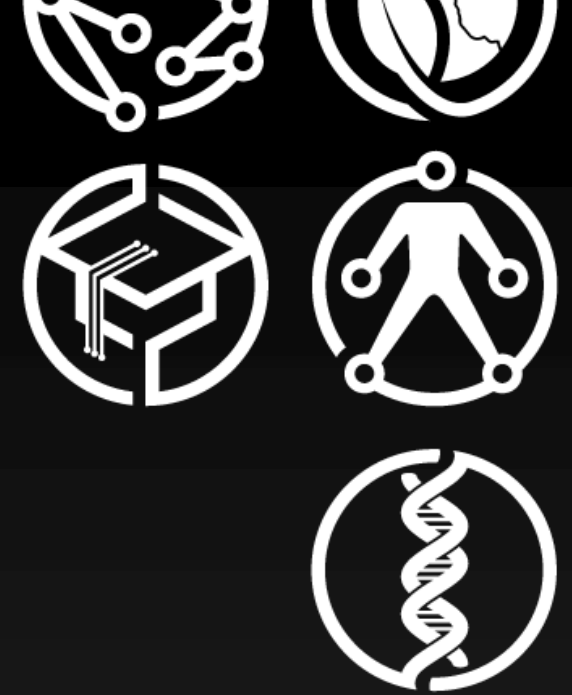


The Changing Nature and Uses of Data

Jim Pinkelman
Senior Director
jimpi@microsoft.com



Changes in Data

Effect and Behavior

Challenges and Needs

Data Size and Speed are Growing



Entire sequence of DNA for the human body, consists of around 3 billion of these base pairs.

The human genome requires ~750 megabytes of storage



Large Hadron Collider

150 million sensors delivering data 40 million times per second.

Data flow: ~700 MB/sec
~15 PB/year

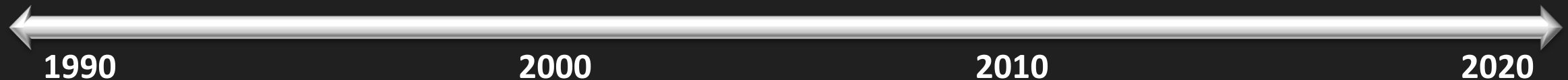
1000's of scientists around the world; Institutions in 34 different countries:



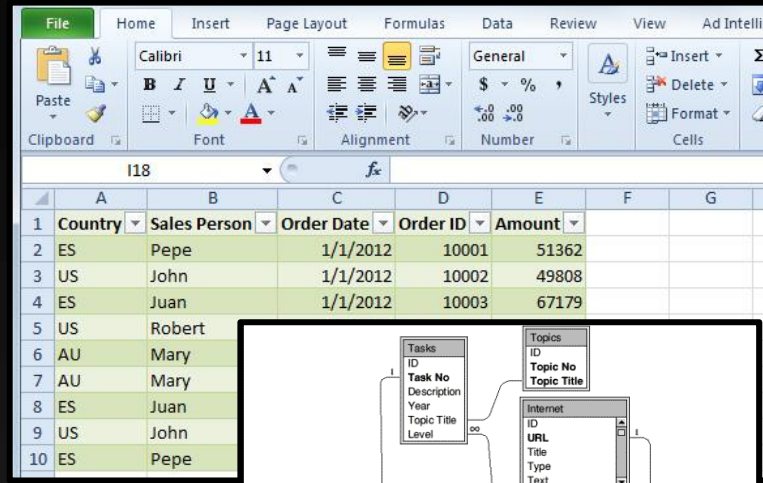
Thousands of small antennas spread over a distance of more than 3000km.

Data flow: ~60 GB/sec
1 Million PB/day

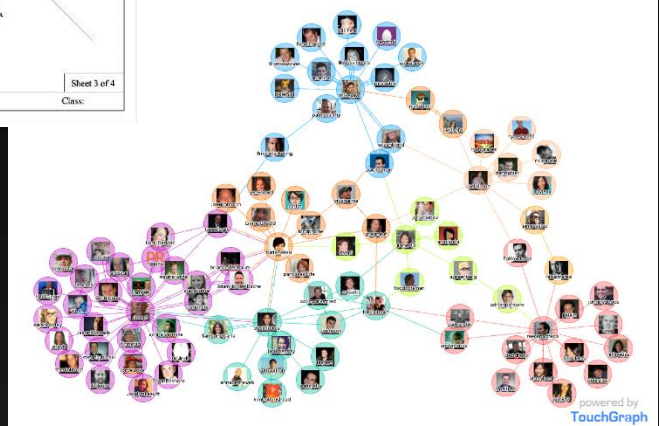
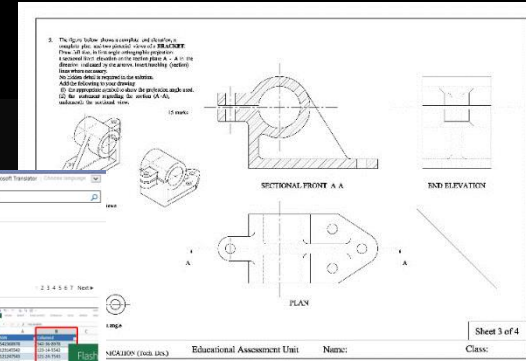
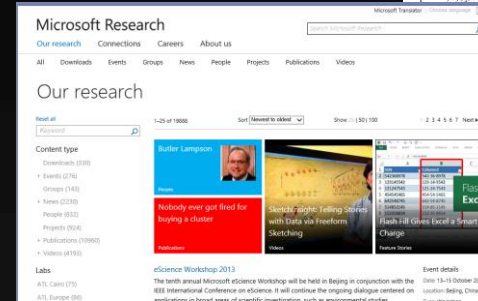
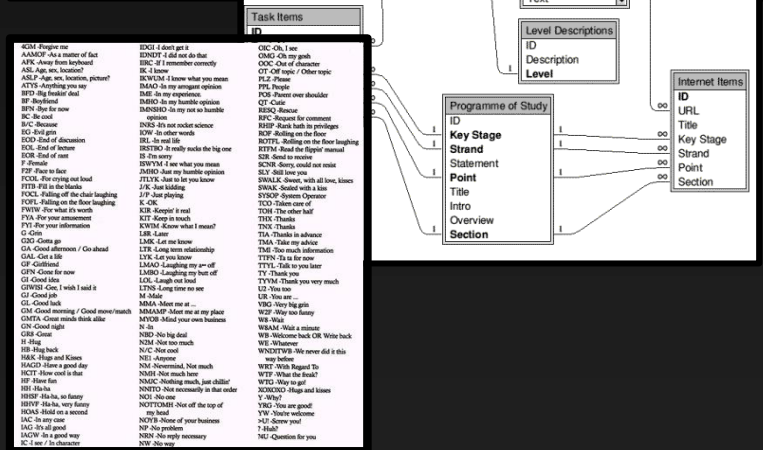
The SKA supercomputer will perform 10^{18} operations per second ~ 100M PCs



Data Complexity is Increasing



	A	B	C	D	E	F	G
1	Country	Sales Person	Order Date	Order ID	Amount		
2	ES	Pepe	1/1/2012	10001	51362		
3	US	John	1/1/2012	10002	49808		
4	ES	Juan	1/1/2012	10003	67179		
5	US	Robert					
6	AU	Mary					
7	AU	Mary					
8	ES	Juan					
9	US	John					
10	ES	Pepe					

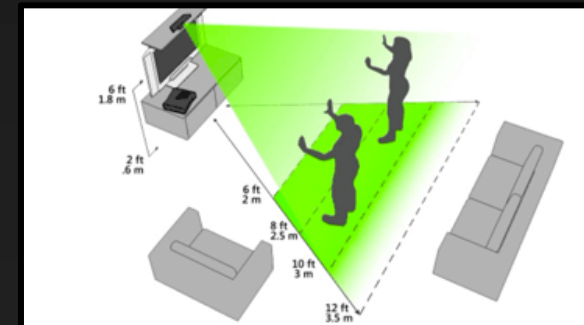
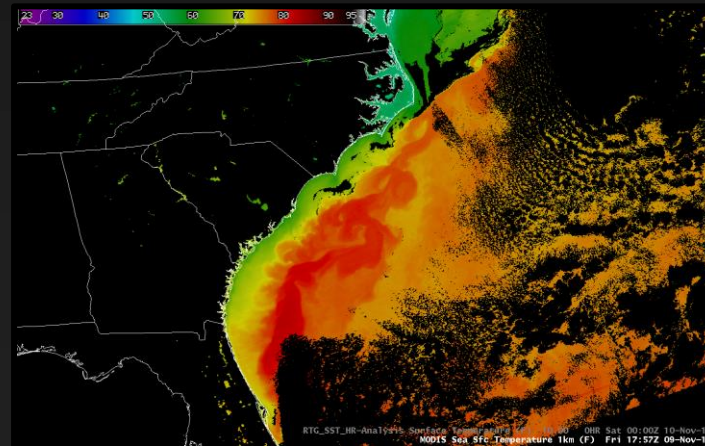
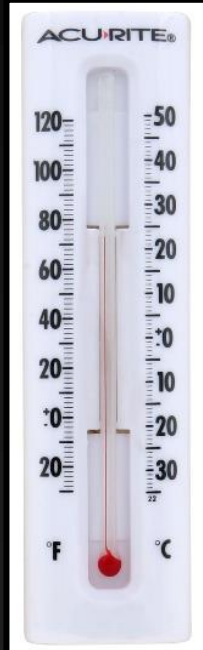


“... ‘data’ means recorded information, regardless of the form or medium on which it may be recorded, and includes writings, films, sound recordings, pictorial reproductions, drawings, designs, or other graphic representations, procedural manuals, forms, diagrams, work flow charts, equipment descriptions, data files, data processing or computer programs (software), statistical records, and other research data.”

The National Institutes of Health (NIH) Grants Policy Statement

http://grants.nih.gov/grants/policy/nihgps_2012/nihgps_ch8.htm

Data Sensors have become Digital



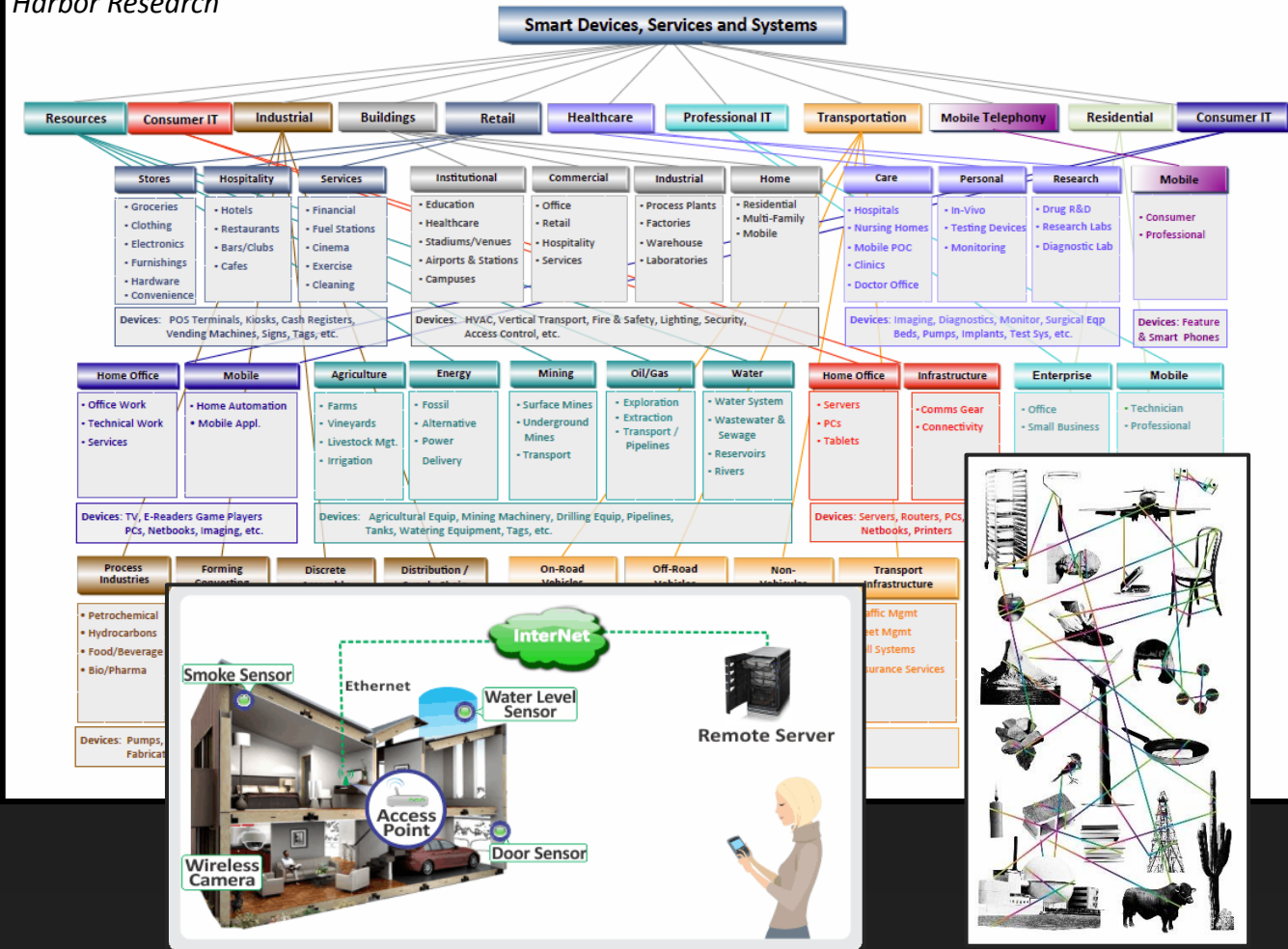
“... vision of processing power so distributed throughout the environment that computers per se effectively disappear”

*Adam Greenfield
Head of design and UI design, Nokia*

Smaller, Inexpensive, Wireless, Sophisticated, Abundant, **Digital**

Data can now be transmitted Wirelessly

Harbor Research



“ ... not only "hard sensors" that track physical attributes such as light, heat, pressure and motion, but also "soft sensors" such as a user's calendar, social network activity and Web browsing habits. “

"What context awareness does is collect all of that, some of which is up-to-the-minute on the physical sensors and some of which is accumulated incrementally over a long expanse of time through these soft senses, to create devices that really anticipate your needs“

Intel CTO Justin Rattner

Distributed, Connected, Flexible, Power Efficient, **Internet of Things**

The Cloud is Available

Omnipresent Services

- Uploading data
- Download commands
- Streaming signals
- Network between Devices

Compute & Storage Elasticity

- Lower barriers to adoption
- Lower barriers to scaling
- Lower overheads

Accelerates Collaboration

- Sharing data
- Sharing algorithms
- Co-authoring
- Reproducible Research



Changes in Data

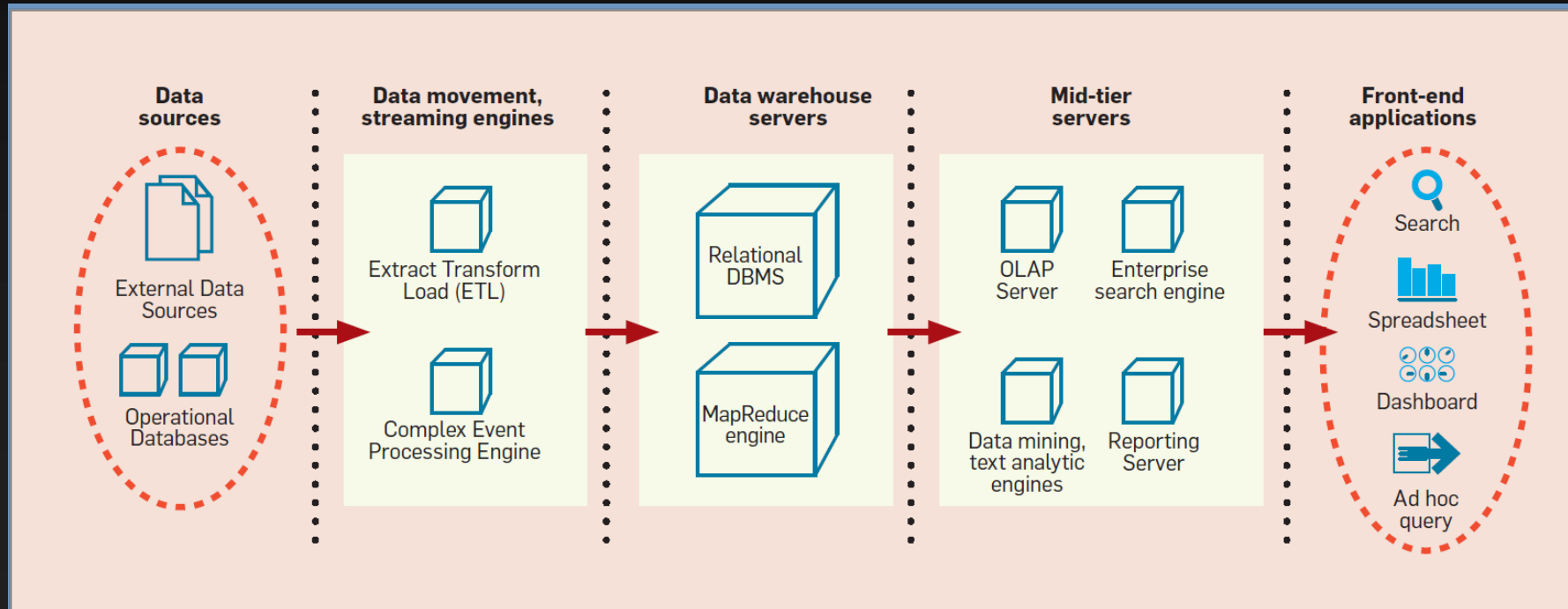
Effect and Behavior

Challenges and Needs

Uses of Analytics continues to Grow

Business Intelligence in Industry

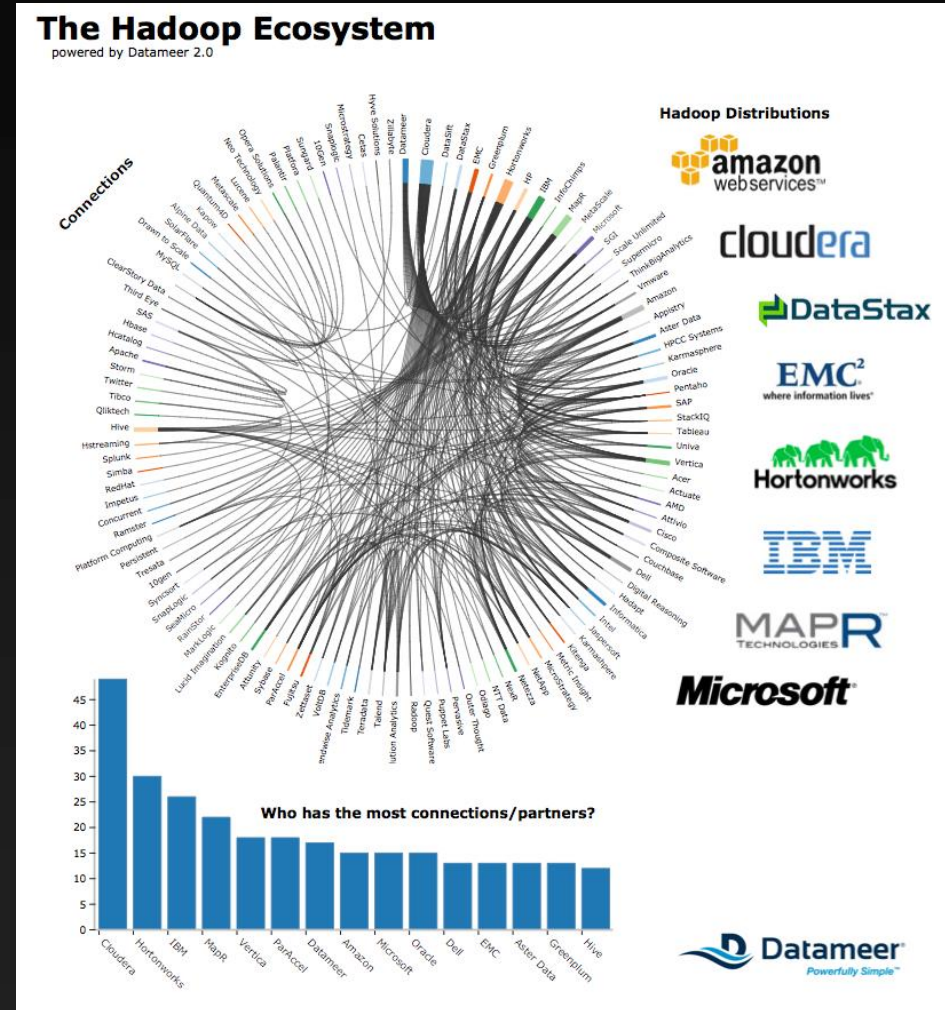
1. More data; Incorporation of External Data Sources
2. Emergence Near Real Time Analytics
3. Less structured data and non-RDBMS 'Data Warehouses'



Unstructured Data is Valuable

MapReduce -> Hadoop

- Targeting advertisements
- Fraud detection
- Financial modeling
- Business analytic
- Predicting markets
- Social network analysis
- Audience sentiment

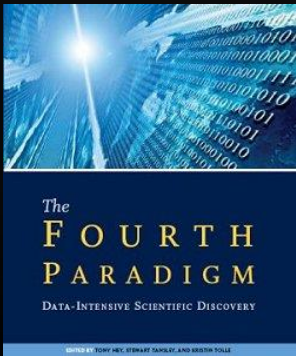
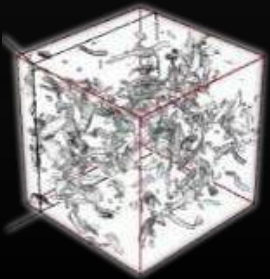


<http://www.datameer.com/blog/uncategorized/the-hadoop-ecosystem-visualized-in-datameer.html>

The Nature Of Research is Changing



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



Experimental

Thousand
years ago

*Description of natural
phenomena*

Theoretical

Last few
hundred years

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

Computational

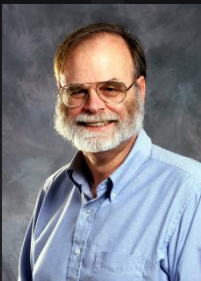
Last
few decades

*Simulation of
complex phenomena*

Data-Intensive

Today and the Future
*Unify theory, experiment
and simulation with large
multidisciplinary data*

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



Jim Gray

DNA and The Tree of Life

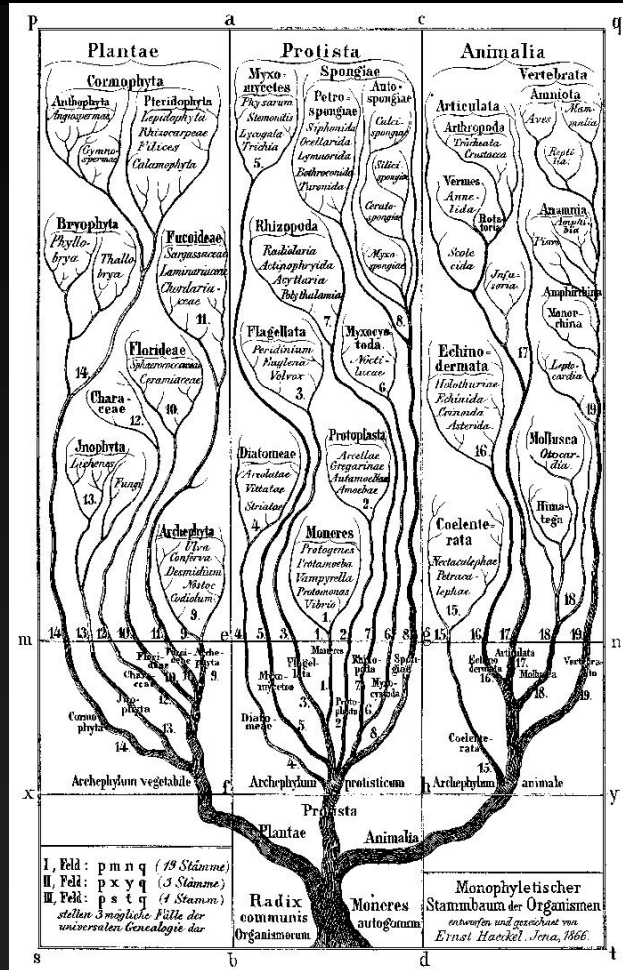
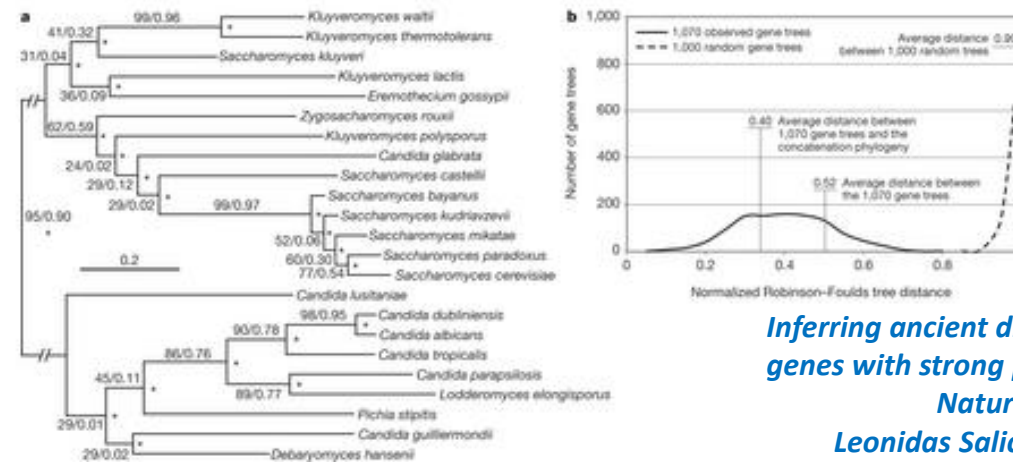


Figure 1: The yeast species phylogeny recovered from the concatenation analysis of 1,070 genes disagrees with every gene tree, despite absolute bootstrap support.



... nearly 150 years later, scientists have vast amounts of data with which to build so-called phylogenetic trees, the modern version of Mivart's structure. Advances in DNA sequencing technology and bioinformatics enable them to compare the sequence of hundreds of genes, sometimes entire genomes, among many different species, creating a tree of life more detailed than ever before.

Emily Singer, *Simons Foundation*

<https://www.simonsfoundation.org/features/science-news/a-new-approach-to-building-the-tree-of-life/>

<http://www.nature.com/journal/v497/n7449/full/nature12130.html>

Science is now Data-Intensive

- Extremely large data sets
 - Expensive to move
 - Domain standards
 - High computational needs
 - Supercomputers, HPC, Grids
- e.g. High Energy Physics, Astronomy*

- Large data sets
 - Some Standards within Domains
 - Shared Datacenters & Clusters
 - Research Collaborations
- e.g. Genomics, Financial*

“A paper from Microsoft Research, aptly titled ‘Nobody ever got fired for buying a cluster,’ which points out that a lot of the problems solved by engineers at even the most data-hungry firms don’t need to be run on clusters. ... there are vast classes of problems for which clusters are a relatively inefficient—or even totally inappropriate—solution.”

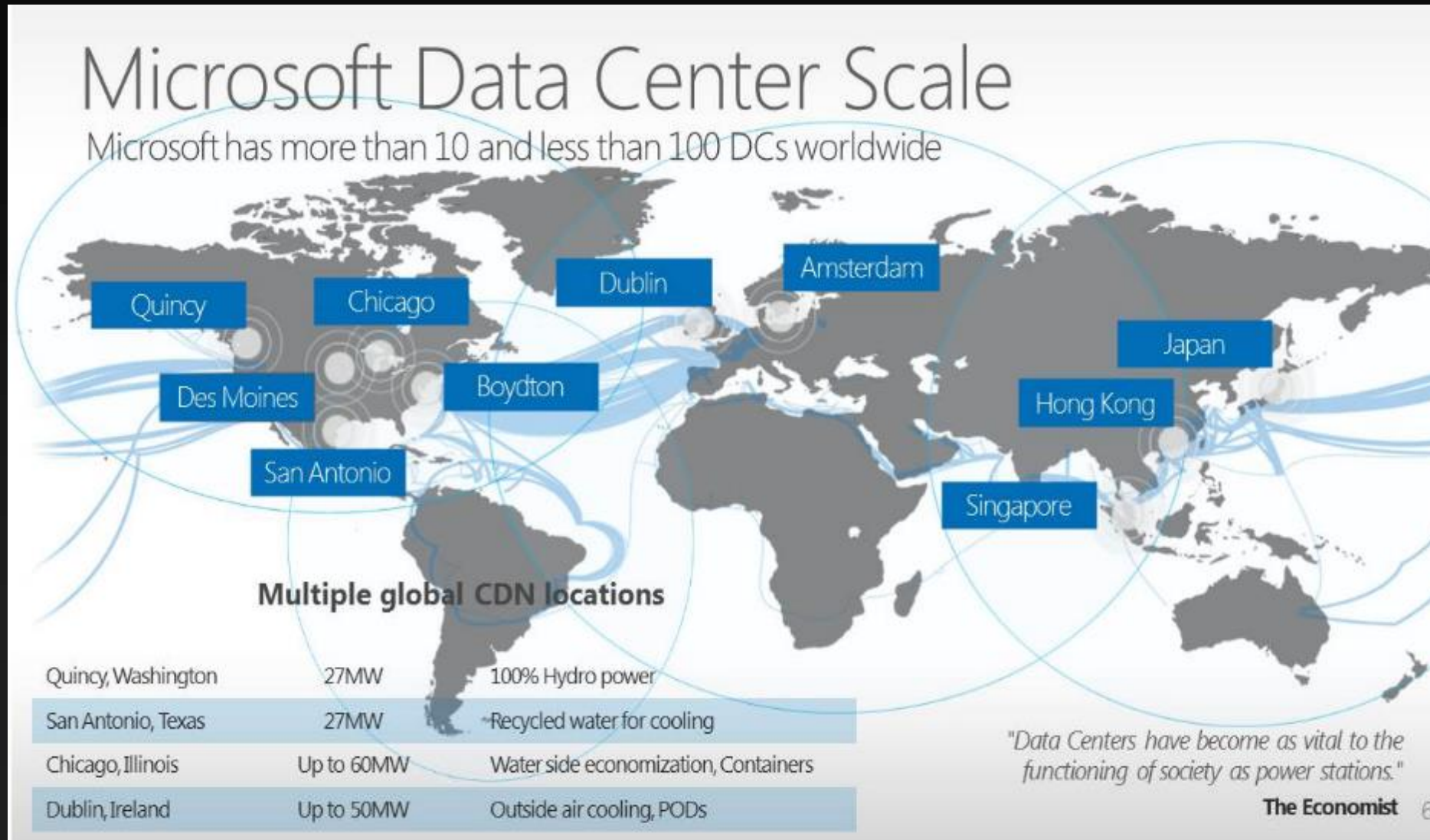
*Christopher Mims
Science and Technology Editor, Quartz*

- Medium & Small data sets
 - Flat Files, Excel
 - Widely diverse data; Few standards
 - Local Servers & PCs
- e.g. Social Sciences, Humanities*

‘The Long Tail of Science’

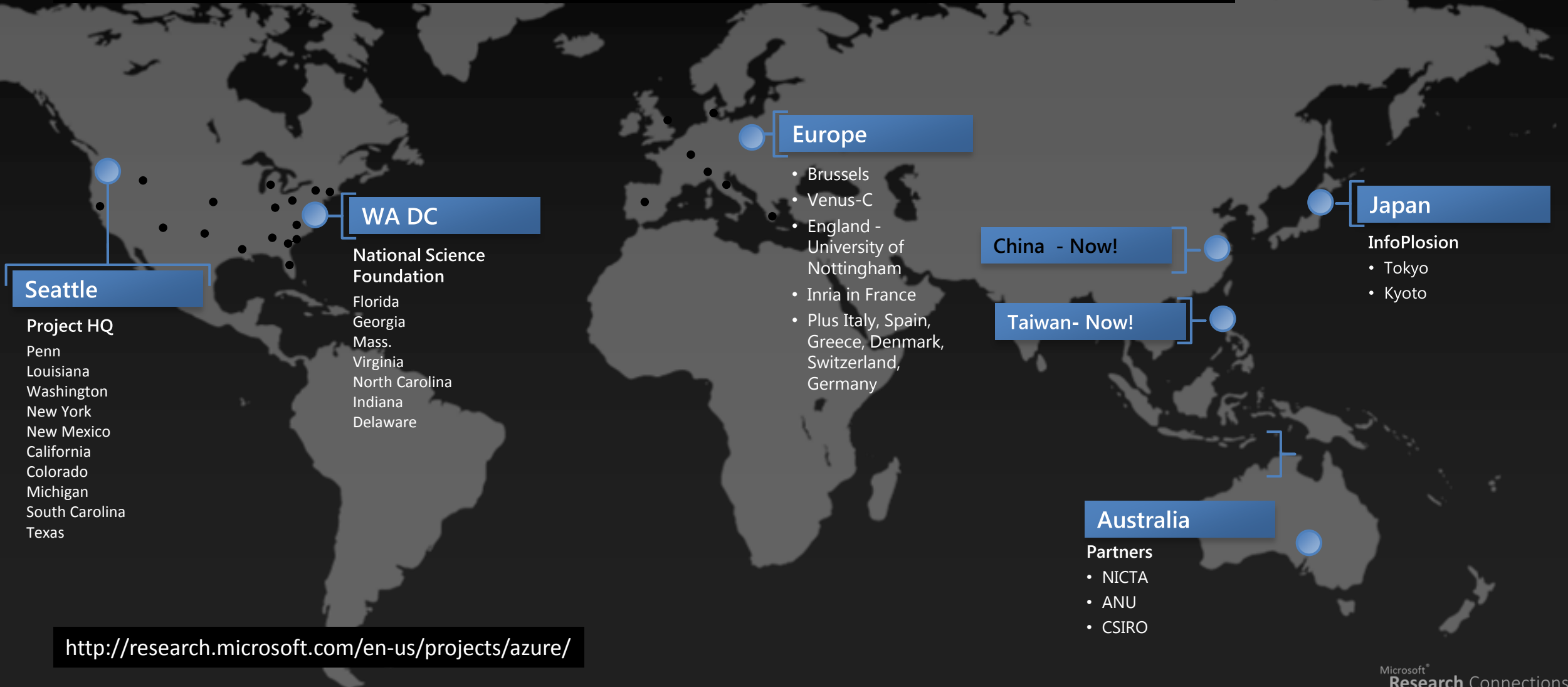
Number of Researchers

Industry is building out massive Infrastructure



Microsoft Cloud Research Engagement Project

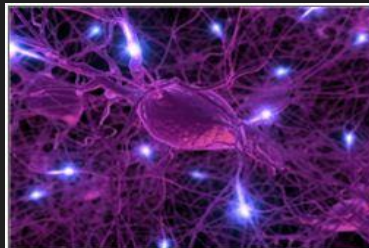
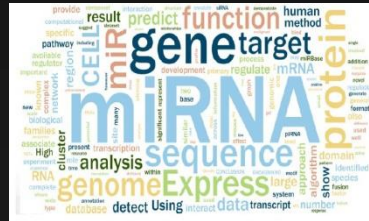
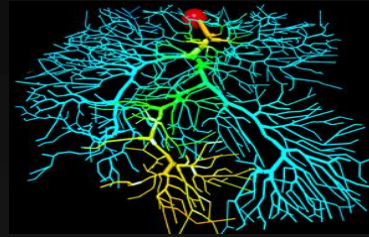
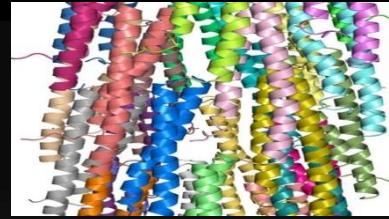
Worked with international funding agencies to grant access to cloud resources to researchers. 90 projects world wide.



<http://research.microsoft.com/en-us/projects/azure/>

Bioinformatics Research on Windows Azure

- Protein Folding
 - The **University of Washington** is studying the ways proteins from salmonella inject DNA into cells. Used 2000 concurrent cores.
- Joint Genetic and Neuroimaging Analysis
 - France's premier research institute **INRIA** is using 1000 cores of Azure to study large cohorts of subjects to understand links between genetic patterns and brain anomalies.
- Comparative Genomics
 - Researchers at the **University of North Carolina** Charlotte are doing large scale operon prediction using Windows HPC Scheduler on Azure using 300 cores to do BLAST analysis. Used 1,000,000 hours.
- Drug Discovery
 - Researchers at **Newcastle University** in the U.K. are using Azure to model the properties (toxicity, solubility, biological activity) of molecules for potential use as drugs.
- Systems Biology
 - The **University of Trento Centre for Computational and Systems Biology** have developed an Azure based tool, BetaSIM for modeling and simulating biological systems.



Machine Learning is becoming more Applicable

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



Used an array of 16,000 processors to create a neural network with more than one billion connections. They then fed it random thumbnails of images, one each extracted from 10 million YouTube videos.

Stanford University Andrew Ng
Google fellow Jeff Dean



... turning my English into Chinese in two steps. The first takes my words and finds the Chinese equivalents, and while non-trivial, this is the easy part. The second reorders the words to be appropriate for Chinese, an important step for correct translation between languages.

Microsoft Research Rick Rashid

Collaboration and Sharing of Data is Expected and Growing



... **expects investigators to share with other researchers**, at no more than incremental cost and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work.

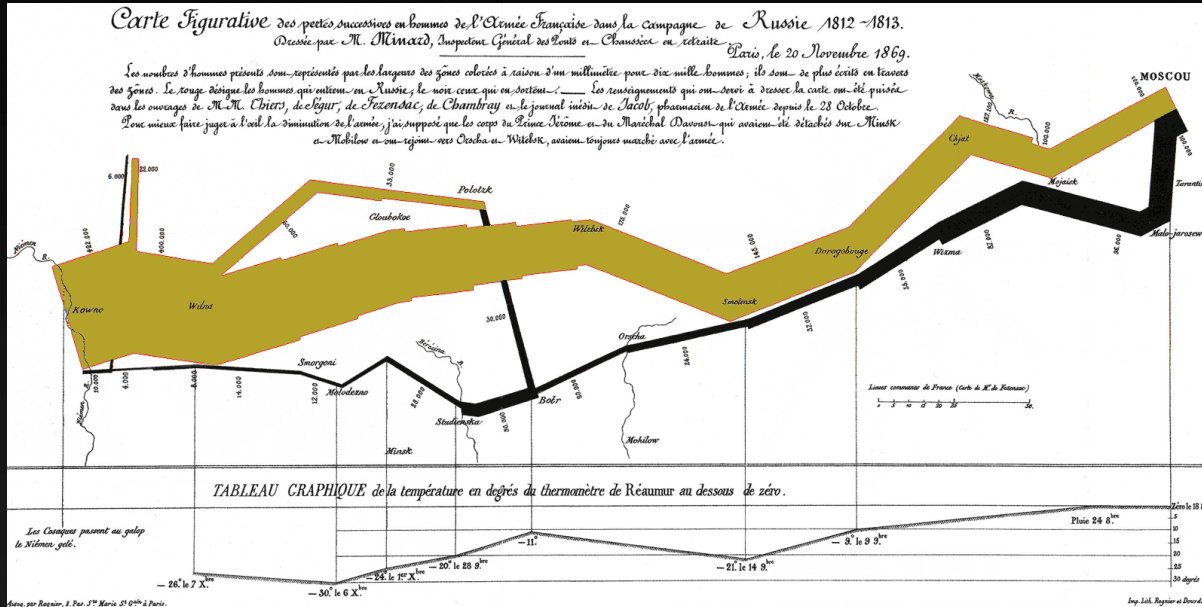


NIH reaffirms its support for the concept of data sharing. We believe that **data sharing is essential for expedited translation of research results into knowledge, products, and procedures to improve human health**. ... The NIH expects and supports the timely release and sharing of final research data from NIH-supported studies for use by other researchers.



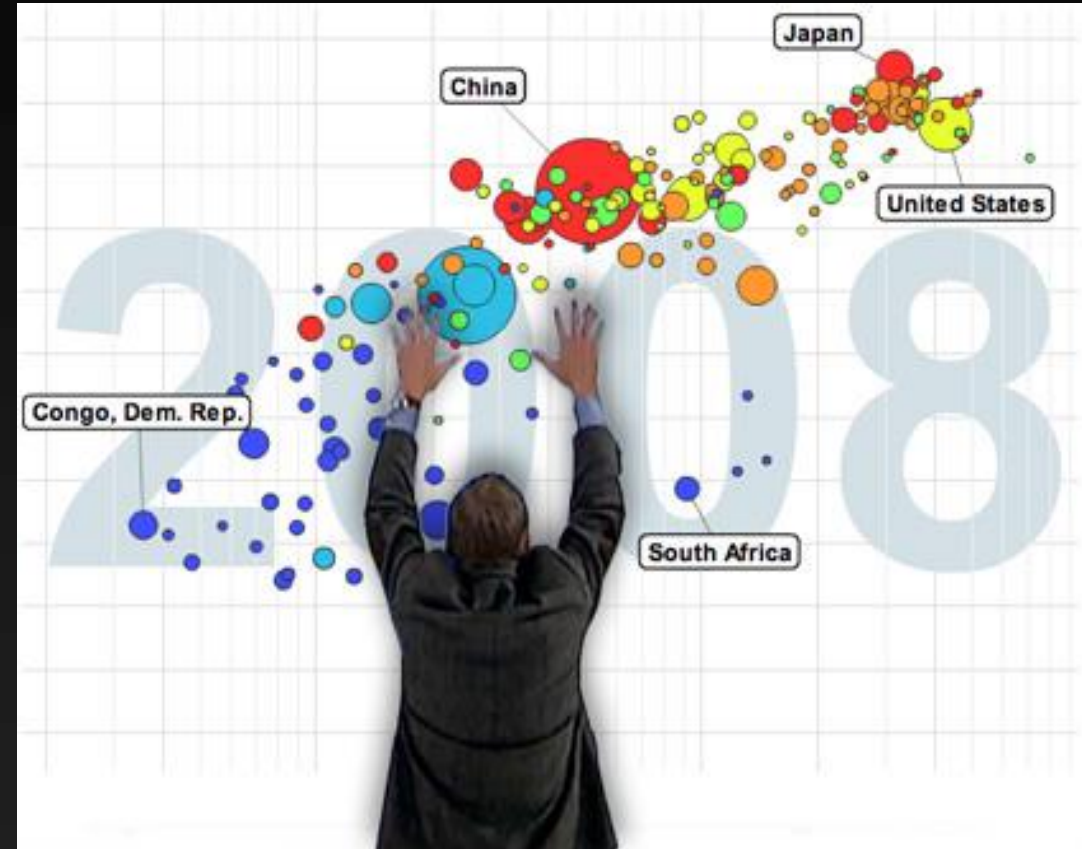
A primary goal of Data.gov is to improve access to Federal data and expand creative **use of those data beyond the walls of government** by encouraging innovative ideas (e.g., web applications). Data.gov strives to make government more transparent and is committed to creating an unprecedented level of openness in Government.

Data Visualization is leading to *Digital Storytelling*



Napoleon's March
Charles Joseph Minard

Edward Tufte: 'Probably the best statistical graphic ever drawn'



Hans Rosling
Karolinska Institute
Gapminder Foundation

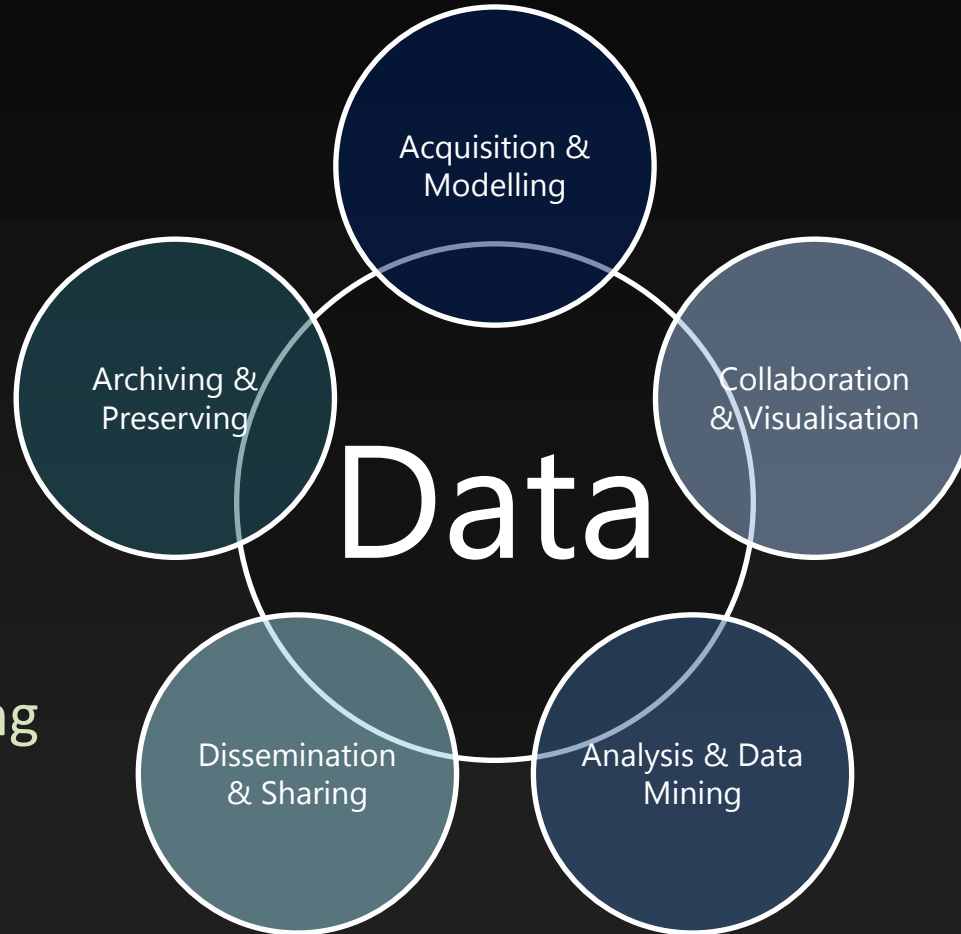
Hans Rosling on CNN: US in a converging world
<http://www.gapminder.org/videos/hans-rosling-on-cnn-us-in-a-converging-world/>

Changes in Data

Effect and Behavior

Challenges and Needs

The Data Lifecycle



Technical Challenges

Privacy & Security

Integrity

Data Portability

Communication Protocols

Metadata Standards

Data Curation and Archiving

Social & Cultural

Economic Sustainability

Trust

Institutional Agility

Talent Pipeline

Scholarly Communications

Semantic Diversity

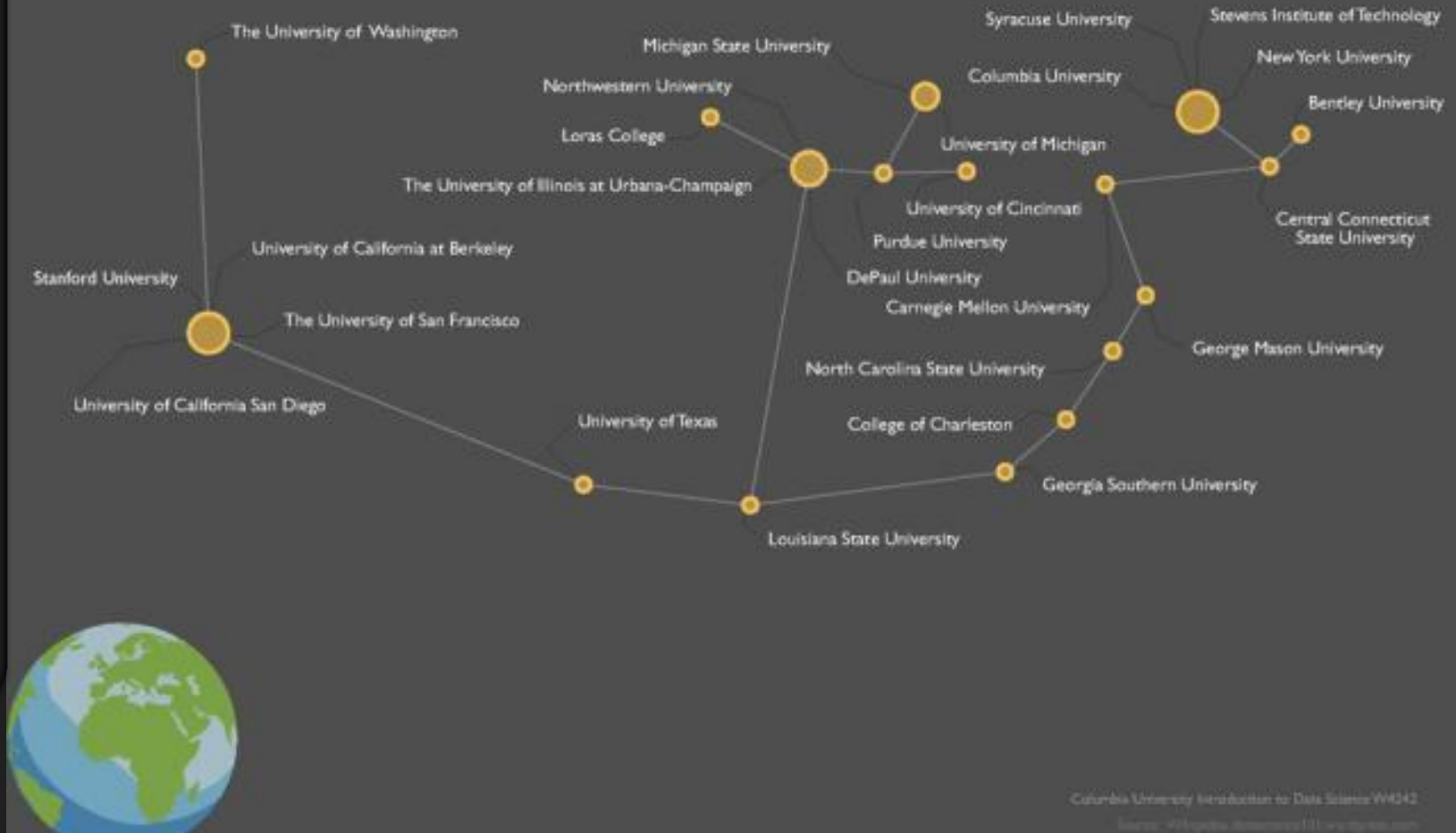
Data Science

Columbia will offer new master's and certificate programs heavy on data. The University of San Francisco will soon graduate its charter class of students with a master's in analytics. Other institutions teaching data science include New York University, Stanford, Northwestern, George Mason, Syracuse, University of California at Irvine and Indiana University.

*Claire Cain Miller
NY Times*

A Constellation Is Born

Data Science classes forming across the country



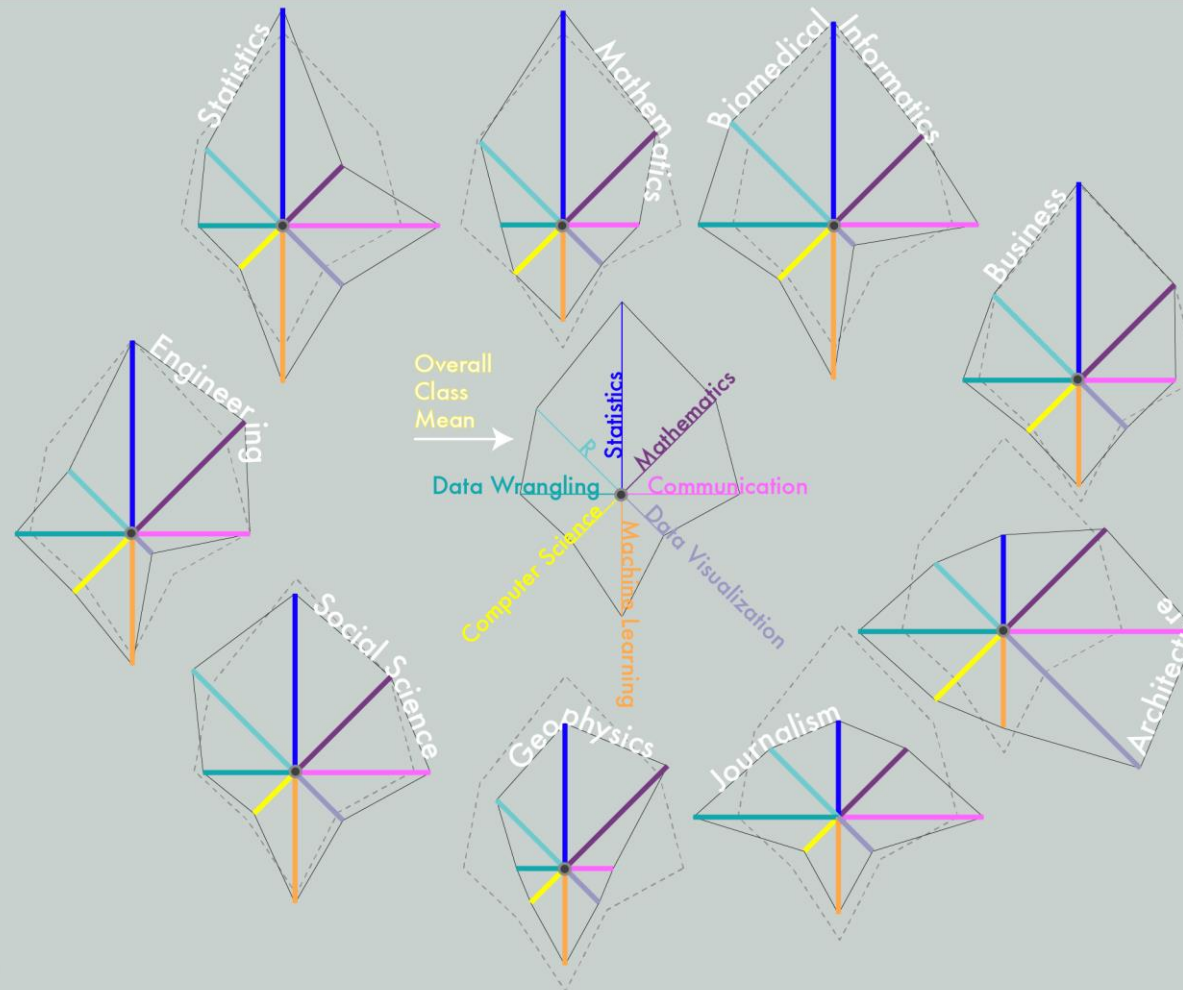
Columbia University Introduction to Data Science W4342
Source: <http://blogs.columbia.edu/entry/2013/04/11/data-science-101-what-is-it/>

<http://wikipedia.datascience101.wordpress.com/>




The Stars of Data Science

Students in Columbia's Introduction to Data Science course came from across the academic spectrum. Their skills are presented here in star charts with spokes representing their skill levels* across the data science skillset: **R**, **statistics**, **mathematics**, **communication**, **data visualization**, **machine learning**, **computer science**, and **data wrangling**. In addition to hovering in the center, the star chart of the overall class mean underlies each academic domain, so you can see students from each academic domain relative to the rest of the class. How would you compose your own intergalactic data science team?

*Skills were assessed by a survey written and administered by a subset of students in the class.



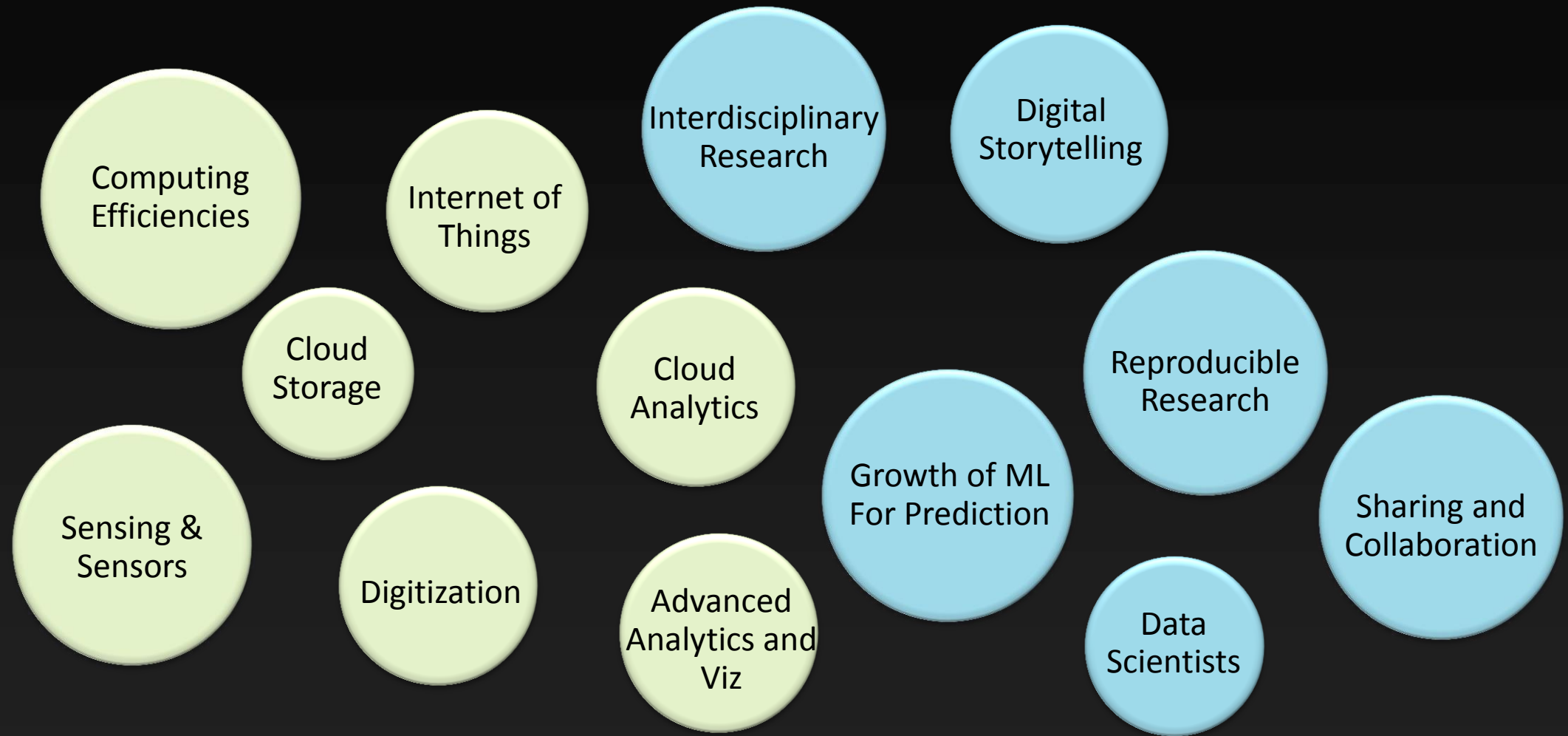
Talent - A Data Scientist?

Data Engineer 	People who are expert at <ul style="list-style-type: none">• Operating at low levels close to the data, write code that manipulates• They may have some machine learning background.• Large companies may have teams of them in-house or they may look to third party specialists to do the work.
Data Analyst 	People who explore data through statistical and analytical methods <ul style="list-style-type: none">• They may know programming; May be an spreadsheet wizard.• Either way, they can build models based on low-level data.• They eat and drink numbers; They know which questions to ask of the data. Every company will have lots of these.
Data Steward 	People who think to managing, curating, and preserving data. <ul style="list-style-type: none">• They are information specialists, archivists, librarians and compliance officers.• This is an important role: if data has value, you want someone to manage it, make it discoverable, look after it and make sure it remains usable.

What is a data scientist? Microsoft UK Enterprise Insights Blog, Kenji Takeda

<http://blogs.msdn.com/b/microsoftenterpriseinsight/archive/2013/01/31/what-is-a-data-scientist.aspx>

Real Evolution



Thank you

Jim Pinkelman
jimpi@microsoft.com

