Google Cloud

How cloud computing and machine learning are transforming science

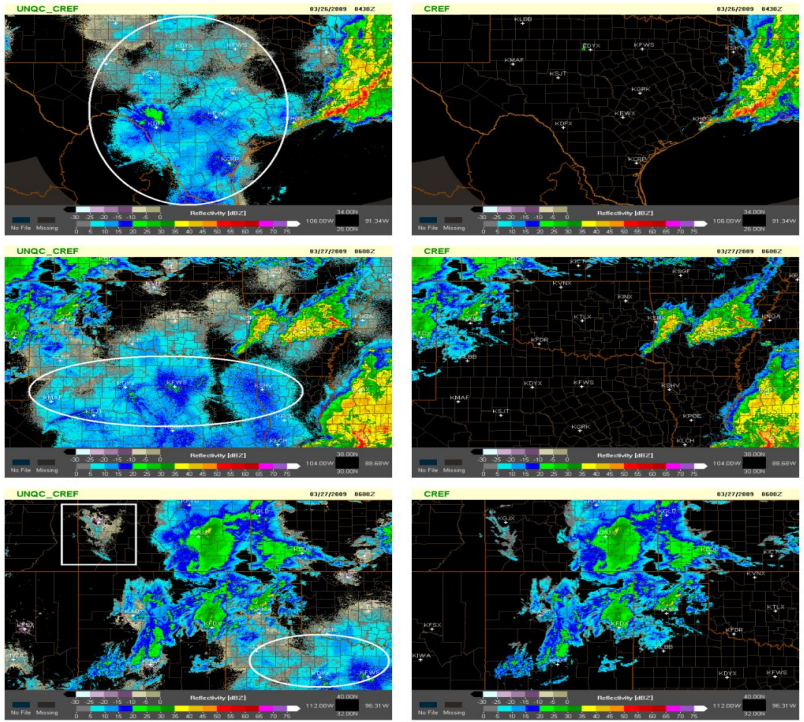
lak@google.com

Lak Lakshmanan
Tech Lead, Big Data & Machine Learning Professional Services
Google Cloud

Google Cloud

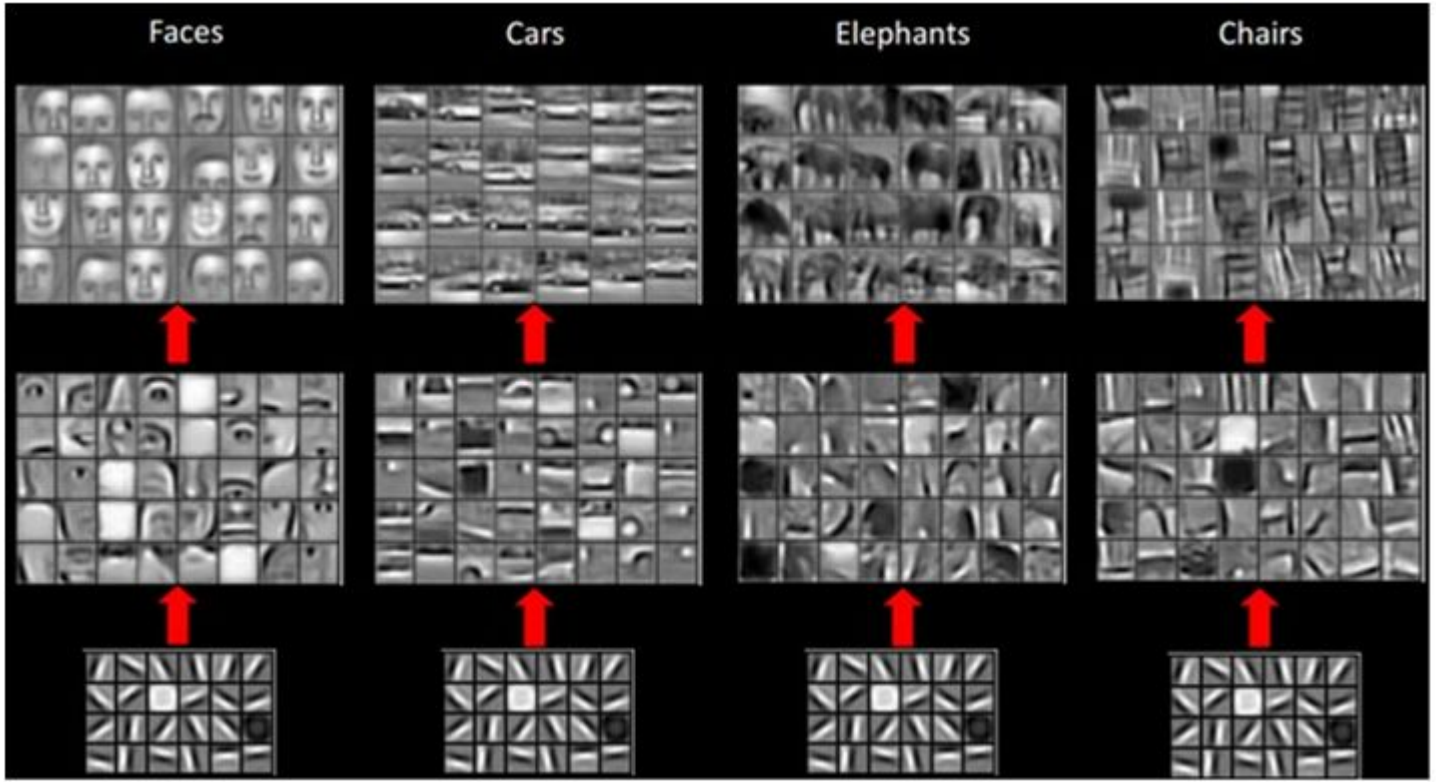
My (unusual) path to Google ...

Much of my work at NSSL involved image processing and pattern recognition

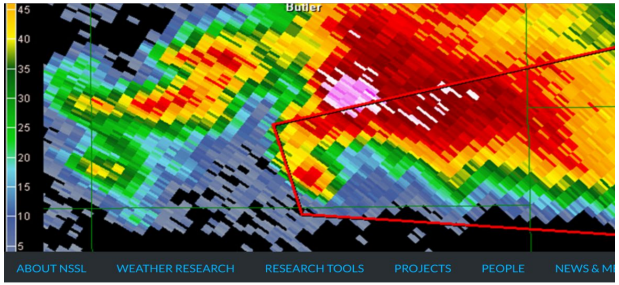


Google Cloud

Deep learning has fundamentally changed what's possible in the realm of images (also: speech, text, …)



At NSSL, it took us four years to create a “multi-year reanalysis of remotely sensed storms”



Home > Research Tools > Warning

RESEARCH TOOLS: WARNING

FACETs

Forecasting a Continuum of Environmental Threats (FACETs) serves as a broad-based framework and strategy to help focus and direct efforts related to next-generation science, technology and tools for forecasting environmental hazards. FACETs will address grid-based probabilistic threats, storm-scale observations and guidance, the forecaster, threat grid tools, useful output, effective response, and verification.

FACETs: A New Warning Paradigm & Framework for Progress (.pptx, 28.6 MB)

MYRORSS

The **Multi-Year Reanalysis Of Remotely-Sensed Storms (MYRORSS** – pronounced “mirrors”) NSSL and the National Climatic Data Center (NCDC) to reconstruct and evaluate numerical model output and radar products derived from 15 years of WSR-88D data over the coterminous U.S. (CONUS). The end result of this research will be a rich dataset with a diverse range of applications, including severe weather diagnosis and climatological information.

Google Cloud

At Climate, we did the equivalent every two weeks

Meteorology @ The Climate Corporation

Our Mission is to develop methods and curate datasets to provide the best available estimates of recent and near-term weather.

Our Team

Inspire one another
Be direct and transparent

Scientists, engineers, statisticians, and data specialists work together to curate comprehensive data sets and develop scalable, production-ready algorithms.

Our Impact

Leave a mark on the world

We provide the best available estimates of precipitation to people who need weather information the most. Our products reported field-specific weather information for 75 million acres in 2015.

Crop health/damage

Field workability

If and when to fertilize

Our Work: Project Example

Find the possible in the impossible

Motivation Drop Size Distribution (DSD) variability introduces substantial uncertainty when predicting rain rate (R) from radar reflectivity (Z).

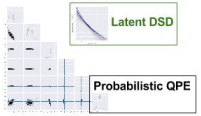
Image Source: The COMET Program

Same Z, different R

Use the DSD as a latent model to fit dual-pol radar data and generate a probabilistic Quantitative Precipitation Estimate (QPE).

Test/calibrate/validate model on best available dataset. The MC3E experiment collected disdrometer, MRR, and NPOL radar data.

Fall speed vs drop size



Bayesian sampling of DSD and rain rate from dual-pol radar observations.

Extend the initial model as needed: *spatio-temporal correlations? coalescence/evaporation of drops? better error models?*

File a patent application

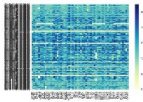
Consider model for use in production

Our Tools

Innovate in all we do

Computing

We routinely process TBs of raw radar data to evaluate precipitation algorithms.



Initial coverage map of AWS NEXRAD archive

Amazon + NOAA provide a historical archive back to 1991 as well as a real-time feed of NEXRAD data.

xray

Methodologies

We rely on a diverse set of tools that are customized to each problem:

- Statistical methods that can account for spatial and temporal correlations and estimate uncertainties (e.g. Gaussian processes)
- Machine learning (e.g. neural networks)
- Physical modeling (e.g. latent DSD model)
- Data assimilation (e.g. radar + rain gauges)
- In-house experiments and datasets

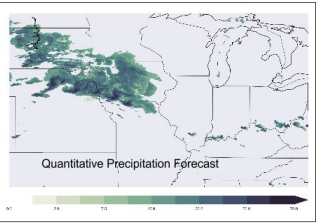
We love the interactive python data analysis ecosystem and contribute to the open source community.

xray + dask task graph

Our Future

With an unprecedented volume of radar data easily available, we are tackling new challenges at the forefront of meteorology. We focus on developing the products of most utility to farmers' operations.

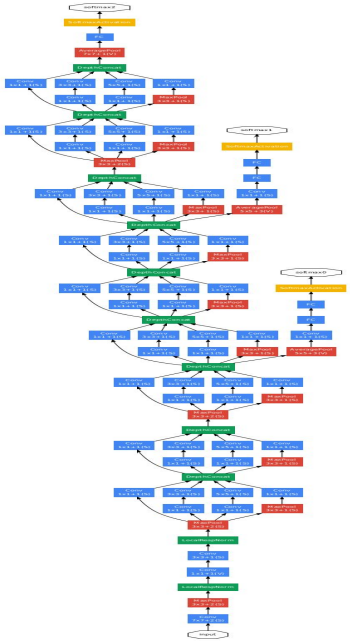
- When and how much will it rain today?
- Did today's storm cause hail damage to my crop?

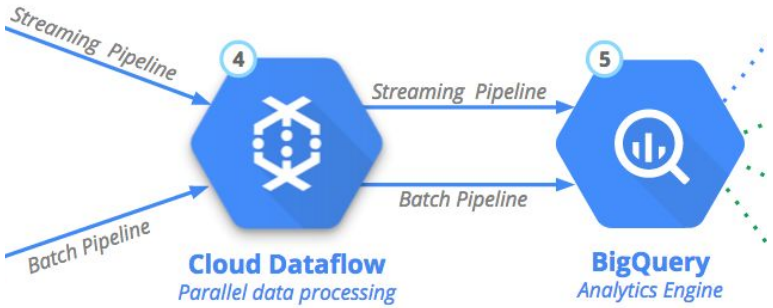


Join us and put your knowledge to work!

Google Cloud

Where would you go if you want to be part of two revolutions?

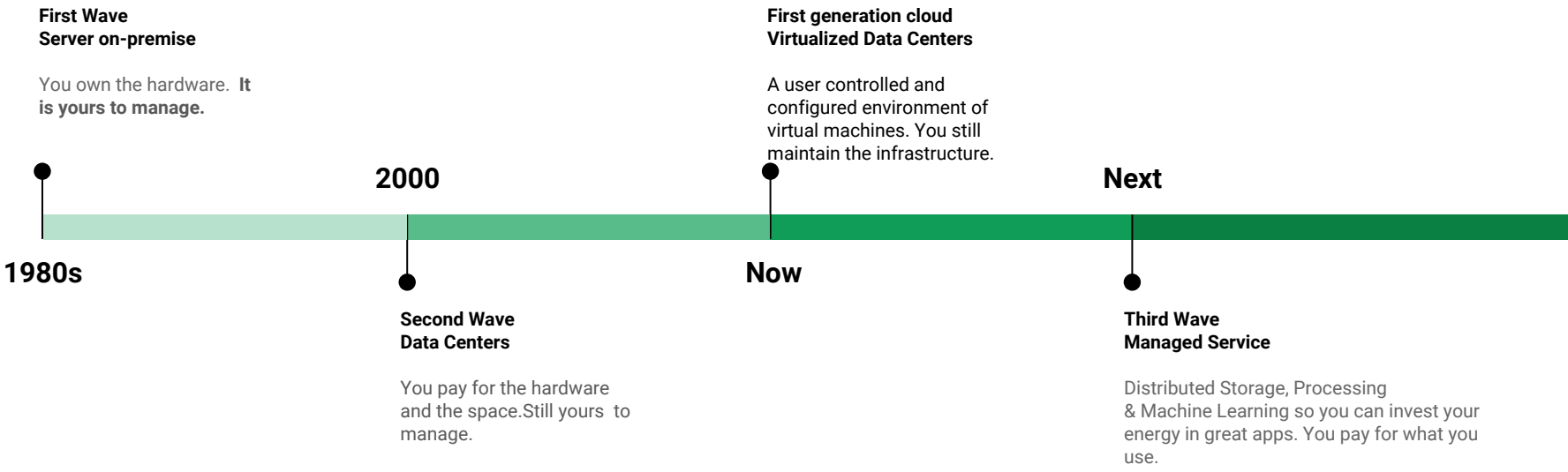




Google Cloud

Cloud computing ...

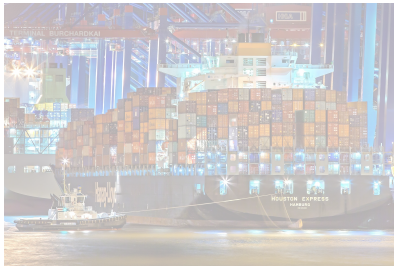
Cloud computing is a continuation of a long-term shift in how computing resources are managed



Cloud Computing brings four key benefits to science

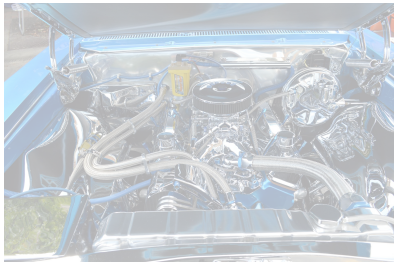
Lower cost

Repeatability



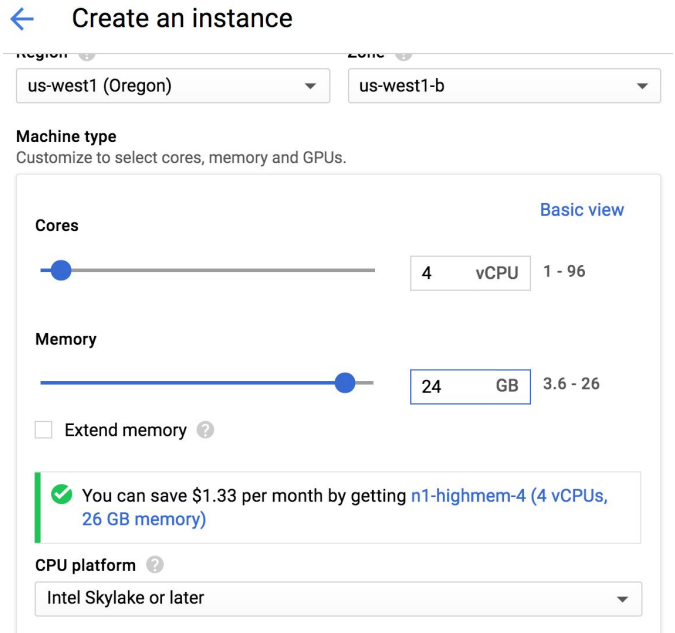
Collaboration

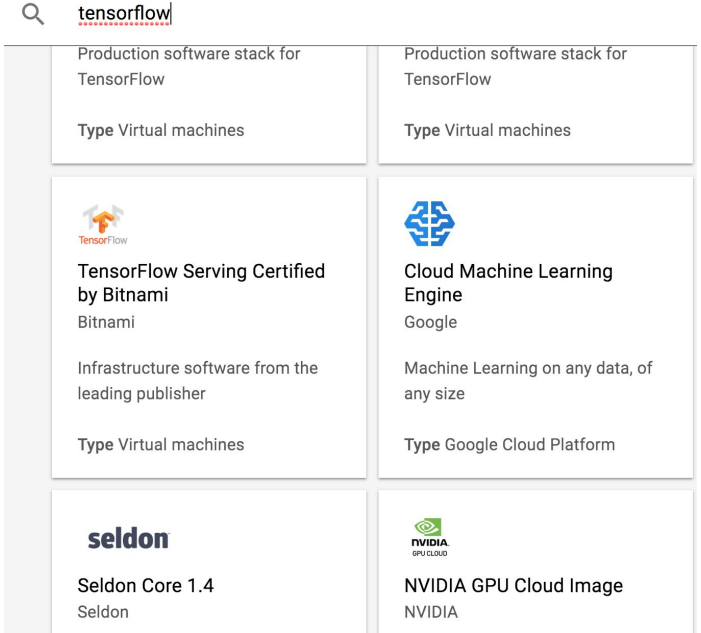
Democratization



Google Cloud

Demo: Google Compute Engine and Cloud Launcher



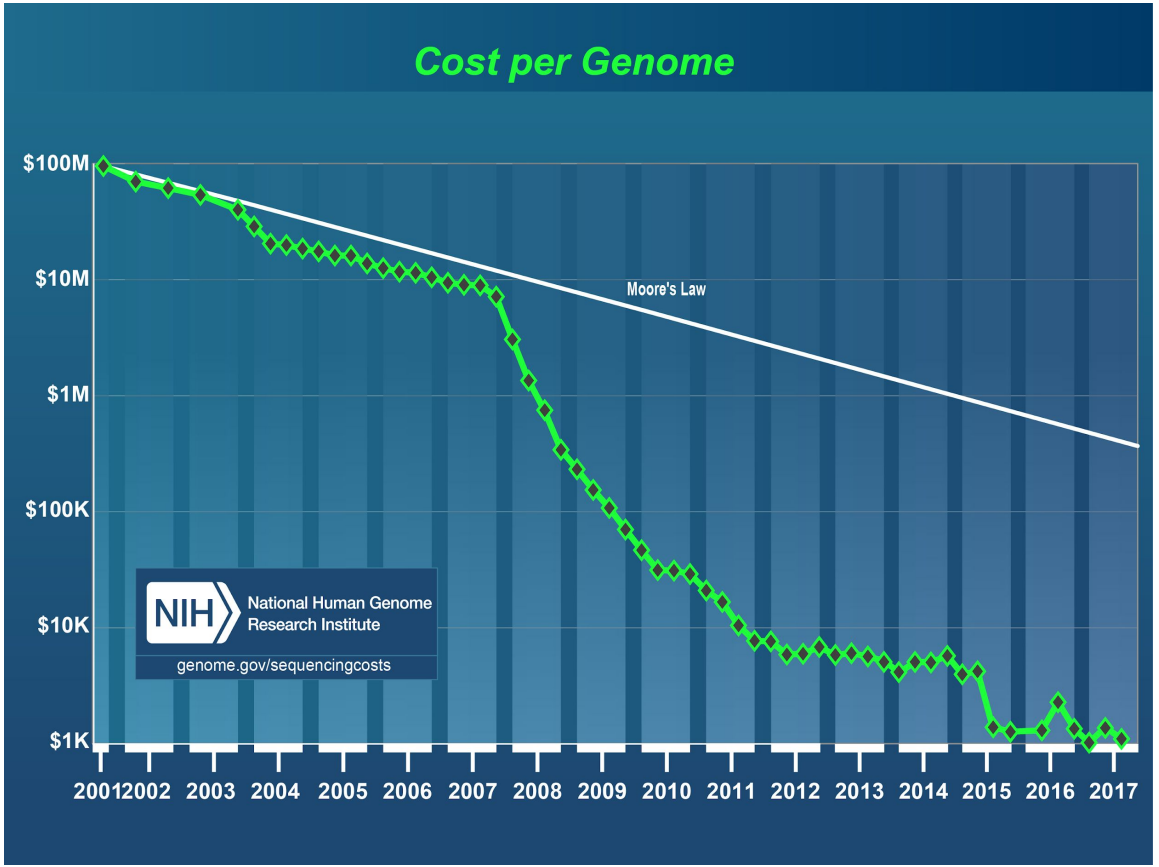


https://console.cloud.google.com/compute/instancesAdd

https://console.cloud.google.com/launcher/browse?q=tensorflow

Google Cloud

The cost of sequencing a genome has fallen dramatically



https://www.genome.gov/27541954/dna-sequencing-costs-data/

Google Cloud

8.4 Billion

The number of connected things in use in 2017, up 31% from 2016*

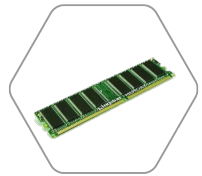
We're generating more data than ever before

Google Cloud

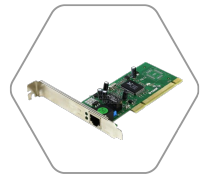
The cloud could be simply rented infrastructure ...

Storage

Processing



Memory



Network

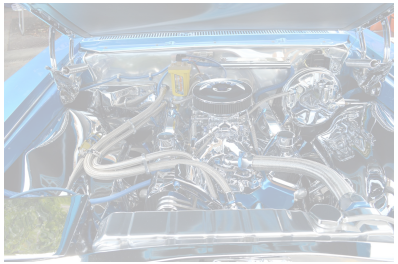
Cloud Computing brings four key benefits to science

Lower cost

Repeatability

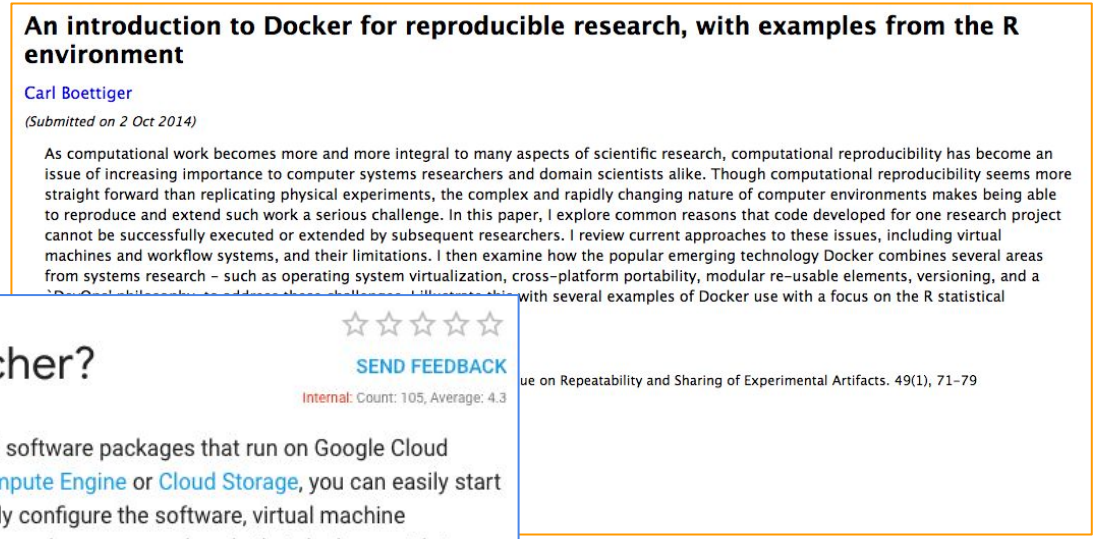
Collaboration

Democratization



Google Cloud

Software, datasets in an executable environment



An introduction to Docker for reproducible research, with examples from the R environment

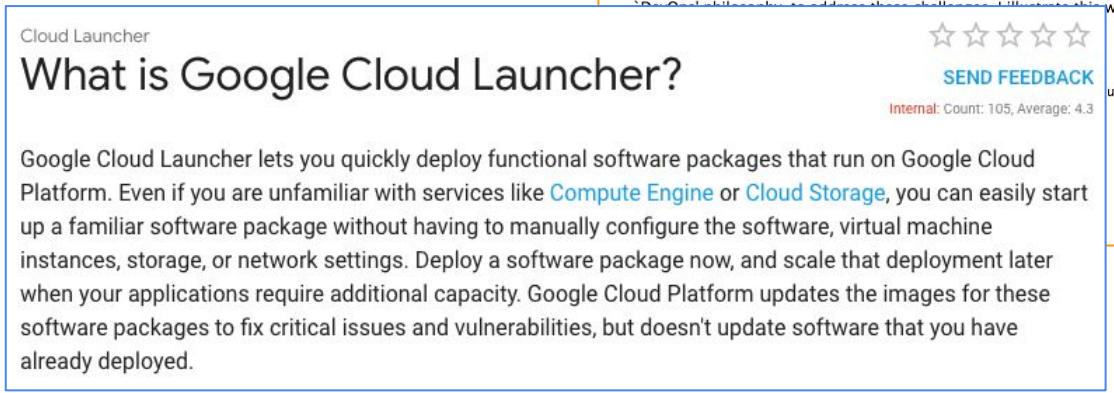
Carl Boettiger

(Submitted on 2 Oct 2014)

As computational work becomes more and more integral to many aspects of scientific research, computational reproducibility has become an issue of increasing importance to computer systems researchers and domain scientists alike. Though computational reproducibility seems more straight forward than replicating physical experiments, the complex and rapidly changing nature of computer environments makes being able to reproduce and extend such work a serious challenge. In this paper, I explore common reasons that code developed for one research project cannot be successfully executed or extended by subsequent researchers. I review current approaches to these issues, including virtual machines and workflow systems, and their limitations. I then examine how the popular emerging technology Docker combines several areas from systems research – such as operating system virtualization, cross-platform portability, modular re-usable elements, versioning, and a

with several examples of Docker use with a focus on the R statistical

ue on Repeatability and Sharing of Experimental Artifacts. 49(1), 71–79



Cloud Launcher

What is Google Cloud Launcher?

☆ ☆ ☆ ☆ ☆

SEND FEEDBACK

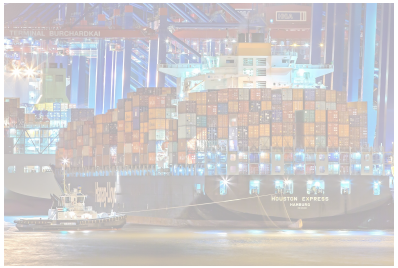
Internal: Count: 105, Average: 4.3

Google Cloud Launcher lets you quickly deploy functional software packages that run on Google Cloud Platform. Even if you are unfamiliar with services like Compute Engine or Cloud Storage, you can easily start up a familiar software package without having to manually configure the software, virtual machine instances, storage, or network settings. Deploy a software package now, and scale that deployment later when your applications require additional capacity. Google Cloud Platform updates the images for these software packages to fix critical issues and vulnerabilities, but doesn't update software that you have already deployed.

Cloud Computing brings four key benefits to science

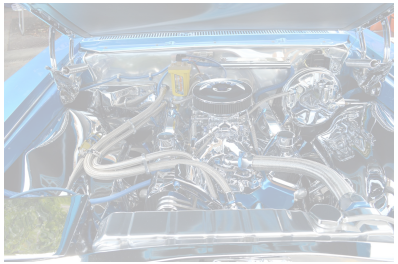
Lower cost

Repeatability



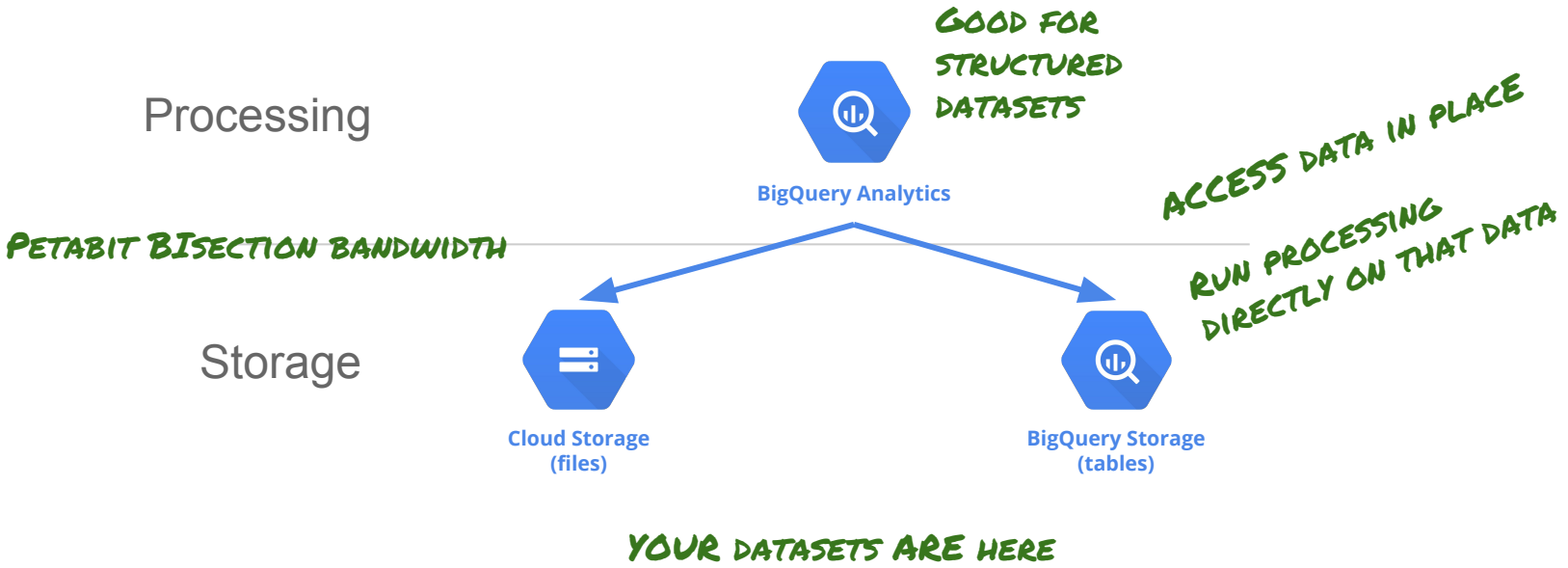
Collaboration

Democratization



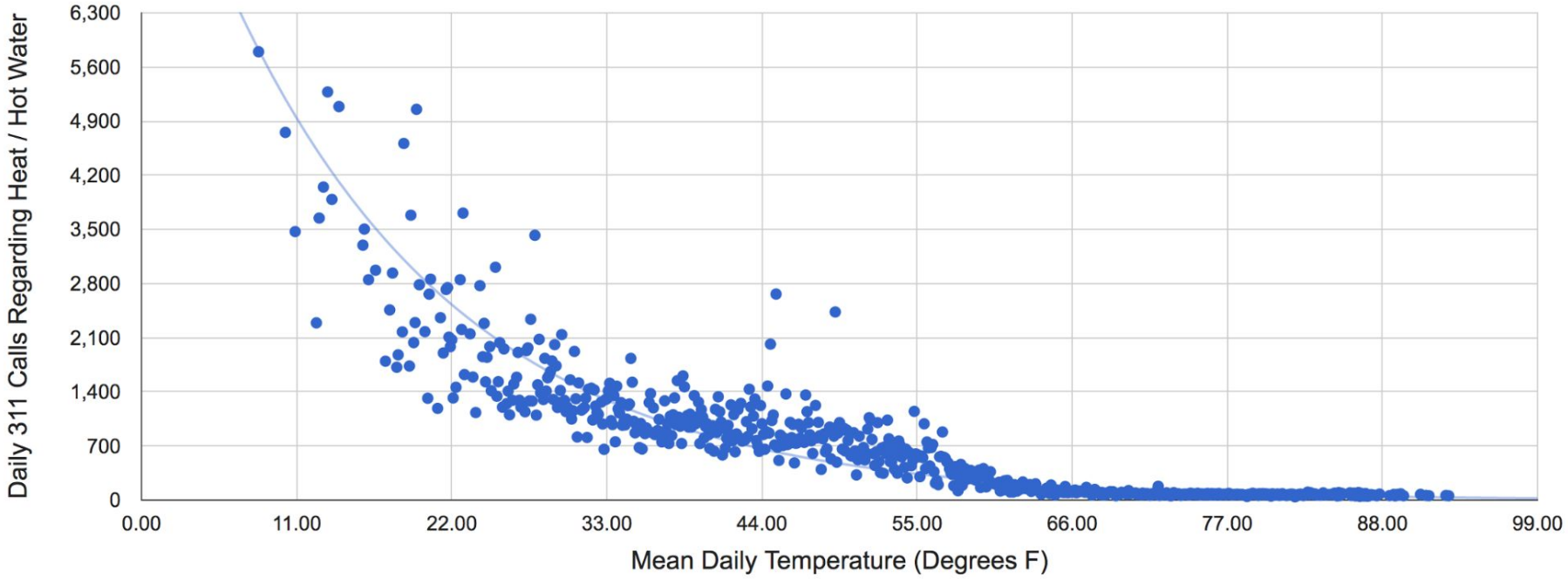
Google Cloud

GCP lets you separate data and compute, allowing for ad-hoc, ephemeral compute



e.g. Municipal complaints & weather in BigQuery

https://codelabs.developers.google.com/codelabs/scd-nycweather/index.html



What just happened?

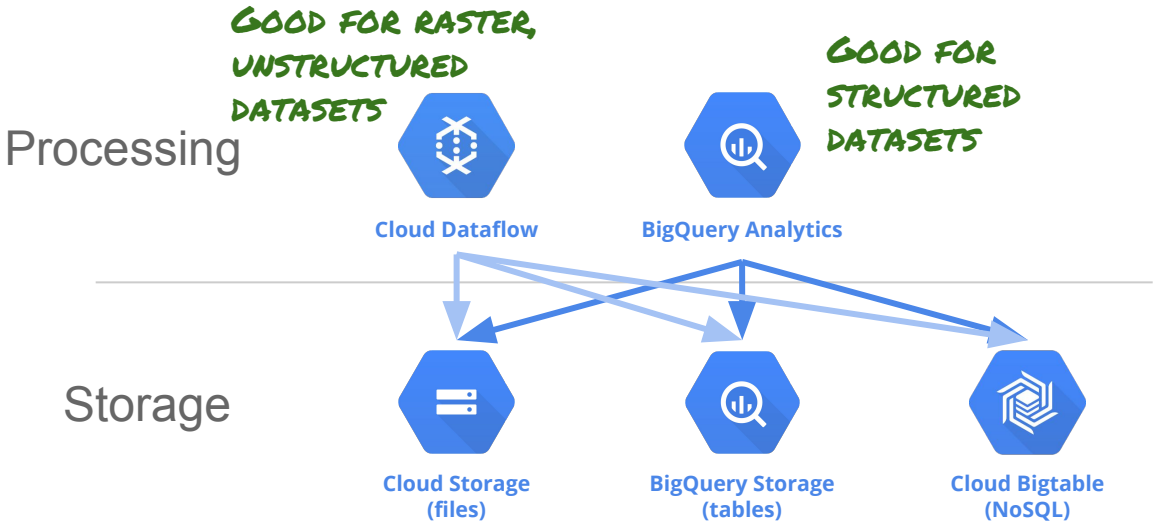
1. **Serverless.** No need to download data to your machine in order to work with it. The dataset will remain on the cloud.

2. **Ease of use.** Run ad-hoc SQL queries on your dataset without having to prepare the data beforehand (in other words, no indexes, etc.). This is invaluable for data exploration.

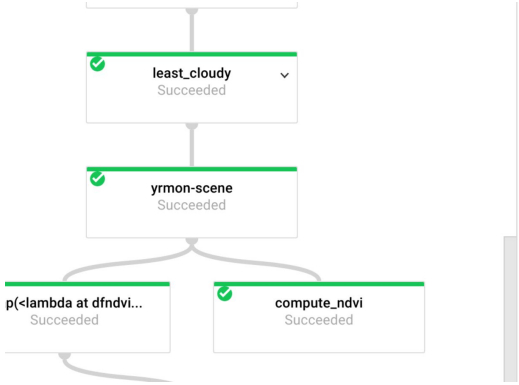
3. **Scale.** Carry out data exploration on extremely large datasets interactively. You don't need to sample the data in order to work with it in a timely manner.

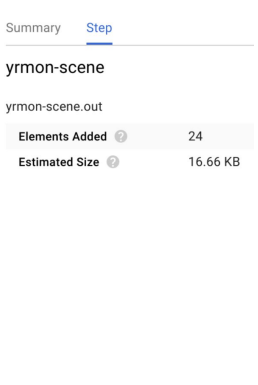
4. **Shareability**. Run queries on data from different datasets -- BigQuery is a convenient way to share datasets.

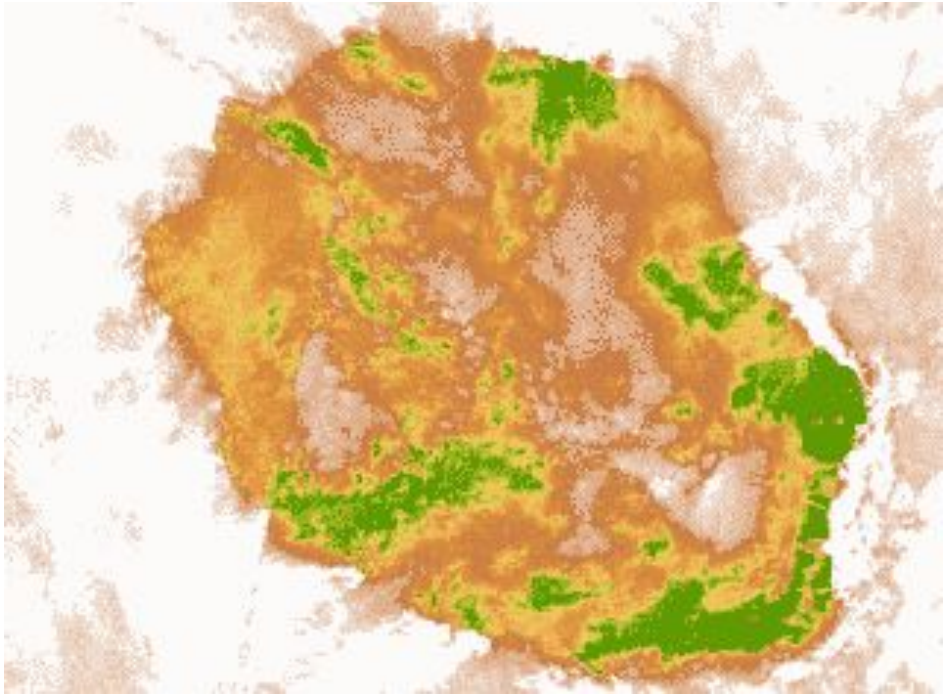
Public datasets are about ad-hoc, ephemeral compute



#4 Distributed processing of geo-imagery on GCP





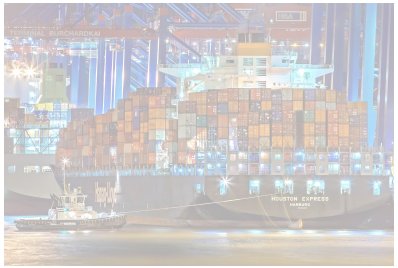


Read [this blog post](#) on what this pipeline does

Cloud Computing brings four key benefits to science

Lower cost

Repeatability



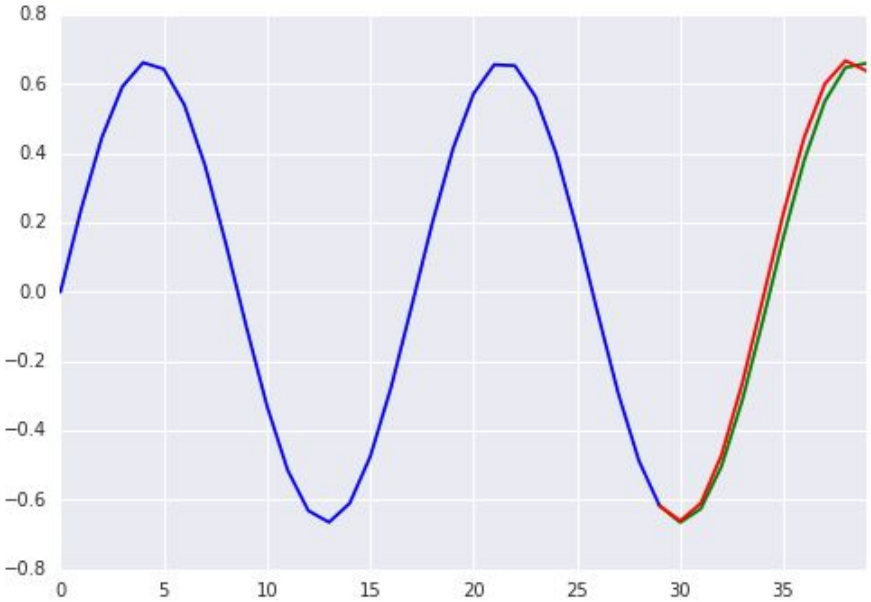
Collaboration

Democratization

Google Cloud

Machine learning ...

Predicting time series using a LSTM model … seems unimpressive ...



https://cloud.google.com/blog/big-data/2017/10/exploring-tensorflow-samples-in-google-cloud-datalab

But map words into numbers and this, too, is a sequence-to-sequence problem …

```
last december the european commission proposed updating the existing customs union with turkey and extending bilateral trade relations once negotiations have been completed the agreement would still have to be approved by the parliament before it could enter into force
```

It produces:

```
last december , the european commission proposed updating the existing customs union with turkey and extending bilateral trade relations once negotiations have been completed . the agreement would still have to be approved by the parliament before it could enter into force .
```

https://cloud.google.com/blog/big-data/2017/10/exploring-tensorflow-samples-in-google-cloud-datalab

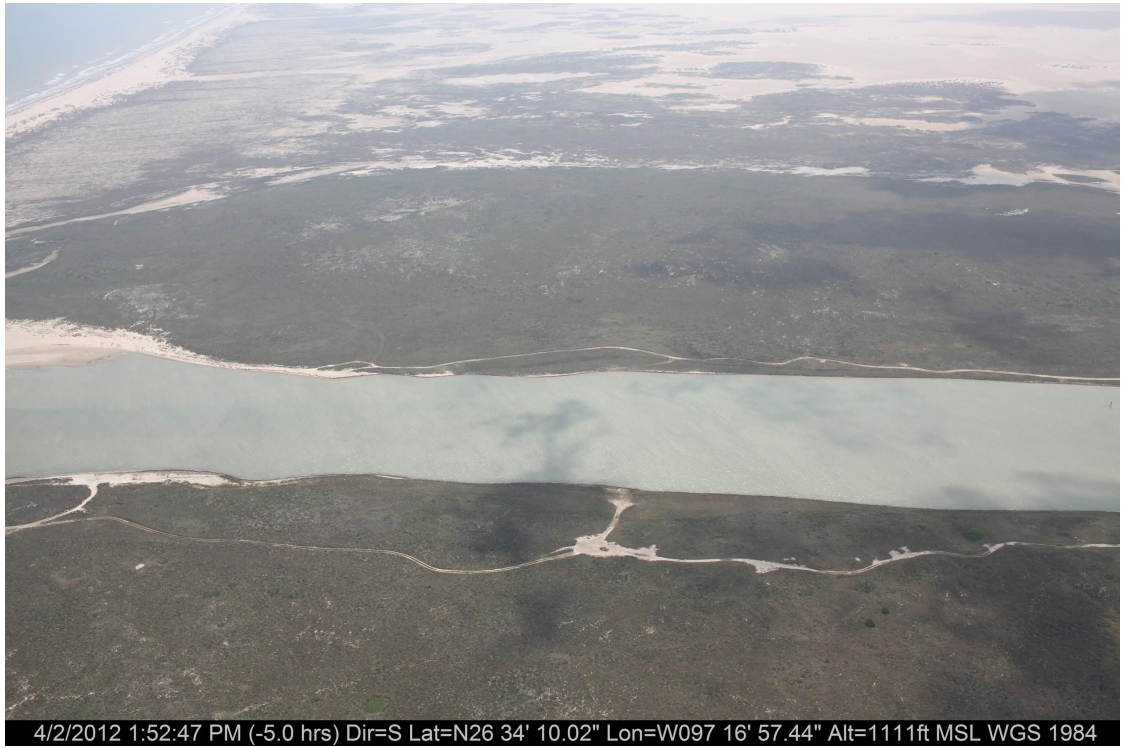
What if you combine the advances in image models (CNNs, etc.) with the advances in sequence modeling?

Generated captions:

- a cat laying on top of a rug next to a cat
- a cat laying on the floor next to a cat
- a cat laying on top of a rug next to a cat

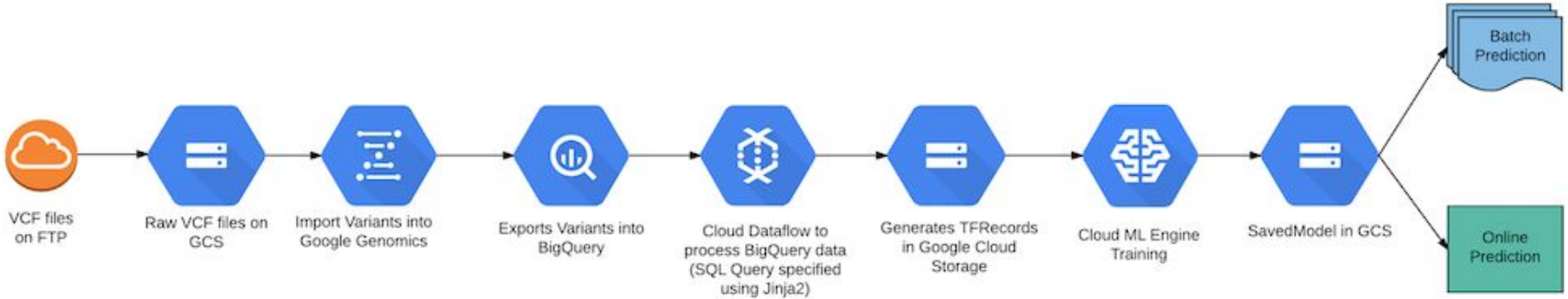
https://cloud.google.com/blog/big-data/2017/10/exploring-tensorflow-samples-in-google-cloud-datalab

ML model to classify coastline images



Google Cloud

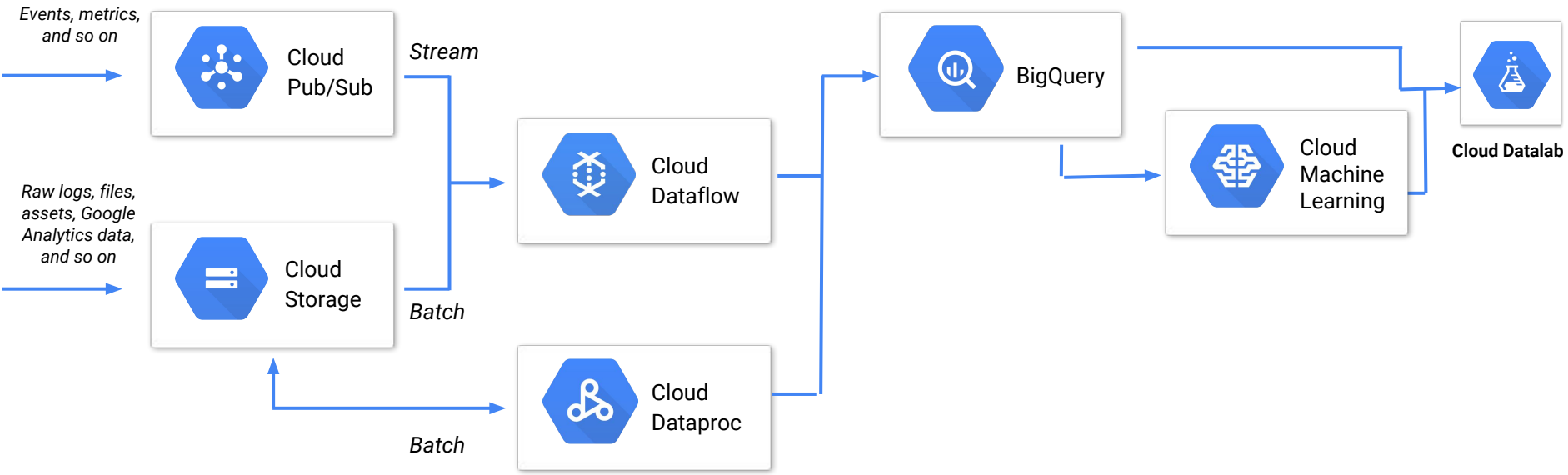
Genomics ancestry inference with deep learning



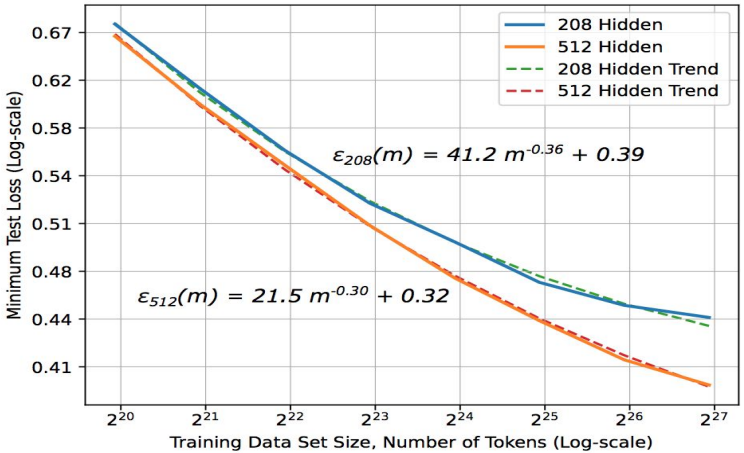
https://cloud.google.com/blog/big-data/2017/09/genomic-ancestry-inference-with-deep-learning

Google Cloud

Autoscaling data pipelines and ML

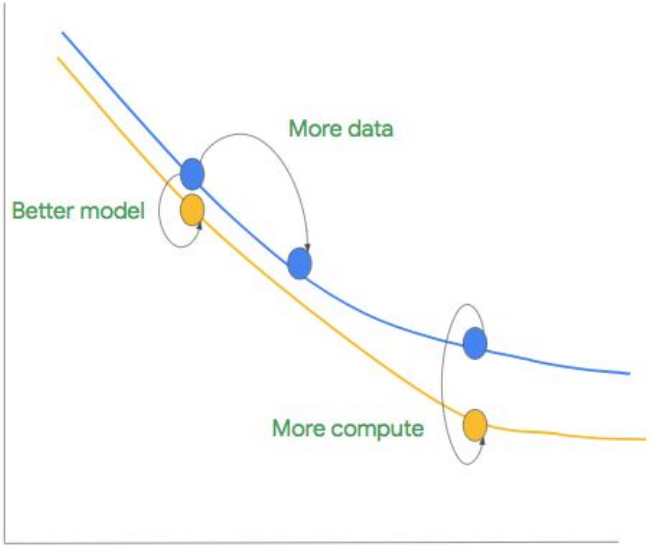


Deep learning works because datasets are large

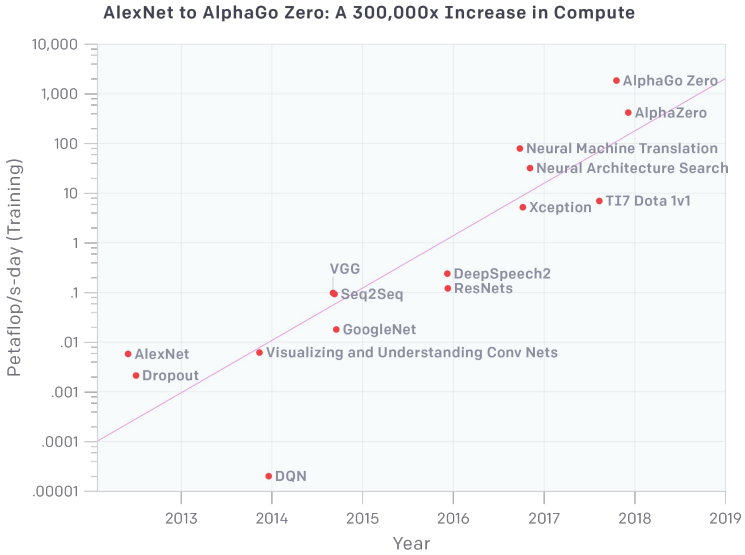


The unreasonable effectiveness of data
https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/35179.pdf

Deep Learning scaling is predictable, empirically
https://arxiv.org/abs/1712.00409

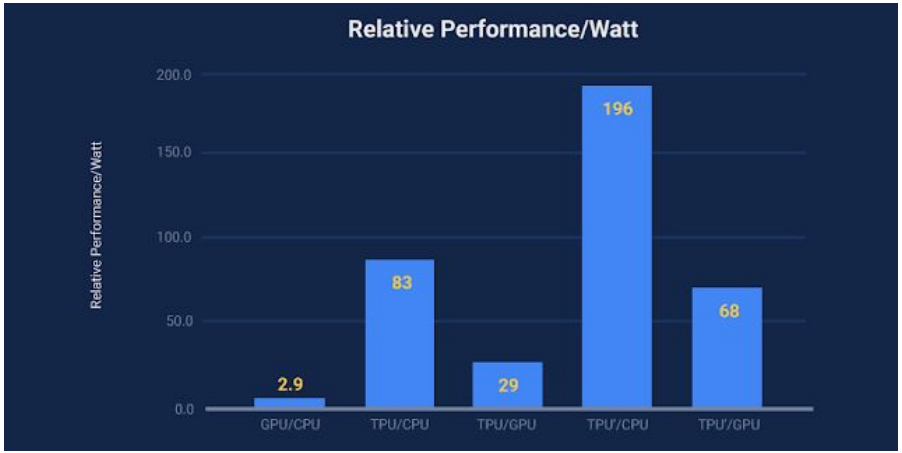


But the compute power needed keeps increasing (exponentially)



https://blog.openai.com/ai-and-compute/

The economic efficiency and collaborative power of Cloud Computing are needed for ML



https://cloudplatform.googleblog.com/2017/04/quantifying-the-performance-of-the-TPU-our-first-machine-learning-chip.html

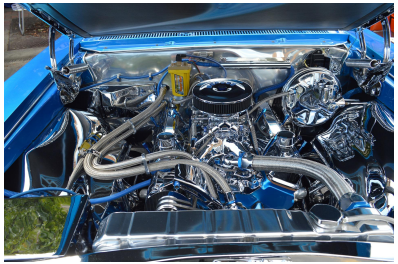
Cloud Computing brings four key benefits to science

Lower cost

Repeatability

Collaboration

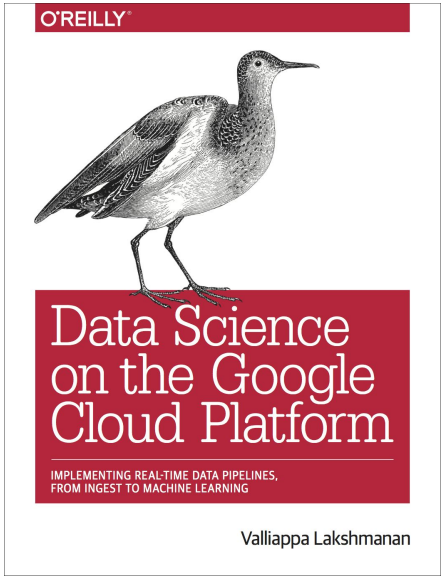
Democratization



Google Cloud

Resources

https://codelabs.developers.google.com/cloud-quest-scientific-data



Google Cloud