Welcome to DATA 151

I'm so glad you're here!

DATA 151: INTRODUCTION TO DATA SCIENCE (WITH R)

CLASS 1B: TAXONOMY OF VARIABLES, DATA ORGANIZATION

NOTES PREPARED BY PROF. KITADA SMALLEY (FALL 2022)

AGENDA

Time	Topics
2:30 – 2:45	Finish syllabus and policies Answer questions
2:45 – 3:00	What is data? What is big data? How can data be used to help people?
3:00 – 3:30	Group work on UN SDGs 13 minutes for group 17 minutes for sharing
3:30 – 3:45	How is data structured? What are observations and variables? Taxonomy of data
3:45 – 4:00	Dear Data and wrap-up!

HELLO DATA

WHAT IS DATA?

WHAT IS DATA?

In the pursuit of knowledge, **data** is a collection of discrete states that convey information, describing quantity, quality, fact, statistics, other basic units of meaning, or simply sequences of symbols that may be further interpreted.

QUICK GRAMMAR LESSON

da·tum

/ˈdādəm,ˈdadəm/

singular

noun

1. a piece of information.
   "the fact is a datum worth taking into account"

da·ta

/ˈdadə,ˈdādə/

plural

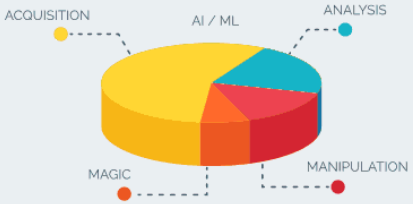
noun

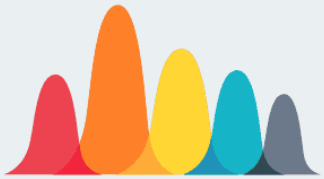
facts and statistics collected together for reference or analysis.
"there is very little data available"

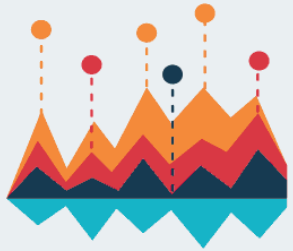
Similar: facts | figures | statistics | details | particulars | specifics | features

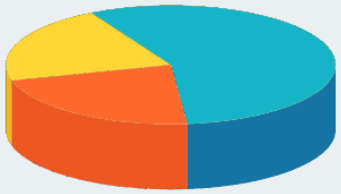
- the quantities, characters, or symbols on which operations are performed by a computer, being stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media.

Data: Singular or Plural?









As a singular mass noun (like *information*)

All the data is available for download.

Our data shows that online businesses grew in 2020.

As the plural of *datum* (esp. in scientific and academic writing)

The collected data are then analyzed.

The data indicate that bias is pervasive across all fields of research.

WHAT IS BIG DATA?

DATA

DATA

BIG DATA – FOUR V'S

- **Volume:** Scale of data
- **Variety:** Different forms
- **Velocity:** How fast sent
- **Veracity:** Trustworthiness
  - *Uncertainty, bias, or inaccuracies in the data make information less valuable for meaningful analysis and decision making*

THE 4 V'S OF BIG DATA

Volume

SCALE OF DATA

40 ZETTABYTES
of data will be created by 2020, an increase of 300 times from 2005

6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION

2.5 QUINTILLION BYTES
of data are created each day

Most companies in the U.S. have at least
100 TERABYTES
of data stored

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES

30 BILLION PIECES OF CONTENT
are shared on facebook every month

4 BILLION + HOURS OF VIDEO
are watched on You Tube each month

4 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
1TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

Veracity

UNCERTAINITY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

27% OF RESPONDENTS
in one survey were unsure of how much of data was inaccurate

Reference : http://www.ibmbigdatahub.com/infographic/four-vs-big-data

READ AND DISCUSS "UNITED NATIONS: BIG DATA FOR SUSTAINABLE DEVELOPMENT"

UN – BIG DATA FOR SUSTAINABLE DEVELOPMENT

United Nations | Peace, dignity and equality on a healthy planet

TAKE AWAYS

BIG DATA

- In 2020, **64.2 zettabytes** (2 to the 70th power) of data were created, that is a 314 percent increase from 2015.
- "**Data exhaust**," or passively collected data deriving from everyday interactions with digital products or services, including mobile phones, credit cards, and social media

TAKE AWAYS

OPPORTUNITIES

- Enable more agile, efficient and **evidence-based decision-making** and can better measure progress on the Sustainable Development Goals (SDGs) in a way that is both inclusive and fair.

RISKS

- Fundamental elements of human rights must be safeguarded to realize the opportunities presented by big data: **privacy, ethics and respect for data sovereignty** require us to assess the rights of individuals along with the benefits of the collective
- There is also a **risk of growing inequality and bias**.
- Major gaps are already opening up between the **data haves and have-nots**.

BIG DATA & THE SDGs

How data science and analytics can contribute to sustainable development

1 NO POVERTY
Spending patterns on mobile phone services can provide proxy indicators of income levels

2 ZERO HUNGER
Crowdsourcing or tracking of food prices listed online can help monitor food security in near real-time

3 GOOD HEALTH AND WELL-BEING
Mapping the movement of mobile phone users can help predict the spread of infectious diseases

4 QUALITY EDUCATION
Citizen reporting can reveal reasons for student drop-out rates

5 GENDER EQUALITY
Analysis of financial transactions can reveal the spending patterns and different impacts of economic shocks on men and women

6 CLEAN WATER AND SANITATION
Sensors connected to water pumps can track access to clean water

7 AFFORDABLE AND CLEAN ENERGY
Smart metering allows utility companies to increase or restrict the flow of electricity, gas or water to reduce waste and ensure adequate supply at peak periods

8 DECENT WORK AND ECONOMIC GROWTH
Patterns in global postal traffic can provide indicators such as economic growth, remittances, trade and GDP

9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
Data from GPS devices can be used for traffic control and to improve public transport

10 REDUCED INEQUALITY
Speech-to-text analytics on local radio content can reveal discrimination concerns and support policy response

11 SUSTAINABLE CITIES AND COMMUNITIES
Satellite remote sensing can track encroachment on public land or spaces such as parks and forests

12 RESPONSIBLE CONSUMPTION AND PRODUCTION
Online search patterns or e-commerce transactions can reveal the pace of transition to energy efficient products

13 CLIMATE ACTION
Combining satellite imagery, crowd-sourced witness accounts and open data can help track deforestation

14 LIFE BELOW WATER
Maritime vessel tracking data can reveal illegal, unregulated and unreported fishing activities

15 LIFE ON LAND
Social media monitoring can support disaster management with real-time information on victim location, effects and strength of forest fires or haze

16 PEACE, JUSTICE AND STRONG INSTITUTIONS
Sentiment analysis of social media can reveal public opinion on effective governance, public service delivery or human rights

17 PARTNERSHIPS FOR THE GOALS
Partnerships to enable the combining of statistics, mobile and internet data can provide a better and real-time understanding of today's hyper-connected world

GLOBAL PULSE

www.unglobalpulse.org
@UNGlobalPulse 2017


BIG
DATA
& THE
SDGs

SMALL GROUP DISCUSSION (10 MINUTES)

- **Groups of 4**
- Were there any parts of the ”UN: Big Data for Sustainable Development” article that stood out to you?

- Each group will randomly be assigned to a sustainable development goal.
  - In addition to ideas posed by the UN, brainstorm other ways that data could be used to meet the sustainable development goal.

1 NO POVERTY

Spending patterns on mobile phone services can provide proxy indicators of income levels

2 ZERO HUNGER

Crowdsourcing or tracking of food prices listed online can help monitor food security in near real-time

3 GOOD HEALTH AND WELL-BEING

Mapping the movement of mobile phone users can help predict the spread of infectious diseases

4 QUALITY EDUCATION

Citizen reporting can reveal reasons for student drop-out rates

5 GENDER EQUALITY

Analysis of financial transactions can reveal the spending patterns and different impacts of economic shocks on men and women

6

CLEAN WATER AND SANITATION

Sensors connected to water pumps can track access to clean water

7 AFFORDABLE AND CLEAN ENERGY

Smart metering allows utility companies to increase or restrict the flow of electricity, gas or water to reduce waste and ensure adequate supply at peak periods

8 DECENT WORK AND ECONOMIC GROWTH

Patterns in global postal traffic can provide indicators such as economic growth, remittances, trade and GDP

9 INDUSTRY, INNOVATION AND INFRASTRUCTURE

Data from GPS devices can be used for traffic control and to improve public transport

10 REDUCED INEQUALITY

Speech-to-text analytics on local radio content can reveal discrimination concerns and support policy response

11 SUSTAINABLE CITIES AND COMMUNITIES

Satellite remote sensing can track encroachment on public land or spaces such as parks and forests

12 RESPONSIBLE CONSUMPTION AND PRODUCTION

Online search patterns or e-commerce transactions can reveal the pace of transition to energy efficient products

13 CLIMATE ACTION

Combining satellite imagery, crowd-sourced witness accounts and open data can help track deforestation

14 LIFE BELOW WATER

Maritime vessel tracking data can reveal illegal, unregulated and unreported fishing activities

15 LIFE ON LAND

Social media monitoring can support disaster management with real-time information on victim location, effects and strength

16 PEACE, JUSTICE AND STRONG INSTITUTIONS

Sentiment analysis of social media can reveal public opinion on effective governance, public service delivery or human rights

17 PARTNERSHIPS FOR THE GOALS

Partnerships to enable the combining of statistics, mobile and internet data can provide a better and real-time understanding of today's hyper-connected world

HOW IS DATA STRUCTURED?

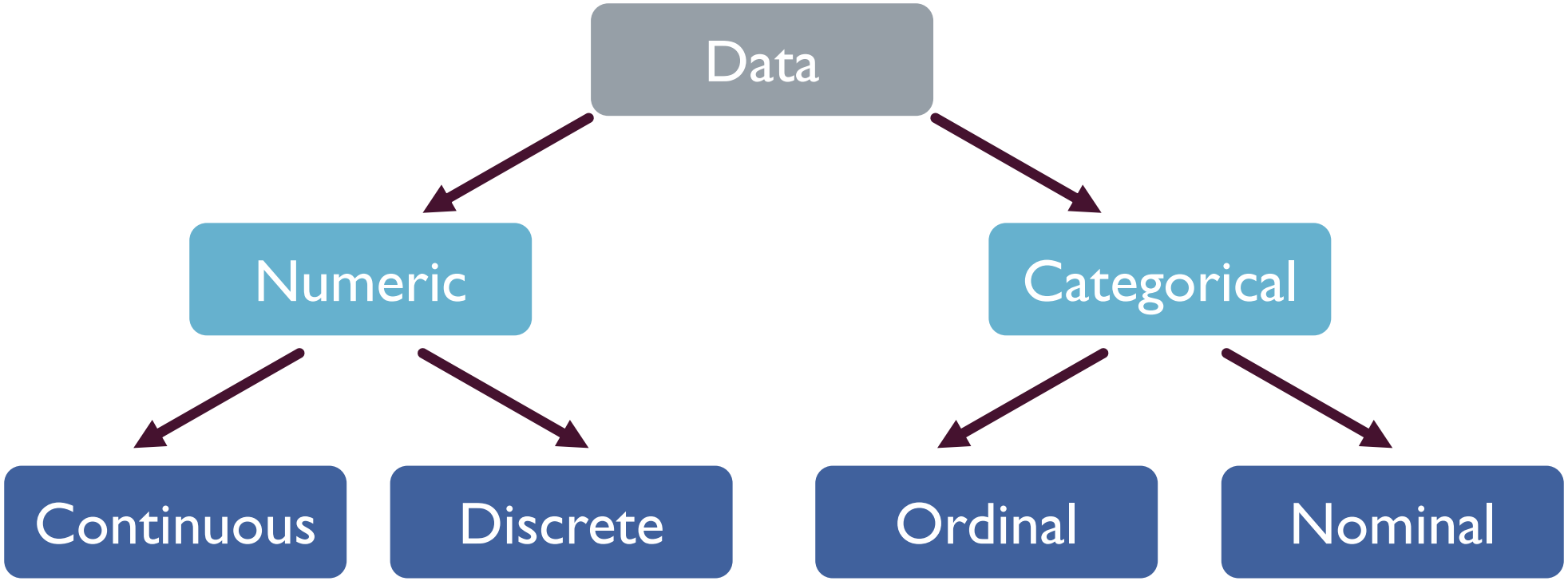
STRUCTURE OF DATA

Structured	Semi-structured	Unstructured
Surveys*	Social network graphs	Texts
Geospatial data	Telemetry	Tweets
Experiments	Internet pages	Google searches
Spreadsheets	Emails	Images and videos
		Web server logs

DEFINITIONS

- **Individual**/Observation/Respondent:
  - An object/unit described by data
    - *Ex: a person, an animal, a household*
- **Variable:**
  - A characteristic that is observed on an individual that can change between individuals
    - *Ex: height, weight, hair color, political party*

TYPES OF VARIABLES



TYPES OF VARIABLES

- **Numeric:** A variable observed as a number
  - **Continuous:** Measured characteristics that can take any value on the real line $(-\infty, \infty)$
  - **Discrete:** Counted items
- **Categorical:** Places individuals into one of several groups or categories
  - **Ordinal:** Groups have inherent ordering
  - **Nominal:** No inherent group ordering
  - **Binary:** (special case) two unordered categories

ORGANIZATION OF DATA

- **Data matrix:**
  - The organization of data into a spreadsheet/table where observations are in rows and their respective variables are in columns

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	cycle	branch	type	matchup	forecastdate	state	startdate	enddate	pollster	grade	samplesize	population	poll_wt	rawpoll_clint	rawpoll_trun
2	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/3/16	11/6/16	ABC News/W	A+	2220	lv	8.720654	47	43
3	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/1/16	11/7/16	Google Cons	B	26574	lv	7.628472	38.03	35.69
4	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/2/16	11/6/16	Ipsos	A-	2195	lv	6.424334	42	39
5	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/4/16	11/7/16	YouGov	B	3677	lv	6.087135	45	41
6	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/3/16	11/6/16	Gravis Marke	B-	16639	rv	5.316449	47	43
7	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/3/16	11/6/16	Fox News/Ar	A	1295	lv	5.218141	48	44
8	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/2/16	11/6/16	CBS News/N	A-	1426	lv	4.881873	45	41
9	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/3/16	11/5/16	NBC News/W	A-	1282	lv	4.836171	44	40
10	2016	President	polls-plus	Clinton vs. Tr	11/8/16	New Mexico	11/6/16	11/6/16	Zia Poll		8439	lv	4.609492	46	44
11	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/4/16	11/7/16	IBD/TIPP	A-	1107	lv	4.520075	41.2	42.7
12	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/4/16	11/6/16	Selzer & Com	A+	799	lv	4.150419	44	41
13	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/1/16	11/4/16	Angus Reid G	A-	1151	lv	4.141083	48	44
14	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/3/16	11/6/16	Monmouth U	A+	748	lv	3.965234	50	44
15	2016	President	polls-plus	Clinton vs. Tr	11/8/16	Virginia	11/3/16	11/4/16	Public Policy	B+	1238	lv	3.923524	48	43
16	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/1/16	11/3/16	Marist Colleg	A	940	lv	3.886329	44	43
17	2016	President	polls-plus	Clinton vs. Tr	11/8/16	Iowa	11/1/16	11/4/16	Selzer & Com	A+	800	lv	3.842234	39	46
18	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/5/16	11/7/16	The Times-Picayune/Lucid		2521	lv	3.792648	45	40
19	2016	President	polls-plus	Clinton vs. Tr	11/8/16	Wisconsin	10/26/16	10/31/16	Marquette U	A	1255	lv	3.789957	46	40
20	2016	President	polls-plus	Clinton vs. Tr	11/8/16	North Carolir	11/4/16	11/6/16	Siena College	A	800	lv	3.774139	44	44
21	2016	President	polls-plus	Clinton vs. Tr	11/8/16	Georgia	11/6/16	11/6/16	Landmark Co	B	1200	lv	3.752961	46	49
22	2016	President	polls-plus	Clinton vs. Tr	11/8/16	Florida	11/3/16	11/6/16	Quinnipiac U	A-	884	lv	3.718829	46	45
23	2016	President	polls-plus	Clinton vs. Tr	11/8/16	North Carolir	11/3/16	11/6/16	Quinnipiac U	A-	870	lv	3.677972	47	45
24	2016	President	polls-plus	Clinton vs. Tr	11/8/16	Virginia	10/27/16	10/30/16	ABC News/W	A+	1024	lv	3.632387	48	42

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	cycle	branch	type	matchup	forecastdate	state	startdate	enddate	pollster	grade	samplesize	population	poll_wt	rawpoll_clint	rawpoll_trun
2	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/3/16	11/6/16	ABC News/W	A+	2220	lv	8.720654	47	43
3	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/1/16	11/7/16	Google Cons	B	26574	lv	7.628472	38.03	35.69
4	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/2/16	11/6/16	Ipsos	A-	2195	lv	6.424334	42	39
5	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/4/16	11/7/16	YouGov	B	3677	lv	6.087135	45	41
6	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/3/16	11/6/16	Gravis Marke	B-	16639	rv	5.316449	47	43
7	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/3/16	11/6/16	Fox News/Ar	A	1295	lv	5.218141	48	44
8	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/2/16	11/6/16	CBS News/N	A-	1426	lv	4.881873	45	41
9	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/3/16	11/5/16	NBC News/W	A-	1282	lv	4.836171	44	40
10	2016	President	polls-plus	Clinton vs. Tr	11/8/16	New Mexico	11/6/16	11/6/16	Zia Poll		8439	lv	4.609492	46	44
11	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/4/16	11/7/16	IBD/TIPP	A-	1107	lv	4.520075	41.2	42.7
12	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/4/16	11/6/16	Selzer & Com	A+	799	lv	4.150419	44	41
13	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/1/16	11/4/16	Angus Reid G	A-	1151	lv	4.141083	48	44
14	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/3/16	11/6/16	Monmouth U	A+	748	lv	3.965234	50	44
15	2016	President	polls-plus	Clinton vs. Tr	11/8/16	Virginia	11/3/16	11/4/16	Public Policy	B+	1238	lv	3.923524	48	43
16	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/1/16	11/3/16	Marist Colleg	A	940	lv	3.886329	44	43
17	2016	President	polls-plus	Clinton vs. Tr	11/8/16	Iowa	11/1/16	11/4/16	Selzer & Com	A+	800	lv	3.842234	39	46
18	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/5/16	11/7/16	The Times-Pi	cayune/Lucid	2521	lv	3.792648	45	40
19	2016	President	polls-plus	Clinton vs. Tr	11/8/16	Wisconsin	10/26/16	10/31/16	Marquette U	A	1255	lv	3.789957	46	40
20	2016	President	polls-plus	Clinton vs. Tr	11/8/16	North Carolir	11/4/16	11/6/16	Siena College	A	800	lv	3.774139	44	44
21	2016	President	polls-plus	Clinton vs. Tr	11/8/16	Georgia	11/6/16	11/6/16	Landmark Co	B	1200	lv	3.752961	46	49
22	2016	President	polls-plus	Clinton vs. Tr	11/8/16	Florida	11/3/16	11/6/16	Quinnipiac U	A-	884	lv	3.718829	46	45
23	2016	President	polls-plus	Clinton vs. Tr	11/8/16	North Carolir	11/3/16	11/6/16	Quinnipiac U	A-	870	lv	3.677972	47	45
24	2016	President	polls-plus	Clinton vs. Tr	11/8/16	Virginia	10/27/16	10/30/16	ABC News/W	A+	1024	lv	3.632387	48	42

PRINCIPLES OF TIDY DATA

- "Tidy" data needed to import data files (.csv, .txt, …) into statistical software packages such as R, SAS, Stata, SPSS, …

- **(Wickham 2014)**
  - 1) Each variable forms a column
  - 2) Each observation forms a row
  - 3) Each cell contains a single value

PRINCIPLES OF TIDY DATA

- **Examples of Messy Data:**
  - Column headers are values, not variable names
  - Multiple variables are stored in one column
  - Variables are stored in both rows and columns
  - Multiple types of observational units are stored in the same table
  - A single observational unit is stored in multiple tables

PRINCIPLES OF TIDY DATA

Before next class please read (at least) pages 1-5 of

"Tidy Data" by

Hadley Wickham (2014)

Journal of Statistical Software

August 2014, Volume 59, Issue 10. *http://www.jstatsoft.org/*

Tidy Data

Hadley Wickham
RStudio

Abstract

A huge amount of effort is spent cleaning data to get it ready for analysis, but there has been little research on how to make data cleaning as easy and effective as possible. This paper tackles a small, but important, component of data cleaning: data tidying. Tidy datasets are easy to manipulate, model and visualize, and have a specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table. This framework makes it easy to tidy messy datasets because only a small

DATA COMES FROM EVERYWHERE….

DATA COMES FROM YOU

DEAR DATA


Dear Data
the book
the project
press
the authors
get in touch
news!
Dear Data is a year-long, analog data drawing project by Giorgia Lupi and Stefanie Posavec, two award-winning information designers living on different sides of the Atlantic.
By collecting and hand drawing their personal data and sending it to each other in the form of postcards, they became friends.

DEAR DATA

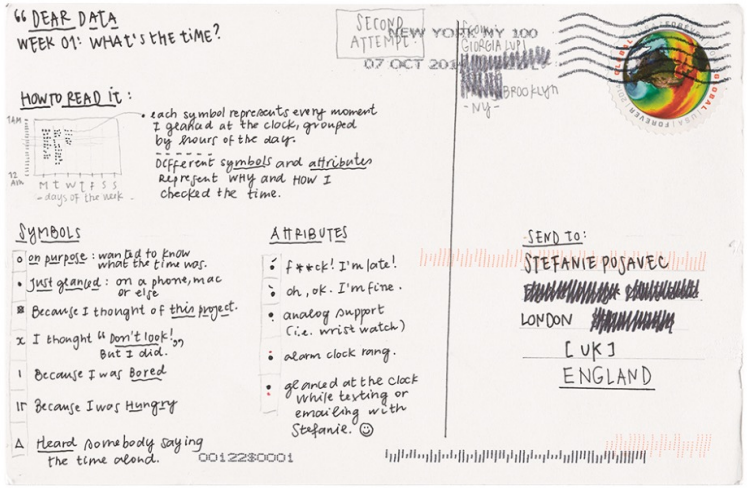
Each week, and for a year, we collected and measured a particular type of data about our lives, used this data to make a drawing on a postcard-sized sheet of paper, and then dropped the postcard in an English "postbox" (Stefanie) or an American "mailbox" (Giorgia)!

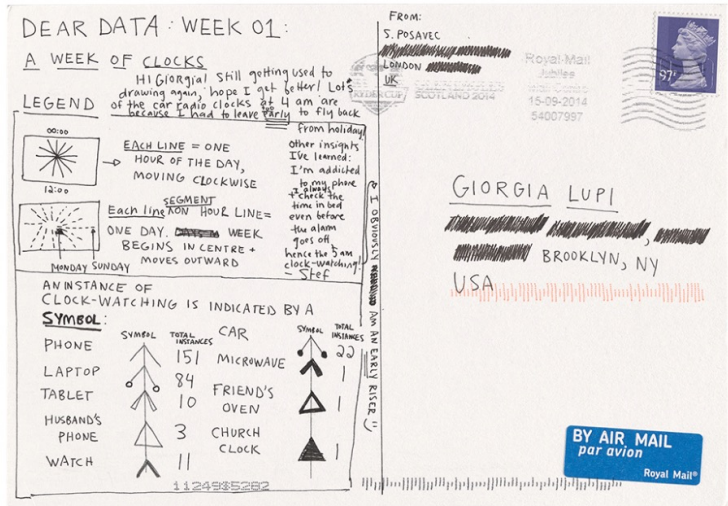
Eventually, the postcard arrived at the other person's address with all the scuff marks of its journey over the ocean: a type of "slow data" transmission.

https://www.youtube.com/watch?v=iqaVe1MCTIA&t=185s

DEAR DATA

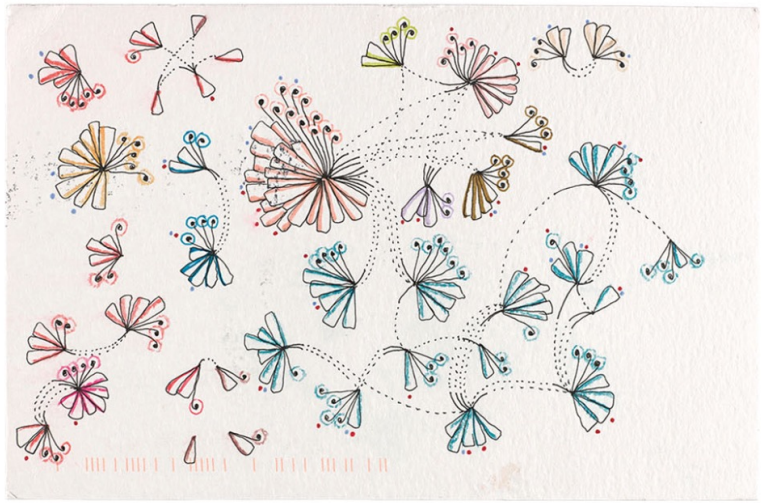
CLOCKS

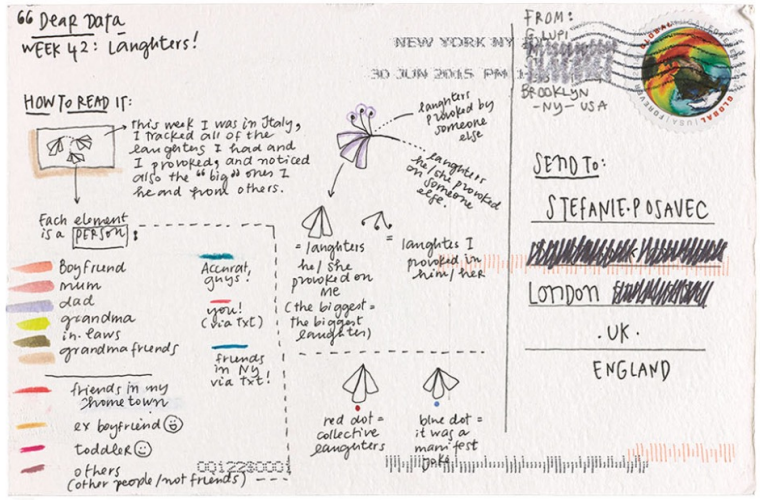


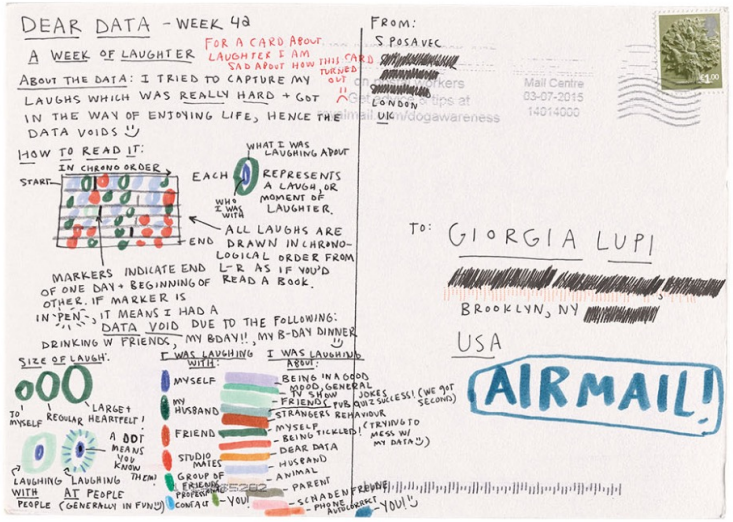


Drawing her first postcard, Giorgia had an idea for her whole collection: from now on every time

LAUGHTER!

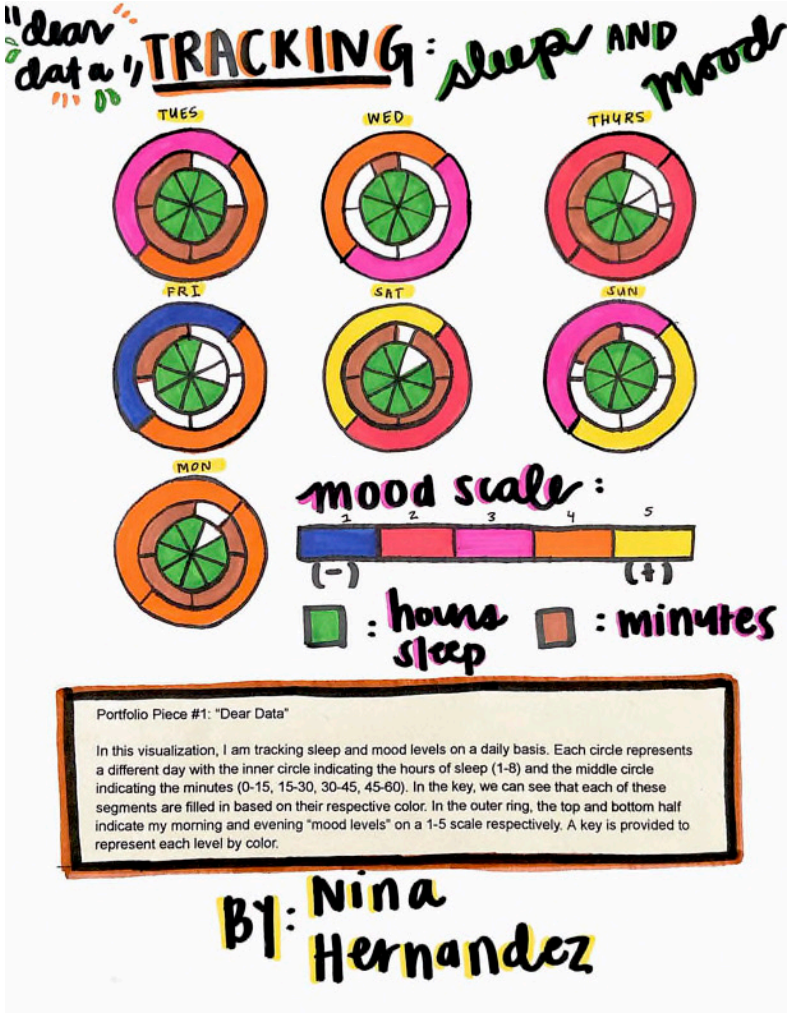


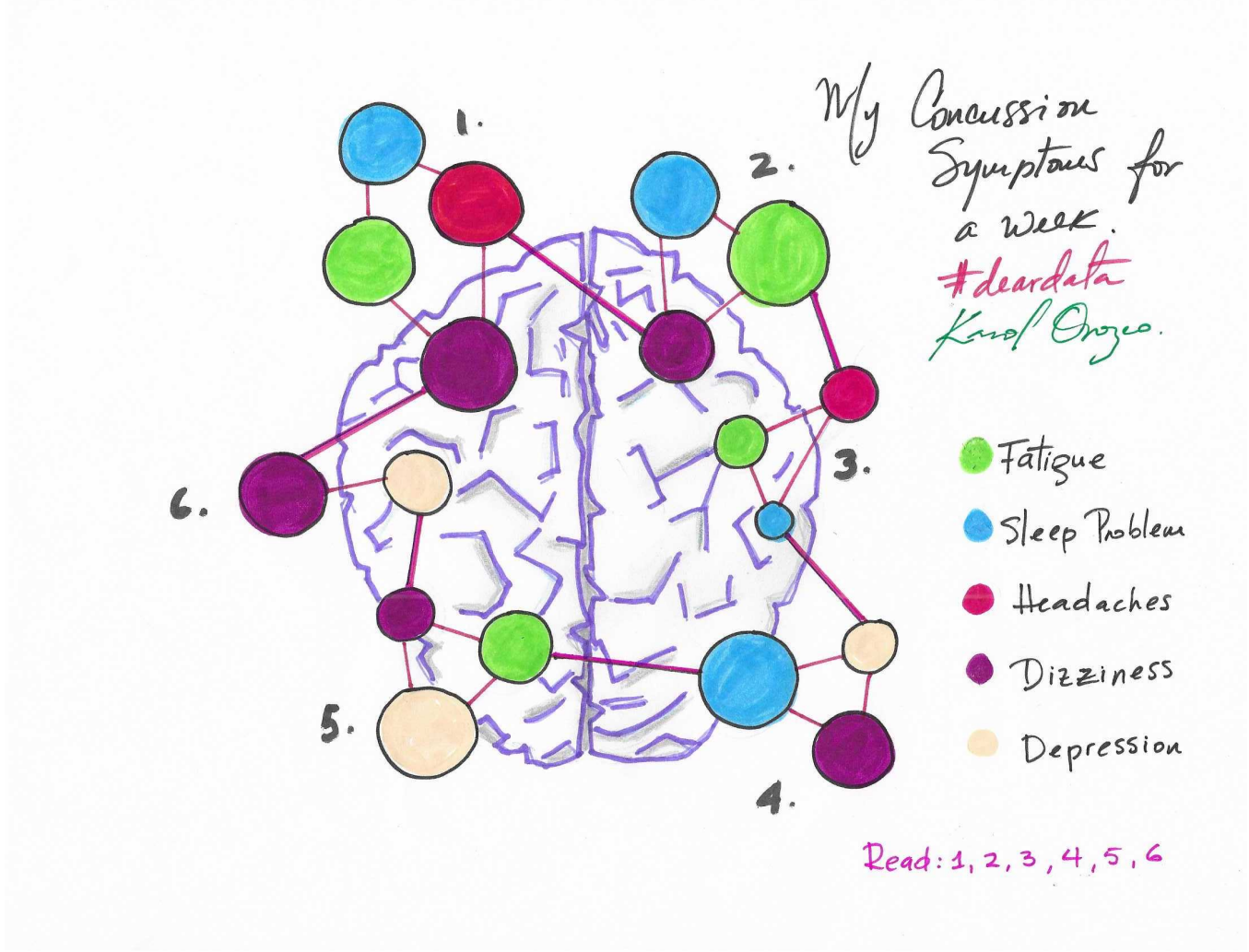




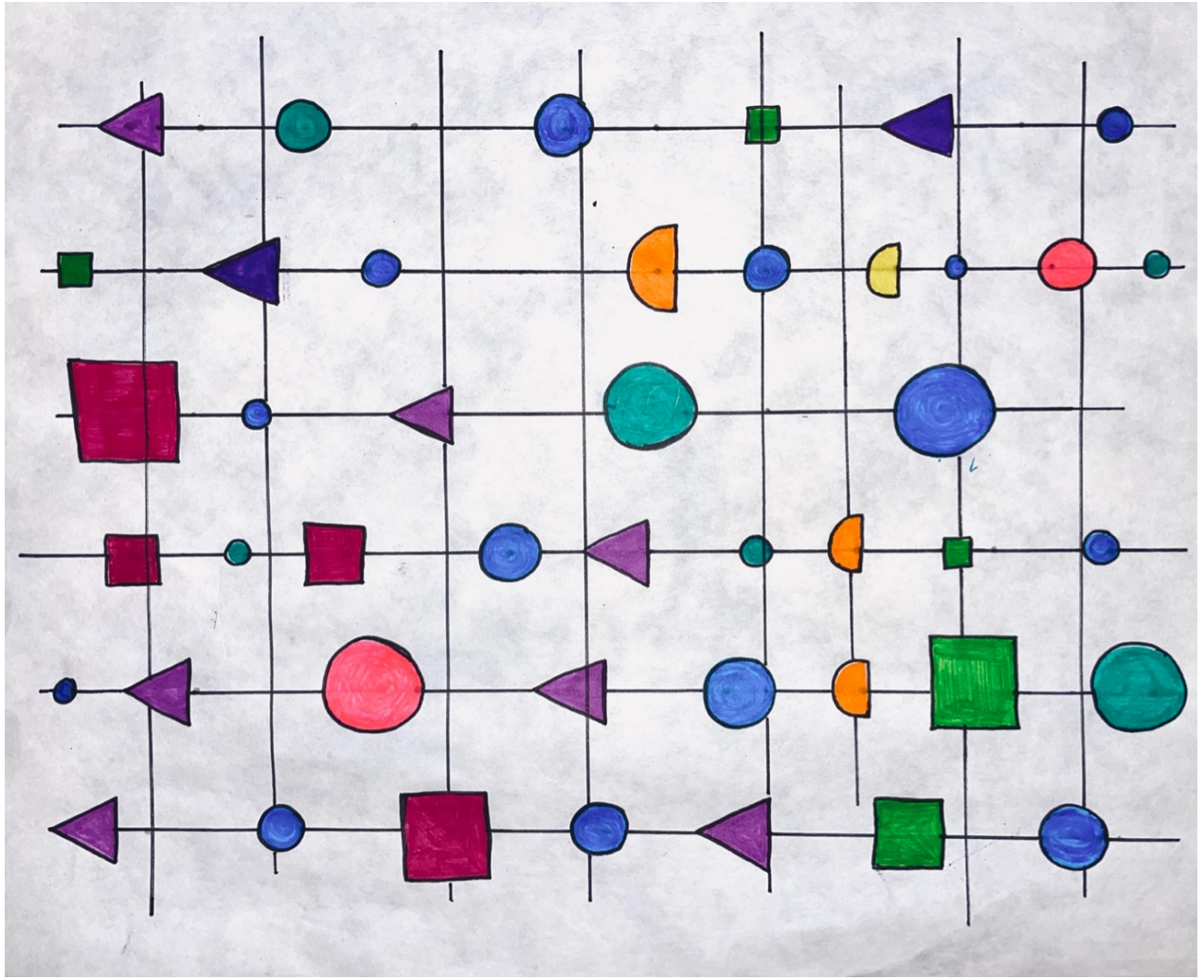
When data-gathering gets in the way: Giorgia laughs and then has to report it. So, is she fully enjoying

EXAMPLES FROM STUDENTS





EXAMPLES FROM STUDENTS



BRAINSTORM

Turn to your partner and brainstorm different ideas for data that you could track for a week.