Introduction to cloud computing

Sudhakaran Prabakaran & Matt Wayland

Lecture 1

Lecture 1: Introduction to cloud computing

1. Background

2. Why Cloud Computing?

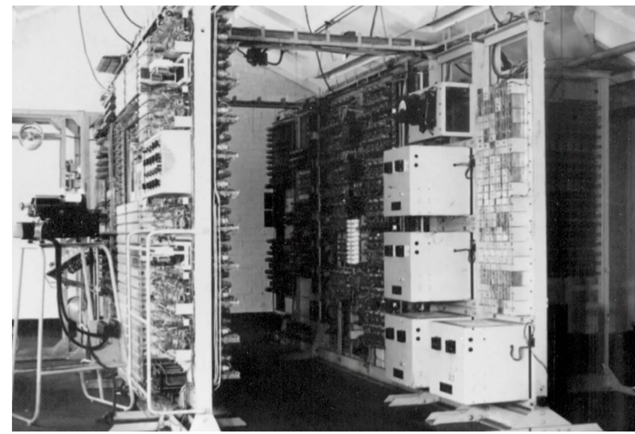
Some history & background

First computers!

1939-42

ABC computer that solved 29 problems at once

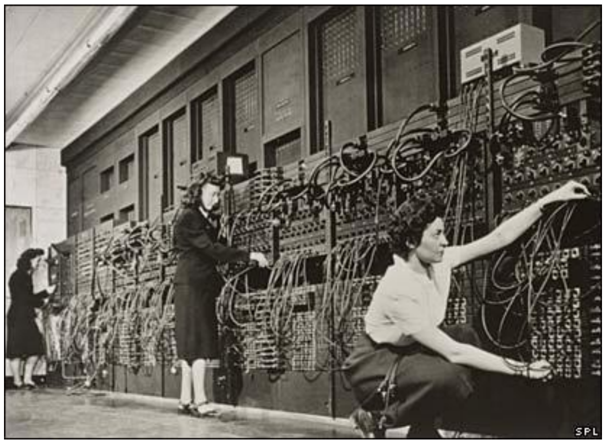
1939-45



British engineers built Colossus to crack WW2 codes

DNA the genetic material 1944

1943-45

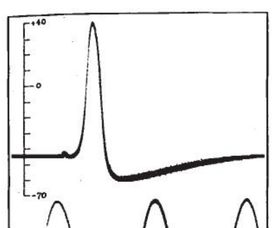


ENIAC computer solved 1000s of MATH problem per second

1964

CDC 6600 can do 3 million tasks per second

Hodgkin-Huxley Model in 1952



Jacob and Monad Lac Operon Model in 1961

1975

Altair 8800

1977

Apple II

1.023 Mhz, 4 kB of RAM

Recombinant DNA technology and automated sequencing

1995

486SX 16 Mhz

H. *influenzae*, first genome sequenced

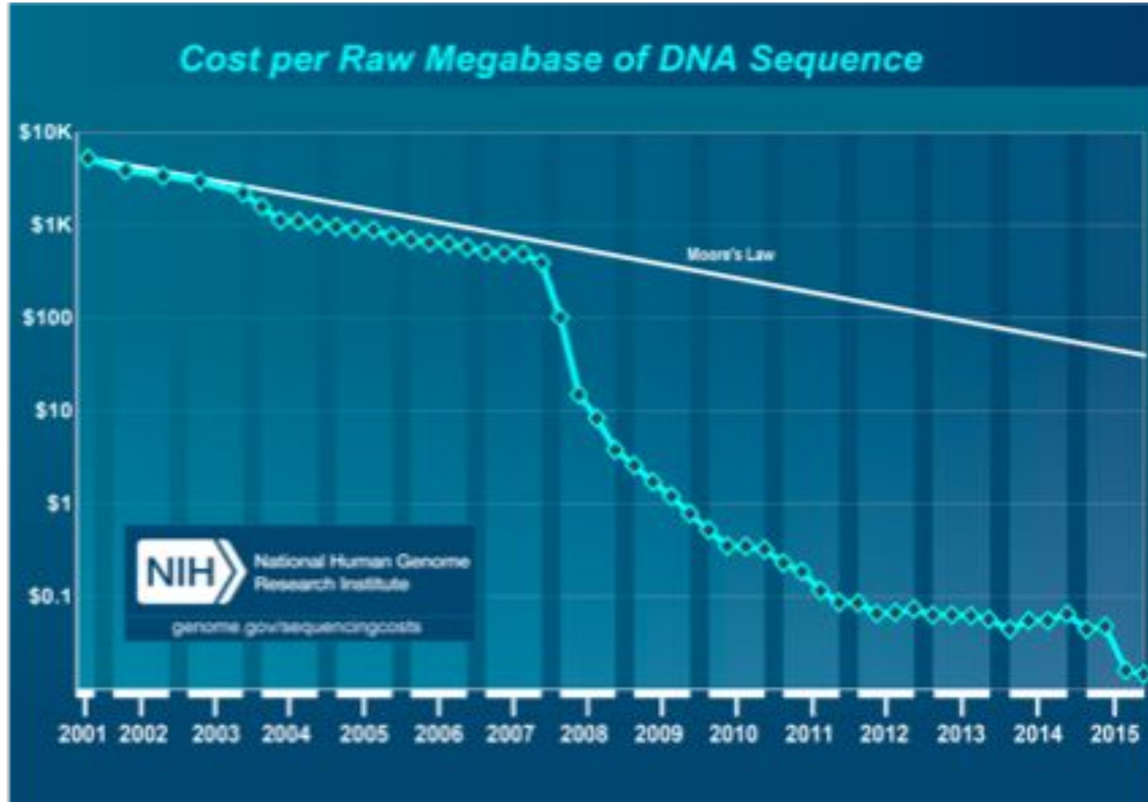
2001

Human genome sequenced
13 years and 3 billion USDs

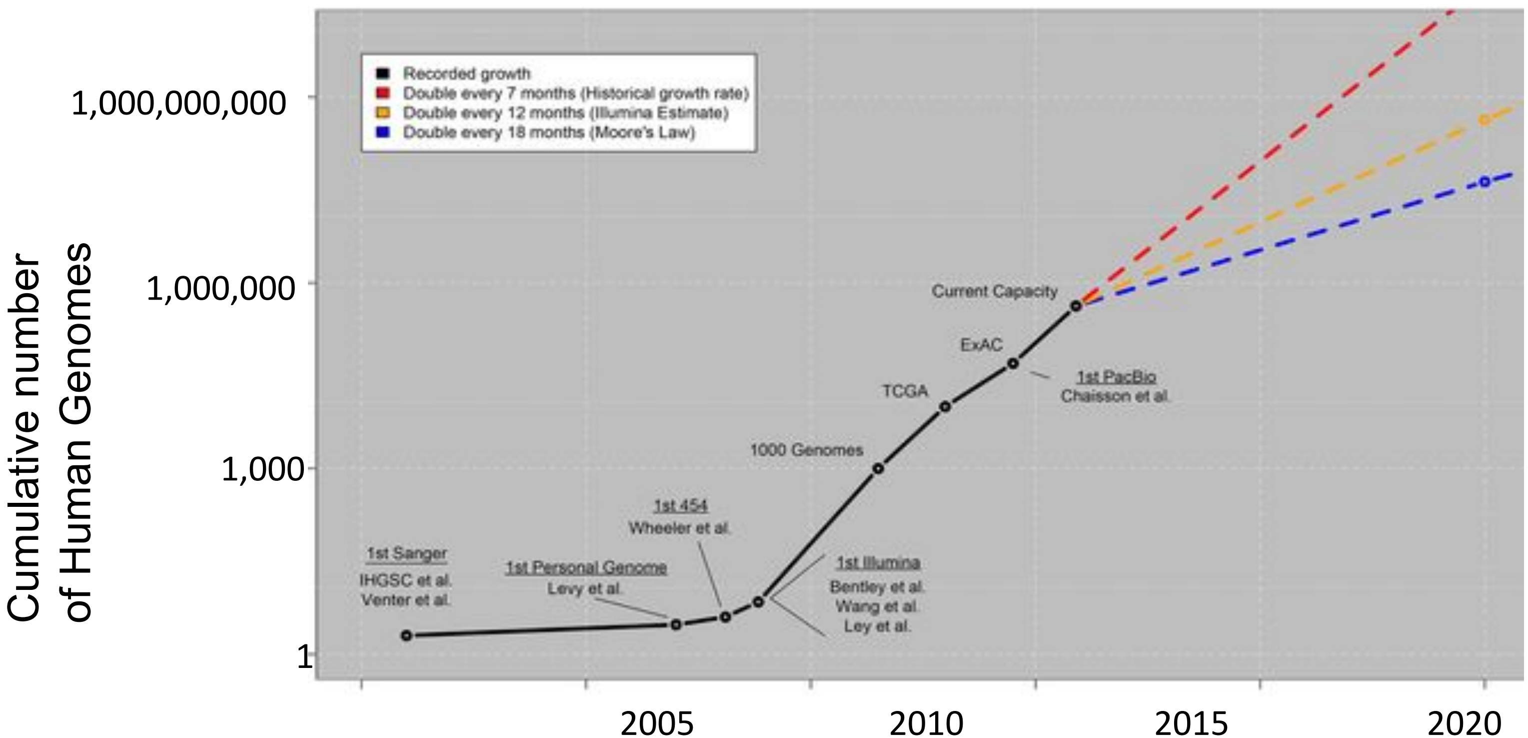
Sequencing data explosion: faster than moore's law over time

DNA sequencing has gone through technological S-curves

- In the early 2000's, improvements in Sanger sequencing produced a scaling pattern similar to Moore's law.
- The advent of NGS was a shift to a new technology with dramatic decrease in cost).



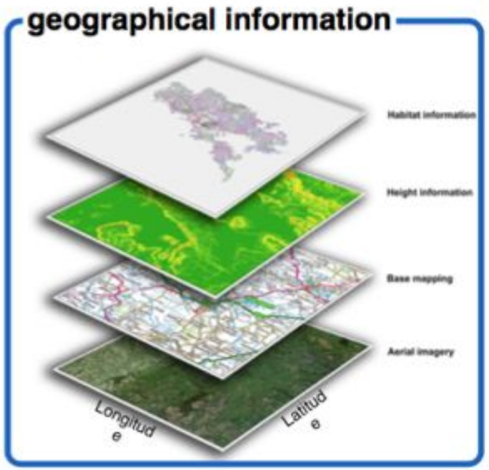
“More human genomes” creates more challenges



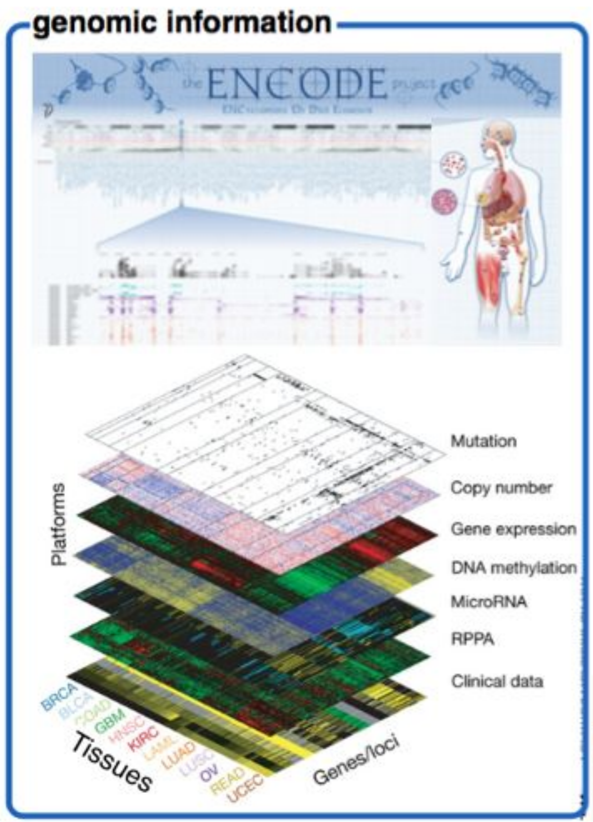
Between 2014-2018, production of new NGS data to exceed ***2 Exabytes***

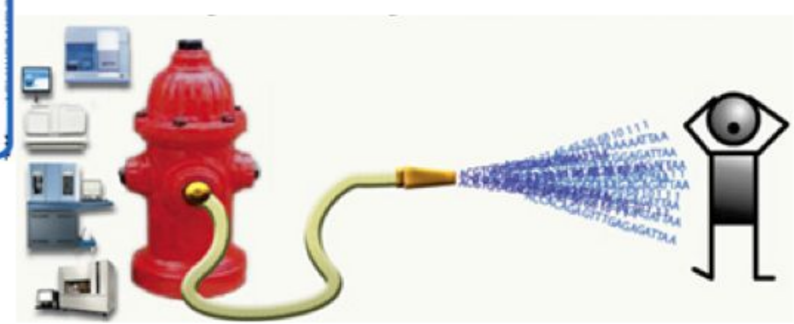
More and more data!

Human genome annotation — a non-intuitive map



- Large-scale organisation providing an overview of the genome
- Integration of heterogeneous data

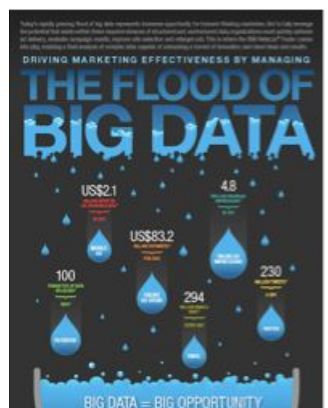




Big data world

IBM estimates that 90% of world’s data has been created in the last two years

Commercial World Data: Financial & Retail Data



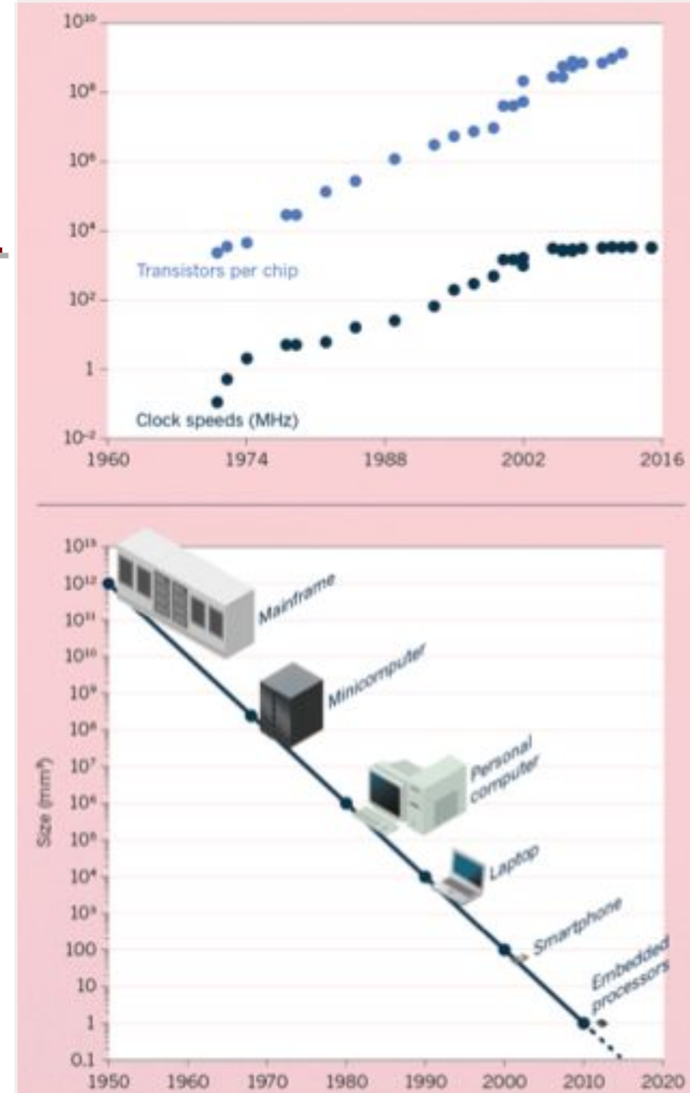
*

Human brain	2.5 PB
Spotify	10 PB
Ebay	90 PB
Facebook	300 PB
Google	15000 PB

***Royal Society website**

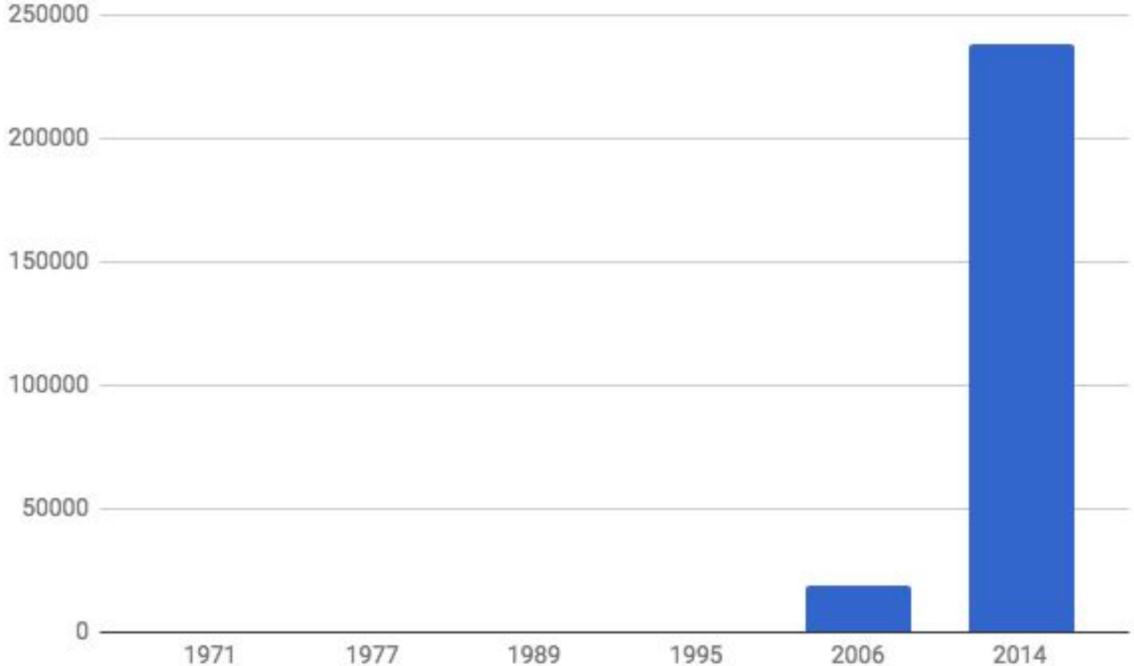
Moore's Law: Exponential Scaling of Computer Technology

- Exponential increase in the number of transistors per chip.
- Led to improvements in speed and miniaturization.
- Drove widespread adoption and novel applications of computer technology.



[Waldrop ('15) Nature]

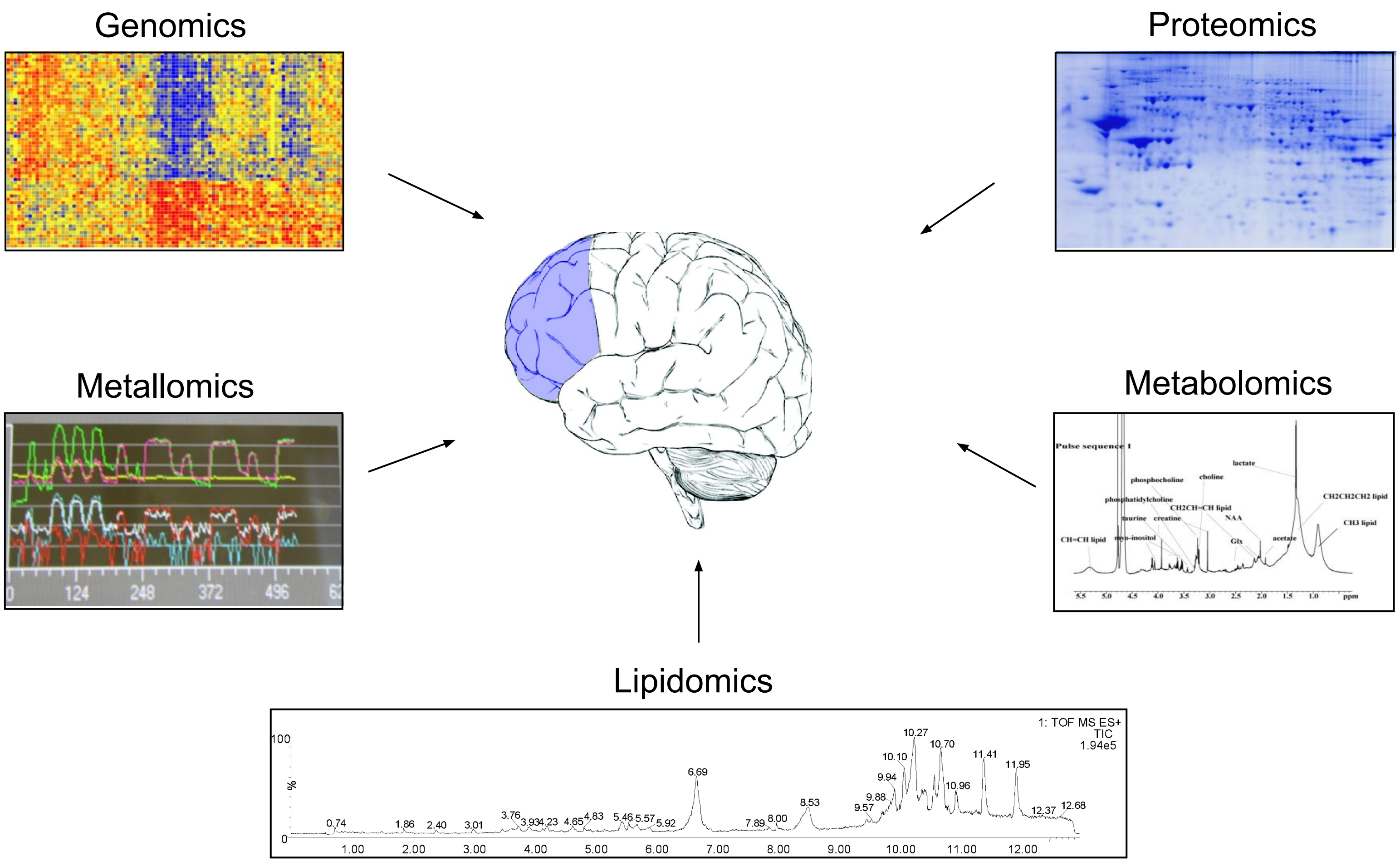
Processing speeds



Intel 4004 | Intel 8085 | Intel 80486 | Pentium Pro | Triple Core | Intel i7

My own path

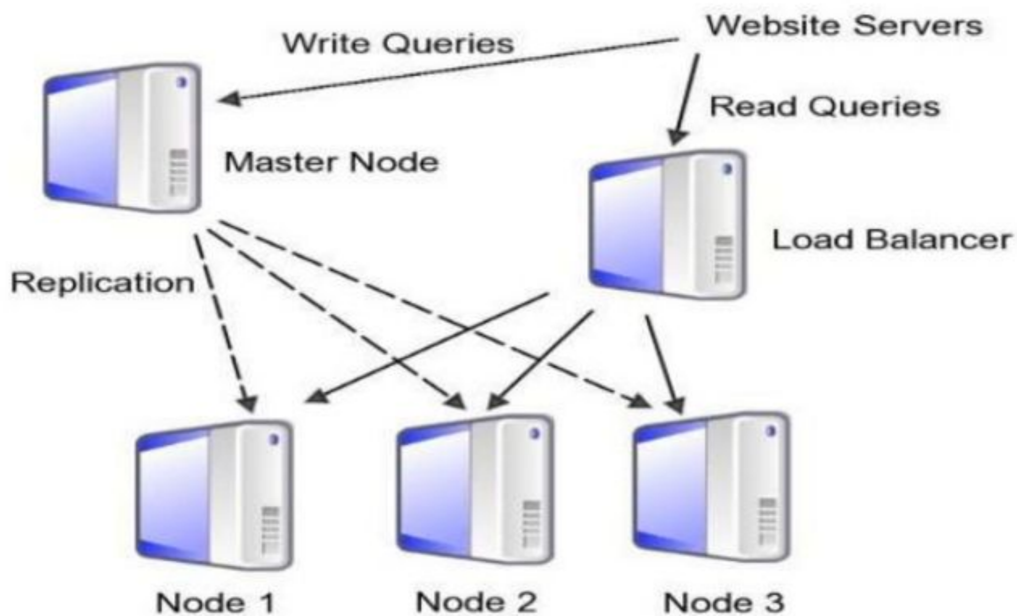
Systems Neuroscience



Prabakaran, S., et al. Mitochondrial dysfunction in schizophrenia: evidence for compromised brain metabolism and oxidative stress. *Mol Psychiatry,* 2004. 9, 684-97, 643.

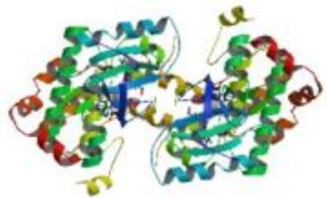
My Ph.D. work started with this one computer in 2002

Soon became five computer study



Postdoctoral work:- Two site phosphorylation network

Mek



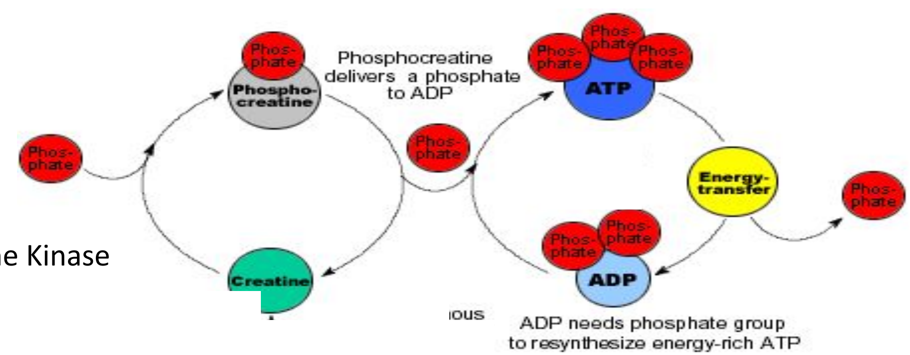
+

Erk

+

Mkp3

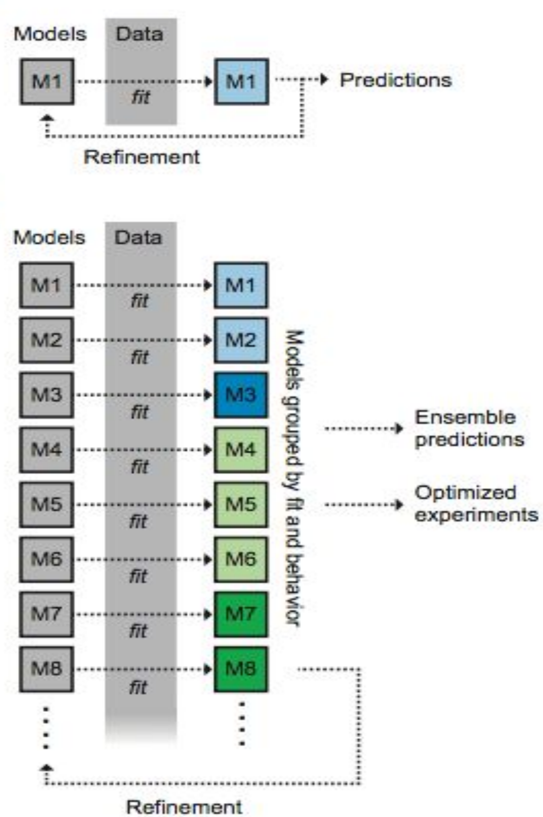
Creatine Kinase



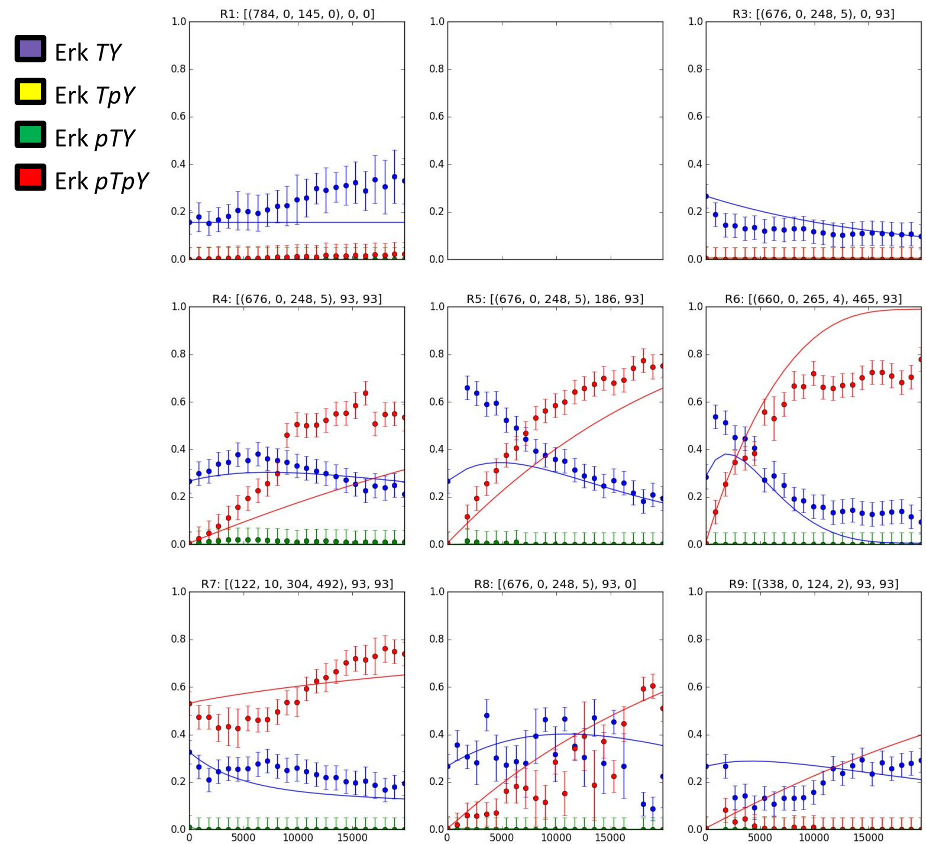
+ 0.1 mg/ml BSA

Ensemble modeling and Bayesian statistics

Ensemble model predictions



Data



Prabakaran, S., Gyori, B. and Gunawardena, J. Regulation of Erk by multisite phosphorylation-a three body problem in biochemistry, Manuscript in preparation, 2016

Thousands of computers in the cluster

Orchestra
HIGH PERFORMANCE COMPUTING CLUSTER

Please see our Orchestra status page for known issues.

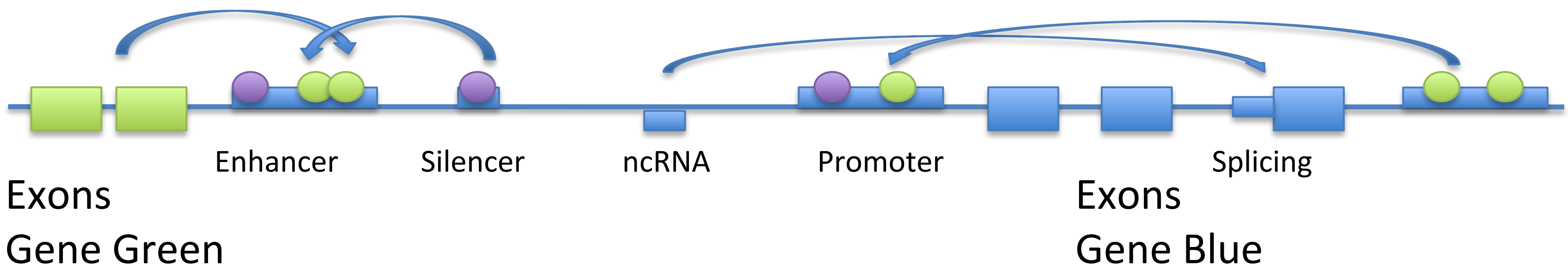
The **Orchestra** platform provides UNIX-based high performance computing, web hosting, and database hosting services at Harvard Medical School.

Orchestra and its associated services are managed by the **Research Computing Group**, part of the HMS Information Technology Department.

2016

Current work

1. Identifying novel translation from noncoding regions using proteogenomics



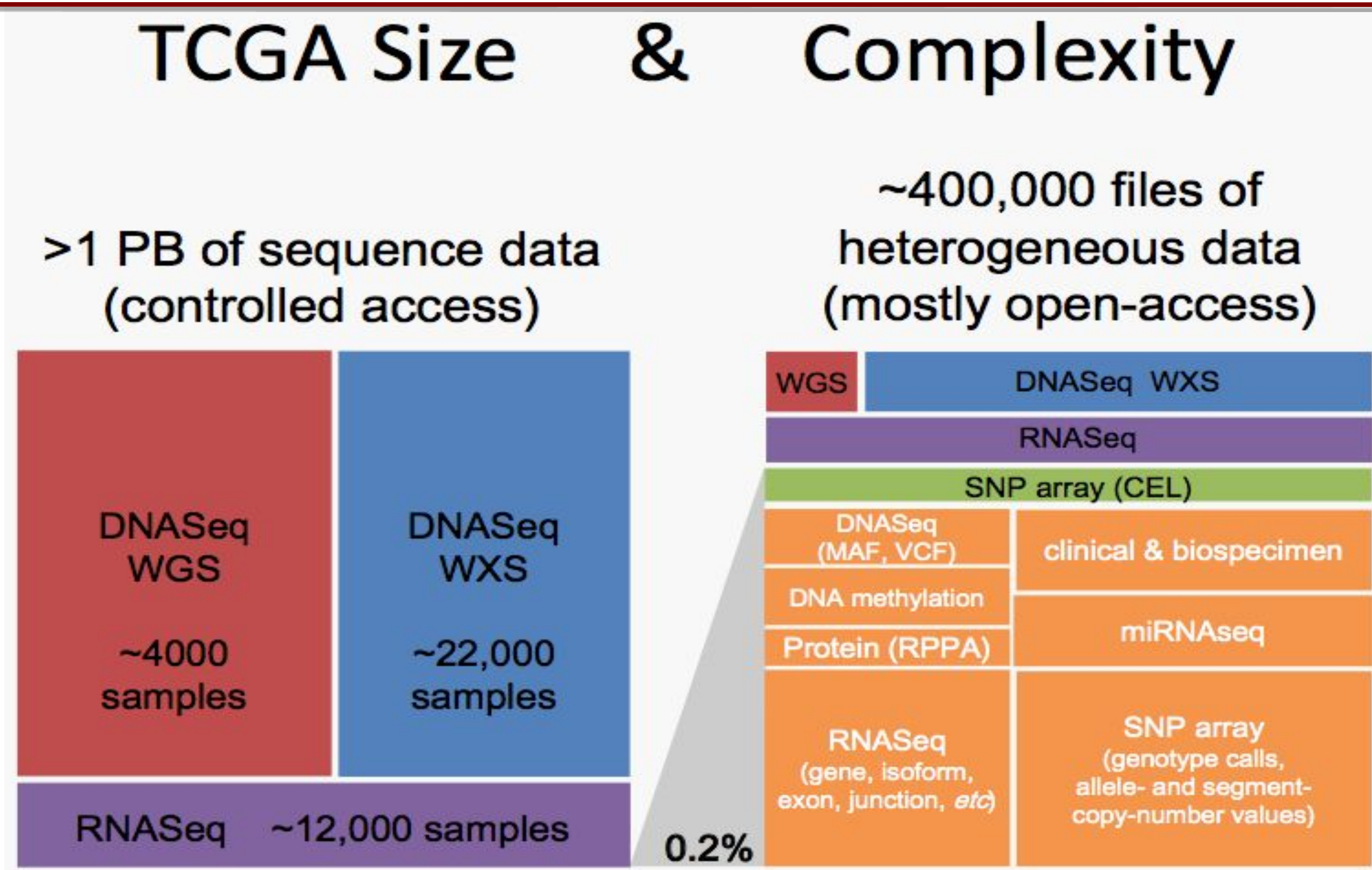
1. Orphan genes
2. *De Novo* genes
3. Pseudogenes
4. sORFs
5. altORFs

2. Prioritizing variants in these regions using machine learning approaches

Inside TCGA


TCGA IS A TREMENDOUS GIFT TO THE CANCER RESEARCH COMMUNITY …
1098 Breast Cancer patients
470 Melanoma patients
185 Pancreatic Cancer patients
More than 11,000 cases representing 33 cancer types

Data types in TCGA



National and International sequencing projects

Examples of >100K projects already in progress

Genomics england

华大基因 BGI

MVP

sequencebio

THE PRECISION MEDICINE INITIATIVE

ancestryDNA

HUMAN LONGEVITY, INC.

برنامج الجينوم البشري السعودي
Saudi Human Genome Program

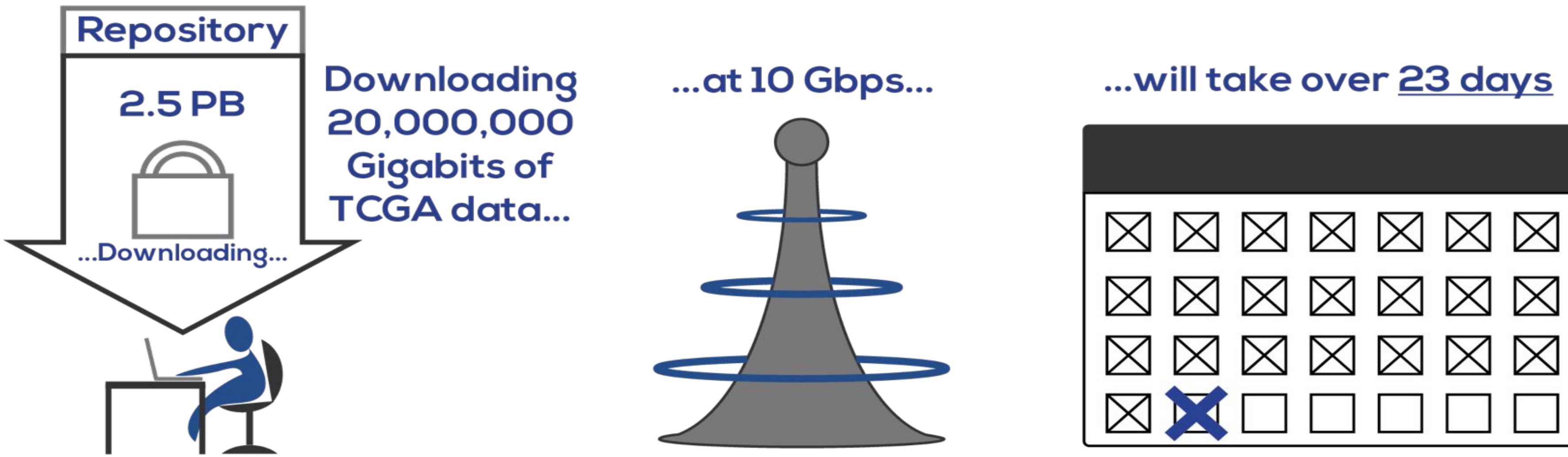
International Cancer Genome Consortium

23andMe

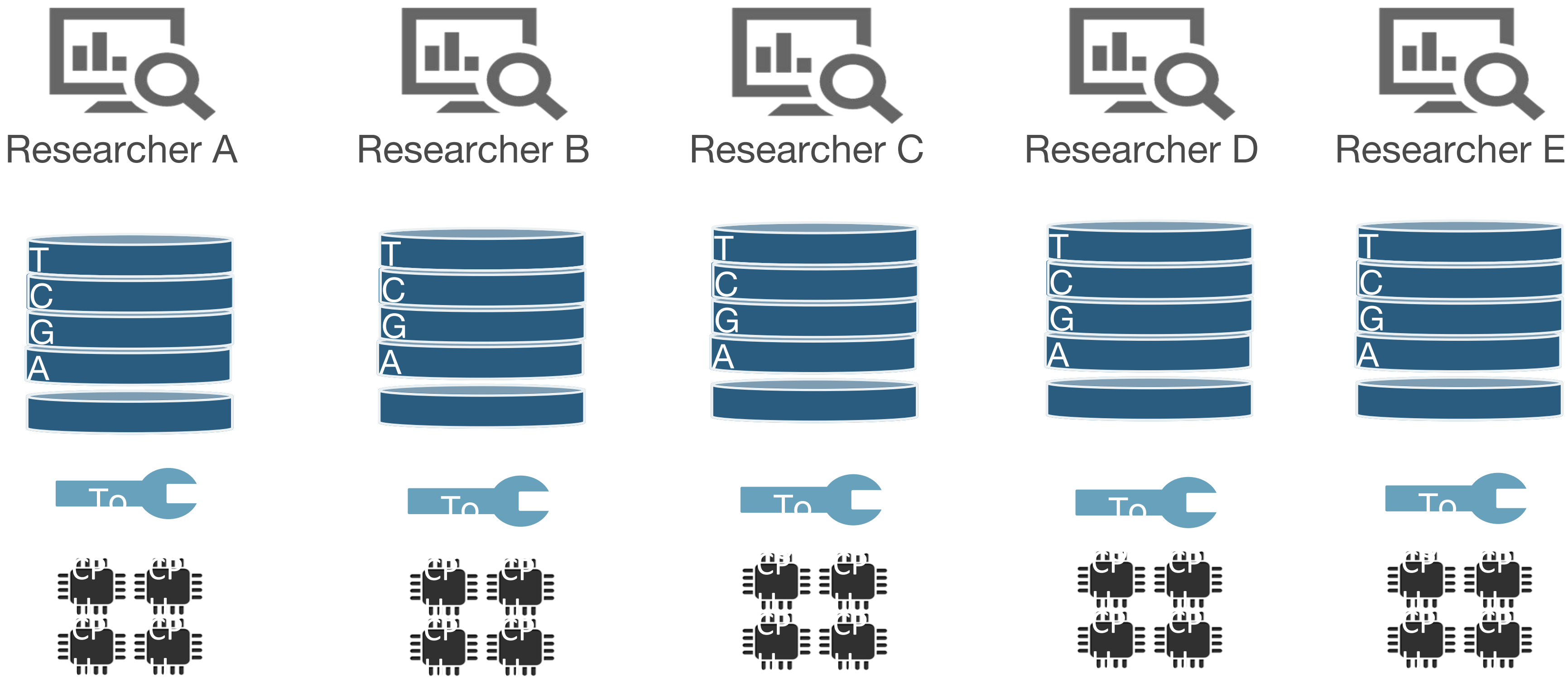
deCODE genetics

REGENERON

Accessing data at petabyte scale

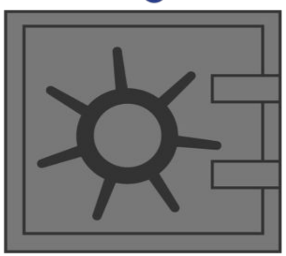


Previous analysis methods

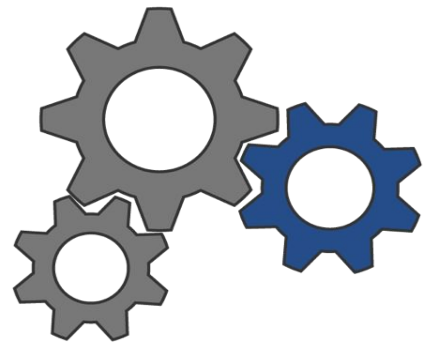


Cost of this model

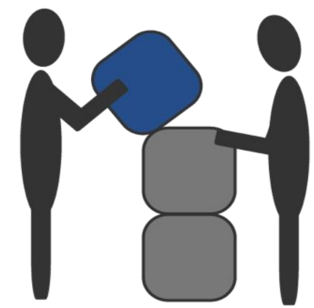
$2M/year
in storage costs



Data is locked away and replicated unnecessarily across many institutions



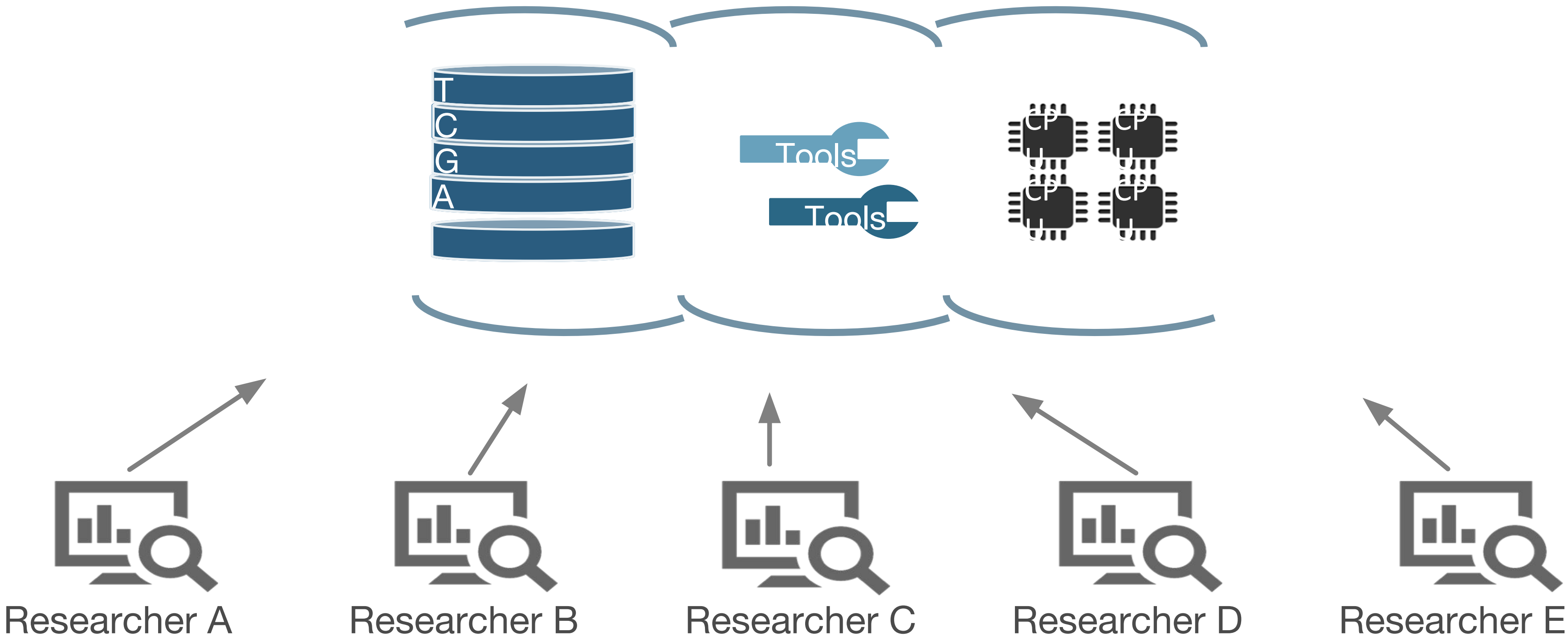
Requires significant computational resources



Collaborating in real-time and sharing reproducible results is challenging.

Biological questions come to data

Current data analysis model



Tools to the data

AS THE AMOUNT OF DATA HAS GROWN, SO TOO HAS THE NUMBER OF TOOLS AVAILABLE TO ANALYZE IT.

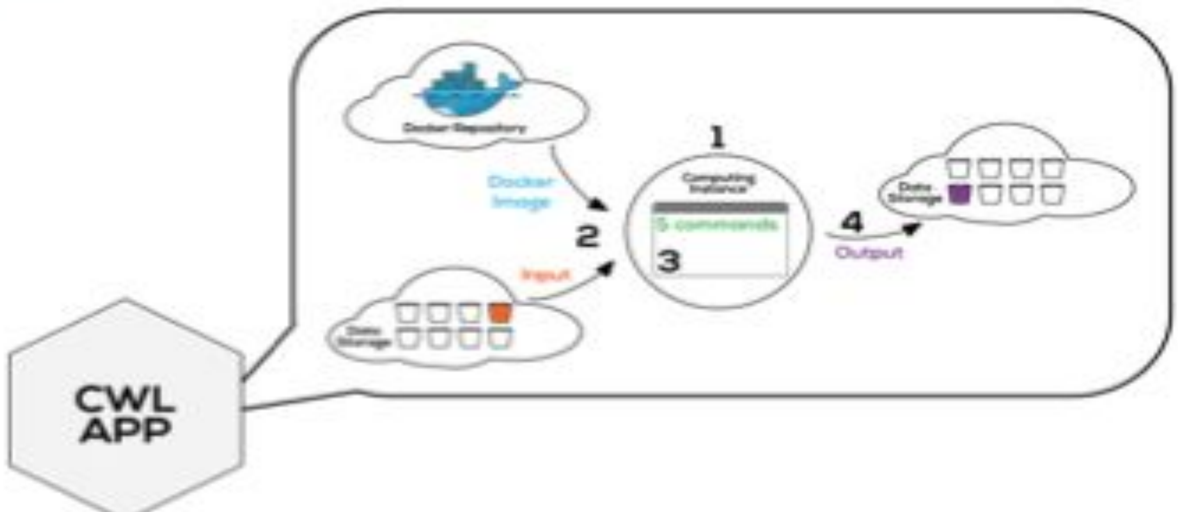
11,000+ -omics data analysis tools*

(each with many versions)

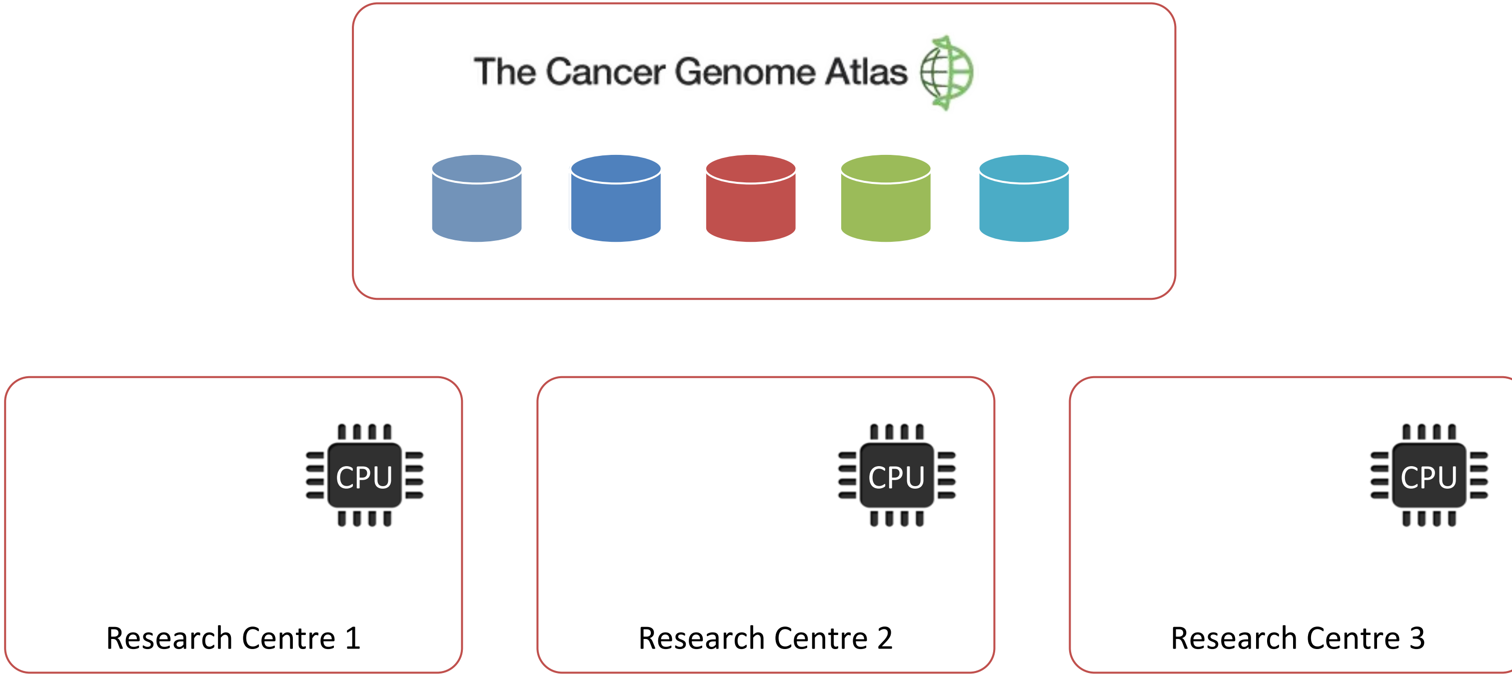
50+ used in a single TCGA marker paper

*omictools.com

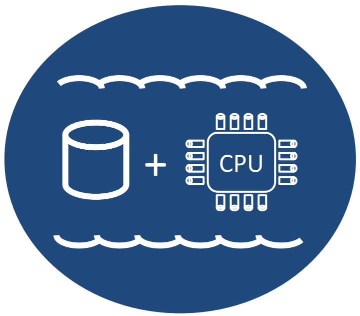
Cloud computing to the rescue



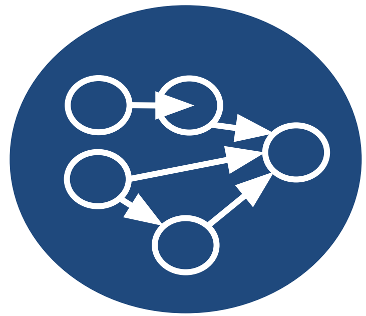
Today the data comes to computation



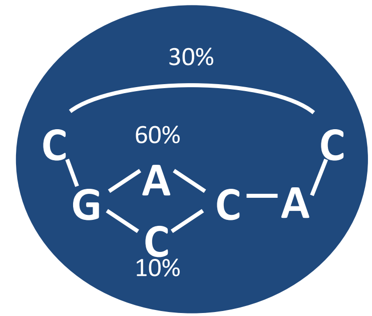
Millions of genomes will shift how computation will be done



COMPUTATION CENTERS
replace data repositories



PORTABLE WORKFLOWS
replace data transfers



ADVANCED DATA STRUCTURES
replace static flat files

image source: Seven Bridges website

Introduction to cloud computing

Lecture 2

What is cloud?

Cloud is not a Buzzword

Web 2.0, Internet of Things are a buzzword! in fact Big Data is a buzzword

it is not a computer in some else’s datacenter

NIST definition of cloud computing

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model is composed of five essential characteristics, three service models, and four deployment models.

Why Cloud Computing?

- Small as well as some large IT companies follows the traditional methods to provide the IT infrastructure. That means for any IT company, we need a **Server Room** that is the basic need of IT companies.

- In that server room, there should be a database server, mail server, firewalls, routers, modem, switches and the maintenance engineers.

- To establish such IT infrastructure, we need to spend lots of money. To overcome all these problems and to reduce the IT infrastructure cost, Cloud Computing came into existence.

What was used before?

1st Generation Business Computing: Mainframes

- Upfront costs: High (millions of dollars)
- Users: specialized IT operators
- I/O: Punched card and Printer

2nd Generation Business Computing: Midrange

- Upfront costs: Moderate (hundreds of thousands of dollars)
- Users: Specialized IT Operators and specialized end-users
- I/O: Keyboard and screen

3rd Generation Business Computing: Client Server

- Upfront costs: Low (thousands of dollars)
- Users: Specialized IT Operators and general end-user
- I/O: Keyboard, mouse and Graphical user interface

4th Generation Business Computing: The Cloud

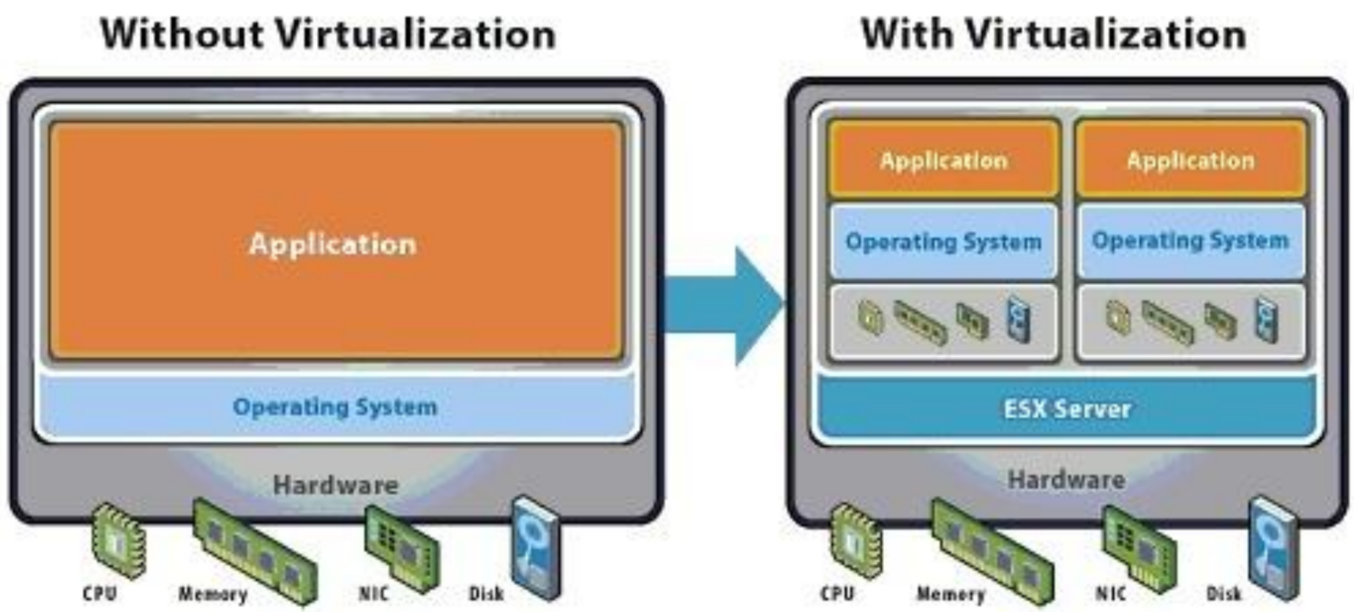
- Upfront costs: Very low or nothing
- Users: IT operators, general end-users, consumers
- I/O: Intelligent devices

Cloud computing service providers

What makes cloud computing possible?

- Virtualization of workloads
- Storage
- Networking

What is virtualisation?



Act of creating something virtual

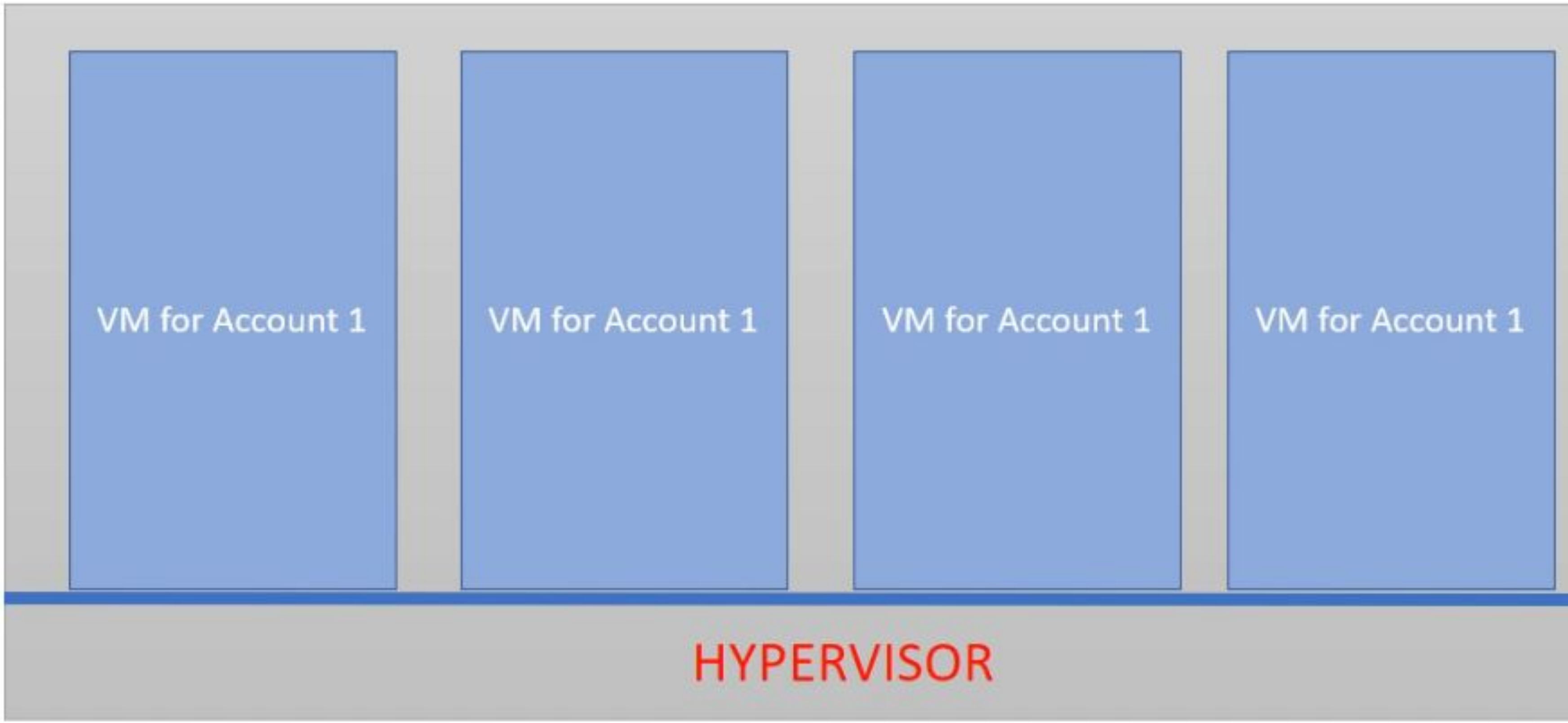
Actual physical resources are hidden from users

operating system directly talking to hardware

hypevisor (ESX server): an intermediate layer between operting system and hardware will present the hardware as a virtual resource

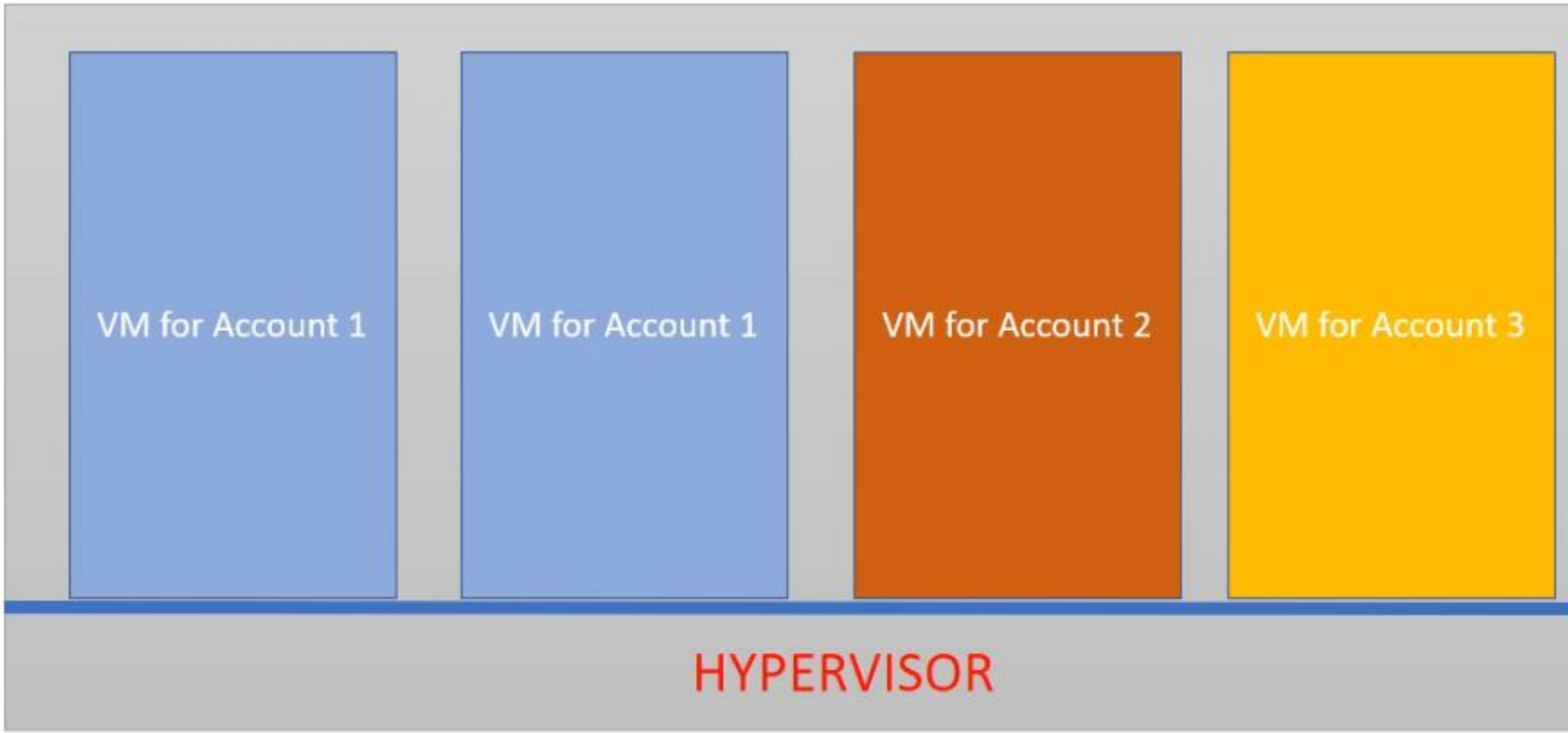
image source: VMware website

Single tenancy

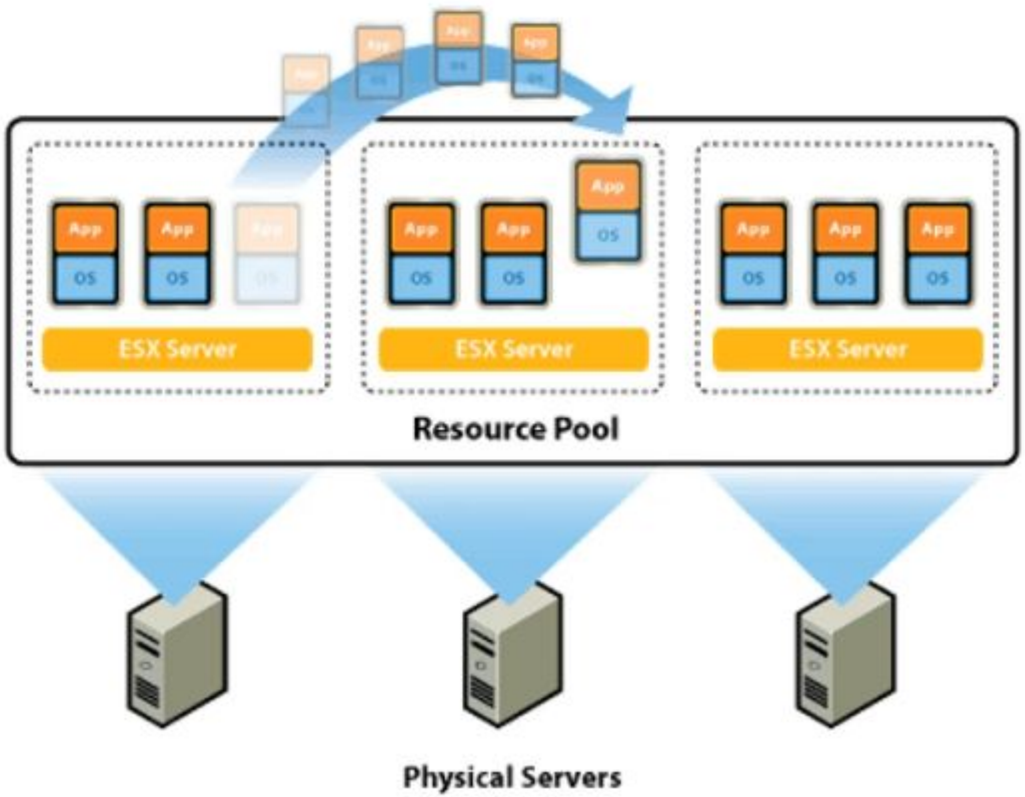


Each VM is an instance

Multi tenancy



Fault tolerant

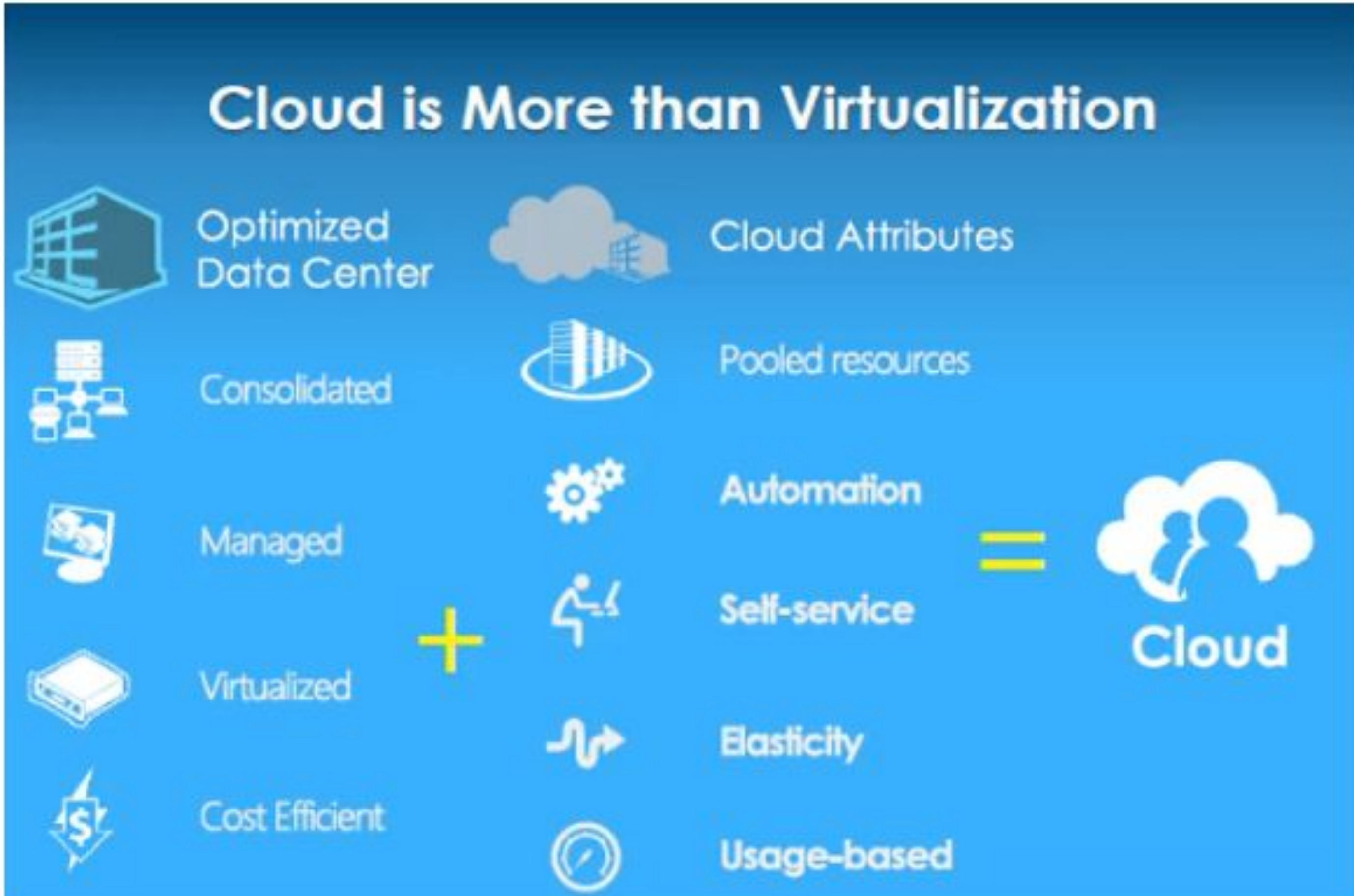


VM can be moved seamlessly if a certain physical machine fails

Advantages of VMs

- Optimal use of server resources (cost saving)
- Better fault tolerant
- Easy cloning
- Optimised load distribution

Data Centers



Docker and containerized workflow

Docker is a tool to create, deploy, and run applications by using containers.

Containers allow a developer to package up an application with all of the parts it needs, such as libraries and other dependencies, and ship it all out as one package. By doing so, thanks to the container, the developer can rest assured that the application will run on any other Linux machine regardless of any customized settings that machine might have that could differ from the machine used for writing and testing the code.

In a way, Docker is a bit like a virtual machine. But unlike a virtual machine, rather than creating a whole virtual operating system, Docker allows applications to use the same Linux kernel as the system that they're running on and only requires applications be shipped with things not already running on the host computer. This gives a significant performance boost and reduces the size of the application.

Common Workflow Language

- Common Workflow Language: An standard for describing a pipeline object model: Data, tools, pipelines, parameters

  https://github.com/common-workflow-language/common-workflow-language

- RABIX: Seven Bridges' open-source implementation of the CWL + custom extensions

  http://rabix.io/

- Nextflow enables scalable and reproducible scientific workflows using software containers. It allows the adaptation of pipelines written in the most common scripting languages. (non containerised)

  https://www.nextflow.io/

Key advantages of cloud computing

Cost planning
companies simply pay for what they use. As long as the company has the network capacity, they don’t have to buy any hardware or other infrastructure to support the solution.

Enterprise level systems
enterprise level systems, which contain all the platform features and scalability needed by large organizations to run their businesses.

Faster go live
Cloud computing eliminates this worry because it is very easy to just select a virtual machine or component from your cloud provider and be up and running in seconds or minutes.

Agile and elastic capacity
It means the solution you deploy in the cloud does not have to stay the way you deployed it. So, elasticity means you can change your capacity requirements, and agility means you can do it as often as you like.

Key advantages of cloud computing

Out of the box integration

High Availability and disaster recovery
high availability (HA) & Disaster recovery (DR) are important features of the cloud

Enhanced security
A cloud platform can offer a range of security features to detect or prevent attacks from the outside, which may be very expensive for smaller companies to implement. This again makes the IT systems in the cloud much more secure than they used to be in-house

Compliance
Many companies must comply with various laws and regulations based on the country/region or industry they work in. These include regulations such as Sarbanes-Oxley (SOX), HIPPA or GDPR. It can be very expensive to put the necessary infrastructure, processes, and management tools in place to ensure the company follow the laws of their government. When companies grow and they want to operate in other countries/regions, they also must adhere to the laws of those countries/regions. This will again mean more costs and management tools for the company. Cloud companies themselves must also follow these regulations and can therefore offer the same capabilities, management tools, and compliance certificates to the companies that use their cloud.

Service models

entire stack is
managed
by someone else

things we install & use

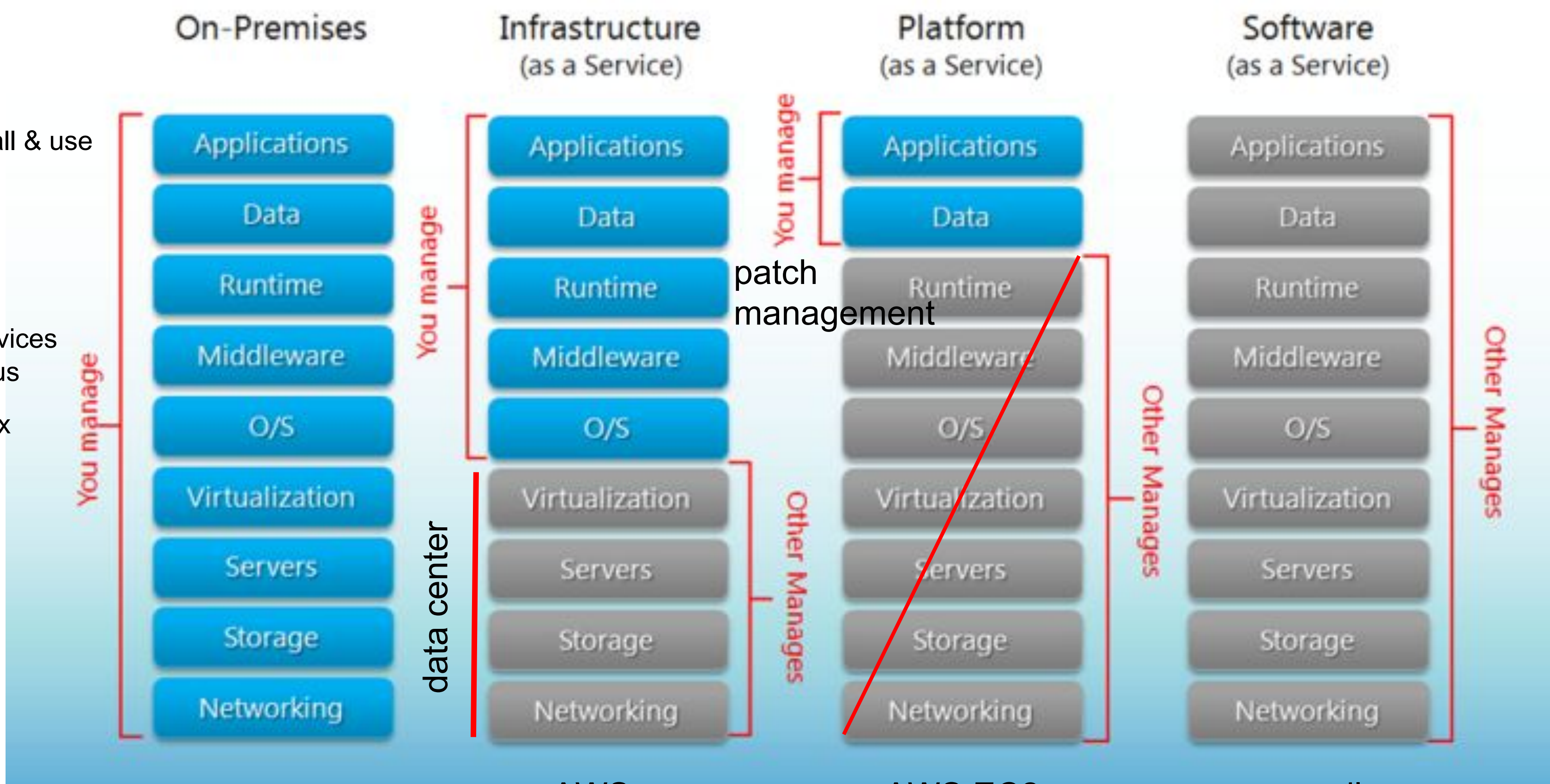
databases

java and .net

messaging services
like service bus

windows & linux

VMware



AWS
Azure

AWS EC2

gmail

image source: Robert Greiner

Essential characteristics of cloud computing

On-demand self-service: A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service provider.

Broad network access: Capabilities are available over the network & accessed through standard mechanisms that promote the use by heterogeneous thin or thick client platforms eg. mobile phones, laptops and workstations

Resource pooling: The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. There is a sense of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or datacenter). Examples of resources include storage, processing, memory, and network bandwidth.

Essential characteristics of cloud computing

Rapid elasticity:

Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be appropriated in any quantity at any time.

save money

Measured service:

Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.

source: NIST

Deployment Models

Private cloud: The cloud infrastructure is provisioned for exclusive use by a single organization comprising multiple consumers (e.g., business units). It may be owned, managed, and operated by the organization, a third party, or some combination of them, and it may exist on or off premises.

Community cloud: The cloud infrastructure is provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns (e.g., mission, security requirements, policy, and compliance considerations). It may be owned, managed, and operated by one or more of the organizations in the community, a third party, or some combination of them, and it may exist on or off premises.

Public cloud: The cloud infrastructure is provisioned for open use by the general public. It may be owned, managed, and operated by a business, academic, or government organization, or some combination of them. It exists on the premises of the cloud provider.

Hybrid cloud: The cloud infrastructure is a composition of two or more distinct cloud infrastructures (private, community, or public) that remain unique entities, but are bound together by standardized or proprietary technology that enables data and application portability (e.g., cloud bursting for load balancing between clouds).

Pros and Cons of Cloud Computing

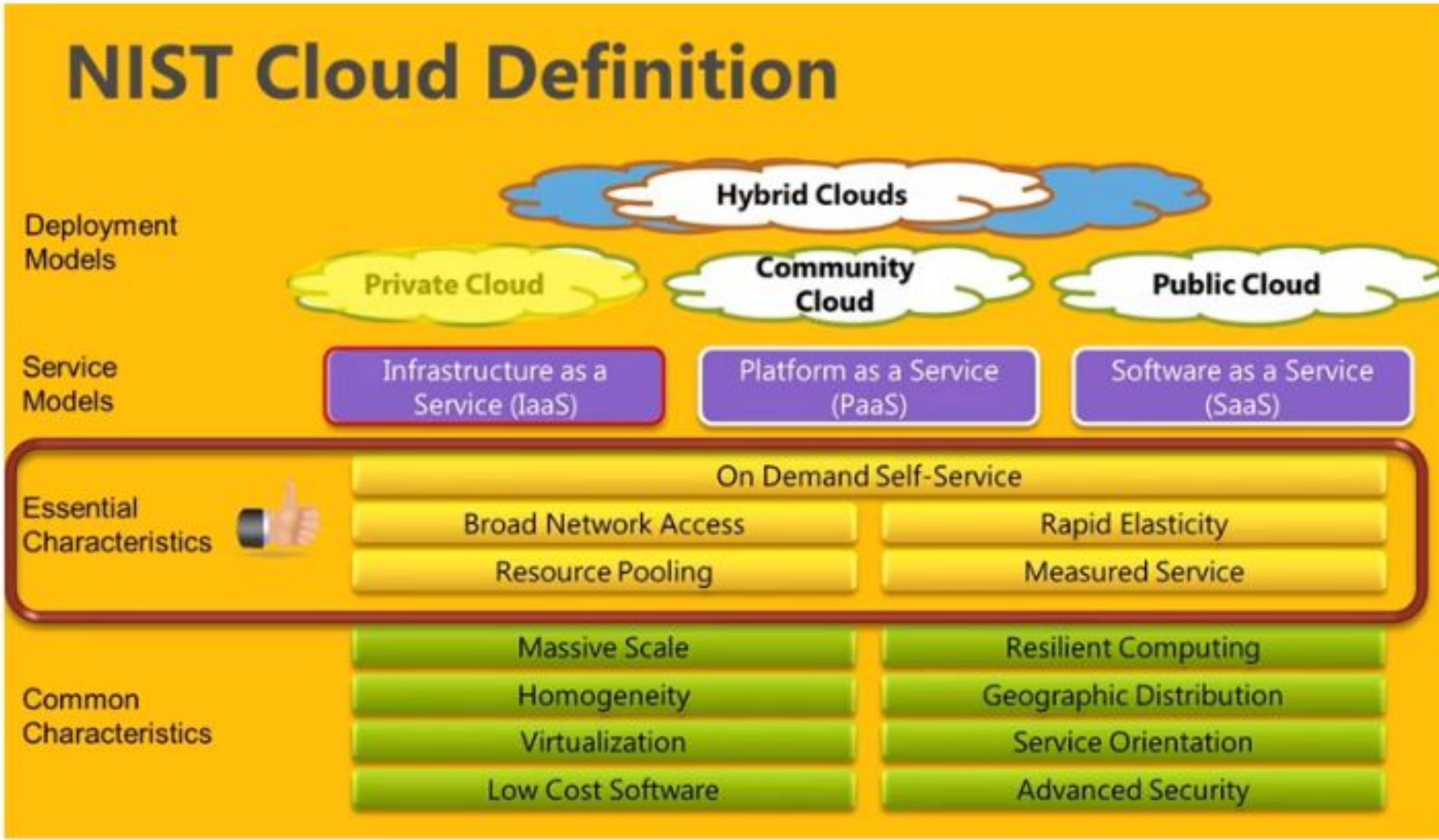
Pros of Cloud Computing

- Lower cost computer for users
- Lower IT infrastructure cost
- Fewer maintenance cost
- Increased computing Power
- Unlimited storage capacity

Cons of Cloud Computing

- Require a constant Internet Connection
- Require High Speed Internet connection

Summary


NIST Cloud Definition
Deployment Models
Hybrid Clouds
Private Cloud
Community Cloud
Public Cloud
Service Models
Infrastructure as a Service (IaaS)
Platform as a Service (PaaS)
Software as a Service (SaaS)
Essential Characteristics
On Demand Self-Service
Broad Network Access
Rapid Elasticity
Resource Pooling
Measured Service
Common Characteristics
Massive Scale
Resilient Computing
Homogeneity
Geographic Distribution
Virtualization
Service Orientation
Low Cost Software
Advanced Security