W4995 Applied Machine Learning

Introduction

01/23/19

Andreas C. Müller

Logistics

Email

andreas.mueller@columbia.edu (NOT amueller who is someone else)

CAs: Pranjal Bajaj, Ujjwal Peshin, Liyan Nie, Yao Fu, Luv Aggarwal, Sukriti Tiwari

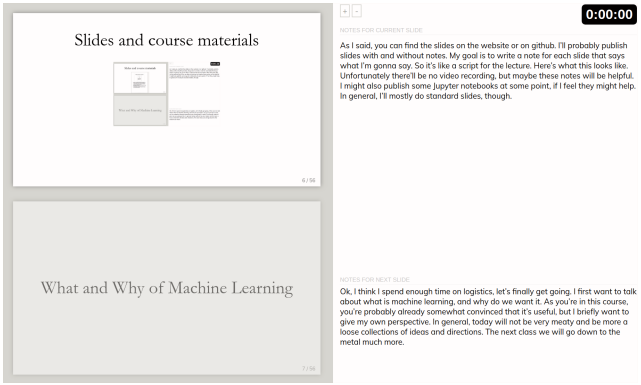
Office Hours

- Andreas Müller Wednesdays 10am-11am, Interchurch 320 K
- CA office hours: TBA

Logistics

- Course website http://www.cs.columbia.edu/~amueller/comsw4995s19/
- Six programming assignments
- Grade: 60% homeworks, 20% first exam, 20% second exam

Slides and course materials



Using markdown with remark. Press "p" for notes.

Lecture Recordings

Plagiarism and Code copying

- Homeworks are checked for plagiarism
- Copied code will result in 0 points for all involved
- Copying from my slides or online sources (Stack overflow, tutorials, etc.) is fine.

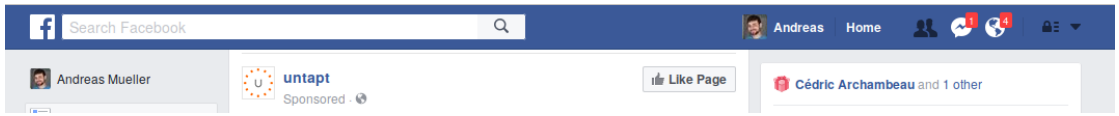
Scikit-learn Development

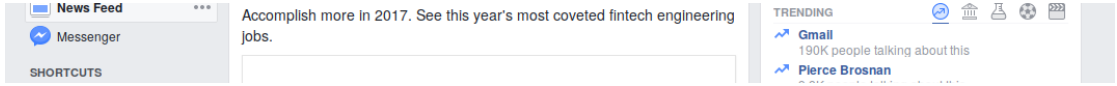
http://scikit-learn.org/dev/developers/contributing.html

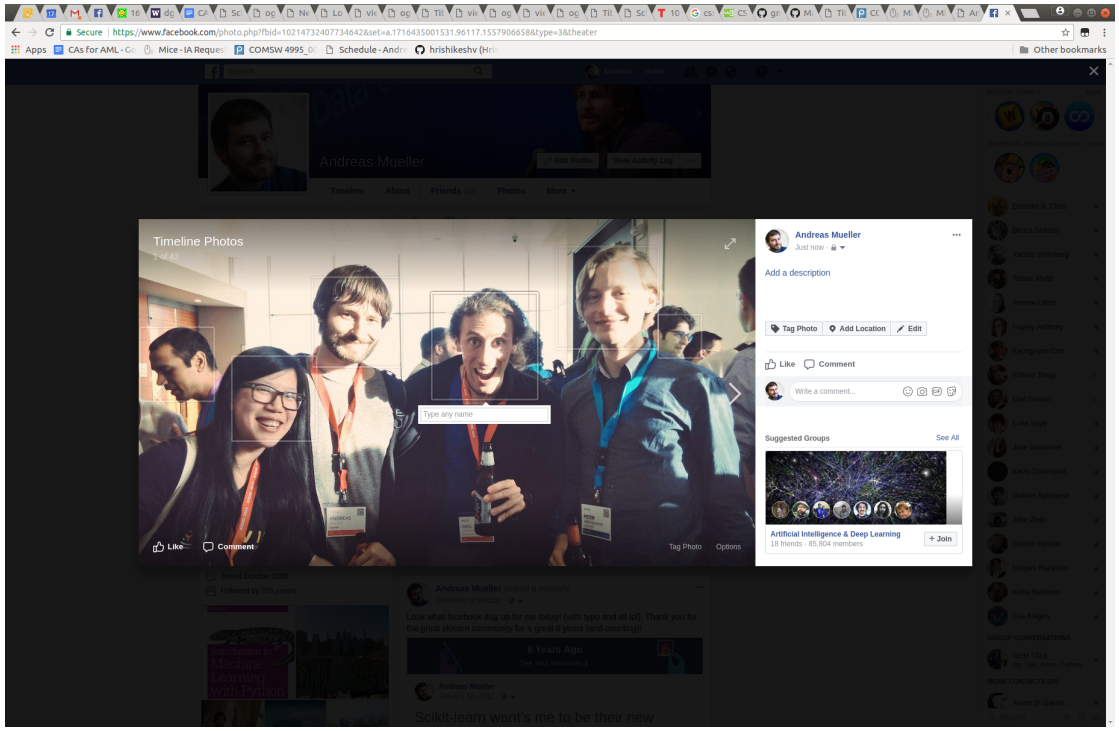
Books

What and Why of Machine Learning

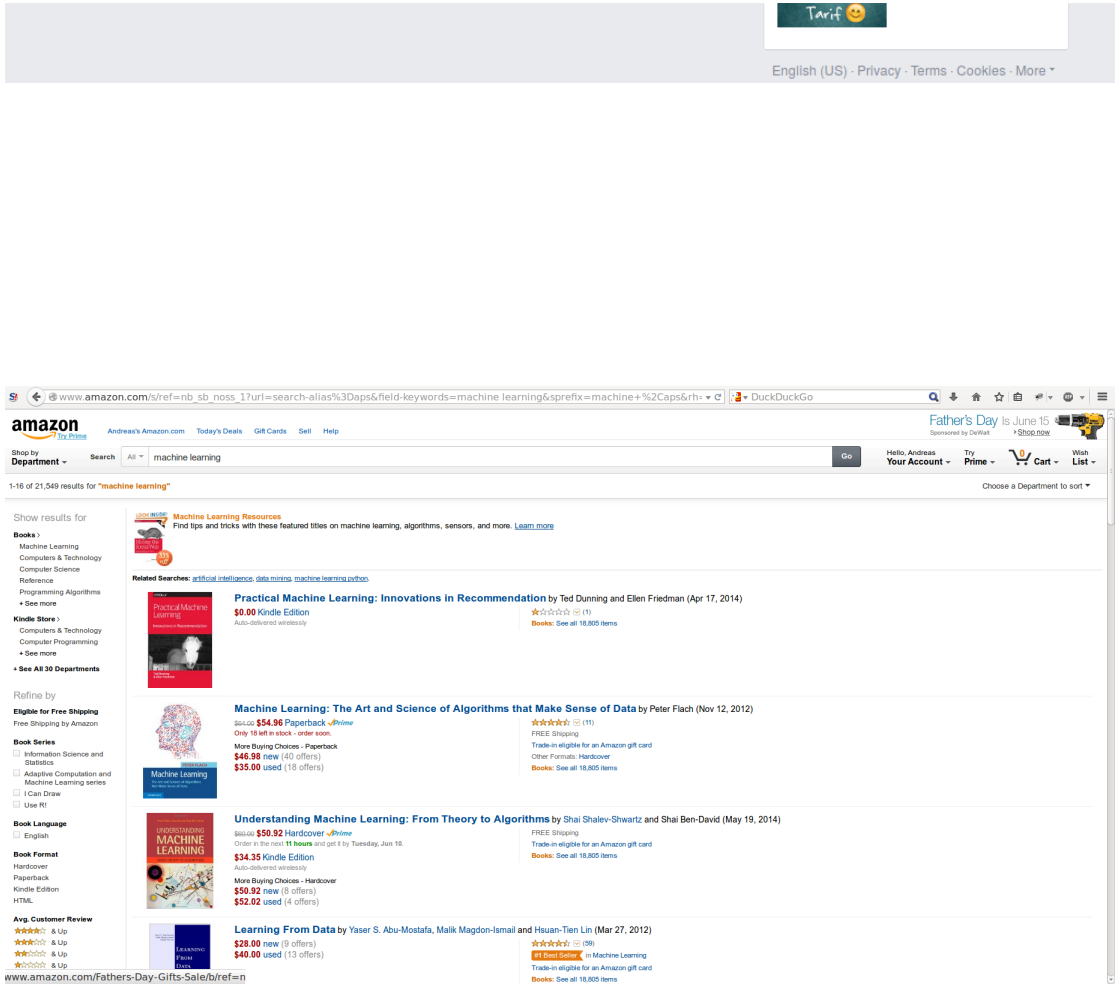
What is machine learning?


Search Facebook
Andreas
Home
Andreas Mueller
untapt
Sponsored
Like Page
Cédric Archambeau and 1 other


News Feed
Messenger
SHORTCUTS
Accomplish more in 2017. See this year's most coveted fintech engineering jobs.
TRENDING
Gmail
190K people talking about this
Pierce Brosnan


Andreas Mueller
Timeline Photos
Type any name
Like
Comment
Tag Photo
Options
Andreas Mueller
Just now
Add a description
Tag Photo
Add Location
Edit
Like
Comment
Write a comment...
Suggested Groups
See All
Artificial Intelligence & Deep Learning
18 friends · 85,804 members
Join


Search for people, places and things
Andreas
Home
Andreas Mueller added 8 new photos.
June 6 at 5:07pm
Bye bye hong kong
Like · Comment · Share
and 6 others like this.
Looks like such a vibrant city! I bet it was awesome
June 6 at 7:39pm · Like
Write a comment...
SPONSORED
See All
Da ist Abwechslung drin!
franziskaner-weissbier.de
3 Sorten Franziskaner Weissbier in einem Pack. Hier klicken und 1 von 100 Packs gewinnen!
Singles auf Facebook
Schau dir Dating-Profile von Singles in deiner Nähe an.
Online Essen bestellen
lieferheld.de
Neu mit Lieferheld: PLZ eingeben, Lieferdienst finden und genießen!
Globaler Chauffeurservice
blacklane.com
Fahren Sie eine Klasse besser zu günstigen Preisen – Blacklane ist weltweit verfügbar!
28,723 people like this
Sanssouci...
Geld vom Staat zurück!
Steuererklärung preiswert für Arbeitnehmer, Azubis Arbeitsuchende, Rentner Pensionäre etc.
Like Page
· 4 people like this page
Der neue WhatsApp Tarif!
eplus.de
WhatsApp SIM – der revolutionäre Prepaid Tarif
Der Prepaid


Tarif
English (US) · Privacy · Terms · Cookies · More
www.amazon.com/s/ref=nb_sb_noss_1?url=search-alias%3Daps&field-keywords=machine learning&sprefix=machine+%2Caps&rh=
DuckDuckGo
amazon
Try Prime
Andreas's Amazon.com
Today's Deals
Gift Cards
Sell
Help
Father's Day Is June 15
Shop now
Shop by Department
Search
machine learning
Go
Hello, Andreas
Your Account
Try Prime
Cart
Wish List
1-16 of 21,549 results for "machine learning"
Choose a Department to sort
Show results for
Books
Machine Learning
Computers & Technology
Computer Science
Reference
Programming Algorithms
+ See more
Kindle Store
Computers & Technology
Computer Programming
+ See more
+ See All 30 Departments
Refine by
Eligible for Free Shipping
Free Shipping by Amazon
Book Series
Information Science and Statistics
Adaptive Computation and Machine Learning series
I Can Draw
Use It!
Book Language
English
Book Format
Hardcover
Paperback
Kindle Edition
HTML
Avg. Customer Review
& Up
Machine Learning Resources
Find tips and tricks with these featured titles on machine learning, algorithms, sensors, and more. Learn more
Related Searches: artificial intelligence, data mining, machine learning python
Practical Machine Learning: Innovations in Recommendation by Ted Dunning and Ellen Friedman (Apr 17, 2014)
$0.00 Kindle Edition
Auto-delivered wirelessly
Books: See all 18,805 items
Machine Learning: The Art and Science of Algorithms that Make Sense of Data by Peter Flach (Nov 12, 2012)
$54.96 Paperback Prime
Only 18 left in stock - order soon.
More Buying Choices - Paperback
$46.98 new (40 offers)
$35.00 used (18 offers)
FREE Shipping
Trade-in eligible for an Amazon gift card
Other Formats: Hardcover
Books: See all 18,805 items
Understanding Machine Learning: From Theory to Algorithms by Shai Shalev-Shwartz and Shai Ben-David (May 19, 2014)
$50.92 Hardcover Prime
Order in the next 11 hours and get it by Tuesday, Jun 10.
$34.35 Kindle Edition
Auto-delivered wirelessly
More Buying Choices - Hardcover
$50.92 new (8 offers)
$52.02 used (4 offers)
FREE Shipping
Trade-in eligible for an Amazon gift card
Books: See all 18,805 items
Learning From Data by Yaser S. Abu-Mostafa, Malik Magdon-Ismail and Hsuan-Tien Lin (Mar 27, 2012)
$28.00 new (9 offers)
$40.00 used (13 offers)
#1 Best Seller in Machine Learning
Trade-in eligible for an Amazon gift card
Books: See all 18,805 items
www.amazon.com/Fathers-Day-Gifts-Sale/b/ref=n

Introduction to Machine Learning with Python: A Guide for Data Scientists 1st Edition

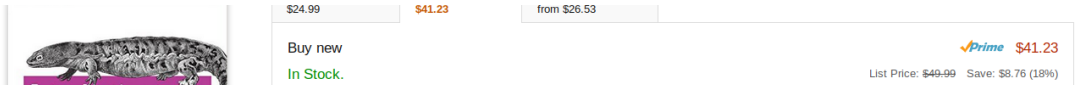
by Andreas C. Müller (Author), Sarah Guido (Author)

9 customer reviews

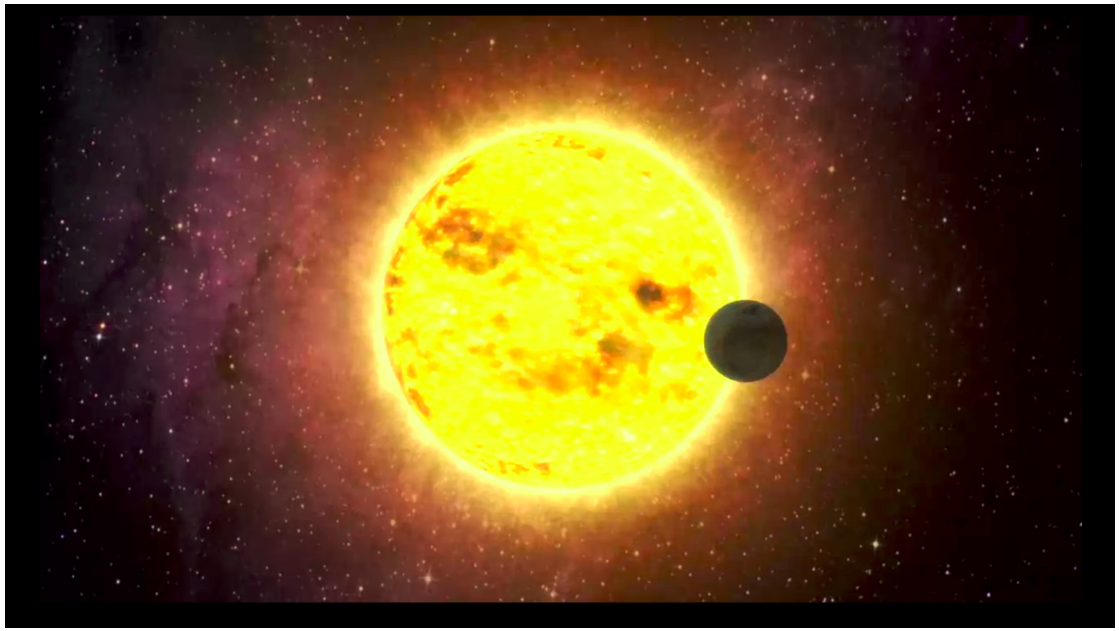
Look inside

O'REILLY

Kindle | Paperback | Other Sellers


$24.99
$41.23
from $26.53
Buy new
In Stock.
Prime $41.23
List Price: $49.99 Save: $8.76 (18%)

Science!



Types of Machine Learning

Types of Machine Learning

- Supervised
- Unsupervised
- Reinforcement

Supervised Learning

$$(x_i, y_i) \propto p(x, y) \text{ i.i.d.}$$

$$x_i \in \mathbb{R}^p$$

$$y_i \in \mathbb{R}$$

$$f(x_i) \approx y_i$$

Generalization

Not only $f(x_i) \approx y_i,$

also for new data: $f(x) \approx y$

Examples of Supervised Learning

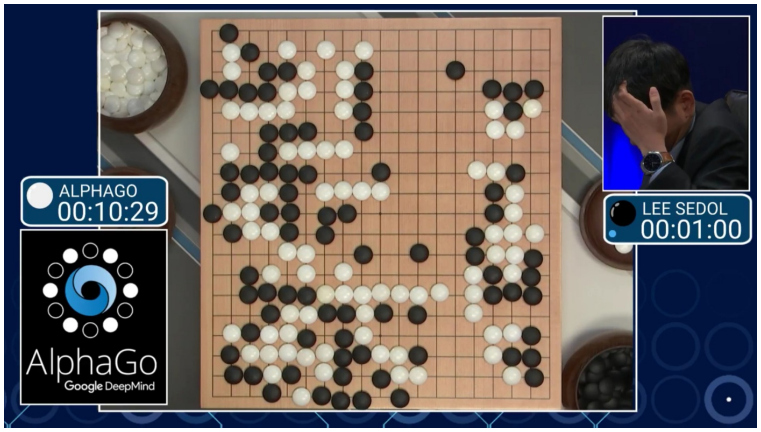
- spam detection
- medical diagnosis
- add click prediction

Unsupervised Learning

$$x_i \propto p(x) \text{ i.i.d.}$$

Learn about p.

Reinforcement Learning



Explore & Learn

Other kinds of learning

- Semi-supervised
- Active Learning
- Forecasting
- ...

Classification and Regression

Classification

- target y discrete
- Will you pass?

Regression

- target y continuous
- How many points will you get in the exam?

Relationship to Statistics

Statistics

- model first
- infence emphasis

Machine learning

- data first
- prediction emphasis

Relationship to Statistics

Statistics

- model first
- infence emphasis

Machine learning

- data first
- prediction emphasis

Guiding Principles in Machine Learning

Goal considerations

The Cost of Complex Systems

Data driven first? yes! (or maybe)

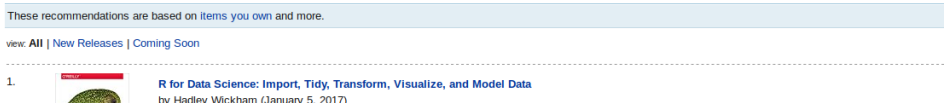
Machine Learning first: No!

Thinking in Context!
What is the baseline?
What is the benefit?

Good and Bad Substitutes

Communicating Results

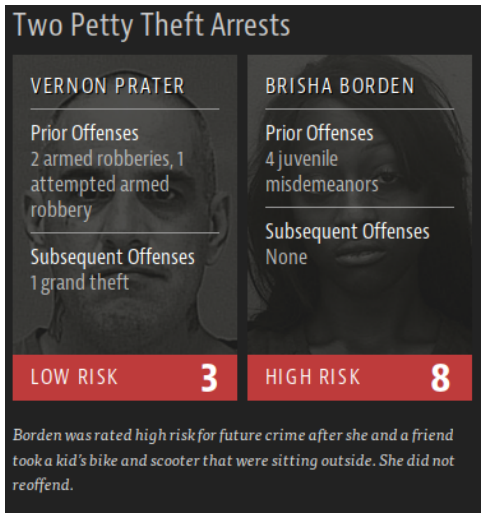
Explainable Results



Average Customer Review: ★★★★★ (4)
In Stock
List Price: ~~$39.99~~
Price: $32.91

Add to Cart Add to Wish List

Sidebar: Ethical Considerations



https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[criminal sentencing]

Ethics: It's in the application!

Data and Data Collection

Free vs Expensive Data

Free

Expensive

Free vs Expensive Data

Free

Predict observable events

- Stock market
- Clicks
- House numbers

Expensive

Automate complex process

- Dignosis
- Drug Trial
- Chip Design

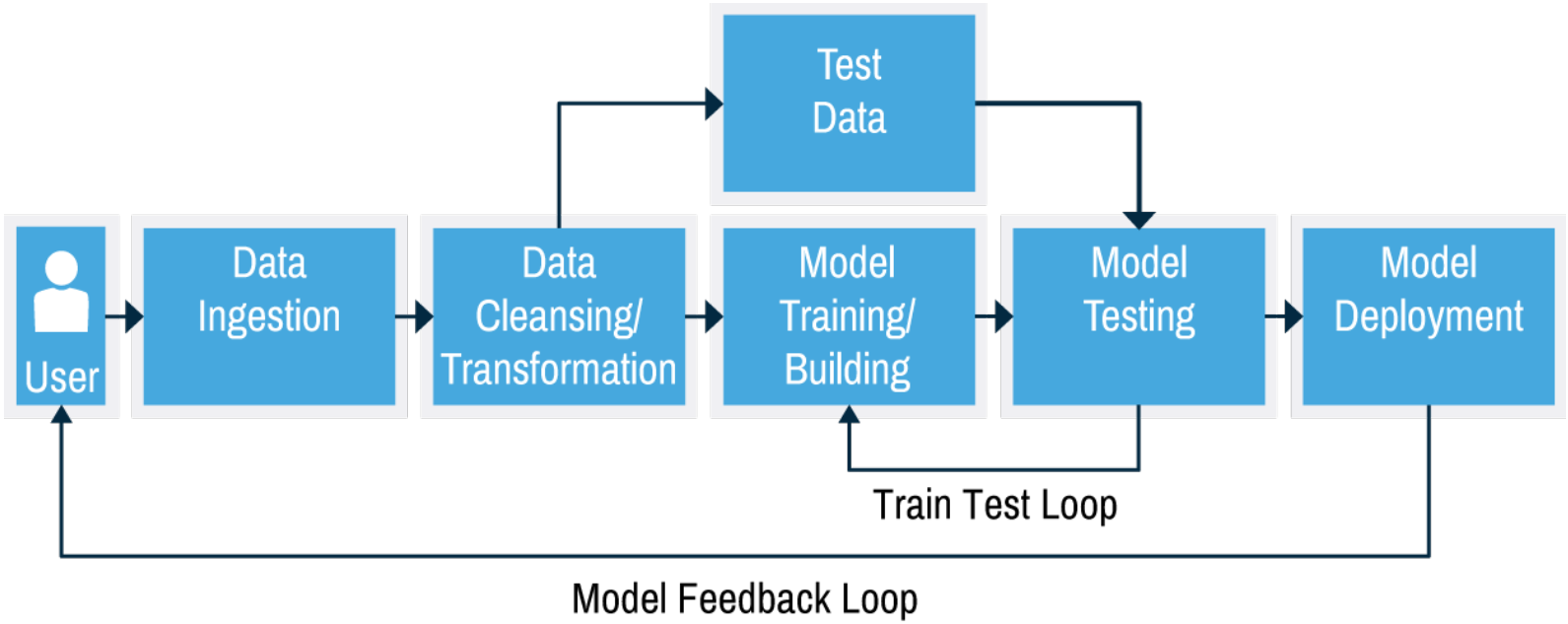
The cost (and benefit?) of BigData

Subsample to RAM (which can be 512gb)

Cornerstones of this Course

- Good software engineering practices
- Problem definition and success measures
- Feature engineering and data cleaning
- Strength and weaknesses of different algorithms
- Model selection best practices

The Machine Learning Work-Flow



Taken from MAPR https://www.mapr.com/ebooks/spark/08-recommendation-engine-spark.html

[recommendation engine sparkmlib]

General coding guidelines

Programs must be written for people to read, and only incidentally for machines to execute.

Harold Abelson (wizard book)

Everyone knows that debugging is twice as hard as writing a program in the first place. So if you're as clever as you can be when you write it, how will you ever debug it?

Brian Kernighan

- Don’t be clever!
- Make it readable!
- Future you is the most likely person to try to understand your code.

- Don’t be clever!
- Make it readable!
- Future you is the most likely person to try to understand your code.
- Avoid writing code.

Python basics

Why Python?

- General purpose language
- Great libraries
- Easy to learn / use
- Contenders: R (Scala? Julia?)

The two language problem

Python is sloooow...

- Numpy: C
- Scipy: C, fortran
- Pandas: Cython, Python
- Scikit-learn: Cython, Python

- CPvthon: C

Python 2 vs Python 3

- “current” : (2.7), 3.6, 3.7
- Don't use Python 2

Python ...

Package management:

- don't use system python!
- use Virtual environments
- understand pip (and wheels)
- probably use Conda (and anaconda or conda-forge)

Python ...

Package management:

- don't use system python!
- use Virtual environments
- understand pip (and wheels)
- probably use Conda (and anaconda or conda-forge)

Pip and conda and upgrades

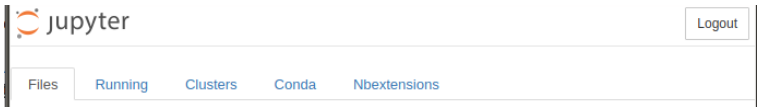
- Pip upgrade works on dependencies (unless you do -no-dep)
- pip has no dependency resolution!
- conda has dependency resolution
- Use conda environments!
- upgrading a conda package with pip (or vice versa) will break stuff!

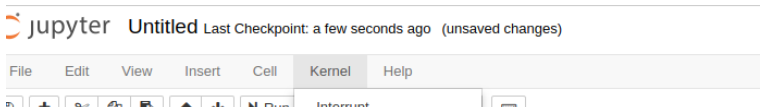
Environments and Jupyter Kernels

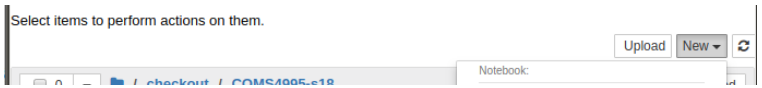
- Environment != kernels
- Use nb_conda_kernels or add environment kernels manually:

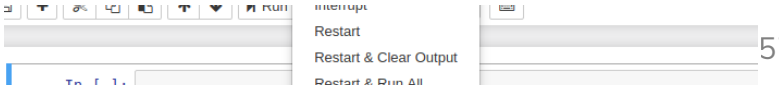
```
source activate myenv
python -m ipykernel install --user --name myenv --display-name "Python (myenv)"
source activate other-env
python -m ipykernel install --user --name other-env --display-name "Python (other-env)"
```

- https://jakevdp.github.io/blog/2017/12/05/installing-python-packages-from-jupyter/









Dynamically typed, interpreted

- Invalid syntax lying around
- Code is less self-documenting

Editors

- Flake8 / pyflake
- Scripted / weak typing: Have a syntax checker!
- write pep8 (according to the standard, not the tool)
- use autopep8 if you have code lying around

Questions ?

60 / 60