



Democratization of data-driven research

Dr. Evelyne Viegas
Director
Microsoft Research Connections

Microsoft Research Connections

HOME KEYNOTES PROGRAMME BIOS INFORMATION



Software Summit 2011

13-15 APRIL, PARIS, FRANCE

CONTACT INFORMATION

For more information please contact us at:
E: softsumm@microsoft.com

REGISTRATION

Registration is now closed. Please contact us at softsumm@microsoft.com if you have any questions.

Speakers: Please download your release form [here](#).

Software underlies everything in our daily lives, from transport to entertainment, from business to social networking. The way we design, engineer, maintain and use software has changed dramatically in the past five years, and will change more in the next. We need to be ready for these changes, and be able to educate the next generation with the real tools and theory that will make them effective software developers, engineers and architects.

To meet this challenge, Microsoft Research is holding the inaugural Software Summit which will bring together thought leaders from academia, research labs, funding agencies and Microsoft to discuss the state of software research and development. Participants will experience a vibrant programme of talks, panels, workshops and demonstrations, and will come away with a much better idea of the integral part that industrial research plays in society, education and technology transfer, and with how they can contribute to this thriving community.

KEYNOTE SPEAKERS



try F# = lightbulb

Learn, create, and explore F#! Try the new Try F# Beta

Venus-C

Register Now

Home News Mission Partners Experts Fact file VENUS-C Platform User Scenarios Document Library Events Media Room

VENUS-C Value Proposition

VENUS-C has a compelling range of applications growing over time through an Open Call to broaden the scope of the project. Andrea Monari, Engineering & VENUS-C Co-ordinator

What's VENUS-C?

VENUS-C draws its strength from a joint co-operation bringing together industrial partners and scientific user communities to develop, test and deploy an industry-quality Cloud Computing service for Europe. VENUS-C is an opportunity to rethink how we approach scientific discovery, setting scientists as scientists by focusing on their core work to seed discovery from new insights.

User Scenarios

VENUS-C User Scenarios currently comprise seven applications across four thematic areas for European research & business communities:

- Biomedicine
- Civil Engineering
- Civil Protection & Emergencies
- Data for Science

Open Call

Codalab ALPHA

Microsoft Research Project Hawaii

Join Now

Personalise your VENUS-C Channel! Register to access resources of interest

Project Hawaii tools and resources for instructors

Project Hawaii tools and resources for students

Using the cloud to enhance

What Is Project Hawaii?

Innovations in WLAN (Wireless Local Area Network) and WWAN (Wireless Wide Area Network) technology bring us to today's connected world, and smartphones are gaining acceptance in

Competitions My_Codalab About

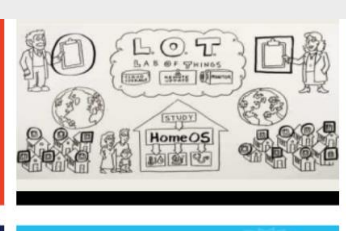
Bringing cloud computing to researchers



WINDOWS AZURE FOR RESEARCH

Microsoft help the discover research. Windows support compute events, location

LoT LAB of THINGS



MACHINE LEARNING SUMMIT

PARIS, APRIL 23 AND 24, 2013



La Gare

Concorde, LaFayette

Microsoft

Codalab = Experiments + Competitions

Experiments

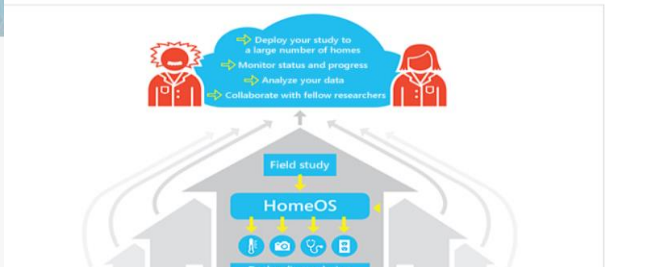
Codalab Experiments enable collaborative research and computational research to be done in an efficient and reproducible manner. By providing modularity, live execution, and inline annotation of code with rich explanations, Codalab enables you to quickly sketch ideas and collaborate with fellow community members.

Competitions

Codalab competitions provide an opportunity for researchers, developers and algorithmists to create solutions for problems across a wide range of domains, and advance the state of the art for their respective areas of interest.

Get started with The Lab of Things →

Who is using The Lab of Things →



Deploy your study to a large number of homes

Monitor status and progress

Analyze your data

Collaborate with fellow researchers

Field study

HomeOS

Deploy diverse devices



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



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>

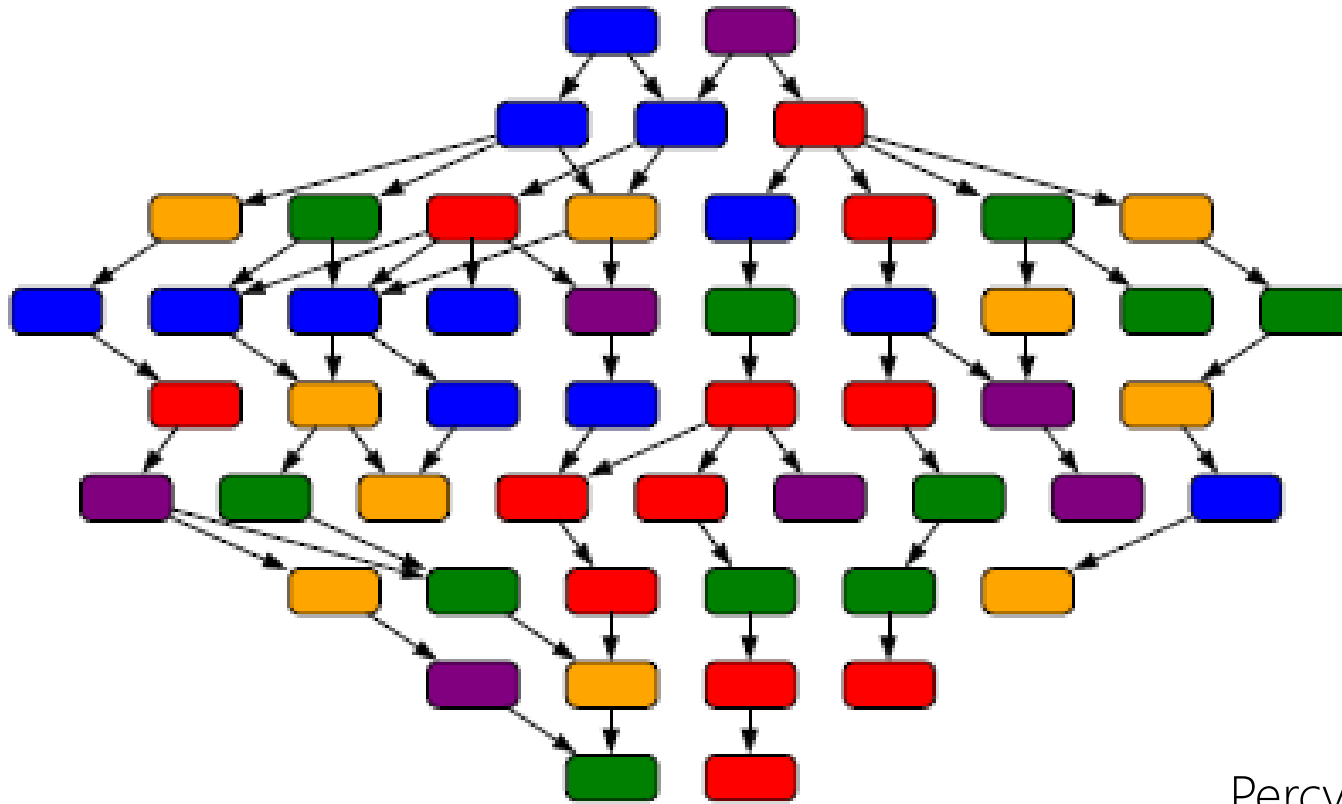


CodaLab Machine Learning Community Site

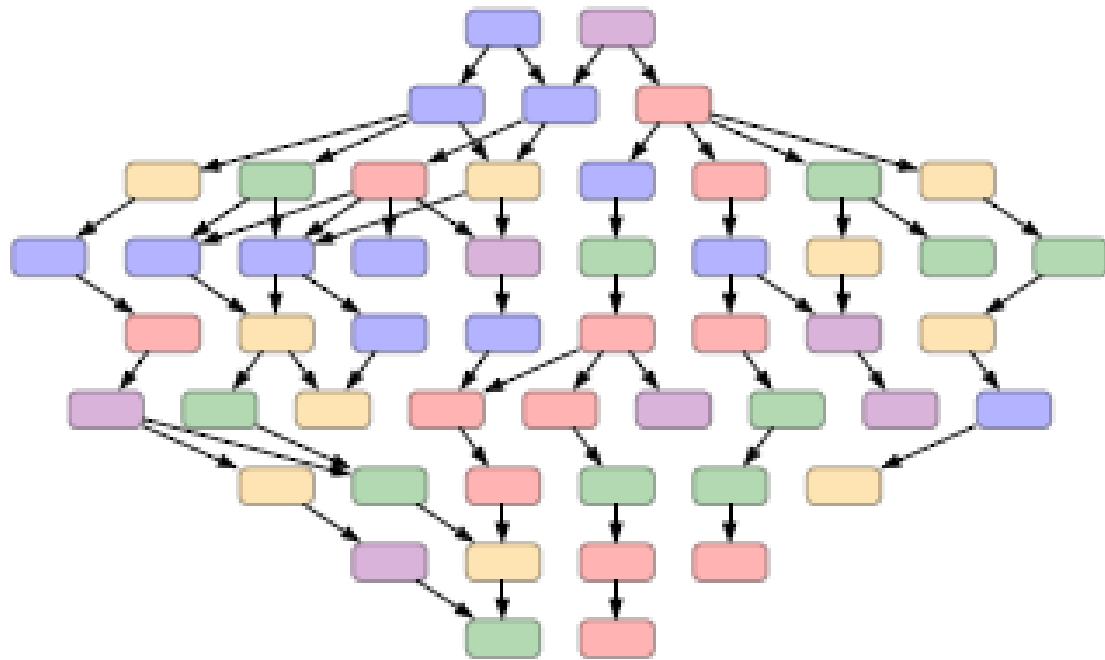
- **Experimentation**
 - Data "repository"
 - Algorithm "repository"
 - Creation, Execution, Sharing of Experiments
 - Collaborative workflows
 - Executable Notebook with code and descriptions (textual, visuals) interleaved
- **Competition**
 - Customized for ML

Principle 1: Modularity

AI problems require efforts of entire community
People specialize, contribute in decentralized way




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



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



Related effort

Name	Description
MLComp	Precursor to CodaLab <ul style="list-style-type: none">• 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[®]