SAND2017-3497C

Building cyberthreat models around genomic security

COREY M. HUDSON

SANDIA NATIONAL LABORATORY

SANDIA NATIONAL LABORATORIES IS A MULTI-PROGRAM LABORATORY OPERATED BY SANDIA CORPORATION, A WHOLLY OWNED SUBSIDIARY OF LOCKHEED MARTIN CORPORATION, FOR THE U.S. DEPARTMENT OF ENERGY'S NATIONAL NUCLEAR SECURITY ADMINISTRATION UNDER CONTRACT DE-AC04-94AL85000.

Genomics and Computing

Primary question of the talk: How has the innovation in Next-Gen sequencing and in Synthetic Biology affected our cybersecurity risk models?

Avoid giving a recipe for mayhem, while still illuminating the realistic risks.

Where does the biological community fit in?

What is a cybersecurity risk model?

Risk models

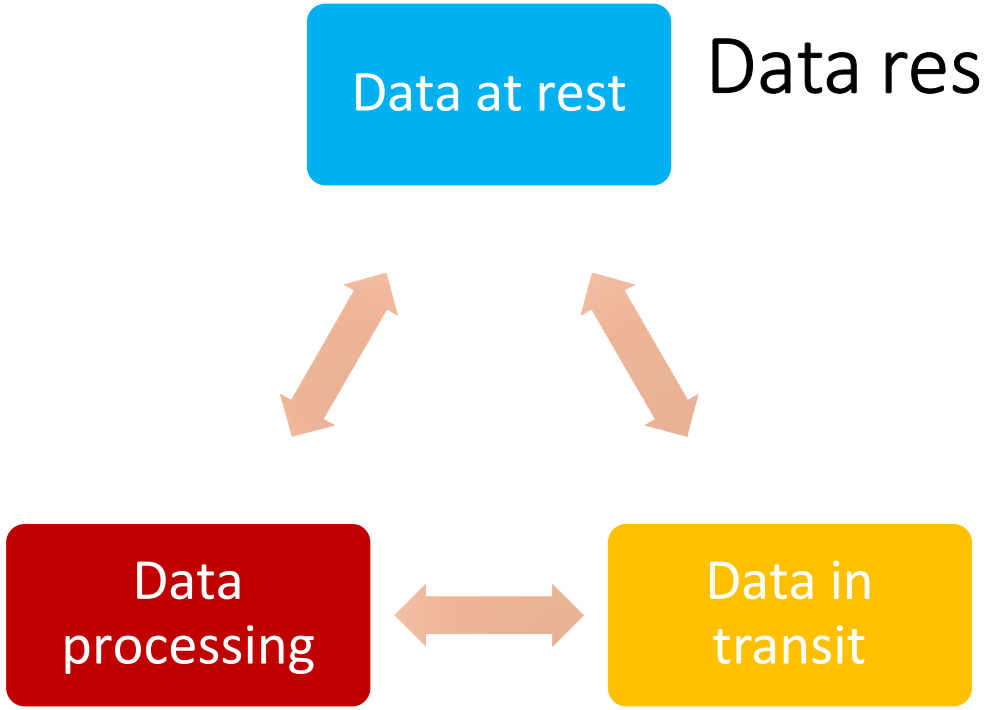
- Used to specify likely modes of attack
- Highlight areas for sensing attack
- Recovery mechanisms in the event of an attack.

System profiling

Computational resources

- Servers
- PCs
- Equipment Computers
- Cloud Resources
- Mobile Computing

Data resources



Equipment

Personnel

What are the risks?

Risk = Vulnerability x Threat x Impact x Probability

Risks at a facility doing **sequencing genomics**:

- Failure to complete work
- Release of protected data
  - Intellectual property
  - Personally identifying information
  - Secret information
- Destruction of data integrity
- Release of operational security and adversarial surveillance

What are the risks?

Risk = Vulnerability x Threat x Impact x Probability

Risks at a facility doing **sequencing genomics** – **Hacking equivalent**

- Failure to complete work – **DDOS**
- Release of protected data – **Man-in-the-middle exfiltration and hacking theft**
  - Intellectual property
  - Personally identifying information
  - Secret information
- Data manipulation – **Social Media Hactivism**
- Release of operational security and adversarial surveillance – **APT**

Issue at hand: Genomics has moved from a scientific technique to an industry

The <$1000 genome has changed sequencing into a consumable.

- Illumina's NovaSeq and new technologies being developed by Complete Genomics (BGI company) suggest this will be ~$100 by year's end.

The technology emerged before the safeguards or training in cyber-risks were in place.

Without an adequate threat model, automation exacerbates, rather than relieves the risk.

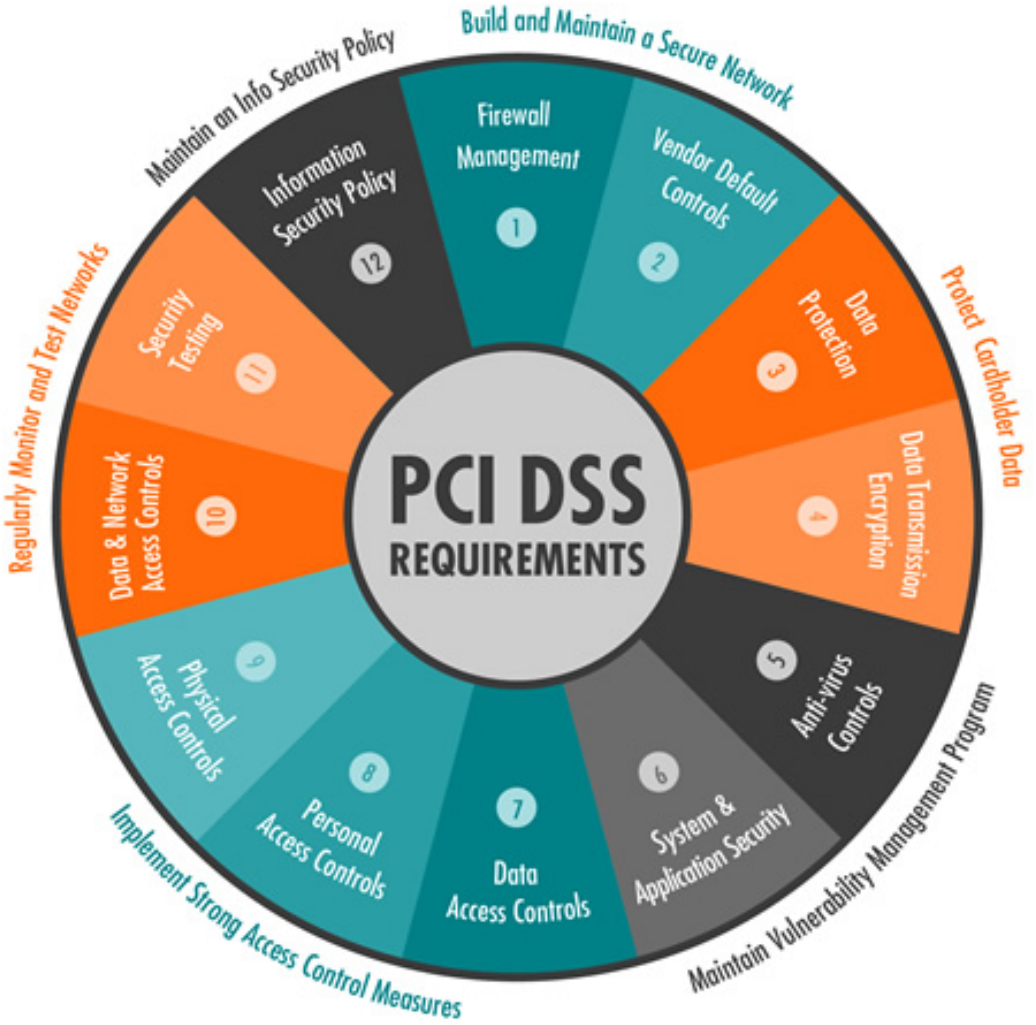


Image:
http://ryantech.com/

How are genomics data different?

Comparison to credit card data

CCs have an established threat model:

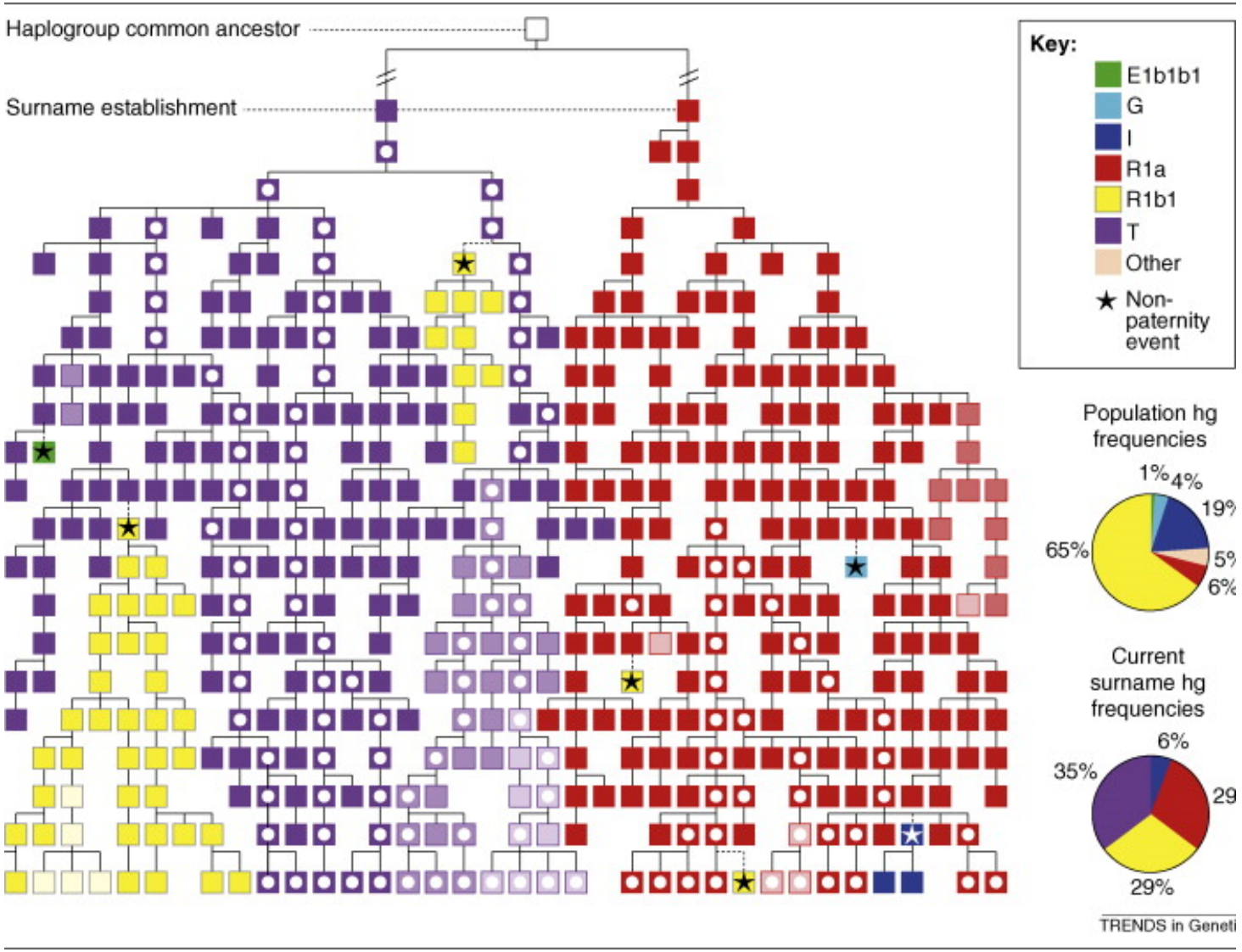
1) Secure data

2) Authenticate over encrypted network

3) Limit access

4) Regular vulnerability audit

5) Legal mechanisms for recuperation



King & Jobling, *Trends in Genetics* 2009 25, 351-360

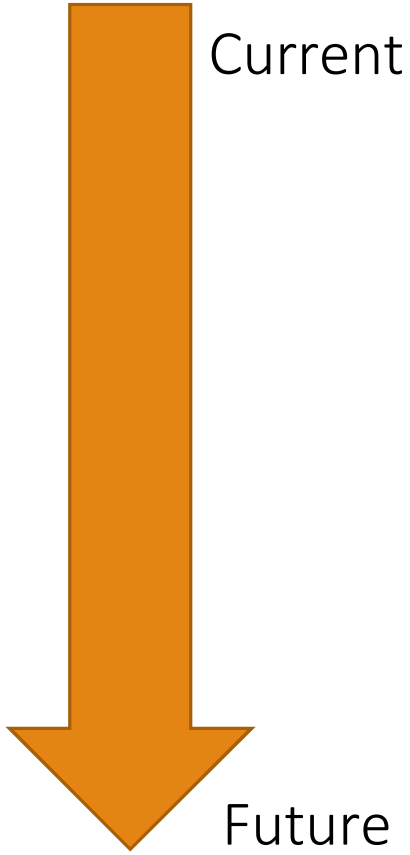
Genomic data are associational

Every leaked genome leaks data about associated family members.

Asymptotically, this means that genomic data cannot be secured indefinitely.

What is the risk space around privacy?

- Paternity breach
- Privacy and identification
- Racial or at-risk subgroup identification
- Legal/forensic identification/manipulation
- Phenotype inference
- Genomic access controls
- Genomic targeting



Recovering from genomic breach

Fundamental maxim of genomic data breach: There is currently no model for recovery from genomic data release.

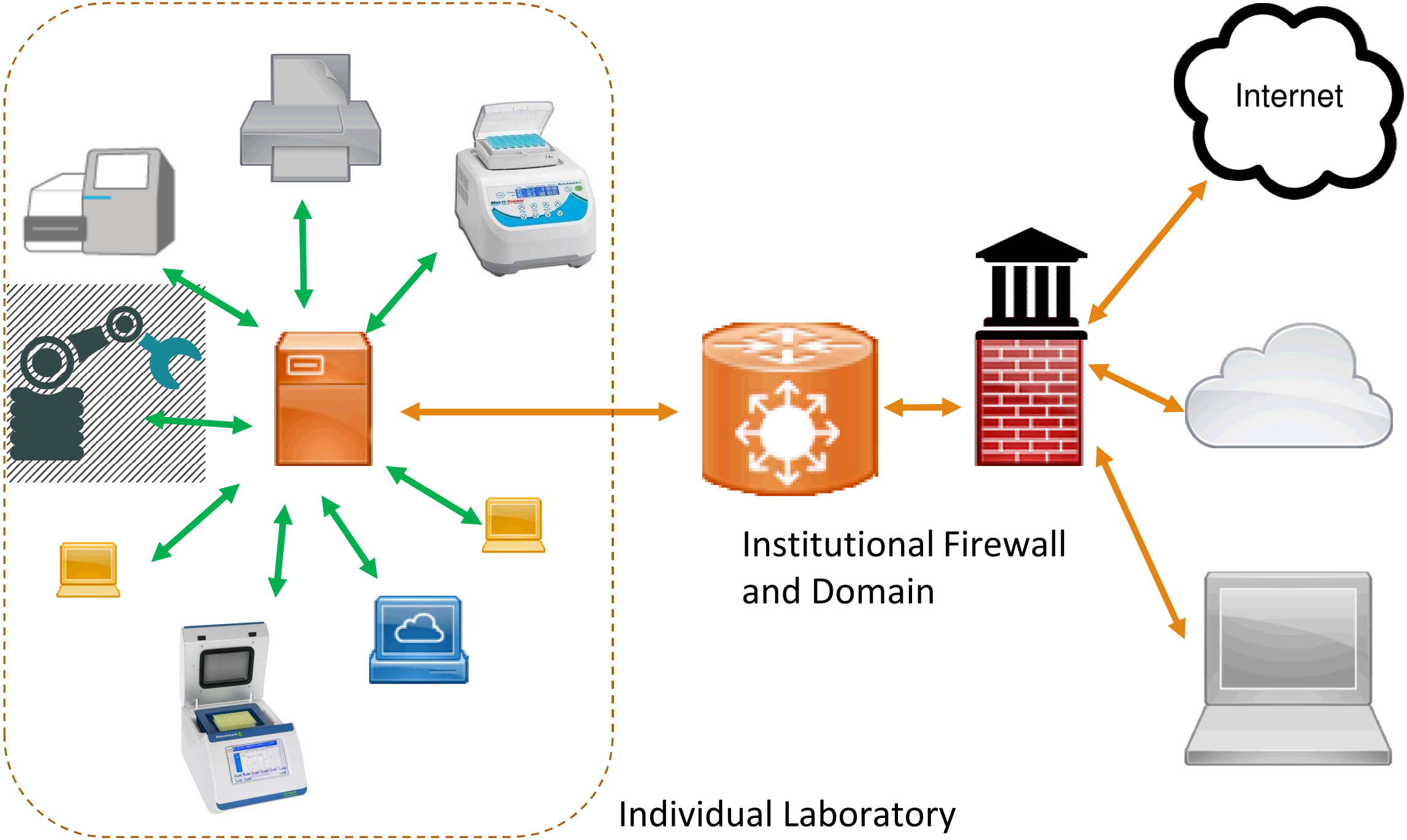
Genomic data are basically unchangeable through the life of the victim.

Leaked data can create new victims, through the associational nature of genomic data.

Threat model 1: Firewall & Forget

Mantra: Business as usual, data and operations are secure and compliant.

Use the security of the institution housing the system.


Internet
Institutional Firewall
and Domain
Individual Laboratory

Advantages

- Rolling your own security system is DANGEROUS
- Institutions have IT departments
- Creates and establishes Access Control protocols
- Maintains compliance with larger institution

Disadvantages

- Sensors are outside realm of activity
- Machine-to-machine communication is assumed secure
- IT department may not be appraised of the level of risk they have signed on to
- To facilitate work, personnel may open unsecure channels to bypass firewall
- Many modern Next-Gen sequencing tools require cloud access

Threat model 2: Security by obscurity

Mantra: The system is too idiosyncratic or unsophisticated to be hacked.

Bizarre names and interactions

Idiosyncratic security protocols

Advantages

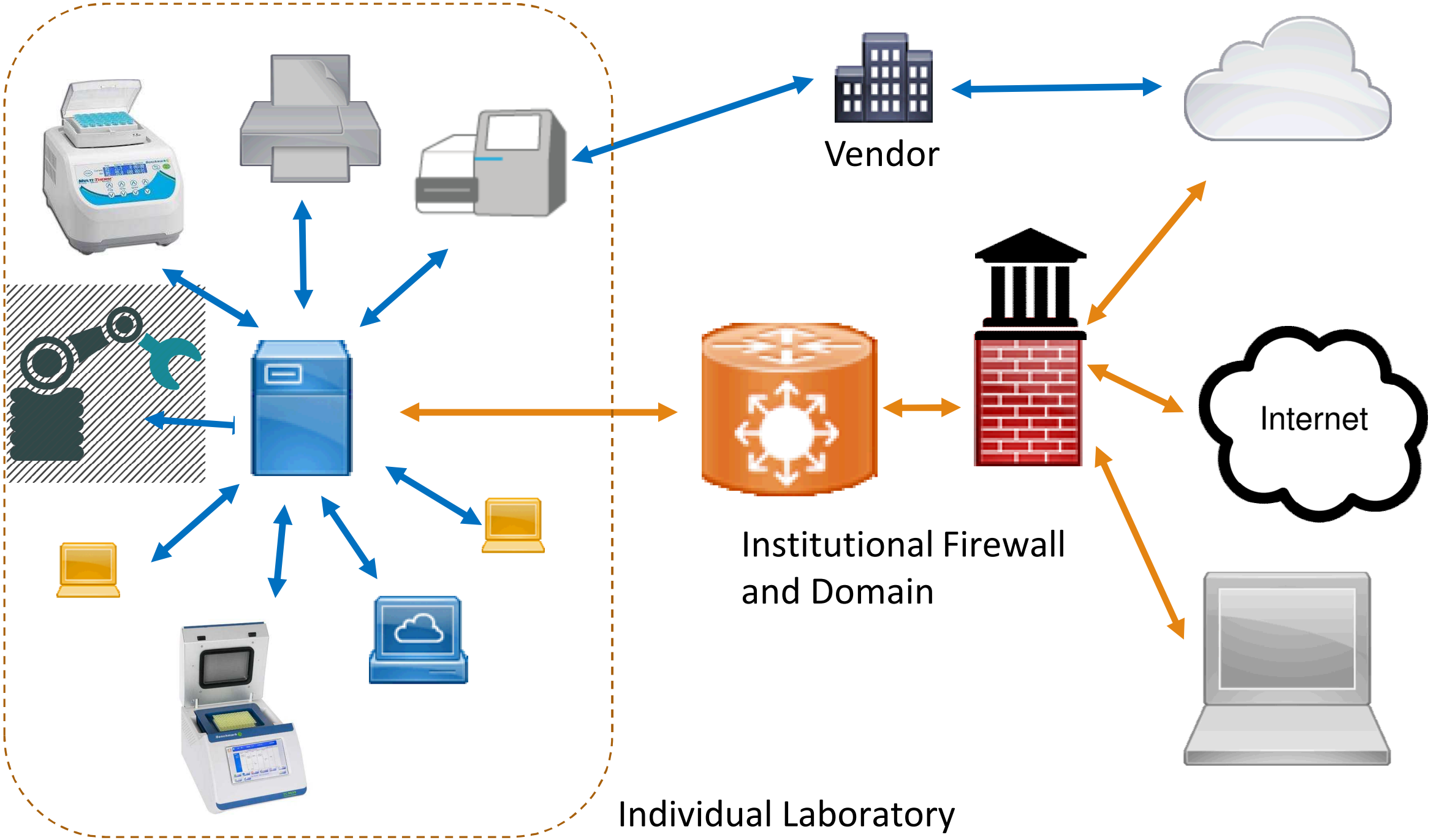
- Level of reconnaissance necessary to do damage may not be worth return
- Conscious thought about threats
- Layered security is generally preferred

Disadvantages

- Rolling your own security is DANGEROUS
- May leave open huge gaps
- Threat model likely not comprehensive
- May not be compliant with institution
- Likely will not have secure ports

Threat model 3: Leave it to the vendor

Mantra: The less the lab interacts with security, the lower the chance that they will wreck it.


Vendor
Internet
Institutional Firewall
and Domain
Individual Laboratory

Advantages

- Able to simultaneously handle lab and institutional compliance
- Vendor has best understanding of machines and potential insecurities
- Allows genomics specific threat model

Disadvantages

- Dependent on service agreement
- A lab may involve multiple vendors
- May not be compliant with data provider's specifications
- What happens after service agreement runs out?

Question: What is the optimal network setup that perfectly solves the sequencing security problem?

Answer: No….

All systems are vulnerable. One possible long term solution to sequence security is *fully homomorphic encryption (FHE)*.

FHE allows queries of encrypted data – but the data has to be fully assembled before it can be encrypted.

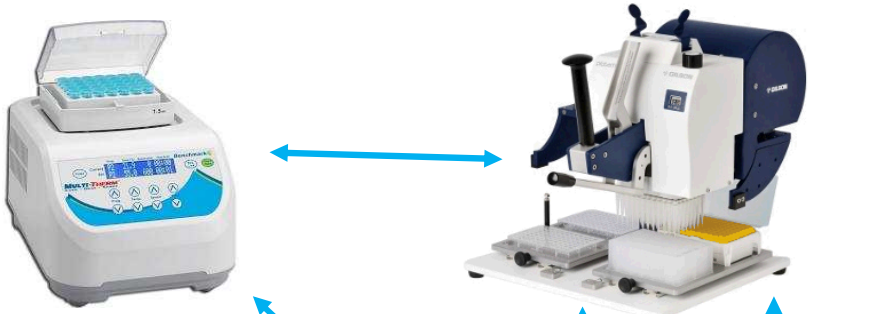
What are the cyber-risks in synthetic biology?

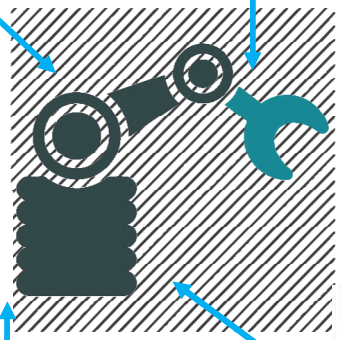
Risk = Vulnerability x Threat x Impact x Probability

Risks at a facility doing **genetic/genomic manufacture**:

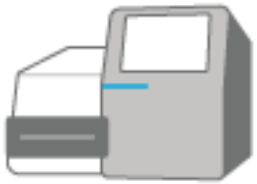
- Sequencing risks present in manufacture as well
- Unintended manufacture

Synthesis, Assembly and Cloning





Screening and Sequencing



Interface

BioThreat Database

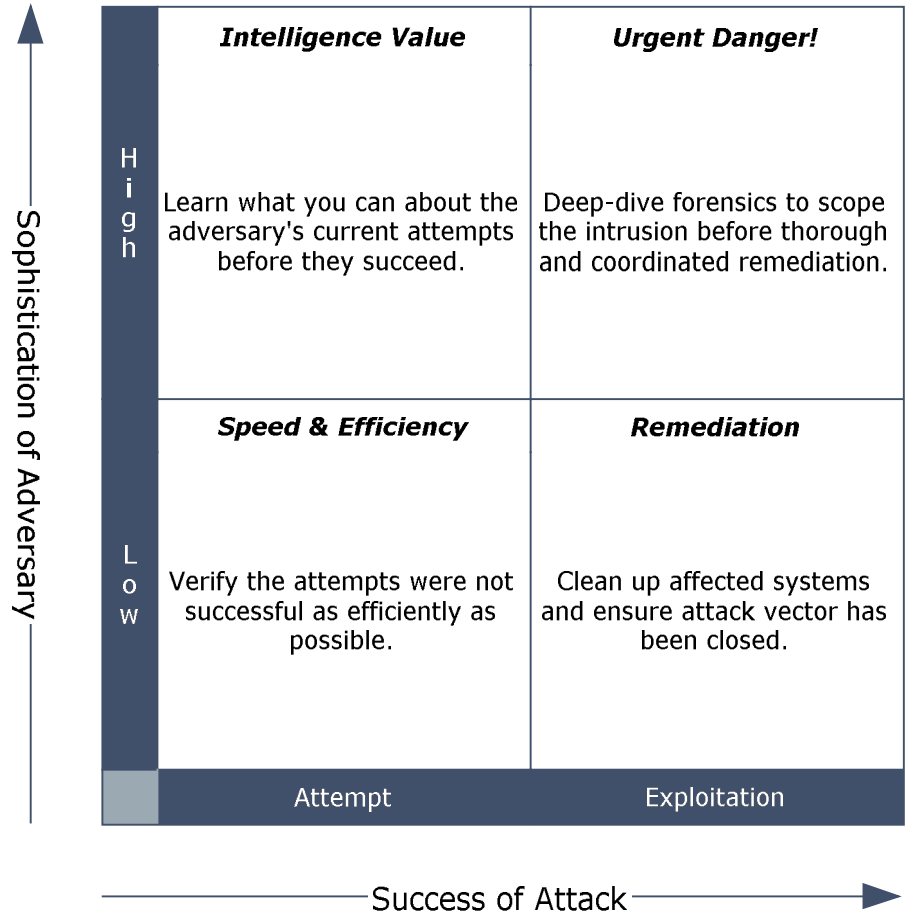
Design

Control Server

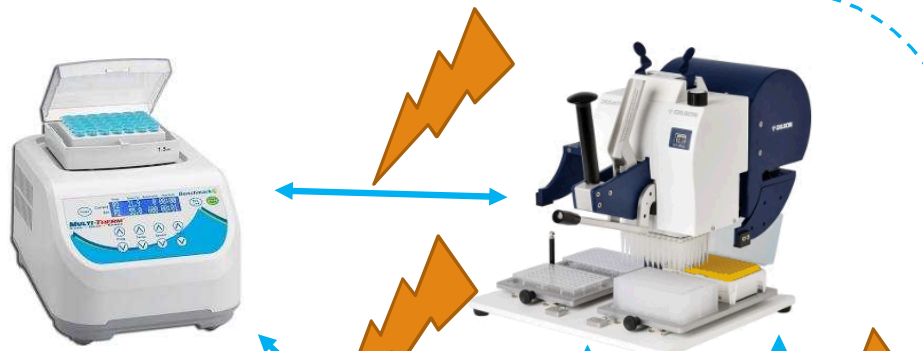
Adversarial sophistication

Threat models typically take into account the sophistication of the adversary and the success of the attack

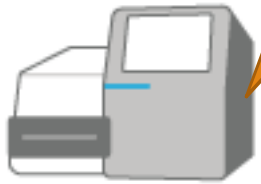
Once an adversary has command and control access, the sophistication of the adversary determines the response



Synthesis, Assembly and Cloning



Screening and Sequencing



Interface

BioThreat Database

Design

Control Server

Conclusions

- Next gen sequencing and synthetic biology has grown at a speed that has outpaced the security implications of the platform
- Desperate need for research on vulnerabilities in NGS systems
- The loss in security of genomic data has implications outside the original sequence
- There is a distinct and underappreciated risk of unintentional manufacture of synthetic biological material