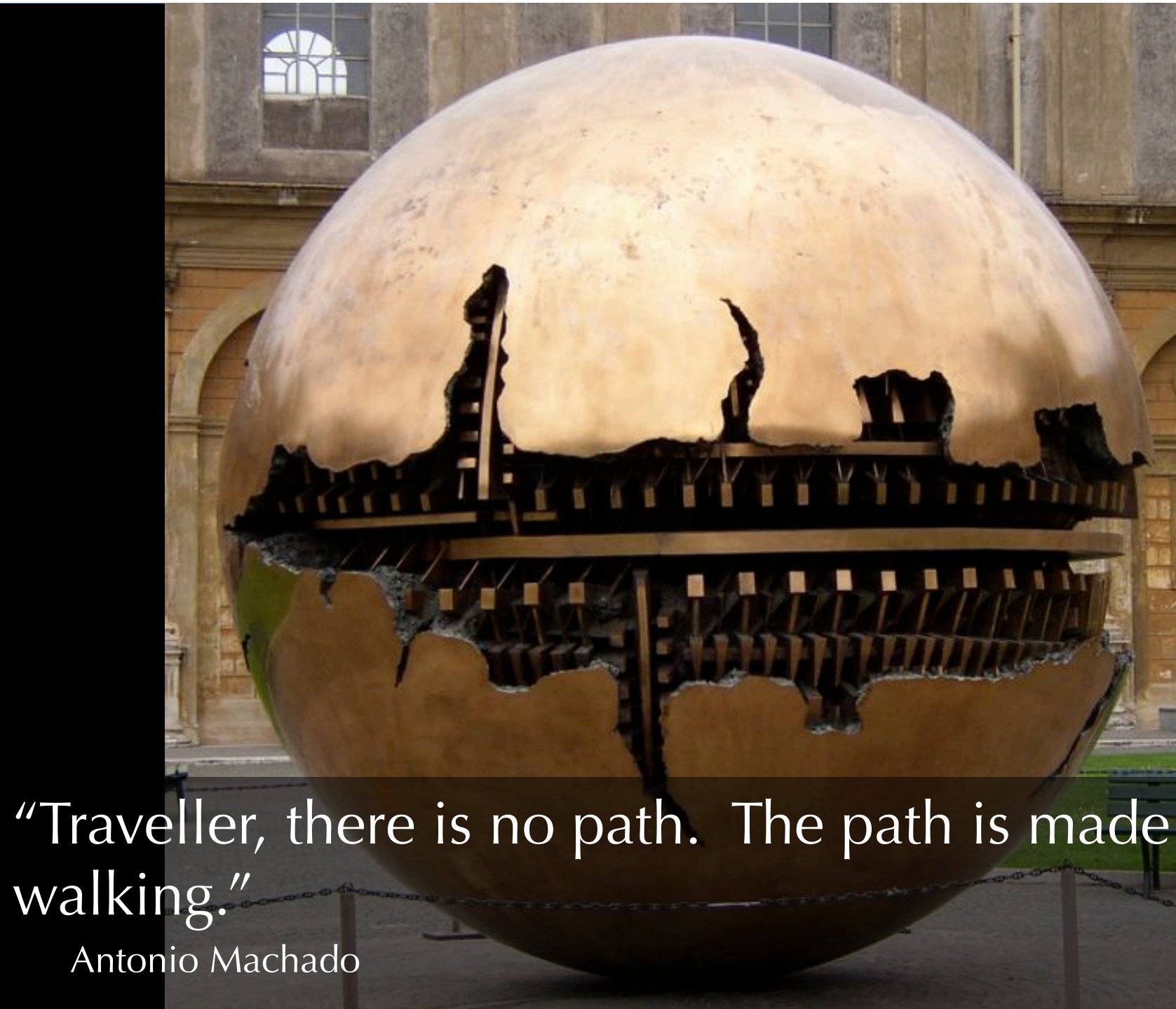

Issues in Genomics At Scale

2012 Consumer Genetics Conference

Chris Dwan cdwan@bioteam.net
Bioteam (<http://bioteam.net>)





“Traveller, there is no path. The path is made by walking.”

Antonio Machado

Issues in Genomics at Scale

- **Data:**
 - Networks, Repositories, Management
 - Data Lifecycle: Production, management, destruction
- **Regulatory Environment:**
 - CLIA / HIPAA
 - Fear, Uncertainty and Doubt
- **The future**
 - Data access via flexible APIs is the single most important thing we can do
 - Infrastructure as a Service (IaaS) and Software as a Service (SaaS)
 - The legislative and regulatory context is terrifying.

INTRODUCTION

The BioTeam Inc.

Independent Consulting Shop

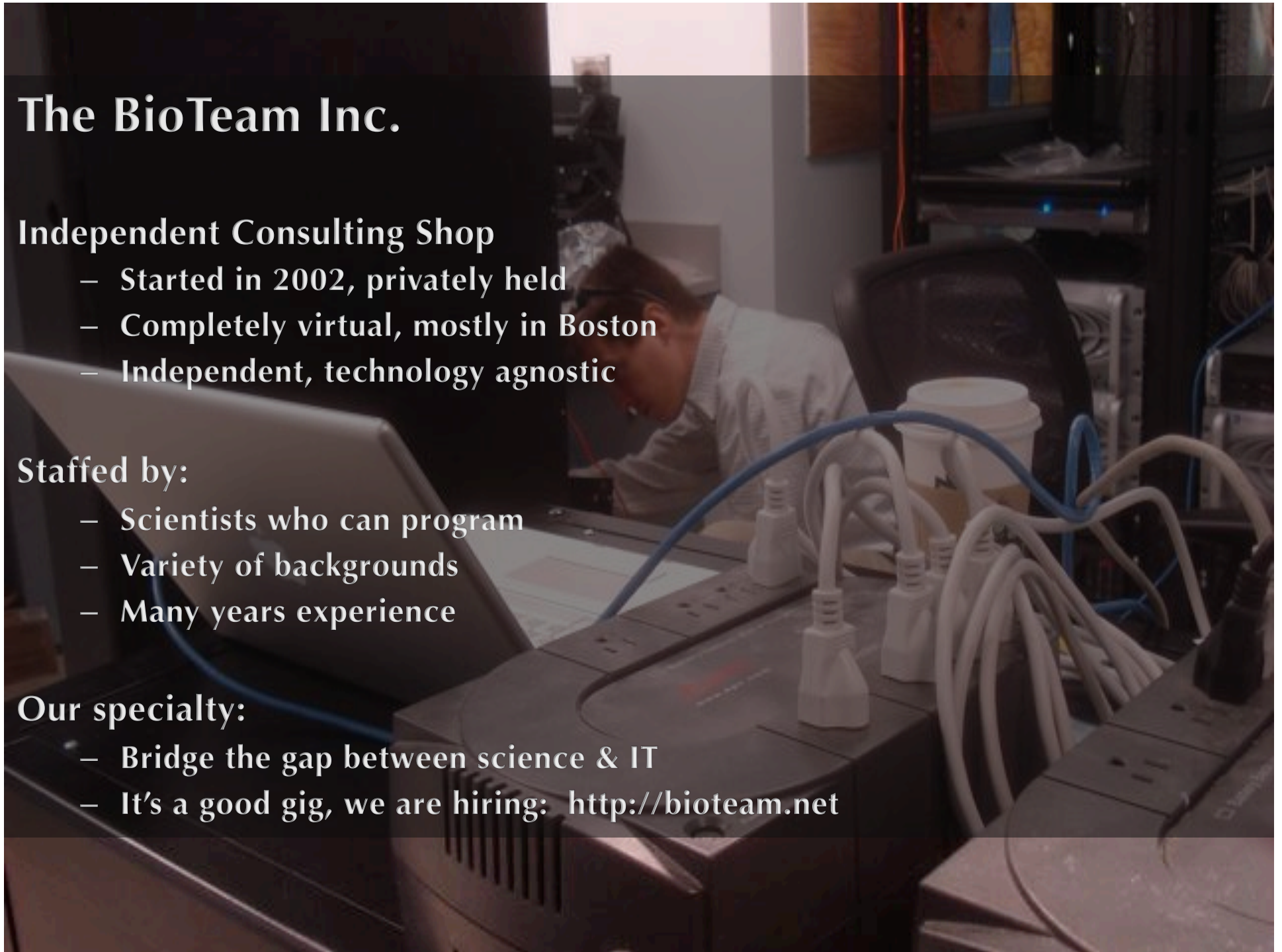
- Started in 2002, privately held
- Completely virtual, mostly in Boston
- Independent, technology agnostic

Staffed by:

- Scientists who can program
- Variety of backgrounds
- Many years experience

Our specialty:

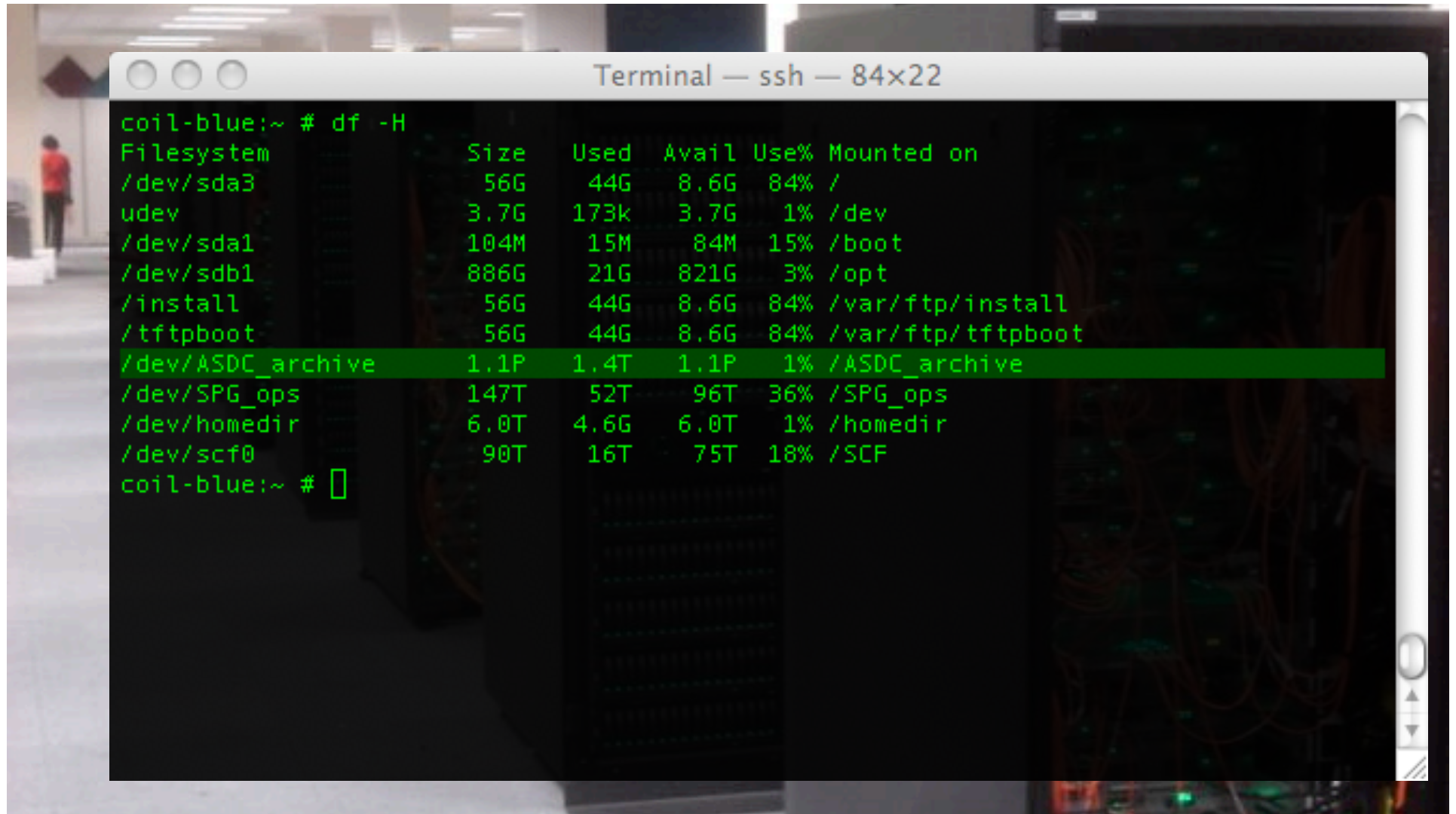
- Bridge the gap between science & IT
- It's a good gig, we are hiring: <http://bioteam.net>



Geek Cred: My First Petabyte, 2008



Geek Cred: My First Petabyte, 2008



Geek Cred: 23andme, 2008



Consumer Genetics Experiences

- **2008:**
 - 23andme (SNP v1)
 - Coriell
- **2009:**
 - 23andme (SNP v2)
- **2012:**
 - 23andme (exome)
- Personal Genome Project, Google Health, ...

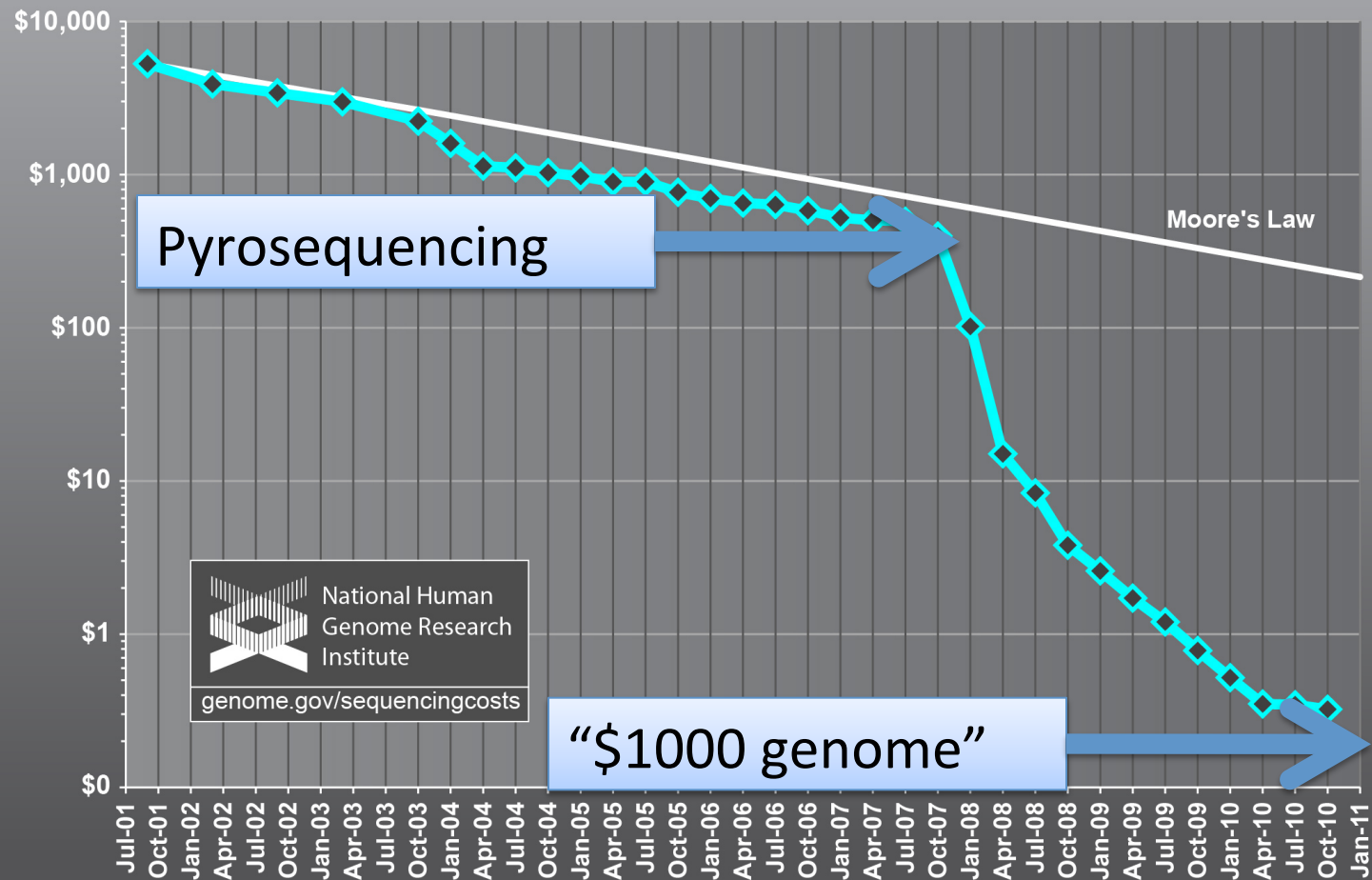
Shoutout: Mike Cariaso (Thursday, 4pm: SNPedia / Promethease)

New York Genome Center

- Collaborative effort between 12 New York institutions
 - >> \$1 x 10⁸ capital raised from academic, corporate and philanthropic sources
 - Novel collaboration between Clinical, Academic, Pharma research
- Potentially the largest facility of its kind in North America
 - Initial footprint of 36+ HiSeq 2000 instruments, scaling to 100+ by 2017
 - 500+ in-house staff, heavily weighted to bioinformatics
 - Initial service offerings in mid 2012
 - Manhattan facility to open in mid 2013
- It's a good gig, we are hiring: <http://nygenome.org>

GENOME SEQUENCING CONTEXT

Cost per Megabase of DNA Sequence



454 Instrument



Linux Server
under the table

Illumina and SOLiD Instruments



Linux Cluster
under the table

Ion Torrent PGM



Linux server (TorrentServer)
somewhere nearby ...

HiSeq 2000



Servers and storage somewhere

...

DNA Sequencing Is Not The Point

Genome sequencing is going in two directions:

- Very small (close to the patient)
- Very large (genome centers)

The smaller “consumer” model DNA sequencers (MiSeq, PGM, ...) will be used by people who do not fundamentally care about DNA sequencing.

- Time to result governed by physical locality
- Once we scrape the matter away from the data, it is **vastly** more useful.

I consider the IT infrastructure, as well as the DNA sequencing operation itself, to be a necessary evil on the road to transforming health care and improving people's lives.

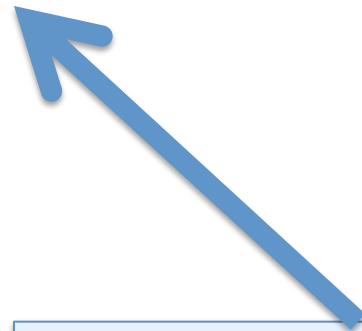
BIG DATA

Issues With Data

- **Repositories:** Can I store it?
- **Networks:** Can I move it / access it?
- **Organization:** Can the end user access it?

Issues With Data

- **Repositories:** Can I store it?
- **Networks:** Can I move it / access it?
- **Organization:** Can the end user access it?



I have no good answer
to this question



Genome Scale Data: The Sky is NOT falling

- ~130GB per “human genome” in raw BAM format
 - 3×10^9 base pairs in a human genome
 - 30x coverage (short reads)
 - 1.25 bytes per base pair (BAM file, including quality information)
 - Nobody stores raw TIFF images anymore
 - Emerging technologies will achieve massive reductions (CRAM)
 - End users actually want interpretation anyway, not reads.
- NYGC:
 - Several batches up to 1,000 genomes this year
 - Exome and RNA-SEQ are “genome lite” in terms of data volume
 - **Integrated offering with analysis and multi-year data storage**

Data access via API is vastly superior to mere data *delivery*

Genome Scale Data: The Sky is Sorta Falling

- **Flat file hierarchies are not going to cut it anymore**
 - Lightweight, semantic databases for metadata
 - Number of files rather than data volume
 - Data access must be compartmentalized by more than just file permissions.
- **Sample management / retention is a nightmare**
- **Data retention policies will eventually become a nightmare**
 - When data are created, they must have an expiration date.
- **The concept of “backup” is a sick joke anymore**
 - Of course we'll have tiered storage. What's a “backup” anyway?

2008, Custom Storage Build



2012: Fresh Storage from the vendor.

2012: Deploying petabyte scale storage is an exercise in requirements, purchasing, and project management.

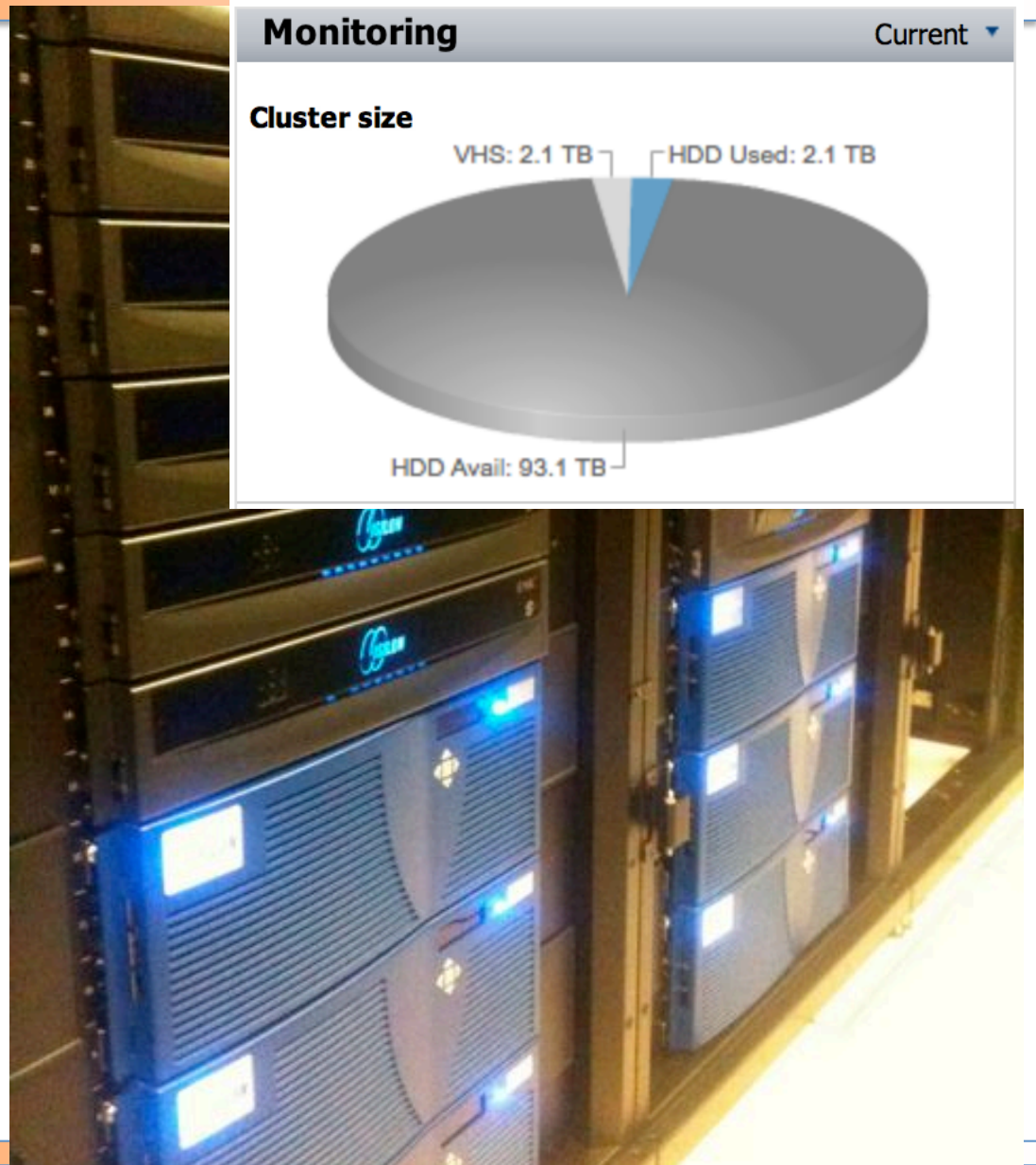
If you find yourself designing custom storage, you are probably doing it wrong.



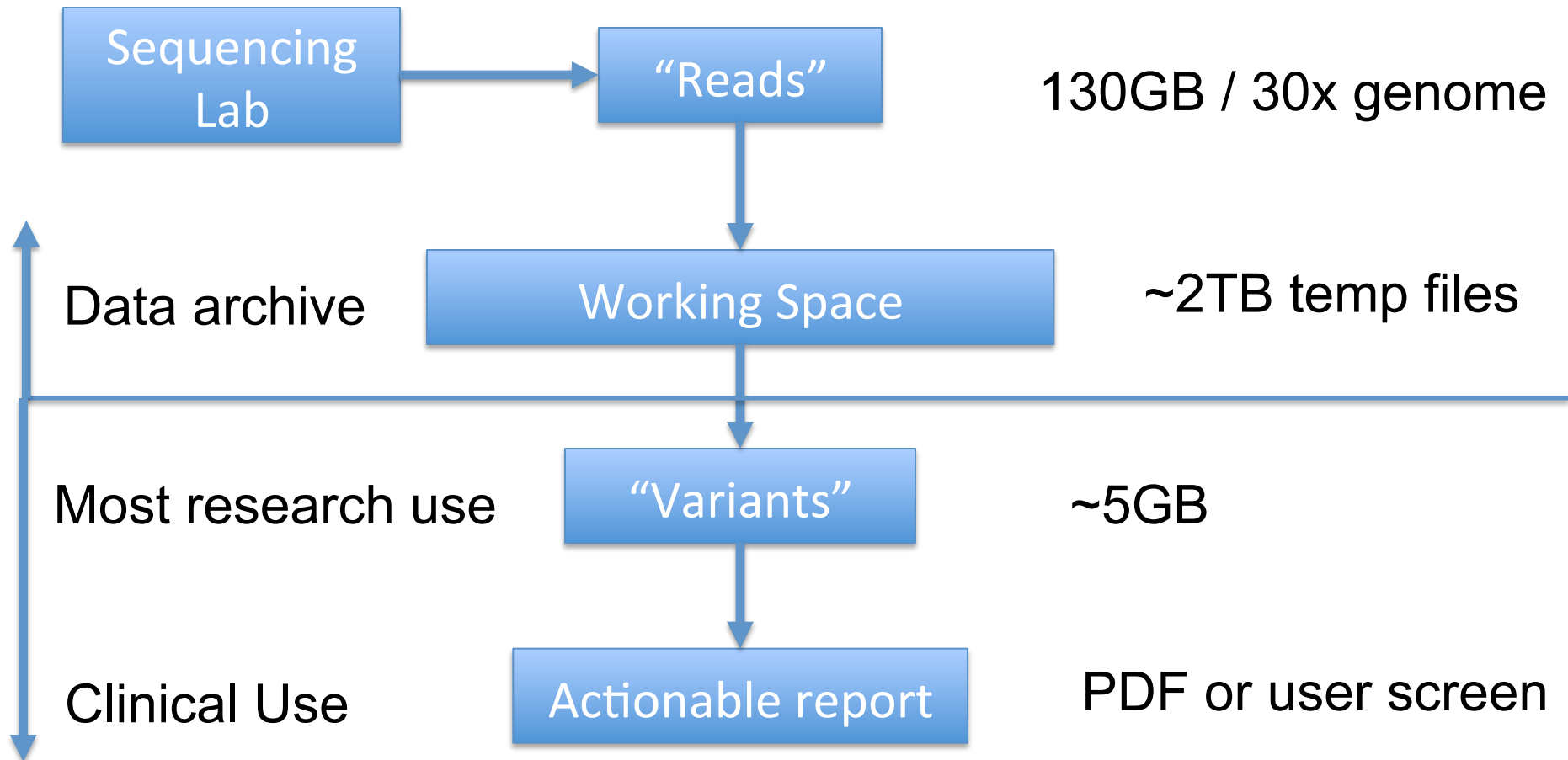
2012: Raw Storage

2012: Deploying petabyte scale storage is an exercise in requirements, purchasing, and project management.

If you find yourself designing custom storage, you are probably doing it wrong.




Genome Sequencing Data Flow



Data Bandwidth

Bandwidth	1 Gigabyte	1 Genome (130GB)	Genomes / day	HiSeq 2000 daily raw output (55GB/day)
T1 business link (12Mb/sec)	11m 22s	24.6h	1	2
T3 business link (45Mb/sec)	3m 10s	6.9h	3.4	9
700Mb/sec	11s	24m	60	134
Gigabit	8 sec	17m	84	192



If we can make full use of the available bandwidth, gigabit networking is sufficient for the long term data motion needs of a large genome center

Aspera Data Motion (coast to coast)



Shipping disks via Fedex

- Assume a 48 hour point to point latency
 - Two sets of checksums
 - Sneakerbot / human interaction on either side
- Very low potential for automation
- Many opportunities for error
- Sadly, this is the state of the art in many places



Data Lifecycle

“Data flows downhill” (Jeff Hammerbacher)

- Plan for data management **before** generating it
- Move data *once* to its final resting place, to a cheap location close to (or colocated with, cf. Hadoop) some decent computing capability.
- Be flexible: Move data to computing, computing to data, or user to both.

Organization, big data analytics, etc are open questions.

CRAM

Fritz, Leinonen, Cochrane, and Birney. “Efficient storage of high throughput DNA sequencing data using reference-based compression”

<http://genome.cshlp.org/content/21/5/734.long>

Significant reduction in long term file storage requirement by semi-lossy compression.

- Quality score bins

- Delta-reference compression

- Aligned read “stacking”

Infrastructure as a Service (IaaS)

- **Please stop saying “cloud” like the word means something.**
 - Clouds are built from servers and storage
 - If you’re going to engineer using them, you have to know what they’re made of.
- Amazon’s S3 data storage is more robust and reliable than what you can build.
 - 99.999999999% durability
 - 500,000 requests per second
- HIPPA / CLIA certifications are possible
 - Don’t let the Fear Uncertainty and Doubt crowd tell you otherwise.
- It is still engineering, but high levels of security and operational availability are possible.

PRIVACY AND REGULATION

“I believe education and legislation aimed less at protecting privacy and more at preventing discrimination will be key”

Eric Schadt, September 11, 2012

Health records, google style

The screenshot shows an XML editor interface with three main panels: Project, Outline, and the main XML view.

Project Panel: Displays a file tree for 'sample.xpr' with subfolders: css, debugger, epub, fo, import, json, jsp, nvd1, and relaxng. A message states 'The Master Files support is disabled' with an 'Enable' button and a 'read more...' link.

Outline Panel: Features an 'Element name filter' and a list of elements: link 'self', link 'edit', author Blue Cross Blue Shield of MA, and ContinuityOfCareRecord 'urn:as'. The ContinuityOfCareRecord element is expanded, showing CCRDocumentObjectID Wb1 and Language English.

Main XML View: Displays XML code for a health record. The code includes a header with ActorRole, Actor, Source, and Actors. The main entry is a ContinuityOfCareRecord with the following details:

- entry** gd:etag="W/"DkEMRHczcSp7ImA9WxJbFkU."";
- id** https://www.google.com/health/feeds/profile/default/zElfDLktsYM
- published** 2009-07-27T08:51:25.989Z
- updated** 2009-07-27T08:51:25.989Z
- app:edited** xmlns:app="http://www.w3.org/2007/app" 2009-07-27T08:51:25.!
- category** term="MEDICATION"
- category** scheme="http://schemas.google.com/g/2005#kind" term="http://schemas.google.com/health/kinds#profile"/>
- title** Health entry zElfDLktsYM
- link** rel="http://schemas.google.com/health/data#complete" type="applic:
- href** "https://www.google.com/health/feeds/profile/default/-/MEDICATION/%7Bhttp%3A%2F%2Fschemas.google.com%2Fhealth%2Fkinds%23profile"/>
- link** rel="self" type="application/atom+xml" href="https://www.google.com/health/feeds/profile/default/zElfDLkts!"
- link** rel="edit" type="application/atom+xml" href="https://www.google.com/health/feeds/profile/default/zElfDLkts!"
- author** <name>Blue Cross Blue Shield of MA</name> <uri>Blue Cross Blue Shield of MA</uri>
- ContinuityOfCareRecord** xmlns="urn:astm-org:CCR" <CCRDocumentObjectID>lrb_cx6h_GIRX8p0470L8QzE1fDLktsYM</CCRDocumentID>
- Language** <Text>English</Text> <Code> <Value>en</Value> <CodingSystem>ISO-639-1</CodingSystem>
- Version** V1.0

Thanks, Google Health

```
.LINK REF- edit type- app
  href="https://www.google.com/health/record/continuityOfCareRecord.xml"
author>
  <name>Chris Dwan</name>
  <email>chrisdwan@gmail.com</email>
author>
continuityOfCareRecord xmlns="http://hl7.org/v3"
  <CCRDObjectID>S
```

Regulation

HIPAA (1996)*: “Health Insurance Portability and Accountability Act”

Deals with data *privacy* and *protection*.

The fact that this law includes the word “insurance” should terrify you.

**Please note spelling*

CLIA

Clinical Lab Improvement Amendments (1988)

“A laboratory is any facility that does laboratory testing on specimens derived from humans to give information for the diagnosis, prevention, treatment of disease, or impairment of, or assessment of health.”

Three semi-independent certifications:

- Clinical director of record
- Laboratory
- Assay

Fundamentally, this is a *process* certification like ISO.

Additional thoughts

The current laws governing genetic and genomic data are severely out of date and ignore fundamental realities

- Genome data cannot be anonymized
- New laws could easily be written based on fear, uncertainty, and doubt.
- It is incumbent on us to communicate thoughtfully and honestly

Additional thoughts

The current laws governing genetic and genomic data are severely out of date and ignore fundamental realities

- Genome data cannot be anonymized
- New laws could easily be written based on fear, uncertainty, and doubt.
- It is incumbent on us to communicate thoughtfully and honestly

Costs that drive to zero over time

- Data storage, Networking, computing
- DNA sequencing

Additional thoughts

The current laws governing genetic and genomic data are severely out of date and ignore fundamental realities

- Genome data cannot be anonymized
- New laws could easily be written based on fear, uncertainty, and doubt.
- It is incumbent on us to communicate thoughtfully and honestly

Costs that drive to zero over time

- Data storage, Networking, computing
- DNA sequencing

Costs that do not drive to zero:

- Novel analysis
 - Patient care
 - Regulatory compliance
-

The Future

- **Patient clinical data *must* be owned by the patient**
 - Germ line genetic information will be as ubiquitous as blood type
 - We must not re-sequence everyone to overcome the failures of our data access policies
- **Access to data and analysis *cannot* be monolithic**
 - Data to compute, compute to data, user to both.
 - Focus on networks and interfaces, not a single massive data archive
- **Certification of analysis pipelines**
 - Leverage Infrastructure as a Service and Software as a Service.

API based access to genome scale data and analysis tools is the key.



end;