UCSB IGERT Bootcamp (1/3)

#IGERTBootcamp14
http://git.io/XxcQAA

Luca Foschini (@calimagna)

Day 1

Preliminaries

- Introductions
- Take the self assessment

http://svy.mk/1uAvTyC

- 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

- The basic of Computer Science, search, sort, index, hash tables
- Algorithmic complexity: Big O notation, examples

Foundations in Python

- Lists, Dict, Set, Efficiency
- Theory and practice: vectorized forms in python, matlab, R

Probability Theory

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

Miniproject

- Randomness and Bernoulli trials
- Scientist dilemma (miniproject)

Day 3

Graphs

- Definition, examples
- Visits
- Restricted classes (trees, planars, sparse vs. dense)

Graph Zoo

- Directed, weighted
- Edge/node costs/labels
- How to generate graphs? Generate restricted graph classes (2-3 colorable)

Measures Modeling

- Diameter, connectivity.
- Shortest paths

Hard vs. Easy

- Problems on Graphs.
- Digression: NP hardness. TSP vs. Eulerian, vertex cover, approximation algorithm
- Sparsification, sampling
- Multi-genre graphs

Other libraries

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

Exercise on Graphs

- Connect to FB/Twitter/LinkedIn/GitHub/Google+ API
- Visualize social graphs derived from the above