

What to do when your data are **large** but not big

Dillon Niederhut

PyBay – the San Francisco Bay Area Python Conference

20 August 2016

Introduction

Motivation

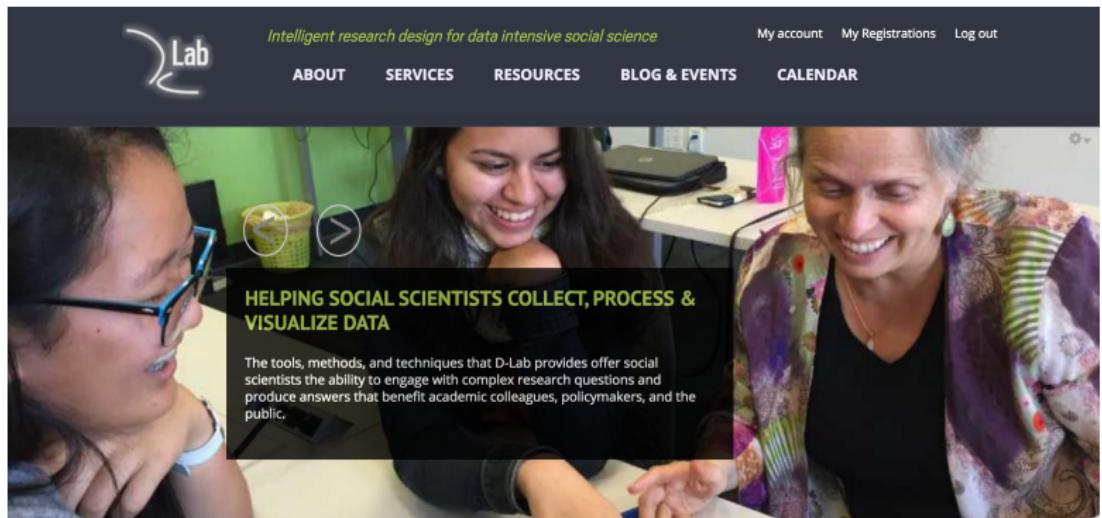
Strategies

Closing

about this talk

- data at github.com/deniederhut/pybay_2016
- python libraries : celery, h5py, numpy, pandas, pymongo
- other libraries : mongodb, rabbitmq, sqlite

about me



- dlab.berkeley.edu
- @DLabAtBerkeley

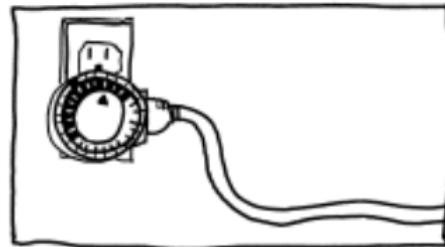
size concerns

FIGURING OUT WHY MY HOME
SERVER KEEPS RUNNING OUT
OF SWAP SPACE AND CRASHING:



1-10 HOURS

PLUGGING IT INTO A LIGHT TIMER
SO IT REBOOTS EVERY 24 HOURS:



5 MINUTES

WHY EVERYTHING I HAVE IS BROKEN

1

¹from xkcd

time concerns

ONLINE PACKAGE TRACKING:

PROS:

CONVENIENT
USEFUL

CONS:

MAKES YOU
CRAZY

REFRESH

| AWW, STILL IN MEMPHIS.

REFRESH

| AWW, STILL IN MEMPHIS.

REFRESH

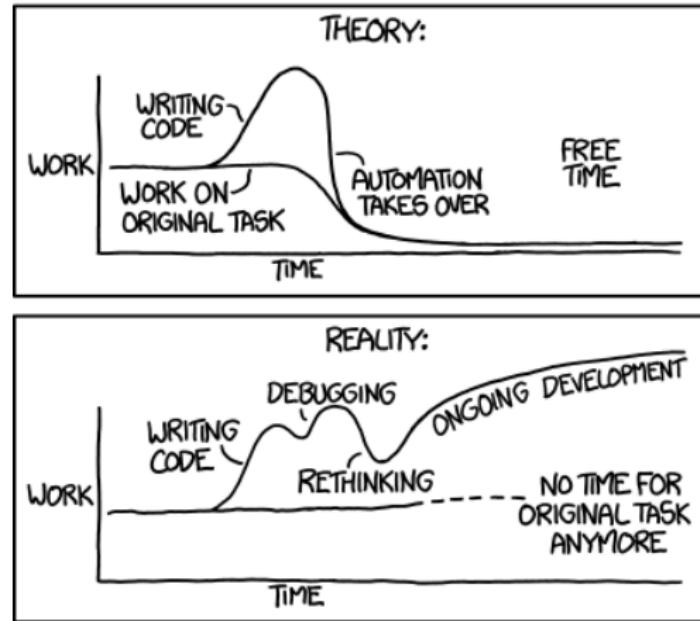
| AWW, STILL IN MEMPHIS.



2

code concerns

"I SPEND A LOT OF TIME ON THIS TASK.
I SHOULD WRITE A PROGRAM AUTOMATING IT!"



3

³thanks Randall!!

sequential processing



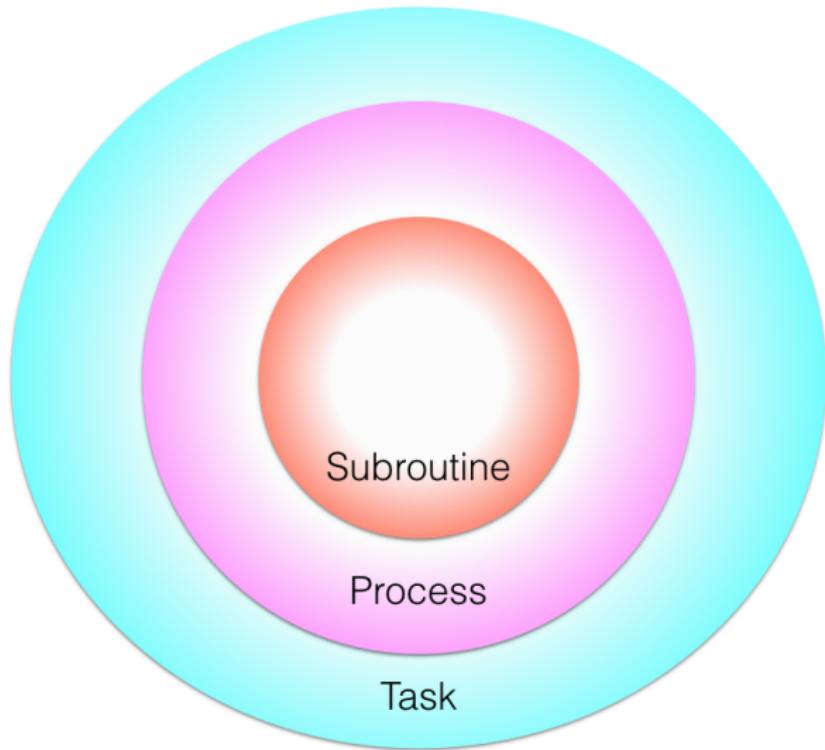
parallel processing

Introduction

Motivation

Strategies

Closing

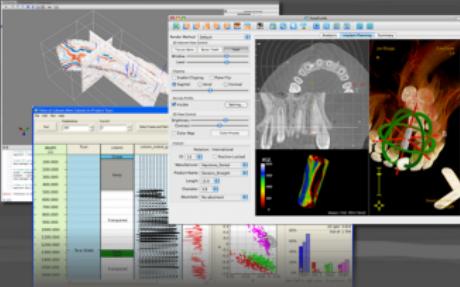


contact

DOWNLOADS: [Canopy](#) | [PyXLL](#) | [View cart \(\\$0\)](#) | [Create Account or Log In](#)

ENTHOUGHT SCIENTIFIC COMPUTING SOLUTIONS

PRODUCTS TRAINING CONSULTING COMPANY CONTACT



Software Consulting and Application Development

Our data analysis, data visualization, and data processing expertise can help you:

- Turn ideas into results
- Translate data into actionable insight
- Fast track innovation

[Learn More >](#)

Python Training on Demand | Enthought Canopy | Python for Excel | Software Consulting

- dillon.niederhut.us
- [@dillonniederhut](https://twitter.com/dillonniederhut)