Data Science Meets Open Science

Jeannette M. Wing
Avanessians Director of the Data Science Institute and Professor of Computer Science, Columbia University
Adjunct Professor of Computer Science, Carnegie Mellon University

Open Science Day
EPFL 50th Anniversary
18 October 2019
Data Life Cycle

generation → collection → processing → storage → management → analysis → visualization → interpretation

privacy and ethical concerns throughout
What is Data Science?

**Definition:** Data science is the study of extracting value from data.
Outline

• The Good, Bad, and Better News
• Technical and Non-Technical Challenges
• A Big Vision: Academic Cloud
The Good News

Open source software systems are indispensable to practicing data scientists, teaching data science, and researchers.
Andrew Gelman, Columbia University
Statistical modeling and analysis platform
Probabilistic programming language
Bayesian inference, MCMC built-in, R-based
Large and diverse user community worldwide
academia, government, industry

parameters {
  real y;
}
model {
  target += -0.5 * y * y;
}

$$\log p(y) = -\frac{y^2}{2} - \log Z$$
Andreas Mueller, Nicholas Hug, Thomas Fan, (core contributors), Columbia University

Machine learning

NumPy, SciPy, matplotlib

Large and diverse user community worldwide

```python
images_and_labels = list(zip(digits.images, digits.target))
for index, (image, label) in enumerate(images_and_labels[:4]):
    plt.subplot(2, 4, index + 1)
    plt.imshow(image, cmap=plt.cm.gray_r, interpolation='nearest')
    plt.title('Training: %i' % label)

# To apply a classifier on this data, we need to flatten the image, to
# turn the data in a (samples, feature) matrix:
num_samples = len(digits.images)
data = digits.images.reshape((num_samples, -1))

# Create a classifier: a support vector classifier
classifier = svm.SVC(gamma=0.001)

# We learn the digits on the first half of the data:
classifier.fit(data[:num_samples // 2], digits.target[:num_samples // 2])

# Now predict the value of the digit on the second half:
predicted = classifier.predict(data[num_samples // 2:])

# for index, (image, label) in enumerate(digits.images[num_samples // 2:]):
#     plt.subplot(2, 4, index + 1)
#     plt.imshow(image, cmap=plt.cm.gray_r, interpolation='nearest')
#     plt.title('Prediction: %i' % predicted[index])
```

0 1 2 3
8 5 4 9
• Ryan Abernathy, Columbia University
• Platform for sharing data, code, and models for geoscience
• Jupyter-in-the-cloud, Python ecosystem
• Applications in meteorology, hydrology, oceanography, climate modeling
The Bad News

Much of the big datasets are locked up in companies.
Government funding agencies don’t pay for our unsung heroes and their work.
Data is Locked Up

• Data is locked up for good reason
  • Privacy of customers
  • Data is company asset, and accrues to its bottom-line

• Consequences for science
  • Industry is ahead of academia, in some areas of science
    • Academia can/should work on problems industry can’t/won’t
    • Academics work inside company, leading to new models of academic-industry relations
Open Source Developers

• Most government funding agencies in the US do not support software engineers
• Academia does not treat them as equal to tenure-track faculty
The Better News

New efforts support data sharing.
New funding sources and culture change for software developers.
Private companies (as well as government, non-profits, and academia) exchange data to create public value.

datacollaboratives.org
Welcome to dataCommons

Publicly available data from open sources (i.e. census.gov, NOAA, data.gov etc) are a vital resource for students and researchers in a variety of disciplines. Unfortunately, processing these datasets is often tedious and cumbersome.
Open Source Developers

- Foundations to the rescue!
  - Chan-Zuckerberg, Moore, Sloan, Schmidt, …
- Forward-looking universities are developing new tracks (or reusing existing ones) to be equivalent to the tenure-track
  - Carnegie Mellon: “systems science”
  - Columbia: “applied data scientist”
Challenges

What should the community focus on next?
What the Community Can Work on Together

• Non-technical
  • Define a “Data IRB”: Institutional Review Board (IRB) process for data
  • Explore new models of academia-industry engagement
  • Create a culture change at universities to acknowledge contributions of open source developers, applied data scientists, etc.

• Technical
  • Make finding and publishing datasets as easy as using the Internet/web/browser
  • Explore “share back” model, to give back control of data to individual. See Inverse Privacy paper, by Gurevich, Hudis, and Wing
A Big Vision
Academic Cloud
Progress


Academic Cloud for CISE workshop, January 8-9, 2018

October 2018

Coordinating “entity” (UW, UC San Diego, UC Berkeley) awarded August 2019
But, it needs to be for all disciplines, not just computer science, and ideally (eventually), for all academic institutions worldwide, not just in the US.

Academic Cloud = the next “Internet”
Thank You