UCSB IGERT Bootcamp (1/3)

#IGERTBootcamp14
http://git.io/XxcQAA

Luca Foschini (@calimagna)

Day 1

Preliminaries

- Introductions
- Take the self assessment

http://svy.mk/1q5wG4x

- You're the experiment

Preliminaries (2)

- Grab the course material:

```
git clone https://github.com/LucaFoschini/IGERTBootcamp.git
```

- Set path

```
cd ~/IGERTBootcamp/scripts
source set_path.sh
```

- Start the notebook:

```
cd ~/IGERTBootcamp/notebooks
ipython notebook
```

Version Control

- Why version control?
- Git and GitHub
- Git for Scientist: A Tutorial

Reproducible Science

- Reproducible science
- One possible approach: Python Notebook
- Mix code, latex, visualization.

Data Science

- Definition(s)
- Presentation on data science
- Data science from command line

Introduction to Python

- Introduction to Python
- Basic data structures
- Read, save, open files

Data Preparation

- Data wrangling in python, pandas
- Selection, grouping, time series, data in-out

Libraries and Integrations

- APIs
- NLTK, NetworkX, scikit-learn
- theano, pyMCMC
- Big Data: python parallel, spark

Miniproject

- Extend the MaxMind Dataset exploration

Day 2

CS Foundation

- Day I survey
- The basic of Computer Science, search, sort, index, hash tables
- Algorithmic complexity
- http://bost.ocks.org/mike/algorithms/#shuffling

Foundations in Python

- Lists, Dict, Set, Efficiency
- Theory and practice: vectorized forms in python, matlab, R
- http://nbviewer.ipython.org/github/rossant/ipython-minibook/blob/master/chapter3/301-vector-computations.ipynb

Probability Theory

- Computing statistics of distribution: average, max, min, top-k, median
- Bernoulli trials, conditioning, paradoxes
- Randomized algorithms, sampling.
- Digression: Distance between distributions. implement EM distance

Statistics

- Correlation, causation
- Significance, validation. p-values and its problems (blog post) compute in R.
- Check your assumptions: stationarity, population size, experiment design, power of test calculation.

Exercise

- Histograms, scatterplots,
- common pitfalls in probability
- Digression: randomness in computers
- Scientist dilemma: Coin flip problem

Day 3

Graphs

- Definition, examples
- Visits
- Generate (ER models)
- http://bost.ocks.org/mike/algorithms/#maze-generation

Graph Zoo

- Directed, indirected, planar, trees, cliques
- Edge/node costs/labels
- Graphs as models,
- Generate restricted graph classes (planar? geometric graphs)

Measures Modeling

- Diameter, connectivity, degree distribution.
- Similarity?
- Shortest paths, landmark
- Digression, time dependent shortest paths

Hard vs. Easy

- Problems on Graphs:
- Digression: NP hardness. TSP vs. Eulerian
- BC distance, sparsification, sampling
- Multi-genre graphs

Other libraries

- Boost Graphs
- pregel, GraphX (spark)
- Mathematica
- simpleNetworkD3.js
- three.js

Exercise on Graphs

- Find the most influential nodes in:
- https://github.com/rossant/ipython-minibook