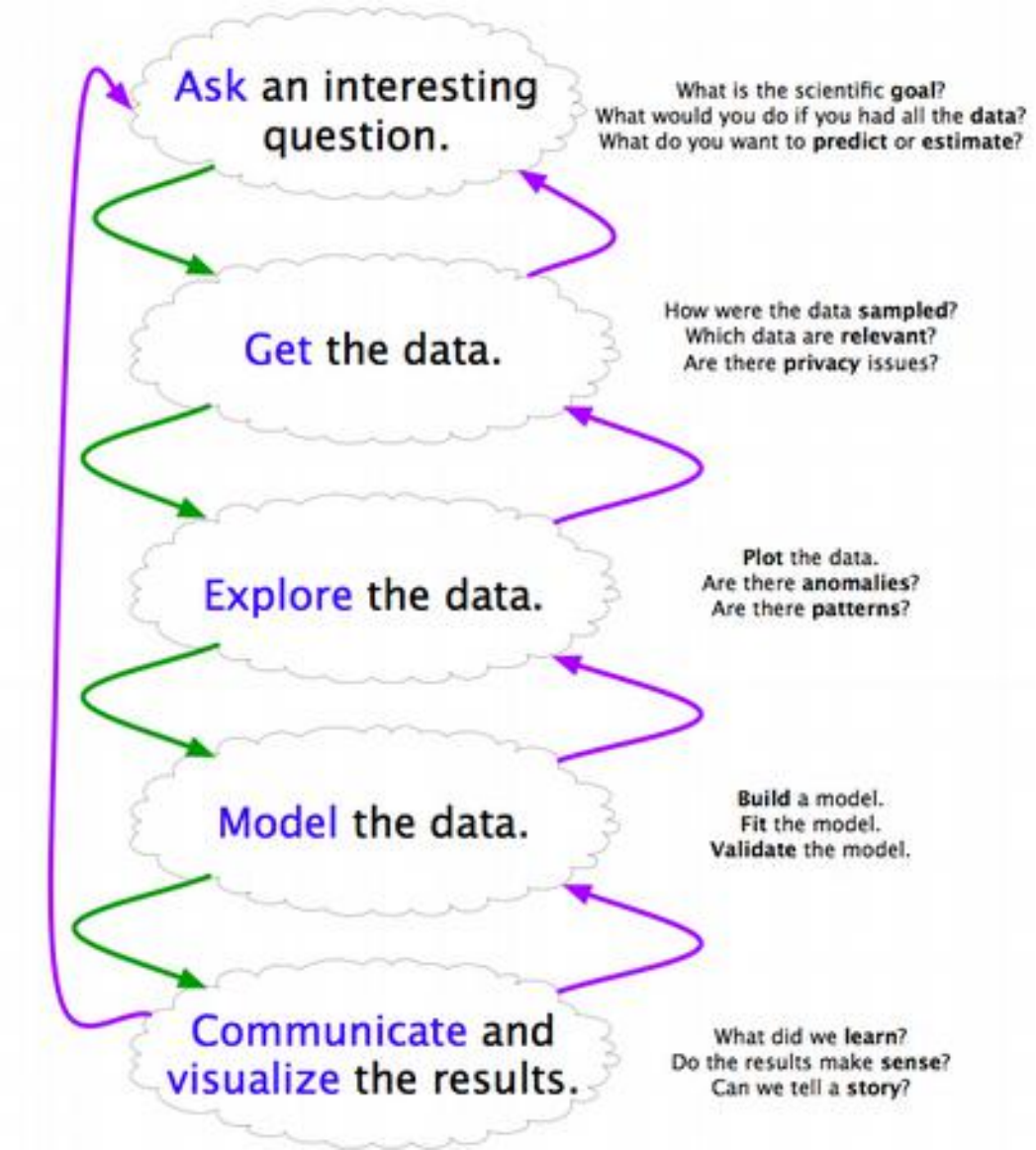
Introduction to Data Science: Course Project

Gerard de Melo
http://gerard.demelo.org

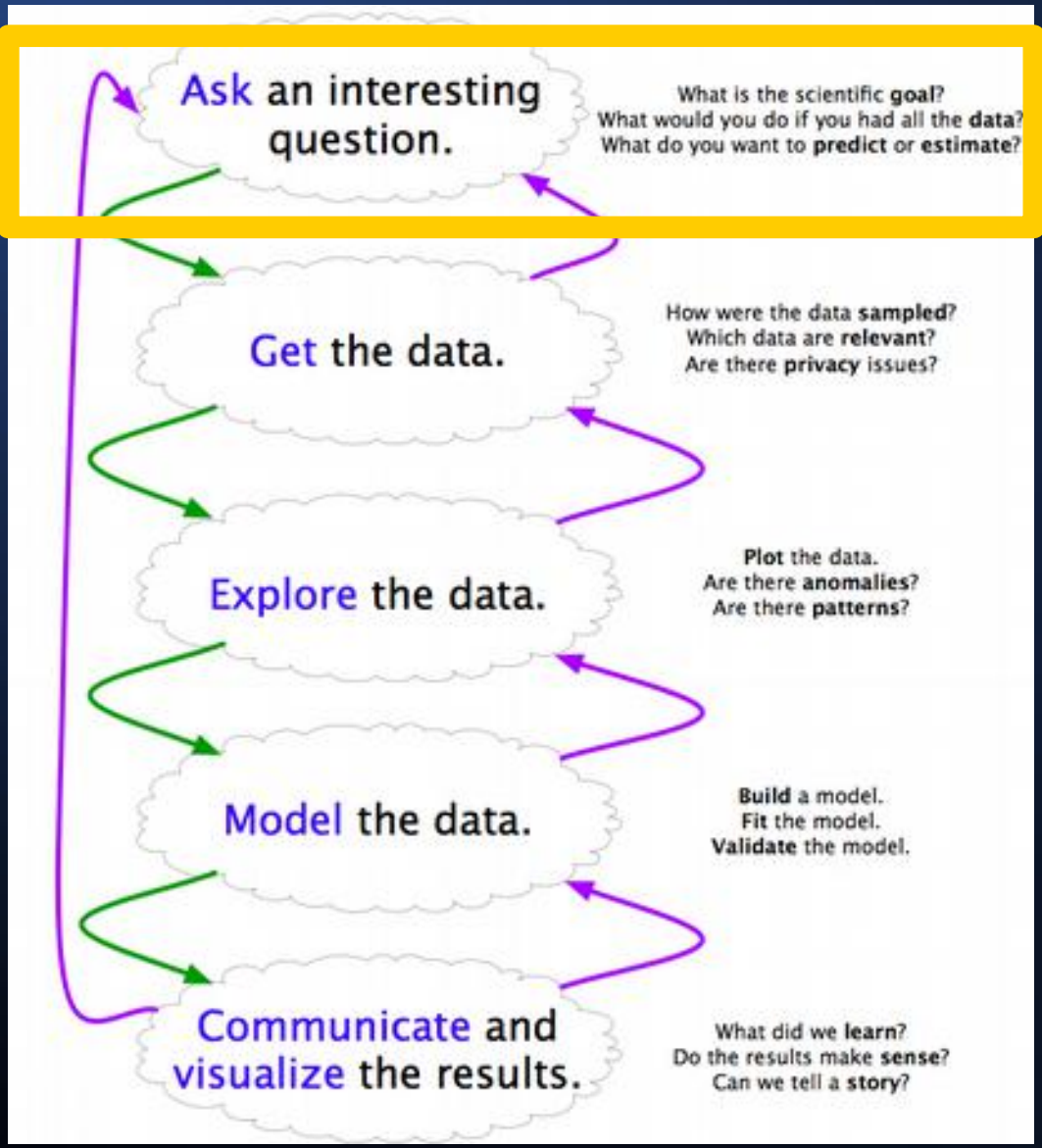
Rutgers University

Course Project



Source: Hanspeter Pfister, Joe Blitzstein, Verena Kaynig. Harvard CS109 2015

Course Project

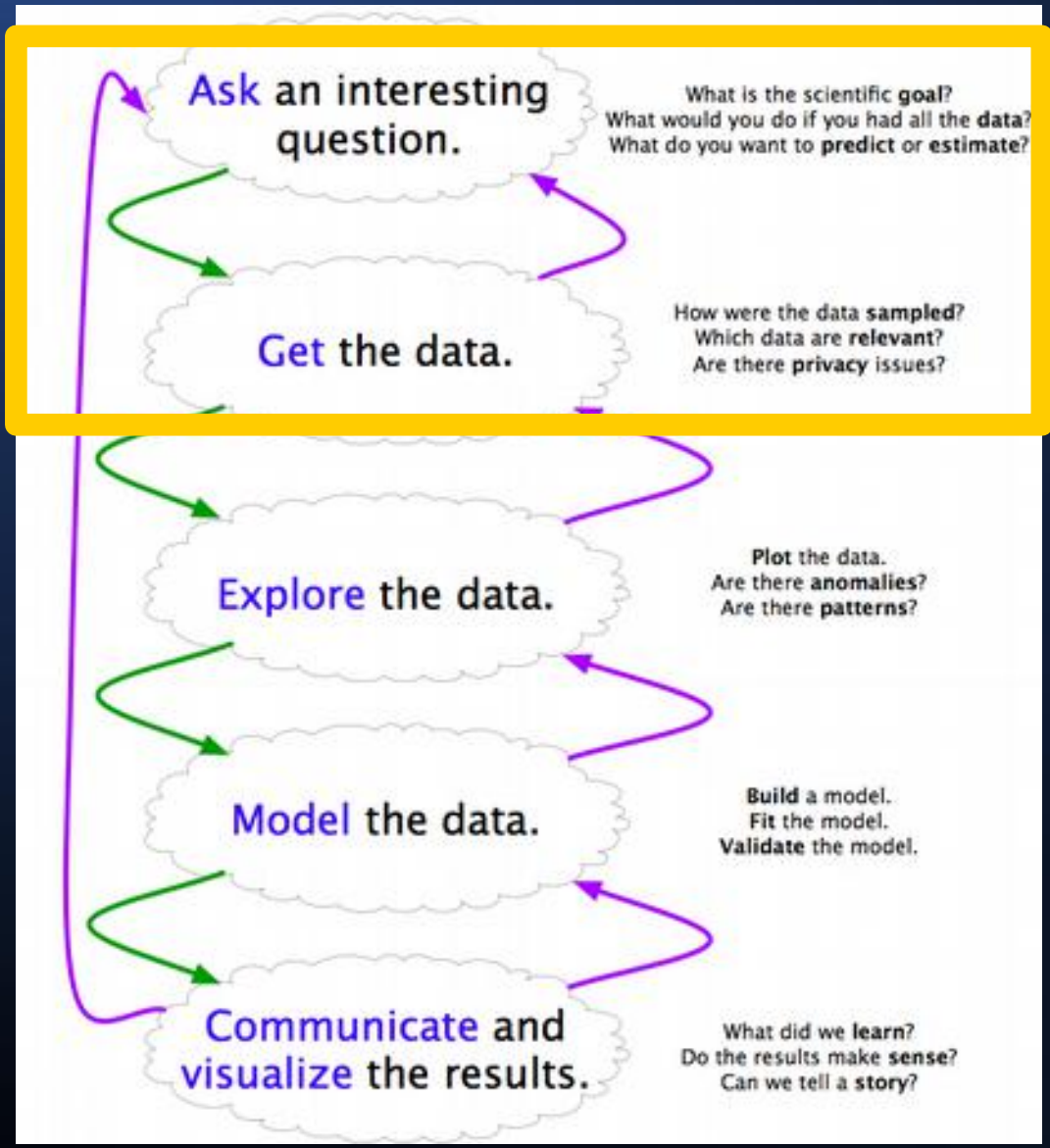


Source: Hanspeter Pfister, Joe Blitzstein, Verena Kaynig. Harvard CS109 2015

Course Project: Topic

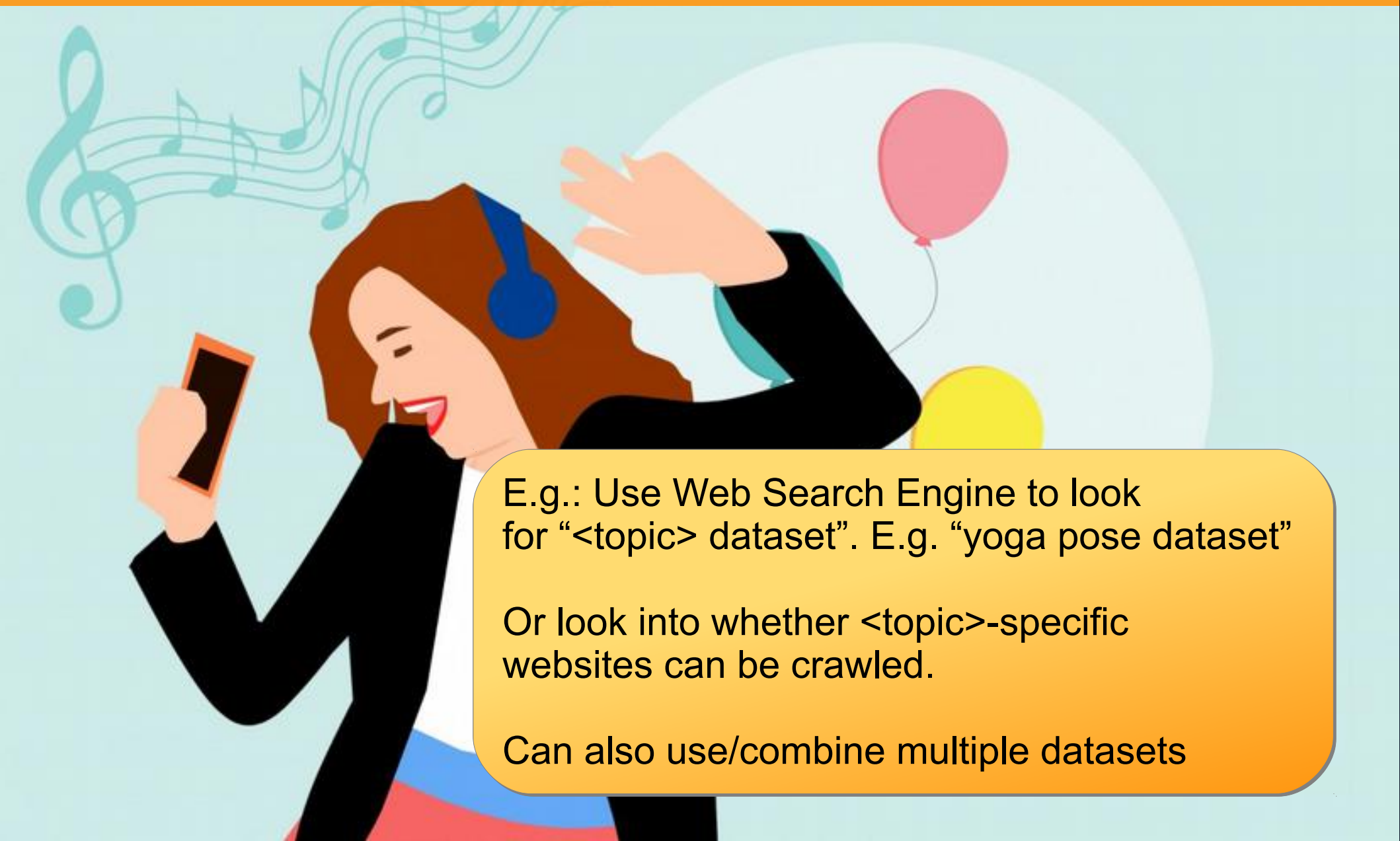
Music?
Travel?
Art?
Gaming?
etc.

Course Project



Source: Hanspeter Pfister, Joe Blitzstein, Verena Kaynig. Harvard CS109 2015

Course Project: Data



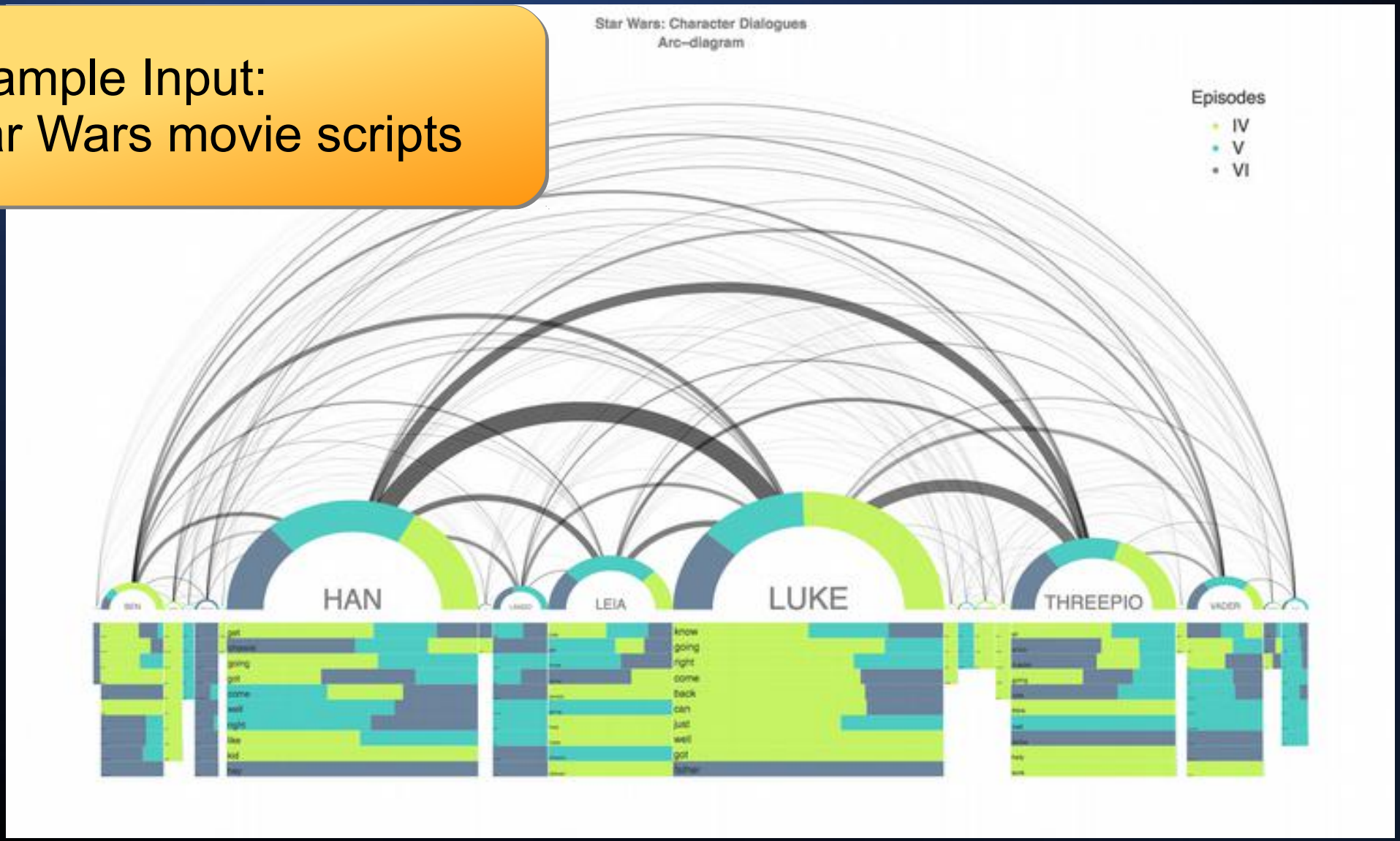
E.g.: Use Web Search Engine to look for “<topic> dataset”. E.g. “yoga pose dataset”

Or look into whether <topic>-specific websites can be crawled.

Can also use/combine multiple datasets

Course Project: Data

Example Input:
Star Wars movie scripts



https://github.com/gastonstat/StarWars

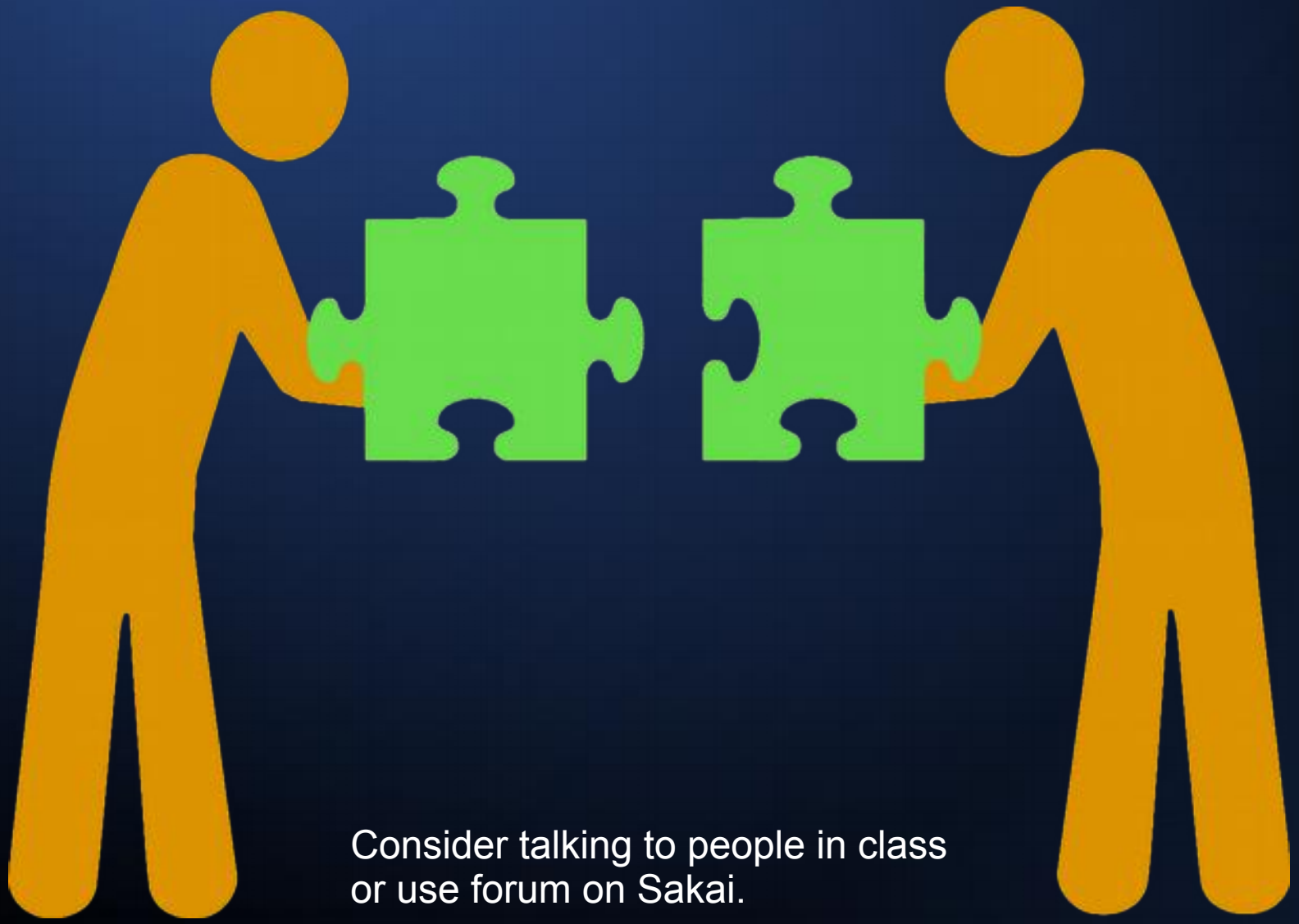
Course Project

Step 1: Short Project Proposal (by Feb. 17)

Just a 2-3 paragraph description of what you are planning to do and hoping to achieve, what data sources you hope to use, etc.

This can still be changed later, with approval from us.

Course Project: Team?



Consider talking to people in class or use forum on Sakai.

Course Project: Team?

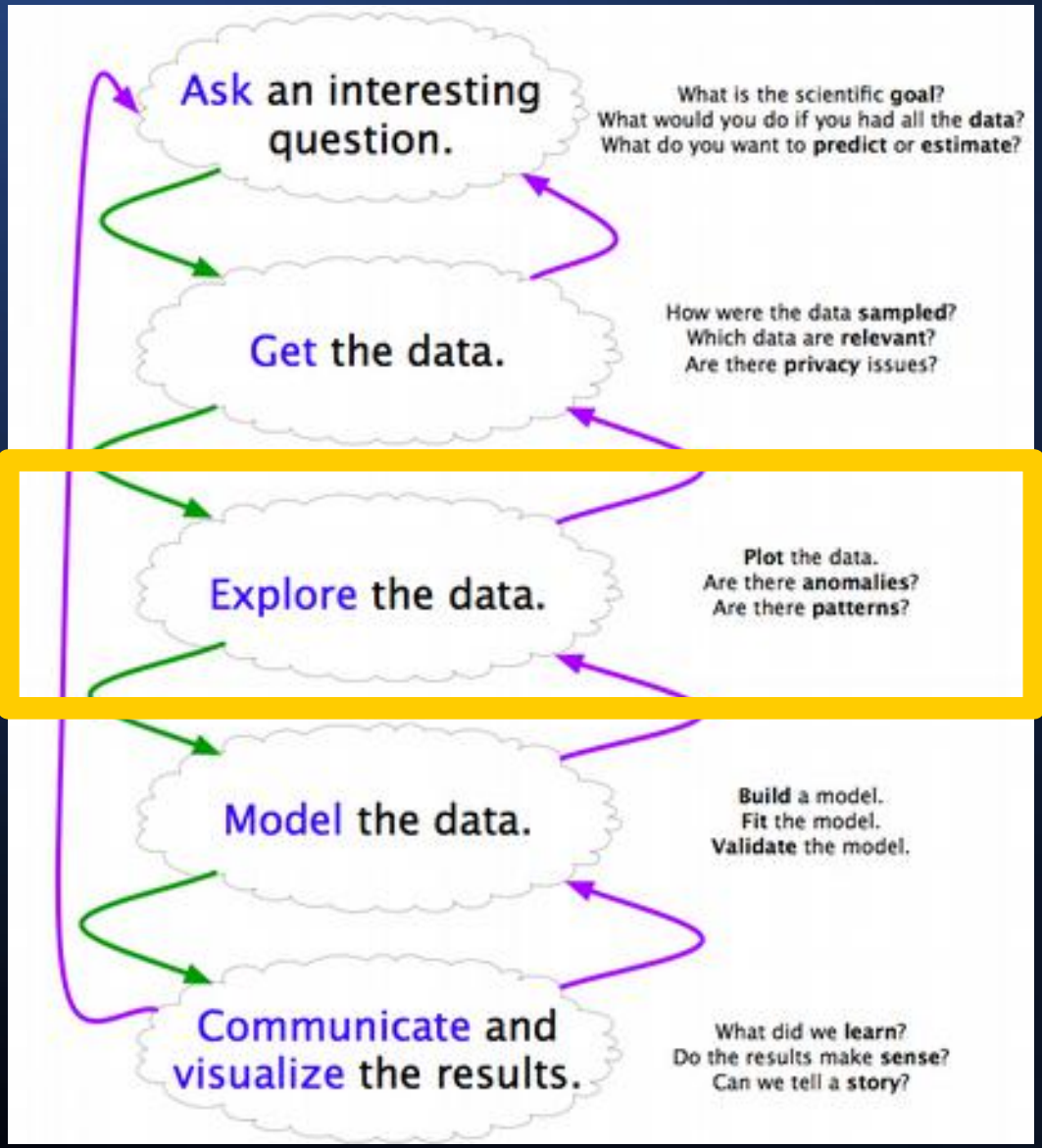
Teams

Team Size: 1 or 2
(exceptions only with prior approval,
for particularly large/challenging projects)

Teams normally should not change after submitting the proposal.

Grade: Equal for all team members
However, we reserve the right to deviate from this if the contributions were particularly unequal.

Course Project



Source: Hanspeter Pfister, Joe Blitzstein, Verena Kaynig. Harvard CS109 2015

Course Project: Deliverables

Intermediate Milestones as Homework

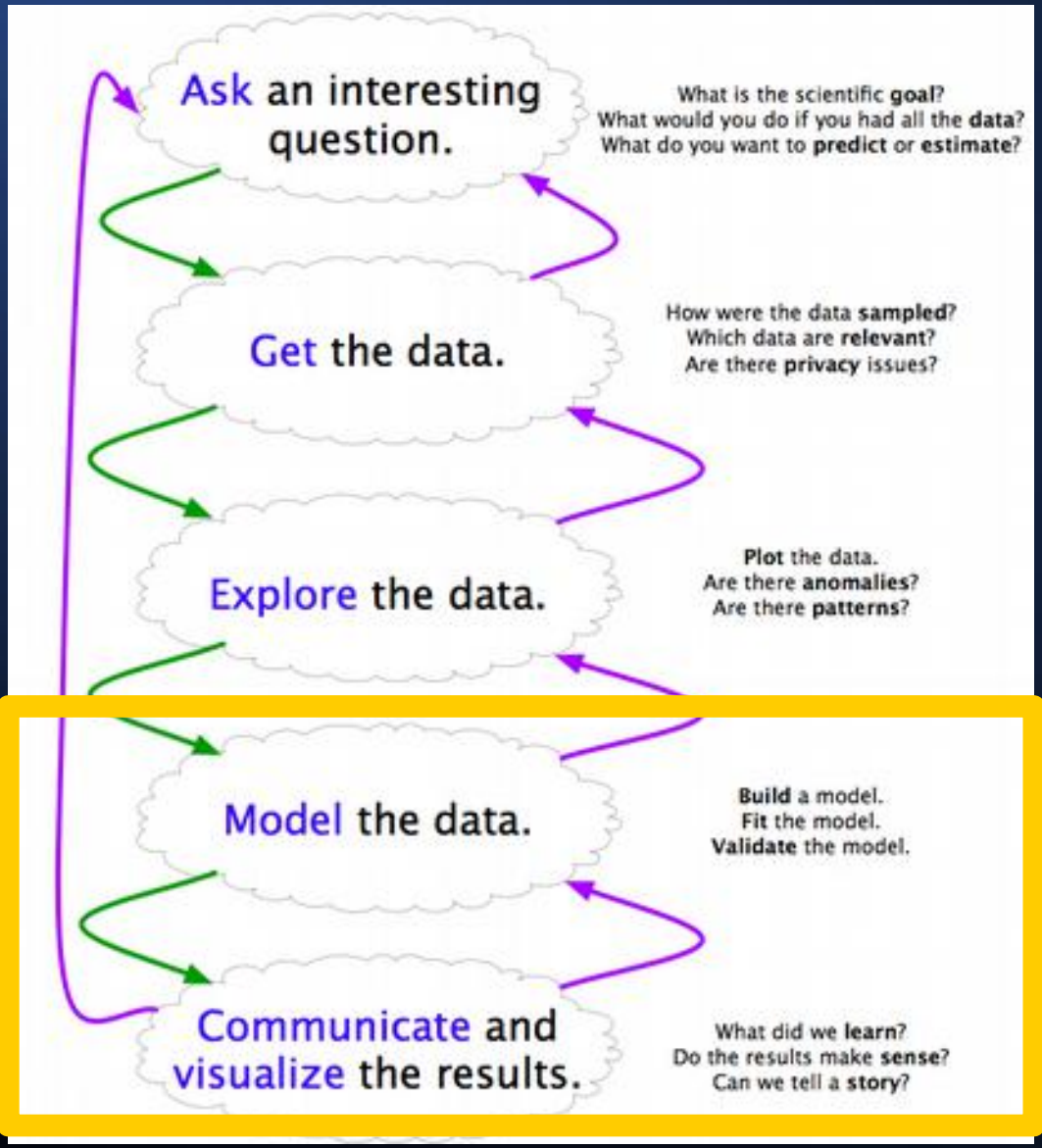
Dates: TBA

E.g.:

Describe data collection/source of data,
data format, data preprocessing.

Conduct data analysis, create visualizations

Course Project



Source: Hanspeter Pfister, Joe Blitzstein, Verena Kaynig. Harvard CS109 2015

Course Project: Deliverables

Short Project Presentations (Likely Apr. 28/30)

Very short (5-10 minute) presentation of your work – probably not for all groups

May use slides or interactively demonstrate your system.

Course Project: Deliverables

Final Report (by May 11)

1. Should include some machine learning experiments. Could also be a practical application.

2. Explain and evaluate your results, numerically or via visualizations. Should show what insights have been gained, or how well your method works.

Course Project: Deliverables

Final Report

3. Describe related work
Describe in what ways your work is different from
what others have done online in blog posts/Kaggle/etc.

4. Conclusion section
What insights did you gain? What worked, what didn't work?
What else would you do if you had more time (or could start over)?

6. Acknowledgments section (mandatory!)
Mention all libraries used, all third-party material used!
You may not use third-party images without attribution.

Course Project: Academic Dishonesty

Image: Taken movie

Course Project: Details

Rules

You may use any external libraries, as long as you explicitly mention this in an “Acknowledgments” section in your report.

Any third-party material used, even if modified or translated from a different programming language, must be mentioned in the “Acknowledgments” section in your report. Clearly indicate the extent of your own contribution.

Course Project: Details

Submission Details

All deadlines refer to 11:59pm Eastern time.
Late submissions at discretion of instructor, but with grade penalty.

1) Include Source Code as Juypter Notebook/ZIP file/etc.

2) Please do not submit large original input data files.

Course Project: What to Focus on

Key Elements for a Good Grade

Put in effort

Sophisticated analysis methods: machine learning, etc. Beautiful visualizations are also nice, but primarily if they support analysis and insights about data.

Interesting idea:
Imagine something that would score highly on Reddit or even that newspapers would write about

Course Project: Topic Examples (just for Inspiration)

stackoverflow

Developer Jobs Hire Developers Search …

Company Code for a Living Insights Engineering Podcasts Developer Hiring Blog

What Programming Languages Are Used Late at Night?

by David Robinson on April 19, 2017

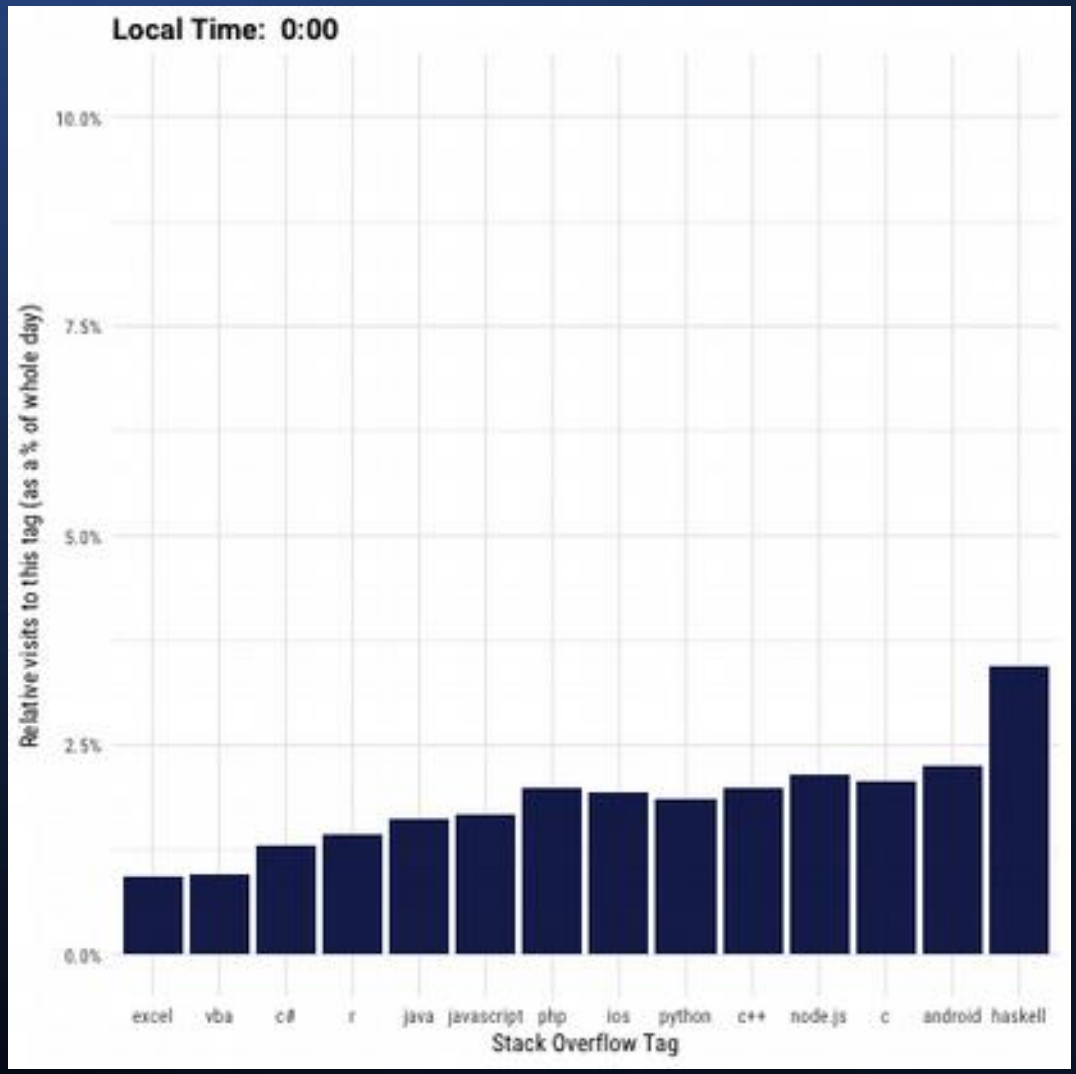
Do you program in the evening? How about late at night?

I enjoy programming as a hobby and I'm a bit of a night owl, so I often code well outside working hours, sometimes past midnight. And whatever time it is, when I run into a bug I visit Stack Overflow.

This means that Stack Overflow data can give us insights into what kinds of developers program in the evening or night, and conversely what programming languages are used only during the workday. In this post, we'll examine what tags are visited at what times a day, and also compare major cities to see how their working hours differ.

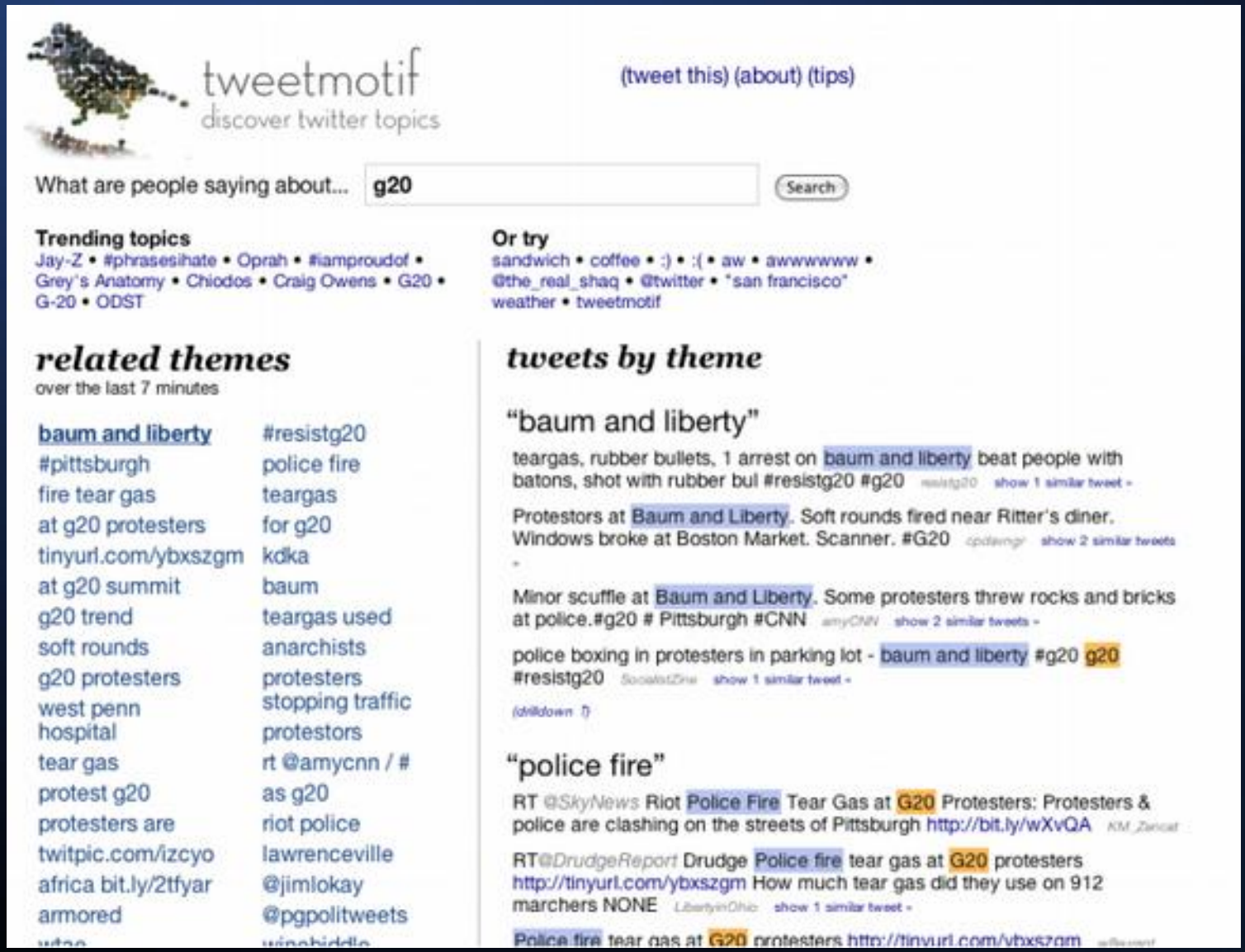
https://stackoverflow.blog/2017/04/19/programming-languages-used-late-night/

Course Project: Topic Examples (just for Inspiration)



https://stackoverflow.blog/2017/04/19/programming-languages-used-late-night/

Course Project: Topic Examples (just for Inspiration)



O'Connor et al. TweetMotif: Exploratory Search and Topic Summarization for Twitter

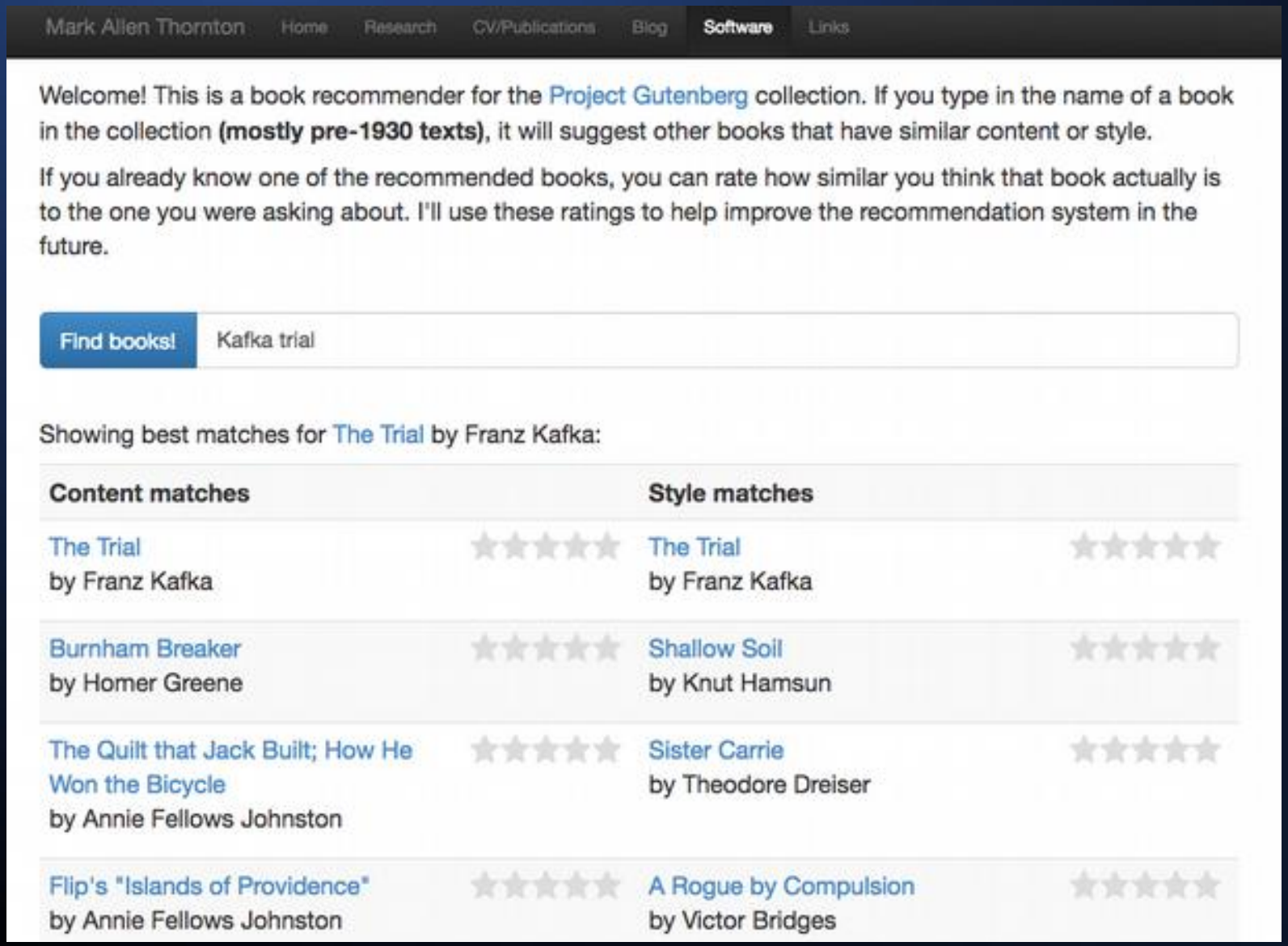
Course Project: Topic Examples (just for Inspiration)

Sample of Word Anomalies

The Bible (King James Edition); Anonymous / Various	
Frequent:	*unto, lord, isreal, shall, god, moses, jesus, david, offering, tabernacle*
Infrequent:	*girl, boy, school, success, condition, listen, princess*
Wonderful Wizard of Oz; Baum, Frank	
Frequent:	*woodman, scarecrow, witch, tin, emerald, monkeys, kansas, brains, winged*
Infrequent:	*mother, money, soul, natural*
White Fang; London, Jack	
Frequent:	*musher, beaver, sled, dogs, cherokee, snarl*
Infrequent:	*letter, person, window, green, sweet, loved, party, paper*
The Republic; Plato	
Frequent:	*guardians, unjust, true, injustice, state, gymnastic, rulers, democractical*
Infrequent:	*miss, girl, boy, prince*
Alice's Adventures In Wonderland; Carroll (C.L. Dodgson), Lewis	
Frequent:	*gryphon, turtle, caterpiller, mock, dodo, mouse, rabbit, hedgehog*
Infrequent:	*death, country, happy, fair, common*
Origin of the Species; Darwin, Charles	
Frequent:	*species, varieties, subaerial, selection, sterility, plants, modification, forms, variability*
Infrequent:	*person, government, love, thinking, god, evil, fire*
Communist Manifesto; Marx, Karl/Engels, Friedrich	
Frequent:	*bourgeois, proletariat, communists, antagonisms, revolutionising, socialism, production, class, feudal, reactionary, exploitation, conditions, crises*
Infrequent:	*said, love, why, heart, mother, poor, felt*
Paradise Lost; Milton, John	
Frequent:	*wonderous, heaven, satan, dominations*
Infrequent:	*country, church, horses, sister*
Apology; Plato	
Frequent:	*corrupter, accusers, demigods, socrates, oracle, indictment*
Infrequent:	*she, work, morning, replied, body*
Gargantua and Pantagruel; Rabelais, Francis	
Frequent:	*codpiece, catchpole, ballocks, dingdong, fart, chitterlings, gymnast, arse*
Infrequent:	*smile, existence, feelings, british, professor, suffering*

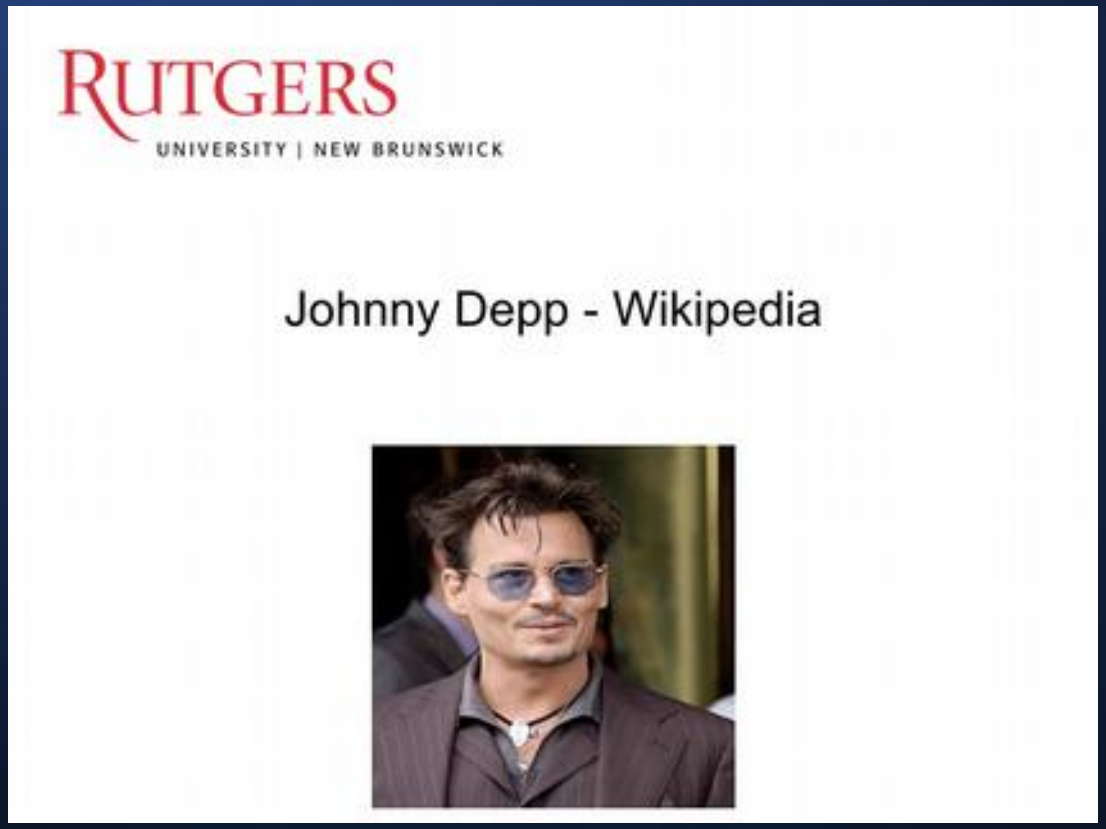
http://www.mine-control.com/zack/guttenberg/

Course Project: Topic Examples (just for Inspiration)



http://markallenthornton.com/recommender/

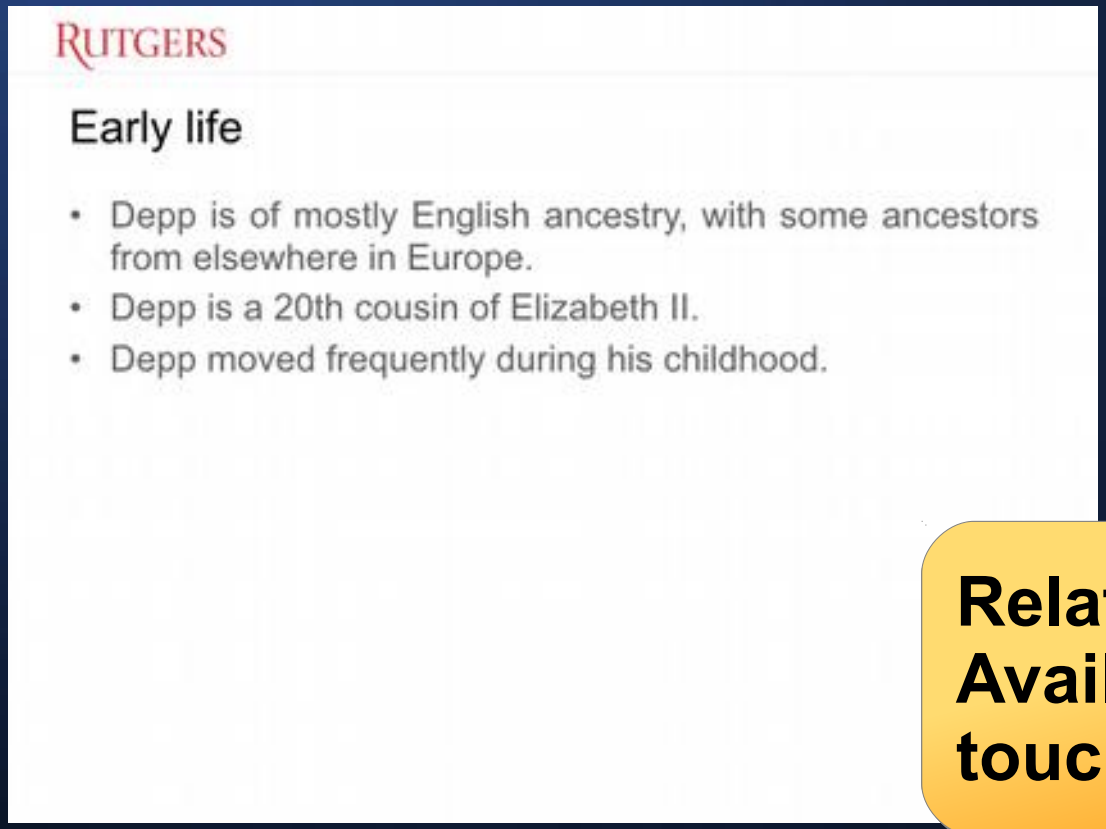
Course Project: Topic Examples (just for Inspiration)



Wikipedia Article → PPT

Darshan Sonagara, Raj Inamdar, Jyoti Chaudhary, CS 674 2017

Course Project: Topic Examples (just for Inspiration)

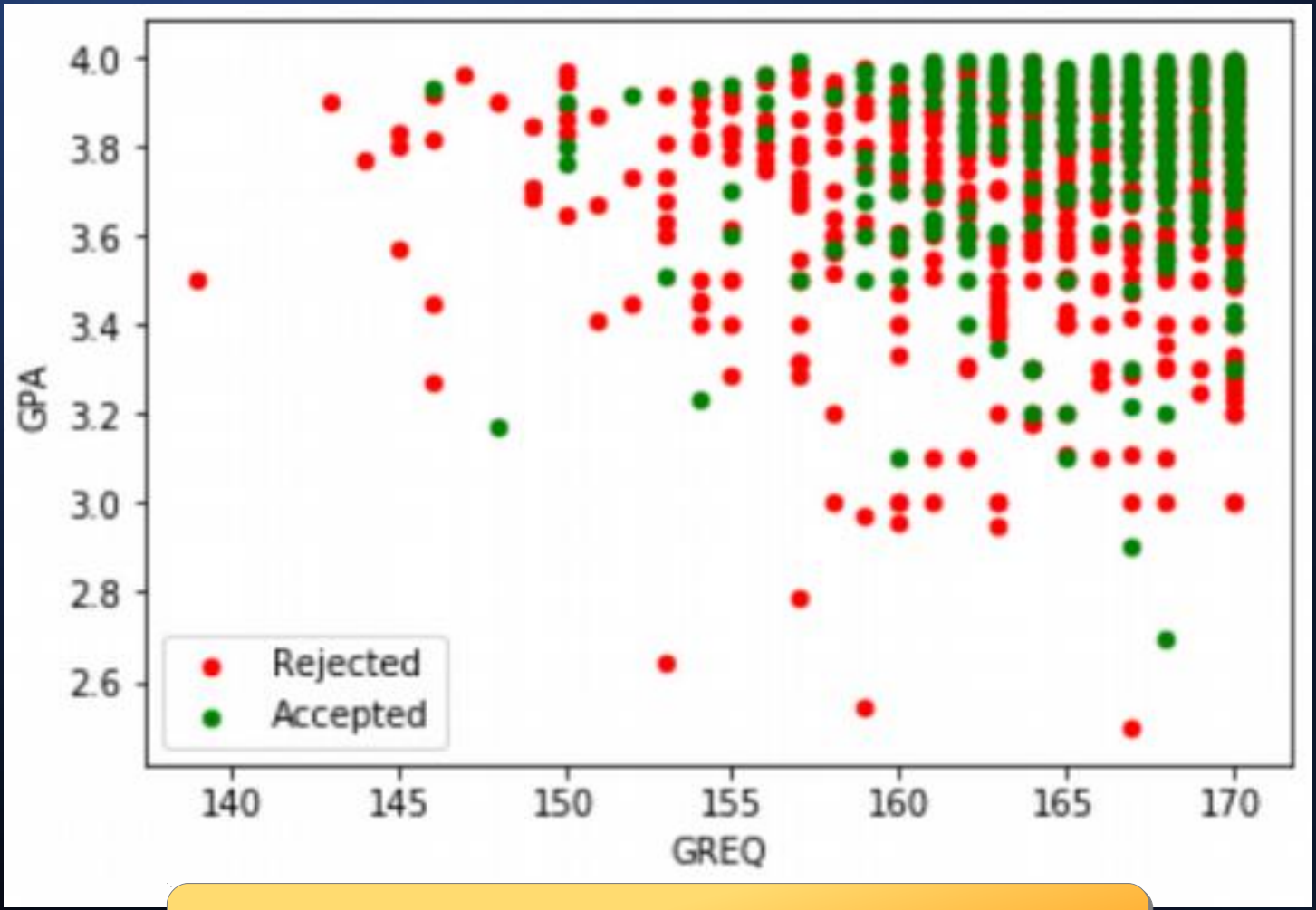


Related Topics Available. Get in touch with me.

Wikipedia Article → PPT

Darshan Sonagara, Raj Inamdar, Jyoti Chaudhary, CS 674 2017

Course Project: Topic Examples (just for Inspiration)

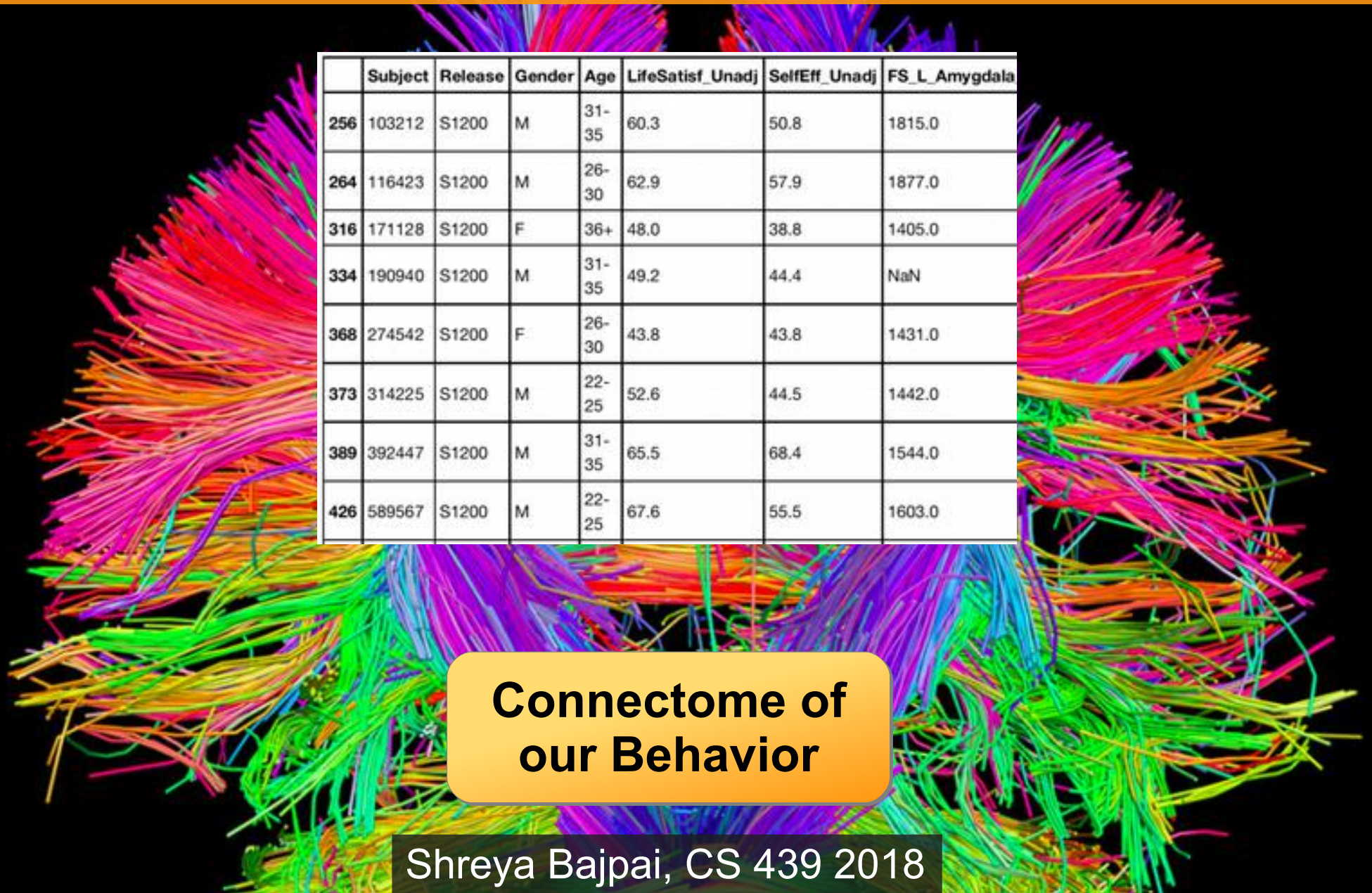


College Acceptance Statistics

Nithin Tammishetti, CS 439 2018

Course Project: Topic Examples (just for Inspiration)

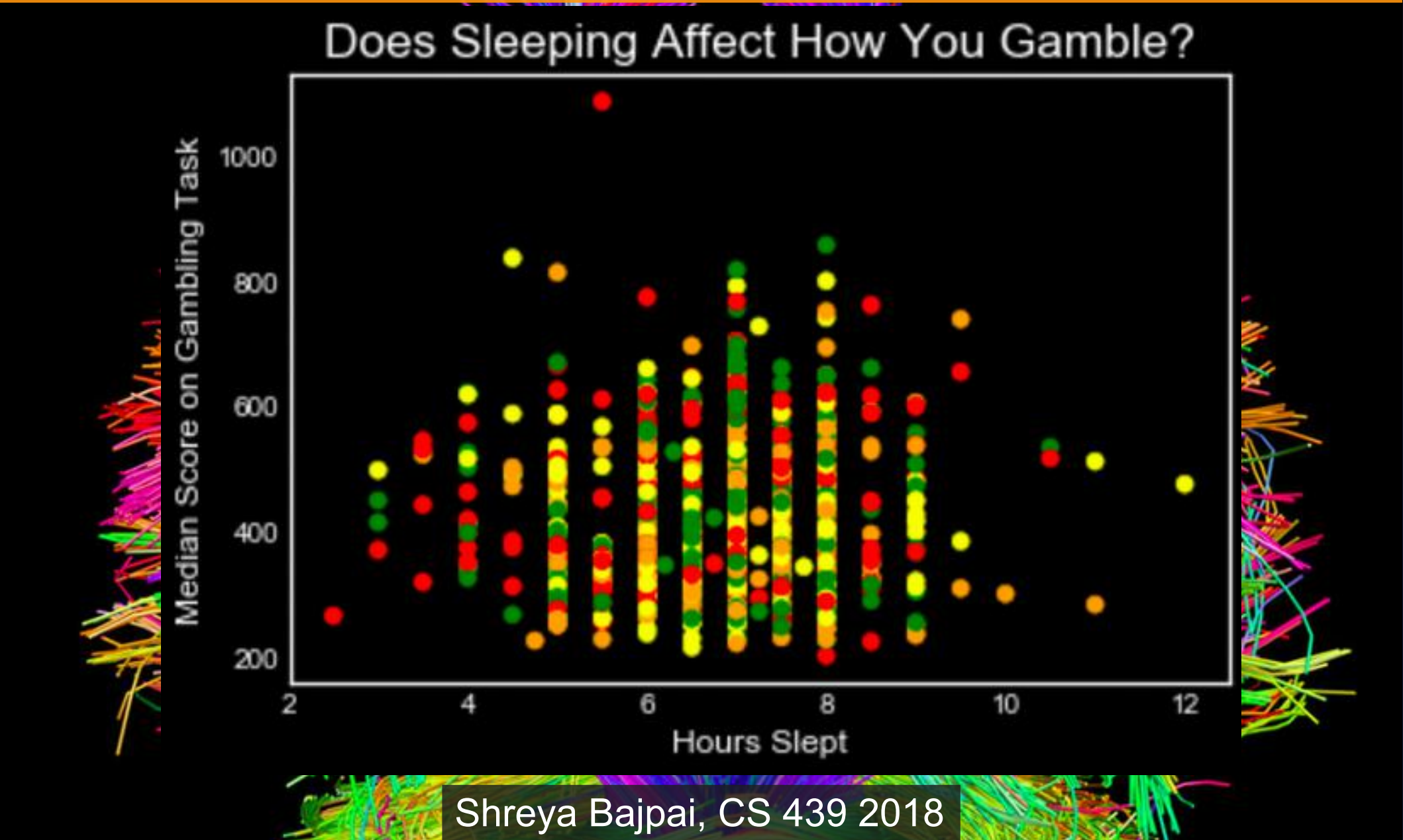
	Subject	Release	Gender	Age	LifeSatisf_Unadj	SelfEff_Unadj	FS_L_Amygdala
256	103212	S1200	M	31-35	60.3	50.8	1815.0
264	116423	S1200	M	26-30	62.9	57.9	1877.0
316	171128	S1200	F	36+	48.0	38.8	1405.0
334	190940	S1200	M	31-35	49.2	44.4	NaN
368	274542	S1200	F	26-30	43.8	43.8	1431.0
373	314225	S1200	M	22-25	52.6	44.5	1442.0
389	392447	S1200	M	31-35	65.5	68.4	1544.0
426	589567	S1200	M	22-25	67.6	55.5	1603.0



Connectome of our Behavior

Shreya Bajpai, CS 439 2018

Course Project: Topic Examples (just for Inspiration)



Shreya Bajpai, CS 439 2018

Course Project