Data Science with Python

Phil Mui
@philmui

Today

- **Hello**
- Data Science: what is it?
- Course Overview
- Python: codeacademy
- Jupyter / ipython
- Git / VirtualEnv
- Check-in

Today

- Hello
- **Data Science: what is it?**
- Course Overview
- Python: codeacademy
- Jupyter / ipython
- Git / VirtualEnv
- Check-in

What is Data Science?

- What words come to mind when you think of Data Science?
- What experience do you have with Data Science?
- Why are you taking an Introduction to Data Science Class?

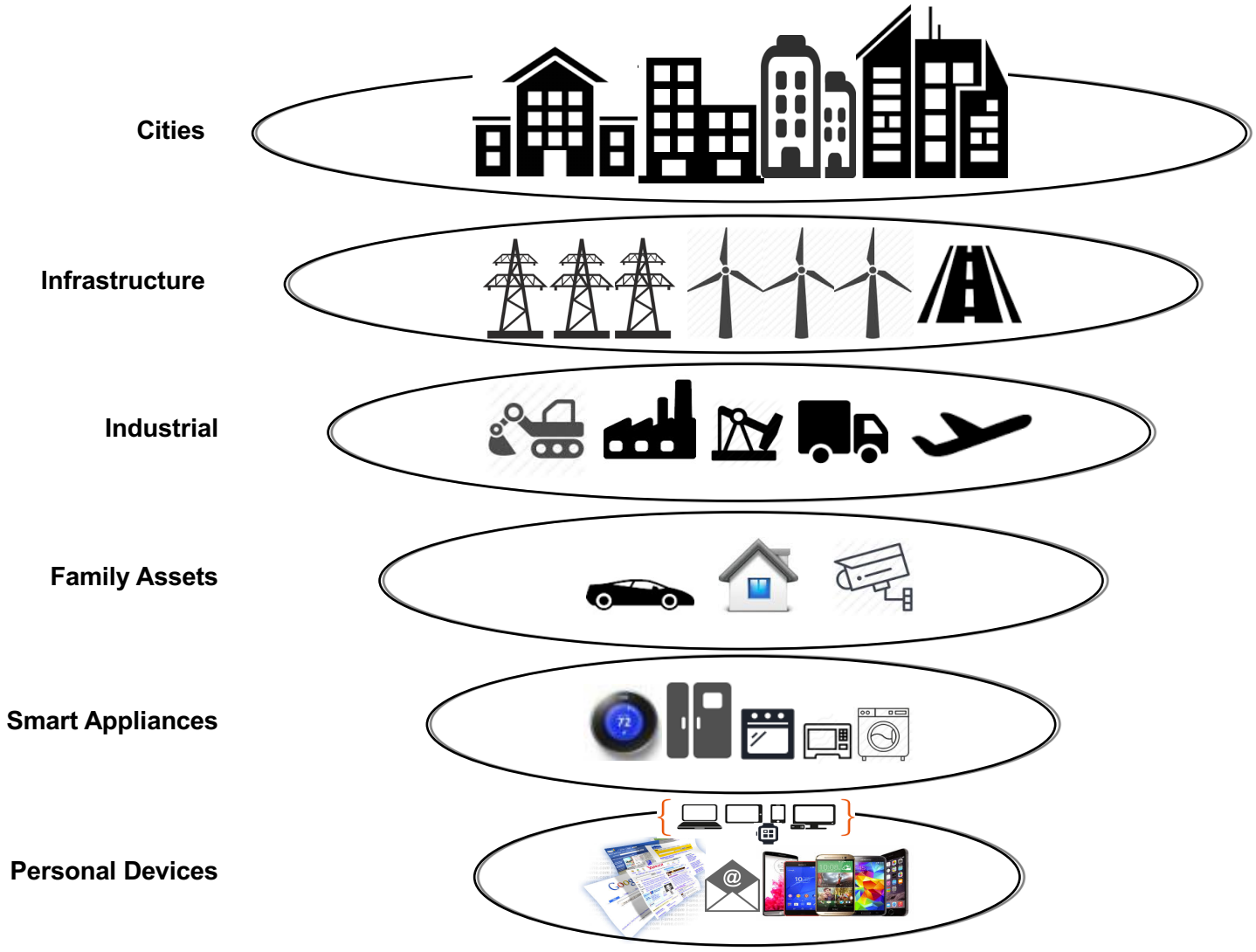
“Data”


msn
YAHOO!
Google
i-une.com


msn
YAHOO!


msn
YAHOO!
10:08
10:35


msn
YAHOO!
10:08
10:35


Cities
Infrastructure
Industrial
Family Assets
Smart Appliances
Personal Devices

HBR.ORG
Harvard Business Review
OCTOBER 2012
REPRINT R1210C

SPOTLIGHT ON BIG DATA

Big Data: The Management Revolution

Exploiting vast new flows of information can radically improve your company's performance. But first you'll have to change your decision-making culture.

by Andrew McAfee and Erik Brynjolfsson

Data Scientist:

The Sexiest Job of the 21st Century


sas
SPSS
AN IBM COMPANY
webtrends
mixpanel
OMNITURE
COREMETRICS
Localytics
REVOLUTION ANALYTICS
R
Google Analytics
DOMO
Spark
ORACLE
SAP
TERADATA
hortonworks
IBM
ASTER
DATA
ClearStory
hadoop
eBureau
GoodData
AnalyticsIQ Inc.
Polk
tableau
cloudera
Greenplum
IXI™ SERVICES
NETEZZA
causa
Quantivo
alteryx
VERTICA
nielsen
Semcasting

What is Data Science?

- "How Companies Learn Your Secrets" *NYT*, by Charles Duhigg, February 16, 2012
- http://bit.ly/targetdata

http://bit.ly/targetdata

What did Target Do?

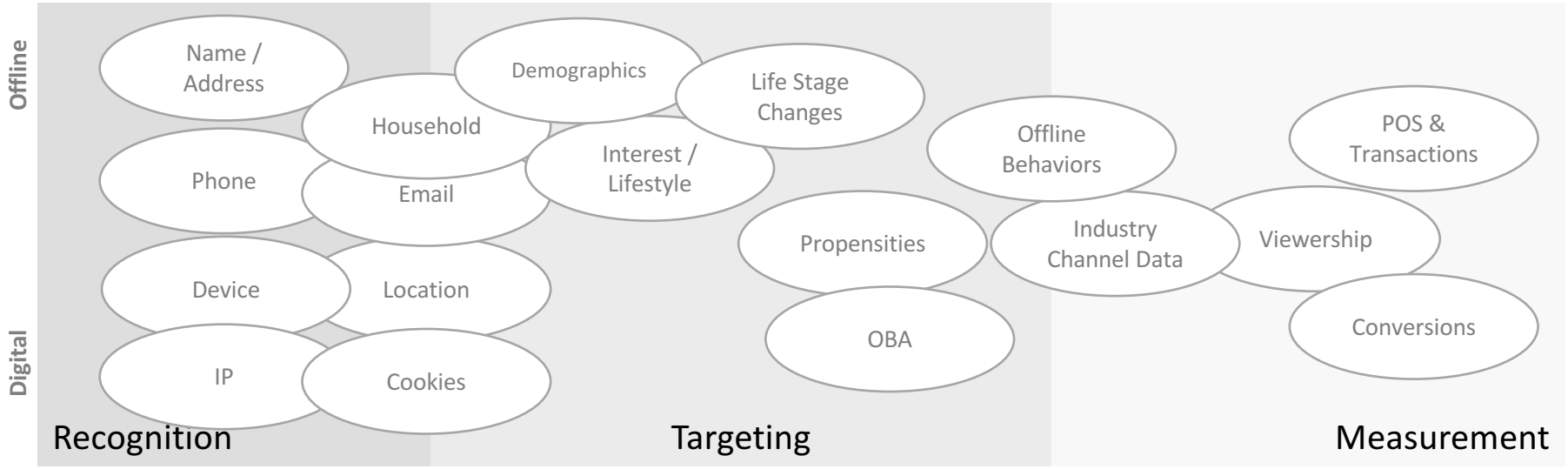
- Mining of data on shopping patterns
  - Specific products purchased
  - Combination of products purchased
  - Combined with demographic and other data
- Psychology and neuroscience
  - Habits:
    - Cue-routine-reward
    - When are habits open to change?

Lessons from Target

- Yes, Data Science is about mining data
- There are deeper theoretical issues involved in understanding what you find
- Left out of that long article are most of the critical steps that precede the analysis
- In short, Data Science > data mining

Consumer data landscape

A framework for understanging consumer data



Consumer data landscape

A framework for understanging consumer data

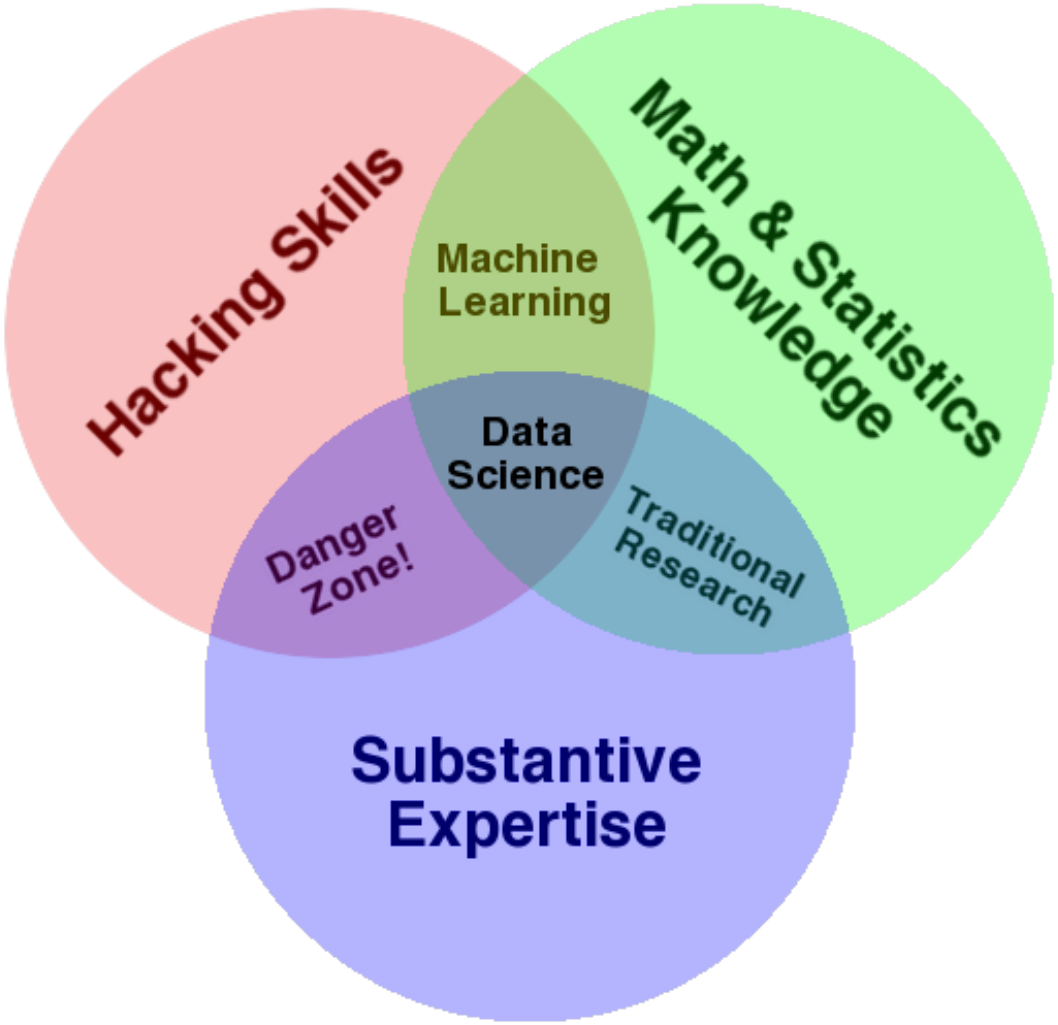
Non-industry specific → Industry specific

Not outcome predictable → High outcome predictable

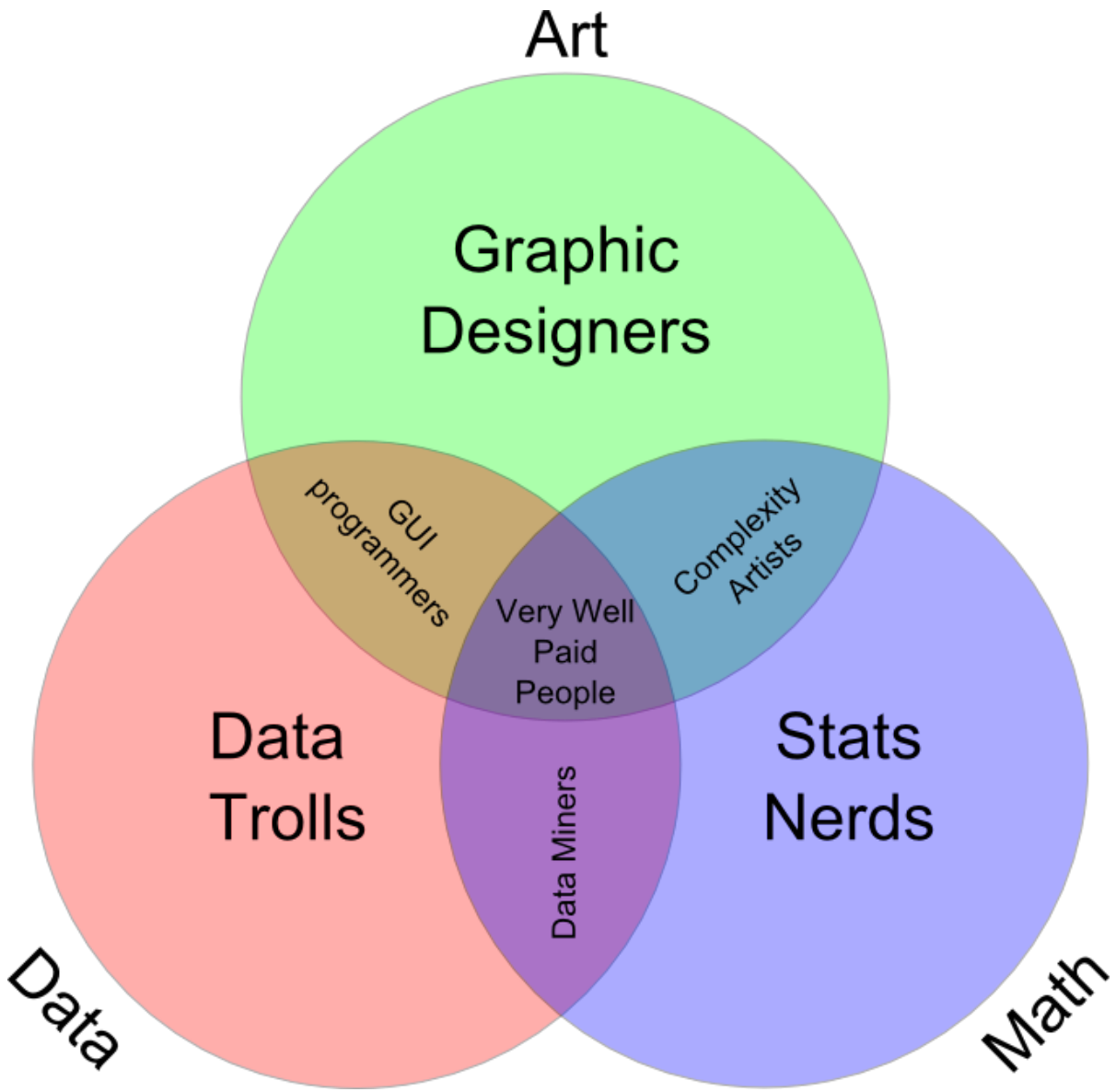
Definition of Data Science

- There are many, but most say data science is:
  - Broad – broader than any one existing discipline
  - Interdisciplinary: Computer Science, Statistics, Information Science, databases, mathematics
  - Applied focus on extracting knowledge from data to inform decision making.
  - Focuses on the skills needed to collect, manage, store, distribute, analyze, visualize, and reuse data.
- There are many visual representations of Data Science

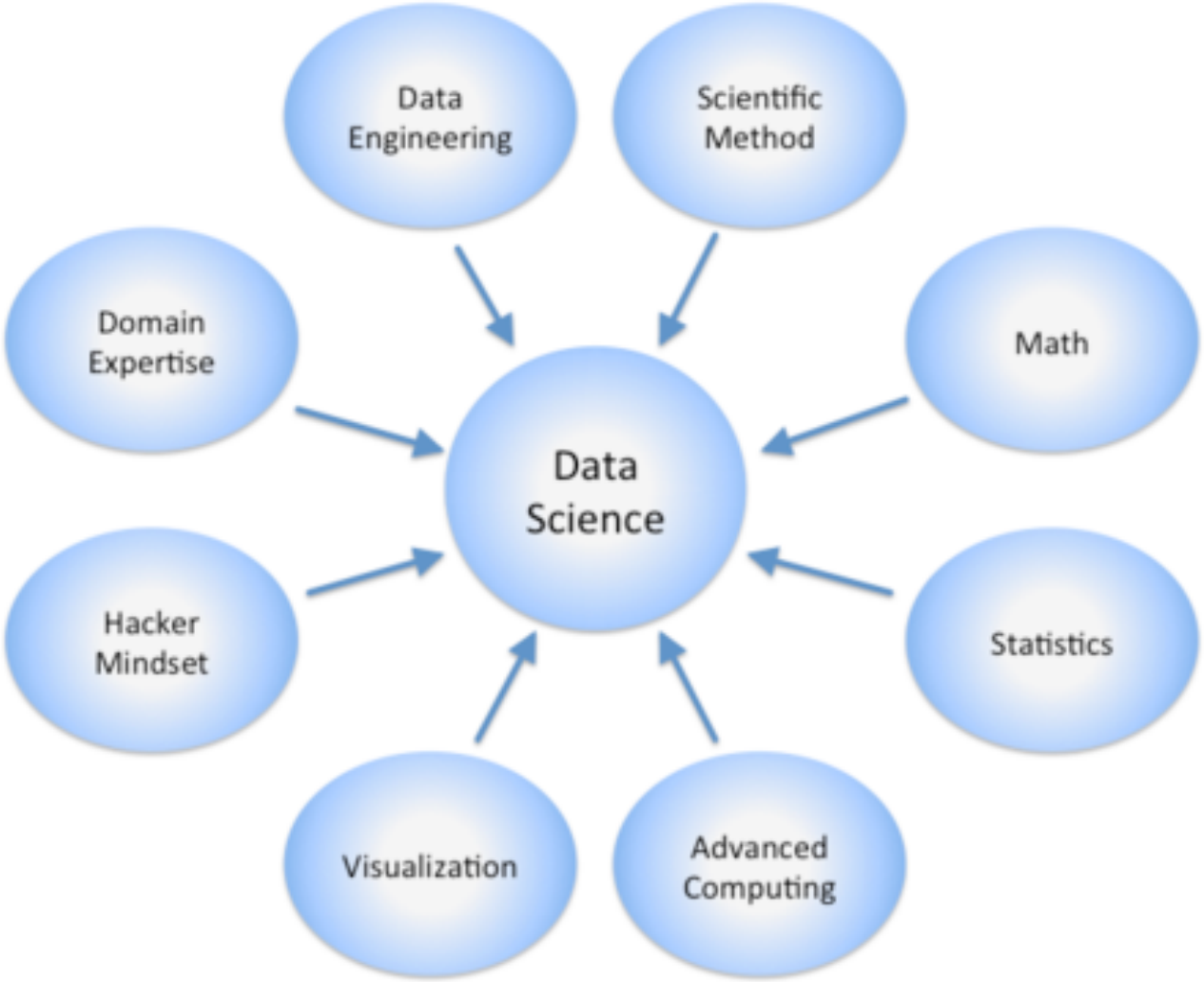
Some definitions link computational, statistical, and substantive expertise.



Other definitions focus more on technical skills alone.



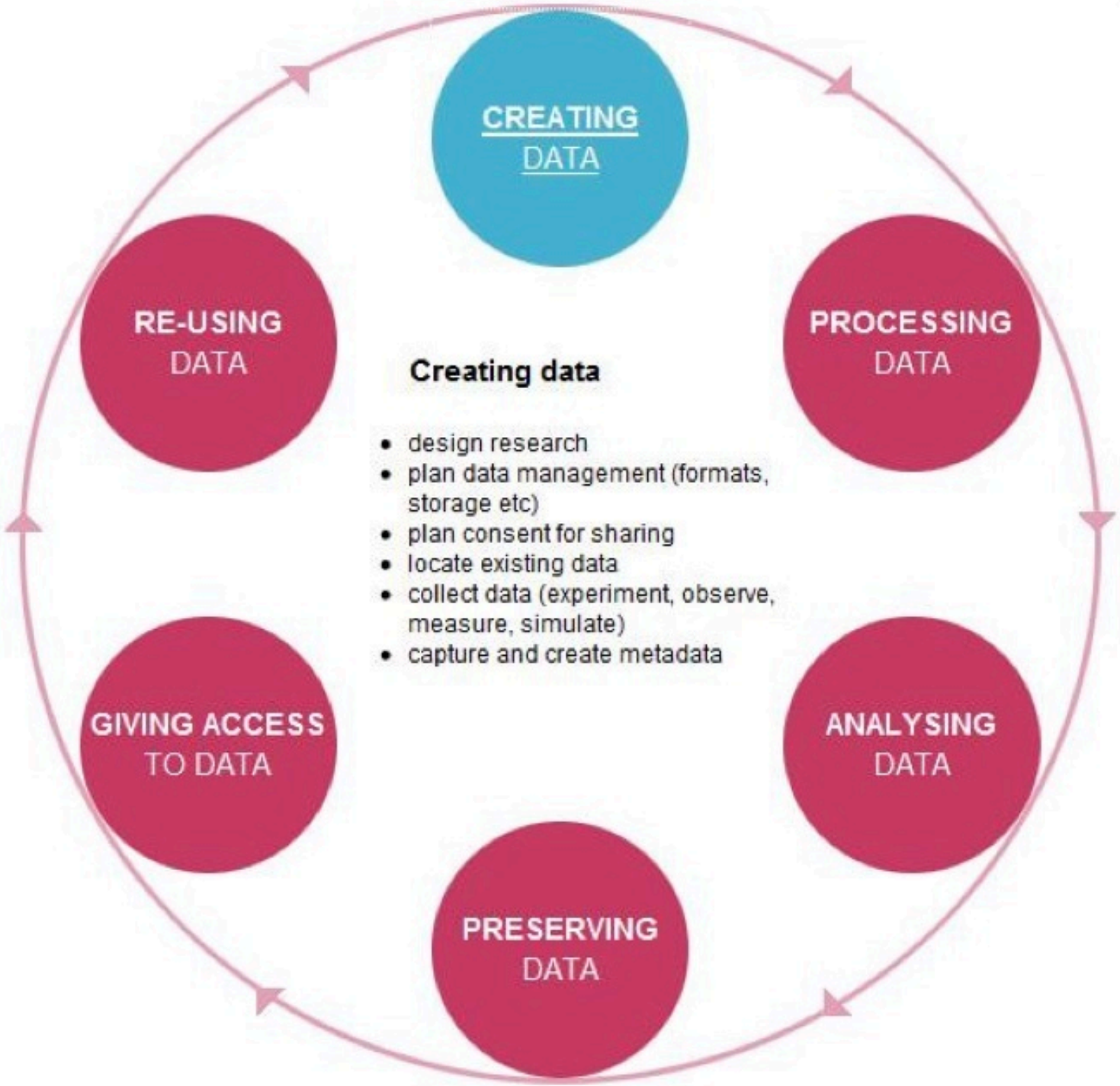
Still other definitions are so broad as to include nearly everything.

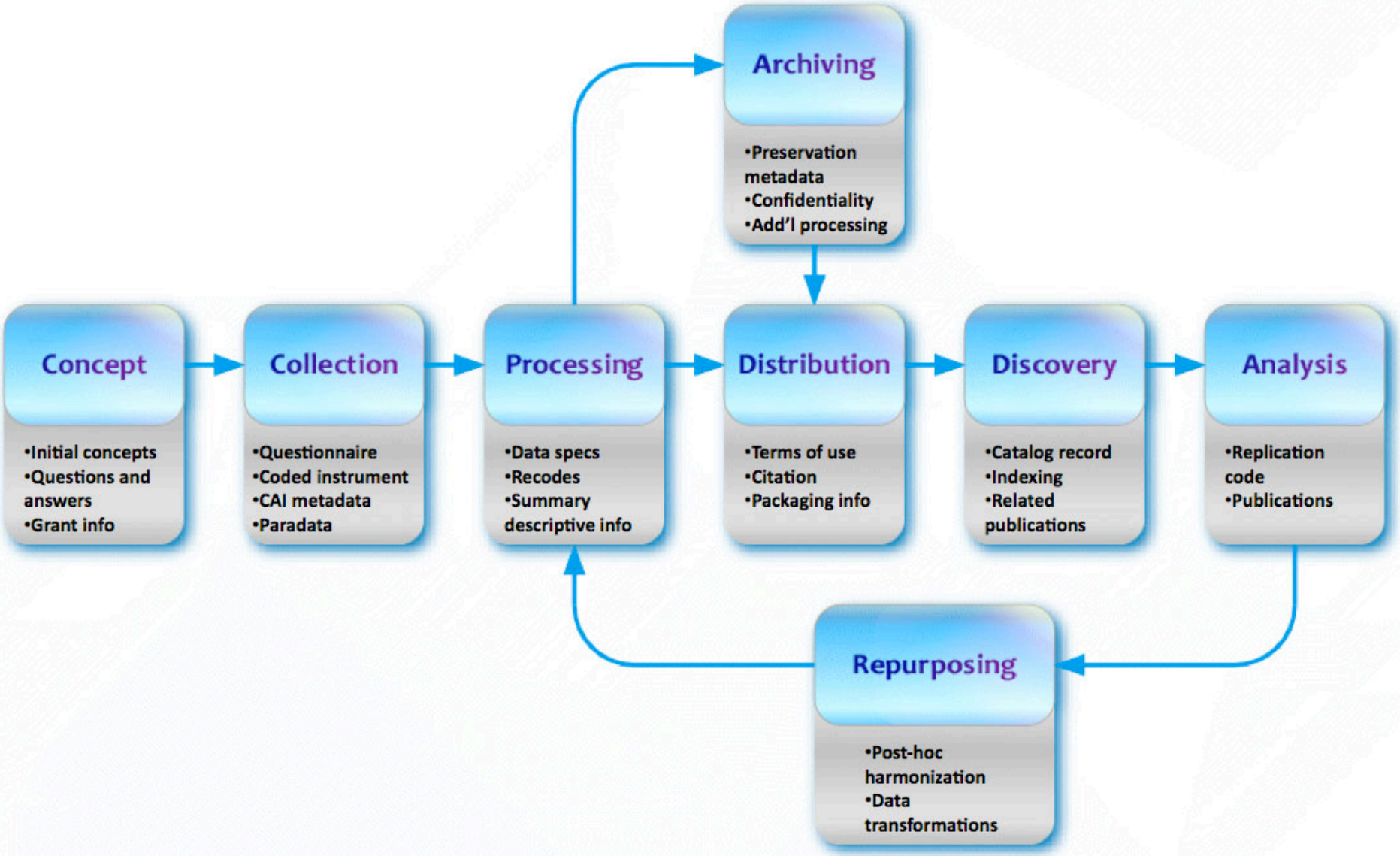


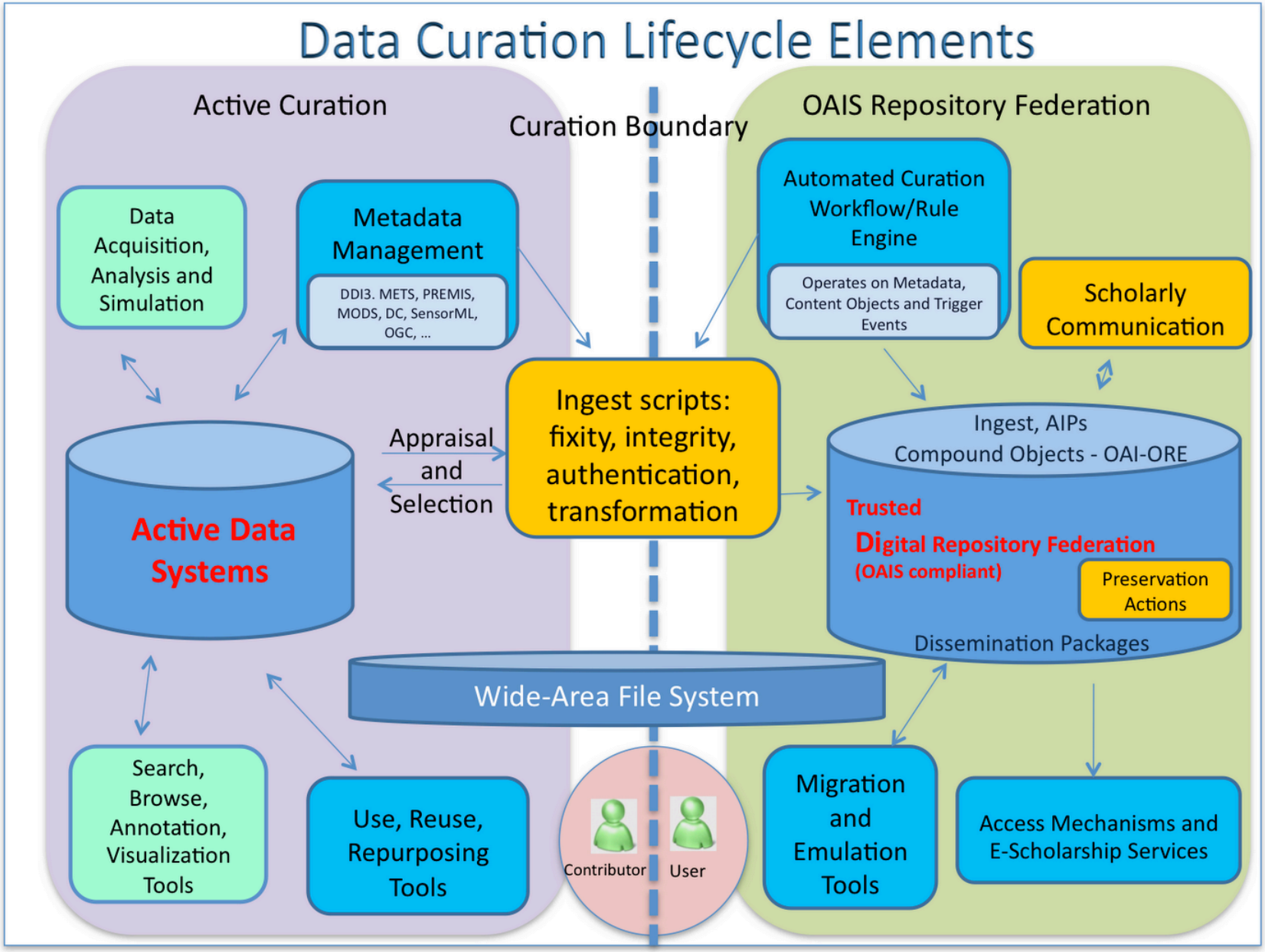
There are many “Word Cloud” representations of Data Science as well.


Databases
Processing
probabilistic
query
queries
system
incomplete
Efficient
Information
Uncertain
World-set
Data
Algorithms
Decompositions
Database
Dynamic
games
Evaluation
Beyond
Streaming
compositional
coordination
Representation
XQuery
Expressiveness
MayBMS
research
XML
Worlds
In-network
multi-query
Better
Youtopia
Relational
Complexity
Second-Order
Problem
Conjunctive
data-driven
Conditioning
Management
Exchange
Simple
logic
Markov
complex
algebra
Update
Systems
Tree
Cooperative
leashed
Trees
Approaches
computer
opportunities
Language
Minimization
Scripts
Buffer
Massively
Visual
XPath
Aggregation
Web
Fast
languages
declarative
Chain
Compiler
Approximate
Approximating
expectations
Automata
Rule-based
Carlo
Monte
Stream
querying
efficiently
Static
confidence
logics
Online
Project
Scaling
tree-structured
Eager
Plans
Course
values
streams
Entangled
Causality
support
pattern
High-Performance
Delta
back
descendant-only
ring
Formal
Homeomorphism
expressive
functional
Main-Memory
predicates
SQL
Wrapper
external
fixpoint
Parameterized
Combined
Extraction
Foundations
Comparison
Complex-Value
XML-DBMS
Tree-to-Tree
computation
Incremental
Tuple-independent
Multiplayer
Generators
framework
Lazy
Agile
Matching
Attribute
enabling


hadoop
science
mining
hive
apache
platform
programming
mapreduce
experience
tutorial
applications
algorithms
hands-on
big
using
use
talk
learn
analytics
crowdsourcing
recommendation
research
provides
predictive
cassandra
hbase
netflix
visualization
microwork
sets
base
code
people
massive
just
learning
building
beyond
hidden
case
distributed
virtualization
real
find
predicting
practical
source
machine
optimization
design
seeking
also
problem
tell
time
describe
world
new
web
needs
normal
cancer
development
real-time
tool
pig
large
carry
tools
cibygrid
introduction
discuss
drawing
session
show


CREATING DATA
PROCESSING DATA
ANALYSING DATA
PRESERVING DATA
GIVING ACCESS TO DATA
RE-USING DATA
Creating data
• design research
• plan data management (formats, storage etc)
• plan consent for sharing
• locate existing data
• collect data (experiment, observe, measure, simulate)
• capture and create metadata


Archiving
•Preservation metadata
•Confidentiality
•Add'l processing
Concept
•Initial concepts
•Questions and answers
•Grant info
Collection
•Questionnaire
•Coded instrument
•CAI metadata
•Paradata
Processing
•Data specs
•Recodes
•Summary descriptive info
Distribution
•Terms of use
•Citation
•Packaging info
Discovery
•Catalog record
•Indexing
•Related publications
Analysis
•Replication code
•Publications
Repurposing
•Post-hoc harmonization
•Data transformations


Data Curation Lifecycle Elements
Active Curation
Curation Boundary
OAIS Repository Federation
Data Acquisition, Analysis and Simulation
Metadata Management
DDI3. METS, PREMIS, MODS, DC, SensorML, OGC, ...
Automated Curation Workflow/Rule Engine
Operates on Metadata, Content Objects and Trigger Events
Scholarly Communication
Ingest scripts: fixity, integrity, authentication, transformation
Appraisal and Selection
Active Data Systems
Ingest, AIPs Compound Objects - OAI-ORE
Trusted
Digital Repository Federation
(OAIS compliant)
Preservation Actions
Dissemination Packages
Wide-Area File System
Search, Browse, Annotation, Visualization Tools
Use, Reuse, Repurposing Tools
Contributor
User
Migration and Emulation Tools
Access Mechanisms and E-Scholarship Services

What is Missing?

- Most definitions of data science underplay or leave out discussions of:
  - Substantive theory
  - Metadata
  - Privacy and Ethics

Today

- Hello
- Data Science: what is it?
- **Git / VirtualEnv**
- Course Overview
- Python: codeacademy
- Jupyter / ipython
- Check-in

Course github site:

https://github.com/philmui/datascience2016fall

```
git clone https://github.com/philmui/datascience2016fall
```

Getting git

Mac:

- brew install git

PC:

- https://git-scm.com/downloads

Today

- Hello
- Data Science: what is it?
- Git / VirtualEnv
- **Course Overview**
- Python: codeacademy
- Jupyter / ipython
- Check-in

docs/Data.Science.Python.Syllabus.ipynb

Today

- Hello
- Data Science: what is it?
- Git / VirtualEnv
- Course Overview
- **Python: codeacademy**
- Jupyter / ipython
- Check-in

Getting on the same page on Python

`https://www.codecademy.com/learn/python`

Goal: finish by October 15th

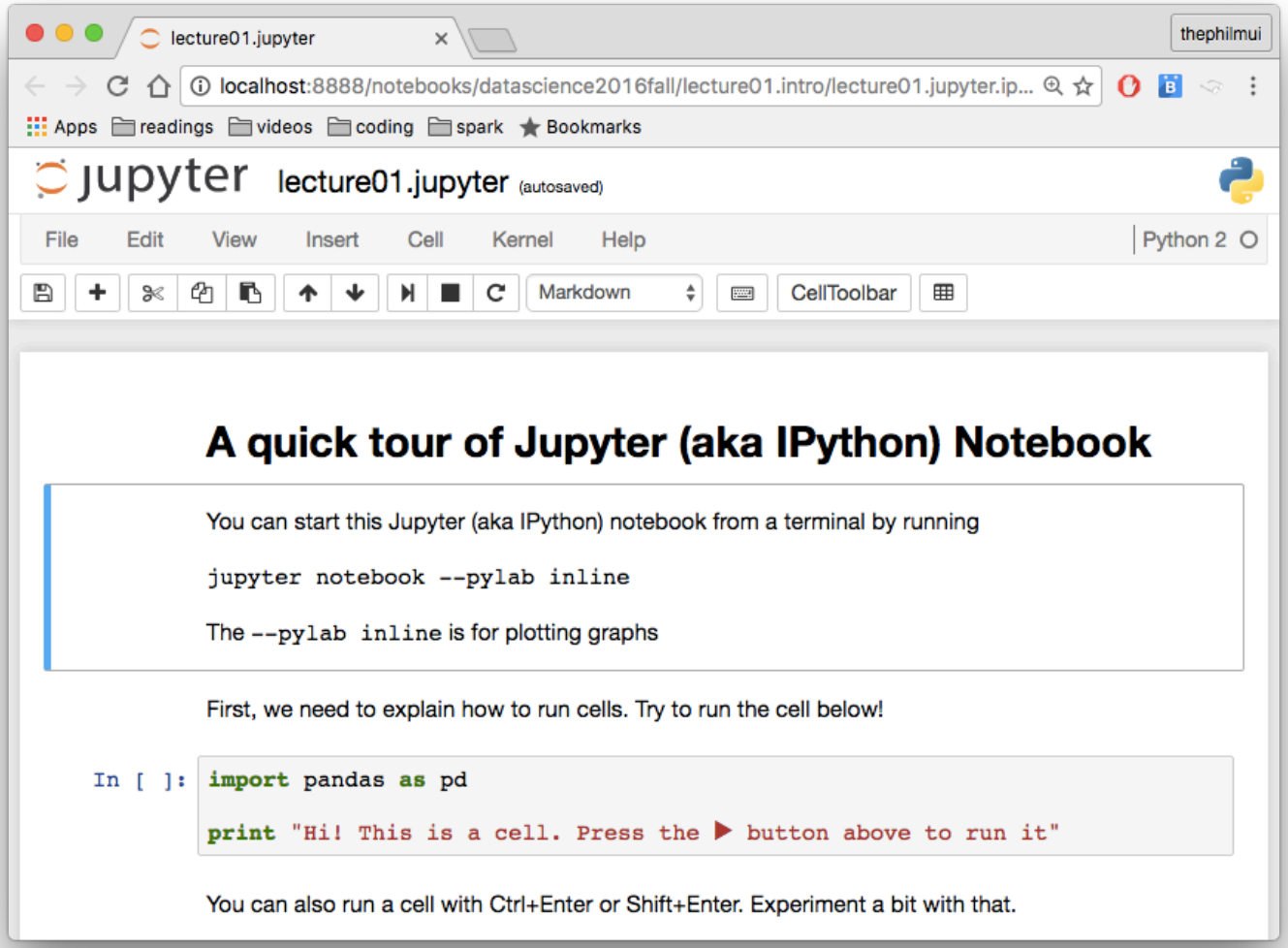
- We have no in-person class on October 15th
- You should use that day to complete the ENTIRE CodeAcademy Python course
- Submission to Camino a screenshot of your completed Python course (no need to pay)

Today

- Hello
- Data Science: what is it?
- Git / VirtualEnv
- Course Overview
- Python: codeacademy
- **Jupyter / ipython**
- Check-in

lecture01/lecture01.intro.ipynb

Jupyter / IPython



Today

- Hello
- Data Science: what is it?
- Git / VirtualEnv
- Course Overview
- Python: codeacademy
- Jupyter / ipython
- **Check-in**

Assignment 1

- Open Jupyter notebook: "`lecture01.intro.ipynb`"
- Examine the EU Revolving Loan data set: "`eu_revolving_loans.csv`"
- Answer the questions:
  - Which EU Country has the largest revolving loan % increase from 2000 – 2016?
  - Which EU Country has the largest revolving loan % decrease from 2000 – 2016?
  - Qualitatively describe in 5 sentences or less why these 2 countries may be seen these big changes during this period.

- Submission: to "Assignment" section of course Camino
- Due date: 10/08/2016 (Sat) 11:59pm PT