A Brief History of BigData Era

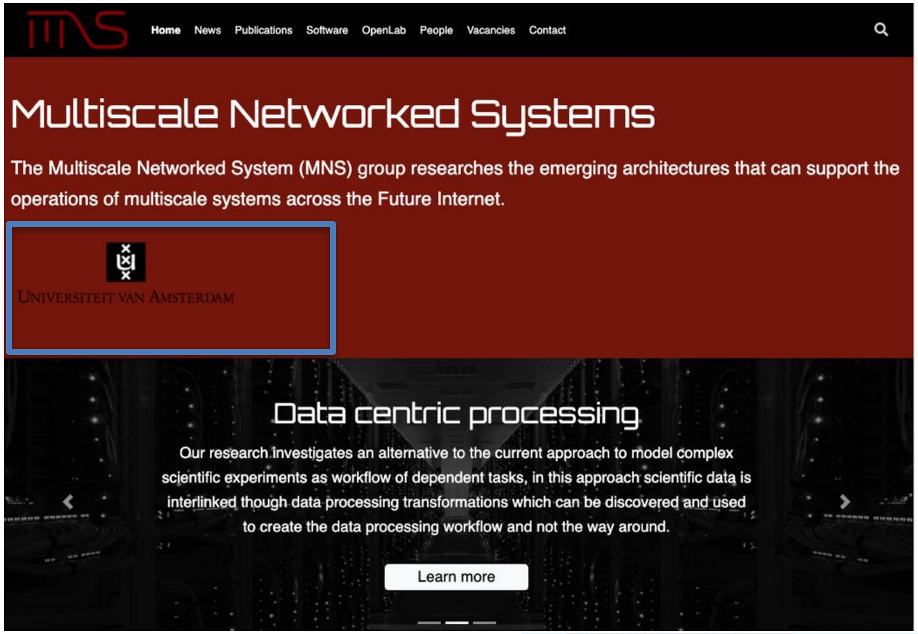
A journey from laptops to supercomputers and beyond

Adam Belloum

Those who own data own the future"
Yuval Noah Harari

From Constantine to Amsterdam via Compiegne

eland
North Sea
United Kingdom
Ireland
London
5 h 35 min
555 km
Compiègne
Vienna
Slovakia
Austria
Hungary
France
Roma
Croatia
Serbia
Italy
Rome
Bulga
Barcelona
Portug
Portugal
Madrid
Spain
Greece
Mc
iterranean Sea
Morocco
Algeria
Libya
ern
ra
ania
lauritania
Mali
Niger
Chad



Multiscale Networked Systems

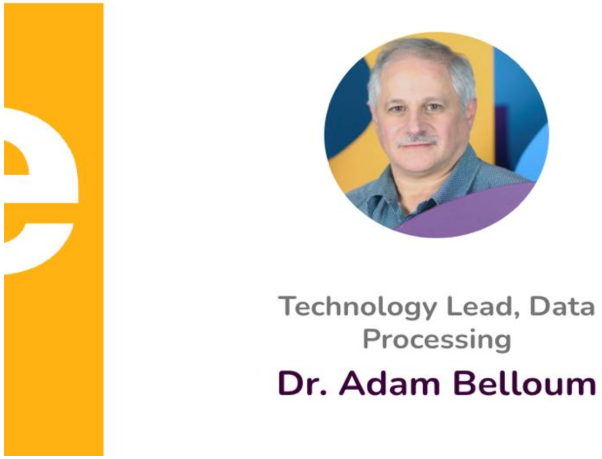
The Multiscale Networked System (MNS) group researches the emerging architectures that can support the operations of multiscale systems across the Future Internet.

Universiteit van Amsterdam

Data centric processing

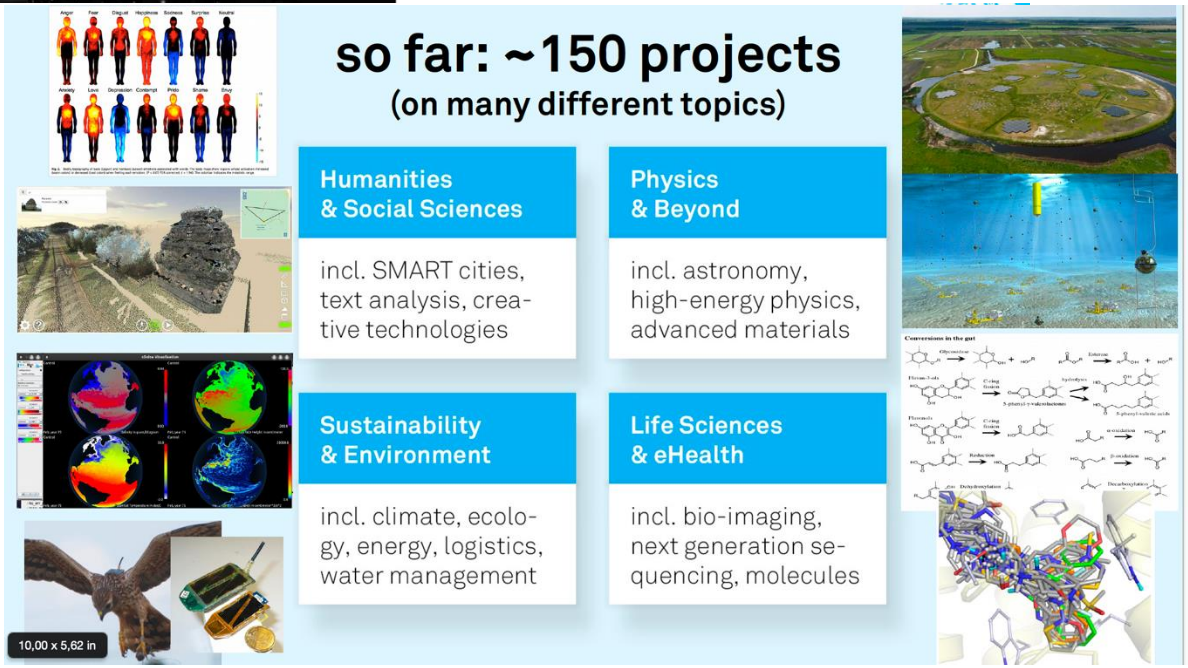
Our research investigates an alternative to the current approach to model complex scientific experiments as workflow of dependent tasks, in this approach scientific data is interlinked though data processing transformations which can be discovered and used to create the data processing workflow and not the way around.

netherlands eScience center



Technology Lead, Data Processing

Dr. Adam Belloum



so far: ~150 projects

(on many different topics)

Humanities & Social Sciences

incl. SMART cities, text analysis, creative technologies

Physics & Beyond

incl. astronomy, high-energy physics, advanced materials

Sustainability & Environment

incl. climate, ecology, energy, logistics, water management

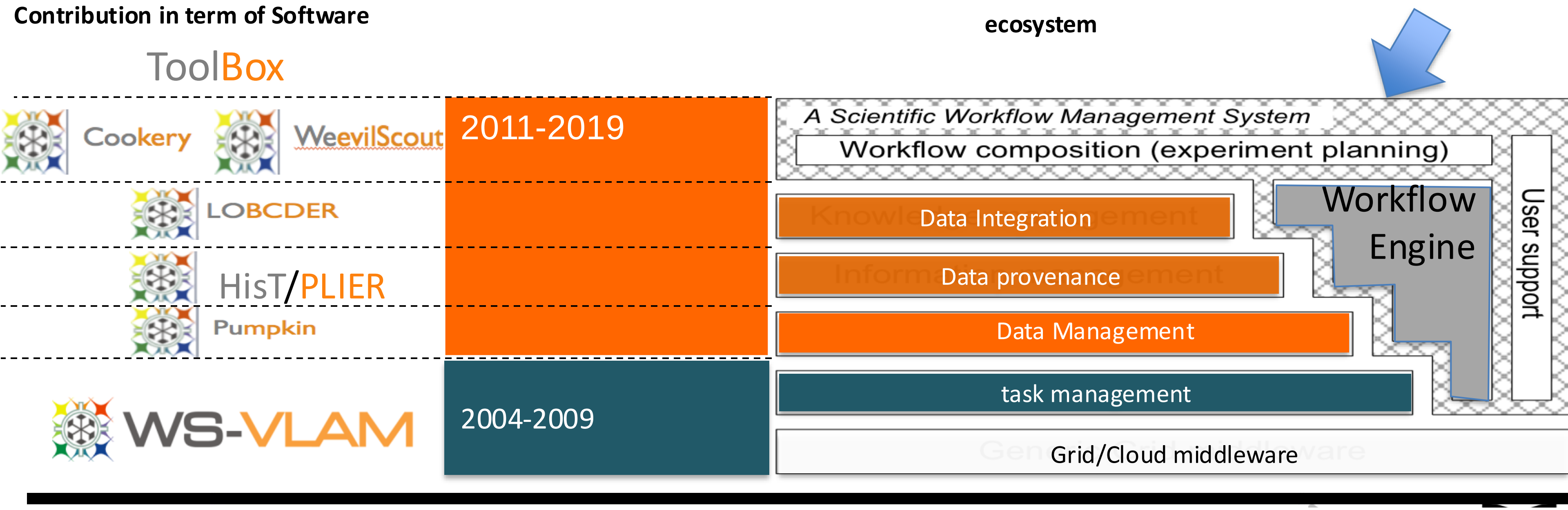
Life Sciences & eHealth

incl. bio-imaging, next generation sequencing, molecules

The research work

- Acquire understanding of the **system as a whole** by " the analyses of individual phenomena and **the** **integration of** different, **interdisciplinary sources** of knowledge about a **complex system**"

Foster, I., Kesselman, C., *Scaling system-level science: Scientific exploration and its implications*. *IEEE Computer 39 (11) 2006*



Contribution in term of Software

ToolBox

Cookery

WeevilScout

LOBCDER

HisT/PLIER

Pumpkin

WS-VLAM

2011-2019

2004-2009

ecosystem

A Scientific Workflow Management System

Workflow composition (experiment planning)

Data Integration

Data provenance

Data Management

task management

Grid/Cloud middleware

Workflow Engine

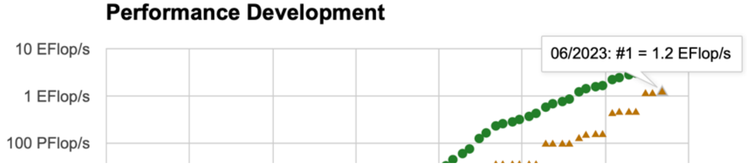
User support

Sponsor

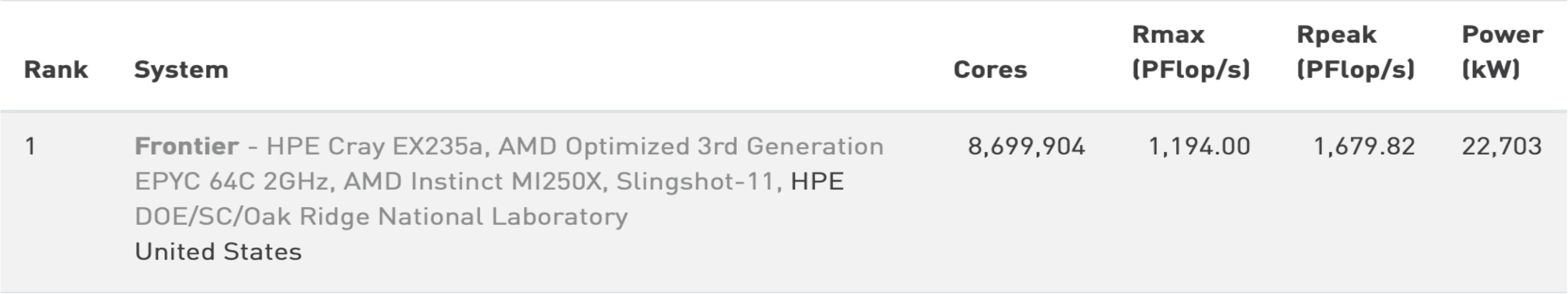
vl-e	2004-2009
Dutch research program COMMIT/	2011-2017
EU-FP7 Project VPH-SHARE	2011-2015
EU H2020 project EDISON building the data science profession	2015-2017
PROCESS	2017-2020

Questions to be “answered” in this talk …

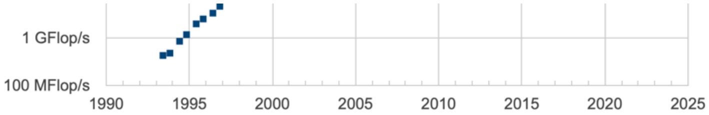
- Why do we need more and more computing power?
  - Does m
  - Do you to run p



Performance Development



Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,699,904	1,194.00	1,679.82	22,703

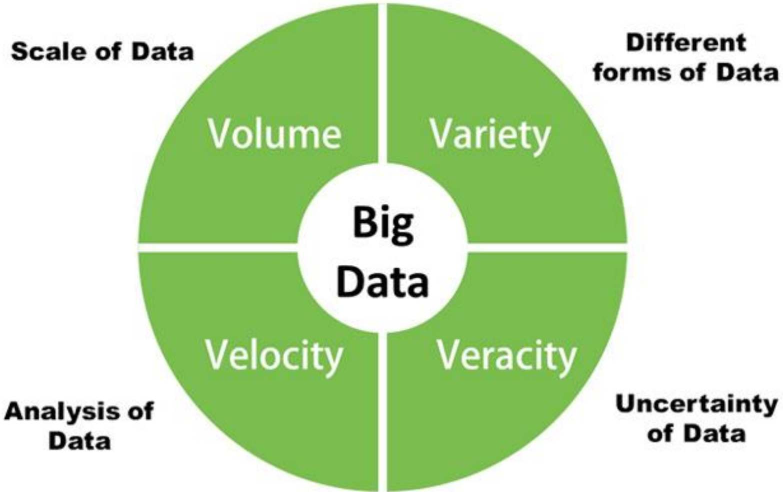


Source https://www.top500.org/statistics/perfdevel/

- What is Big data?
  - A Terabyte of **Storage Space**: How much is it?
  - How much does it take to **sort** one Terabyte?
  - How much does it take to **move** Terabyte/Exabyte over the internet?

- How can we build system beyond Supercomputer “Cloud system”?

- What is the connection between AI and Bigdata?
- …



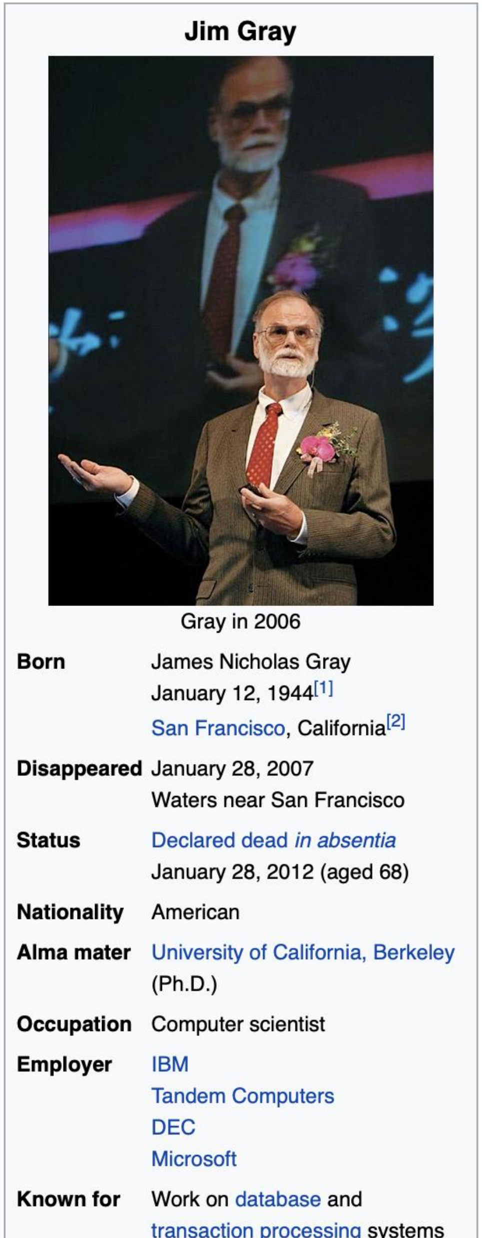
- **Storage**
- **Processing**
- **Movement**

Big data Era

- “ We have to **do better at producing tools** to support the **whole research cycle**—from **data capture and data curation to data analysis and data visualization**. **Today, the tools** for capturing data both at the mega-scale and at the milli-scale are just **dreadful**. After you have captured the data, you need to curate it before you can start doing any kind of data analysis, and **we lack good tools for both data curation and data analysis**.”
- “Then comes the **publication** of the results of your research, and the published literature is just the tip of the data iceberg. By this I mean that people collect a lot of data and then reduce this down to some number of column inches in Science or Nature—or 10 pages if it is a computer science person writing. **So what I mean by data iceberg is that there is a lot of data that is collected but not curated or published in any systematic way**. “

Based on the transcript of a talk given by Jim Gray

to the NRC-CSTB1 in Mountain View, CA, on January 11, **2007**



Jim Gray

Gray in 2006

Born	James Nicholas Gray January 12, 1944[1] San Francisco, California[2]
Disappeared	January 28, 2007 Waters near San Francisco
Status	Declared dead *in absentia* January 28, 2012 (aged 68)
Nationality	American
Alma mater	University of California, Berkeley (Ph.D.)
Occupation	Computer scientist
Employer	IBM Tandem Computers DEC Microsoft
Known for	Work on database and transaction processing systems

- **Storage**
- **Processing**
- **Movement**

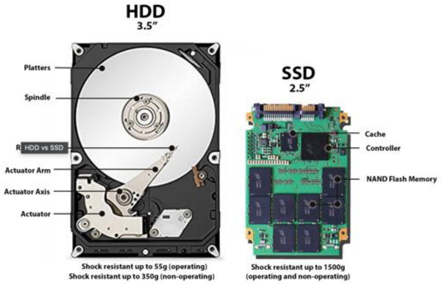
Big Data

Note: Kilo is exactly 1024 ~ 1000

- YottaByte (YB) = 10^{24} Byte
- ZetaByte (ZB) = 10^{21} Byte
- ExaByte (EB) = 10^{18} Byte
- PetaByte (PB) = 10^{15} Byte
- **TeraByte (TB) = 10^{12} Byte**
- **GigaByte (GB) = 10^{9} Byte**
- **MegaByte (MB) = 10^{6} Byte**
- **KiloByte (KB) = 10^{3} Byte**
- **Byte = 8 bits**

capacity

R/W speed


HDD
3.5″
Platters
Spindle
R HDD vs SSD
Actuator Arm
Actuator Axis
Actuator
Shock resistant up to 55g (operating)
Shock resistant up to 350g (non-operating)
SSD
2.5″
Cache
Controller
NAND Flash Memory
Shock resistant up to 1500g
(operating and non-operating)

- **1+ ZB - Internet size in bytes**
- **Radio astronomy- SKA-Phase 3+ EFlops**

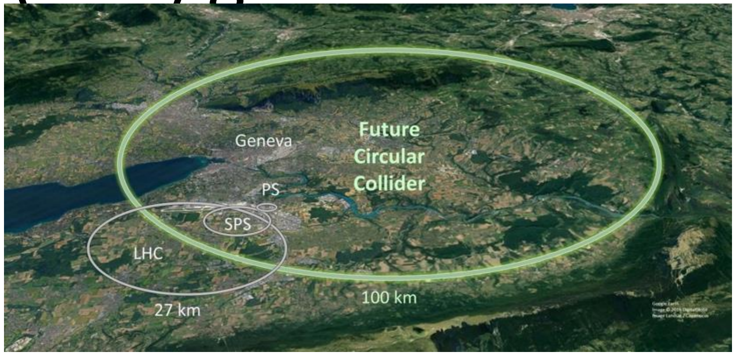
- **1 TB HDD/~60$ - Storage technology**
- **18 TB HDD/~600$ - Storage technology**

- **Storage**
- **Processing**
- **Movement**

Big Data

In Industry and science around 2009

- Google processes **20 PB a day**
- Wayback Machine has 3 PB **+ 100 TB/month**
- Facebook has 2.5 PB of user data **+ 15 TB/day**)
- eBay has 6.5 PB of user data **+ 50 TB/day**
- CERN's Large Hydron Collider (LHC) generates **15 PB/year**



Note: 1 TB = 1,000 (10^3) gigabytes (GB) or 1,000,000 (10^6) megabytes (MB)

Souce: https://aimblog.uoregon.edu/2014/07/08/a-terabyte-of-storage-space-how-much-is-too-much/

A Terabyte of Storage Space: How Much …?

- **Storage**
- **Processing**
- **Movement**

personal usage

- ➢ ~200,000 average songs, High-Quality Compressed Audio (~17,000 hours of music)

- ➢ ~256 Standard DVD Movies 120 minutes long (~500 hours of movies)

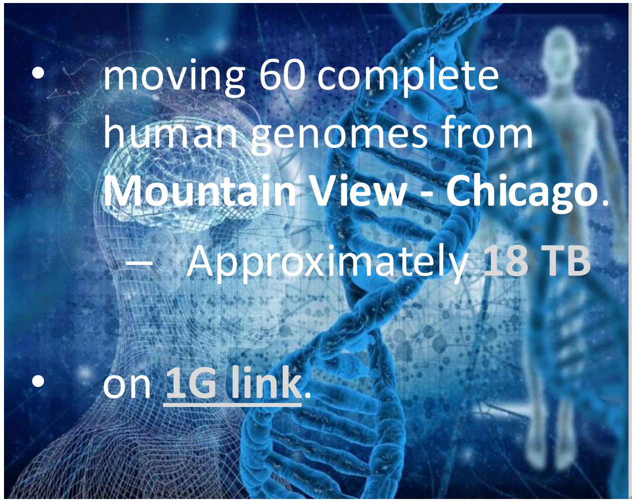
- ➢ ~310,000 Standard-Resolution Photos

Note: 1 TB = 1,000 (10^3) gigabytes (GB) or 1,000,000 (10^6) megabytes (MB)

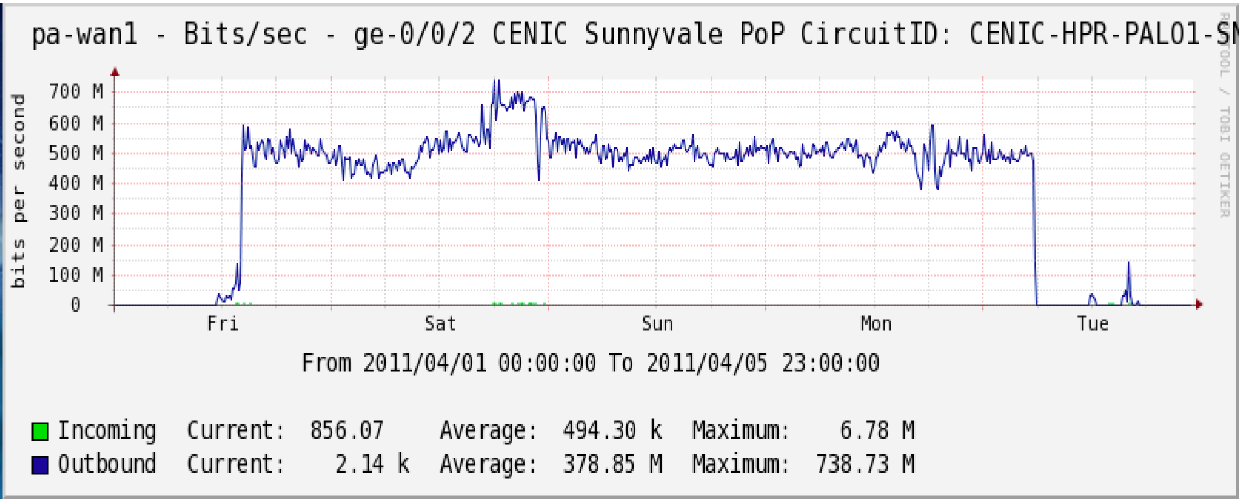
Souce: https://aimblog.uoregon.edu/2014/07/08/a-terabyte-of-storage-space-how-much-is-too-much/

How much take to move 18 TB over the internet ?

- **Storage**
- **Processing**
- **Movement**



- moving 60 complete human genomes from **Mountain View - Chicago**.
  - Approximately 18 TB
- on 1G link.



From 2011/04/01 00:00:00 To 2011/04/05 23:00:00

```
🟩 Incoming   Current:  856.07     Average:  494.30 k   Maximum:    6.78 M
🟦 Outbound   Current:    2.14 k   Average:  378.85 M   Maximum:  738.73 M
```

Credit: Robert Grossman University of Chicago Open Data Group, November 14, 2011

https://delaat.net/sc/sc17/demo02/index.html

- **Storing**
- **Processing**
- **Movement**

FedEx Express Has More Bandwidth Than the Internet—and When That'll Change

- If you're looking to transfer **hundreds of gigabytes** of data, it's still—weirdly—faster to ship hard drives via FedEx than it is to transfer the files over the internet.

CISCO estimates that total internet traffic averages **167 terabits per second**.

FedEx Express has a fleet of 654 aircraft with a lift capacity of 26.5 million pounds daily.

- A solid-state laptop drive weighs about 78 grams and can hold up to a terabyte.
- FedEx is capable of transferring **150 exabytes of data per day**, or **14 petabits per second**—**almost a hundred times the throughput of the internet in 2013**.

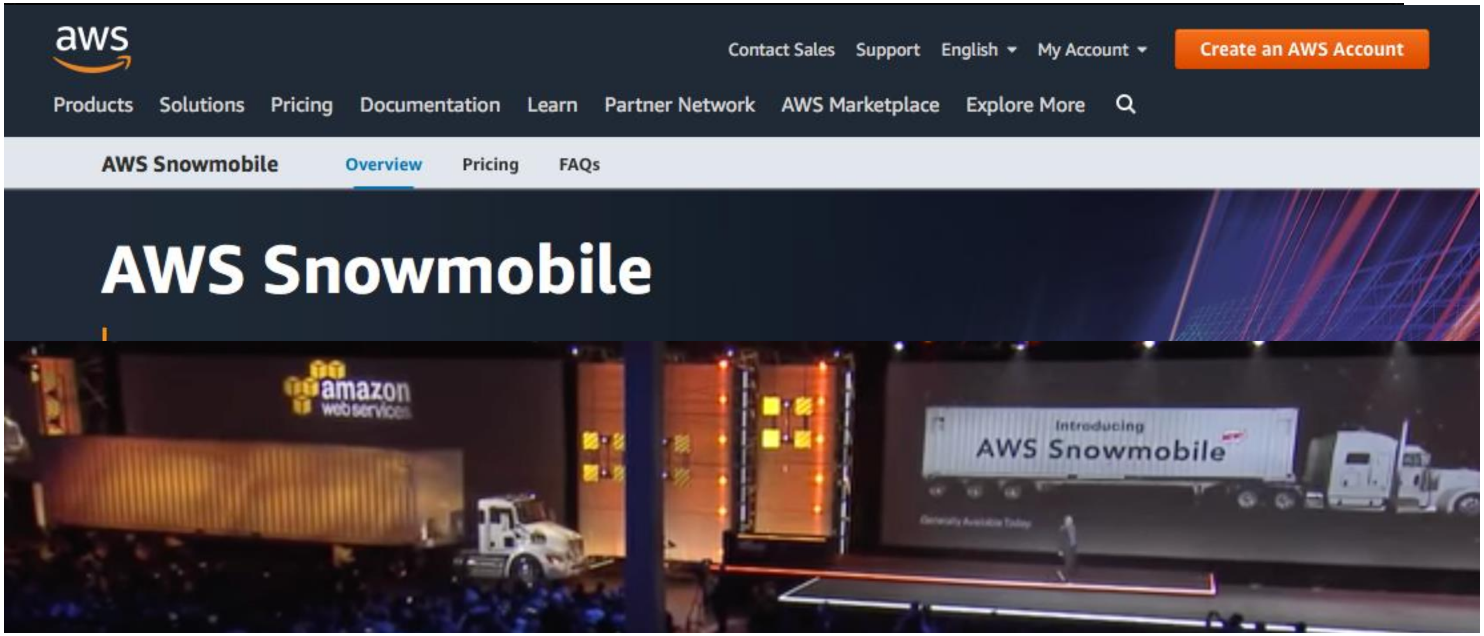
By[Jamie Condliffe](#) PublishedFebruary 5, **2013**

Source: http://gizmodo.com/5981713/how-fedex-has-more-bandwidth-than-the-internetand-when-thatll-change

- Storage
- Processing
- Movement

How can we move one exabyte over the internet?

Over **10Gbs** line it will take ~ **26 years**

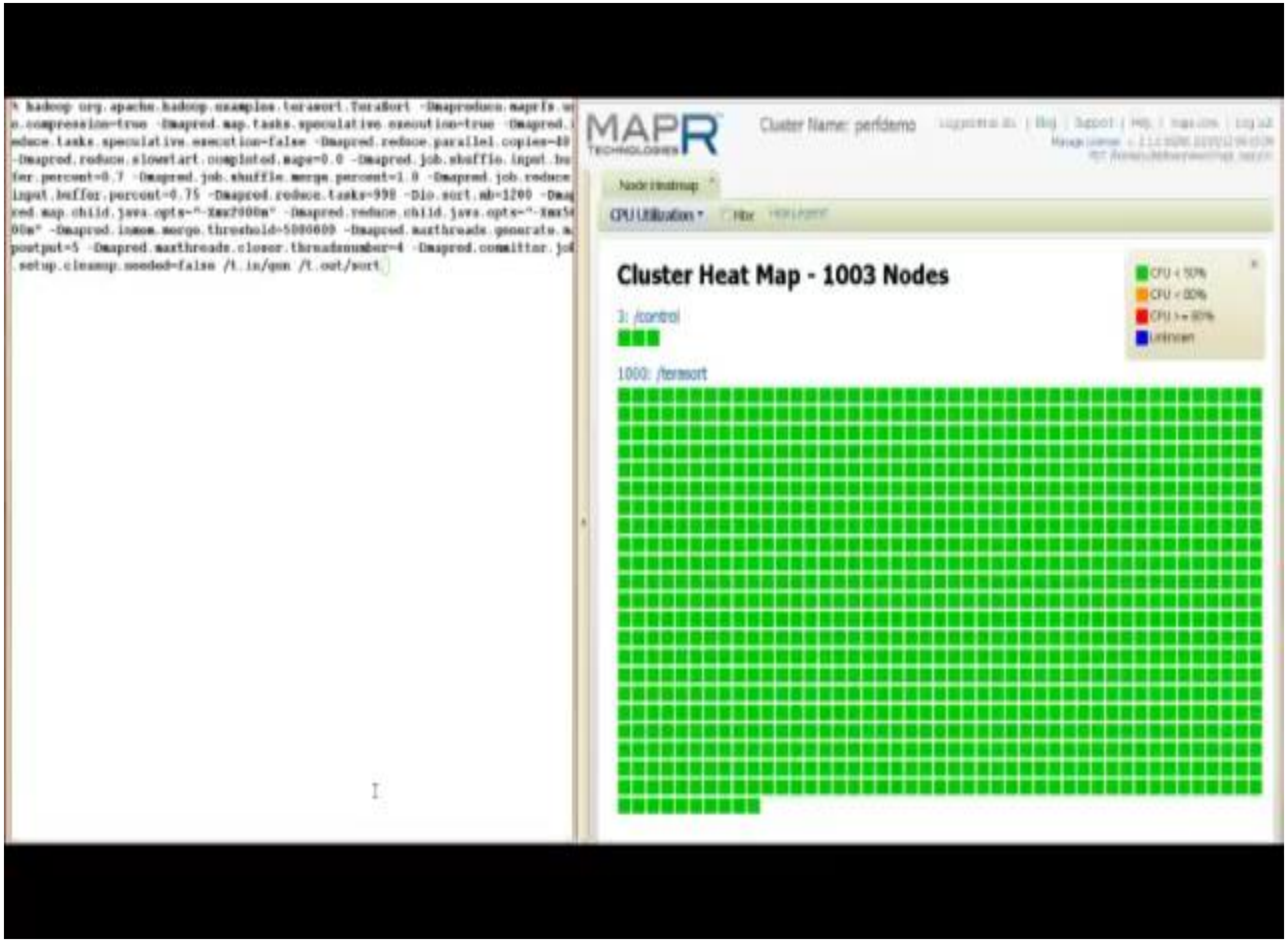


Note: 1 exa-Byte = **1,000 (10^3) petabytes (PB)**
or **1,000,000 (10^6) terabytes (TB)**
or **1,000,000, 000 (10^9) gigabytes (GB)**
or **1,000,000, 000, 000 (10^{12}) megabytes (MB) …**

Source: https://aws.amazon.com/snowmobile/

- Storage
- Processing
- Movement

Sorting 1 TB of DATA



```
% hadoop org.apache.hadoop.examples.terasort.TeraSort -Dmapreduce.maprfs.u
e.compression=true -Dmapred.map.tasks.speculative.execution=true -Dmapred.
educe.tasks.speculative.execution=false -Dmapred.reduce.parallel.copies=10
-Dmapred.reduce.slowstart.completed.maps=0.0 -Dmapred.job.shuffle.input.bu
fer.percent=0.7 -Dmapred.job.shuffle.merge.percent=1.0 -Dmapred.job.reduce
input.buffer.percent=0.75 -Dmapred.reduce.tasks=998 -Dio.sort.mb=1200 -Dma
red.map.child.java.opts="-Xmx2000m" -Dmapred.reduce.child.java.opts="-Xmx5
00m" -Dmapred.inmem.merge.threshold=5000000 -Dmapred.maxthreads.generate.m
poutput=5 -Dmapred.maxthreads.closer.threadnumber=4 -Dmapred.committer.jo
.setup.cleanup.needed=false /t.in/gen /t.out/sort
```

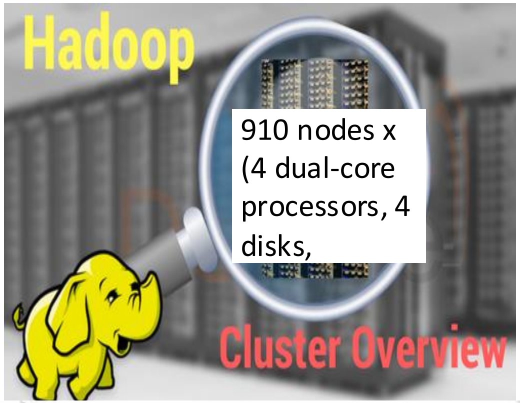
Cluster Name: perfdemo

Cluster Heat Map - 1003 Nodes

3: /control

1000: /terasort

🟩 CPU < 50%
🟧 CPU < 80%
🟥 CPU >= 80%
🟦 Unknown



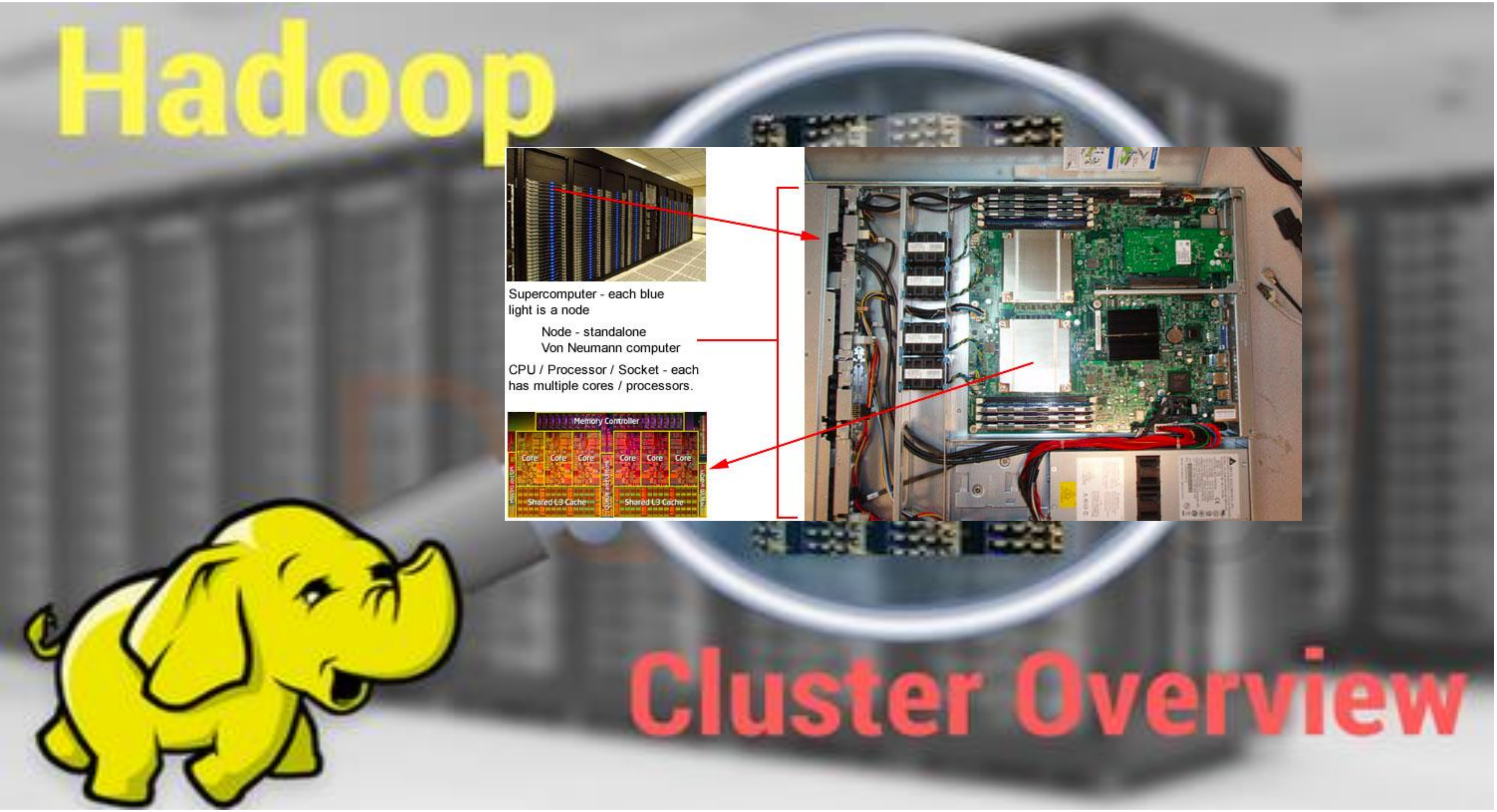
910 nodes x (4 dual-core processors, 4 disks,

http://sortbenchmark.org/

(*)https://googleblog.blogspot.com/2008/11/sorting-1pb-with-mapreduce.htr

Does more CPUs imply faster execution times?

- Storage
- **Processing**
- Movement



Supercomputer - each blue light is a node

Node - standalone Von Neumann computer

CPU / Processor / Socket - each has multiple cores / processors.

- How CPU works http://www.youtube.com/watch?v=cNN_tTXABUA
- Richard Feynman Computer Heuristics Lecture http://www.youtube.com/watch?v=EKWGGDXe5MA

Using more CPUs imply faster execution times!

- Storage
- **Processing**
- Movement

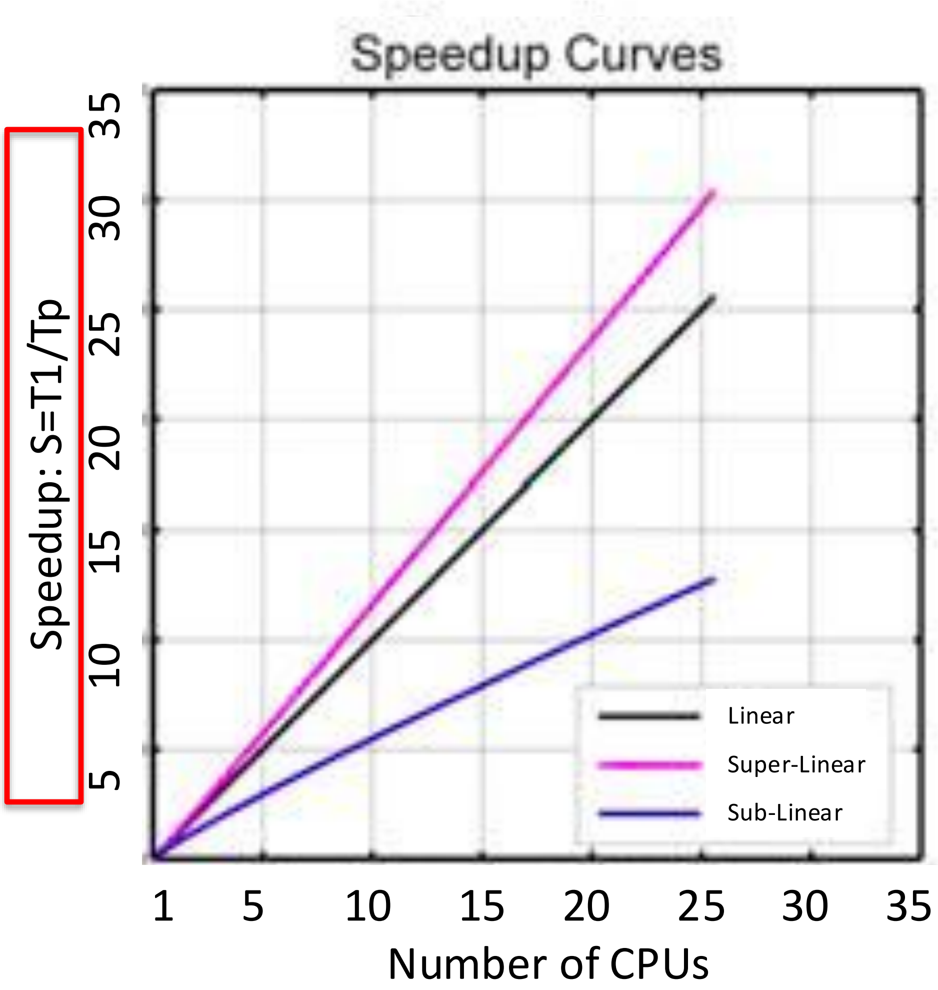
- Speedup
  - Superlinear
  - Linear
  - Sublinear
  - Other?

Best

Worst

You have to learn Parallel programming [(*)]

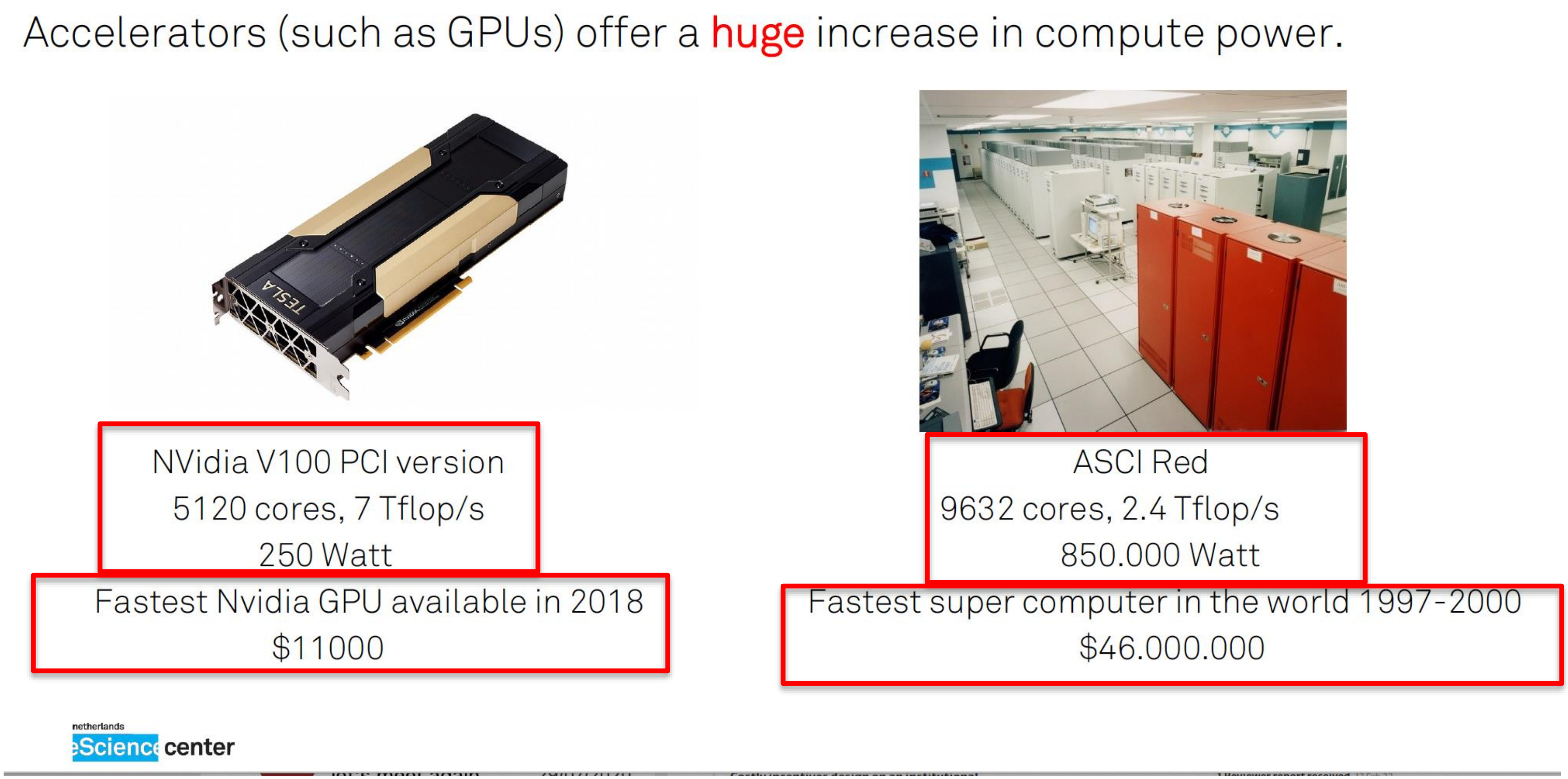
MPI, OpenMP, …



Credit: Jon Johansson Academic ICT Copyright © 2006 University of Alberta

[(*)] Computer Science profile

Do we need always need a Supercomputer to get some Speedup?



Accelerators (such as GPUs) offer a huge increase in compute power.

NVidia V100 PCI version
5120 cores, 7 Tflop/s
250 Watt

Fastest Nvidia GPU available in 2018
$11000

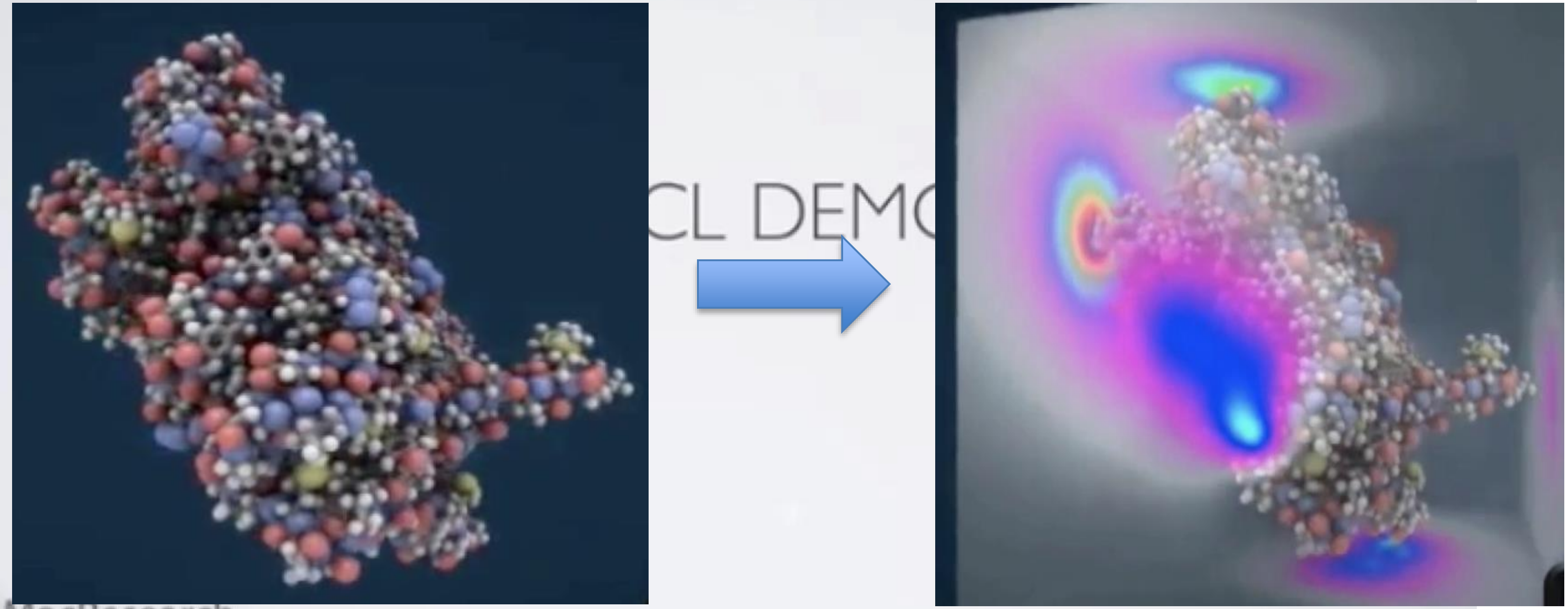
ASCI Red
9632 cores, 2.4 Tflop/s
850.000 Watt

Fastest super computer in the world 1997-2000
$46.000.000

netherlands eScience center

Do we need always a Super computer to get some Speedup?

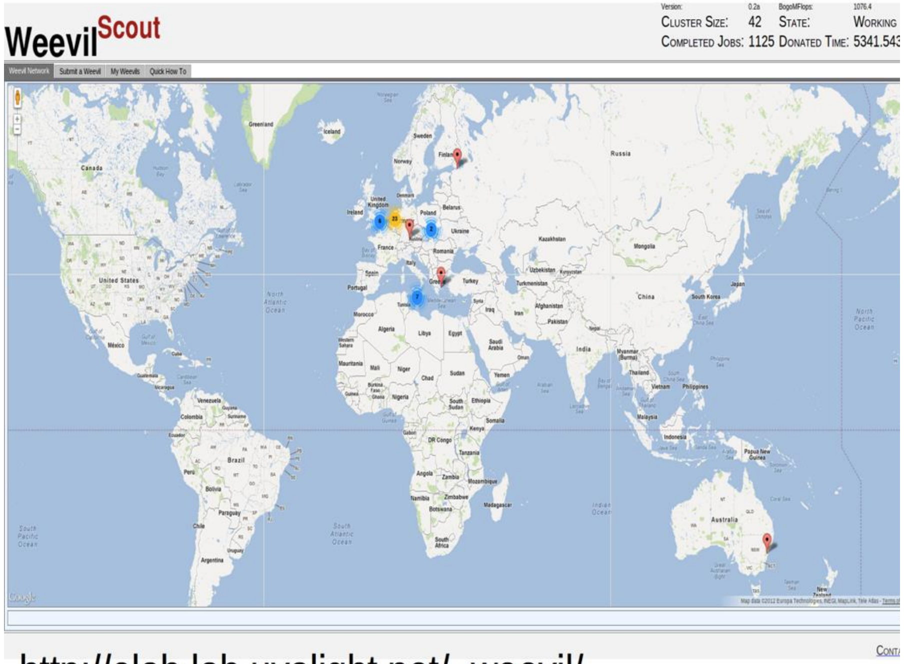
- Not necessary ➔ Do you have a Game computer?
- Demo: Software the electrostatic properties of biological molecules
  - **Usage**: drug discovery
  - **Calculation** of the boundary value condition (quite slow).
  - **GPU** : EVGA GeForce GTX 285 1GB(~ 400$)
  - Programming Language: OpenCL



Do we need a Supercomputer or GPU to get some Speedup?

- Not necessary ➔ Poor man's supercomputer

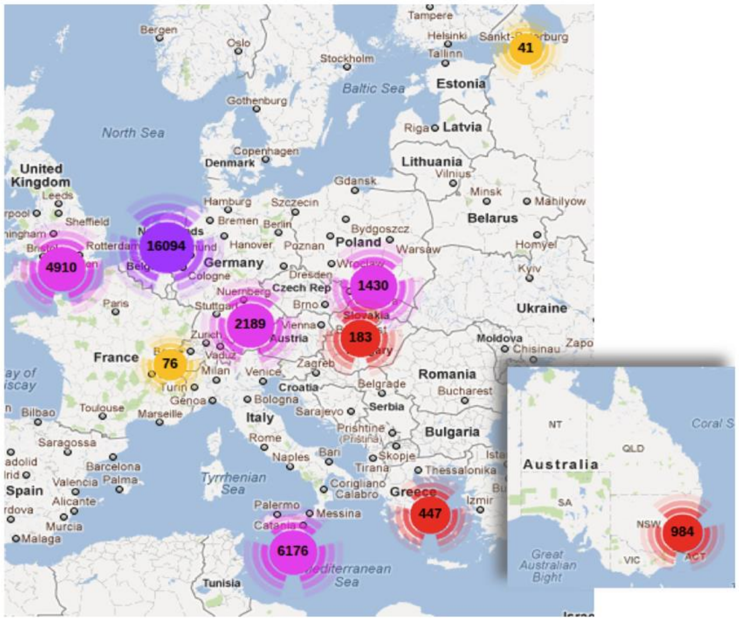
Browser computing resources



Version: 0.2a BogoMFlops: 1076.4
Cluster Size: 42 State: Working
Completed Jobs: 1125 Donated Time: 5341.543

http://elab.lab.uvalight.net/~weevil/

Browser computing resources

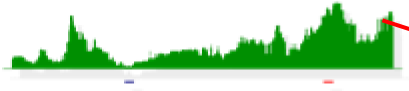


Distributed Computing on an Ensemble of Browsers, *R. Cushing, G.a Putra, S. Koulouzis, A.S.Z Belloum, M.T. Bubak, C. de Laat* IEEE Internet Computing, 10.1109/MIC.2013.3, January 2013

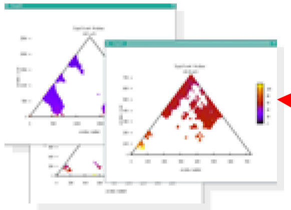
- Storing
- Processing
- Movement

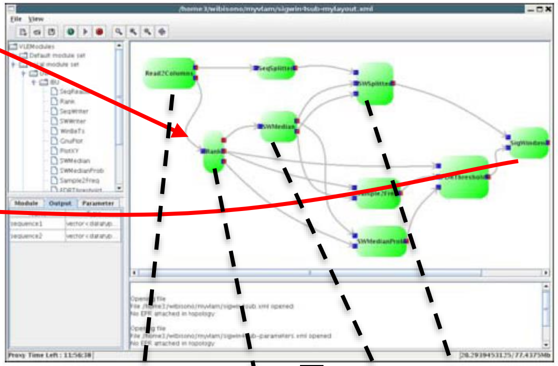
Why Use supercomputers?

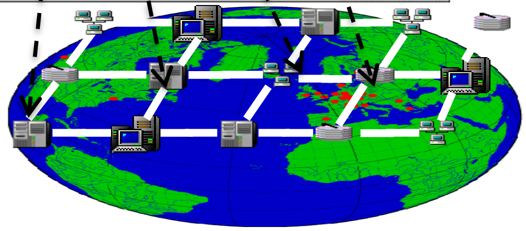
- To solve larger problems
- To use of non-local resources
- To save time and/or money



Human transcriptome map







DreamWorks Presents the Power of Supercomputing

http://www.youtube.com/watch?v=TGSRvV9u32M&feature=fvwp

Content

- Why we need Supercomputers ?
  - Big Data
- SuperComputers for everyone
  - Cloud systems
- AI a different approach to programming
  - Supervised/Unsupervised/Reinforcement Learning
  - Deep Learning
  - Limits and Challenges
- Examples
  - AWS Amazon
  - regional sea-level changes (caused by climate change)
  - GÉANT Open Cloud eXchange (gOCX)

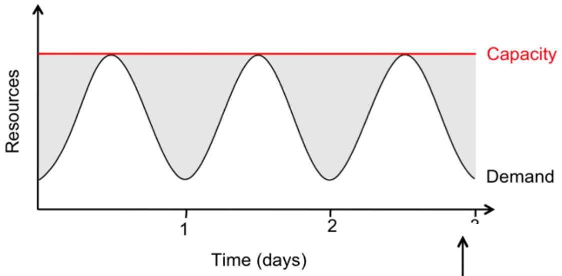
Go to Cloud part

Go to AI part

Go to Example part

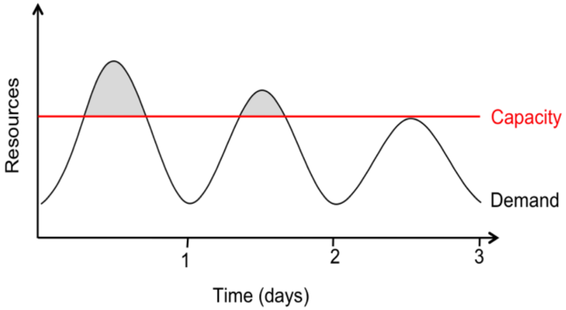
The provisioning problem

Capacity vs Demand



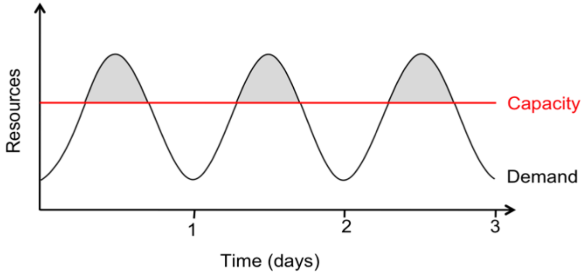
Users 🙂

providers ☹️



Users 😐

providers 😐

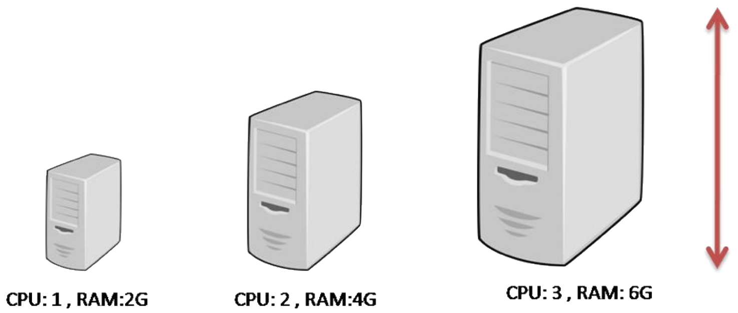


Users 😠

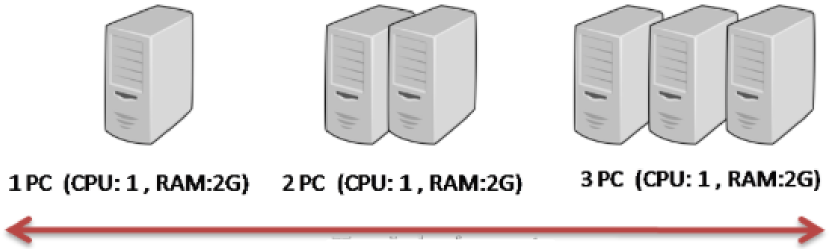
providers 🤔

Time

Elastic approach to resource provisioning



Vertical scaling / scale up

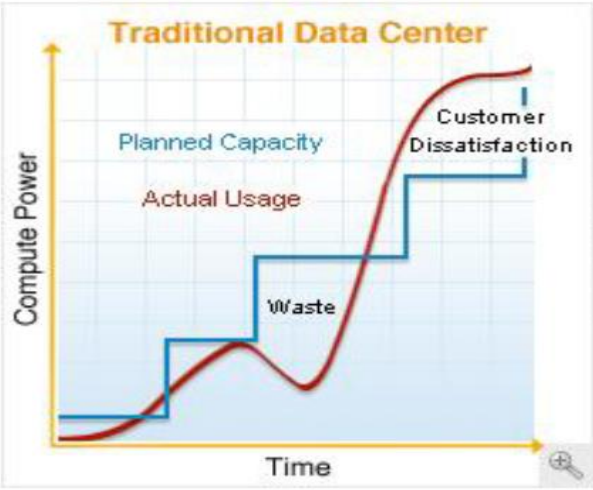


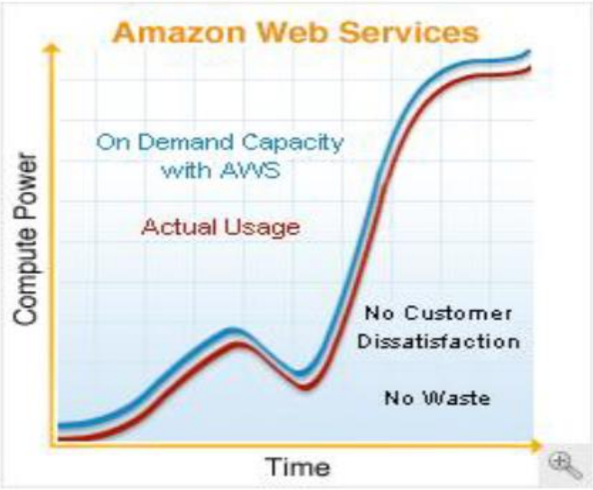
horizontal scaling / scale out

Time

Months to years

Clouds → Seconds to minutes



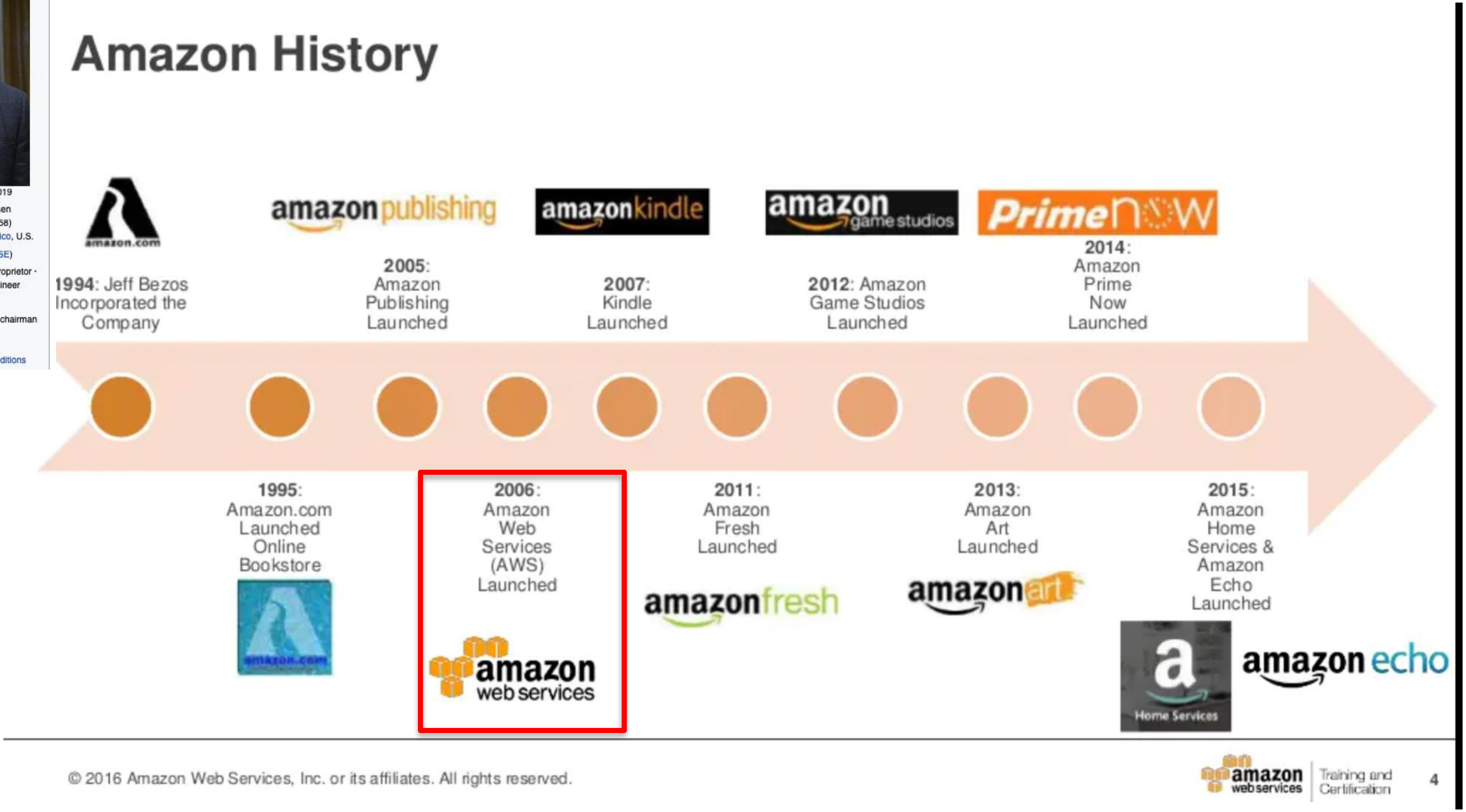


Amazon Web Services: The Pioneer in Cloud Computing

Jeff Bezos

Bezos visits LAAFB SMC in 2019

Born	Jeffrey Preston Jorgensen January 12, 1964 (age 58) Albuquerque, New Mexico, U.S.
Education	Princeton University (BSE)
Occupations	Entrepreneur · media proprietor · investor · computer engineer
Years active	1986–present
Title	Founder and executive chairman of Amazon Founder of Blue Origin Founder of Bezos Expeditions



Amazon History

1994: Jeff Bezos Incorporated the Company

1995: Amazon.com Launched Online Bookstore

2005: Amazon Publishing Launched

2006: Amazon Web Services (AWS) Launched

2007: Kindle Launched

2011: Amazon Fresh Launched

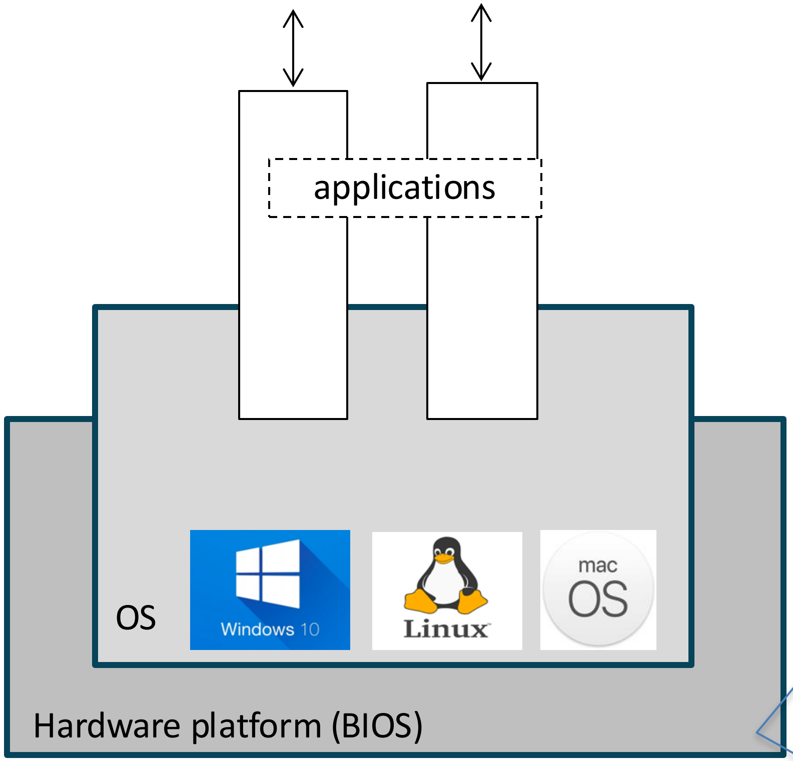
2012: Amazon Game Studios Launched

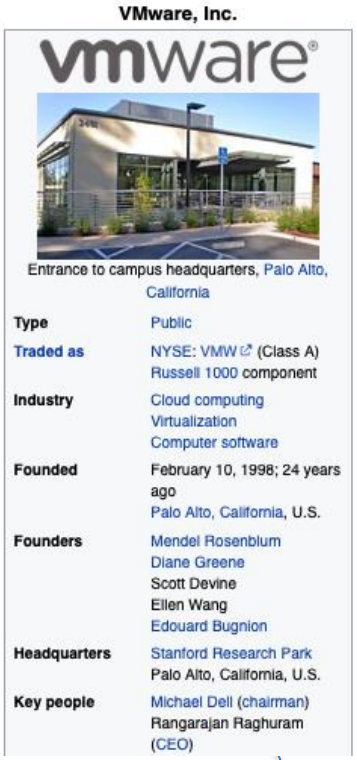
2013: Amazon Art Launched

2014: Amazon Prime Now Launched

2015: Amazon Home Services & Amazon Echo Launched

How can build such a system?



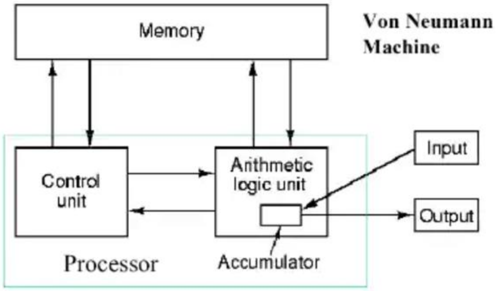


VMware, Inc.

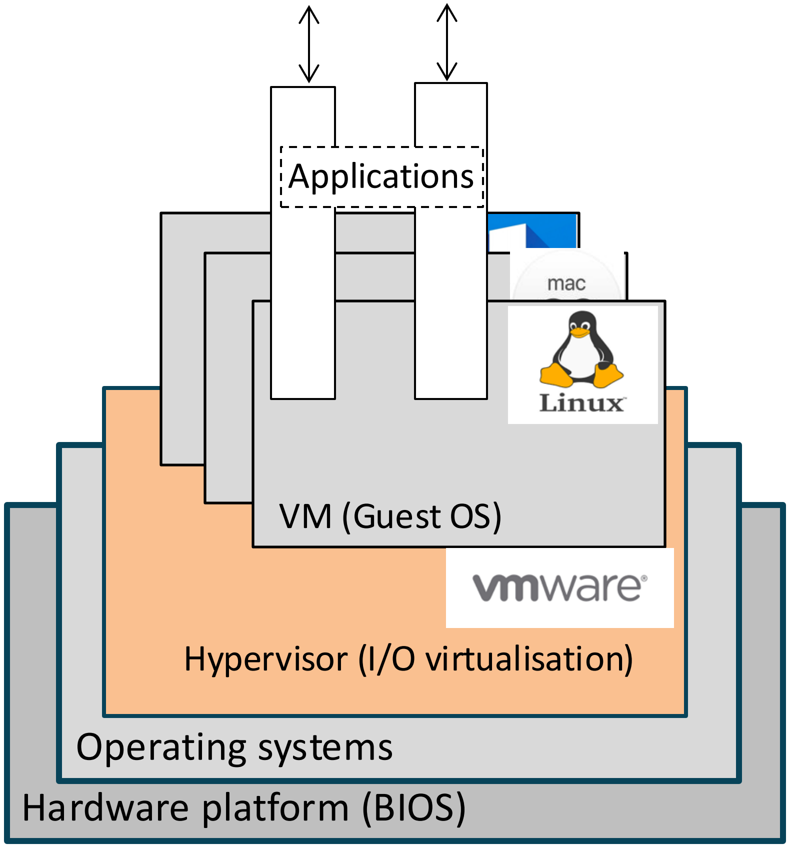
Entrance to campus headquarters, Palo Alto, California

Type	Public
Traded as	NYSE: VMW (Class A) Russell 1000 component
Industry	Cloud computing Virtualization Computer software
Founded	February 10, 1998; 24 years ago Palo Alto, California, U.S.
Founders	Mendel Rosenblum Diane Greene Scott Devine Ellen Wang Edouard Bugnion
Headquarters	Stanford Research Park Palo Alto, California, U.S.
Key people	Michael Dell (chairman) Rangarajan Raghuram (CEO)

VON NEUMANN ARCHITECTURE
(1945)



Virtualisation: Cloud Systems



Applications

mac

Linux

VM (Guest OS)

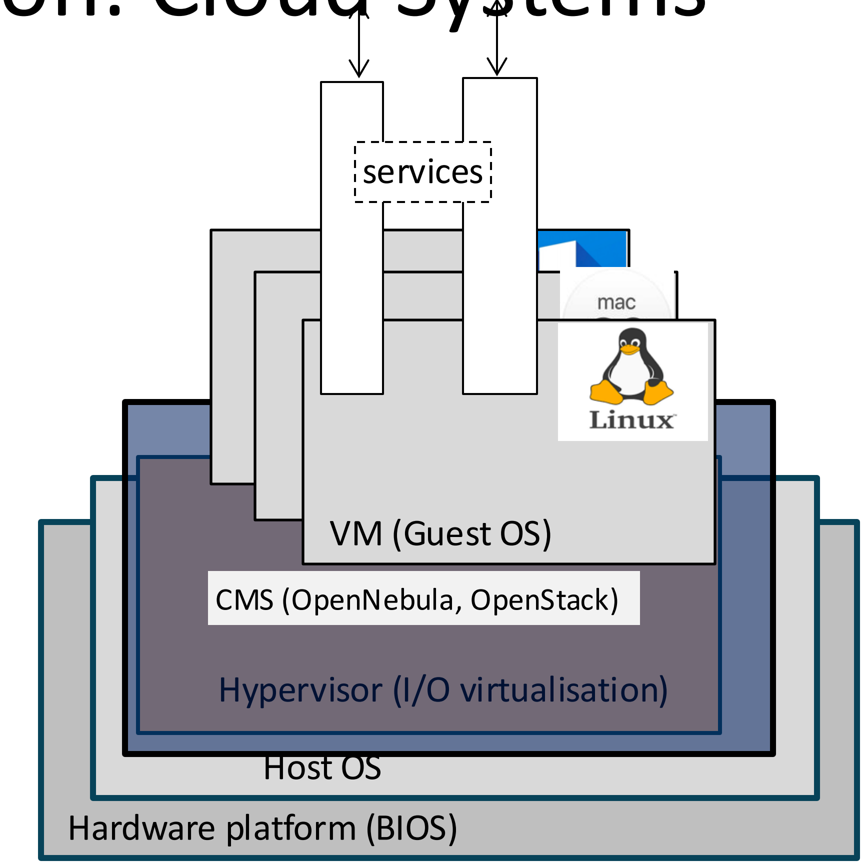
vmware

Hypervisor (I/O virtualisation)

Operating systems

Hardware platform (BIOS)

Simple Virtualization model



services

mac

Linux

VM (Guest OS)

CMS (OpenNebula, OpenStack)

Hypervisor (I/O virtualisation)

Host OS

Hardware platform (BIOS)

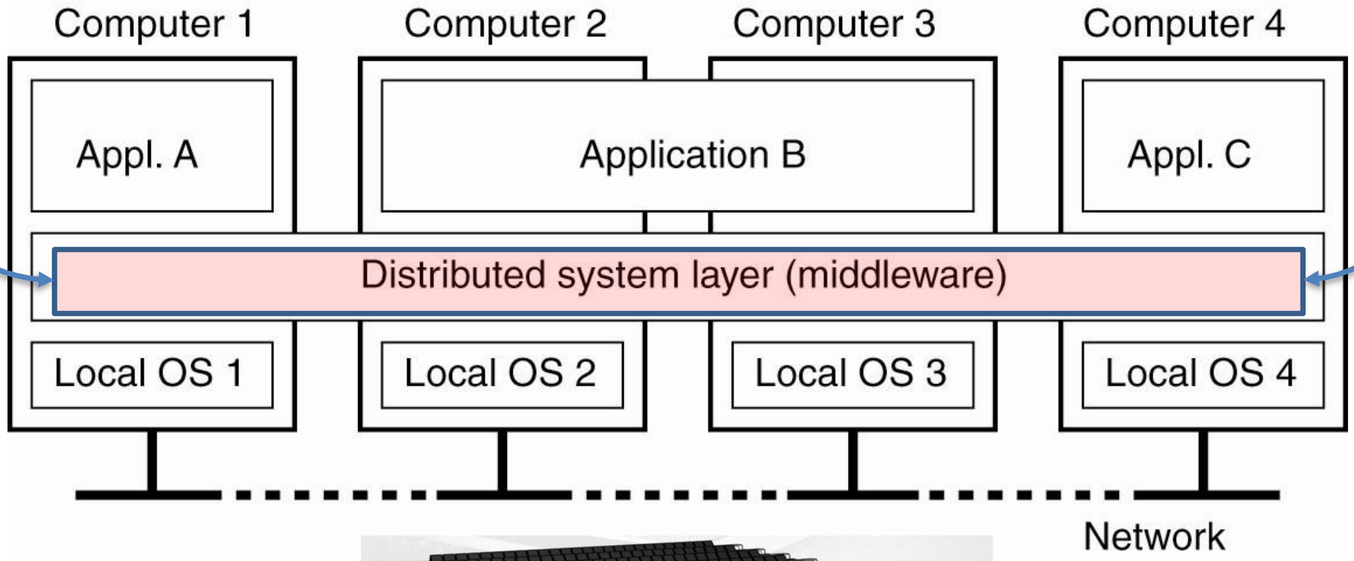
A **few** scientists run their computations on big machines.
(like this 560640 core,17 PFlop/s Titan at Oak Ridge National Laboratory)

Cloud Services model

Cloud Systems

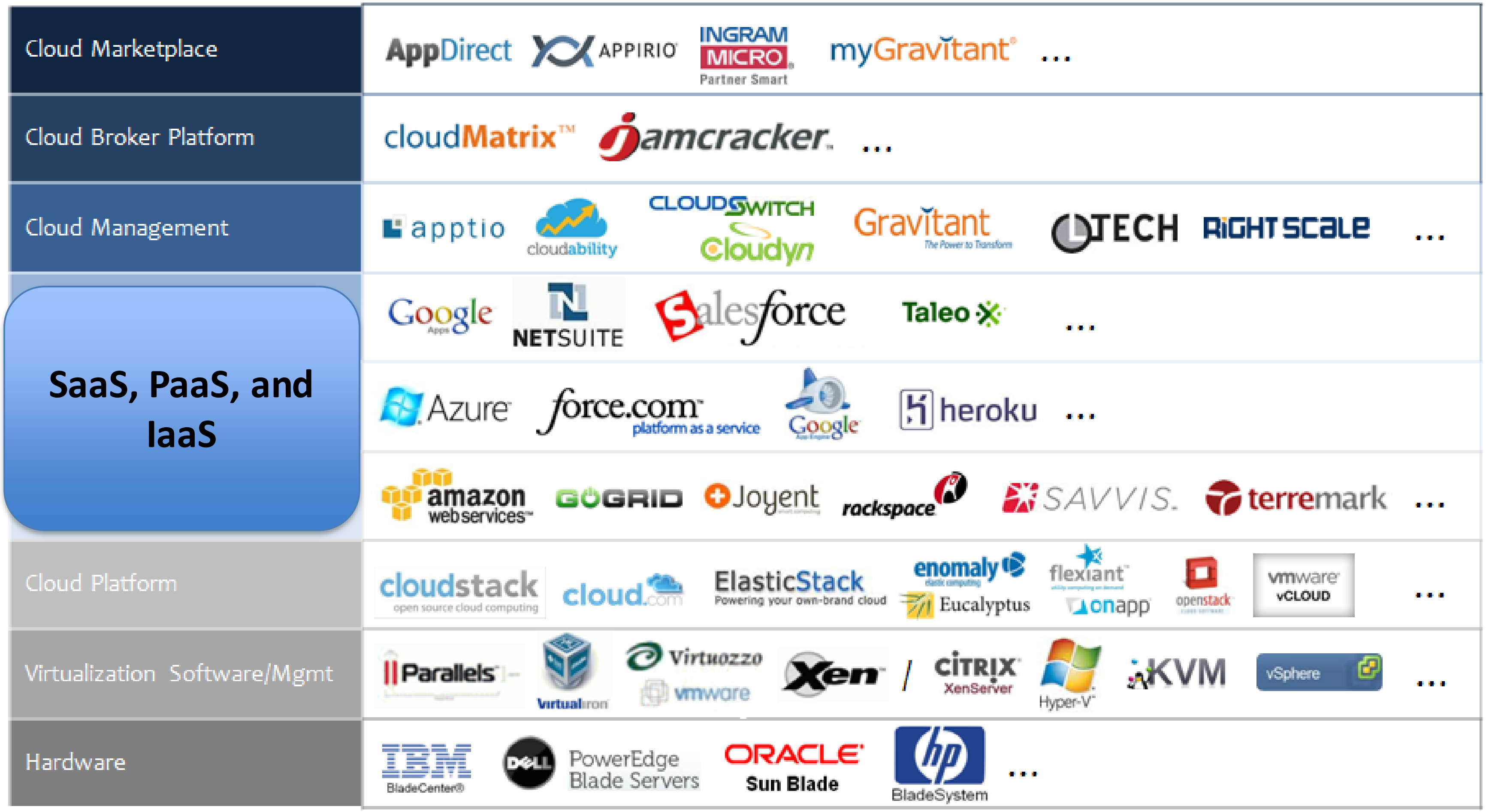
Closed source

Open Source



A **few** scientists run their computations on big machines.
(like this 560640 core,17 PFlop/s Titan at Oak Ridge National Laboratory)

Cloud provider landscape



Cloud Marketplace	AppDirect, APPIRIO, INGRAM MICRO Partner Smart, myGravitant …
Cloud Broker Platform	cloudMatrix™, jamcracker …
Cloud Management	apptio, cloudability, CLOUDSWITCH, Cloudyn, Gravitant The Power to Transform, LTECH, RightScale …
SaaS, PaaS, and IaaS	Google Apps, NETSUITE, Salesforce, Taleo …
	Azure, force.com platform as a service, Google App Engine, heroku …
	amazon web services, GOGRID, Joyent, rackspace, SAVVIS, terremark …
Cloud Platform	cloudstack open source cloud computing, cloud.com, ElasticStack Powering your own-brand cloud, enomaly elastic computing, Eucalyptus, flexiant, onapp, openstack, vmware vCLOUD …
Virtualization Software/Mgmt	Parallels, Virtualiron, Virtuozzo, vmware, Xen / CITRIX XenServer, Hyper-V, KVM, vSphere …
Hardware	IBM BladeCenter®, DELL PowerEdge Blade Servers, ORACLE Sun Blade, hp BladeSystem …

Content

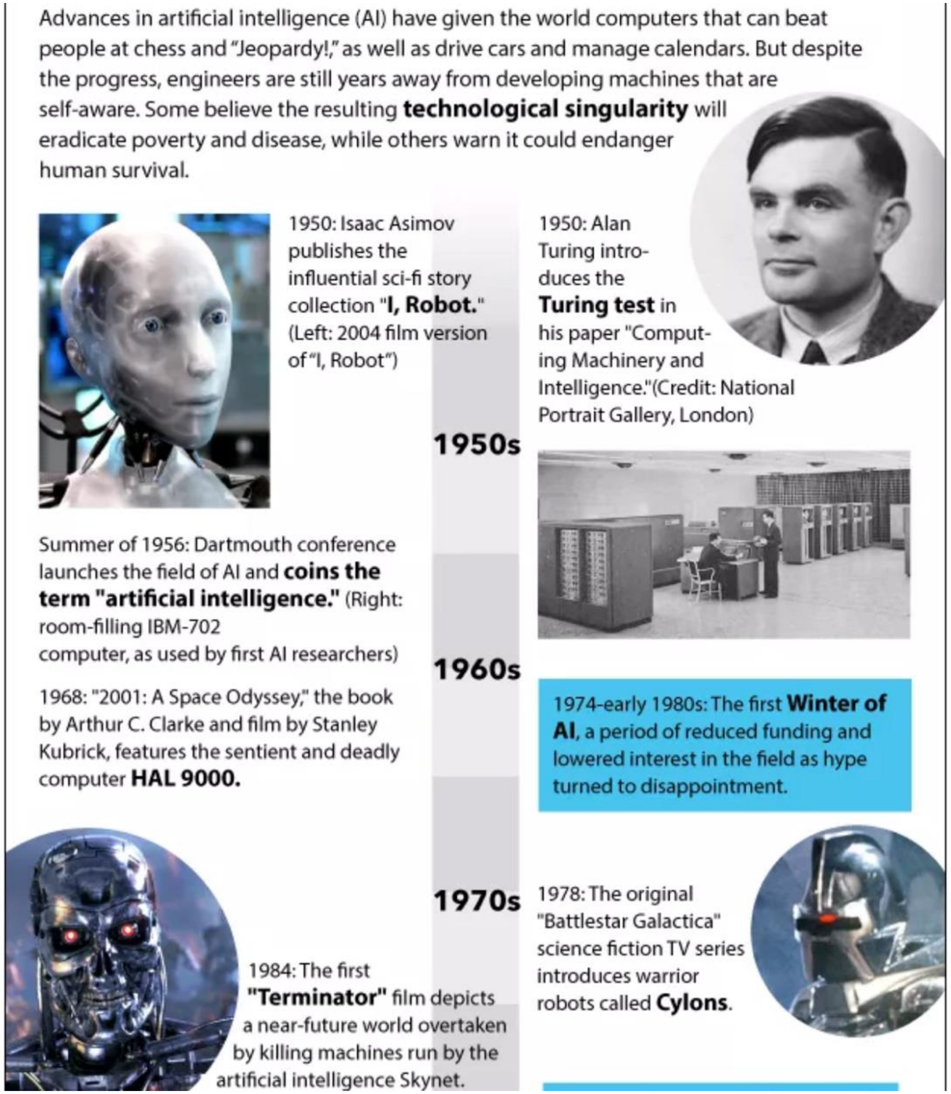
- Why we need Supercomputers ?
  - Big Data
- SuperComputers for every one
  - Cloud systems
- AI a different approach to programming
  - Supervised/Unsupervised/Reinforcement Learning
  - Deep Learning
  - Limits and Challenges
- Examples
  - AWS Amazon
  - regional sea-level changes (caused by climate change)
  - GÉANT Open Cloud eXchange (gOCX)

Go to Cloud part

Go to AI part

Go to Example part

Artificial Intelligence



Advances in artificial intelligence (AI) have given the world computers that can beat people at chess and "Jeopardy!," as well as drive cars and manage calendars. But despite the progress, engineers are still years away from developing machines that are self-aware. Some believe the resulting **technological singularity** will eradicate poverty and disease, while others warn it could endanger human survival.

1950: Isaac Asimov publishes the influential sci-fi story collection "**I, Robot.**" (Left: 2004 film version of "I, Robot")

1950: Alan Turing introduces the **Turing test** in his paper "Computing Machinery and Intelligence."(Credit: National Portrait Gallery, London)

1950s

Summer of 1956: Dartmouth conference launches the field of AI and **coins the term "artificial intelligence."** (Right: room-filling IBM-702 computer, as used by first AI researchers)

1960s

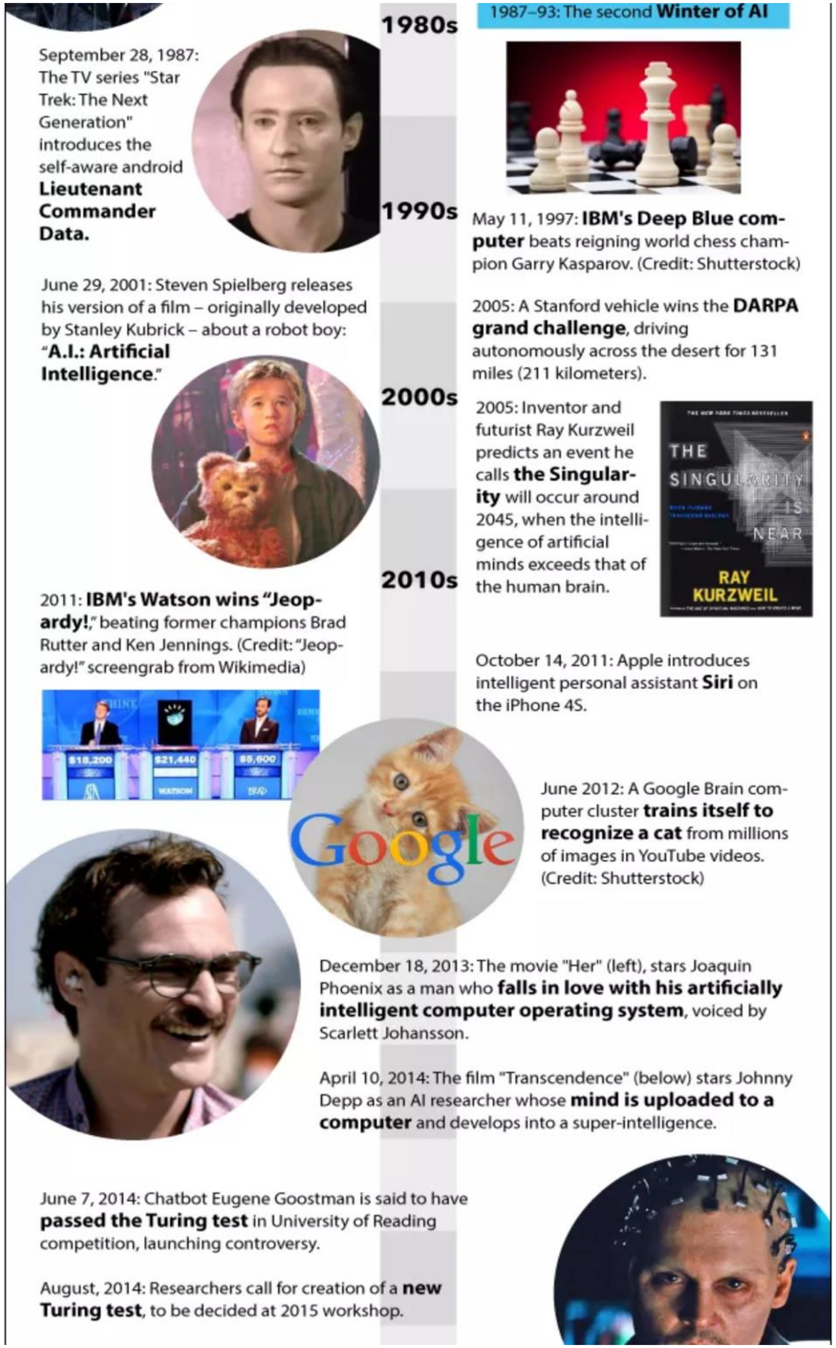
1968: "2001: A Space Odyssey," the book by Arthur C. Clarke and film by Stanley Kubrick, features the sentient and deadly computer **HAL 9000.**

1974-early 1980s: The first **Winter of AI**, a period of reduced funding and lowered interest in the field as hype turned to disappointment.

1970s

1978: The original "Battlestar Galactica" science fiction TV series introduces warrior robots called **Cylons**.

1984: The first **"Terminator"** film depicts a near-future world overtaken by killing machines run by the artificial intelligence Skynet.



1980s

1987–93: The second **Winter of AI**

September 28, 1987: The TV series "Star Trek: The Next Generation" introduces the self-aware android **Lieutenant Commander Data.**

1990s

May 11, 1997: **IBM's Deep Blue computer** beats reigning world chess champion Garry Kasparov. (Credit: Shutterstock)

June 29, 2001: Steven Spielberg releases his version of a film – originally developed by Stanley Kubrick – about a robot boy: "**A.I.: Artificial Intelligence**."

2005: A Stanford vehicle wins the **DARPA grand challenge**, driving autonomously across the desert for 131 miles (211 kilometers).

2000s

2005: Inventor and futurist Ray Kurzweil predicts an event he calls **the Singularity** will occur around 2045, when the intelligence of artificial minds exceeds that of the human brain.

2010s

2011: **IBM's Watson wins "Jeopardy!,"** beating former champions Brad Rutter and Ken Jennings. (Credit: "Jeopardy!" screengrab from Wikimedia)

October 14, 2011: Apple introduces intelligent personal assistant **Siri** on the iPhone 4S.

June 2012: A Google Brain computer cluster **trains itself to recognize a cat** from millions of images in YouTube videos. (Credit: Shutterstock)

December 18, 2013: The movie "Her" (left), stars Joaquin Phoenix as a man who **falls in love with his artificially intelligent computer operating system**, voiced by Scarlett Johansson.

April 10, 2014: The film "Transcendence" (below) stars Johnny Depp as an AI researcher whose **mind is uploaded to a computer** and develops into a super-intelligence.

June 7, 2014: Chatbot Eugene Goostman is said to have **passed the Turing test** in University of Reading competition, launching controversy.

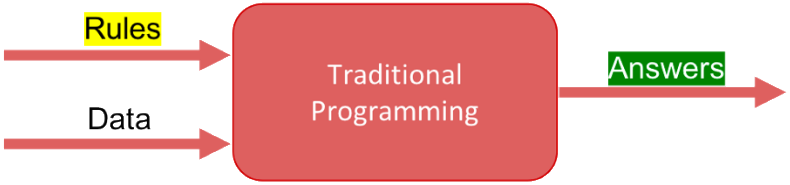
August, 2014: Researchers call for creation of a **new Turing test**, to be decided at 2015 workshop.

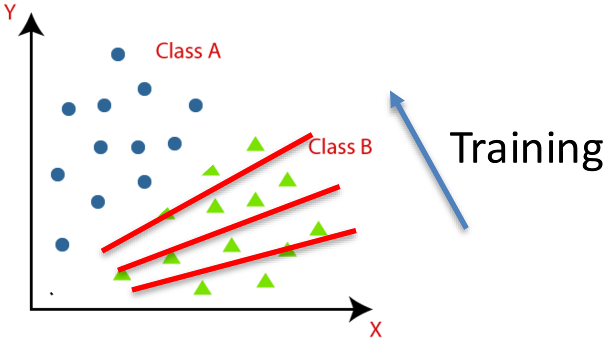
History of A.I.: Artificial Intelligence (Infographic)

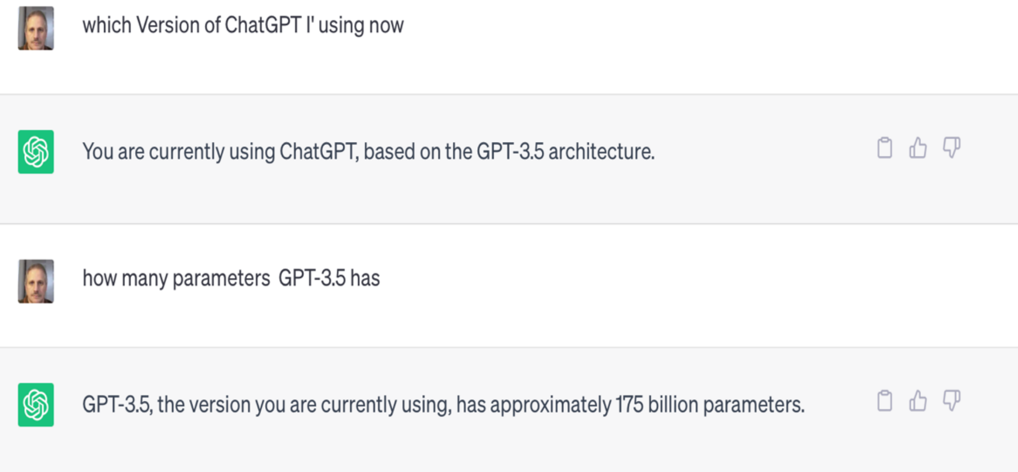
https://www.livescience.com/47544-history-of-a-i-artificial-intelligence-infographic.html

AI / Machine Learning / Big Data ?

- “ML is a scientific discipline that deals with the construction and study of algorithms that can learn from data. Such Algorithms operate in 2 steps:
  1. building a model based on the data
  2. using the model make predictions and decision rather that following explicitly programmed instructions “





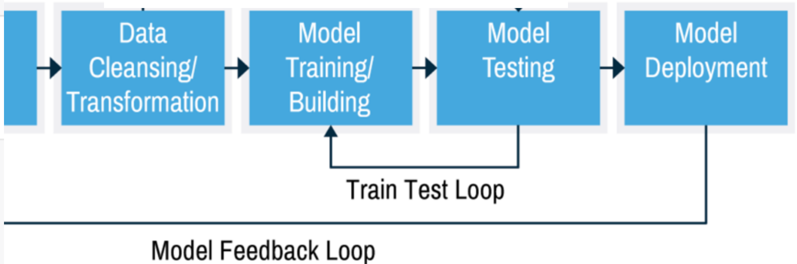


which Version of ChatGPT I' using now

You are currently using ChatGPT, based on the GPT-3.5 architecture.

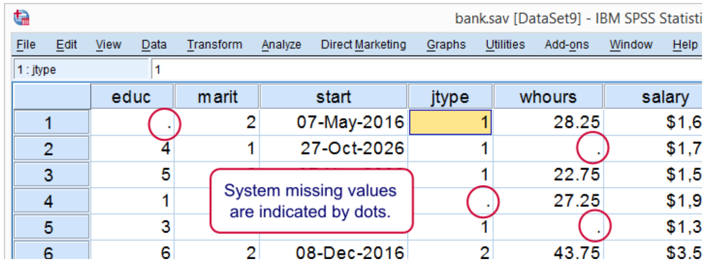
how many parameters GPT-3.5 has

GPT-3.5, the version you are currently using, has approximately 175 billion parameters.



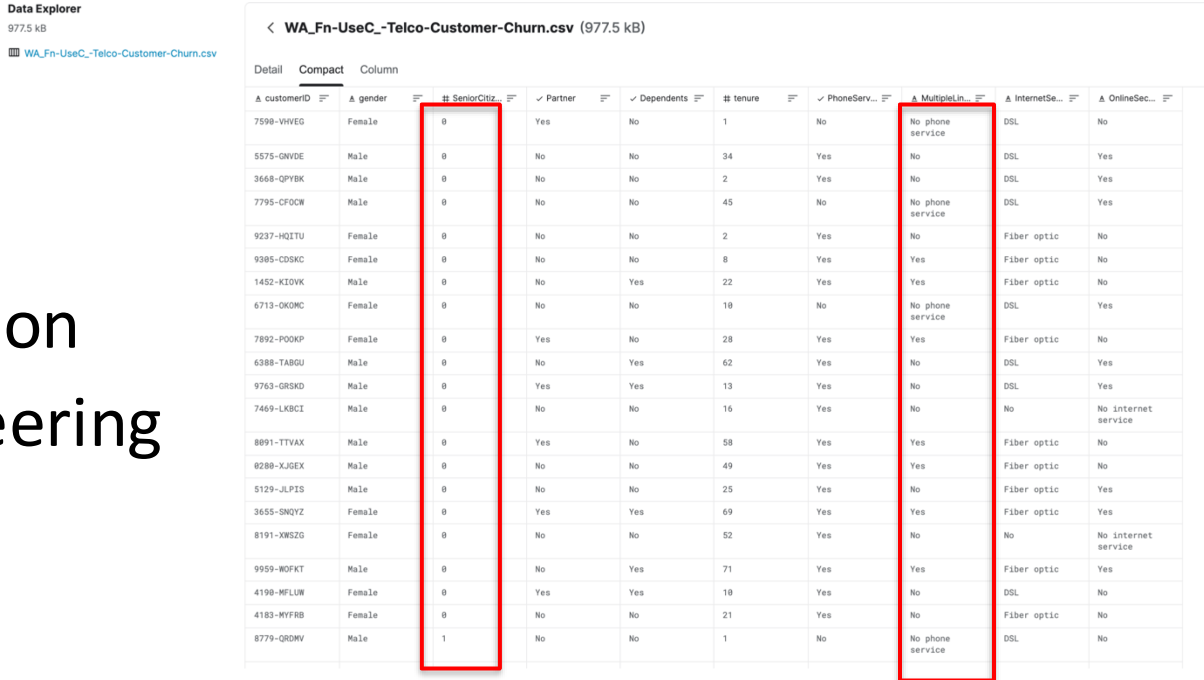
Is the data ready to be processed?

- Not always:
  - Data not in the correct format: images, voice, text…
  - Dealing with missing values
  - Dealing with noise (errors) in the data values


bank.sav [DataSet9] - IBM SPSS Statisti
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
1 : jtype 1

	educ	marit	start	jtype	whours	salary
1	.	2	07-May-2016	1	28.25	$1,6
2	4	1	27-Oct-2026	1	.	$1,7
3	5			1	22.75	$1,5
4	1			.	27.25	$1,9
5	3			1	.	$1,3
6	6	2	08-Dec-2016	2	43.75	$3.5

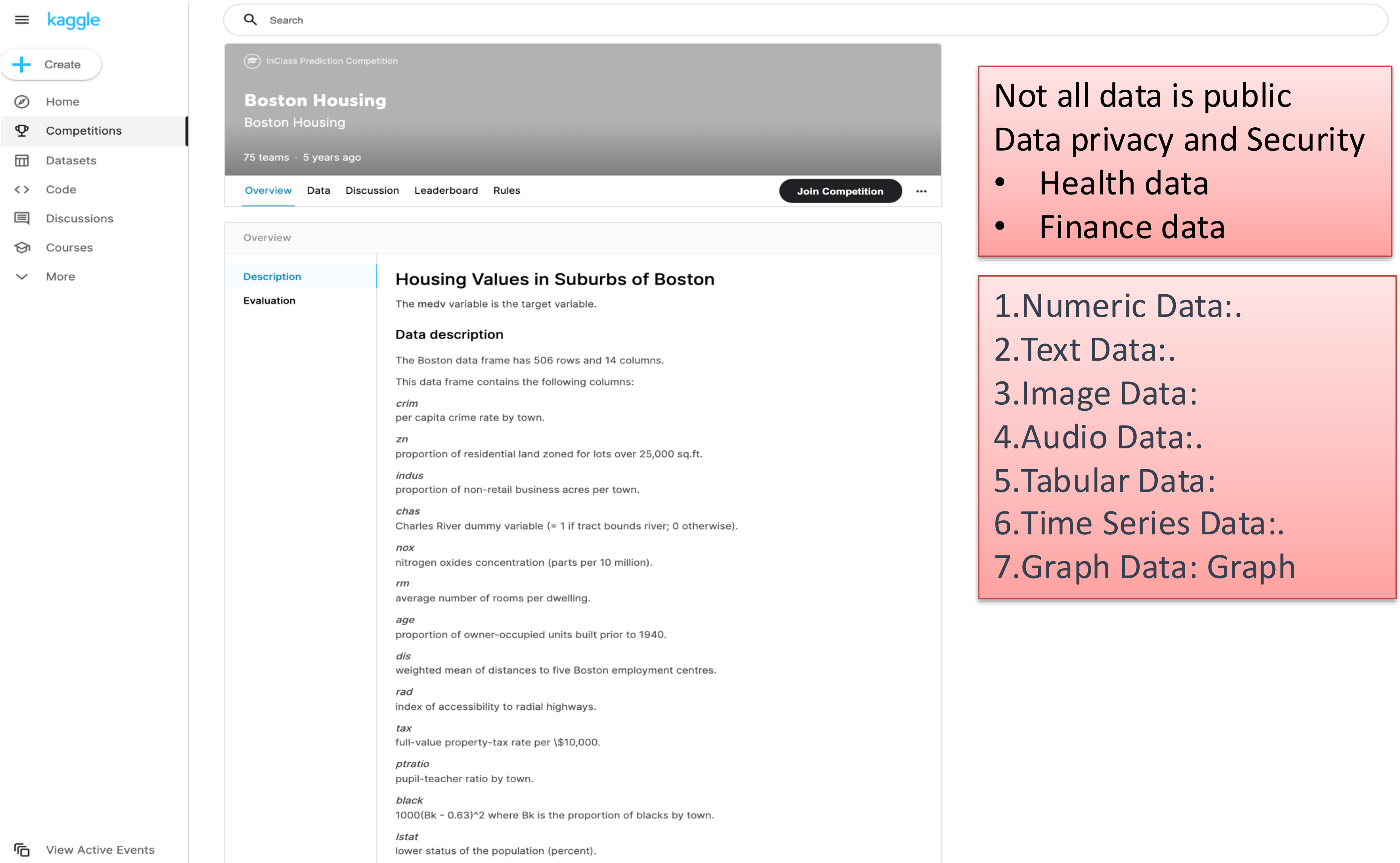
System missing values are indicated by dots.


Data Explorer
977.5 kB
WA_Fn-UseC_-Telco-Customer-Churn.csv
< WA_Fn-UseC_-Telco-Customer-Churn.csv (977.5 kB)
Detail Compact Column

customerID	gender	SeniorCitiz...	Partner	Dependents	tenure	PhoneServ...	MultipleLin...	InternetSe...	OnlineSec...
7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No
5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes
3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes
7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes
9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No
9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber optic	No
1452-KIOVK	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No
6713-OKOMC	Female	0	No	No	10	No	No phone service	DSL	Yes
7892-POOKP	Female	0	Yes	No	28	Yes	Yes	Fiber optic	No
6388-TABGU	Male	0	No	Yes	62	Yes	No	DSL	Yes
9763-GRSKD	Male	0	Yes	Yes	13	Yes	No	DSL	Yes
7469-LKBCI	Male	0	No	No	16	Yes	No	No	No internet service
8091-TTVAX	Male	0	Yes	No	58	Yes	Yes	Fiber optic	No
0280-XJGEX	Male	0	No	No	49	Yes	Yes	Fiber optic	No
5129-JLPIS	Male	0	No	No	25	Yes	No	Fiber optic	Yes
3655-SNQYZ	Female	0	Yes	Yes	69	Yes	Yes	Fiber optic	Yes
8191-XWSZG	Female	0	No	No	52	Yes	No	No	No internet service
9959-WOFKT	Male	0	No	Yes	71	Yes	Yes	Fiber optic	Yes
4190-MFLUW	Female	0	Yes	Yes	10	Yes	No	DSL	No
4183-MYFRB	Female	0	No	No	21	Yes	No	Fiber optic	No
8779-QRDMV	Male	1	No	No	1	No	No phone service	DSL	No

- Pre-processing
  - Feature Selection
  - Feature engineering

Example of Data Set: Boston housing set …



Boston Housing

Boston Housing

75 teams · 5 years ago

Housing Values in Suburbs of Boston

The **medv** variable is the target variable.

Data description

The Boston data frame has 506 rows and 14 columns.

This data frame contains the following columns:

crim
per capita crime rate by town.

zn
proportion of residential land zoned for lots over 25,000 sq.ft.

indus
proportion of non-retail business acres per town.

chas
Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).

nox
nitrogen oxides concentration (parts per 10 million).

rm
average number of rooms per dwelling.

age
proportion of owner-occupied units built prior to 1940.

dis
weighted mean of distances to five Boston employment centres.

rad
index of accessibility to radial highways.

tax
full-value property-tax rate per \$10,000.

ptratio
pupil-teacher ratio by town.

black
1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.

lstat
lower status of the population (percent).

Not all data is public

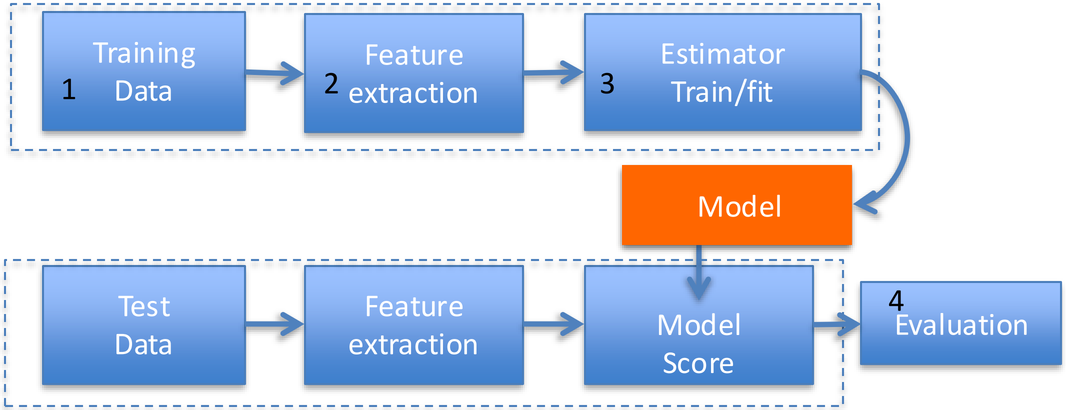
Data privacy and Security

- Health data
- Finance data

1.Numeric Data:.
2.Text Data:.
3.Image Data:
4.Audio Data:.
5.Tabular Data:
6.Time Series Data:.
7.Graph Data: Graph

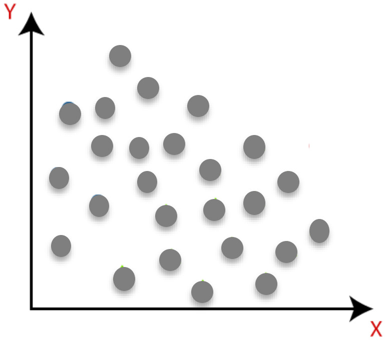
The Machine Learning WorkFlow

1. building a **model** based on the data

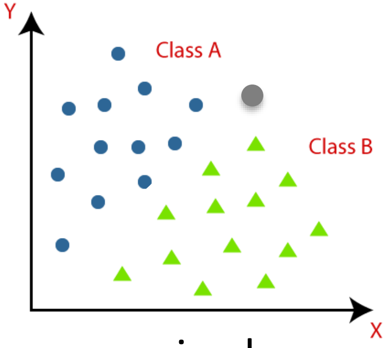


2. using **the model** make **predictions** and **decisions** rather that following explicitly programmed instructions “

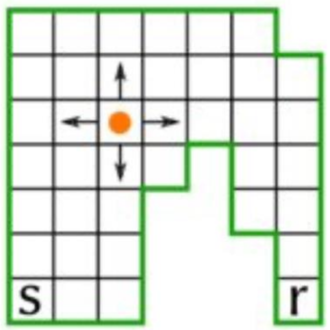
The types of Machine Learning



Unsupervised



supervised

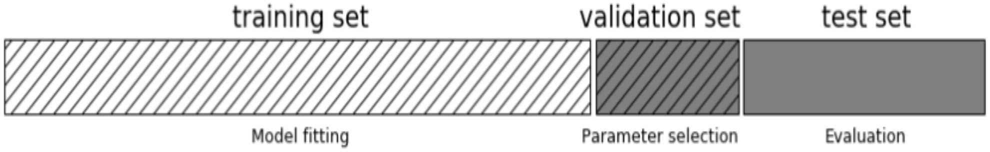


Reinforcement learning

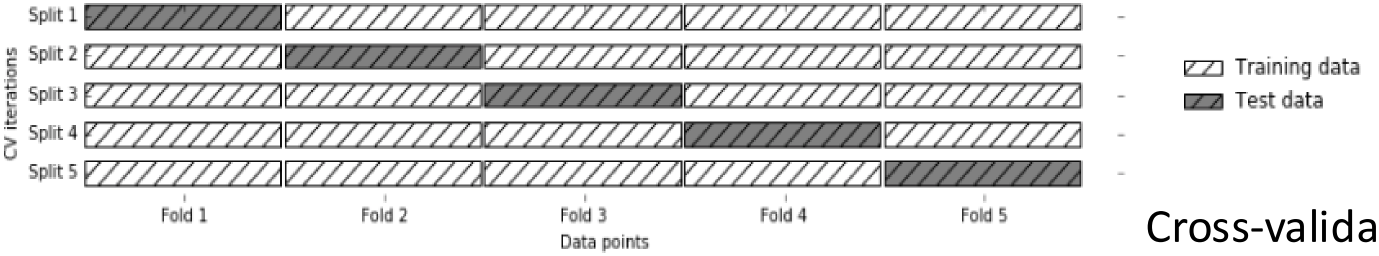
How to split the input dataset into: training data and test data?

Simple Answer ➔ There many ways

- **Simple split (train, test) ➔ (default 75%, 25%) or any proportion**
- Threefold split (train, test, validate)
- Cross-validation:
  - Nested cross-validation, Stratified cross-validation, TimeSeriesSplit



➢ Threefold split

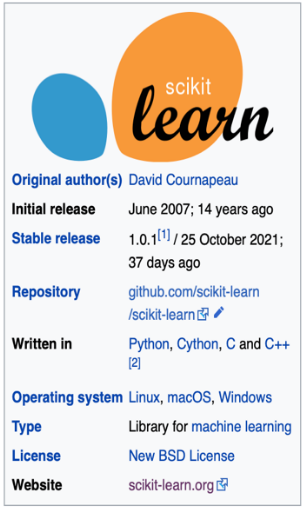


Cross-validation

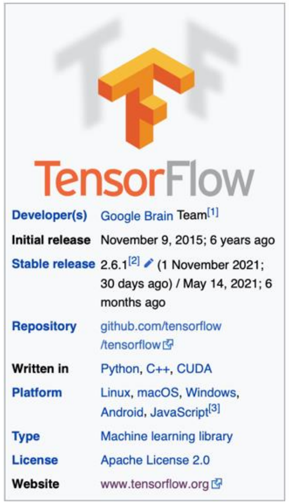
Which ML Algorithms to use?

There are many Machine learning Algorithms (Models) with different: Model complexity, computational Complexity, memory usage,

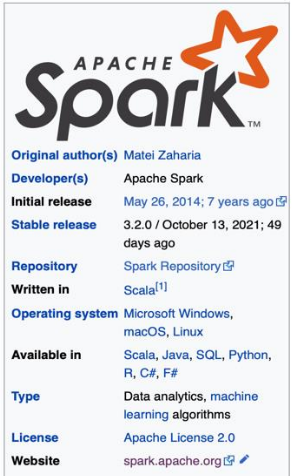
- Which one to use? ➔ depends on the application
- **Basic models**
  1. Nearest Neighbours,
  2. Nearest Centroid
  3. Linear Classification and Regression
  4. Logistic Regression
- **Non-Linear models**
  6. Support Vector Machines and Kernels
  7. Decision Trees
  8. Random Forests
  9. Gradient Boosting
  10. Model Calibration



Original author(s)	David Cournapeau
Initial release	June 2007; 14 years ago
Stable release	1.0.1[1] / 25 October 2021; 37 days ago
Repository	github.com/scikit-learn/scikit-learn
Written in	Python, Cython, C and C++ [2]
Operating system	Linux, macOS, Windows
Type	Library for machine learning
License	New BSD License
Website	scikit-learn.org



Developer(s)	Google Brain Team[1]
Initial release	November 9, 2015; 6 years ago
Stable release	2.6.1[2] (1 November 2021; 30 days ago) / May 14, 2021; 6 months ago
Repository	github.com/tensorflow/tensorflow
Written in	Python, C++, CUDA
Platform	Linux, macOS, Windows, Android, JavaScript[3]
Type	Machine learning library
License	Apache License 2.0
Website	www.tensorflow.org

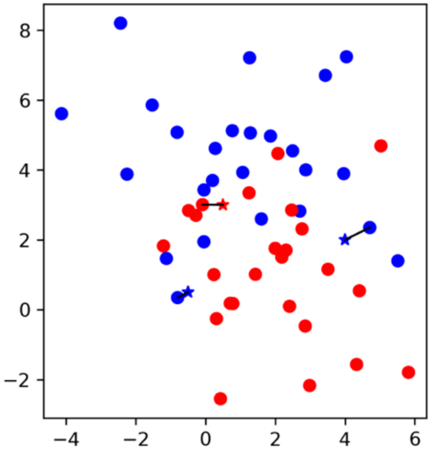


Original author(s)	Matei Zaharia
Developer(s)	Apache Spark
Initial release	May 26, 2014; 7 years ago
Stable release	3.2.0 / October 13, 2021; 49 days ago
Repository	Spark Repository
Written in	Scala[1]
Operating system	Microsoft Windows, macOS, Linux
Available in	Scala, Java, SQL, Python, R, C#, F#
Type	Data analytics, machine learning algorithms
License	Apache License 2.0
Website	spark.apache.org

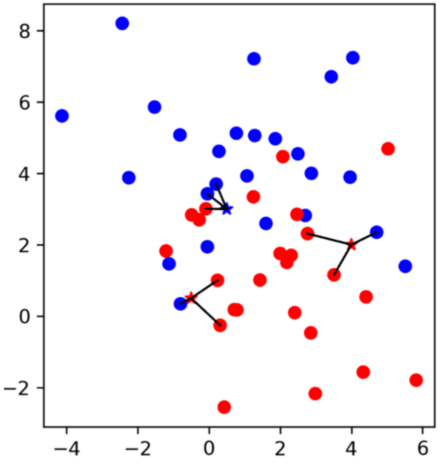
Nearest Neighbours

The algorithm

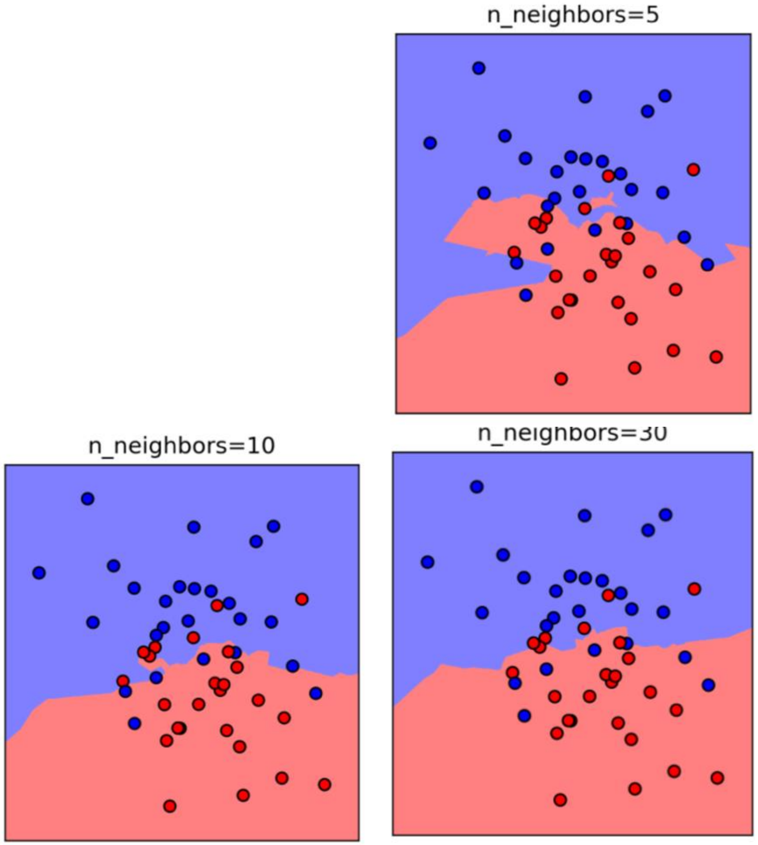
1neighbour



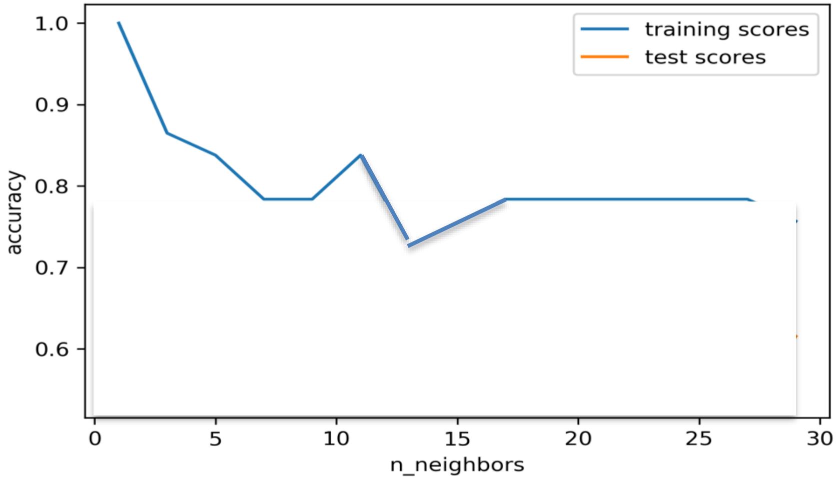
3 neighbours



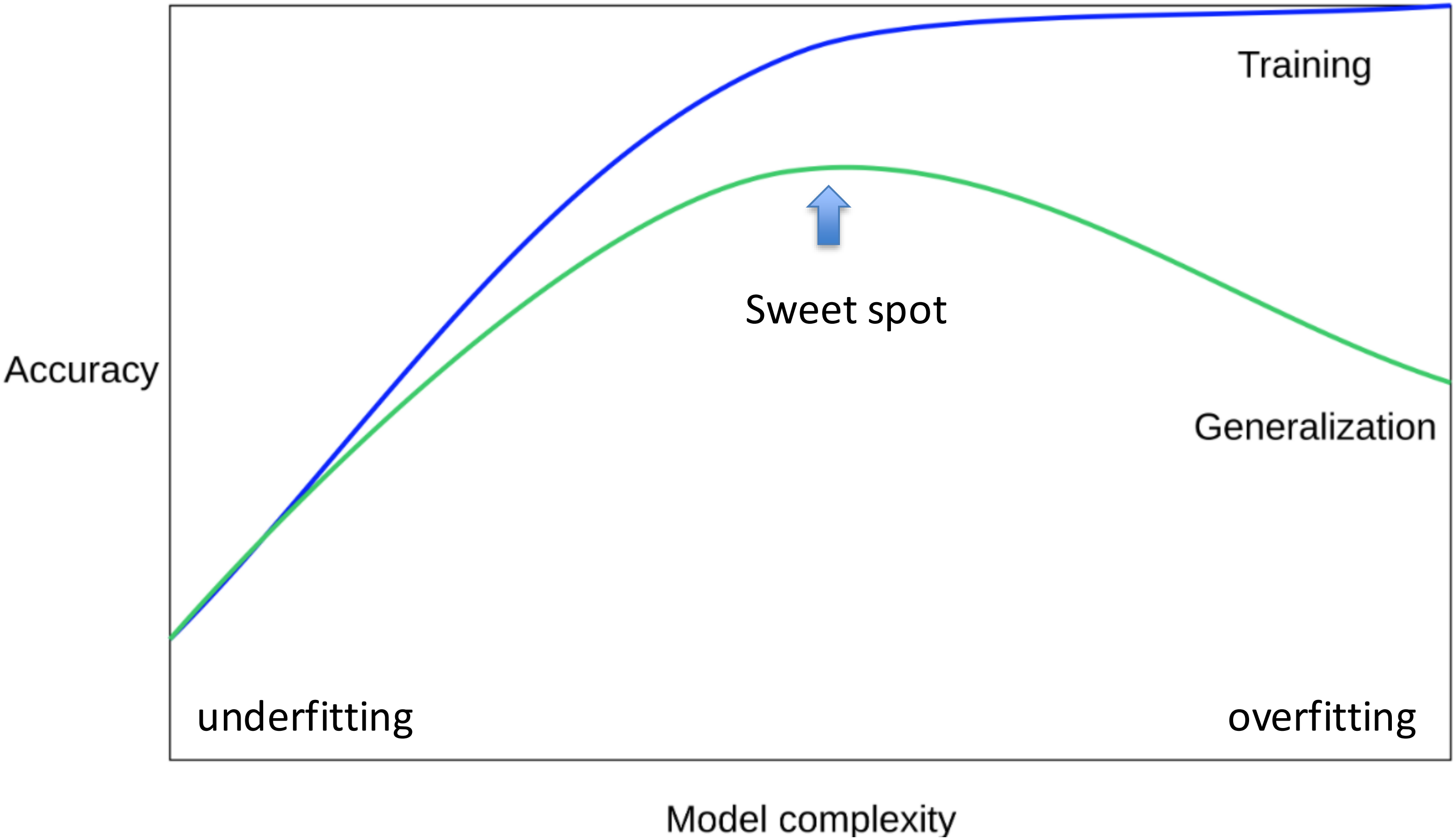
The model



Complexity of the model



Accuracy of the model



Computational properties of the models

In the era of BigData the ML Model will likely work large data sets:

- What the **computational complexity** of the **training**?
- What is the **computational complexity** of the **prediction**?
- What the **memory consumption**?

- Some ChatGPT commentators have estimated that if ChatGPT was to be trained on a single NVIDIA Tesla V100 'Graphics Processing Unit' (GPU) that it would take around 355 years to complete ChatGPT's training on its training dataset.
  - However, OpenAI reportedly used **1,023 A100 GPUs** to train ChatGPT, so **it is possible that the training process was completed in as little as 34 days**. (Source: Lambda Labs.)

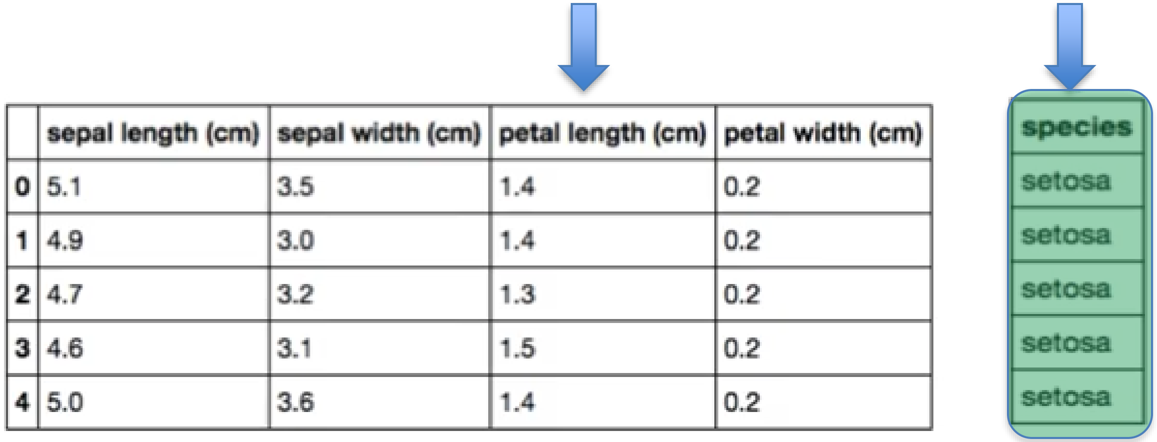
- The costs of training ChatGPT is estimated to be just under $5 million dollars. (Source: Lambda Labs.)

Supervised learning

- Input data (training, test) is labelled
  - Predictor variables/features and a target variable

Predictor variables

target variables



	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2

species
setosa
setosa
setosa
setosa
setosa

Iris Data Set

- Aim: predict the target variable given the predictor variables

Supervised Learning

Example Question	Training Data
How much is a home worth?	Previous home sales
Will a customer default on a loan?	Previous loan that were paid/defaulted
How many customers will apply for a loan next month?	Previous months of loans applications
Is this cancer Malignant?	Previous Stats of benign /malignant cancers

Unsupervised learning

- Data is not labeled
- Goal is to **uncover hidden patterns** in the data
- Example: grouping customers into distinct categories (Clustering)

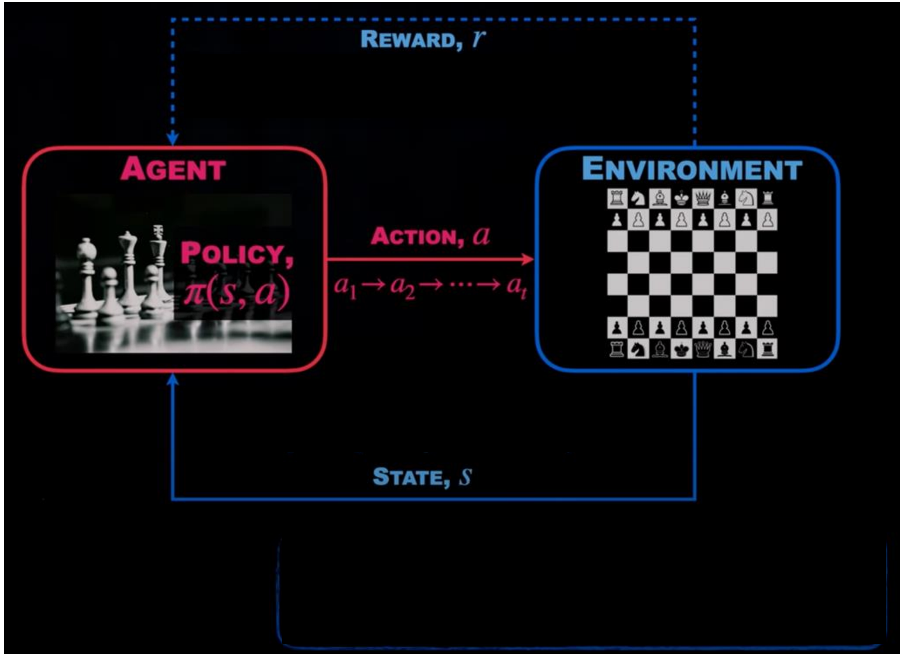
Date	Customer	Account	Auth	Class	Zip	amount	
Mon	Bob	3421	Pin	Clothes	46140	135	
Tue	Bob	3421	Sign	Food	*46140*	401	
Tue	Alice	2456	Pin	Food	12222	234	
Wed	Sally	6788	Pin	Gas	26339	94	similar
Wed	Bob	3421	Pin	tech	21350	2459	Anomaly detection
Wed	Bob	3421	Pin	gas	26339	83	similar
thr	Sally	6788	Sign	food	46140	51	

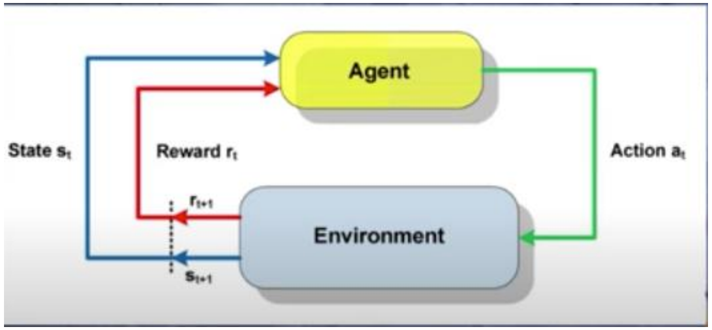
Supervised Learning

Example Question	Training Data
Are certain customers similar?	Customer profiles
Is a transaction Unusual?	Previous transactions
Are certain products purchased together?	Example of previous purchases

Reinforced learning

- Software agents interact with an environment
  - Learn how to optimize their behavior
  - Given a system of rewards and punishments



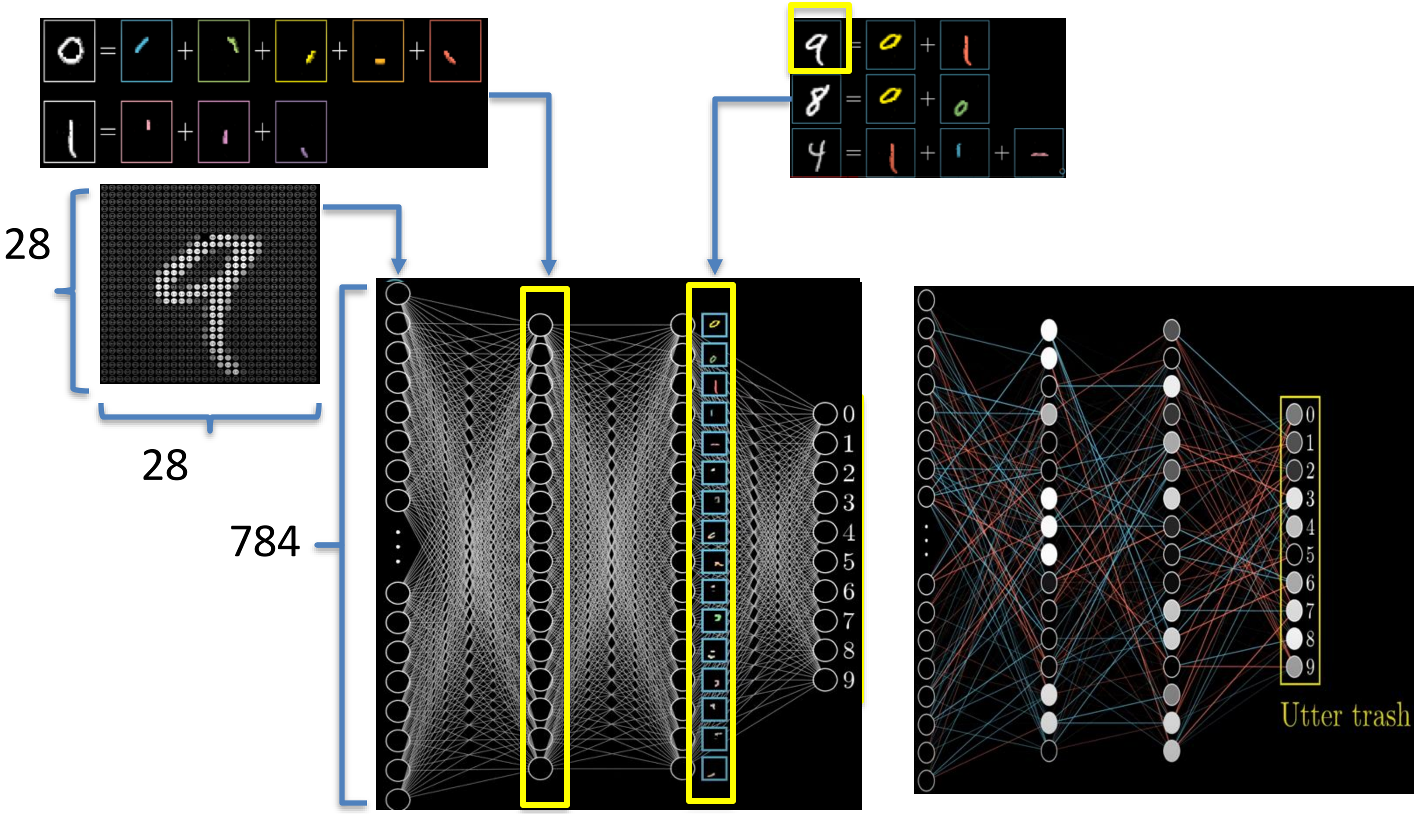


Example

AlphaGo. First computer to defeat the world champion in Go

Deep Learning

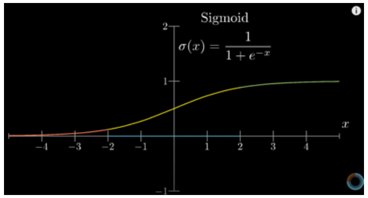
- Neuron Network multiple Layers
  - Neuron → Function that outputs a number (0-1)

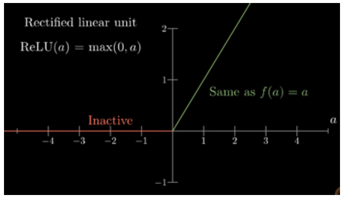


Multilayer Perceptron

Deep Learning

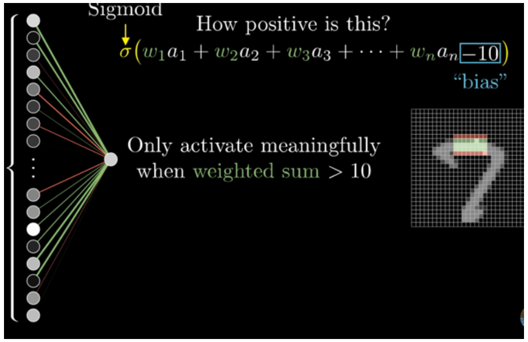
- Neuron Network multiple Layers
  - Neuron → Function that outputs a number (0-1)
    - Activation Function



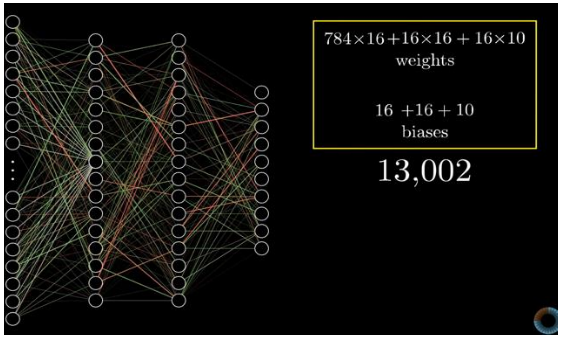


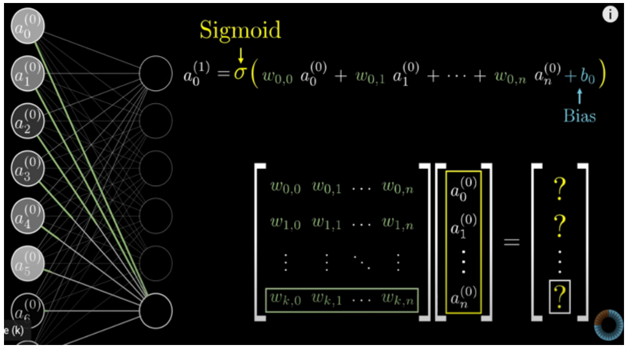
    - In/out connection

For each neuron


Sigmoid
How positive is this?
$\sigma(w_1a_1 + w_2a_2 + w_3a_3 + \cdots + w_na_n \boxed{-10})$
"bias"
Only activate meaningfully when weighted sum > 10

For all neurons in all layers




Sigmoid
$a_0^{(1)} = \sigma(w_{0,0}\, a_0^{(0)} + w_{0,1}\, a_1^{(0)} + \cdots + w_{0,n}\, a_n^{(0)} + b_0)$
Bias
$\begin{bmatrix} w_{0,0} & w_{0,1} & \cdots & w_{0,n} \\ w_{1,0} & w_{1,1} & \cdots & w_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{k,0} & w_{k,1} & \cdots & w_{k,n} \end{bmatrix} \begin{bmatrix} a_0^{(0)} \\ a_1^{(0)} \\ \vdots \\ a_n^{(0)} \end{bmatrix} = \begin{bmatrix} ? \\ ? \\ \vdots \\ ? \end{bmatrix}$
$a_0^{(0)}$ $a_1^{(0)}$ $a_2^{(0)}$ $a_3^{(0)}$ $a_4^{(0)}$ $a_5^{(0)}$ $a_6^{(0)}$

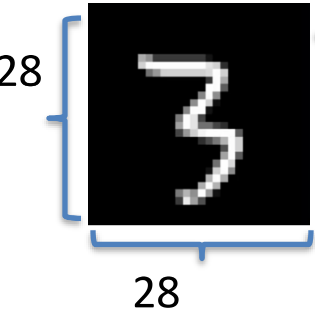
Note: Initial setting **weights** and **Bias** are random values (initial output is just a random output)
The network need to be trained in other word find better values for the **Weights** and **Bias**

Deep Learning

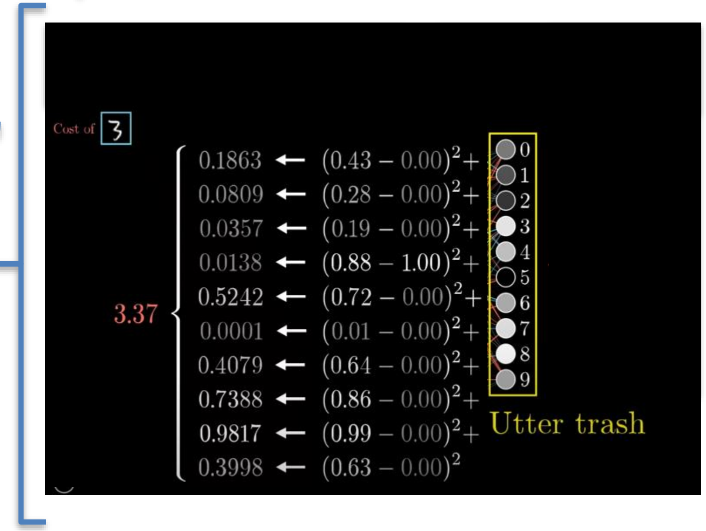
- Neuron Network multiple Layers
  - Neuron → Function that outputs a number (0-1)
    - Activation Function
    - In/out connection
  - Cost Function
  - Backpropagation

Needed for the training of the network

Training will a lot of Data (numbers)

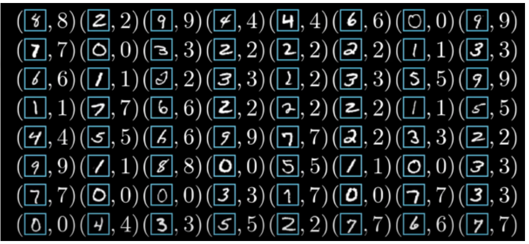


784



$$3.37 \left\{ \begin{array}{l} 0.1863 \leftarrow (0.43 - 0.00)^2 + \\ 0.0809 \leftarrow (0.28 - 0.00)^2 + \\ 0.0357 \leftarrow (0.19 - 0.00)^2 + \\ 0.0138 \leftarrow (0.88 - 1.00)^2 + \\ 0.5242 \leftarrow (0.72 - 0.00)^2 + \\ 0.0001 \leftarrow (0.01 - 0.00)^2 + \\ 0.4079 \leftarrow (0.64 - 0.00)^2 + \\ 0.7388 \leftarrow (0.86 - 0.00)^2 + \\ 0.9817 \leftarrow (0.99 - 0.00)^2 + \\ 0.3998 \leftarrow (0.63 - 0.00)^2 \end{array} \right.$$

Calculate the **average cost** of all the training data

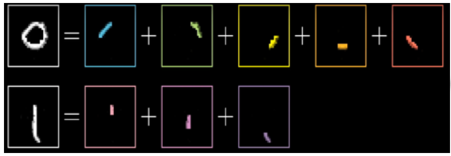


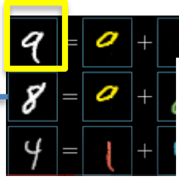
Keep some data for the testing phase

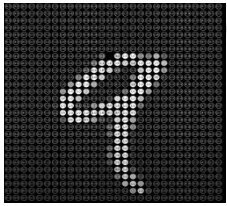
MNIST database contains

- 60,000 training images
- 10,000 testing images

Recap



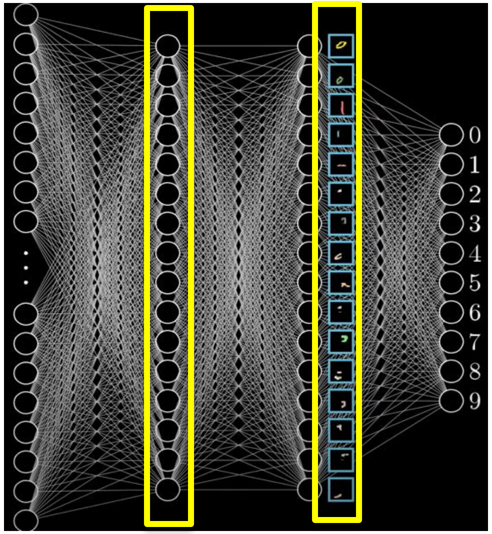




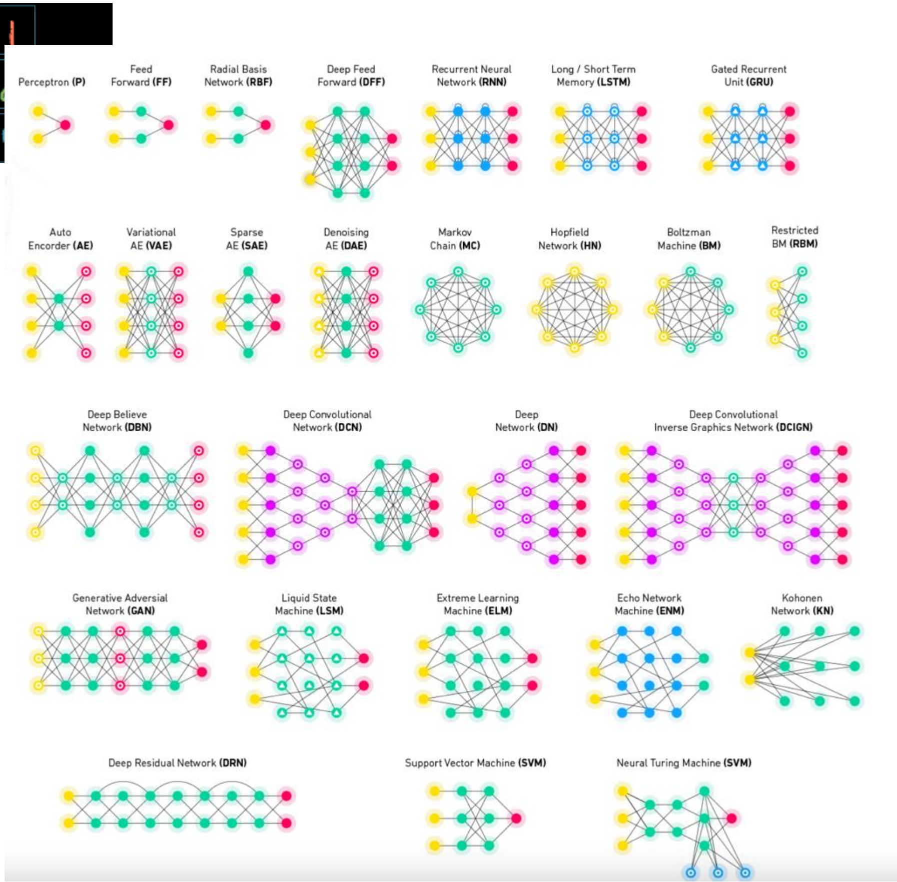
28

28

784



Multilayer Perceptron


Perceptron (P)
Feed Forward (FF)
Radial Basis Network (RBF)
Deep Feed Forward (DFF)
Recurrent Neural Network (RNN)
Long / Short Term Memory (LSTM)
Gated Recurrent Unit (GRU)
Auto Encorder (AE)
Variational AE (VAE)
Sparse AE (SAE)
Denoising AE (DAE)
Markov Chain (MC)
Hopfield Network (HN)
Boltzman Machine (BM)
Restricted BM (RBM)
Deep Believe Network (DBN)
Deep Convolutional Network (DCN)
Deep Network (DN)
Deep Convolutional Inverse Graphics Network (DCIGN)
Generative Adversial Network (GAN)
Liquid State Machine (LSM)
Extreme Learning Machine (ELM)
Echo Network Machine (ENM)
Kohonen Network (KN)
Deep Residual Network (DRN)
Support Vector Machine (SVM)
Neural Turing Machine (SVM)

Old technology

Convolutional NN
LSTM

3Blue1Brown series S3 • E1 But what is a neural network? | Chapter 1,2, Deep learning

Content

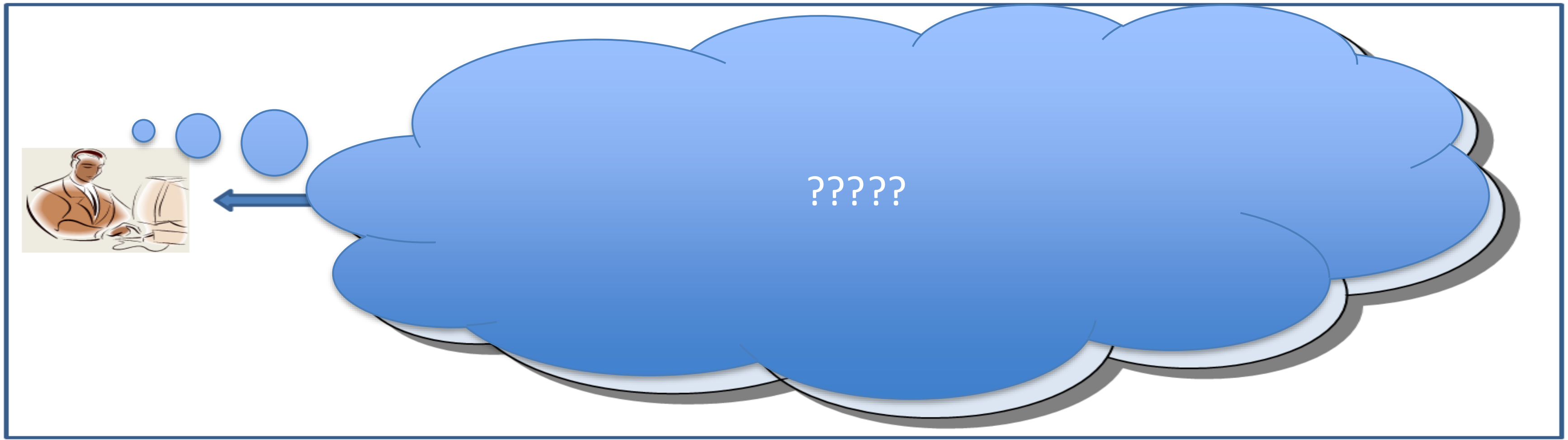
- Why we need Supercomputers ?
  - Big Data
- Supercomputers for every one
  - Cloud systems
- AI a different approach to programming
  - Supervised/Unsupervised/Reinforcement Learning
  - Deep Learning
  - Limits and Challenges

- Examples
  - AWS Amazon
  - regional sea-level changes (caused by climate change)
  - GÉANT Open Cloud eXchange (gOCX)

Amazon Simple Queue Service

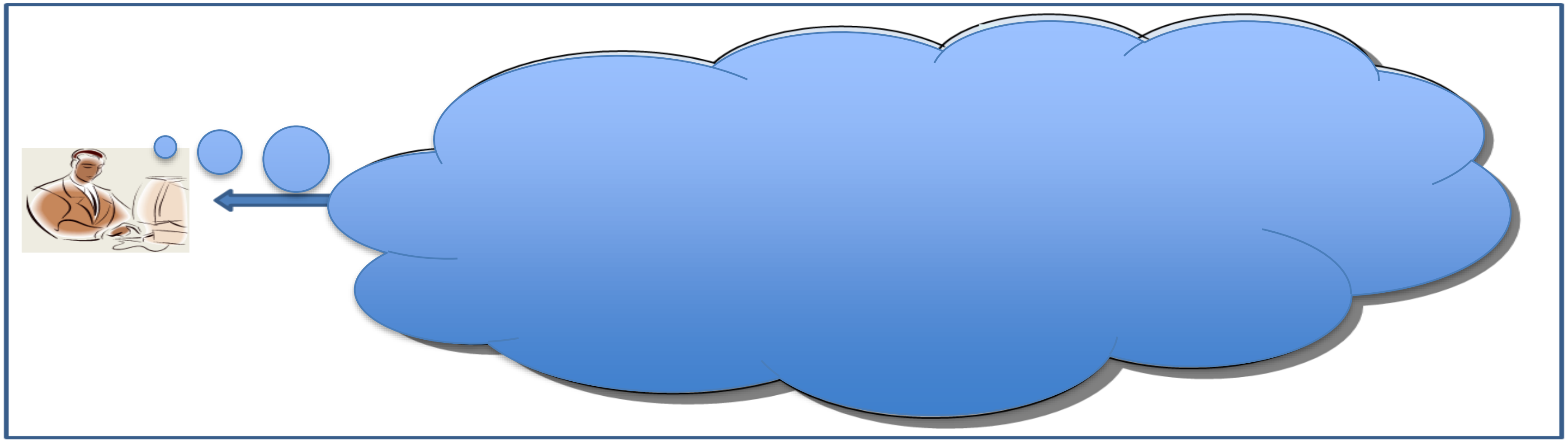
online photo processing service

- Requirements
  - Upload a few or hundreds of photos
  - specify the tasks to be performed on the photos
  - Processing might take a few second or several minutes



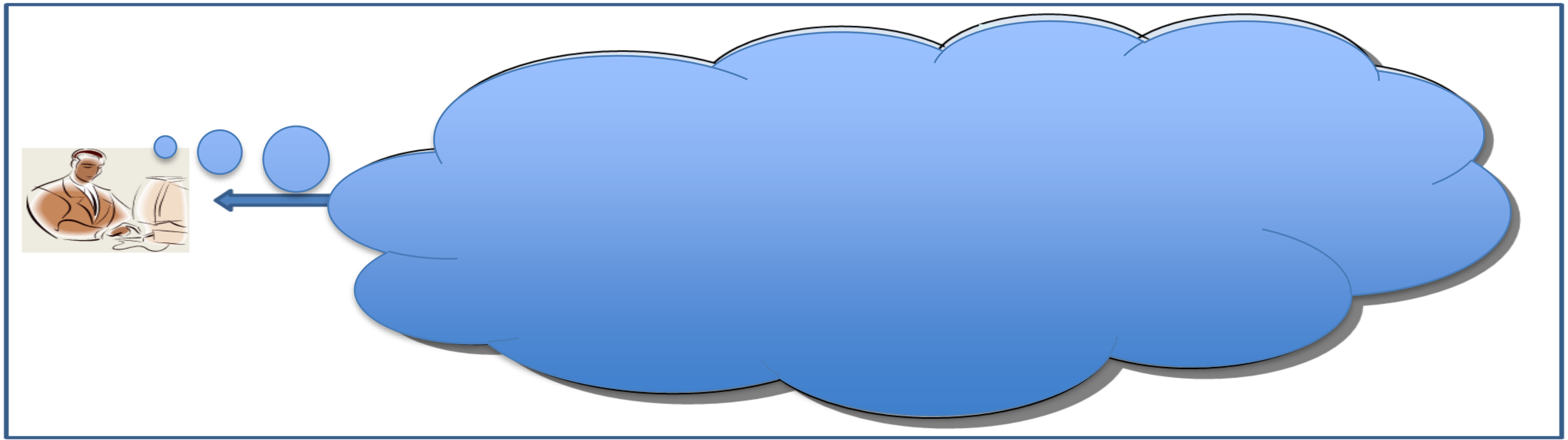
What can go wrong with online photo processing service?

- Photo **Processing Server crashes**
  - **With** SQS failure is transparent to the end user.
  - Users can continue to upload photos to the web site
  - Web server continues to send messages to the SQS Request queue



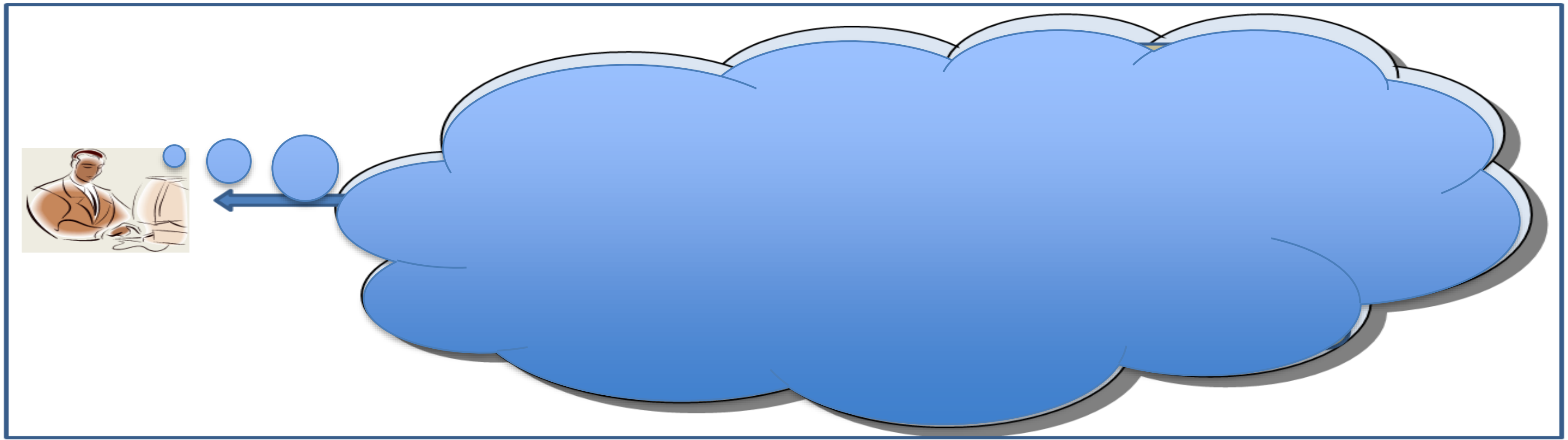
What can go wrong with online photo processing service?

- Photo Processing Server **cannot be restarted**.
  - SQS makes it possible to just drop in a replacement server.



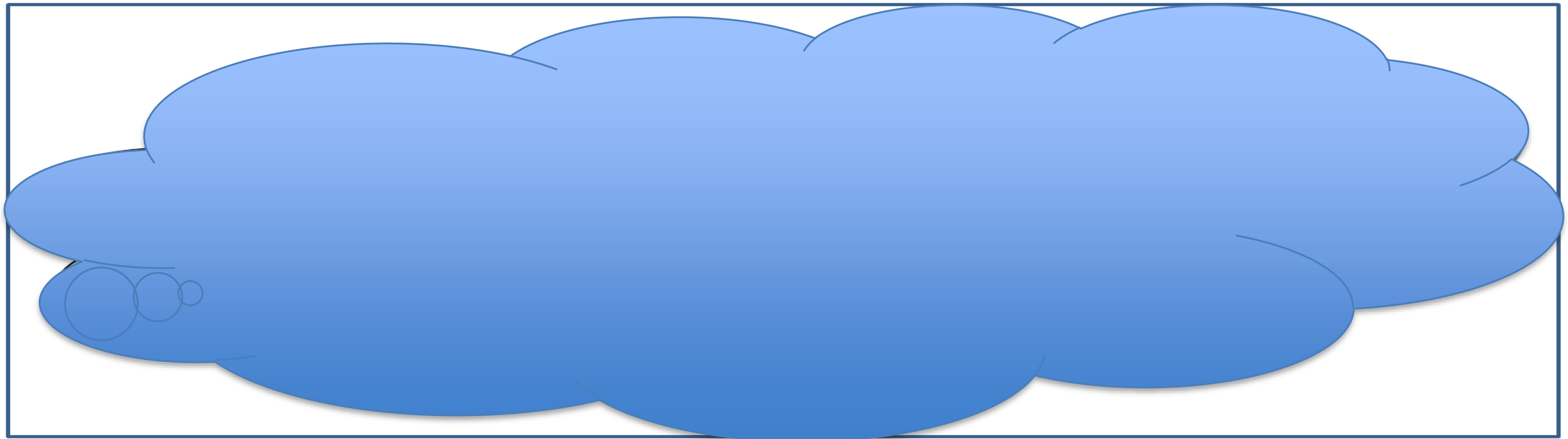
What can go wrong with online photo processing service?

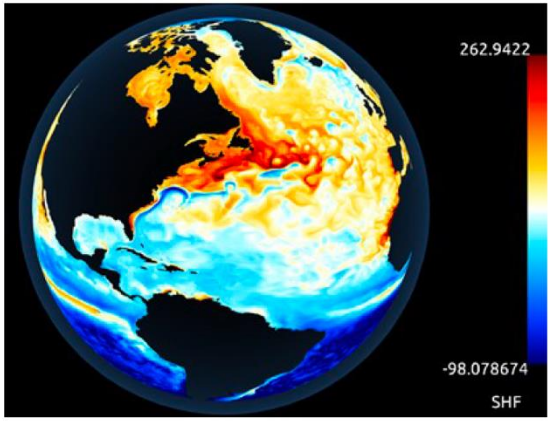
- Photo Processing Server is **overloaded**
  - A single SQS queue can be shared by multiple servers.
  - A server that is processing a message can prevent other servers from processing the same



A better design of the online photo processing service?

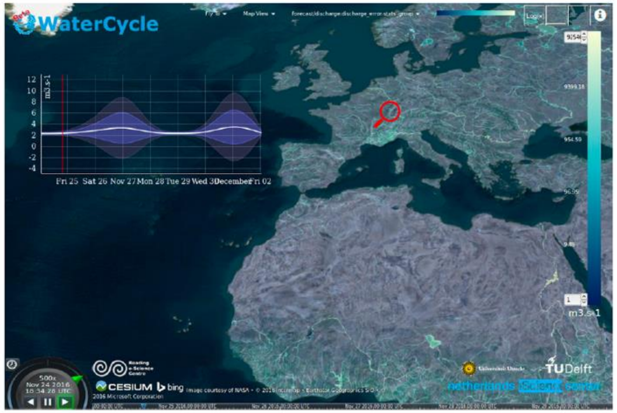
- If we know that **some** of the photo processing **operations take significantly longer** time than the rest,
  - you want to implement these longer-running operations in a separate, dedicated server.





eSalsa

Summer in the city

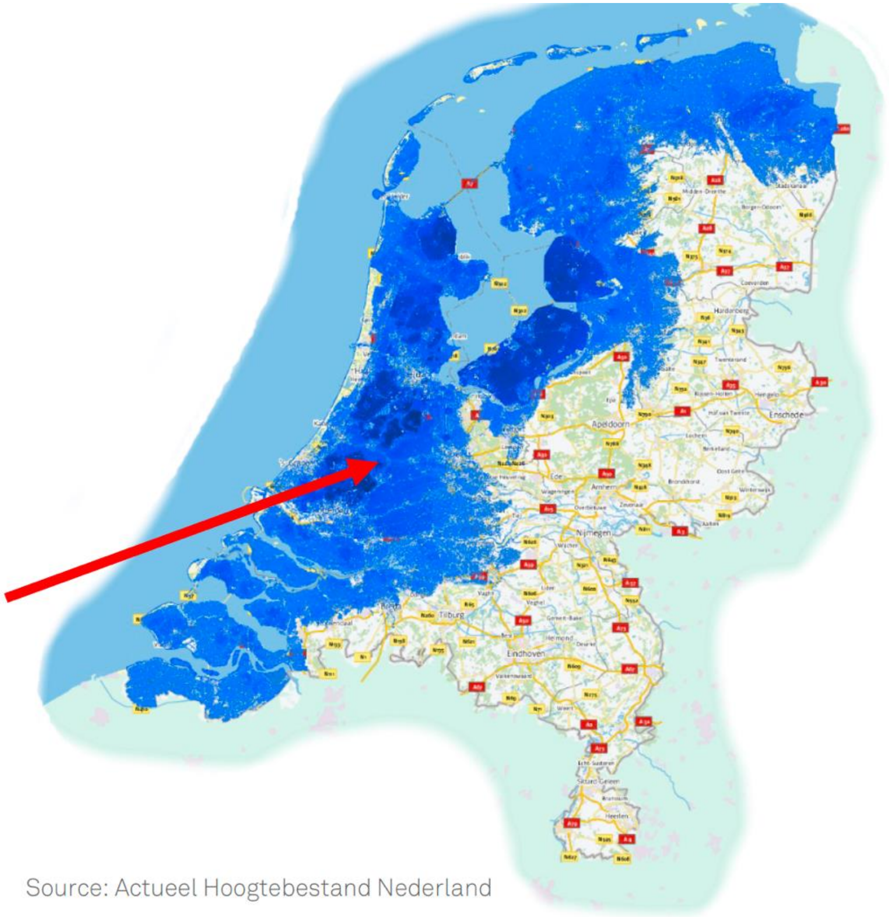


eWaterCycle

Many of our "traditional HPC" projects have a climate focus. They need to increase the resolution of their simulations, couple models, integrate observation data, after which they have trouble with load balancing or the large amounts of data they need to store

The eSalsa Project

Gain insight into **regional** sea-level changes (caused by climate change) by simulating the oceans with an unprecedented level of detail.



26% to 55% below sea level

Source: Actueel Hoogtebestand Nederland

IMAU Institute for Marine and Atmospheric research Utrecht

Universiteit Utrecht

VU UNIVERSITY AMSTERDAM

COMMIT/

netherlands eScience center

Sea levels are changing...

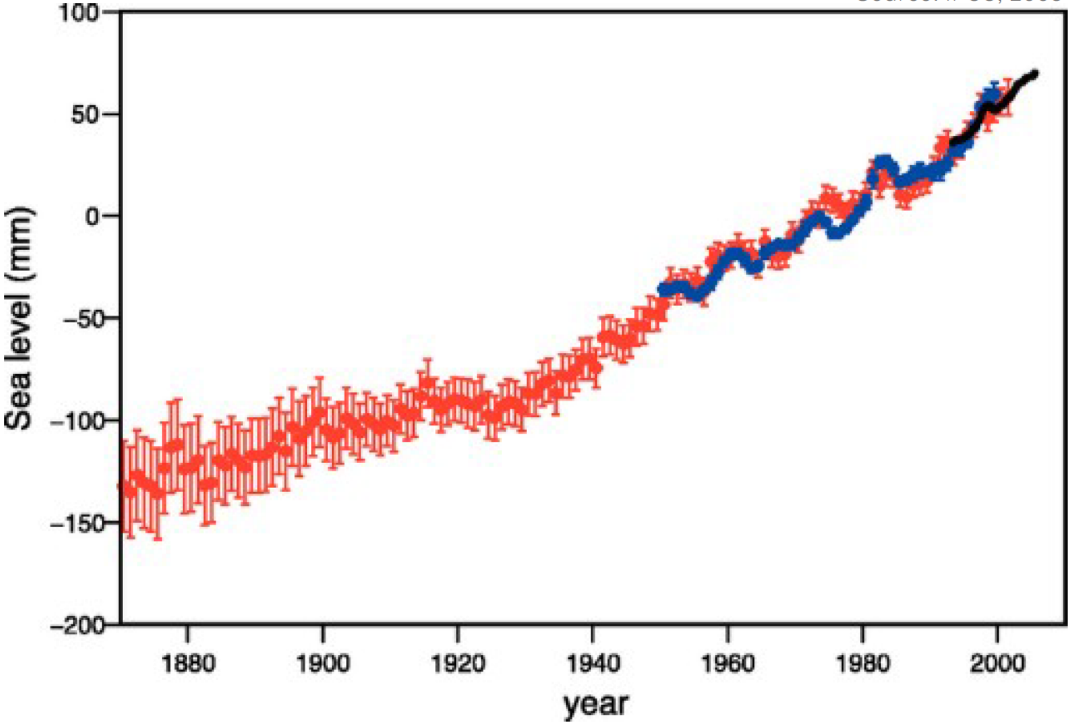
red -- historical records

blue -- tidal gauges

black -- satellite observations

All data show an upward trend!

Source: IPCC, 2003

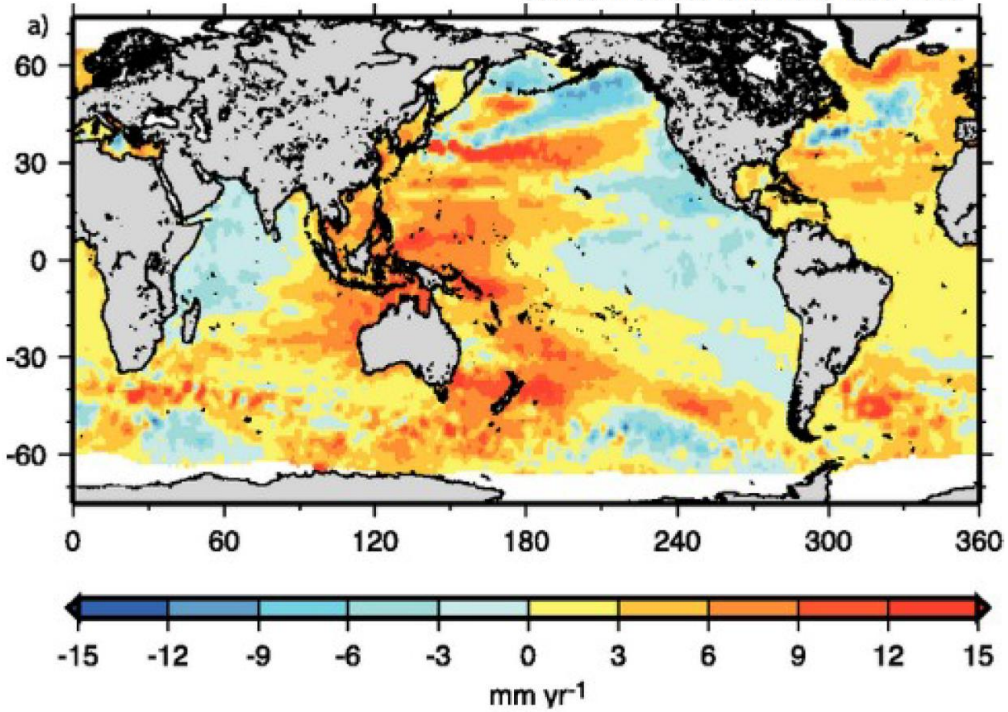


Junk Email

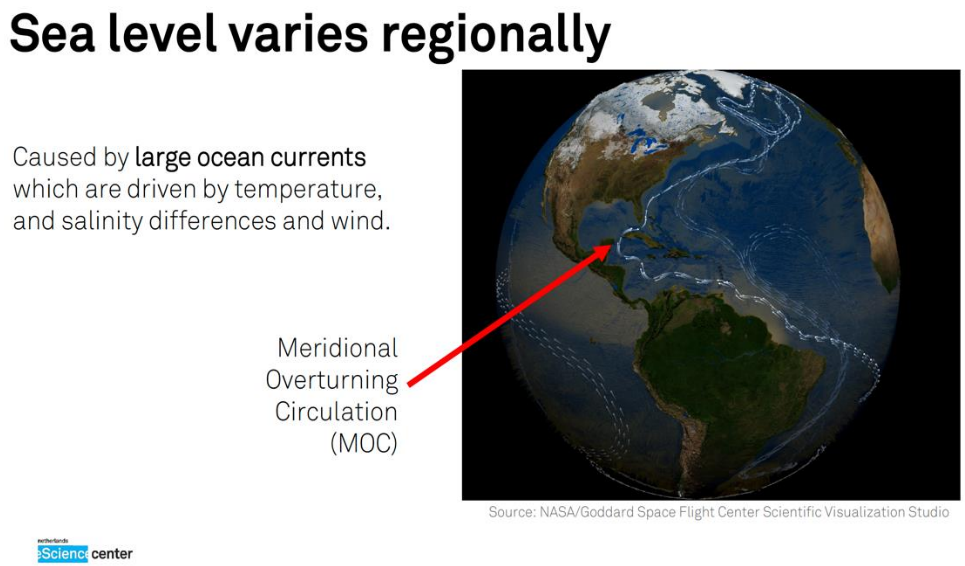
…but the change is not uniform!

Satellite observations show large regional variations in sea-level change.

Source: IPCC Sea level trend 1993-2003



Oceanography in one Slide …

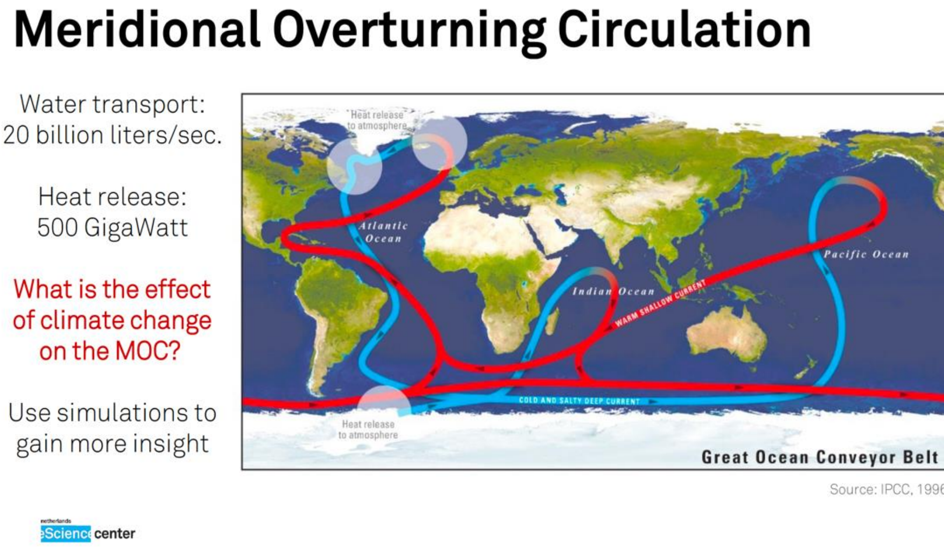


Sea level varies regionally

Caused by **large ocean currents** which are driven by temperature, and salinity differences and wind.

Meridional Overturning Circulation (MOC)

Source: NASA/Goddard Space Flight Center Scientific Visualization Studio



Meridional Overturning Circulation

Water transport: 20 billion liters/sec.

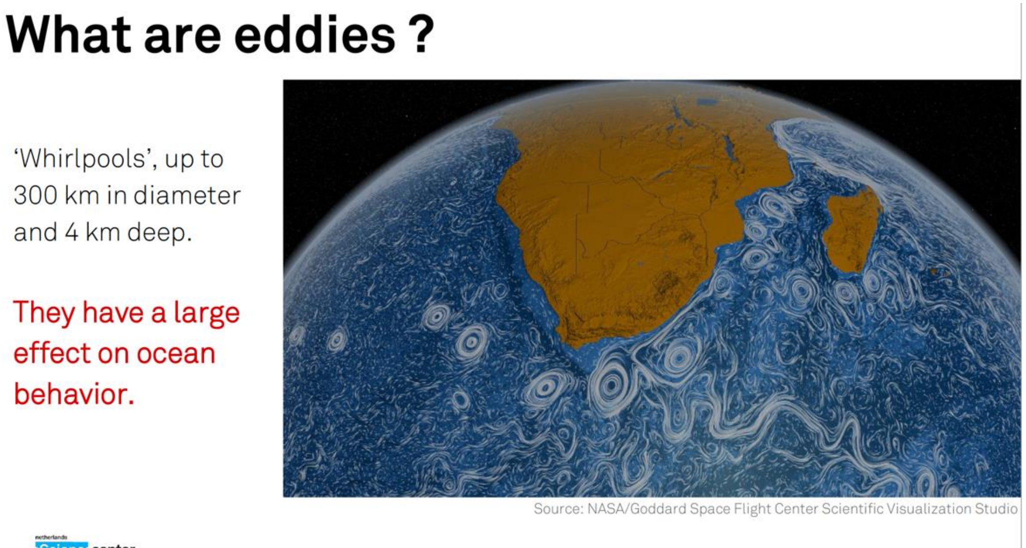
Heat release: 500 GigaWatt

What is the effect of climate change on the MOC?

Use simulations to gain more insight

Heat release to atmosphere
Atlantic Ocean
Pacific Ocean
Indian Ocean
WARM SHALLOW CURRENT
COLD AND SALTY DEEP CURRENT
Heat release to atmosphere
Great Ocean Conveyor Belt

Source: IPCC, 1996

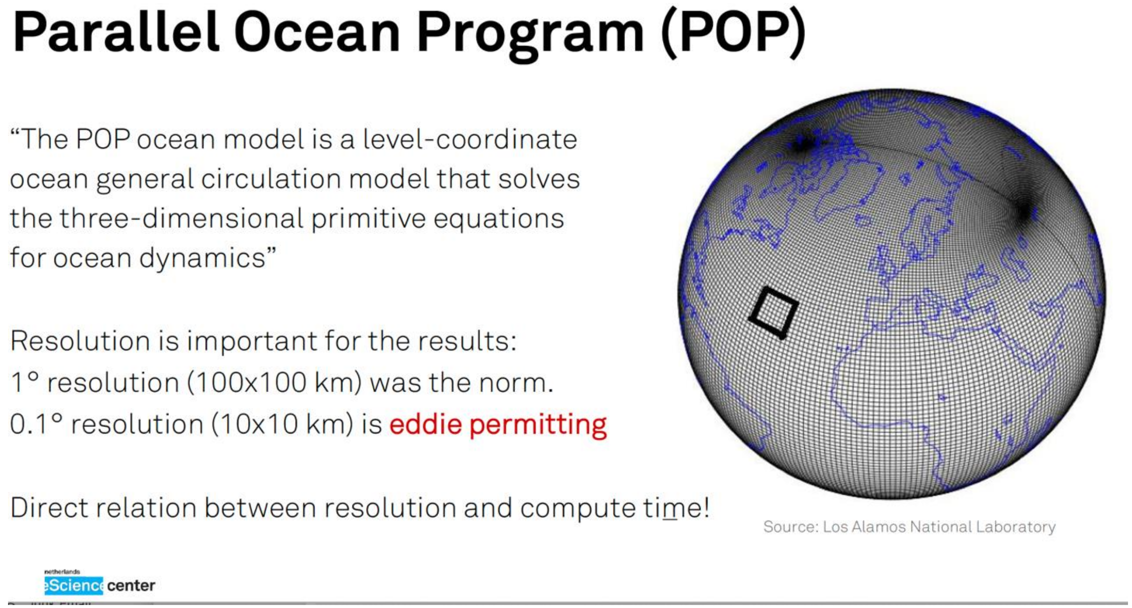


What are eddies ?

'Whirlpools', up to 300 km in diameter and 4 km deep.

They have a large effect on ocean behavior.

Source: NASA/Goddard Space Flight Center Scientific Visualization Studio



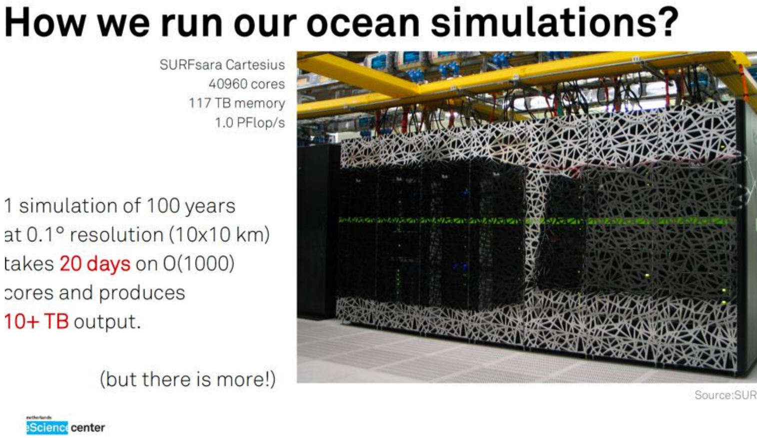
Parallel Ocean Program (POP)

"The POP ocean model is a level-coordinate ocean general circulation model that solves the three-dimensional primitive equations for ocean dynamics"

Resolution is important for the results:
1° resolution (100x100 km) was the norm.
0.1° resolution (10x10 km) is eddie permitting

Direct relation between resolution and compute time!

Source: Los Alamos National Laboratory



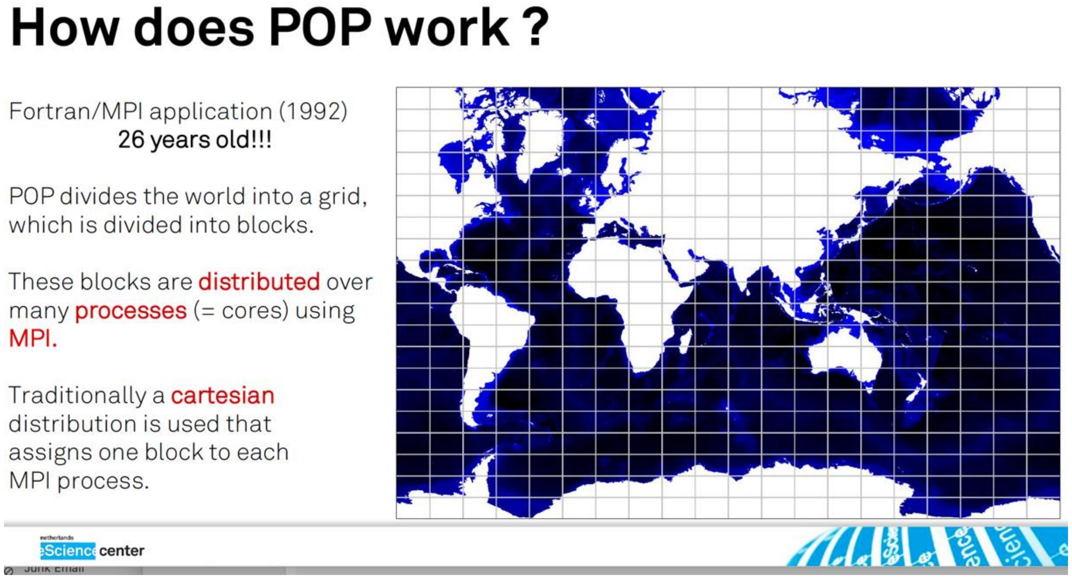
How we run our ocean simulations?

SURFsara Cartesius
40960 cores
117 TB memory
1.0 PFlop/s

1 simulation of 100 years
at 0.1° resolution (10x10 km)
takes 20 days on O(1000)
cores and produces
10+ TB output.

(but there is more!)

Source:SUR



How does POP work ?

Fortran/MPI application (1992)
26 years old!!!

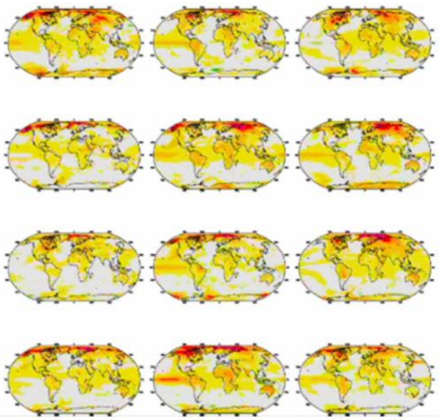
POP divides the world into a grid, which is divided into blocks.

These blocks are distributed over many processes (= cores) using MPI.

Traditionally a cartesian distribution is used that assigns one block to each MPI process.

Ensembles

We don't run 1 simulation but an ensemble of 16, each using a slightly different forcing.

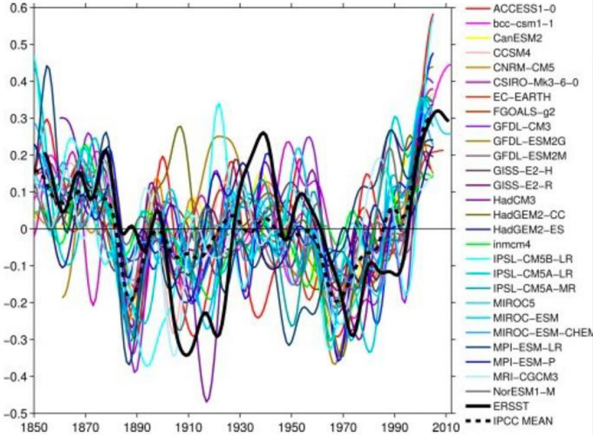


16x increase in compute time

Why ensembles?

Climate is a chaotic system: a small change in forcing, model or starting conditions may change the outcome significantly.

By running many simulations and/or different models, we get many results and do statistics on them to determine the certainty of the results.



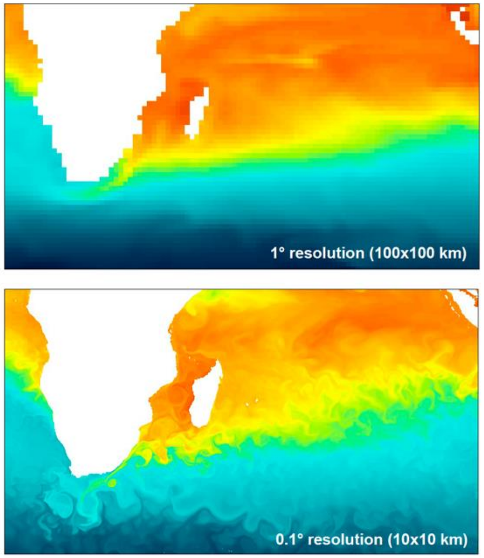
Source: L. Zhang, C. Wang, DOI: 10.1002/jgrc.20390

Higher Resolution

0.1° resolution (10x10 km) is only the start! We want to increase the model resolution even further to get more detailed results.

Ultimate goal (last time I asked):
0.01° resolution (1x1 km)
(fully eddie resolving)

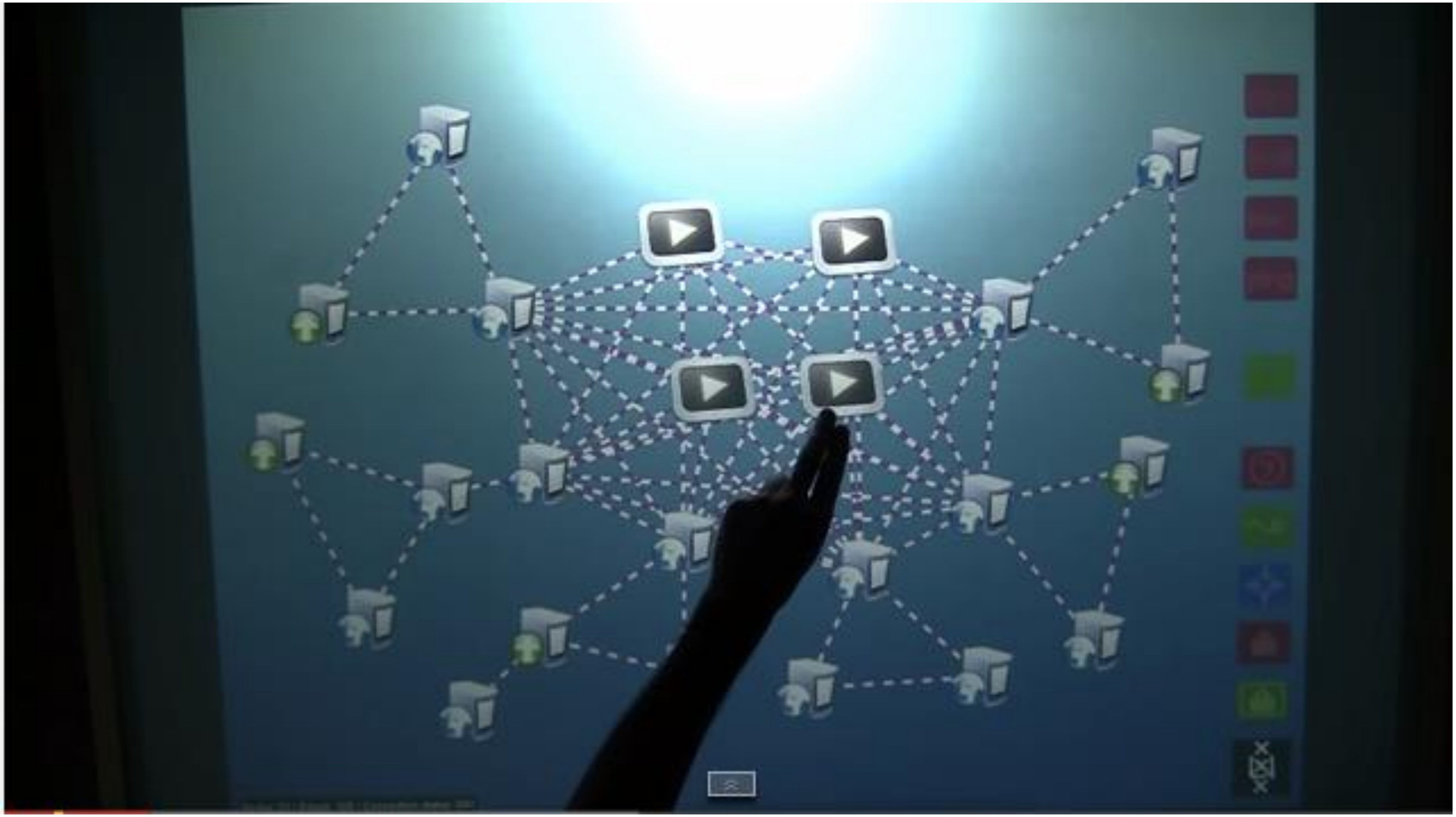
100x increase in compute time!



Source: eSalsa results

Interactive Networks: creation of the virtual network in which the video streams can be manipulated

Super Computing 2011, Seattle , WA



Video available on youtube →https://www.youtube.com/watch?v=nGljMqqCUVA

GÉANT Open Cloud eXchange (gOCX)

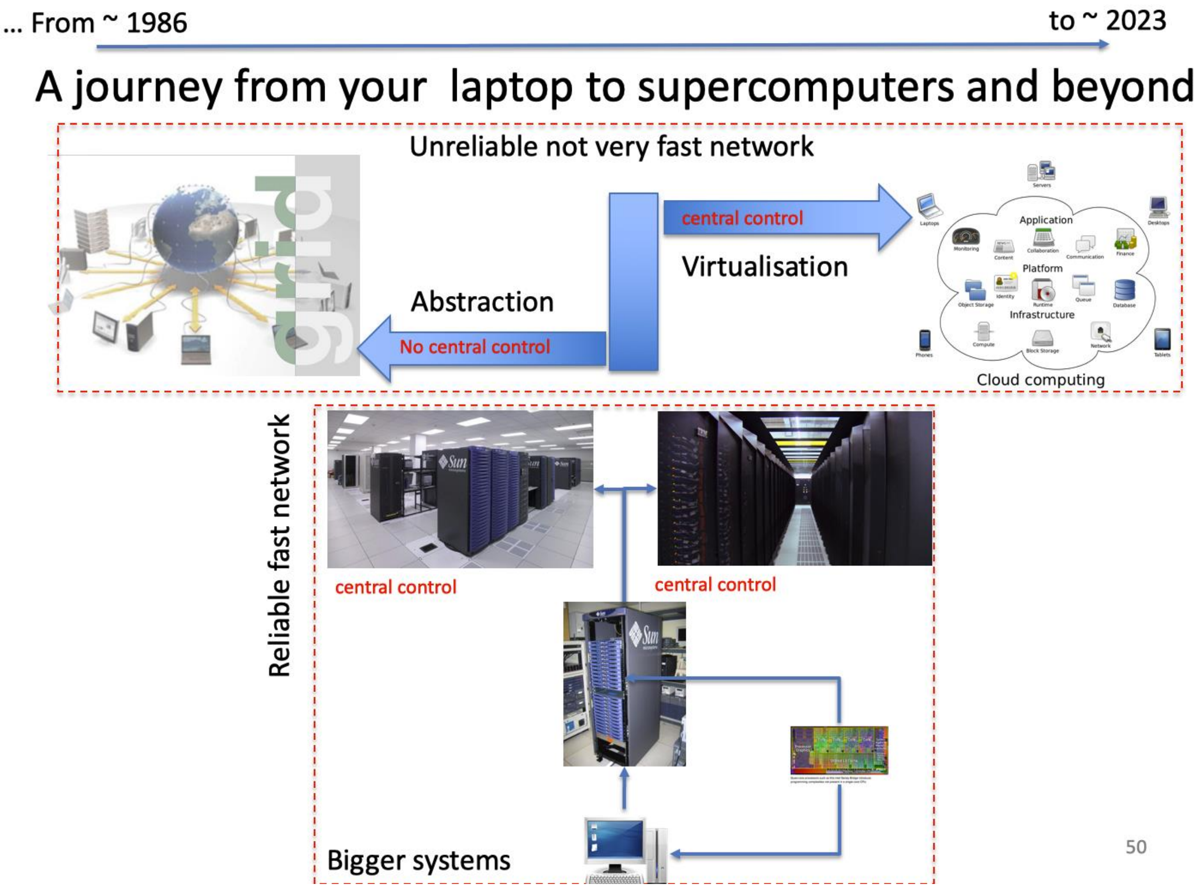
GÉANT tv, Augut 2014,

Video available on YouTube → https://www.youtube.com/watch?v=q7IAAFUcTY0

Other demos around Data management

- policy Auditing in **Data Exchange** Systems.
  - https://dl4ld.nl/2021-02-10/ICT-demo-Xin.mp4

- User Friendly **Data Transfers** with DTNs.
  - https://delaat.net/sc/sc19/demo02/movie-s.m4v



... From ~ 1986 to ~ 2023

A journey from your laptop to supercomputers and beyond

Unreliable not very fast network

central control

Virtualisation

Abstraction

No central control

Cloud computing

Reliable fast network

central control

central control

Bigger systems

50

MORE INFORMATION

1. Email: A.S.Z.Belloum@uva.nl
2. Web page: https://ivi.fnwi.uva.nl/sne/wsvlam2/
3. Demos: https://youtube.com/playlist?list=PLCEhEFHyv3IjGJlIXfIV4OpB4uLH4lm7f