BIOINFORMATICS

Driving force for innovation in Cloud Storage Technology

Jay Etchings
Solutions Architect, Health-Care Life Sciences
Dell Enterprise

Disclosure(s)

- #IWORK4DELL
- Data Management / Casino Analytics consultant for casino properties
- Recovery Audit Contractor for Medicare / Medicaid (CMSRAC-Payer)
- Vendor Agnostic but not without personal preference
- Solutions Consultant to Major Universities and HCLS start-ups

ASU School of Life Sciences

ASU Ira A. Fulton Schools of Engineering

The University of New Mexico

Northern Arizona University

NIH

The University of Arizona

Arizona Health Sciences Library
Quality Information for Quality Healthcare

tgen

Bright Computing
Advanced cluster management made easy

Caesars Entertainment

Dell HPC

What-How-Why & What now?

- What is this and why do I care?
  - How exactly did we get here?
    - What is the solution
      - Potential Strategies
        - Close & Questions

Jay's 60 Second Big Data History

1944 Wesleyan University Librarian estimates American university libraries were doubling in size every sixteen years. Given this growth rate, the Yale Library in 2040 will have "approximately 200,000,000 volumes, occupy over 6,000 miles of shelves, staff of over six thousand persons."

1971 Arthur Miller writes "Too many information handlers seem to measure a man by the number of bits of storage capacity."

1975 The Ministry of Posts and Telecommunications in Japan conducting Information Flow Census, tracking the volume of information circulating in Japan

1980 I.A. Tjomsland: "Where Do We Go From Here?" in which he says '**Data expands to fill the space available**'….

Belief that large amounts of data are being retained because users have no way of identifying obsolete data;

"The penalties for storing obsolete data are less than are the penalties for discarding potentially useful data."

1986 Hal B. Becker publishes "Can users really absorb data at today's rates? Tomorrow's?" in *Data Communications*.

1990 "Saving all the Bits" in *American Scientist."* Imperative to save all the bits forces us into an impossible situation:

The rate and volume of information flow overwhelm our networks, storage devices and retrieval systems, as well as the human capacity for comprehension… *(Sounds Like 3-V's?)*

What machines shall we build to monitor the data stream of an instrument, or sift through a database of recordings, propose for a statistical summary?

1996 Digital storage more cost-effective for storing data than paper according to R.J.T. Morris.

Jay's 60 Second Big Data History(cont.)

1997 Lesk publishes "How much data is in the World?" There may be a few thousand petabytes of information, production of tape and disk will reach that level by the year 2000. In only a few years we will save everything!

1998 Chief Scientist at SGI, presents a paper titled **"Big Data and the next wave of InfraStress."**

1999 Publication "Visually exploring gigabyte data sets in real time". It is the first CACM article to use the term "Big Data" (Big Data for Scientific Visualization)

2001 Laney publishes a research note titled "3D Data Management: Controlling Data Volume, Velocity, and Variety."

First use of 3-V's Volume, Variety, Velocity

2008 Swanson and Gilder publish "Estimating the Exaflood," They project that U.S. IP traffic could reach one Zettabyte by 2015 & the U.S. Internet of 2015 will be at least 50 times larger than it was in 2006

2009 Study finds that in 2008, "Americans consumed information for about 1.3 trillion hours, an average of alm

Consumption totaled 3.6 Zettabytes and 10,845 trillion words

2012 Boyd and Crawford publish "Critical Questions for Big Data".

2013 Phil Simon's "Too Big too Ignore" The Case for Big Data is published.

2014 Speaker at SNIA Conference Ravages the history of Big Data.

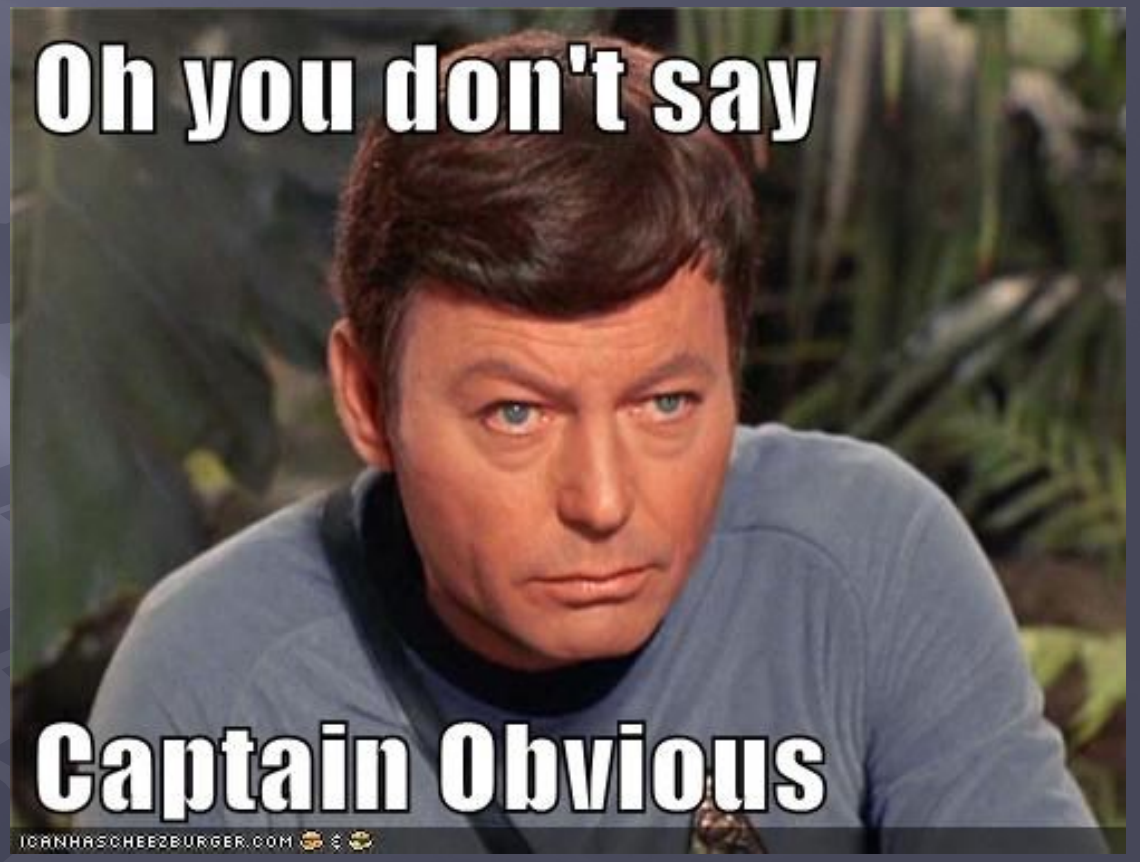
The world of data is changing

4.3 → connected devices per adult

85% → from new data types

10X → increase every five years

27% → use social media



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikimedia Shop

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item

Article | Talk | Read | Edit | View history | Search

Data wrangling

From Wikipedia, the free encyclopedia

Data munging or **data wrangling** is loosely the process of manually converting or mapping data from one "raw" form into another format that allows for more convenient consumption of the data with the help of semi-automated tools. This may include further munging, data visualization, data aggregation, training a statistical model, as well as many other potential uses. Data munging as a process typically follows a set of general steps which begin with extracting the data in a raw form from the data source, "munging" the raw data using algorithms (e.g. sorting) or parsing the data into predefined data structures, and finally depositing the resulting content into a data sink for storage and future use.[1] Given the rapid growth of the internet [2] such techniques will become increasingly important in the organization of the growing amounts of data available.

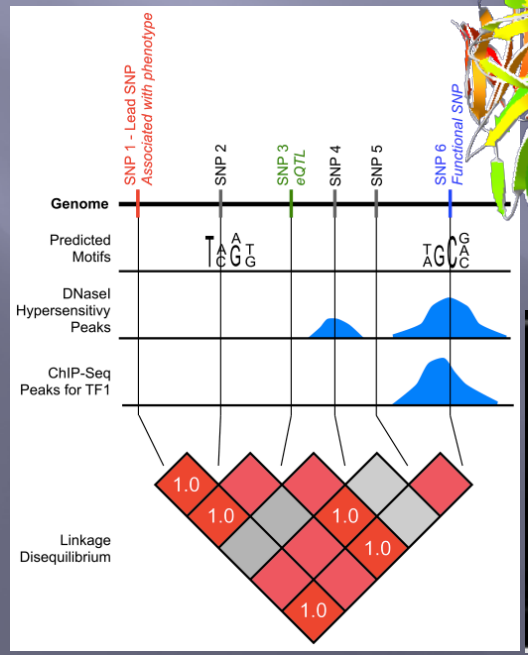
A **data wrangler** is the person performing the wrangling. In the scientific research context, the term often refers to a person responsible for gathering and organizing disparate data sets collected by many different investigators, often as part of a field campaign. In this sense, the term could be credited to Donald Cline during the NASA/NOAA Cold Lands Processes Experiment.[3] It specifies duties typically handled by a **storage administrator** for working with large amounts of data. This can occur in areas like major research projects and the making of films with a large amount of complex computer-generated imagery. In research, this involves both data transfer from research instrument to storage grid or storage facility as well as data manipulation for re-analysis via high performance computing instruments or access via cyberinfrastructure-based digital libraries.

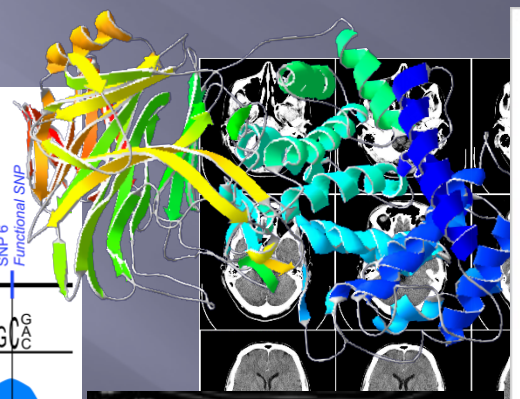
The "wrangler" non-technical term is often said to derive from work done by the United States Library of Congress's National Digital Information Infrastructure and Preservation Program (NDIIPP) and their program partner the Emory University Libraries based MetaArchive Partnership. The term "mung" has roots in munging as described in the Jargon File[4] The term "Data Wrangler" was also suggested as the best analogy to coder for code for someone working with data.[5]

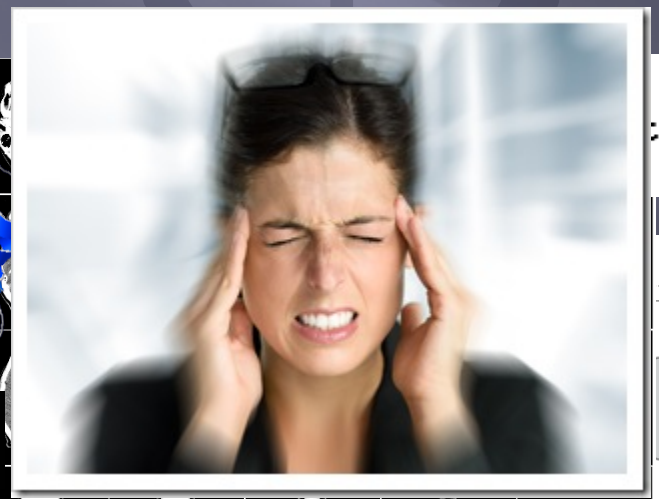
Why do I care?

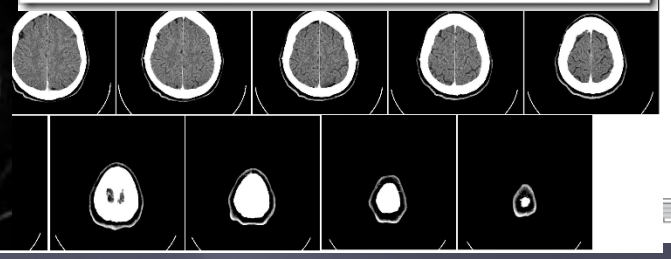
Big Data Presents Huge Challenges in Biomedicine

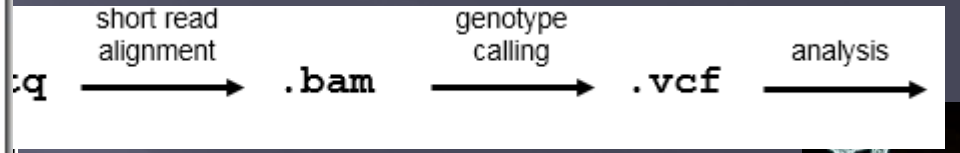
- Biomedical knowledge (data) is growing in size and complexity.
- Translation of complex data and sources has an unknown path
- Aggregation of source data is a **PROBLEM**

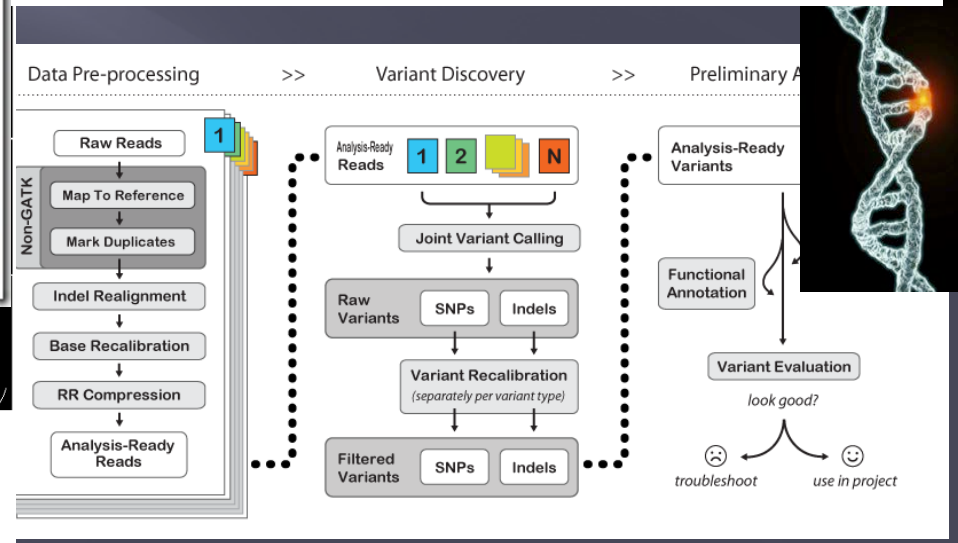


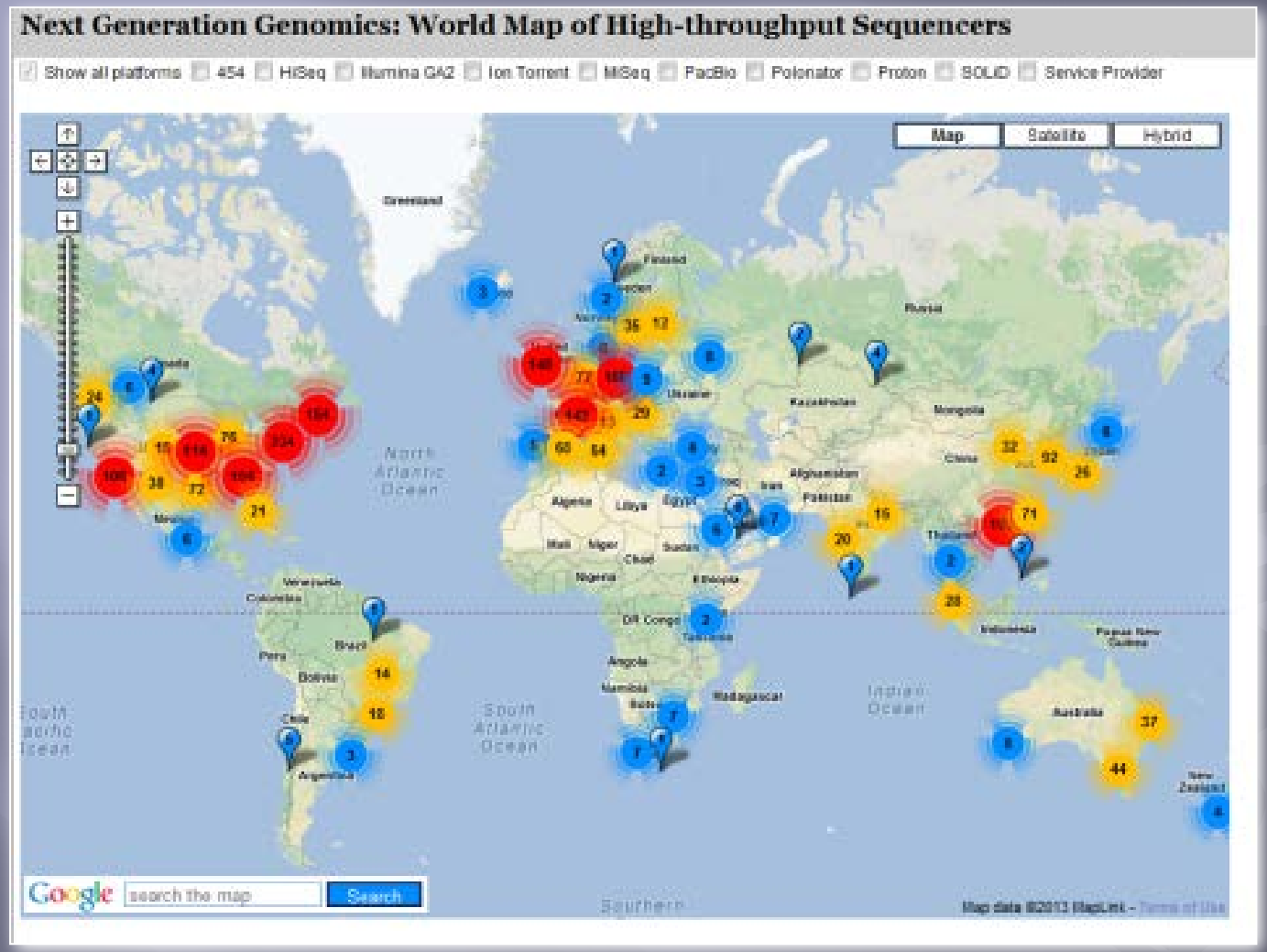












Widespread adoption of NGS systems

Sequencing cost $3K-5K/genome

Generation time < 1-2 days

"With the imminent arrival of the $1,000 genome and continuing advances in global IT infrastructure**,**
we expect whole genome sequencing and analysis to quickly become ubiquitous"

Alan S. Louie, Ph.D., IDC

"Gene Sequencing on its way to being Free"

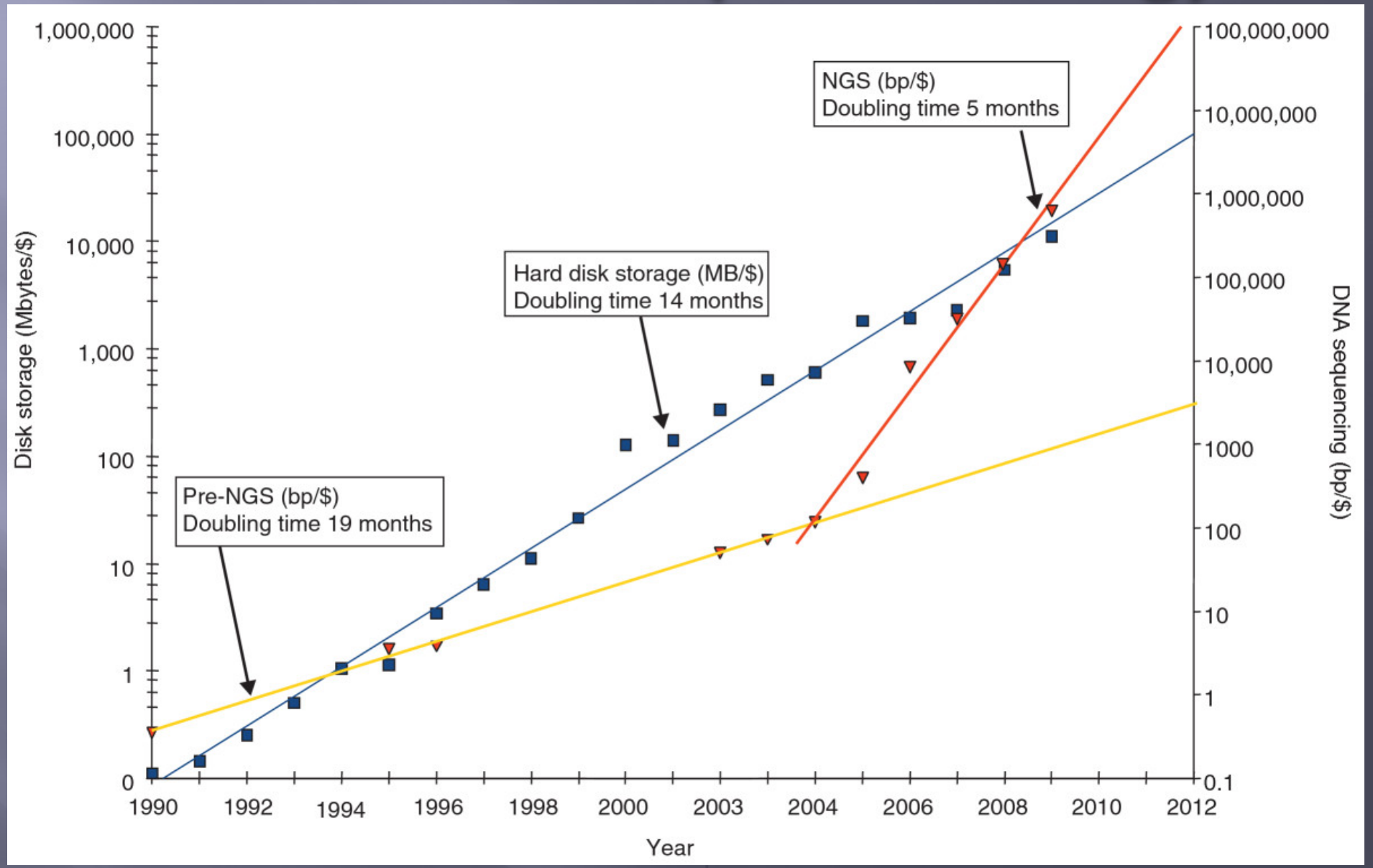
Allen Day, PhD MapR Technologies

Source: Alan S. Louie, Ph.D., IDC, Perspective: From Promise to Practice- Translational Medicine at the Consumer Doorstep, #H1238752

Source: http://omicsmaps.com/ Crowdsourced map of NGS systems conceived by James Hadfield (Cancer Research UK, Cambridge) and built by Nick Loman (University of Birmingham).

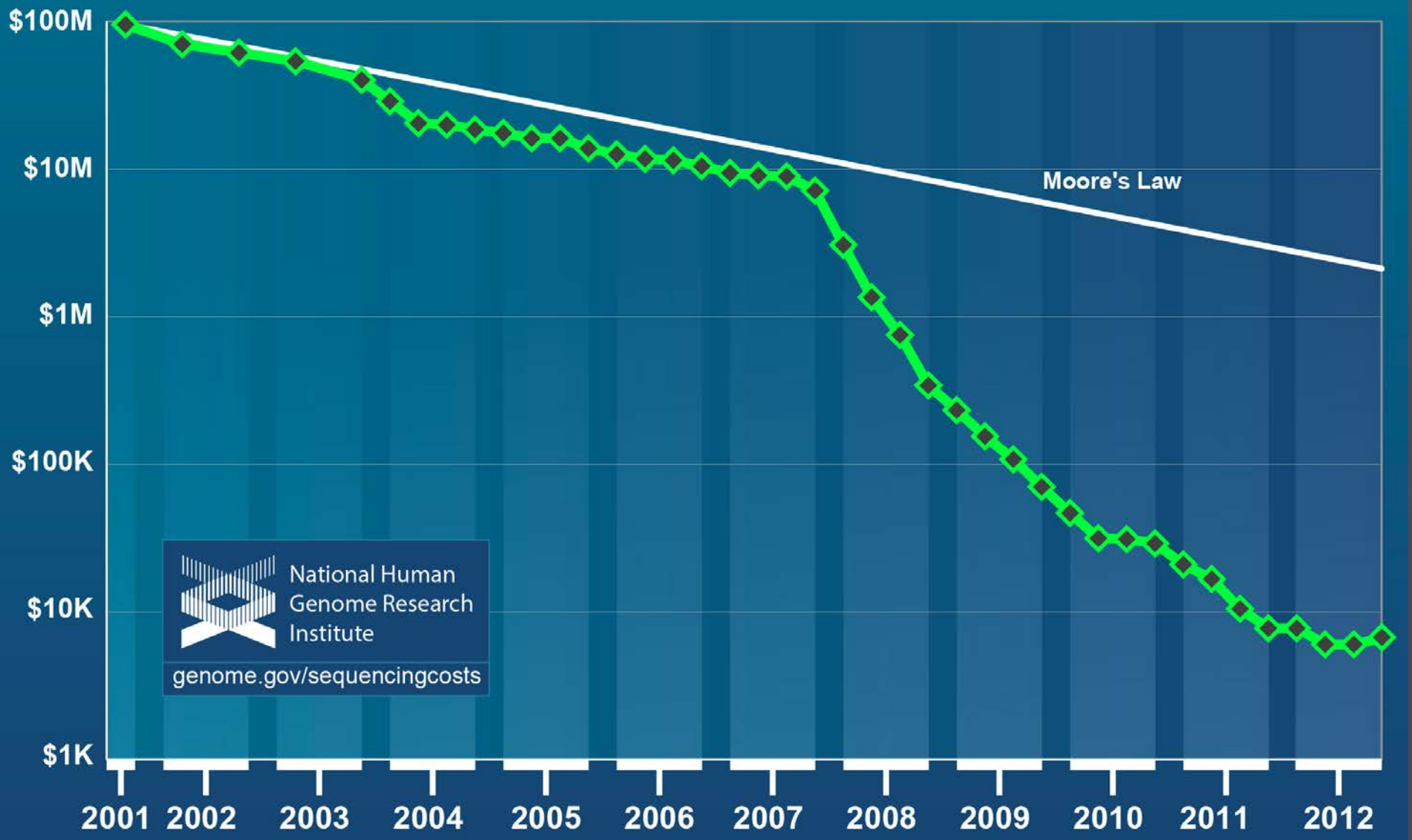
Source: https://www.genome.gov/images/illustrations/hgp_measures.pdf NIH/NHGRI

NGS relationship to technology



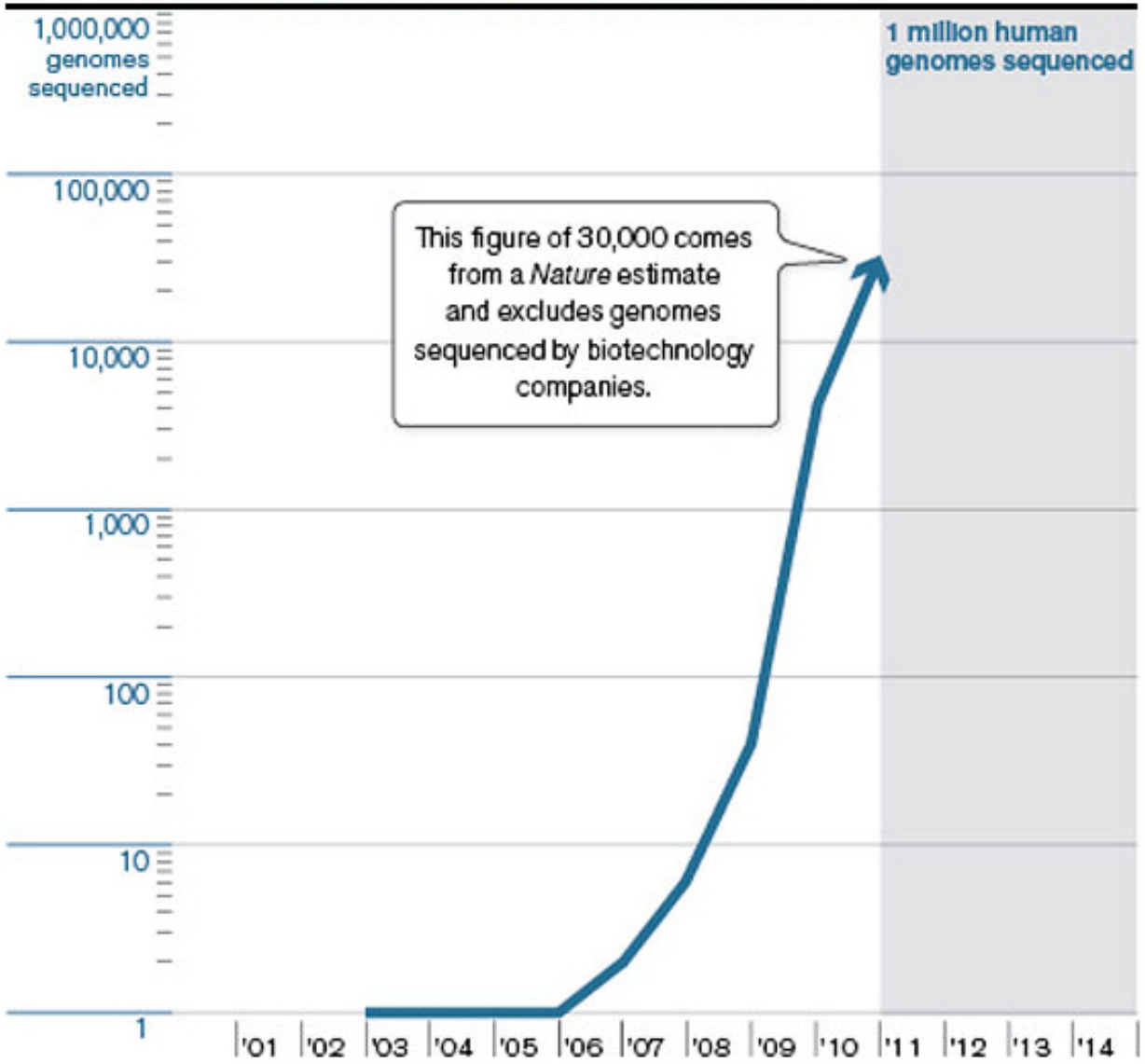
Stein, L. D. The case for cloud computing in genome informatics. Genome Biology (2010).

Cost per Genome



Output Skyrocketing

Number sequenced



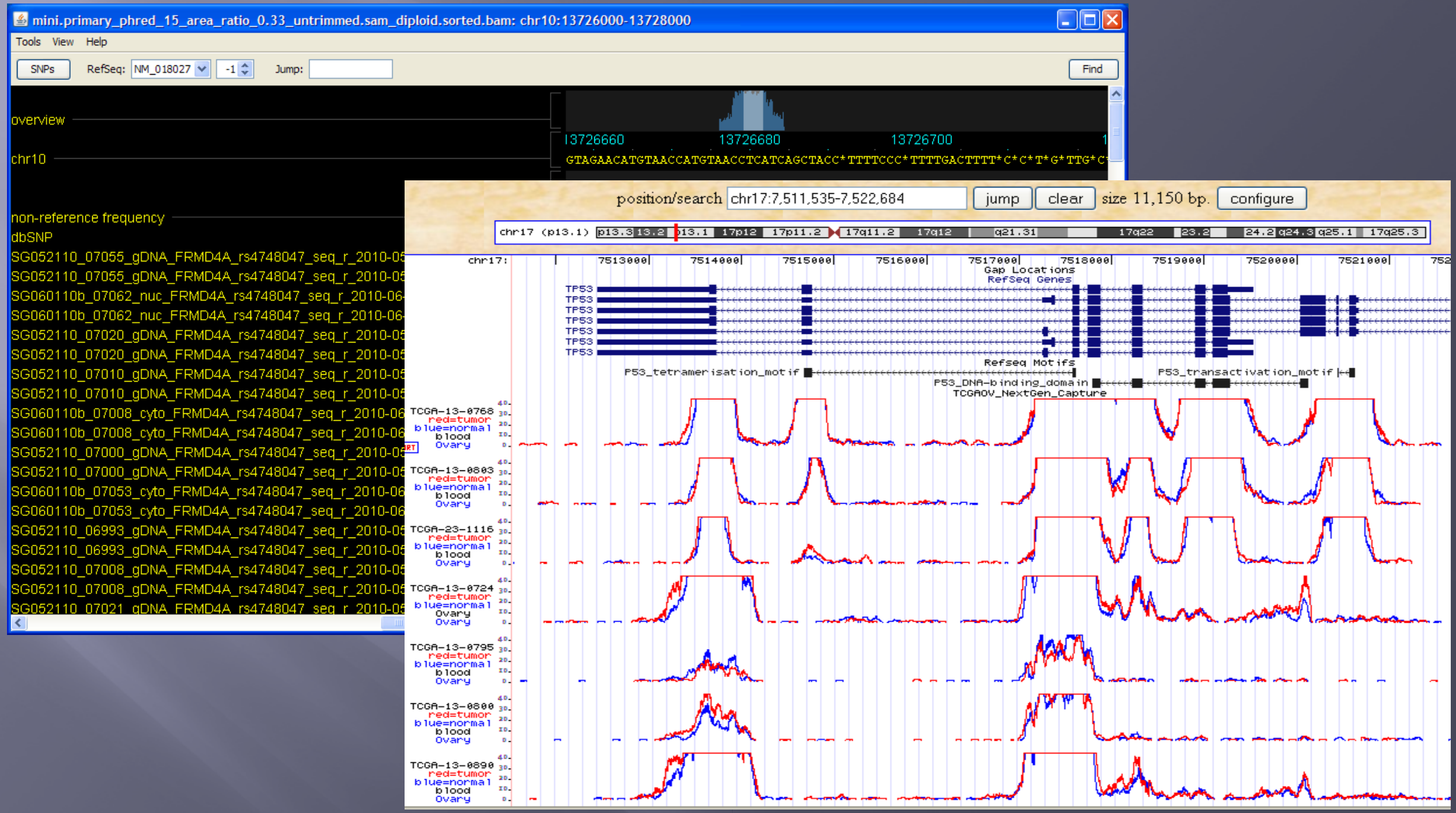
Orcutt Technology Review 2012, data NHGRI

Alternative Representation

- Using two bits for each nucleotide a human genome can be encoded by **800 megabytes**
- Alternatively, if one only represented differences from a single reference genome, one could reduce the stored information to **4 megabytes**
- Practically, to account for the lack of precision and accuracy of current technology and complexity of actual genomes it is common to store **100 gigabytes**

Orcutt Technology Review 2012, data NHGRI

Why the Differences?


mini.primary_phred_15_area_ratio_0.33_untrimmed.sam_diploid.sorted.bam: chr10:13726000-13728000
Tools View Help
SNPs RefSeq: NM_018027 -1 Jump: Find
overview
chr10
13726660 13726680 13726700
GTAGAACATGTAACCATGTAACCTCATCAGCTACC*TTTTCCC*TTTTGACTTTT*C*C*T*G*TTG*C
non-reference frequency
dbSNP
SG052110_07055_gDNA_FRMD4A_rs4748047_seq_r_2010-05
SG060110b_07062_nuc_FRMD4A_rs4748047_seq_r_2010-06
SG052110_07020_gDNA_FRMD4A_rs4748047_seq_r_2010-05
SG052110_07010_gDNA_FRMD4A_rs4748047_seq_r_2010-05
SG060110b_07008_cyto_FRMD4A_rs4748047_seq_r_2010-06
SG052110_07000_gDNA_FRMD4A_rs4748047_seq_r_2010-05
SG060110b_07053_cyto_FRMD4A_rs4748047_seq_r_2010-06
SG052110_06993_gDNA_FRMD4A_rs4748047_seq_r_2010-05
SG052110_07008_gDNA_FRMD4A_rs4748047_seq_r_2010-05
SG052110_07021_gDNA_FRMD4A_rs4748047_seq_r_2010-05
position/search chr17:7,511,535-7,522,684 jump clear size 11,150 bp. configure
chr17 (p13.1) p13.3 p13.2 p13.1 17p12 17p11.2 17q11.2 17q12 q21.31 17q22 23.2 24.2 q24.3 q25.1 17q25.3
chr17: 7513000 7514000 7515000 7516000 7517000 7518000 7519000 7520000 7521000
Gap Locations
RefSeq Genes
TP53
Refseq Motifs
P53_tetramerisation_motif
P53_transactivation_motif
P53_DNA-binding_domain
TCGAOV_NextGen_Capture
TCGA-13-0768 red=tumor blue=normal blood Ovary
TCGA-13-0803 red=tumor blue=normal blood Ovary
TCGA-23-1116 red=tumor blue=normal blood Ovary
TCGA-13-0724 red=tumor blue=normal Ovary Ovary
TCGA-13-0795 red=tumor blue=normal blood Ovary
TCGA-13-0800 red=tumor blue=normal blood Ovary
TCGA-13-0890 red=tumor blue=normal blood Ovary

Whole Genome Analysis

440GB/sample*

200GB/sample

100MB/sample

Bowtie
BWA

24 cores,
72 hours

GATK,
Bambino

16 cores,
24 hours

ANNOVAR

8 cores,
15 minutes

** Assuming 60X coverage*

The Cancer Genome Atlas (TCGA)

2.5 petabytes of data!!!

Phenome

Genome

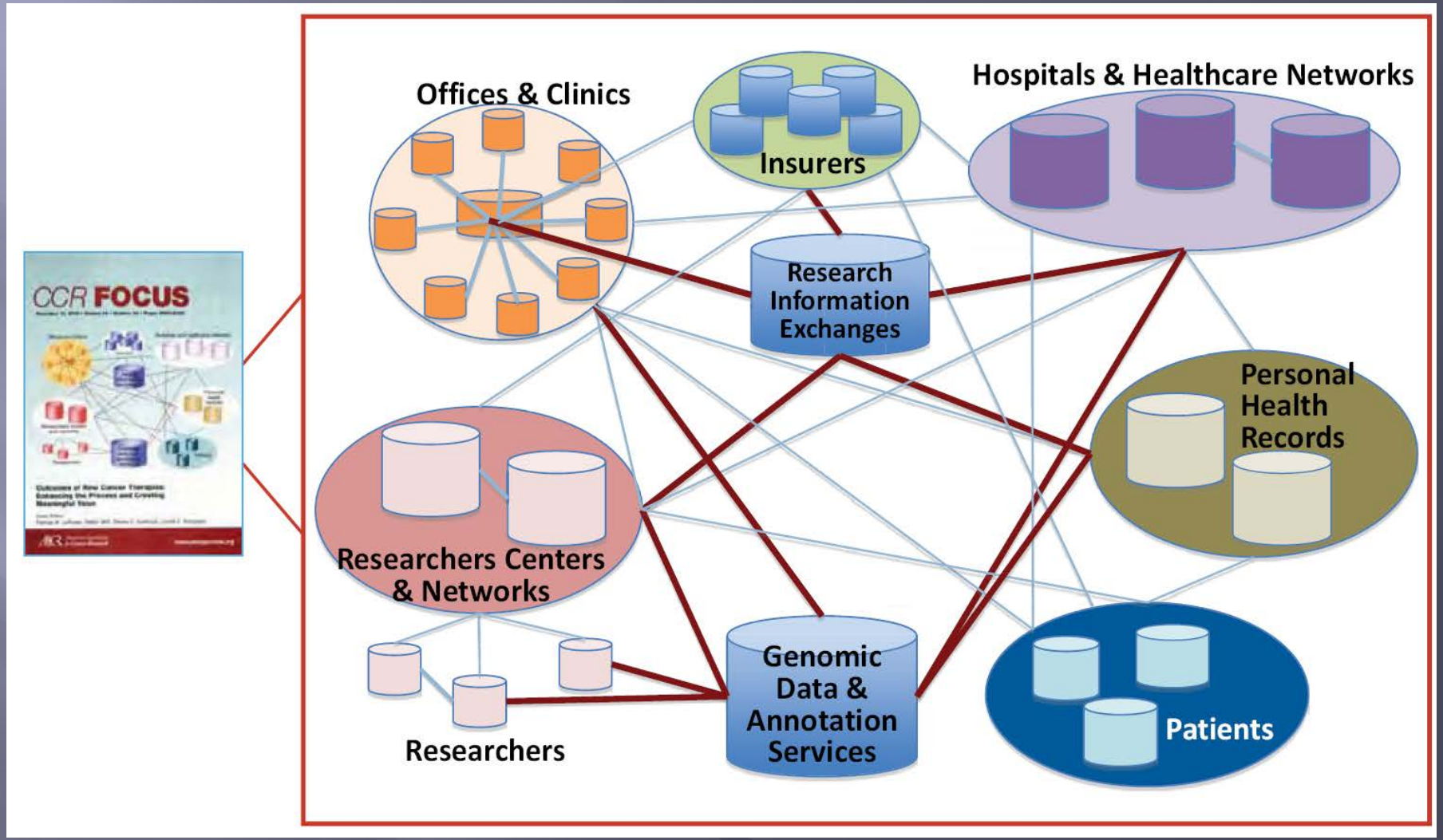
“Big Data”

Exposome

Phenome Data

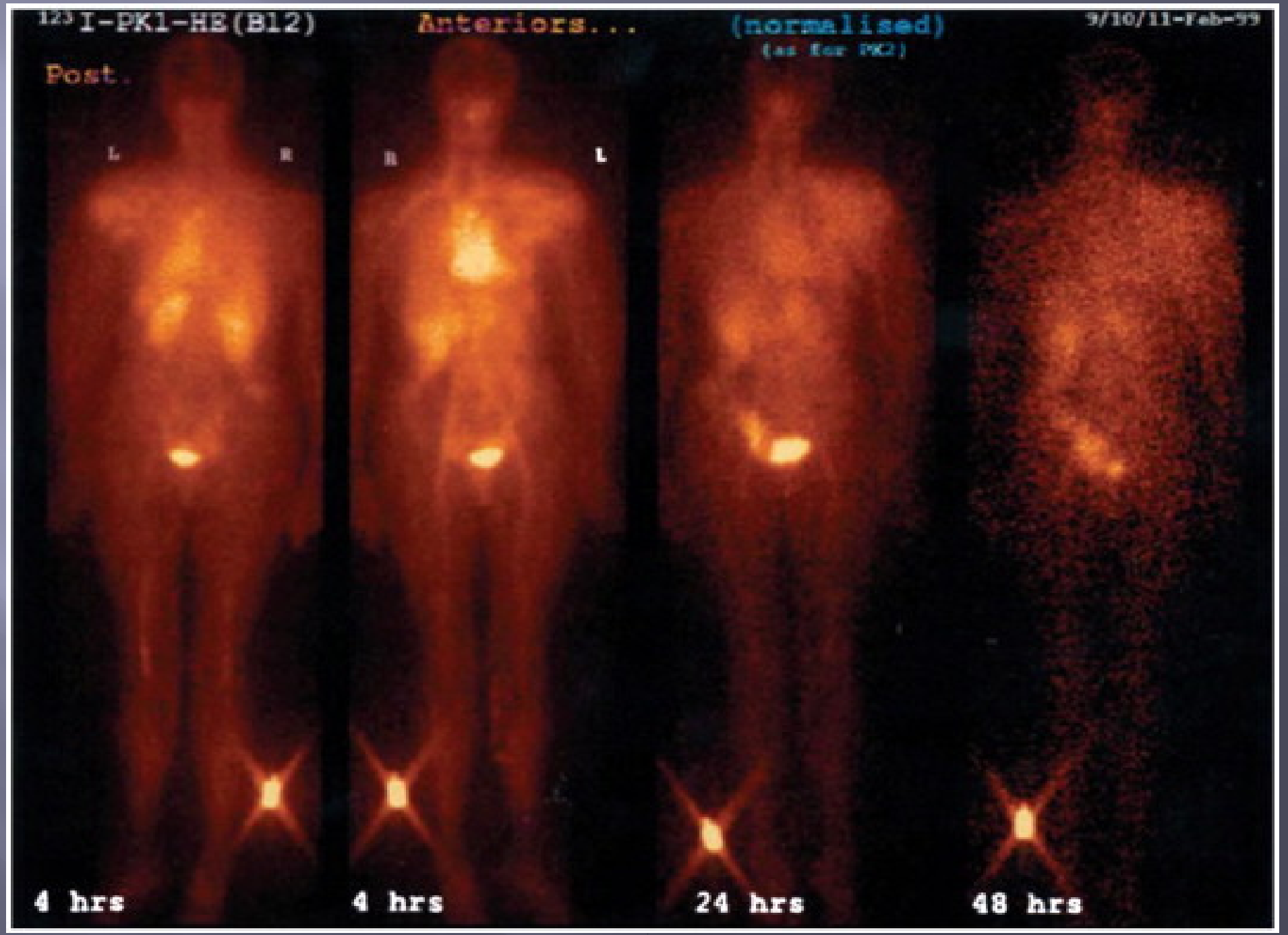
- Diverse types
  - Clinical Observation
  - Clinical Laboratory
  - Imaging
  - Registry
  - Biospecimens
  - Reference
- Distributed sources
  - Research Center
  - Care Delivery Setting
    - Hospital
    - Practice
    - Laboratory
  - Registry
  - Industry
  - Consumer

Distributed sources are *the rule* in biomedicine



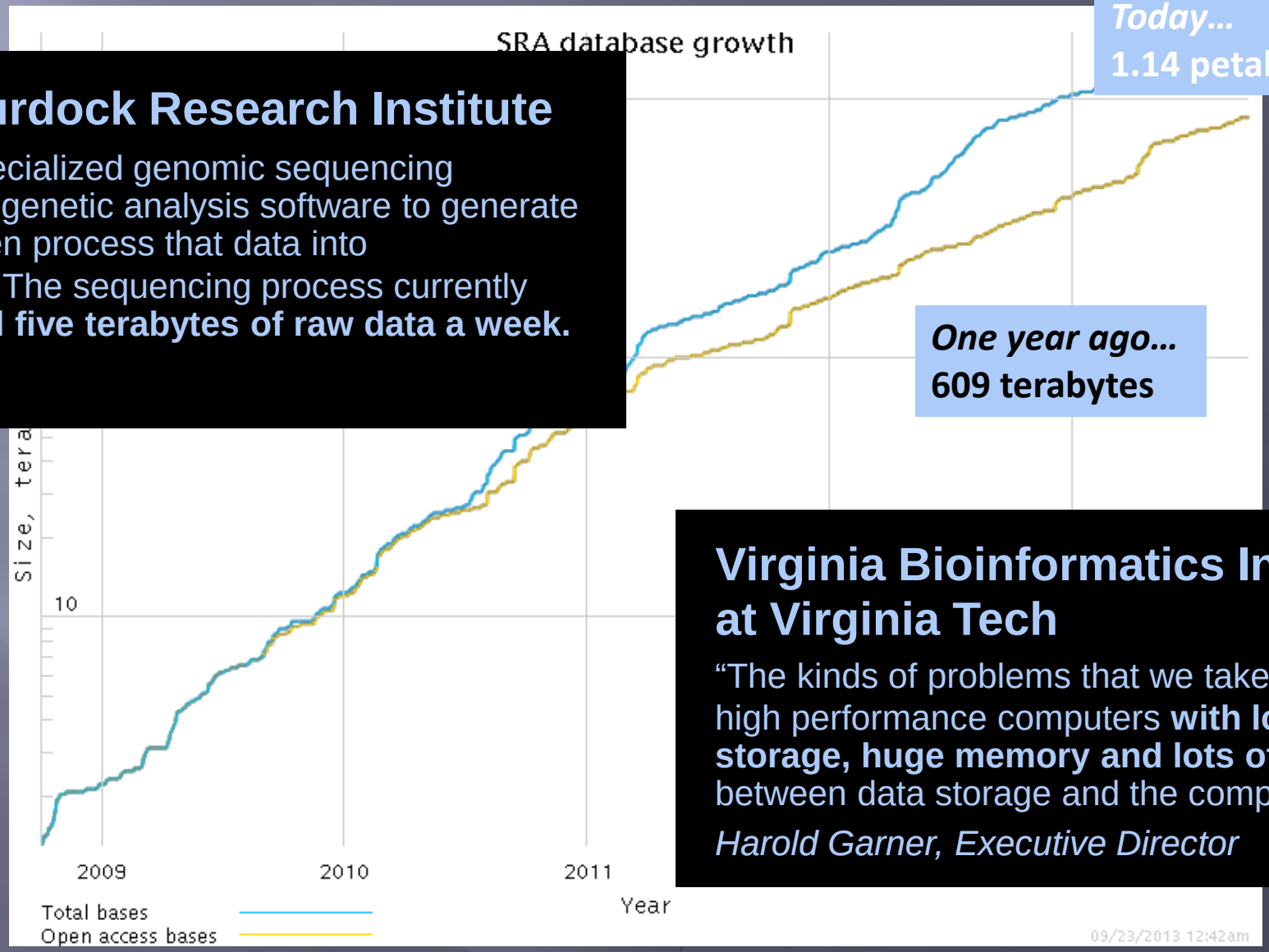
W. Dalton et al, Clinical Cancer Res; 16 (24) December 15, 2010

Amplifying the complexity



Lu, Advanced Drug Delivery Reviews 2010

NCBI Sequence Read Archive (SRA)



Today…
1.14 petabytes

One year ago…
609 terabytes

David H. Murdock Research Institute

DHMRI uses specialized genomic sequencing instruments and genetic analysis software to generate raw data and then process that data into a usable format. The sequencing process currently produces around **five terabytes of raw data a week.**

Virginia Bioinformatics Institute at Virginia Tech

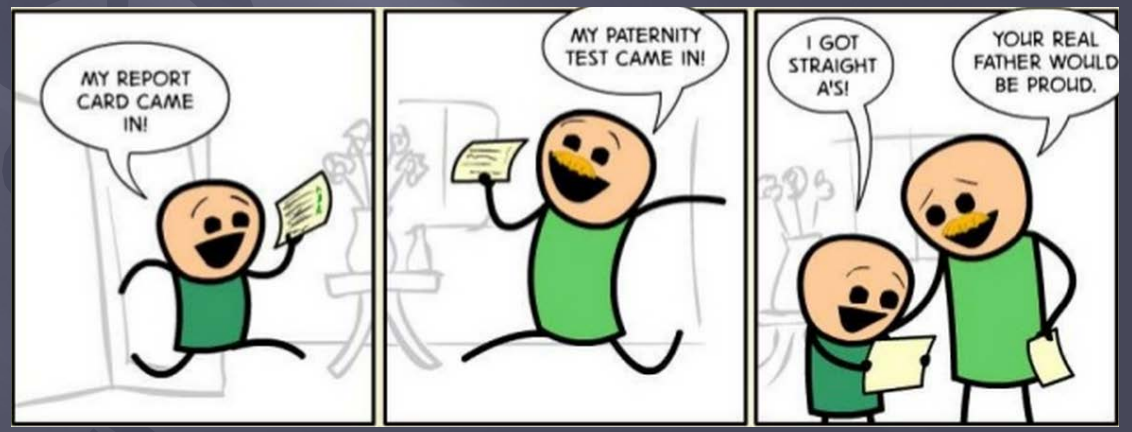
"The kinds of problems that we take on require high performance computers **with lots of data storage, huge memory and lots of bandwidth** between data storage and the compute clusters."

Harold Garner, Executive Director

http://www.ncbi.nlm.nih.gov/

Next Generation Genome Sequencing (NGS)

What your mother thinks it is…….

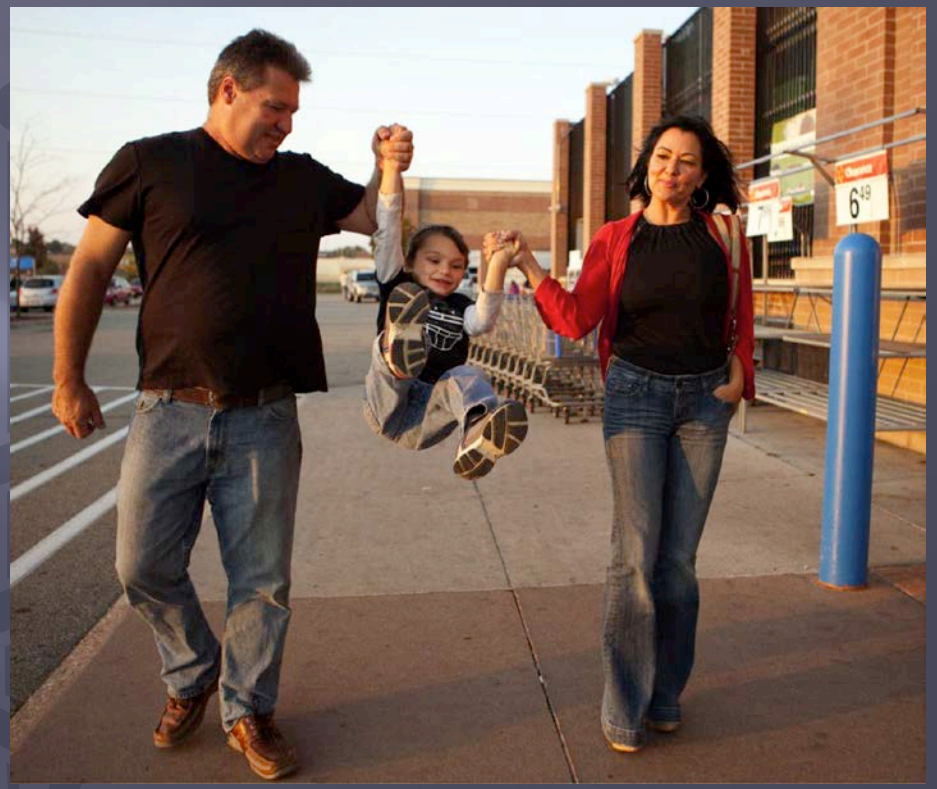


What your friends think it is…….

http://en.wikipedia.org/wiki/Paternity_fraud
http://www.nytimes.com/2008/04/29/world/asia/29singapore.html?pagewanted=2&_r=1&ref=asia
Images courtesy of Cyanide and Happiness / Star Wars.

Next Generation Genome Sequencing (NGS)

What it is…….



Better Patient Outcomes!

Rare Disease Case Study : Nicholas Volker

http://raregenomics.org/rare-disease-case-study-nicholas-santiago-volker-alive-because-of-genomics/

Electronic Health Records EHR / EMR

Electronic Health Records EHR / EMR

When health care providers have access to complete and accurate information, patients receive better medical care. Electronic Health Records or Electronic Medical Record (EHRs/EMRs) can improve the ability to diagnose diseases and reduce—even prevent—medical errors, improving patient outcomes.

A national survey of doctors who are ready for meaningful use offers important evidence:

94% of providers report that their EHR/EMR makes records readily available at point of care.
88% report that their EHR/EMR produces clinical benefits for the practice.
75% of providers report that their EHR/EMR allows them to deliver better patient care.

Find the Right Expert Resource

Find the Right drug/Trial for personalized care

Jamoom, E., Patel, V., King, J., & Furukawa, M. (2012, August). National perceptions of EHR/EMR adoption: Barriers, impacts, and federal policies. National conference on health statistics.
BCouch, James B. "CCHIT certified electronic health records may reduce malpractice risk," Physician Insurer. 2008.

Electronic Health Records EHR / EMR

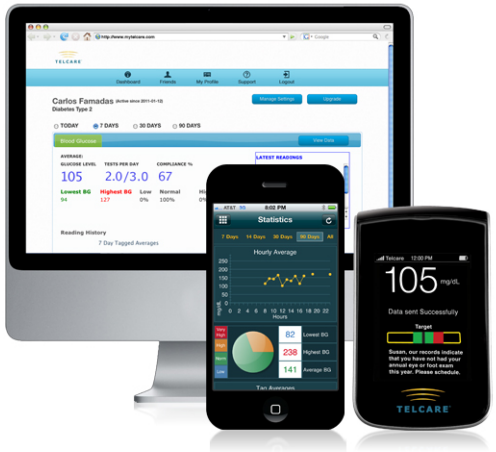
An IDC Health Insights study predicts there will be explosive, double-digit growth in spending on ambulatory and inpatient electronic medical record (EMR) and electronic health record (EHR) software between 2009 and 2015.

EHR Spending To Hit $3.8 Billion In 2015

http://www.informationweek.com/healthcare/electronic-health-records/ehr-spending-to-hit-$38-billion-in-2015/d/d-id/1095366?

Real time consumer data: the *next* Big Data Challenge

Cardiovascular



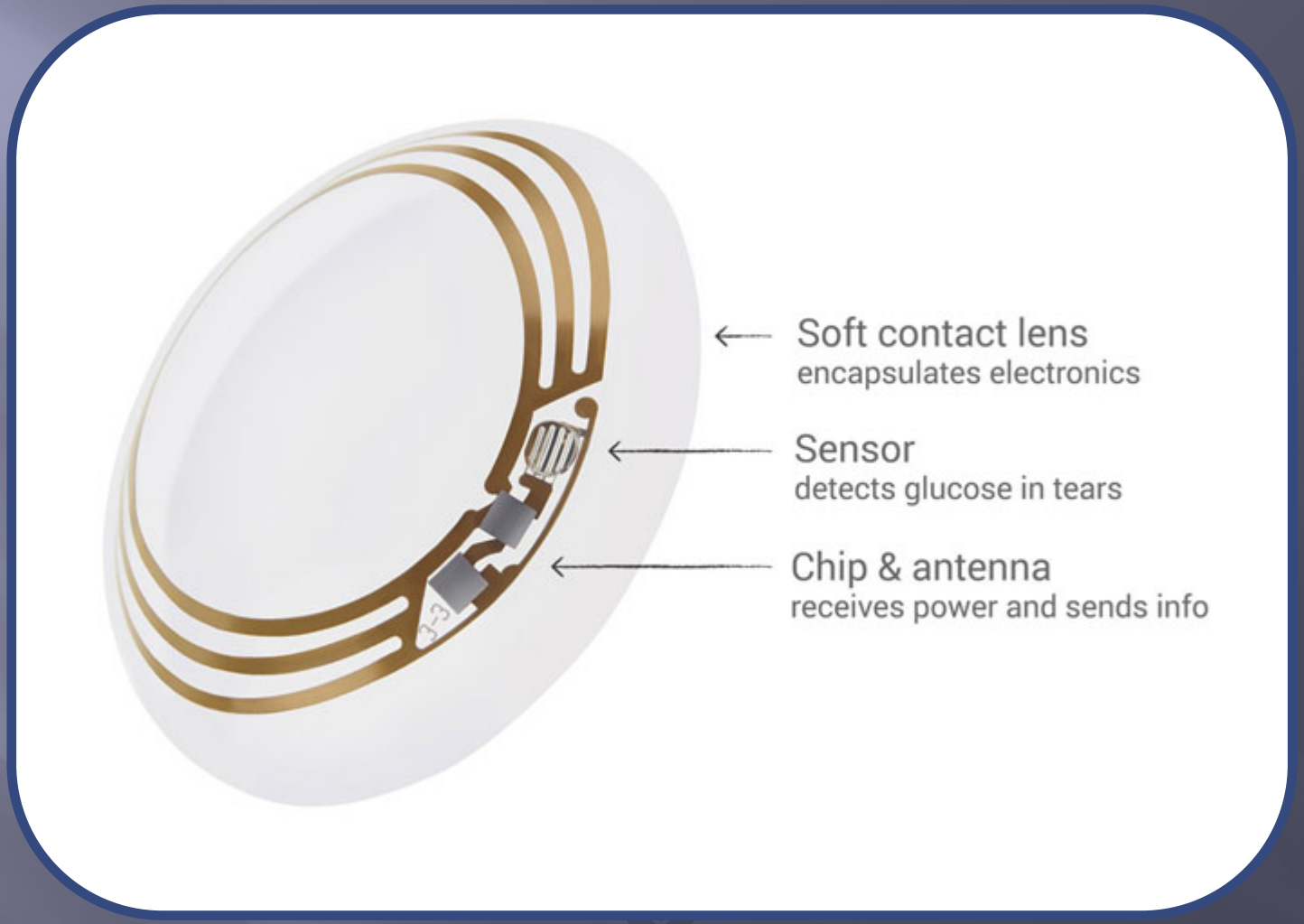
Diabetes

Fitness

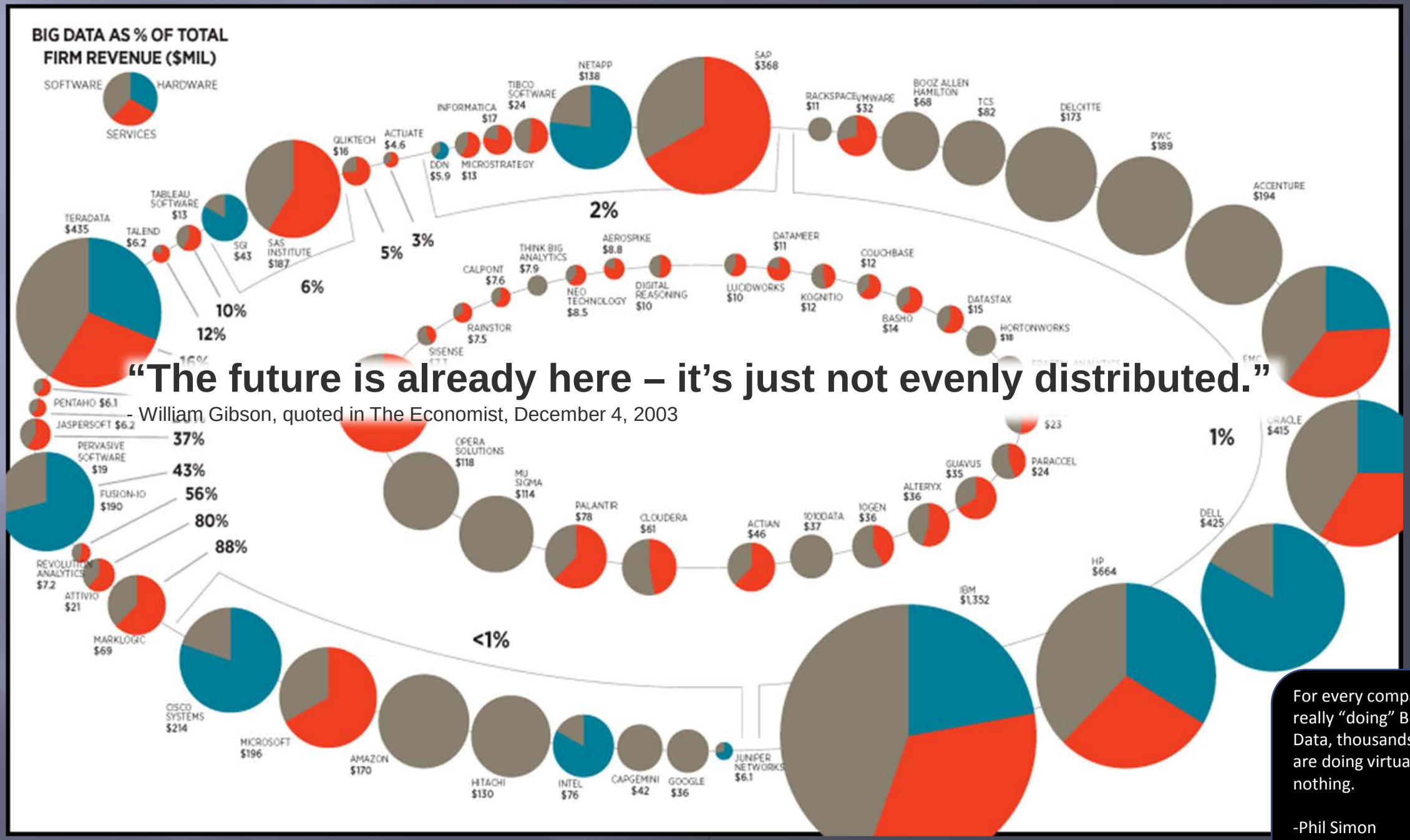
***Shipments* of TeleHealth devices grow to about 7 million by 2018**

Source: www.ihs.com World Market for TeleHealth – 2014 Edition

Google's "Smart" Contact Lens to measure glucose levels



http://www.engadget.com/2014/01/17/google-health-smart-contact-lenses-diabetes/



“The future is already here – it’s just not evenly distributed.”

- William Gibson, quoted in The Economist, December 4, 2003

For every company really “doing” Big Data, thousands more are doing virtually nothing.

-Phil Simon
Too Big to Ignore

http://www.aerospike.com/wp-content/uploads/2013/05/Forbes-Big-Data-Graph.png

Real time consumer data: the *next* Big Data Challenge

VIEW

2 COMMENTS

Emerging Technology From the arXiv

August 18, 2014

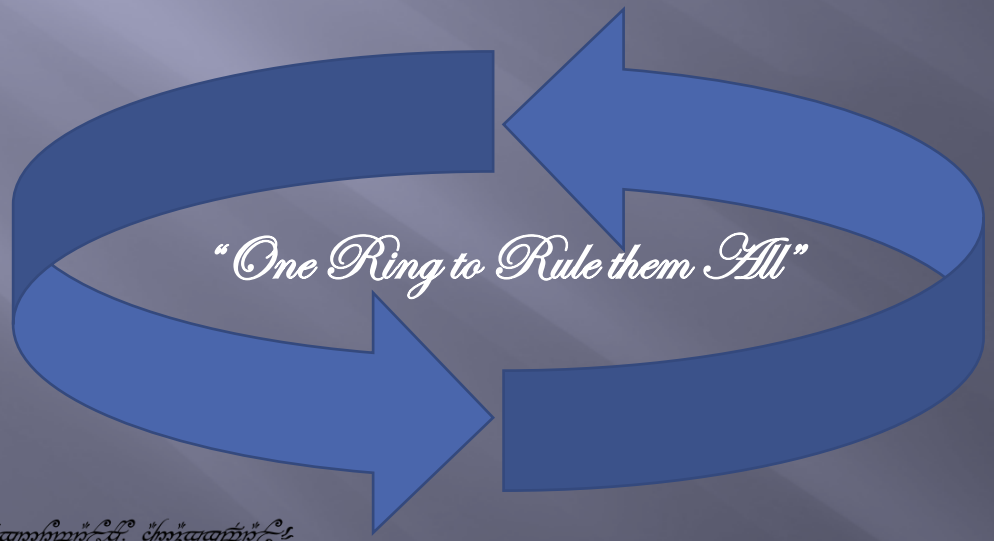
The Emerging Pitfalls Of Nowcasting With Big Data

Statisticians have boasted of the benefits of big data. Now they're discovering the weaknesses.

Source: http://www.technologyreview.com/view/530131/the-emerging-pitfalls-of-nowcasting-with-big-data

Primary Target Types of Data Sources

- Shared Access/ Single Source
- Patient identity portability
- Complete Life-Cycle
- Opt-In (Privacy)



NGS

- Known Resources (TCGA, NCBI, NCI, Etc.)
- Newly shared research data (Cloud)

EHR/EMR

- Patient Records Portability
- Unique Identifiers (Life Cycle)

Pharma

- Clinical Trials Data (Succeed/Failed)

What about MapReduce?

- **Problems with MR**:
  - **Very low-level**: requires a lot of code to do simple things
  - **Very constrained**: everything must be described as “map” and “reduce”.

Powerful but sometimes difficult to think in these terms.

- We don’t like to work in JAVA constrained JVM Operations
- We solved the SPOF issues with **MRv2 and YARN**

Can we improve on MR?

- Two approaches to improve on MapReduce:

1. Special purpose systems to solve one problem domain well.
   - Giraph / Graphlab (graph processing)
   - Storm (stream processing)
   - Impala (real-time SQL)

2. Generalize the capabilities of MapReduce to provide a richer foundation to solve problems.
   - Tez, MPI, Hama/Pregel (BSP), Dryad (arbitrary DAGs)

- *Both are viable strategies depending on the problem but what about this Spark thingy?!?!?*

Hadoop for Real-Time Big Data ???

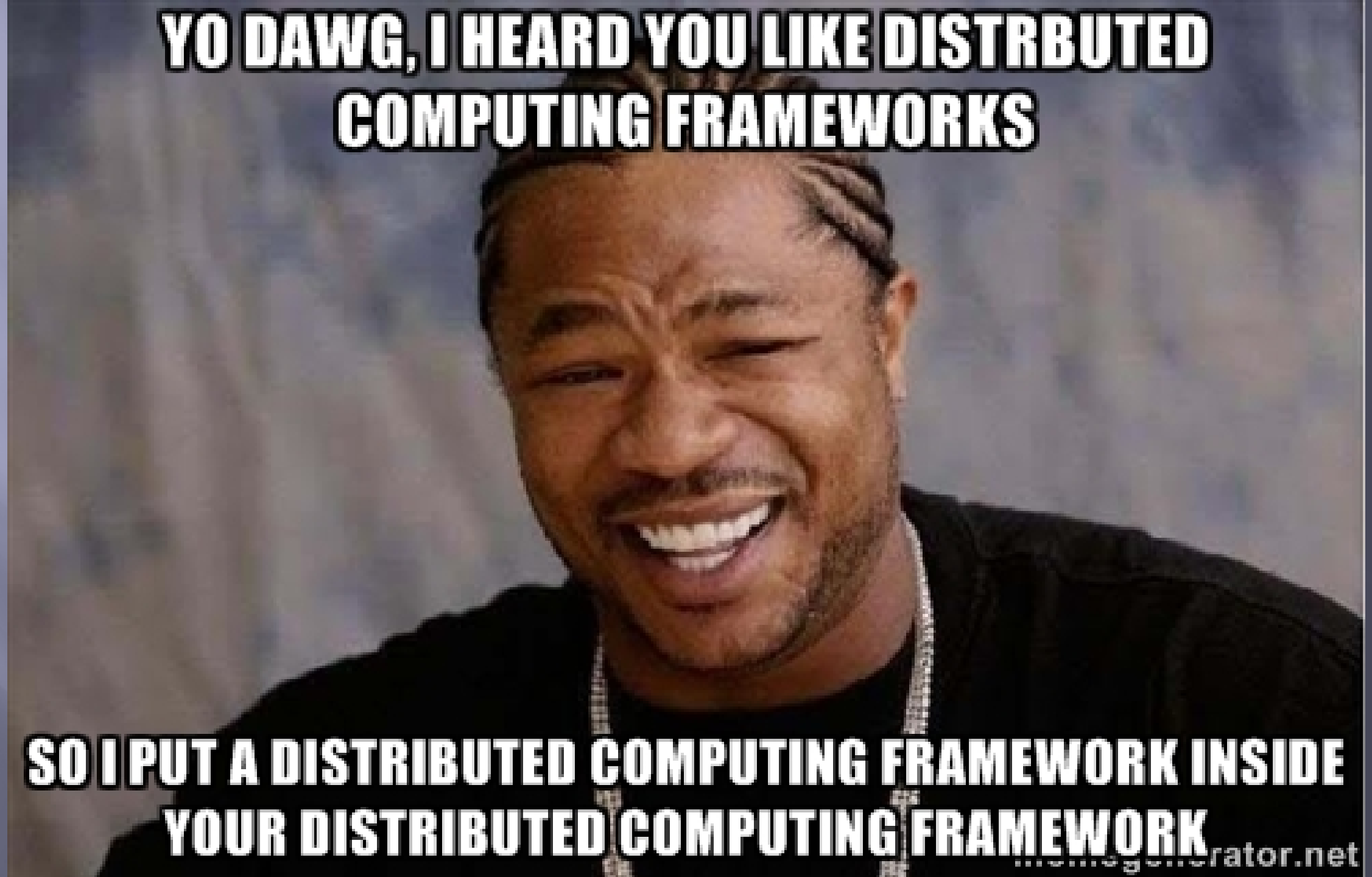
That's so 2005 Dude!

Ancient Aliens Created Big Data

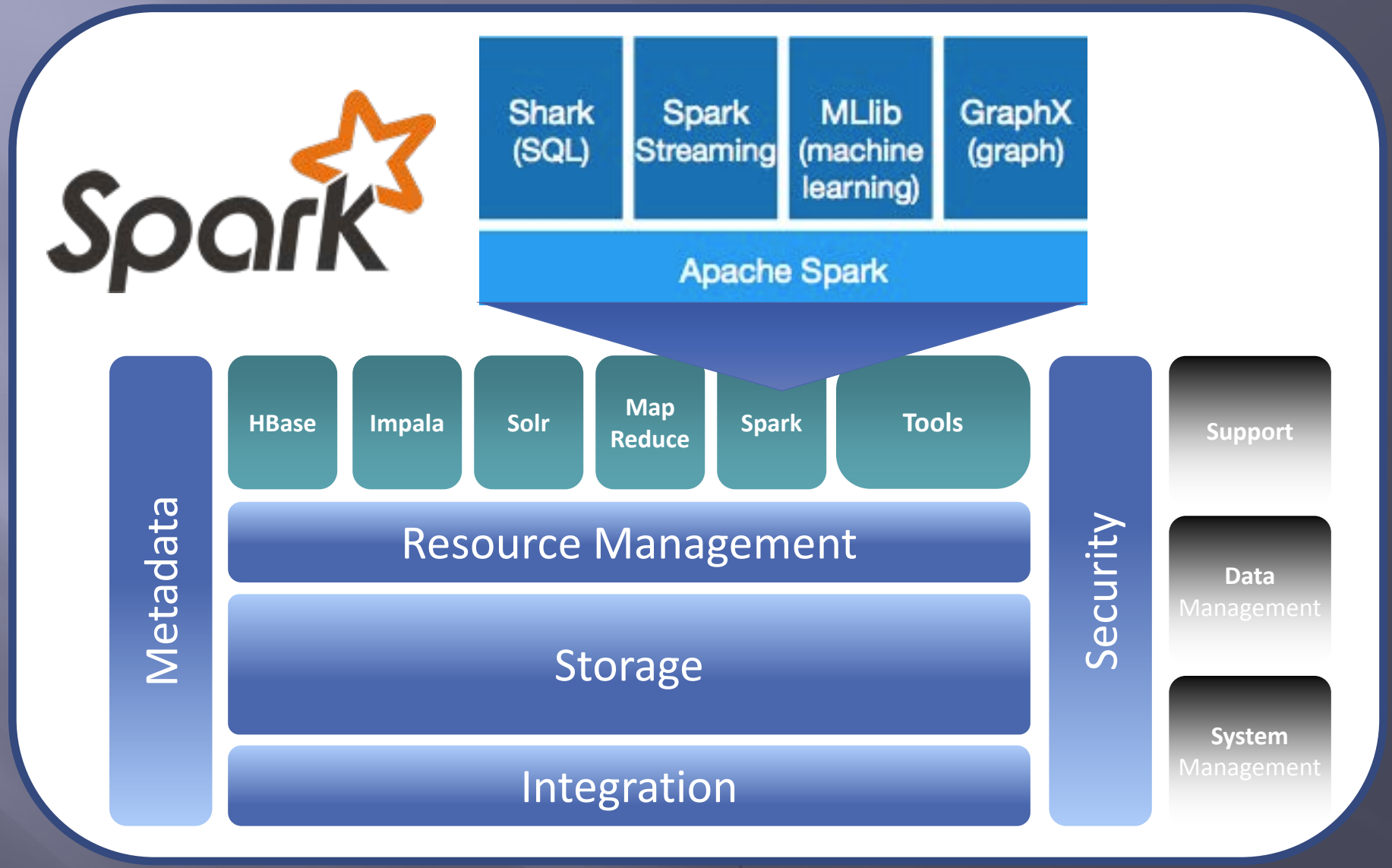
What is Apache Spark?

- **Spark is a general purpose computational framework**
- **Retains the advantages of MapReduce:**
- Linear scalability
- Fault-tolerance
- Data Locality based computations

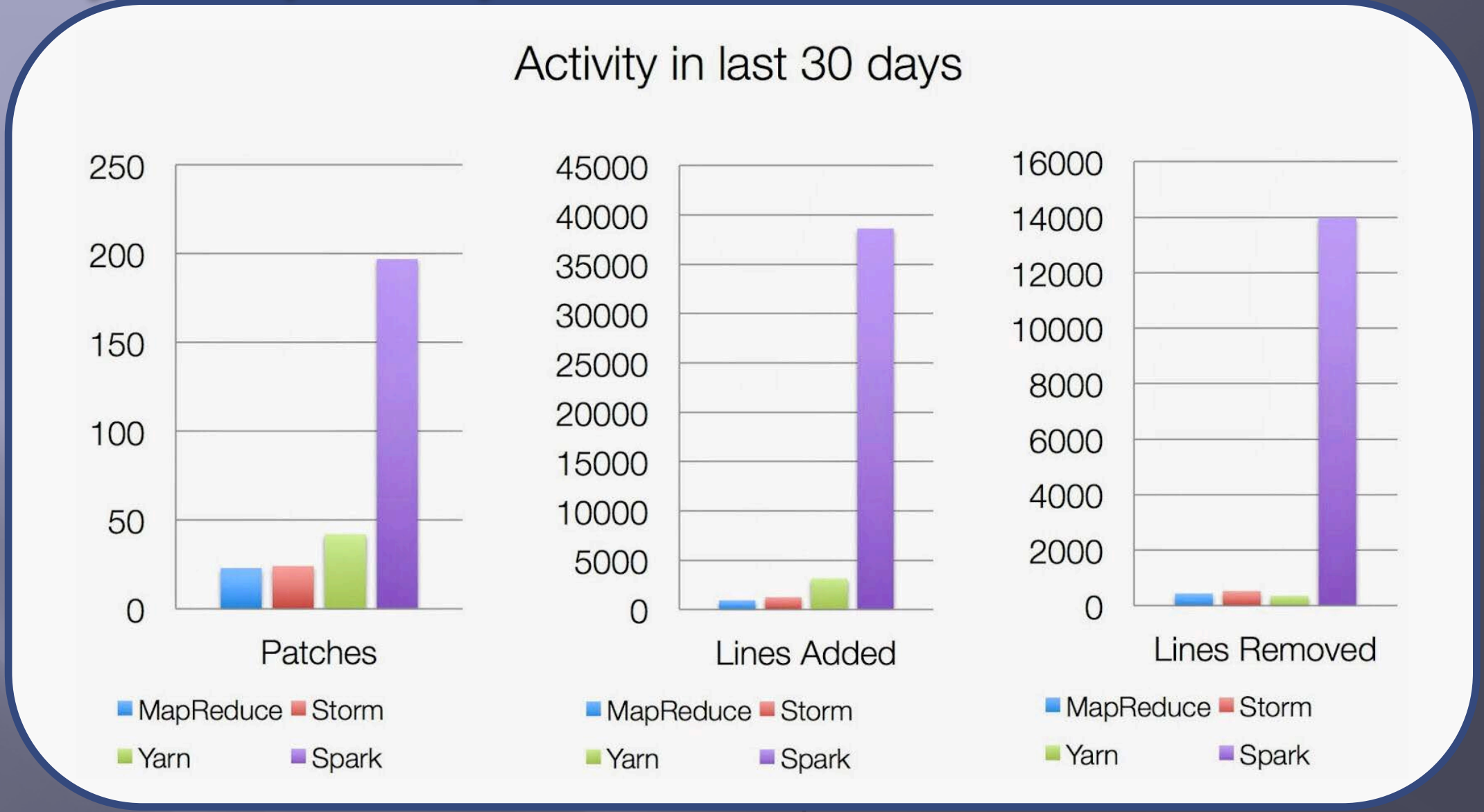
- **…but offers so much more:**
- Leverages distributed memory for better performance
- Supports iterative algorithms that are not feasible in MR
- Improved developer experience
- Full Directed Graph expressions for data parallel computations
- Comes with libraries for machine learning, graph analysis, etc.


YO DAWG, I HEARD YOU LIKE DISTRBUTED COMPUTING FRAMEWORKS
SO I PUT A DISTRIBUTED COMPUTING FRAMEWORK INSIDE YOUR DISTRIBUTED COMPUTING FRAMEWORK
rator.net

In Memory Database Blazing Fast



Project Popularity



Mahout Project has moved to Spark as Development Platform

Word Count in MapReduce

Word Count in Spark

```
package org.myorg;

import java.io.IOException;
import java.util.*;

import org.apache.hadoop.fs.Path;
import org.apache.hadoop.conf.*;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapreduce.*;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;

public class WordCount {

 public static class Map extends Mapper<LongWritable, Text, Text, IntWritable> {
   private final static IntWritable one = new IntWritable(1);
   private Text word = new Text();

   public void map(LongWritable key, Text value, Context context) throws IOException,
InterruptedException {
       String line = value.toString();
       StringTokenizer tokenizer = new StringTokenizer(line);
       while (tokenizer.hasMoreTokens()) {
          word.set(tokenizer.nextToken());
          context.write(word, one);
       }
   }
 }
```

```
public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable> {

  public void reduce(Text key, Iterable<IntWritable> values, Context context)
   throws IOException, InterruptedException {
     int sum = 0;
     for (IntWritable val : values) {
        sum += val.get();
     }
     context.write(key, new IntWritable(sum));
  }
}

public static void main(String[] args) throws Exception {
  Configuration conf = new Configuration();

     Job job = new Job(conf, "wordcount");
  job.setOutputKeyClass(Text.class);
  job.setOutputValueClass(IntWritable.class);

  job.setMapperClass(Map.class);
  job.setReducerClass(Reduce.class);

  job.setInputFormatClass(TextInputFormat.class);
  job.setOutputFormatClass(TextOutputFormat.class);

  FileInputFormat.addInputPath(job, new Path(args[0]));
  FileOutputFormat.setOutputPath(job, new Path(args[1]));

  job.waitForCompletion(true);
}

}
```

```
sc.textFile(“words”)
  .flatMap(line => line.split(" "))
  .map(word=>(word,1))
  .reduceByKey(_+_).collect()
```

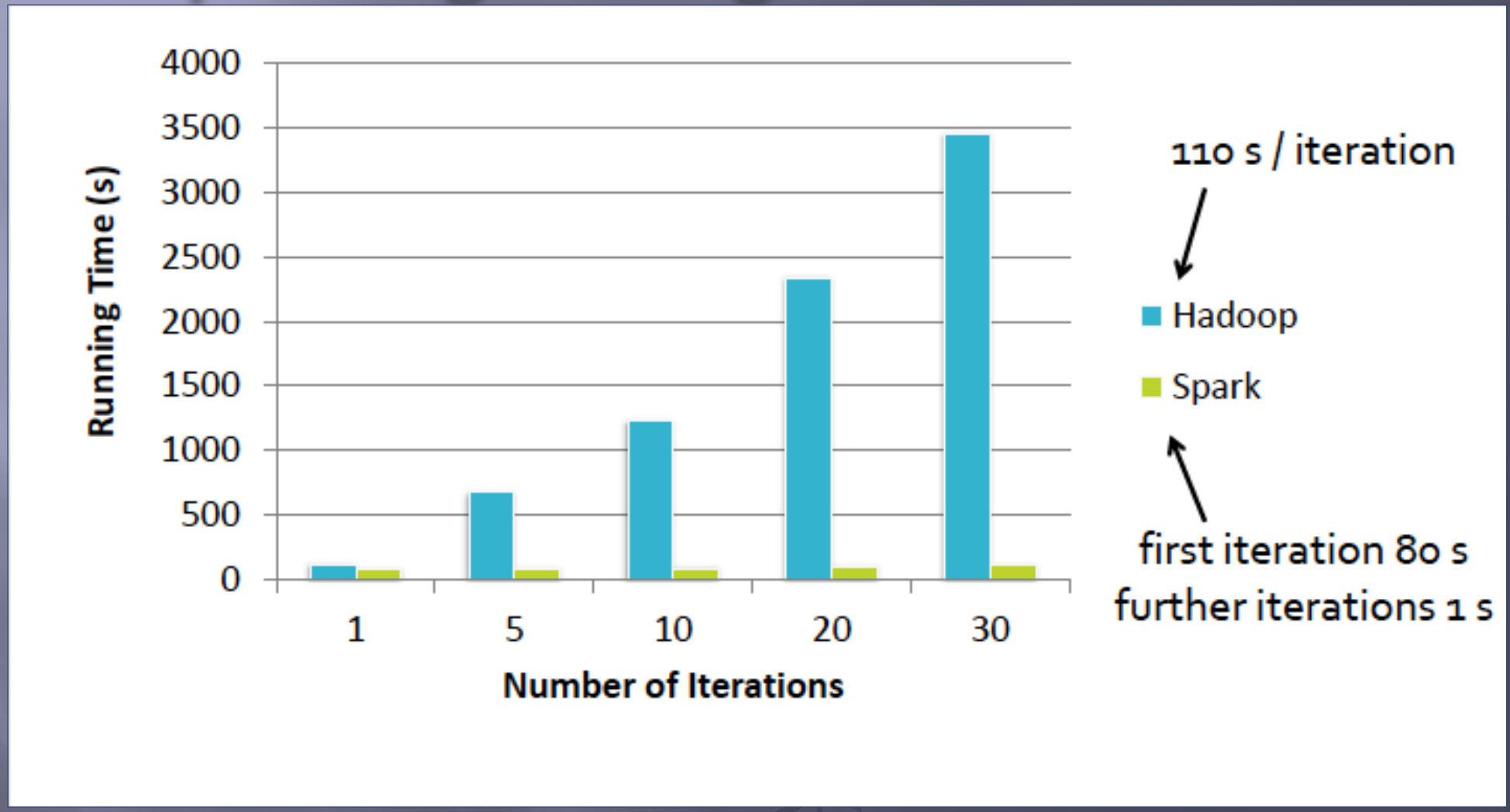
So fast we can do it again!

```
sc.textFile(“words”)
    .flatMap(line => line.split(" "))
    .map(word=>(word,1))
    .reduceByKey(_+_).collect()
```

*Systems analysis utilizing pathway interactions identifies **sonic hedgehog pathway** as a primary biomarker and oncogenic target in hepatocellular carcinoma.*

Source: http://www.ncbi.nlm.nih.gov/pubmed/24712101

Spark Logistic Regression Performance



What needs to be on the roadmap?

- ✓ Singularity of namespace across potentially multi-petabyte structures
- ✓ Geographically distributed filesystems that are Highly Available and WAN Optimized
- ✓ POSIX compliant object based storage | Unified file and object storage (UFOS)
- ✓ Big Data as a Service with WAN optimization (Deduplication for HDFS?)
- ✓ Software defined abstraction layers to commoditize storage (SDS-SDC)
- ✓ Big Data security models (HIPPA, FISMA, FERPA, DISA-STIG, PCI-DSS)
- ✓ In memory databases to aligning storage and compute
- ✓ **Big Data that is Open Big Data**

Bioinformatics as a Service

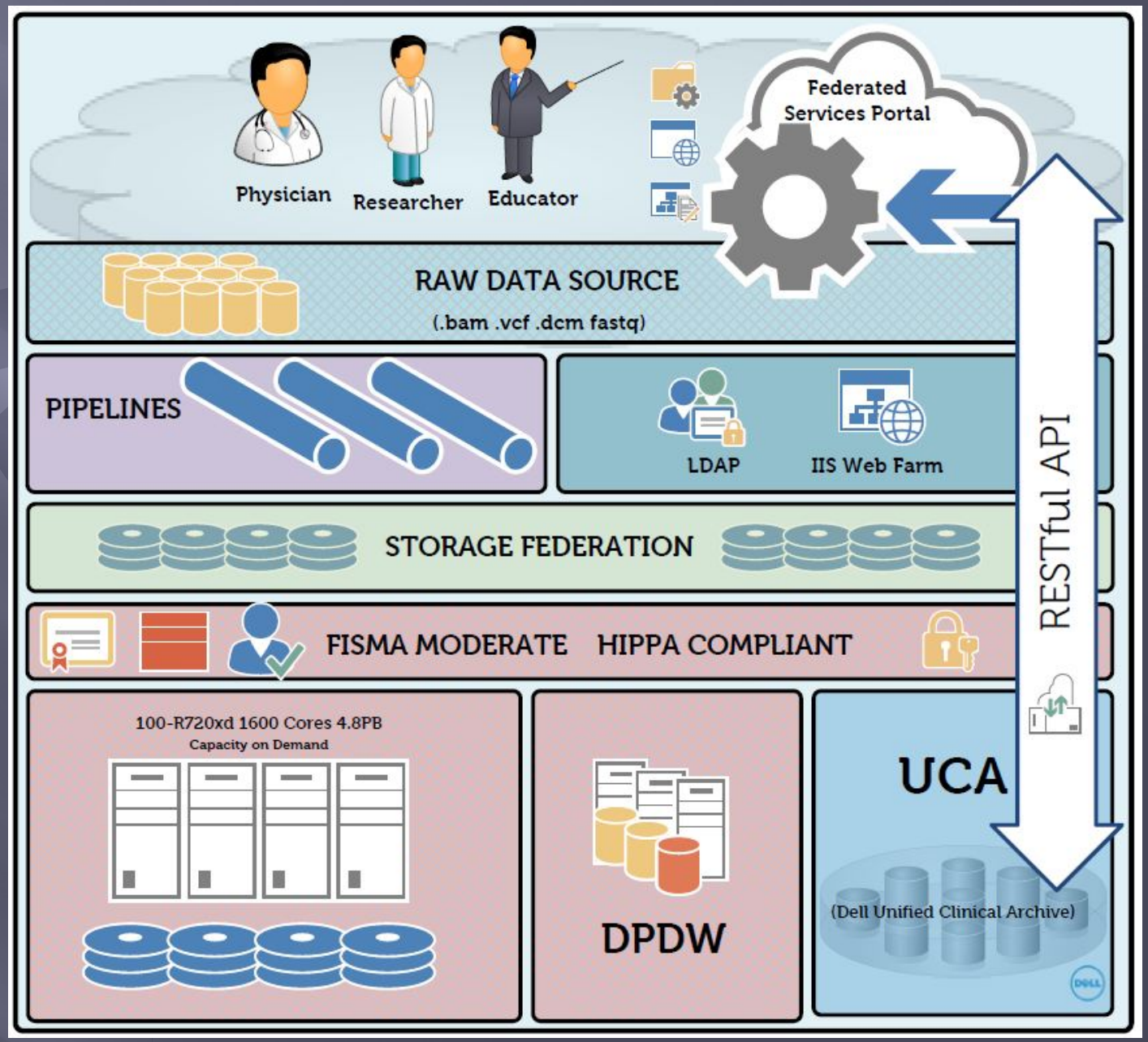
One potential solution is the much discussed “Bioinformatics as a Service” model. This model requires sharing of data as a prerequisite to accessing data.

This model has already found success in the academic research model where researchers are required to share their results to access independent research data from their peers.

The graphic depicts one suggested model positioned to unify a communication method that would become *The Open Source glue for the Bioinformatics cloud.*

From the 2011 Bio-IT World conference,

Ken Buetow noted that "technology is transforming every area of the economy while we in biomedicine are still pretty much a backwater." It's not that technologies don't exist, but rather that they exist in isolation. The industry is an "interconnected collection of different sources of information," from electronic health records and social media to wireless devices and smart phones. No single source holds all the data.



http://cabig.cancer.gov/resources/news/
http://www.jayetchings.com/datamanagementmodels

Unlocking the Power of Personalized Medicine

First Generation Data Science Research Instrument

”One may then fairly question the sanity of the biomedical enterprise – stuck by complex forces in existing paradigms we continue to hope for new outcomes such as personalized medicine”

Dr. Kenneth Buetow PhD. | Genomicist

- Creation of the Next Generation Cyber Capabilities Platform
- Advanced Genomic, Proteomic analysis on an Open Data Platform
- 2K+ server cores combining HPC and Big Data in one Ecosystem
- Scalable solution supports 100% annual growth in data volume

2014 Big Data Impact Awards Nominee

Taking aim at pediatric cancer

Customer-inspired innovation

"With diseases like neuroblastoma, hours matter. Our new Dell HPC cluster allows us to do the processing we need to get a meaningful result **in a clinically relevant amount of time.**"

Jason Corneveaux, Bioinformatician

- 12-fold improvement in processing power for patient data
- Reduced genomic analysis time from 7 days to a few hours
- 800 server cores managed by one IT administrator
- Scalable solution supports 100% annual growth in data volume

Future of healthcare

- Go beyond treating symptoms
- Integrate phenotypic and genotypic data
- Personalized care and preventative medicine strategies
- Drive efficiencies, increase speed
- Improve cost benefit and patient satisfaction

Questions?

@JayEtchings

Jay Etchings