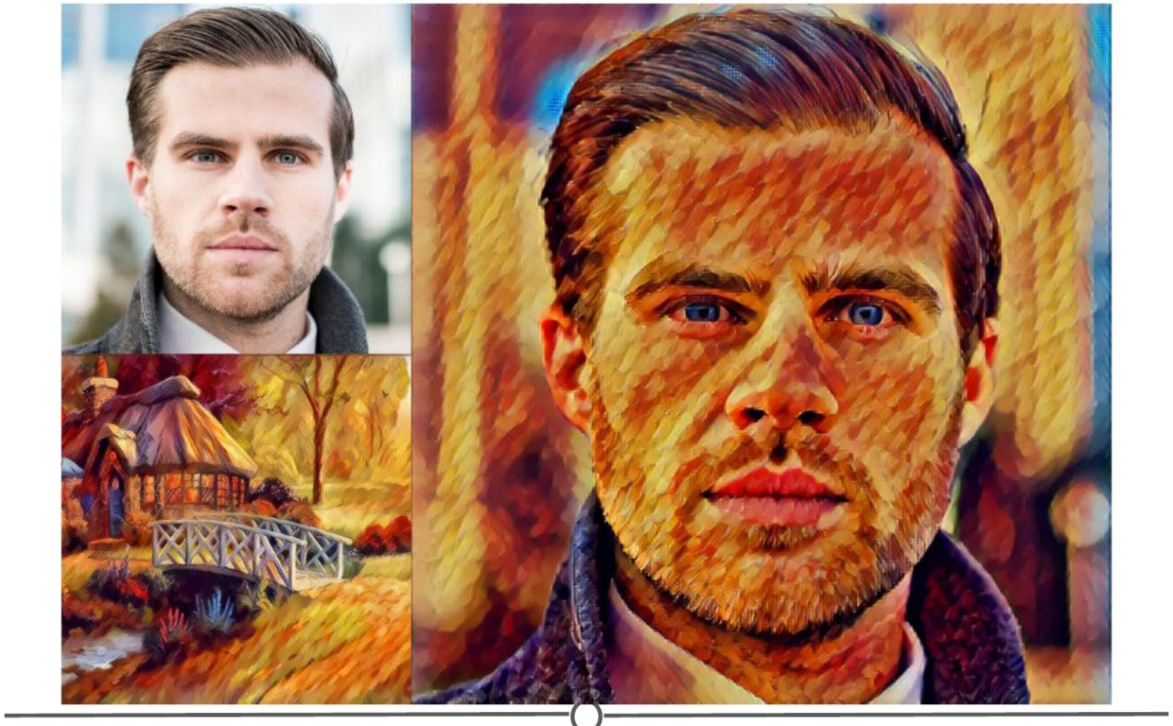
https://www.census.gov/popclock/

Messy Data (cont’d)

- *Messy Data Examples: EMR*
- Utilizing electronic medical records for insights.
- Extracting allergy information and structuring it for analysis.
- *Messy Data Examples: Image Analysis*
- Extracting insights from images/videos.
- **Examples:** Facebook face recognition, DeepDream software, Google's image analysis initiative.

Messy Data (fun example)

- A fun example you can play with is the DeepDream software that was originally designed to detect faces in an image but has since moved on to more *artistic* pursuits.
- The DeepDream software is trained on your image and a famous painting, and your provided image is then rendered in the style of the famous painter:



https://deepdreamgenerator.com/#tools

Data is of secondary importance! (?)

- Recognizing that we’ve spent a lot of time going over what data is, we need to reiterate - Data is important, but it is secondary to your question. A good data scientist asks questions first and seeks out relevant data second.
- Admittedly, often the available data will limit, or perhaps even disable, certain questions you are trying to ask. In these cases, you may have to reframe your question or answer a related question, but the data itself does not drive the question asking.

Conclusion

- Recap of the lesson's focus on data and its attributes.
- Importance of question-first approach in data science.

Example 1: Simple Stats

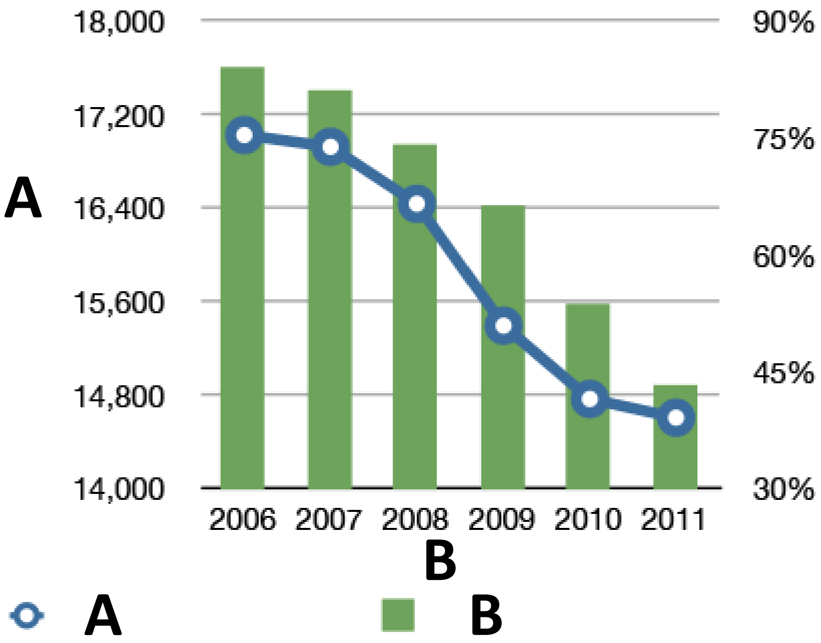
- Suppose X is a collection of 99 integers (positive or negative) and we observe that Mean(X) > 0

- How many elements of X are > 0?

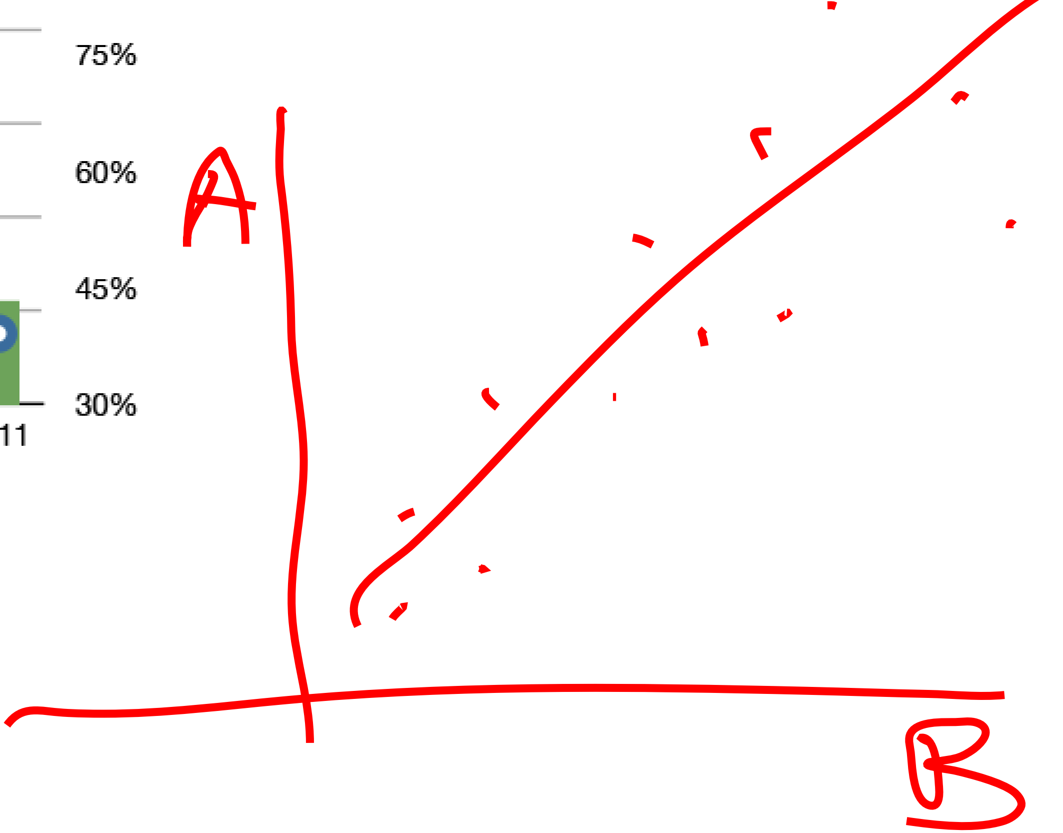
- How about when Median(X) > 0?

$$X = \{\underbrace{3, 5, -2, \ldots, 89}_{99 \text{ many}}\}$$

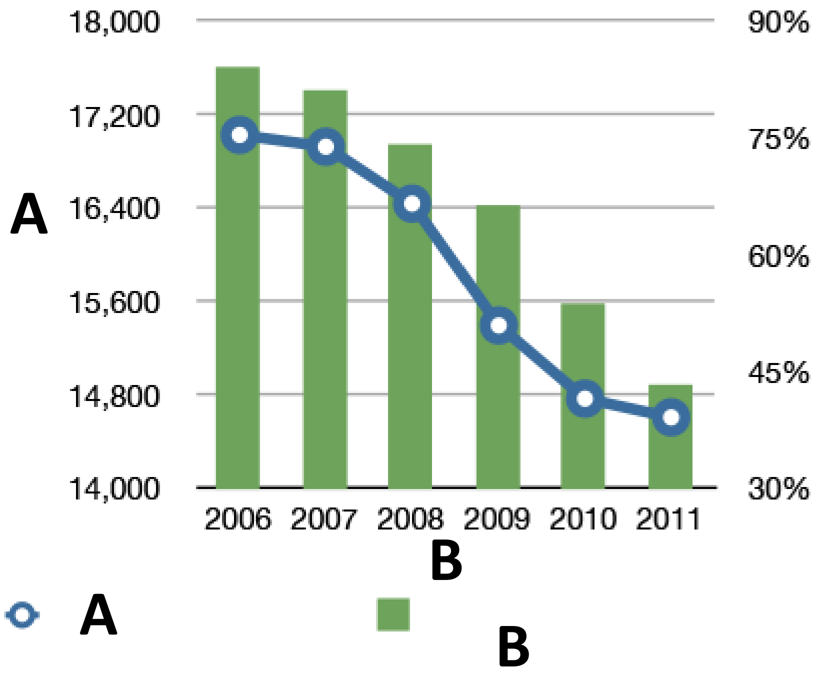
Example 2: Correlation vs. Causation



Q1: Are A and B correlated?



Example 2: Correlation vs. Causation



Q2: Which of the following is true

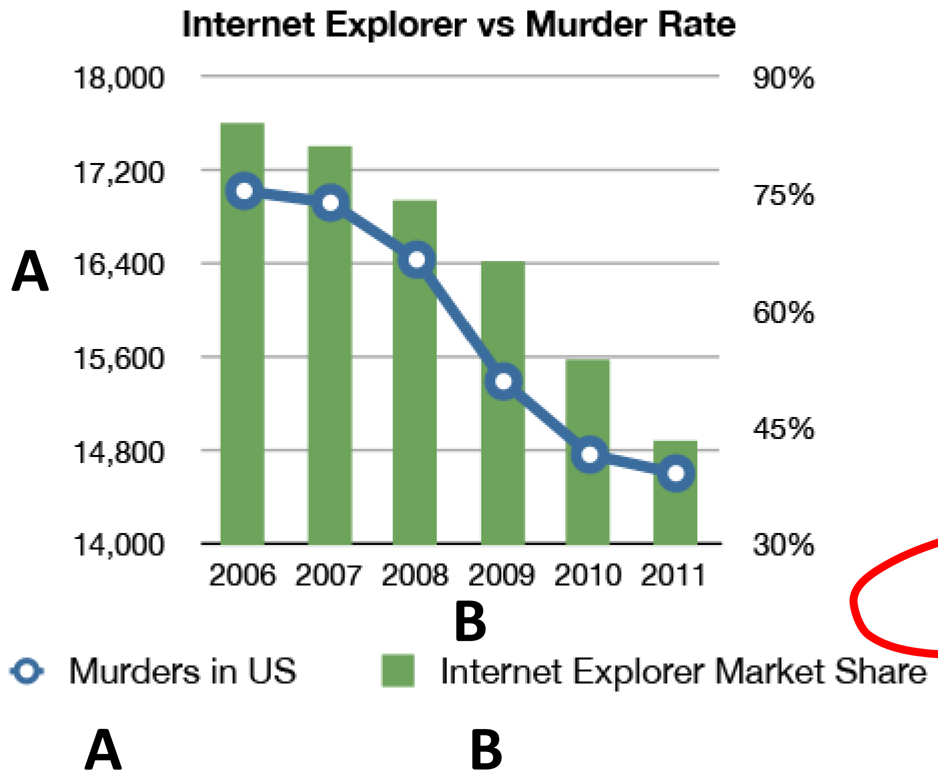
(i) A causes B

(ii) B causes A

(iii) Either (i) or (ii)

(iv) None of the above

Example 2: Correlation vs. Causation



Q2: Which of the following is true

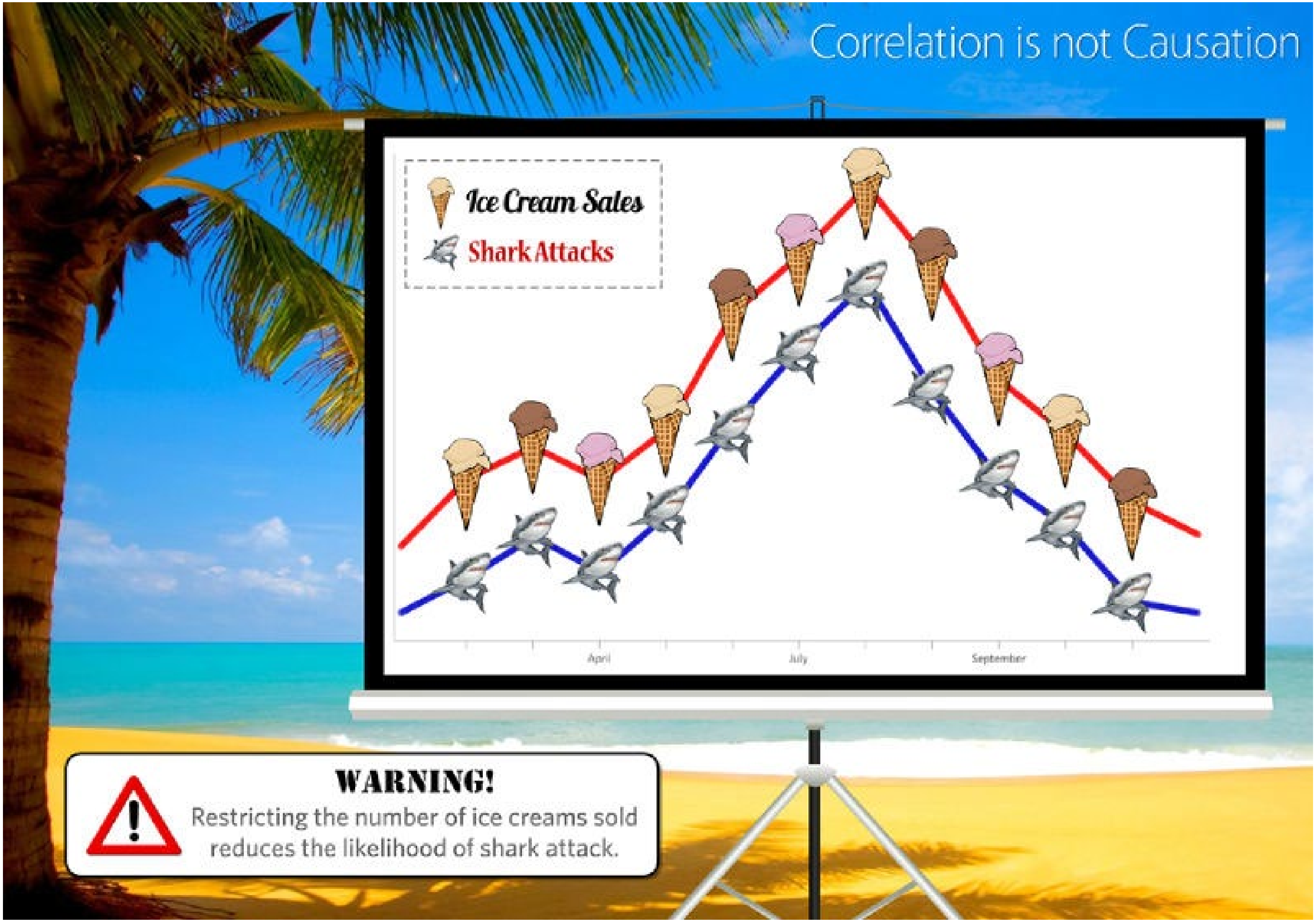
(i) A causes B

(ii) B causes A

(iii) Either (i) or (ii)

[✓ (iv) None of the above]

Example 2: Correlation vs. Causation

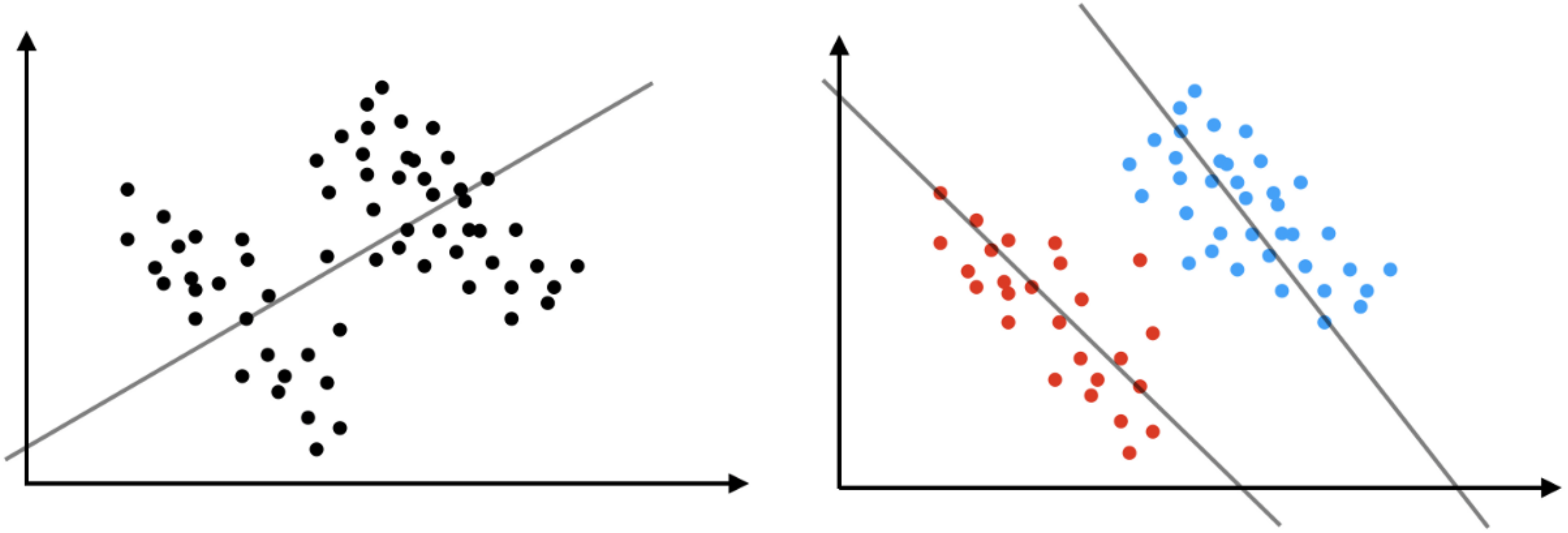


WARNING!

Restricting the number of ice creams sold reduces the likelihood of shark attack.

Example 3: Simpson’s Paradox

Simpson's Paradox is a statistical phenomenon where an association between two variables in a population emerges, disappears or reverses when the population is divided into groups.



Example 3: Simpson's Paradox

- **Ex 3(a) Strawberry vs Peach**
- Suppose we're in the soft drinks industry and we're trying to choose between two new flavors we've produced. We could sample public opinion on the two flavors — let's say we choose to do so by setting up two sampling stalls for each flavor in a busy area and asking 1000 people at each stall if they enjoy the new flavor.

Example 3: Simpson’s Paradox

- **Ex 3(a) Strawberry vs Peach**

Overall

Flavor	# people	# liked the flavor	% liked the flavor
Strawberry	1000	800	80%
Peach	1000	750	75%

Women

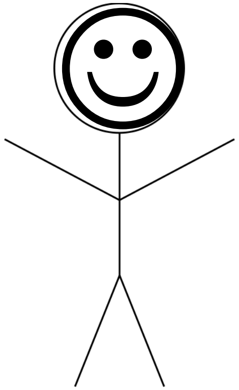
Flavor	# women	# liked the flavor	% liked the flavor
Strawberry	100	40	40%
Peach	300	150	50%

Men

Flavor	# people	# liked the flavor	% liked the flavor
Strawberry	900	760	84.4%
Peach	700	600	85.7%

Example 3(b): Simpson’s Paradox

Earns
above-average
income in A



☹

Earns
below-average
income in B

Developing Nations (A)

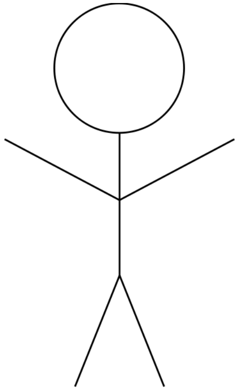
Average income of (A+B) goes down!!

Developed Nations (B)

Average income
of A goes up ☺

Average income
of B goes up ☺

Example 3(b): Simpson's Paradox



Earns
above-average
income in A

Developing
Nations (A)

Person 1: 20K
Person X: 40K

Average income of (A+B)
Before: 5357
After: 5305

Earns
below-average
income in B

Developed
Nations (B)

Person 2: 100K
Person X: 80K

Example 3(b): Simpson’s Paradox

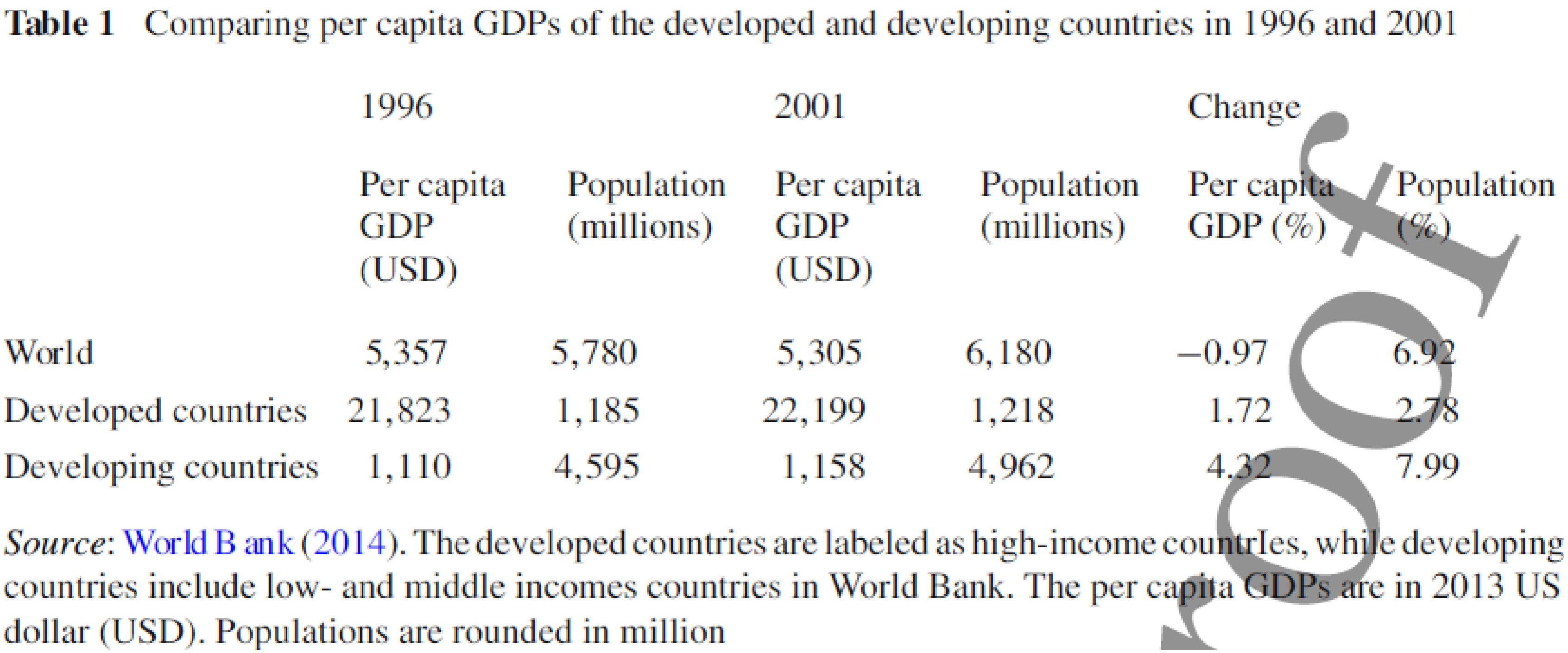


Table 1 Comparing per capita GDPs of the developed and developing countries in 1996 and 2001

	1996		2001		Change	
	Per capita GDP (USD)	Population (millions)	Per capita GDP (USD)	Population (millions)	Per capita GDP (%)	Population (%)
World	5,357	5,780	5,305	6,180	–0.97	6.92
Developed countries	21,823	1,185	22,199	1,218	1.72	2.78
Developing countries	1,110	4,595	1,158	4,962	4.32	7.99

Source: World B ank (2014). The developed countries are labeled as high-income countries, while developing countries include low- and middle incomes countries in World Bank. The per capita GDPs are in 2013 US dollar (USD). Populations are rounded in million

Ma, Zee. (2015). Simpson’s paradox in GDP and per capita GDP growths. Empirical Economics. 49. 10.1007/s00181-015-0921-3.

Example 3(b): Simpson’s Paradox

Since 2000, the median US wage has **risen** about 1% (adjusted)

But over the same period, the median wage for:

- high school dropouts,
- high school graduates with no college education,
- people with some college education, and
- people with Bachelor’s or higher degrees

have *all* decreased.

In other words, within *every* educational subgroup, the median wage is **lower** now than it was in 2000.

How can both things be true?

Example 3: Simpson’s Paradox

How can both things be true?

In this case, the explanation lies in the changing educational profile of the workforce over the past 13 years: there are now many more college graduates (who get higher-paying jobs) than there were in 2000, but wages for college graduates collectively have fallen at a much slower rate (down 1.2%) than for those of lower educational attainment (whose wages have fallen precipitously, down 7.9% for high school dropouts). ***The growth in the proportion of college graduates swamps the wage decline for specific groups.***

Next

- Probability Review - 1
  - Basics: sample space, outcomes, probability
  - Events: mutually exclusive, independent
  - Calculating probability: sets, counting, tree diagram