data science for (physical) scientists I

I: epistemological concepts and working environment

dr.federica bianco | *fbb.space* |  *fedhere* |  *fedhere*

1 what is data science

2 the scientific method

epistemology

- falsifiability
- probabilistic induction
- reproducibility

3 data science tools

- github
- python
- jupyter notebooks
- google colab
- stackoverflow

this slide deck

http://bit.ly/dsps2019_1

0 some administrative stuff

Syllabus

http://bit.ly/dspssyllabus

Learning Outcomes

By the end of this class you should be able to formulate an appropriate analysis plan for a research question, select, gather, and prepare data for analysis, and choose and apply machine learning methods to the data.

Syllabus

http://bit.ly/dspssyllabus

The instructors is: Dr. **Federica Bianco** fbianco@udel.edu

office hours: Tentatively: Tuesday 3:30-5PM (or by appointment).

The Class assistants are:

Grader: **Yuqi Kong** kongyq@udel.edu

Office Hours: Monday 5PM Smith Hall 220

Technical and coding questions

Physics instructor: Dr. **Alexandre David-Uraz** adu@udel.edu

Office Hours: TBD Sharp 101B physics help center

Physics and Method-Application questions

Syllabus

http://bit.ly/dspssyllabus

- 10% pre-class questions
- 10% class performance and participation
- 20% homeworks
- 25% midterm
- 35% final

quizz

http://bit.ly/dspssyllabus

- 10% pre-class questions
- 10% class performance and participation
- 20% homeworks
- 25% midterm
- 35% final

from beginning of class to 5 minutes past (be on time!)

questions on previous class material and reading assignments

participation

http://bit.ly/dspssyllabus

- 10% pre-class questions
- 10% class performance and participation
- 20% homeworks
- 25% midterm
- 35% final

ask questions

answer questions

get up and code

extra credit assignments

homework

http://bit.ly/dspssyllabus

- 10% pre-class questions
- 10% class performance and participation
- 20% homeworks
- 25% midterm
- 35% final

Homework projects must be turned in as *jupyter notebooks* by checking them into your github account in a DSPS_<firstinitialLastname> repo and the project directories HW<hw number> (unless otherwise stated).

<finitialLastname> is e.g. fBianco

homework

Please work in groups of up to 5 people on homework as a collaborative projects.

Individual notebooks must be returned for each homework. Different group members should lead different aspects of the work. A statement **must be included in the README** explaining each team member's contribution (similar to an acknowledge of contribution you would find in a *Nature* letter see, for example these contributions).

- **10% pre-class questions**
- **10% class performance and participation**
- **20% homeworks**
- **25% midterm**
- **35% final**

homework

Please work in groups of up to 5 people on homework as a collaborative projects.

Individual notebooks must be returned for each homework. Different group members should lead different aspects of the work. A statement **must be included in the README** explaining each team member's contribution (similar to an acknowledge of contribution you would find in a *Nature* letter see, for example these contributions).

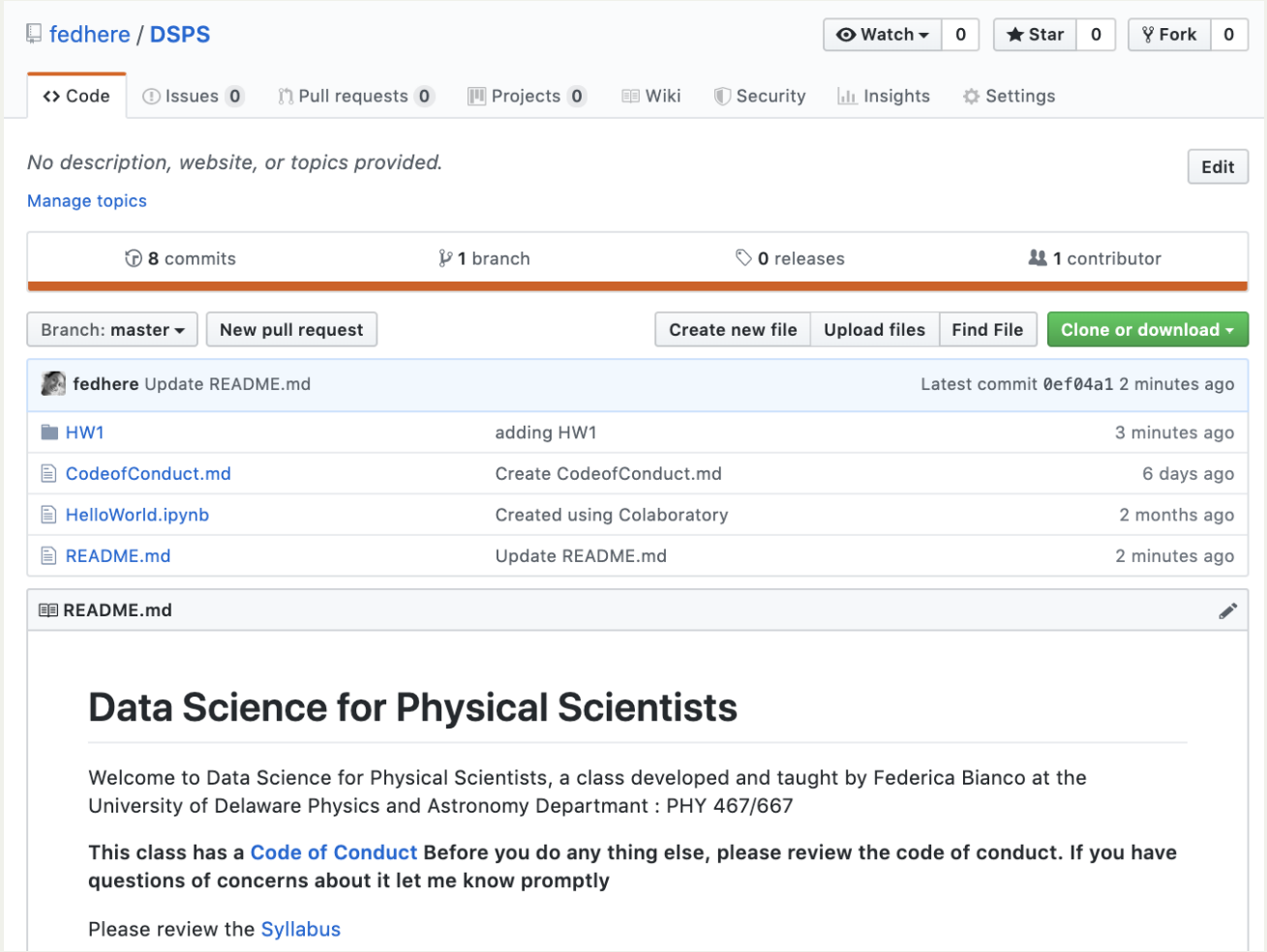
- 10% pre-class questions
- 10% class performance and participation
- 20% homeworks
- 25% midterm
- 35% final

Contributions

K.T. led the project and reduced the ALMA data. K.T. and D.I. wrote the manuscript. M.S.Y. reduced the Large Millimeter Telescope data and edited the final manuscript. Other authors contributed to the interpretation and commented on the ALMA proposal and the paper.

homework

- 10% pre-class questions
- 10% class performance and participation
- 20% homeworks
- 25% midterm
- 35% final

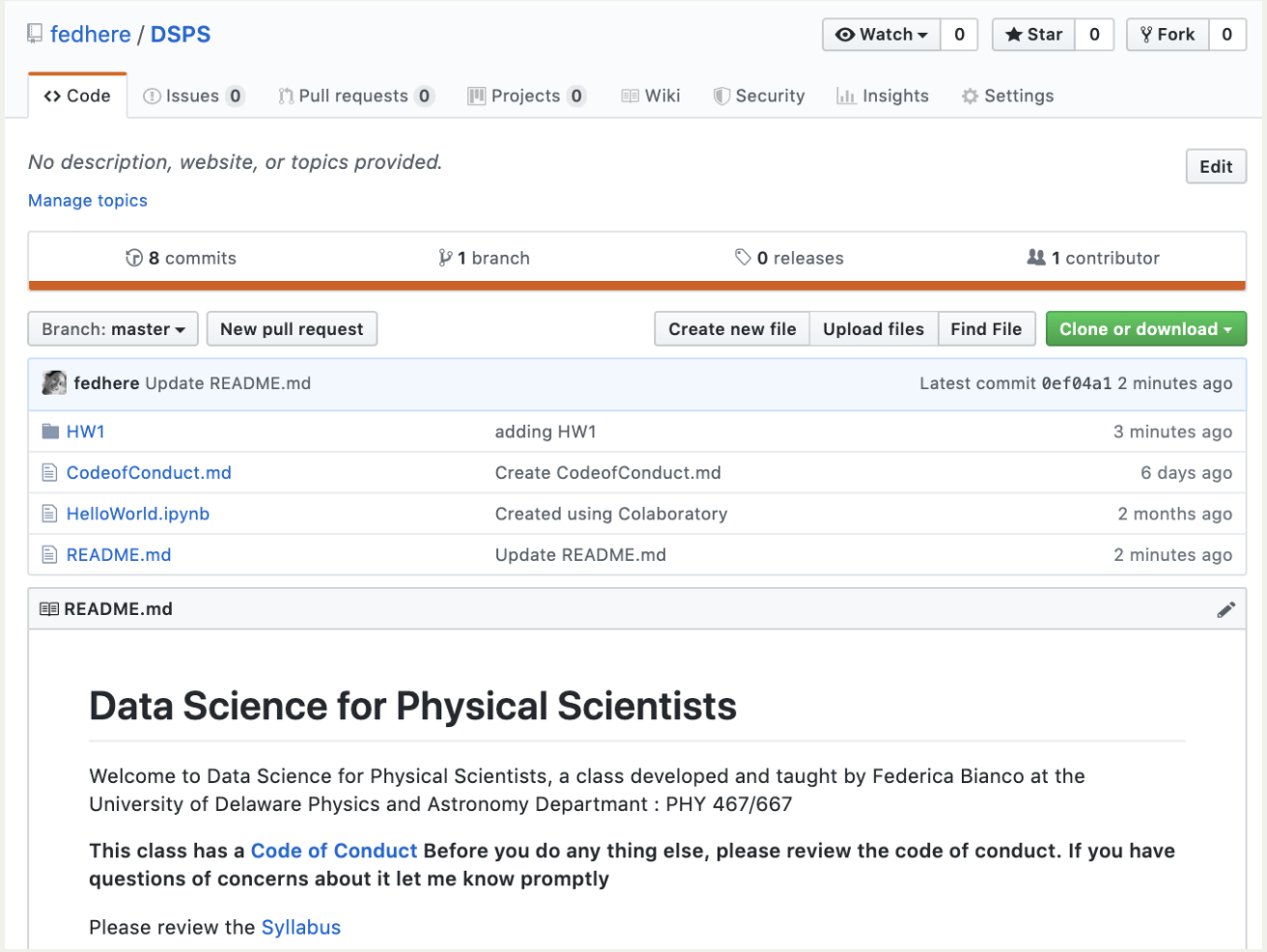


instructions will be here

https://github.com/fedhere/DSPS

homework

- 10% pre-class questions
- 10% class performance and participation
- 20% homeworks
- 25% midterm
- 35% final



instructions will be here

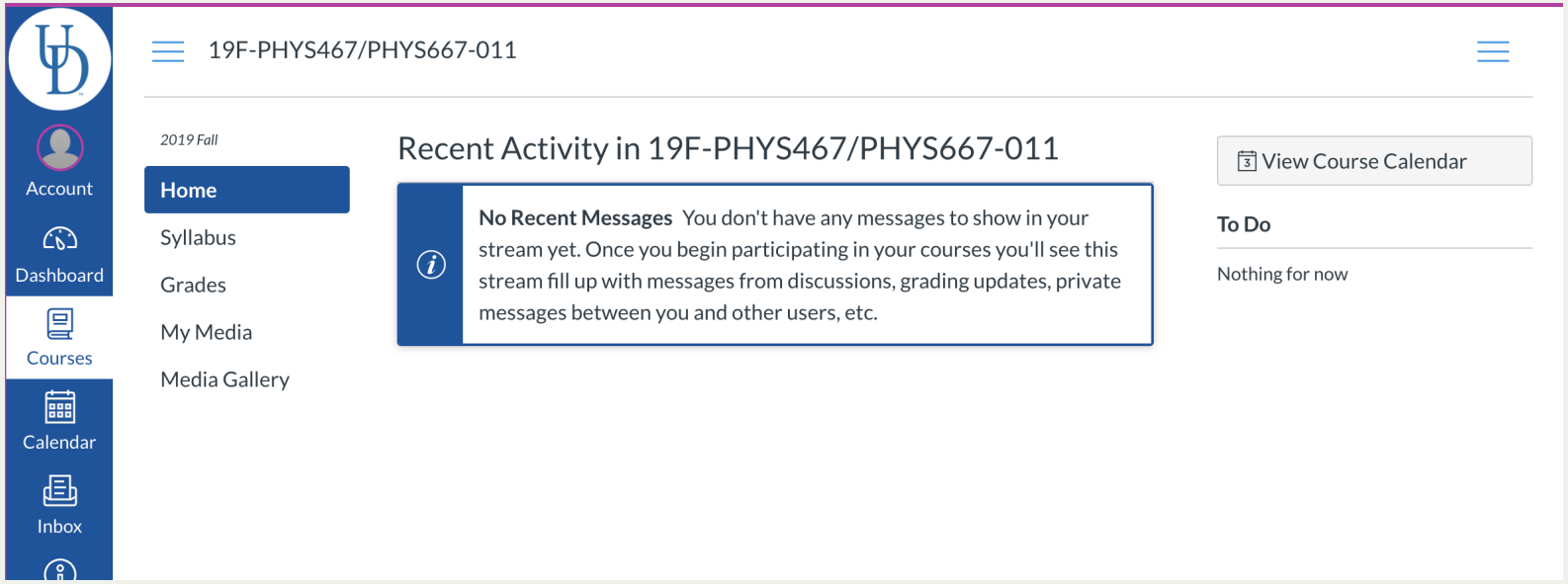
https://github.com/fedhere/DSPS

online help: http://bit.ly/2HtkQoc

please sign up asap!!

slack

homework



- 10% pre-class questions
- 10% class performance and participation
- 20% homeworks
- 25% midterm
- 35% final

of course there is also UD Canvas, which will be used to give you grades and occasionally post messages (I am still learning how to use it tho!)

HW 1 is posted already

midterm

For the *Midterm* and the *Final* you are responsible for material in the labs, the reading, and the homework. **In preparing for the exams, use the homework as a guide to which material is essential. In the Midterm and Final YOU WILL BE EXPECTED TO WORK INDIVIDUALLY.**

- **10% pre-class questions**
- **10% class performance and participation**
- **20% homeworks**
- **25% midterm**
- **35% final**

Midterm... probably in class

issues: stereotype thread - working under derass is not necessarily a required skill

advantages: interviews for jobs are often timed

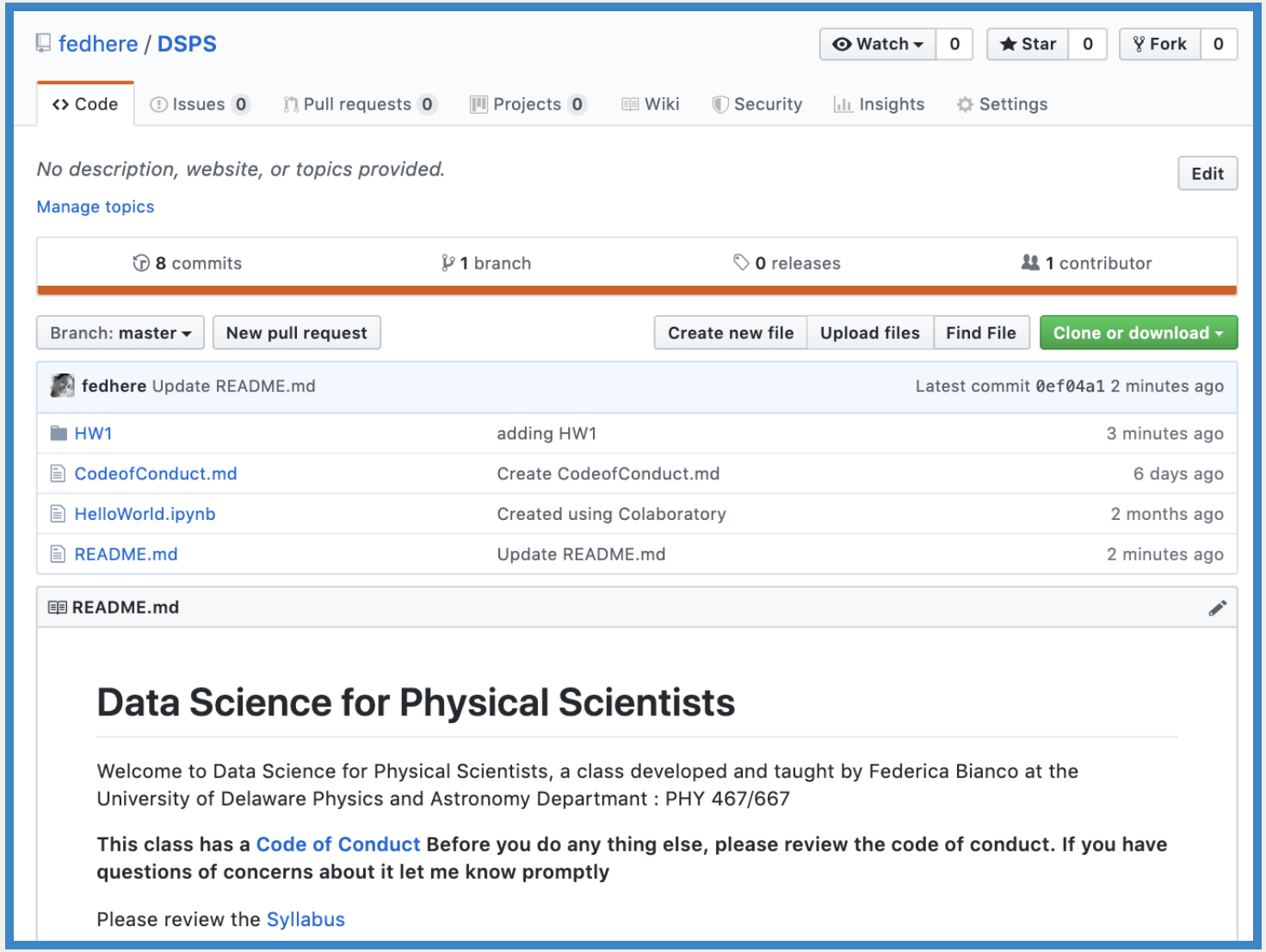
final

For the *Midterm* and the *Final* you are responsible for material in the labs, the reading, and the homework. **In preparing for the exams, use the homework as a guide to which material is essential. In the Midterm and Final YOU WILL BE EXPECTED TO WORK INDIVIDUALLY.**

- 10% pre-class questions
- 10% class performance and participation
- 20% homeworks
- 25% midterm
- 35% final

Final: take home, multiple days.

Resources



https://github.com/fedhere/DSPS

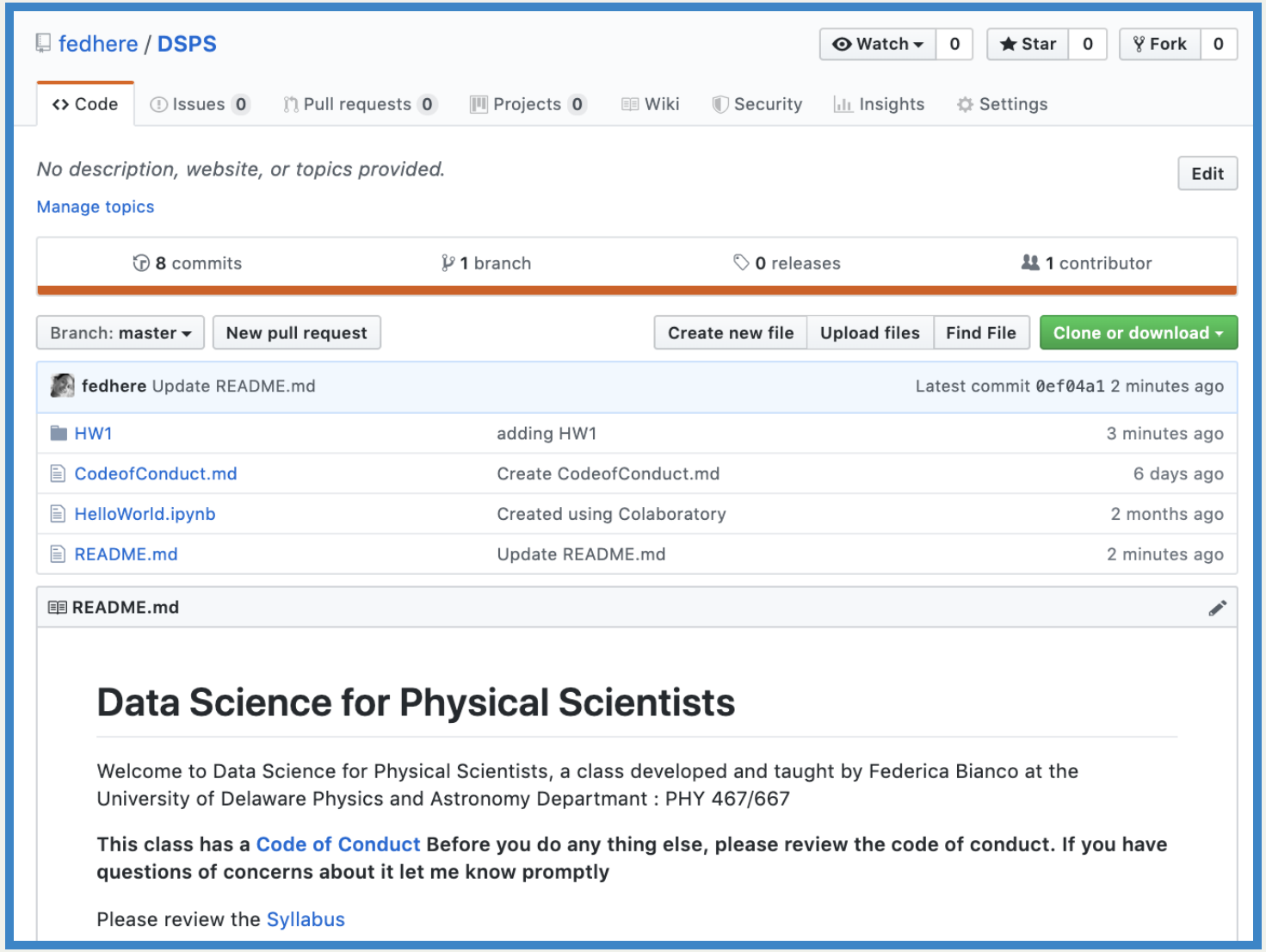
Resources

https://github.com/fedhere/DSPS

- SLIDES here
- HOMEWORK INSTRUCTIONS here
- RESOURCES here

If notebooks do not display use

https://nbviewer.jupyter.org



Resources

Resources

The primary textbooks are:

- **Elements of Statistical Learning,** Hastie,Tibshirani,Friedman, Springer 2001
- **Python Data Science Handbook,** Jake VanderPlas, O'Reilly Media [https://www.oreilly.com/library/view/python-data-science/9781491912126/]
- **Statistics, Data Mining, and Machine Learning in Astronomy,** Ivezic, Connoly, VanderPlas, Gray, Princeton Press

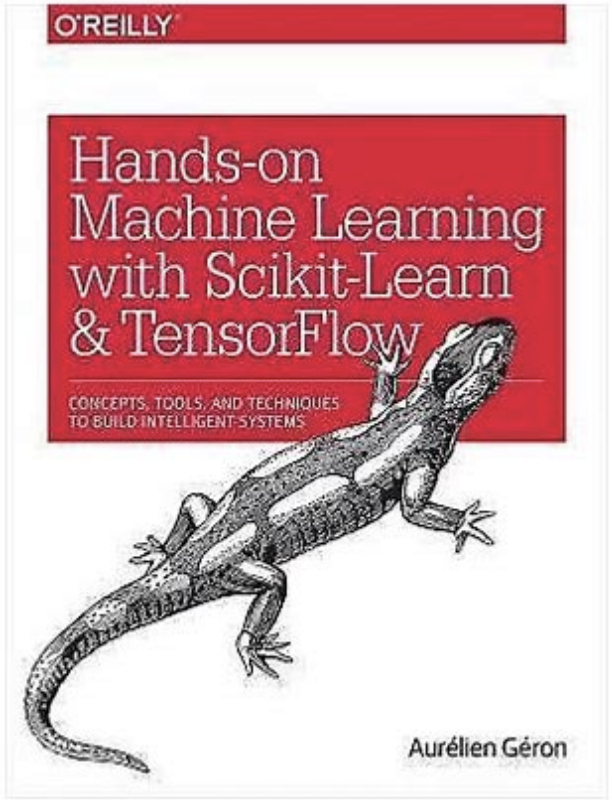
In addition, depending on your familiarity with coding, statistics, and visualization

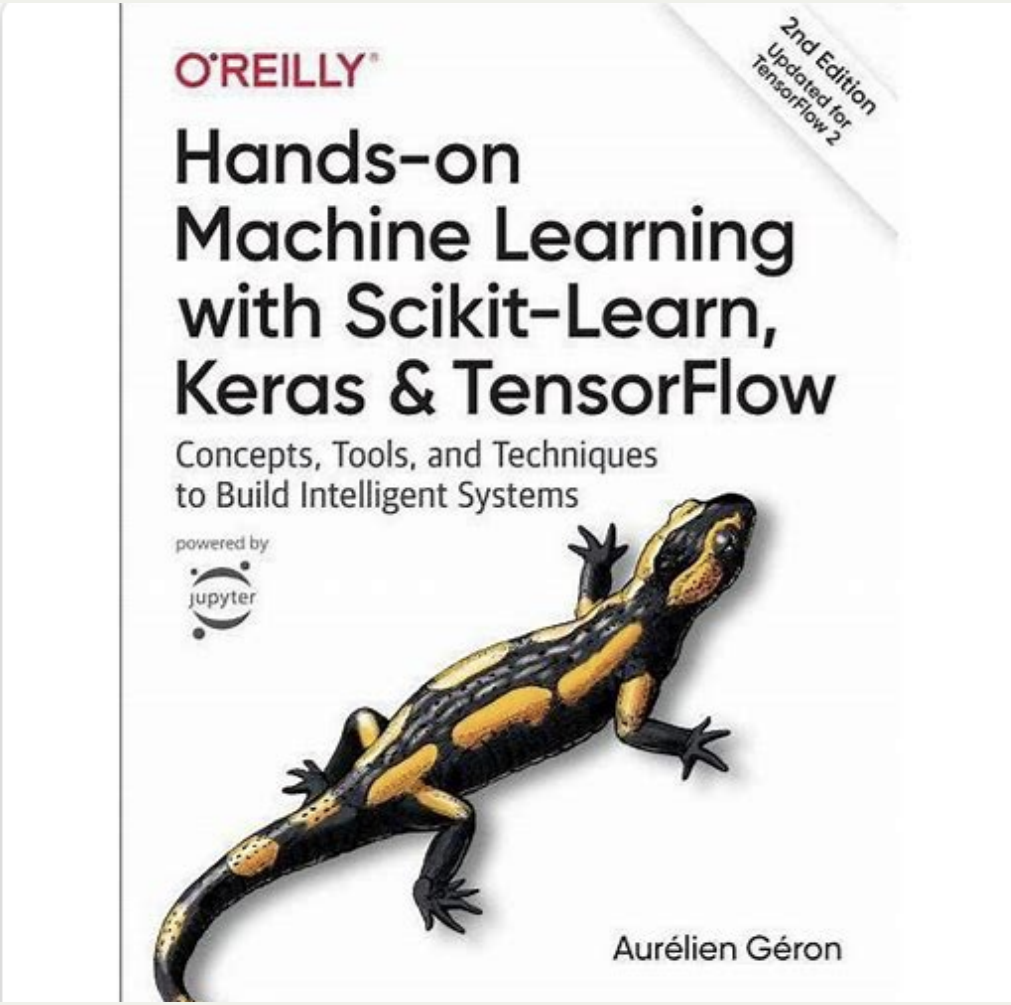
- **ML in python: Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow** probably the book that is closer to the syllabus in terms of techniques, but don't buy it, because the second edition is due to come out imminently and the deep learning chapters of the previous edition are out of date now
- computing and coding: **Beginning Python Visualization, 2009**
- data analysis: **Statistics in a nutshell, S. Boslaugh, O'Reilly Media**
- **Interactive Data Visualization, S. Murray, O'Reilly Media**
- Visualizations: **Visualizations Analysis and Design, T. Munzer, 2014**

Resources

https://github.com/ageron/handson-ml

dont buy this! the second edition will come out soon and TensorFlow has changed significantly!





Resources

Jake Vanderplas is a physicst-data scientists

http://vanderplas.com/

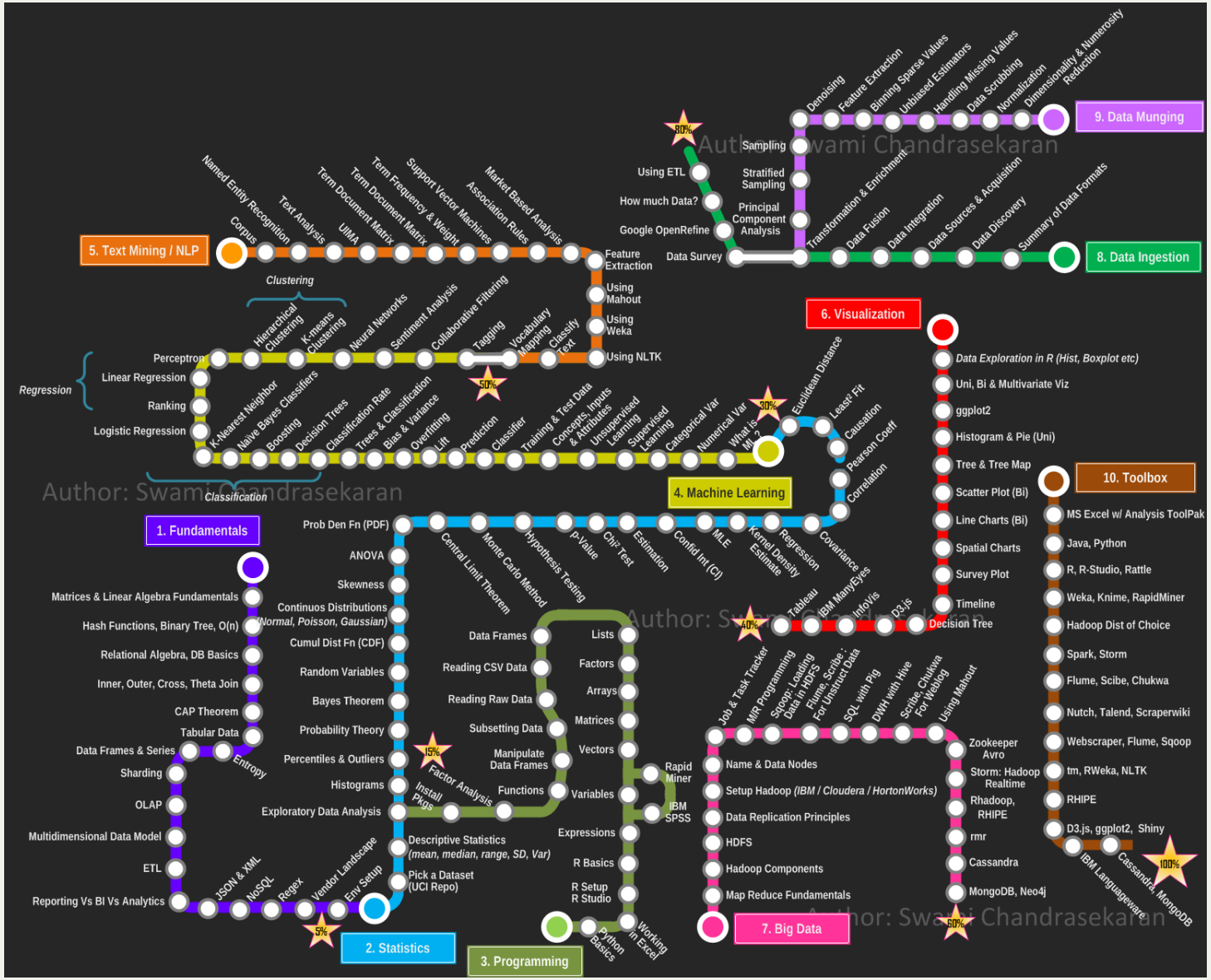
http://vanderplas.com/

Title Text

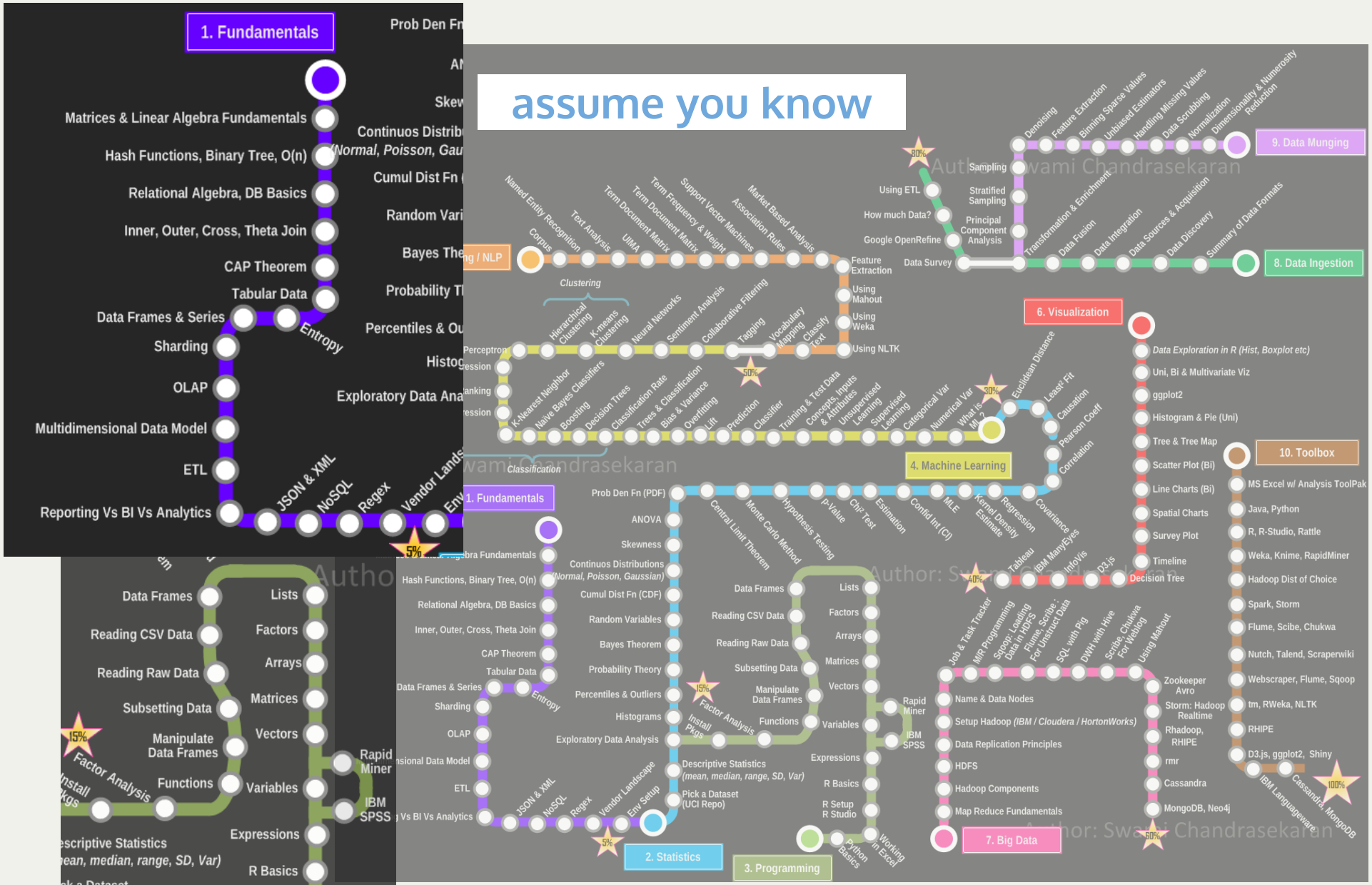
Subtitle Text

https://www.youtube.com/embed/ZyjCqQEUa8o?start=310&enablejsapi=1

1 what is data science?

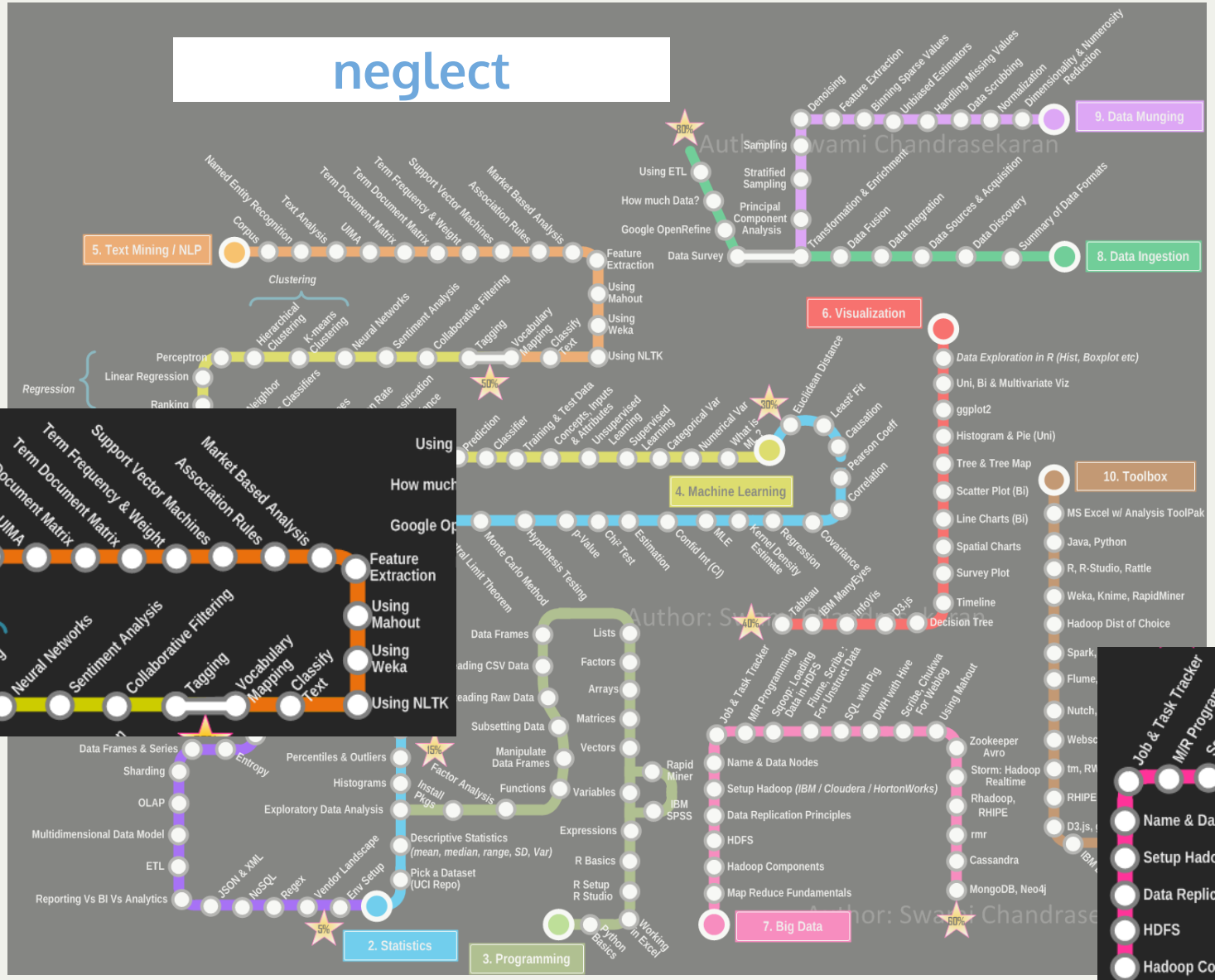


http://nirvacana.com/thoughts/2013/07/08/becoming-a-data-scientist



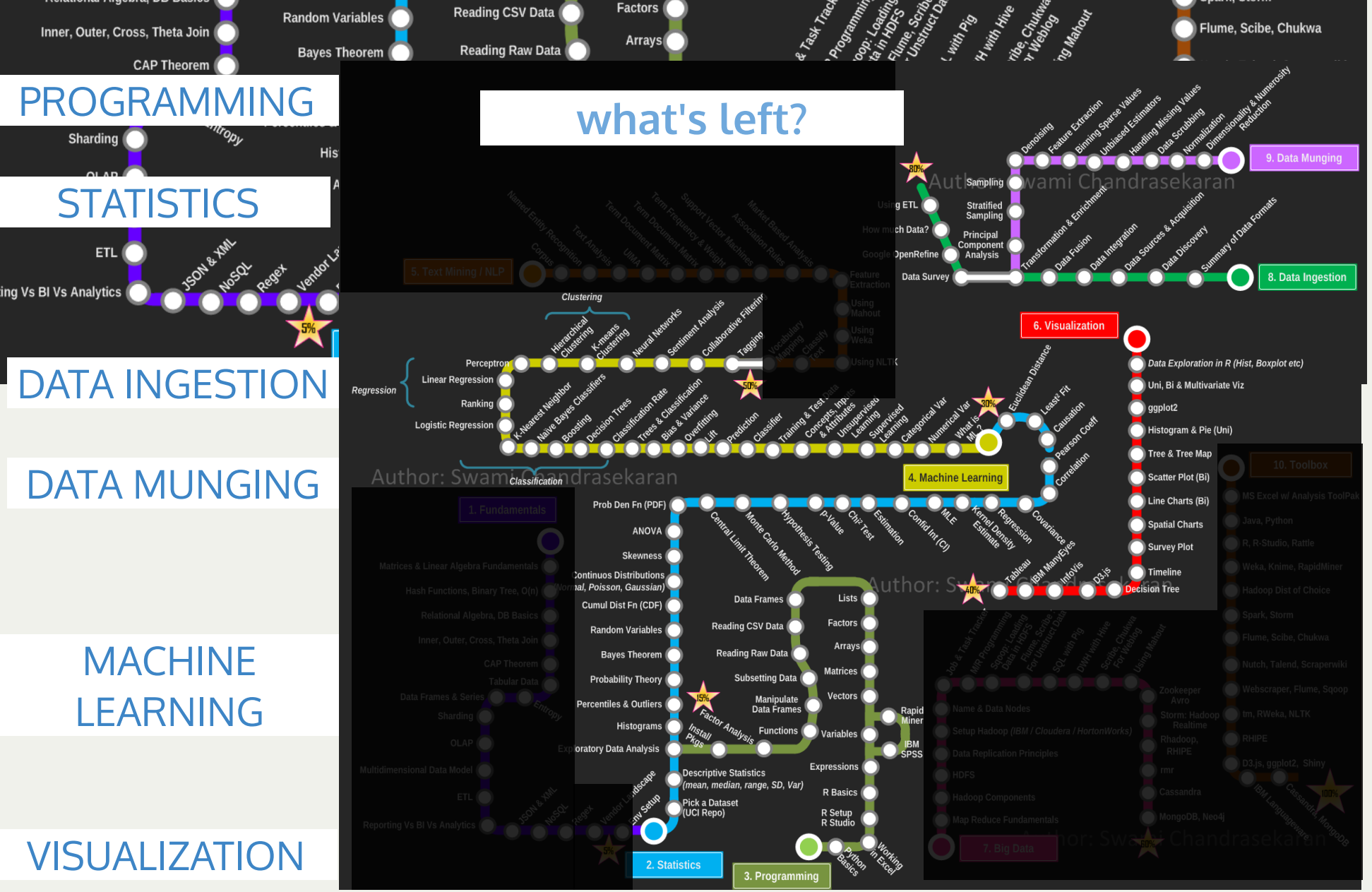
http://nirvacana.com/thoughts/2013/07/08/becoming-a-data-scientist

neglect



http://nirvacana.com/thoughts/2013/07/08/becoming-a-data-sci

what's left?



http://nirvacana.com/thoughts/2013/07/08/becoming-a-data-scientist

PROGRAMMING

STATISTICS

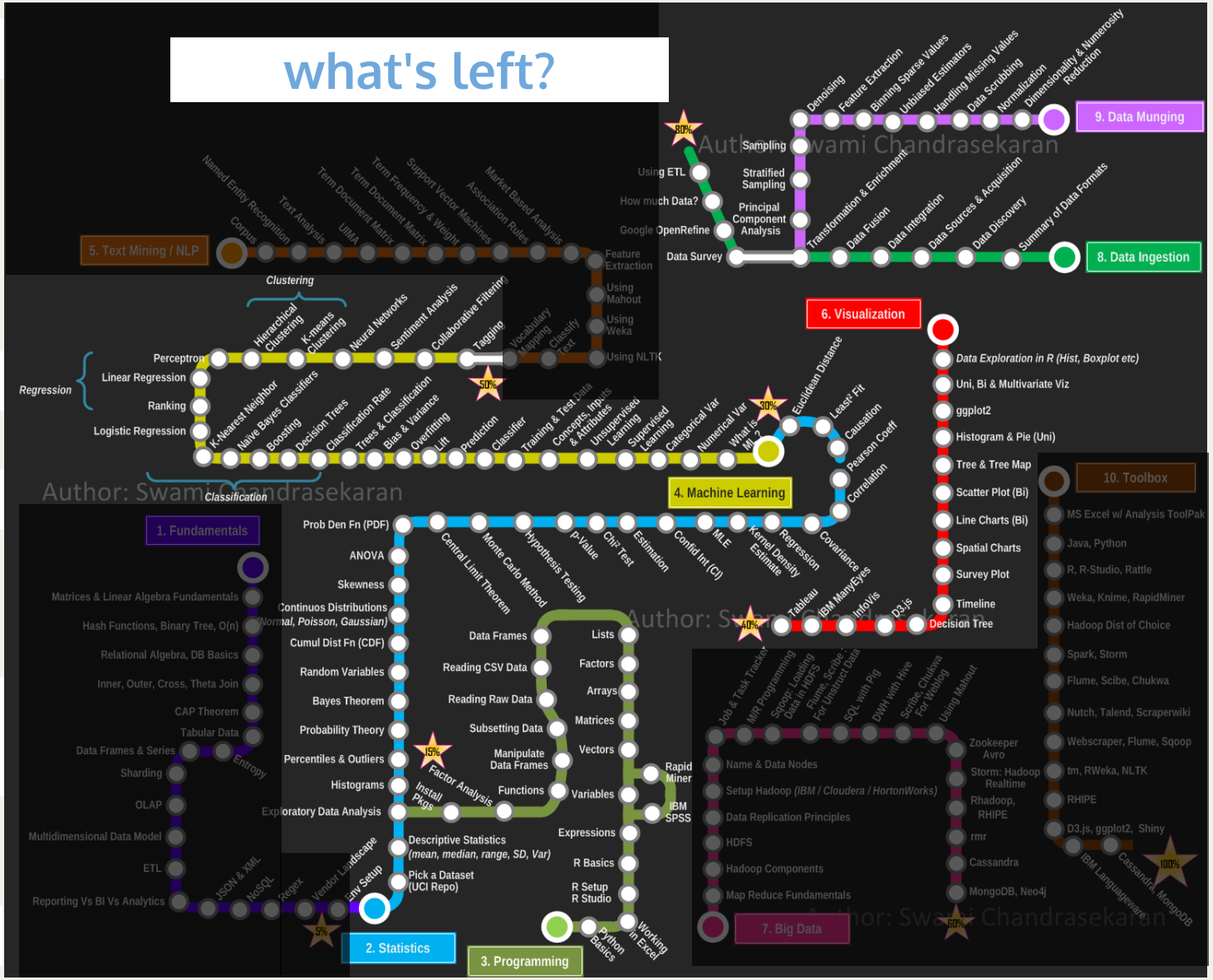
DATA INGESTION

DATA MUNGING

MACHINE LEARNING

VISUALIZATION

python



http://nirvacana.com/thoughts/2013/07/08/becoming-a-data-scientist

PROGRAMMING

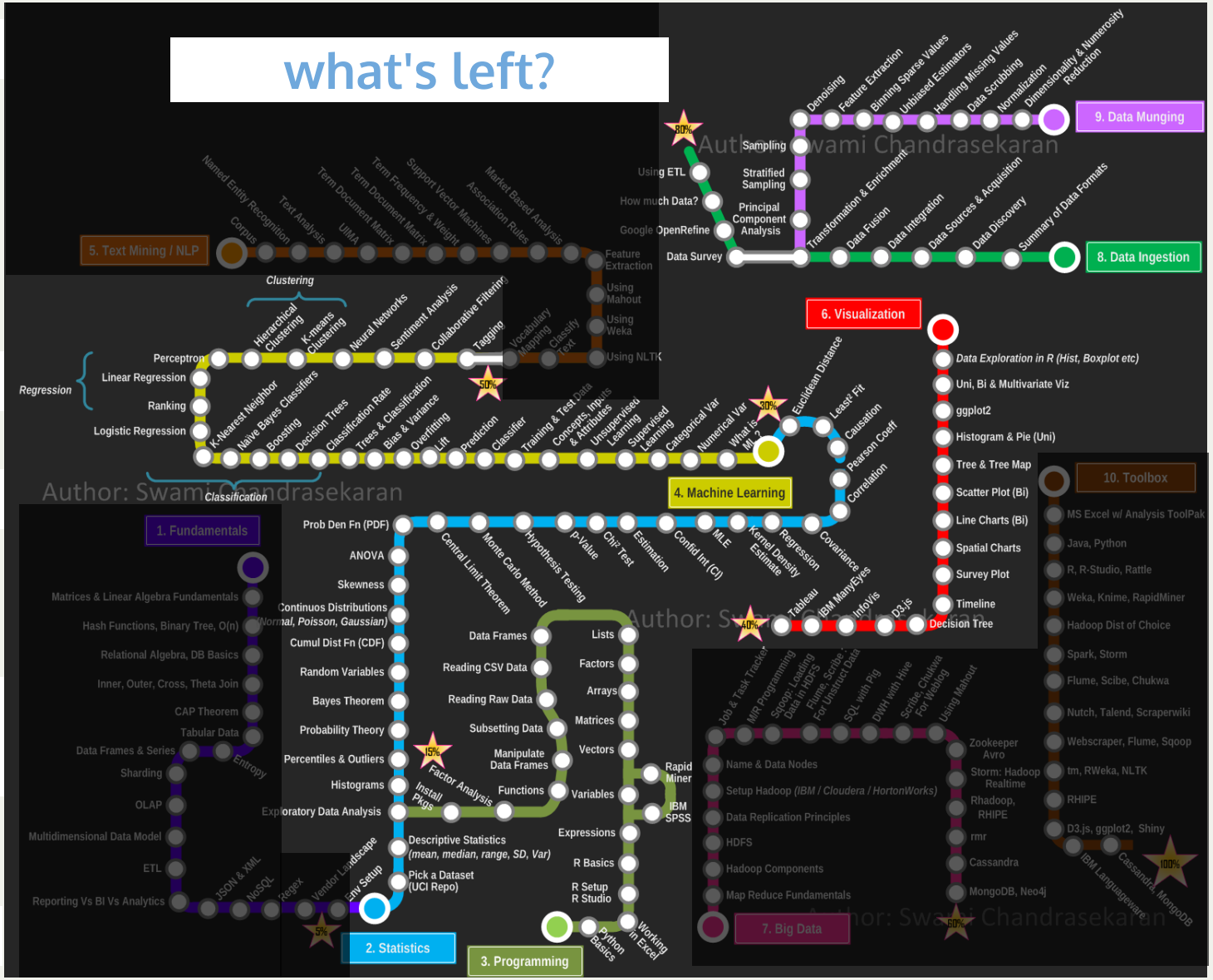
STATISTICS

DATA INGESTION

DATA MUNGING

MACHINE LEARNING

VISUALIZATION



http://nirvacana.com/thoughts/2013/07/08/becoming-a-data-scientist

python

probability distributions

p-values

uncertainties

MCMC

PROGRAMMING

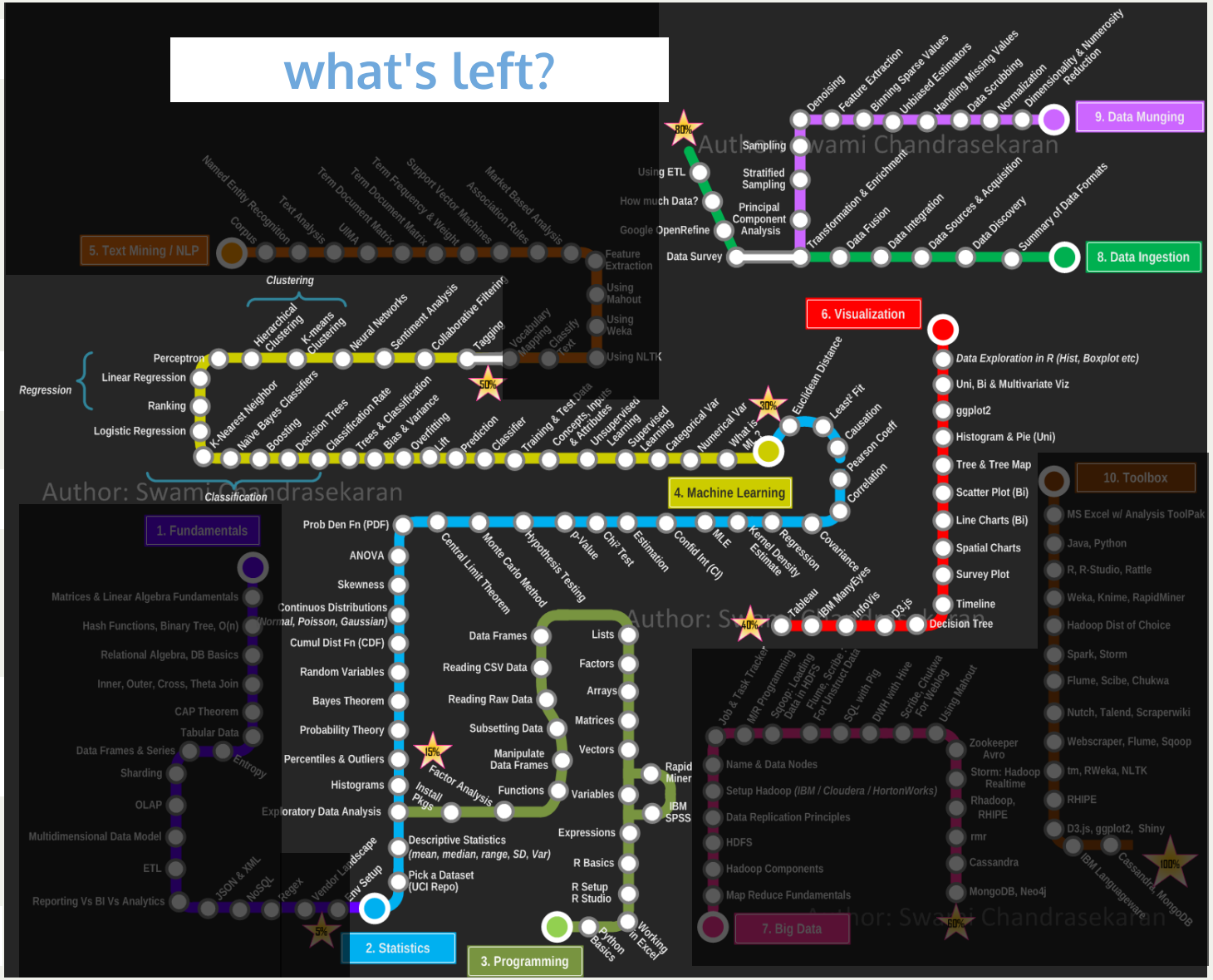
STATISTICS

DATA INGESTION

DATA MUNGING

MACHINE LEARNING

VISUALIZATION



http://nirvacana.com/thoughts/2013/07/08/becoming-a-data-scientist

python

probability distributions

p-values

uncertainties

MCMC

regression (linear, template)

classification (trees, neural networks)

clustering

2 the scientific method (what is science?)

epistemology:

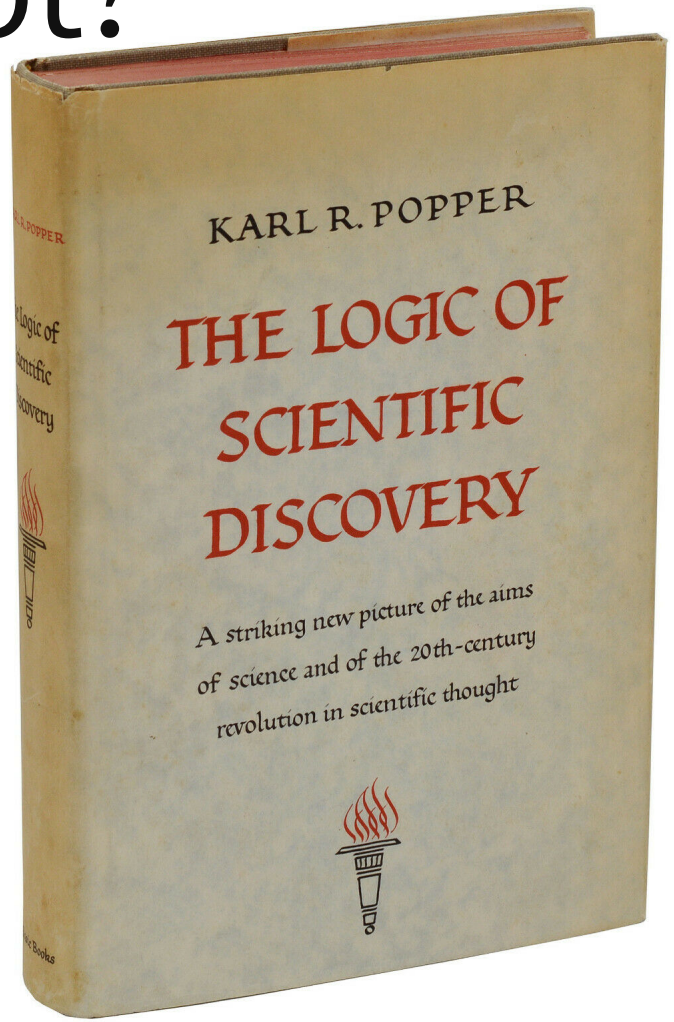
the philosophy of science and of the scientific method

the *demarcation* problem:

what is science? what is not?

My proposal is based upon an *asymmetry* between verifiability and **falsifiability**; an asymmetry which results from the logical form of universal statements. For these are never derivable from singular statements, but can be contradicted by singular statements.

—Karl Popper, *The Logic of Scientific Discovery*



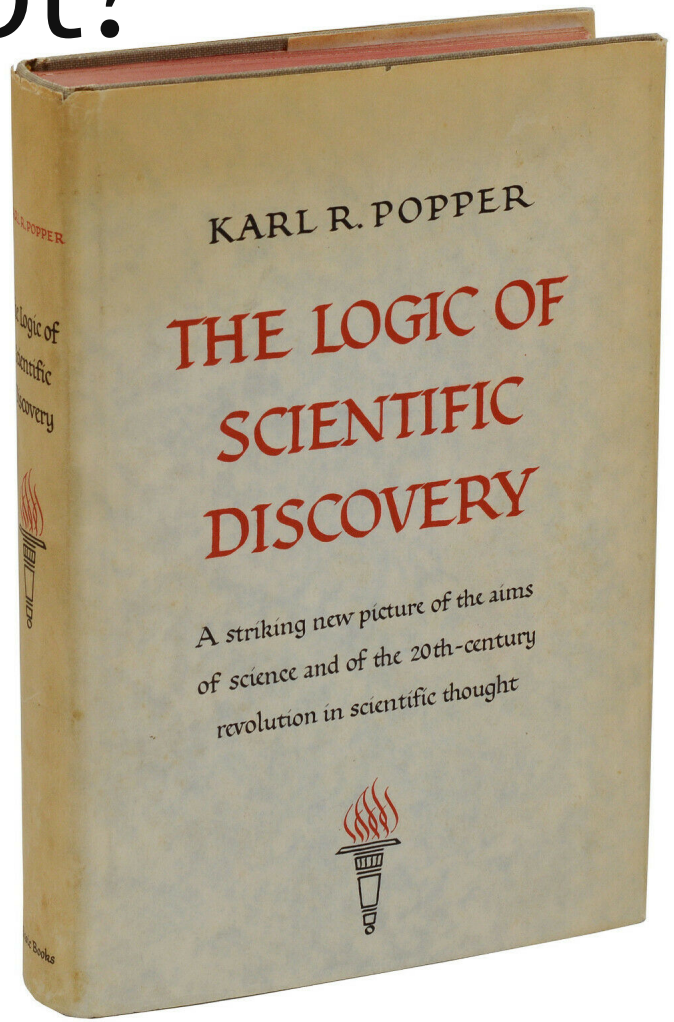
the *demarcation* problem:

what is science? what is not?

My proposal is based upon an *asymmetry* between verifiability and **falsifiability**; an asymmetry which results from the logical form of universal statements. For these are never derivable from singular statements, but can be contradicted by singular statements.

—Karl Popper, *The Logic of Scientific Discovery*

a scientific theory must be *falsifiable*



the *demarcation* problem

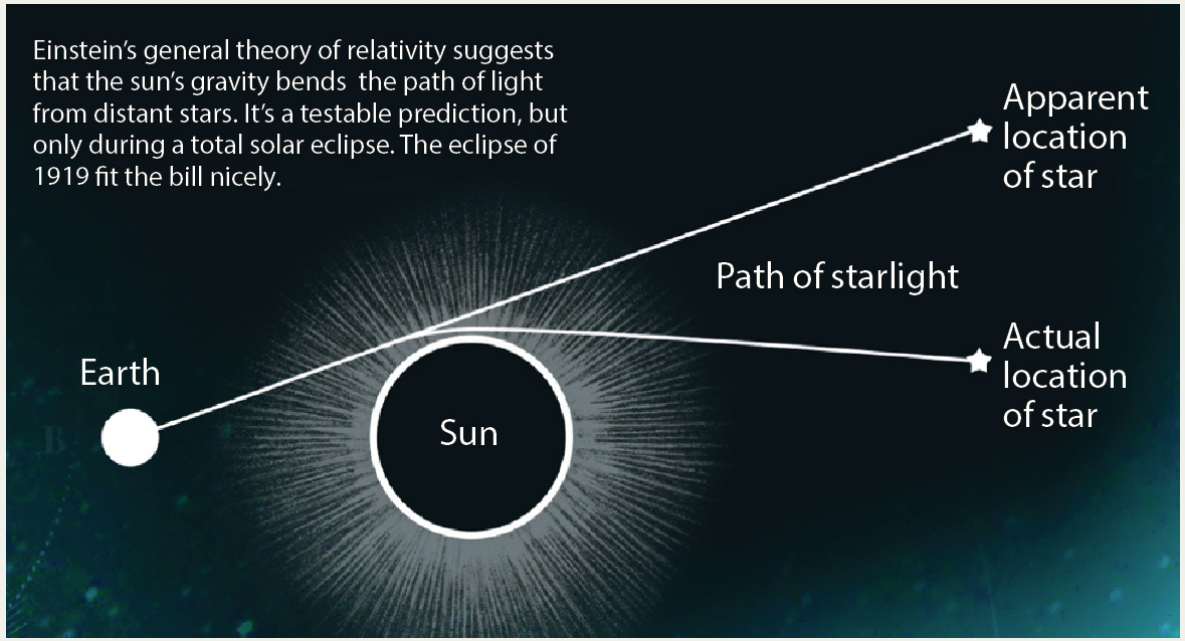
model  **prediction**

the *demarcation* problem

model → **prediction**

Einstein GR

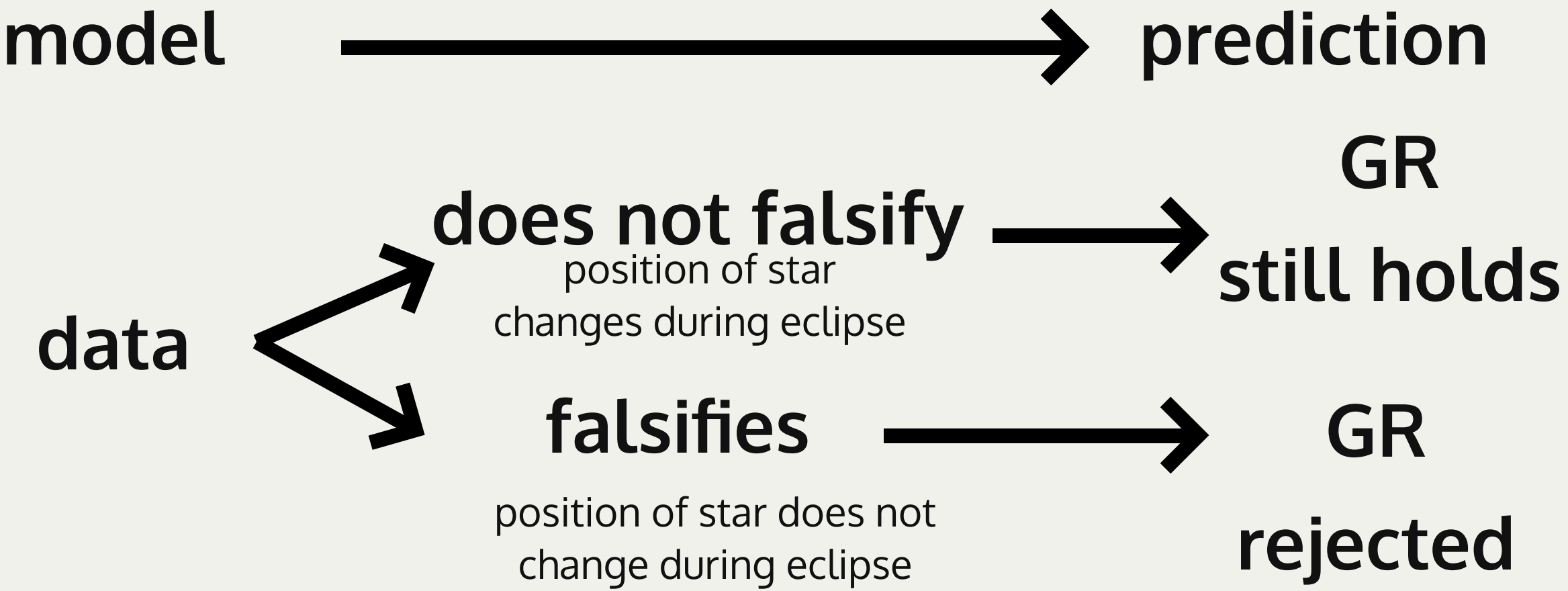
$$R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} = 8\pi G T_{\mu\nu}$$



Light rays are deflected by mass

http://discovermagazine.com/2019/may/why-it-took-the-1919-solar-eclipse-for-physicists-to-believe-einstein

the *demarcation* problem



the *demarcation* problem

is phsychology a science?

DISCUSS!

the *demarcation* problem

things can get more complicated though:

most scientific theories are actually based largely on *probabilistic induction* and modern *inductive inference* (Solomonoff, frequentist vs Bayesian methods...)

Reproducibility

Reproducible research means:

the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results.

https://acmedsci.ac.uk/viewFile/56314e40aac61.pdf

Reproducibility

Reproducible research means:

the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results.

https://acmedsci.ac.uk/viewFile/56314e40aac61.pdf

why?

1

assures a result is grounded in evidence

#openscience

#opendata

Reproducibility

Reproducible research means:

the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results.

https://acmedsci.ac.uk/viewFile/56314e40aac61.pdf

why?

facilitates scientific progress by avoiding the need to duplicate unoriginal research

Reproducibility

Reproducible research means:

the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results.

https://acmedsci.ac.uk/viewFile/56314e40aac61.pdf

why?

3

facilitate collaboration and teamwork

Reproducibility

Reproducible research means:

the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results.

https://acmedsci.ac.uk/viewFile/56314e40aac61.pdf

Reproducible research in practice:

all numbers in a data analysis can be recalculated exactly (down to stochastic variables!)

using the **code** and **raw data** provided by the analyst.

Claerbout, J. 1990,

Active Documents and Reproducible Results, Stanford Exploration Project Report, 67, 139

Repdorucibility

Reproducible research means:

the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results.

https://acmedsci.ac.uk/viewFile/56314e40aac61.pdf

Reproducible research in practice:

all numbers in a data analysis can be recalculated exactly (down to stochastic variables!)

using the **code** and **raw data** provided by the analyst.

- provide raw data and code to reduce it to all stages needed to get outputs
- provide code to reproduce all figures
- provide code to reproduce all number outcomes

3 the tools

github

reproducibility

https://github.com

Reproducible research means:

all numbers in a data analysis can be recalculated exactly (down to stochastic variables!) using the **code** and **raw data** provided by the analyst.

Claerbout, J. 1990,

Active Documents and Reproducible Results, Stanford Exploration Project Report, 67, 139

allows reproducibility through code distribution

github

version control

https://github.com

the Git software

is a distributed *version control system:*

a version of the files on your local computer is made also available at a central server.

The history of the files is saved remotely so that any version (that was checked in) is retrievable.

allows version control

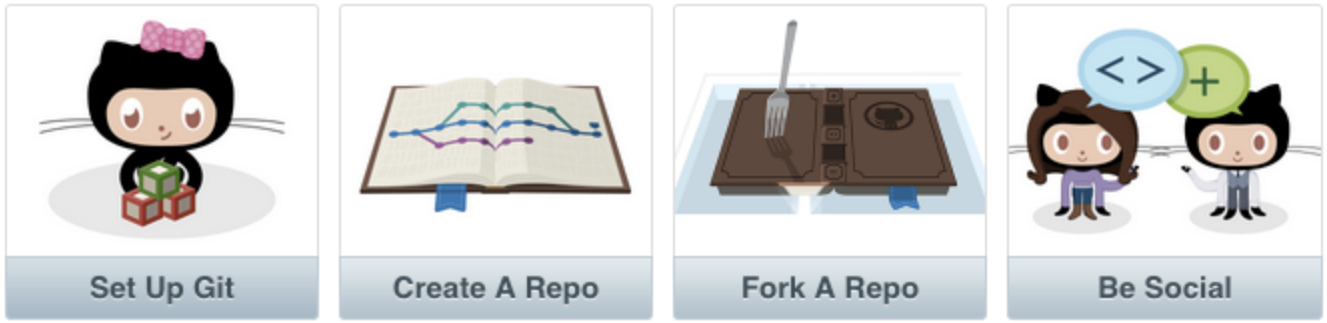
github

collaborative platform

https://github.com

collaboration tool

by fork, fork and pull request, or by working directly as a collaborator

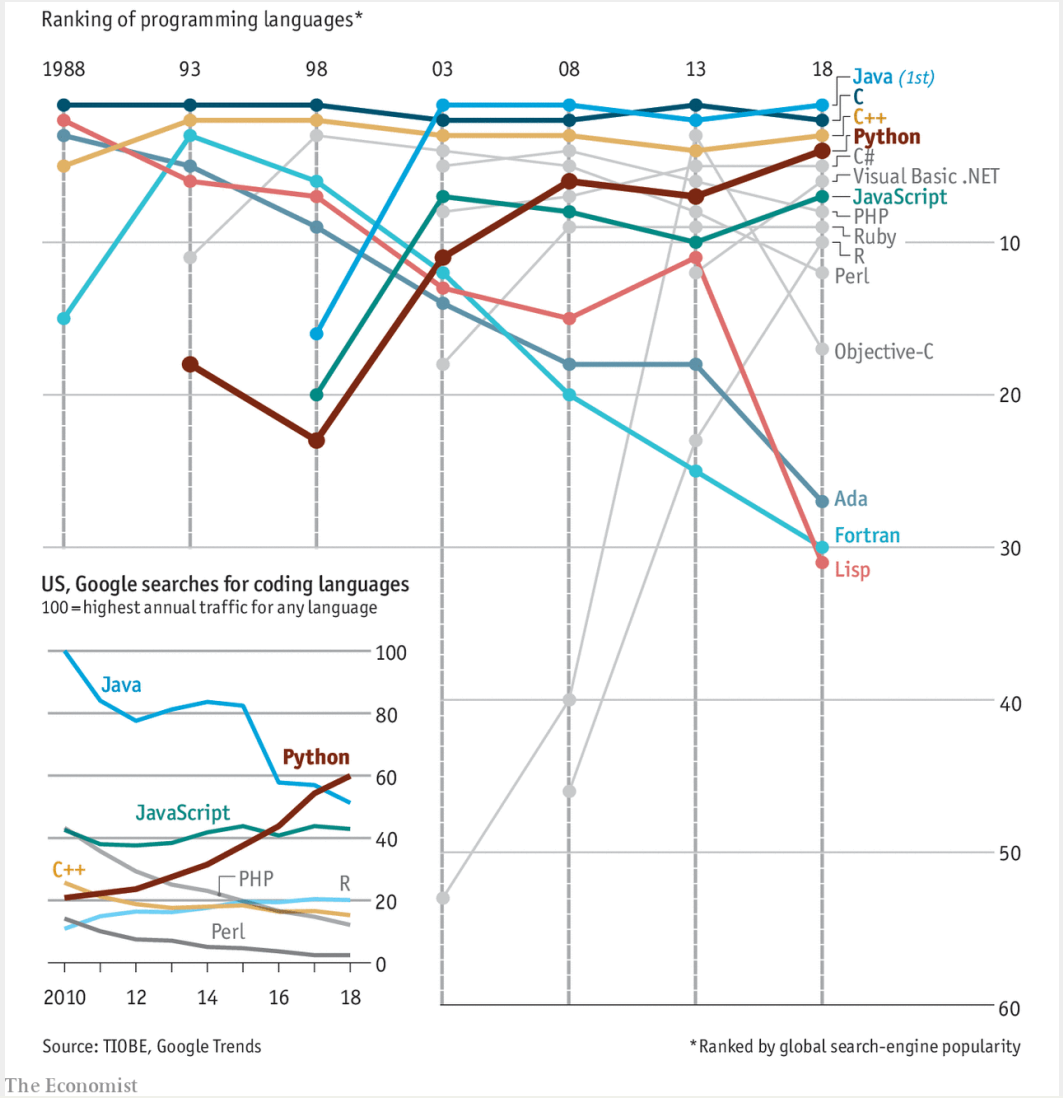


allows effective collaboration

python

- intuitive and readable
- open source
- support **C** integration for performance
- packages designed for science:
  - scipy
  - statsmodels
  - numpy (computation)
  - sklearn (machine learning)

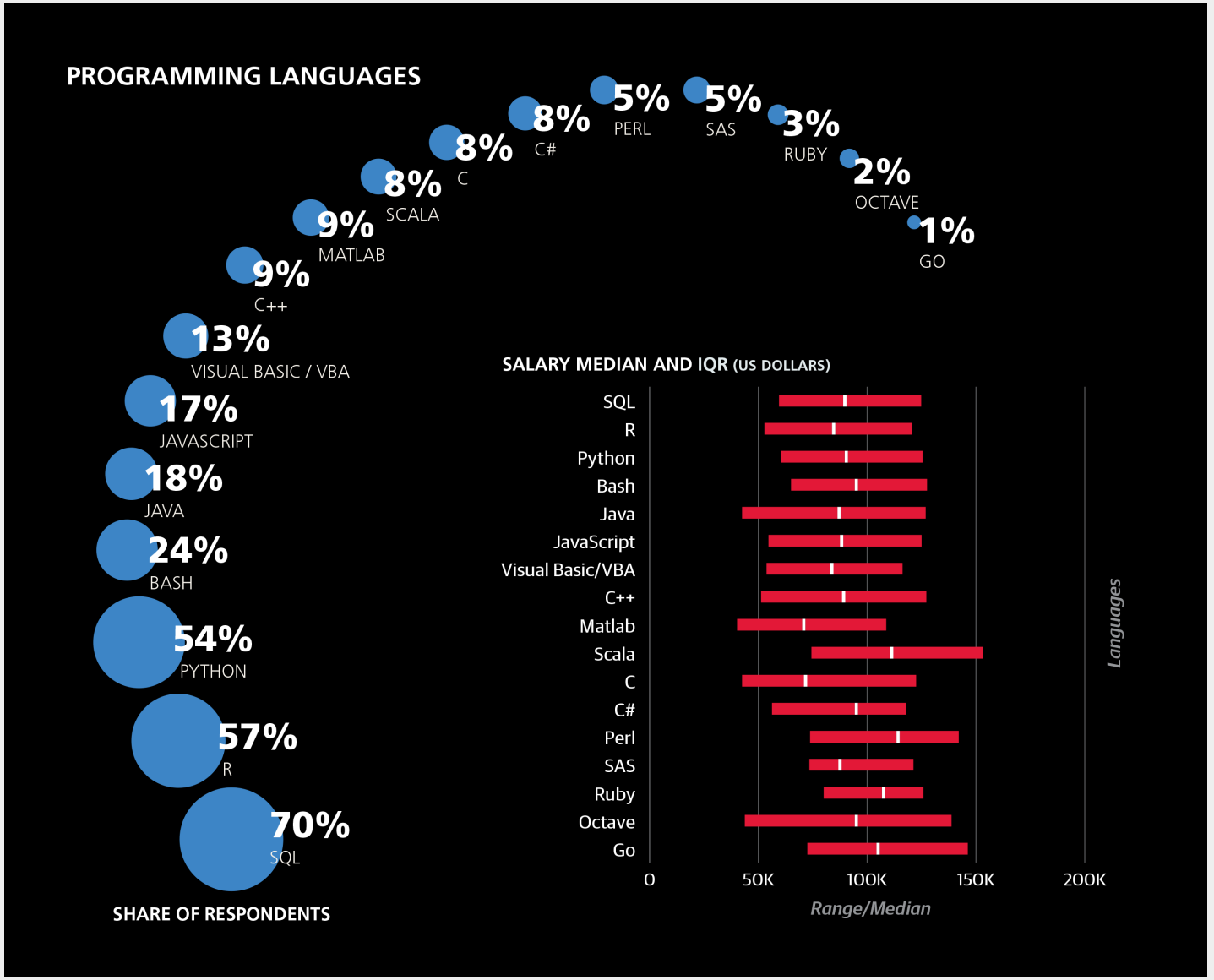
https://www.economist.com/graphic-detail/2018/07/26/python-is-becoming-the-worlds-most-popular-coding-language



python

- intuitive and readable
- open source
- support **C** integration for performance
- packages designed for science:
  - scipy
  - statsmodels
  - numpy (computation)
  - sklearn (machine learning)

https://www.oreilly.com/ideas/2016-data-science-salary-survey-results



python

https://sharmamohit.com/work/tutorials/ucsl/

series of notebooks designed for Urban Science students by Dr. Mohit Sharma (in consultation with me)

recommanded if you are brand new to python and coding or are serious about cleaning up your foundamentals

python

quick bootcamp

recommanded if you know some python or if you know some other conding language reasonably proficiently

https://github.com/fedhere/PyBOOT

Table of Contents


```
from __future__ import print_function, division
#importing this to make code python2&3 compatible
# overwrites the default print to require parenthesis
# overwrites the default / (division operator) behavior
#   so division of 2 integers returns a float
```

python

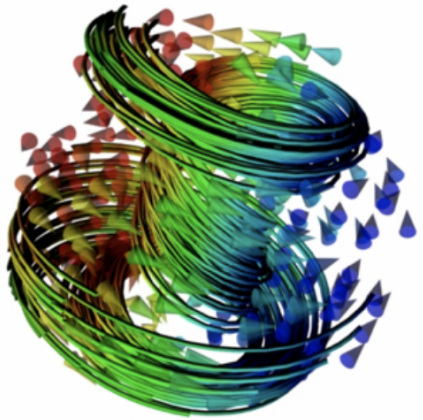
online book

https://www.southampton.ac.uk/~fangohr/training/python/pdfs/Python-for-Computational-Science-and-Engineering.pdf

Introduction to

Python for Computational Science and Engineering

(A beginner's guide)



Hans Fangohr
Faculty of Engineering and the Environment
University of Southampton

September 7, 2015

python

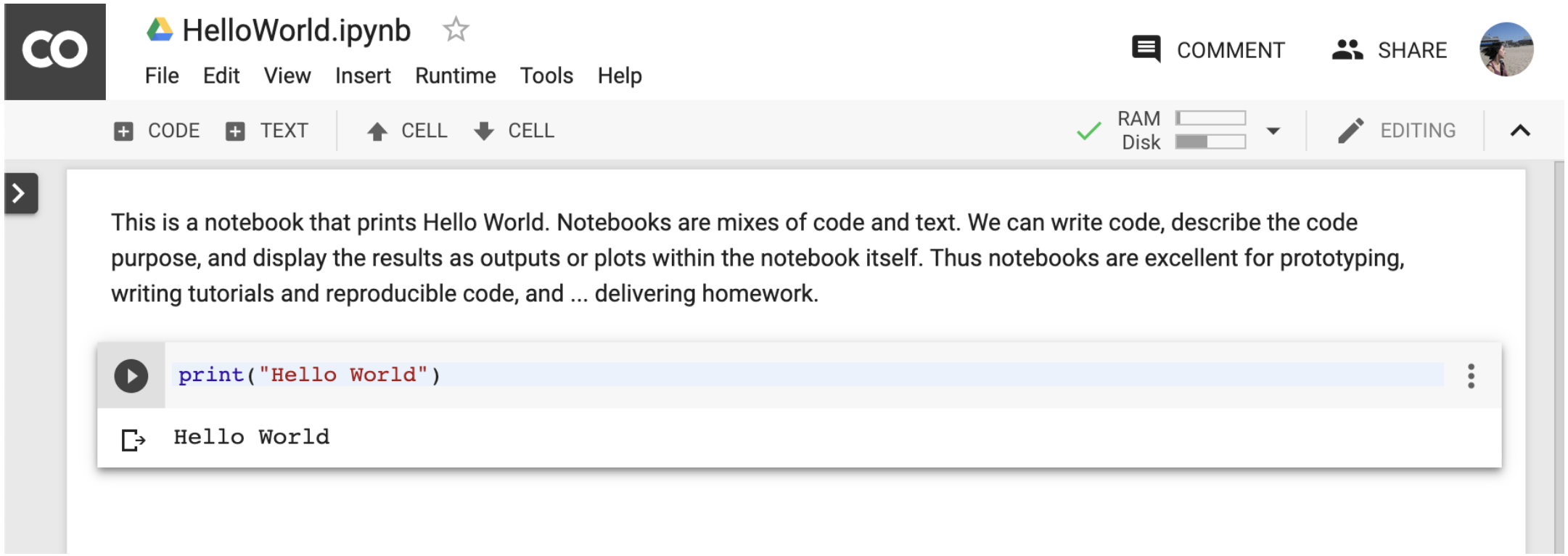
PEP8: Python Enhancement Proposals 8

"This document gives coding conventions for the Python code comprising the standard library in the main Python distribution."

Indentation, *Tabs vs Spaces*, *Maximum Line Length*, *Blank Lines*, *Source File Encoding*, *Imports*, *Whitespace in Expressions and Statements* , *Imports*, *Comments Bookeeping*, *Naming*

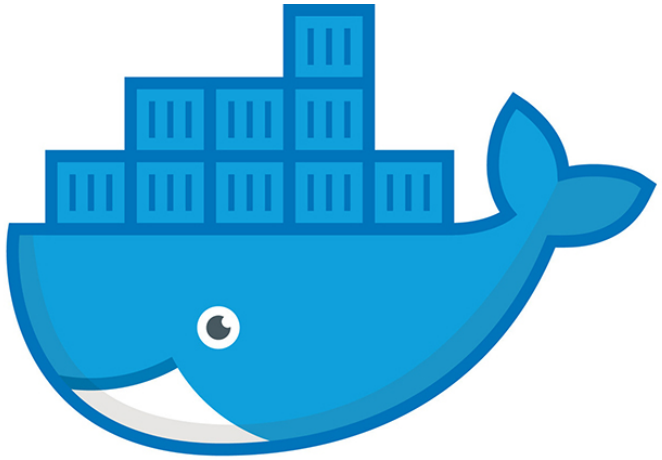
Jupyter Notebook
Google Colaboratory

https://colab.research.google.com/notebooks/welcome.ipynb#



Jupyter Notebook

local setup



docker

1

install docker image
from my account here

Jupyter Notebook
local setup

handle your own installation with python or anaconda (or whatever else on linux and windows) but make sure results are reproducible on google colab

stackoverflow

for when you need help

https://stackoverflow.com/

you can ask coding questions, installation questions, colab questions...

How to type list comprehensions

Ask Question

0

I have the following list comprehensions in Python:

```
from typing import cast

# everything is fine
print([value for value in [1, 2, 3, 4]])

# on the first "value": Expression type contains "Any" (has type "List[Any]")
print("{}".format([value for value in [1, 2, 3, 4]]))

# on the "cast": Expression type contains "Any" (has type "List[Any]")
print("{}".format([cast(int, value) for value in [1, 2, 3, 4]]))
```

Why does using `format` cause Mypy to give me back errors? As you can see, I tried to use casting and it still failed.

This question looks similar, but my particular case is weird because Mypy seems to be fine as long as I'm not using the `format` function (yet it's always okay with the `print` function).

Is there anything I can do to not have the lines with formatting give me errors? (Or should I just `# type: ignore` them?)

python | python-3.x | list-comprehension | typing | mypy

stackoverflow

for when you need help

https://stackoverflow.com/

you can ask coding questions, installation questions, colab questions...

Multiple output regression or classifier with one (or more) parameters with Python

Ask Question

5

I wrote a simple linear regression and decision tree classifier code with Python's Scikit-learn library for predicting the outcome. It works good.

My question is, Is there a way to do this backwards, to predict the best combination of parameter values based on imputed outcome (parameters where accuracy will be the best).

Or I can ask like this, is there a classification, regression or some other type of algorithm (decision tree, SVM, KNN, logistic regression, linear regression, polynomial regression...) that can predict multiple outcomes based on one (or more) parameter/s?

I have tried to do this with putting multivariate outcome, but it shows the error:

```
ValueError: Expected 2D array, got 1D array instead:
array=[101 905 182 268 646 624 465].
Reshape your data either using array.reshape(-1, 1) if your data has a single feature
```

This is the code that I wrote for regression:

```
import pandas as pd
from sklearn import linear_model
from sklearn import tree

dic = {'par_1': [10, 30, 13, 19, 25, 33, 23],
       'par_2': [1, 3, 1, 2, 3, 3, 2],
       'outcome': [101, 905, 182, 268, 646, 624, 465]}
```

stackoverflow

for when you need help

it can be a toxic environment...

https://stackoverflow.com/

you can ask ciding questions, installation questions, colab questions...

Multiple output regression or classifier with one (or more) parameters with Python

Ask Question

5

I wrote a simple linear regression and decision tree classifier code with Python's Scikit-learn library for predicting the outcome. It works good.

My question is, Is there a way to do this backwards, to predict the best combination of parameter values based on imputed outcome (parameters where accuracy will be the best).

Or I can ask like this, is there a classification, regression or some other type of algorithm (decision tree, SVM, KNN, logistic regression, linear regression, polynomial regression...) that can predict multiple outcomes based on one (or more) parameter/s?

I have tried to do this with putting multivariate outcome, but it shows the error:

```
ValueError: Expected 2D array, got 1D array instead:
array=[101 905 182 268 646 624 465].
Reshape your data either using array.reshape(-1, 1) if your data has a single feature
```

This is the code that I wrote for regression:

```
import pandas as pd
from sklearn import linear_model
from sklearn import tree

dic = {'par_1': [10, 30, 13, 19, 25, 33, 23],
       'par_2': [1, 3, 1, 2, 3, 3, 2],
       'outcome': [101, 905, 182, 268, 646, 624, 465]}
```

key concepts

Science and Data Science

Falsifiability

Reproducibility

1

homework

- make an account on GitHub if you do not have one yet
- Create a repository called DSPS_<firstinitialLastname>
- upload your repo URL on canvas as HW 1

2

homework

- make an account on GitHub if you do not have one yet
- Create a repository called DSPS_<firstinitialLastname>

https://github.com/fedhere/DSPS/tree/master/HW1

Jeff Leek & Rodger Peng.
2015,
What is the Question?

http://fbb.space/dsps/The%20Research%20Question-2015-Leek-1314-5.pdf

reading

the original link:
https://science.sciencemag.org/content/347/6228/1314.summary
is link nees access to science magazine, but ou can use the link above which is the same file)

STATISTICS

What is the question?

Mistaking the type of question being considered is the most common error in data analysis

By **Jeffery T. Leek** *and* **Roger D. Peng**

additional reading

Karl Popper, J. 1934,

The Logic of Scientific Discovery

http://strangebeautiful.com/other-texts/popper-logic-scientific-discovery.pdf

Claerbout, J. 1990,

Active Documents and Reproducible Results, Stanford Exploration Project Report, 67, 139

http://sepwww.stanford.edu/data/media/public/docs/sep67/jon2/paper_html/