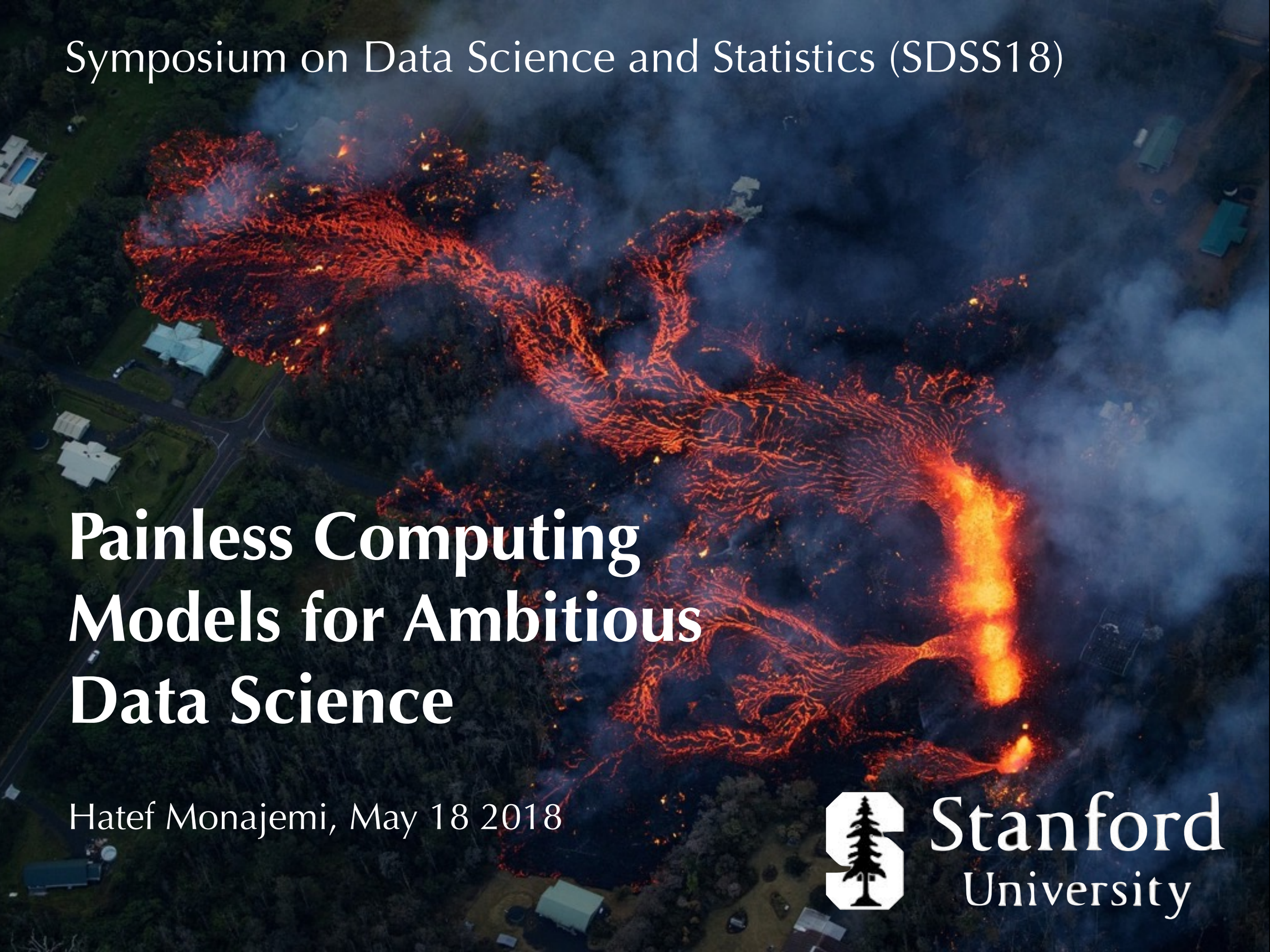
Symposium on Data Science and Statistics (SDSS18)



Painless Computing Models for Ambitious Data Science

Hatef Monajemi, May 18 2018

Stanford University

Coauthors

David Donoho
(Stanford)


Riccardo Murri
(UZH)


Percy Liang
(Stanford)

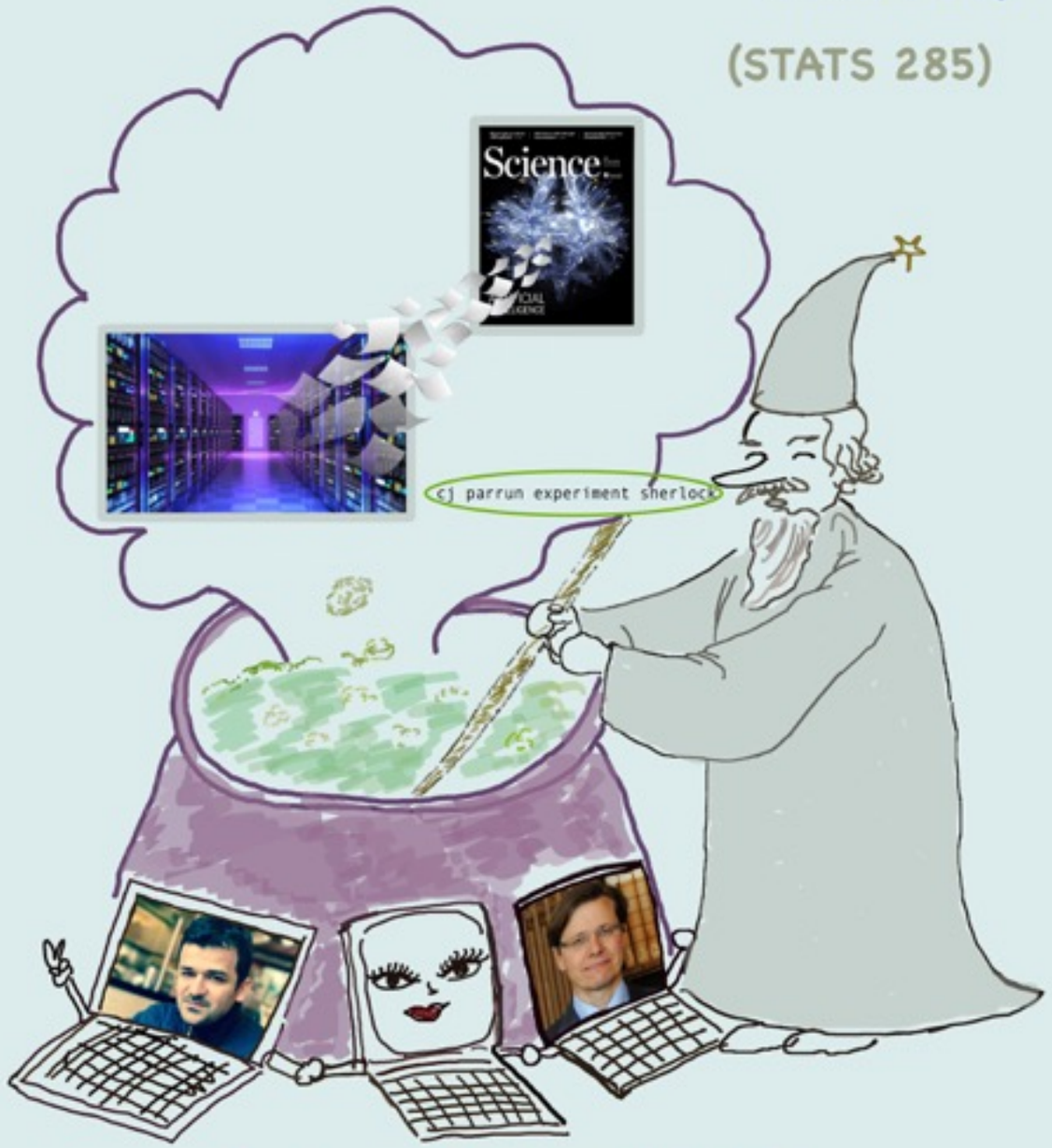
Eric Jonas (UCB)

Victoria Stodden
(UIUC)

@stats285

Massive Computational Experiments, Painlessly

(STATS 285)



Time: Monday 3:00 - 4:20
Place: Thornt110
Website: stats285.github.io

2012

The world changed

How to advance knowledge?

use a better mathematical model

1800

experiment until you find a winner

2012

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

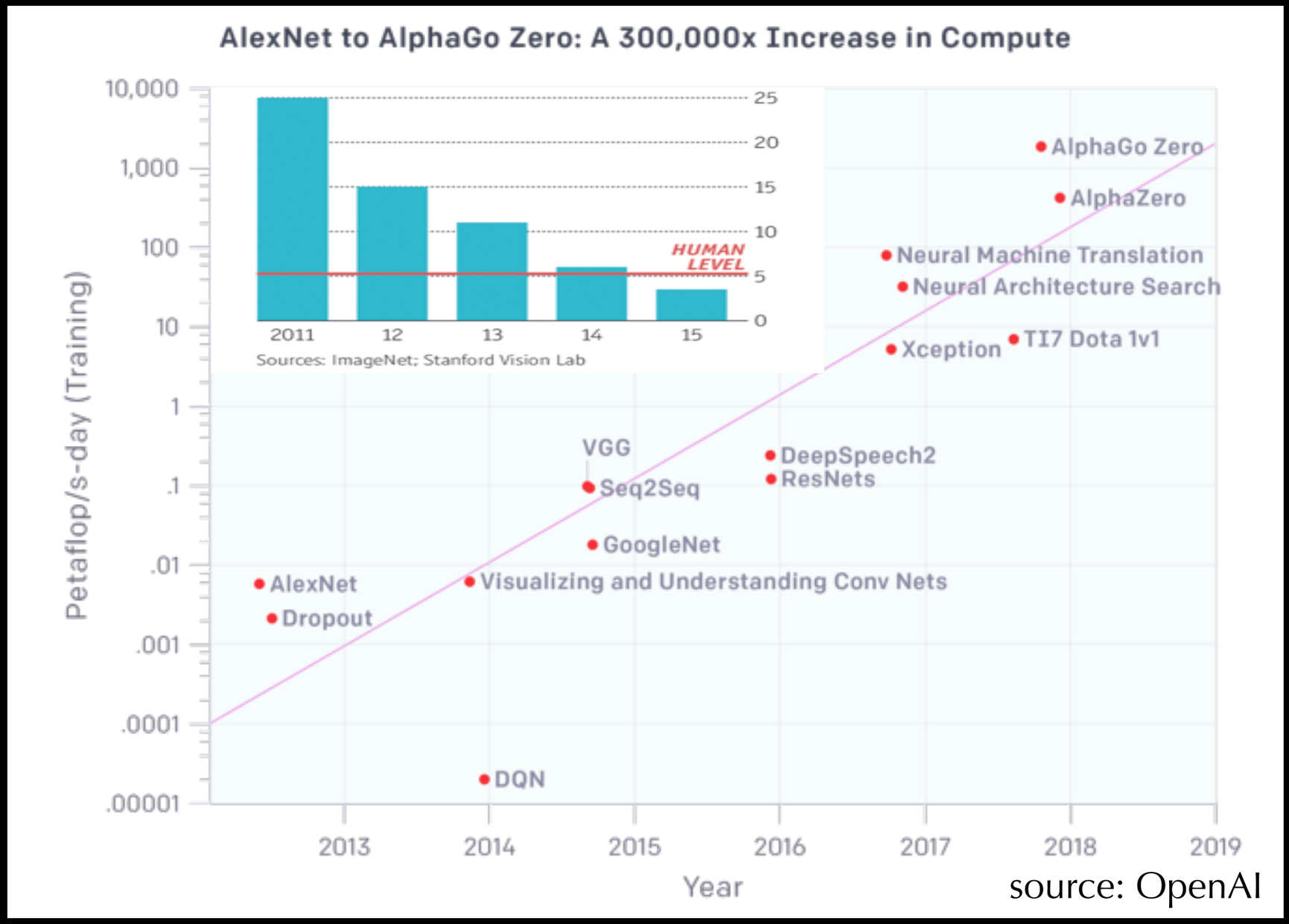
- Traditionally

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

- Emerging new approach

    3. Massive Computational Experiments (MCE)

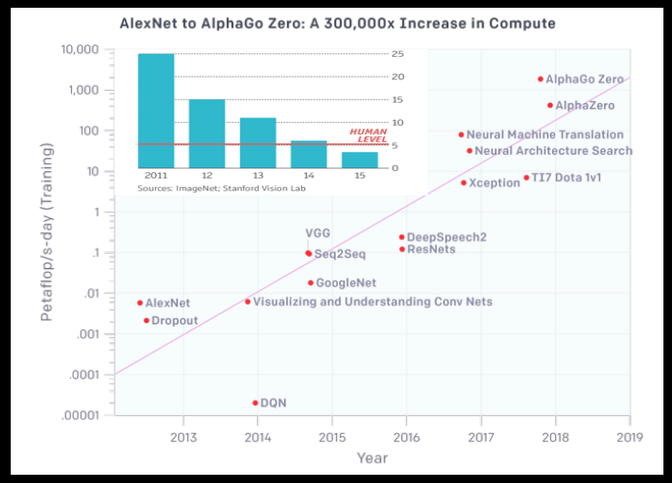
MCE Transforming Science

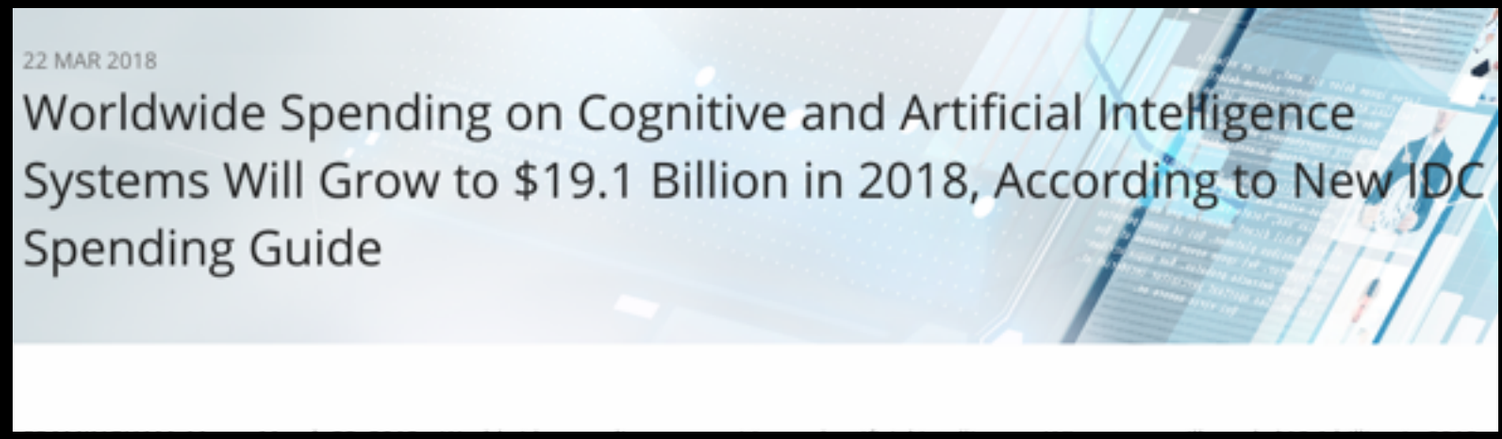


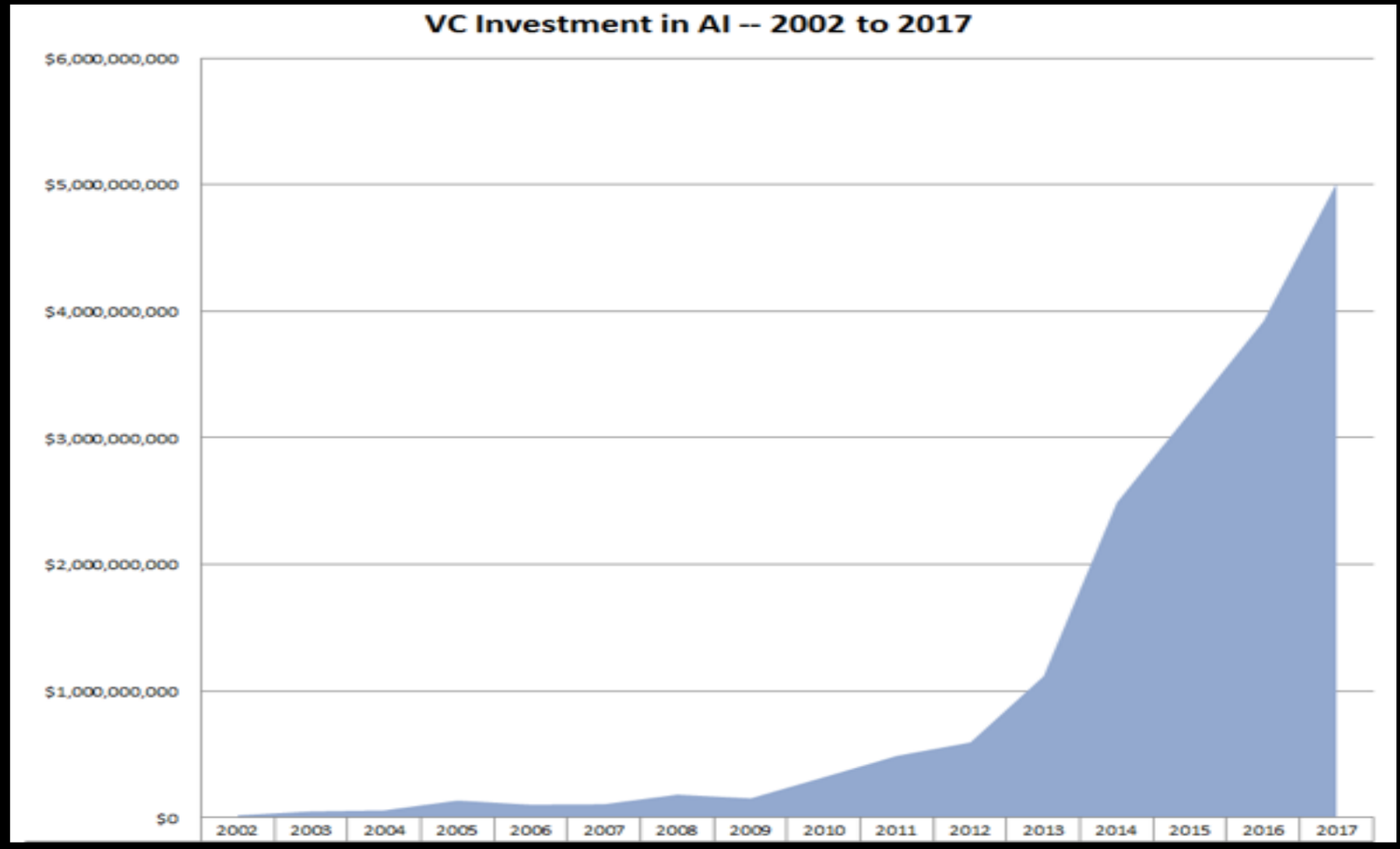
amount of available compute doubles every 3.5 month

300,000x since 2012

MCE Transforming the world



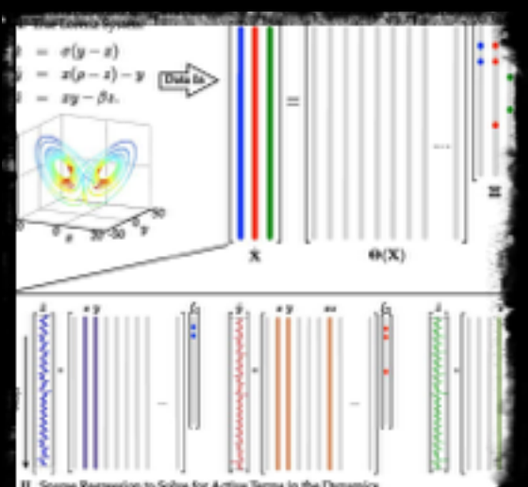


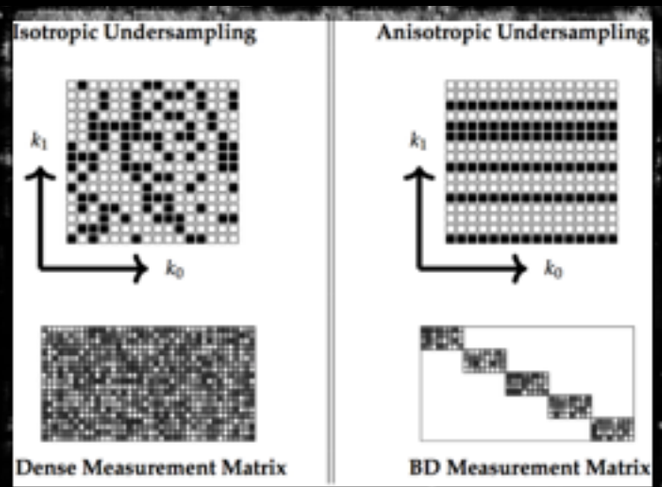


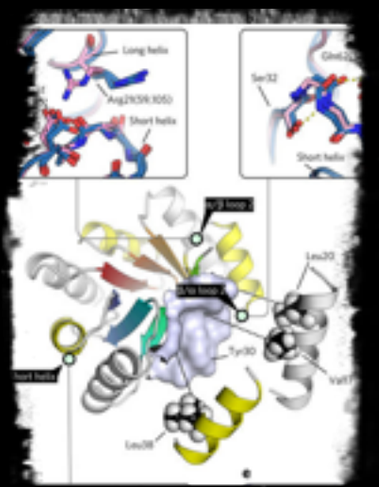
https://www.datanami.com/2018/05/11/inside-one-vc-firms-hands-on-approach-to-ai/

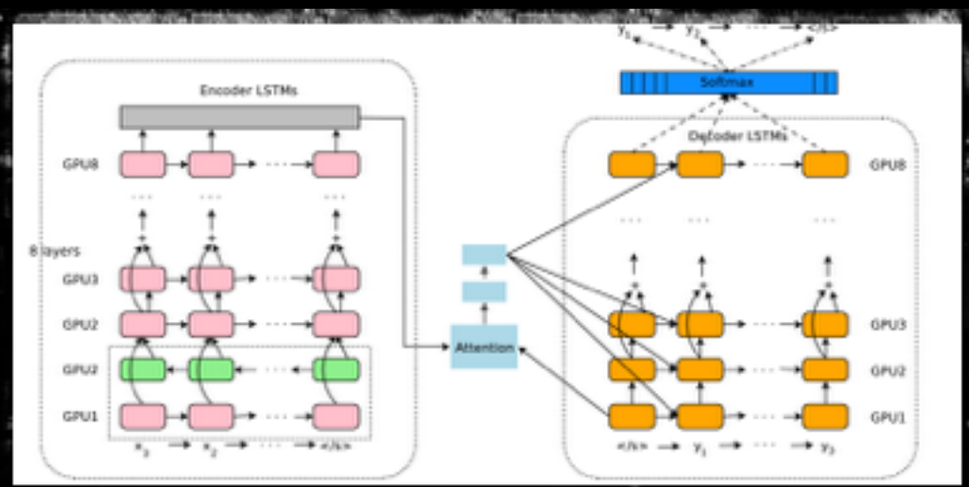
MCEs everywhere

- 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

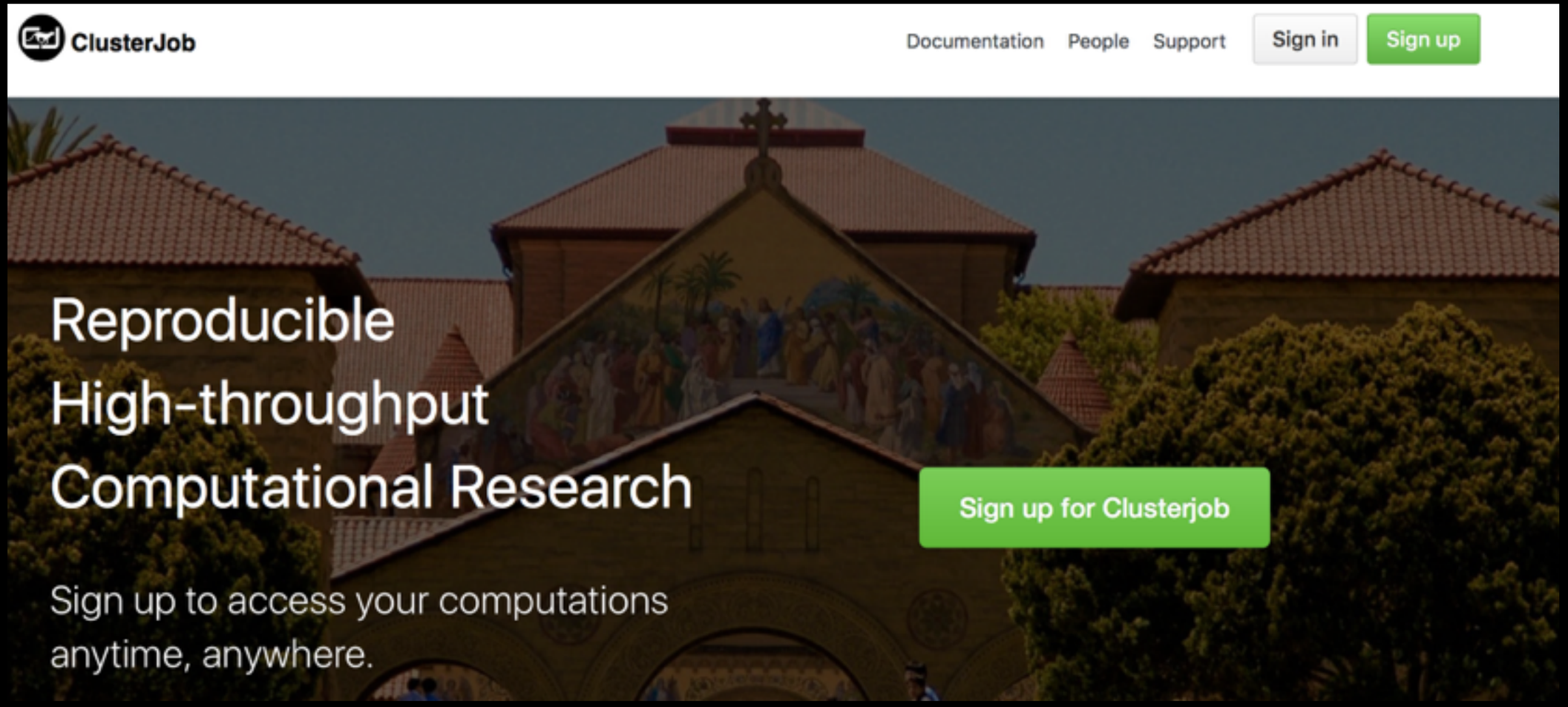
The grand challenges of #datascience2018

1. Conduct MCEs, crush other scientists, win prizes

The grand challenges of #datascience2018

1. Conduct MCEs, crush other scientists, win prizes

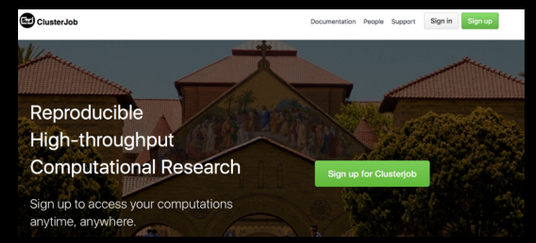
2. Enable MCEs, win admiration of other scientists



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

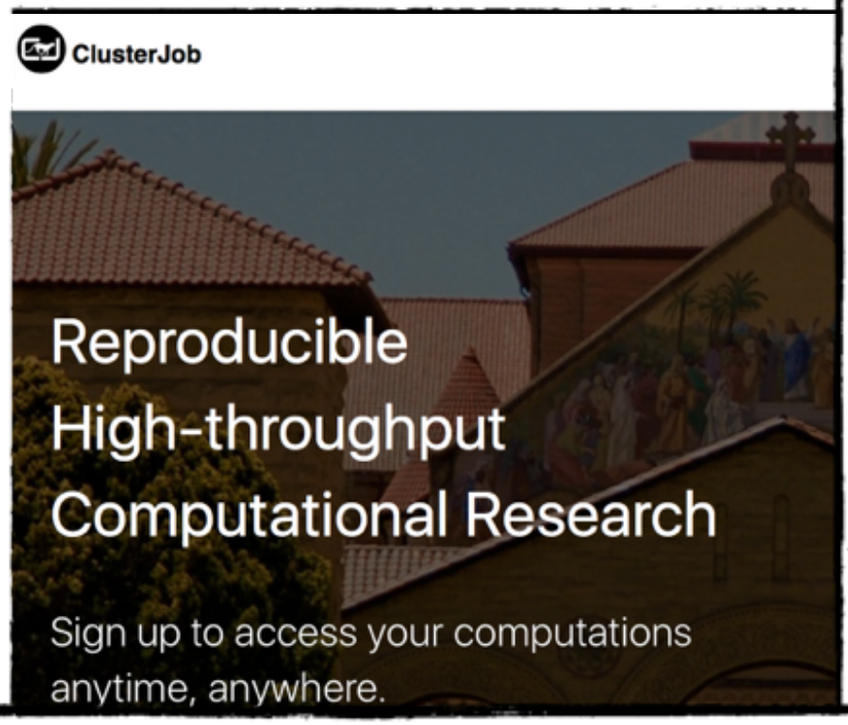
Painlessly?

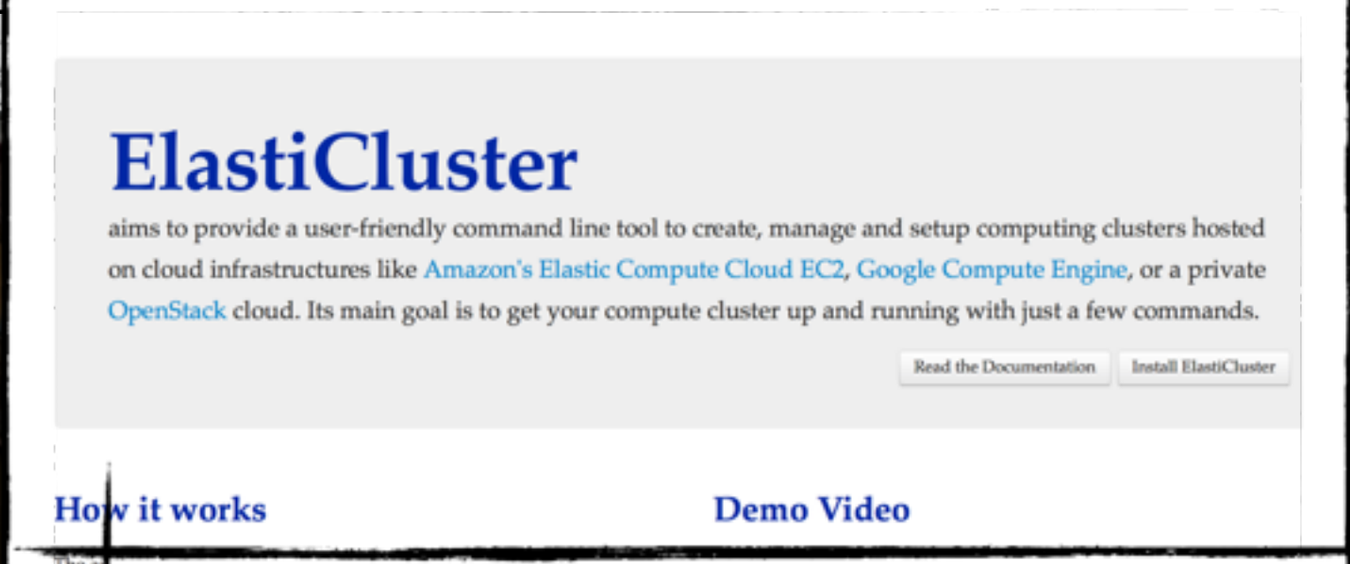
Experiment Management System

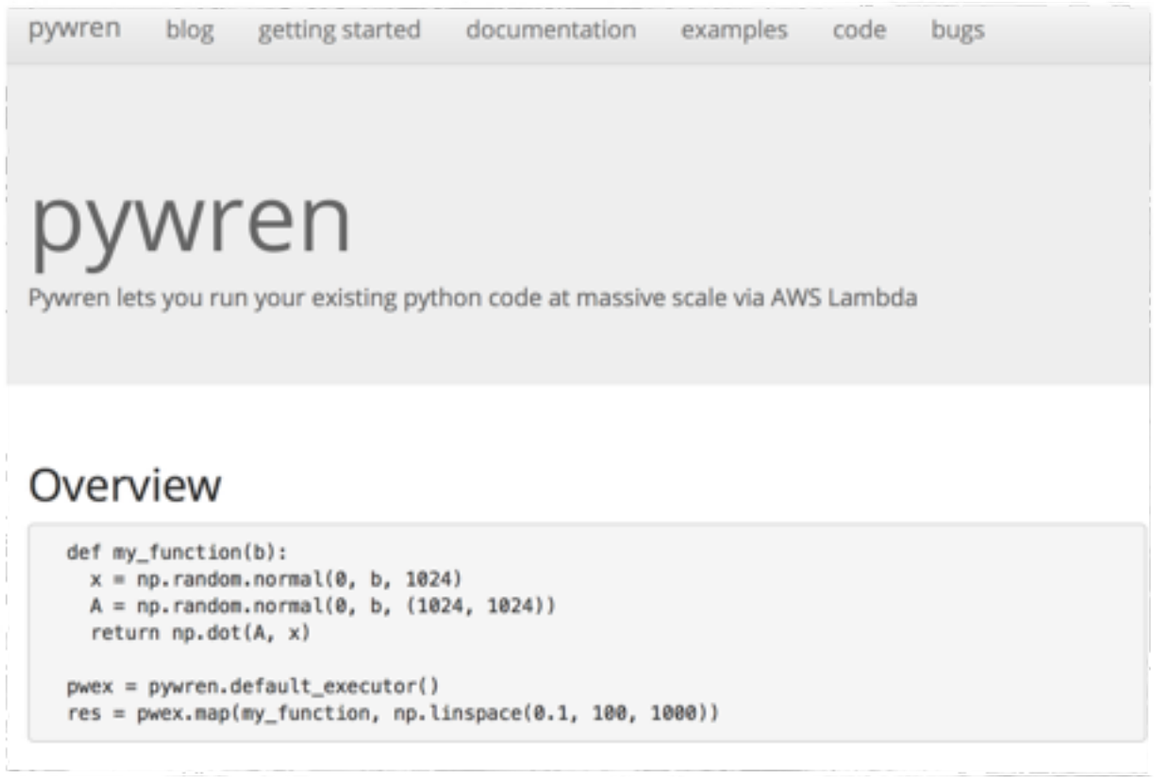
(Painless Frameworks for Massive Experiments)

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





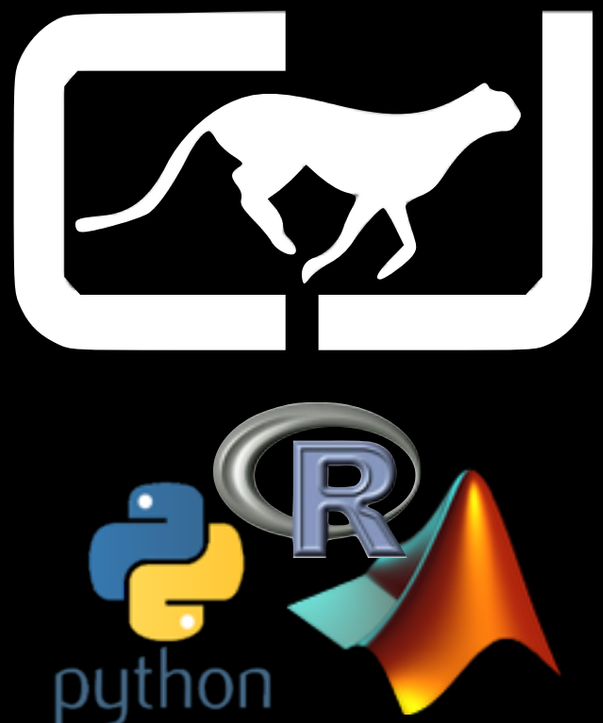


3 models

3 abstractions

Monajemi-Murri Model

ElasticCluster



MCEs push-button, *Literally*!

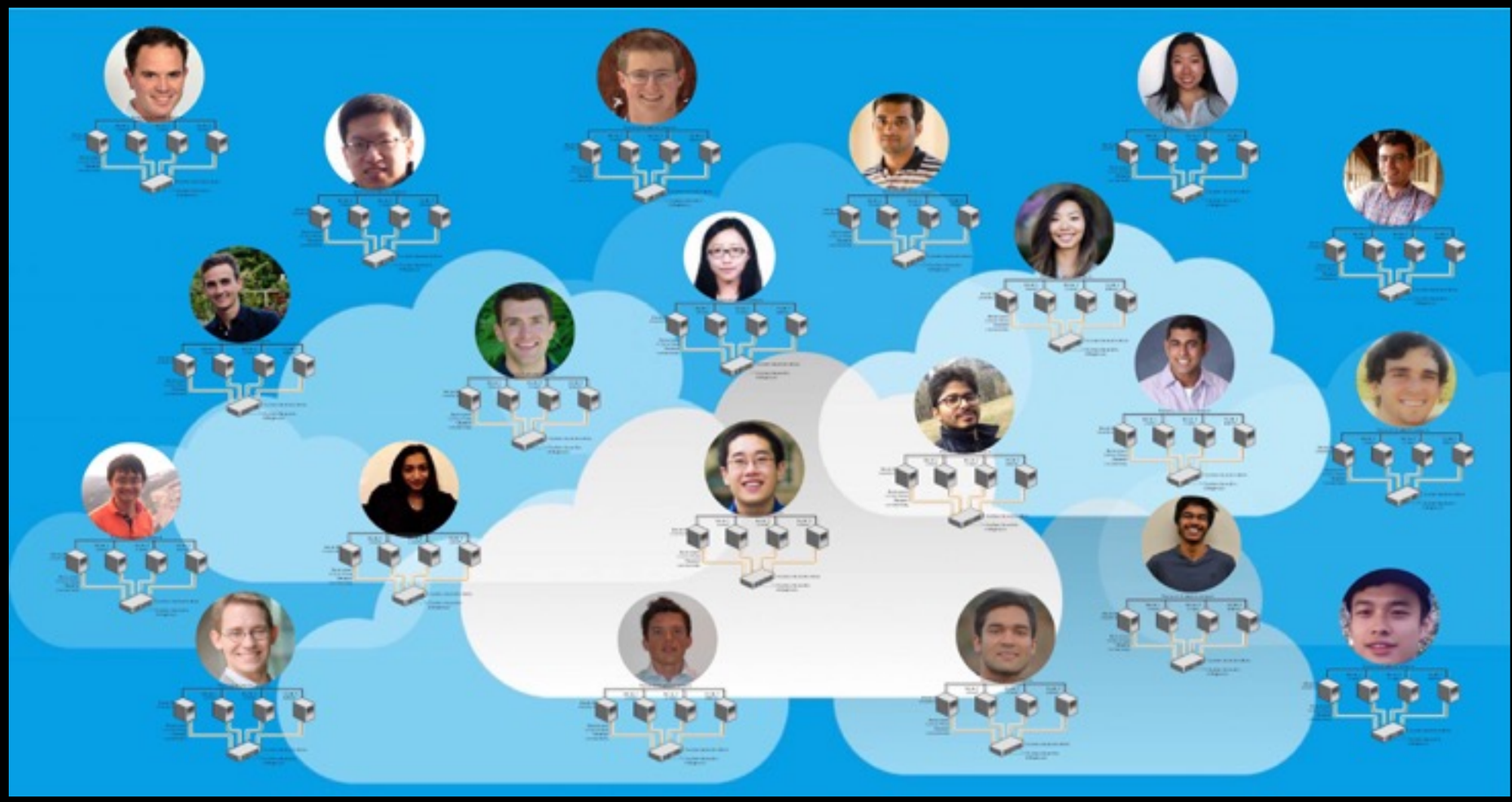
1. build personal CPU/GPU cluster (~20 min)

```
elasticluster start gce
```

2. Fire up 1000's of jobs

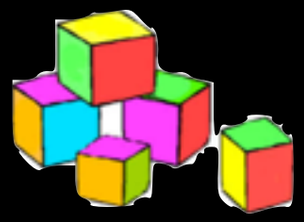
```
cj parrun train.py gce
```

Stats285 discovers math in the cloud



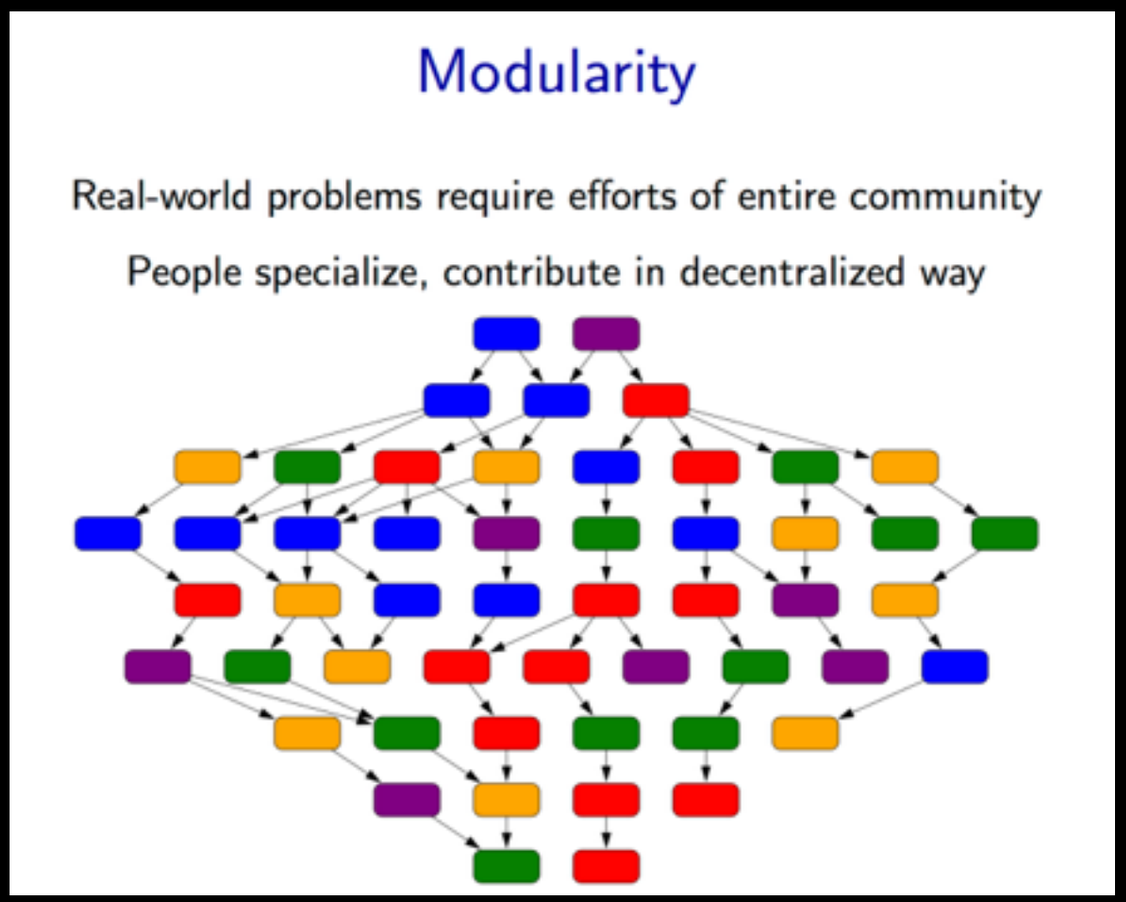
- 50 students trained 1500 Deep Nets in one computing day
- Each build his/her GPU cluster on Google Cloud
- collectively discovered new phenomena in Deep Learning
- PNAS paper in progress …

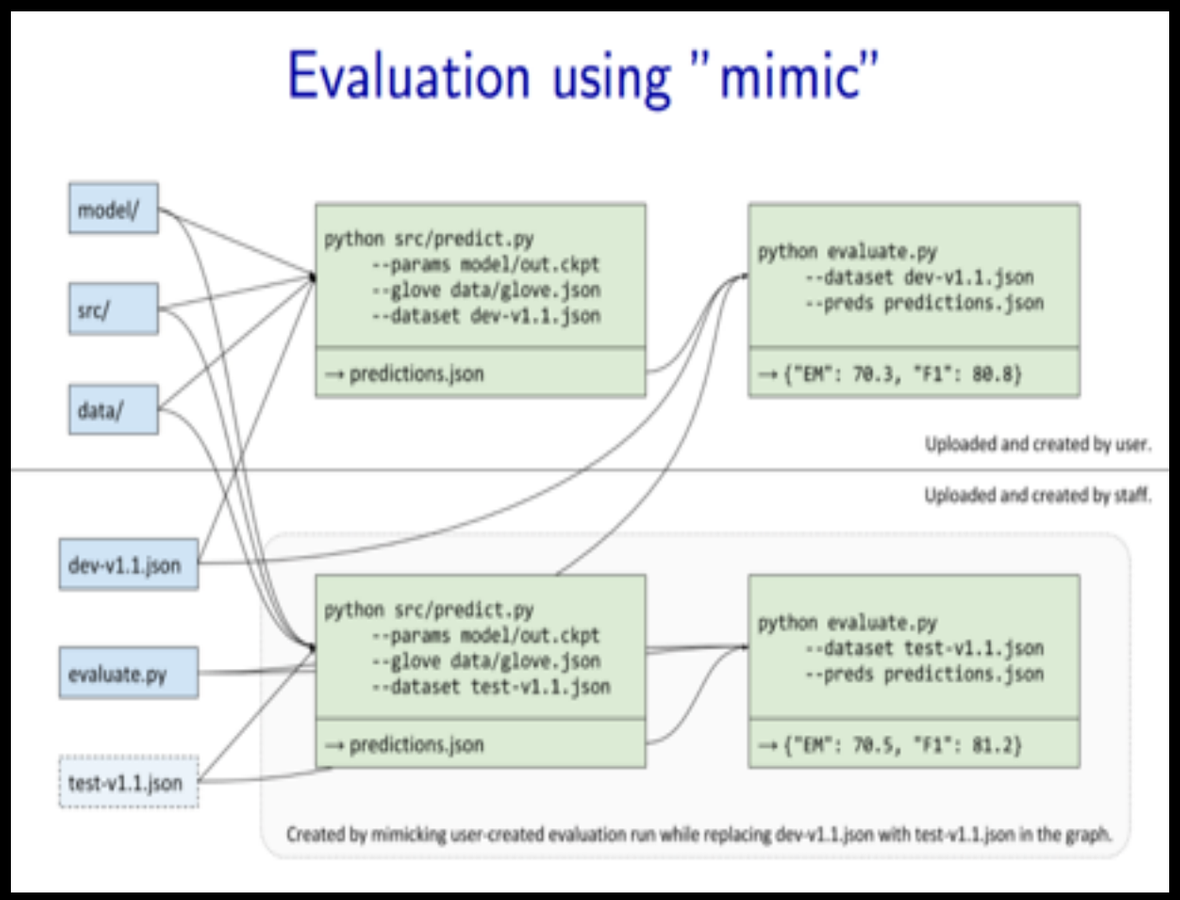
CodaLab Model



Bundles (Immutable)

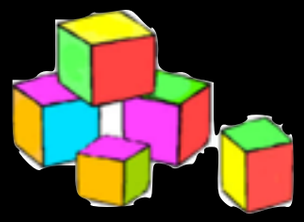
Worksheets





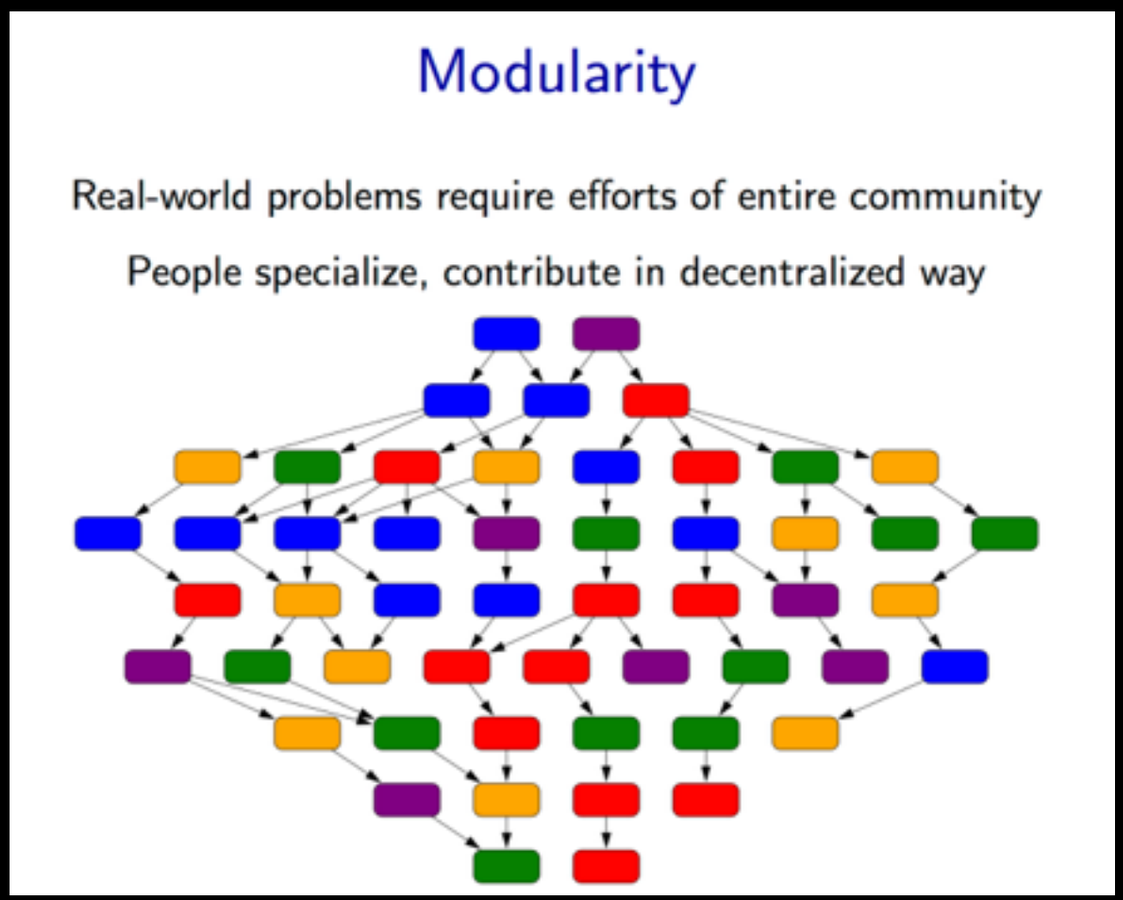
https://competitions.codalab.org

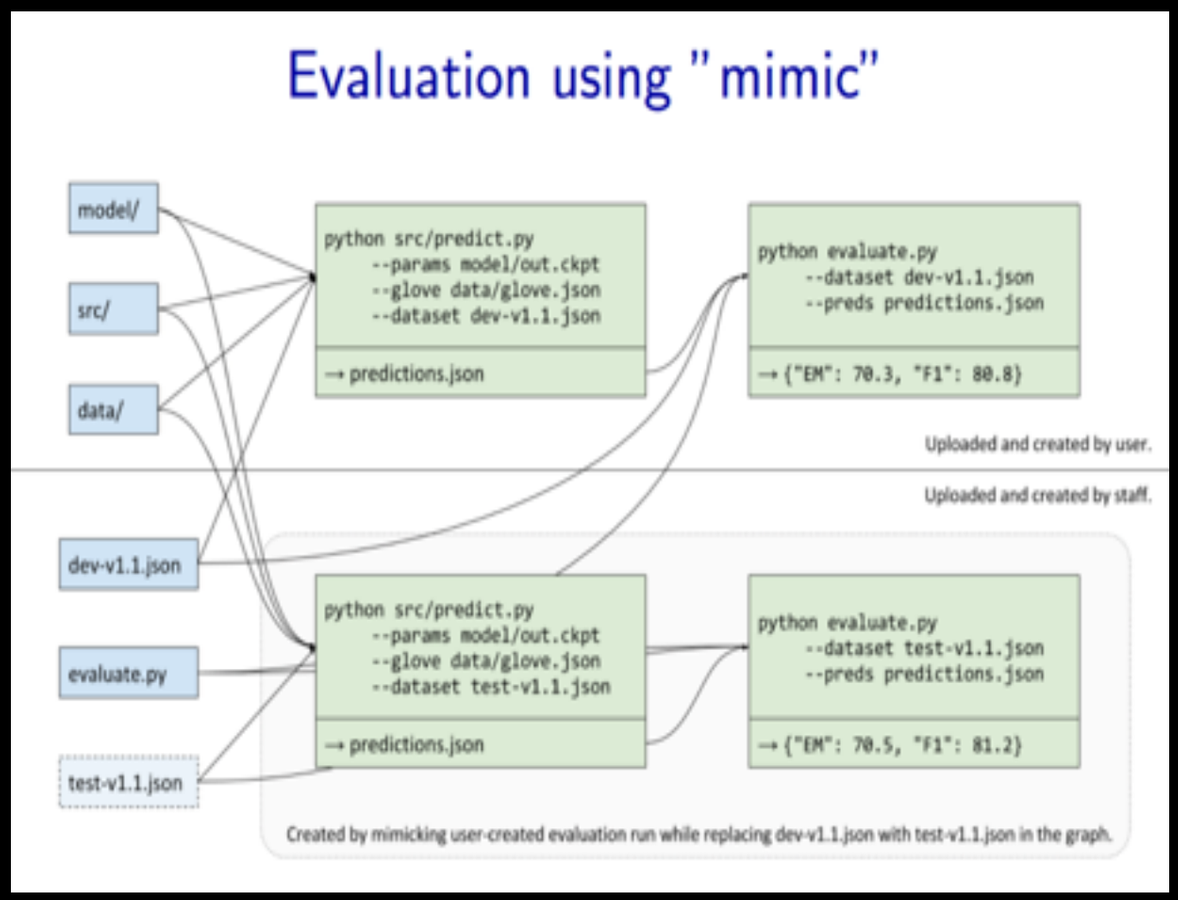
CodaLab Model



Bundles (Immutable)

Worksheets





More at https://stats285.github.io

Serverless Computing: PyWren

Abstract away server provisioning

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

Serverless Computing: PyWren

Abstract away server provisioning

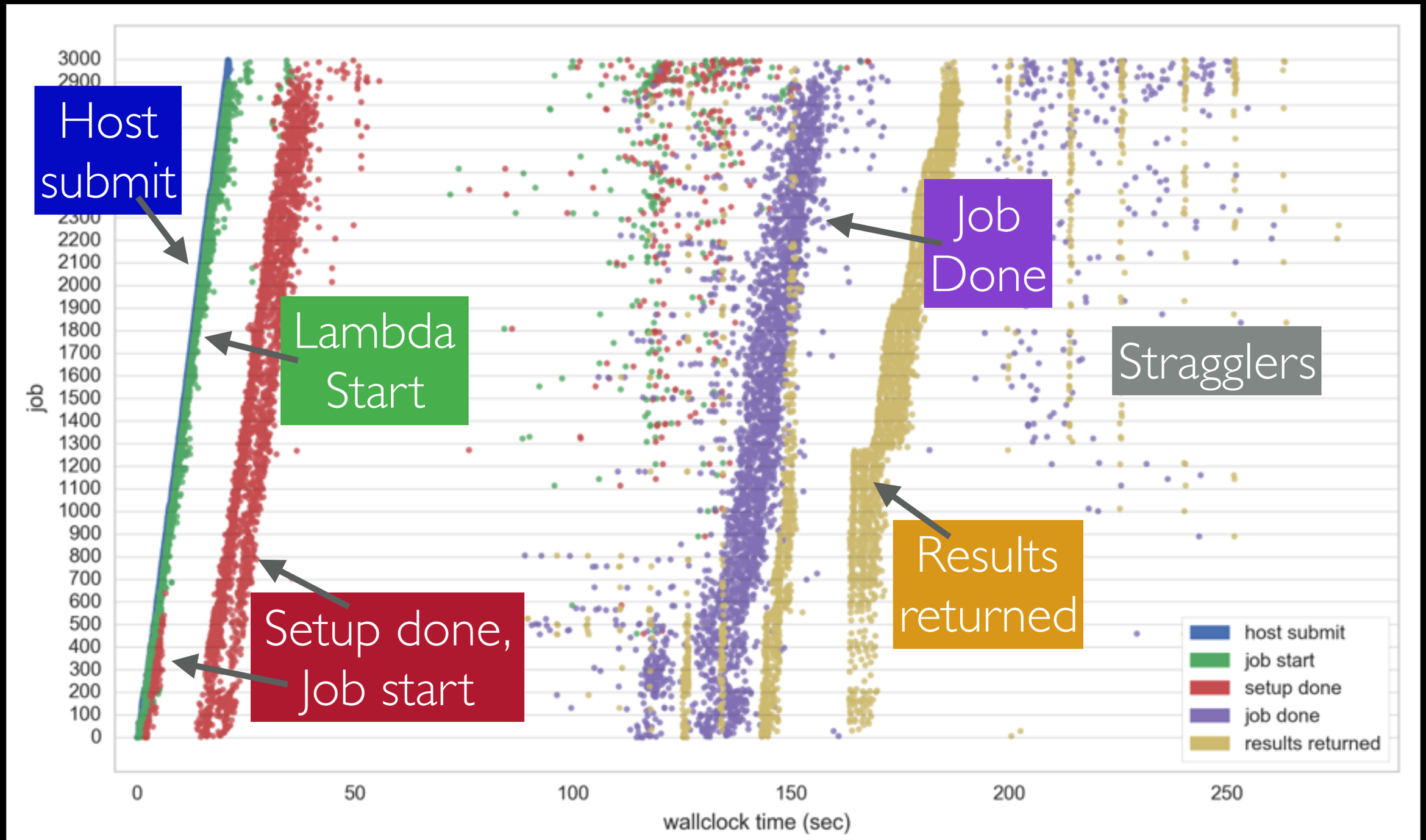
short-lived (< 5 min),
small (50 MB/deployment)

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 codalab.org pywren.io