Introduction to Data Programming

Michael Ernst and Bill Howe

Summer 2012

What this course is

- An introduction to core programming concepts with an emphasis on real data manipulation tasks from science, engineering, and business.

What this course is not

- A “skills course” in Python
  - …though you will become proficient in the basics of the Python programming language
  - …and you will gain experience with some important Python libraries
- A data analysis / “data science” / data visualization course
  - There will be very little statistics knowledge assumed or taught
- A “project” course
  - the assignments are “real,” but are intended to teach specific programming concepts
- A “big data” course
  - Datasets will all fit comfortably in memory
  - No parallel programming

4 September 2008 | www.nature.com/nature | £10
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
FRANK OHLHORST
Beautiful Data
O'REILLY
Economist
The economic shift from West to East
Genetically modified crops blossom
model
technologies
DATA

"It's a great time to be a data geek."
-- Roger Barga, Microsoft Research

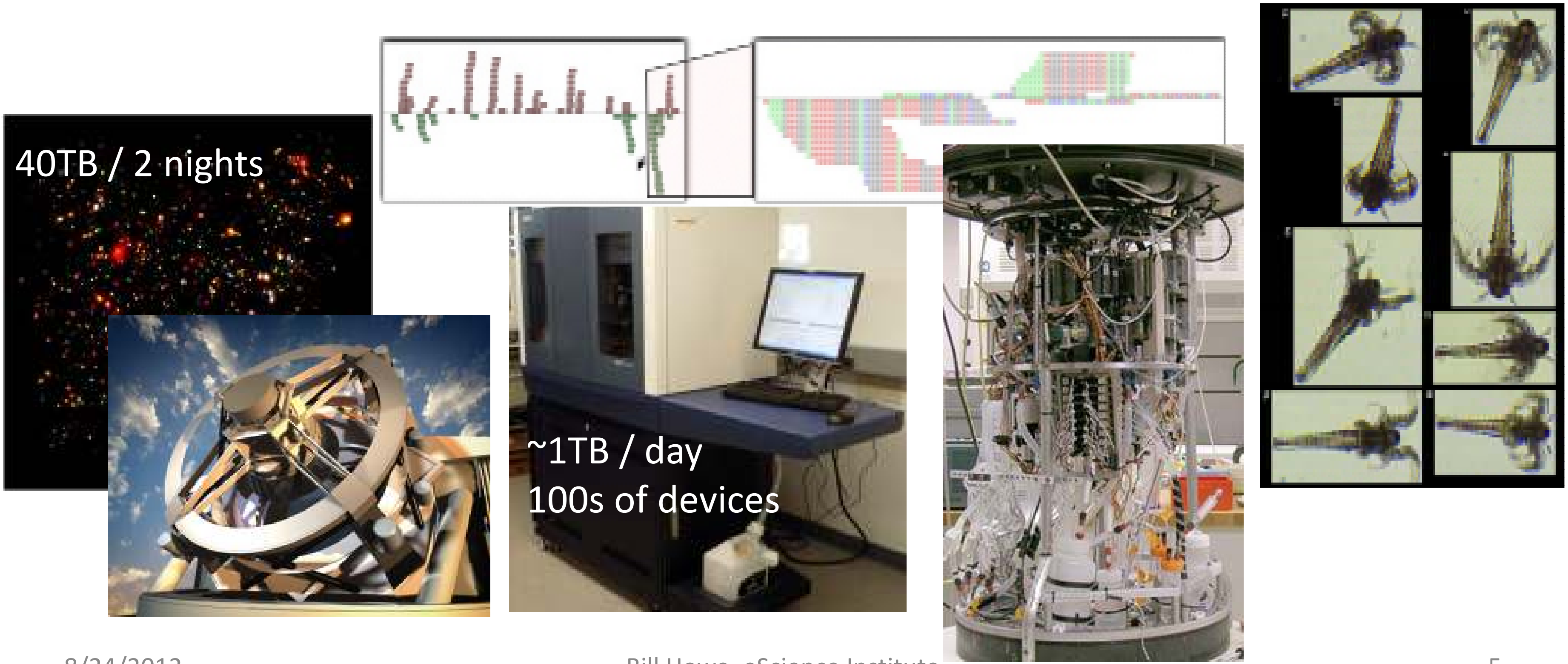
"The greatest minds of my generation are trying to figure out how to make people click on ads"
-- Jeff Hammerbacher, co-founder, Cloudera

All of science is reducing to computational data manipulation

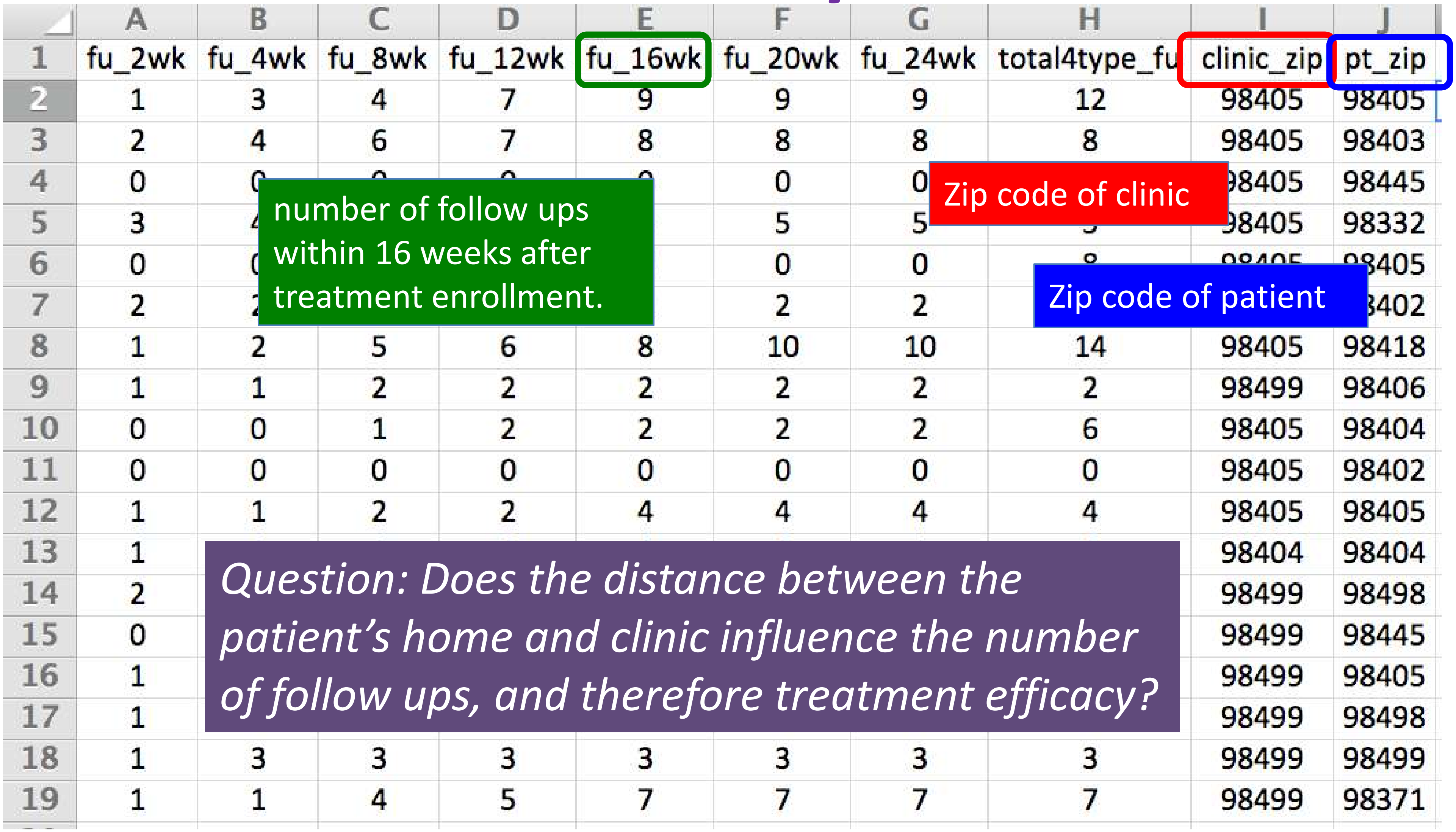
Old model: "Query the world" (Data acquisition coupled to a specific hypothesis)

New model: "Download the world" (Data acquisition supports many hypotheses)

- **Astronomy: High-resolution, high-frequency sky surveys (SDSS, LSST, PanSTARRS)**
- **Biology: lab automation, high-throughput sequencing,**
- **Oceanography: high-resolution models, cheap sensors, satellites**



Example: Assessing Treatment Efficacy



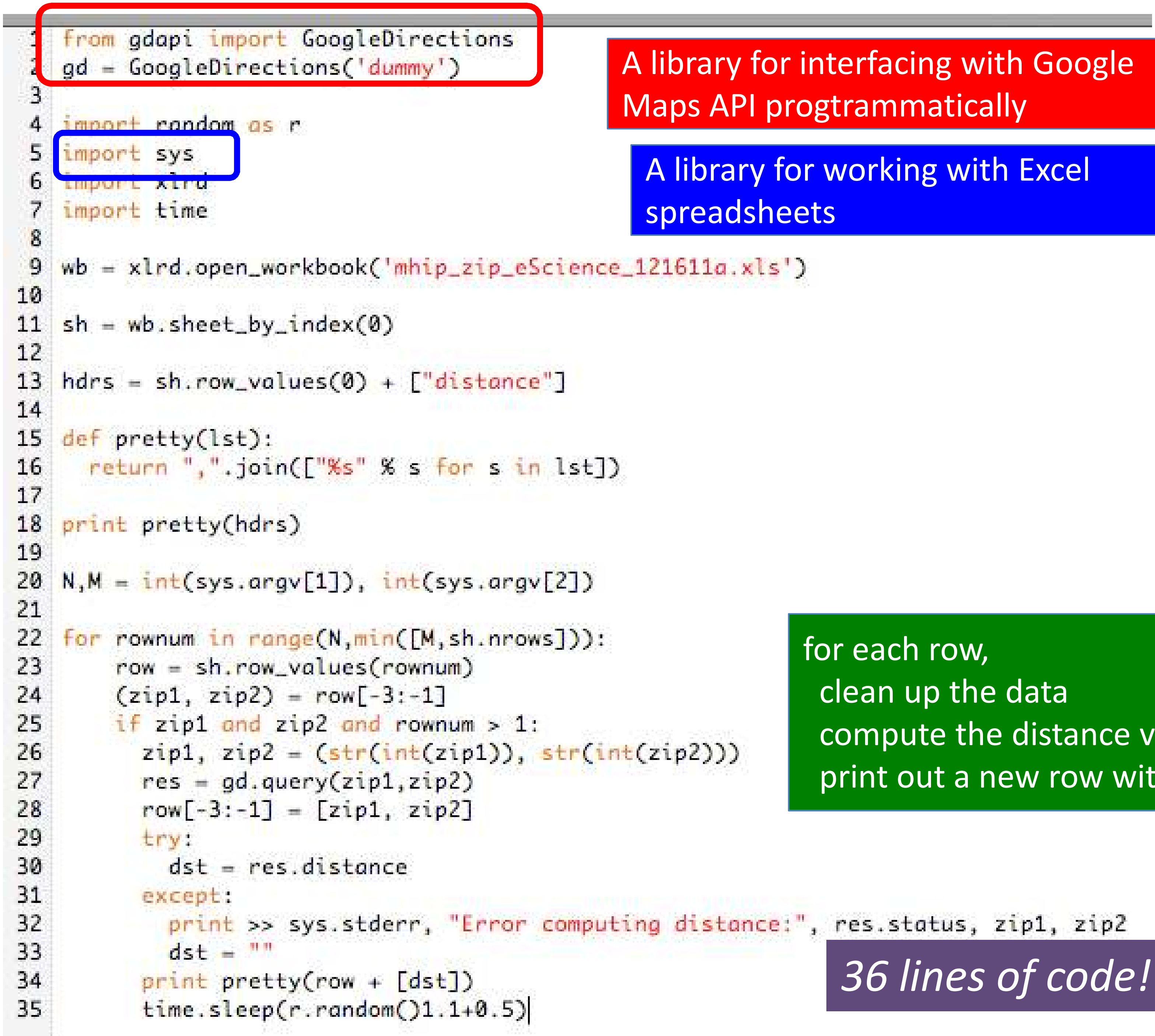
	A	B	C	D	E	F	G	H	I	J
1	fu_2wk	fu_4wk	fu_8wk	fu_12wk	fu_16wk	fu_20wk	fu_24wk	total4type_fu	clinic_zip	pt_zip
2	1	3	4	7	9	9	9	12	98405	98405
3	2	4	6	7	8	8	8	8	98405	98403
4	0	0	0	0	0	0	0		98405	98445
5	3	4				5	5	5	98405	98332
6	0	0				0	0	8	98405	98405
7	2	2				2	2			98402
8	1	2	5	6	8	10	10	14	98405	98418
9	1	1	2	2	2	2	2	2	98499	98406
10	0	0	1	2	2	2	2	6	98405	98404
11	0	0	0	0	0	0	0	0	98405	98402
12	1	1	2	2	4	4	4	4	98405	98405
13	1								98404	98404
14	2								98499	98498
15	0								98499	98445
16	1								98499	98405
17	1								98499	98498
18	1	3	3	3	3	3	3	3	98499	98499
19	1	1	4	5	7	7	7	7	98499	98371

number of follow ups within 16 weeks after treatment enrollment.

Zip code of clinic

Zip code of patient

Question: Does the distance between the patient’s home and clinic influence the number of follow ups, and therefore treatment efficacy?



```
from gdapi import GoogleDirections
gd = GoogleDirections('dummy')

import random as r
import sys
import xlrd
import time

wb = xlrd.open_workbook('mhip_zip_eScience_121611a.xls')

sh = wb.sheet_by_index(0)

hdrs = sh.row_values(0) + ["distance"]

def pretty(lst):
  return ",".join(["%s" % s for s in lst])

print pretty(hdrs)

N,M = int(sys.argv[1]), int(sys.argv[2])

for rownum in range(N,min([M,sh.nrows])):
    row = sh.row_values(rownum)
    (zip1, zip2) = row[-3:-1]
    if zip1 and zip2 and rownum > 1:
      zip1, zip2 = (str(int(zip1)), str(int(zip2)))
      res = gd.query(zip1,zip2)
      row[-3:-1] = [zip1, zip2]
      try:
        dst = res.distance
      except:
        print >> sys.stderr, "Error computing distance:", res.status, zip1, zip2
        dst = ""
      print pretty(row + [dst])
      time.sleep(r.random()1.1+0.5)
```

A library for interfacing with Google Maps API progtrammatically

A library for working with Excel spreadsheets

for each row,
  clean up the data
  compute the distance via Google
  print out a new row with dist.

36 lines of code!

Demo: Twitter Sentiment Analysis

- Do people have a favorable opinion of your product or company? Can we quantify this?
- A very simple model: Words and phrases can be assigned a positive/negative valence.
- Add up the valences of all the words in the tweet.
- Get the current sentiment for a search term
- Plot the sentiment of all tweets for a given period.

Demo: Twitter message length distribution

- Plot the message length vs. time

Demo: Twitter Top K

- Count the number of unique values for
  - words
  - user mentions
  - hashtags
  - places (of the user)
- Plot the top k values

Demo: Twitter Tagcloud

- Given the histogram of words, visualize the top 100 or so by mapping count to font size.

Other Possible Twitter Exercises

- Map of geocoordinates showing location
- Continuously updating plots

Learning Objectives (1)

- Computational problem-solving
  - Writing a program will become your "go-to" solution for data analysis tasks
- Basic Python proficiency
  - Including experience with relevant libraries for data manipulation, scientific computing, and visualization.
- Experience working real datasets
  - astronomy, biology, linguistics, oceanography, open government, social networks, and more.
  - You will see that these are easy to process with a program, and that doing so yields insight.

Learning Objectives (2)

- By the end of the quarter, students will be able to take a data source and a problem description and independently write a complete, useful program to solve the problem. Students will learn problem-solving and produce end-to-end solutions.
- All assignments will use real data, at realistic scale and complexity.
- Students will learn enough programming concepts to permit them to continue to more advanced computer science classes such as CSE 143, if desired. However, students will be independent enough not to need that for simple day-to-day data analysis and manipulation.

Book

- Think Python: How to Think Like a Computer Scientist
- Freely available online:
  - http://www.greenteapress.com/thinkpython/

Think Python: How to Think Like a Computer Scientist

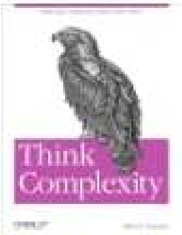
Allen B. Downey

Version 1.6.6

May 2012

Are you using one of our books in a class?

We'd like to know about it. Please consider filling out this short survey.



Assessment

- Assignments 60%
- Exams 30%
- Participation 10%

Late policy

- Each student is permitted 4 late days to use during the quarter.
- Each late day permits you to submit an assignment up to 24 hours late.
- You may use up to 2 late days per assignment.
- Each late day is atomic; for example, you cannot use 8 hours of a single late day on each of three assignments.

Homework 1