Democratization of data-driven research

Dr. Evelyne Viegas
Director
Microsoft Research Connections

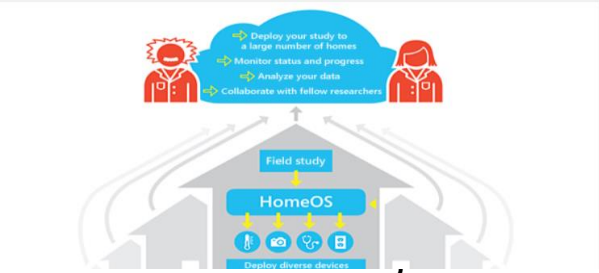
Microsoft Research Connections

WINDOWS AZURE FOR RESEARCH

Microso
help the
discover
research
Window
support
computi
events,
location

Get started with The Lab of Things

Who is using The Lab of Things



Predictive Analytics

Competing based on Analytics:

- If you can predict it, you can own:
  - Targeting
  - Fraud detection

Predictive Analytics in Action:

- Big Data and Death at America's Racetracks http://thorotrends.com/news-and-views-20526/109-big-data-and-death-at-americas-racetracks
- How Target Figured Out A Teen Girl Was Pregnant http://onforb.es/SokL3j

Data Science

- Strategic shortage of data scientists in industry, a role many were not aware of 2 years ago

- Signature Skills needed for Data science
  - Data curiosity; Data exploration; Data visualisation
  - Math, Statistics
  - Machine Learning
  - Software Development

Too complex, won't scale!

Need for accessible tools to support Algorithmic Creativity and Exploration

Research in support of AI

Data has become a first class citizen

IT'S A DATA-DRIVEN WORLD

MACHINE LEARNING AS AN ENABLER OF DECISION MAKING

It's a data-driven world

- Spell Checking
- Machine Translation
- Search queries + click through
- Online games skill matching

Data logs behaviours in more reliable ways than demographic studies or surveys to study/predict trends

(Banko and Brill, 2001) – effectiveness of statistical NLP techniques is highly susceptible to the **data size** used to develop them

Challenge in Data-driven Research

- Lot of the data needed for data-driven research in industry
  - **Reason: scale; privacy, business sensitivity**

How to make real world large scale data available to researchers to nurture innovation and perform valid experimentation, while maintaining privacy?

Machine Learning Services

CodaLab – community service to democratize machine learning, enable better benchmarks and help the data scientist

What are the issues?

Duplication of Effort: People spend a lot of time on the empirical evaluation of an idea: finding and pre-processing datasets, finding and implementing competing methods for comparison, running all the experiments, creating tables and figures to summarize the results

Reproducibility: Empirical results in papers are difficult to reproduce, because the code and data are rarely made available in a usable form

Comparable Baselines: Empirical results in papers are difficult to compare, because methods are often evaluated on different datasets or even different versions of the same dataset

Solving
Large Scale
Real World Problems with
CodaLab
The Hub for Data-driven Research
and Scientific Advances

Machine Learning – Enabler of Decision Making

The good

- Data is available to stimulate innovation and enable new forms of collaboration and knowledge creation

- Effort in community and government to have reproducible research

It is not just about the data. It is also about the algorithm, the transparency and reproducibility of the research process

The research process

- Find data, clean it, convert between formats
- Find code, compile it, email authors, reimplement
- Run experiments, keep track of multiple versions

Non-exhaustive comparisons

	Previous method	Our method
Dataset 1	88% accuracy	92% accuracy
Dataset 2	72% accuracy	77% accuracy
Dataset 3	?	?
Dataset 4	?	?
Dataset 5	?	?
Dataset 6	?	?
...	?	?

Uncontrolled comparisons

Previous method	Our method
88% accuracy	92% accuracy
using sampling	using optimization
L_2 regularization	L_1 regularization
5-fold cross-validation	10-fold cross-validation
one set of bugs	another set of bugs

Percy Liang, 2013

What is needed:

Datasets

Programs

5.6	6.2	2.0	5.6	4.7	3.0	8.1	7.5	7.2	7.0	2.1	5.2	4.6
1.1	1.4	5.0	10.0	1.2	7.8	1.1	5.7	8.6	9.1	6.2	0.9	4.8
5.5	4.5	0.1	0.6	7.3	1.7	0.8	0.6	7.2	9.2	0.1	1.8	1.7
0.4	1.5	2.7	0.4	7.5	5.7	8.2	3.3	9.0	8.3	5.1	0.8	9.5
9.0	8.9	3.1	9.5	9.6	6.0	6.3	3.1	4.4	7.8	0.7	6.6	3.9
3.5	5.2	1.6	4.6	9.3	7.0	7.0	2.0	2.2	4.1	6.1	2.5	9.5
1.9	7.4	2.9	1.5	1.2	9.7	6.3	0.0	6.4	1.3	2.3	1.0	0.9
3.3	9.5	9.8	7.1	8.3	6.4	1.1	3.4	8.9	2.5	9.5	2.2	3.9

Machine Learning – Enabler of Decision Making

- The wanted
  - Reduce amount of time spent per researcher on preprocessing datasets, writing evaluation and visualization scripts, getting other people's code to run
  - Lower the barrier to create experiments via agile exploration of data and code
  - Accelerate the pace of innovation by creating an online community around sharing and executing modules, enabling the creation of "executable papers"

CODALAB

An Open Source Platform which lets communities create and explore Experiments together and engage in Competitions to advance the state of the art in Machine Learning

Community Leads:

- Percy Liang, Stanford University, USA
- Isabelle Guyon, ChaLearn, USA

http://codalab.org

CodaLab Machine Learning Community Site

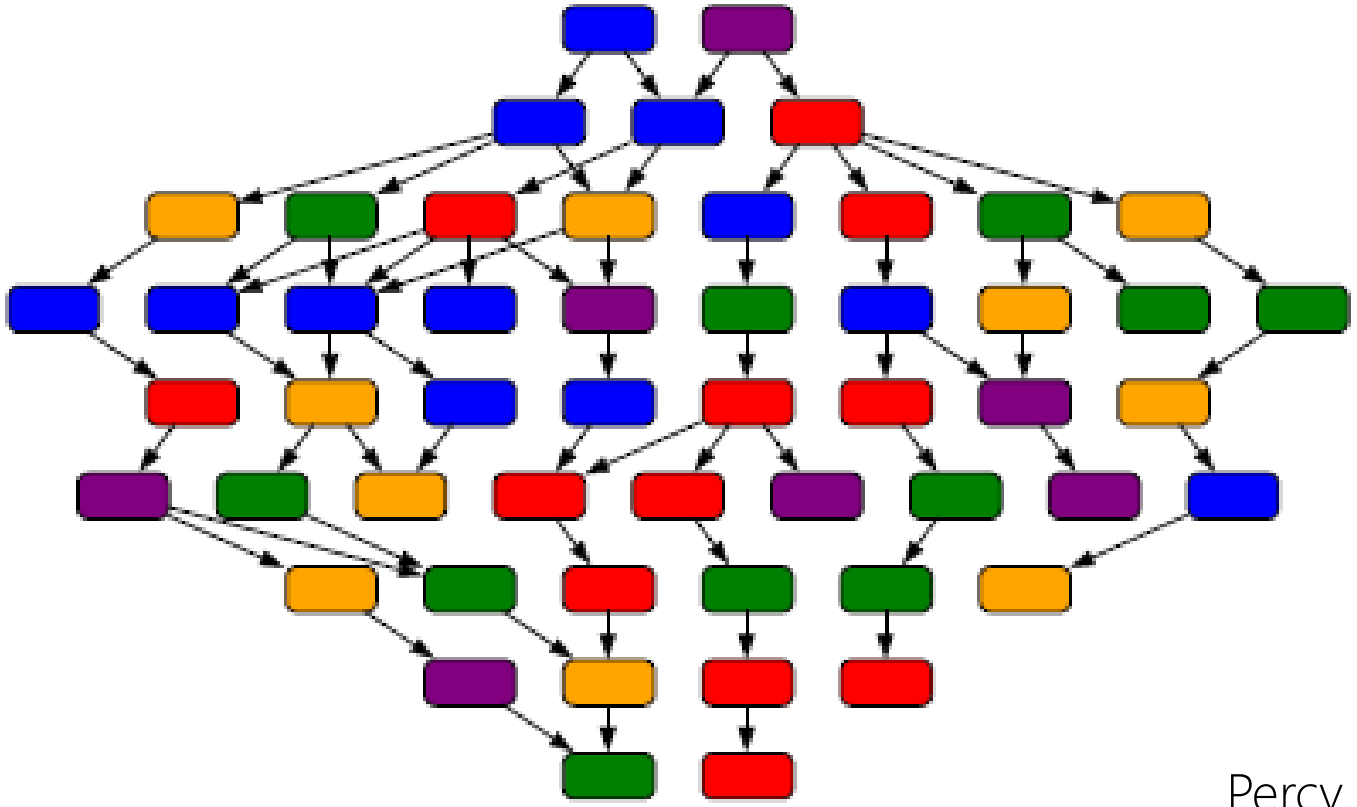
- **Experimentation**
  - Data “repository”
  - Algorithm “repository”
  - Creation, Execution, Sharing of Experiments
  - Collaborative workflows
  - Executable Notebook with code and descriptions (textual, visuals) interleaved

- **Competition**
  - Customized for ML

Principle 1: Modularity

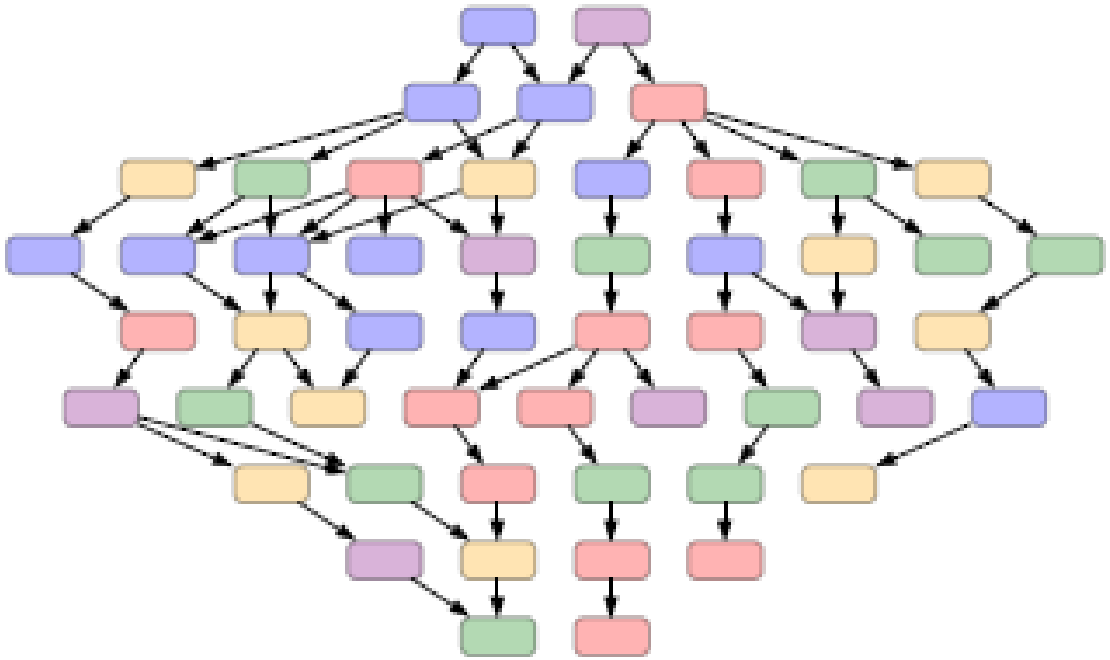
AI problems require efforts of entire community

People specialize, contribute in decentralized way



Percy Liang, 2013

Principle 2: Immutability



Inspiration: Git version control system

- All programs/datasets/runs are write-once
- Enable collaboration without chaos
- Capture the research process in a **reproducible** way

Percy Liang, 2013

Principle 3: Literacy

CodaLab
Executable Notebook
(interpretation
of the experiment)

We now train the classifier with more data.

Program :	**SVMlight**
Arguments :	-n 2000
Dataset :	**thyroid**
Error :	2.6%
Time :	1 second

Notice that the error remains the same, suggesting that we've saturated our model.

Use cases:

- Informal blog posts
- Formal executable papers

Percy Liang, 2013

Related effort

Name	Description
MLComp	Precursor to CodaLab • No workflow
BigML, Google Prediction API	Provide fixed set of programs People submit (private) data
runmycode.org, myexperiment.org, Weka	Run computer codes associated with a scientific publication Require specific formats
UCI ML, MLData	Repository of data
MLoss	Repository of code
Mathematica, IPython	Interleave code with text descriptions No notion of immutability across people

CodaLab Goals

- Advance the state of the art in Machine Learning Research
  - Democratize Machine Learning via data benchmarking, algorithm comparison
  - Enable repeatability and transparency of experimentation
- Create a Meta-Learning Matching Platform
  - Get scientific insights into what techniques worked on which datasets
  - Given a brand new problem, predict techniques which will work well
- Build Multi-year Challenges
  - Causality
  - Text Understanding
  - Medical Imaging

Next?

- How to better engage with academia to drive data-driven research?
- What else can industries do to help democratize large scale data-driven research?

Contact: evelynev@microsoft.com

THANK YOU

Microsoft®

© 2012 Microsoft Corporation. All rights reserved. Microsoft, Windows, and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries.
The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation. MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.