Data analysis in Python - come in, don't get lost

Pietro Battiston
University of Milan Bicocca

Kaunas, May 5, 2018
Pycon LT

About me

Pietro Battiston (github: **@toobaz**)

About me

Pietro Battiston (github: **@toobaz**)

- Background in maths/complexity

About me

Pietro Battiston (github: **@toobaz**)

- Background in maths/complexity
- Moved to the dark side of the force (Economics)

About me

Pietro Battiston (github: **@toobaz**)

- Background in maths/complexity
- Moved to the dark side of the force (Economics)
- Researcher on...

About me

Pietro Battiston (github: **@toobaz**)

- Background in maths/complexity
- Moved to the dark side of the force (Economics)
- Researcher on. . . data

About me

Pietro Battiston (github: **@toobaz**)

- Background in maths/complexity
- Moved to the dark side of the force (Economics)
- Researcher on. . . data
- pandas core dev

About me

Pietro Battiston (github: **@toobaz**)

- Background in maths/complexity
- Moved to the dark side of the force (Economics)
- Researcher on. . . data
- pandas core dev (in love with MultiIndex)

About me

Pietro Battiston (github: **@toobaz**)

- Background in maths/complexity
- Moved to the dark side of the force (Economics)
- Researcher on. . . data
- pandas core dev (in love with MultiIndex)

- love music, mountains

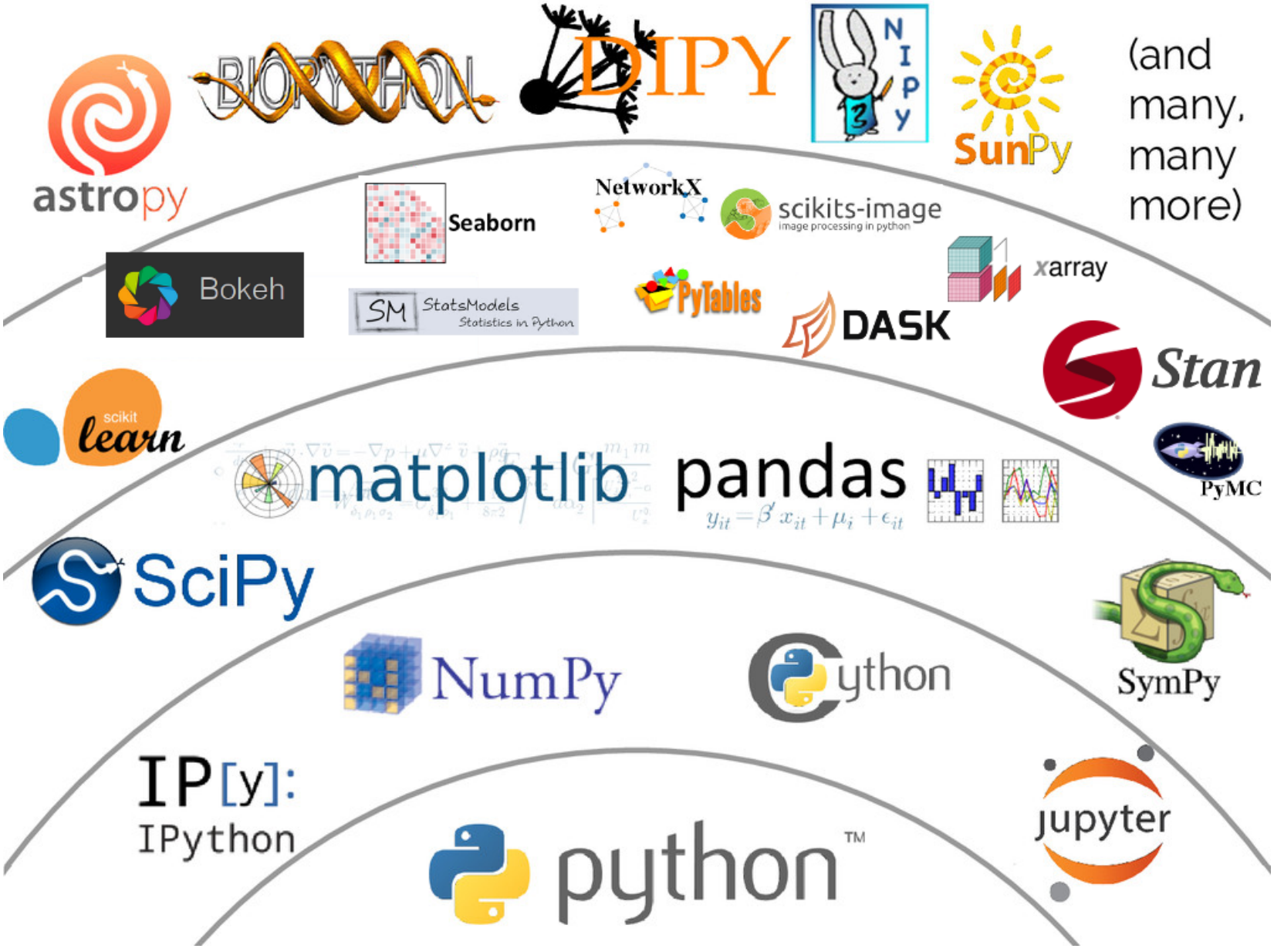
About me

Pietro Battiston (github: **@toobaz**)

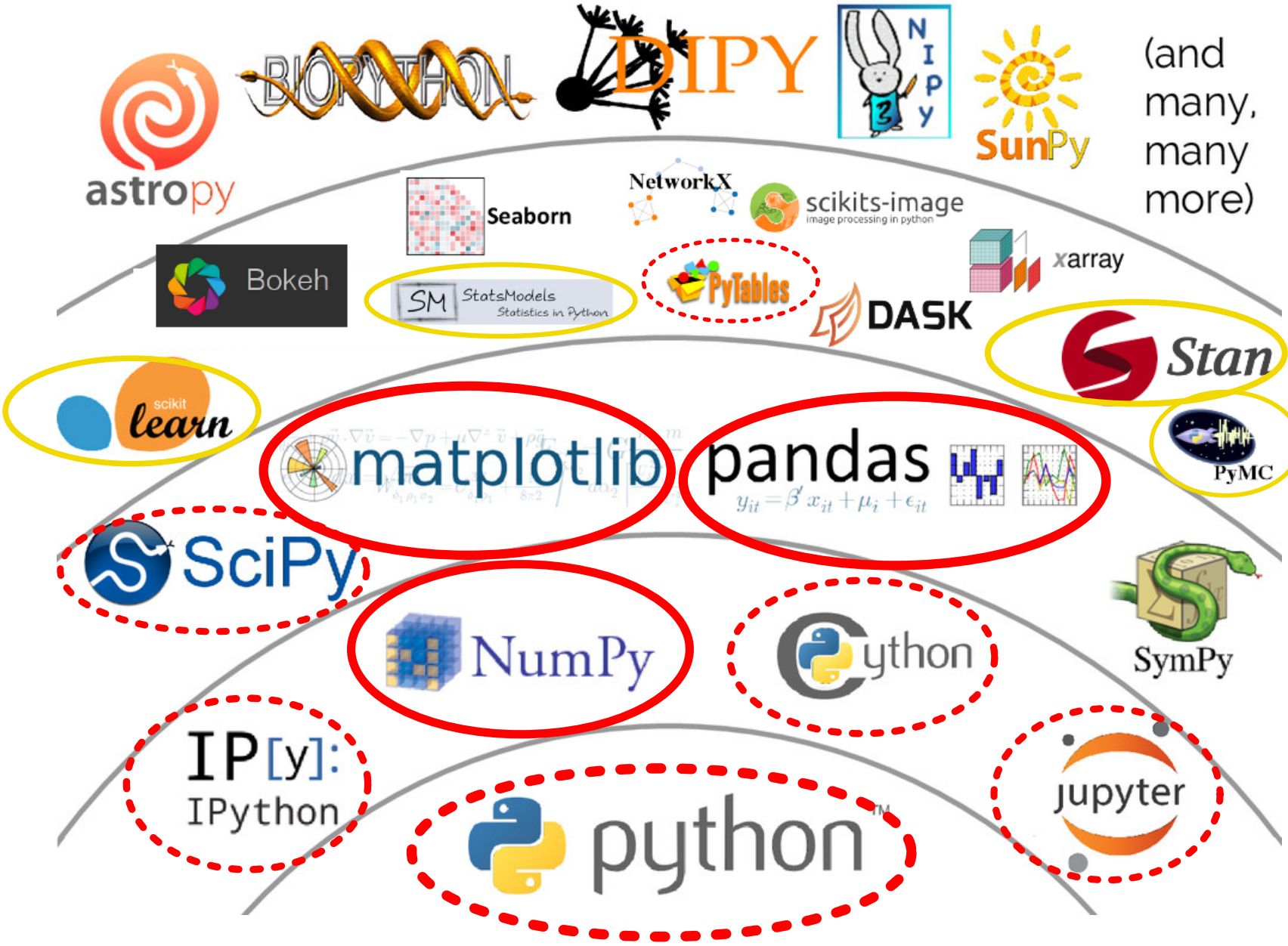
- Background in maths/complexity
- Moved to the dark side of the force (Economics)
- Researcher on. . . data
- pandas core dev (in love with MultiIndex)

- love music, mountains, MultiIndex

A map


astropy
BIOPYTHON
DIPY
NIPY
SunPy
(and many, many more)
Seaborn
NetworkX
scikits-image
image processing in python
Bokeh
StatsModels
Statistics in Python
PyTables
xarray
DASK
Stan
scikit learn
matplotlib
pandas
PyMC
SciPy
NumPy
Cython
SymPy
IP[y]:
IPython
python
jupyter

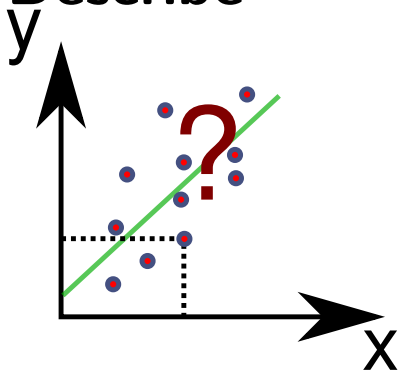
Let's get oriented



What can we do with data?

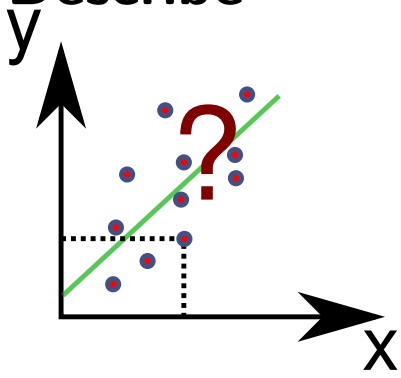
What can we do with data?

- **Describe**

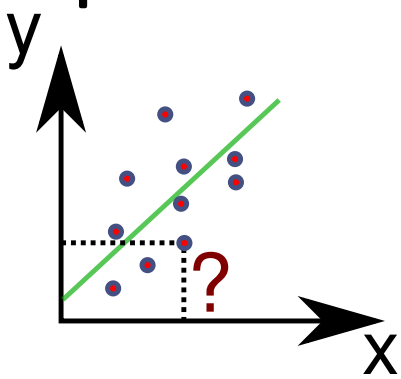


What can we do with data?

- **Describe**

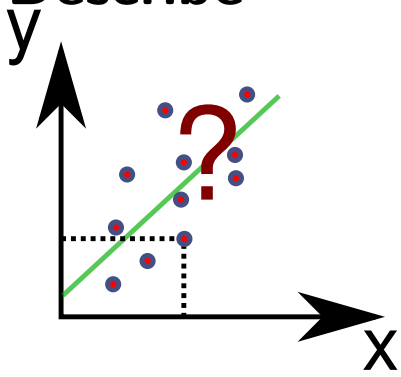


- **Explain**

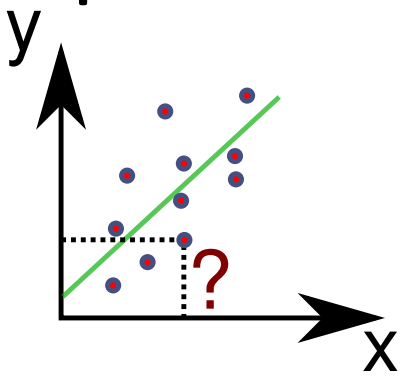


What can we do with data?

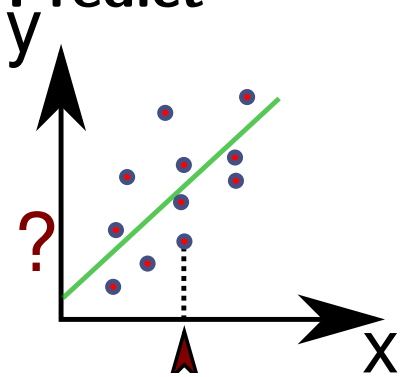
- **Describe**



- **Explain**



- **Predict**



First: load the data

pandas - library for handling *labeled* data

First: load the data

pandas - library for handling *labeled* data

~~Copied~~ inspired from **R**

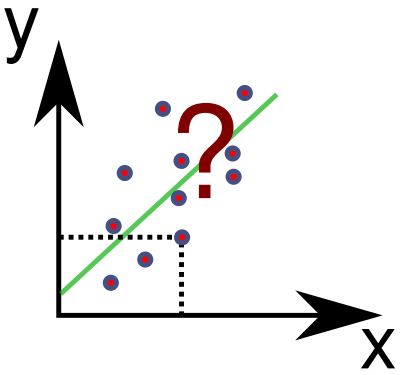
First: load the data

pandas - library for handling *labeled* data

~~Copied~~ inspired from **R**

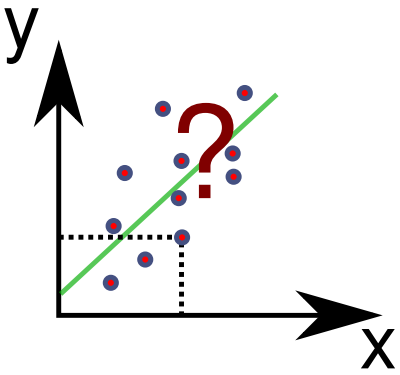
Code!

Describe



matplotlib - librery for data *visualization*

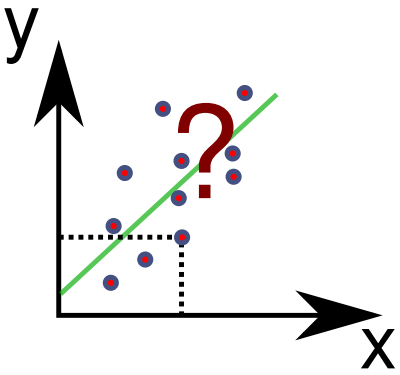
Describe



matplotlib - librery for data *visualization*

bokeh - on the web, **seaborn** - higher level

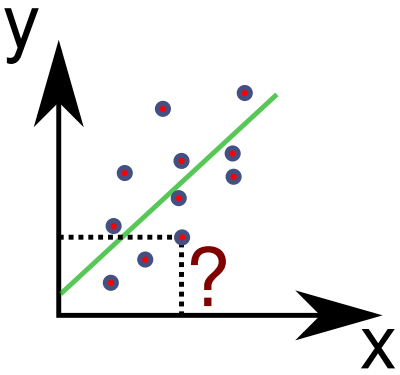
Describe



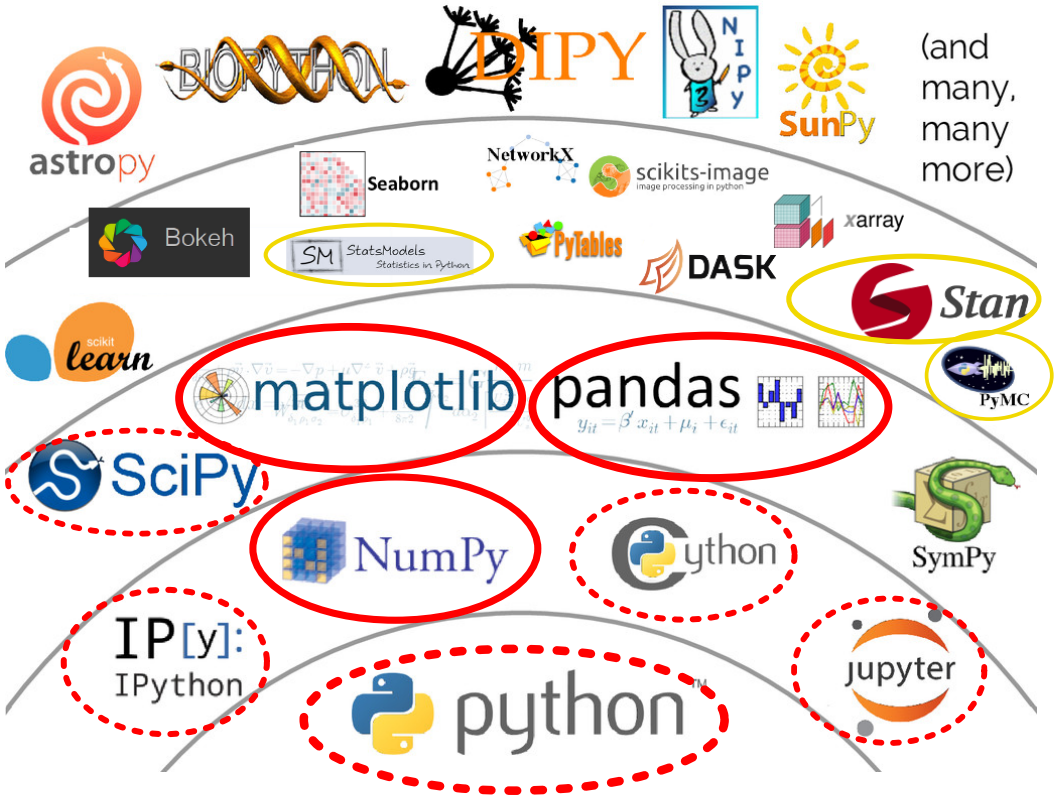
matplotlib - librery for data *visualization*

bokeh - on the web, **seaborn** - higher level **Code!**

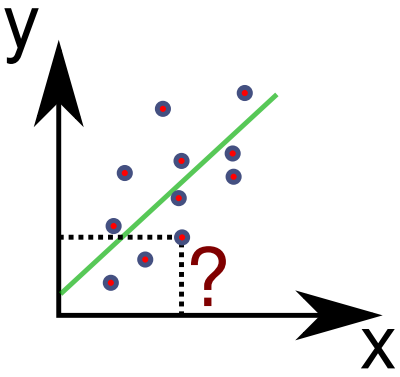
Explain



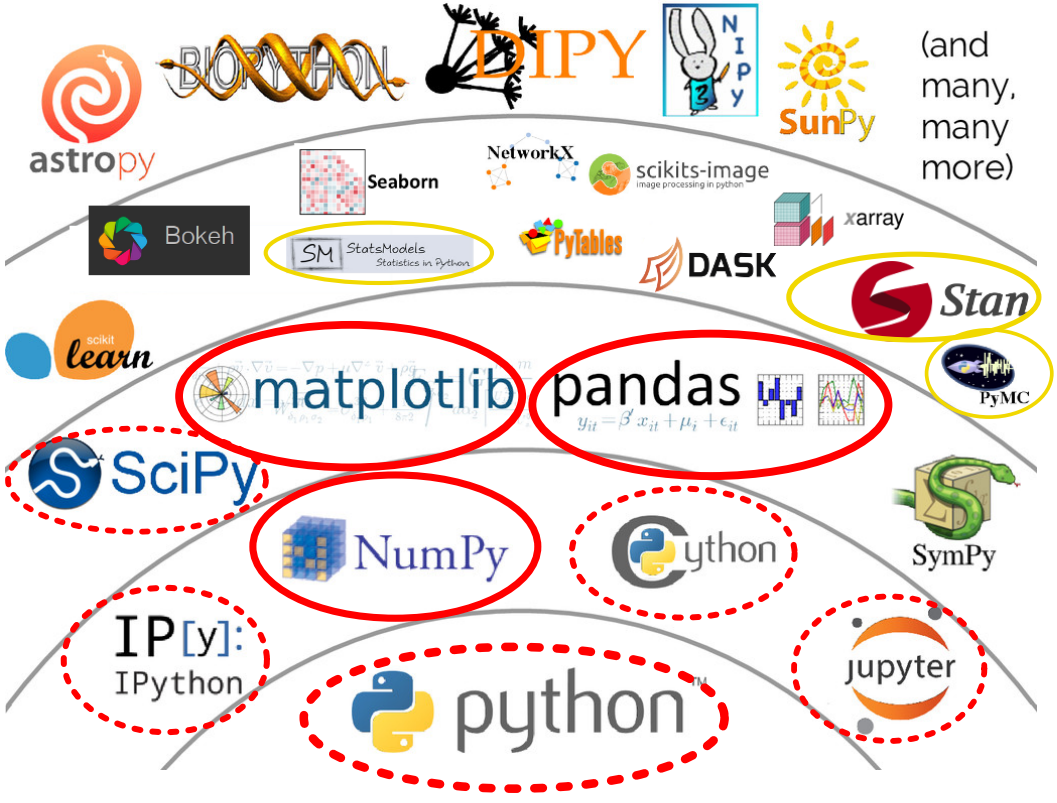
statsmodels - library for *classical* statistics



Explain

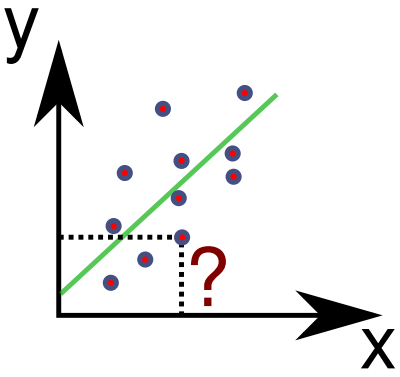


statsmodels - library for *classical* statistics

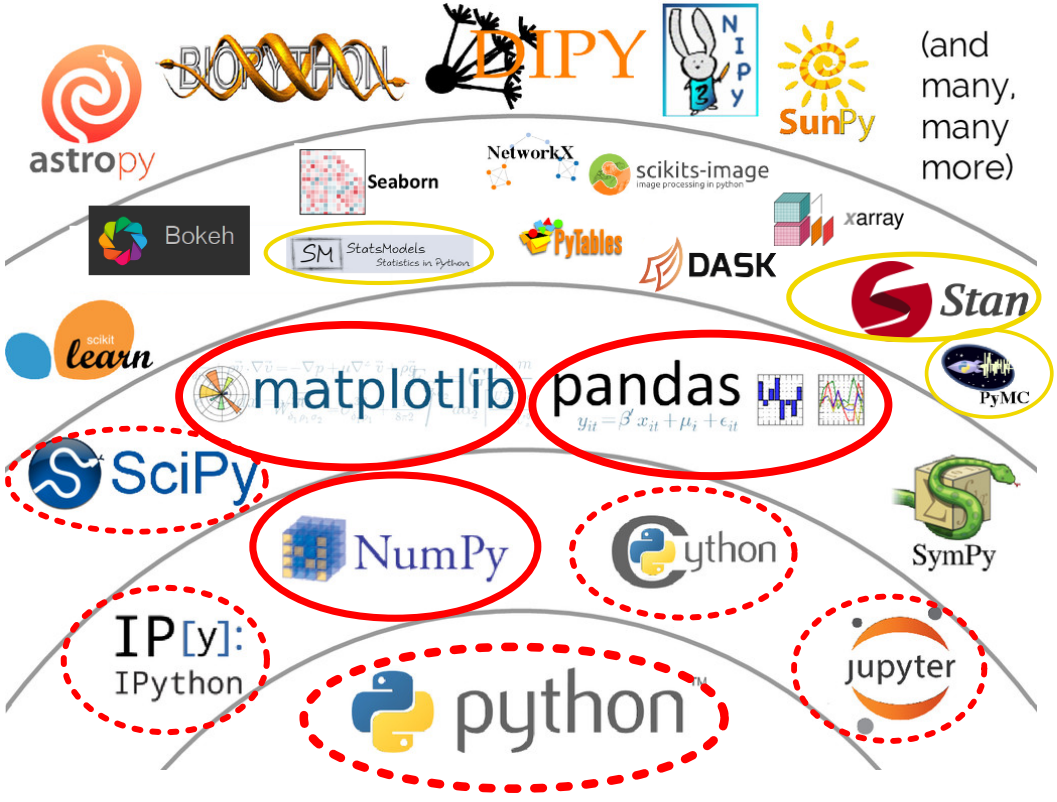


pymc2, **pystan** - libraries for *Bayesian* statistics

Explain

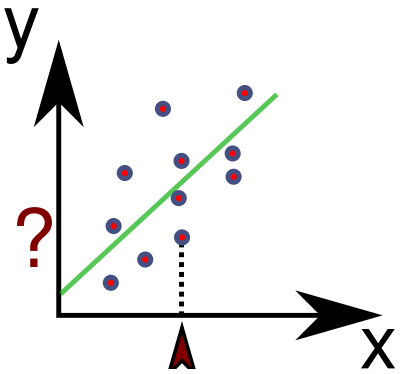


statsmodels - library for *classical* statistics

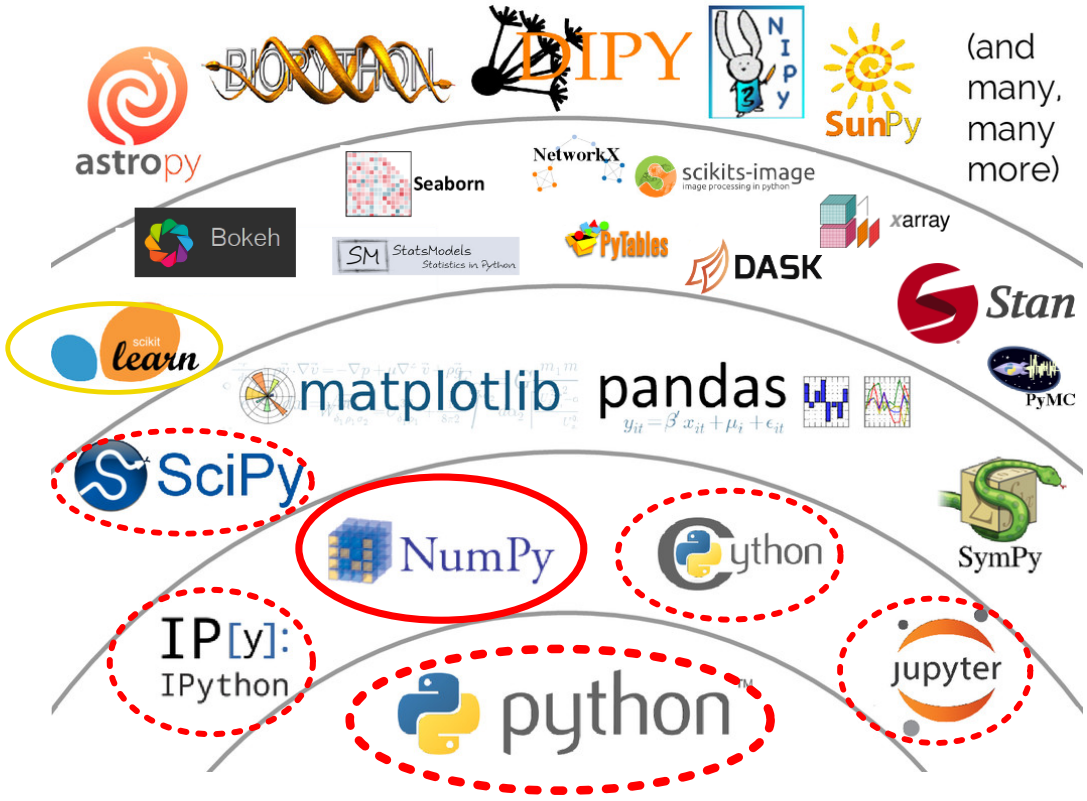


pymc2, **pystan** - libraries for *Bayesian* statistics **Code!**

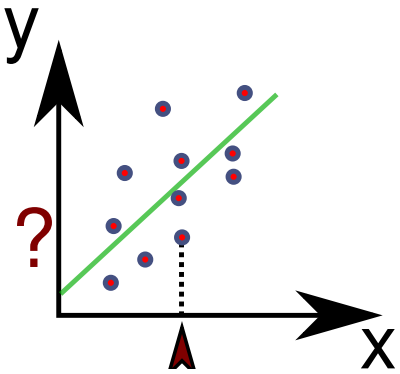
Predict



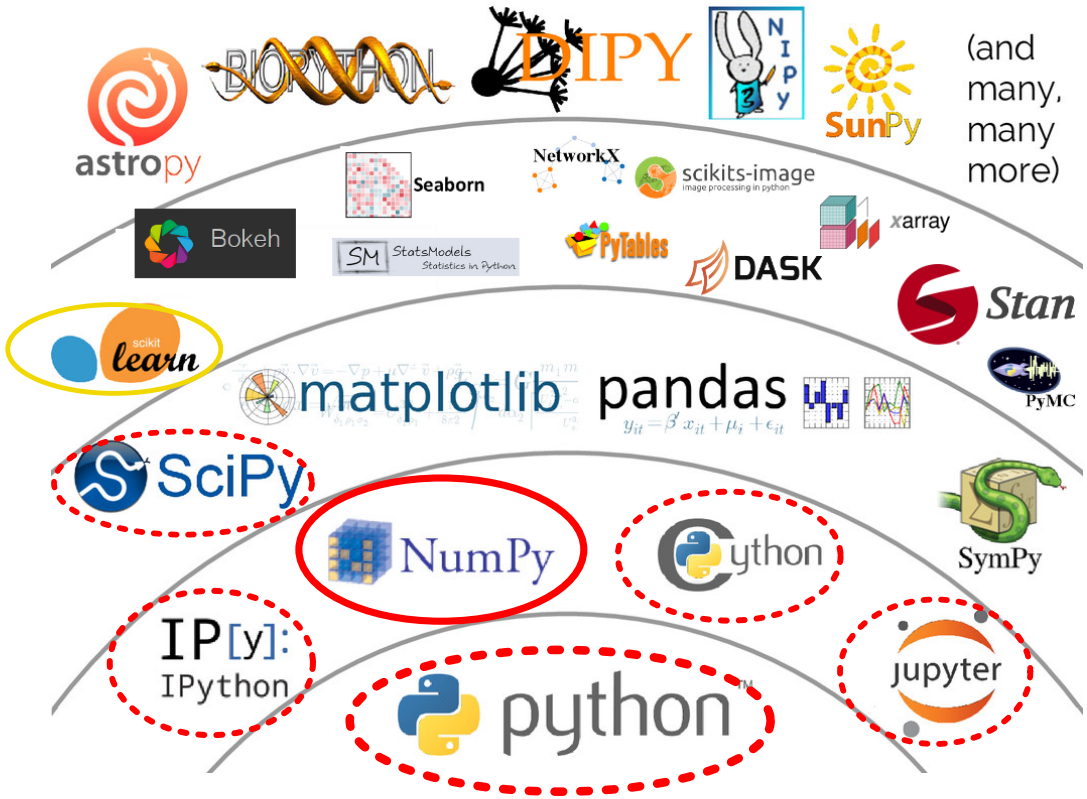
scikit-learn - library for machine learning



Predict

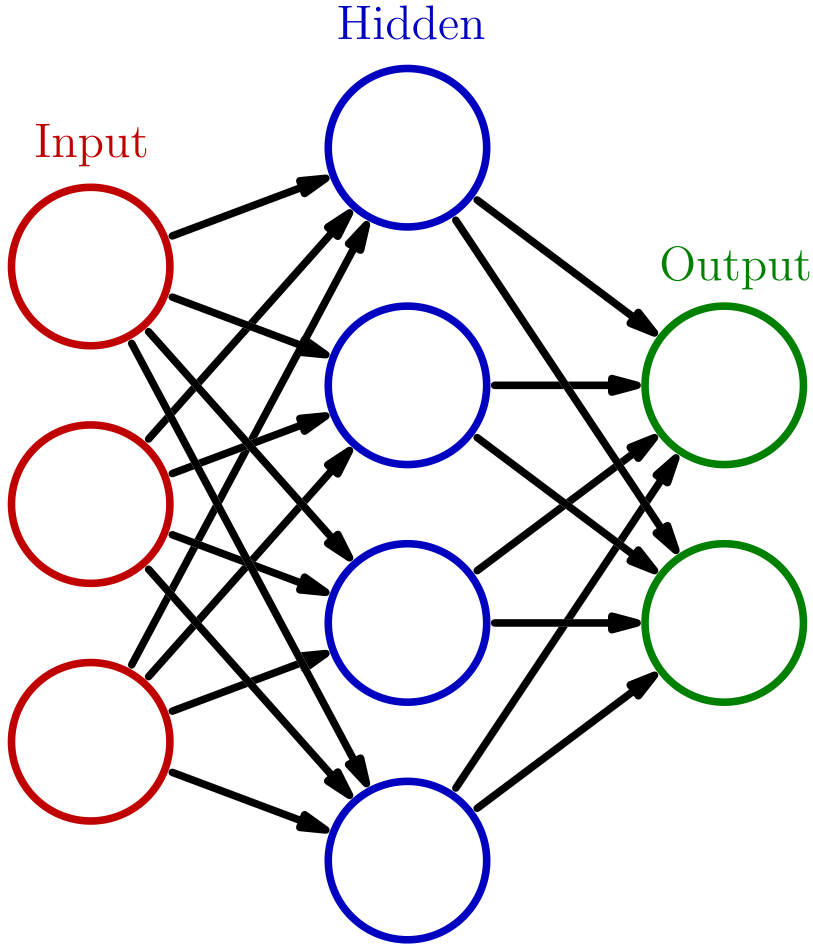


scikit-learn - library for machine learning



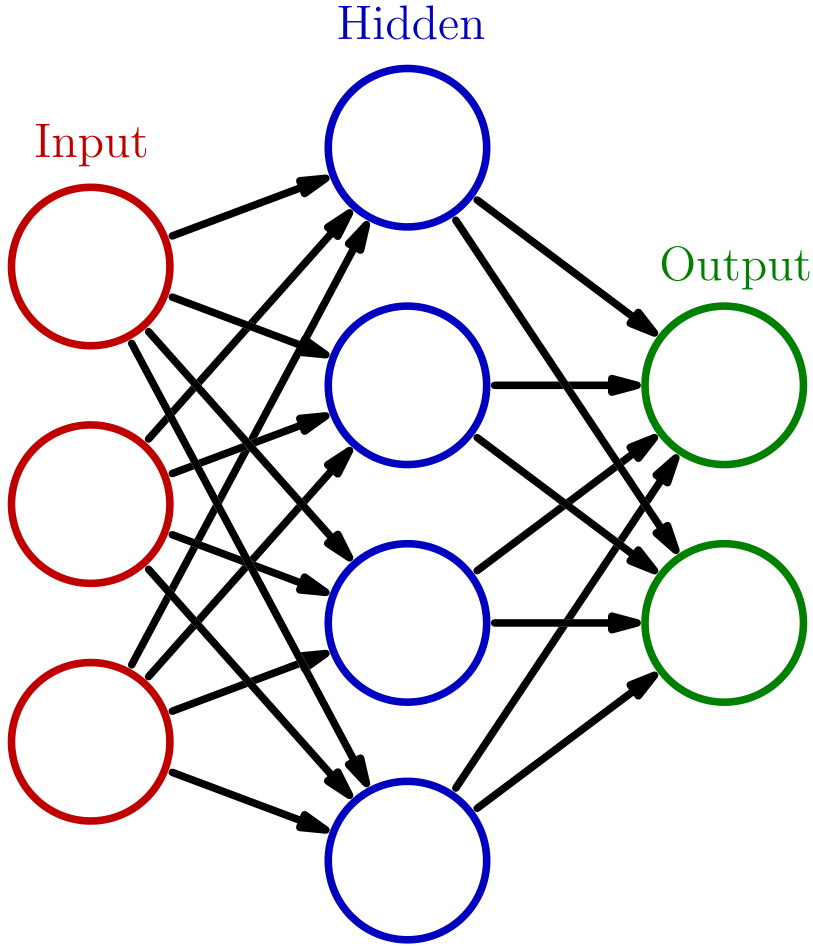
Code!

Predict - with neural networks



Tensorflow, Theano, Keras, Caffe, Torch... - specific for neural networks

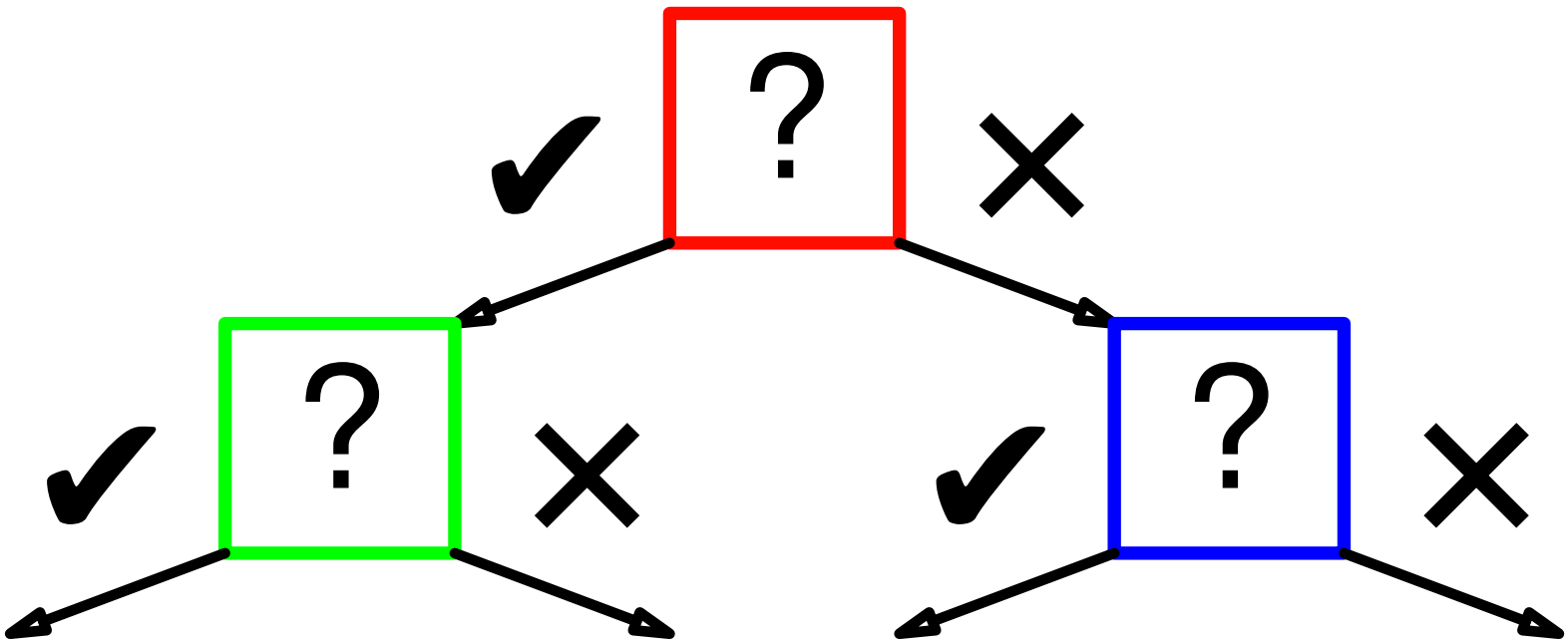
Predict - with neural networks



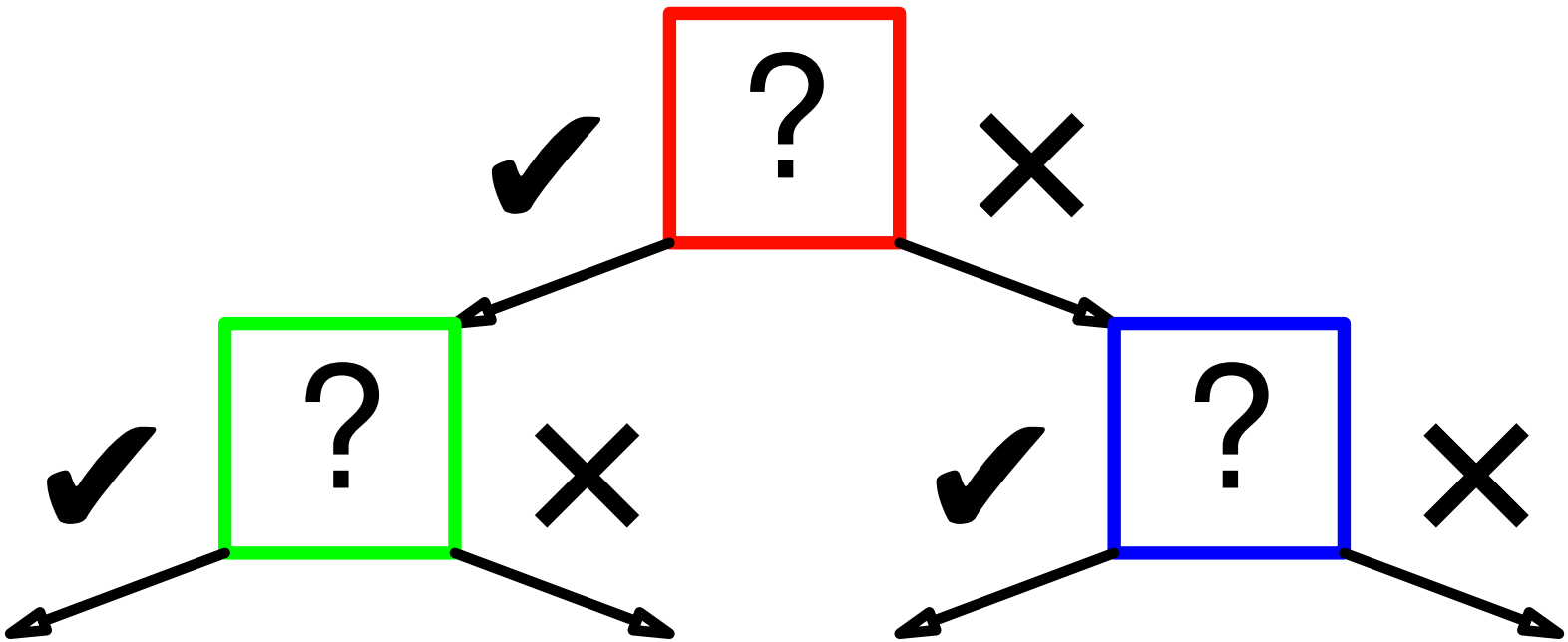
Tensorflow, Theano, Keras, Caffe, Torch... - specific for neural networks

Code!

Predict - with a decision tree



Predict - with a decision tree



Code!

Python or R?

Both - rpy2

Credits

- Map of the pydata stack: adapted from **Jake VanderPlas**'s one:
  `https://speakerdeck.com/jakevdp/the-state-of-the-stack-scipy-2015-keynote`
- **Wikipedia** for the neural network image:
  `https://en.wikipedia.org/wiki/File:Colored_neural_network.svg`