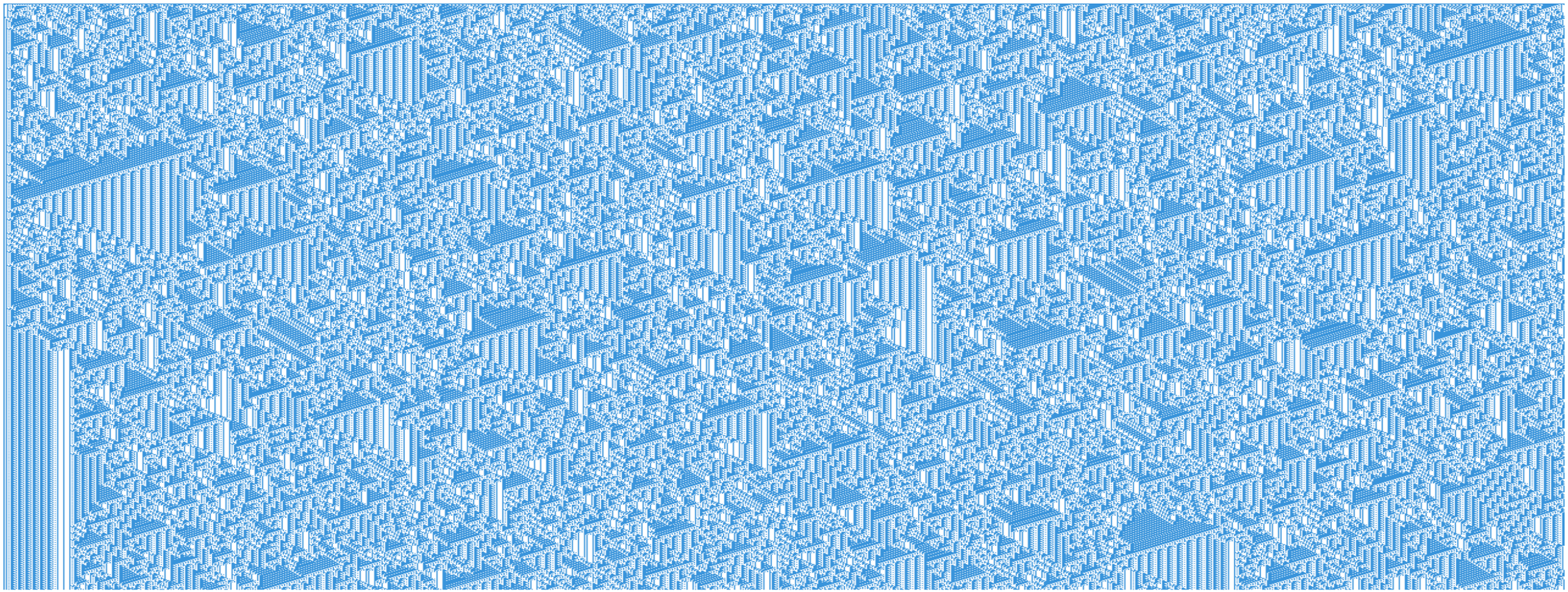


TEACHING FUTURE BIG DATA ANALYSTS

Joshua Eckroth
Stetson University
EduPar-17

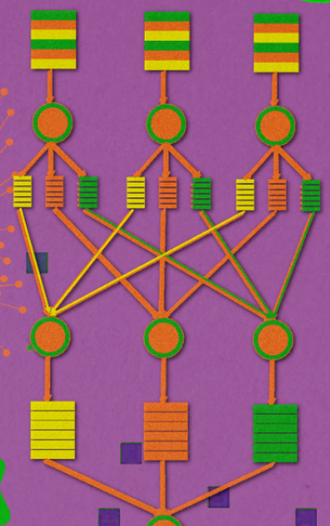
CINF401

Big Data Mining and Analytics

with Dr. Eckroth

Spring 2015

MW 12-1:15
R 11:30-12:45



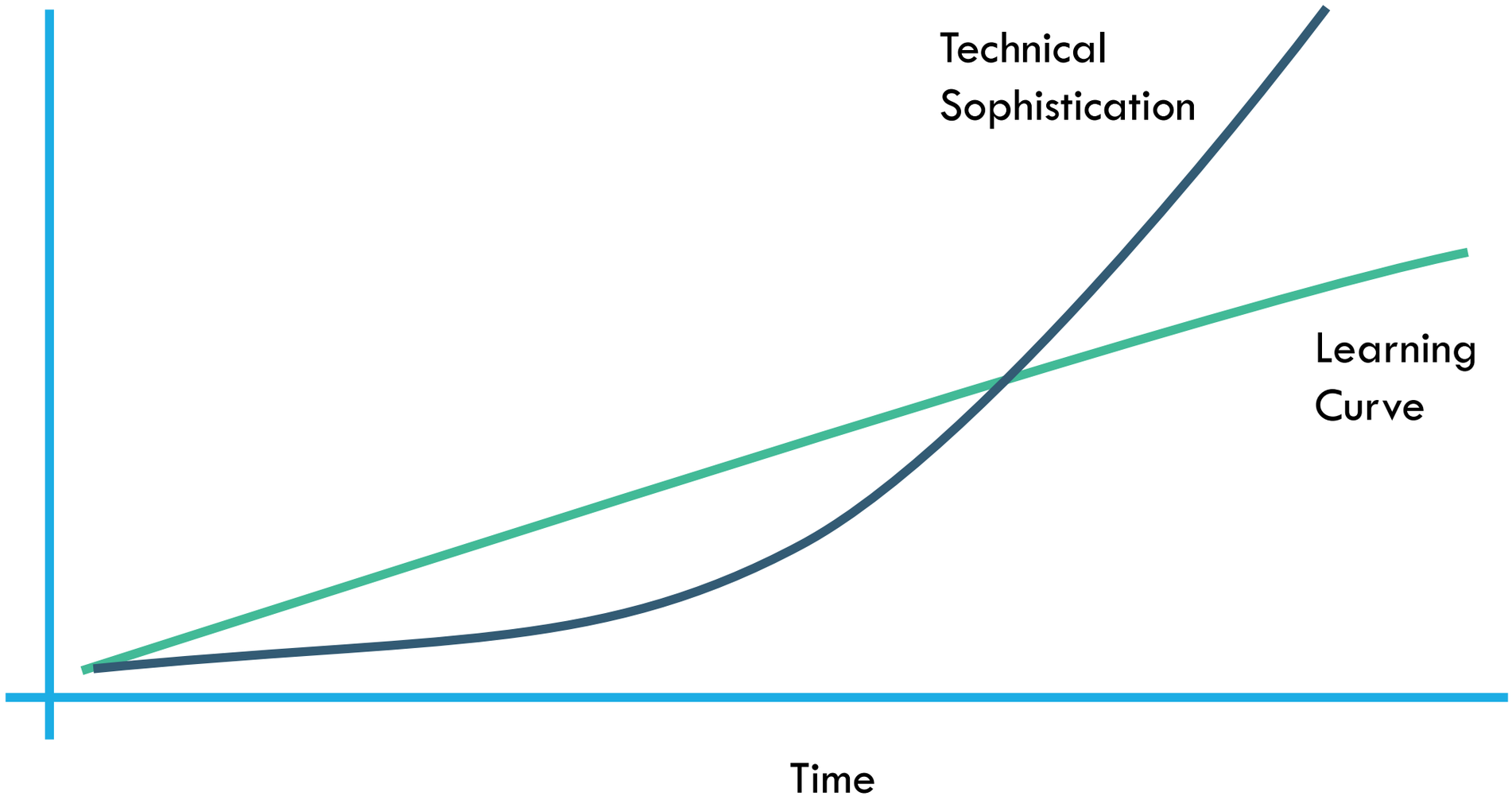
http://cinf401.artifice.cc

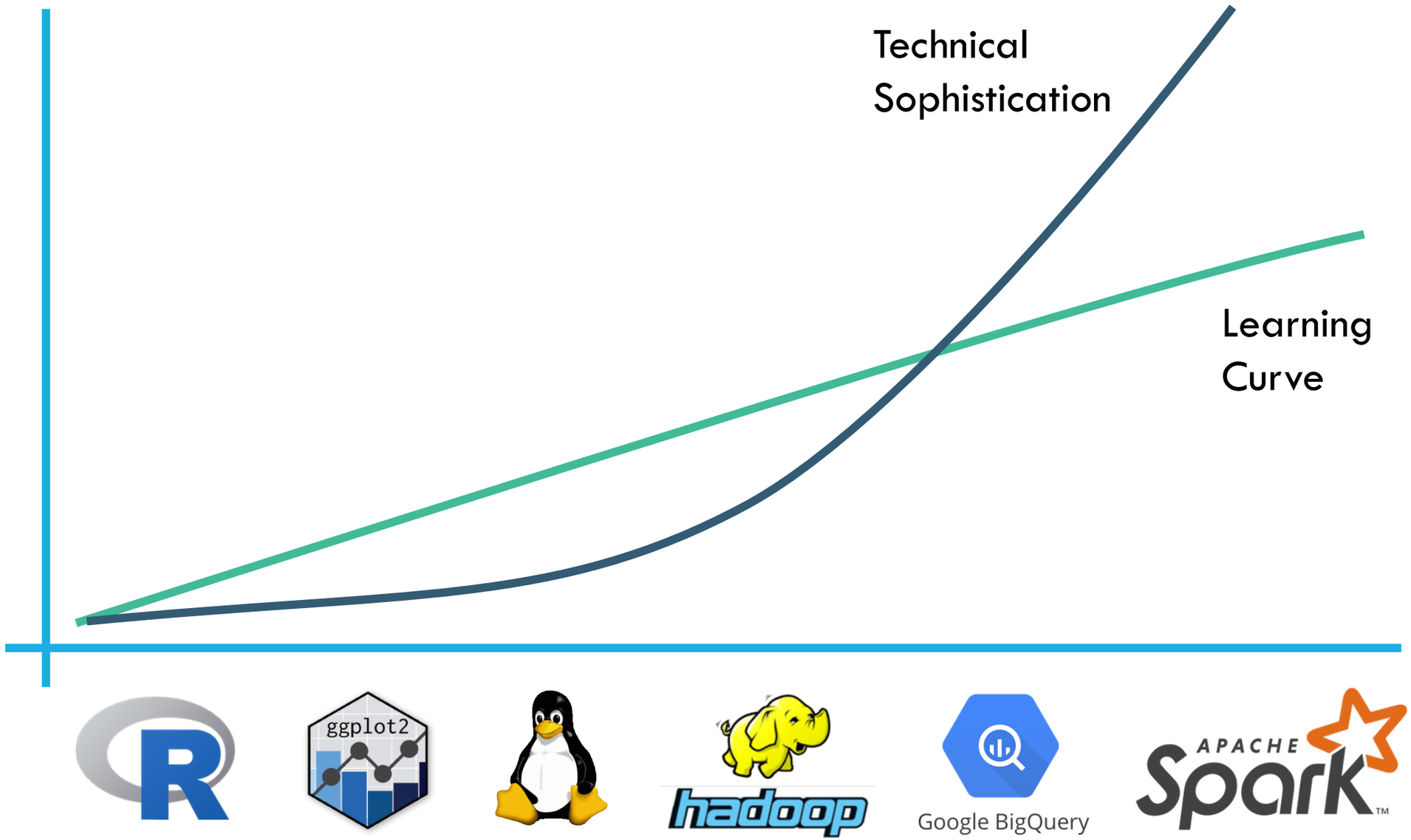
LEARNING OBJECTIVES

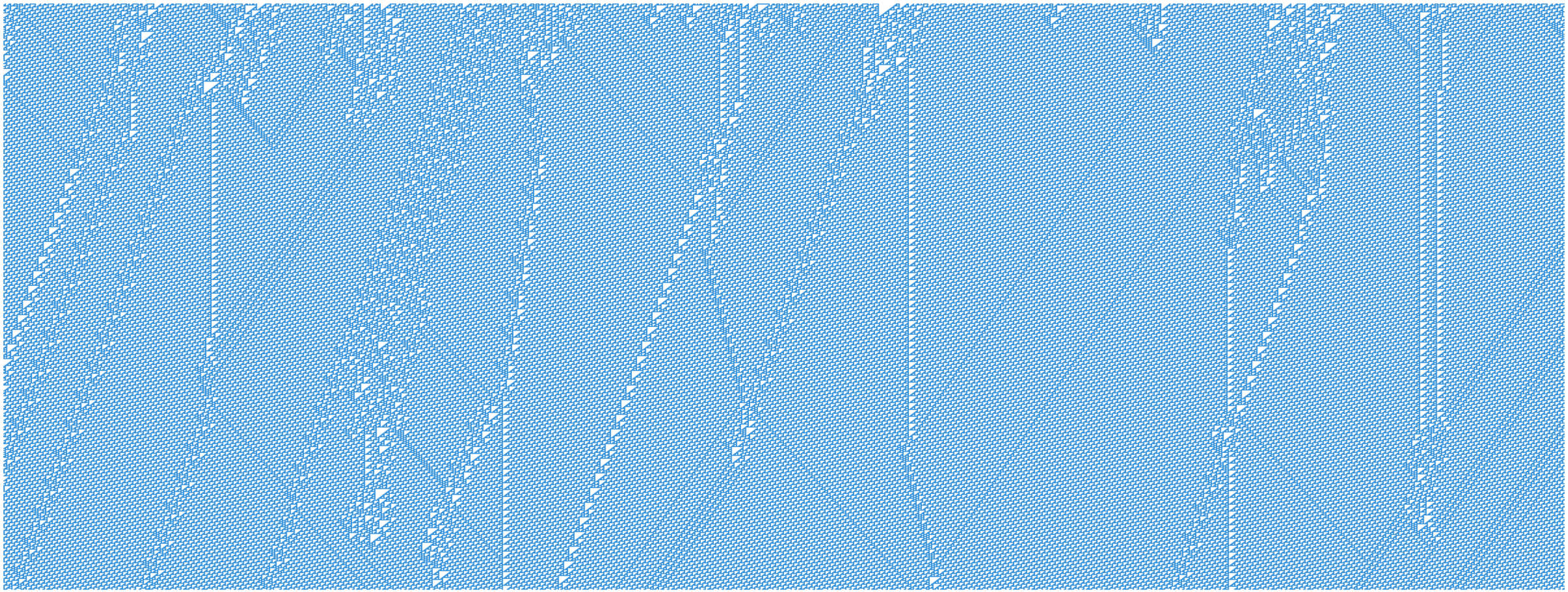
Determine whether a data analysis task requires “big data” tools and techniques, or can be done with traditional methods, and identify appropriate tools to perform the analysis.

Skillfully make use of tools to perform data processing and analysis.

Communicate the outcomes of data analysis with convincing arguments involving, as appropriate, text, tables, and plots.


Technical
Sophistication
Learning
Curve
Time


Technical
Sophistication
Learning
Curve
ggplot2
hadoop
Google BigQuery
APACHE
Spark



PROJECTS

PROJECT 1: BACKBLAZE HARD DRIVE FAILURES

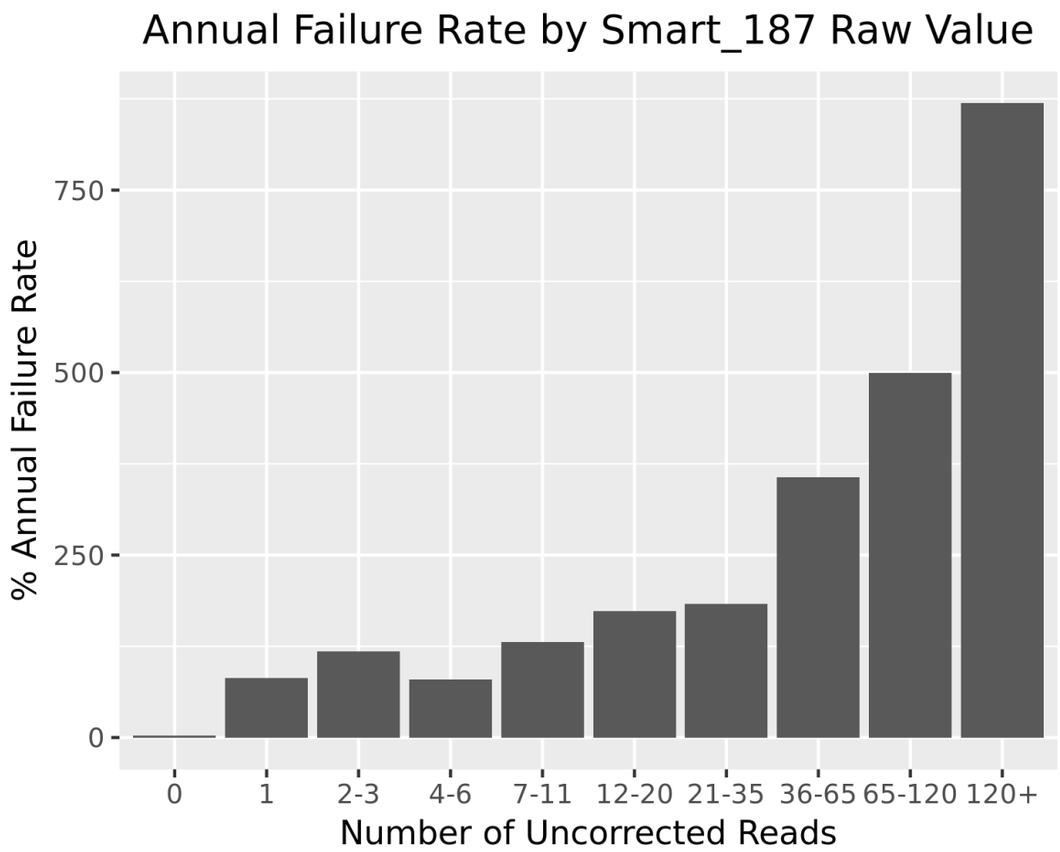
Analyze BackBlaze's hard drive monitoring data to determine if some metric could be used to predict hard drive failure.

Learning objectives:

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

The dataset was 3 GB in size.

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



PROJECT 1: BACKBLAZE HARD DRIVE FAILURES

Résumé note:

Hard Drive Failure Analysis & Visualization
(R, CSVKit, ggplot)

I analyzed 60 millions records of daily hard drive performance statistics provided by Backblaze to discover daily and yearly total storage capacity, yearly failure rate per manufacturer, and a strong positive statistical correlation between the SMART 187 hard drive metric and imminent drive failure. I preprocessed the data with CSVKit to extract a relevant subset, then analyzed and visualized these findings using R and ggplot.

PROJECT 2: STACKEXCHANGE ARCHIVE

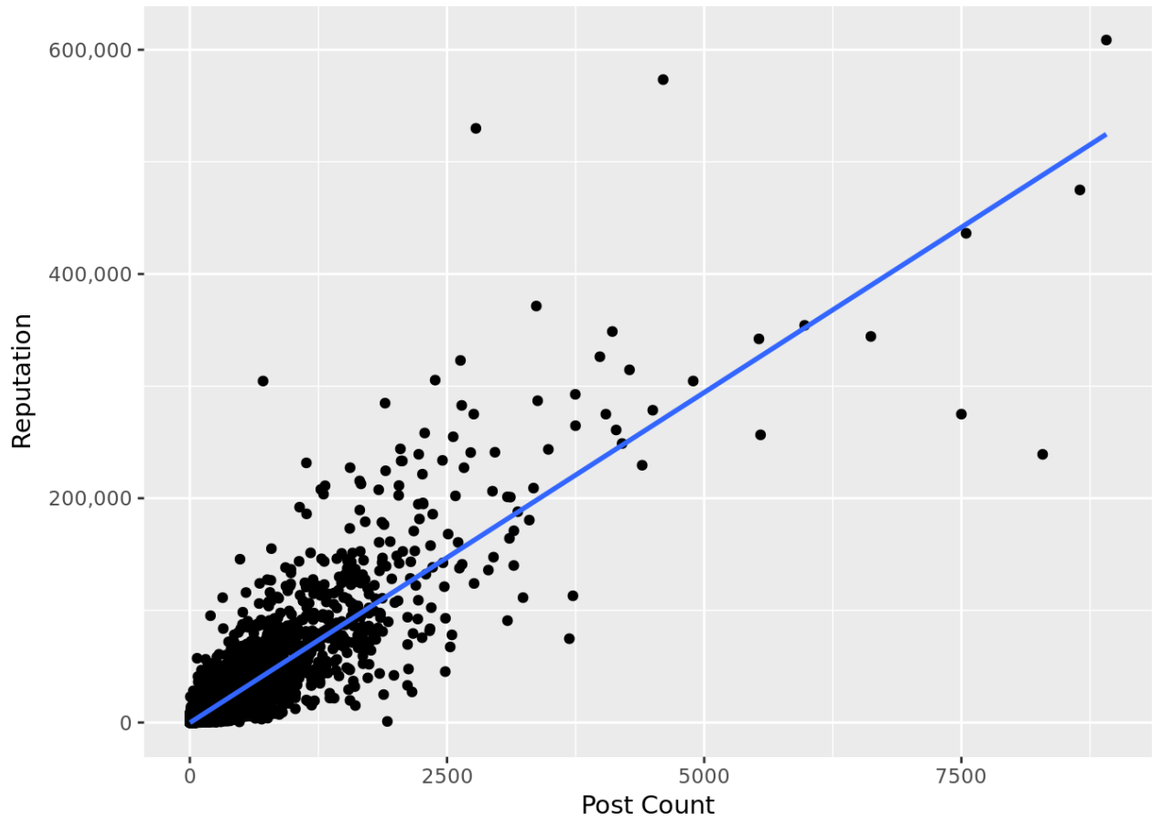
Analyze StackExchange's entire archive of questions and answers to determine if users with high reputation typically answer more questions, among other summary analyses.

Learning objectives:

1. Merge two big data sets (posts + users) via multiple MapReduce jobs.

The dataset was about 116 GB in size.

https://archive.org/details/stackexchange



PROJECT 2: STACKEXCHANGE ARCHIVE

Résumé note:

Social Q&A Site Analysis & Visualization
(Java, MapReduce, R, ggplot)

I developed a parallel and distributed data processing pipeline using Apache Hadoop and the MapReduce architecture to study various aspects of all StackExchange sites, including StackOverflow. The data consisted of 6.5 million users and 27.5 million posts stored across hundreds of XML files in an HDFS cluster. I found that most users are between 20 and 30 years old, users with more posts have higher reputation, and the most common tag on StackOverflow is “Java,” among other findings.

PROJECT 3: NYC YELLOW TAXI DATA

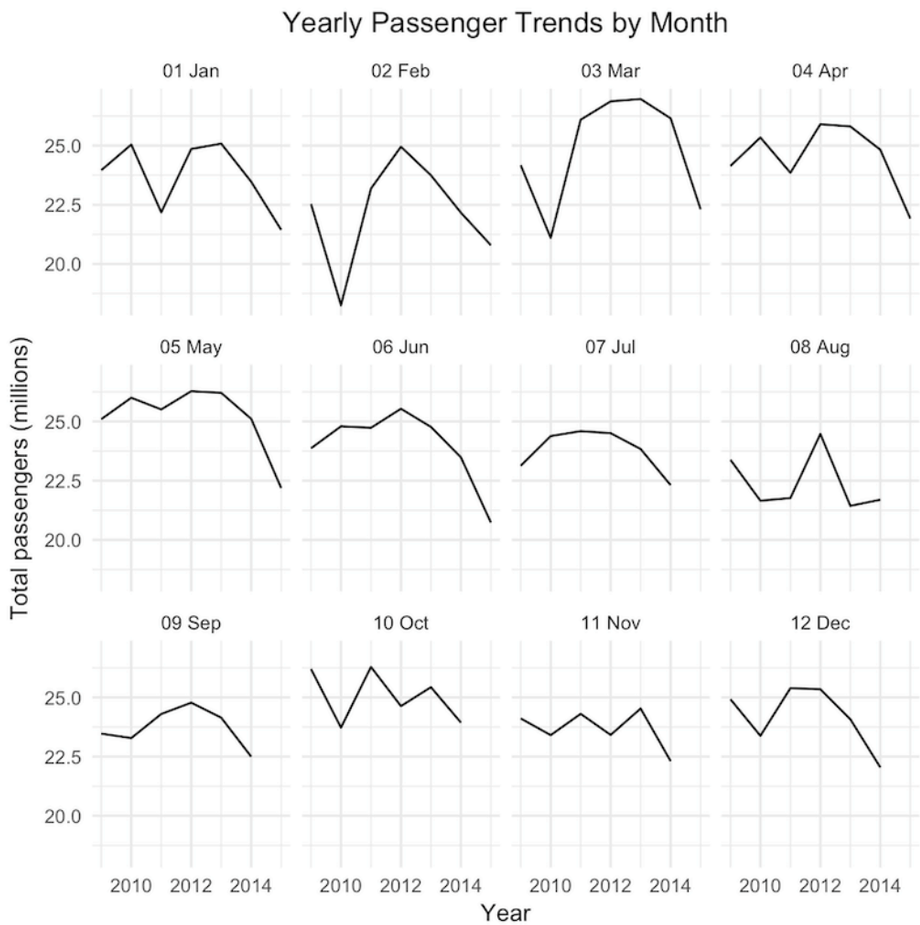
Summarize and visualize various aspects of NYC's public Yellow Taxi dataset.

Learning objectives:

1. Utilize Google BigQuery to quickly transform big data into small data.
2. Use ggplot to produce sophisticated and elegant plots.

The dataset was about 200 GB in size.

https://cloud.google.com/bigquery/public-data/nyc-tlc-trips



PROJECT 3: NYC YELLOW TAXI DATA

Summarize and visualize various aspects of NYC's public Yellow Taxi dataset.

Learning objectives:

1. Utilize Google BigQuery to quickly transform big data into small data.
2. Use ggplot to produce sophisticated and elegant plots.

The dataset was about 200 GB in size.

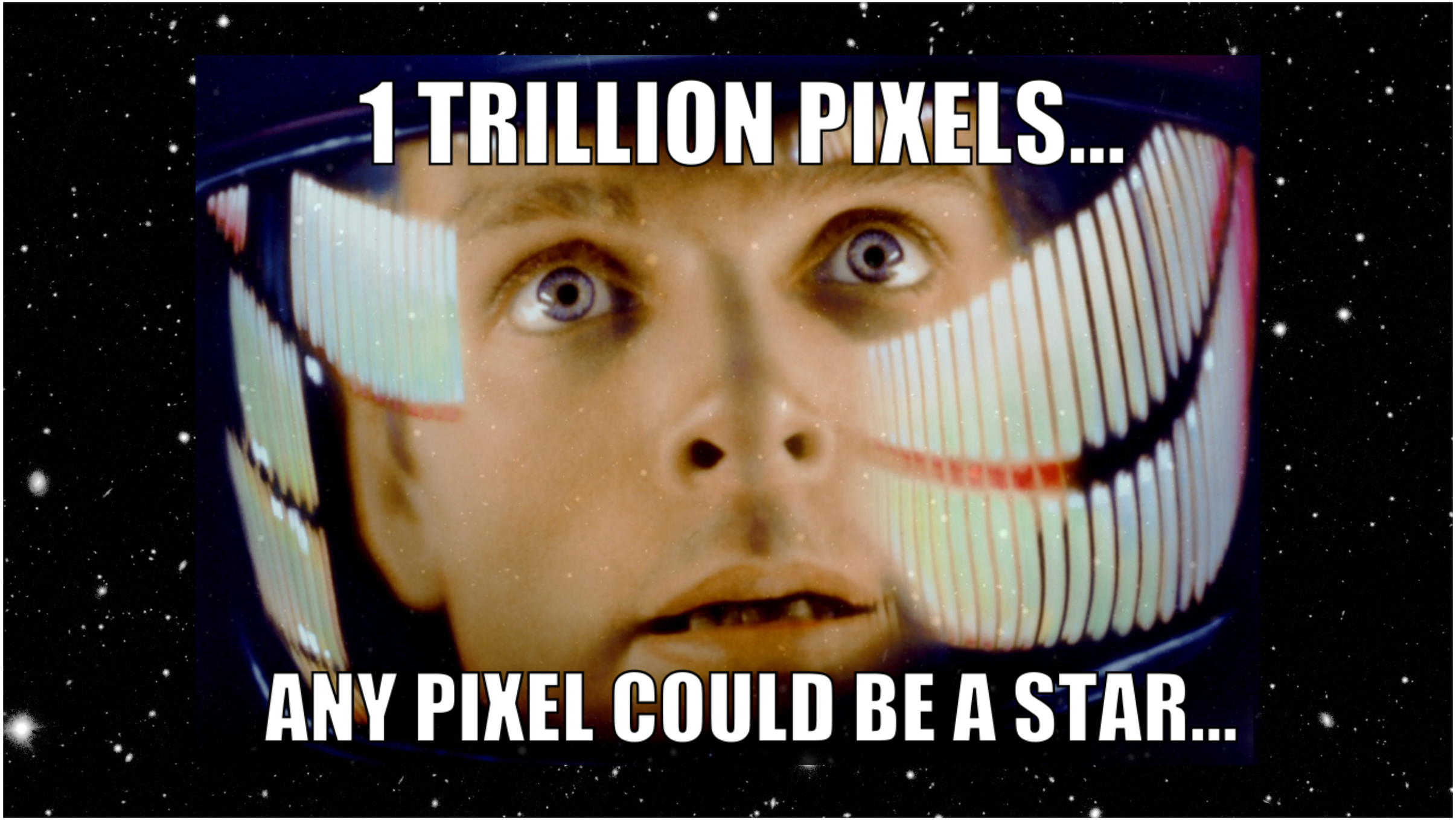
https://cloud.google.com/bigquery/public-data/nyc-tlc-trips

PROJECT 3: NYC YELLOW TAXI DATA

Résumé note:

NYC Taxi Usage and Revenue Analysis & Visualization
(Google Cloud Platform, SQL, R, ggplot)

I performed a deep analysis on more than a billion rows of NYC Yellow Taxi data to visualize every taxi ride for morning and evening commutes drawn on a map of Midtown Manhattan. The most frequent origin and destination points were highlighted on the map and each trip was drawn as a line connecting the starting and ending points. I also produced plots representing the frequency of trips for each hour of each day of the week, average distance traveled each month of each year, and the overall revenue trend over the last decade.


1 TRILLION PIXELS...
ANY PIXEL COULD BE A STAR...

PROJECT 4: STAR FINDER

Find the 100 largest stars according to pixel area and report their location in RA/Dec coordinates.

Learning objectives:

1. Utilize Spark for parallel processing and OpenCV for image processing.
2. Find an efficient trade-off in processing parallelism (thread count) and CPU/GPU usage.

The dataset was about 535 GB in size.

https://confluence.stsci.edu/display/PANSTARRS

PROJECT 4: STAR FINDER

Résumé note:

Star Identification from Wide-Field Astronomical Imaging
(Apache Spark, OpenCV, Python)

I built a parallel processing pipeline using Apache Spark to analyze more than 15,000 images totaling more than one trillion pixels provided by the Panoramic Survey Telescope and Rapid Response System. Using Python and OpenCV, I wrote an algorithm that eliminated noise from each image and detected each star. Ultimately, I discovered more than 10 million stars in the images and reported the “Right Ascension / Declination” coordinates for the top 100 brightest stars in the survey.

PROJECT 5: UNIQUE PER GROUP

Define your own data analysis problem.

During our “final exam time,” you and your partner are required to present your findings. Your grade will depend on the quality of your report and presentation and the appropriateness and insight in your analysis. As a final requirement, you must complete a “decision matrix” that shows how you decided to use certain tools for your analysis. You will explain your decision matrix during your presentation.

This project is 25% of your grade.

PROJECT 5: UNIQUE PER GROUP

Group 1: “Do highly controversial subreddits experience more growth than less controversial subreddits? And are certain words in titles correlated with controversy?”

Group 2: “Do Hacker News posts about programming languages correlate with an increase in questions related to that programming language on Stack Overflow?”

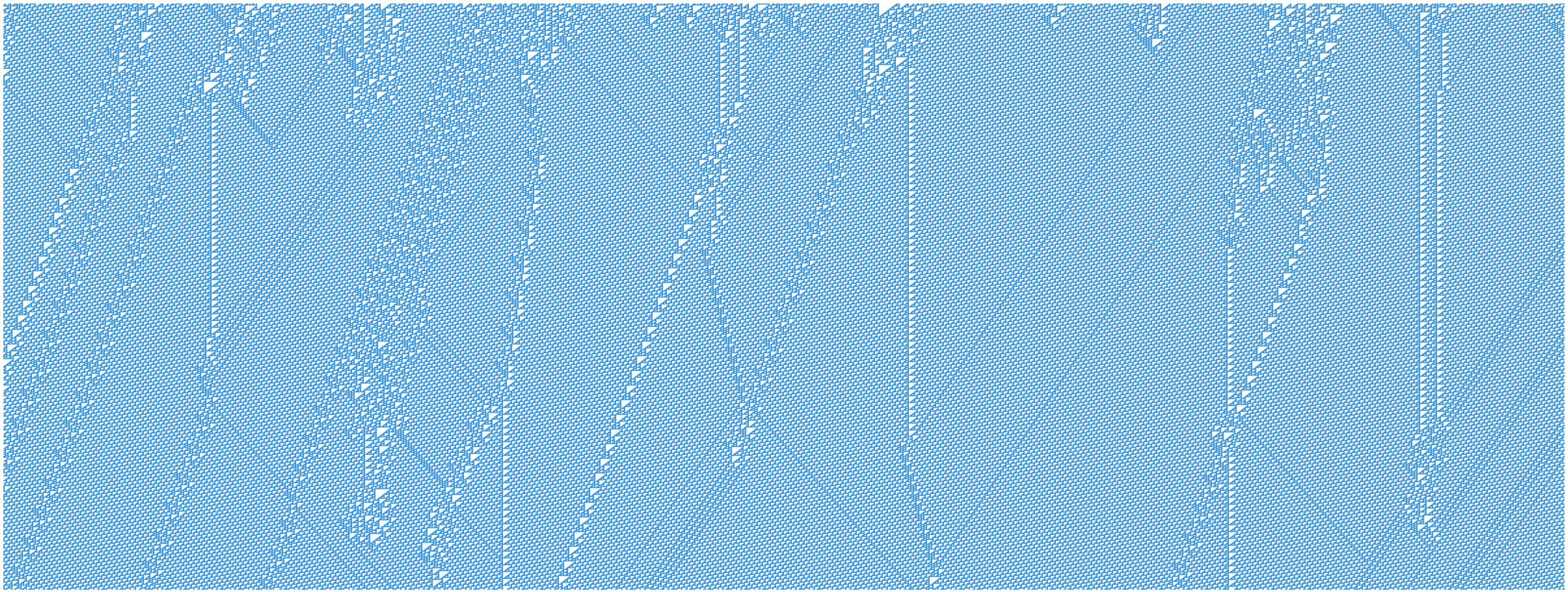
Group 3: “Generate a 2016 ‘annual report’ for the San Francisco Bike Share program.”

PROJECT 5: UNIQUE PER GROUP

Group 4: “Can we group Reddit users by certain attributes such as comment frequency, comment length, subreddit subscriptions, etc.?”

Group 5: “Can we predict Bitcoin price based on sentiment in Twitter messages?”

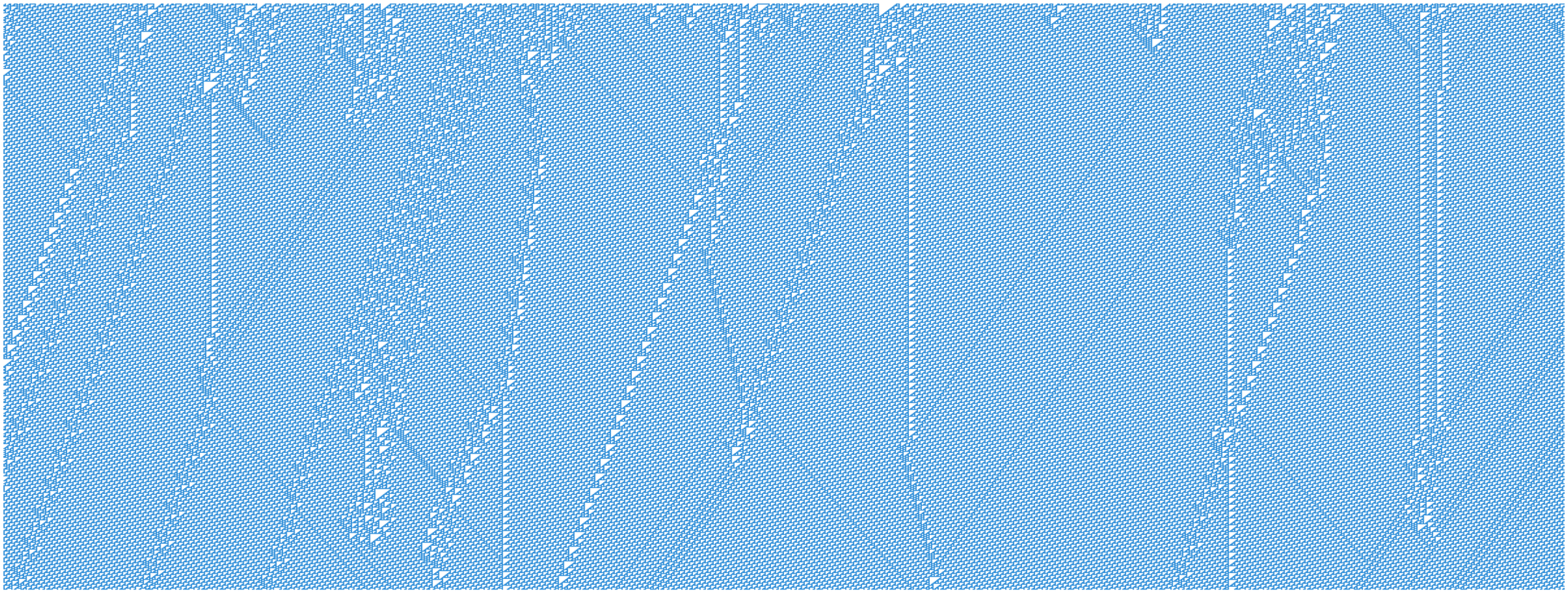
Group 6: “What are the most common video tags for the top YouTube channels?”



DECISION MATRIX

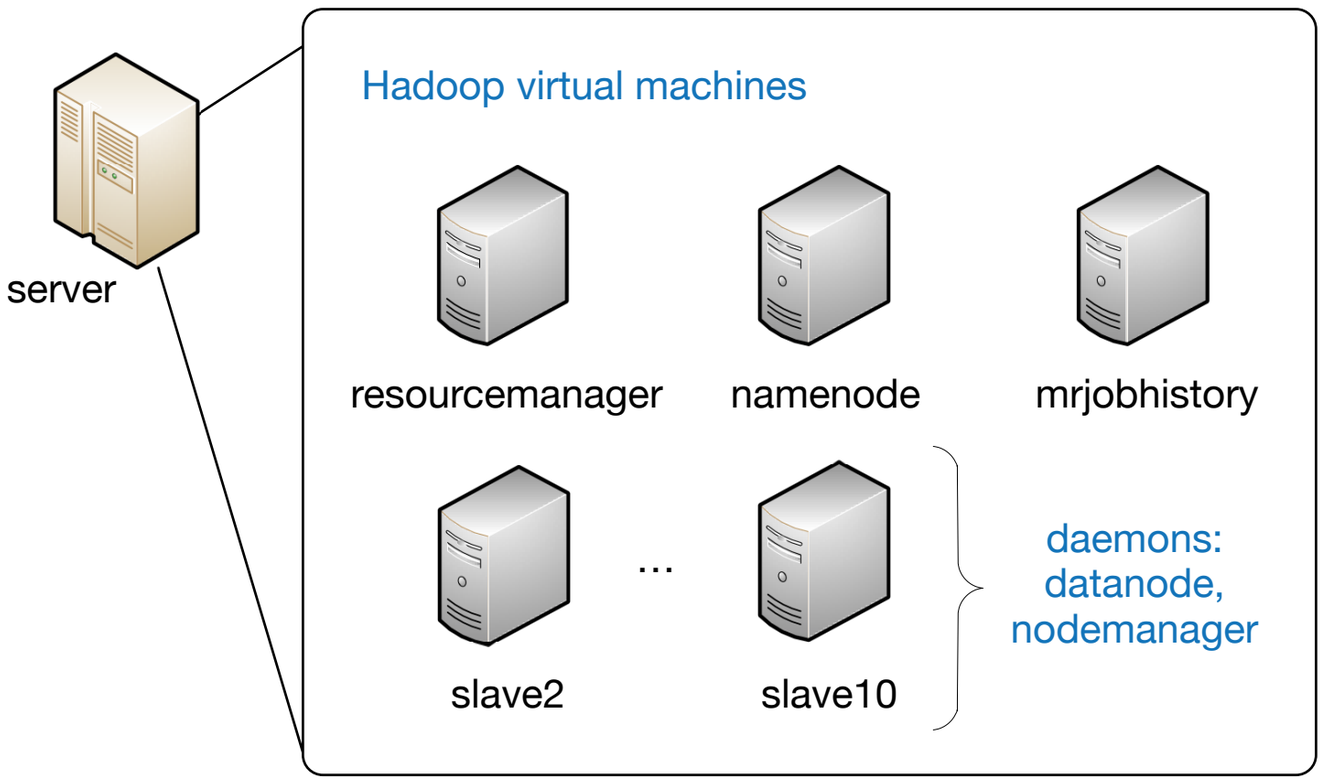
Task	Unix tools	Excel	R	MySQL	BigQuery	MapReduce	Spark	Spark MLlib	Weka	OpenCV
Data acquisition	X									
Exploratory analysis	X									X
Plotting			X							
SQL-like queries										
Distributed workers							X			
Numeric/string processing							X			
Machine learning										
Image processing										X

My example ratings for Stars assignment.



HARDWARE

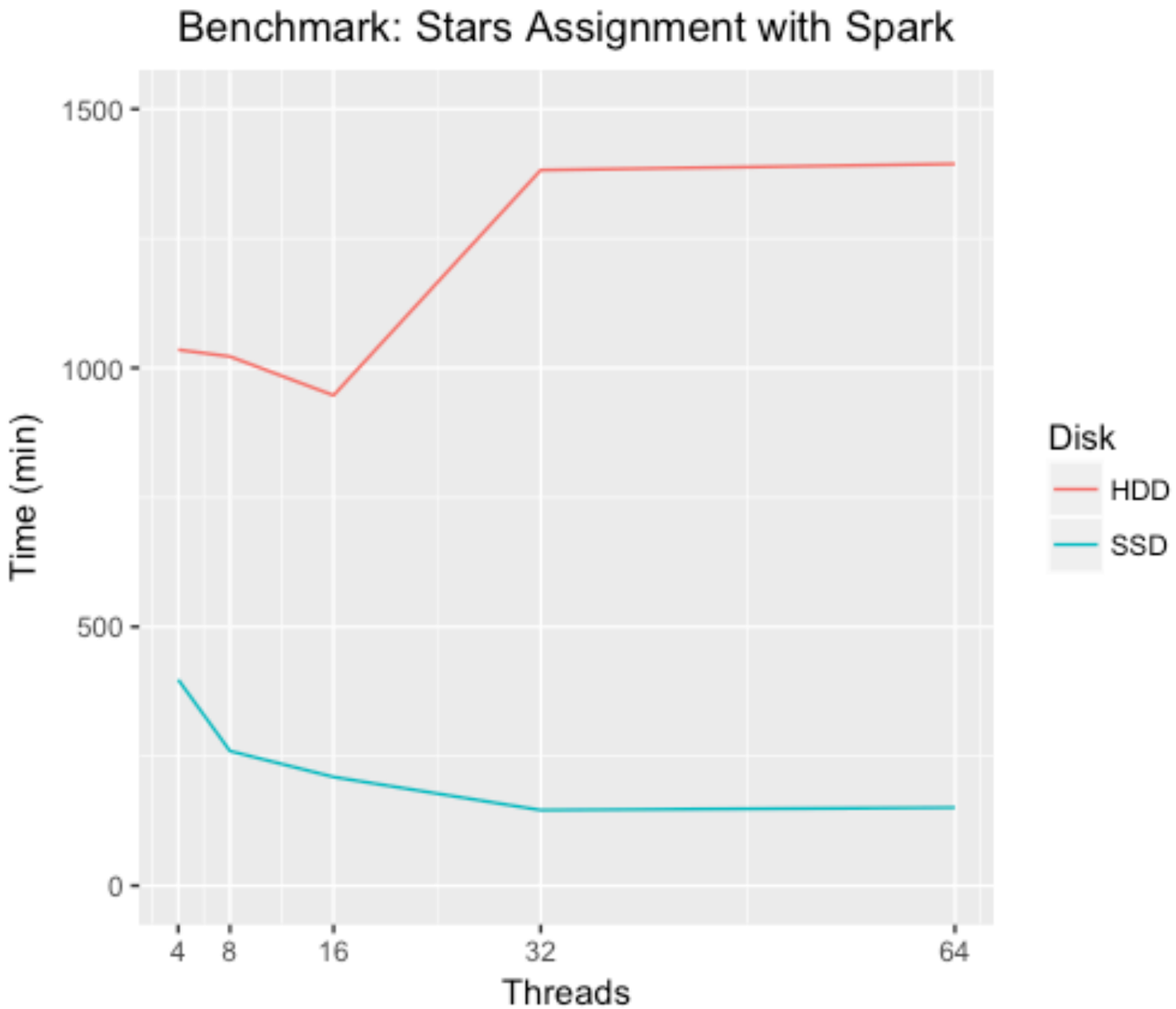
PHYSICAL AND VIRTUAL MACHINES

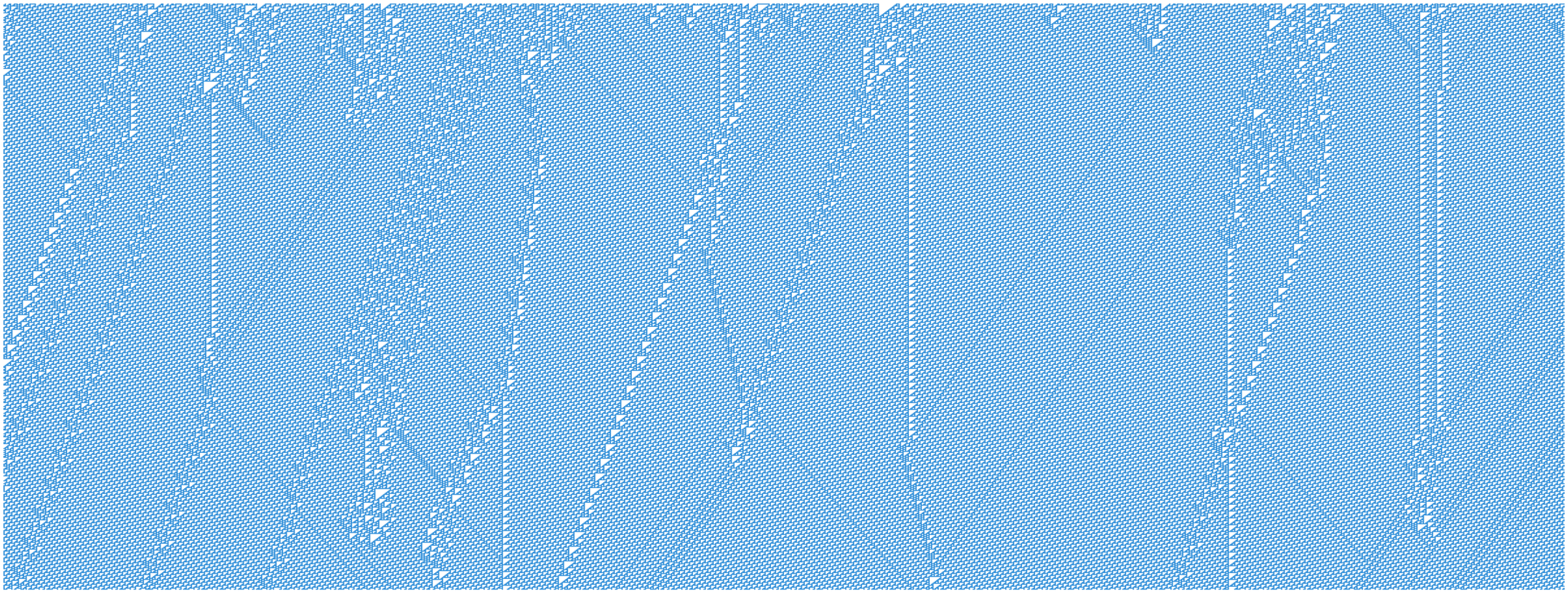


AMD Opteron 6272 x2
(32 cores, 2.1 GHz)
128 GB memory
Hard drives totaling 28 TB
NVIDIA GTX 1080

Arch Linux host OS
Libvirt + Qemu/KVM
Ubuntu 14.04 guest OS

HDD VS. SSD


Benchmark: Stars Assignment with Spark
1500
1000
500
0
Time (min)
4
8
16
32
64
Threads
Disk
HDD
SSD



STUDENT FEEDBACK

STUDENT ANONYMOUS EVALUATIONS

Spring 2015, N = 17 out of 20

Prompt	Mean	Std. dev.
Gaining factual knowledge	4.29	0.89
Learning fundamental principles	4.18	0.86
Learning to apply course material	4.18	0.78
Developing specific skills needed by professionals	4.19	0.96
Learning how to find and use resources	4.24	0.81
Acquiring an interest in learning more	4.18	0.92

Spring 2017, N = 8 out of 12

Prompt	Mean	Std. dev.
Gaining factual knowledge	5.00	0.00
Learning fundamental principles	5.00	0.00
Learning to apply course material	5.00	0.00
Developing specific skills needed by professionals	5.00	0.00
Learning how to find and use resources	5.00	0.00
Acquiring an interest in learning more	5.00	0.00

STUDENT ANONYMOUS EVALUATIONS

Spring 2015:

"**Which aspects of this course helped you learn the most?**"

"[...] mostly the projects and how they were structured, they allowed us to take the content from the class and apply it."

"The hands on labs and extremely well put together projects."

"The practice with the various tools and software used in the class."

"Hands on experience with current trends in the industry."

"The projects that involved finding our own solution to a data analytics problem were the most challenging and I learned a lot from them. It was also nice to learn new techniques from the student presentations."

STUDENT ANONYMOUS EVALUATIONS

Spring 2017:

"**Which aspects of this course helped you learn the most?**"

"Constantly added new and innovative content throughout the year."

"Slowly building up to bigger data sets made me understand big data clearly. Having multiple small and big project with different tools was the best part of this class. I can say that I have learned a lot from this course, and projects were really challenging and really fun."

"The bulk of the grades in this course come from our projects, which really allow us to get familiar with the tools we have learned in the class."

STUDENT OUTCOMES

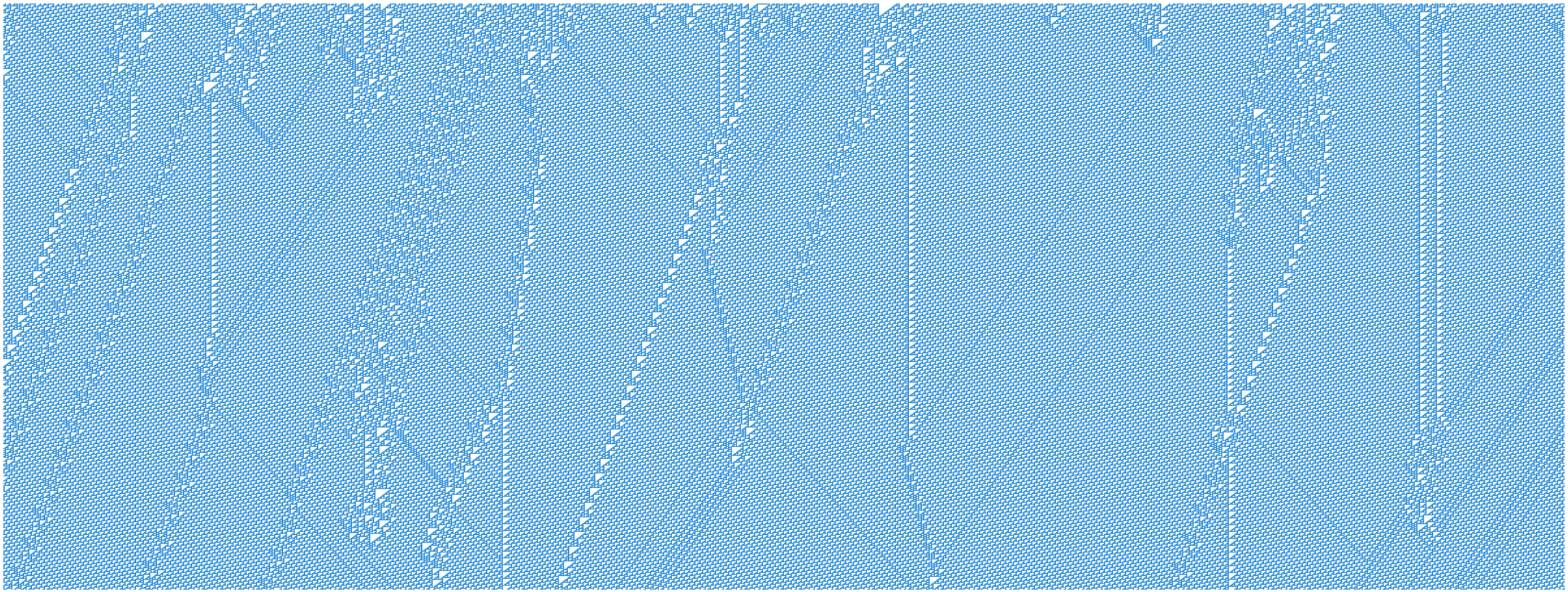
GEICO X 2

Pacific Northwest NATIONAL LABORATORY X 2

Carnegie Mellon University summer research intern

facebook summer intern

Florida Institute of Technology summer REU



FUTURE WORK

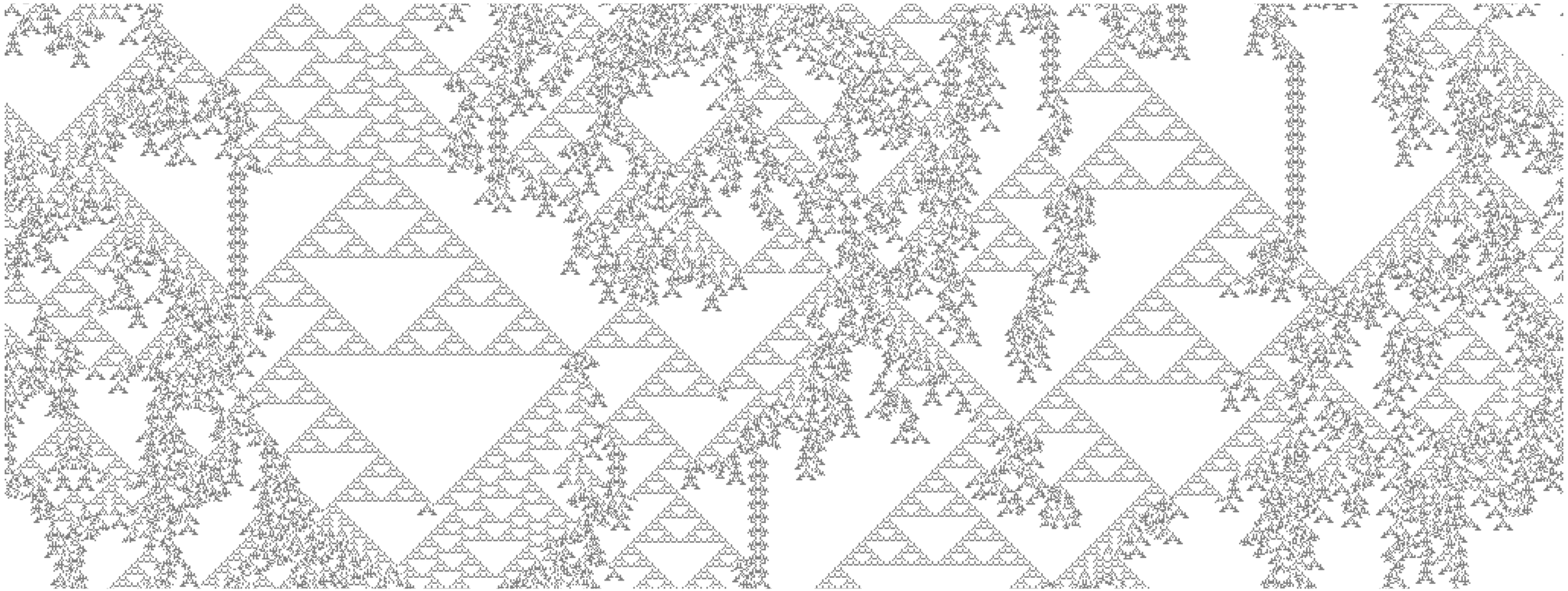
IDEAS FOR SPRING 2018

Introduce Spark earlier in lieu of a redundant R/ggplot project.

Include a stream processing project (Apache Kafka?).

Find a big image processing dataset that benefits from GPU processing. Demonstrate CPU vs. GPU processing.

Possibly: hold a competition for fastest data processing, e.g., C++ vs. Spark, local server vs. cloud.



QUESTIONS?