Data Science 1000: How Data Shape Our World

Josh Clinton & Will Doyle

8/26/2021

Today's Agenda

1. Meet the Instructor(s): Prof. Clinton: *josh.clinton@vanderbilt.edu* Prof. Doyle: *w.doyle@vanderbilt.edu*

Meet the Teaching Assistants: Mellissa Meisels: *mellissa.b.meisels@vanderbilt.edu* Qi Xu: *qi.xu.1@vanderbilt.edu*

2. Course Objectives

- Content: Critical Thinking, Analysis, Presentation
- Skills: Computing and Analysis using R

3. Course Structure

- Lectures
- Workshops
- Evaluations

Too Much Information? (even in 2017...)

Course Motivation via “Tweets"

- “It is a capital mistake to theorize before one has data.” Sherlock Holmes
- “Data is the sword of the 21st century, those who wield it the samurai.” Jonathan Rosenberg, former Senior Vice President of Products at Google and current advisor to Alphabet Inc. CEO Larry Page
- “Torture the data, and it will confess to anything.” Ronald Coase, Nobel Prize Laureate in Economics
- “Here’s an open secret of the big data world: all data is dirty. All of it.” Meredith Broussard, *Artificial Unintelligence: How Computers Misunderstand the World*
- “Big Data processes codify the past. They do not invent the future.” Cathy O’Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*

Course Motivation

Data are more prevalent than ever!

But does quantification equal precision? Or knowledge? Or understanding?

- Numbers without context are meaningless.
- *How* is data collected? What is actually being *measured*?
- What might be wrong/misleading/misinterpreted?
- Theory-driven? Data-Driven?

Data-driven decisions and conclusions require multiple skills... ...and such skills are in high demand.

The Data Science Venn Diagram

The Venn diagram of data science ...

Everyone is doing it?

Everyone is doing it?

For nearly everything.

Is Shakespeare by Shakespeare?

Big debate about Shakespeare finally settled by big data: Marlowe gets his due

Can it tell us who will fall in love?

Body sway predicts romantic interest in speed dating

Andrew Chang, Haley E Kragness, Wei Tsou, Dan J Bosnyak, Anja Thiede, Laurel J Trainor

Social Cognitive and Affective Neuroscience, Volume 16, Issue 1-2, January-February 2021, Pages 185–192, https://doi-org.proxy.library.vanderbilt.edu/10.1093/scan/nsaa093

Published: 18 July 2020 **Article history**

PDF Split View Cite Permissions Share

Abstract

Social bonding is fundamental to human society, and romantic interest involves an important type of bonding. Speed dating research paradigms offer both high external validity and experimental control for studying romantic interest in real-world settings. While previous studies focused on the effect of social and personality factors on romantic interest, the role of non-verbal interaction has been little studied in initial romantic interest, despite being commonly viewed as a crucial factor. The present study investigated whether romantic interest can be predicted by non-verbal dyadic interactive body sway, and enhanced by movement-promoting ('groovy') background music. Participants' body sway trajectories were recorded during speed dating. Directional (predictive) body sway coupling, but not body sway similarity, predicted interest in a long-term relationship above and beyond rated physical attractiveness. In addition, presence of groovy background music promoted interest in meeting a dating partner again. Overall, we demonstrate that romantic interest is reflected by non-verbal body sway in dyads in a real-world dating setting. This novel approach could potentially be applied to investigate non-verbal aspects of social bonding in other dynamic interpersonal interactions such as between infants and parents and in non-verbal populations including those with communication disorders.

Can it tell us who will give us a pizza?

How to Ask for a Favor: A Case Study on the Success of Altruistic Requests

Tim Althoff
Stanford University

Cristian Danescu-Niculescu-Mizil
Max Planck Institute SWS

Dan Jurafsky
Stanford University

Keywords: case study, success of requests, altruistic requests, helping behavior, linguistic analysis, reddit

Abstract

Name the 1400 people most likely to cause or experience gun violence in Chicago?

CHICAGO DATA PORTAL

Browse Tutorial F

Strategic Subject List - Historical

Public Safety

View Data | Visualize | Export | API | ...

The program described below ended in 2019. This dataset is being retained for historical reference.

The information displayed represents a de-identified listing of arrest data from August 1, 2012 to July 31, 2016, that was used by the Chicago Police Department's Strategic Subject Algorithm, created by the Illinois Institute of Technology and funded through a Department of Justice Bureau of Justice Assistance grant, to create a risk assessment score known as the Strategic Subject List or "SSL." These scores reflect an individual's probability of being involved in a shooting incident either as a victim or an offender. Scores are calculated and placed on a scale ranging from 0 (extremely low risk) to 500 (extremely high risk).

Based on this time frame's version of the Strategic Subject Algorithm, individuals with criminal records are ranked using eight attributes, not including race or sex. These attributes are: number of times being the victim of a shooting incident, age during the latest arrest, number of times being the victim of aggravated battery or assault, number of prior arrests for violent offenses, gang affiliation, number of prior narcotic arrests, trend in recent criminal activity and number of prior unlawful use of weapon arrests.

Please note that this data set includes fields that are not used to calculate SSL, for example, neither race nor sex are used in the Strategic Subject Algorithm. Portions of the arrest data are de-identified on the basis of privacy concerns. The attributes used in the Strategic Subject Algorithm were revised on an ongoing basis during the lifetime of the program.

Less

Updated
September 25, 2020

Data Provided by
City of Chicago

Figure out who is sick or not?

Let's take a look at one (details on heatmap calculation given at end):

Patient with cardiomegaly: heatmap of CNN on left, original image on right

Confuse the Public?

How the U.S. Economy Is Actually Doing, in 9 Charts

The unemployment rate doesn't tell the whole story, so we talked to a panel of economists to find out what other measures can shed light.

What we will do. . .

- How to think about data. (Theory-driven vs. ‘data-driven)
- How to work with data in a modern, widely-used, statistical programming language.
- How to analyze? How to present? [Reproducible!]
- Empower you to collect, and analyze new data to describe and understand the world

What we will NOT do...

- Talk (too much) about any one topic.
- Provide a comprehensive understanding of any subject.
- Cover the latest and greatest Data Science tools and methods.

Some of what we will talk about:

- Predicting Elections
- Predicting College Admission Decisions (and Enrollment)
- Visualizing Pandemic Data
- Analyzing Consumer Choice

Preliminaries

- Where to find things
- How class will work
- Expectations: attendance, assignments, guided exercises
- Communication
- Health and Safety
- Honor Code
- Technology

Your first assignment

- Go and get R!
- Go and get Rstudio!
- Open them up!
- Take a picture!