

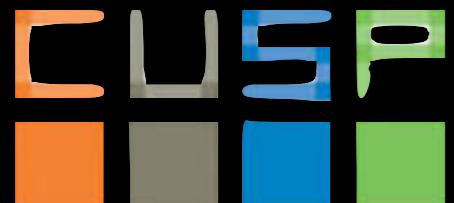
# Urban Informatics

Fall 2015

dr. federica bianco [fb55@nyu.edu](mailto:fb55@nyu.edu)



@fedhere



# Urban Informatics

Dr. federica bianco [fb55@nyu.edu](mailto:fb55@nyu.edu)

Office hours: Tue/Fri 1:30-3:30

Office: CUSP 1928, NYU Physics 545

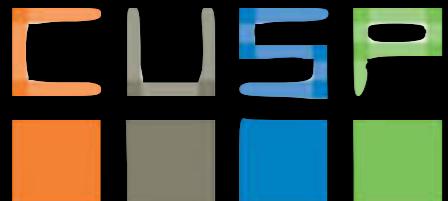
TAs:

Himanshu Kumawat (Tue) [hk1953@nyu.edu](mailto:hk1953@nyu.edu)

office hours: Tue/Wed 3-5

Saubreen Syedmajeed (Thu) [ss9570@nyu.edu](mailto:ss9570@nyu.edu)

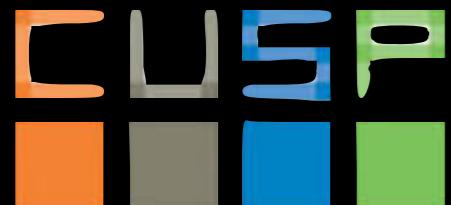
office hours: Mon/Thu 2-4

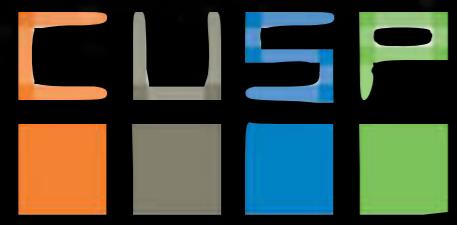


# Urban Informatics

Dr. federica bianco [fb55@nyu.edu](mailto:fb55@nyu.edu),  
astrophysicist

[http://blogs.teradata.com/international/  
sciences-loss-gain-data-science/](http://blogs.teradata.com/international/sciences-loss-gain-data-science/)



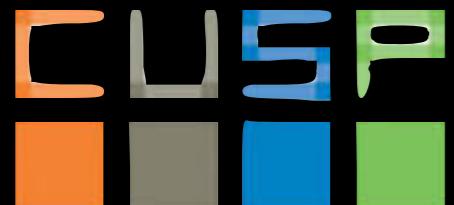


Urban Observatory

# Urban Informatics

Class website:

<http://cosmo.nyu.edu/~fb55/PUI2016>

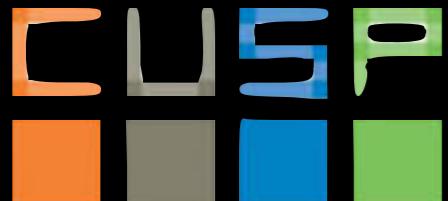


# Urban Informatics

Class: 3 hours, lecture + lab

Grade

- 5% on pre-class question
- 10 % class performance and participation
- 25 % homework
- 25 % midterm
- 35 % final



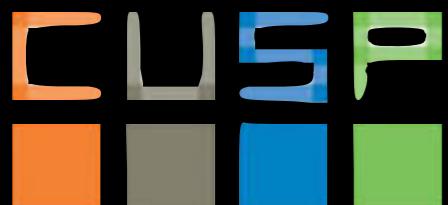
# Urban Informatics

Class: 3 hours, lecture + lab

Grade

- 5% on pre-class question
- 10 % class performance and participation
- 25 % homework
- 25 % midterm
- 35 % final

*from beginning of class to 5 minutes past the hour (be on time!)  
questions on previous class material AND READING ASSIGNMENTS*



# Urban Informatics

Class: 3 hours, lecture + lab

Grade

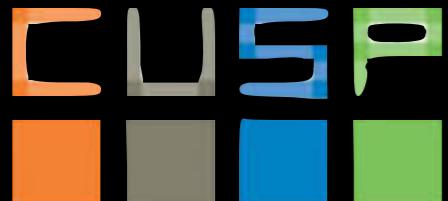
- 5% on pre-class question
- 10 % class performance and participation
- 25 % homework
- 25 % midterm
- 35 % final

ask questions

answer questions

get up and code

extra credit assignments



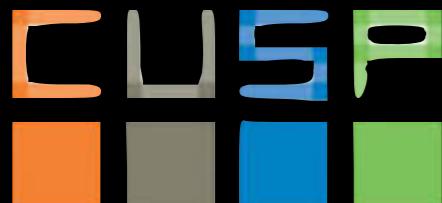
# Urban Informatics

Class: 3 hours, lecture + lab

Grade

- 5% on pre-class question
- 10 % class performance and participation
- 25 % homework
- 25 % midterm
- 35 % final

Homework projects must be turned in as iPython notebooks by checking them into your github account in the PUI2016\_<netID> repo and the project directories HW<hw number>\_<netID> (unless otherwise stated). <nyuid> is e.g. fb55



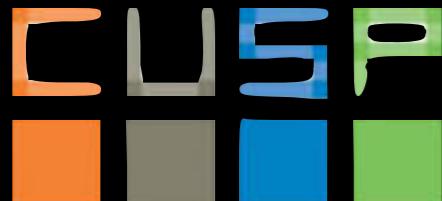
# Urban Informatics

Class: 3 hours, lecture + lab

Grade

- 5% on pre-class question
- 10 % class performance and participation
- 25 % homework
- 25 % midterm
- 35 % final

we encourage you to work in groups! but as a collaborative project where different group members lead different aspects of the work.  
A statement to describing your contribution to the project MUST be included in the README (a la Nature Magazine).



# Light echoes reveal an unexpectedly cool η Carinae during its nineteenth-century Great Eruption

A. Rest, J. L. Prieto, N. R. Walborn, N. Smith, F. B. Bianco, R. Chornock, D. L. Welch, D. A. Howell, M. E. Huber, R. J. Foley, W. Fong, B. Sinnott, H. E. Bond, R. C. Smith, I. Toledo, D. Minniti & K. Mandel

Affiliations | Contributions

Nature 482, 375–378 (16 Fe  
Received 26 August 2011 | A  
Brief Communication Arisin

## Contributions

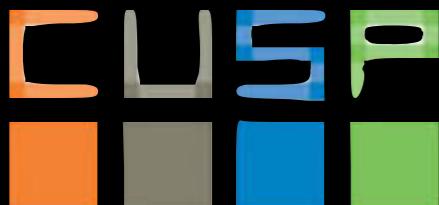
All authors contributed to the drafting of the paper. A.R., N.S. and R.C.S. imaged the area around η Car. A.R. and M.E.H. reduced the imaging data. H.E.B. provided images of the echoes that guided our spectroscopic pointings. J.L.P., R.C., R.J.F. and W.F. obtained the spectra and reduced them. A.R. and J.L.P. performed spectral analysis and interpretation. A.R., N.R.W. and F.B.B. performed spectral classification. F.B.B. and K.M. correlated the spectra. A.R., D.L.W. and B.S. modelled the light echo. I.T. and D.M. provided imaging of η Car. F.B.B. and D.A.H. provided the FTS images, and F.B.B. and A.R. reduced them.

Class: 3 hours, lecture + lab

Grade

- 5% on pre-class question
- 10 % class performance and participation
- 25 % homework
- 25 % midterm
- 35 % final

we encourage you to work in groups! but as a collaborative project where different group members lead different aspects of the work.  
A statement to describing your contribution to the project MUST be included in the README (a la Nature Magazine).



# Example of a README.md for a PUI homework: missing the README.md costs you 10% of the grade!

The README.md is a MarkDown (md) file. The syntax of a MarkDown is rather simple: <https://github.com/adam-p/markdown-here/wiki/Markdown-Cheatsheet>. Also the MD syntax can be used in Jupyter notebook cells to include text (not code) that is automatically formatted (which you will need to do over and over...)

## CitiBike HW - v1

### Question

Are CitiBike's easing commuter journey's across the East River?

### Hypothesis

- H0: The probability of a citibike subscriber crossing the East River in a given month is independent of whether the trip is taken during rush hour
- H1: The probability of a citibike subscriber crossing the East River in a given month is not independent of whether the trip is taken during rush hour

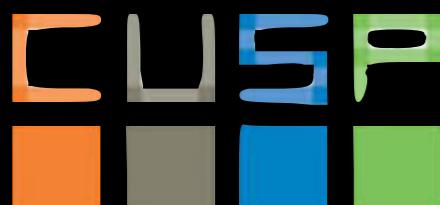
### Project work balance

#### hypothesis generation

Max, Arno, Clayton discussed and equally shared hypothesis generation. Max had the original idea of looking at bridges as he is an avid CitiBike user

#### Tasks

1. Clayton is tagging trips as cross east river or not
2. Max is defining historic hours as "on peak" or "not on peak"
3. Arno completes a chi-square test of our hypothesis



# Urban Informatics

Class: lecture + lab

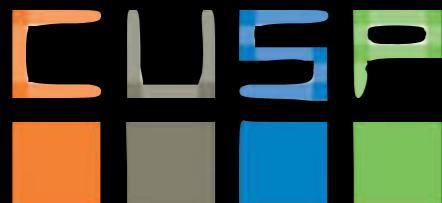
Grade

- 5% on pre-class question
- 10 % class performance and participation
- 25 % homework
- 25 % midterm
- 35 % final

After the midterm projects will be code reviewed by your peers.

We'll have 1 multi-week homework project from proposal to peer review.

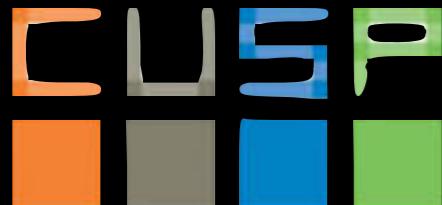
<https://blog.fogcreek.com/increase-defect-detection-with-our-code-review-checklist-example/>

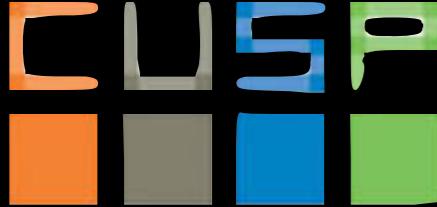


**Class:** lecture + lab  
**Grade**

- 5% on pre-class question
- 10 % class performance and participation
- 25 % homework
- 25 % midterm
- 35 % final

**Midterm and Final will include aspects of the work developed in the homework sessions. Failing to actively participate in the homework will result in not being able to get the Midterm and Final done.**





- Class: Grade
- 5%
  - 10
  - 25
  - 25
  - 35

This repository Search Pull requests Issues Gist

Unwatch 1 Star 0 Fork 0

Code Issues 0 Pull requests 0 Wiki Pulse Graphs Settings

No description or website provided. — Edit

61 commits 1 branch 0 releases 1 contributor

Branch: master New pull request Create new file Upload files Find file Clone or download

**fedhere committed on GitHub** Update README.md Latest commit a183019 2 minutes ago

**HW1\_fb55** Update README.md 21 hours ago

**Lab1\_fb55** Delete github\_create\_repo\_cmds.md 21 hours ago

**PEP8MinimalRequirements.md** Update PEP8MinimalRequirements.md an hour ago

**README.md** Update README.md 2 minutes ago

## PUI2016\_fb55

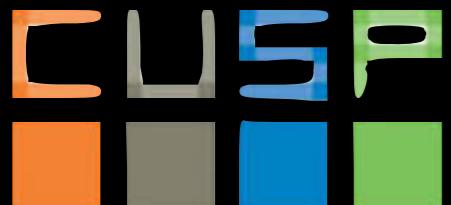
This repository contains the assignments for NYU CUSP Principles of Urban Informatics 2016. Check here for the new assignments, and for the solutions to be posted.

### GRADING GUIDELINES

- Each HW must be turned in as a directory in PUI2016\_<netID>.
- The directory HW<hw\_number>\_<netID> must have a README.md which states the student's participation. No penalty if the student declares not to have had any contribution but to have just followed and learned. However missing the README.md, about who the student worked with and what they did, or inconsistencies between the statements of students within the group that cannot be easily reconciled by asking will cost them 10% of the grade.
- Each assignment turned in as a notebook must have rendered plots with axis labels and captions. Each missing/non rendered plot, or plot without axes labels or caption will cost 10% of the grade.
- The notebook must be executables: the TA must download the notebook and run it cell by cell without errors. If

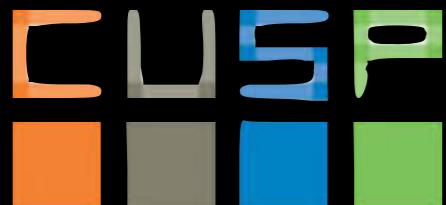
# Urban Informatics

## GOALS



# The workflow of a data driven project

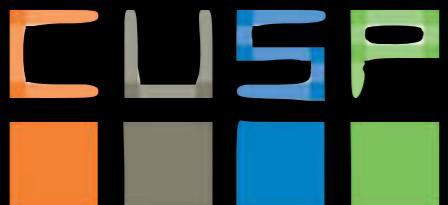
- IDEA
- dataset
  - define ideal data
  - figure out best data available
  - figure out if you can get new data
  - obtain data (including policy issues + technical issues)
- data handling
  - joining databases
  - formatting data
- exploratory data analysis
  - machine learning (clustering? dimensionality reduction?)
- statistics
  - models (regression)
  - prediction
  - validation (simulations)
- interpretation
- presentation
  - visualization
  - write a paper!



I: Good scientific practice  
& work flow

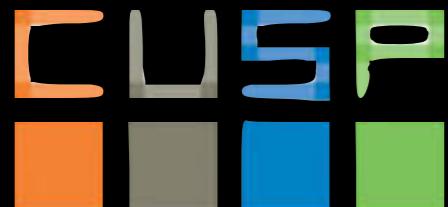
# The workflow of a data driven project

- IDEA
- dataset
  - define ideal data
  - figure out best data available
  - figure out if you can get new data
  - obtain data (including policy issues + technical issues)
- data handling
  - joining databases
  - formatting data
- exploratory data analysis
  - machine learning (clustering? dimensionality reduction?)
- statistics
  - models (regression)
  - prediction
  - validation (simulations)
- interpretation
- presentation
  - visualization
  - write a paper or give a talk



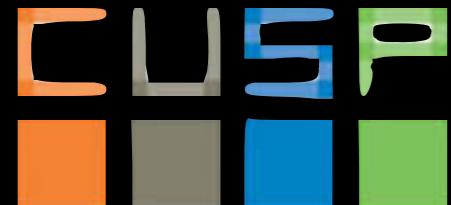
I: Good scientific practice  
& work flow

# *The phylosofical side of things*



I: Good scientific practice  
& work flow

# what is a scientific theory?

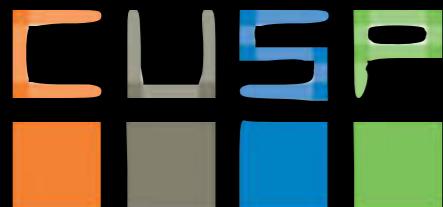


I: Good scientific practice  
& work flow

## The Demarcation Problem: a scientific theory must be *falsifiable*

My proposal is based upon an *asymmetry* between verifiability and falsifiability; an asymmetry which results from the logical form of universal statements. For these are never derivable from singular statements, but can be contradicted by singular statements.

— Karl Popper, *The Logic of Scientific Discovery*



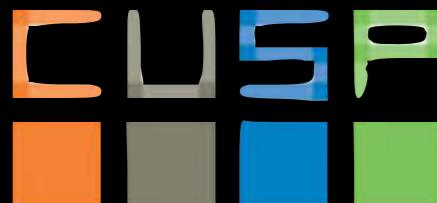
## The Demarcation Problem: a scientific theory must be *falsifiable*

My proposal is based upon an *asymmetry* between verifiability and falsifiability; an asymmetry which results from the logical form of universal statements. For these are never derivable from singular statements, but can be contradicted by singular statements.

— Karl Popper, *The Logic of Scientific Discovery*

things can get more complicated though:

most scientific theories are actually based largely on *probabilistic induction* and modern *inductive inference* (Solomonoff, frequentist vs Bayesian methods...)



**Ockham's razor: *Pluralitas non est ponenda sine necessitate***

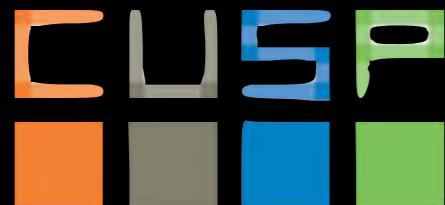
or “the law of parsimony”

William of Ockham (logician and Franciscan friar) 1300ca

but probably to be attributed to John Duns Scotus (1265–1308)

“Complexity needs not to be postulated without a need for it”

“Between 2 theories choose the simpler one”



the earth is round, and it orbits around the sun



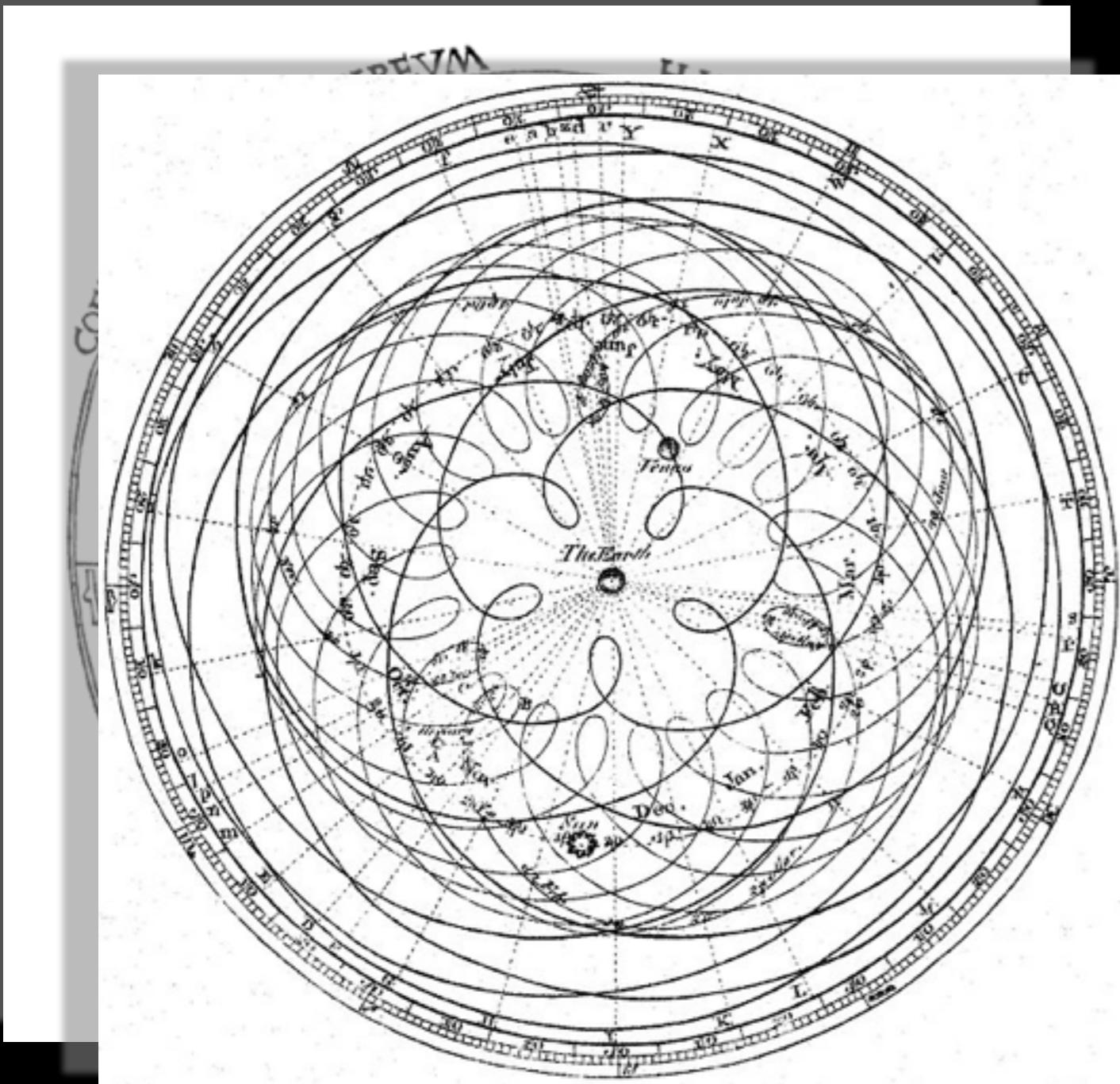
<http://en.wikipedia.org/wiki/File:Ptolemaicsystem-small.png>

Peter Apian, *Cosmographia*, Antwerp, 1524

from Edward Grant, "Celestial Orbs in the Latin Middle Ages", *Isis*, Vol. 78, No. 2. (Jun., 1987).

Geocentric models are natural:  
from our perspective  
we see the Sun  
moving, while we stay  
still

the earth is round, and it orbits around the sun



Source Encyclopaedia Britannica 1st Edition

Author Dr Long's copy of Cassini, 1777

As observations  
improve  
this model cannot fit  
the data anymore!  
*not easily anyways...*

the earth is round, and it orbits around the sun

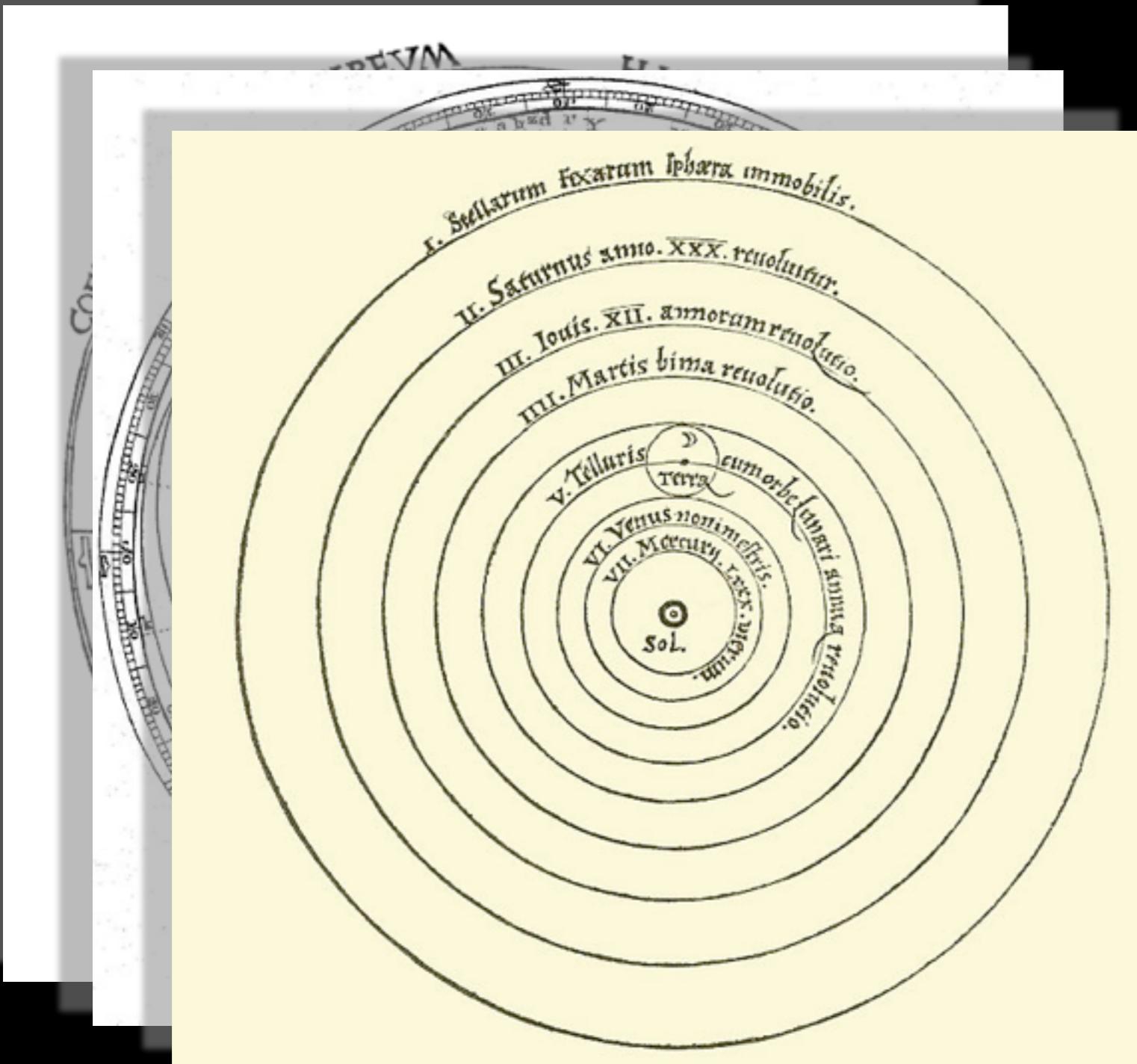


Image of heliocentric model from Nicolaus Copernicus' "De revolutionibus orbium coelestium".

A new model that is much simpler fit the data just as well (perhaps though only until better data comes...)

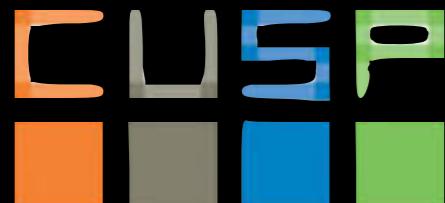
# Ockham's razor: *Pluralitas non est ponenda sine necessitate* or the law of parsimony

William of Ockham (logician and Franciscan friar) 1300ca

but probably to be attributed to John Duns Scotus (1265–1308)

“Complexity needs not to be postulated without a need for it”

“Between 2 theories choose the simpler one”

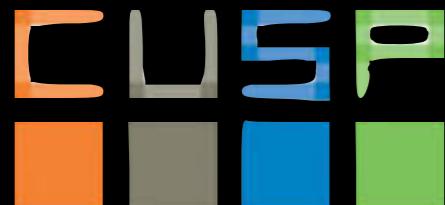


# Ockham's razor: *Pluralitas non est ponenda sine necessitate* or the law of parsimony

William of Ockham (logician and Franciscan friar) 1300ca  
but probably to be attributed to John Duns Scotus (1265–1308)

“Complexity needs not to be postulated without a need for it”

“Between 2 theories choose the one with fewer parameters!”



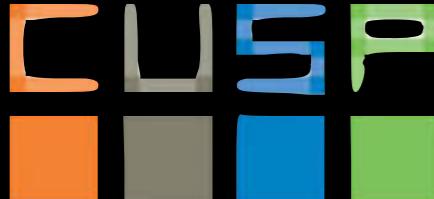
What is the question?

*the data speaks, if you know how to listen...*

Leek&Rodgers 2015 in Science

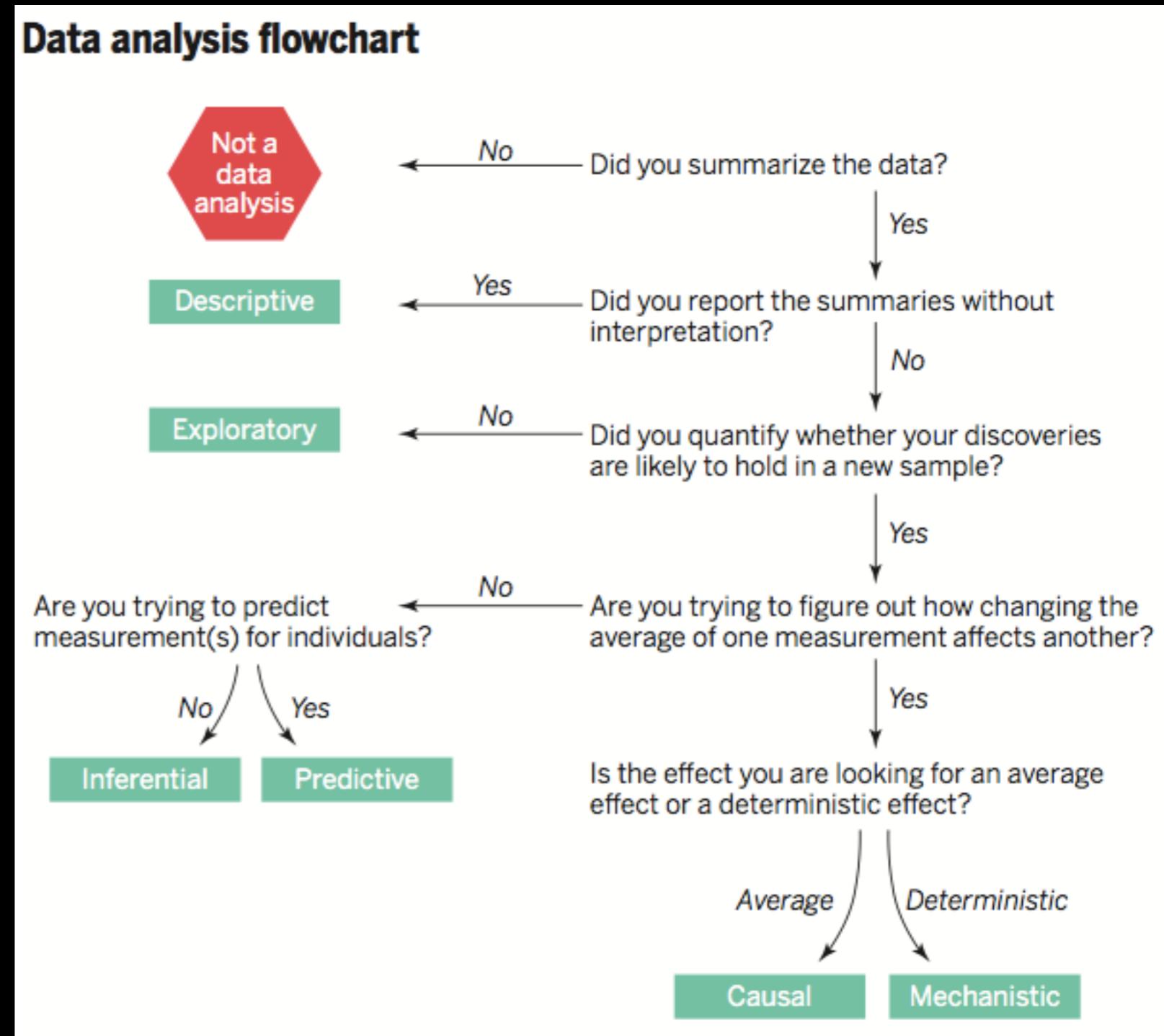
<http://www.sciencemag.org/content/347/6228/1314.full.pdf>

<http://moscow.sci-hub.bz/4d3cf57483ccf211f66cad18440023cd/10.1126%40science.aaa6146.pdf>



I: Good scientific practice  
& work flow

# What is the question?

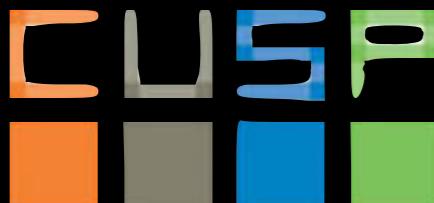


Leek&Rodgers 2015 in Science

<http://www.sciencemag.org/content/347/6228/1314.full.pdf>

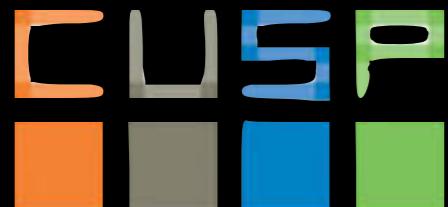
<http://moscow.sci-hub.bz/4d3cf57483ccf211f66cad18440023cd/>

10.1126%40science.aaa6146.pdf



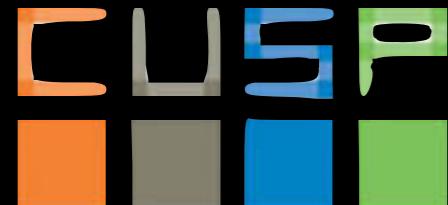
# I: Good scientific practice & work flow

# *The practical side of things*



I: Good scientific practice  
& work flow

# workflow: your environment



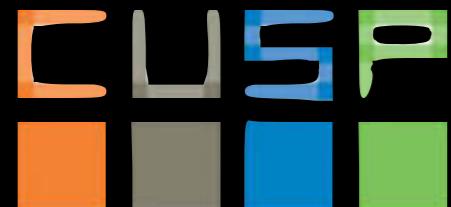
I: Good scientific practice  
& work flow

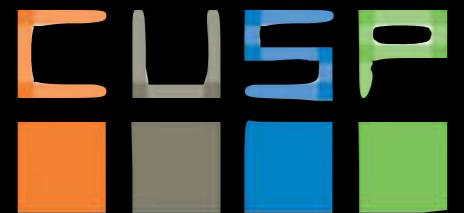
# Operating System:

Android, BlackBerry, BSD, Chrome OS, OS X, iOS, QNX, Linux  
Steam OS

Microsoft Windows, Windows Phone, and z/OS.

FreeRTOS, Micrium, and VxWorks





I: Good scientific practice  
& work flow

# Operating System:

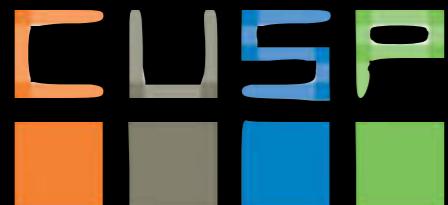
Android, BlackBerry, BSD, Chrome OS, OS X, iOS, QNX, Linux  
Steam OS



Where there is a shell, there is a way.

Microsoft Windows, Windows Phone, and z/OS.

FreeRTOS, Micrium, and VxWorks



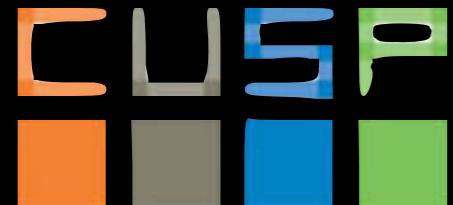
I: Good scientific practice  
& work flow

# Operating System:

OS X

Linux **UNIX**

Where there is a shell, there is a way.



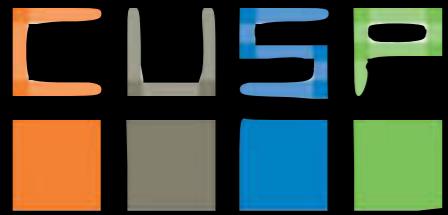
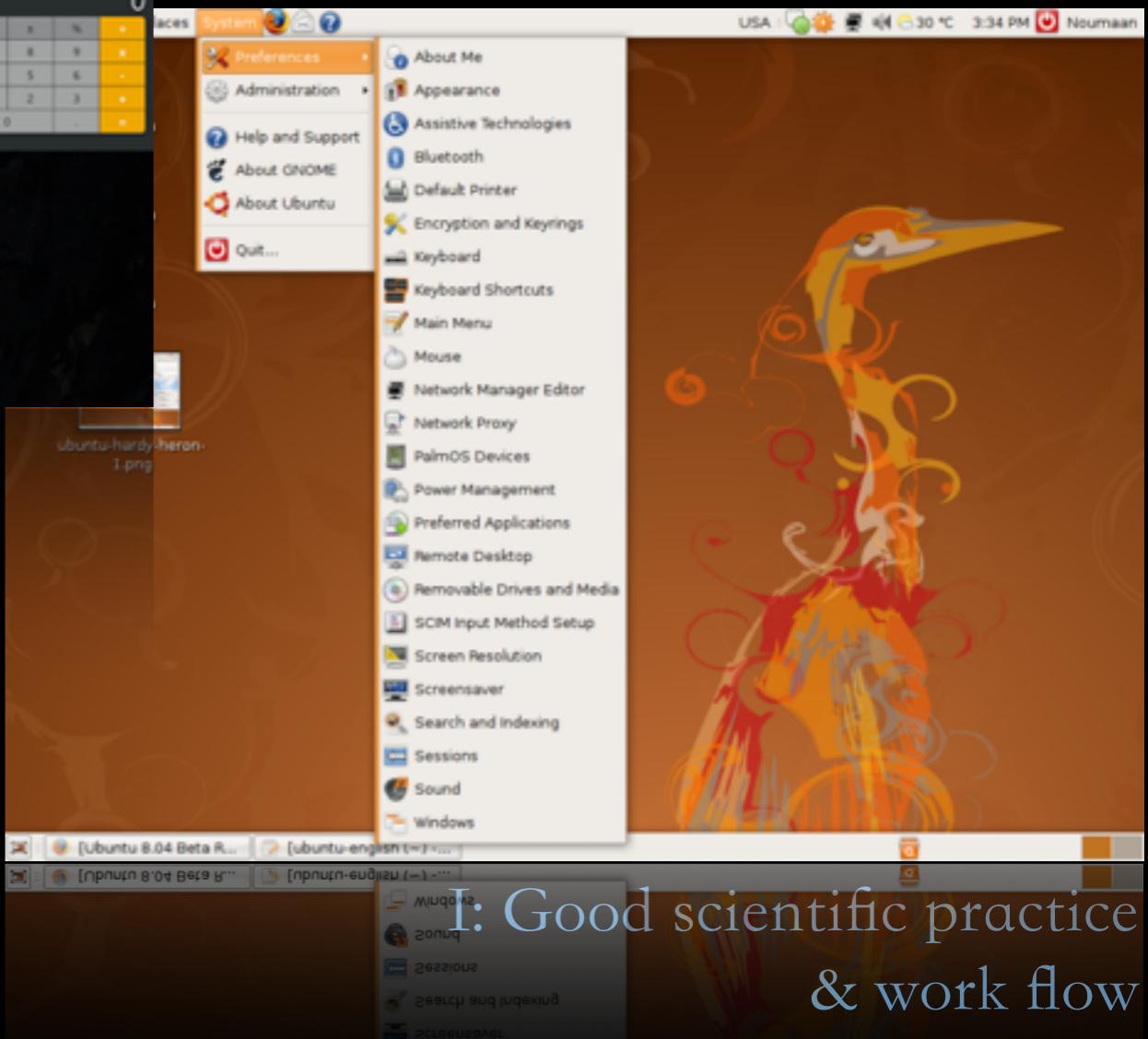
I: Good scientific practice  
& work flow

# Operating System:

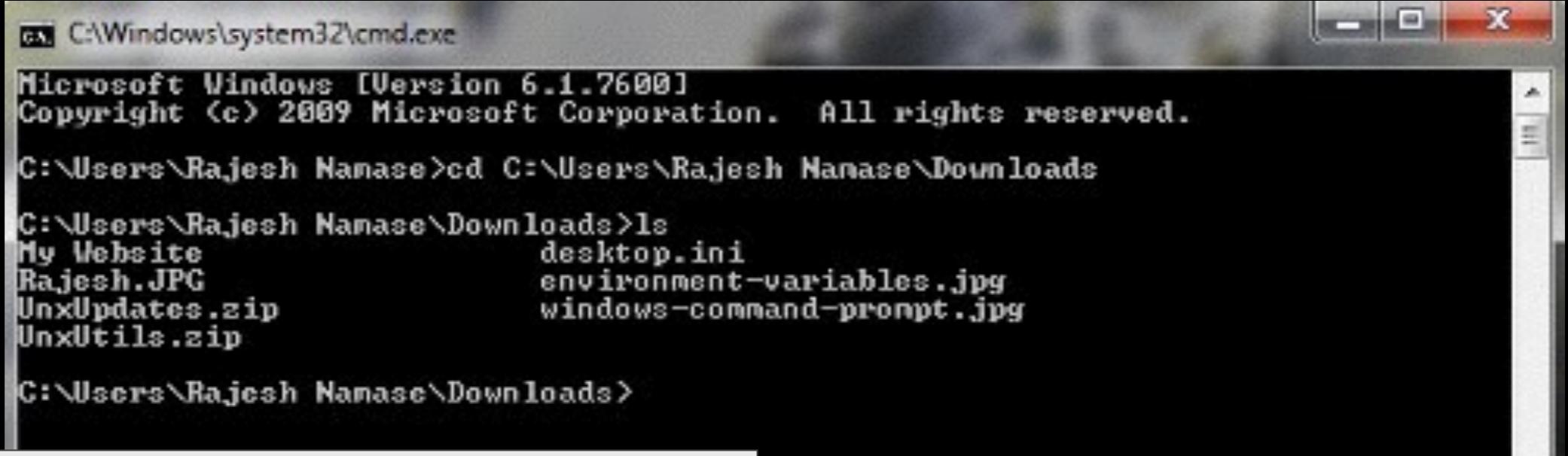
OS X



Linux



# Operating System:

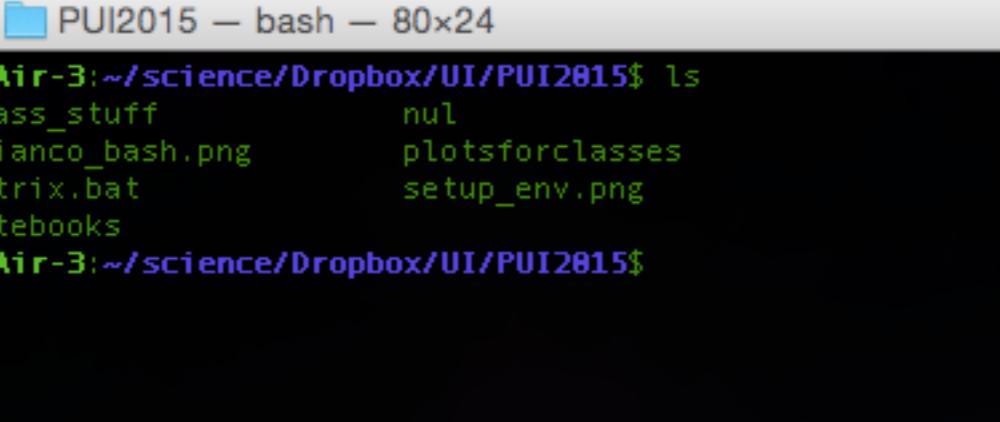


```
C:\Windows\system32\cmd.exe
Microsoft Windows [Version 6.1.7600]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.

C:\Users\Rajesh Namase>cd C:\Users\Rajesh Namase\Downloads

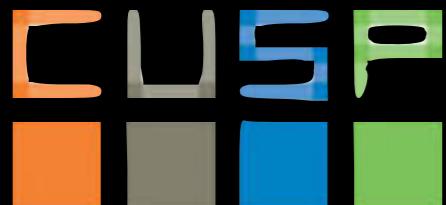
C:\Users\Rajesh Namase\Downloads>ls
My Website                               desktop.ini
Rajesh.JPG                                environment-variables.jpg
UnxUpdates.zip                           windows-command-prompt.jpg
UnxUtils.zip

C:\Users\Rajesh Namase\Downloads>
```

```
PUI2015 – bash – 80x24
fbianco@Federicas-MacBook-Air-3:~/science/Dropbox/UI/PUI2015$ ls
CityChallenge      class_stuff        nul
PUI2015labs        fbianco_bash.png   plotsforclasses
README.md          matrix.bat       setup_env.png
class_slides       notebooks
```

<https://speakerdeck.com/62gerente/bash-introduction>



I: Good scientific practice  
& work flow

# You have a *bash* shell on Compute (which is Linux)

- and you have access to compute via ssh

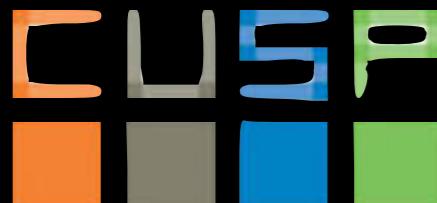
```
ssh -X -A -t cuspid@gw.cusp.nyu.edu ssh -A -X compute
```

- or in the Green Environment of the CUSP data facility via remote desktop

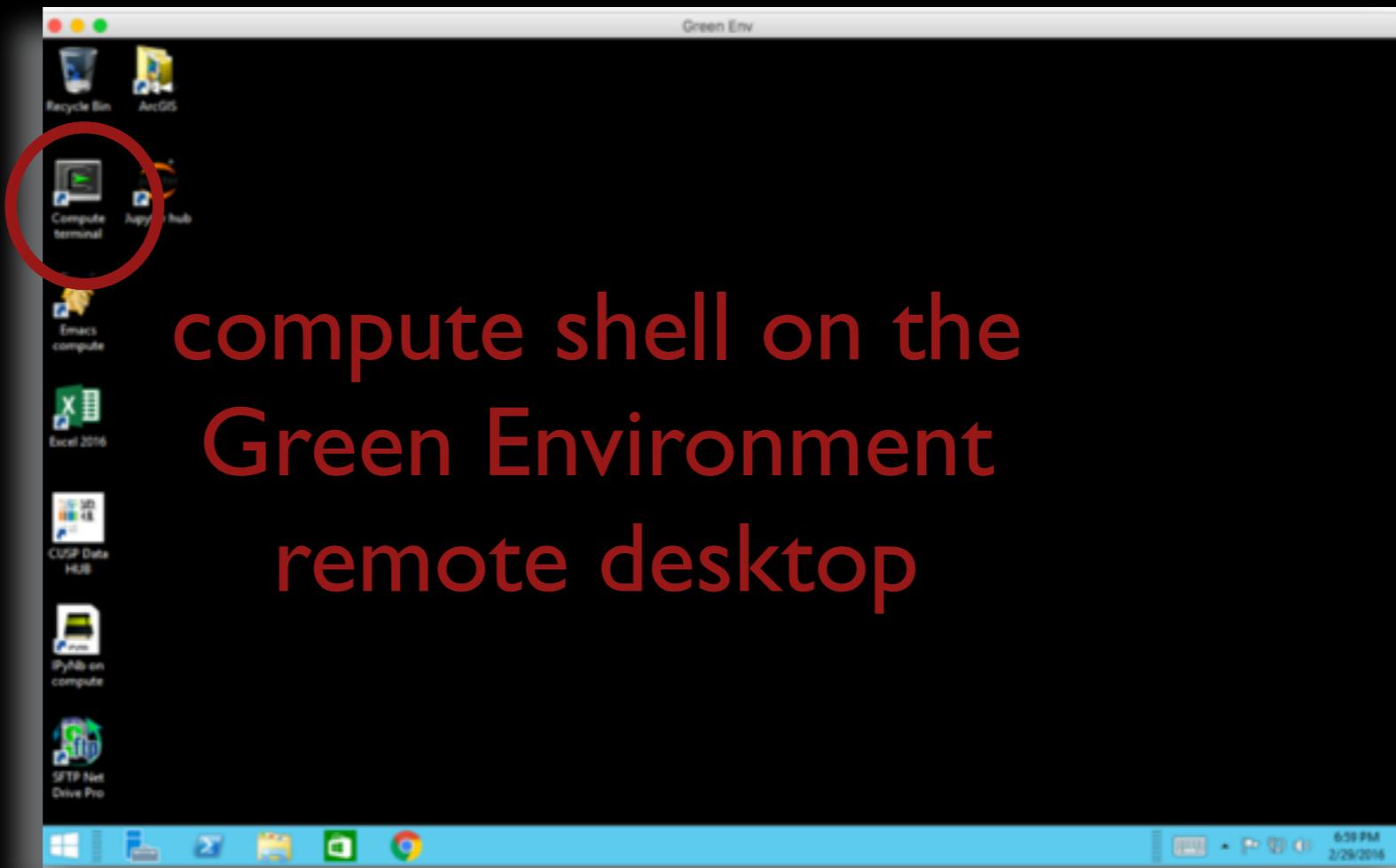
```
ssh cuspid@gw.cusp.nyu.edu -L9000:wingrdp.cusp.nyu.edu:3389
```

followed by remote desktop to localhost:9000

[https://datahub.cusp.nyu.edu/  
computing.html#accessing\\_the\\_workspace](https://datahub.cusp.nyu.edu/computing.html#accessing_the_workspace)

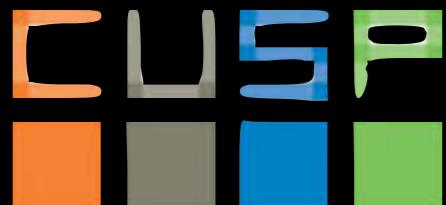


# You have a *bash* shell on Compute (which is Linux)



compute shell on the  
Green Environment  
remote desktop

Remote desktop acces in Green environmen



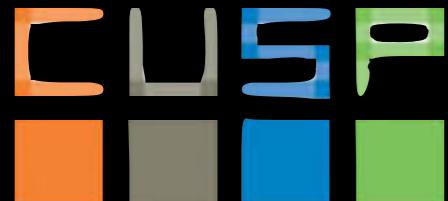
I: Good scientific practice  
& work flow

Operating System:

# Shell commands & Environmental variables demo



<https://speakerdeck.com/62gerente/bash-introduction>



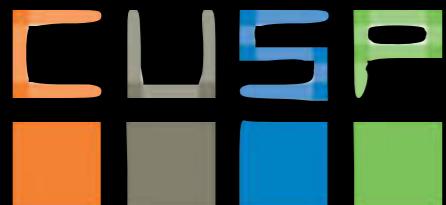
I: Good scientific practice  
& work flow

# Operating System:

# Shell commands & Environmental variables demo

# essential commands

pwd	ssh
ls	scp
mkdir	rsync
cd	df
touch	du
cp	
mv	alias
rm	export
less	Wildcards
echo	I/O Redirection
head/tail	Standard Output
top	Standard Input
ps	Pipes
bg	Listing your processes
chmod	Killing a process
chown	
grep	



**getting to Code:**

**Python, iPython, iPython notebooks**

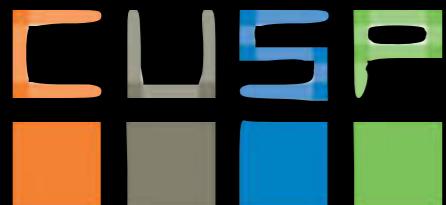
**You should be fluent in *at least Python or R***

**to be competitive on the job market**

**In this class we will only work in Python.**

**All homework should be developed in**

**python and delivered through github.**



I: Good scientific practice  
& work flow

# getting to Code: Python, iPython, iPython notebooks

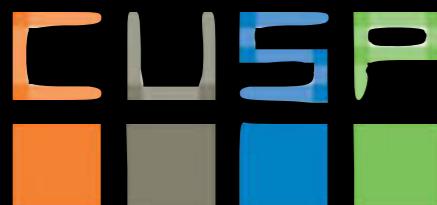
Python 2.7 vs Python 3.0

I will write in Python2.7 for compatibility (e.g. GeoPandas)

I will use the future package

```
from __future__ import print_function  
__author__ = "Federica B. Bianco, CUSP NYU 2016"
```

to make the code forward compatible with Python 3.0  
(though some lines of code may be broken in Python 3.0)



# Choosing a text Editor: Integrated or not?

**Emacs:** “extensible, customizable, self-documenting real-time display editor”

customizable via the .emacs file

can be run without tunneling with `emacs -nw`

the command sequences are tricky (customize emacs for python

<https://realpython.com/blog/python/emacs-the-best-python-editor/>)

**PyCharm, Sublime, Brakets:** Integrated Development

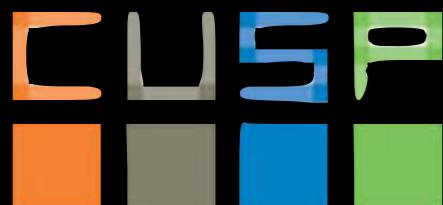
Environment: checks syntax and standard compliancy

can be run directly within the developing window

recognizes the syntax for other codes (Django)

integrated with version control

**Jupyter Notebooks:** Browser-based interactive computational



environment  
**demo time!**

I: Good scientific practice  
& work flow

# Choosing a text Editor: Integrated or not?

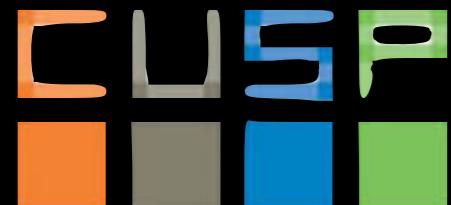
Emacs:

PyCharm:

demo time!

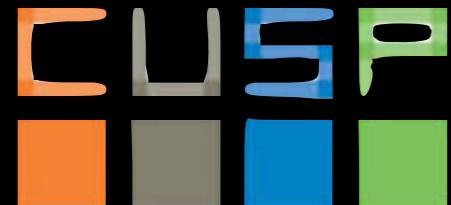
• jupyter

Jupyter Notebooks:



I: Good scientific practice  
& work flow

# reproducible research

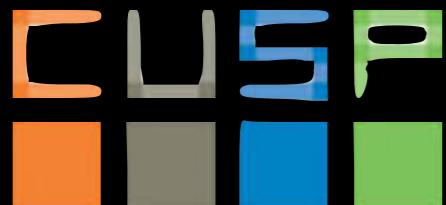


I: Good scientific practice  
& work flow

## Reproducible research means:

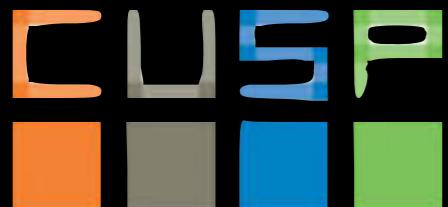
all numbers in a data analysis can be recalculated exactly (down to stochastic variables!) using the **code** and **raw data** provided by the analyst.

Claerbout, J. 1990,  
Active Documents and Reproducible Results, Stanford Exploration Project  
Report, 67, 139



# Reproducible research means:

code      raw data



I: Good scientific practice  
& work flow

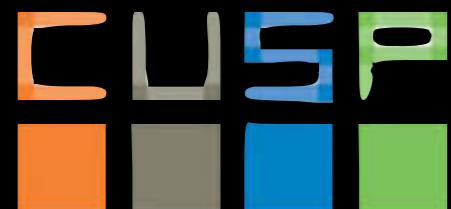
# Reproducible research means:

code

raw data



<https://github.com/>



I: Good scientific practice  
& work flow

## Reproducible research means:

code

raw data



<https://github.com/>

distributed version control system:  
a version of the files on your local computer is  
made also available at a central server.  
The history of the files is saved remotely so  
that any version (that was checked in) is  
retrievable.

Others can access and generate their versions  
of the files enabling collaborative work.

# Reproducible research means:

code

raw data



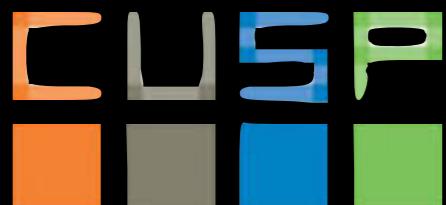
other version control systems:  
RCS

CVS (Centralized version control system)  
Subversion  
SVN

Git (<https://github.com>)

Mercurial (<https://bitbucket.org/>)

<https://git-scm.com/book/en/v2/Getting-Started-About-Version-Control>



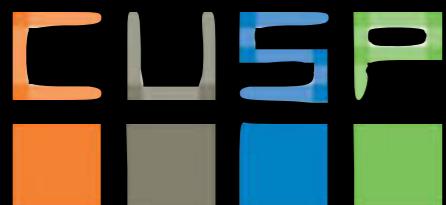
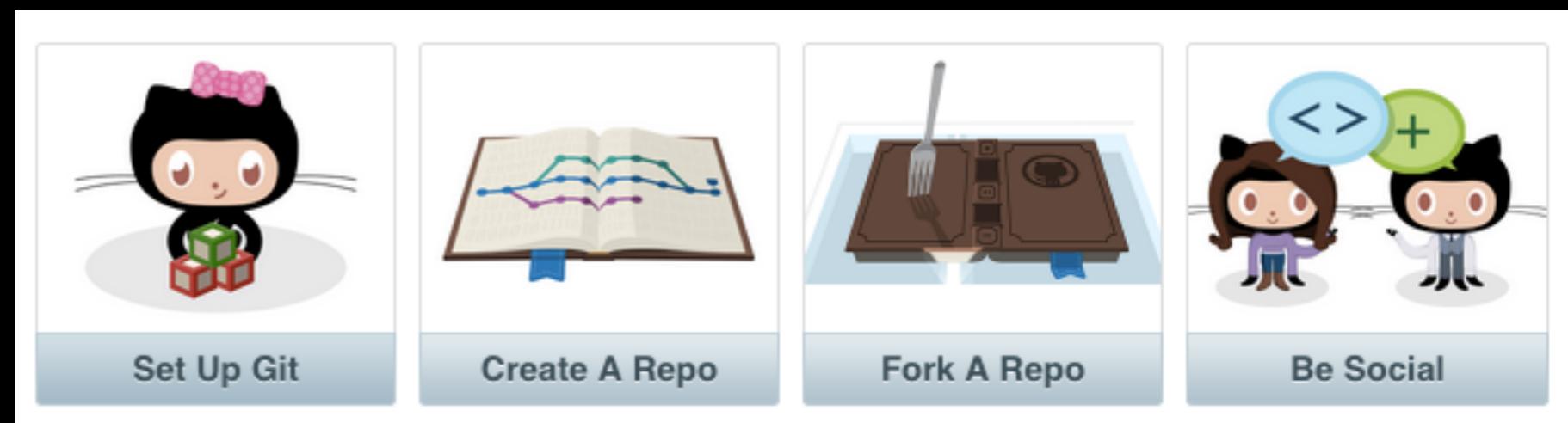
# Reproducible research means:

code

raw data



<https://github.com/>



I: Good scientific practice  
& work flow

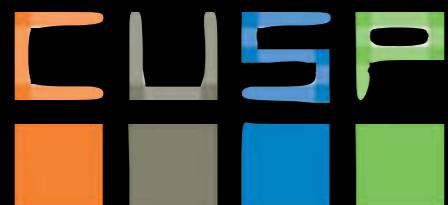
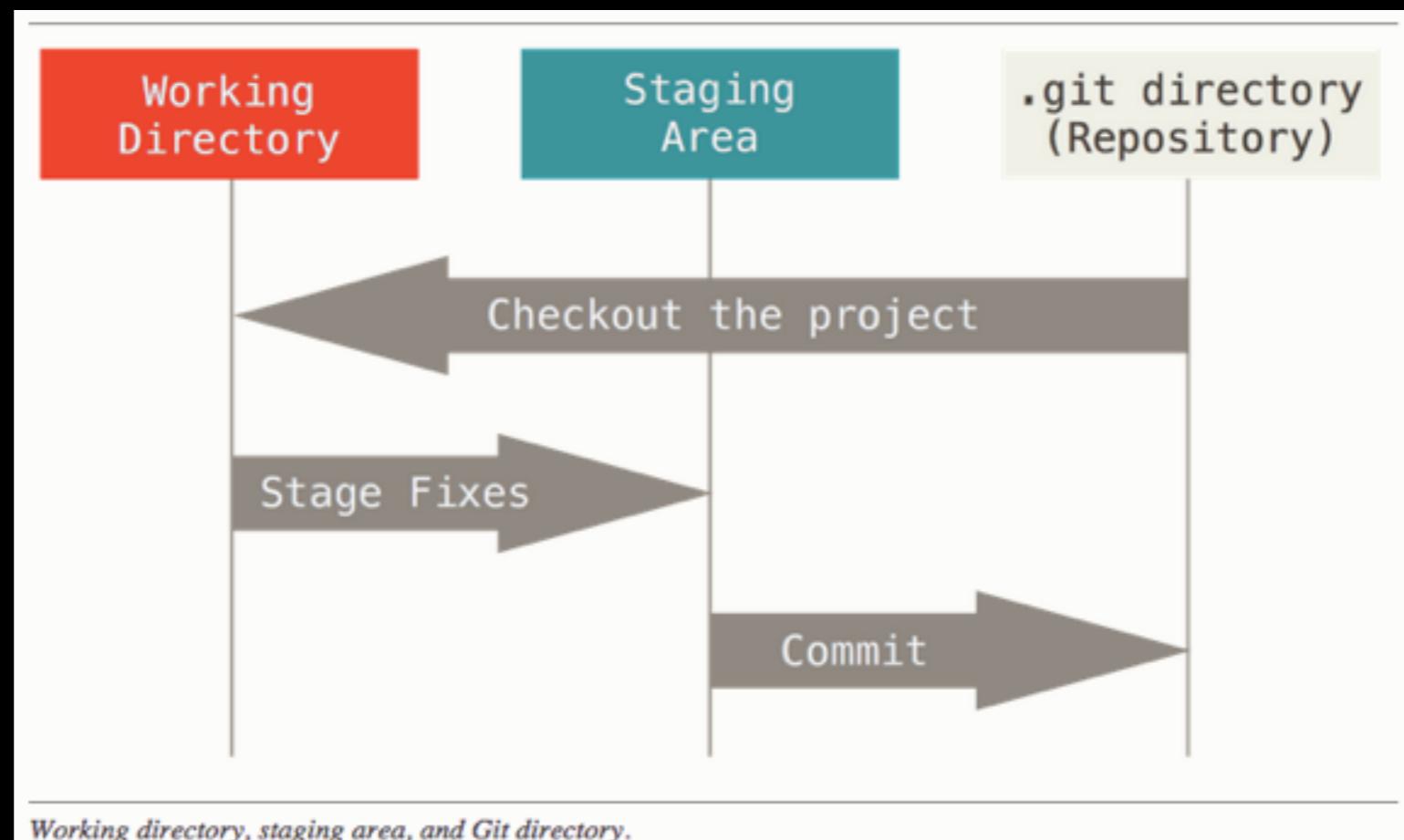
# Reproducible research means:

code

raw data



<https://github.com/>



I: Good scientific practice  
& work flow

# Reproducible research means:

code

raw data



<https://github.com/>

markdowns & standards:

**in order for your research to be  
reproducible it has to be understandable:**

- Paper or slides
- Repository Markdown files
- Understandable (PEP8 compliant) code -  
explicit declare the version of the code!

# Reproducible research means:

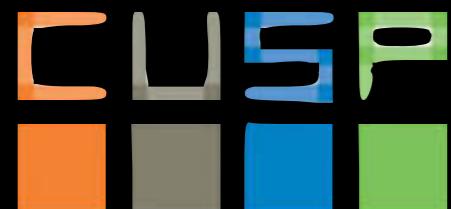
code

raw data



<https://github.com/>

markdowns & standards:



I: Good scientific practice  
& work flow

# Reproducible research means:

code

raw data



<https://github.com/>

markdowns & standards:



This is a markdown file guiding you through the very first steps to create and manage a git repo with github.

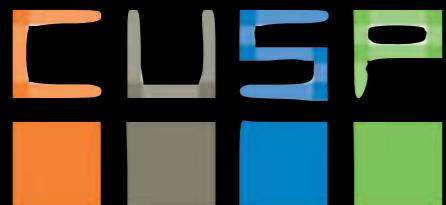
Lets start on your bash shell

Create a directory

```
Franco@Federicas-MacBook-Air-2:~/science/Dropbox/81S mkdir gittest_fbianco
```

```
mkdir gittest_fbianco
```

[https://github.com/fedhere/PUI2016\\_fb55/blob/master/Lab1\\_fb55/githubCreateRepoCmds.md](https://github.com/fedhere/PUI2016_fb55/blob/master/Lab1_fb55/githubCreateRepoCmds.md)



: Good scientific practice  
& work flow

# Reproducible research means:

code

raw data



<https://github.com/>

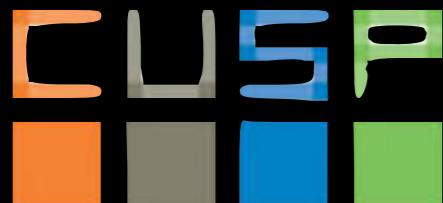
markdowns & standards:



PEP8: Python Enhancement Proposals 8

“This document gives coding conventions for the Python code comprising the standard library in the main Python distribution.”

**Readability counts.**



I: Good scientific practice  
& work flow

# Reproducible research means:

code

raw data



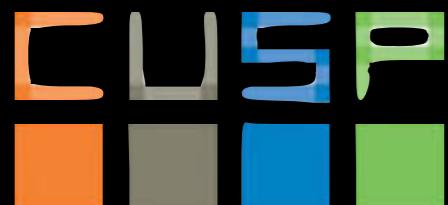
<https://github.com/>

markdowns & standards:



PEP8: Python Enhancement Proposals 8

Indentation, Tabs or Spaces?, Maximum Line Length, Blank Lines, Source File Encoding, Imports, Whitespace in Expressions and Statements, Comments Bookkeeping, Naming



I: Good scientific practice  
& work flow

# Reproducible research means:

code

raw data



<https://github.com/>



a good video tutorial

<https://www.youtube.com/watch?v=ZDR433b0HJY>

# Reproducible research means:

code

raw data



<https://github.com/>

let's make a repo!

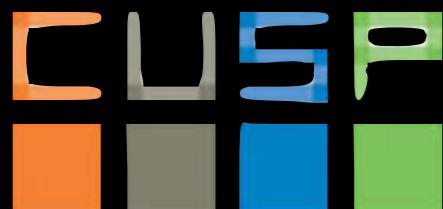


# Reproducible research means:

code      raw data

privacy concerns

in order for your research to be reproducible  
the data you use must be accessible BUT  
THAT IS NOT ALWAYS POSSIBLE:  
CUSP has access to data that has restricted  
access. Share your data when possible!



I: Good scientific practice  
& work flow

# Reproducible research means:

code raw data

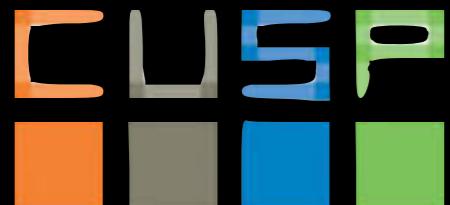
## Remove sensitive data

Some day you or a collaborator may accidentally commit sensitive data, such as a password or SSH key, into a Git repository. Although you can remove the file from the latest commit with `git rm`, the file will still exist in the repository's history. Fortunately, there are other tools that can entirely remove unwanted files from a repository's history. This article will explain how to use two of them: `git filter-branch` and the [BFG Repo-Cleaner](#).

**Danger:** Once you have pushed a commit to GitHub, you should consider any data it contains to be compromised. If you committed a password, change it! If you committed a key, generate a new one.

This article tells you how to make commits with sensitive data unreachable from any branches or tags in your GitHub repository. However, it's important to note that those commits may still be accessible in any clones or forks of your repository, directly via their SHA-1 hashes in cached views on GitHub, and through any pull requests that reference them. You can't do anything about existing clones or forks of your repository, but you can permanently remove all of your repository's cached views and pull requests on GitHub by contacting [GitHub support](#).

<https://help.github.com/articles/remove-sensitive-data/>



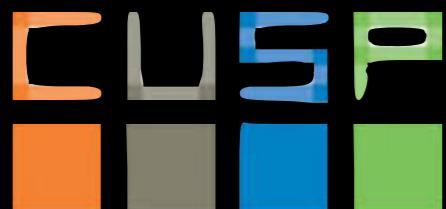
I: Good scientific practice  
& work flow

# Reproducible research:

## How to share your data

- Share/Reference the source of your raw data
- Share the “tidy” data
- Share the code used to process the data at each step

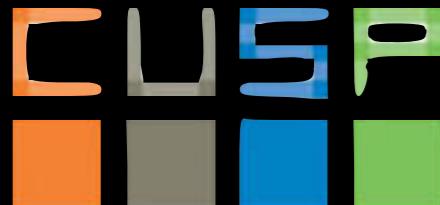
example time



I: Good scientific practice  
& work flow

## Key Concepts:

- falsifiability and law of parsimony
- types of scientific questions
- reproducible research
- PEP8 and style standards
  
- work with github
- understand how to set up your environment
- basic bash commands
- creating and checking into github an ipython notebook



## Resources:

Karl Popper, J. 1934,

**The Logic of Scientific Discovery**

<http://strangebeautiful.com/other-texts/popper-logic-scientific-discovery.pdf>

Jeff Leek & Rodger Peng. 2015,

**What is the Question? ASSIGNED READING**

<http://www.sciencemag.org/content/347/6228/1314.summary>

<moscow.sci-hub.bz/4d3cf57483ccf211f66cad18440023cd/10.1126@science.aaa6146.pdf>

Claerbout, J. 1990,

**Active Documents and Reproducible Results,  
Stanford Exploration Project Report, 67, 139**

[http://sepwww.stanford.edu/data/media/public/docs/sep67/jon2/paper\\_html/](http://sepwww.stanford.edu/data/media/public/docs/sep67/jon2/paper_html/)

Jeff Leek, 2015

**The Elements of Data Analytic Style**

<https://leanpub.com/datastyle> (\$10.00) and <https://github.com/jtleek/datasharing>

Guido van Rossum, Barry Warsaw, Nick Coghlan, 2001

Proposal Enhancement for Python

<https://www.python.org/dev/peps/pep-0008/>

