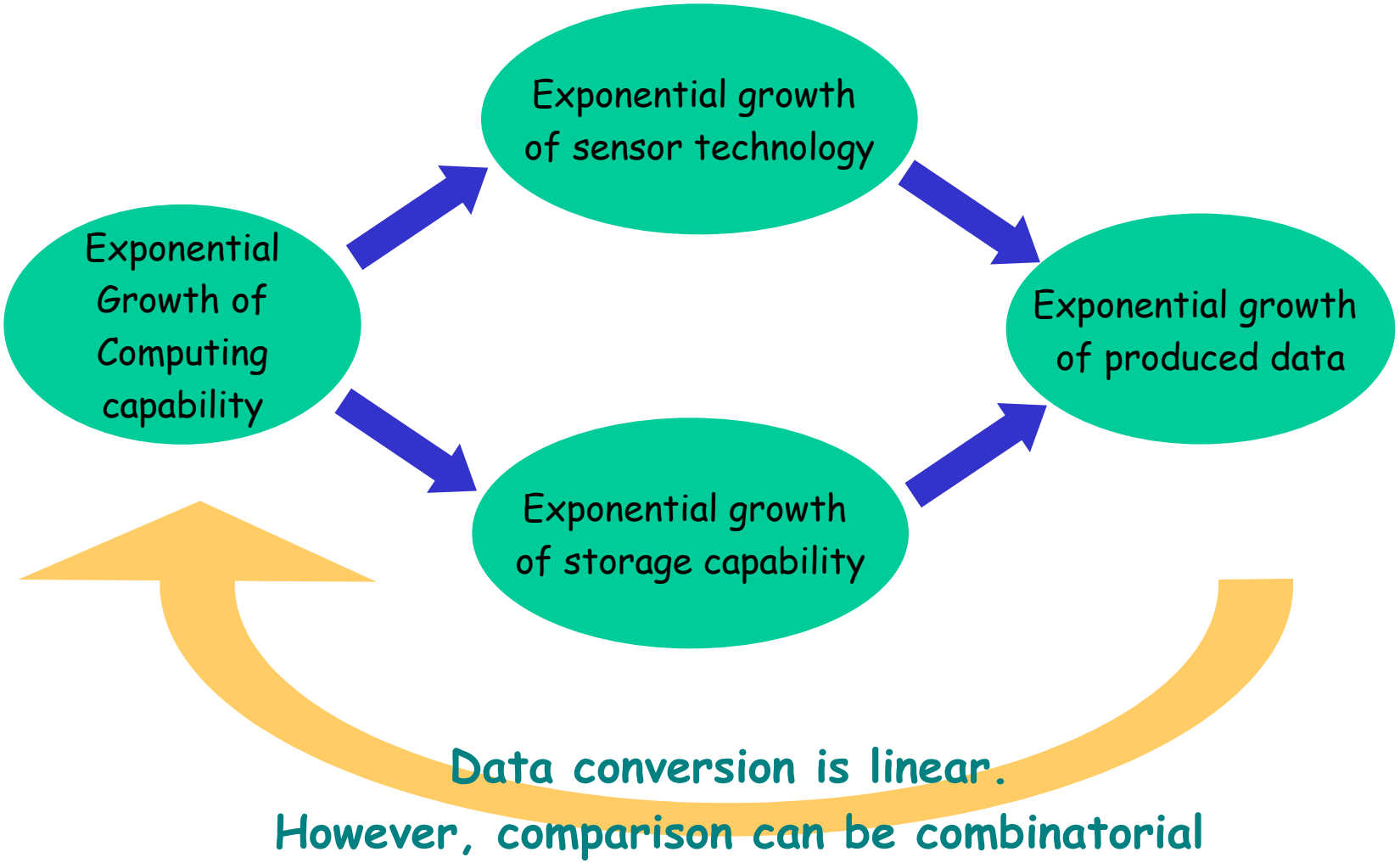
Data deluge
(and its applications)

Gianluigi Zanetti

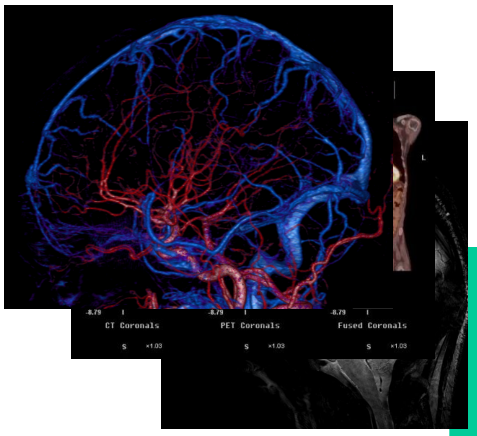
Prologue

- Data is becoming cheaper and cheaper to produce and store
- Driving mechanism is parallelism on sensors, storage, computing
- Data directly produced are complex objects
  - lots of implicit information
  - could be non trivial to extract

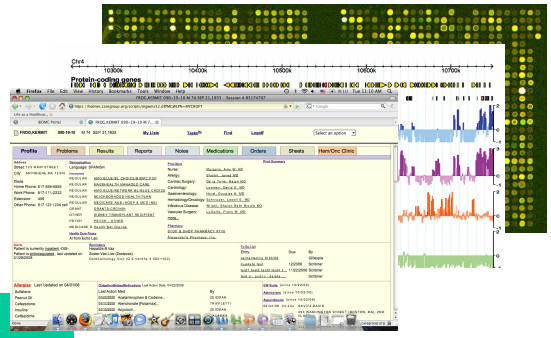
Drivers of exponential data growth



Clinical institutions repository of massive amounts of heterogeneous and complex data

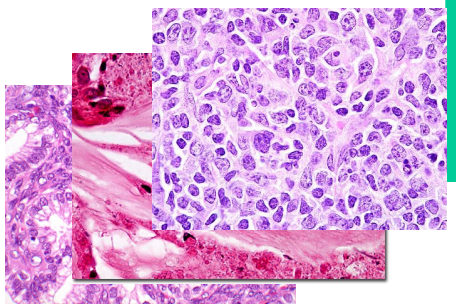


Imaging modalities
(3D/4D high res datasets)

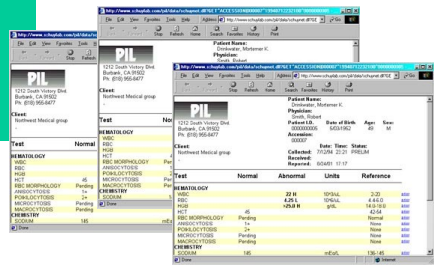


Clinical -omics
(multiple data objects)

Electronic Health Record
(patient centric aggregation)



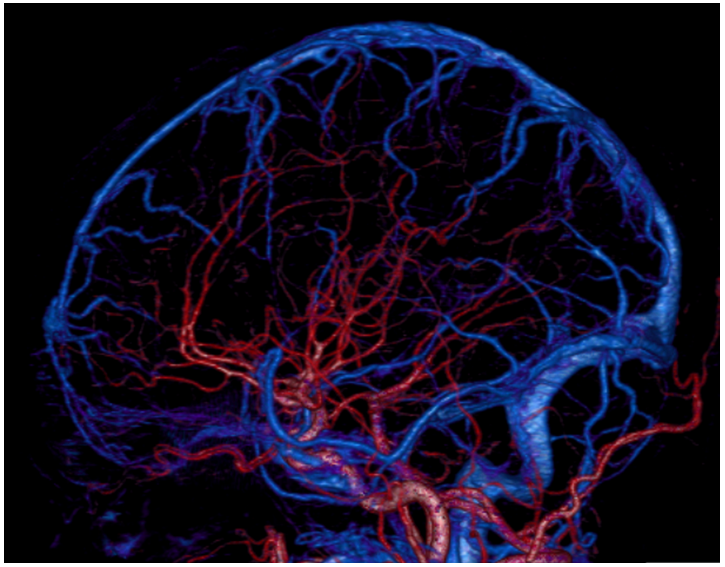
Digital pathology
(ultra high res images)

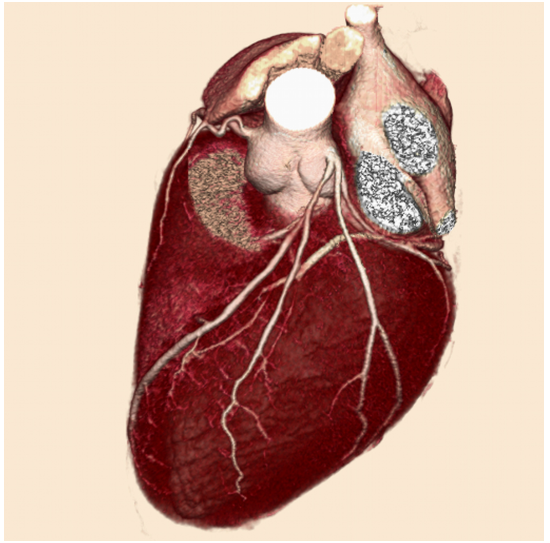


Lab exam

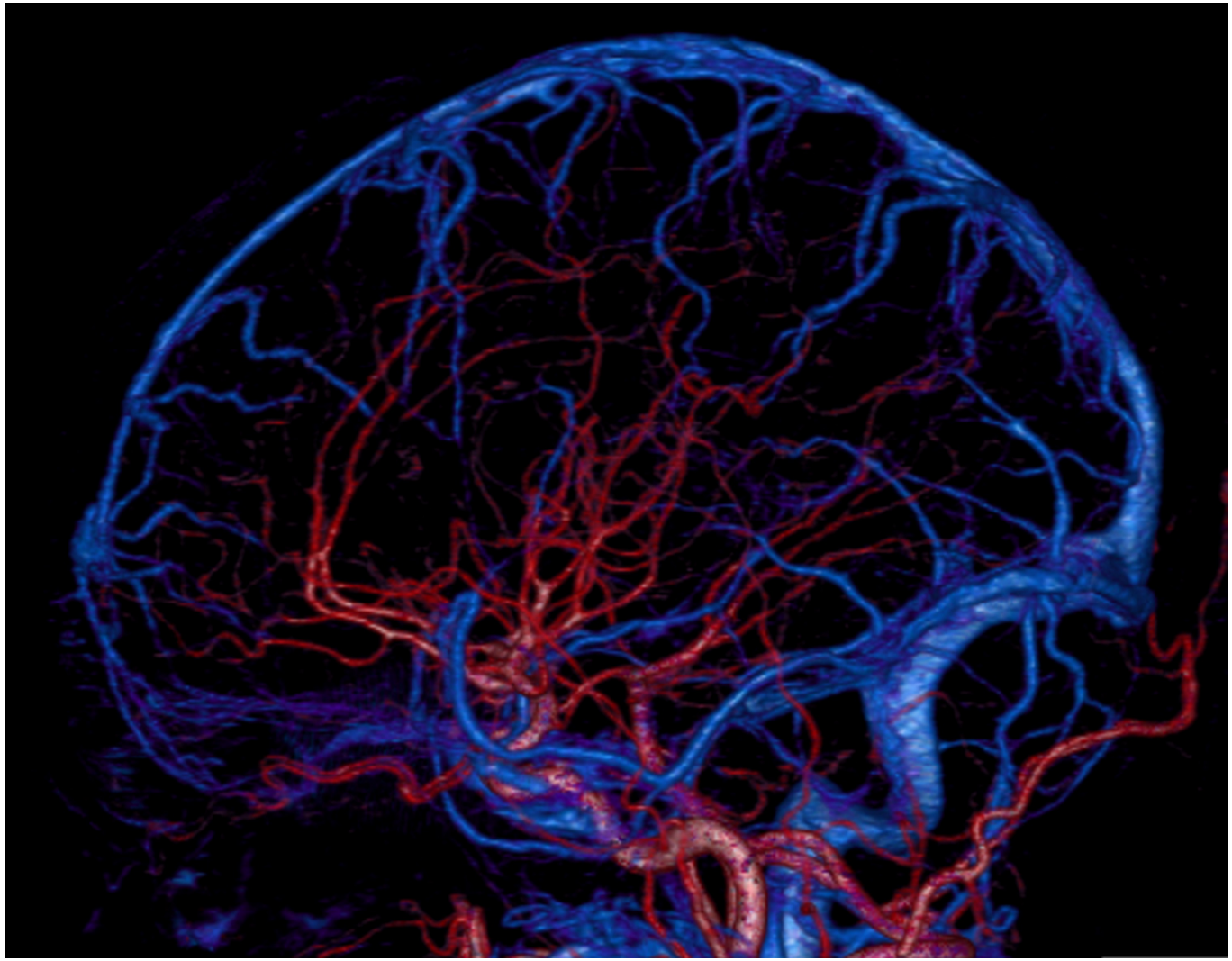
Drivers of data growth
Next generation imaging modalities

- Multi slice CT (Toshiba AquilionOne)
  - 320 detector rows (0.5 mm in width), 16 cm single
  - ~180 msec temporal resolution
  - (for a 64-CT, only 3.2 centimeters, up to 10 seconds.)



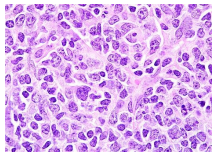


A computable picture

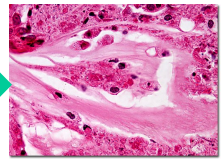


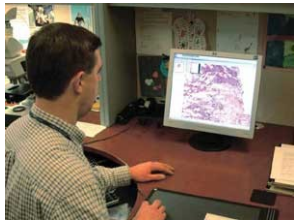
Drivers of data growth
Next generation imaging modalities

Robotized scanners
120-200 slides tray
(overnight scanning)

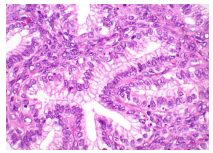


Annotation, sharing,
integration





telepathology

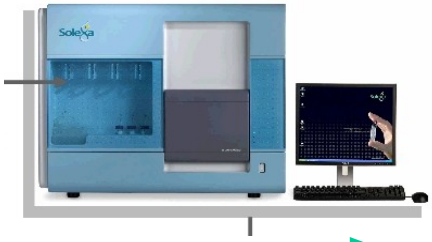


Computer aided analysis

Typical size ~ 50GB per case
Can reach 1TB
(40x 5 focal planes)
[cheating:
jpeg compression ~ 30:1]

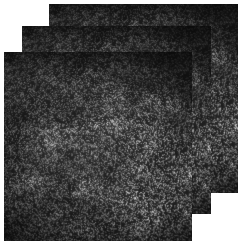
High throughput genomic
Parallel sensors

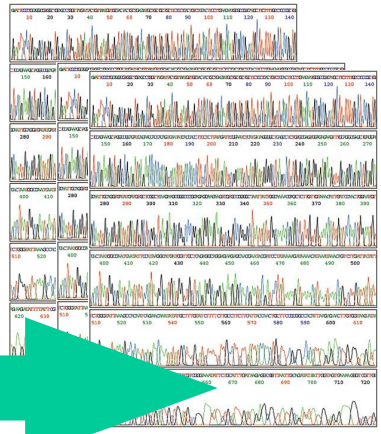
Cluster Growth



3Tbytes per week

From analog to digital





From images to sequences

Wal-Mart & its data

"We know how many 2.4-ounce tubes of toothpaste sold yesterday, and what was sold with them. Our database grows because we capture data on every item, for every customer, for every store, every day"

Dan Phillips, VP of IS, Wal-Mart [From InformationWeek, Jan. 2006]

~3600 U.S. stores and ~500 TB

Wal-Mart keeps watching and react very quickly…

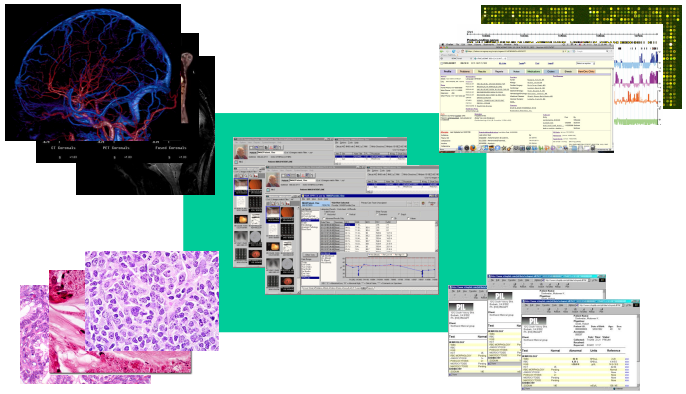
Two approaches to knowledge extraction

- Hypothesis driven
  - Verify if the data available support my idea
- Exploratory (serendipity)
  - Check if it is possible to extract correlations from the data
- Different approaches, different techniques
  - Hypothesis driven
    - Statistical analysis
  - Exploratory analysis relies on massive data mining
    - Supervised and unsupervised learning/classification
- Distributed, multidisciplinary data
  - rising rapidly and correlations needed

Clinical institutions repository of massive amounts of heterogeneous and complex data

Humans

Dealing with patients



Wall Mart is small, and its data is trivial.

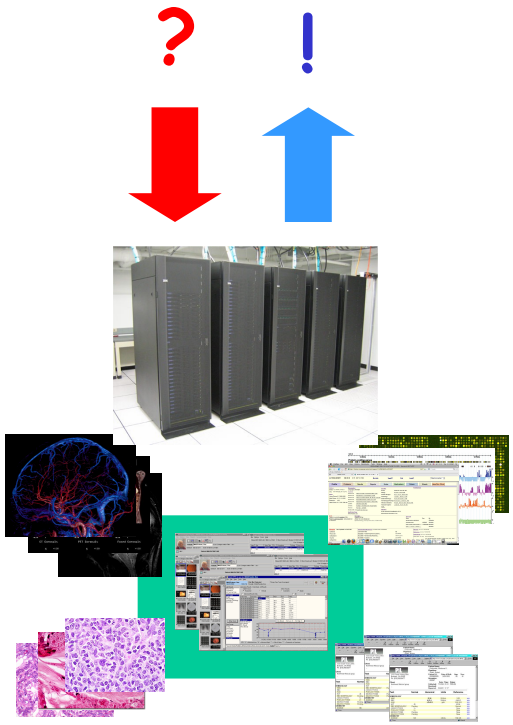
Machines

Real time control, CAD traning, and exploratory analysis, …

Examples

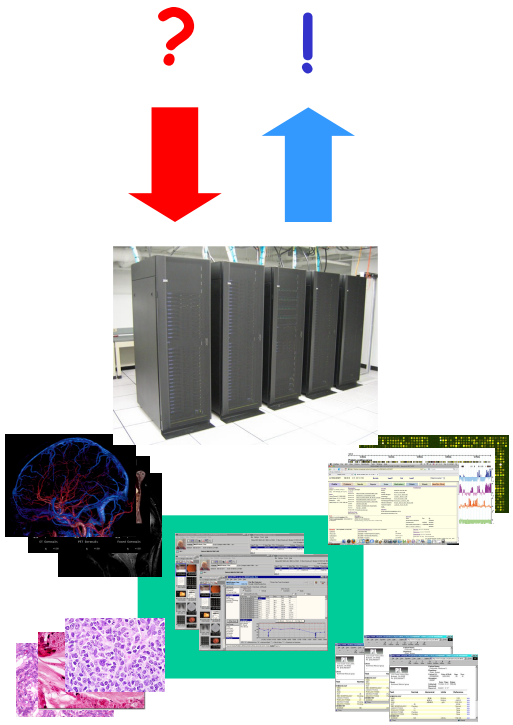
- Training a CAD system
  - Extract a set of digital mammography images of breast lesions that have been confirmed by pathology to be a tumor
  - Extract a set of control images
  - Teach a classifier system to recognize the difference.
- Check for 3-sigma micro-epidemiological fluctuations
  - Do real time gathering of, e.g., lab results and check if it is consistent with the 'expected' results

Data are lots and machines are stupid



- Moving data is becoming non trivial
  - Move computation to data
- Data should be described by computable meta-information
  - A recipe that a program can understand on how one could extract the desired data

Data are lots and machines are stupid



- Moving data is becoming non trivial
  - Move computation to data
- Data should be described by computable meta-information
  - A recipe that a program can understand on how one could extract the desired data

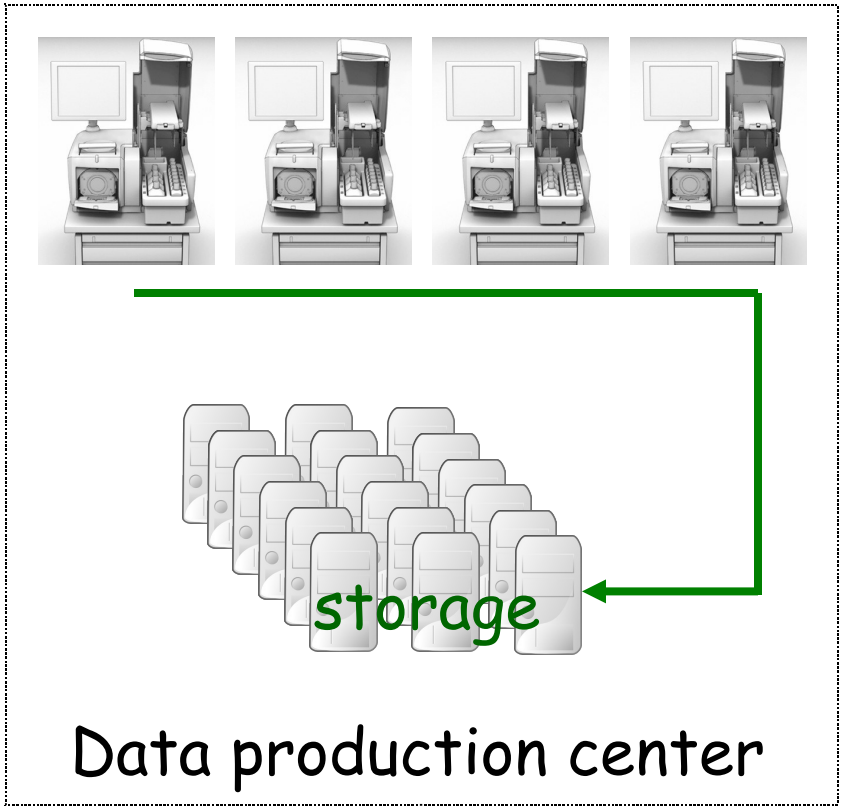
Moving data is becoming non-trivial

An example: Configurable computational facilities for parallel sequencing

- Large data accumulations in productions sites
  - E.g., Sanger Center can produce 300TB/month
- Deep, highly non-linear, information content
  - Complex data analysis based on whole dataset
  - Analysis method itself object of research
  - Multiple computational cultures & approaches

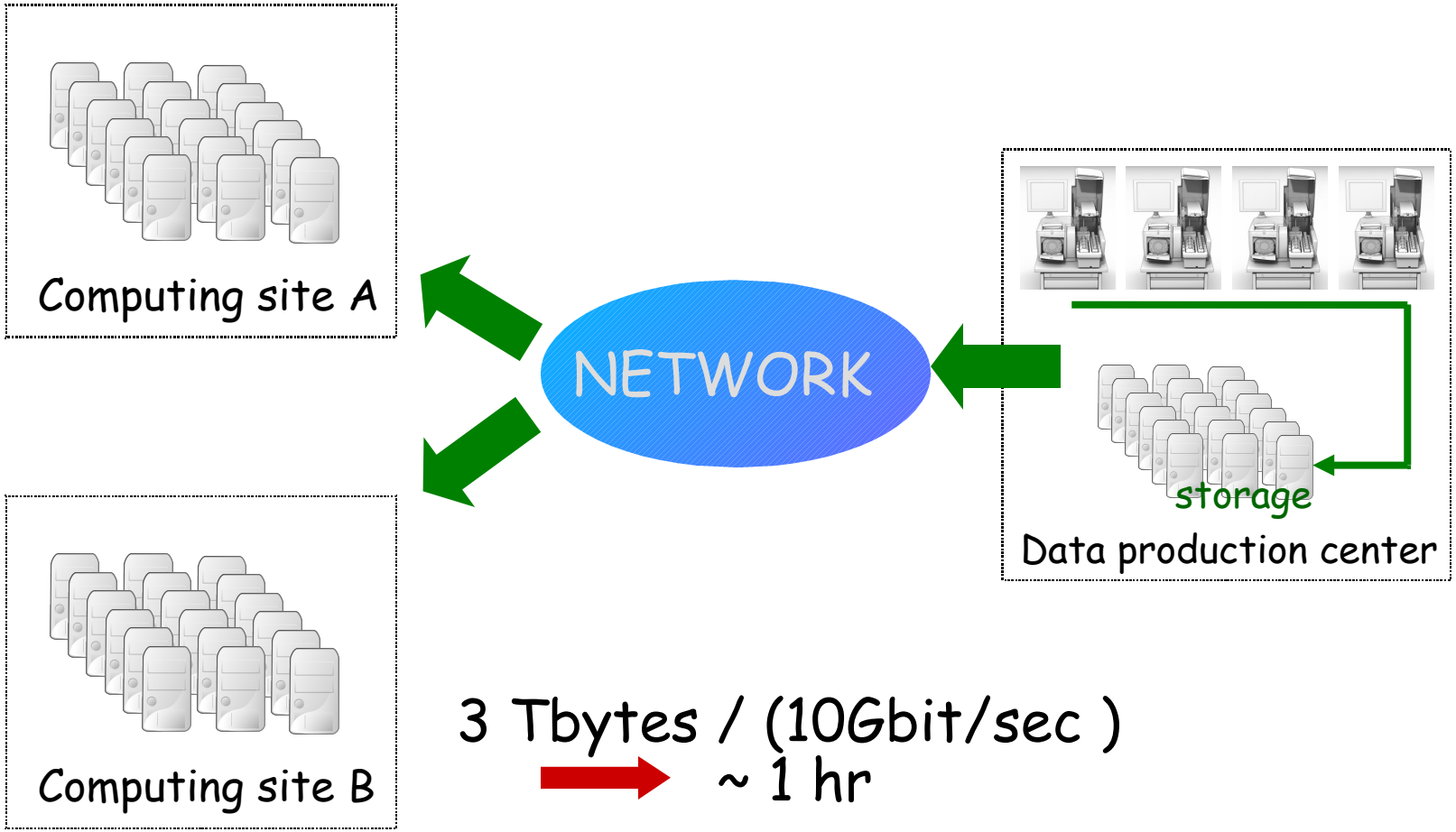
Example from current research

Configurable computational facilities for parallel sequencing



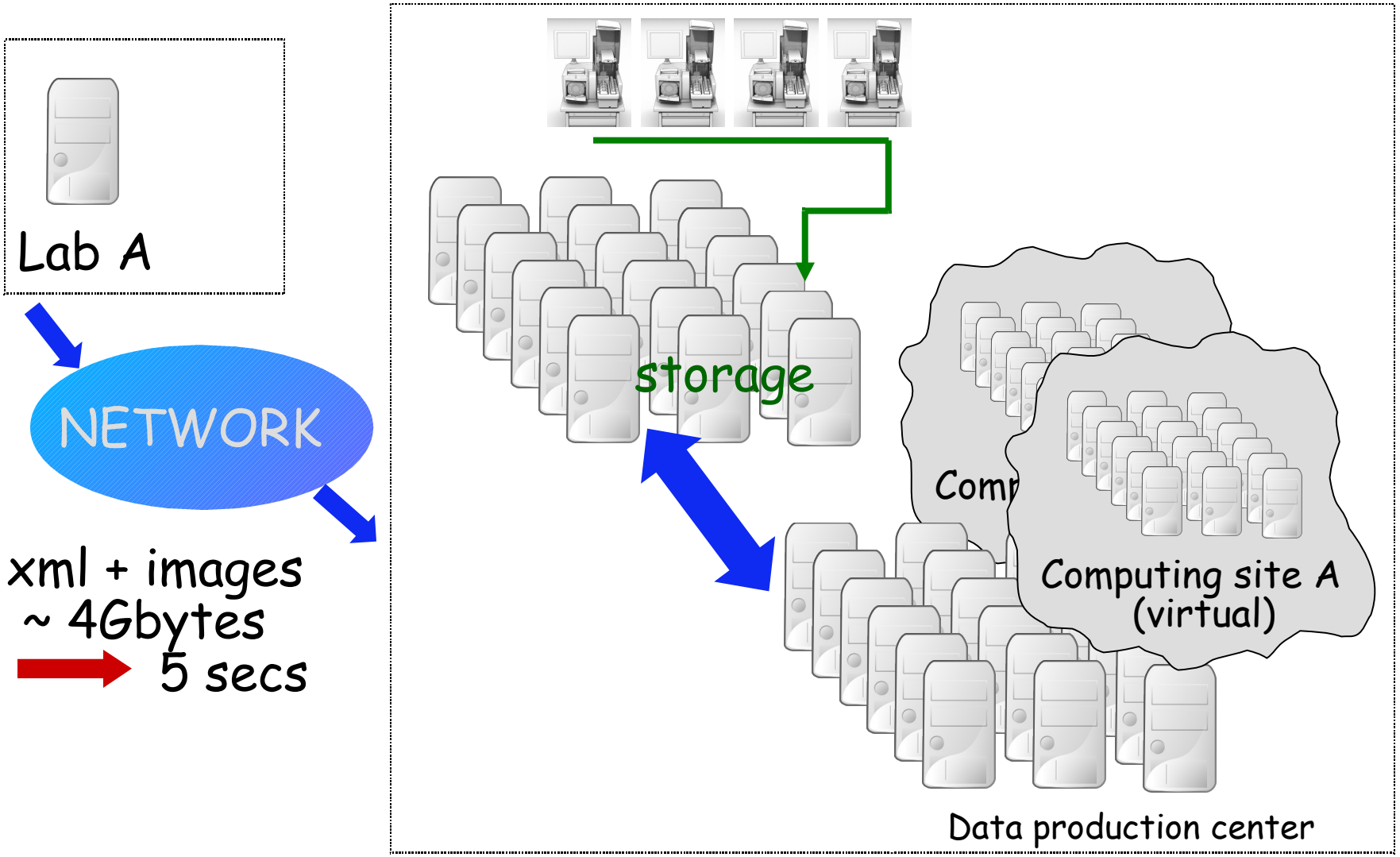
Example from current research

Configurable computational facilities for parallel sequencing

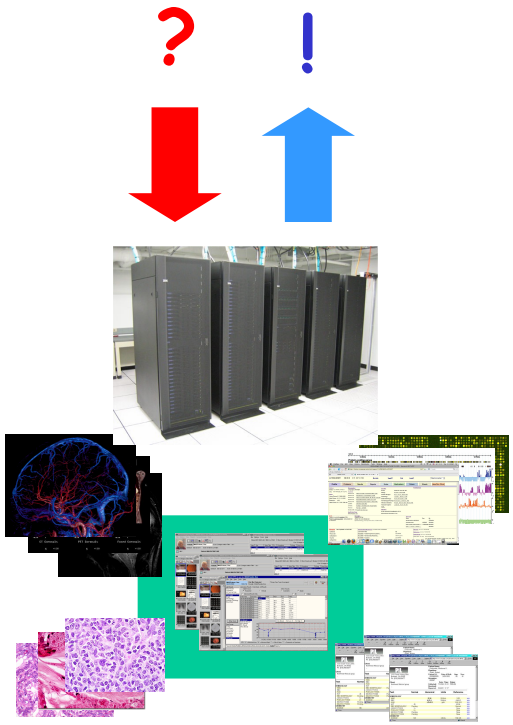


Example of current research

Configurable computational facilities for parallel sequencing



Data are lots and machines are stupid



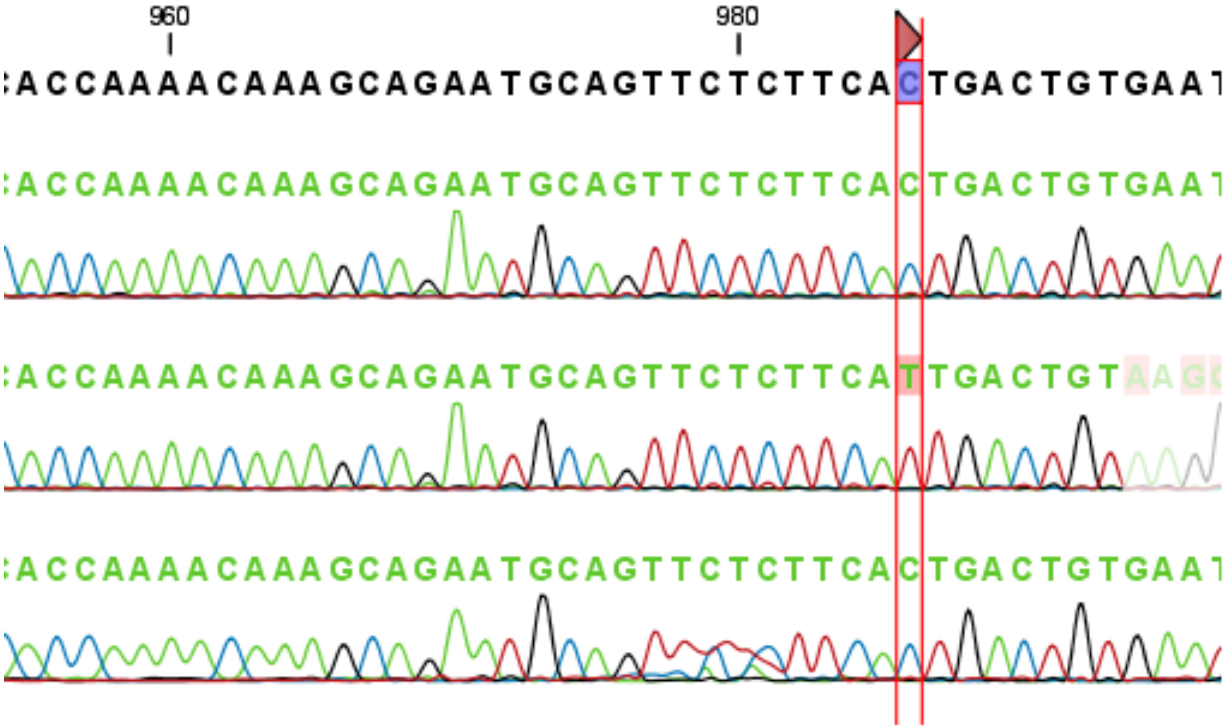
- Moving data is becoming non trivial
  - Move computation to data
- Data should be described by computable meta-information
  - A recipe that a program can understand on how one could extract the desired data

Example from current research

Whole genome association study and computable metadata

- Trying to correlate genotype signature to phenotype
  - Very high resolution genotyping (~ 1M snp on genome)
  - Scale of problem (thus far) order of thousand of individuals
- Heavy computational problem
  - Processing pipeline uses parallel computing on clusters

SNP genotyping



More of 4 Million SNP known

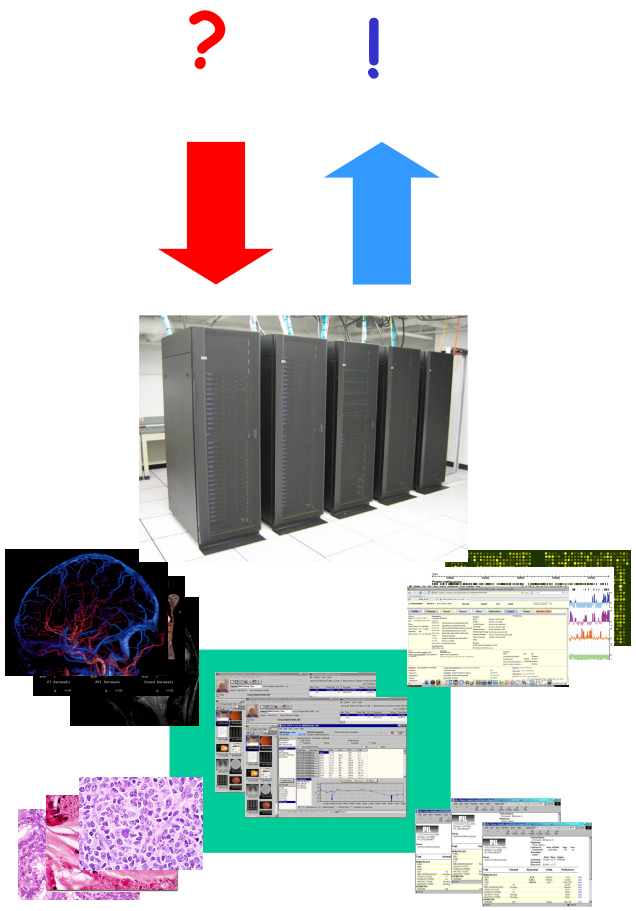
Example from current research

Whole genome association study and computable metadata

- Technically disgusting things in current pipelines:
  - Clinical information saved in (!) ugly excel files
  - Genotyping information saved in vendor specific format
  - Sizable amount of code: ad hoc data conversion
- Clinical information uglyness can be probably fixed by using modern data format with robust computational semantic
  - Portable querying
- Need something similar for datasets!
  - However, it should support heavy duty computing

Last slide…

- Clinical institutions are becoming repository of massive amounts of heterogeneous and complex data
- Wall-mart can explore its data and get out something useful, we should be able to do it to!
- However, it will be very difficult to do systematic data exploration unless there is a robust computable meta-description model for all the data involved




Thank you!

Drivers of exponential data growth: parallel sensors

- High throughput genomic
  - Possible consequence of 'low cost sequencing'
    - Plot curve of reduced cost
  - Applications ranging from profiling to genome-wide transcription
  - LOTS of data
  - Their main focus, however, is using next-gen sequencing for rare variant detection. The goal is to develop clinical assays for early detection of cancer from blood, saliva, stool, or urine samples by looking for somatic mutations carried by a small number of cancer cells. The ratio of mutant to wild type signal is presumably low at early cancer stages, but theoretically detectable with sufficient cancer coverage.

Evaluation of an Ultra-deep Sequencing Method to Identify Minority Sequence Variants in the HIV-1 env Gene from Clinical Samples

  - Another clinical application of this approach is the detection of "heteromorphisms" in mitochondrial DNA. Here, the idea is to detect diseases associated with mitochondrial mutations (encephalopathies, neuropathies, oxidative phosphorylation disorders, etc.). Apparently mtDNA has a high mutation rate and

- General problem harder:
- Example application:
  - Systematic search for correlation between genotype/expression data and resistance to treatment
  - Need to integrate data from different sources
  - Tale from bioinformatics
  - The meta structure of clinical data should be computable examples: multiple clinical structures…. (example openehr)
  - Direction: computable abstract meta description of data that can be compiled in concrete links to real data (a la openEHR) should encompass also 'big' data objects? Describe data type so that I can express a selection on it and that it can be directly computed
  - [end with a wish on what we would like to have]