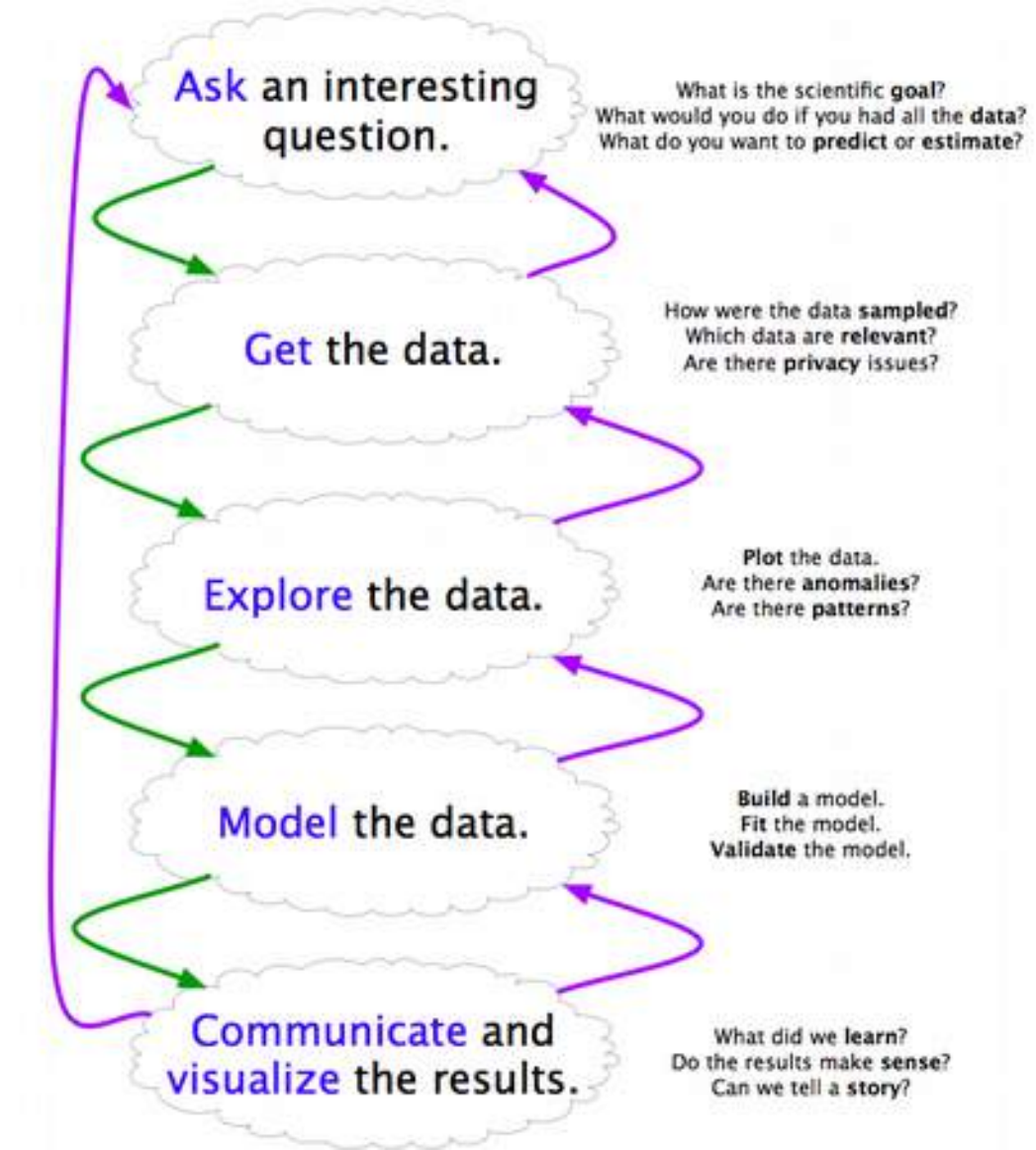
Massive Data Storage and Retrieval: Course Project

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

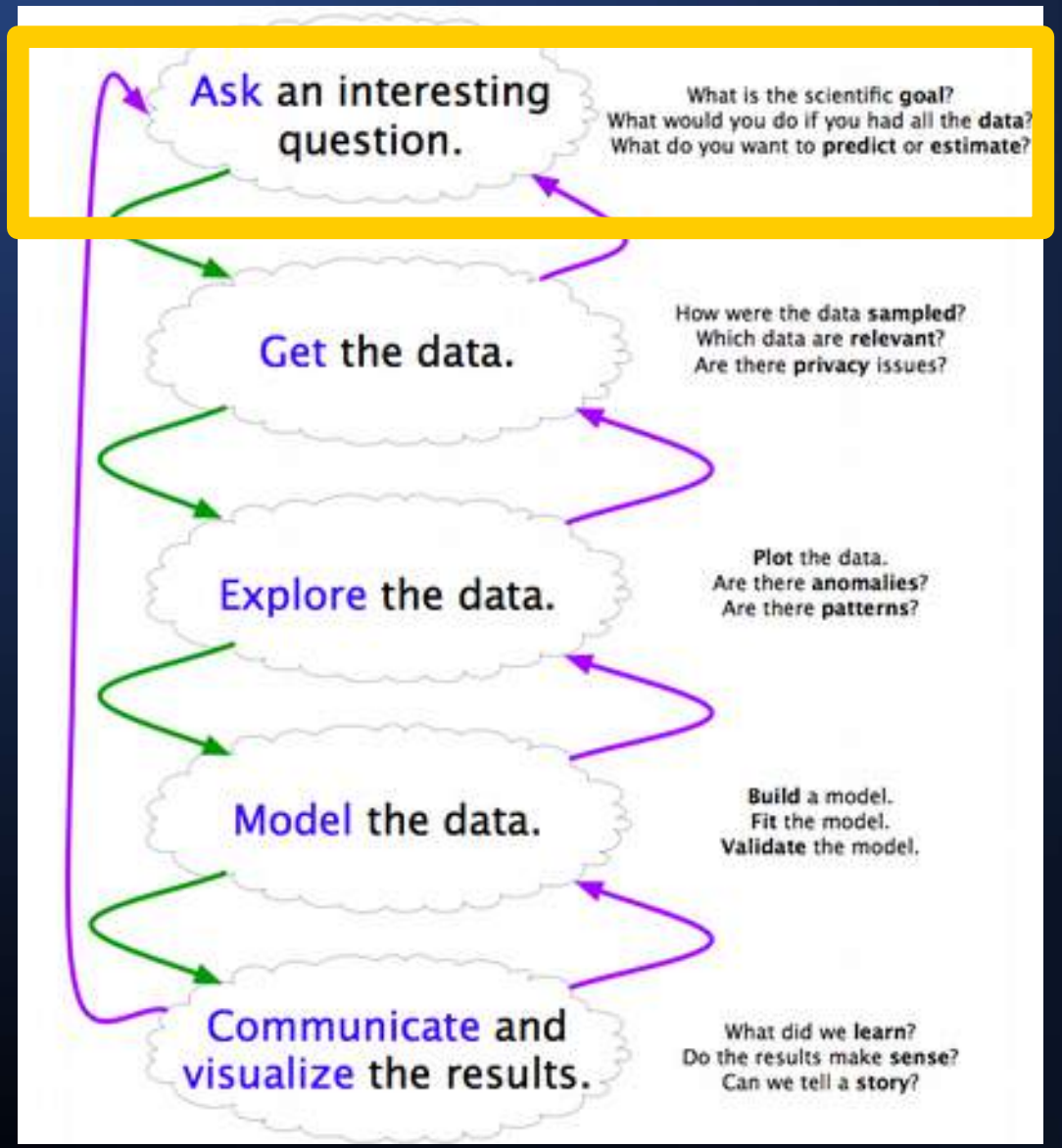
Rutgers University

Course Project



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

Course Project



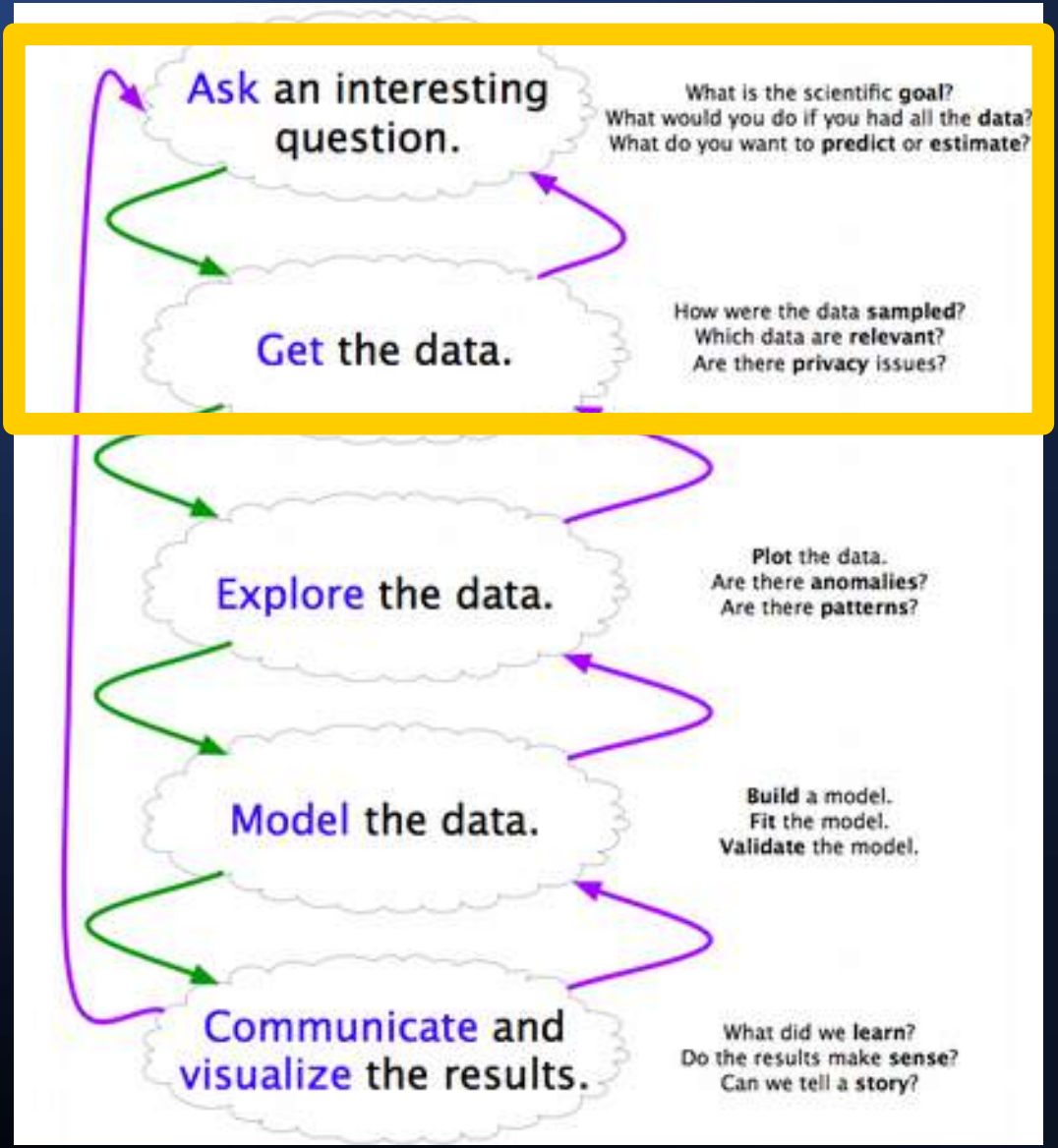
Source: Hanspeter Pfister, Joe Blitzstein, Verena Kaynig. Harvard CS109 2015

Course Project: Topic

What are your hobbies?
What are you most passionate about?

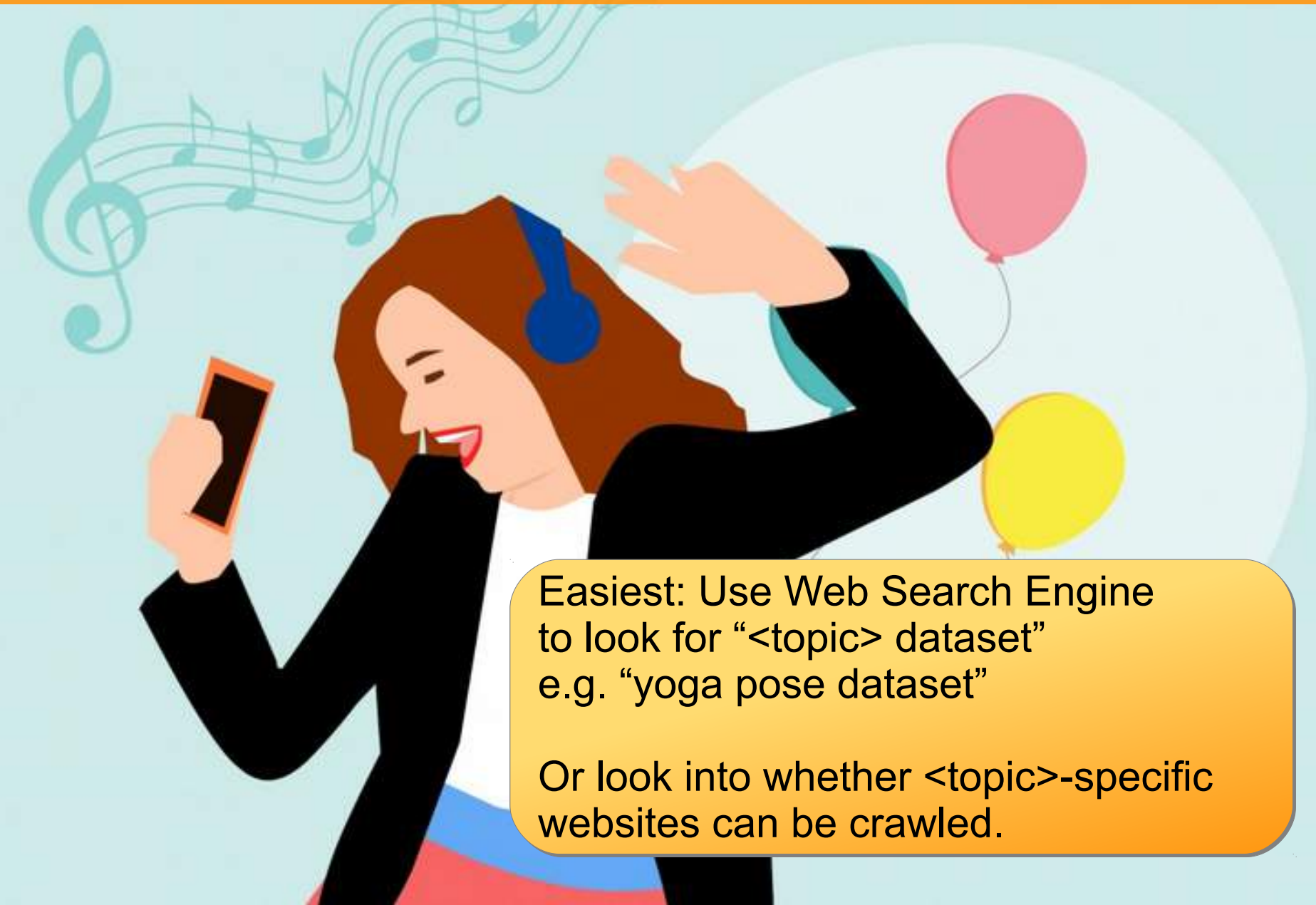
Music?
Travel?
Art?
Gaming?
etc.

Course Project



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

Course Project: Data

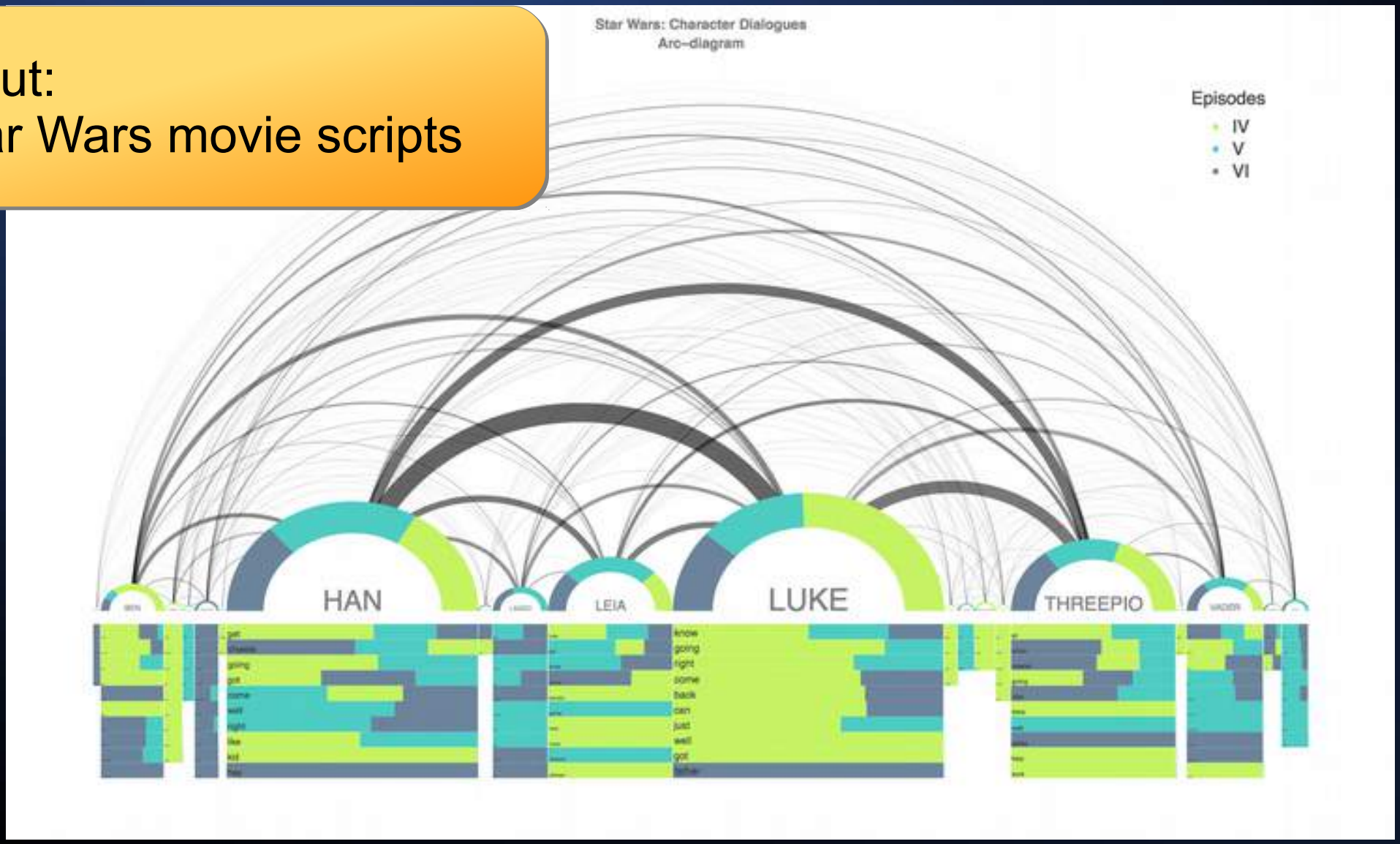


Easiest: Use Web Search Engine
to look for “<topic> dataset”
e.g. “yoga pose dataset”

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

Course Project: Topic Examples (just for Inspiration)

Input:
Star Wars movie scripts



https://github.com/gastonstat/StarWars

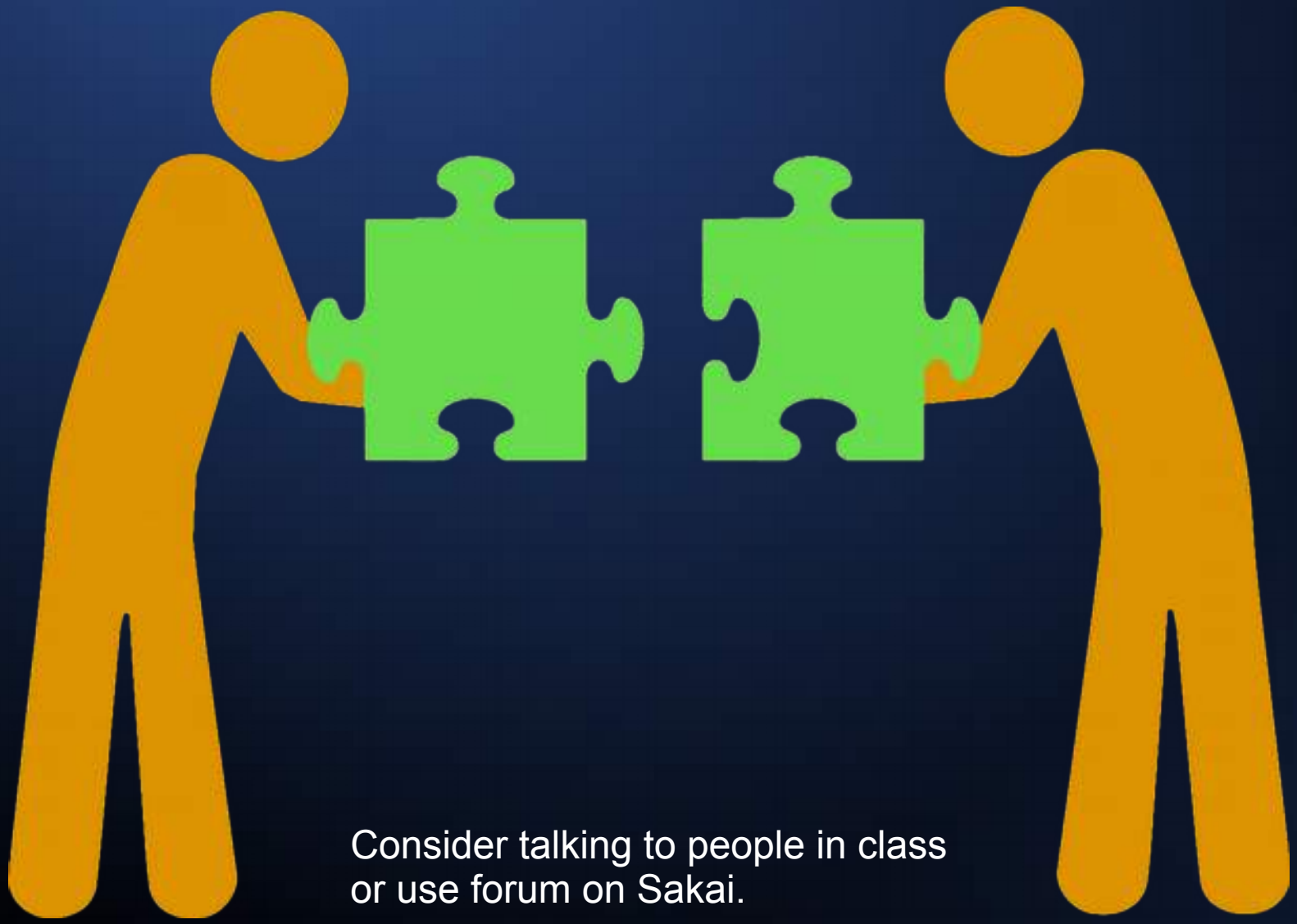
Course Project

Step 1: Short Project Proposal (by Sep. 22)

Just a short one paragraph description of
what you are planning to do and hoping to achieve.

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?

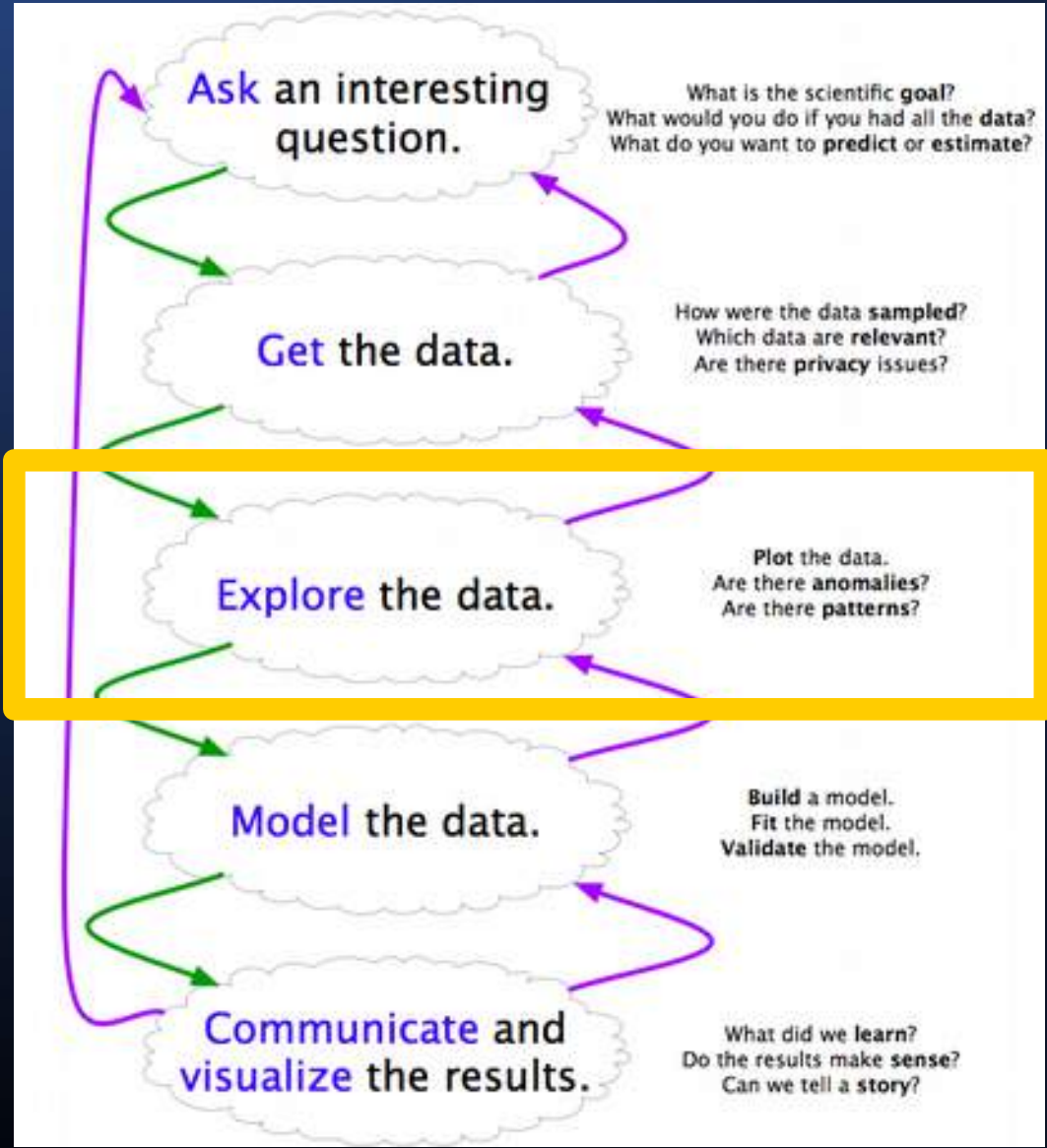
Teams

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

Milestone: Intermediate Report (by Oct. 16)
Will count as Homework!

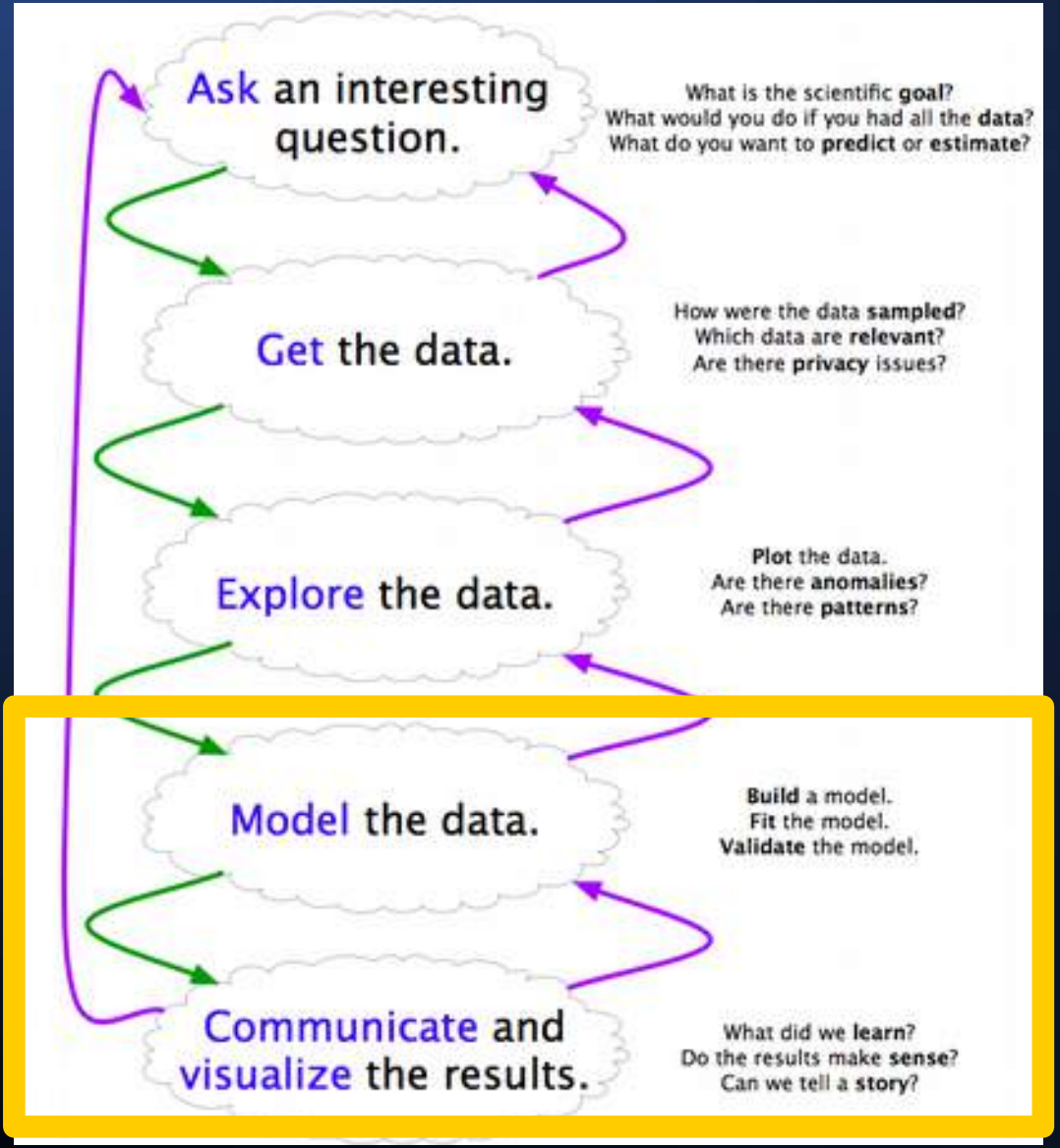
1. Describe project goals and why it is interesting

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

3. Describe contents of data in detail
(must use Spark to analyse it; optionally visualize it)

4. Describe possible applications of this data,
Including your ideas for the next phase.

Course Project



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

Course Project

Short Project Presentations (December 11)

Very short (5-10 minute) presentation of your work – only for some groups

May use slides or interactively demonstrate your system.

Course Project

Final Report (by December 14)

1. Improve on intermediate report.
The final report supersedes the intermediate report,
so all crucial results from the intermediate one
should be repeated.
Note: You can also analyse multiple related datasets.

2. Conduct machine learning experiments
on your data. Ideal goal: Practical application.

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

Course Project

Final Report (by December 14)

4. Describe related work
Cite related research papers.

5. 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

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.

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

Course Project

Report Format

Option 1) PDF:
Written like an academic technical report, typically at least 4 pages.
Recommendation: ACM or ACL 2017 LaTeX stylesheets.

Option 2) Online Notebook:
Notebook with detailed descriptions integrated (i.e. as much text as would be in a regular report!)
Important: Provide both PDF and notebook files!

Course Project

Attachments to Submit

1) Source Code as ZIP
unless all of your source code is already in the notebook

2) Example Outputs of your System
i.e. examples of the websites (or slides, videos) that your program generates, in original form as output by your program, without any manual post-editing (instead consider modify any templates that your program use as inputs).

Course Project: What to Focus on

<u>Key Elements for a Good Grade</u>

Sophisticated analysis methods (machine learning, deep learning, etc.)

Beautiful visualizations are nice but should <u>not</u> be the main focus. Analysis and insights about data are the main goals here.

Interesting idea (imagine something that would score highly on Reddit or even that newspapers would write about)

Course Project: Topic

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

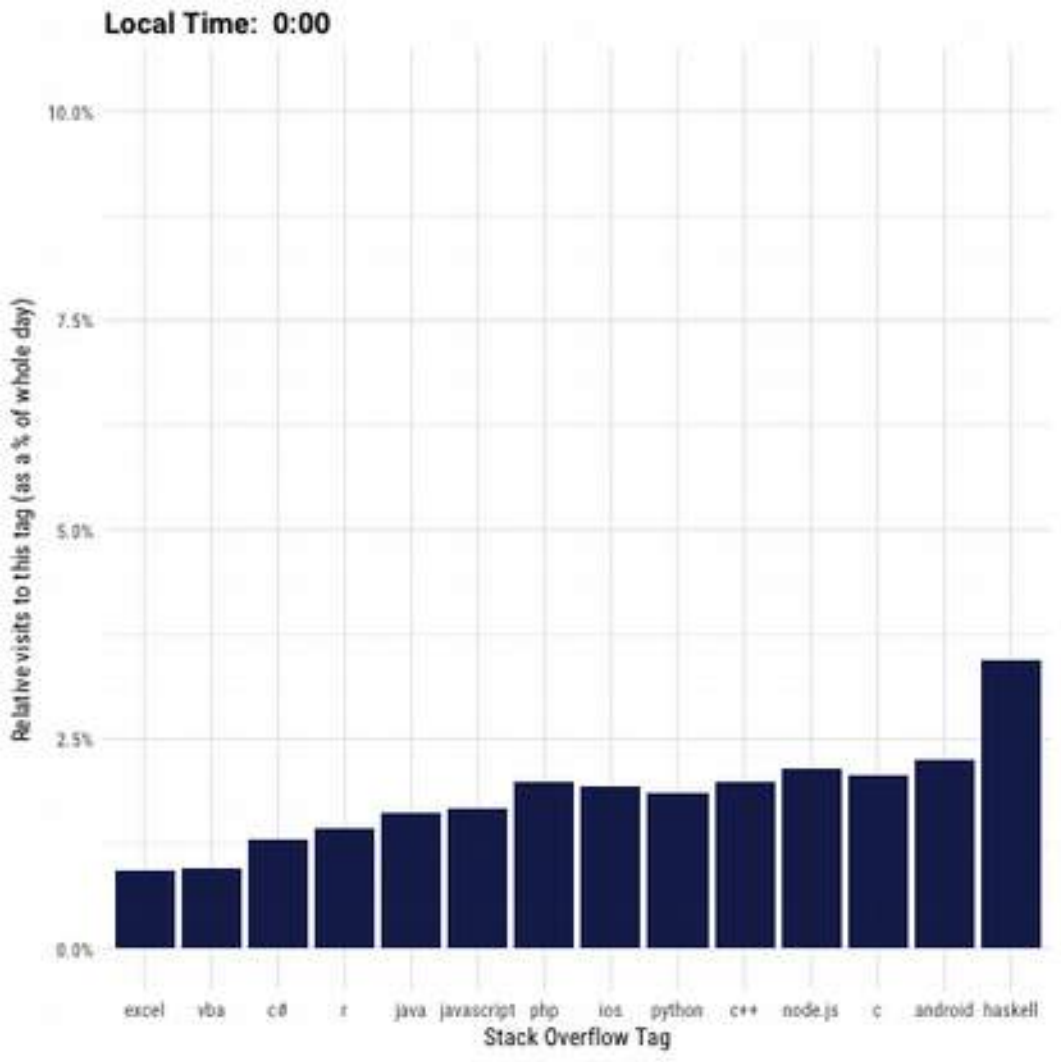
Do you program in the evening? How about late at night?

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.

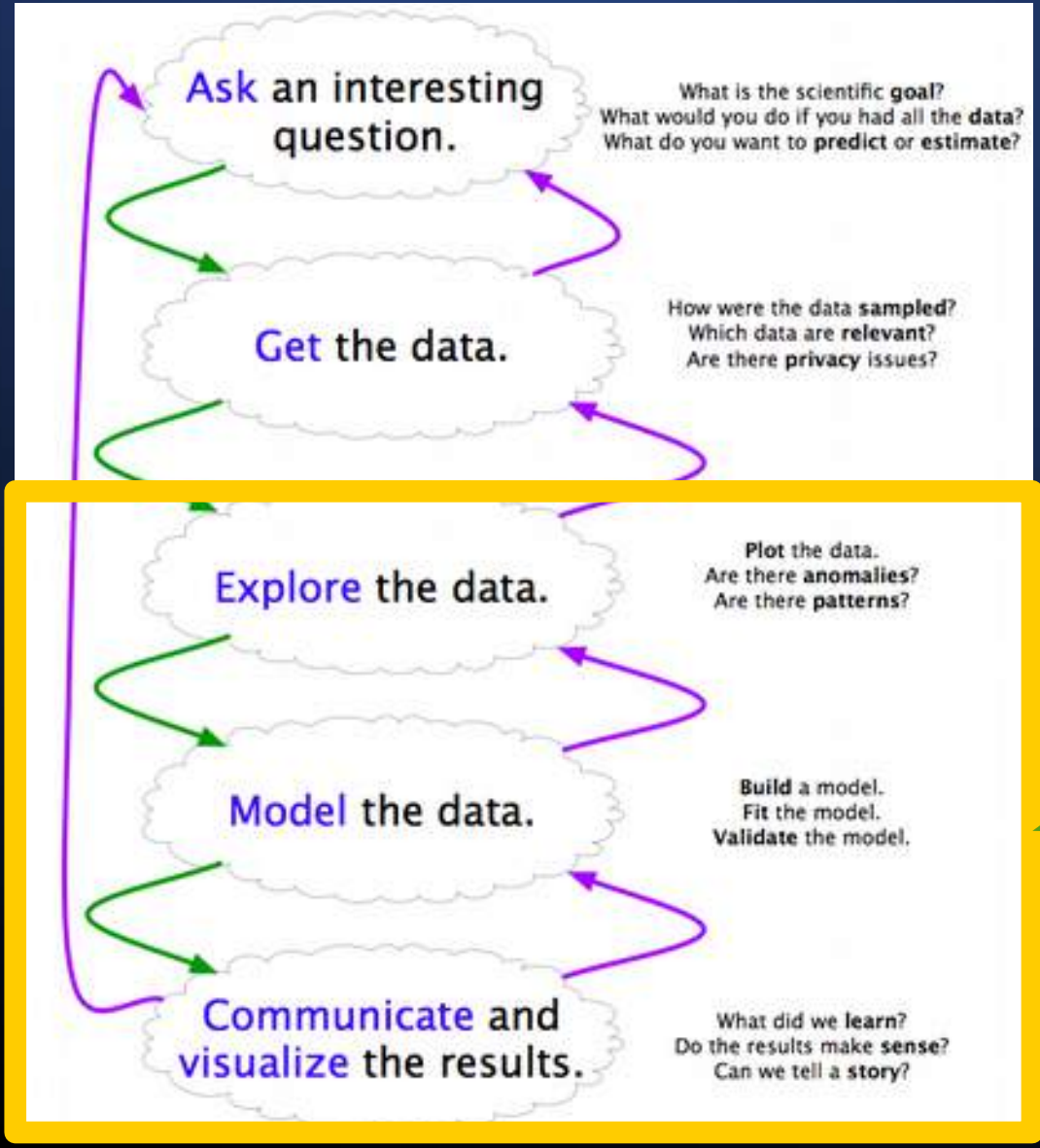
https://stackoverflow.blog/2017/04/19/programming-languages-used-late-night/

Course Project: Topic



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

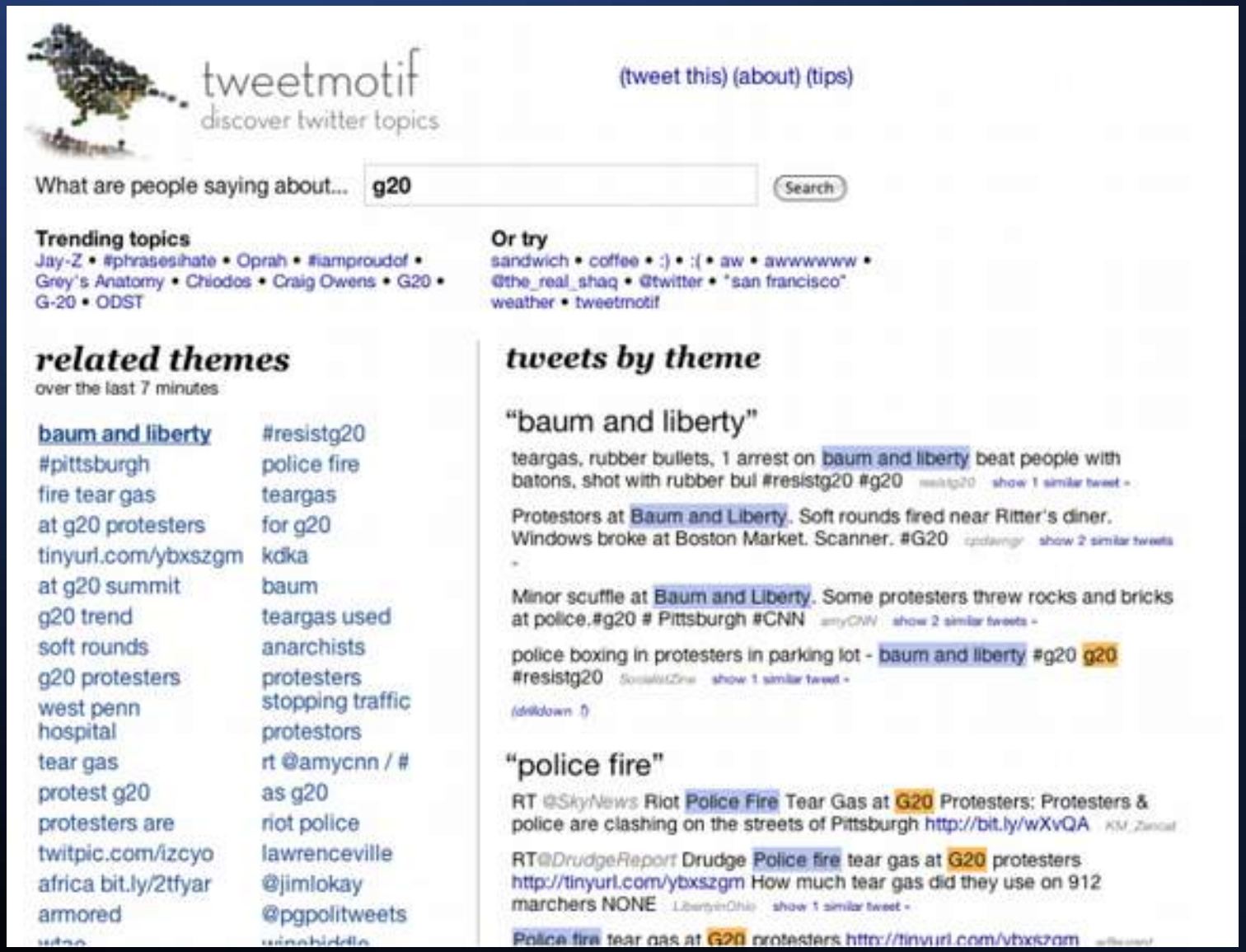
Course Project: Topic



Can we use AI to help us here?

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

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)

Mark Allen Thornton | Home | Research | CV/Publications | Blog | Software | Links

Welcome! This is a book recommender for the Project Gutenberg collection. If you type in the name of a book in the collection **(mostly pre-1930 texts)**, it will suggest other books that have similar content or style.

If you already know one of the recommended books, you can rate how similar you think that book actually is to the one you were asking about. I'll use these ratings to help improve the recommendation system in the future.

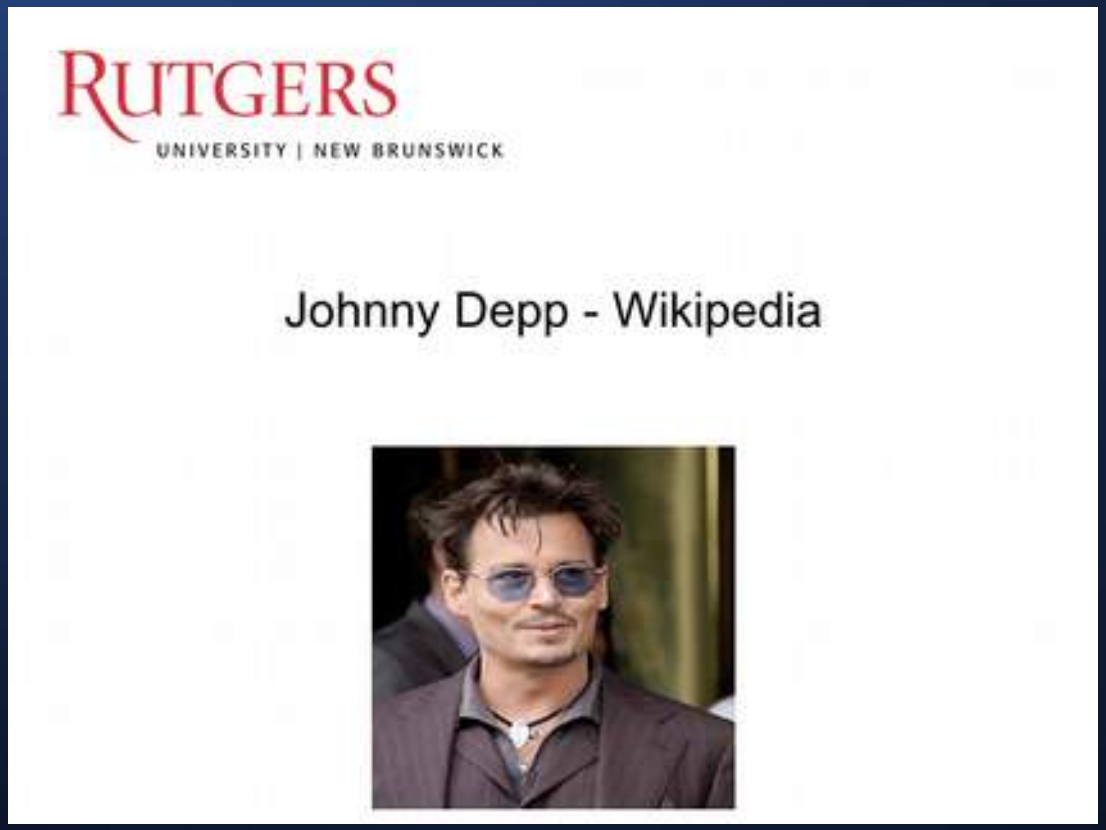
Find books! | Kafka trial

Showing best matches for The Trial by Franz Kafka:

Content matches	Style matches
The Trial by Franz Kafka	The Trial by Franz Kafka
Burnham Breaker by Homer Greene	Shallow Soil by Knut Hamsun
The Quilt that Jack Built; How He Won the Bicycle by Annie Fellows Johnston	Sister Carrie by Theodore Dreiser
Flip's "Islands of Providence" by Annie Fellows Johnston	A Rogue by Compulsion by Victor Bridges

http://markallenthornton.com/recommender/

Course Project: Topic Examples



Wikipedia Article → PPT

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

Course Project: Topic Examples

RUTGERS

Early life

- Depp is of mostly English ancestry, with some ancestors from elsewhere in Europe.
- Depp is a 20th cousin of Elizabeth II.
- Depp moved frequently during his childhood.

Related Topics Available. Get in touch with me.

Wikipedia Article → PPT