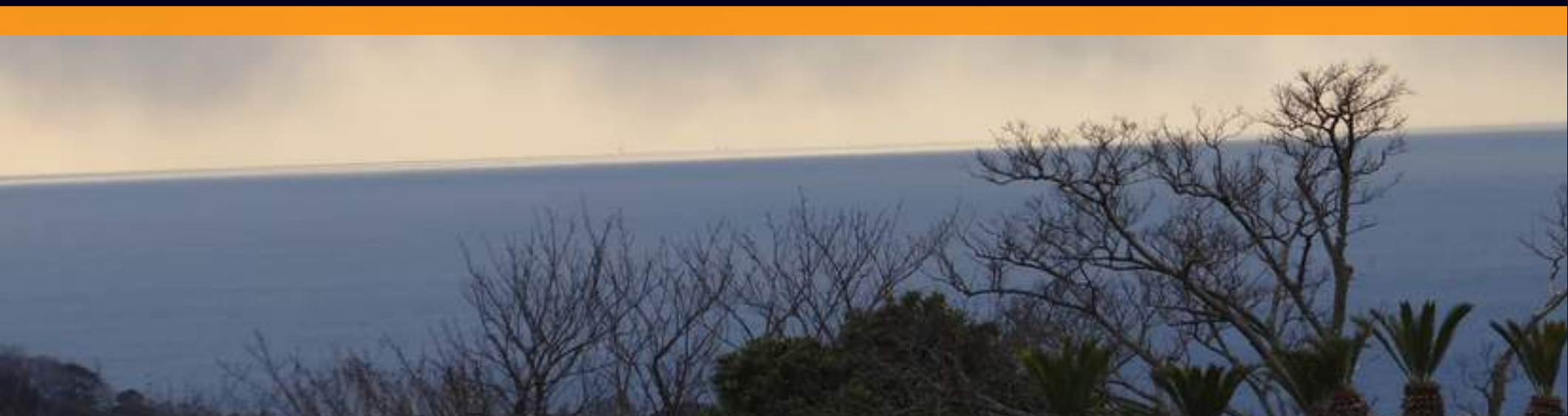


Massive Data Storage and Retrieval: Course Project

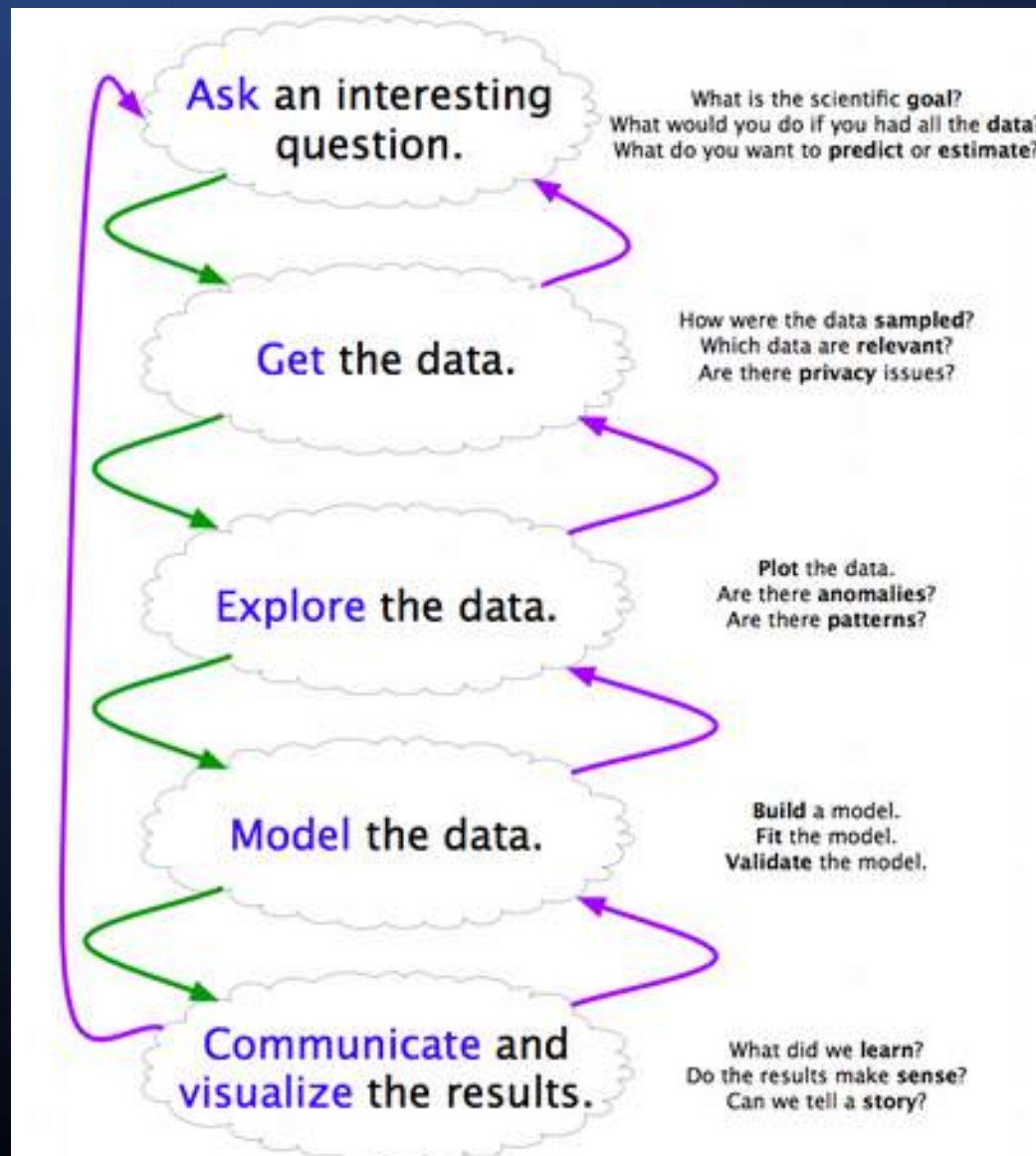
Gerard de Melo

<http://gerard.demelo.org>

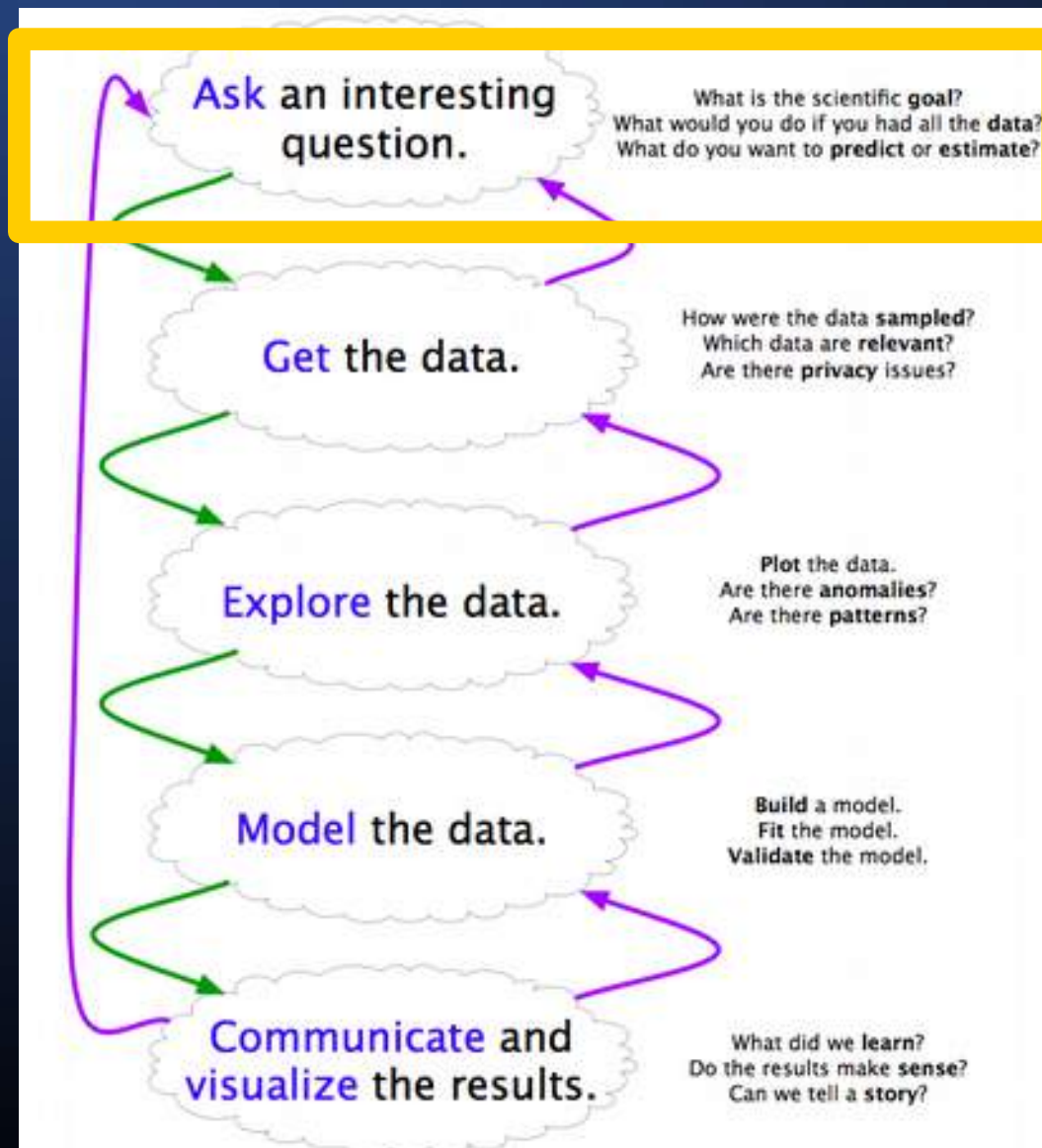
Rutgers University



Course Project



Course Project



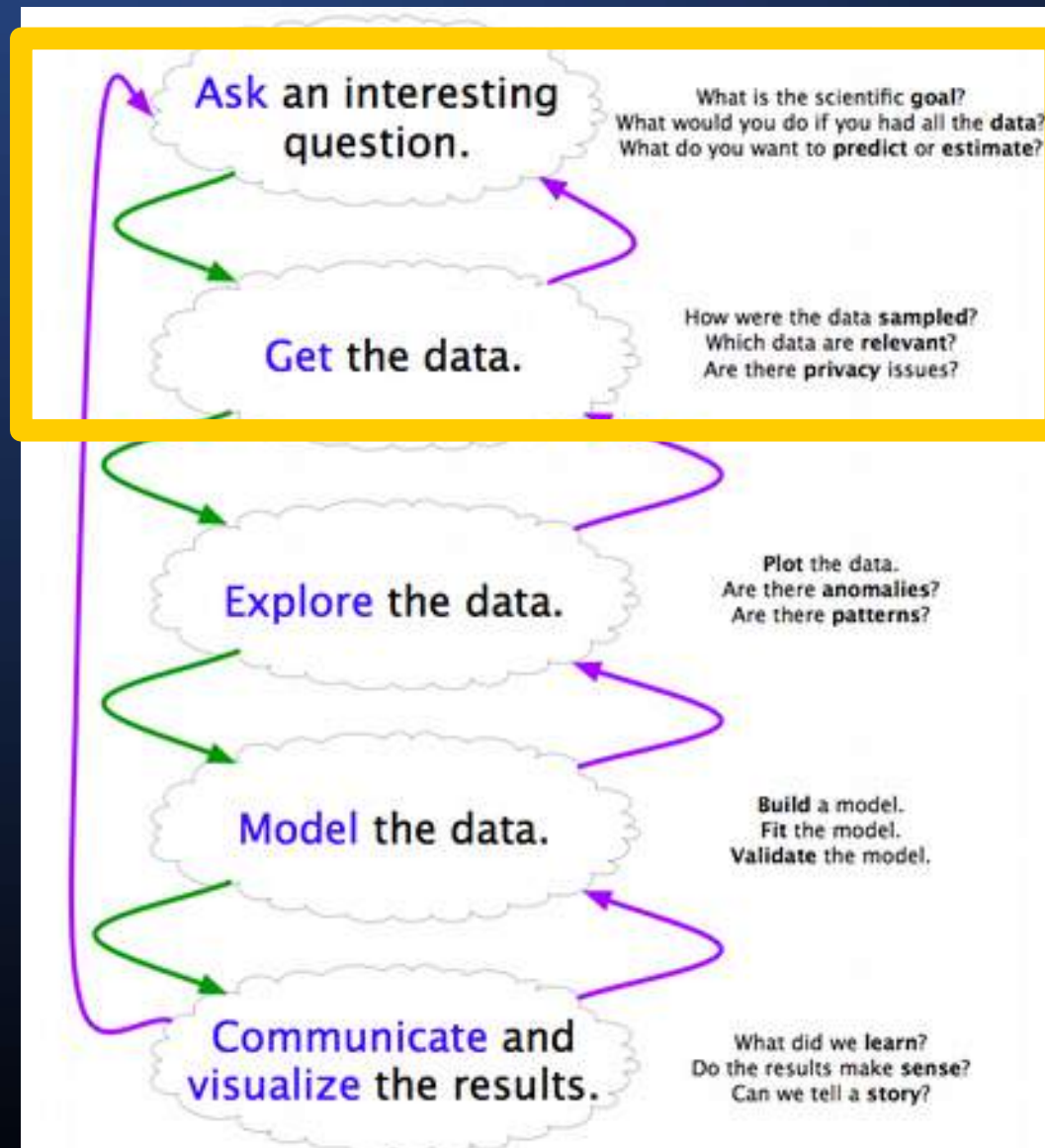
Course Project: Topic



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

Music?
Travel?
Art?
Gaming?
etc.

Course Project



Course Project: Data

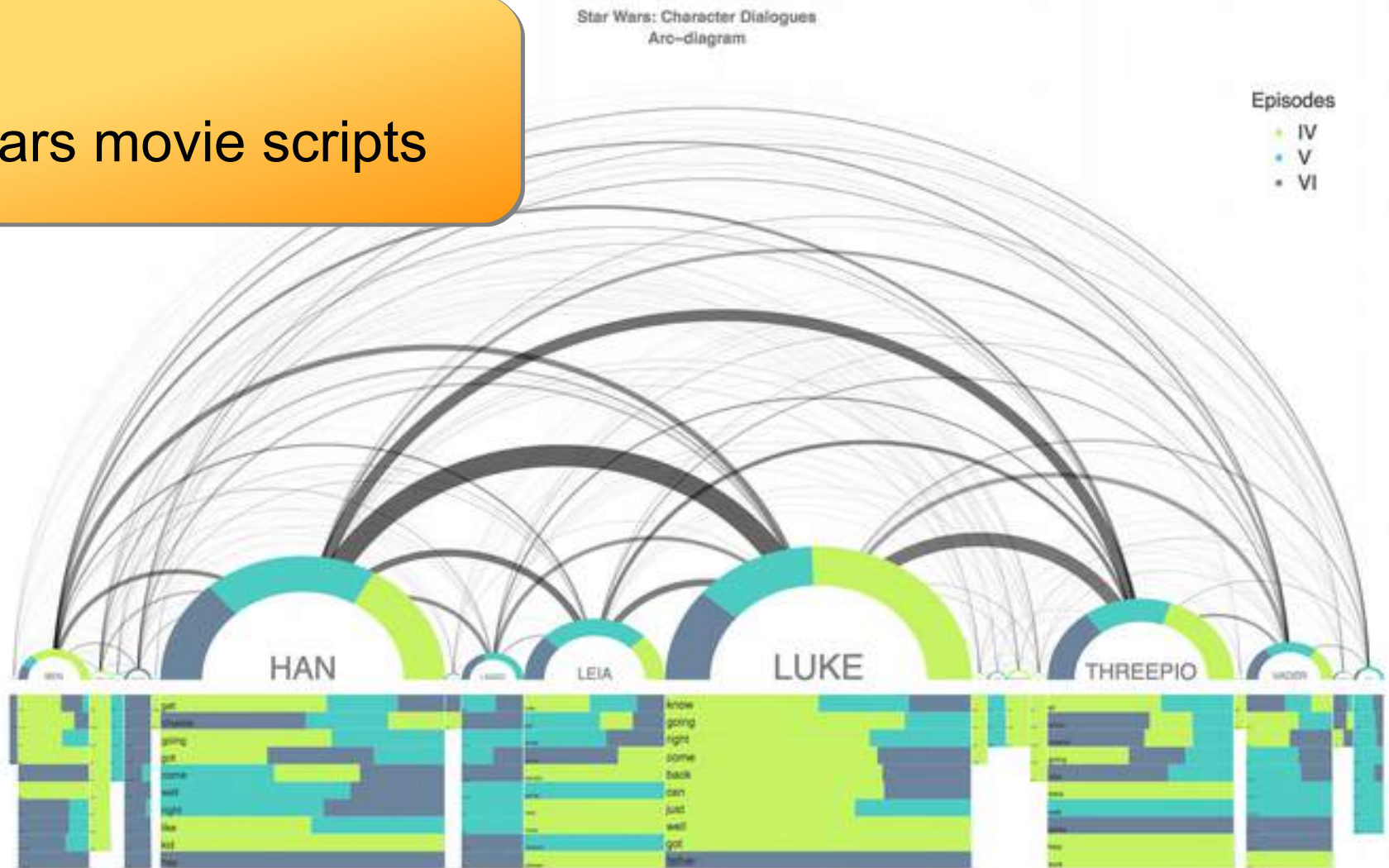


Easiest: Use Web Search Engine to look for “<topic> dataset”
e.g. “yoga pose dataset”

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

Course Project: Topic Examples (just for Inspiration)

Input:
Star Wars movie scripts



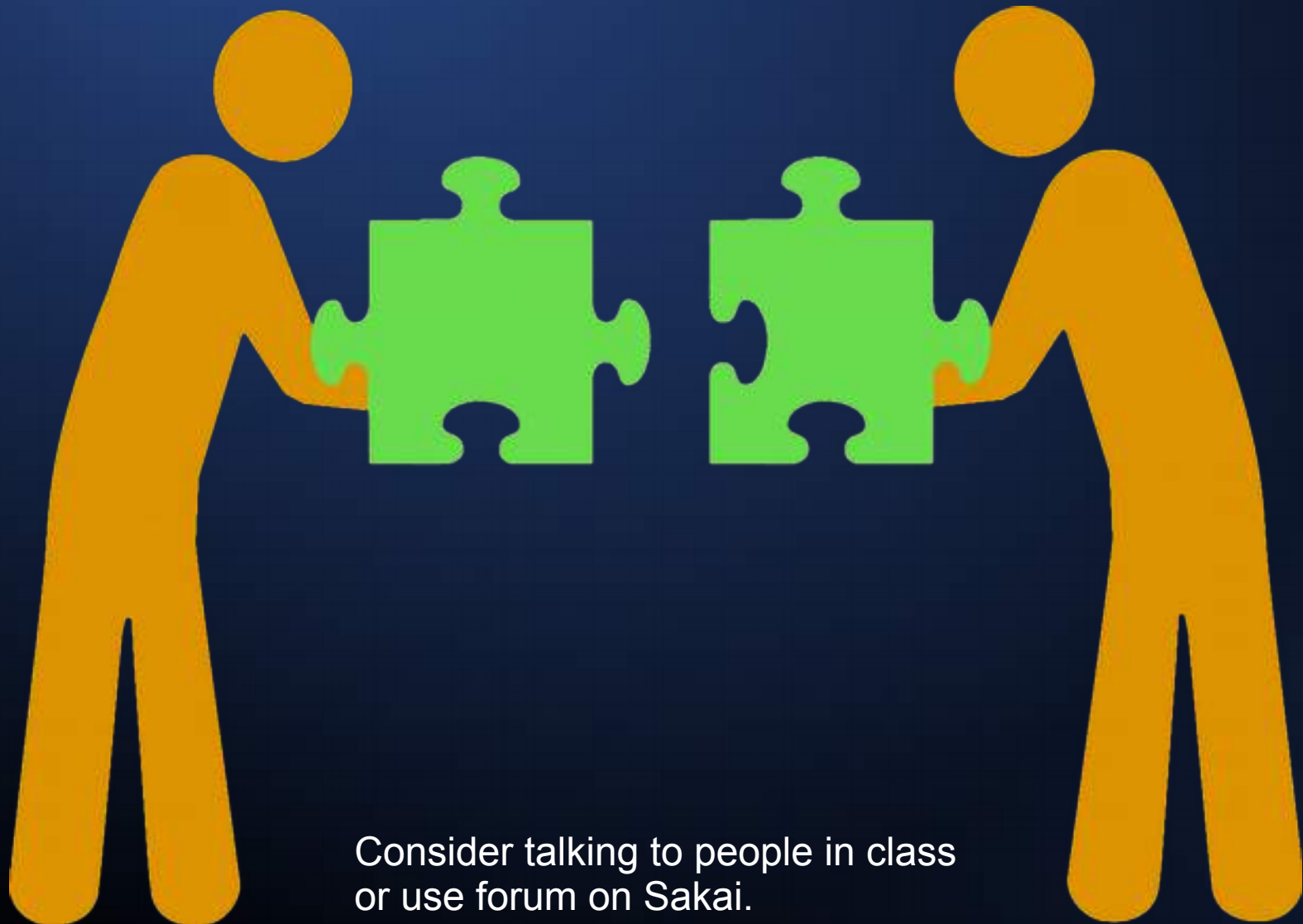
Course Project

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?

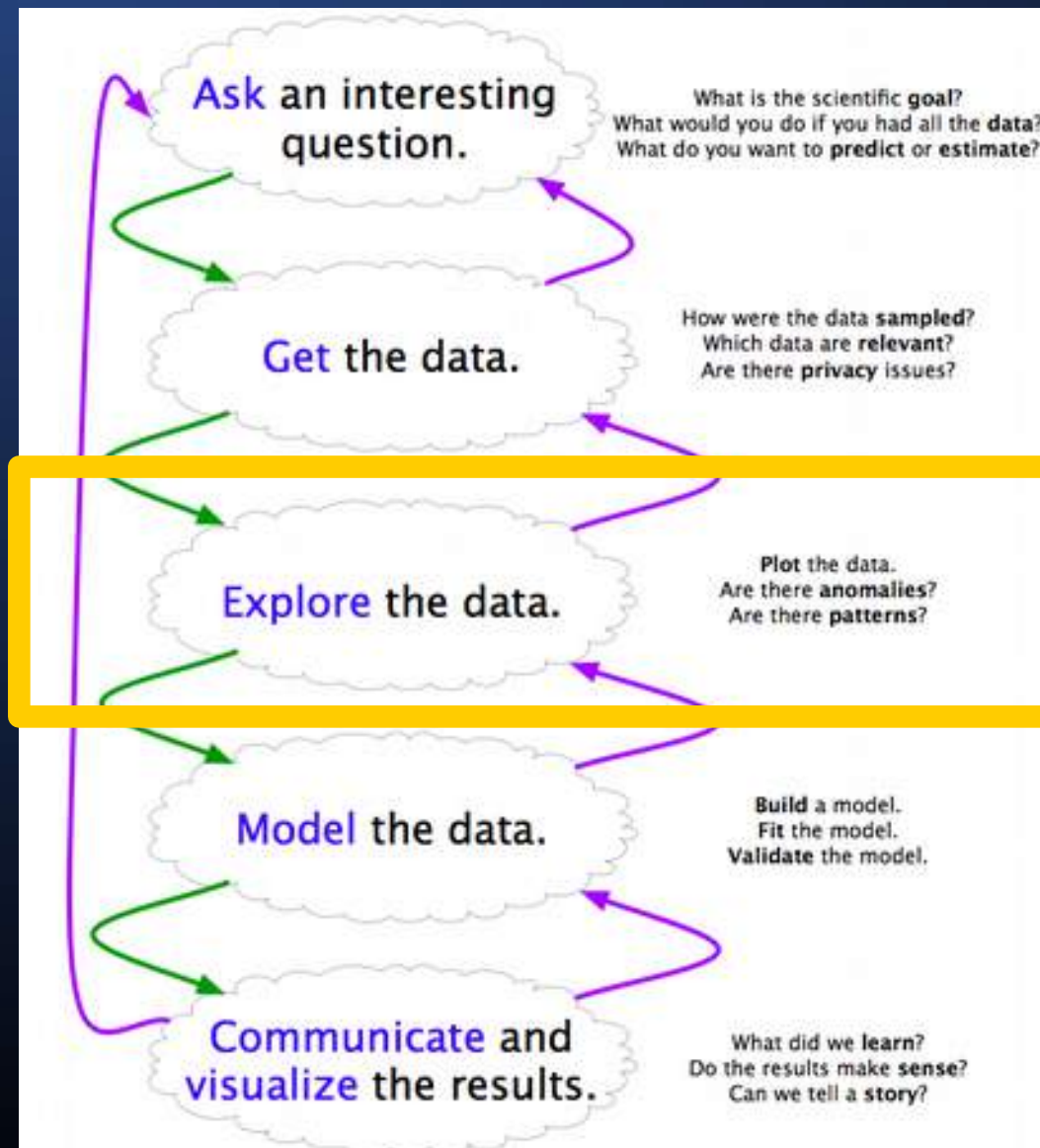
Teams

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

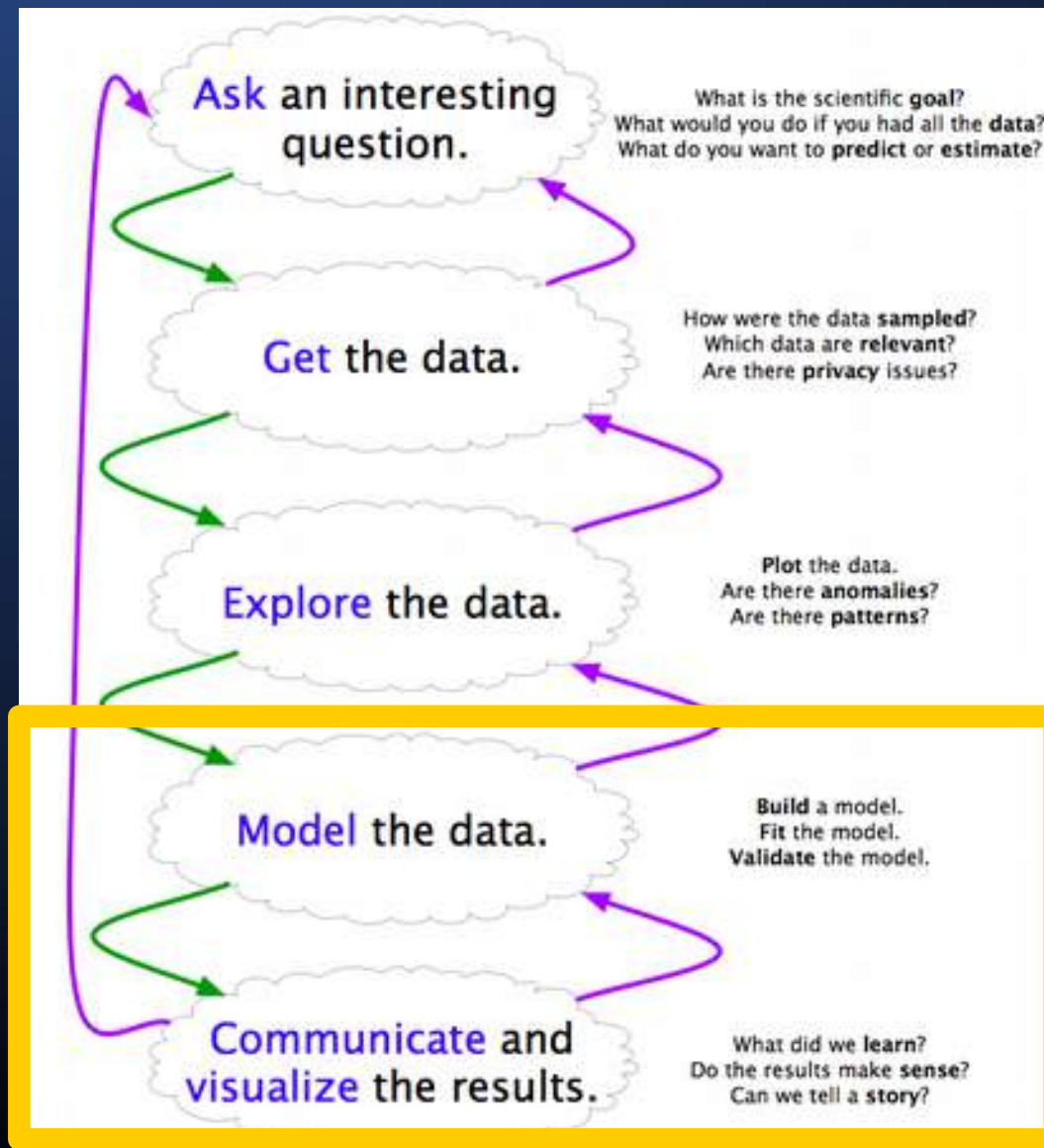


Course Project

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



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

I will find you...

And I will report you!



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

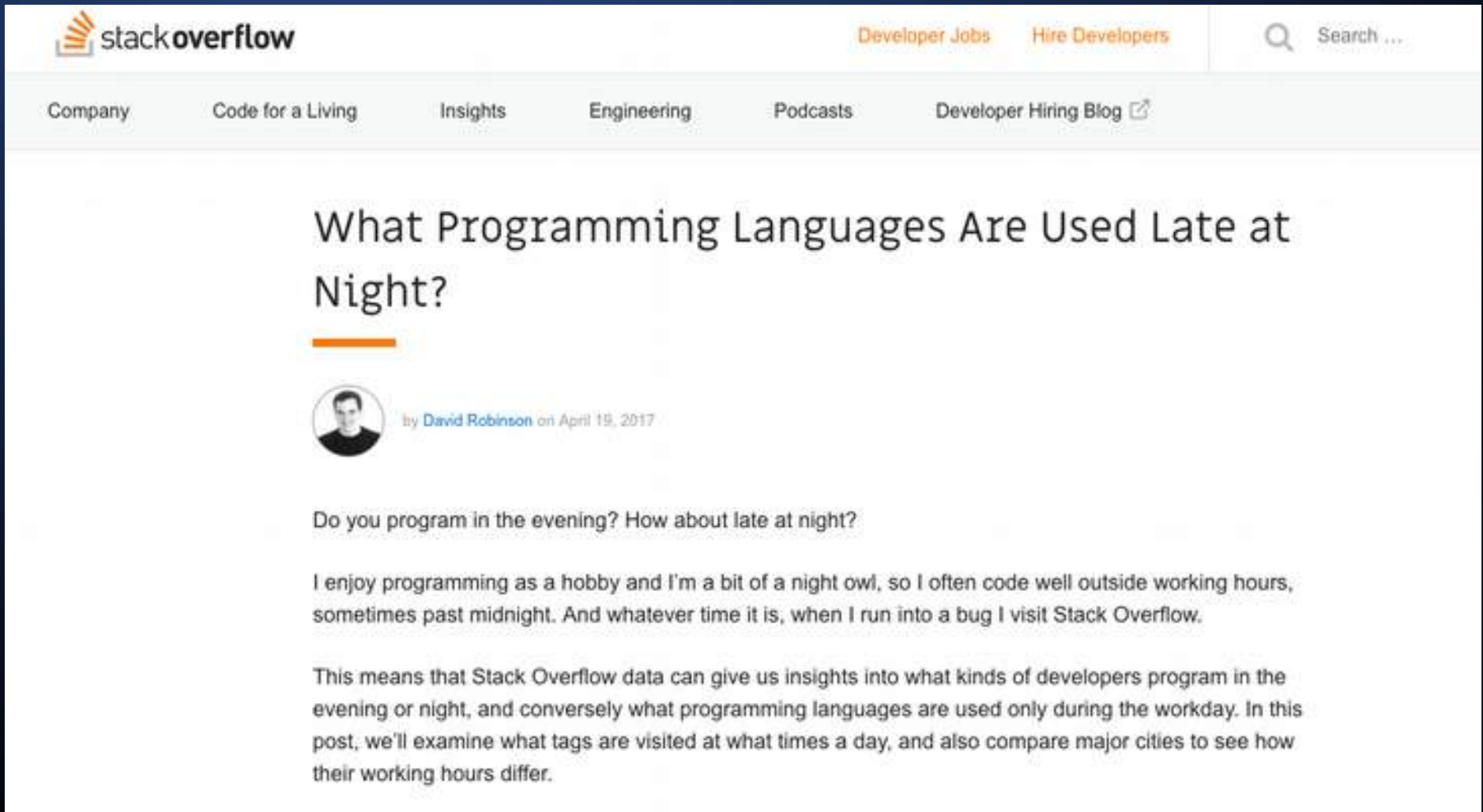
Key Elements for a Good Grade

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

Beautiful visualizations are nice but should not 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



The screenshot shows the Stack Overflow website header with the logo, navigation links like 'Developer Jobs' and 'Hire Developers', and a search bar. Below the header is a secondary navigation bar with links for 'Company', 'Code for a Living', 'Insights', 'Engineering', 'Podcasts', and 'Developer Hiring Blog'. The main content area features a blog post titled 'What Programming Languages Are Used Late at Night?' by David Robinson, dated April 19, 2017. The post begins with the question 'Do you program in the evening? How about late at night?' and discusses the author's interest in programming as a hobby and their use of Stack Overflow for debugging. It also mentions that the post will analyze Stack Overflow data to see what programming languages are used late at night and compare working hours across different cities.

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

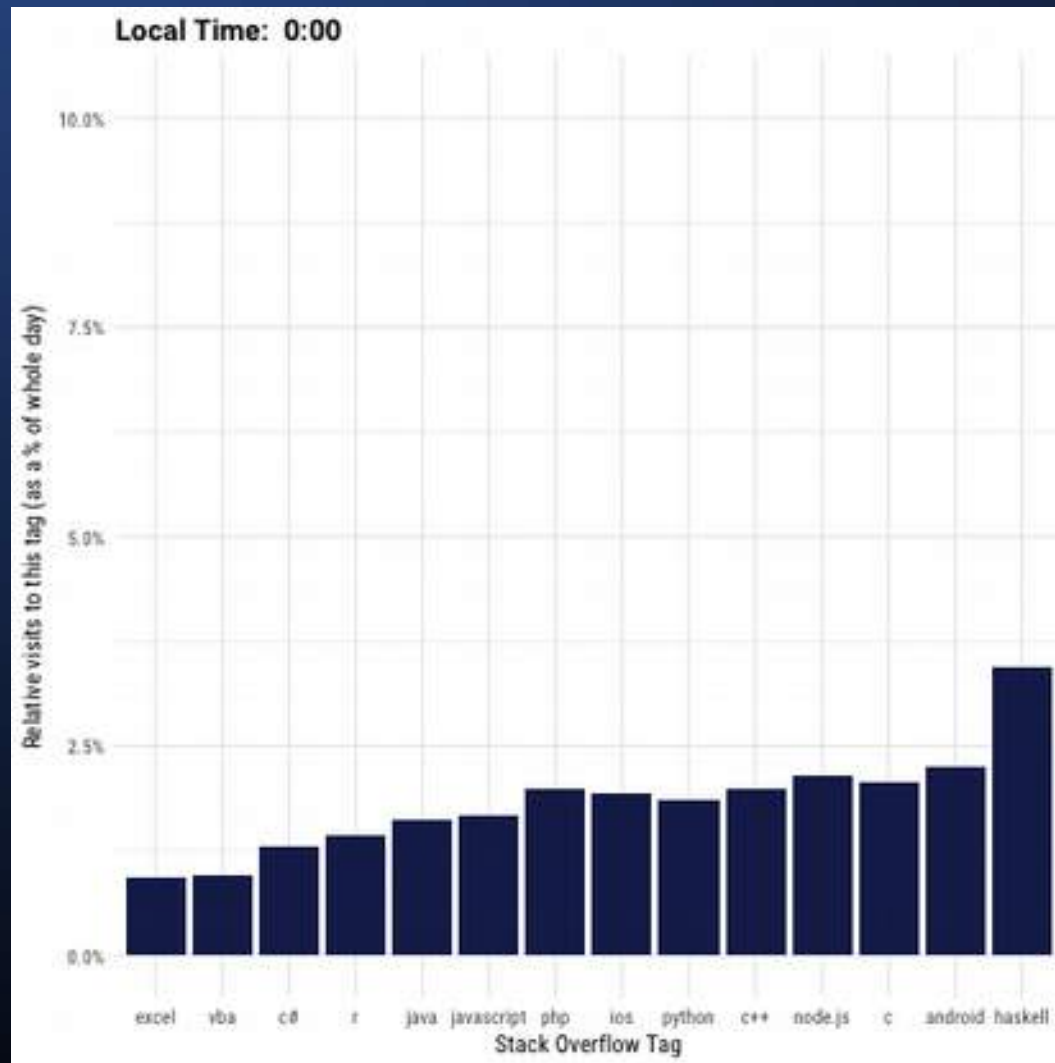
Do you program in the evening? How about late at night?

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.

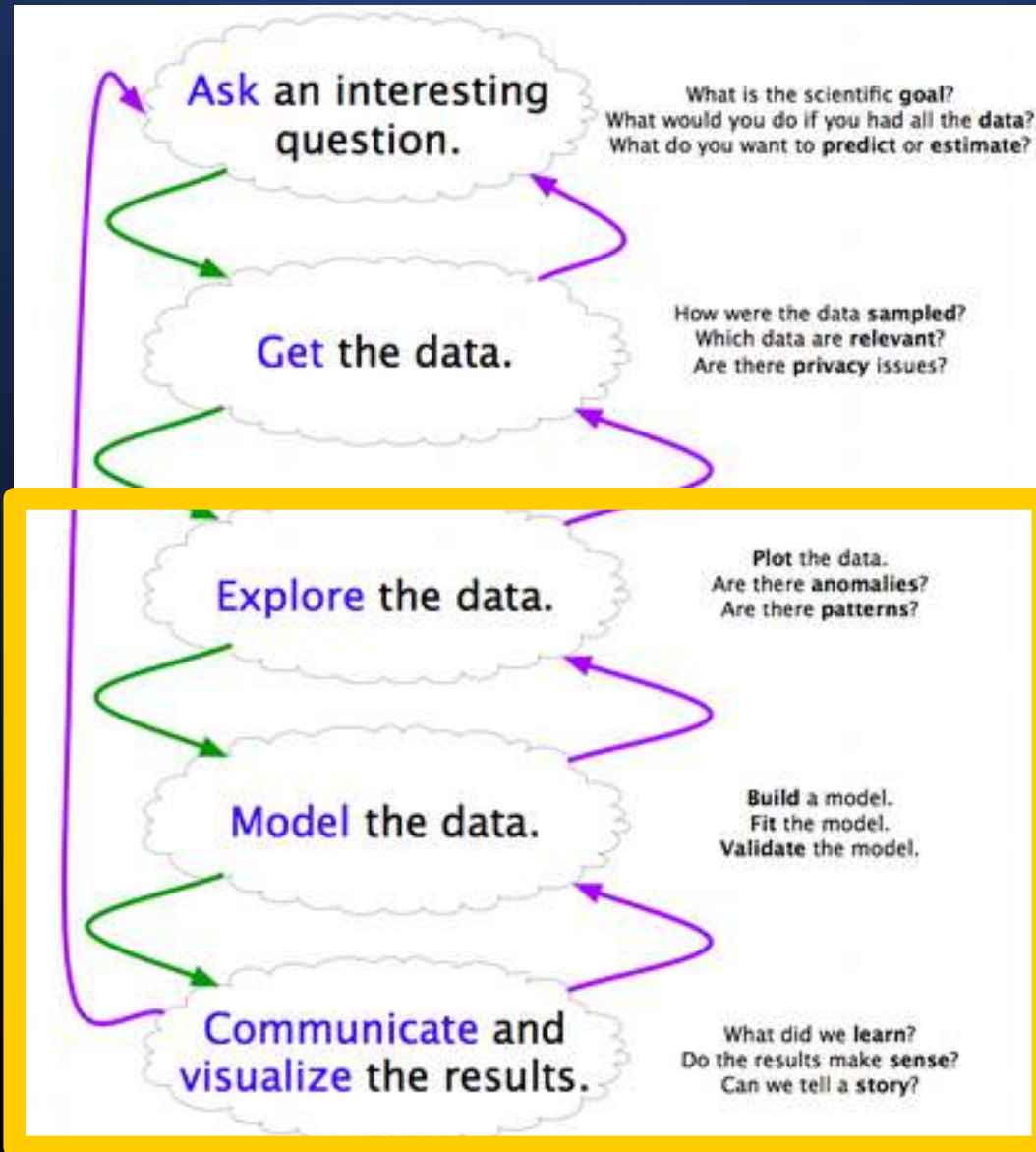
This means that Stack Overflow data can give us insights into what kinds of developers program in the evening or night, and conversely what programming languages are used only during the workday. In this post, we'll examine what tags are visited at what times a day, and also compare major cities to see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

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

Course Project: Topic

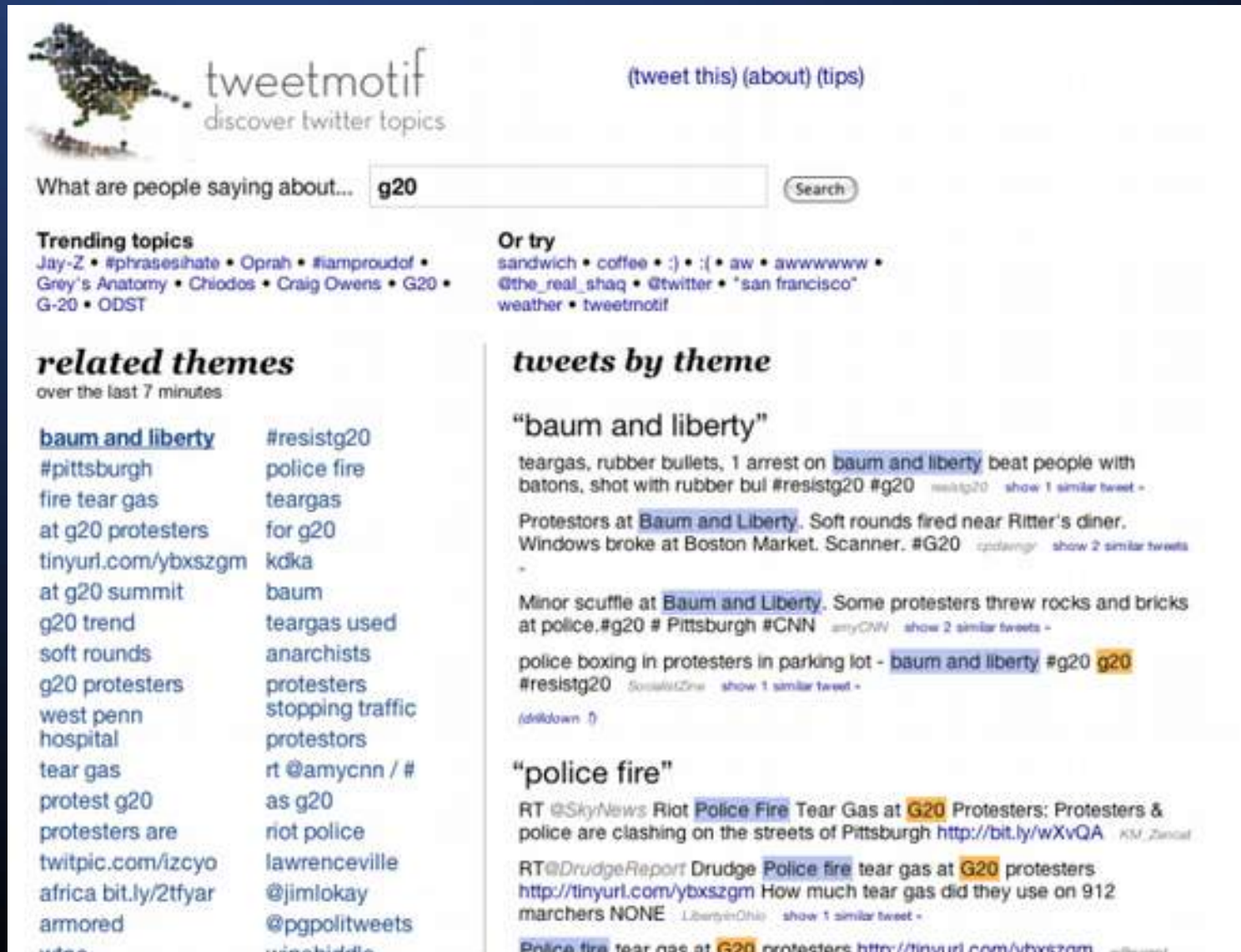


Course Project: Topic



Can we use AI to help us here?

Course Project: Topic Examples (just for Inspiration)



The screenshot shows the TweetMotif website interface. At the top left is the logo with a bird made of tweets and the text 'tweetmotif discover twitter topics'. To the right are links '(tweet this) (about) (tips)'. Below the logo is a search bar with the text 'What are people saying about...' and the input 'g20'. A 'Search' button is to the right. Below the search bar, there are two columns of trending topics. The left column is titled 'Trending topics' and lists: Jay-Z • #phrasesihate • Oprah • #iamproudof • Grey's Anatomy • Chiodos • Craig Owens • G20 • G-20 • ODST. The right column is titled 'Or try' and lists: sandwich • coffee • :) • :(• aw • awwwww • @the_real_shaq • @twitter • "san francisco" • weather • tweetmotif. Below the trending topics, there are two main sections. The left section is titled 'related themes' with the subtitle 'over the last 7 minutes'. It contains two columns of related terms. The first column lists: baum and liberty, #pittsburgh, fire tear gas, at g20 protesters, tinyurl.com/ybxszgm, at g20 summit, g20 trend, soft rounds, g20 protesters, west penn hospital, tear gas, protest g20, protesters are, twitpic.com/izcyo, africa bit.ly/2tfyar, armored. The second column lists: #resistg20, police fire, teargas, for g20, kdka, baum, teargas used, anarchists, protesters, stopping traffic, protestors, rt @amycnn / #, as g20, riot police, lawrenceville, @jimlokay, @pgpolitweets. The right section is titled 'tweets by theme' and contains two sub-sections. The first is titled '"baum and liberty"' and contains three tweets. The first tweet is: 'teargas, rubber bullets, 1 arrest on baum and liberty beat people with batons, shot with rubber bul #resistg20 #g20' with links 'resistg20' and 'show 1 similar tweet'. The second tweet is: 'Protestors at Baum and Liberty. Soft rounds fired near Ritter's diner. Windows broke at Boston Market. Scanner. #G20' with link 'cpdawnr' and 'show 2 similar tweets'. The third tweet is: 'Minor scuffle at Baum and Liberty. Some protesters threw rocks and bricks at police.#g20 # Pittsburgh #CNN' with link 'amycnn' and 'show 2 similar tweets'. The second sub-section is titled '"police fire"' and contains two tweets. The first tweet is: 'RT @SkyNews Riot Police Fire Tear Gas at G20 Protesters: Protesters & police are clashing on the streets of Pittsburgh http://bit.ly/wXvQA' with link 'KM_Zencal'. The second tweet is: 'RT@DrudgeReport Drudge Police fire tear gas at G20 protesters http://tinyurl.com/ybxszgm How much tear gas did they use on 912 marchers NONE' with link 'LibertyinOhio' and 'show 1 similar tweet'.

tweetmotif
discover twitter topics

(tweet this) (about) (tips)

What are people saying about...

Trending topics
Jay-Z • #phrasesihate • Oprah • #iamproudof • Grey's Anatomy • Chiodos • Craig Owens • G20 • G-20 • ODST

Or try
sandwich • coffee • :) • :(• aw • awwwww • @the_real_shaq • @twitter • "san francisco" • weather • tweetmotif

related themes
over the last 7 minutes

baum and liberty
#pittsburgh
fire tear gas
at g20 protesters
tinyurl.com/ybxszgm
at g20 summit
g20 trend
soft rounds
g20 protesters
west penn hospital
tear gas
protest g20
protesters are
twitpic.com/izcyo
africa bit.ly/2tfyar
armored

#resistg20
police fire
teargas
for g20
kdka
baum
teargas used
anarchists
protesters
stopping traffic
protestors
rt @amycnn / #
as g20
riot police
lawrenceville
@jimlokay
@pgpolitweets

tweets by theme

"baum and liberty"
teargas, rubber bullets, 1 arrest on baum and liberty beat people with batons, shot with rubber bul #resistg20 #g20 [resistg20](#) [show 1 similar tweet](#)
Protestors at Baum and Liberty. Soft rounds fired near Ritter's diner. Windows broke at Boston Market. Scanner. #G20 [cpdawnr](#) [show 2 similar tweets](#)
Minor scuffle at Baum and Liberty. Some protesters threw rocks and bricks at police.#g20 # Pittsburgh #CNN [amycnn](#) [show 2 similar tweets](#)
police boxing in protesters in parking lot - baum and liberty #g20 [g20](#) [#resistg20](#) [SocialistZone](#) [show 1 similar tweet](#)
[\(drilldown\)](#)

"police fire"
RT @SkyNews Riot Police Fire Tear Gas at G20 Protesters: Protesters & police are clashing on the streets of Pittsburgh <http://bit.ly/wXvQA> [KM_Zencal](#)
RT@DrudgeReport Drudge Police fire tear gas at G20 protesters <http://tinyurl.com/ybxszgm> How much tear gas did they use on 912 marchers NONE [LibertyinOhio](#) [show 1 similar tweet](#)
Police fire tear gas at G20 protesters <http://tinyurl.com/ybxszgm> [cpdawnr](#)

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, caterpillar, 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: sald, 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

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!

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 ★★★★★

Course Project: Topic Examples



Johnny Depp - Wikipedia



Wikipedia Article
→ **PPT**

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**