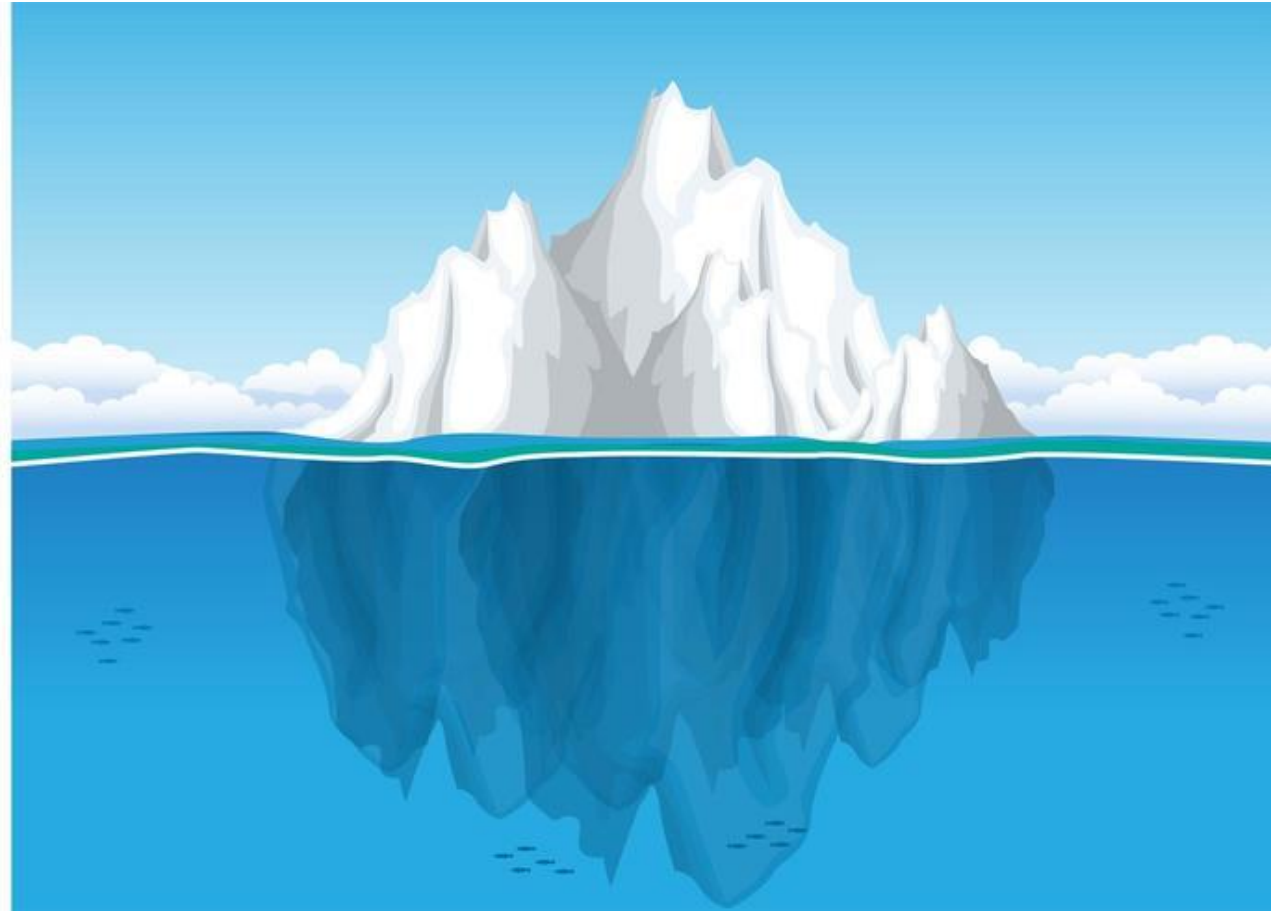


# The Life course of a Computational Sociology Research Project



Alina Arseniev-Koehler and Bernard Koch

# Outline: 3h

- 25 min Concepts and Example CompSoc Projects
- 20 min Developing your CompSoc Project Workflow
- 15 min break
- Remainder ( $\sim 2$ h): Data Wrangling in Python (hands-on)

# Storing and Managing Data

- *Always* keep a **raw** version if permitted by IRB
- Ideal workflow is transparent, replicable, efficient
  - Do as much cleaning with code rather than Excel/by hand
  - Use Github!
    - \*maybe, design as a tutorial ☺
  - Use Box!
- Think carefully about how to name your data files (e.g., by year? by ID?)

# Raw

ID	Birthdate	Gender	Depression	Notes
104	Sep 1992	M	3	
241	October 1998	F	6	
125	999	999	13	
347	04/1995	M	16	
835	September 1993	F	0	

# Data Key

Variable	Description
ID	ID# of participant
Birthdate	Birthdate of participant. Missing: 999
Gender	Male, Female. Missing: 999
Depression	Depression Score, According to PHQ-8. Ranges from 0 (low risk of depression) to 24 (high risk of depression). Missing: 999.
Notes	Any notes.

“Raw” means something different for each analysis/research project. CSV vs XLS

# Clean

ID	Age	Gender	Dep_Score	Dep_Binary	Notes
104	26	M	3	0	
241	20	F	6	0	
125	999	999	13	1	
347	23	M	16	1	
835	25	F	0	0	

# Data Key

Variable	Description
ID	ID# of participant
Age	Age (years) of participant. Missing: 999
Gender	Male, Female. Missing: 999
Depression_Score	Depression Score, According to PHQ-8. Ranges from 0 (low risk of depression) to 24 (high risk of depression). Missing: 999.
Depression_Binary	Has depression risk or not (0/1), according to PHQ-8. Score >9 on PHQ-8 suggests depression risk. Missing=999.
Notes	Any notes.

“Clean” means something different for each analysis/research project

# Thinking about the production of data

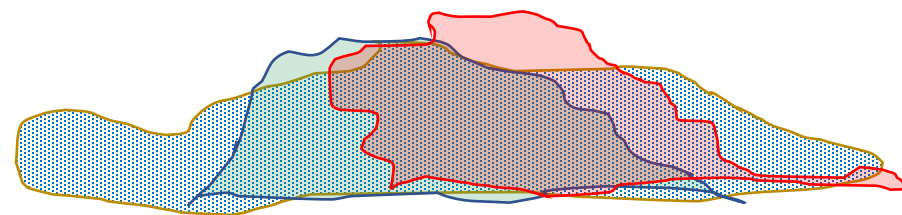
- Where do your variables come from? (e.g., “depression score”)
  - Some social science/health abstraction from psychiatrists abstraction of latent mental distress
  - Continuous vs binary
- How was the data collected/produced?
  - E.g., self report vs interview (interview with human or with virtual agent?)

# Thinking about the production of data

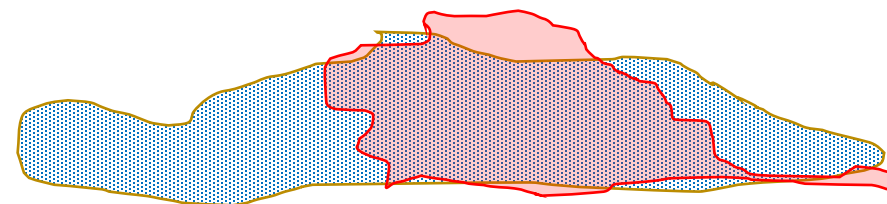
0 or 1: “Depression” or not



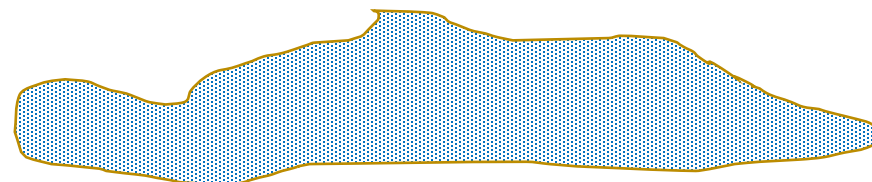
Self-Report Measure, e.g. PHQ



Psychiatrists' Evaluation



Depression (latent)



# Abstracted variables vs human experience

**“I can’t even fathom joy”**

**“yup, uh, I was diagnosed with depression a long while back, i guess it’s one of those things that you gotta keep in check throughout your entire life”**

**“it...it...it's like everyone can just tell and i'm just a time bomb of sadness”**

“0” or “1” for depression





# Missing Data

- No crystal-clear rules about how to deal with, but **important**
- Types: missing completely at random, missing at random, missing not at random
  - Most models assume no missing data/missing completely at random
- Always:
  - check how much is missing
  - relationship of missingness with other variables
- Some imperfect solutions:
  - Impute (e.g., predict missing values, mean imputation, etc.)
  - Delete

# Raw vs Clean: Text mining example

- RAW: ~245 .txt files with 100k articles
- Excel notes while collecting articles:

› GSRM › LexisNexis Data › DataCollection\_Cleaning\_Sampling › NYT- TXTs ›

Name	Date modified	Type	Size
6_NYT_obesityORobese_2001.2002.TXT	7/22/2016 3:12 PM	Text Document	2,679 KB
7_NYT_obesityORobese_2003.2003.TXT	7/22/2016 3:19 PM	Text Document	2,423 KB
8_NYT_obesityORobese_2004.2004.TXT	7/22/2016 3:23 PM	Text Document	2,449 KB
9_NYT_obesityORobese_2005.2005.TXT	7/22/2016 3:29 PM	Text Document	2,879 KB
10_NYT_obesityORobese_2006.2006.T	7/22/2016 3:43 PM	Text Document	3,628 KB
11_NYT_obesityORobese_2007.2007.T	7/22/2016 7:21 PM	Text Document	2,875 KB
12_NYT_obesityORobese_2008.2008.T	7/22/2016 7:29 PM	Text Document	2,688 KB
13_NYT_obesityORobese_2009.2009.T	7/22/2016 7:33 PM	Text Document	3,071 KB
14_NYT_obesityORobese_2010.2010.T	7/22/2016 7:38 PM	Text Document	2,852 KB

Data Collection Tracking.xlsx - Excel

File Home Insert Draw Page Layout Formulas Data Review View Help Power Pivot Tell me what you want to do

C127 (overweight AND NOT fat AND NOT fatty AND NOT fattier AND NOT fatter AND NOT obese AND NOT obesity)

	A	B	C	D	E	F	G	H
1	Date Downloaded	Content Source	Content Search Terms	Content Dates (Inclusive)	N.Articles	Text Document ID	(IDs are first number on text file in	
152	8/10/2016	The New York Times	(diet OR fitness AND NOT overweight A	Dates: 1/1/1986 - 12/31/1986		14	151 1001-1015	
153	8/10/2016	The New York Times	(diet OR fitness AND NOT overweight A	Dates: 1/1/1987 - 12/31/1987		500	152 1-500	
154	8/10/2016	The New York Times	(diet OR fitness AND NOT overweight A	Dates: 1/1/1987 - 12/31/1987		438	153 501-939	
155	8/10/2016	The New York Times	(diet OR fitness AND NOT overweight A	Dates: 1/1/1988-12/31/1988		500	154 1-500	
156	8/10/2016	The New York Times	(diet OR fitness AND NOT overweight A	Dates: 1/1/1988-12/31/1988		477	155 501-978	
157	8/10/2016	The New York Times	(diet OR fitness AND NOT overweight A	Dates: 1/1/1989 - 12/31/1989		500	156 1-500	

- Sample data: “I flew a KITE. I flew it yesterday! It was pretty fun.”
- “Cleaned” for Word2Vec algorithm:

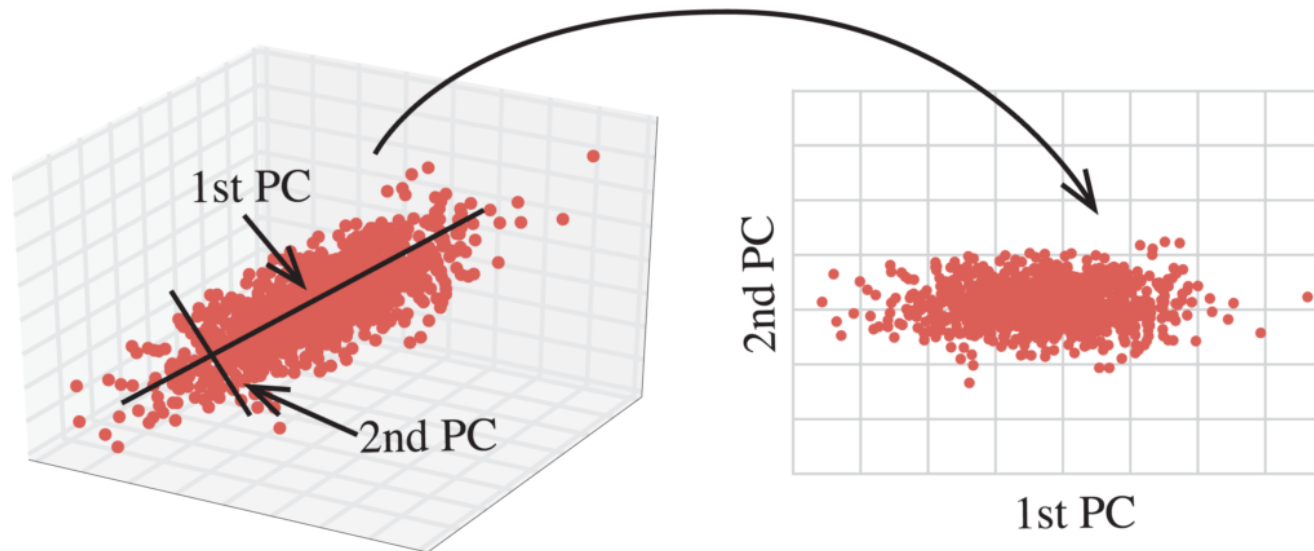
```
sentences= [[i, flew, a, kite], [i, flew, it, yesterday], [it, was, pretty, fun]
```

# Feature Engineering

- Sometimes, *too much* data! (or noise). Or, you want to create a new variable from existing ones
- From theory, previous empirical research (e.g., depression as binary)
- Traditionally, feature engineering was secret sauce to most of data science
- Contrast feature engineering with dimensionality reduction/*learning* features, which is **inductive** (next)


# Dimensionality Reduction

- Let an algorithm *learn* the important features
- Reduce with:
  - Principal Component Analysis, Singular Value Decomposition, etc.



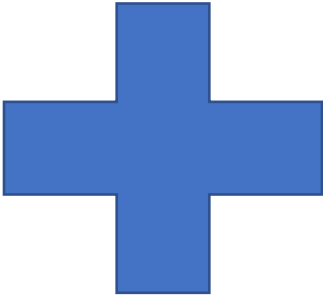
# Long vs Wide Data

	A	B	C	D	E	F	G	H
1	GroupVar	Xvar	Yvar					
2	1	1	34		Xvar	GroupVar1	Groupvar2	GroupVar3
3	1	2	4		1	34	3	14
4	1	3	13		2	4	6	17
5	1	4	15		3	13	12	12
6	1	5	14		4	15	8	11
7	1	6	15		5	14	9	8
8	2	1	3		6	15	12	9
9	2	2	6					
10	2	3	12					
11	2	4	8					
12	2	5	9					
13	2	6	12					
14	3	1	14					
15	3	2	17					
16	3	3	12					
17	3	4	11					
18	3	5	8					
19	3	6	9					
20								
21								



# Merging Data

ID	Age	Gender	Dep_Score	Dep_Binary	Notes
104	26	M	3	0	
241	20	F	6	0	
125	999	999	13	1	
347	23	M	16	1	
835	25	F	0	0	



ID	Twitter_bio_Descript
104	“
241	999
125	
347	999
835	“Happy”



ID	Age	Gender	Dep_Score	Dep_Binary	Notes	Twitter_bio_Descript
104	26	M	3	0		
241	20	F	6	0		
125	999	999	13	1		
347	23	M	16	1		
835	25	F	0	0		

# Getting to know your data vs fishing

- What's the difference?
- The importance of “hanging out with your data”
  - Read, manually
  - Descriptive summaries (e.g., tables, cross tabs, frequencies)
  - Visualizations (may be deceptive!)

# Replicability

- What's the issue?
- Statistical significance is socially constructed
- File-drawer issue
- Types of replicability
  - Can the research be replicated on the same dataset? (but code/resources change...)
  - Is it replicable (generalizable) to other data sets? (but society/ppl change...)
- Pre-registration
  - Inductive vs deductive?
- Triangulation vs replication
- Open-source code, detailed appendices, ideally open source data



# Errors and outliers

Accuracy rates of models to “detect depression” from language:  
>90%\*

What is the most costly error? *Who are the errors?*

	Has Depression	No Depression
Depression Detected		
No Depression Detected		

“Accuracy” according to algorithms vs psychiatrists

# Sample 3 CompSoc Project Workflow (Alina)

Time	Activities
Winter 2015	Interested in Word2Vec (Jacob suggested) Read about it online
Spring 2016	<i>Idea:</i> How is obesity portrayed in news? <i>Method:</i> Word2Vec language model <i>Data:</i> New York Times, from Lexis Nexis
July-August 2016	Collect 100k news articles from LexisNexis. Struggle to clean (~2 months).
August-Sept 2016	Mess around with Word2Vec on news a LOT, seemed promising. Read about obesity/news.
Winter 2016	Presented Word2Vec to CompSoc
Fall 2016-Spring 2017 <i>Spring: MA Thesis due</i>	Narrow RQs. Test out with Word2Vec/news data Check robustness many times, get confused many times, make visuals in R Wrote up results (theory still vague)
Spring 2017-Spring 2018	Revise paper, theory, more robustness checks, ideas for next projects
June 2018	Prettify code, put on Github, Present to CompSoc
Now:	Minor revisions, submitting for peer review
Throughout:	Reading, getting confused, etc.

# Sample 3 CompSoc Abstract (Alina + Jacob)

- Over half (64%) of Americans want to lose weight; eating disorders and weight-based discrimination run rampant. These overwhelmingly negative conceptions of fat are often attributed to media influence, suggesting this is a process of cultural learning. But it remains unclear exactly how public culture becomes private culture.
- We provide a **computational account of this cultural learning, showing how schemata about obesity can be learned from news reporting.** We extract these schemata from New York Times articles with Word2Vec, a model that learns language in ways that are inspired by our own cognition. We identify several cultural schemata around obesity, linking it to femininity, immorality, poor health, and low socioeconomic class.
- Such schemata may be subtly but pervasively activated by our language; thus, language may be one vehicle for the reproduction of biases around body weight and health. Finally, findings validate concerns that machine-learned algorithms may encode, and reproduce, negative biases.

# Sample 3 CompSoc File Management

- (See sample project set up in Alina's dropbox)
- How could this set-up have been better? What is good?

# Developing your CompSoc project

- What kinds of topics are you interested?
  - What kinds of methods are you interest in?
  - What kinds of data sets or types are you interested in?
    - What is the data structure?
    - What structure does it need to be in for it to be “clean” (for analyses)?
  - What theories stand out to you?
  - What is a possible workflow for a project you are (or want to be) working on?
- 
- 15 min: Come up with ~3-5 research areas or questions either on own or with neighbor
  - 5 min: Discussion

# Questions

- Next: Data Wrangling and Cleaning in Python (hands-on)