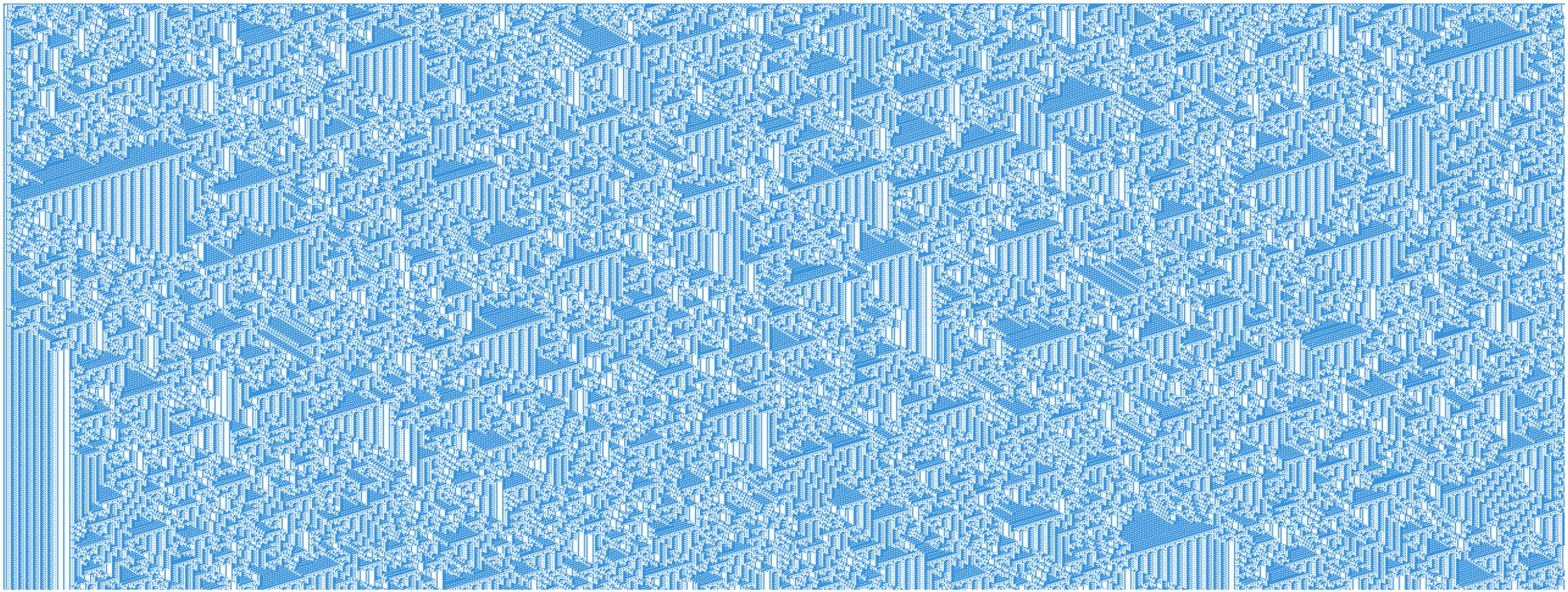


TEACHING BIG DATA WITH A VIRTUAL CLUSTER

Joshua Eckroth
Stetson University
SIGCSE 2016

A DEFINITION OF "BIG DATA"

"Data mining and analysis require 'big data' techniques when the data have such **high volume** or **high velocity** that **more than one machine** are required to store and/or process the data."

GOAL: GIVE STUDENTS HANDS-ON EXPERIENCE

Students understand theory better when they apply it

Familiarity with industry-standard tools helps their prospects in the job market

Comfort with esoteric and complex tools is part of being a good computer scientist

INDUSTRY-STANDARD TOOLS

Hadoop
- HDFS
- Map-Reduce
- Hive
- Mahout

Our focus today

Spark

R, ggplot

Weka

THE CHALLENGE OF TEACHING BIG DATA

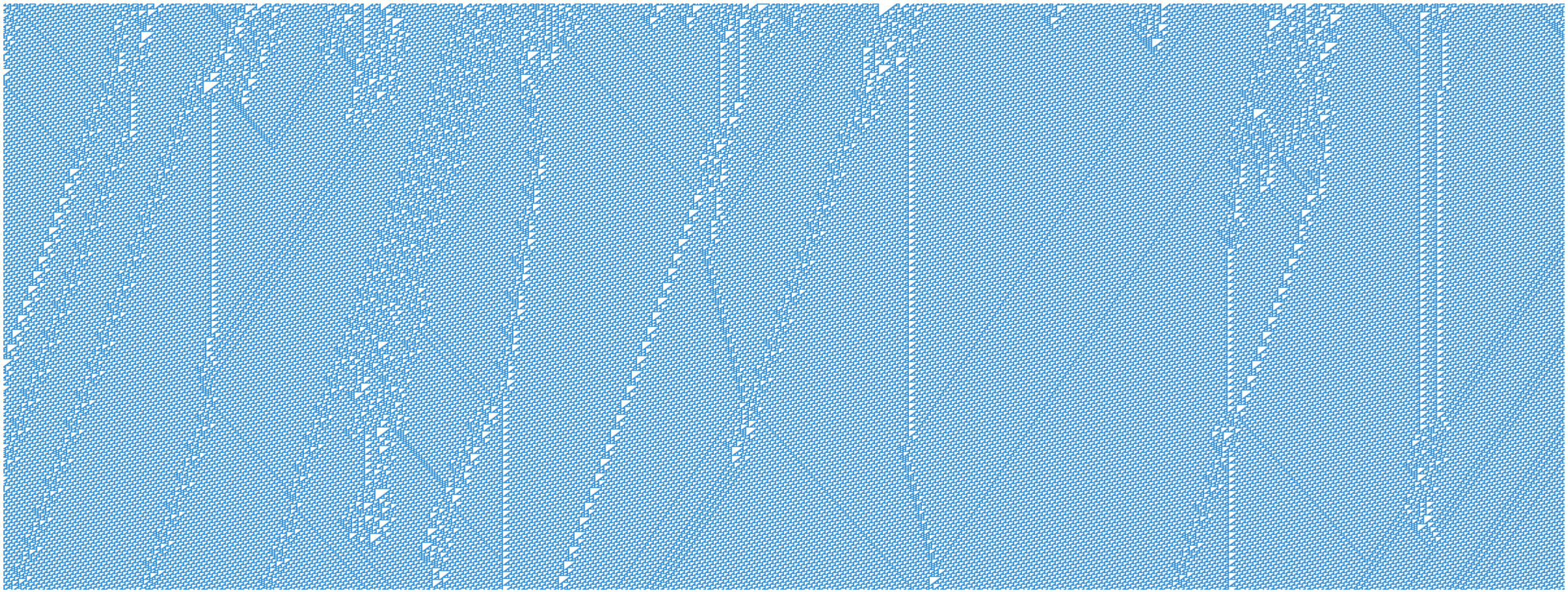
The challenge is not due to a lack of data

- Census.gov
- National Climactic Data Center
- Weather Underground
- IPEDS Data Center (data about US universities)
- IRS Tax Stats
- Lending Club loans
- Internet census
- Google Cluster measurements
- GitHub archive
- Million Song dataset
- StackExchange archive
- Enron emails
- Project Gutenberg texts
- Spam datasets
- Westbury Lab Usenet archive
- Google US Patent archive
- Google US Trademark archive
- Kaggle competitions
- Data.gov
- Etc.

THE CHALLENGE OF TEACHING BIG DATA

Rather, the challenge is providing the **infrastructure** for more-or-less **realistic** big data processing tasks.

This challenge is greater for smaller schools.



A VARIETY OF ALTERNATIVES

OPTION 1: LOCAL INSTALLATION

Pros:

- Students do not compete for computational resources
- Easy to debug (breakpoints, etc.)
- Very cheap (for the school)

Cons:

- Requires that each student has a performant, stable computer
- Hadoop for Windows was very difficult to build!
- Lots of free disk space is required for “big” jobs
- Very slow, little ability to exploit parallelism
- Less realistic (except when debugging)

OPTION 2: HIGH-PERFORMANCE PHYSICAL CLUSTER

Pros:

- Realistic scenario
- Very performant

Cons:

- Expensive if building from scratch
- Possibly difficult to manage

Prior work:

- Ngo, et al. (Clemson University) 2014

OPTION 3: LOW-POWER PHYSICAL CLUSTER

Pros:

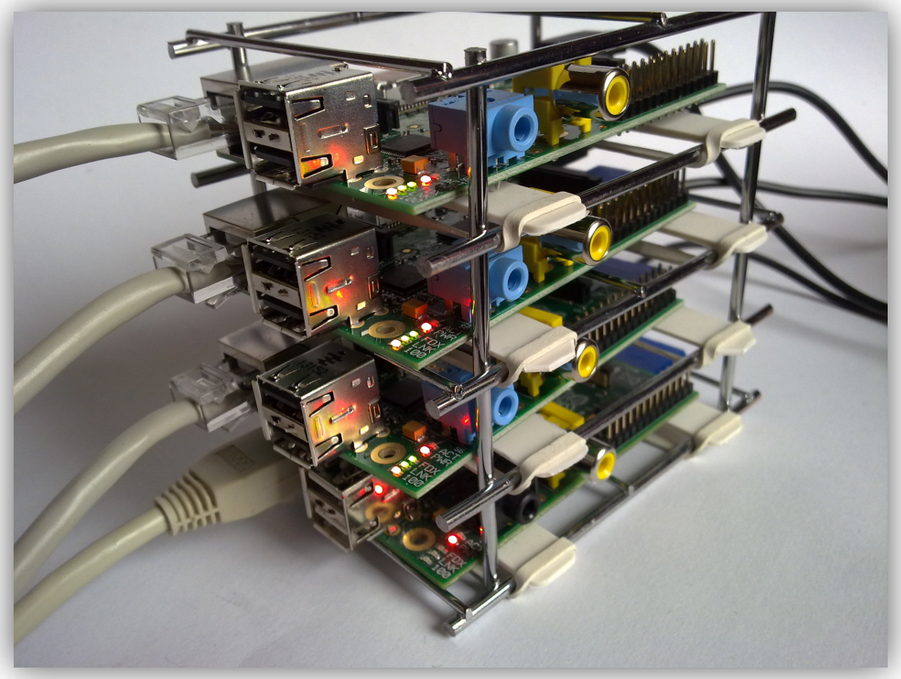
- Students can help design and build the cluster
- Relatively inexpensive per “core”

Cons:

- Less performant than a single server (Cox, et al.)
- Somewhat fiddly

Prior work:

- Cox, et al. (University of Southampton, UK) 2014



OPTION 4: CLOUD COMPUTING

Pros:

- Realistic scenario
- Can be very performant
- Many providers (Amazon, Google, Microsoft, etc.)
- Cheap (for the school) if students pay; grants may be available

Cons:

- Students may be asked to bear the cost, and they might not like that
- One mistake (e.g., infinite loop) could cost a lot

Prior work:

- Rabkin, et al. (UC Berkeley) 2012
- González-Martinez, et al. (survey paper) 2015

OPTION 5: VIRTUAL CLUSTER

Pros:

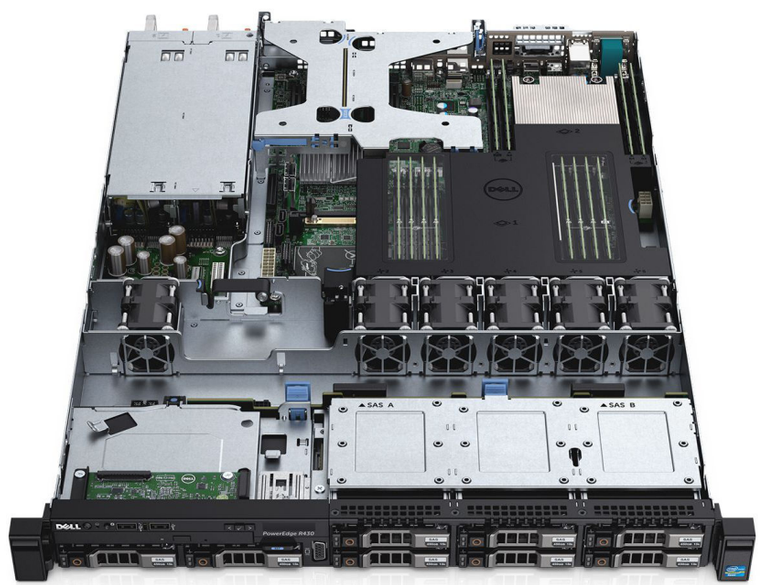
- Somewhat realistic; it's still a "cluster"
- Easy to set up and maintain and reconfigure
- Students can assist with design and maintenance
- Cheaper than a physical cluster

Cons:

- Less performant than physical cluster or cloud computing

Prior work:

- Johnson, et al. (St. Olaf College, MN) 2011
- Brown & Shoop (St. Olaf & Macalester Colleges, MN) 2013

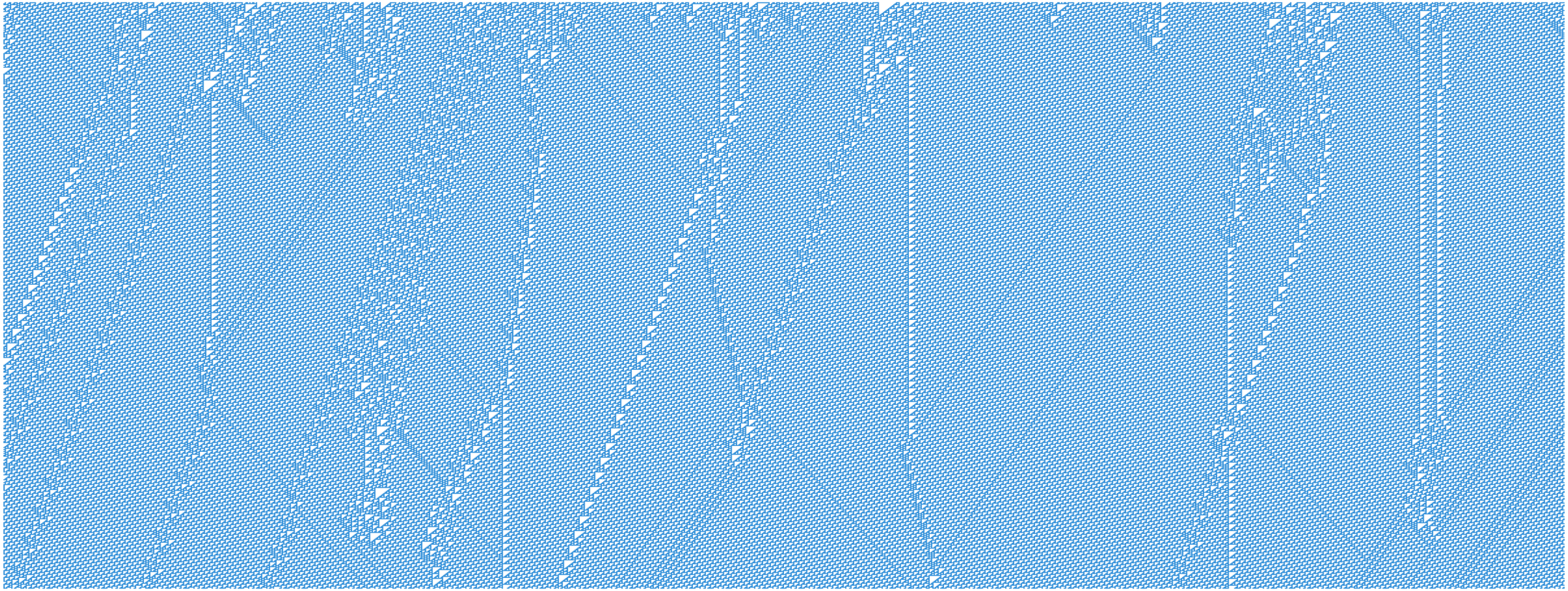


OUR CHOICE

We chose option 5: a virtual cluster. (In fact, we only needed to upgrade a machine.)

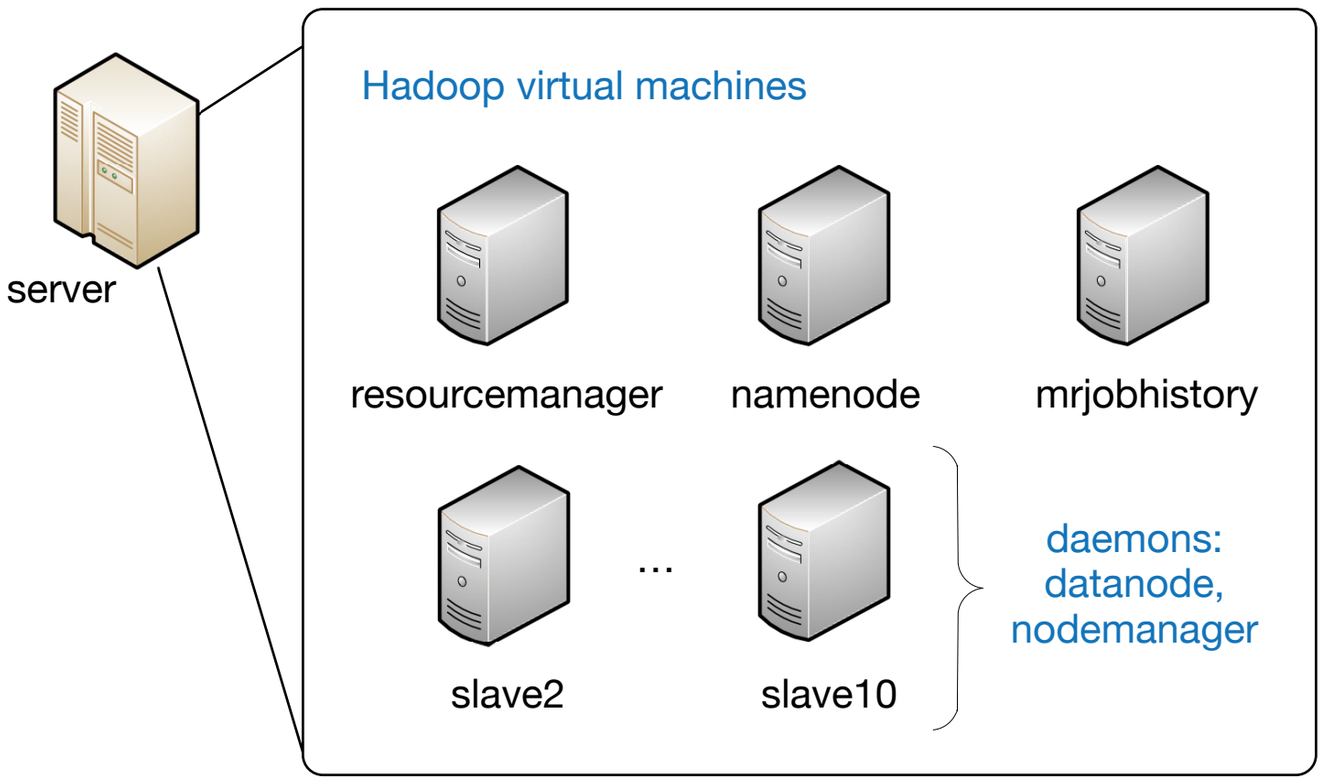
The virtual cluster has some advantages:

- We have some institutional knowledge about how it works.
- Students need not stress about costs.
- Students who are interested can look behind the curtain and learn how it was set up.



VIRTUAL CLUSTER DETAILS

PHYSICAL AND VIRTUAL MACHINES



Dell PowerEdge R430
Intel Xeon E5-2640
128 GB memory
Hard drives totaling 28 TB

CentOS 6.7 host OS
Libvirt + Qemu/KVM
Ubuntu 14.04 guest OS

DEPLOYMENT

A single command builds/resumes all the nodes:

```
vagrant up
```

Behind the scenes, Vagrant sets up the VMs and executes Ansible scripts to install the software, configure users, etc.

Another command configures the host's Apache server for proxies, and starts Hadoop on the VMs:

```
setup-hadoop.sh
```

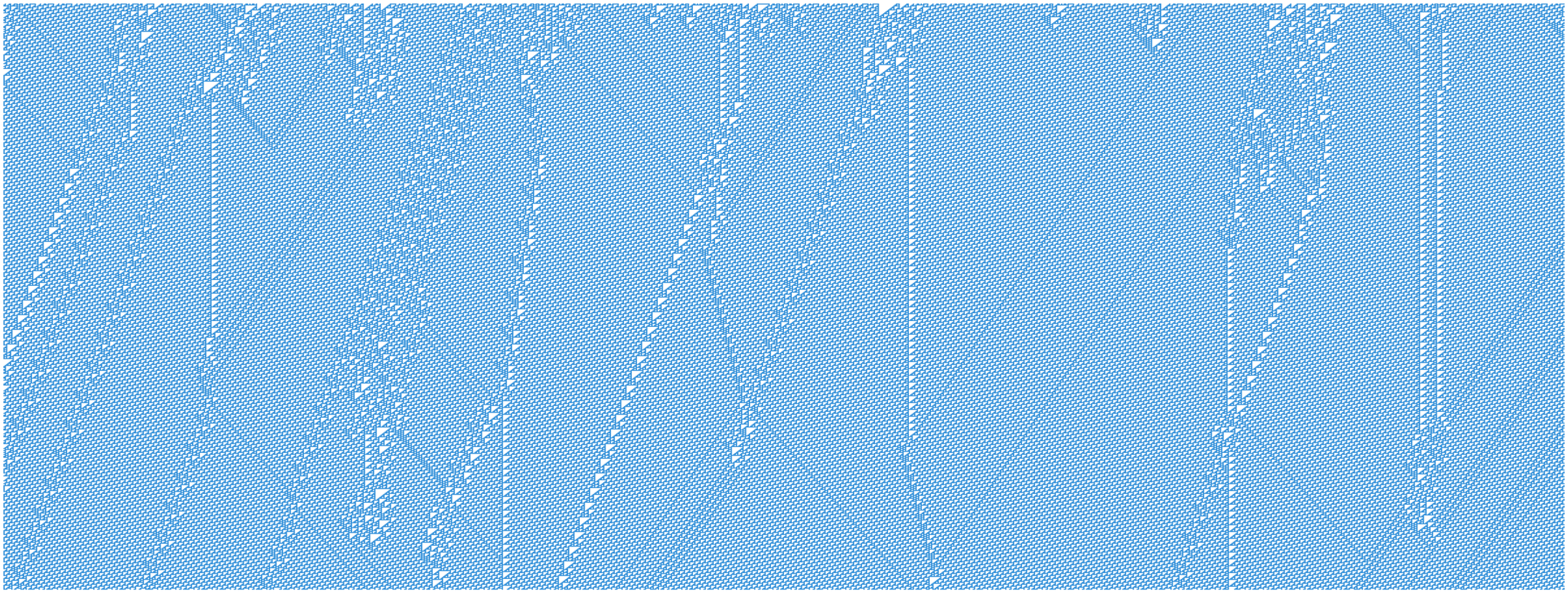
The code is freely available: https://github.com/StetsonMathCS/hadoopvirtualcluster

STUDENT EXPERIENCE

Students develop their code on their own machine, and (sometimes) test it with a local Hadoop installation. Then, they copy the JAR to the server and log in via SSH. They run this command as a normal user on the server:

```
hadoop jar mycode.jar MyClass /input /output
```

A web interface is available on the main server (which proxies into the VMs) to monitor active jobs, review historical jobs, display log files including crash backtraces, and download output files.



EXAMPLE PROJECTS AND STUDENT FEEDBACK

PROJECT 1: WORD COUNT

Typical Map-Reduce word count exercise with a large text file.

Learning objectives:

1. Acquire familiarity with the Hadoop platform and MapReduce design pattern.

PROJECT 2: BACKBLAZE HARD DRIVE FAILURES

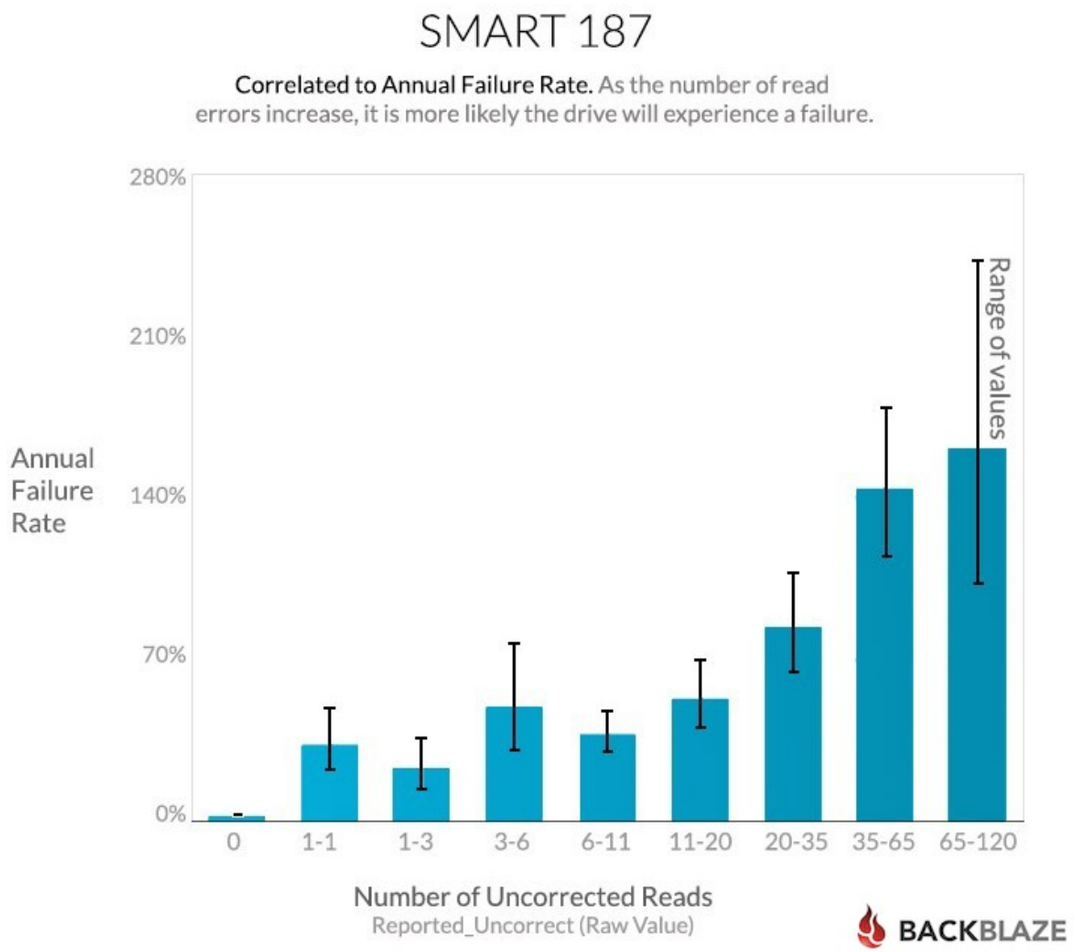
Analyze BackBlaze's hard drive monitoring data to determine if some metric could be used to predict hard drive failure.

Learning objectives:

1. Apply filtering to transform big data into small data.
2. Analyze (small) data in R.

The dataset was 4.2 GB in size.

https://www.backblaze.com/hard-drive-test-data.html



PROJECT 3: STACKEXCHANGE ARCHIVE

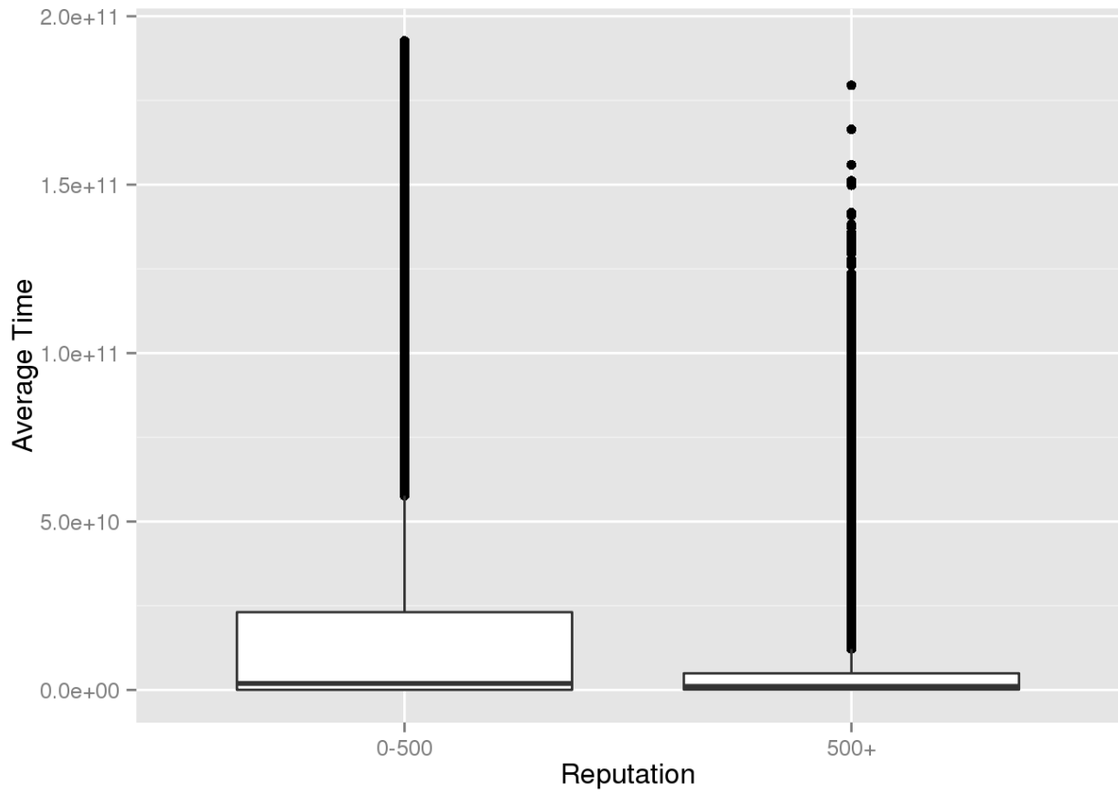
Analyze StackExchange's entire archive of questions and answers to determine if persons with high reputation typically answer questions faster.

Learning objectives:

1. Merge two big data sets (posts + users) via multiple MapReduce jobs (e.g., with Hive).

The dataset was about 116 GB in size.

https://archive.org/details/stackexchange



PROJECT 4: CAT CLUSTERS

"There are different breeds of cats represented in the dataset. Verify."

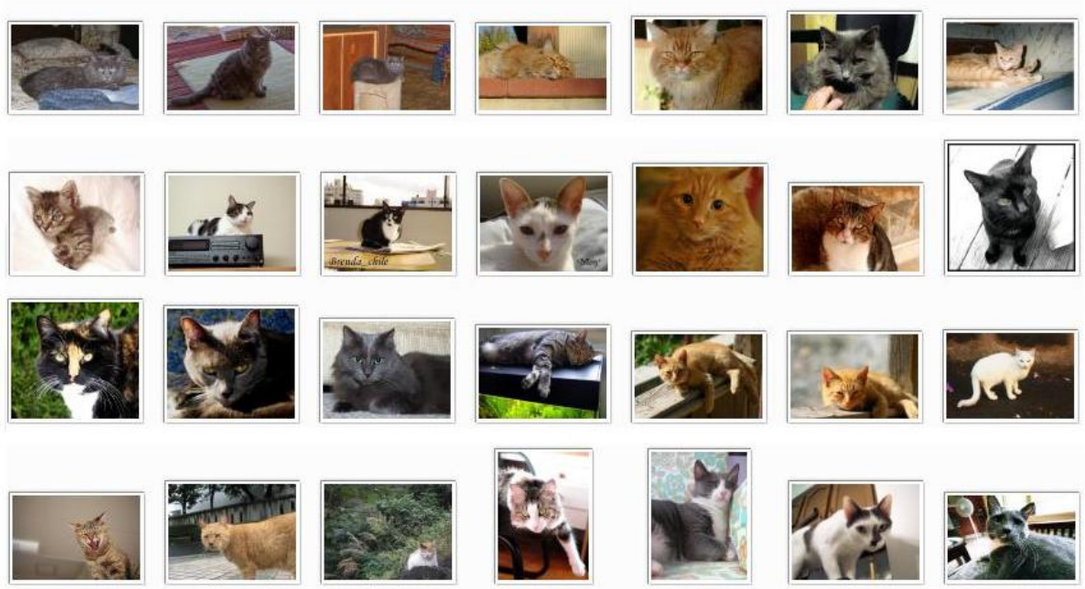
Learning objectives:

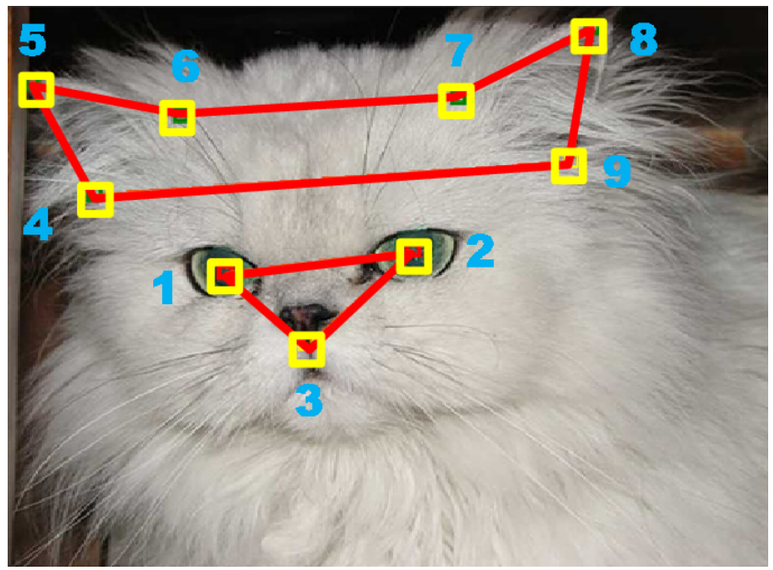
1. Apply big data processing to images rather than text.
2. Independently design a creative approach (such as k-means clustering on colors) and proper analysis to solve an under-specified problem.

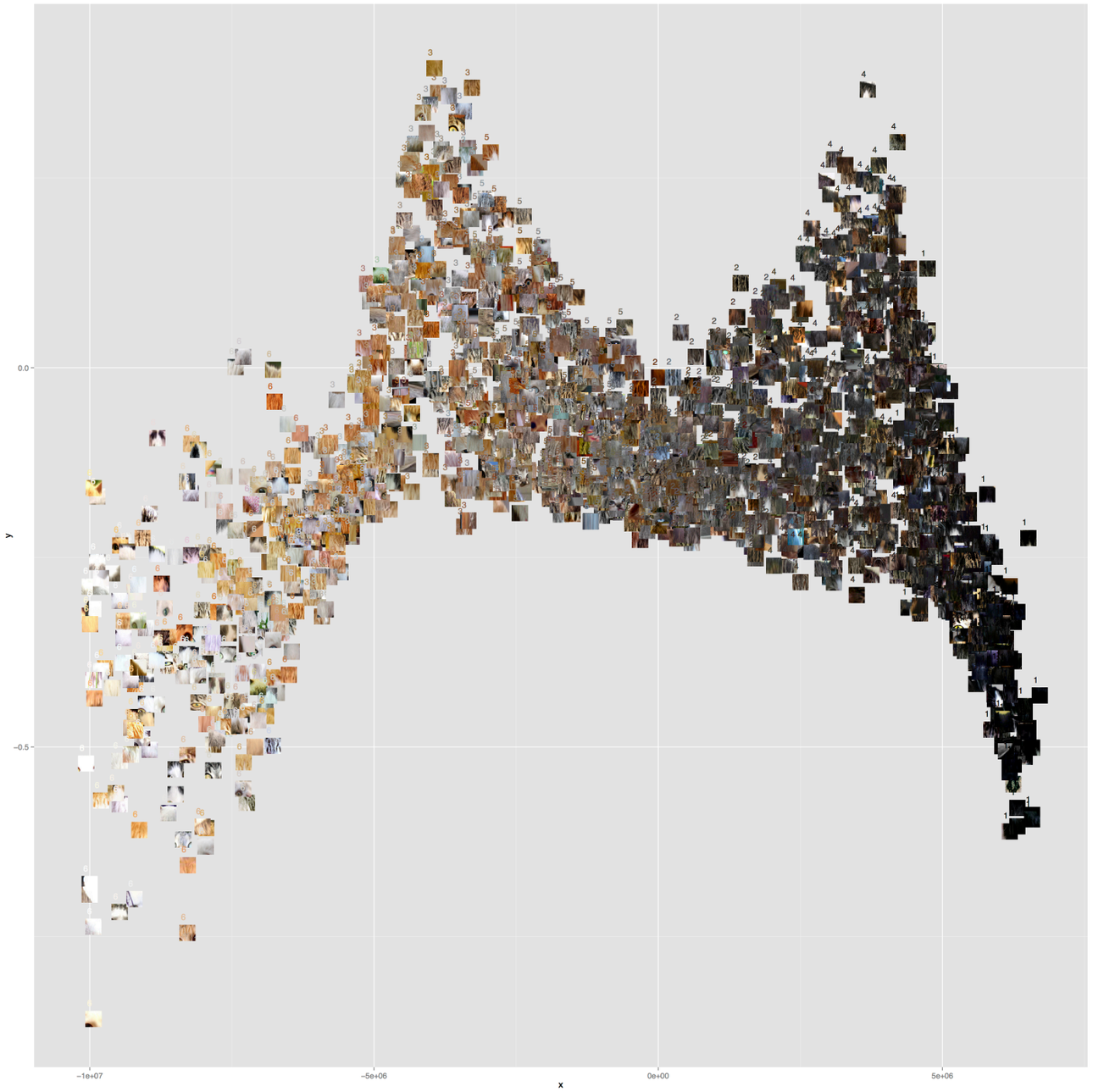
The dataset was 4.2 GB in size.

http://137.189.35.203/WebUI/CatDatabase/catData.html (now offline)

PROJECT 4: CAT CLUSTERS






0.0
−0.5
y
−1e+07
−5e+06
0e+00
5e+06
x

PROJECT 5: SPAM CLASSIFICATION

Build a spam classifier model for 75,419 email messages in the TREC 2007 dataset.

Learning objectives:

1. Apply text processing (e.g., bag-of-words) with a MapReduce design pattern.
2. Utilize Mahout for large-scale machine learning.
3. Use appropriate methodology (cross-validation) to evaluate models.

The dataset was about 0.7 GB in size.

http://plg.uwaterloo.ca/~gvcormac/treccorpus07/

STUDENT FEEDBACK

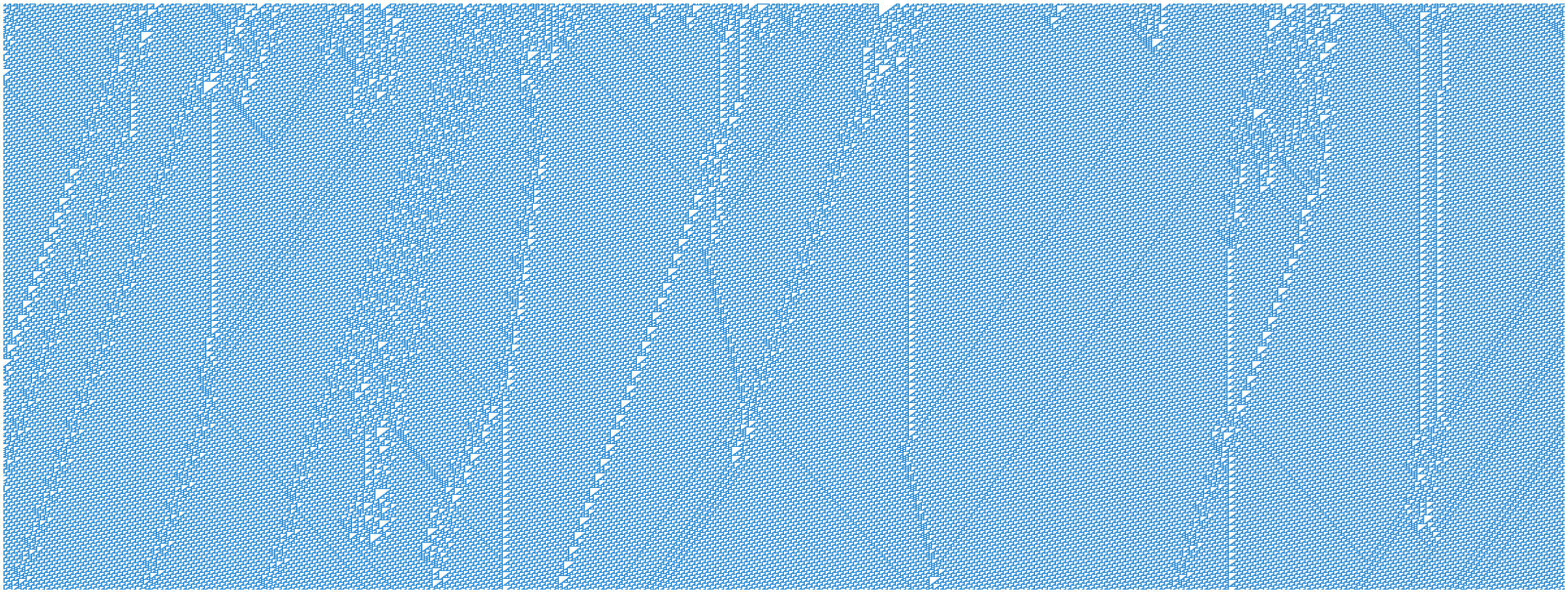
Question: “Which aspects of this course helped you learn the most?”

Answers:

- “The **practice with the various tools and software** used in the class.”
- “**Hands on experience** with current trends in the industry.”

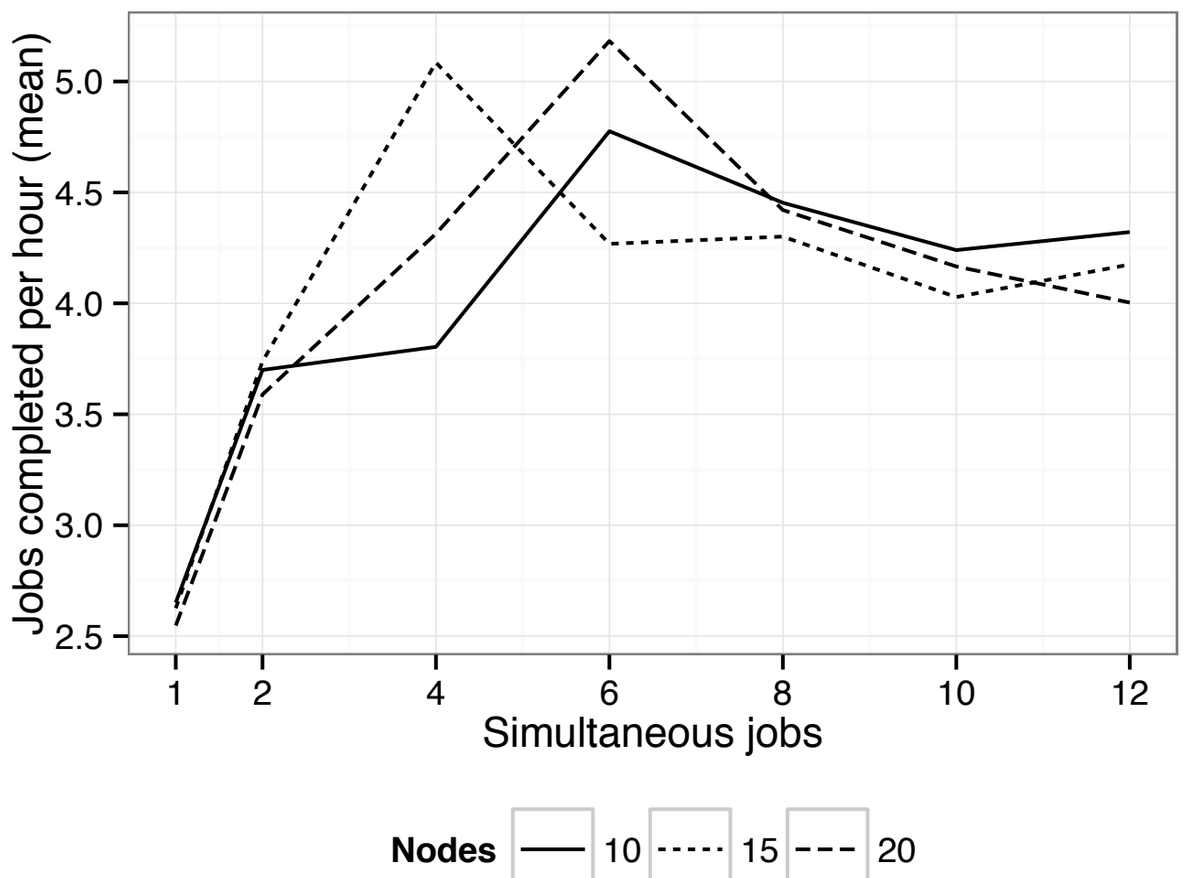
Another comment: “Add additional servers for student use.”

One student was hired as a Data Scientist at Pacific Northwest National Laboratory (PNNL). Another has an on-site interview with State Farm in Atlanta for a data scientist position.

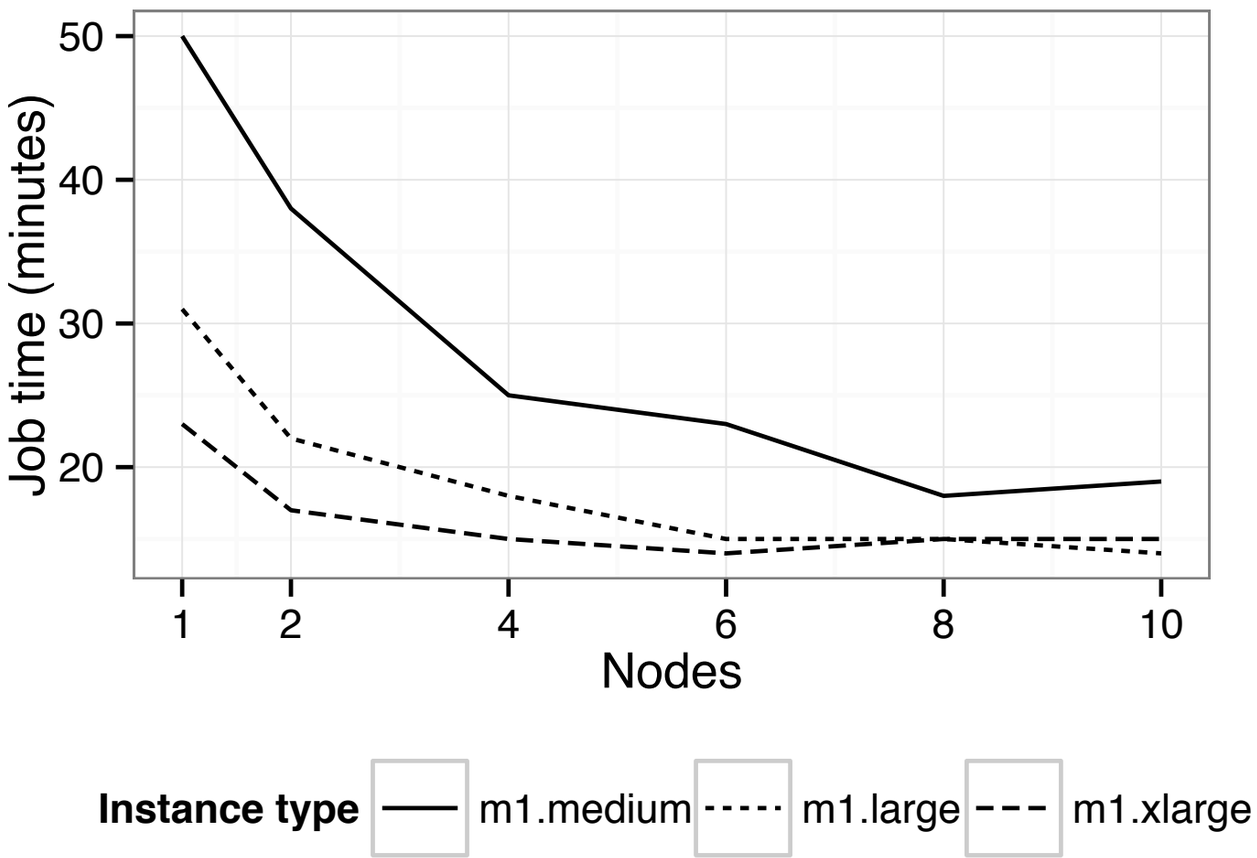


COST/BENEFIT OF VIRTUAL VS. CLOUD CLUSTERS

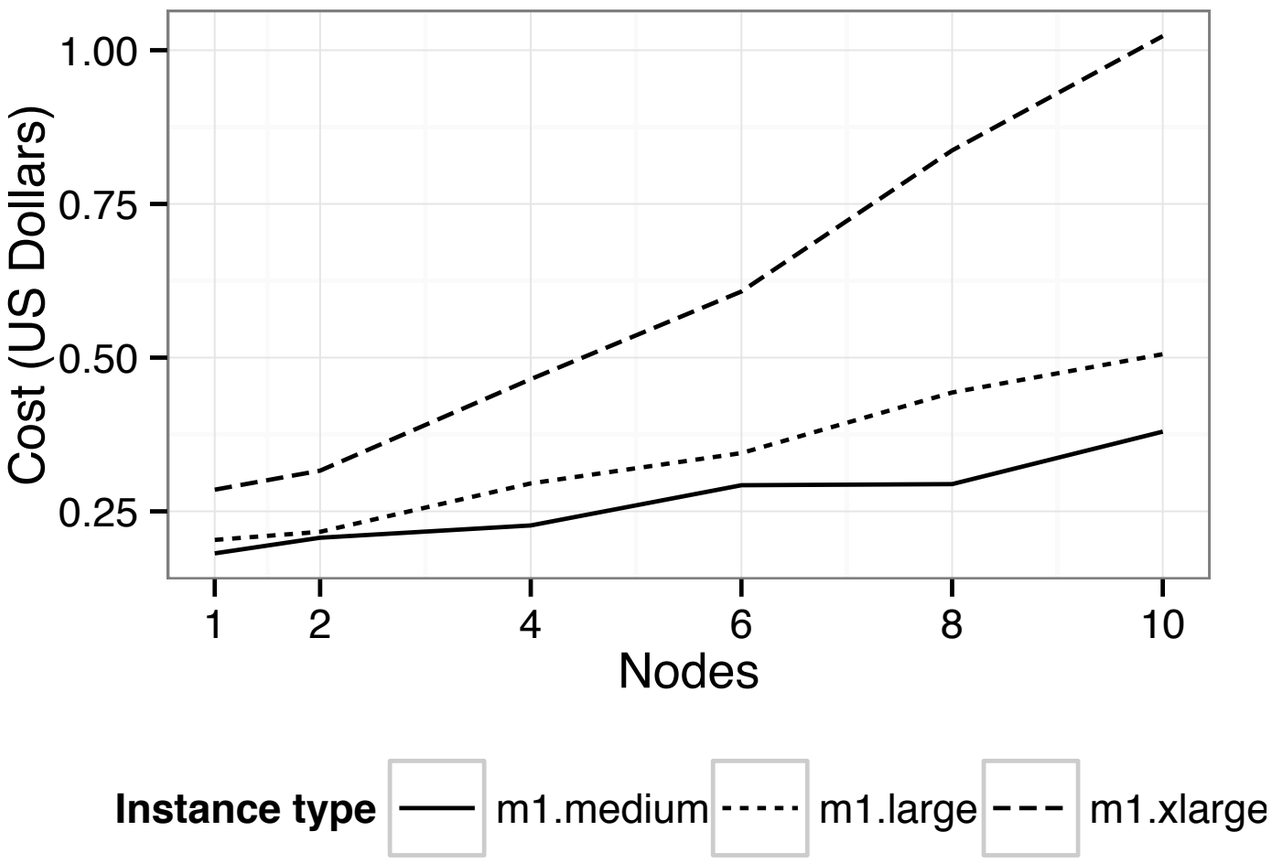
VIRTUAL CLUSTER: JOBS COMPLETED PER HOUR



AMAZON AWS: PROCESSING TIME PER JOB



AMAZON AWS: COST PER JOB



COMPARATIVE COSTS

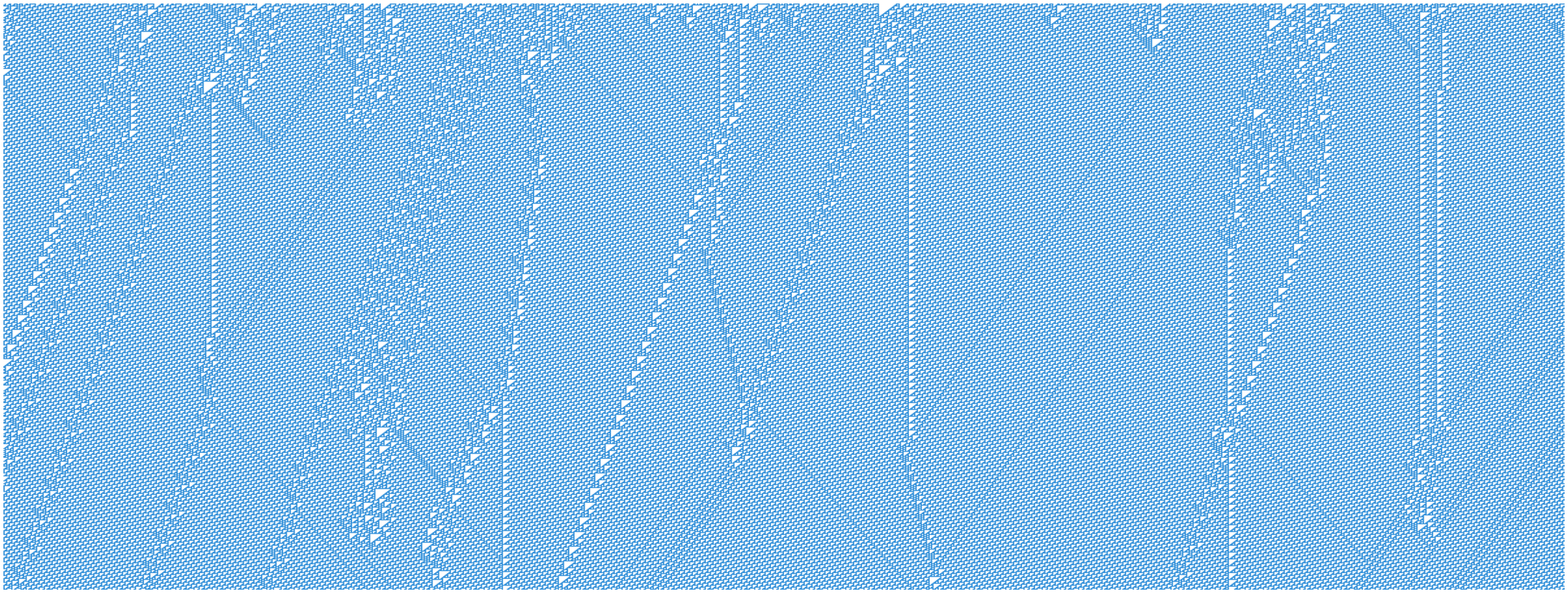
Estimated cost to build a new physical cluster with 10 nodes: **$20,192**

Estimated cost to build a new virtual cluster with 10 virtual nodes: **$7,400**

Assume six semesters before replacement, 20 students, five projects, five submissions per project, 3000 jobs overall.

Running the same job (37GB word count) on different platforms, we get costs per job shown in the table.

Platform (10 nodes)	Job time	Job cost
Virtual cluster	334 min	$2.82
Physical cluster	-	$7.27
Amazon EMR, m1.medium	101 min	$2.02
Amazon EMR, m1.large	56 min	$2.02
Amazon EMR, m1.xlarge	34 min	$2.32



CONCLUSION

CONCLUSION

Our virtual cluster served the needs of about 20 students.

- Performance was severely hindered by having only one channel for disk I/O.

Not worrying about costs was a great feeling.

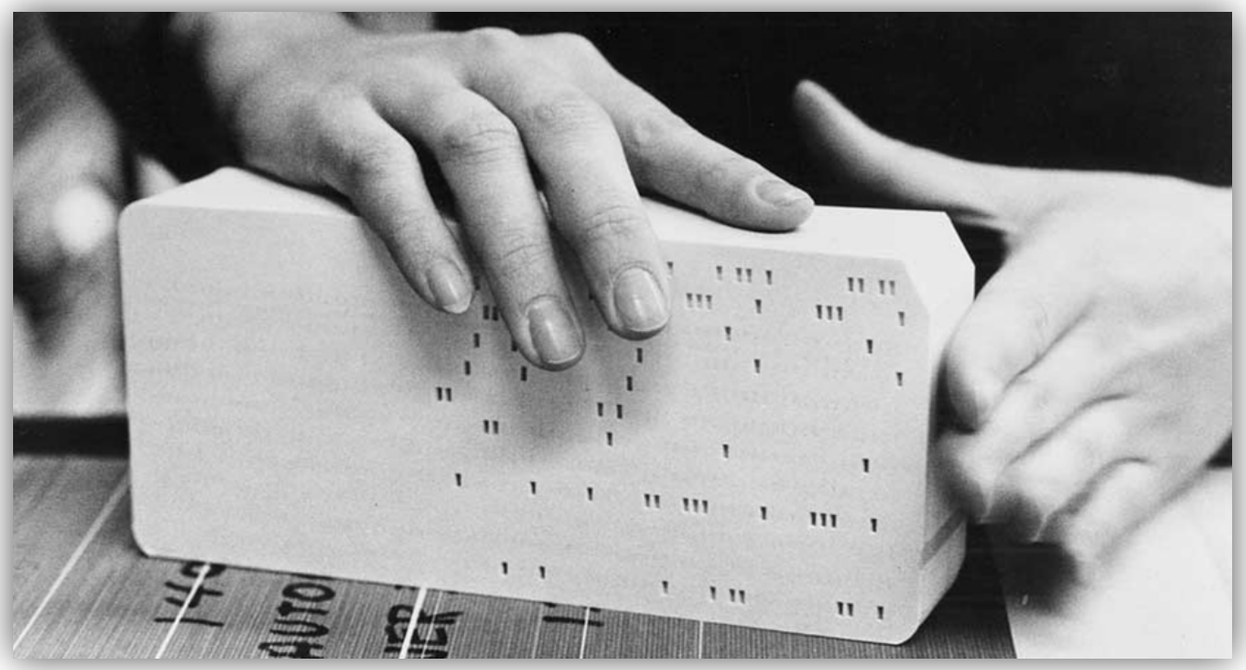
- …from both perspectives: student, faculty/department/school.

But cloud computing is very attractive:

- Much faster job completion
- Much larger jobs
- It's cheaper, if you actually plan to upgrade your cluster
- Very realistic

THE CLOUD COMPUTING FUTURE?

Unanswered question: How do students behave when they are asked to pay for each job submission? Do they avoid running their code until the very end?



REFERENCES

Brown, R. & Shoop, E.
Teaching undergraduates using local virtual clusters
IEEE International Conference on Cluster Computing (CLUSTER), **2013,** 1-8

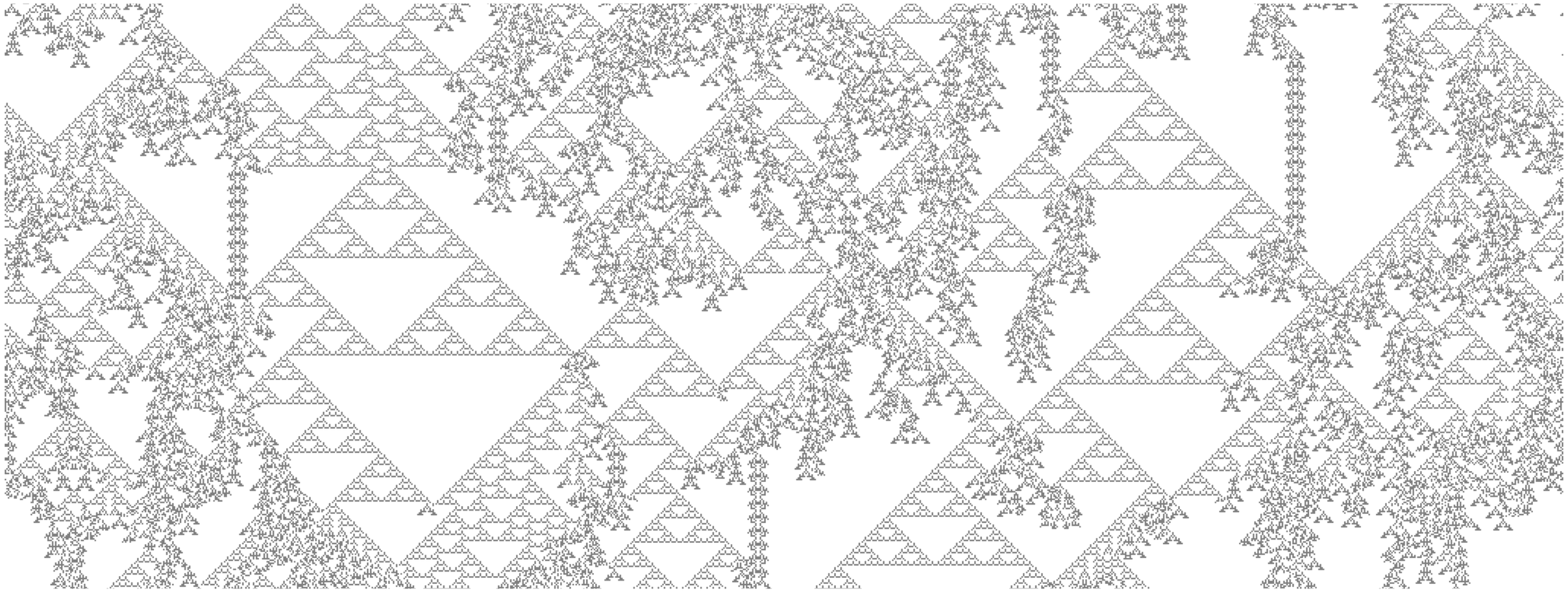
Cox, S. J.; Cox, J. T.; Boardman, R. P.; Johnston, S. J.; Scott, M. & O'Brien, N. S.
Iridis-pi: a low-cost, compact demonstration cluster
Cluster Computing, Springer, **2014**, *17*, 349-358

González-Martinez, J. A.; Bote-Lorenzo, M. L.; Gómez-Sánchez, E. & Cano-Parra, R.
Cloud computing and education: A state-of-the-art survey
Computers & Education, Elsevier, **2015**, *80*, 132-151

Johnson, E.; Garrity, P.; Yates, T.; Brown, R., et al.
Performance of a Virtual Cluster in a General-Purpose Teaching Laboratory
IEEE International Conference on Cluster Computing (CLUSTER), **2011,** 600-604

Ngo, L. B.; Duffy, E. B. & Apon, A. W.
Teaching HDFS/MapReduce Systems Concepts to Undergraduates
Parallel & Distributed Processing Symposium Workshops (IPDPSW), 2014 IEEE International, **2014,** 1114-1121

Rabkin, A. S.; Reiss, C.; Katz, R. & Patterson, D.
Experiences teaching MapReduce in the cloud
Proceedings of the 43rd ACM Technical Symposium on Computer Science Education, **2012,** 601-606



QUESTIONS?

OTHER ASSIGNMENTS

Ngo, et al.:

- Analyze 171GB of Google Data Center system logs to find job with largest number of task resubmissions.
  - http://googleresearch.blogspot.com/2011/11/more-google-cluster-data.html
- Implement descriptive statistics of moving rating database.
  - http://grouplens.org/datasets/movielens/
  - 22mil ratings