

Cloud Evolution and Scientific Data Analytics

Dennis Gannon

MSR retired / emeritus prof CS Indiana Univ.

gannon@Indiana.edu

Dennis.gannon@outlook.com

Outline

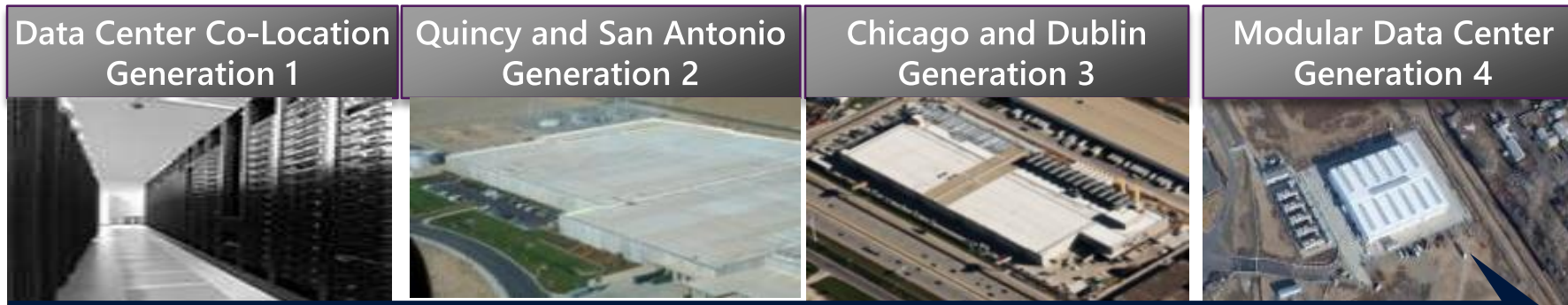
- Why Clouds? (An “industry perspective”)
- Cloud Evolution
- Science and the 4th Paradigm (Science perspective)
- Examples:
 - Urban Informatics
 - The Cloud and the Environment
- Research Opportunities

Why Clouds? (An Industry Perspective)

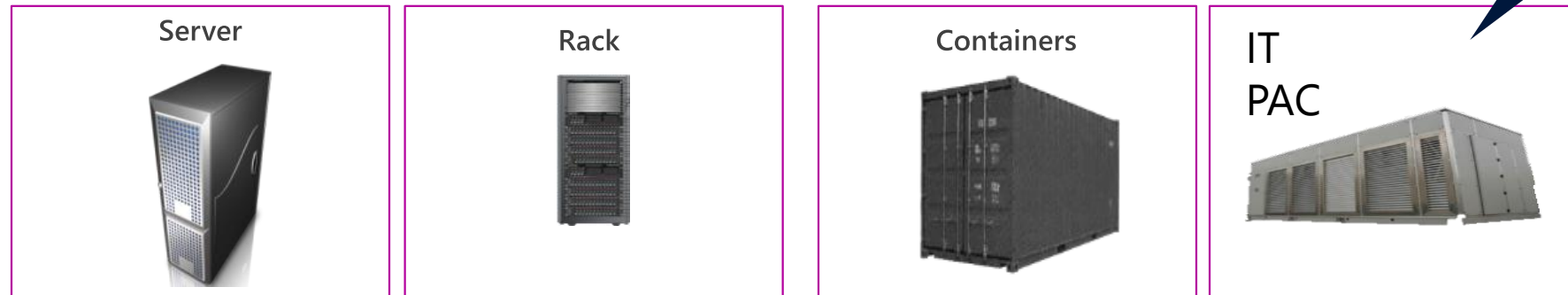
- Initially big data centers needed to support
 - On-line services
 - Building the search engine.
- Then to understand “user intent”
 - Massive buildup of analytics and machine learning capability
- Generalize the infrastructure and make it efficient
- Enter the public cloud business
 - Massive advantage for startups
 - Powerful opportunity for big enterprise
 - Public/Private Cloud
 - Offload burst challenges

Public Cloud Evolution

- Racks to Containers to Custom SOCs & FPGAs



Deployment Scale Unit



Server



Capacity

Rack



Density &
Deployment

Containers



Scalability &
Sustainability

IT
PAC



Time to Market
Lower TCO



Doug Burger's Bing
Catapult FPGA

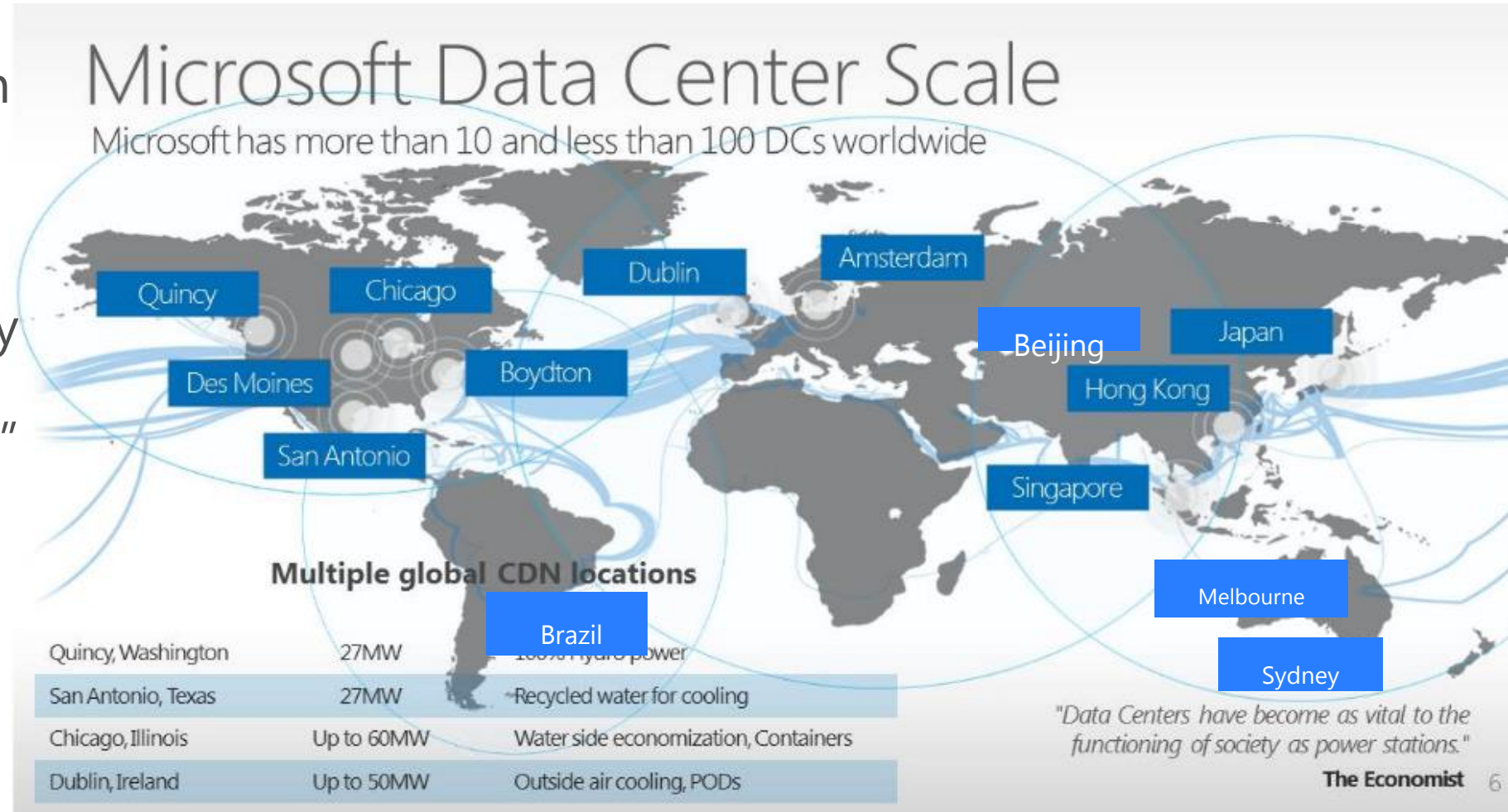
Modular Cloud Construction



Global Scale

For the industry as a whole

- Millions of servers.
- Each data center is a several hundred million dollar investment
- Counting Amazon, Microsoft, Google only there are approximately 50 major data centers and hundreds of "edge" centers.



Cloud Software Evolution

- The drive towards efficient management at scale
 - The economics of the public cloud = economies of scale
 - Goal: reduce the cost to manage infrastructure to near zero.
 - **Accomplish this with uniform abstractions and separation of concerns**
- IaaS to PaaS and SaaS
 - raising the level of abstraction
- Make the cloud a programmable platform for Big Data analytics

Programming tools: Scala, IPython, Azure ML, ...

Frameworks: Spark, Hadoop, Yarn, HDInsight, Reef, Twister, Brisk

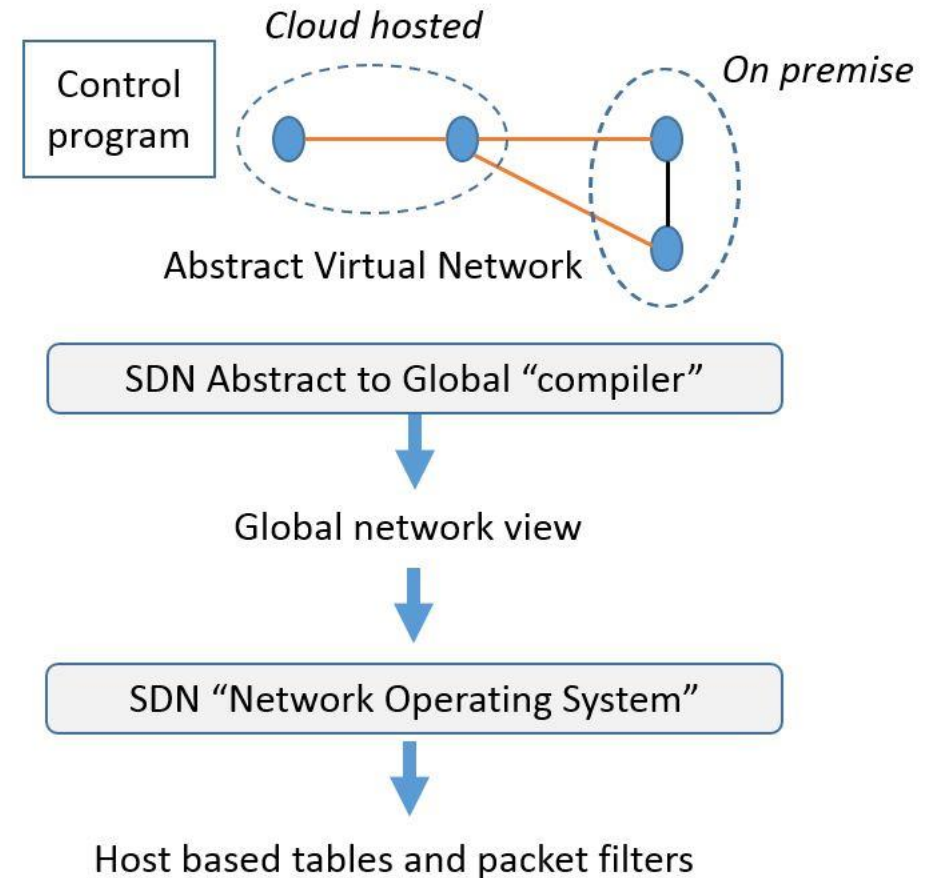
Software Defined Storage

Software Defined Networks

Hardware Abstraction/Virtualization

Software Defined Networks

- From racks and TORS to software defined systems
 - Goal: reduce the cost to manage infrastructure to near zero.
 - How: rebuild the datacenter with SDN and more ...
- Redefining the network control plane
 - Uniform abstractions
 - Separation of concerns
- The Network Operating System
 - Move control functions from switches to the hosts.
- Defining programming abstractions
 - Topology, policy (security, authorization, etc) , optimization
 - See Scott Shenker's excellent lectures <http://tce.technion.ac.il/files/2012/06/Scott-shenker.pdf>
 - See Albert Greenberg discussion of Azure www.opennetsummit.org/pdf/2013/presentations/albert_greenberg.pdf
 - See Pyretic (<http://www.cs.princeton.edu/~jrex/papers/pyretic-login13.pdf>)



Software Defined Storage

- Abstractions and Separation of Concerns?
 - Not as well defined as SDN, but some API convergence
 - S3, Azure, Ceph, Swift Object stores not that far apart.
 - Distributed replication and erasure code recovery schemes
- Misfit with distributed analysis frameworks
 - Need memory-centric, high performance, distributed storage to support the cloud analysis tools.
 - Tachyon has the right idea.
- How can Chameleon & Cloudlab help?

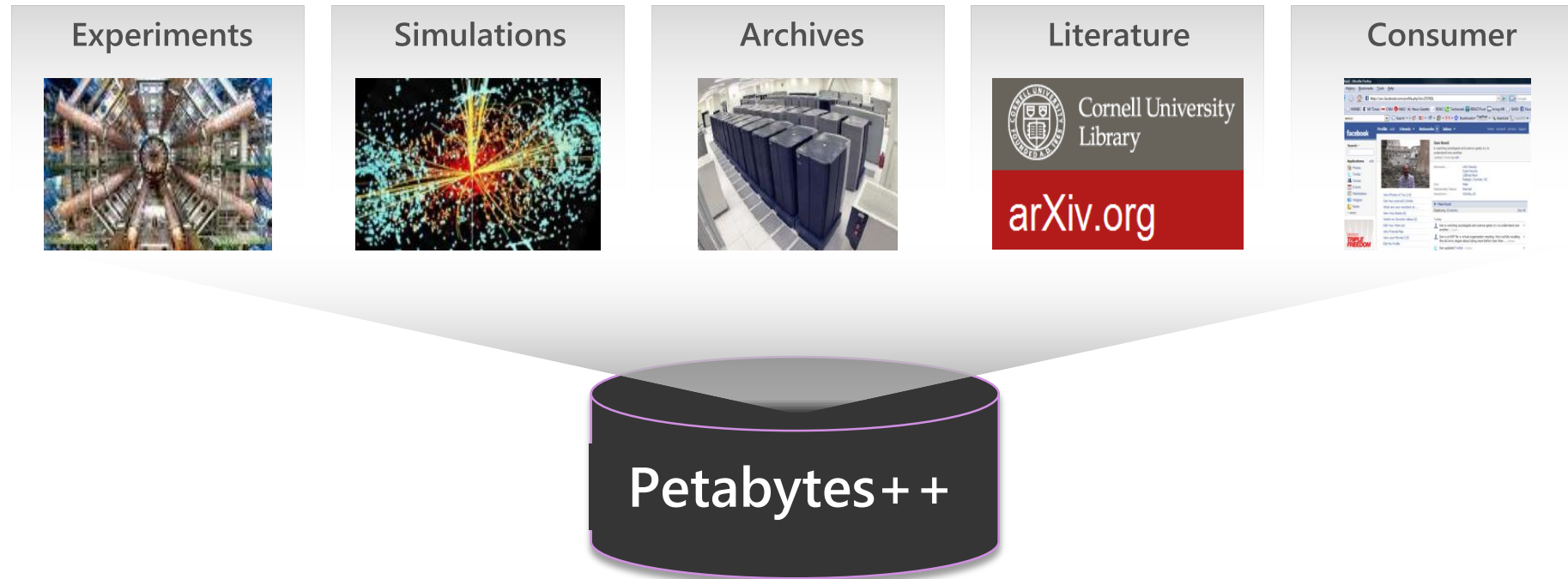


From Haoyuan Li et. al. Spark Summit talk

The Science Perspective

- Now everything generates data.
- What are the implications for scientific research?
- How can Chameleon Help?

Data Explosion is Transforming Science

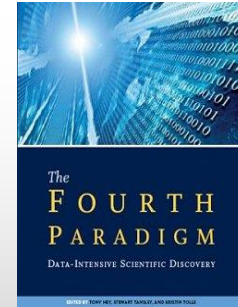
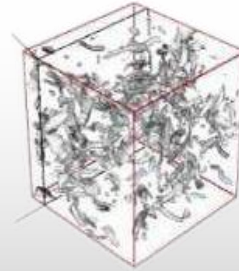


Every research field is now a data science field

The Changing Nature Of Research



$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K \frac{c^2}{a^2}$$



Experimental

Thousand
years ago

*Description of natural
phenomena*

Theoretical

Last few
hundred years

*Newton's laws,
Maxwell's equations...*

Computational

Last
few decades

*Simulation of
complex phenomena*

The Fourth Paradigm

Today and the Future

*Unify theory, experiment and
simulation with large
multidisciplinary Data*

*Using data exploration and
data mining
(from instruments, sensors,
humans...)*

Distributed Communities

Examples: Environmental Science in the Cloud

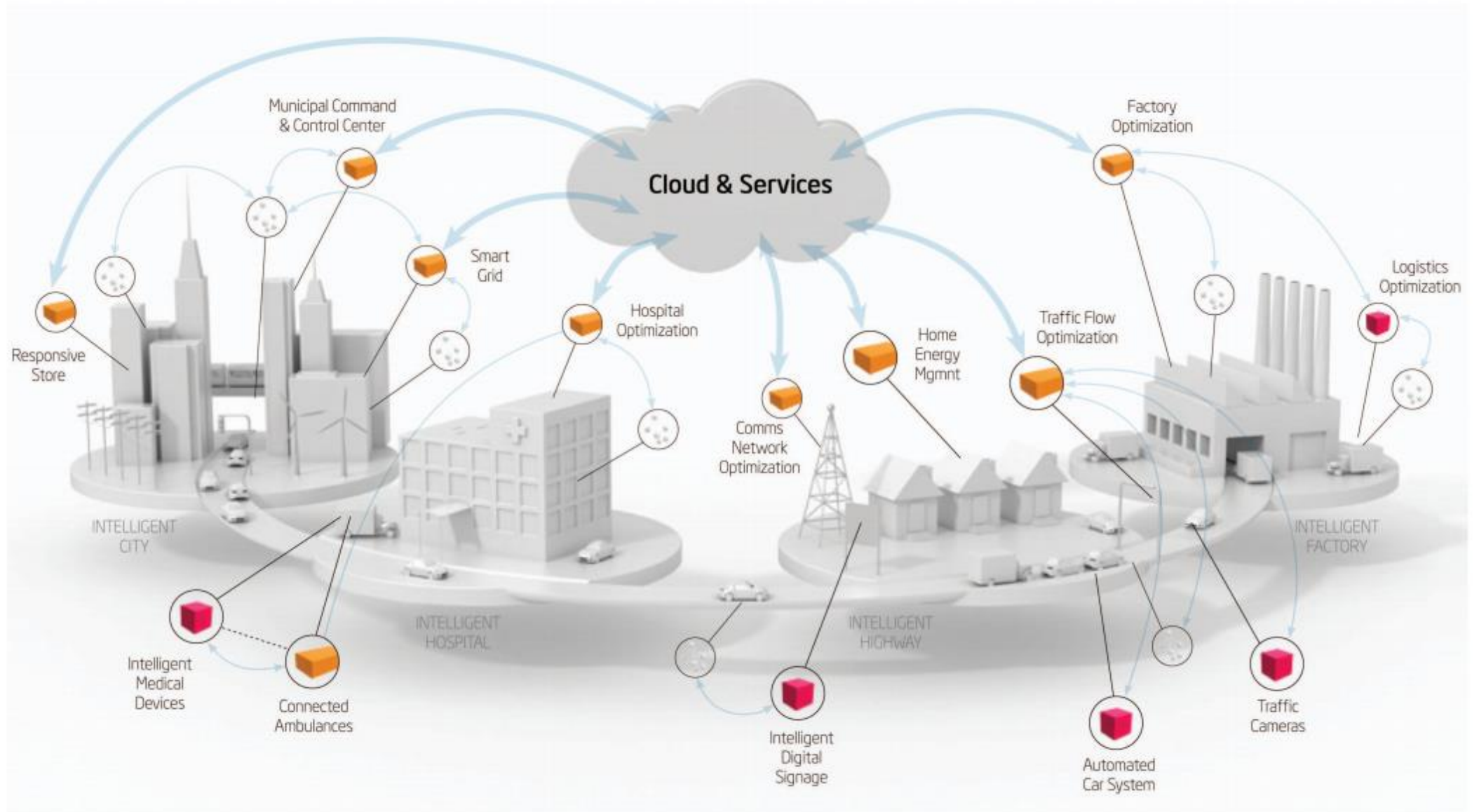
Fighting Fire with Data

- Kostas Kalabokidis @ Univ of the Aegean
- Web portal integrating satellite data, ground sensors, gps signals, weather models
- Real-time data management & fire response

[Video Link](#)



Urban Science and the Cloud



Urban Science Projects



- The Goals

- Improve traffic, energy use, city planning, safety, pollution, emergency response, etc.
- Big project in New York, Chicago, Beijing and more.
- China investing heavily.

- Sample new projects

- Brazil "Cloud Support to Crowdsensing Applications in Smart Cities Scenarios"
- US "Investigating feedbacks between climate and air travel at a global scale"
- France "Recommendation engine for daily human commuting behaviors in Paris Area through mobile phone data analysis"
- Switzerland "Decision Support System (DSS) that leverages 3D digital urban data to facilitate environmental analyses in cities"

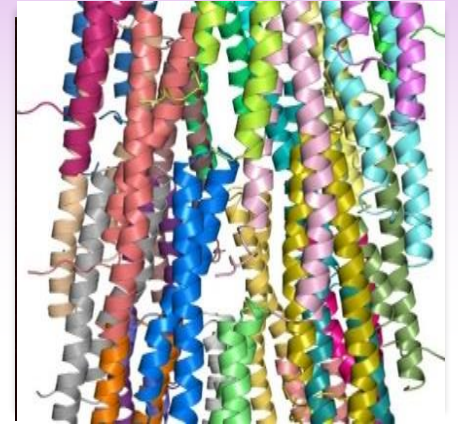


Life Science and Map Reduce

- *Do the same computation on a large collection of inputs and summarize the results.*
- Protein Folding
 - Post-doctoral researcher Nikolas Sgourakis of Baker Lab at Univ. of Washington studied ways in which proteins from Salmonella fold.
 - Used 2000 concurrent cores on Azure
- INRIA fMRI Brain Image Analysis
 - Joint genetic and neuroimaging data analysis on large cohorts of subjects
 - Built custom Map-Reduce and distributed storage layer to solve the problem.

Protein Folding

The University of Washington
Baker Laboratory



INRIA "Azure Brain"

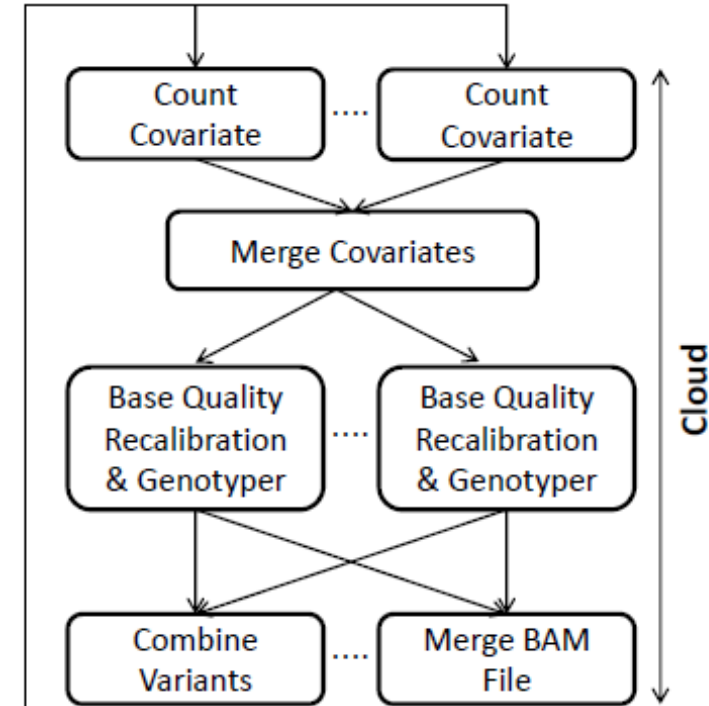
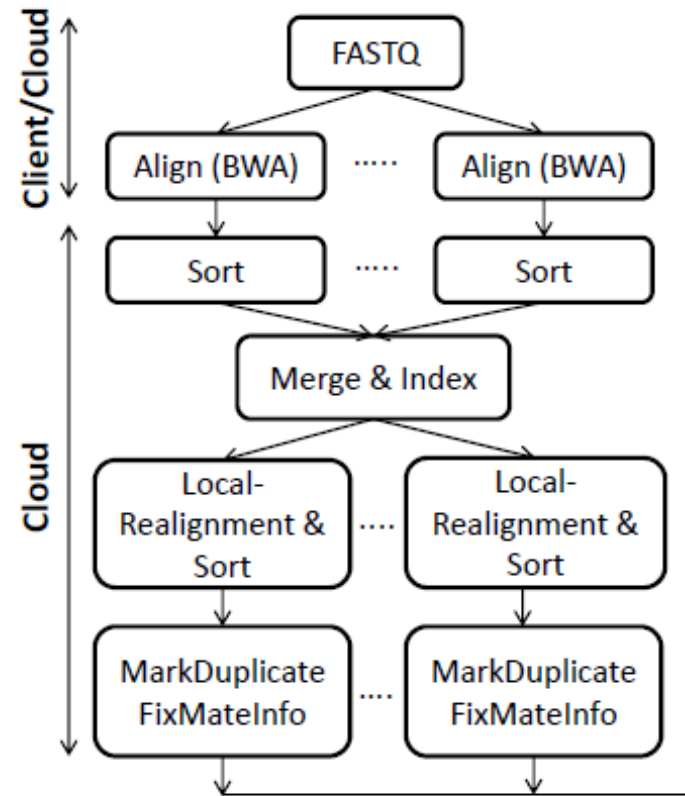
Premier information
technology research
laboratory



More Examples

- HDInsight and Bioinformatics

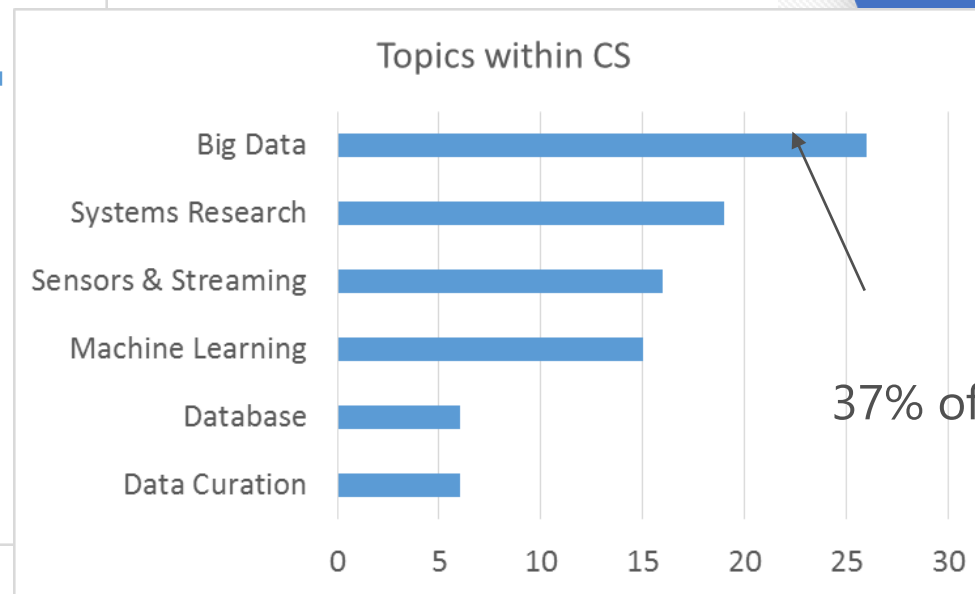
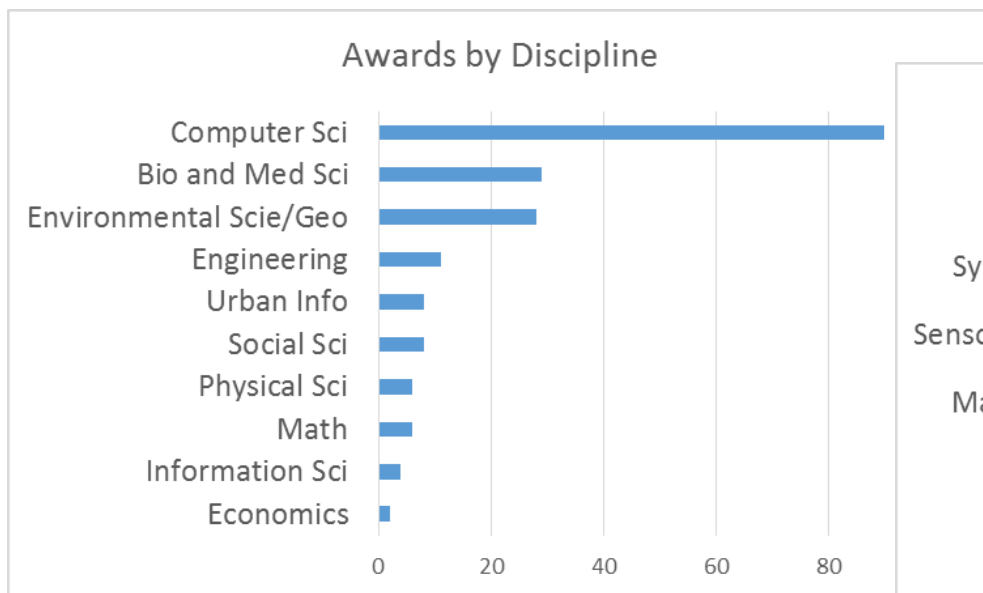
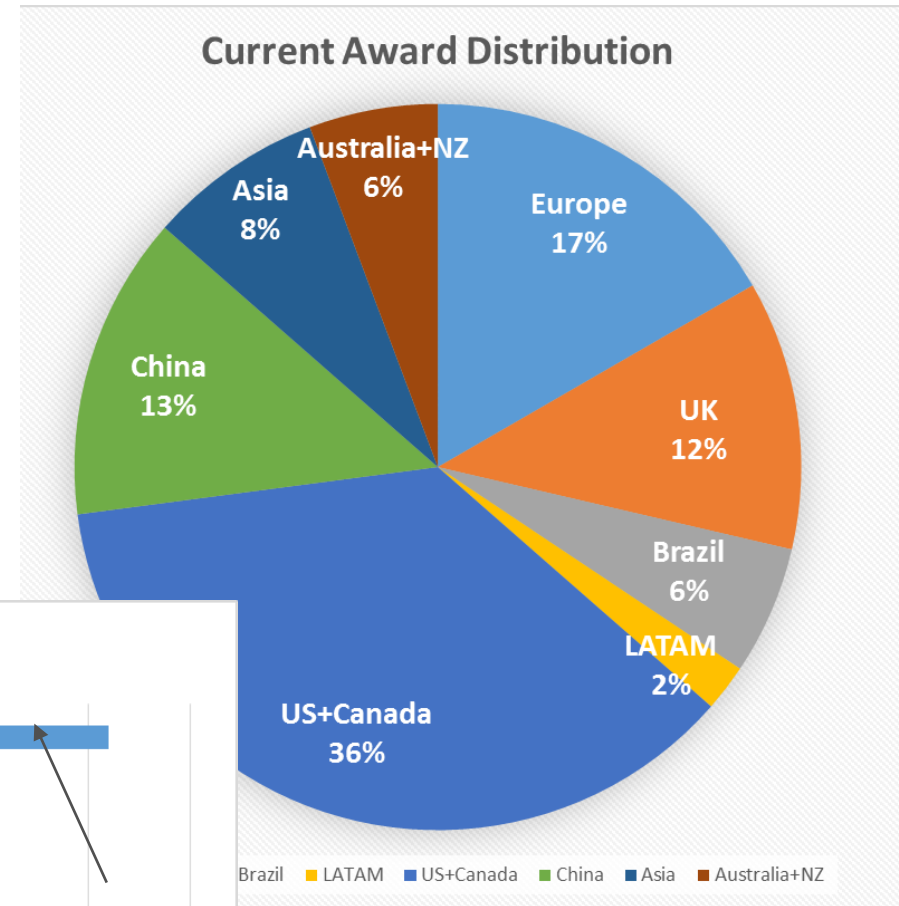
- Wu Feng, Virginia Tech modified Genome Analysis ToolKit (GATK) to use Hadoop Map Reduce (via HDInsight on Azure)



- N. Mohamed, H. Lin, and W. Feng, "Accelerating Data- Intensive Genome Analysis in the Cloud," in Proceedings of the 5th International Conference on Bioinformatics and Computational Biology (Honolulu, Hawaii, March 2013), pp. 297–304

Supporting Science: MS Azure Research Grants

- 300 awards made as of Sept 2014
- Awards made 6 times a year
- Over 700 applications so far
- Each award is \$40K in Azure

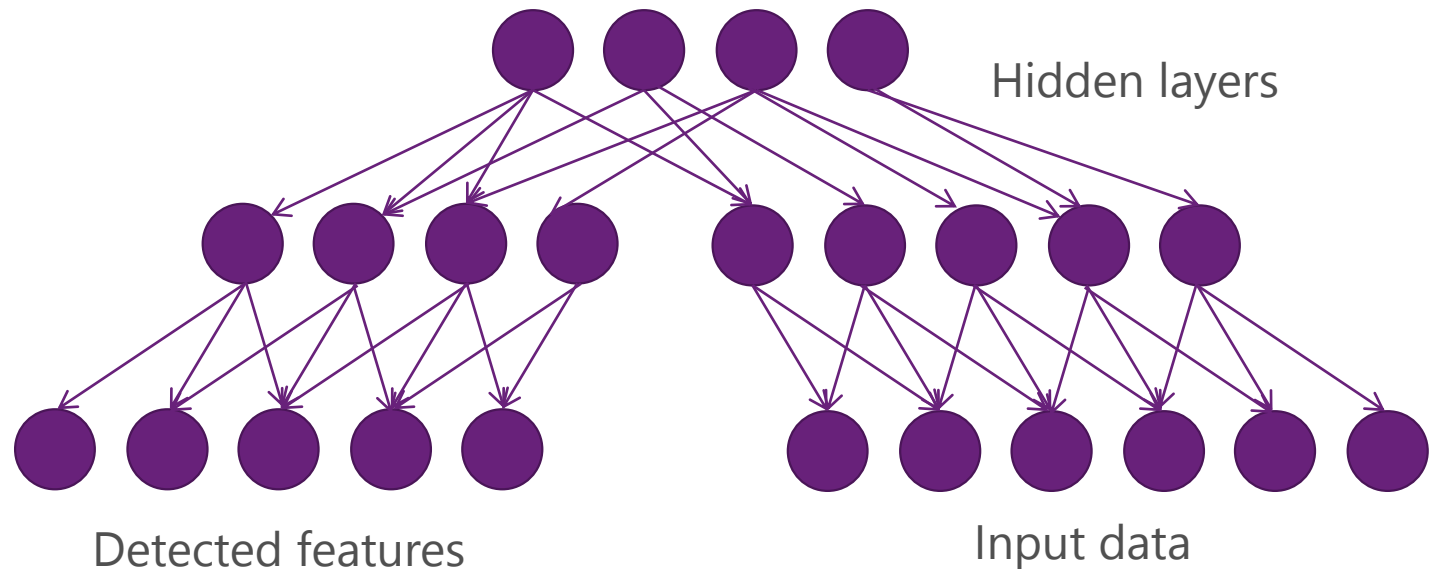
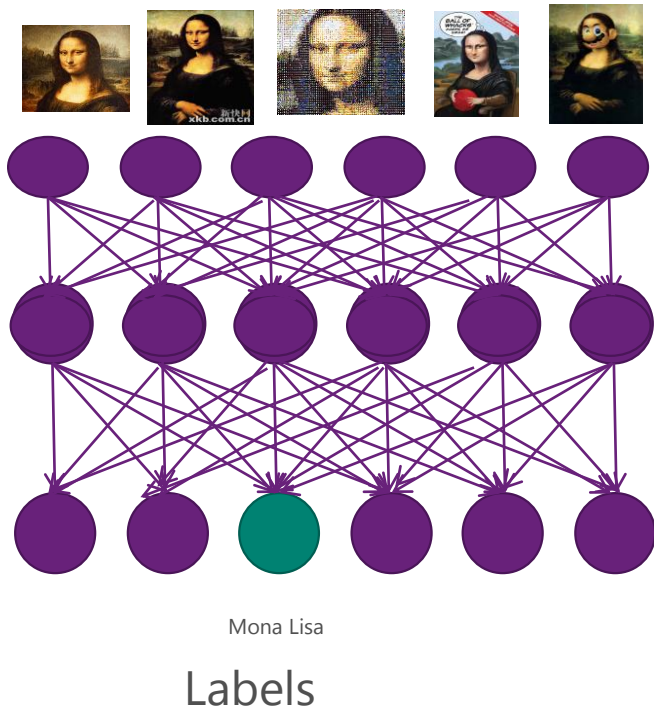


37% of these are genomics

The Machine Learning Revolution

- Big data and massive parallelism change the game.
 - Supervised Machine Learning - inferring knowledge from labeled training data
 - Unsupervised – finding the hidden structure in data without labels

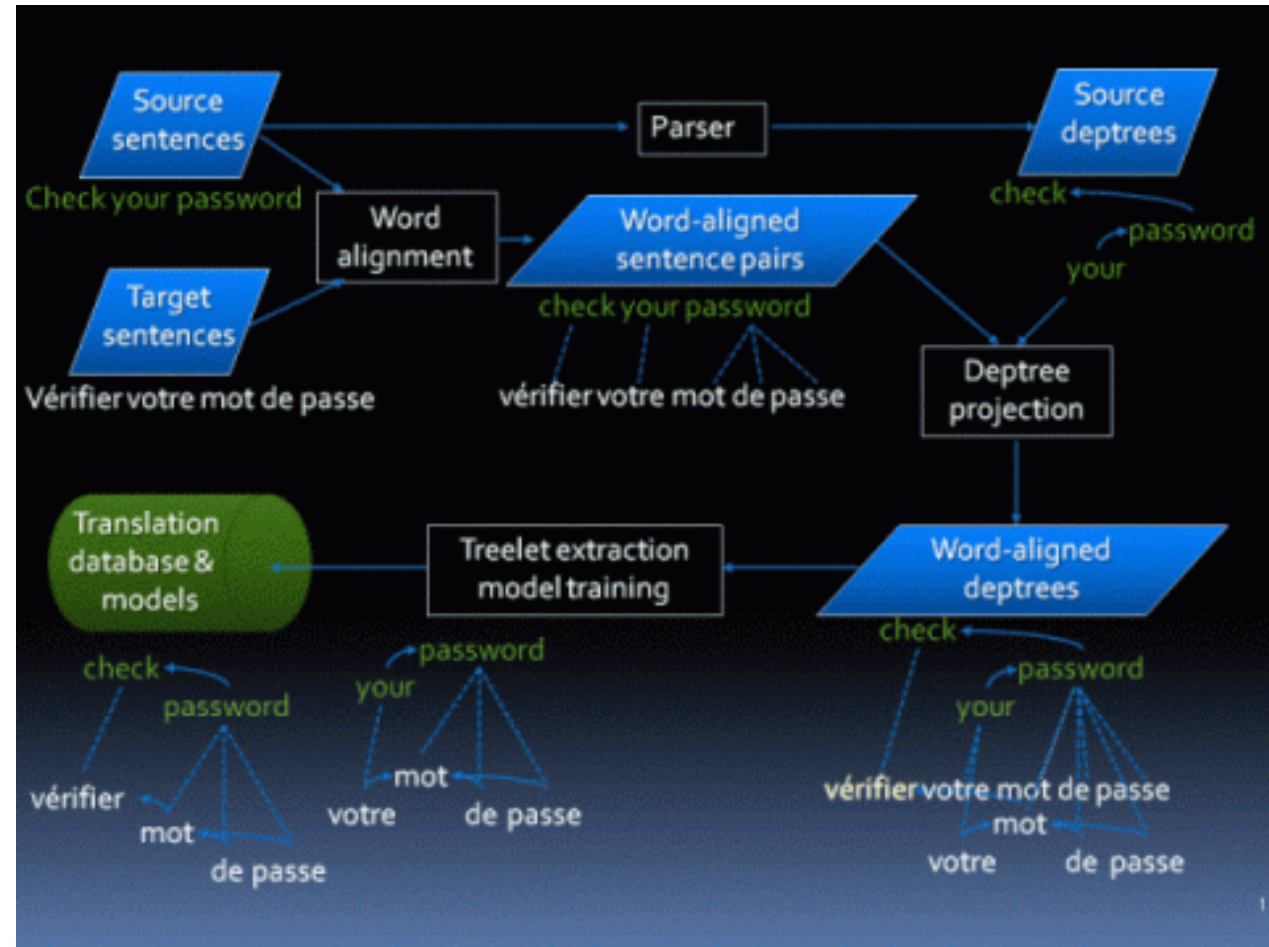
Inputs (training data)



Machine Learning in the Cloud

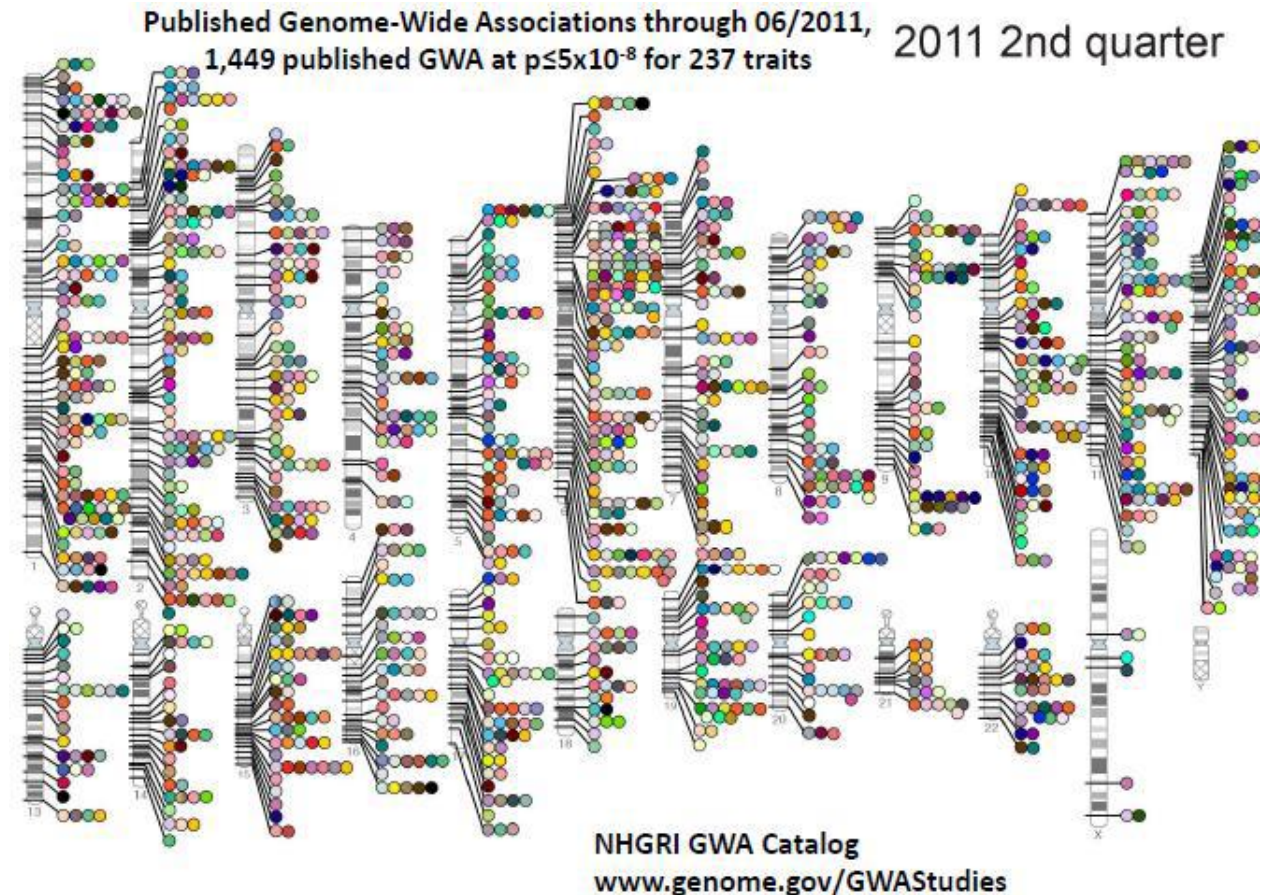
Natural Language Translation

- Given a sufficiently large collection of translated text we can “learn” to translate.
- Bing and Google have fairly good on-line translators
- Both syntax-based and phrase-based statistical machine translation
- Other applications
 - N-grams for query completion
 - “I can’t get no ...”
 - ESL grammar assistant
 - Generate a summary of a text
- Now Skype voice-to-voice real time translation.



Big Data Analytics in Medicine

- The Genetic Causes of Disease (David Heckerman)
 - Wellcome Trust for a GWAS for a large population
 - Looking for causes for seven common diseases (bipolar, r. arthritis, coronary, hypertension,)
 - Confounding is a problem. Needed a new algorithm.
 - Ran on Azure cloud using 35,000 cores in 3 weeks.



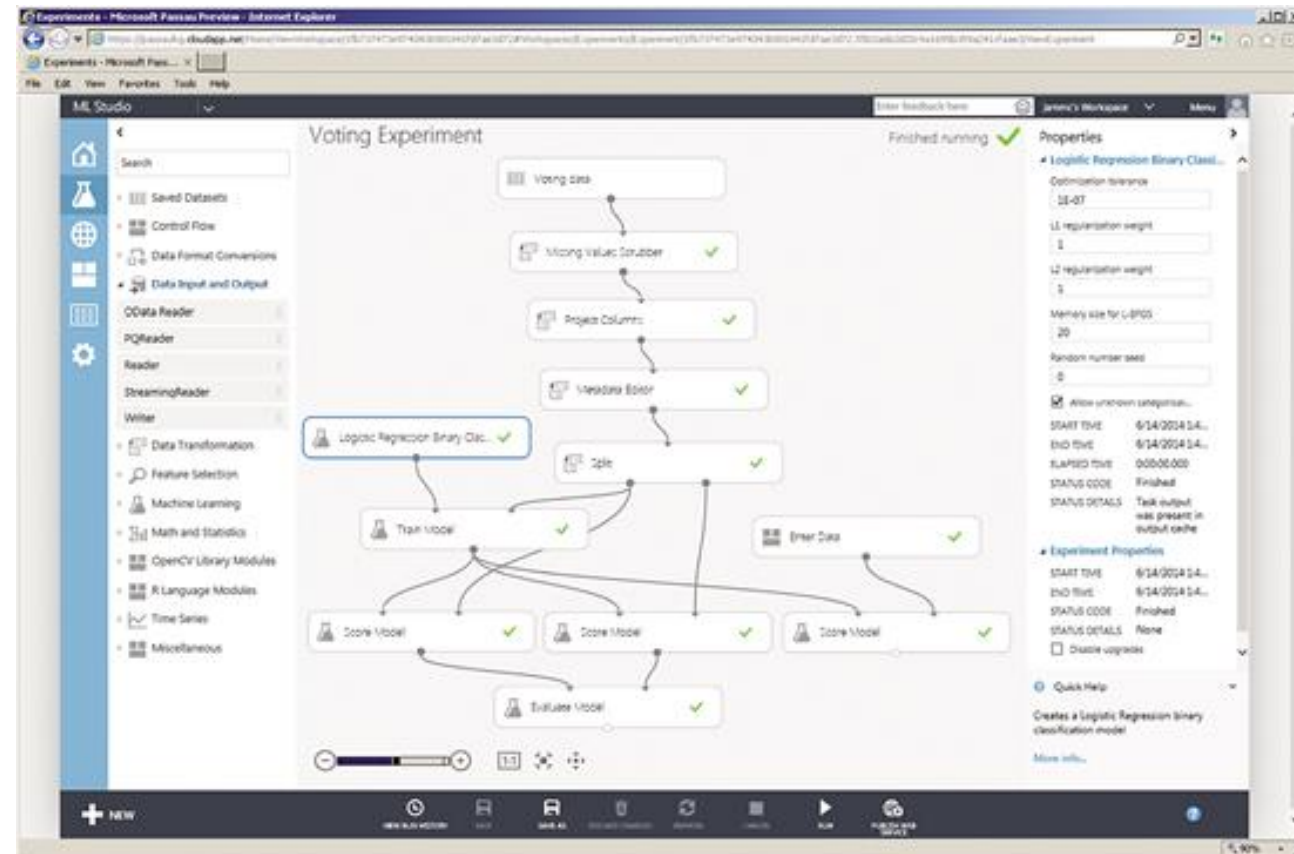
Deep Learning

- Deep neural network concepts pioneered by Geoffrey Hinton
 - *Building High-level Features Using Large Scale Unsupervised Learning*, Quoc V. Le, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg S. Corrado, Jeffrey Dean, and Andrew Y. Ng
 - Google cluster of 16,000 cores with a model of 1 billion connections, 10 million unlabeled images
- Microsoft Research demonstrates Project Adam
 - Large scale deep NN cluster
 - Cortana dog breed recognizer



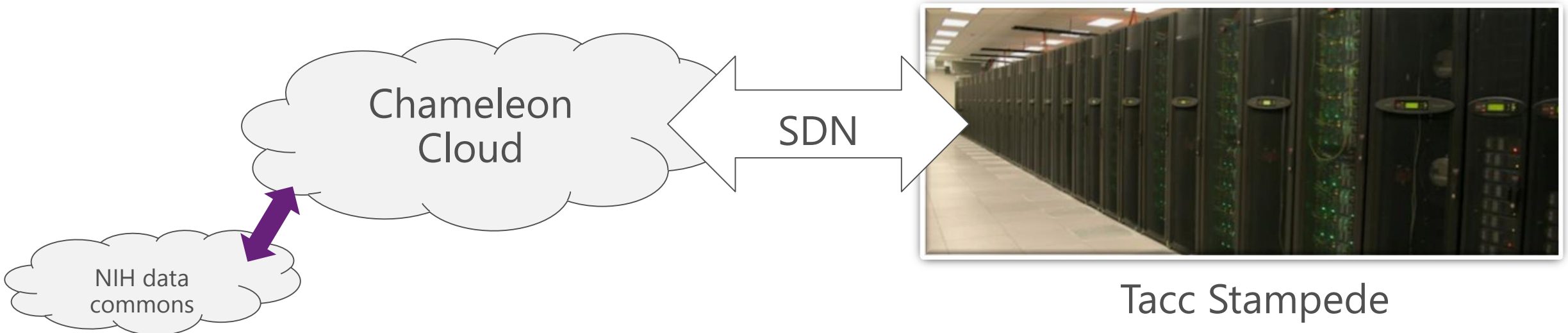
Conclusion: The Next Challenges

- Make Chameleon the place for large scale ML experiments in Science
- Make it ease to do the easy stuff.
- Make it programmable!
 - Azure ML shows you how it can look
- But there is much more ...



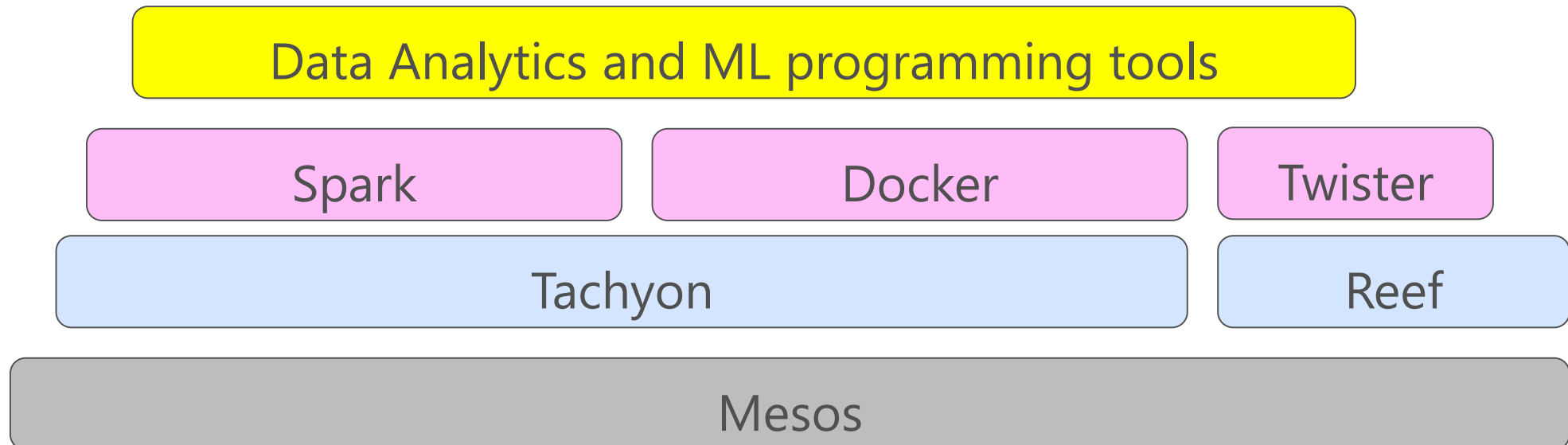
The HPC Data Analytics Challenge

- Extend the cloud seamlessly to the Supercomputer?
 - Supercomputer has vastly superior interconnect and CPUs to support data intensive parallel algorithms
 - Supercomputer is a perfect backend for cloud if...
 - Use SDN to extend TACC network to Chameleon.
 - Build a Tachyon and Spark that screams on Stampede
 - Migrate cloud built analysis transparently to Stampede for batch execution
 - Build efficient data staging pipeline from other sources to chameleon to stampede.



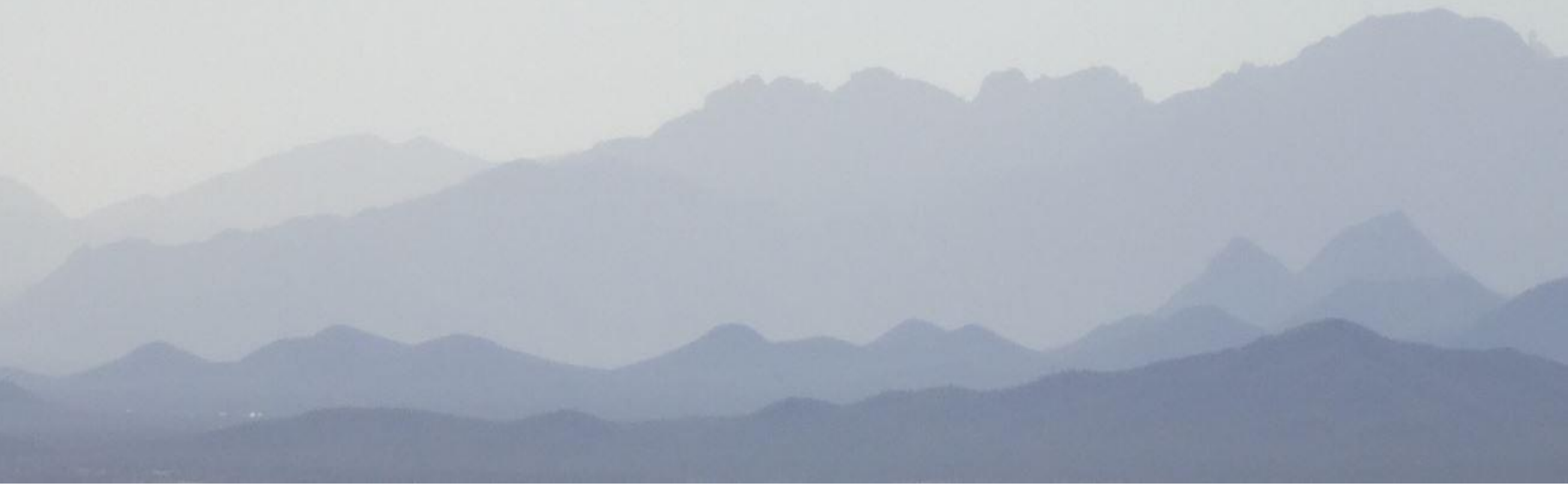
Raise the Abstraction Level for Data Analytics

- Azure ML and Scikit-Learn are fantastic ... but not parallel
 - Yarn and Reef possible. Spark and Twister essential for scalable iterative algorithms.
 - AMPLab stack, Apache stack are right directions
- The problem with rapid deployment of VMs
 - Straight VM deployment is too slow for science experiments
 - Docker + Kubernetes looks very good!
- An experimental Cloudlab stack?



Thanks

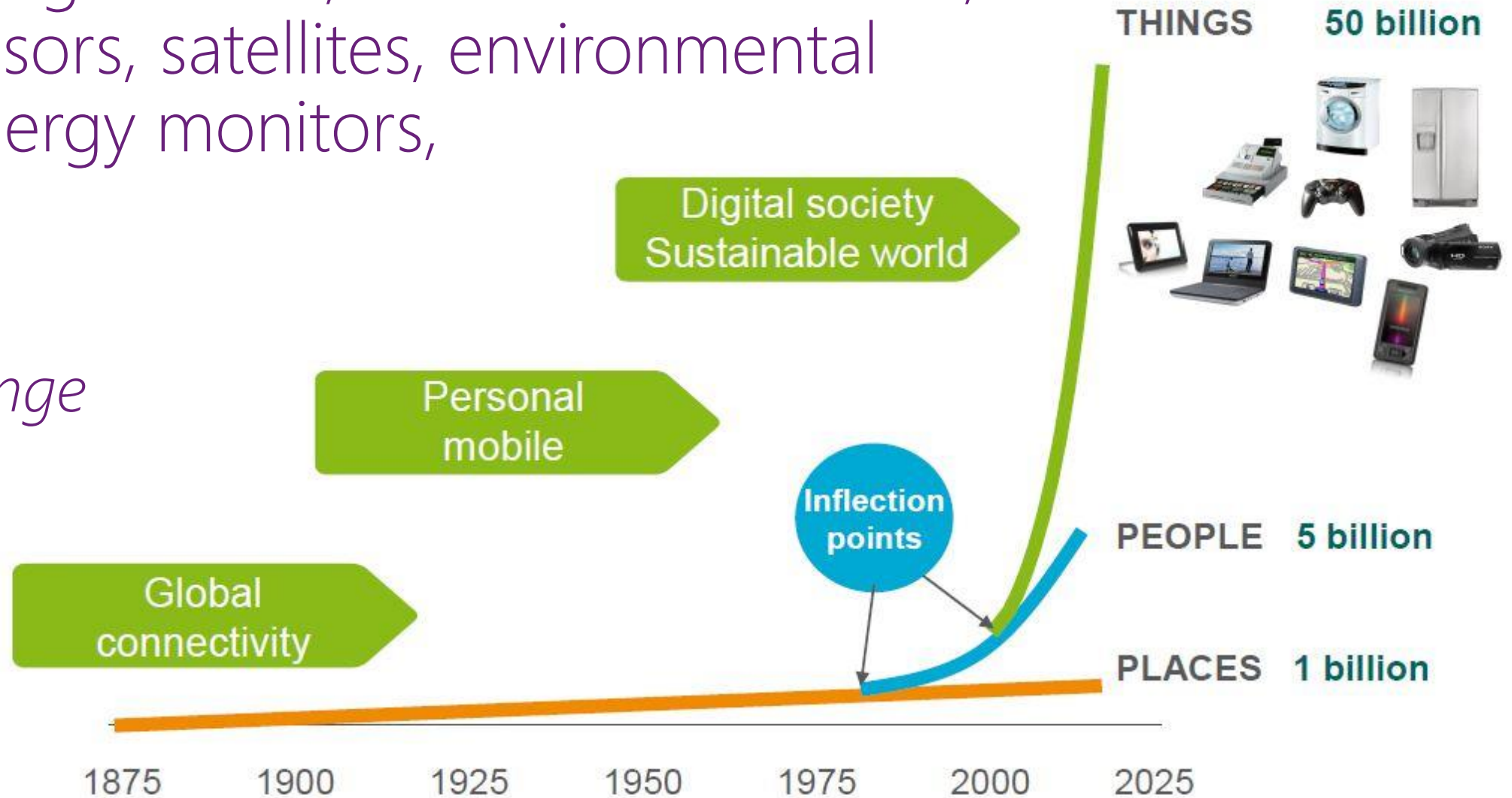
Questions?



The Internet of Things

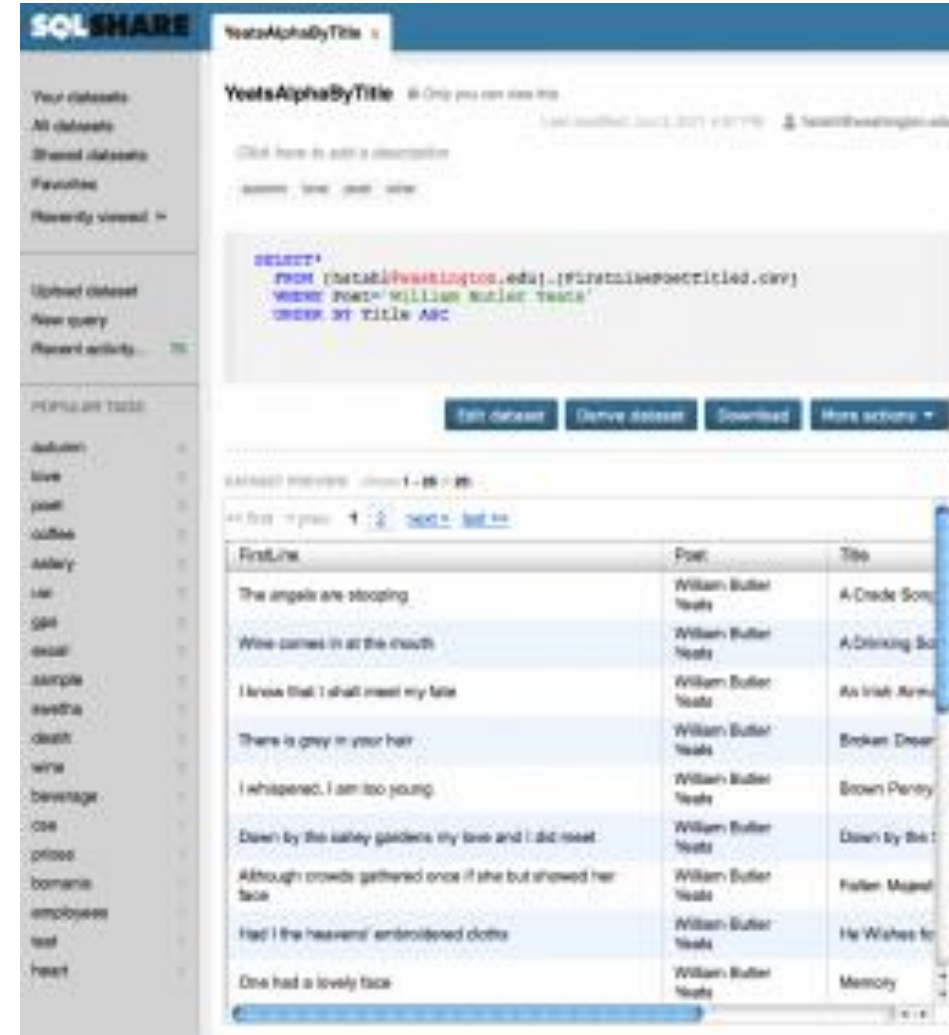
Cars, building sensors, medical instruments, factory sensors, satellites, environmental sensors, energy monitors, robots,

This will change everything



Managing the Data.

- SQL Share: Database-as-a-Service for Long Tail Science
 - Bill Howe, University of Washington. To simplify the challenge of managing research data The UW e-Science Institute, has built a cloud-based relational data sharing and analysis platform called SQLShare that allows users to upload their spreadsheet data and immediately query it using SQL
- DataUP: curating spreadsheet data
 - California Digital Library and Microsoft.
 - Web tool and excel plugin.



The screenshot shows the SQL Share web interface. On the left is a navigation sidebar with options like 'Your datasets', 'Upload dataset', and 'Popular tags'. The main area displays a dataset named 'YeatsAlphaByTitle' with a SQL query: `SELECT * FROM (dataset@uwashington.edu).(firstnamesocritics.csv) WHERE Poet='William Butler Yeats' ORDER BY Title ASC`. Below the query is a table of results with columns 'FirstLine', 'Poet', and 'Title'. The results list several lines of poetry by William Butler Yeats, such as 'The angels are stooping' and 'Wine comes in at the mouth'.

FirstLine	Poet	Title
The angels are stooping	William Butler Yeats	A Cradle Song
Wine comes in at the mouth	William Butler Yeats	A Drinking Song
I know that I shall meet my fate	William Butler Yeats	An Irish Air
There is grey in your hair	William Butler Yeats	Broken Tower
I whispered, I am too young	William Butler Yeats	Brown Penny
Down by the salley gardens my love and I did meet	William Butler Yeats	Down by the Salley Gardens
Although crowds gathered once if she but showed her face	William Butler Yeats	Fallen Maudslowi
Had I the heavens' embroidered cloths	William Butler Yeats	He Wishes for the Changeling
One had a lovely face	William Butler Yeats	Memory

Community Data Collections

- Many Examples

- Genomics, Astronomy, Geosciences, NEON, IPlant, earthCube

- The Challenge: sustainability

- Financial – providing data to public is expensive
 - Disk, curator, format conversions, indexing, etc.
- Who pays? Government? Subscription?
 - Azure data market

