Data Science with Python

Summer 2016

Today

- Hello
- Data Science: what is it?
- Course Overview
- Python: check-in on codeacademy
- Data Ingestion
- 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?

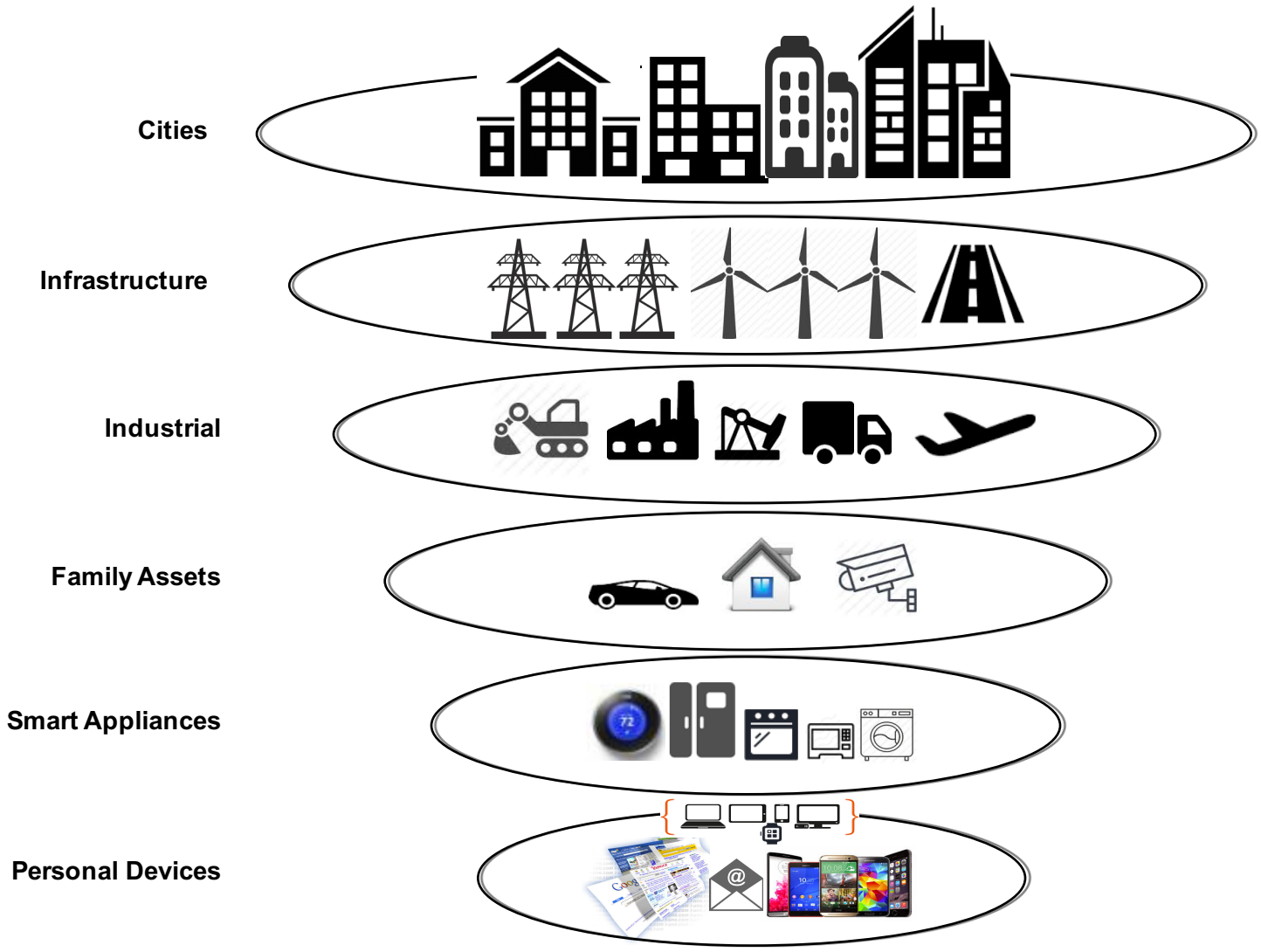
“Data”


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
REVOLUTION ANALYTICS
Localytics
R
Google Analytics
DOMO
Spark
ORACLE
SAP
TERADATA
hortonworks
IBM
ASTER
DATA
ClearStory
hadoop
eBureau
GoodData
cloudera
AnalyticsIQ Inc.
Polk
tableau
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?

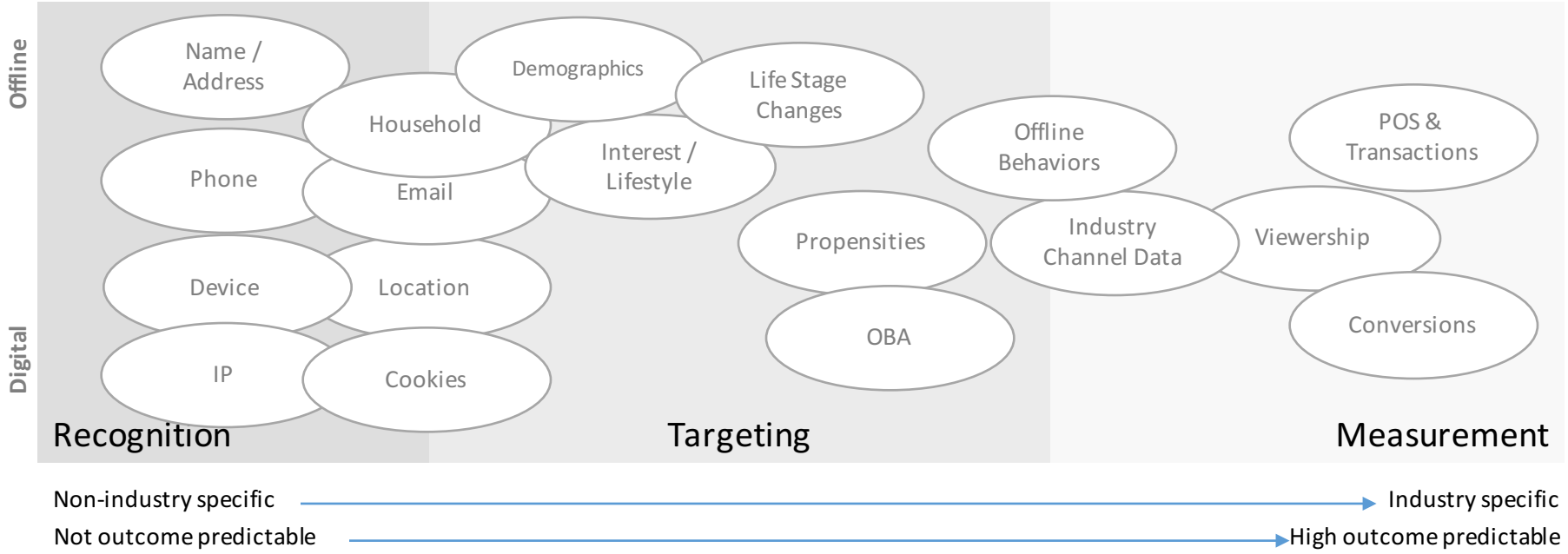
- Mining of data on shopping patterns
  - Specific products purchased
  - Combination of products purchased
  - Combined with demographic and other data
- Psychology and neuroscience
  - Habits:
    - Cue-routine-reward
    - When are habits open to change?

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

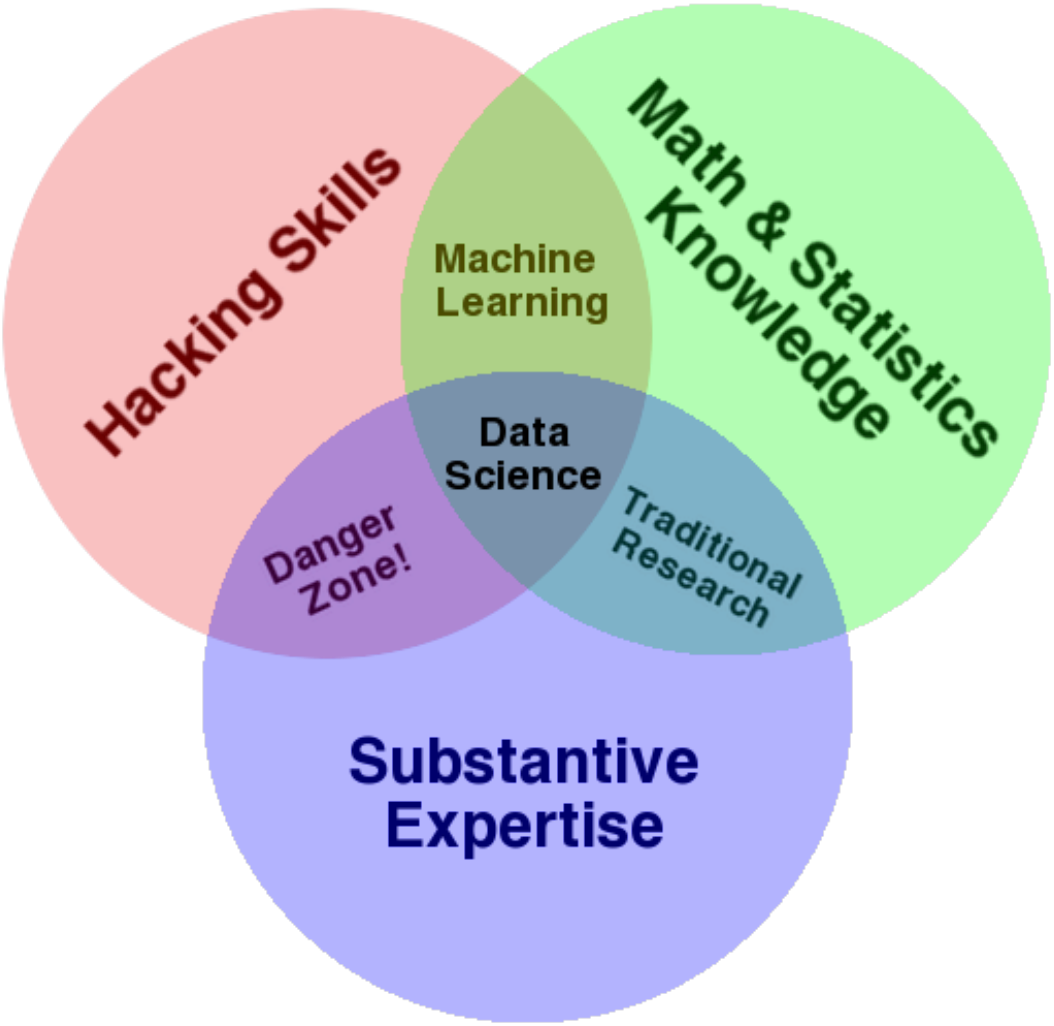
Non-industry specific → Industry specific

Not outcome predictable → High outcome predictable

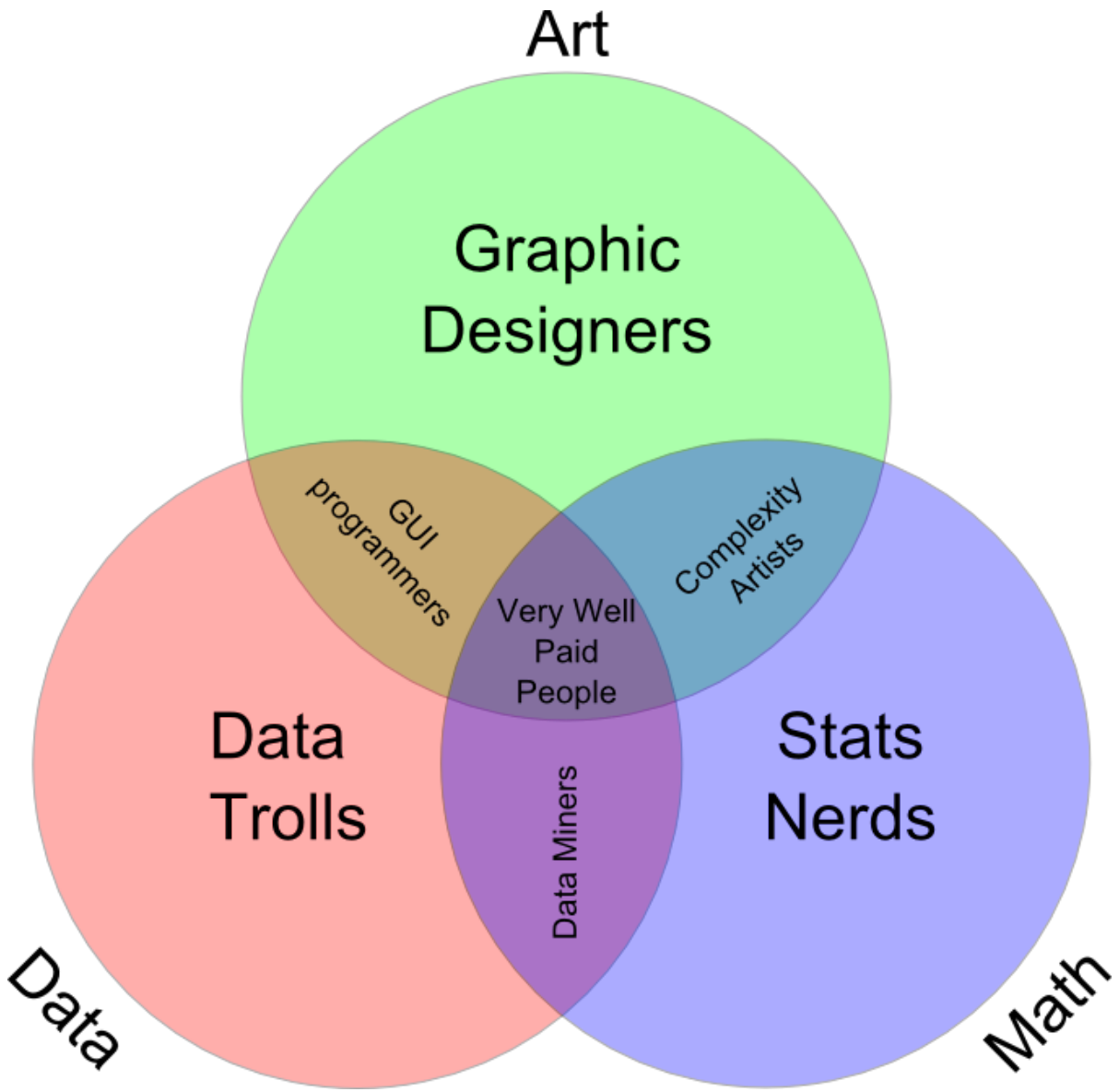
Definition of Data Science

- 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

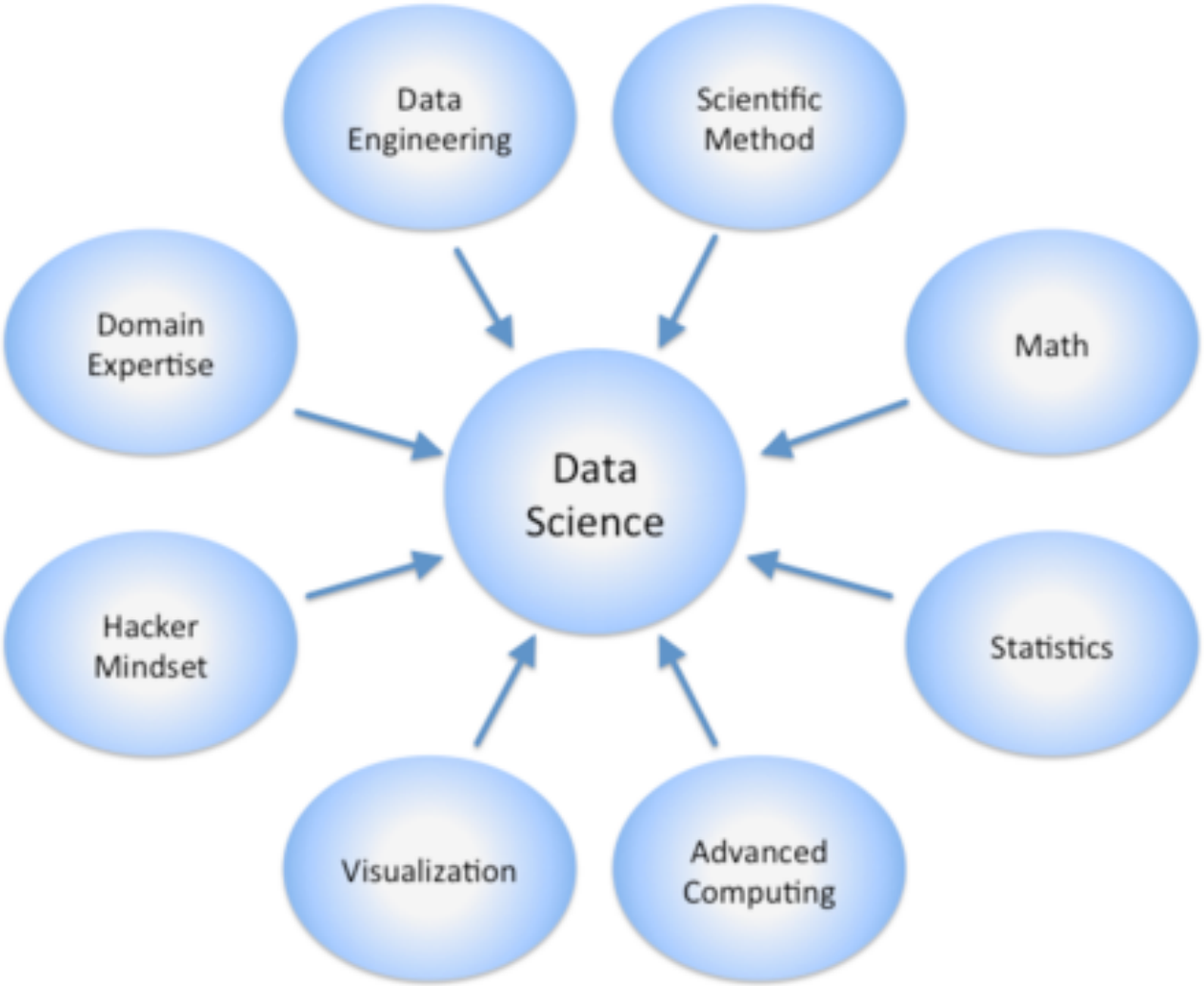
Some definitions link computational, statistical, and substantive expertise.



Other definitions focus more on technical skills alone.



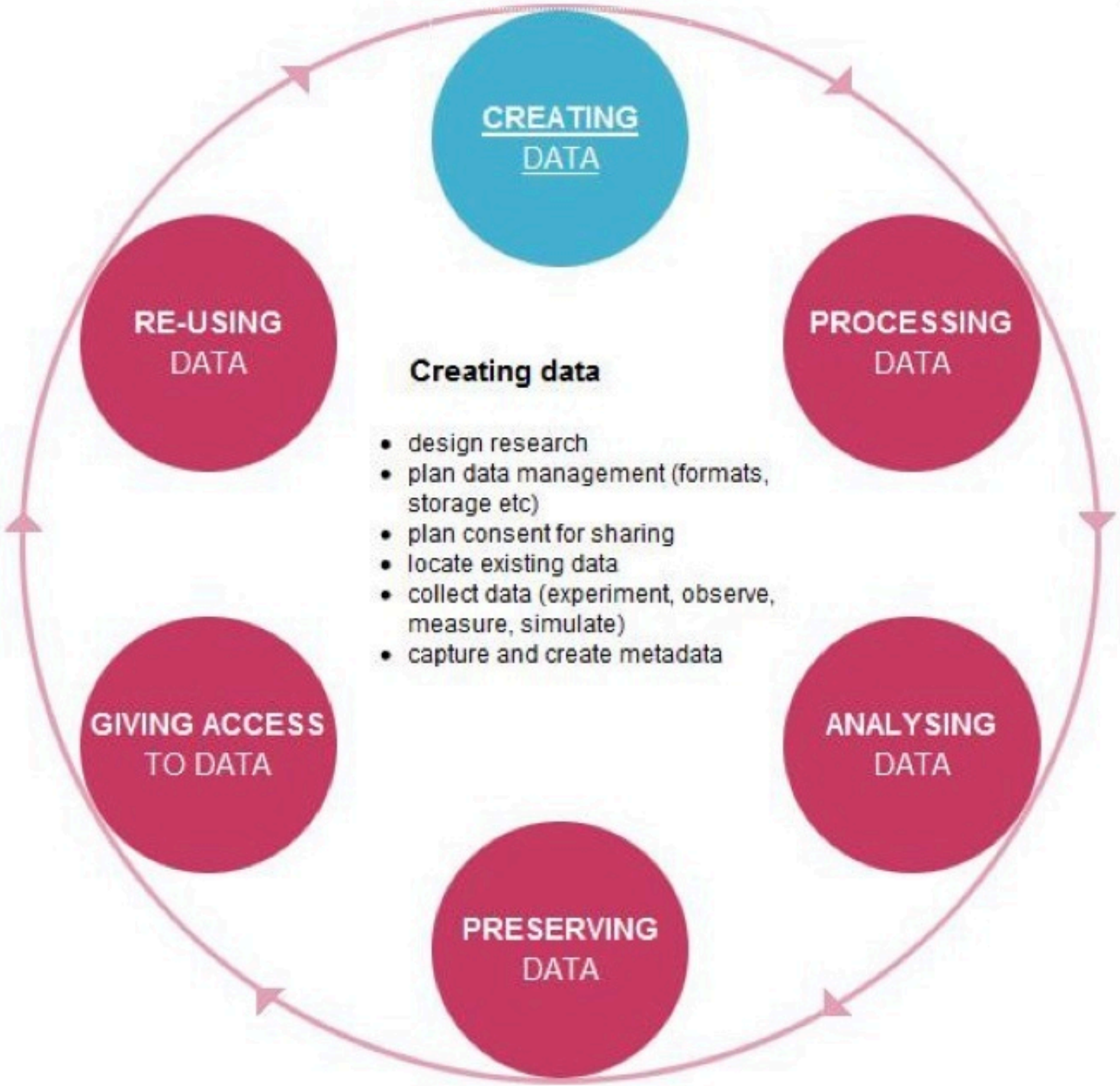
Still other definitions are so broad as to include nearly everything.

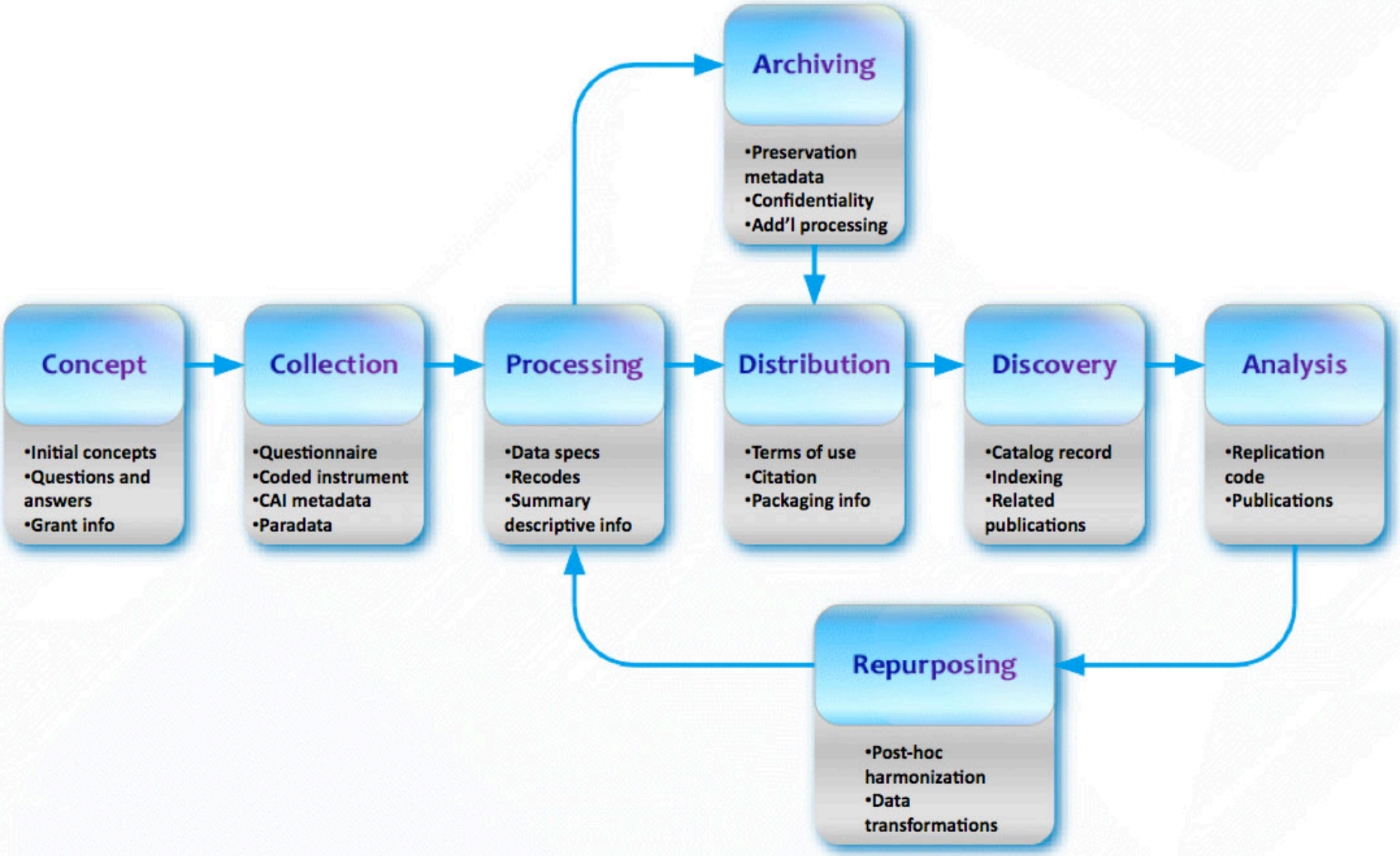


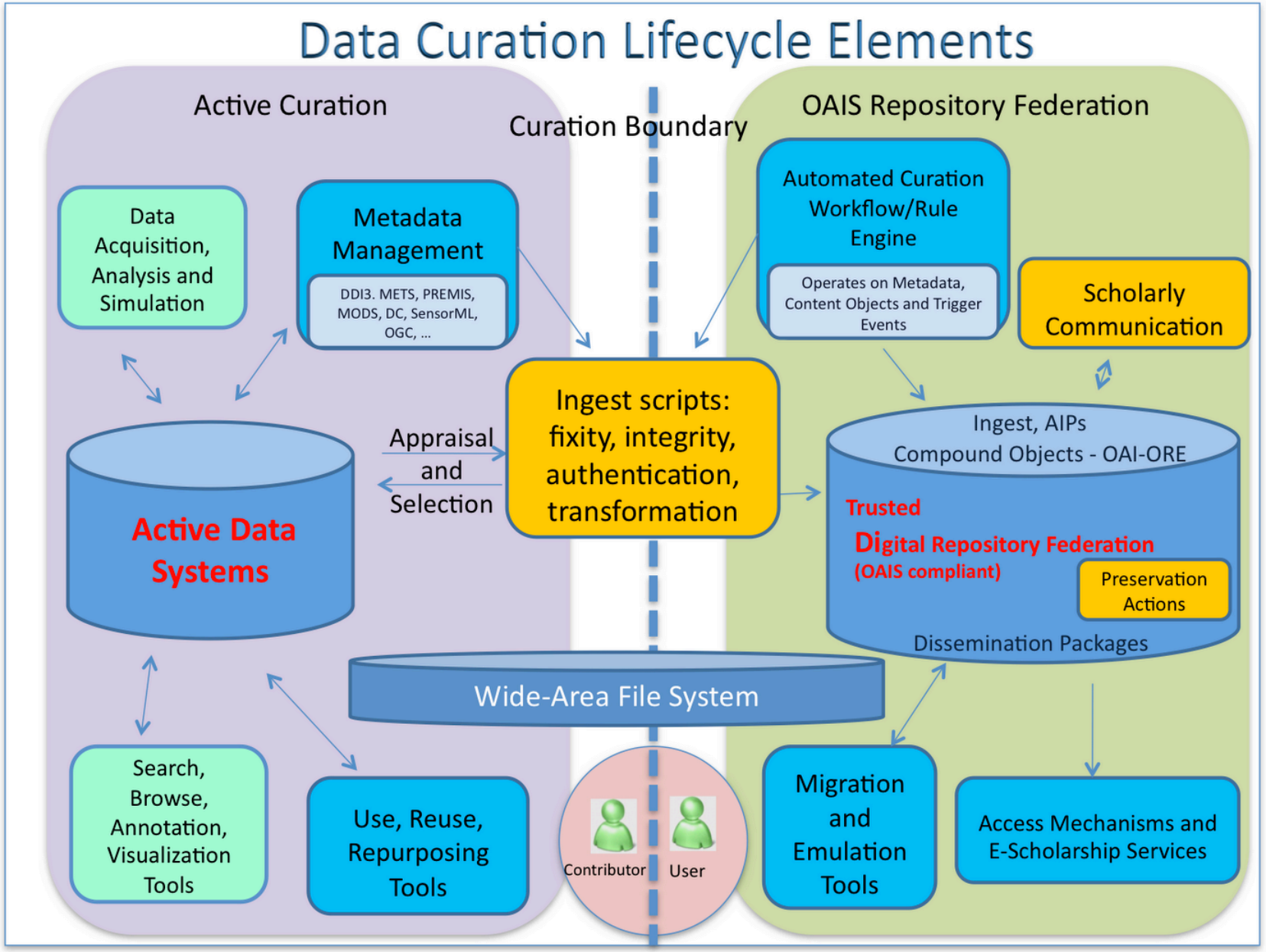
There are many “Word Cloud” representations of Data Science as well.


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


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


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?
- **Course Overview**
- Python: check-in on codeacademy
- Data Ingestion
- Jupyter / ipython
- Git / VirtualEnv
- Check-in

Course github site:

https://github.com/philmui/datascience

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

Getting git

Mac:

- brew install git

PC:

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

APPENDIX