

# Community Information Workbench for Global Biosurveillance

---

Data, Information, Knowledge,  
Analytic Algorithms, Users

**NDIA, Biosurveillance Conference**  
**August 27-28, 2012**

**Los Alamos National Laboratory**  
Helen Cui, Ben MacMahon, Patrick Chain,  
Tracy Erkkila, Harshini Mukundan  
[hhcui@lanl.gov](mailto:hhcui@lanl.gov), 505-665-1994

# LANL: A National Security Science Laboratory Serving the National Interest

- Anticipation, innovation, and delivery of solutions
- Discovery to Applied Science to Prototypes
- Leveraging outstanding science, technology & engineering expertise for national needs



Roadrunner  
Supercomputer

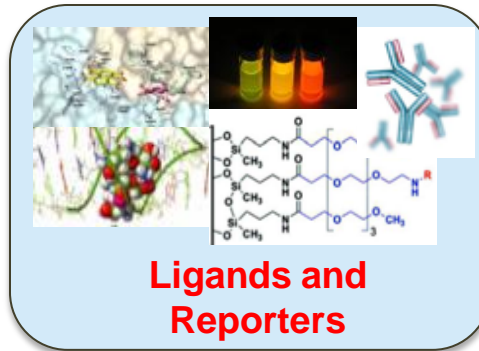


# LANL Biosurveillance Capabilities



**Biomarker Discovery**

This panel features a collage of images related to biomarker discovery. It includes a laboratory setting with a person working at a bench, a 3D molecular model of a protein, a colorful array of spots, and a large data plot with multiple columns and rows of numerical values.



**Ligands and Reporters**

This panel displays various biological and chemical elements. It includes a microscopic view of cells, a test tube with a color gradient, a 3D molecular model of a protein, and a chemical structure diagram of a complex organic molecule.



**Assay Development**

This panel shows a series of images related to assay development, including a 3D molecular model, a grid of colored spots, a diagram of a multi-step process, and a microscopic view of cells.

Biosurveillance  
Integration of  
Diverse Technologies



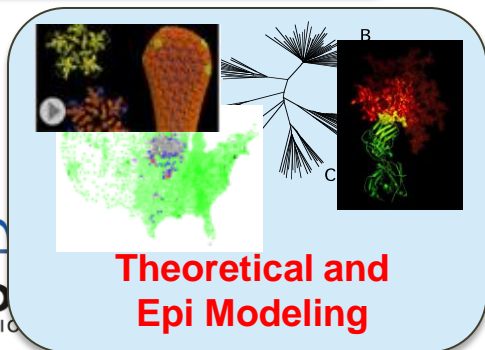
**Engineering Sensors**

This panel illustrates various engineering sensors and devices, including a handheld electronic device, a blue water filtration unit, and other laboratory equipment.



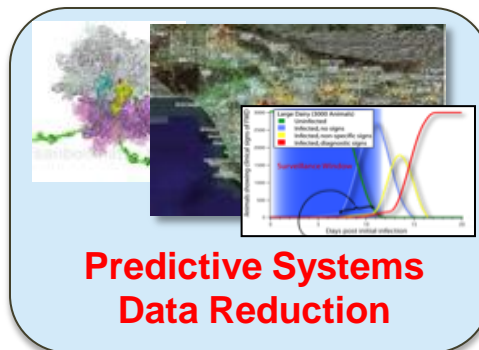
**Next Gen Sequencing**

This panel features a DNA double helix, a colorful grid of data points, and a series of colored wavy lines representing sequencing data.



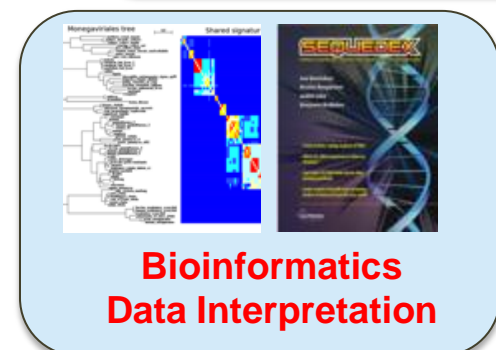
**Theoretical and Epi Modeling**

This panel includes a map of the United States, a 3D molecular model, and a diagram showing a network of nodes and connections, likely representing an epidemiological model.



**Predictive Systems Data Reduction**

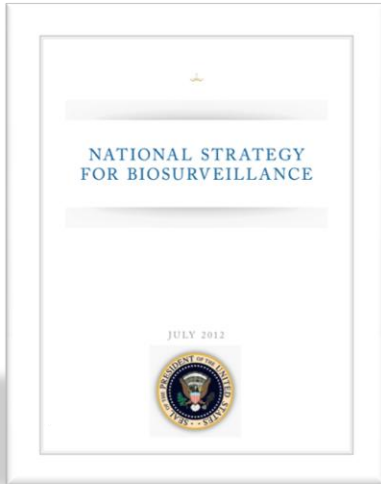
This panel shows a 3D molecular model, a microscopic view of cells, and a line graph with multiple colored curves representing data reduction results.



**Bioinformatics Data Interpretation**

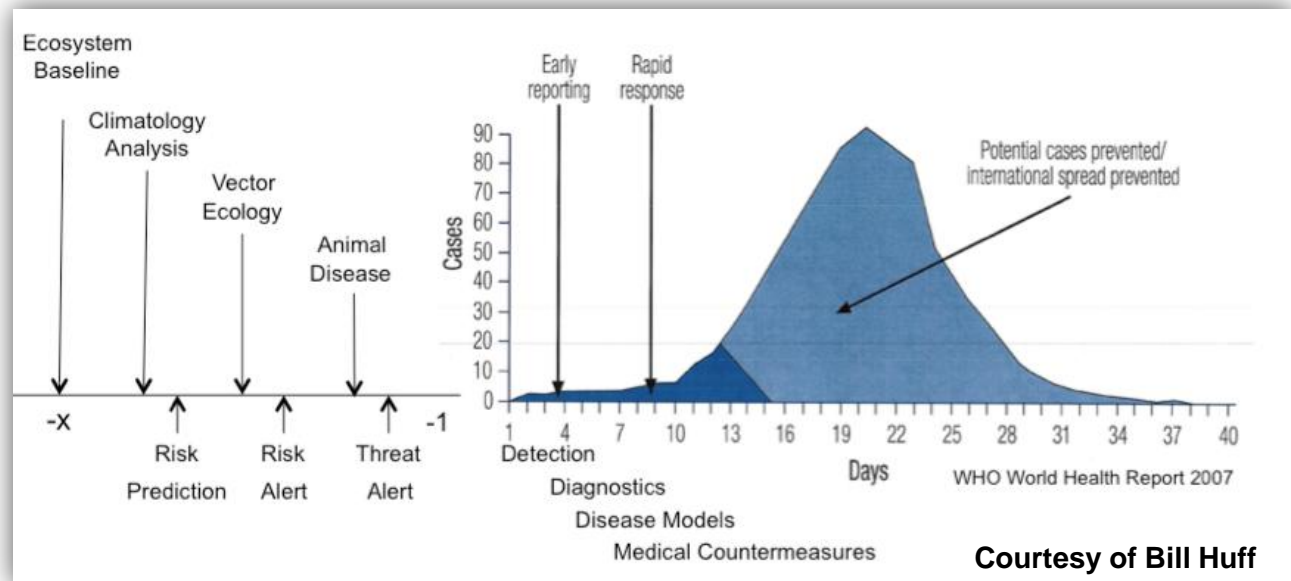
This panel displays a phylogenetic tree, a heatmap, and a DNA double helix, representing various bioinformatics data interpretation tools and results.

# Guiding Principles & Core Functionalities



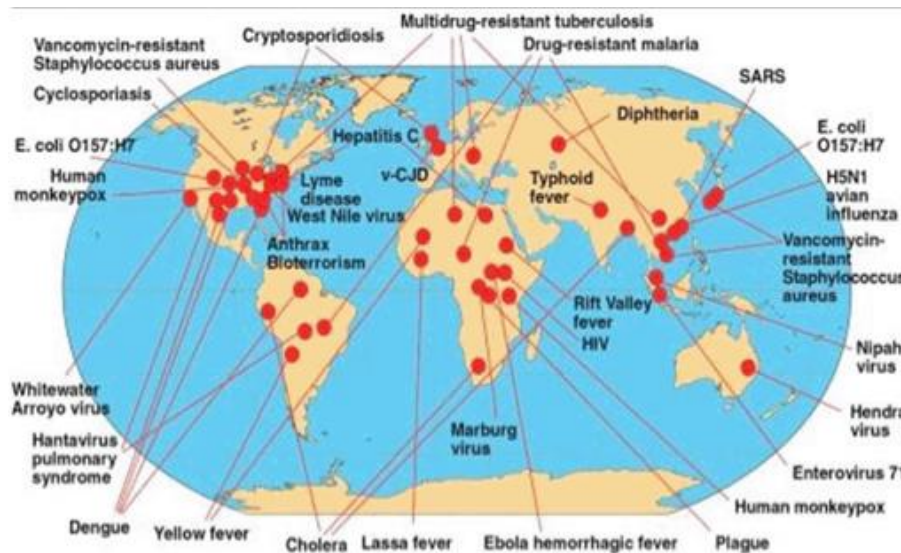
## Providing essential information

- Leverage existing capabilities
- Add values for all participants
- Identify & integrate essential information
- Alert, inform, forecast and advise
- Essential questions & critical answers



# Global Biosurveillance – The Challenge

**Very large, diverse data sets – and growing!**

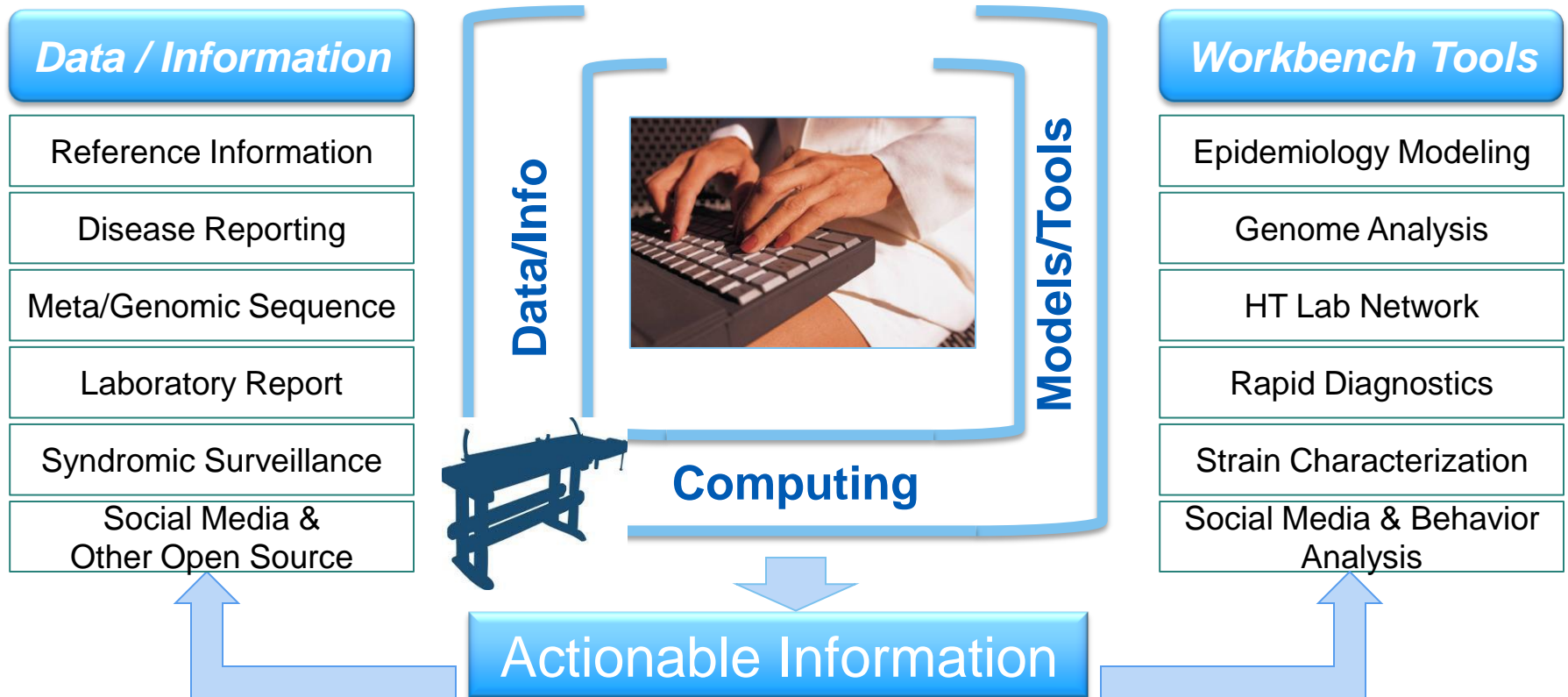


- **Distributed data and tools**
- **Rapidly evolving technology**
- **User needs**
- **Intelligent data reduction and analysis**

**A core interagency capability is required**

- To provide access to existing information and analysis tools
- To generate actionable inference and recommendations

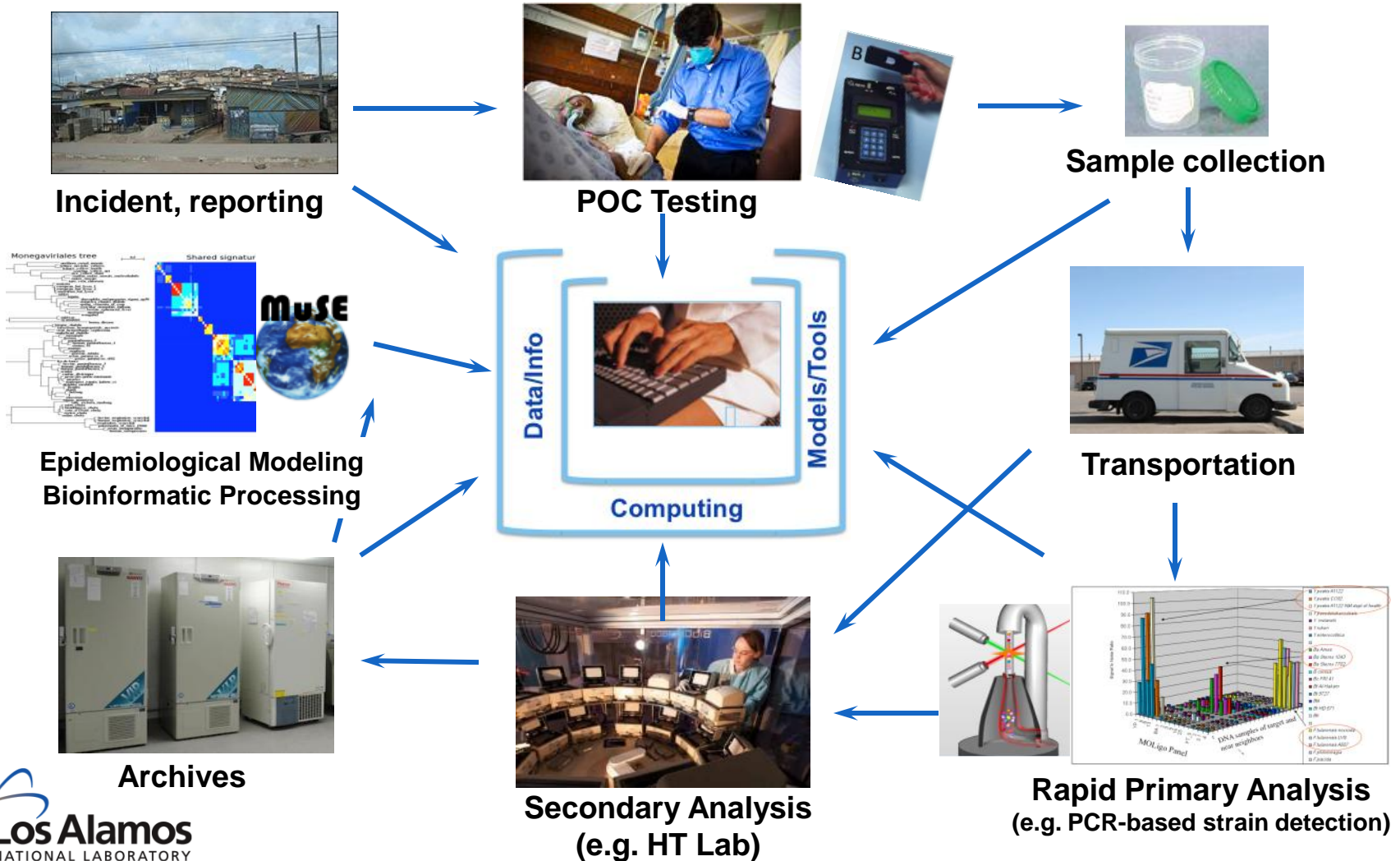
# Workbench Enables Biosurveillance



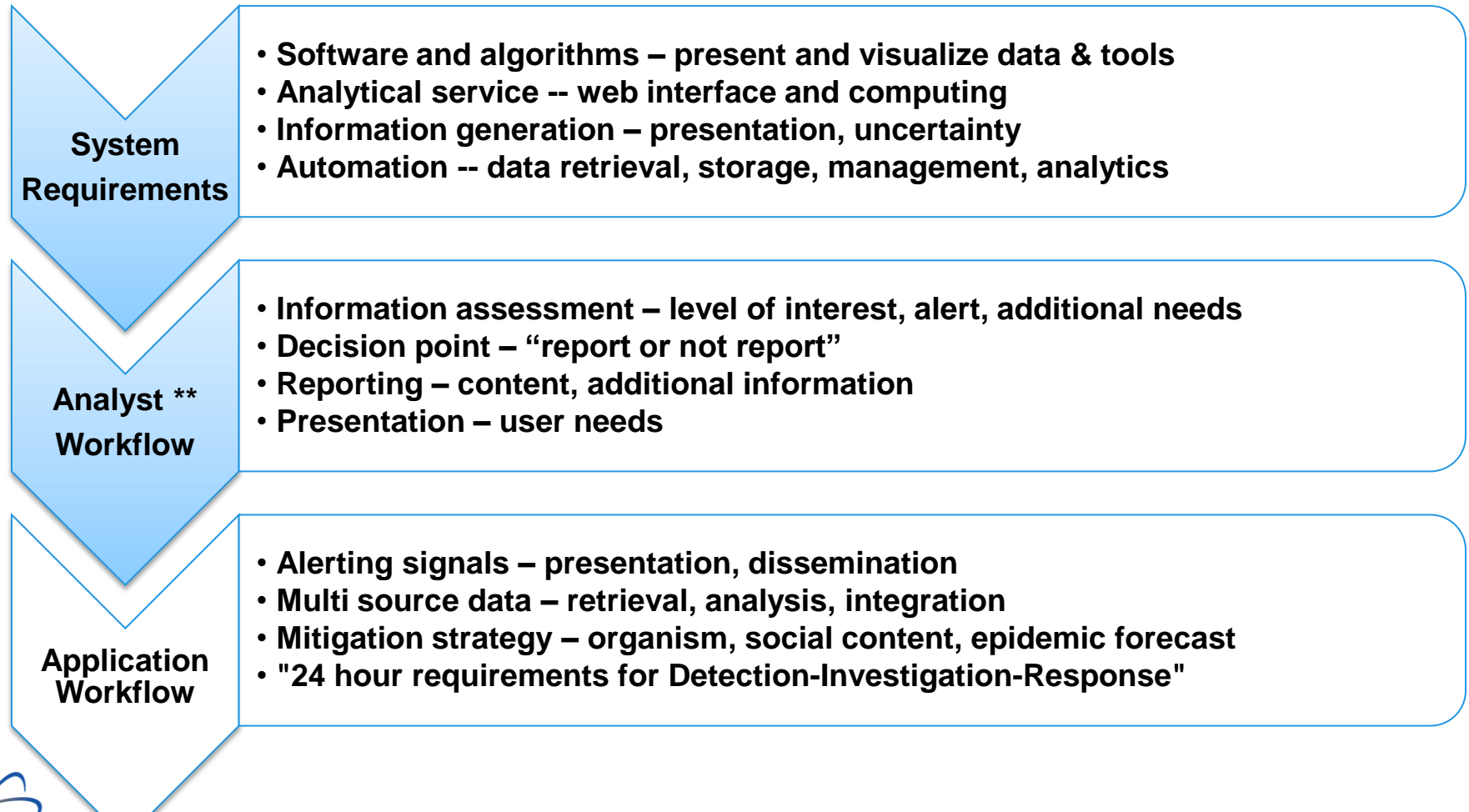
## A core interagency capability:

- User friendly for different analytic requirements
- Adaptable to tailored interagency needs
- Scalable and flexible to address increasing data volume and type

# The Integration Challenge: An example



# Workbench Workflow Considerations





# Reference Information

- **Available Genetic Data**

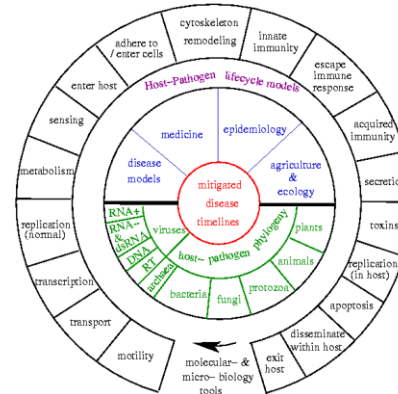
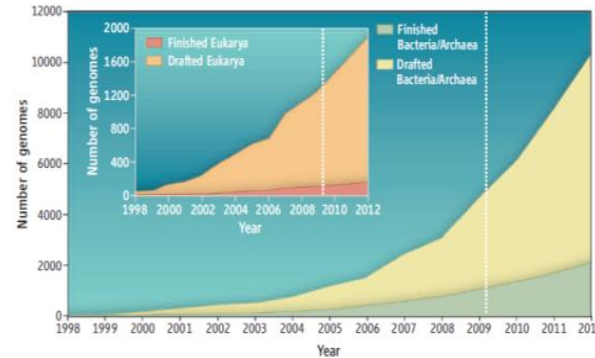
- Genome references
- Metagenome datasets
- Pathogen signatures

- **Host-Pathogen Models**

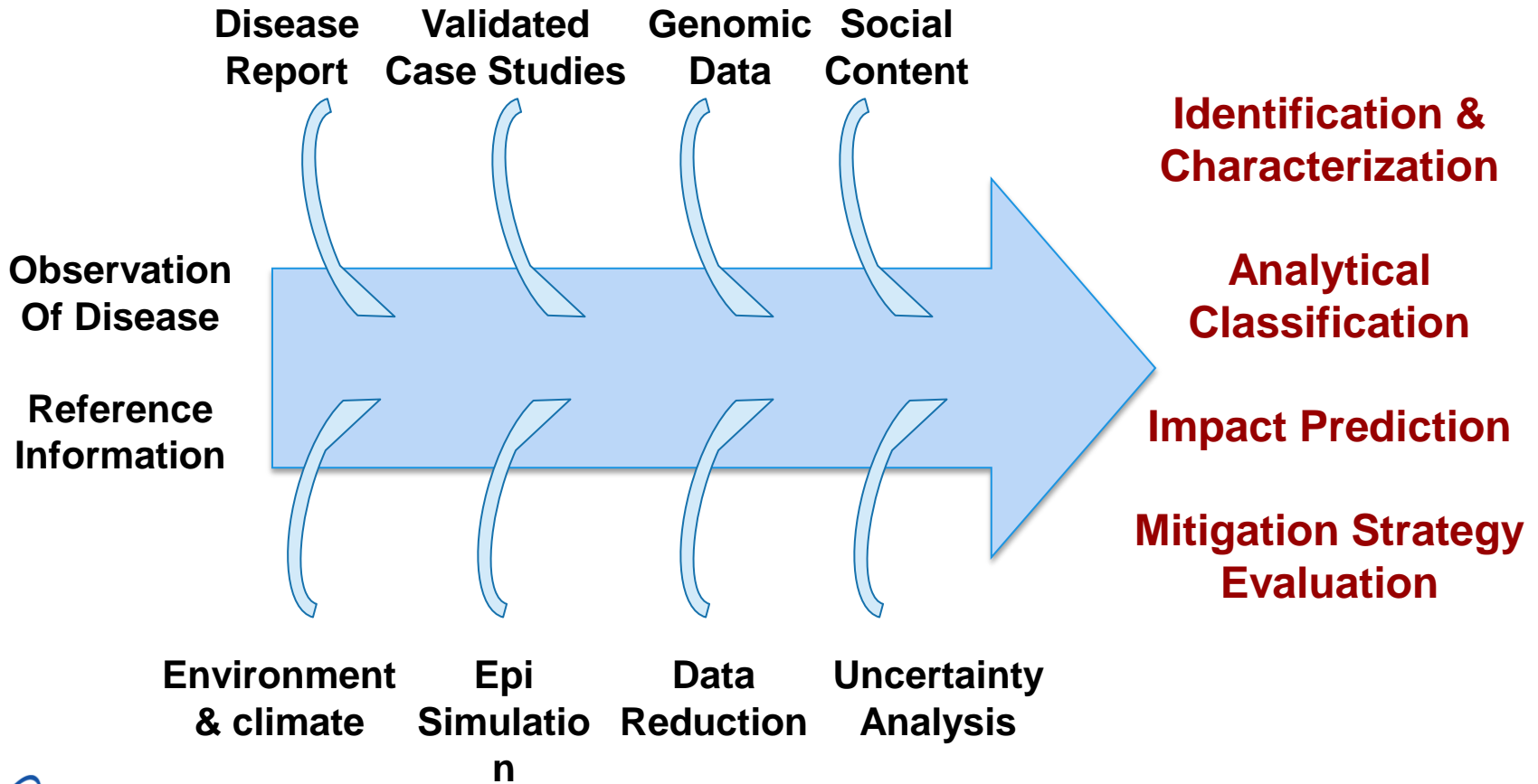
- Disease models
- Virulence genes/pathways
- Person-person spread

- **Historic data**

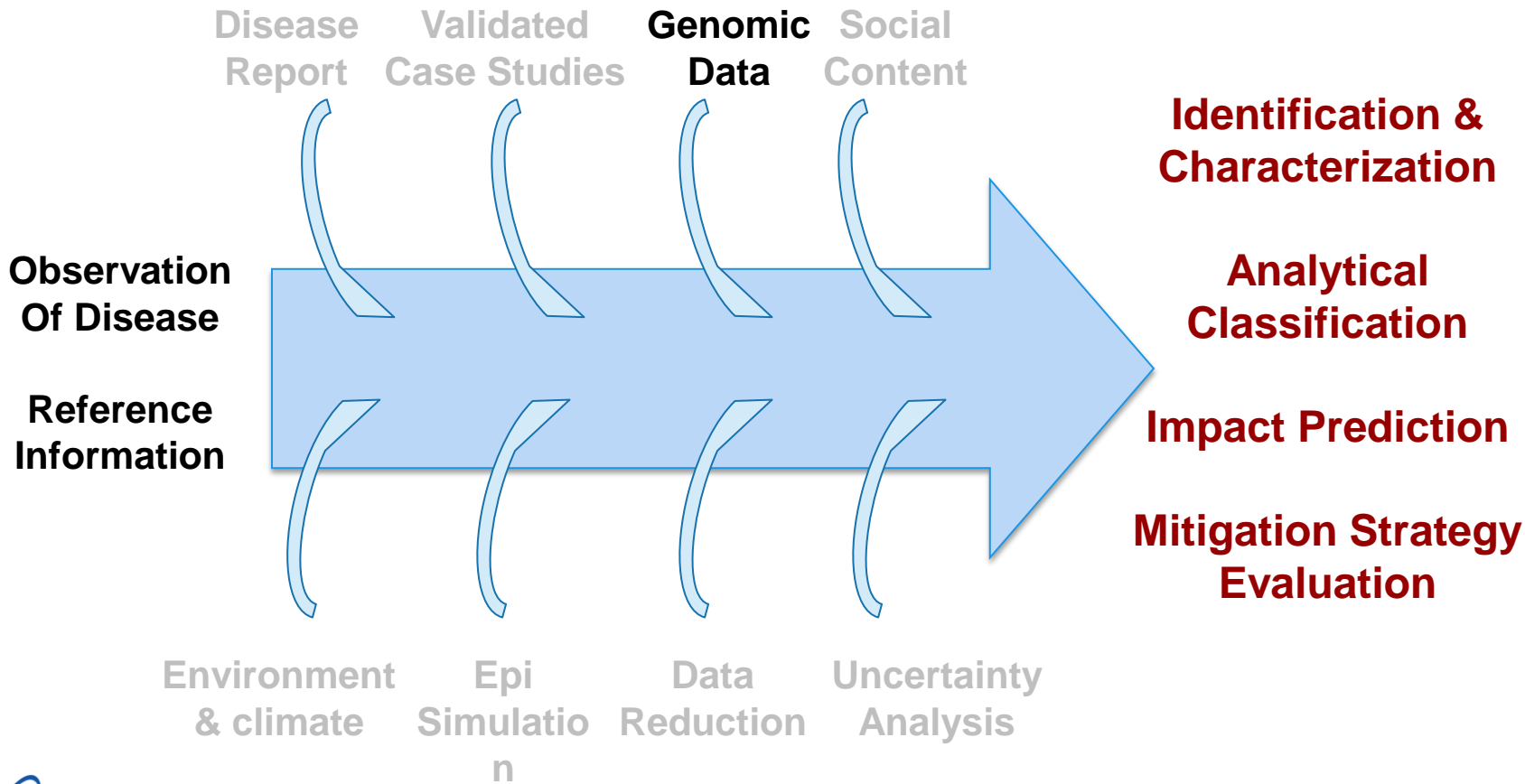
- Environmental background
- Climate
- Epidemiological data



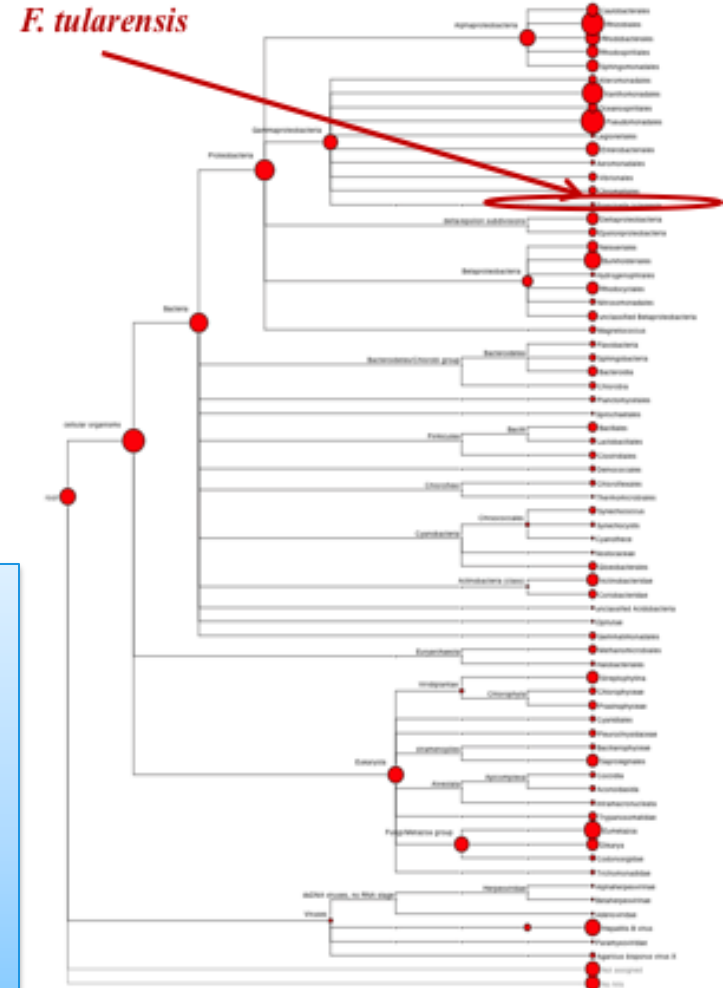
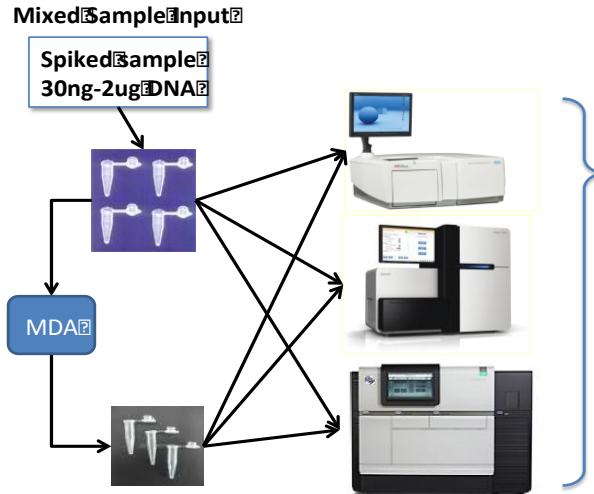
# Active Data Integration



# Active Data Integration

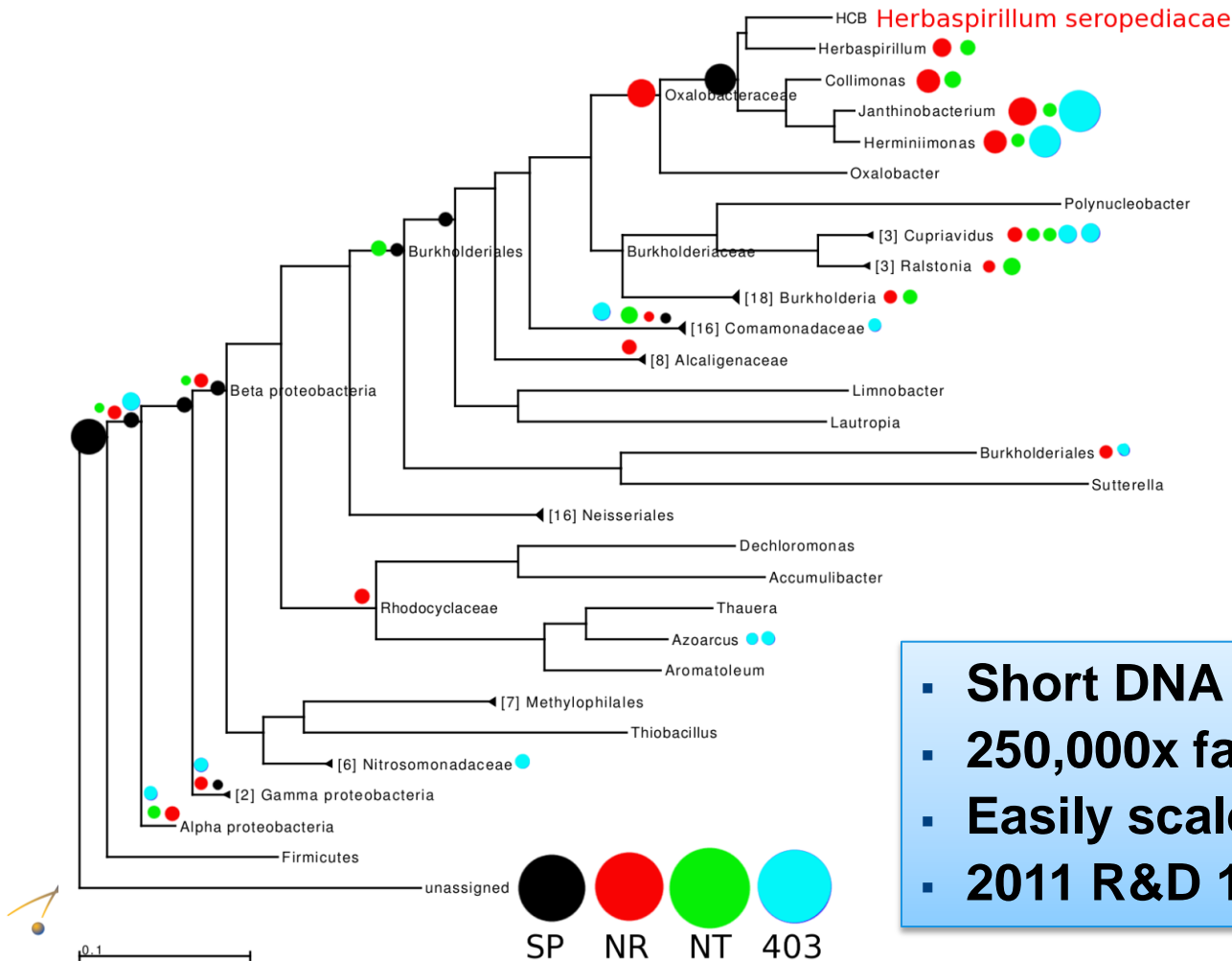


# Rapid Sample-to-ID for Outbreak Scenarios



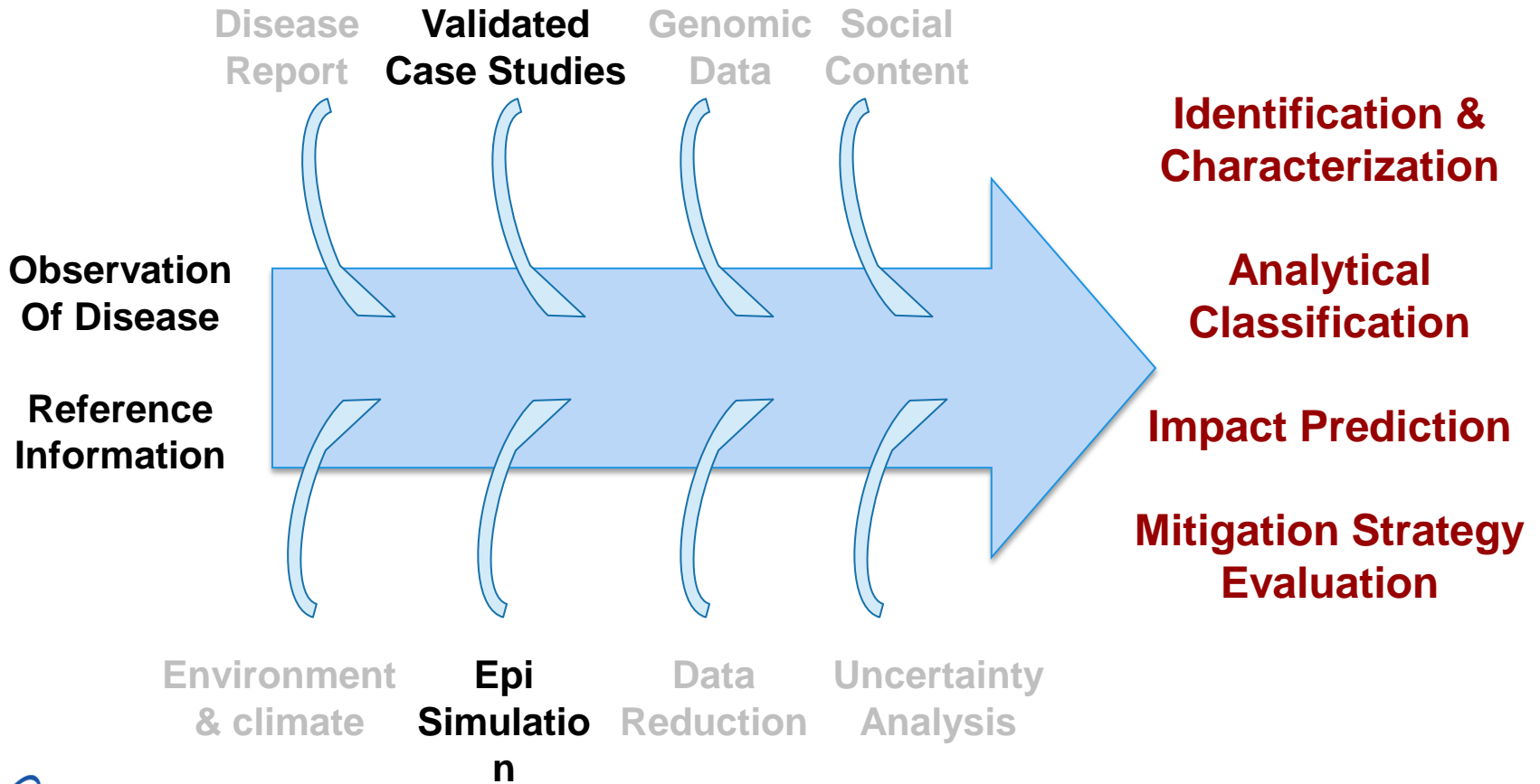
- Rapid detection within 24hrs
- Positive strain discrimination at very low abundance (0.3%)
- Improvements including speed, have been made to library preparation and analysis

# Novel Algorithms for Sequence Data Analysis



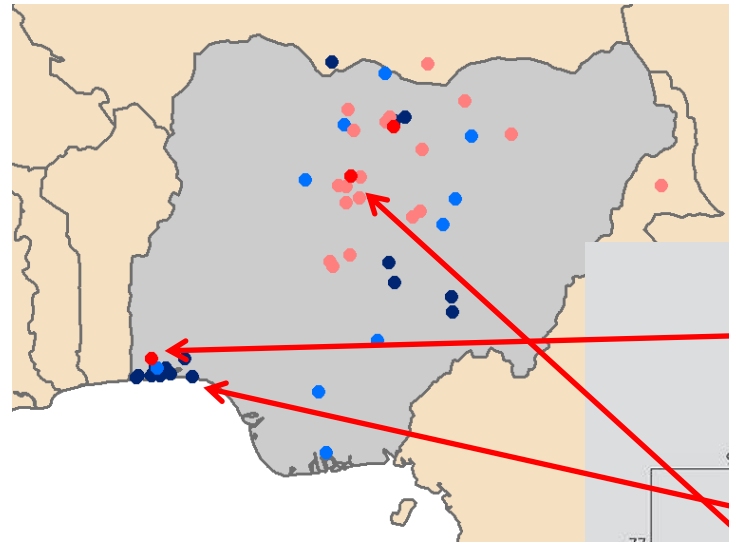
- Short DNA sequences ( $\geq 30$  bp)
- 250,000x faster
- Easily scale up
- 2011 R&D 100 Award

# Active Data Integration

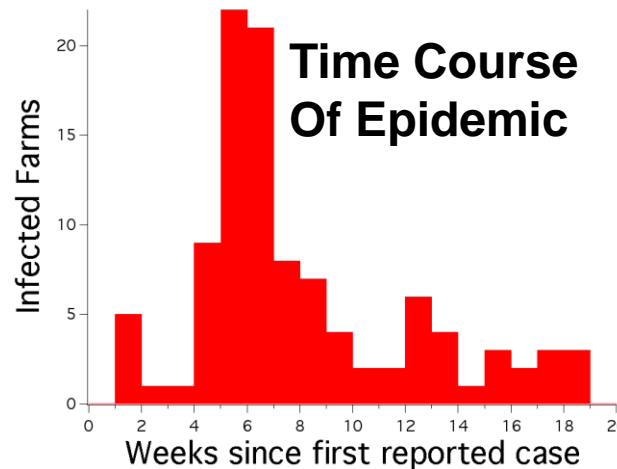


# Epidemiology – Outbreak Identification

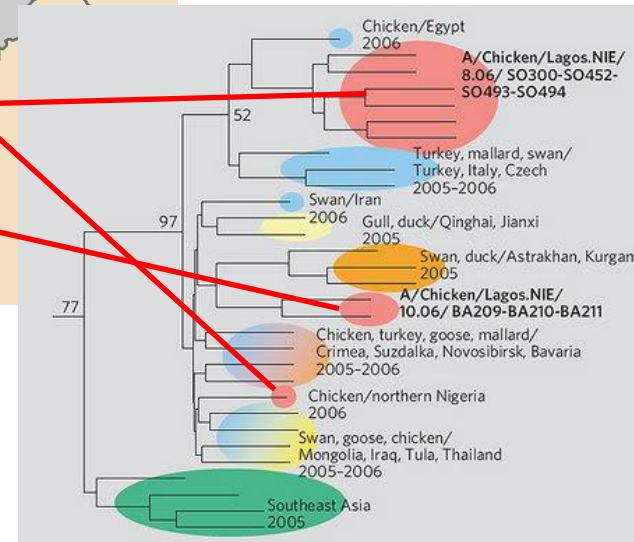
- Nigerian avian influenza outbreak in 2006
- Positive cases: 248
- Depopulated birds: 1,500,000
- Biogeography points to multiple introductions



**Geography**



**Time Course Of Epidemic**

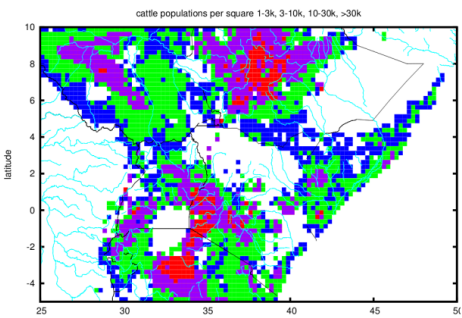
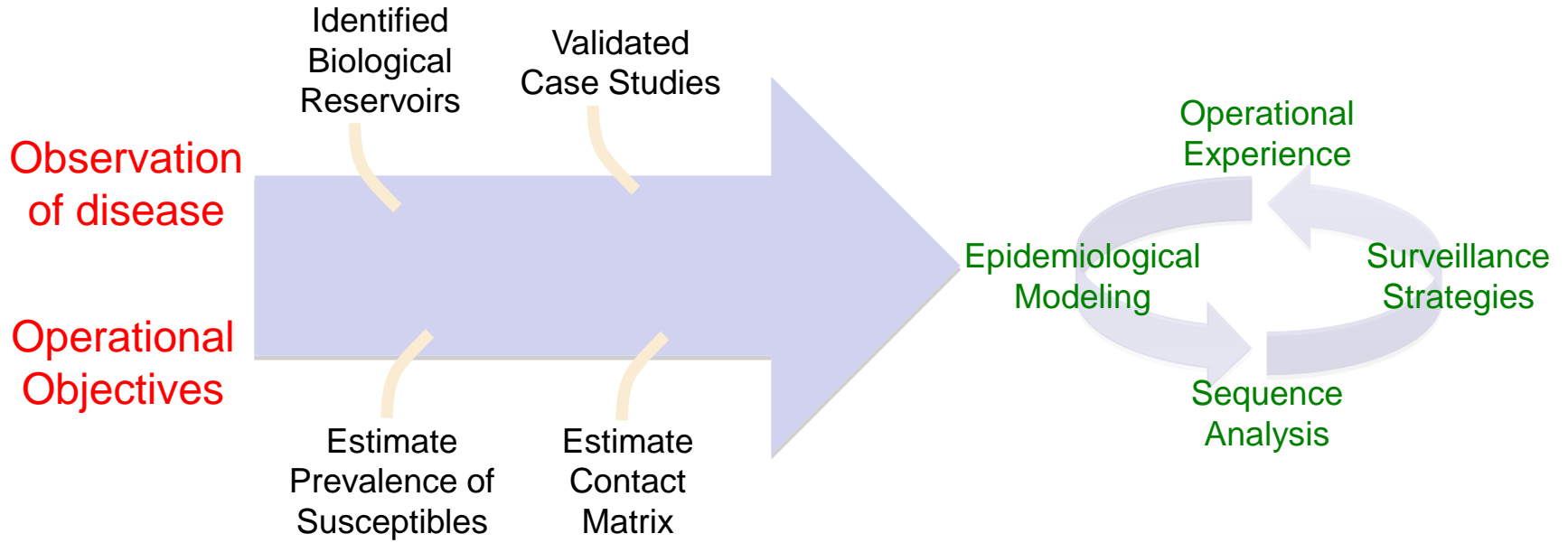


**Sequence Data**

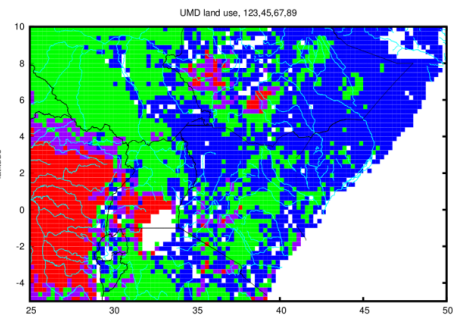
Jeanne Fair et al

# Role of Case-Studies in Epidemiology

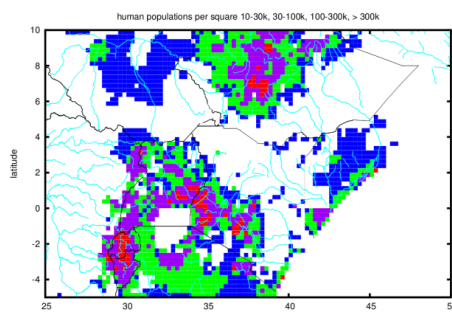
Jeanne Fair



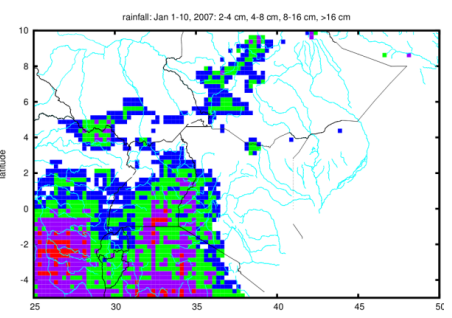
Cattle populations



Land use



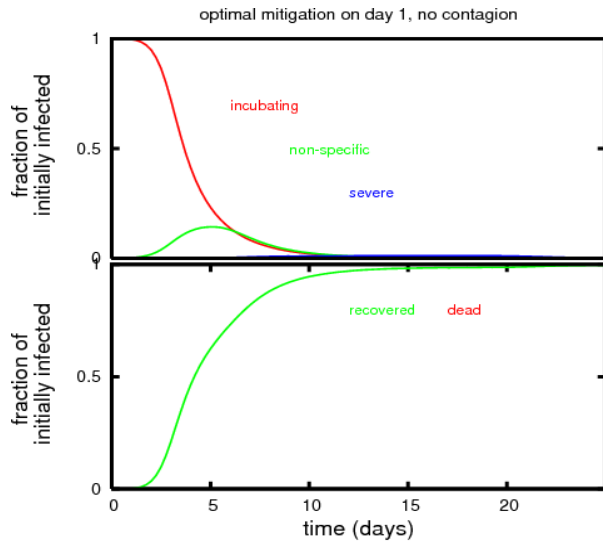
Human populations



Rainfall

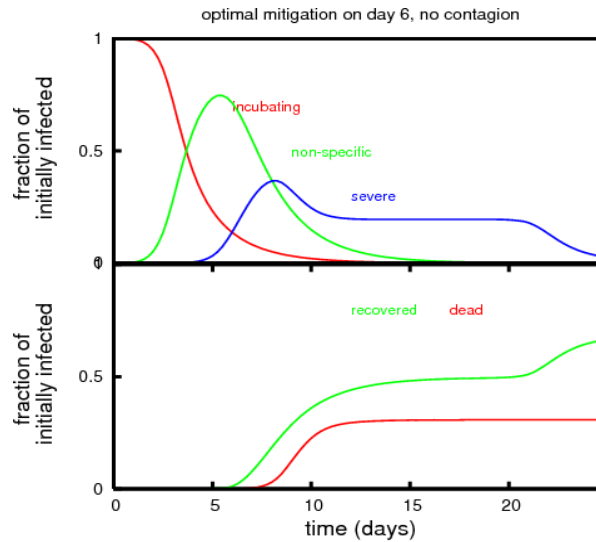


# Disease Progression and Mitigation Characteristics



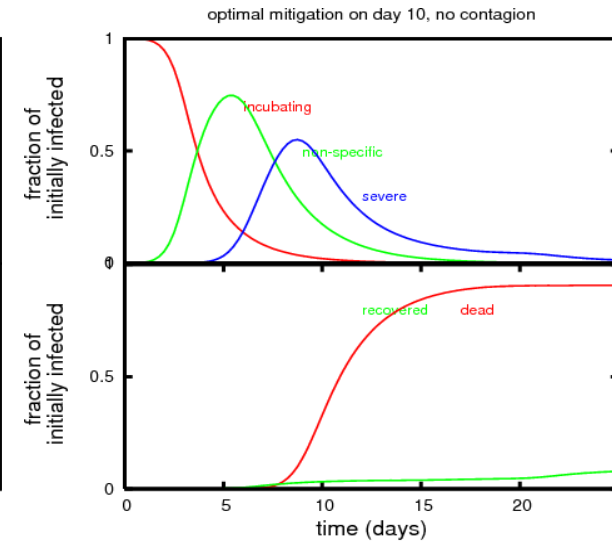
1 2 3 4 5 6 7 8 9 never

Click for new patient mitigation time (days)



1 2 3 4 5 6 7 8 9 never

Click for new patient mitigation time (days)

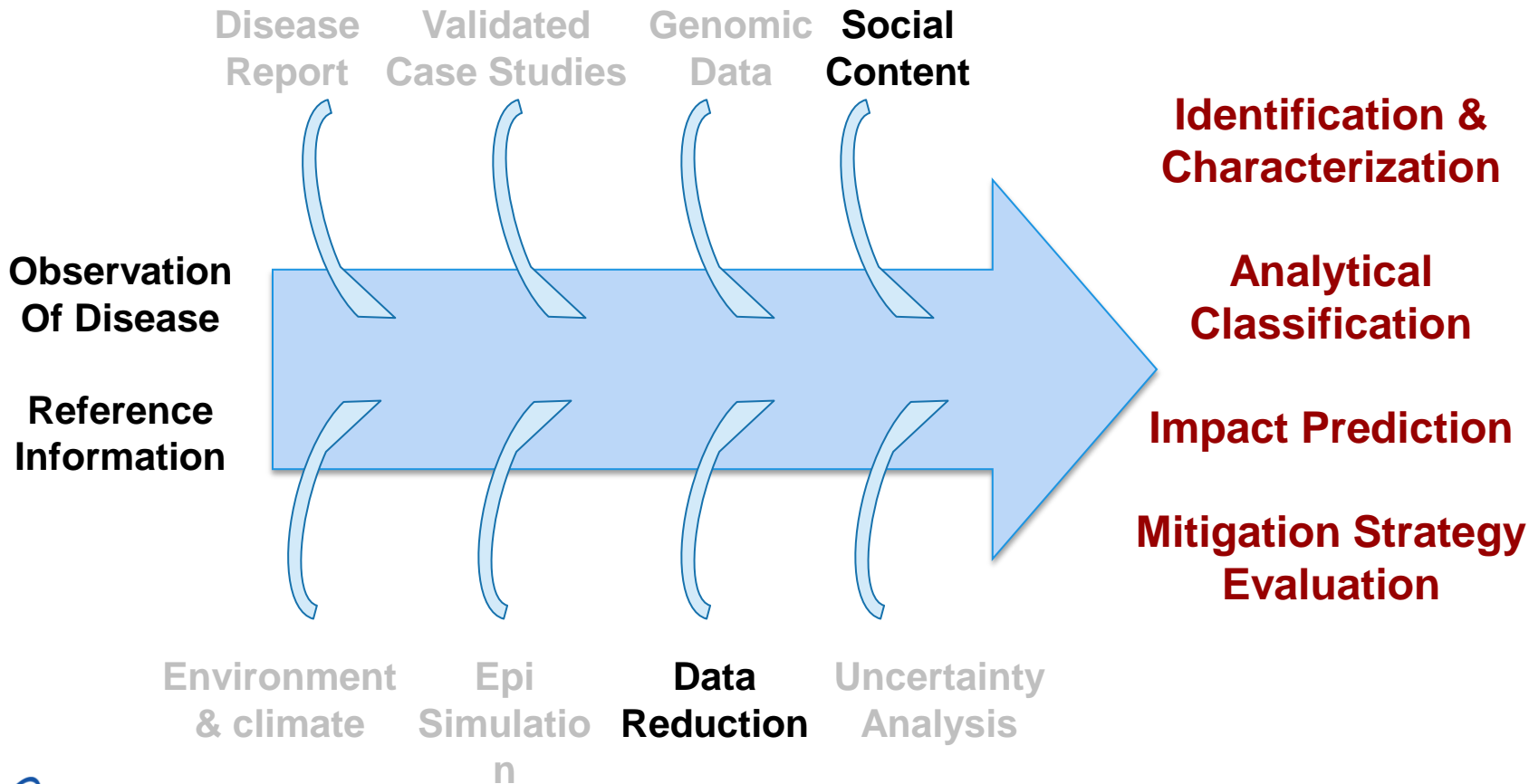


1 2 3 4 5 6 7 8 9 never

Click for new patient mitigation time (days)

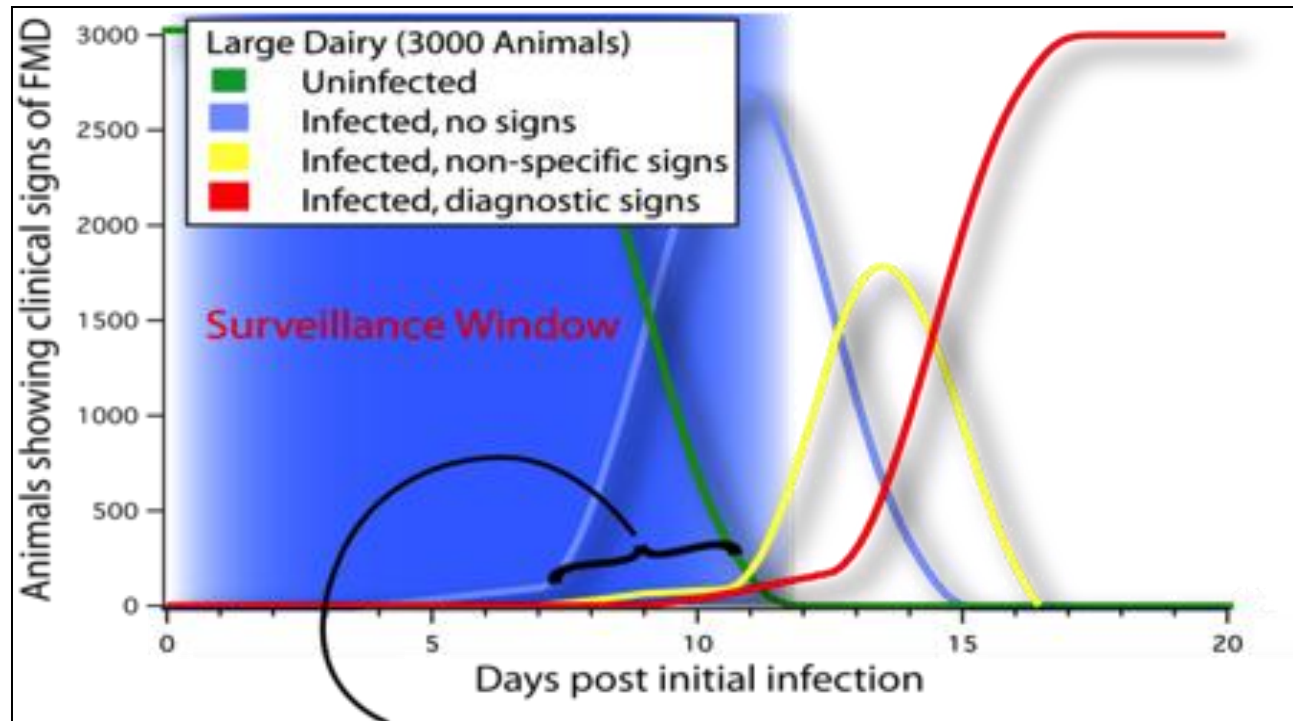
- Host, pathogen, and dose-dependence of disease severity
- Mitigative efficacies and effectiveness
- Distribution of incubation times and duration of prodromal phase

# Active Data Integration



# Data Reduction

Alina Deshpande



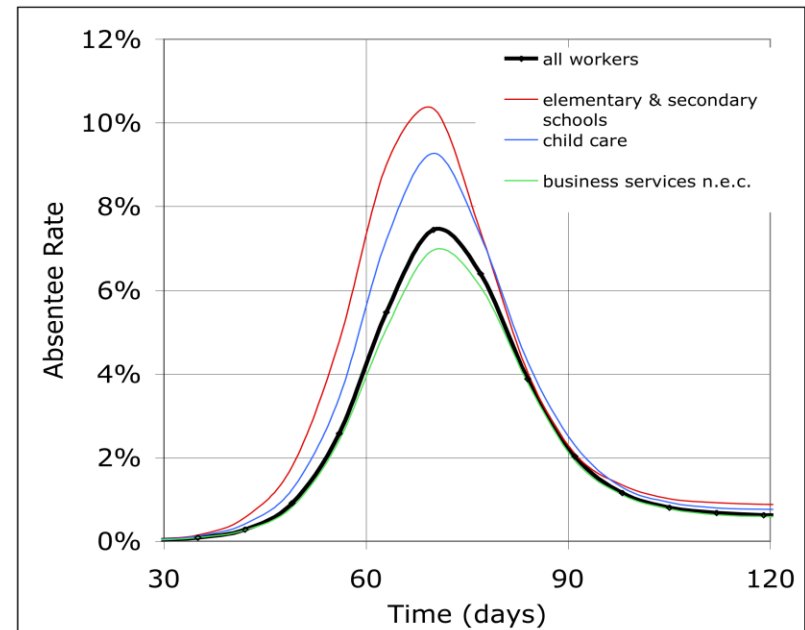
- Ranking potentially useful data streams
- Evaluation of existing algorithms and frameworks for integrating data streams

# Agent-Based Models, Social Media Data Integration



**Pandemic influenza attack rate by census tract. Hot-spots are strongly correlated with household size.**

**Workforce absenteeism by industry classification H5N1 pandemic simulation**



# Information Technology Strategy

- **Basic IT Architecture:**

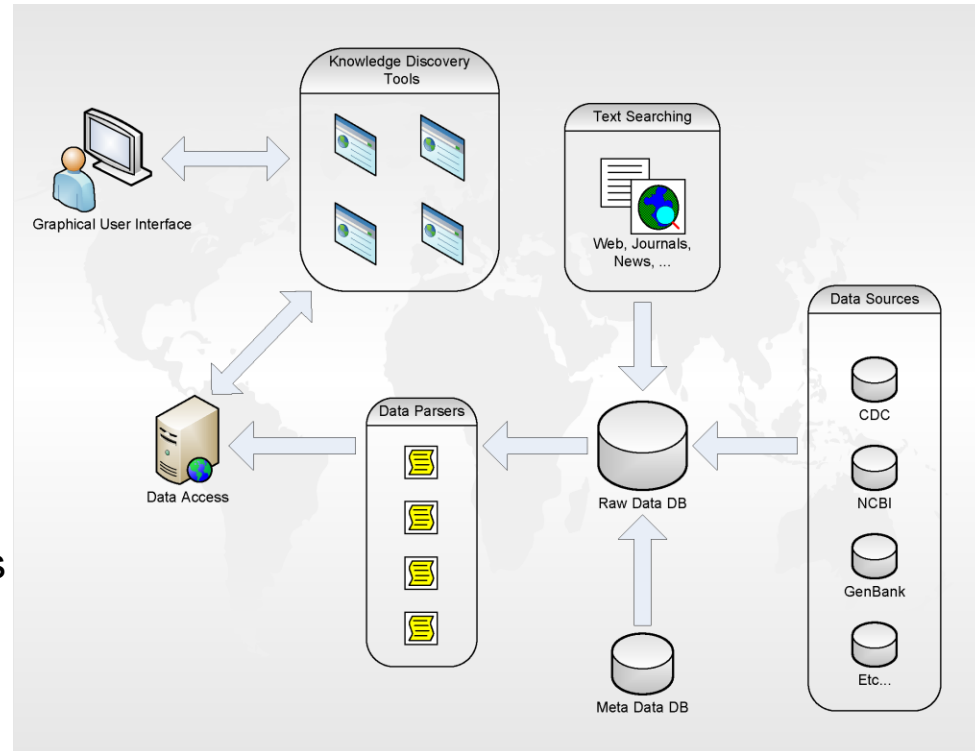
- Data retrieval, cache/storage
- Format/process
- Result presentation
- Flexible computer/programming

- **Tool Integration:**

- Service Oriented Architecture
- Existing/developing software
- Integrating programming languages
- Decouple

- **User Friendly Interface:**

- No client-side install and maintenance
- Available on any platform
- Server-side updates, transparent to user
- Intense computation performed server-side, little computational resources required on client-side



# Leverage Explorative Development

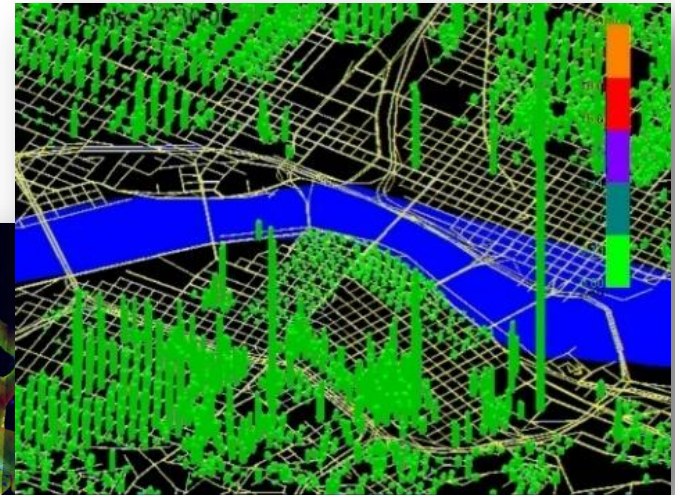
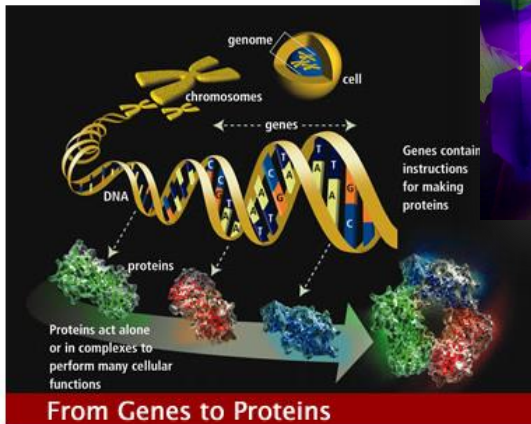
**BioPASS**  
Logged in as: hcui [Logout](#)

- **Rapid organism identification**
- **Key microbial virulence factors**
- **Phylogenetic placement highlighting anomalies**
- **Phylogenetic - geographic location correlation**
- **Disease progression and epidemic spread potential**
- **Microblogging info indication**
- **Biothreat pathogen and infectious disease knowledge**

Barillus cereus NVH0597:99 (1 assembly)

The system aims serve the dual role of identifying well-known pathogenic mechanisms and highlighting potentially novel attributes of the emerging pathogens

# Integrated BSV by Innovation & Delivery



Accelerating discovery-to-innovation for technology delivery and mission impact

# A Highly Accomplished Multidisciplinary Team

## ■ LANL Team:

- *Systems & Programming:* Craig Blackhart, Bob Funkhouser, Chen He,
- *Disease mechanisms:* Jennifer Harris, *Metagenomics:* Ben MacMahon, Patrick Chain, Nick Hengartner
- *Bioinformatics:* Carla Kuiken, Chris Stubben, Jian Song, Jason Gans
- *Genomics:* Chris Detter
- *Epidemiology modeling:* Ben MacMahon, Jeanne Fair, Brent Denial
- *Uncertainty analysis:* Mac Hyman
- *Social content:* Sara Del Valle
- *Data reduction:* Alina Deshpande
- *Students:* Amanda Minnich, Catherine Chen
- *Program & Strategy:* Tom Terwilliger, Gary Resnick, Cathy Cleland, Harshini Mukundan, Jurgen Schmidt, Nan Sauer, Tony Redondo. Frank Alexander, Randy Erickson

## ■ Sponsors:

- Department of State
- LANL LDRD
- DTRA

## ■ LANL Support:

- Center for Biosecurity
- Center for Information Science
- High Performance Computing

## ■ Current & Future Collaborators:

- Paul Keim
- Rutgers University
- PNNL, ORNL
- Gary Simpson
- Many other friends

## ■ Interagency Adaptation/Development Discussion:

- DOD (DTRA, JPEO), DHS, IC



# Sequence Comparison for Identification

The screenshot shows the BioPASS web interface. At the top left is the Los Alamos National Laboratory logo and the Department of State seal. The main header is "BioPASS". A left sidebar contains navigation links: Home, Tools, News, Contact, and About. The main content area is titled "BlastOrg - Identify Organism(s)". It includes a "Purpose" section explaining that the tool identifies genetic sequences by organism. Below is an "Input" section with a "Choose File" button (showing "no file selected") and a large text area containing a FASTA format nucleotide sequence. At the bottom of the input area are "submit" and "reset" buttons. Below the input area are two links: "Sample FASTA file" and "Example Results".

**BioPASS**

**BlastOrg - Identify Organism(s)**

**Purpose** - This web tool accepts nucleotide sequences in FASTA format and identifies the source of the genetic sequences by organism (or substance). For more information, see the [help file](#).

**Input** - Paste your FASTA format nucleotide sequence here

no file selected

```
gaaggcaataattgtactactcatggtagtaacatccaatgcagatcgaa
tctgcactgggataacatcttcaaactcacctcatgtggtcaaaacagct
actcaaggggagggtcaatgtgactggtgtgataccactgacaacaacacc
aacaaaaatcttattttgcaaatctcaaaggaacaaggaccagagggaaac
taagtgtcaacaggtcaggtgtgcatcaatcaagataacatgaagatgctcgaccctagt
cagatagaggaaccagaagaggcaccaggatgaacatcgggtatccacctgaaacaaggt
tatgccggactgtctcaactgtacagatcggatgtggccttgggagg
ccaatgtgtgtggggaccacaccttctgctaaagcttcagtaactccacga
agtcagactgttacatccgggtgctttcctataatgcagacagaacaa
aaatcaggcaactacccaatctctcagaggatagaaaaatcagggtta
tcaacc caaacgttatcgatgcagaaaaagcaccaggaggaccctacag
acttggaaacctcaggatcttgccctaacgctaccagtaaaatcggatttt
tcgcaacaatggcttgggctgtccaaaggacaactacaaaaatgcaacg
aaccactaacagtagaagtaccataattgtacagaagggggaagacca
aattactgtttgggggttcattcagataacaaaacc caaatgaagaacc
```

[Sample FASTA file](#) [Example Results](#)

# Similarity Returns, Virulence Factor Hits

Contact

About

Category	<i>Lysinibacillus sphaericus</i>	<i>Bacillus cereus</i> ATCC 10987	<i>Bacillus anthracis</i> Sterne
% Bacterial hits	84	95	97
% Firmicutes	66	90	94
% Bacilli (class)	54	86	92
% Bacillales (order)	50	85	92
% Bacillus (genus)	26	82	91
% <i>B. cereus</i> (~5 species)	6	76	83
% <i>Bacillus anthracis</i>	2	36	61
# hits Lethal factor	20	25	0
# hits Protective antigen	13	24	0
# hits Edema factor	11	22	0
# hits Cap A, B, C	28	28	0
# hits Drug resistance transporters	36	142	139
# hits Penicillin-binding protein	19	144	136
# Total Toxin, Virulence factor, Ab hits	220	925	965

# Phylogeny Analysis of the New Organism

  **BioPASS**  
Logged in as: hcui [Logout](#)

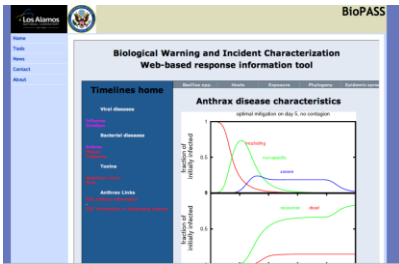
[Home](#)  
[Tools](#)  
[News](#)  
[Contact](#)  
[About](#)

## Bacterial Results

[Show BLAST Results](#) [Links to BLAST score summary table]

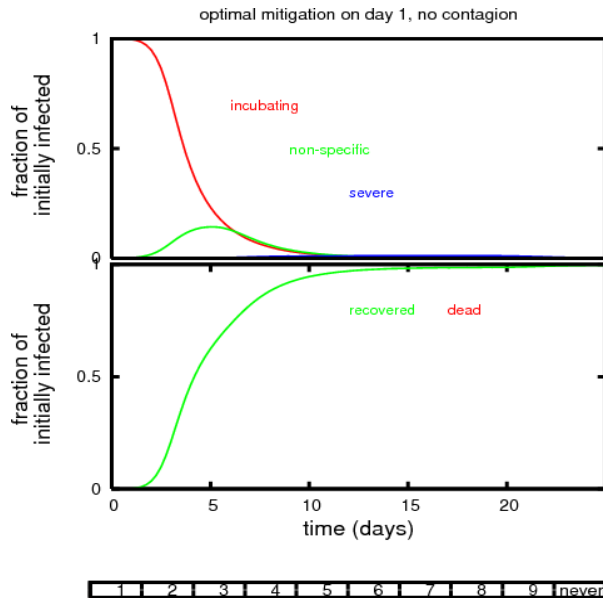
- Bacteria (193 assemblies)
  - Bacteroidetes (1 assembly)
  - Flavobacteria (1 assembly)
    - Flavobacteriales (1 assembly)
      - Flavobacteriaceae (1 assembly)
        - Croceibacter (1 assembly)
          - Croceibacter atlanticus HTCC2559 (1 assembly)**
- Firmicutes (42 assemblies)
  - Bacillales (24 assemblies)
    - Bacillaceae (18 assemblies)
      - Bacillus (18 assemblies)
        - Bacillus cereus group (18 assemblies)
          - Bacillus anthracis Tsiankovskii-I (1 assembly)
          - Bacillus anthracis str. 'Ames Ancestor' (1 assembly)
          - Bacillus anthracis str. A1055 (1 assembly)
          - Bacillus anthracis str. Australia 94 (1 assembly)
          - Bacillus anthracis str. CNEVA-9066 (1 assembly)
          - Bacillus anthracis str. Kruger B (1 assembly)
          - Bacillus anthracis str. Vollum (1 assembly)
          - Bacillus anthracis str. Western North America USA6153 (1 assembly)**
          - Bacillus cereus 03BB108 (1 assembly)
          - Bacillus cereus AH1134 (1 assembly)
          - Bacillus cereus AH187 (1 assembly)
          - Bacillus cereus AH820 (1 assembly)**
          - Bacillus cereus B4264 (1 assembly)
          - Bacillus cereus G9241 (1 assembly)
          - Bacillus cereus G9842 (1 assembly)
          - Bacillus cereus H3081.97 (1 assembly)
          - Bacillus cereus NVH0597-99 (1 assembly)
          - Bacillus cereus W (1 assembly)**
  - Listeriaceae (3 assemblies)
    - Listeria (3 assemblies)
  - Staphylococcus (3 assemblies)
  - Clostridia (14 assemblies)
  - Lactobacillales (4 assemblies)
    - Streptococcaceae (4 assemblies)
      - Streptococcus (4 assemblies)
  - Proteobacteria (116 assemblies)

# Infectious Disease Progression

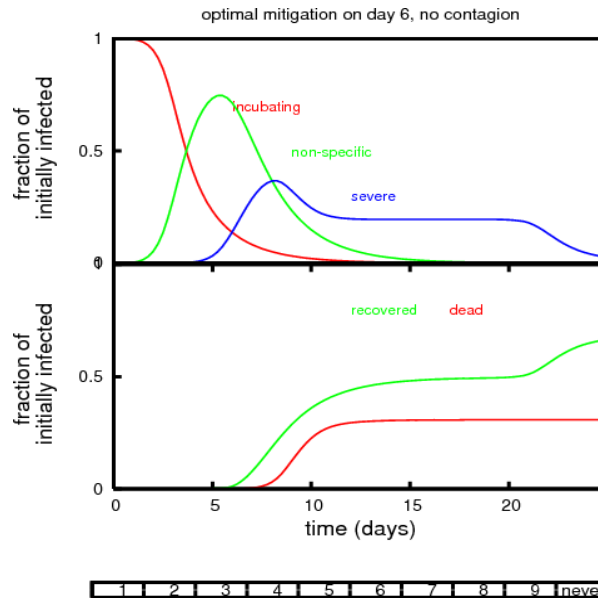


## Disease models available:

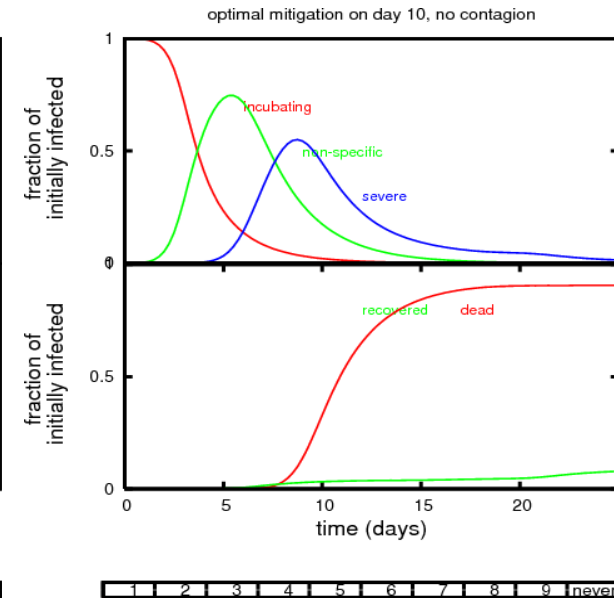
- Enhanced pathogen prediction
- Operational/mitigation planning



Click for new patient mitigation time (days)



Click for new patient mitigation time (days)



Click for new patient mitigation time (days)

# Twitter Disease Tracking



- Worldwide tweets
- Example plotted by location
- Word variants included in the search
- Zoom in for finer location resolution
- Predefined searches are available, and users can type in their own keywords
- Developing repository to follow trends and track spikes

# Phylogeny and Geo-location Correlation

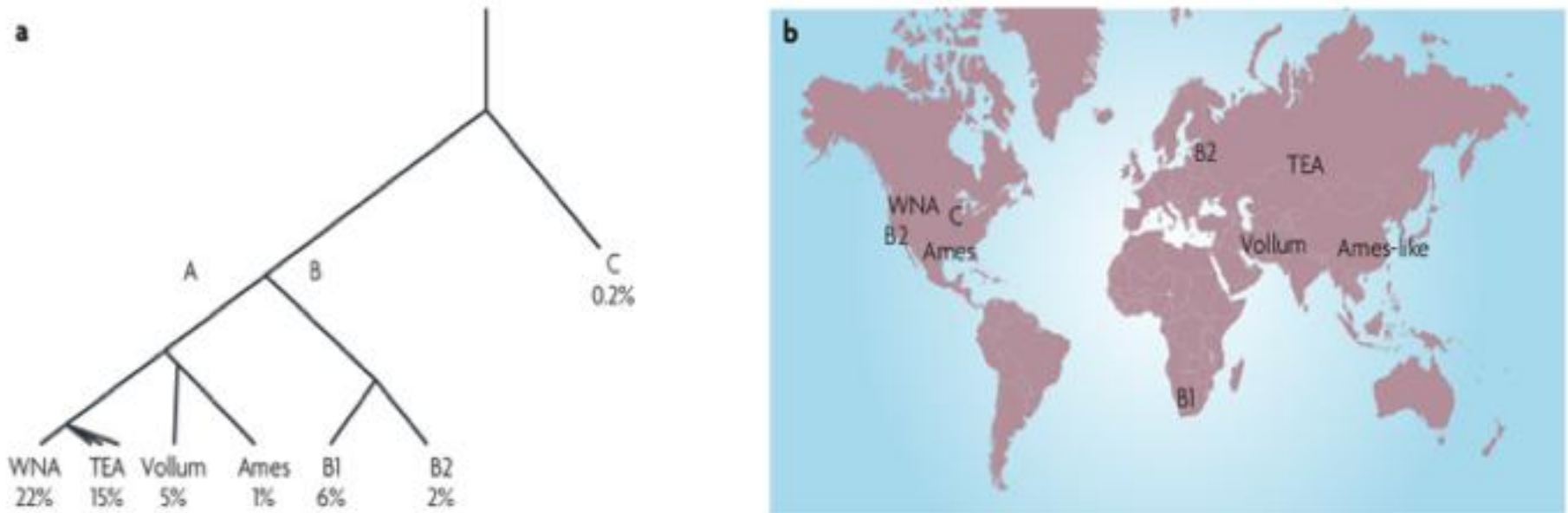
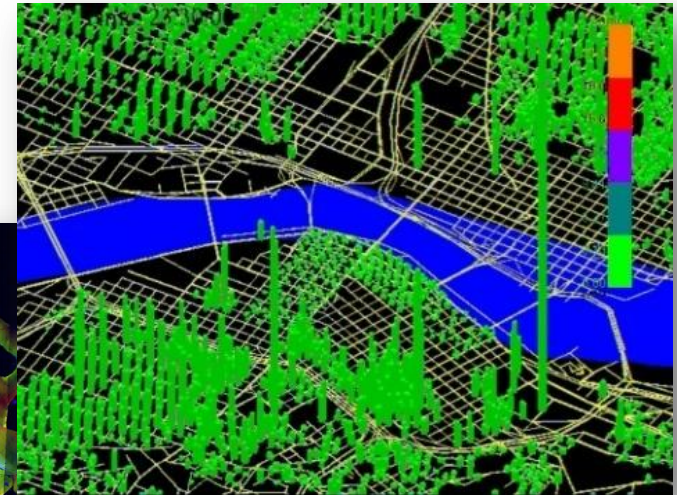
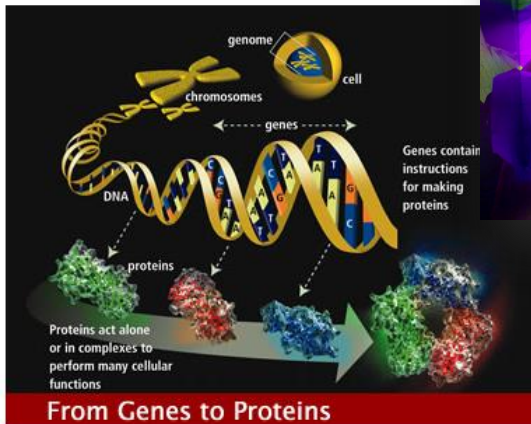


Figure 2 | **Phylogeography of *Bacillus anthracis*.** **a** | The population structure of *Bacillus anthracis* revolves around three major groups (A, B and C). **b** | The group A bacteria are found in all parts of the world and are very common, whereas the B1, B2 and C group bacteria are rarer and mostly restricted to subcontinental locations. Highly successful clonal lineages exist even within group A. TEA, trans-Eurasian; WNA, western North American. Keim and Wagner (2009)



# Integrated BSV by Innovation & Delivery



Accelerating discovery-to-innovation for technology delivery and mission impact