Introduction

BCB 5200 Introduction to Bioinformatics I

Fall 2017

Tae-Hyuk (Ted) Ahn
Department of Computer Science
Saint Louis University

SAINT LOUIS UNIVERSITY™

EST. 1818

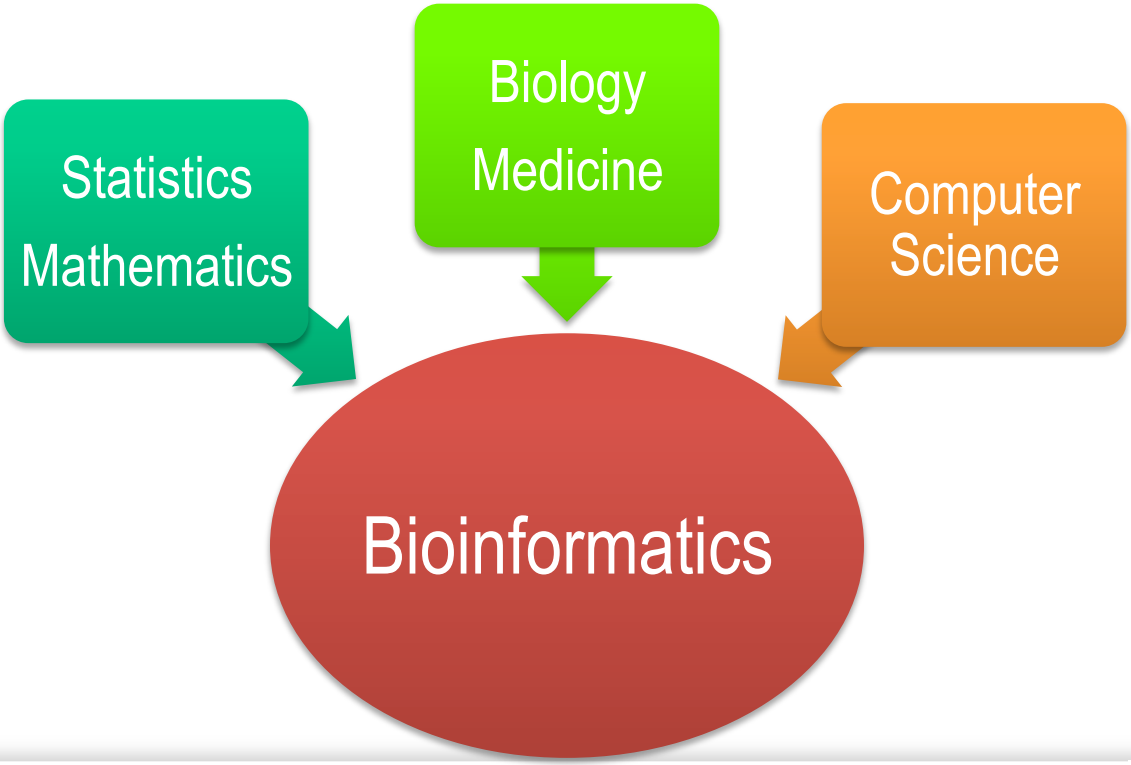
Intro

- Please read the syllabus carefully!
- We will use Blackboard for posting lectures/assignments and grading.
- Please keep practicing Linux and editor (VI, emacs, nano)
  - Contact us if you have any problem!!!!

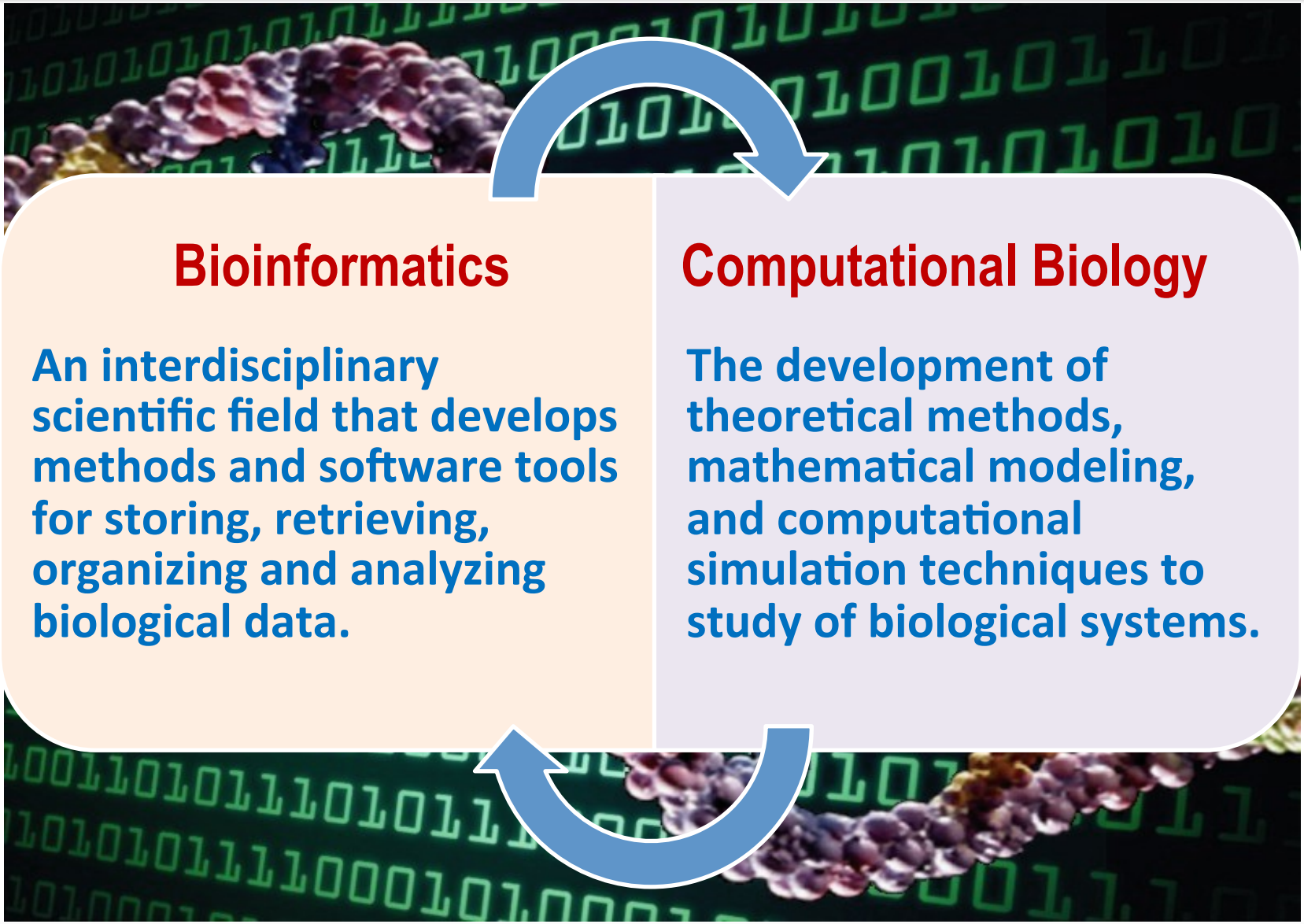
Bioinformatics?

What is Bioinformatics?

Application of techniques from computer science to problems from biology



Bioinformatics vs Computational Biology



Bioinformatics

An interdisciplinary scientific field that develops methods and software tools for storing, retrieving, organizing and analyzing biological data.

Computational Biology

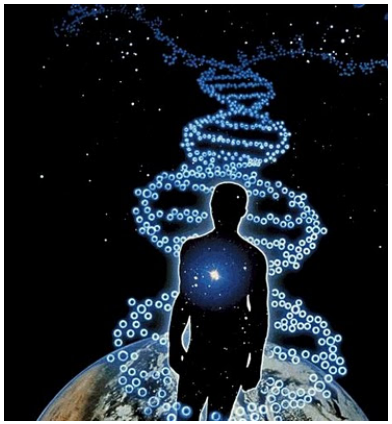
The development of theoretical methods, mathematical modeling, and computational simulation techniques to study of biological systems.

Computational Goals of Bioinformatics

- Learn & Generalize: Discover conserved patterns (models) of sequences, structures, interactions, metabolism & chemistries from well-studied examples.
- Prediction: Infer function or structure of newly sequenced genes, genomes, proteins or proteomes from these generalizations.
- Organize & Integrate: Develop a systematic and genomic approach to molecular interactions, metabolism, cell signaling, gene expression…
- Simulate: Model gene expression, gene regulation, protein folding, protein-protein interaction, protein-ligand binding, catalytic function, metabolism…
- Engineer: Construct novel organisms or novel functions or novel regulation of genes and proteins.
- Gene Therapy: Target specific genes, or mutations, RNAi to change a disease phenotype.

http://cmgm.stanford.edu/biochem218/01Genomics&Bioinformatics.pdf

Myriad Applications of Bioinformatics

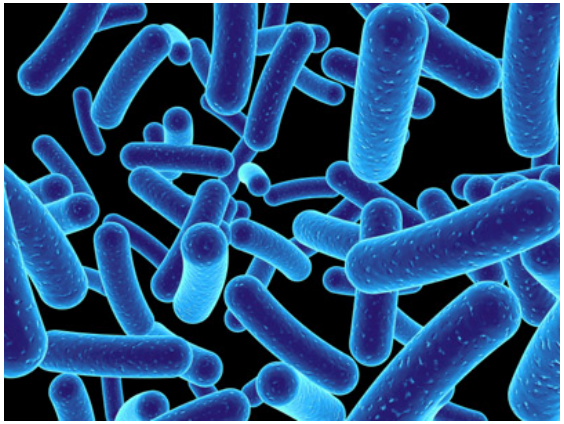


Human

Agriculture

Livestock

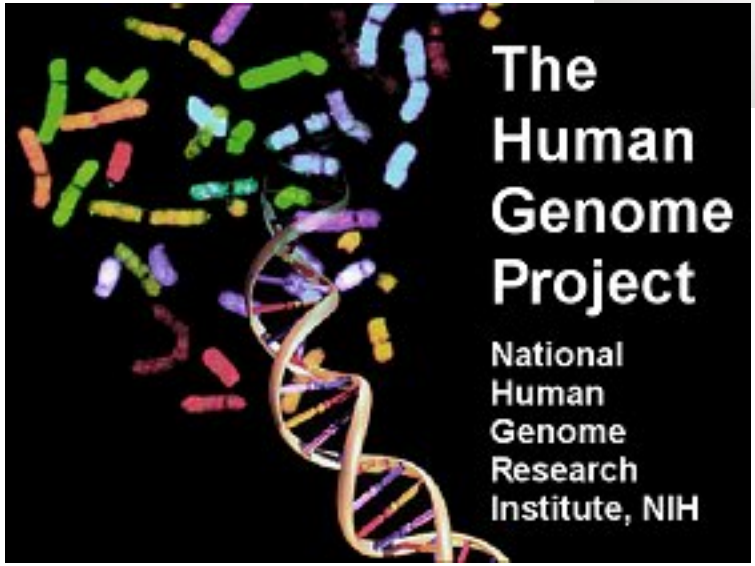
Ancestry

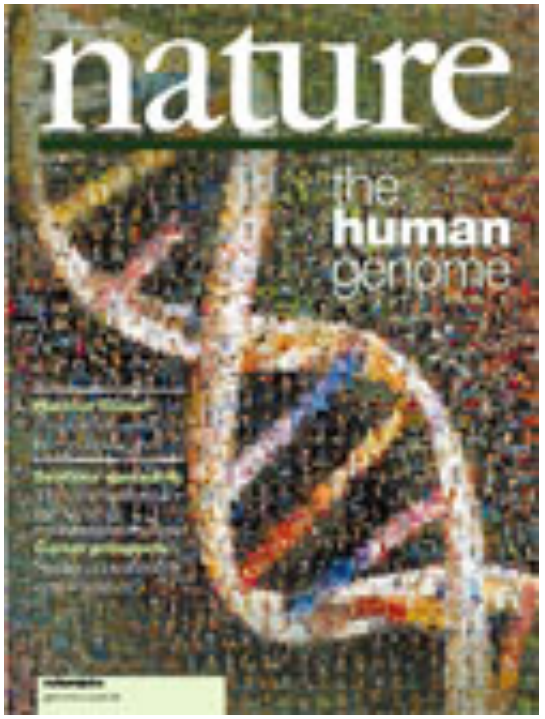


Microbes

Bioenergy

Human Genome Project (1990 – 2003)





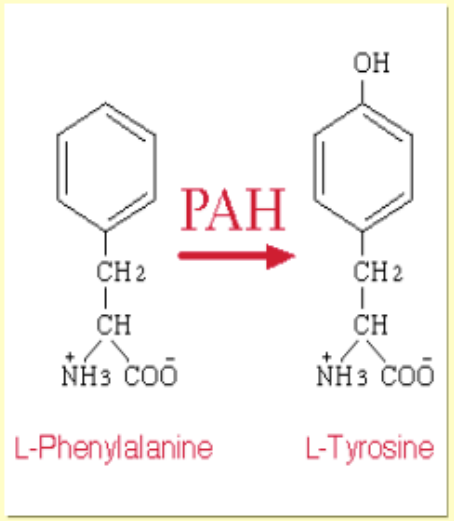
How to Sequence the Human Genome (Mac)

Health, Diseases, and Medicine

Central Paradigm of Bioinformatics

Genetic Information → Molecular Structure → Biochemical Function → Phenotype (Symptoms)

MVHLTPEEKT
AVNALWGKVN
VDAVGGEALG
RLLVVYPWTQ
RFFESFGDLS
SPDAVMGNPK
VKAHGKKVLG
AFSDGLAHLD
NLKGTFSQLS
ELHCDKLHVD
PENFRLLGNV
LVCVLARNFG
KEFTPQMQAA
YQKVVAGVAN
ALAHKYH



http://cmgm.stanford.edu/biochem218/01Genomics&Bioinformatics.pdf

Genomic Medicine

- An emerging medical discipline that involves using genomic information about an individual as part of their clinical care (e.g., for diagnostic or therapeutic decision-making) and the other implications of that clinical use.

The Path to Genomic Medicine

Personal Genome Project

http://my.biotechlife.net

Human Genome Project

Personal Genome

NIH (National Institutes of Health)

- The National Institutes of Health (NIH) is a biomedical research facility primarily located in Bethesda, Maryland.
- An agency of the United States Department of Health and Human Services, it is the primary agency of the United States government responsible for biomedical and health-related research.

- Yearly Budget: $ 30 Billion
- More than 5000 researchers are working

NIH NATIONAL CANCER INSTITUTE

1-800-4-CANCER Live Chat Dictionary

ABOUT CANCER | CANCER TYPES | RESEARCH | GRANTS & TRAINING | NEWS & EVENTS | ABOUT NCI | search

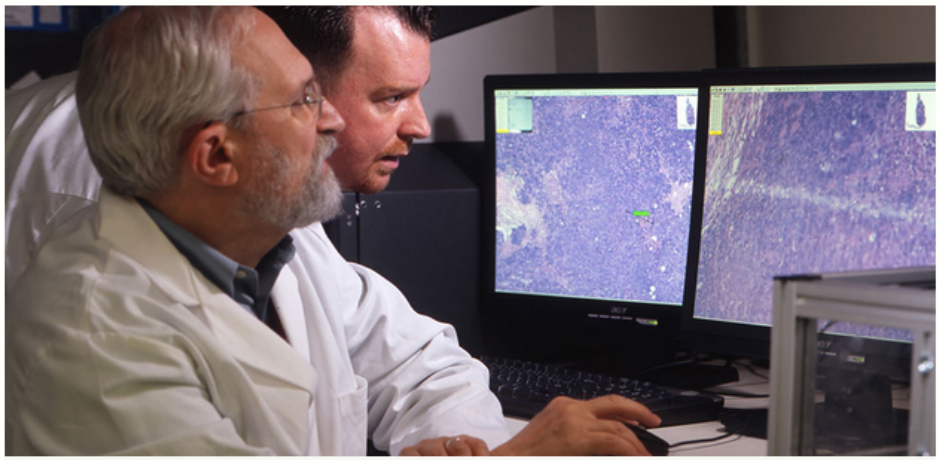
Home > About NCI > Budget and Appropriations > Building on Opportunities in Cancer Research > Building on the National Cancer Program

Bioinformatics to Accelerate Research

Bioinformatics, which enables the management and use of very large sets of molecular and clinical data, has become a core component of the NCI's research enterprise.

The National Cancer Informatics Program (NCIP), part of the NCI Center for Biomedical Informatics and Information Technology (CBIIT), is the institute's main bioinformatics initiative.

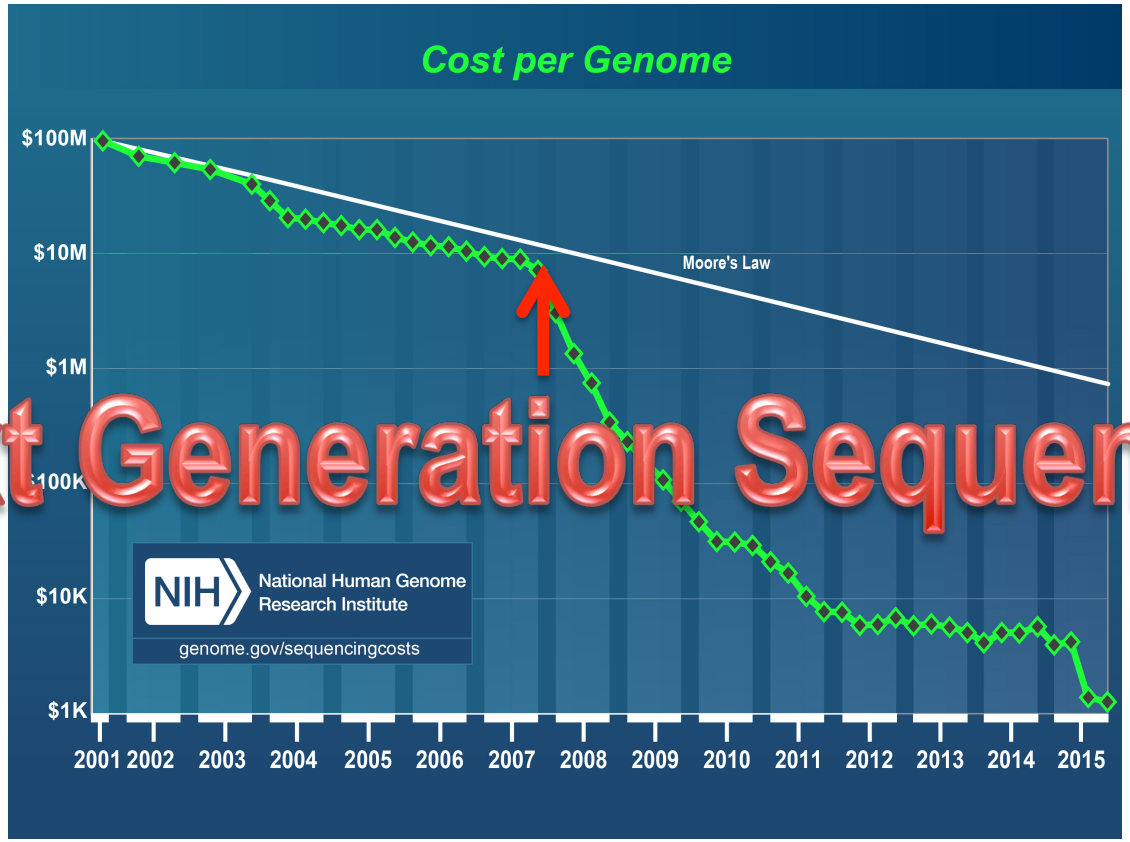
The collection, analysis, storage, retrieval, and distribution of "big data" are essential for many aspects of cancer research—especially for cancer genomics, in which millions of data points are frequently collected on each patient—and the monitoring of clinical trials.



DNA Sequencing Cost

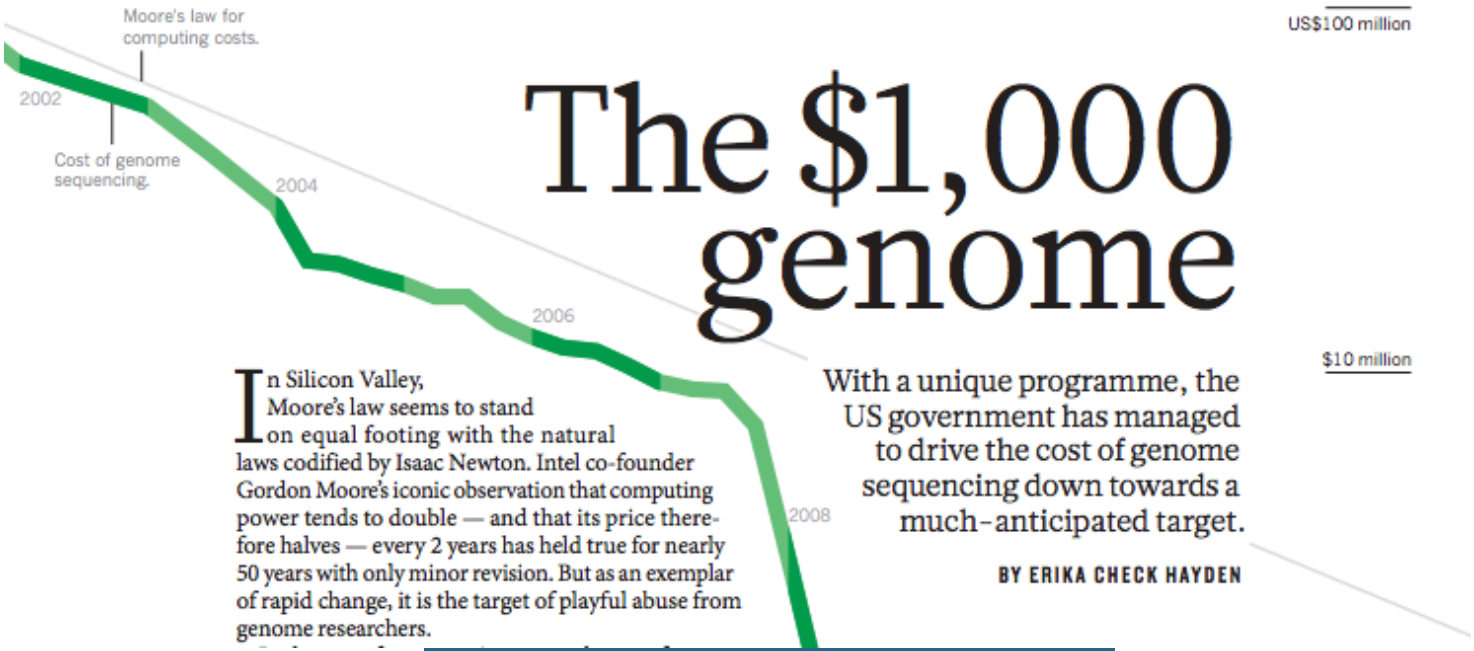
- Cost of sequencing a human-sized genome

  $1,000,000,000 US dollars (2001) →$1,000 (2015)



Next Generation Sequencing

The $1000 Genome

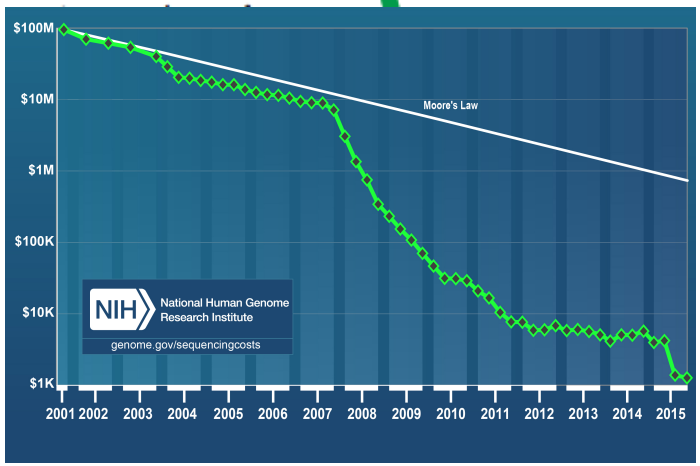


The $1,000 genome

With a unique programme, the US government has managed to drive the cost of genome sequencing down towards a much-anticipated target.

BY ERIKA CHECK HAYDEN

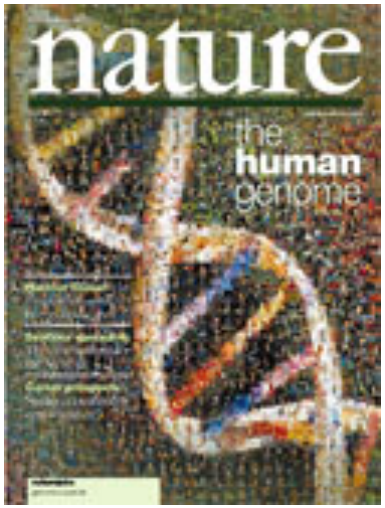
In Silicon Valley, Moore's law seems to stand on equal footing with the natural laws codified by Isaac Newton. Intel co-founder Gordon Moore's iconic observation that computing power tends to double — and that its price therefore halves — every 2 years has held true for nearly 50 years with only minor revision. But as an exemplar of rapid change, it is the target of playful abuse from genome researchers.



Sequencing a Human Genome

Human Genome Project

2003



~$1B

~ 6-8 years

Today

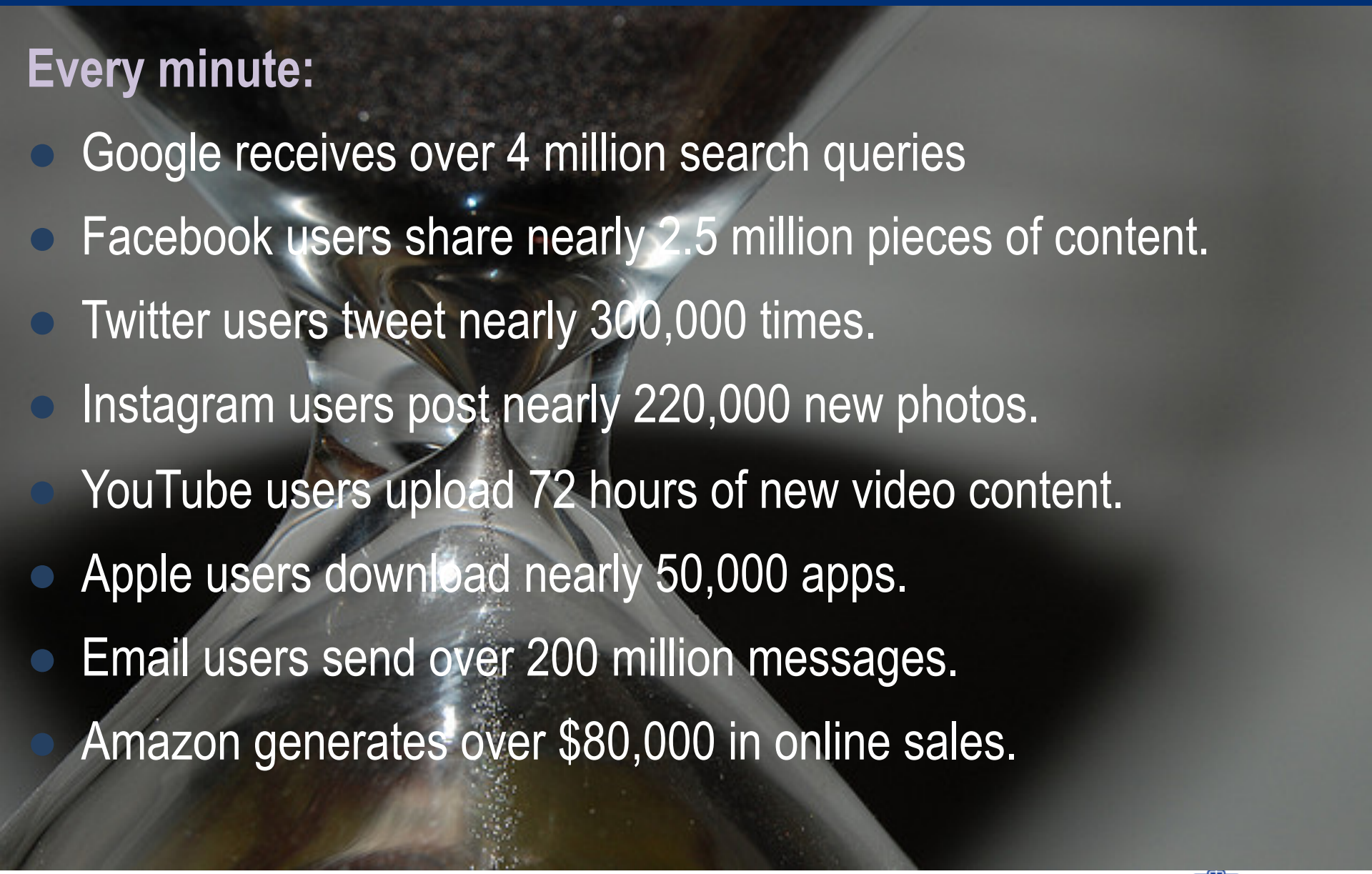
2017

~$2-4K

~ 1-3 days

The Current Bottleneck of Bioinformatics

Data Explosion!



Every minute:

- Google receives over 4 million search queries
- Facebook users share nearly 2.5 million pieces of content.
- Twitter users tweet nearly 300,000 times.
- Instagram users post nearly 220,000 new photos.
- YouTube users upload 72 hours of new video content.
- Apple users download nearly 50,000 apps.
- Email users send over 200 million messages.
- Amazon generates over $80,000 in online sales.

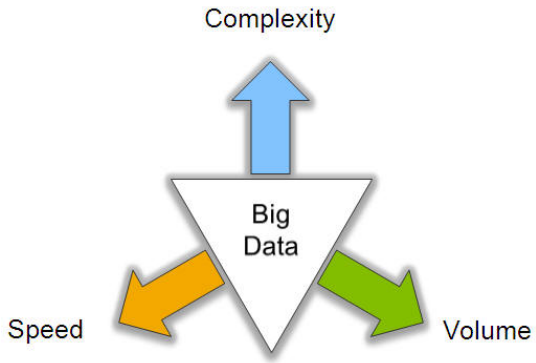
Big Data

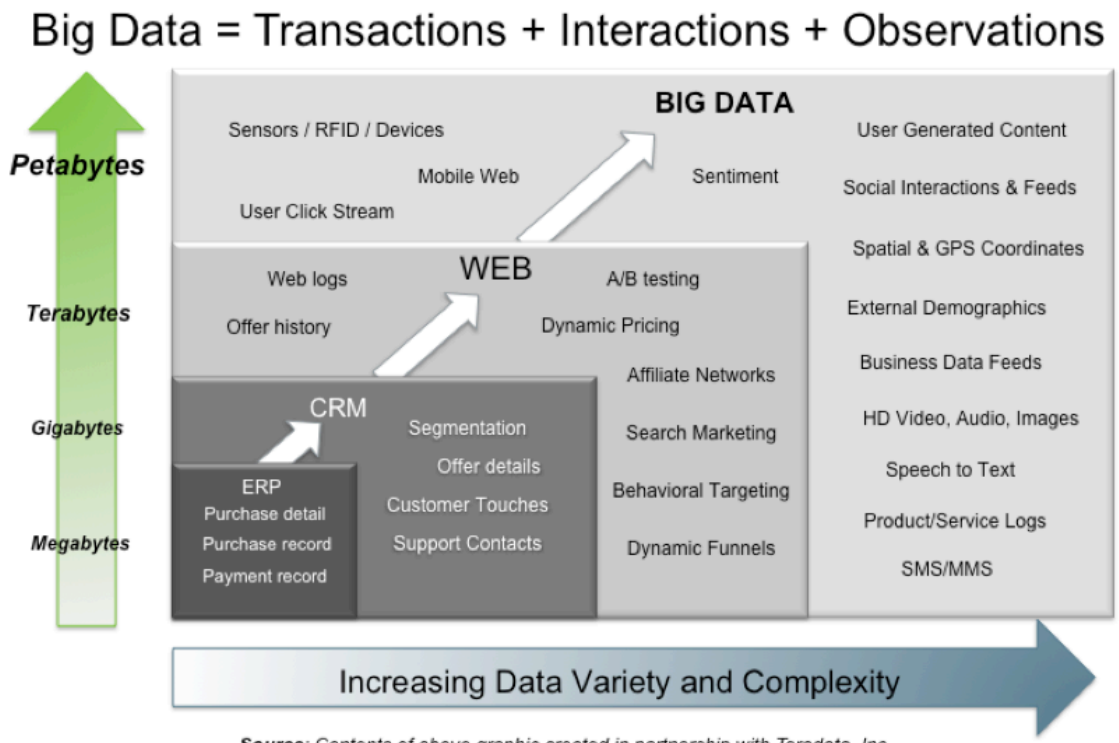

nature
BIG DATA
SCIENCE IN THE PETABYTE ERA


Science
data
AAAS


big
data

Big Data 3V's





***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

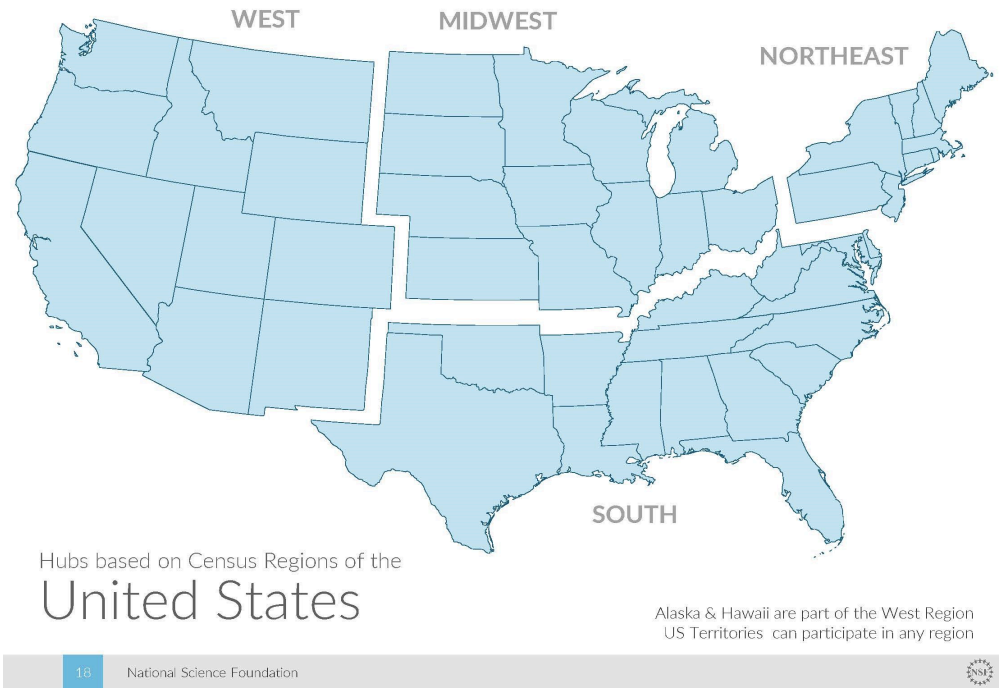
Data Analysis Bottleneck

Big Data to Knowledge (BD2K)

Thursday, October 9, 2014

NIH invests almost $32 million to increase utility of biomedical research data

NSF Big Data Research Initiative



4 Regional Big Data Hub

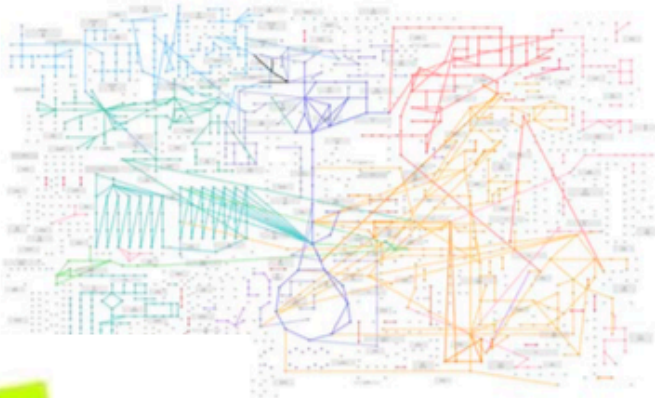
Challenges or Opportunities?

- ✓ Large amount of data.
- ✓ The need for the relevant models.
- ✓ The computational complexity of these models.
- ✓ Various simulation methodologies for these models.

Biological, mathematical and computational challenges related to algorithm development and efficient computation are essential.

Not Just Nucleotide Sequence Data…

Protein st

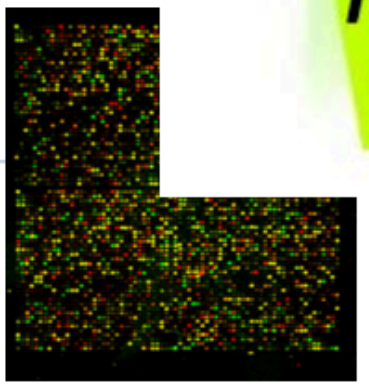


olic pathway data

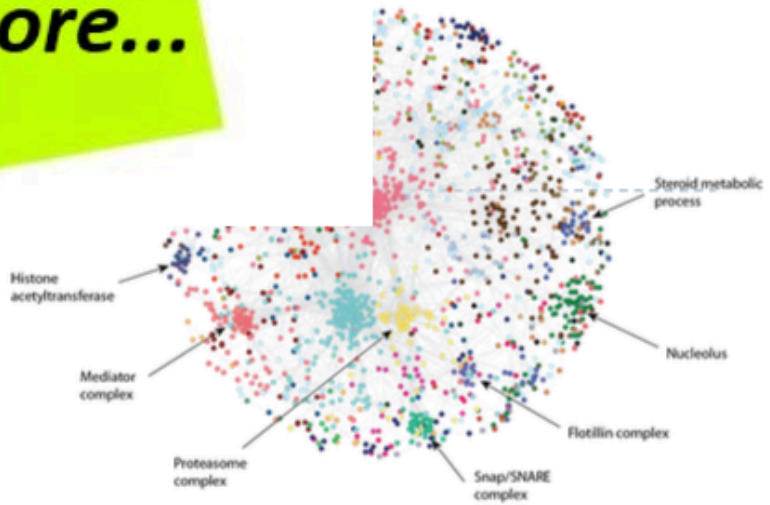
and many
More…

MKLTLKNLSMAIMMSIVMGS
SAMAADSNEK GAS
GYLPEHTLF
ADYLEQD
LHDHYLDI
DRARKDG
DEIKSLKF
QTYPGRFPMG
HTFEEEIEFVQGLNHSTGK
NIGIYPEIKAPWFHQEGKDI
AAKTLEVLKKYGYTGKDDKV

Protein sequence data



Gene expression data



Protein interaction network data

Research Initiative

Genome assembly

Genome reconstruction

Metagenomics

Genetic variation

RNA-Seq

Protein function prediction

Applications

Computing Technologies

Multicores

GPUs

Clusters

Clouds

Supercomputers

Next-Gen Sequencing Technologies

454

Illumina

Solid

PacBio

Oxford Nanopore

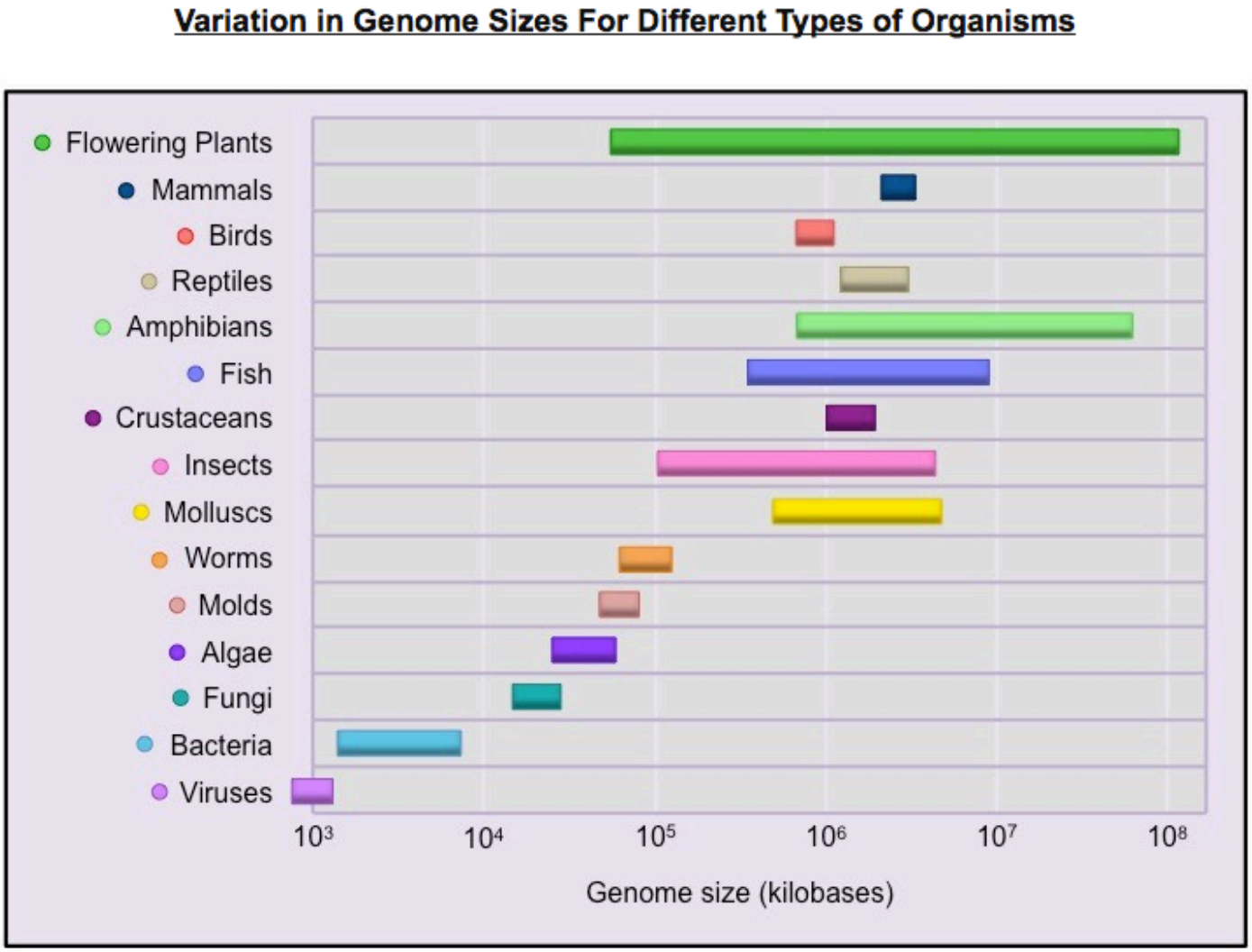
Genome Size

Comparison of Genome Size in Different Organisms

Species	*T2 phage*	*Escherichia coli*	*Drosophila melanogaster*	*Homo sapiens*	*Paris japonica*
Genome Size					
Common Name	 Virus	 Bacteria	 Fruit fly	 Human	 Canopy Plant

http://ib.bioninja.com.au/standard-level/topic-3-genetics/32-chromosomes/genome-size.html

Genome Size



http://ib.bioninja.com.au/standard-level/topic-3-genetics/32-chromosomes/genome-size.html

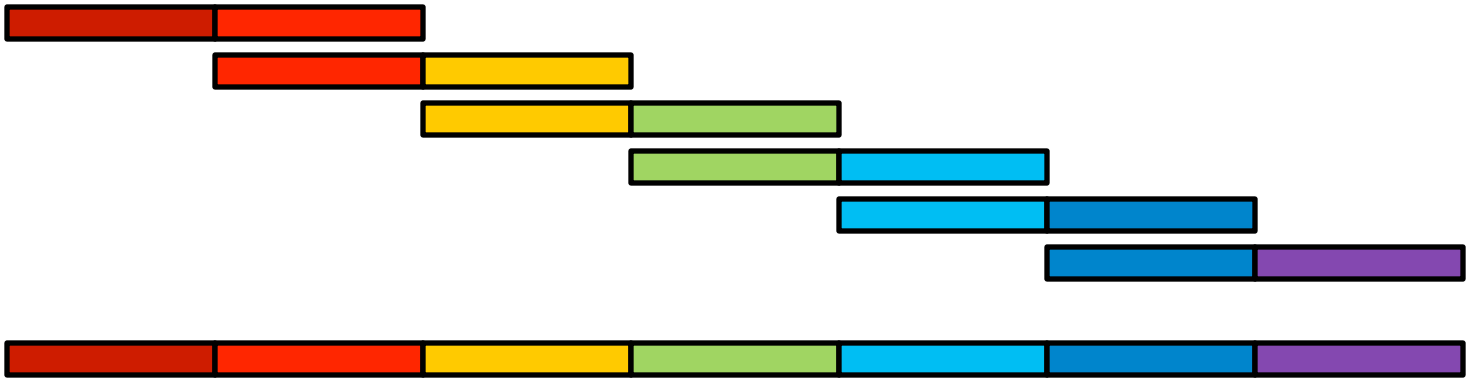
Assembling a Genome

1. Shear & Sequence DNA

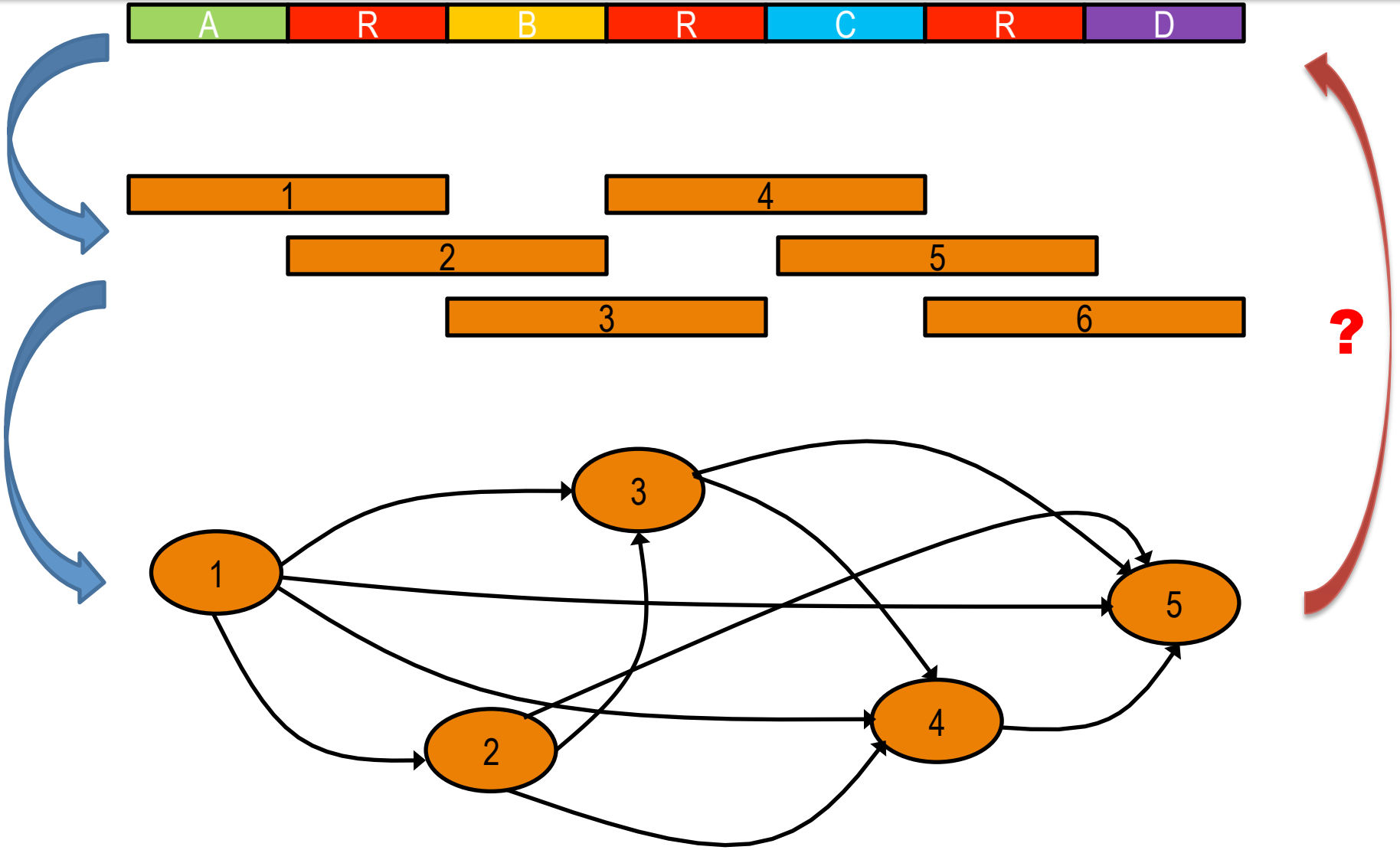
2. Construct assembly graph from overlapping reads

…AGCCTAGGATGCGCGACACGT
TGCGCGACACGTCGCATATCCGGTTGAT
TCCGGTTGATCACGAATACG…

3. Simplify assembly graph



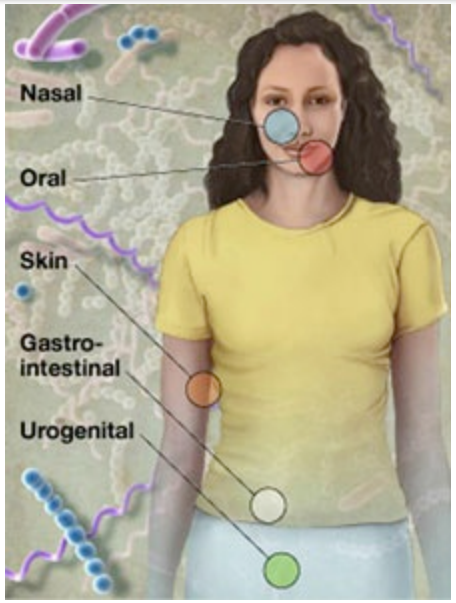
Assembly Complexity

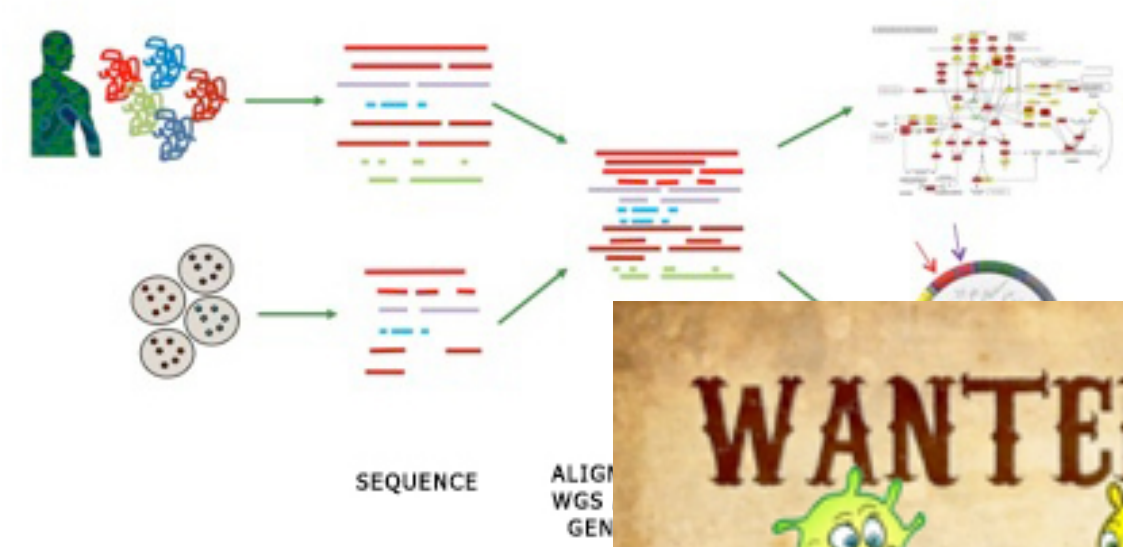


What is Metagenomics?

Culture-independent genomic analysis of a community of microorganisms from **environmental sample**.

Human Microbiome Project





NIH HUMAN MICROBIOME PROJECT

- From 242 healthy U.S
- More than 5,000 sam
- 3000 reference geno genomes are available

image courtesy of the NIH Common Fund

2011 German *E. coli* Outbreak.

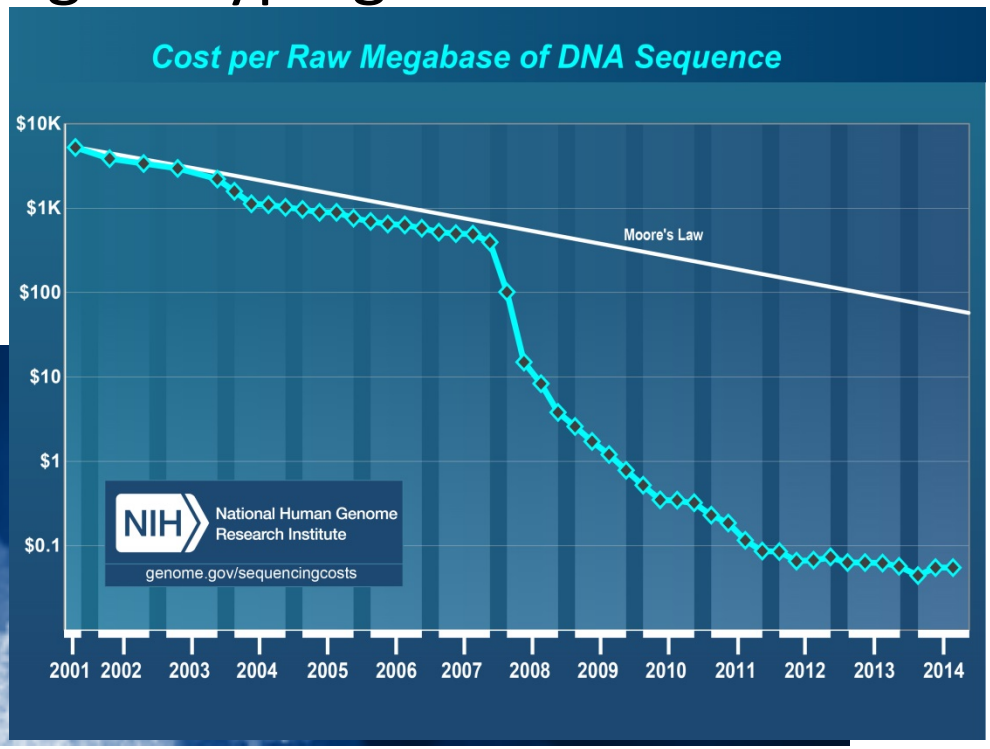
In May through June 2011, a novel strain of *Escherichia coli* O104:H4 bacteria caused a serious outbreak of foodborne illness.

In all, 3,950 people were affected and 53 died.

2014 Ebola Virus Disease

Metagenomics and Biosurveillance

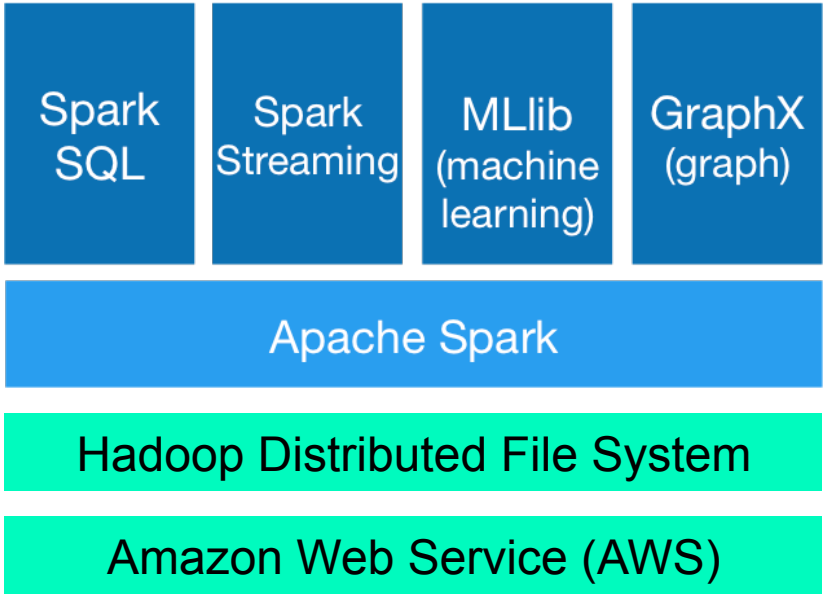
- Conventional methods to identify pathogens: laboratory culture, immunoassays, and genotyping



- NGS metagenomic analysis can be a promising new method for biosurveillance.

New Big Data Technology - Spark

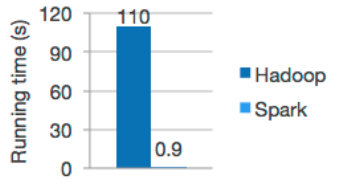
- An emerging in-memory processing framework
  - Iterative machine learning
  - Interactive data analytics
  - Scala based implementation
  - Master-Slave, HDFS, Fault Tolerance



Speed

Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.

Spark has an advanced DAG execution engine that supports cyclic data flow and in-memory computing.



Logistic regression in Hadoop and Spark

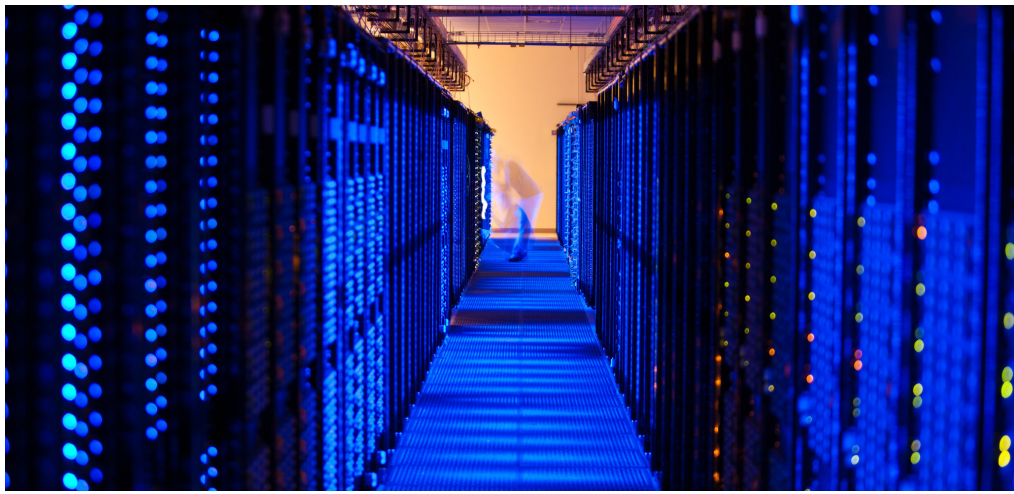
Ease of Use

Write applications quickly in Java, Scala or Python.

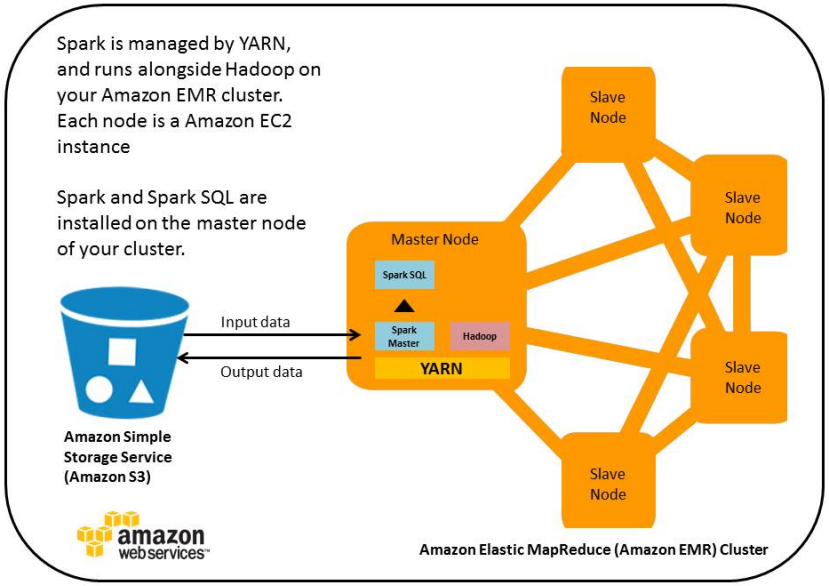
```
file = spark.textFile("hdfs://...")

file.flatMap(lambda line: line.split())
    .map(lambda word: (word, 1))
    .reduceByKey(lambda a, b: a+b)
```

Spark GraphX

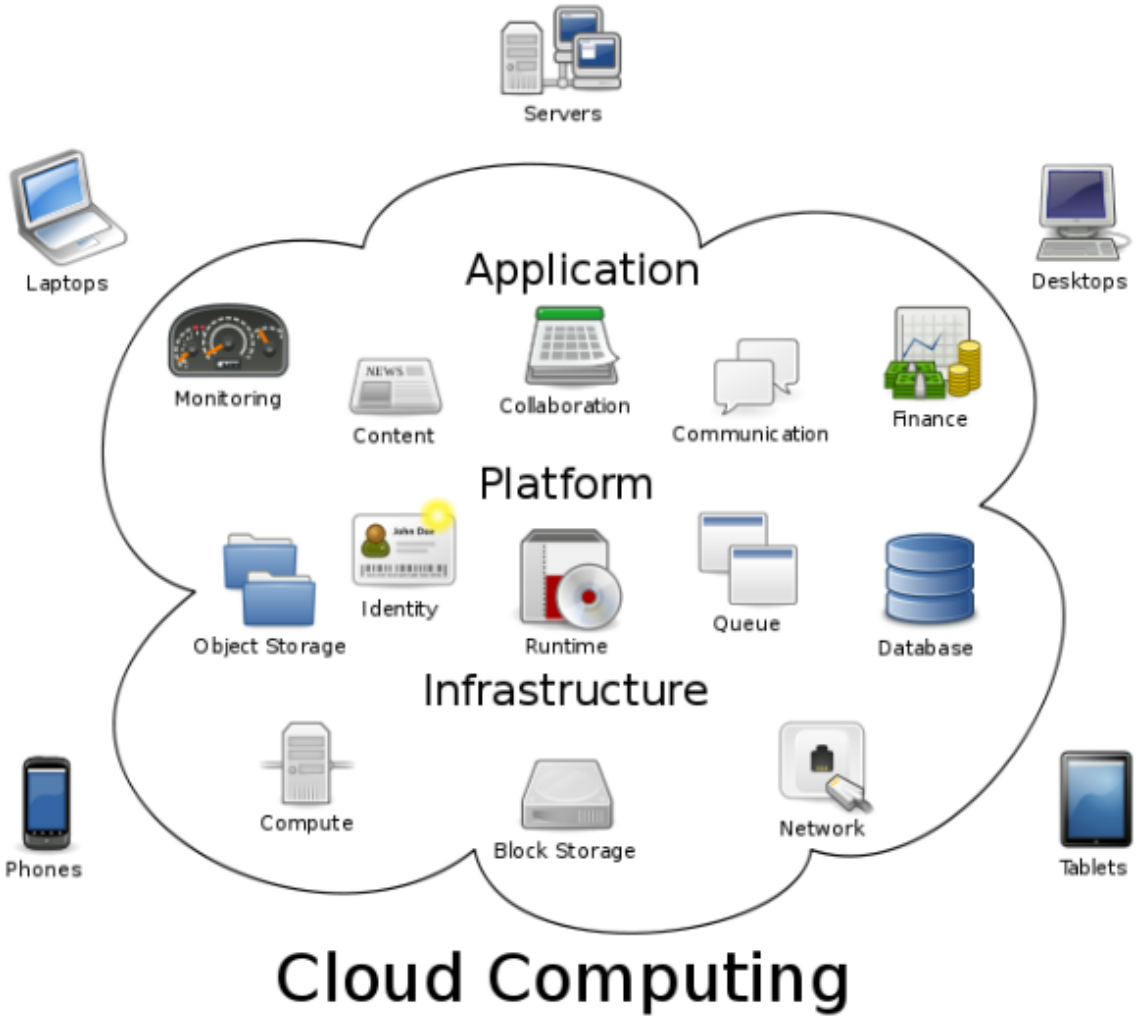
Cloud (Amazon Web Service) + Spark



Lightning-Fast Cluster Computing



Cloud Computing

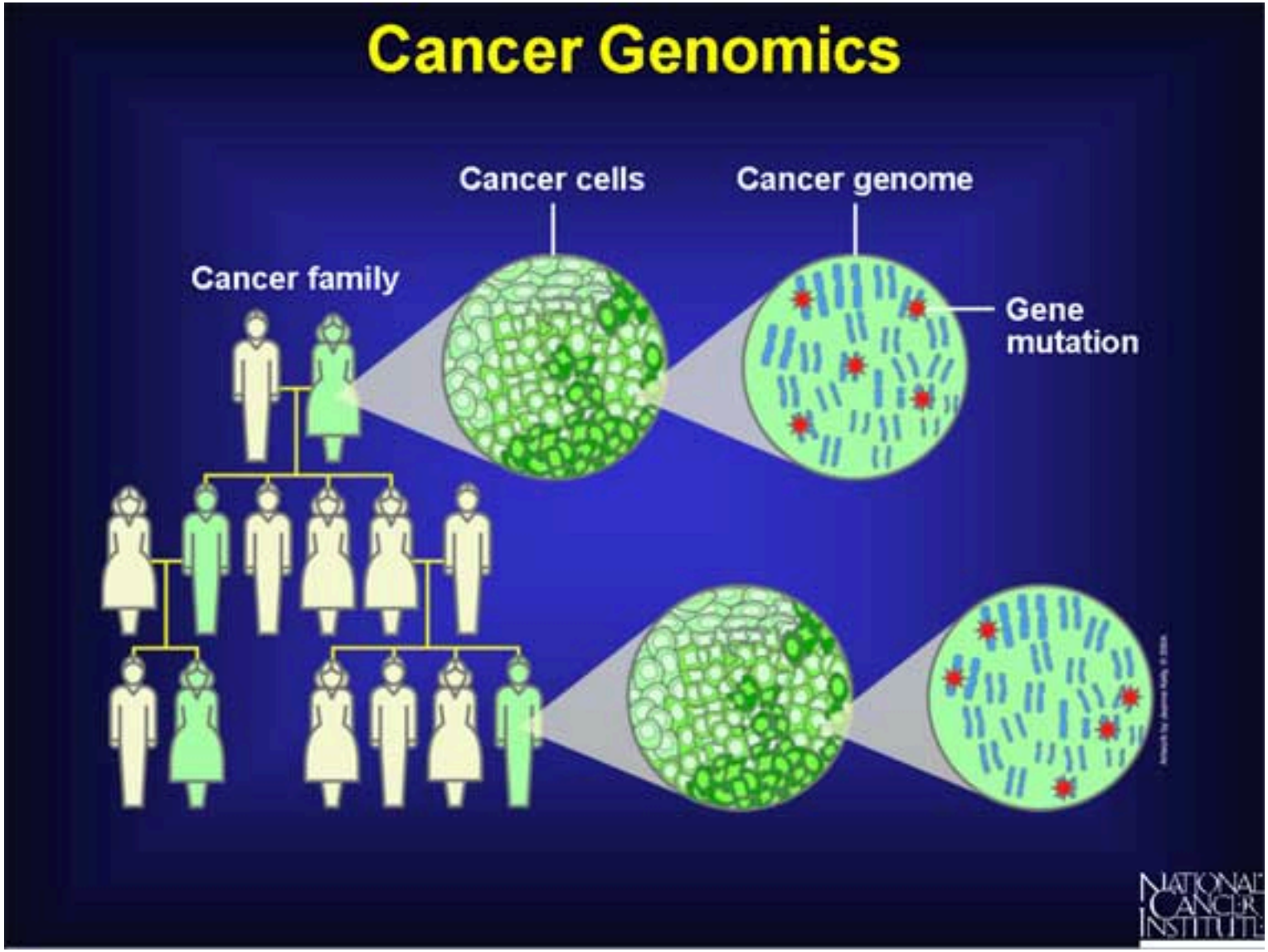


wikipedia: Cloud Computing

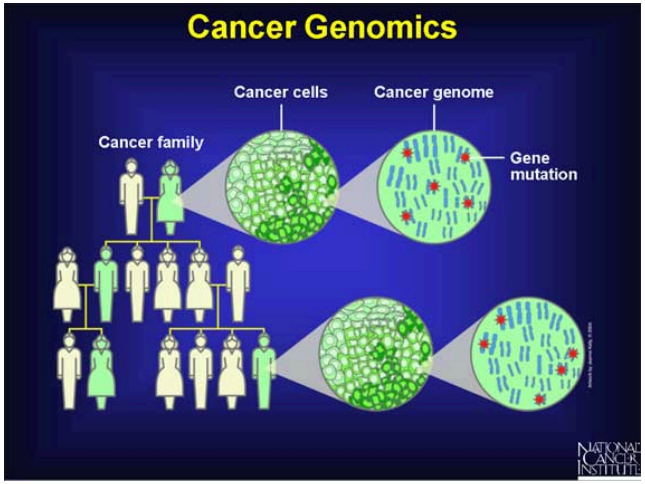
Benefits

- Cost & management
  - Economies of scale, “out-sourced” resource management
- Reduced Time to deployment
  - Ease of assembly, works “out of the box”
- Scaling
  - On demand provisioning, co-locate data and compute
- Reliability
  - Massive, redundant, shared resources
- Sustainability
  - Hardware not owned

Genomics and Cancer

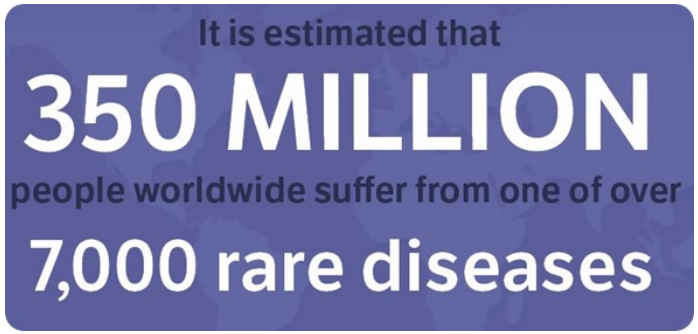


Hot Research Topics in Genomic Medicine



Cancer Genomics

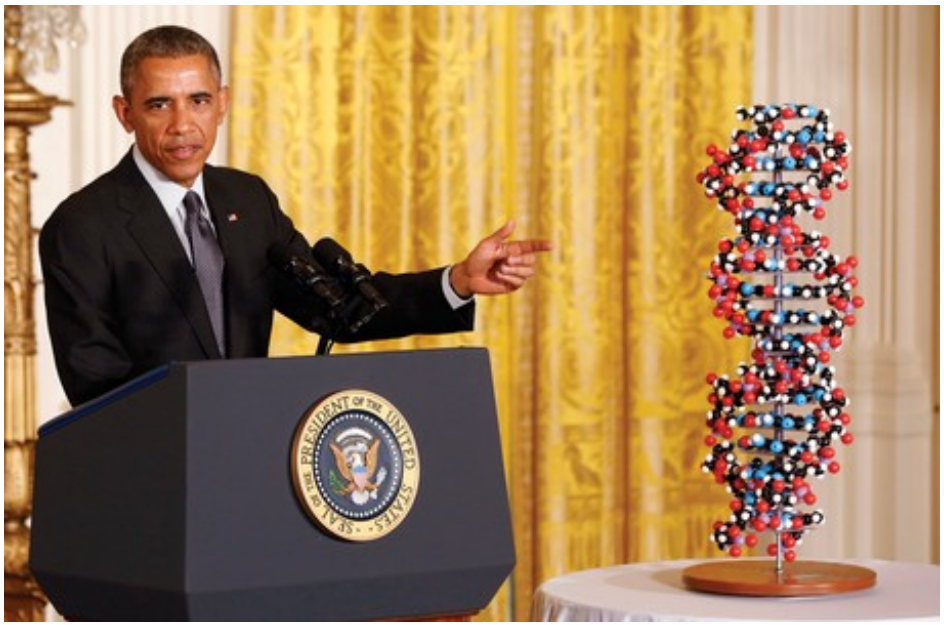
Pharmacogenomics



Ultra-Rare Disease Diagnosis

Prenatal & Newborn Genomic Analysis

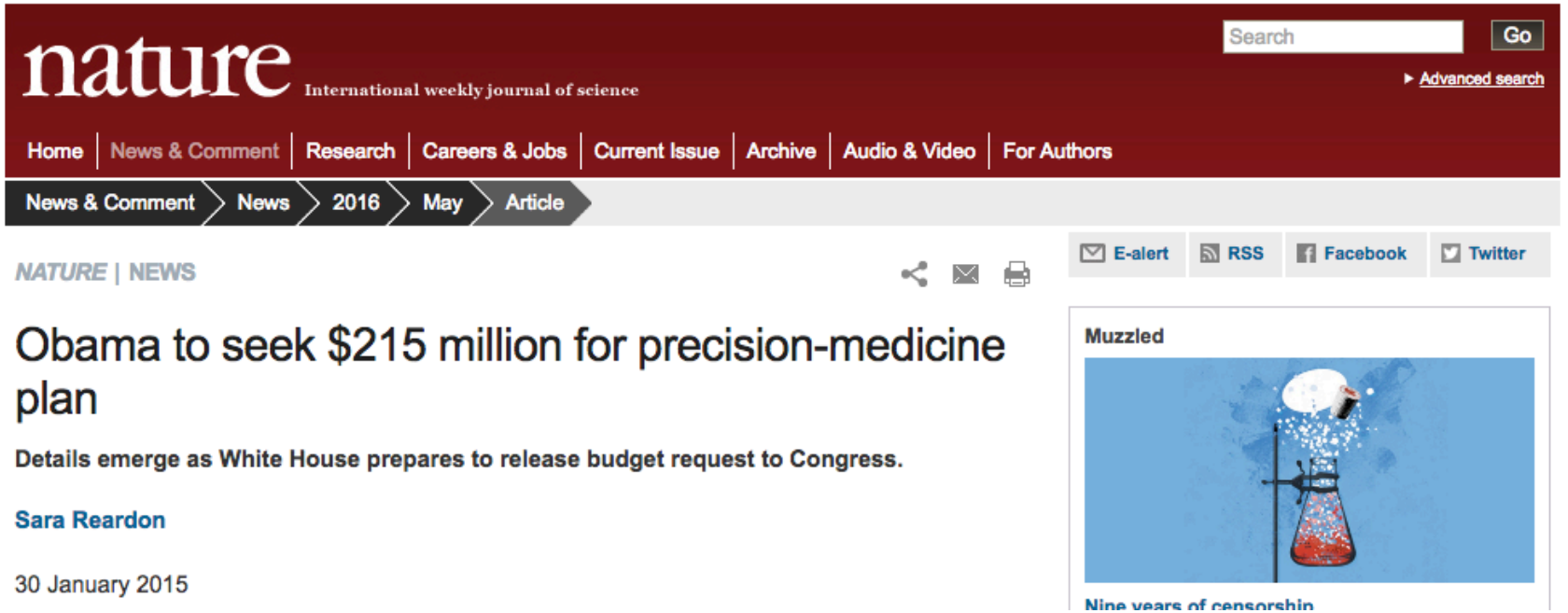
Precision Medicine



- A broader context for ‘individualizing’ medical care to advanced human health

Precision Medicine

- Today: most medical care based on expected response of the average patient
- Tomorrow: medical care based on individual genomic, environmental, and lifestyle differences that enable more precise ways to prevent and treat diseases.



An integrative bioinformatics/biomedical knowledge of big data

- It is not easy to analyze the large-scale data to derive meaningful biological correlations by
  - absence of easy-to-use knowledge-based bioinformatics tools
  - difficulty to select appropriate software tools
  - lack of computing system resources

Bioinformatics + Bio Tech + Big Data + HPC = Gold Mine!!

Question?

Thank you for listening and I will be happy to take questions!