ML for physical and natural scientists 2023 1

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

https://slides.com/federicabianco/mlpns23_1

who am I?

astrophysics -> data science

astrophysics stands our as an *observational*, rather than experimental science

to "observe" the natural status of a system provides advantages. To inquire the system about its status may provide a biased response

(e.g. surveys have many biases)

who am I?

ural status of a
vantages. To
bout its status
ed response
many biases)

https://muonetwork.github.io/

Historical perspective: Big Data

"Data larger than can be analyzed with typical tool"

"Data that stresses the infrastructure"

"Data that does not fit in memory"

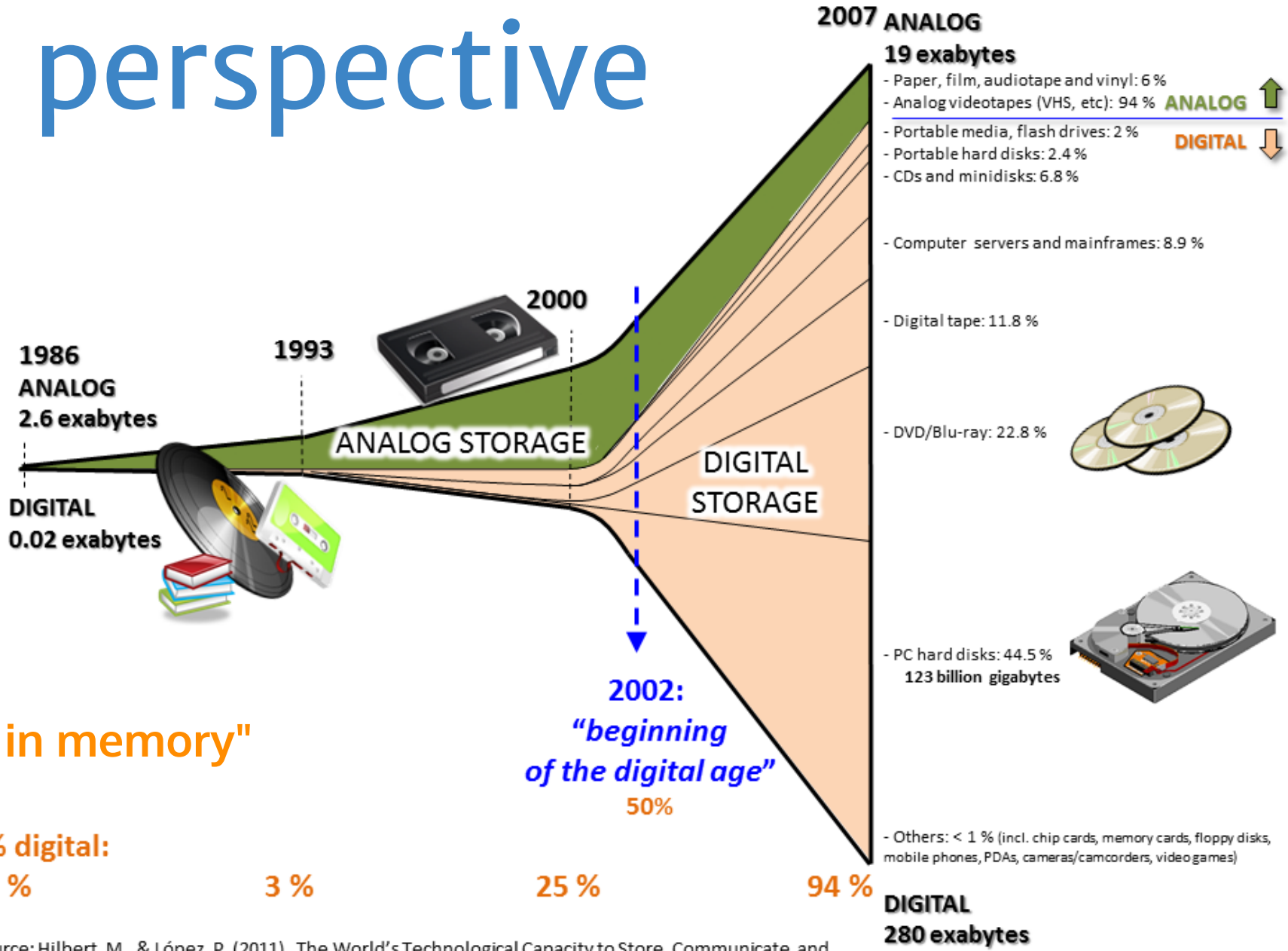
Historical perspective: Big Data

"Data larger than can be analyzed with typical tool"

John R. Mashey Chief Scientist, SGI, mid-1990s

https://static.usenix.org/event/usenix99/invited_talks/mashey.pdf

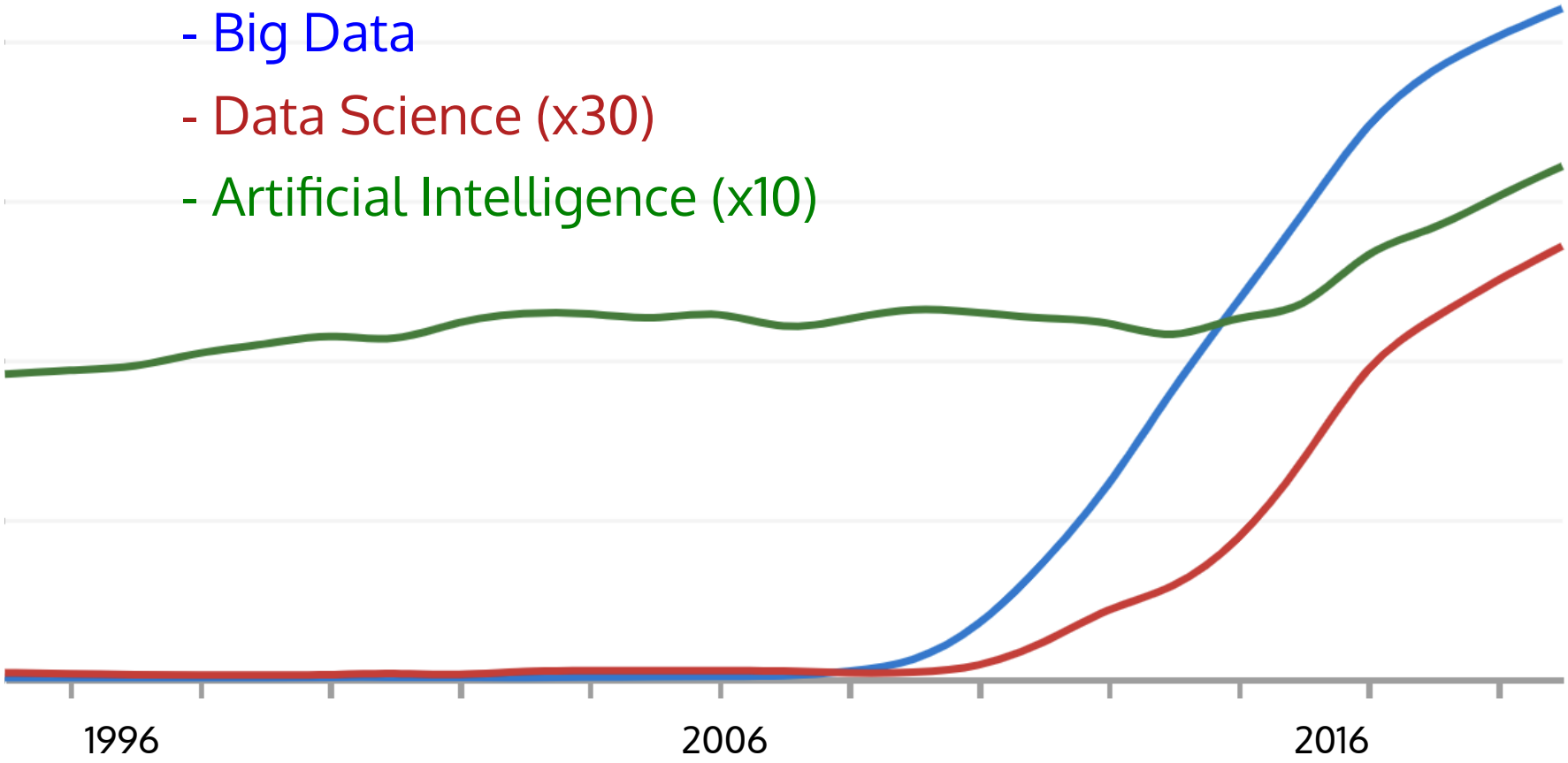
Historical perspective



"Data that does not fit in memory"

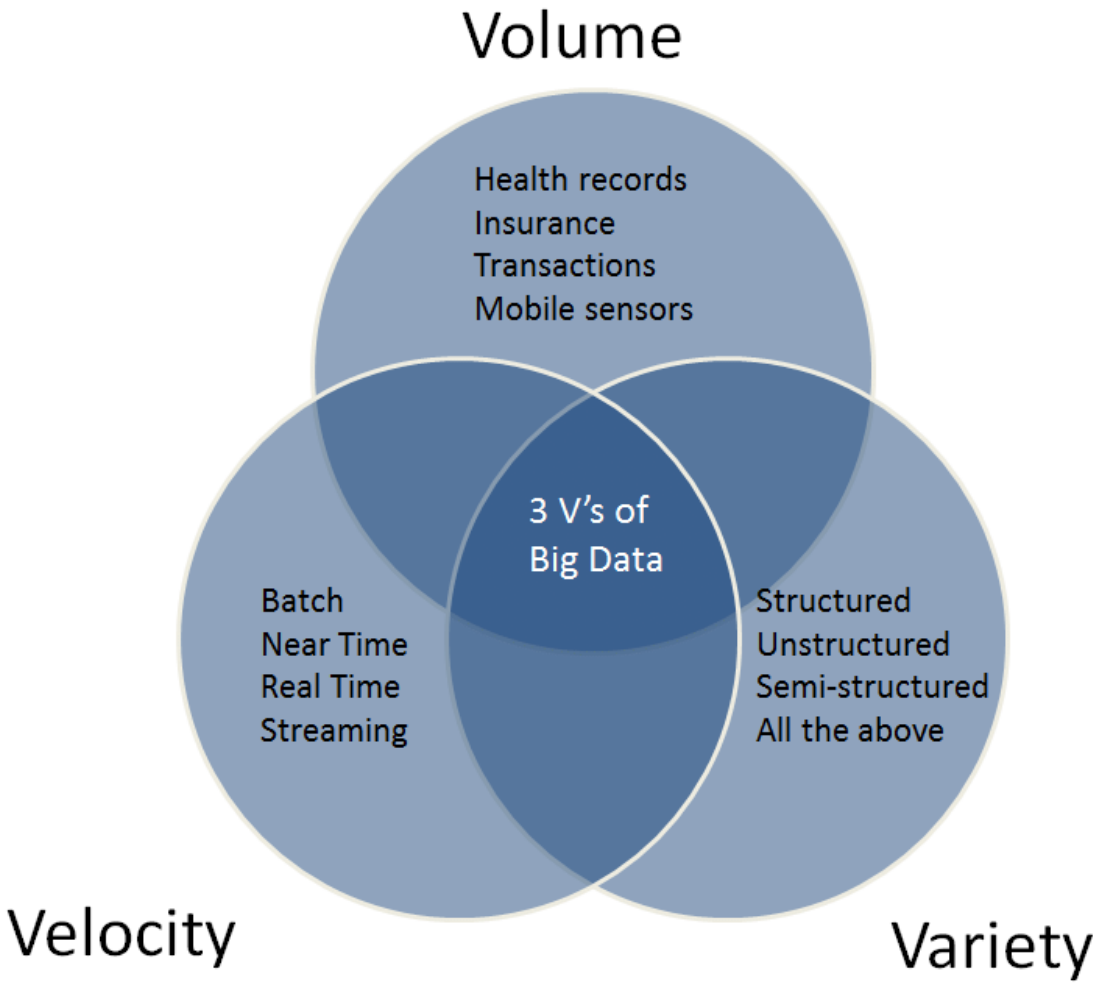
Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

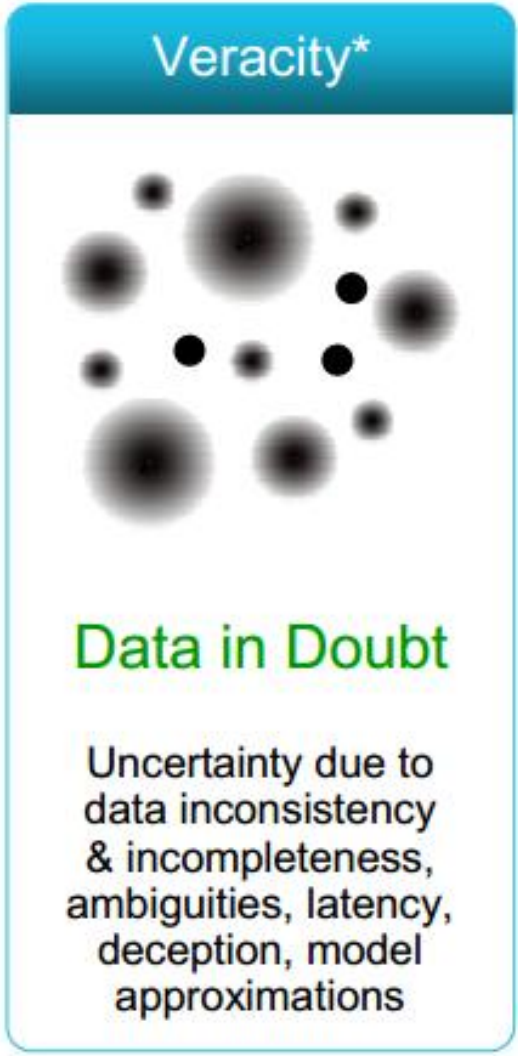
Historical perspective



occurrence of term in Google-books corpus https://books.google.com/ngrams

Historical perspective





Gartner report 2001

4-V of Big Data

V1: Volume

Number of bites

Number of pixels

Number of rows in a data table x number of columns for catalogs

V2: Variety

Diverse science return from the same dataset.

Multiwavelength

Multimessenger

Images and spectra

V3: Velocity

real time analysis, edge computing, data transfer

V4: Veracity

This V will refer to both data quality and availability (added in 2012)

Gartner report 2001

Big Data: Astronomical or Genomical?

Zachary D. Stephens, Skylar Y. Lee, Faraz Faghri, Roy H. Campbell, Chengxiang Zhai, Miles J. Efron, Ravishankar Iyer, Michael C. Schatz, Saurabh Sinha, Gene E. Robinson

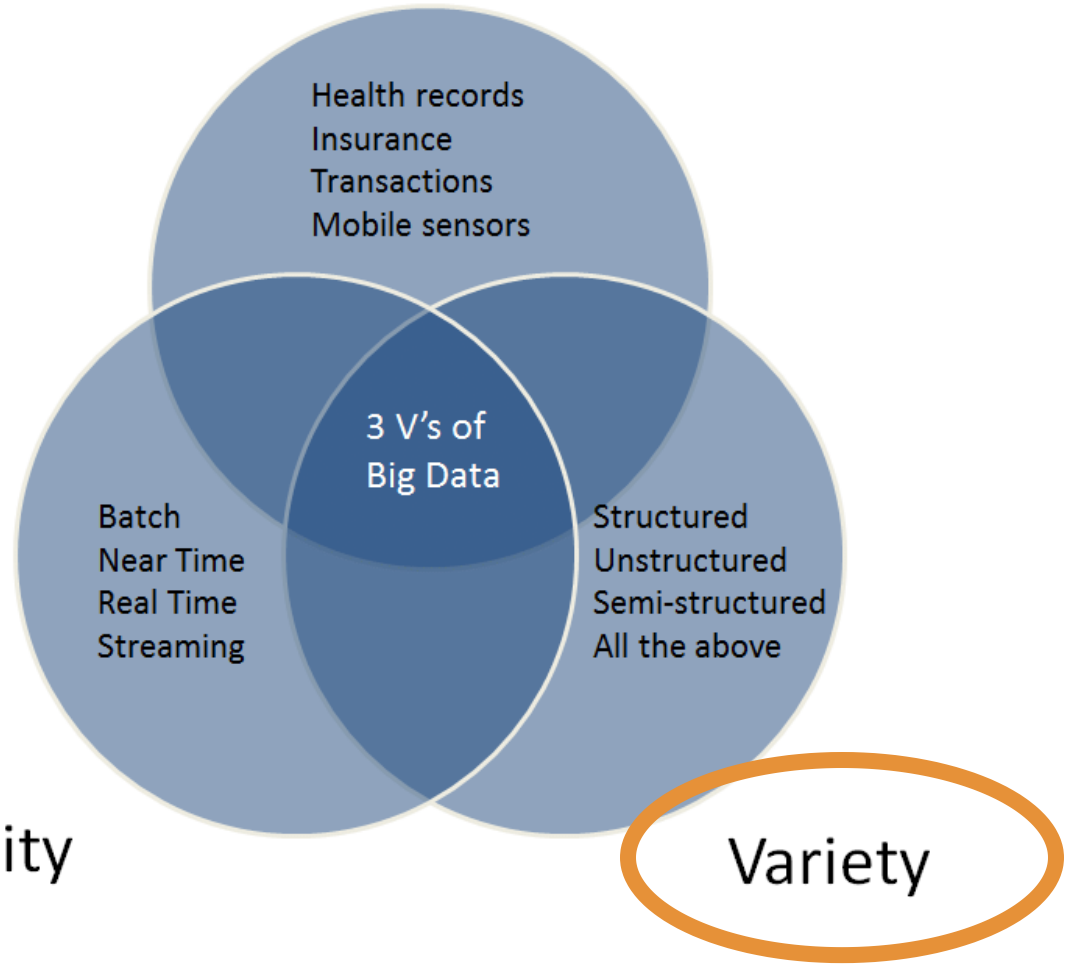
Published: July 7, 2015 • https://doi.org/10.1371/journal.pbio.1002195

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

doi:10.1371/journal.pbio.1002195.t001

Historical perspective

Text Volume



Velocity

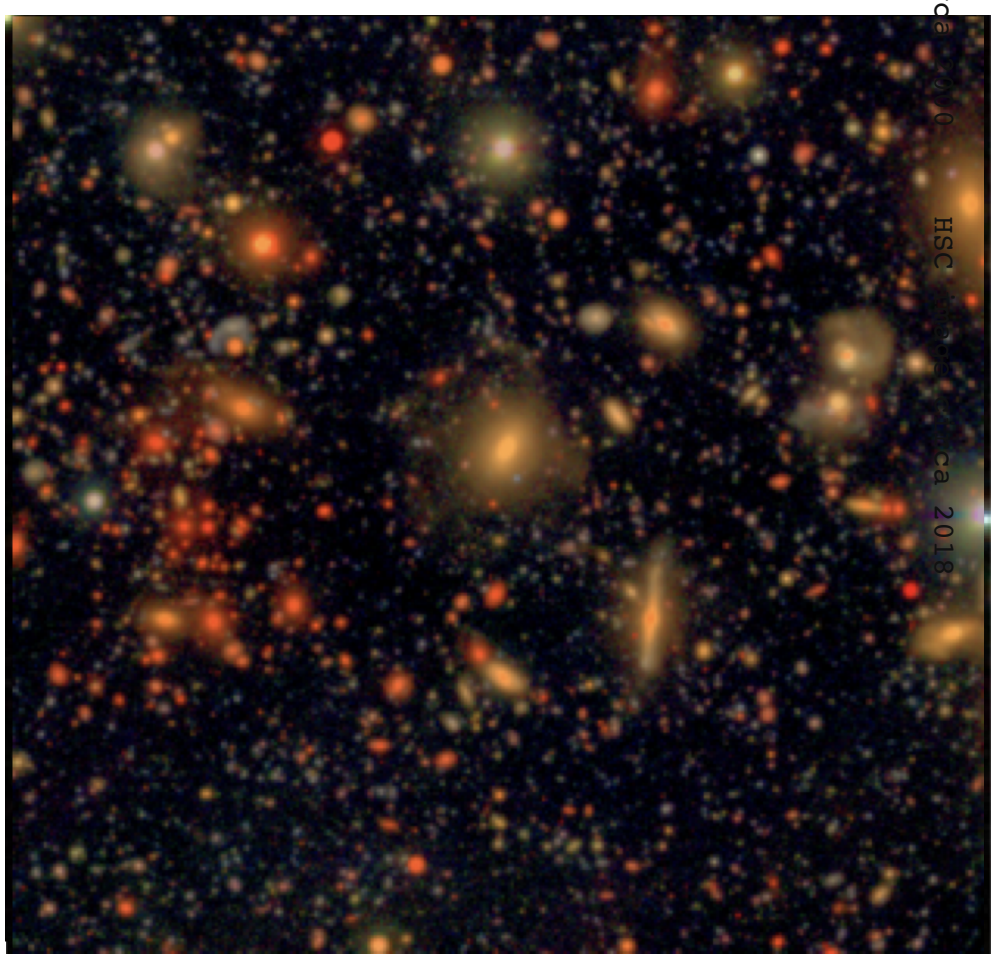
Variety

whe

eve

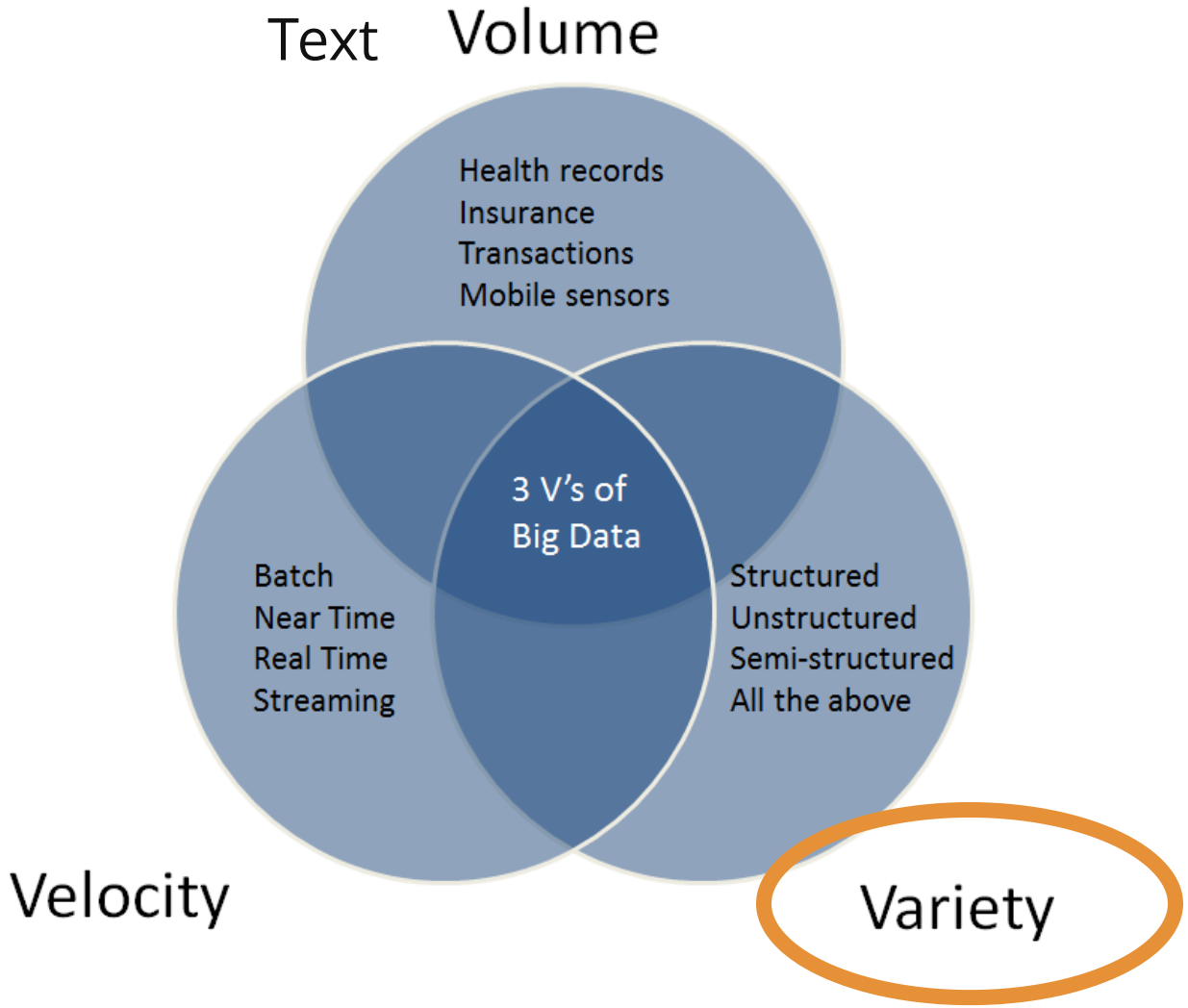
Gartner report 2001

complexity



SDSS image circa 2000 HSC circa 2018

Historical perspective



Gartner report 2001

complexity

= Astronomical data mainly include ***images, spectra, time-series data, and simulation data***.

Most of the data are saved in catalogues or databases. The ***data from different telescopes or projects have their own formats***, which causes difficulty with integrating data from various sources in the analysis phase. In general, ***each data item has a thousand or more features***; this cause a large dimensionality problem. Moreover, data have many data types: structured, semi structured, unstructured, and mixed.

SDSS image circa 2000

HST image circa 2018

why Data Science?

astrophysics -> data science

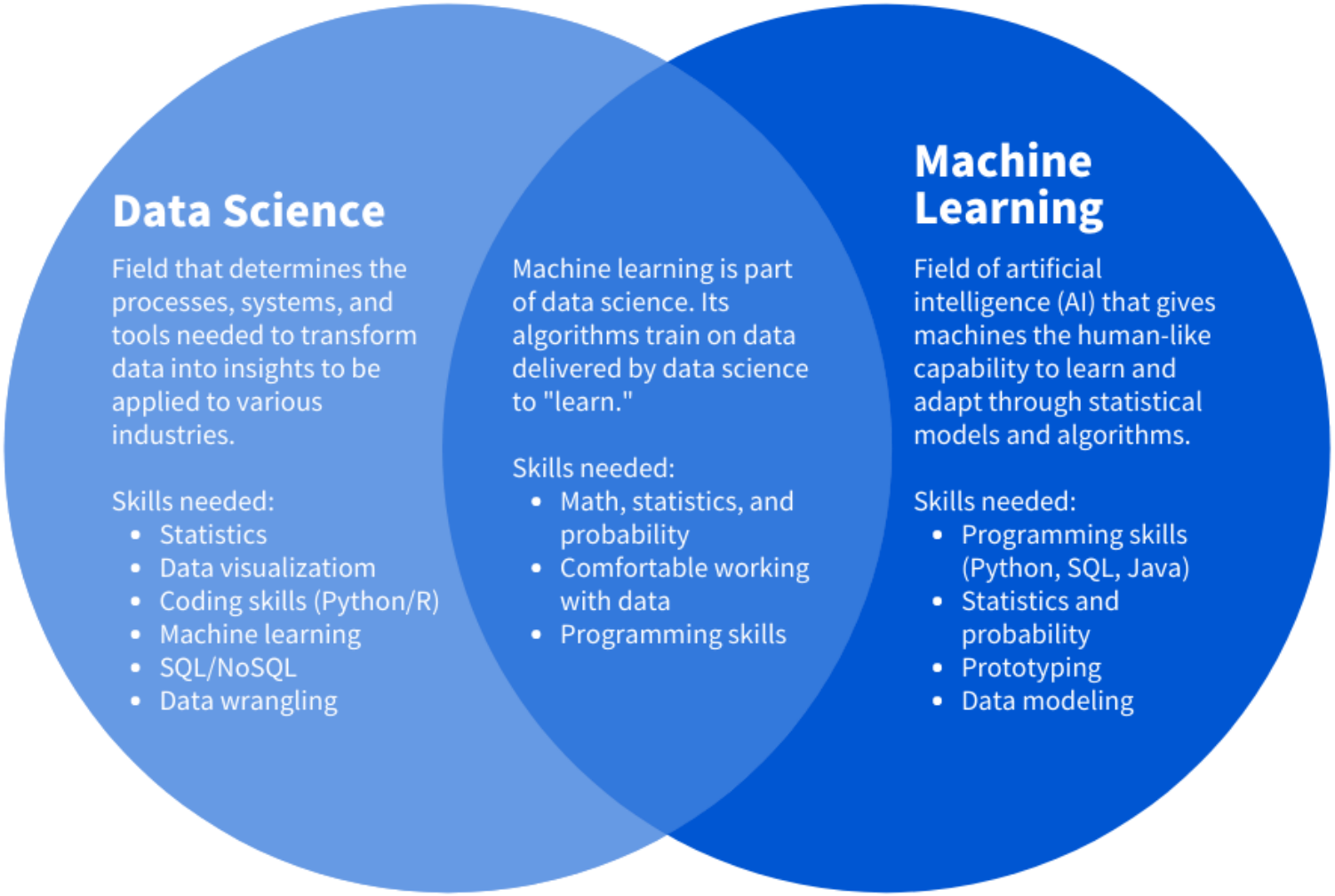
UD Data Science Institute - Inaugural event

what is data science? we have been using data in science the whole time, but with the volume, rate, and complexity of the current data we have to worry about things that we would neglect until now: what happens if our data has errors, what happens if we have missing data?

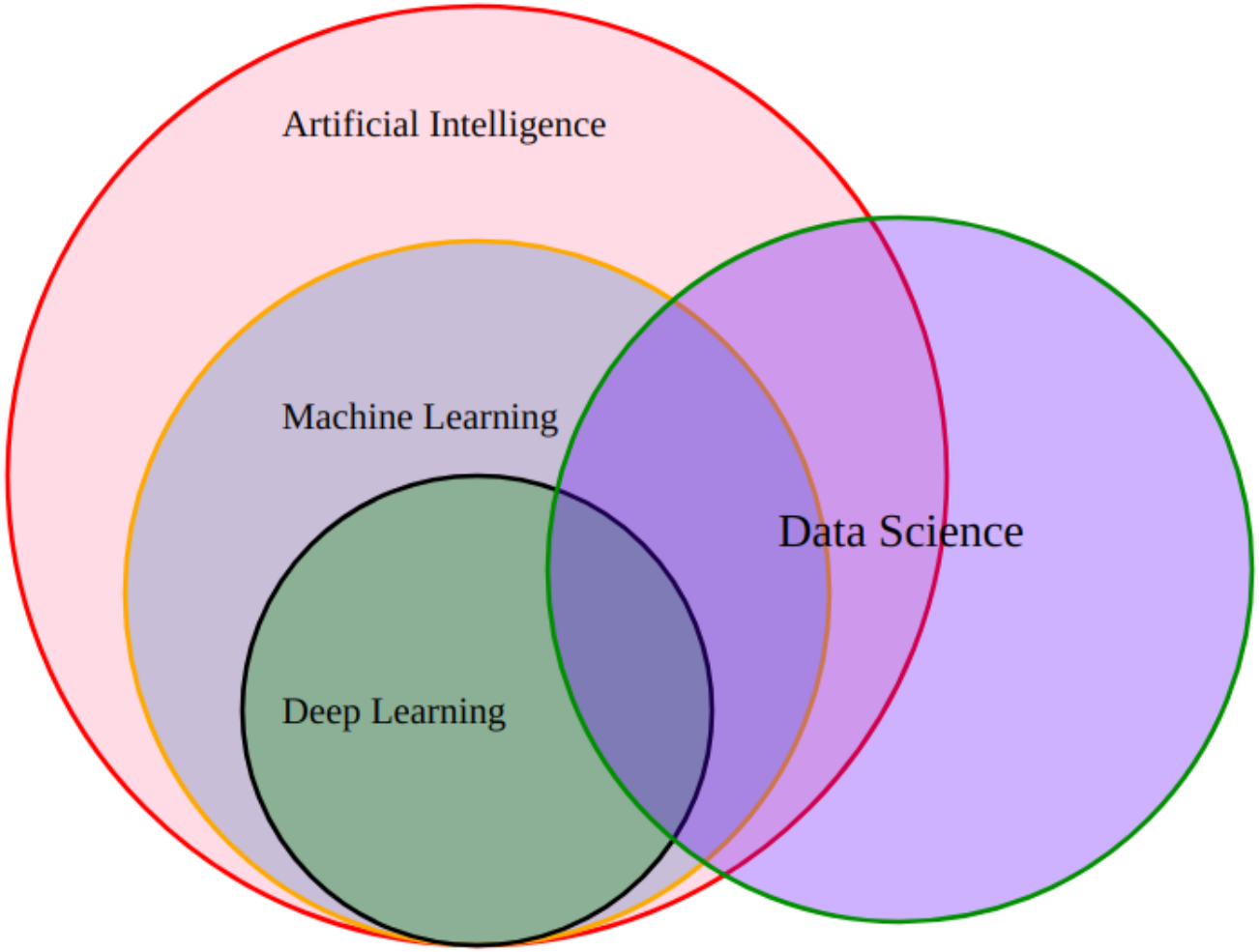
Lou Rossi, Mathematical Sciences Chairperson, UD

(astrophysicists have always worried about that)

why Data Science?



why Data Science?



what is machine learning?

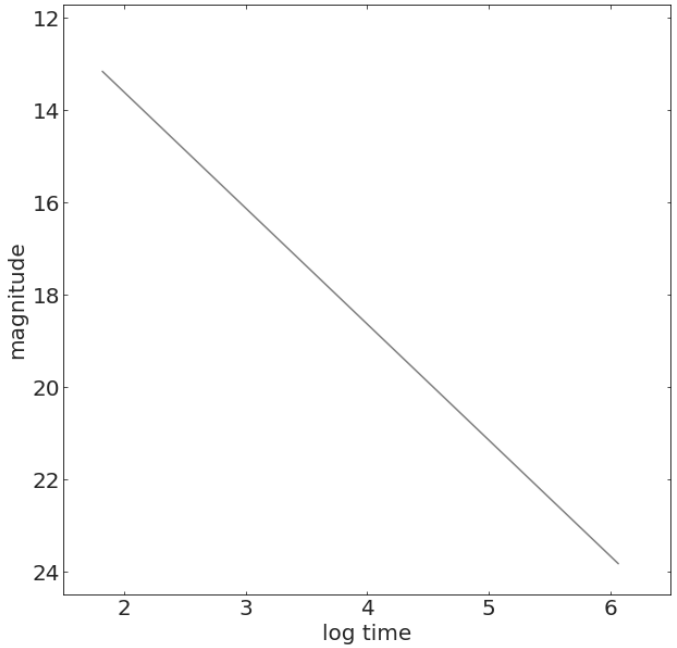
[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed.

Arthur Samuel, 1959

what is machine learning?

[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed.

Arthur Samuel, 1959



Model:

a mathematical formula
with parameters

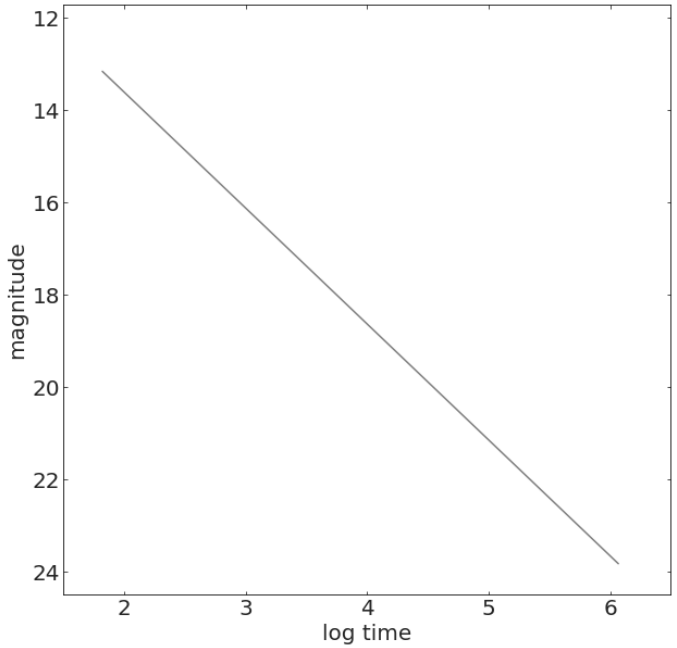
model $y = ax + b$

parameters: slope (*a*), intercept (*b*)

what is machine learning?

[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed.

Arthur Samuel, 1959



Model:

a mathematical formula
with parameters

model

parameters: slope (*a*), intercept (*b*)

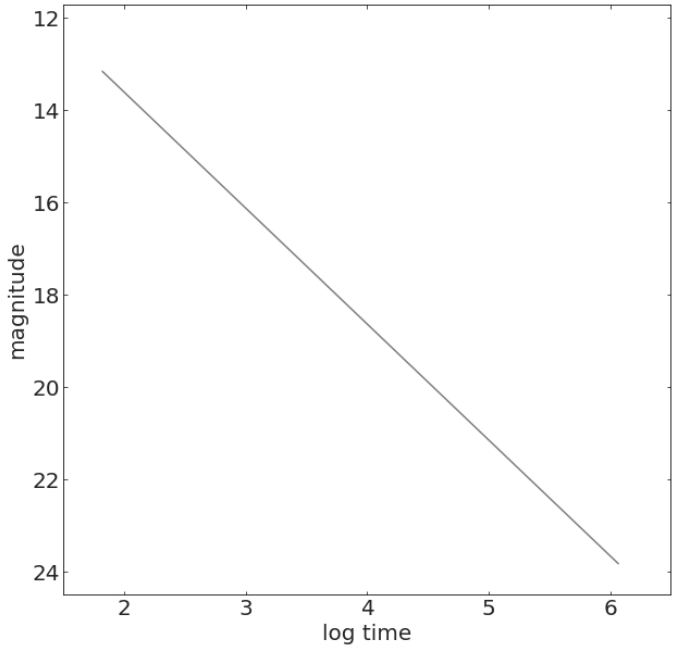
$$y = ax + b$$

model variable: x - for us this will always be time

what is machine learning?

[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed.

Arthur Samuel, 1959

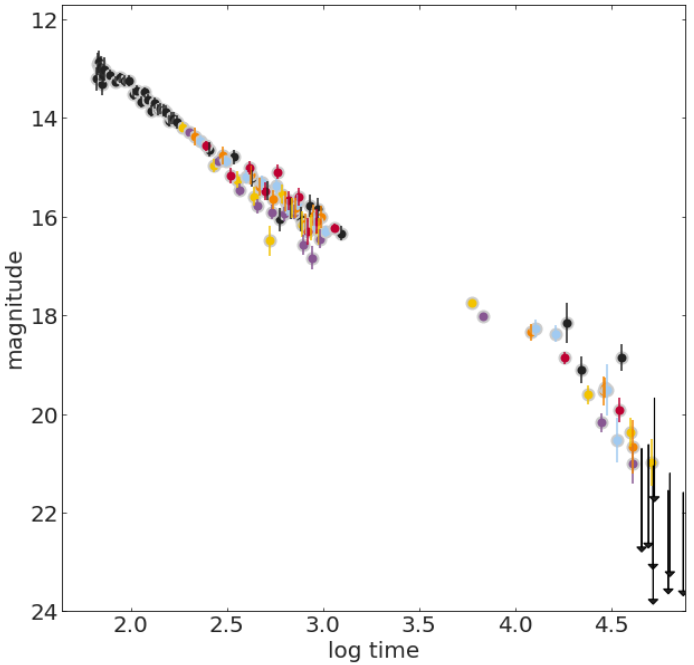


Model:

a mathematical formula with parameters

Data:

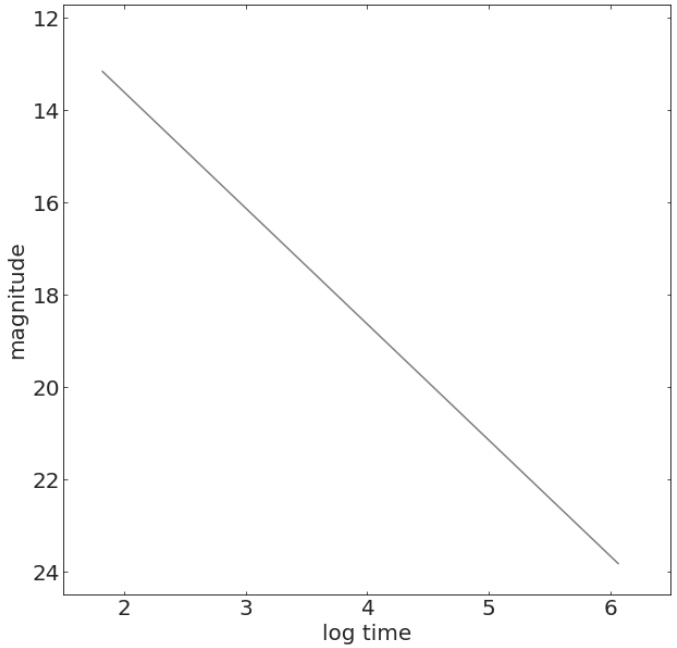
a set of observations



what is machine learning?

[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed.

Arthur Samuel, 1959

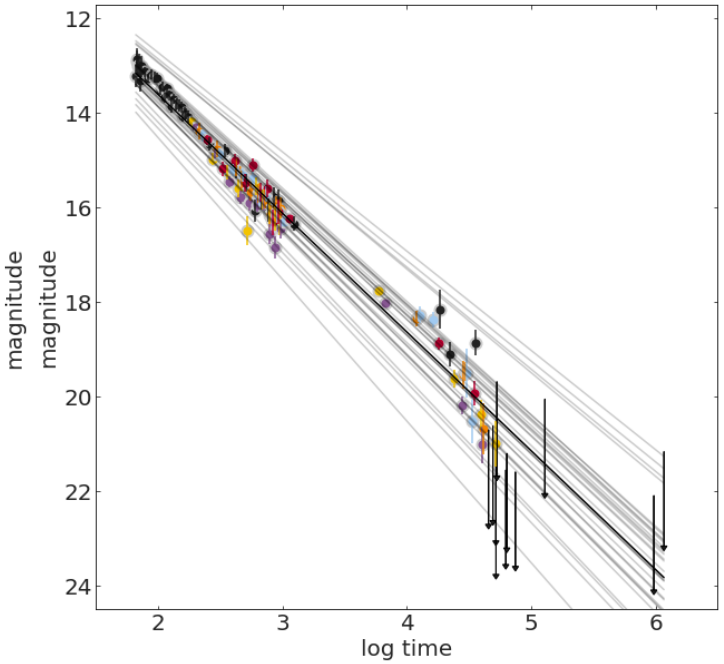


Model:

a mathematical formula with parameters

Data:

a set of observations

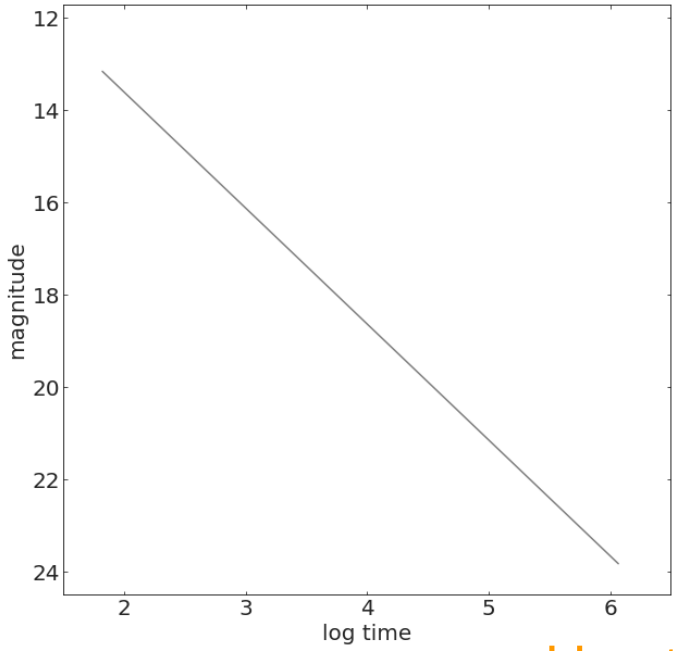


for every parameter there are an infinity of models

what is machine learning?

[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed.

Arthur Samuel, 1959

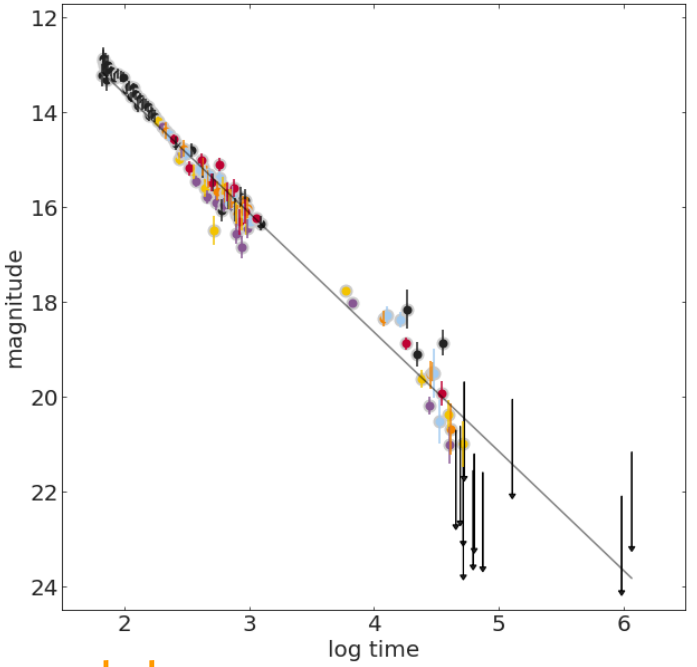


Model:

a mathematical formula with parameters

Data:

a set of observations



Use the data to *learn* the parameters of the model

what is machine learning?

Machine Learning models are parametrized representation of "reality" where the parameters are learned from finite sets of realizations of that reality

Machine Learning is the disciplines that conceptualizes, studies, and applies those models.

Key Concept

bit.ly/MLPNS23_intro

Week 1: Probability and statistics (stats for hackers)
Week 2: linear regression - uncertainties
Week 3: unsupervised learning - clustering
Week 4: kNN | CART (trees)
Week 5: Neural Networks - basics
Week 6: CNNs
Week 7: Autoencoders
Week 8: Physically motivated NN | Transformers

Tuesday: "theory"
Thursday: "hands on work"
Friday: "recap and preview"

Somewhere I will also cover:
notes on visualizations
notes on data ethics

slidocom
#2492 113

0 some administrative stuff

Syllabus

https://bit.ly/MLPNS23_syllabus

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

https://bit.ly/MLPNS23_syllabus

- 10% pre-class questions
- 10% class participation
- 25% midterm
- 15% final written
- 50% final interview

quiz

https://bit.ly/MLPNS23_syllabus

pre-class questions

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

questions on previous class material and reading assignments

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 participation**
- **25% midterm**
- **15% final written**
- **50% final interview**

Midterm... probably in class

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

advantages: interviews for jobs are often timed

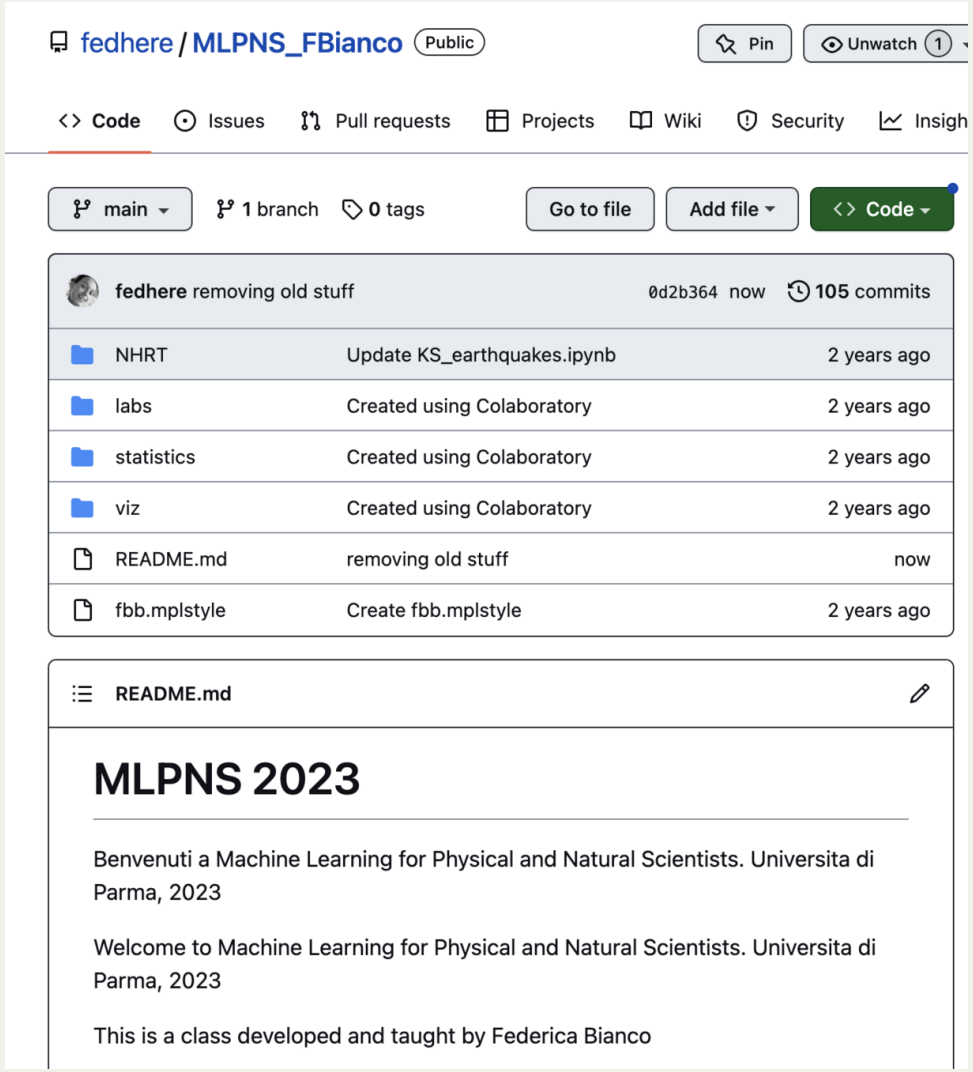
final

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, 3 days, 30 min "interview" after the last day

Resources



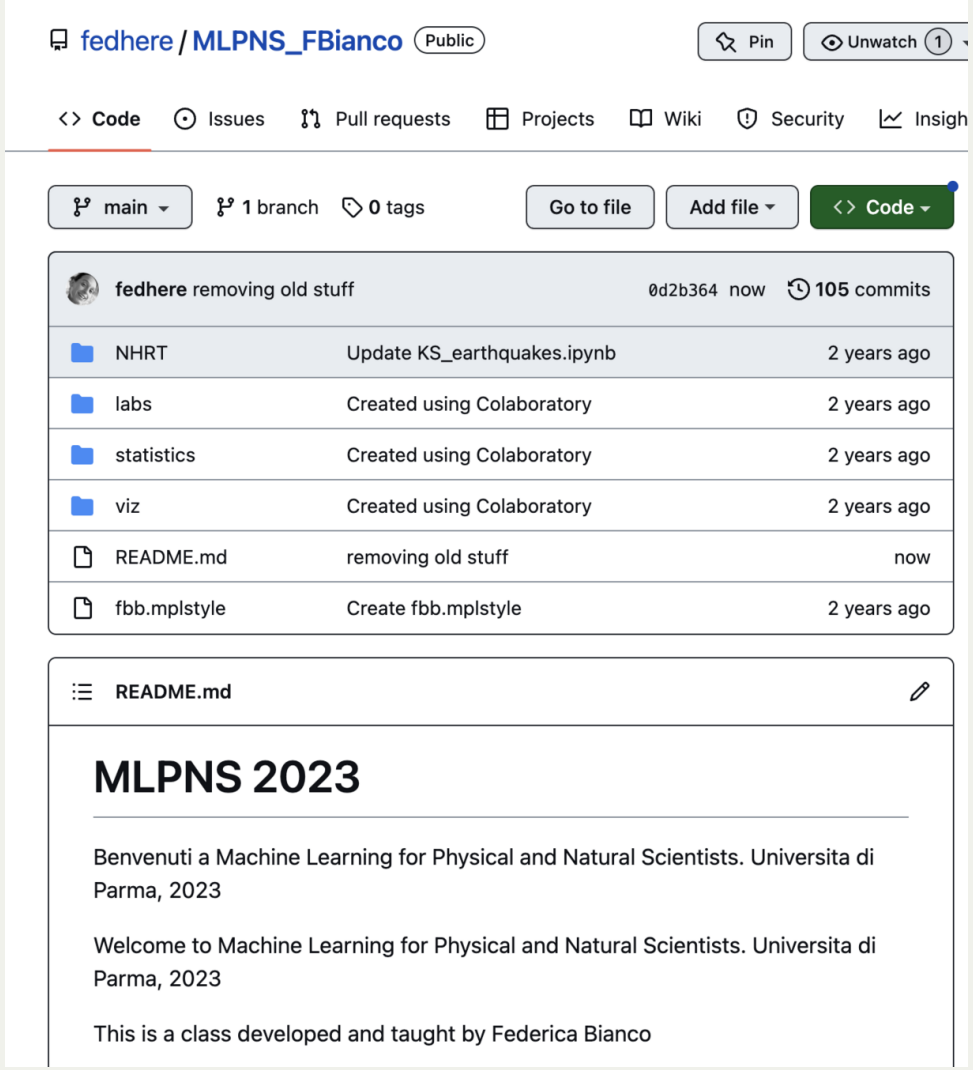
https://github.com/fedhere/MLPNS_FBianco

Resources

- SLIDES here in PDF form
- EXERCISES INSTRUCTIONS here
- RESOURCES here

If notebooks do not display use

https://nbviewer.jupyter.org



https://github.com/fedhere/MLPNS_FBianco

Resources

Resources

The primary textbooks are:

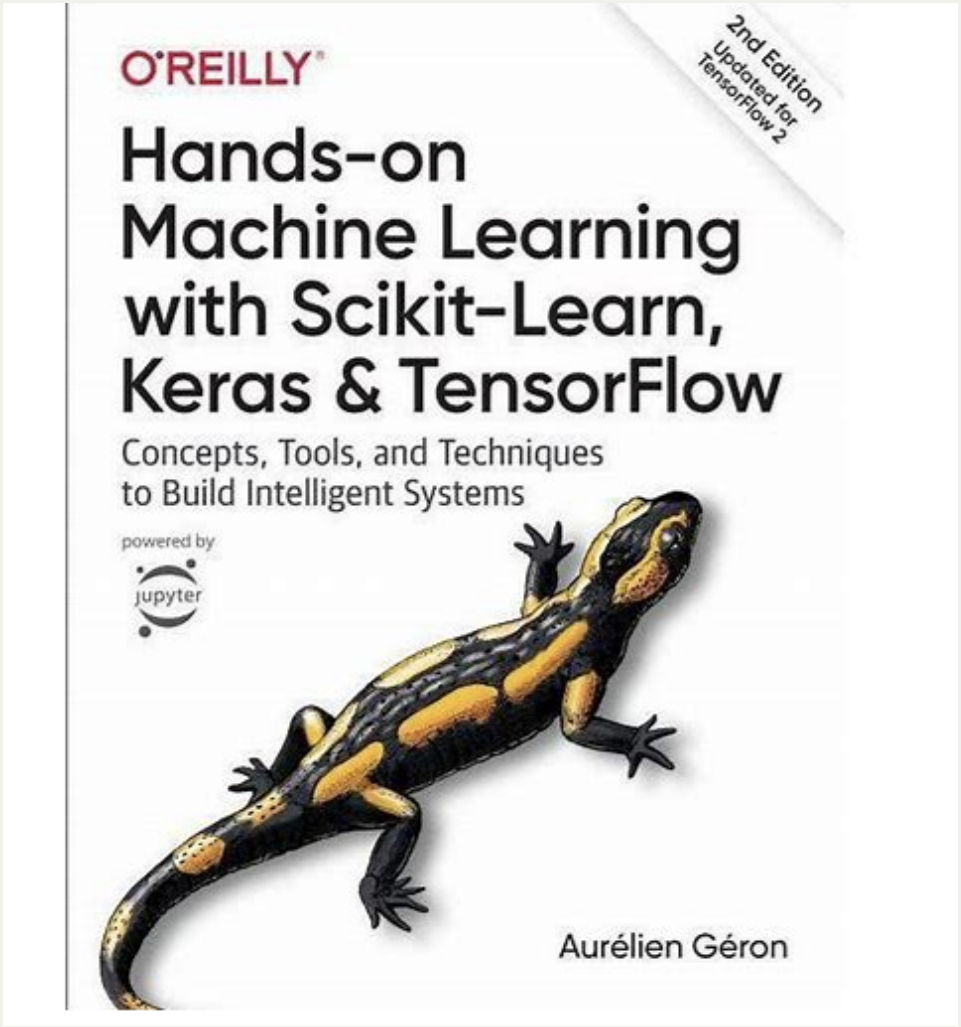
- **Elements of Statistical Learning,** Hastie,Tibshirani,Friedman, Springer 2001
- **Python Data Science Handbook,** Jake VanderPlas, O'Reilly Media [https://www.oreilly.com/library/view/python-data-science/9781491912126/]
- **Statistics, Data Mining, and Machine Learning in Astronomy,** Ivezic, Connoly,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



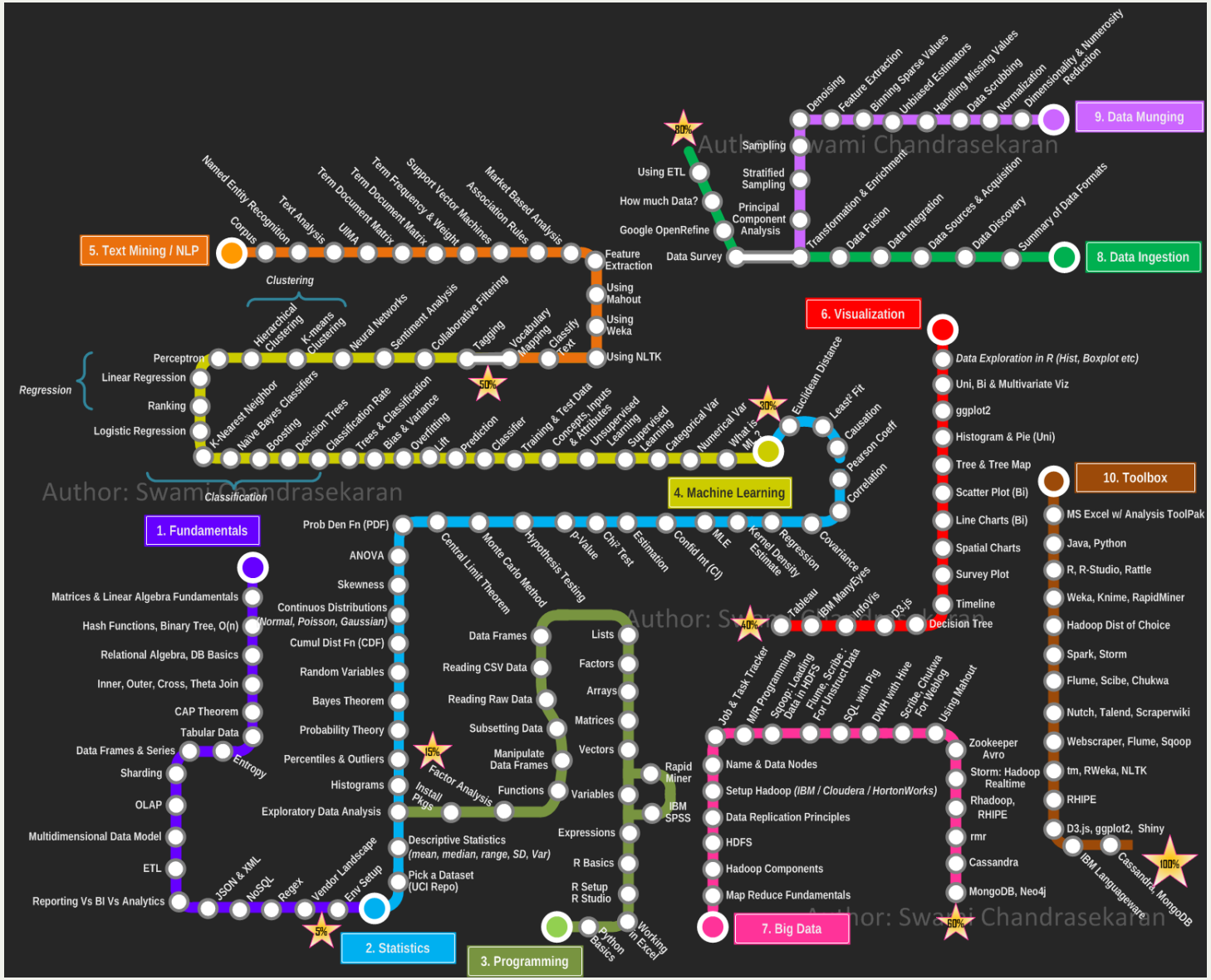
Resources

http://vanderplas.com/

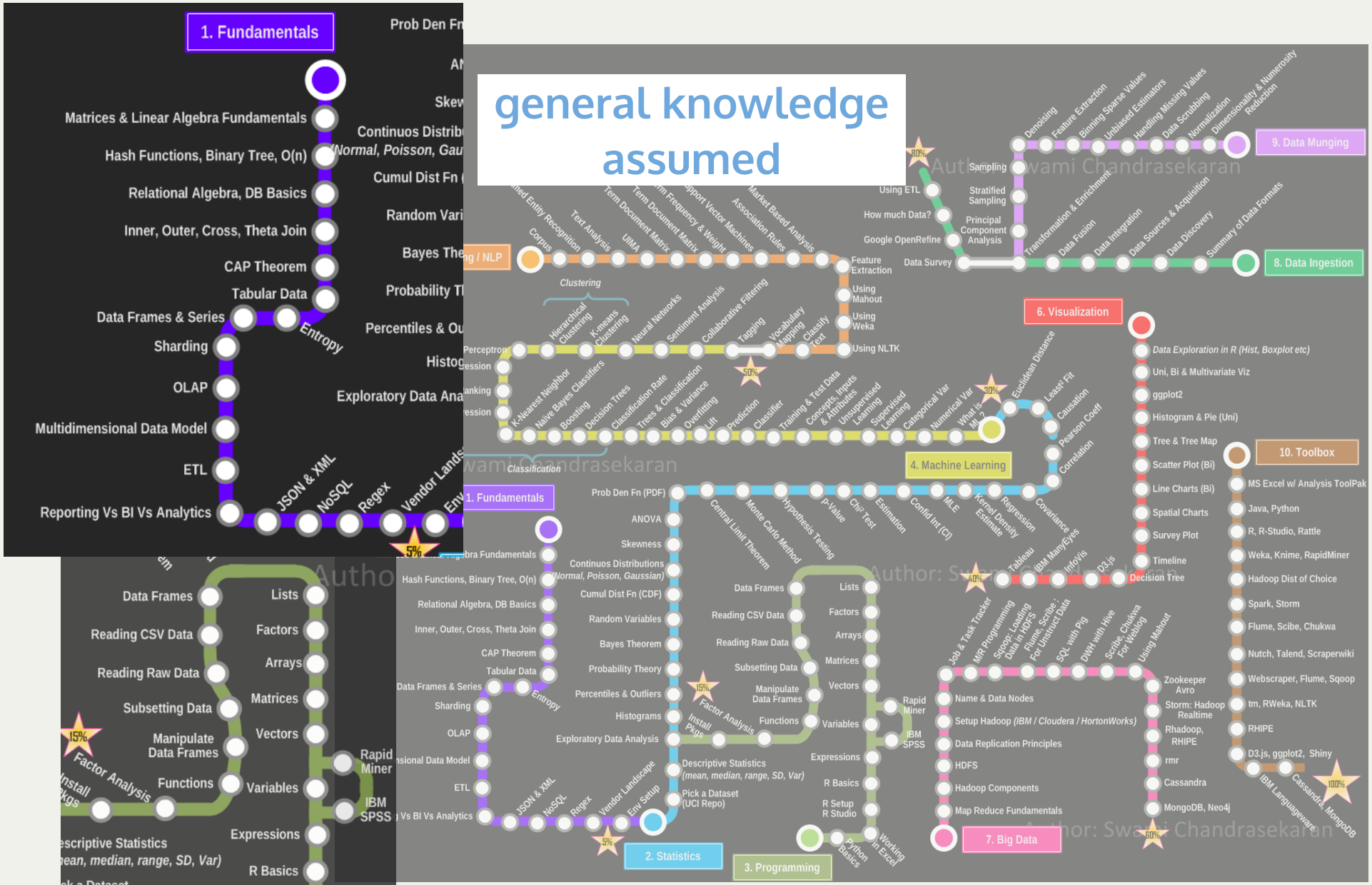
Jake Vanderplas is a physicst-data scientists

https://www.youtube.com/watch?v=ZyjCqQEUa8o

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

most of this will be neglected

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

PROGRAMMING

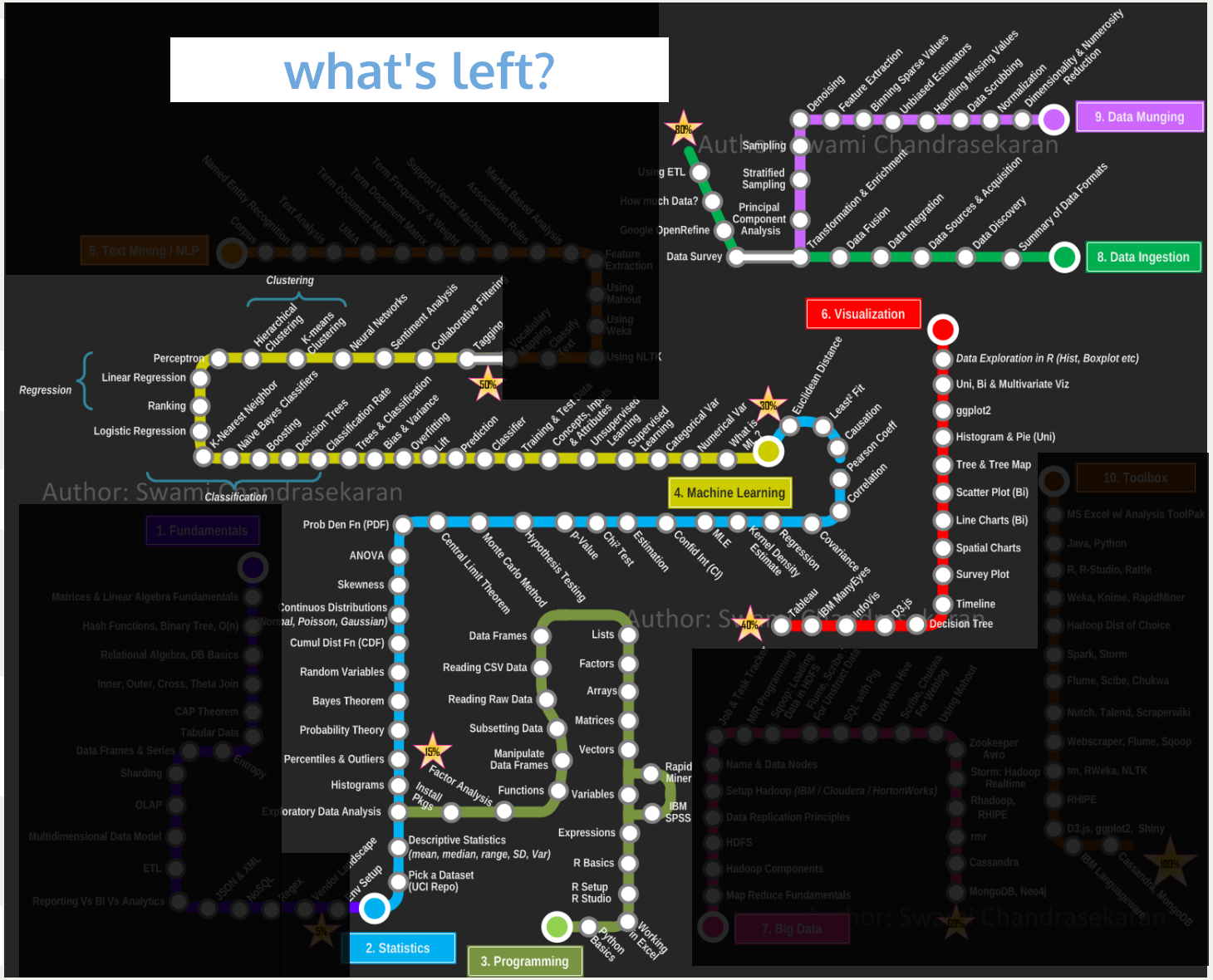
STATISTICS

DATA INGESTION

DATA MUNGING

MACHINE LEARNING

VISUALIZATION



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

PROGRAMMING

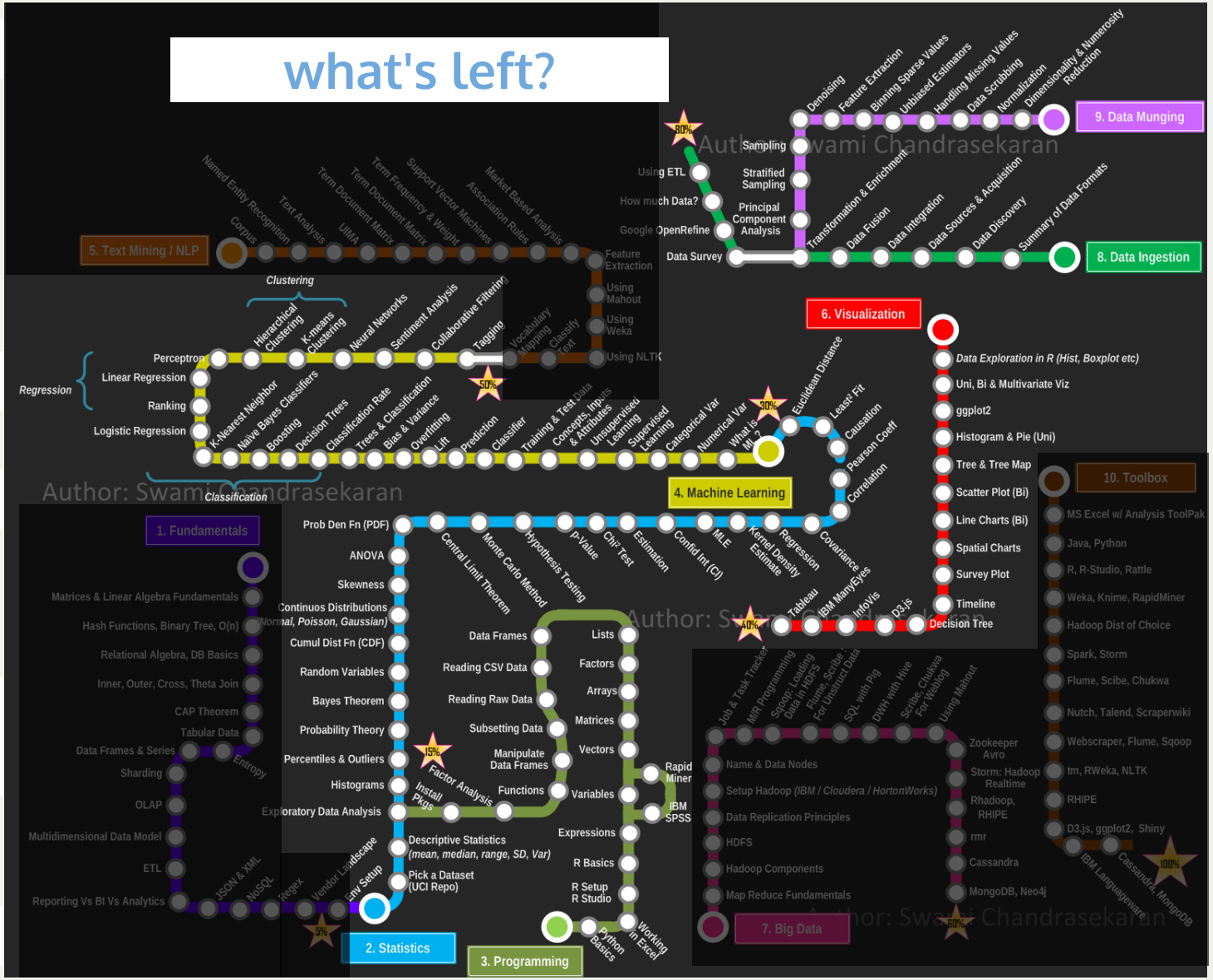
STATISTICS

DATA INGESTION

DATA MUNGING

MACHINE LEARNING

VISUALIZATION



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

python

PROGRAMMING

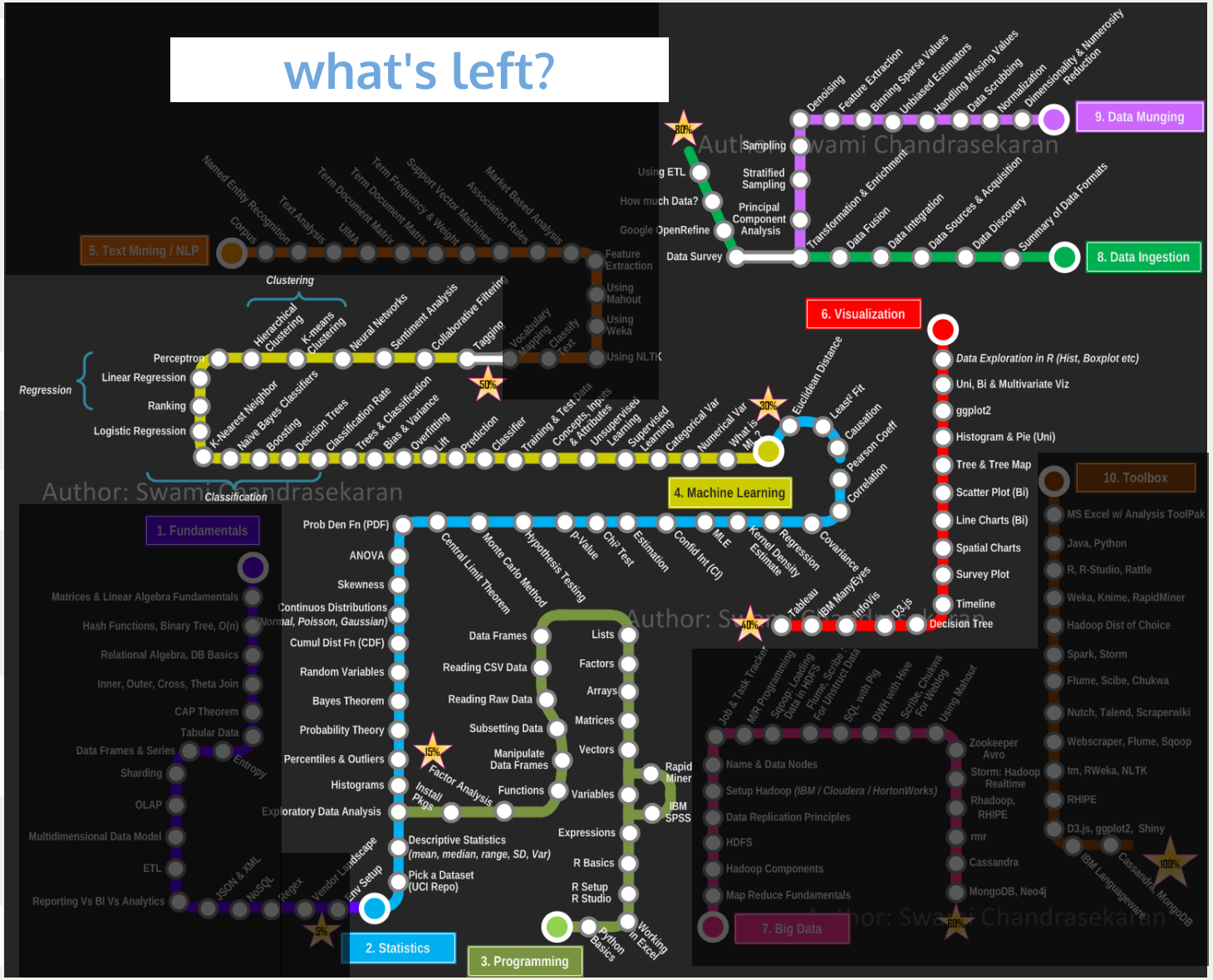
STATISTICS

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

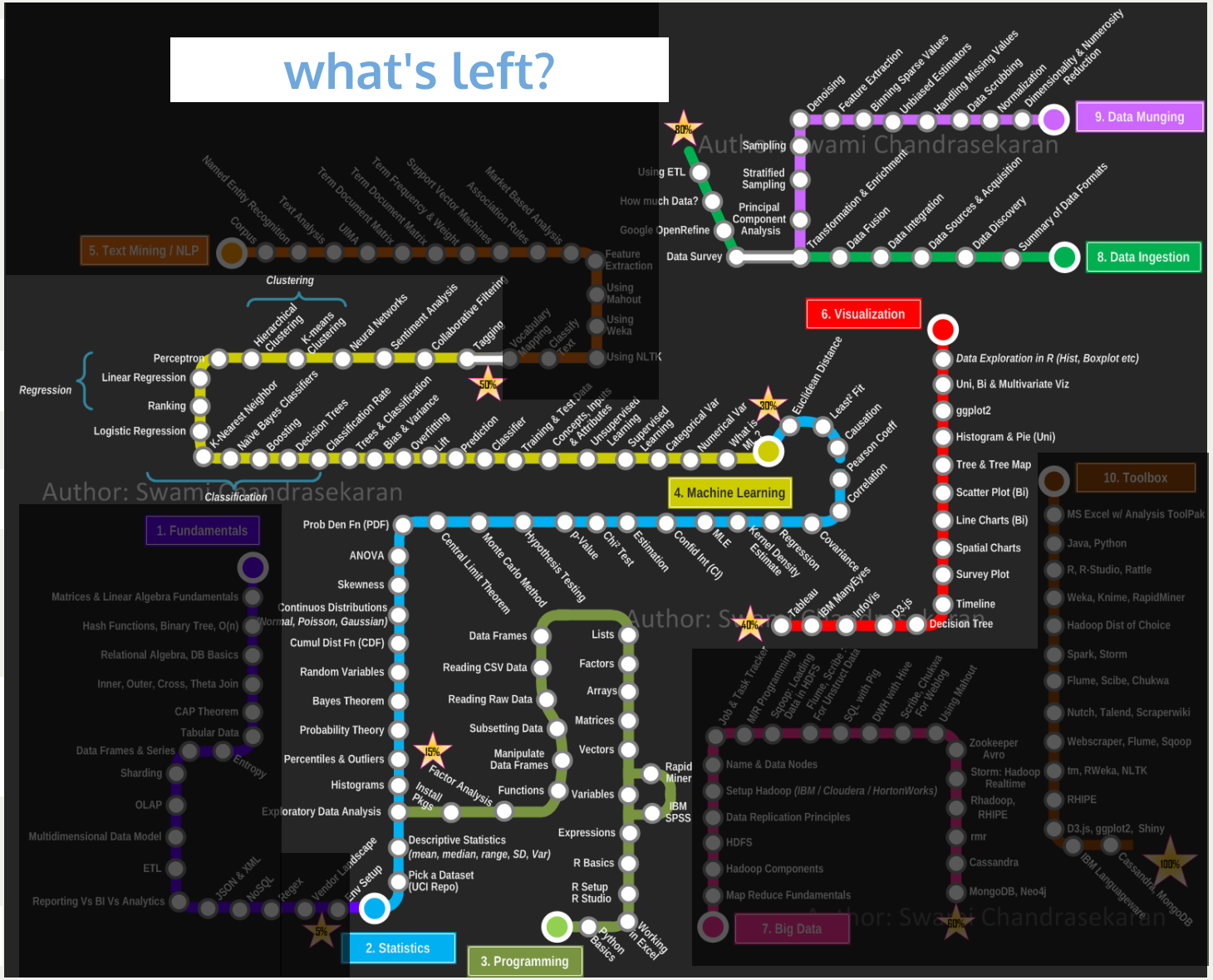
STATISTICS

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

Time series analysis

Geospacial analysis?

2 the scientific method (what is science?)

epistemology:

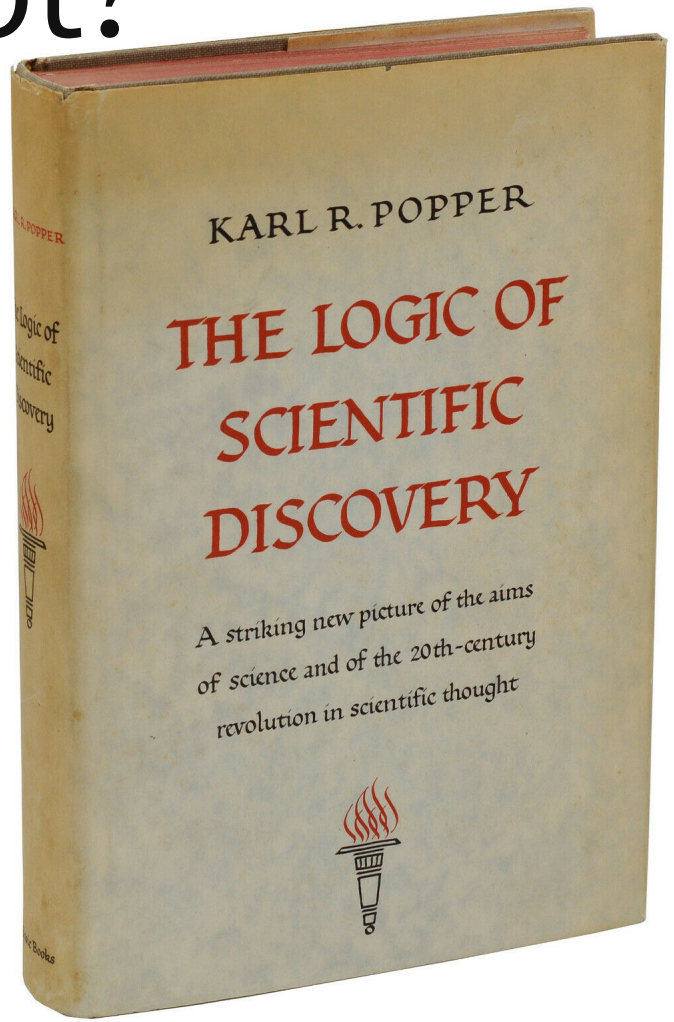
the philosophy of science and of the scientific method

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*



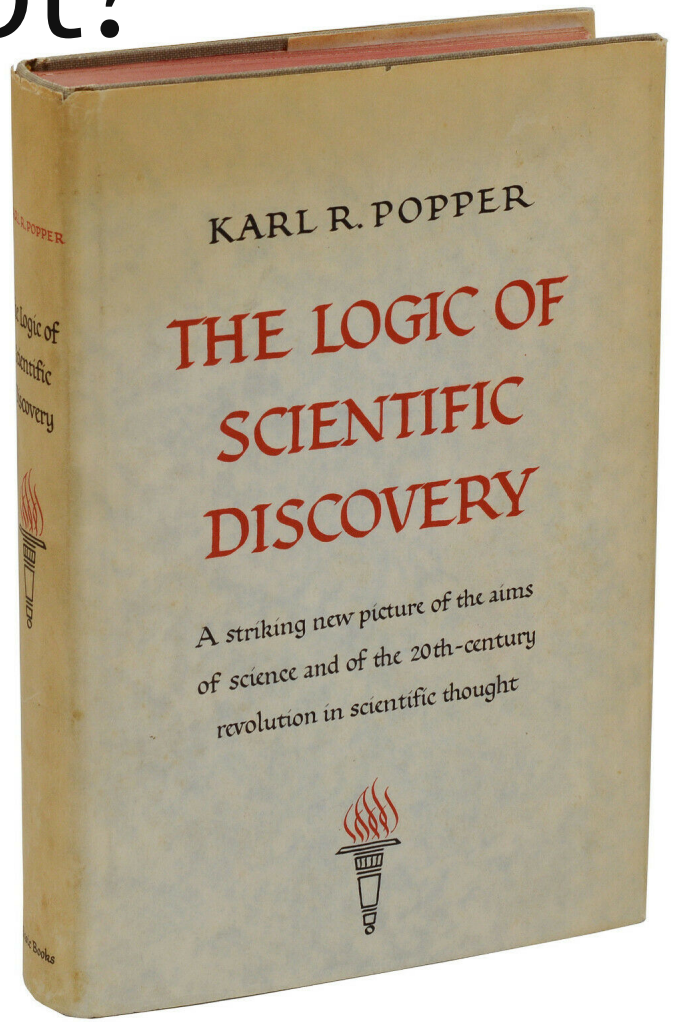
the *demarcation* problem:

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

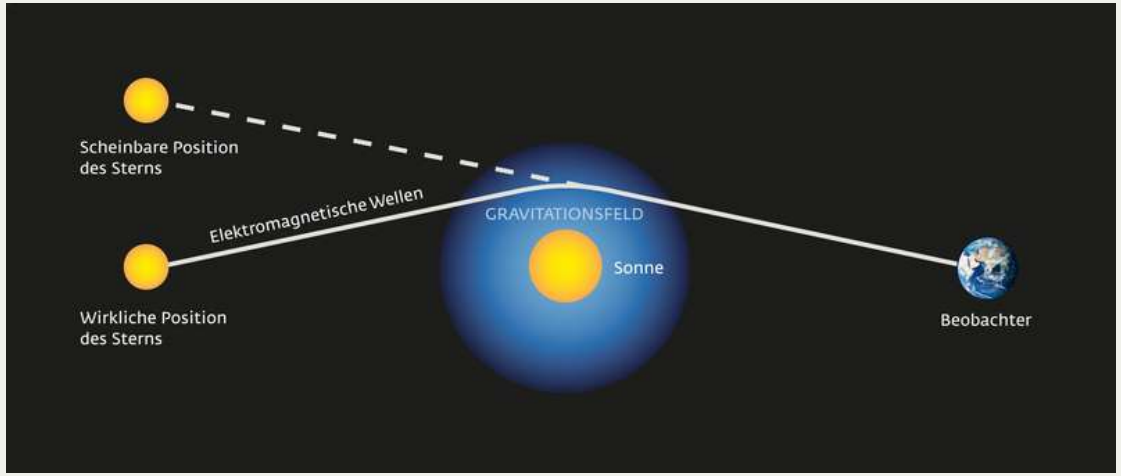
the *demarcation* problem

model → **prediction**

Einstein GR

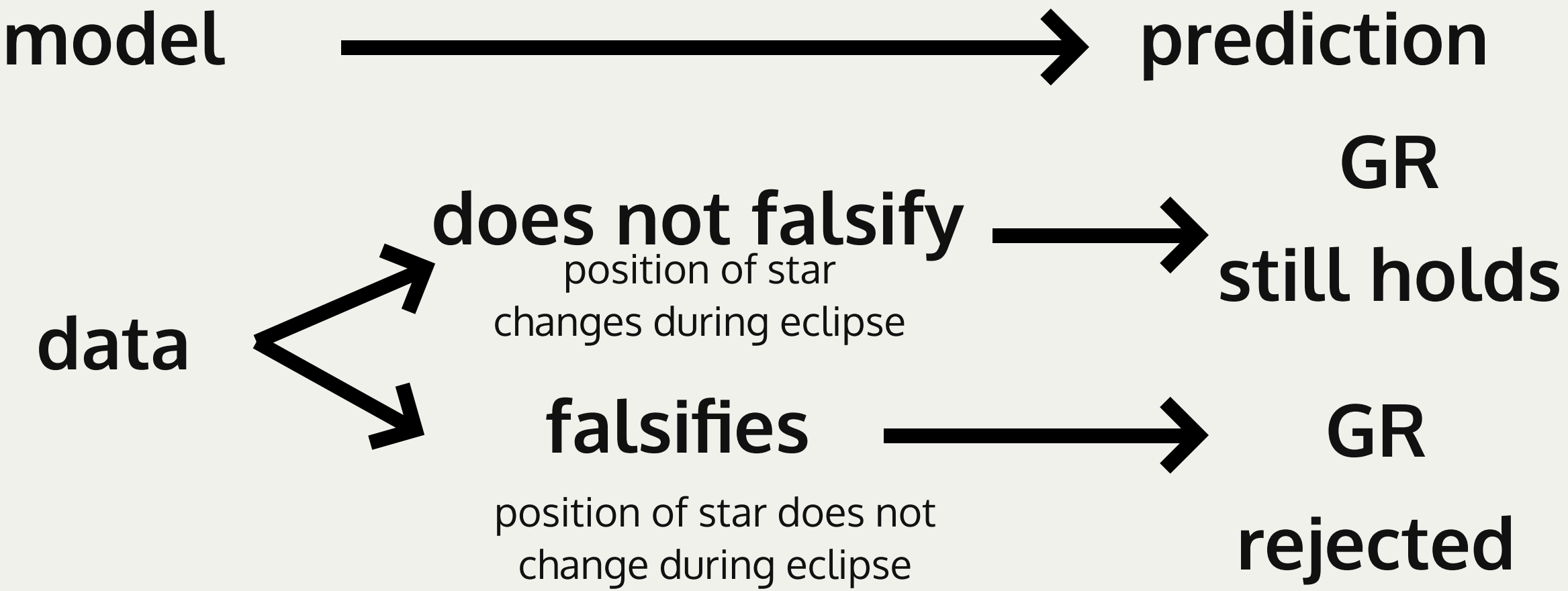
$$R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} = 8\pi G T_{\mu\nu}$$

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

A theory can be said to be scientific if it makes falsifiable predictions

Experiments should be designed to falsify the predictions

Key Concept

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...)

everything has **some** probability of happening. But it might be very small

traditional statistics works as follows:

- if the probability is smaller than some arbitrary cut (e..g p~0.05) then I will say that it is not true

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...)

Text

everything has **some** probability of happening. But it might be very small

traditional statistics works as follows:

- if the probability is smaller than some arbitrary cut (e..g p~0.05) then I will say that it is not true

- what about ML?? all it does is (1) make predictions (2) find structure in data

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

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.

- 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

Reproducibility

A research product is reproducible if all numbers can be reproduced exactly be applying the same code to the same raw data.

It is the responsibility of the researcher to provide the data and code that make a research product reproducible

Key Concept

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

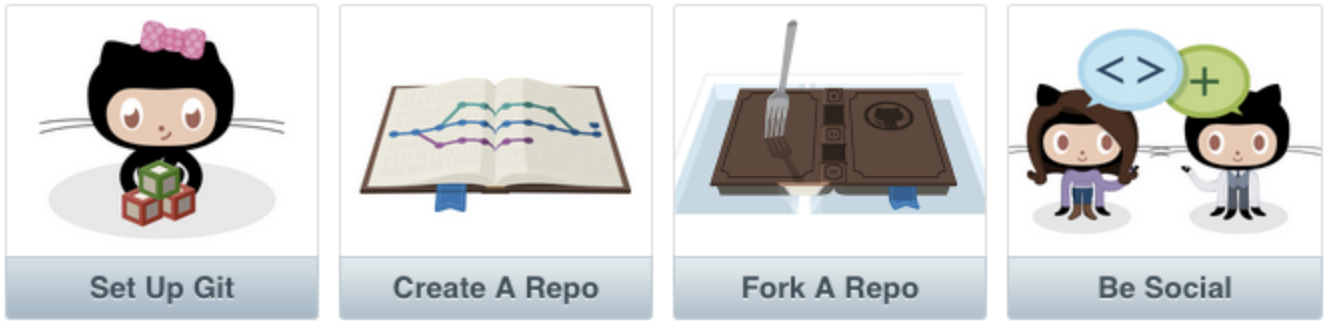
github

collaborative platform

https://github.com

collaboration tool

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

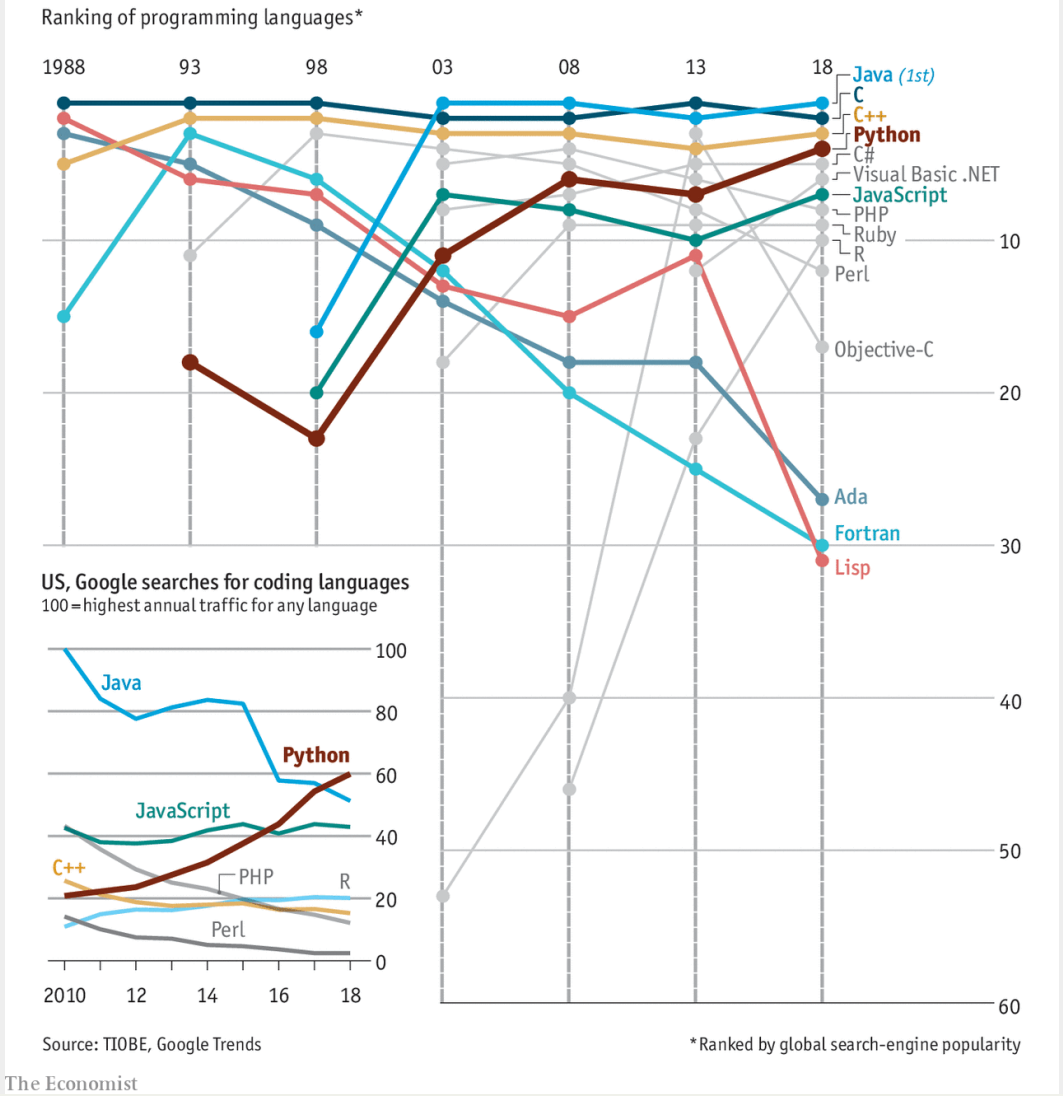


allows effective collaboration

python

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

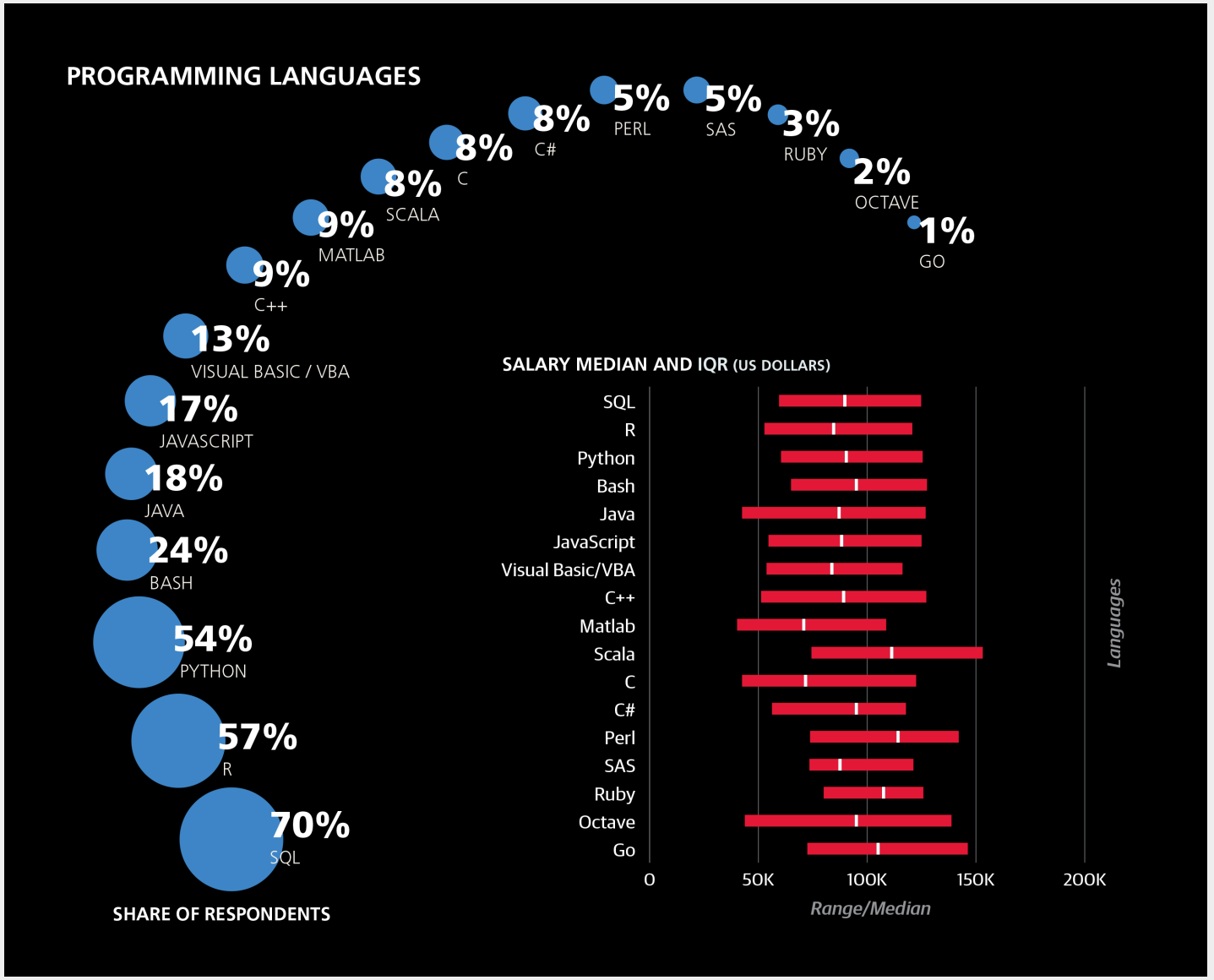
https://www.economist.com/graphic-detail/2018/07/26/python-is-becoming-the-worlds-most-popular-coding-language



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/courses/ucsl/

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

recommended if you are brand new to python and coding or are serious about cleaning up your fundamentals

ignore the references to the CUSP working environment and work on https://colab.research.google.com/notebooks instead

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

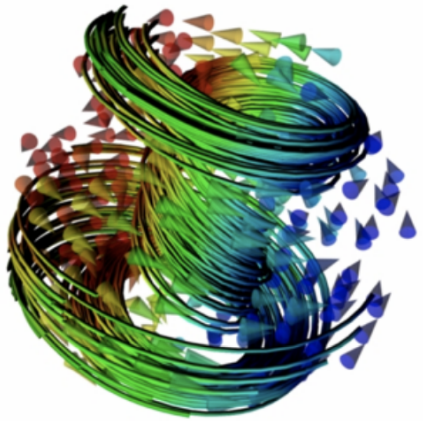
online book

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*

python

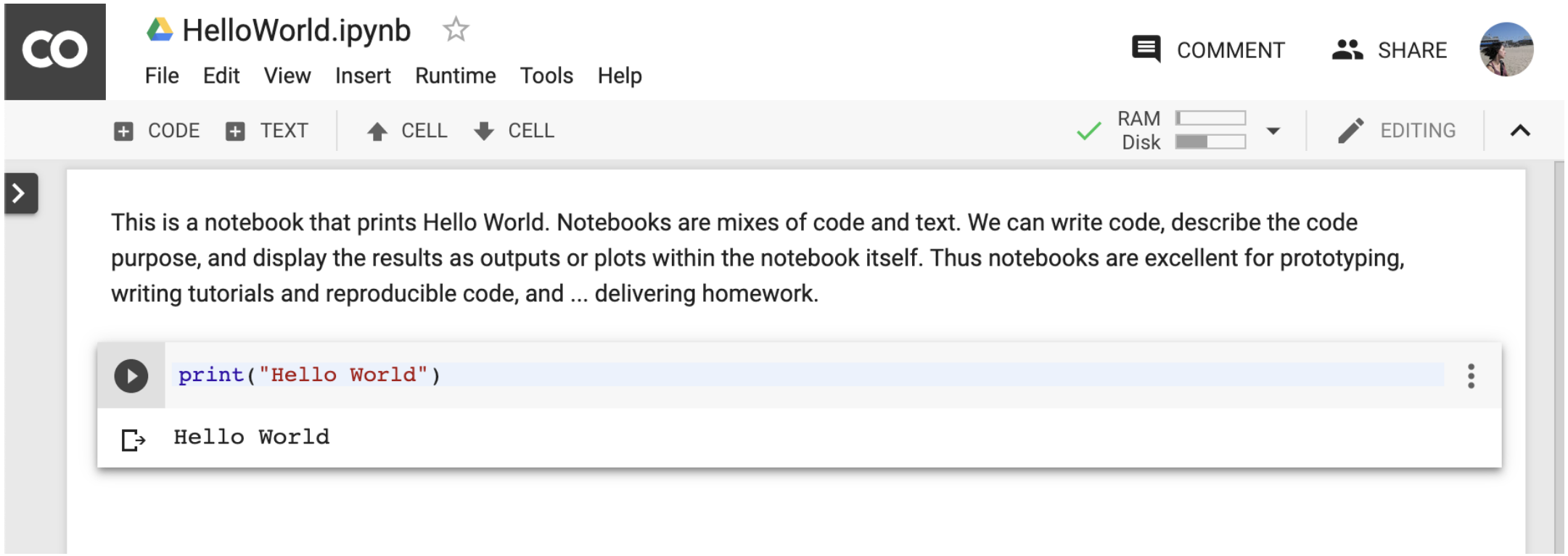
https://github.com/fedhere/pyboot

A Python Bootcamp

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

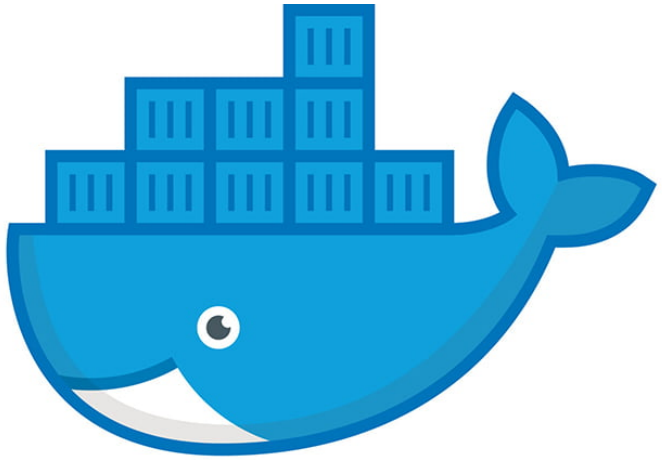
Jupyter Notebook
Google Colaboratory

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 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]}
```

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 MLPNS_<firstinitialLastname>
- use the form to confirm you read the Code of Conduct and deliver your repo link
- write a Readme.md file to state what this repo is for, what your motivation to take this class is, what you hope to learn. At the end of the course we will reflect on these early expectations

https://forms.gle/i7eFWCUt3YeYS9Qc7

TBD

2

homework

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
(this 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/