Outer Space:
Source -- Cosmic Microwave Background (CMB): the oldest light in our universe, Planck spacecraft, 2013 (Copyright: ESA and the Planck Corporation)

The Clinical Galaxy
Validation Plan proposal

Sanjay Joshi
CTO Life Sciences, EMC² Isilon Storage Division

Inner Space:
STORY IN 2 PARTS:

- Rationale
- Proposal
**Population Explosion**

~5000 to ~1000 years ago

→ Rare Variants


---

**Effect vs. Variant Freq.**

**Clinical Effect “chasm”**

**Synthetic Causality?**

- Mendelian diseases
- Family-based approach
- Common diseases
- Rare variants
- Substantial effect size
- Common variants
- Small effect size

*Personal genome (exome and whole genome resequencing)*

*Genome-wide association studies (GWAS)*

**Variant frequency**

**Very rare**

**Common**
~7 \times 10^9 \text{ humans}

\text{range\{1, 240\} \times 10^{12} \text{ cells/human}^*}

\text{range\{0.07, 16\} \times 10^{23} \text{ human cells}}

0.7 to 160 \times 10^{23} \text{ human microbes on earth?}

\text{Avogadro Number : 6.023 \times 10^{23}}

p-values do not work*

In a Population of $7 \times 10^9$, with Confidence Interval of 0.025, Confidence Level of 95% and 100 degrees (genes) of freedom, Sample Size = ~15 million


*Ziliak ST and McCloskey DN, "The Cult of Statistical Significance", Section on Statistical Education, JSM 2009
Data Center Tradeoffs

- **High Availability, Archival**
- **CAPEX vs. OPEX**
- **Temp-Space: Local vs. Scale-Out**
  (define “writes”, understand process)
- **CPU + Acceleration on same node**
- **CPU speed control, High-density disk nodes**

Higher Priority

Lower Priority

Derived from:
Stuart Feldman, Google

Data Intensive Workflows
OPEX

Genomics

Storage: Choose Two

Cost

Performance

Speed
Validation Initiative Proposal
Data Management

The Galaxy philosophy:

• Data is never overwritten
• Data is never deleted
MOTIVATION
Private & Community Cloud?


<table>
<thead>
<tr>
<th>Public</th>
<th>Infrastructure Managed By</th>
<th>Infrastructure Owned By</th>
<th>Infrastructure Located</th>
<th>Accessible and Consumed By</th>
</tr>
</thead>
<tbody>
<tr>
<td>Third Party Provider</td>
<td>Third Party Provider</td>
<td>Off-Premise</td>
<td>Untrusted</td>
<td></td>
</tr>
</tbody>
</table>

| Private/Community | Organization | Third Party Provider | On-Premise | Trusted |

| Hybrid          | Both Organization & Third Party Provider | Both Organization & Third Party Provider | Both On-Premise & Off-Premise | Trusted & Untrusted |

15 hacks per day in 2012**
versus 10 in 2011
** Ponemon Institute Oct 2012

79% of respondents are concerned about Cloud Security*
*Feb 2012

21M data breaches in 18-month period 2010-11 ($2.25M/breach)
- US Govt. Office of Civil Rights, HHS
- £1.79B fines in UK for NHS 2012 breaches

4M Common Variants: Trace DNA

Veracity

Y-STR genome
Surnames from ancestry data

Scientific Reports, 3, Mar 2013

PGP Encrypted De-ID reverse engineered +
+ Data Privacy Lab, CMU, Apr 2013

4 position points (GPS) determine identity
### Access Controls

<table>
<thead>
<tr>
<th>Regulations</th>
<th>Functions</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>45CFR 164.312(a)</strong></td>
<td><strong>System Access:</strong> Unique Name/Id, Role-based perms, Single Sign-On SAML</td>
<td>FDA, HIPAA, AES, SAML, WS-Trust</td>
</tr>
<tr>
<td><strong>21CFR 11.10(b)(d)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>System Entry:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45CFR 164.312(d)</td>
<td><strong>System Entry:</strong> Enterprise User Auth. (EUA)</td>
<td>FDA, HIPAA, ATNA, CLIA-CAP</td>
</tr>
<tr>
<td>IETF RFC 4120</td>
<td>Cross Enterprise User Auth. (XUA)</td>
<td></td>
</tr>
<tr>
<td>IHE-ITI-TF</td>
<td>Logging</td>
<td></td>
</tr>
<tr>
<td>MOL.34960, 34968, 34970</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Audit Controls

<table>
<thead>
<tr>
<th>Regulations</th>
<th>Functions</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>45CFR 164.312(d)</strong></td>
<td><strong>System Entry:</strong> Enterprise User Auth. (EUA)</td>
<td>FDA, HIPAA, ATNA, CLIA-CAP</td>
</tr>
<tr>
<td>IETF RFC 4120</td>
<td>Cross Enterprise User Auth. (XUA)</td>
<td></td>
</tr>
<tr>
<td>IHE-ITI-TF</td>
<td>Logging</td>
<td></td>
</tr>
<tr>
<td>MOL.34960, 34968, 34970</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Data Integrity

<table>
<thead>
<tr>
<th>Regulations</th>
<th>Functions</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>45CFR 164.312(c)</strong></td>
<td><strong>Data @rest:</strong> Encryption, Key Management</td>
<td>HIPAA, SHA2, CLIA-CAP, ASTM-Auth</td>
</tr>
<tr>
<td>FIPS PUB 180-2 SHA-224</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASTM Std E1762-95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MOL.34966</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### De-ID, Re-ID

<table>
<thead>
<tr>
<th>Regulations</th>
<th>Functions</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>45CFR 164.312(d)</strong></td>
<td><strong>PHI:</strong> Code for de-ID, Code for re-ID, Pedigree for Genomics</td>
<td>HIPAA de-ID, HIPAA re-ID, Pseudo-randomization, HL7v3 Pedigree</td>
</tr>
<tr>
<td>IETF RFC 4120</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IHE-ITI-TF</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Transmission Security

<table>
<thead>
<tr>
<th>Regulations</th>
<th>Functions</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>45CFR 164.312(d)</strong></td>
<td><strong>Security:</strong> HTTPS (web), Crypto Message Syntax, Compression</td>
<td>FDA, HIPAA, SHA2, TLS, Compression, TLS, CLIA-CAP</td>
</tr>
<tr>
<td>FIPS 180-2, 197</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IETF RFC 2246, 3546, 2630, 3852</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MOL.34972</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
VALIDATION PROCESS

Story in 4 steps:
- Clinical Use Survey
- Architecture Review
- Quality Systems Review
- Validation
STEP 0: Clinical Use Survey

TBD
STEP 1: Architecture Review
Architecture Review

- Network Security
- OS Hardening
- Deny all first (user/install), allow and use only when needed
- Applications config. (Galaxy, PostgreSQL, Apache, ...)
- Physical server access (BIOS, GRUB, SSH, ...)
- Internet access protocols (SSH, sftp, API...)
STEP 2: Quality System Review
Quality System Review (QSR)

- Logging and Server Management
- Risk Management
  Business, Functional and Application
- Human factors
  Training, SOPs, Change Control
- Software Development Life Cycle
- IT Processes
  “The Cloud is your mess managed by someone else”*

*Source: Halamka J., “Life as a Healthcare CIO” blog
STEP 3: Implement
STEP 4: Validate

TBD
The Quantified Patient

- Demographics
- Genetics
- Family History
- Imaging
- Medications
- Clinical Narratives
- Labs
- Molecular Diagnostics
- Environment
- Sensors

Available at Point of Care
Available Data
Not Available

Source: EMC Healthcare

THANK YOU!