

Symposium on Data Science and Statistics (SDSS18)

# Painless Computing Models for Ambitious Data Science

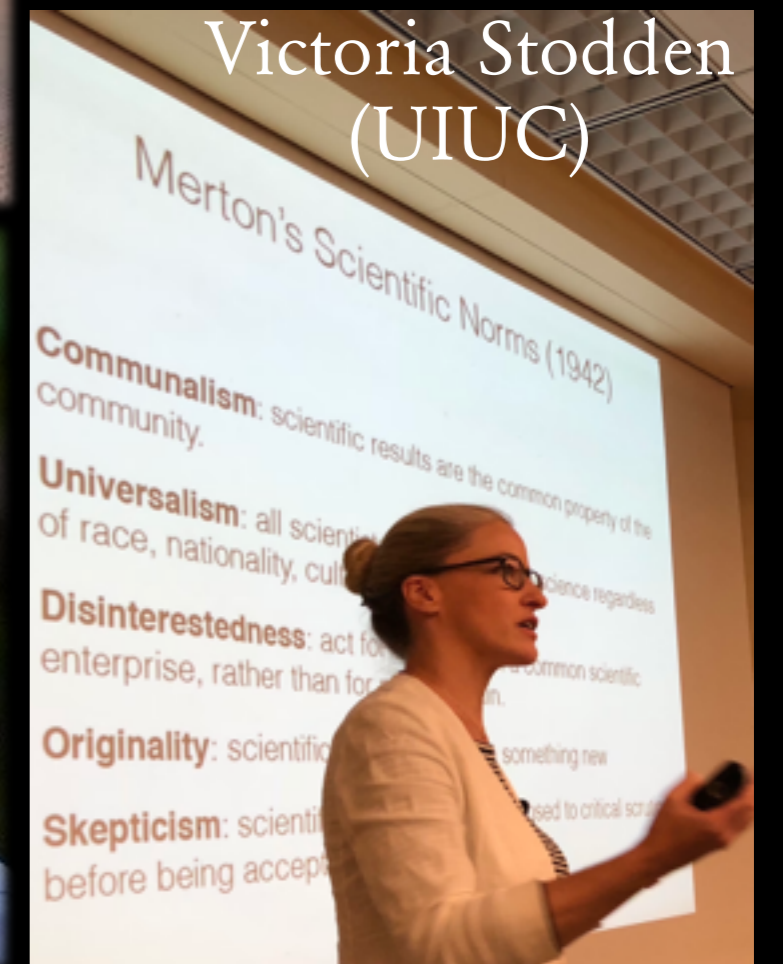
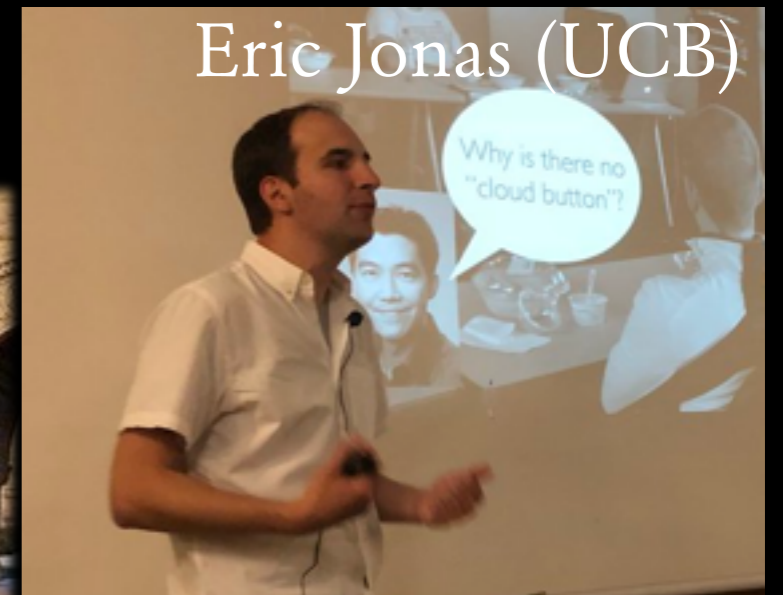
Hatef Monajemi, May 18 2018



Stanford  
University



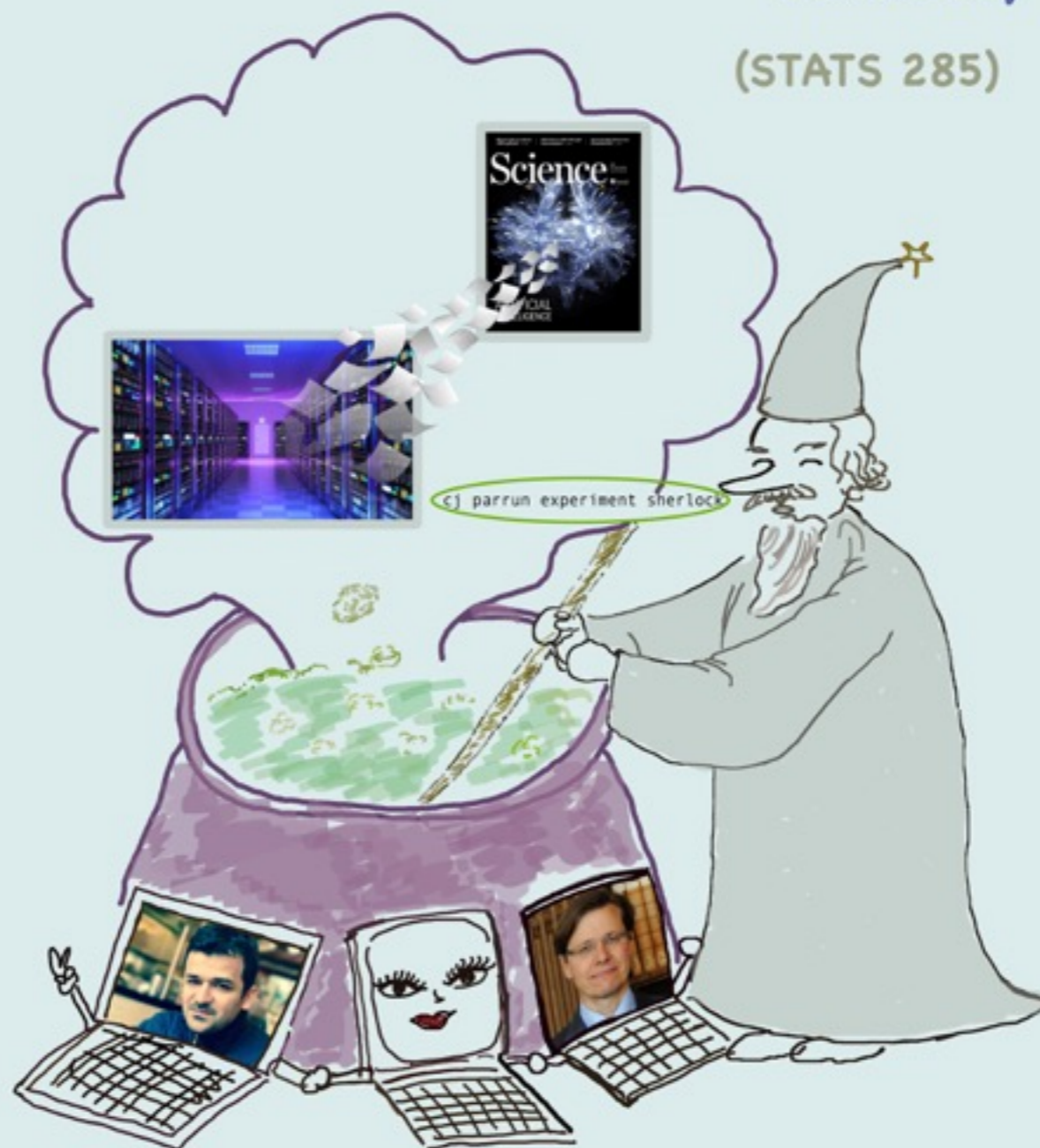
# Coauthors





Massive Computational Experiments,  
Painlessly

(STATS 285)



Time: Monday 3:00 - 4:20  
Place: Thornt110  
Website: [stats285.github.io](http://stats285.github.io)

2012

The world changed

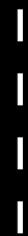


# How to advance knowledge?



use a better mathematical model

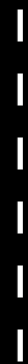
*1800*



*2012*



experiment until you find a winner



*A.I.*

*Apocalypse*

What happened?

# The Great IT Enrichment



“Six decades into the computer revolution, four decades since the invention of microprocessors, and two decades into the rise of modern internet, **all of the technology required to transform industries through software finally works** and can be widely delivered at a global scale.”

Marc Andreessen, *why software is eating the world*,  
WSJ, 2011

# The Great IT Enrichment

"Six decades into the computer revolution, four decades since the invention of microprocessors, and two decades into the rise of modern internet, **all of the technology required to transform industries through software finally works** and can be widely delivered at a global scale."

Marc Andreessen, *why software is eating the world*,  
WSJ, 2011



- Cloud provides **millions of servers** globally
  - ✓ **same-day** delivery of 10k-100k of CPU hours
  - ✓ 3 cents per CPU hour, 45 cents per GPU hour
- **Open-source** Software and Frameworks galore
- **High-Speed Internet**



# Science goes digital

- Traditionally
  1. Deduction (Math proofs)
  2. Induction (Physical sciences)

# Science goes digital

- Traditionally

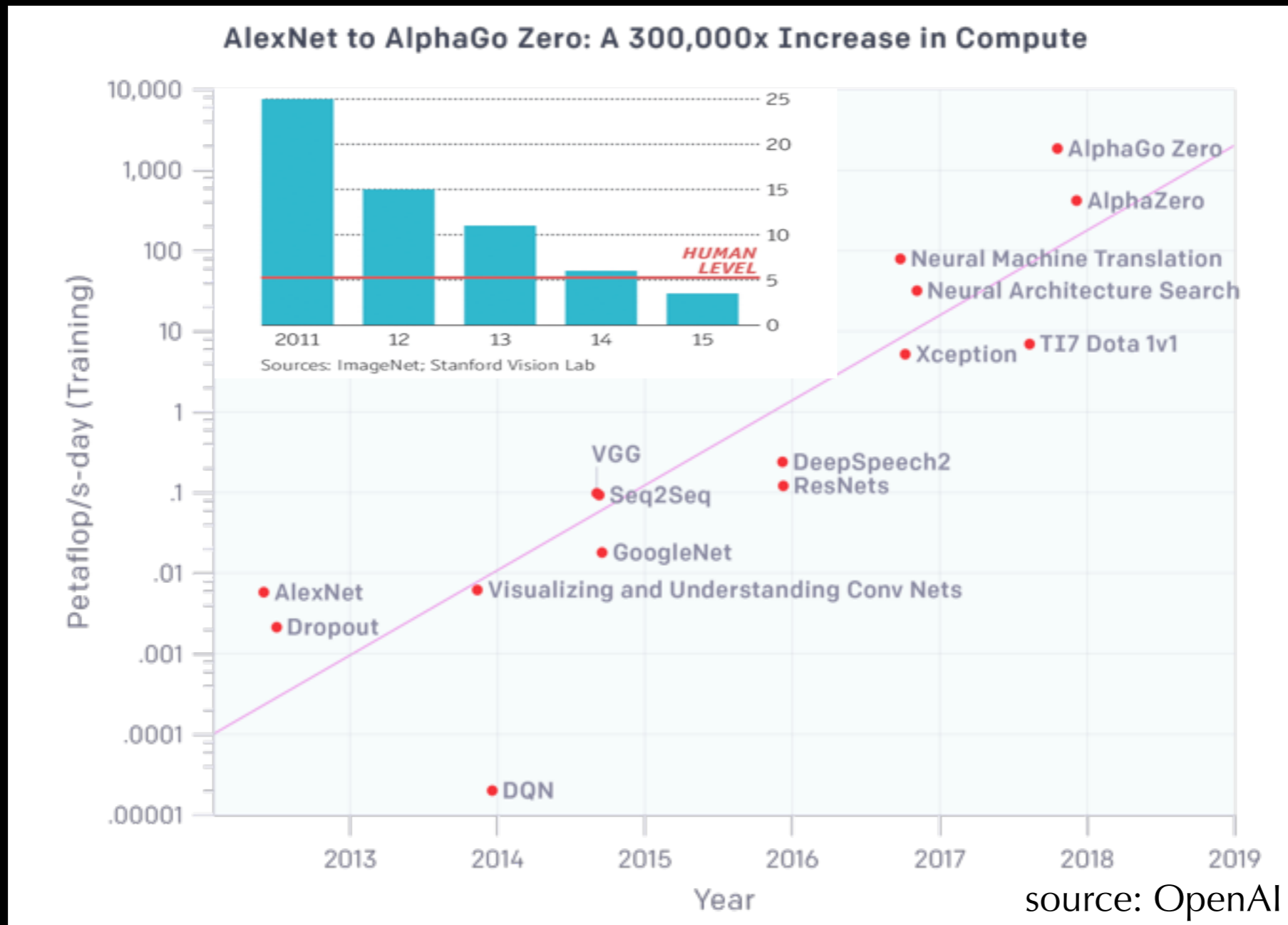
1. Deduction (Math proofs)

2. Induction (Physical sciences)

- Emerging new approach

3. Massive Computational Experiments (MCE)

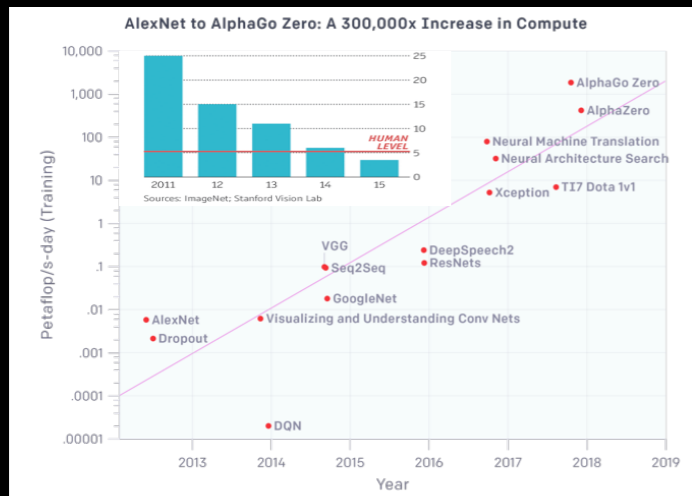
# MCE Transforming Science



amount of available compute doubles every 3.5 month  
300,000x since 2012



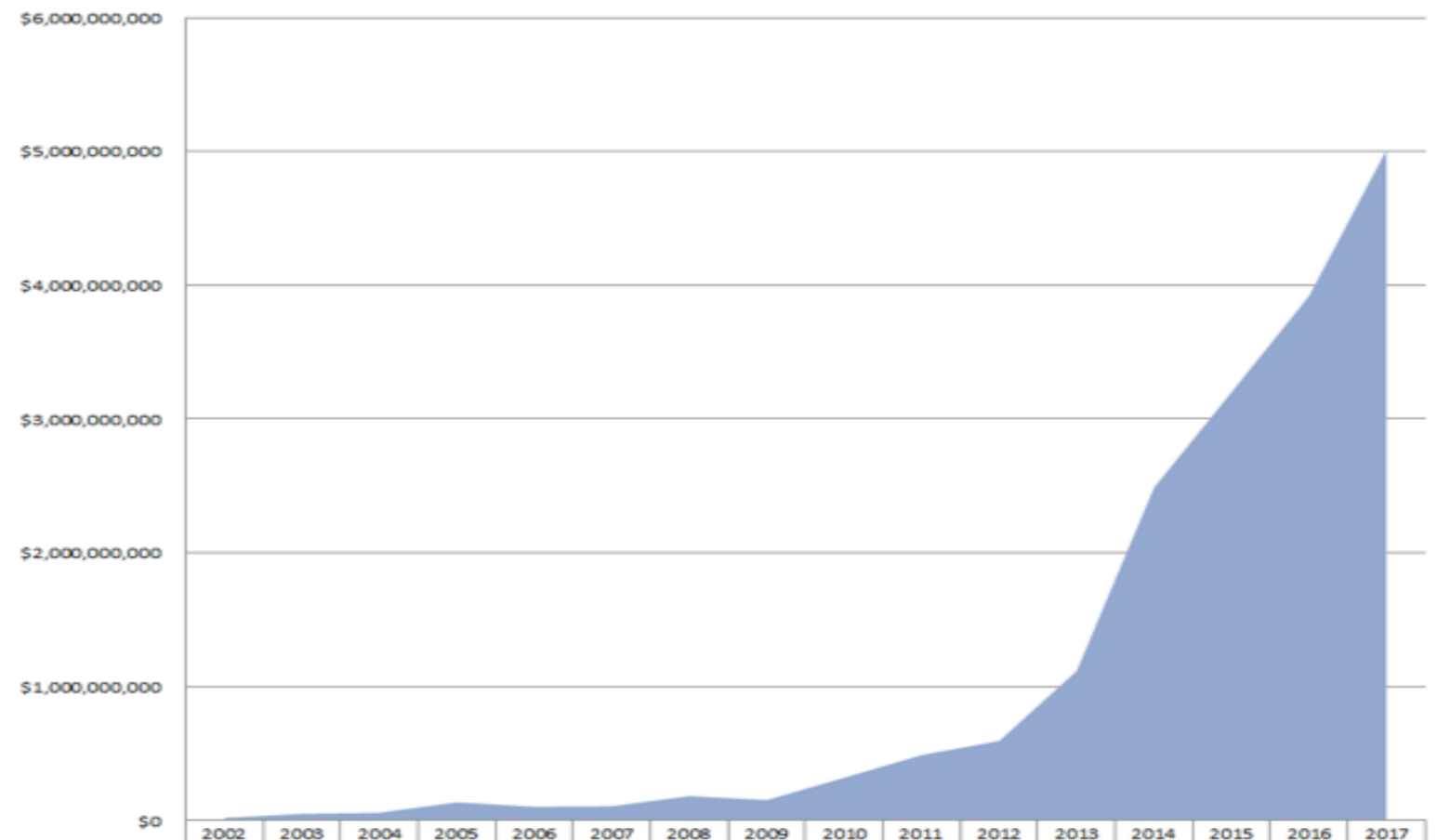
# MCE Transforming the world



22 MAR 2018

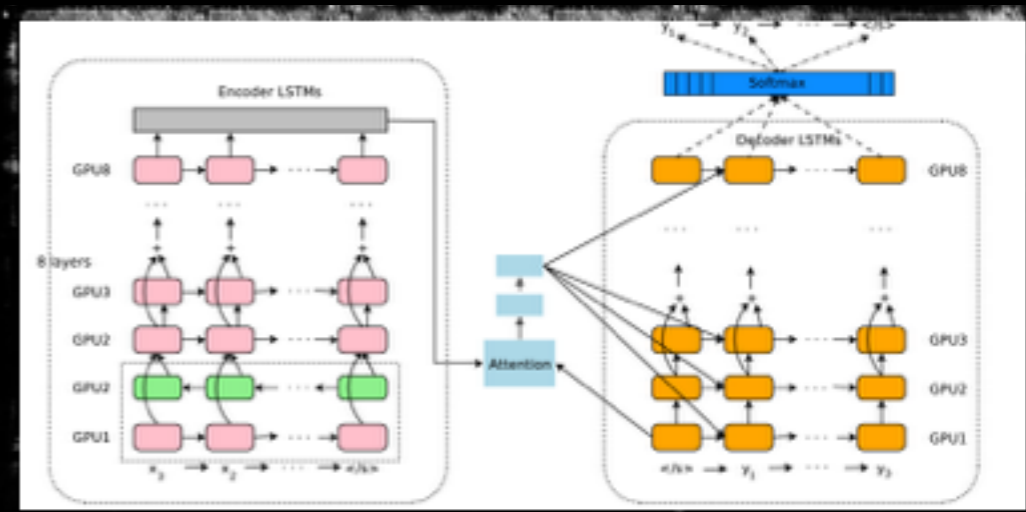
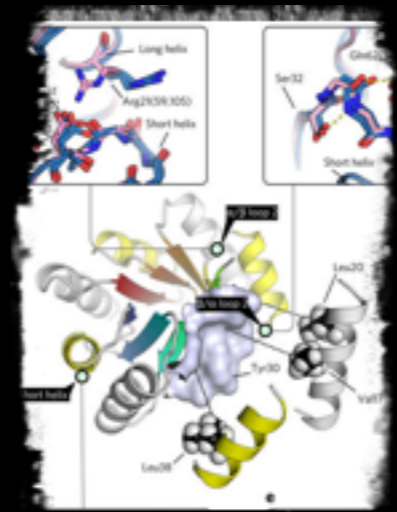
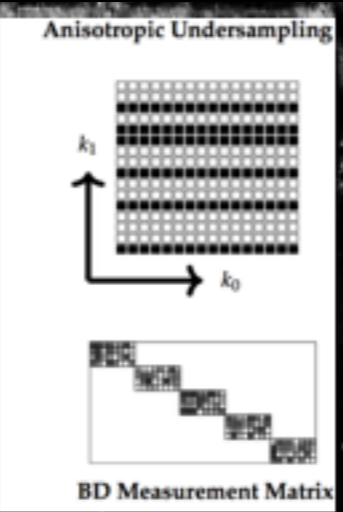
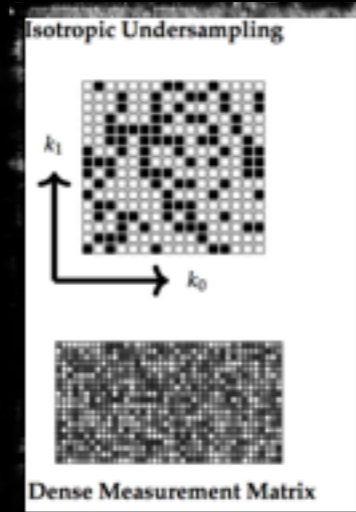
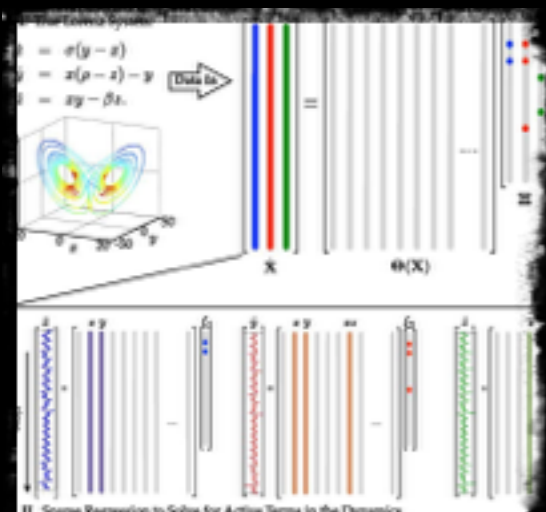
Worldwide Spending on Cognitive and Artificial Intelligence Systems Will Grow to \$19.1 Billion in 2018, According to New IDC Spending Guide

**VC Investment in AI – 2002 to 2017**



# MCEs everywhere

- Deep Learning related
  - ✓ NMT, Tesla, computer vision, etc.
- Applied Mathematics
  - ✓ Computer-aided proofs, compressed sensing
- Other areas
  - ✓ Protein design, dynamical systems, oil field dev
  - ✓ Psychology (Choosing Prediction Over Explanation in Psychology, Yarkoni 2017)



# IT-enriched Science

How does it look like?

What are the grand challenges?



# Data Science

## #21stCenturyScience

*Massive*

*Computational*

*Experiments*

*Theory*

*for guidance/interpretation*

# The grand challenges of #datascience2018

1. Conduct MCEs, crush other scientists, win prizes





# The grand challenges of #datascience2018

1. Conduct MCEs, crush other scientists, win prizes
2. Enable MCEs, win admiration of other scientists



ClusterJob

Documentation People Support Sign in Sign up

Reproducible  
High-throughput  
Computational Research

Sign up for Clusterjob

Sign up to access your computations  
anytime, anywhere.

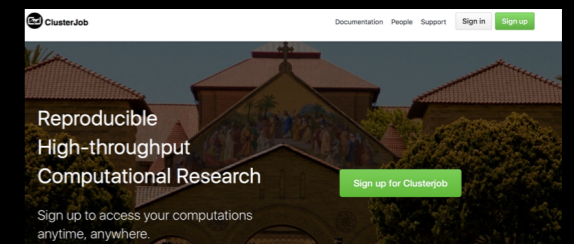


# The grand challenges of #datascience2018

1. Conduct MCEs, crush other scientists, win prizes



2. Enable MCEs, win admiration of other scientists



# Typical Data Science Workflow

1. Precise **specification** of experiments
2. **Distribution and monitoring** of all jobs
3. **Harvesting** data
4. **Analysis** of data
5. Inductive **iterations** of 1-4 (suggested/required by 4)
6. **Dissemination** of acquired knowledge

How can you do  
MCEs  
*Painlessly?*




# Experiment Management System

(Painless Frameworks for Massive Experiments)

1. Systematic structure to coding/experiment definition
2. Automatic access to the cloud/HPC-clusters
3. Automatic harvesting and analysis using defined tools
4. Automatic reproducibility
5. Easy sharing/collaboration/dissemination

# Examples of Painless Framework

 ClusterJob

Reproducible  
High-throughput  
Computational Research

Sign up to access your computations  
anytime, anywhere.

## ElastiCluster

aims to provide a user-friendly command line tool to create, manage and setup computing clusters hosted on cloud infrastructures like [Amazon's Elastic Compute Cloud EC2](#), [Google Compute Engine](#), or a private [OpenStack](#) cloud. Its main goal is to get your compute cluster up and running with just a few commands.

[Read the Documentation](#)

[Install ElastiCluster](#)

[How it works](#)

[Demo Video](#)

The a  
define  
specif  
Using  
cluste

[pywren](#) [blog](#) [getting started](#) [documentation](#) [examples](#) [code](#) [bugs](#)

## pywren

Pywren lets you run your existing python code at massive scale via AWS Lambda

### Overview

```
def my_function(b):  
    x = np.random.normal(0, b, 1024)  
    A = np.random.normal(0, b, (1024, 1024))  
    return np.dot(A, x)  
  
pwex = pywren.default_executor()  
res = pwex.map(my_function, np.linspace(0.1, 100, 1000))
```

# CodaLab

Accelerating reproducible computational research.

[Worksheets](#)

Share notebooks and create executable papers

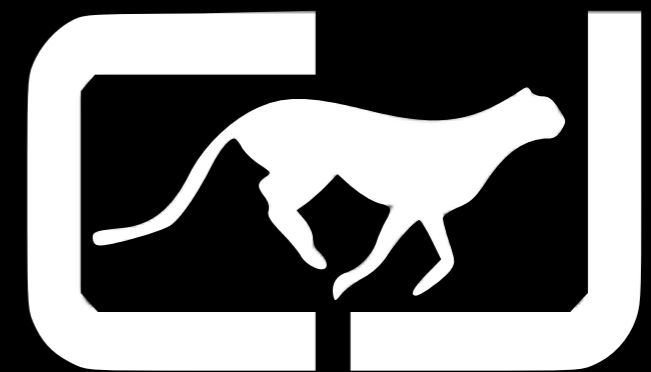
[Competitions](#)

Enter an existing competition to solve challenging problems, or host your own.

3 models  
3 abstractions

# Monajemi-Murri Model

*Elasticcluster*





# MCEs push-button, *Literally!*

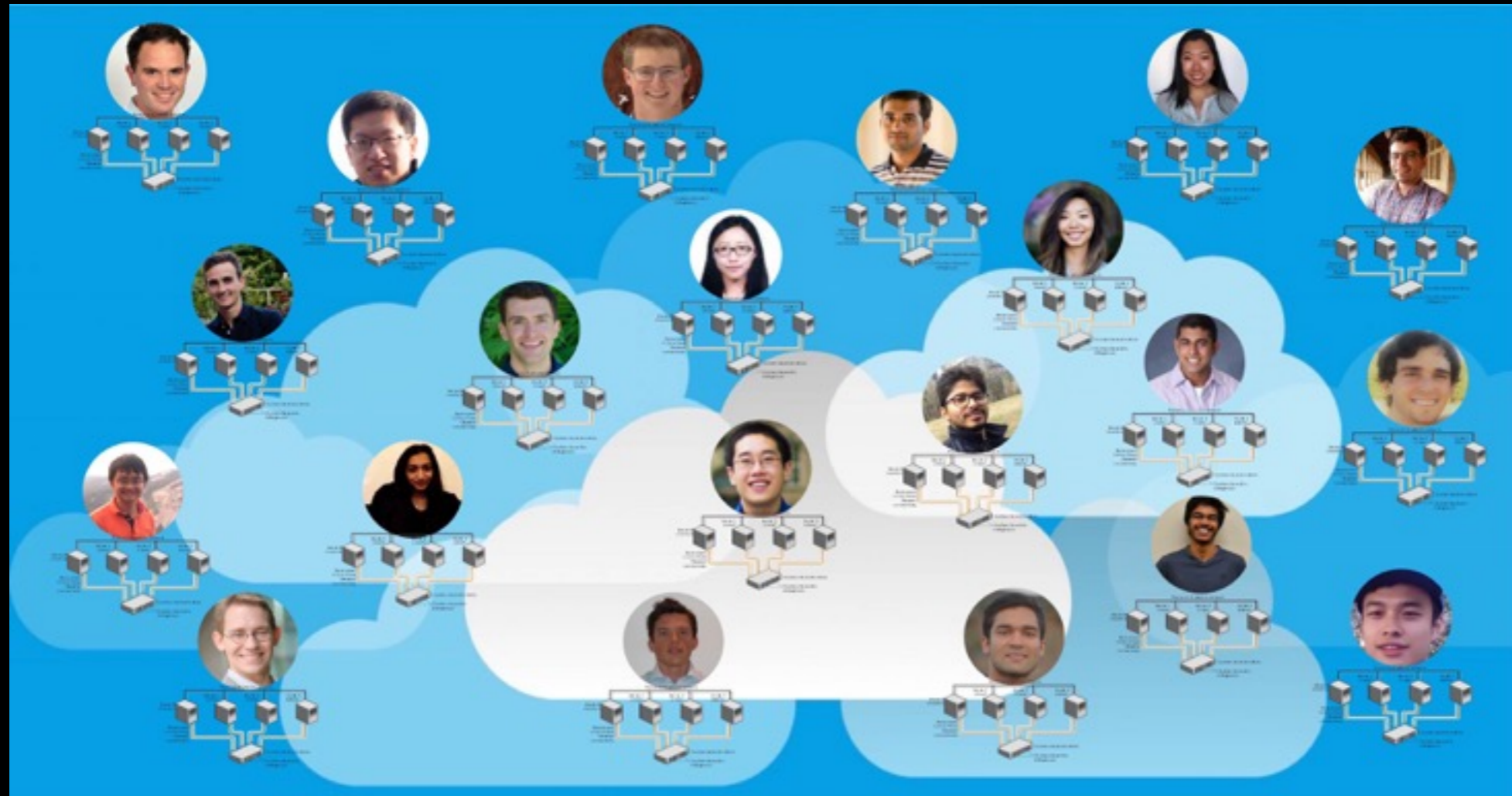
1. build personal CPU/GPU cluster (~20 min)

```
elasticcluster start gce
```

2. Fire up 1000's of jobs

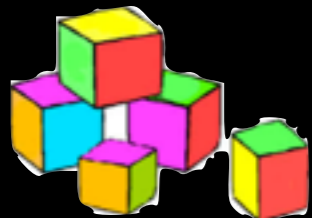
```
cj parrun train.py gce
```

# Stats285 discovers math in the cloud



- 50 students trained 1500 Deep Nets in one computing day
- Each build his/her GPU cluster on Google Cloud
- collectively discovered new phenomena in Deep Learning
- PNAS paper in progress ...

# CodaLab Model



Bundles (Immutable)

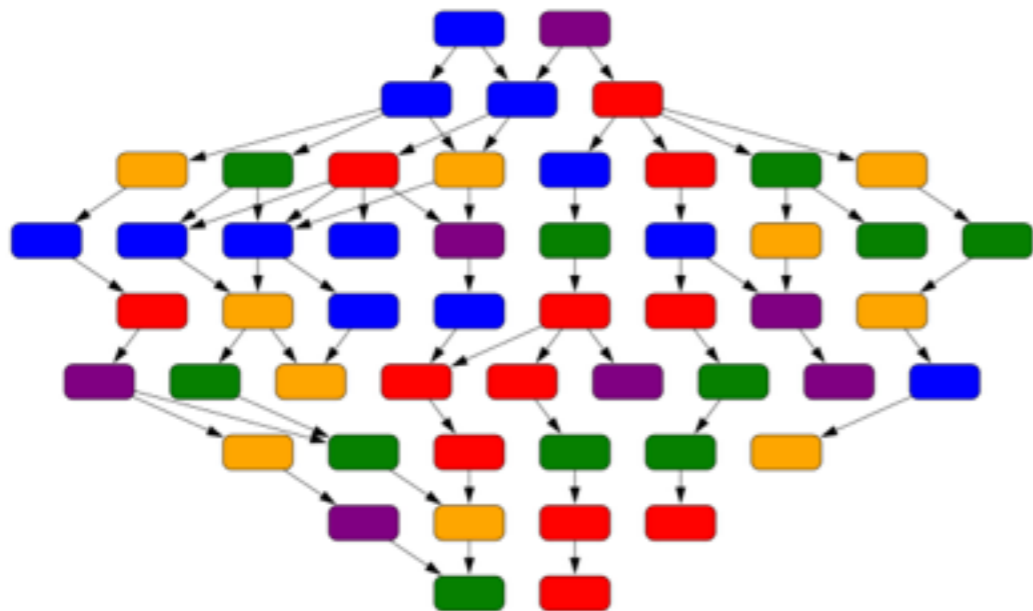


Worksheets

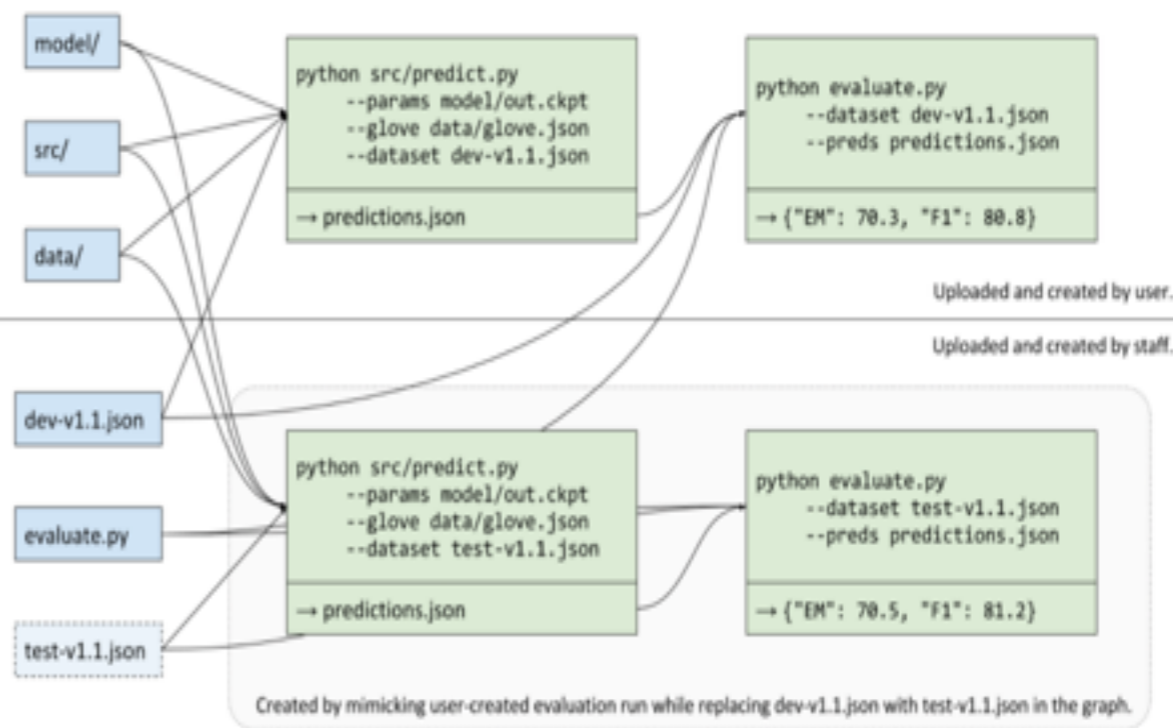
## Modularity

Real-world problems require efforts of entire community

People specialize, contribute in decentralized way

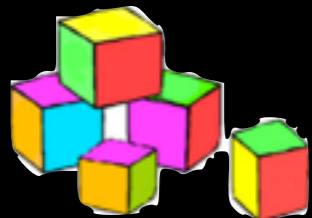


## Evaluation using "mimic"



<https://competitions.codalab.org>

# CodaLab Model



Bundles (Immutable)



Worksheets

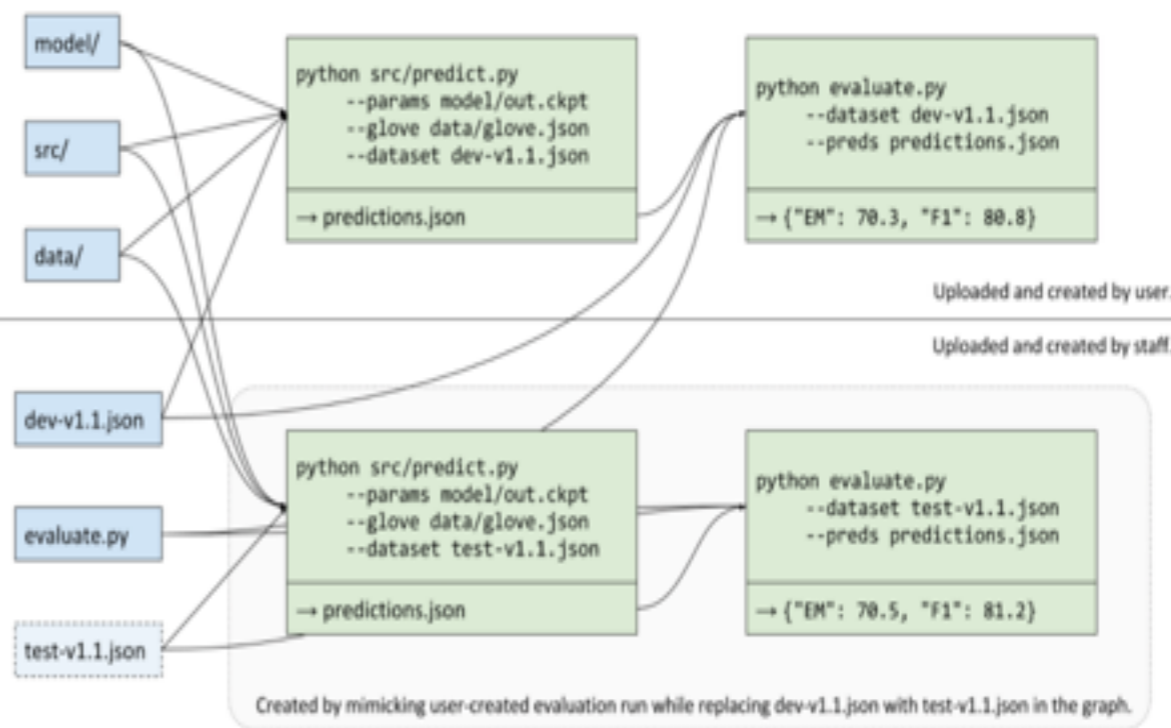
## Modularity

Real-world problems require efforts of entire community

People specialize, contribute in decentralized way



## Evaluation using "mimic"

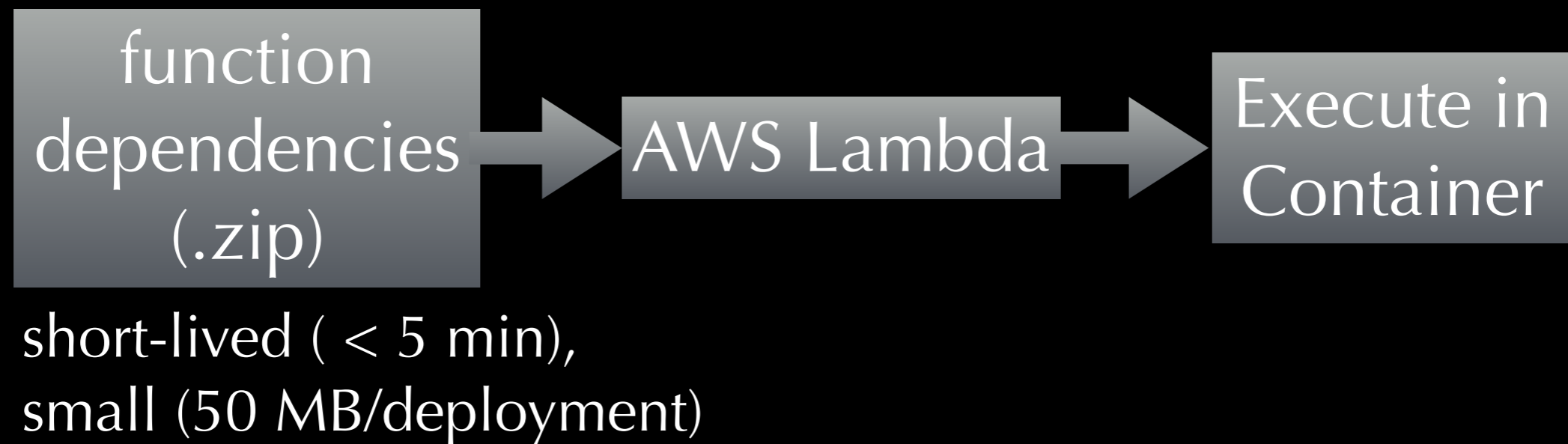


More at <https://stats285.github.io>



# Serverless Computing: PyWren

Abstract away server provisioning



# Serverless Computing: PyWren

Abstract away server provisioning



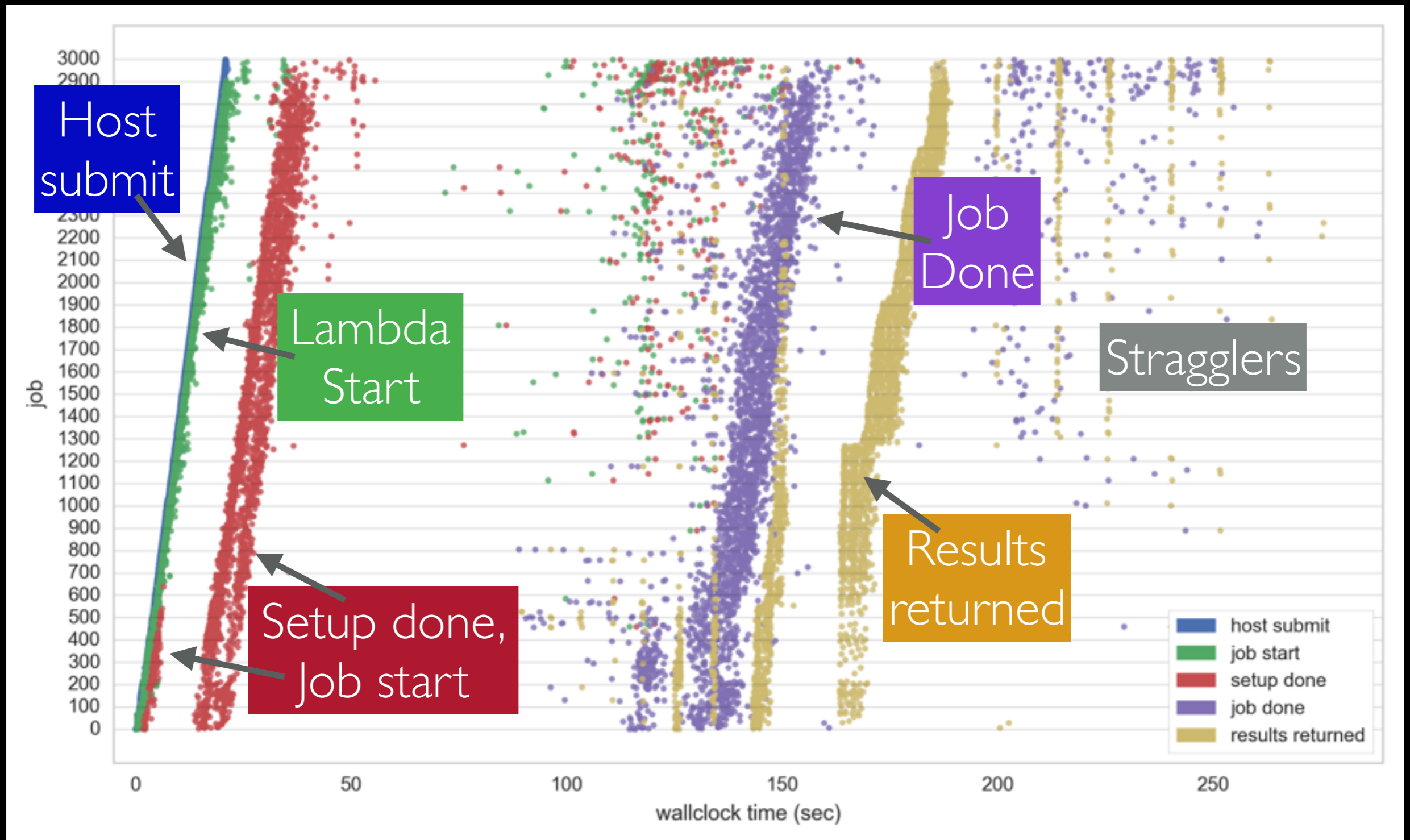
short-lived (< 5 min),  
small (50 MB/deployment)

**PyWren does all the work for you**

```
futures = exec.map(function, data)
```

```
answer = exec.reduce(reduce_func, futures)
```

# Lots of small jobs



More at <https://stats285.github.io>

# Conclusion

MCEs are transforming Science

MCEs can be made painless and transparent through EMS

We are excited to be an enabler of this transformation

[clusterjob.org](http://clusterjob.org)

[codalab.org](http://codalab.org)

[pywren.io](http://pywren.io)