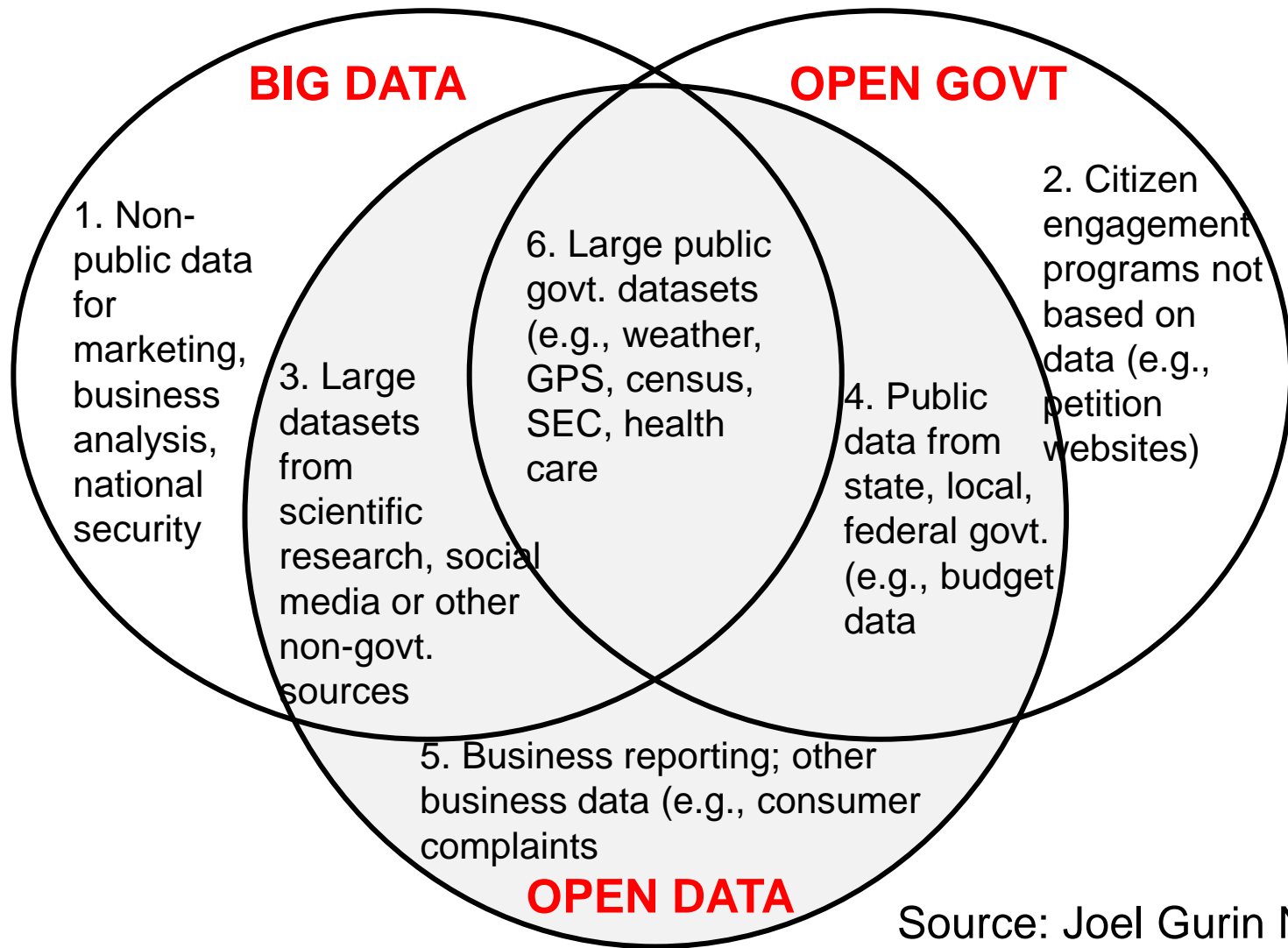


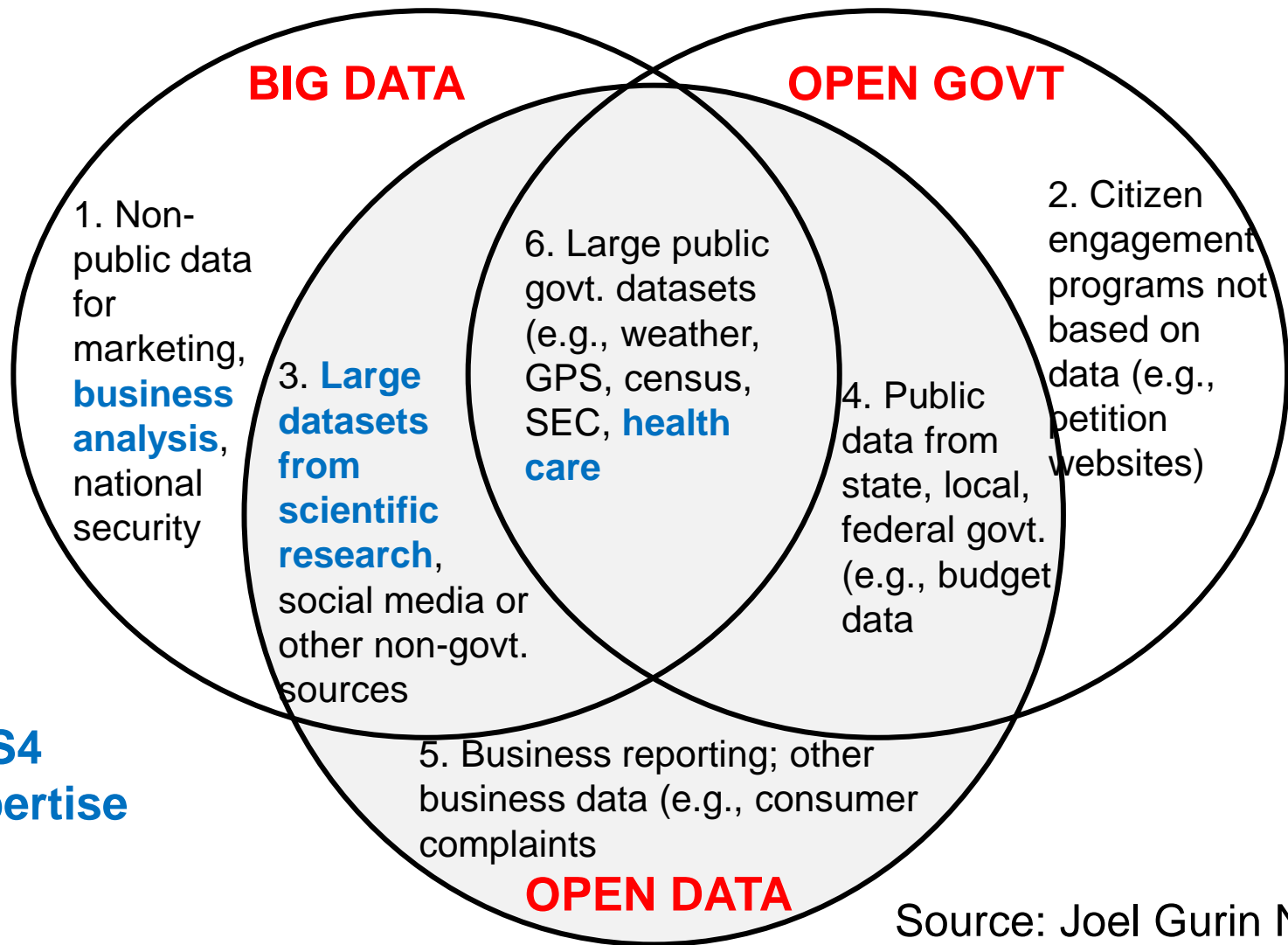
Big data support to open data initiatives

Gianluigi Zanetti
CRS4, Italy



Source: Joel Gurin NYU

**CRS4
Expertise**





www.crs4.it

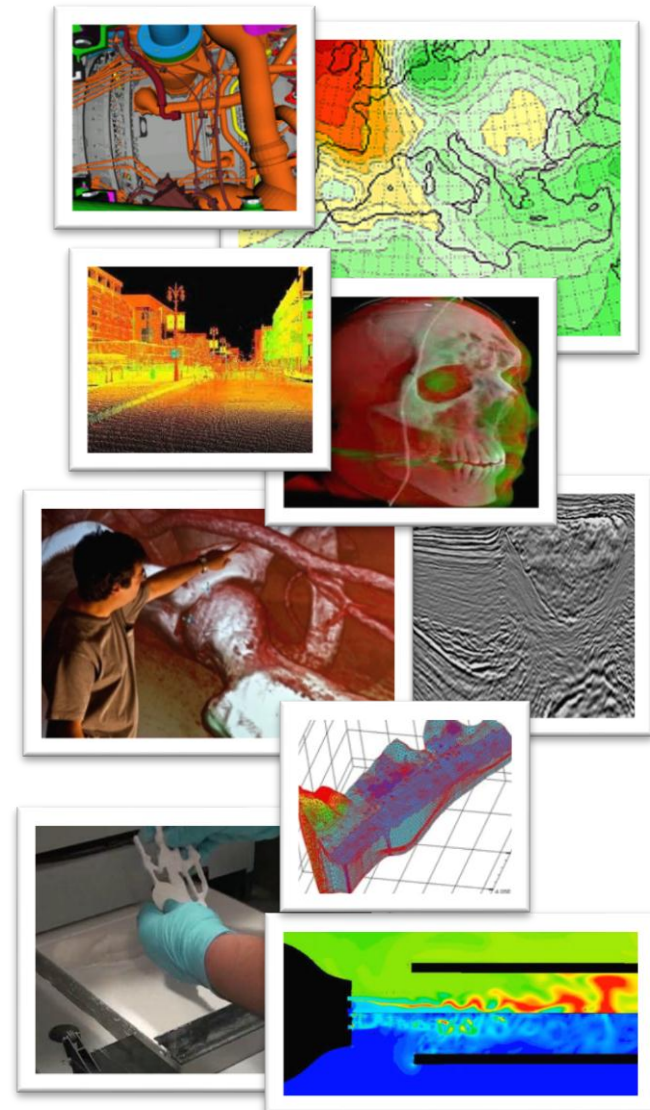
Centro di Ricerca, Sviluppo e Studi Superiori in Sardegna



- Interdisciplinary research center focused on computational sciences
- Located in the POLARIS Science and Technology Park (Pula, Sardinia, Italy)
- Operational since 1992
- RTD staff of ~150 people

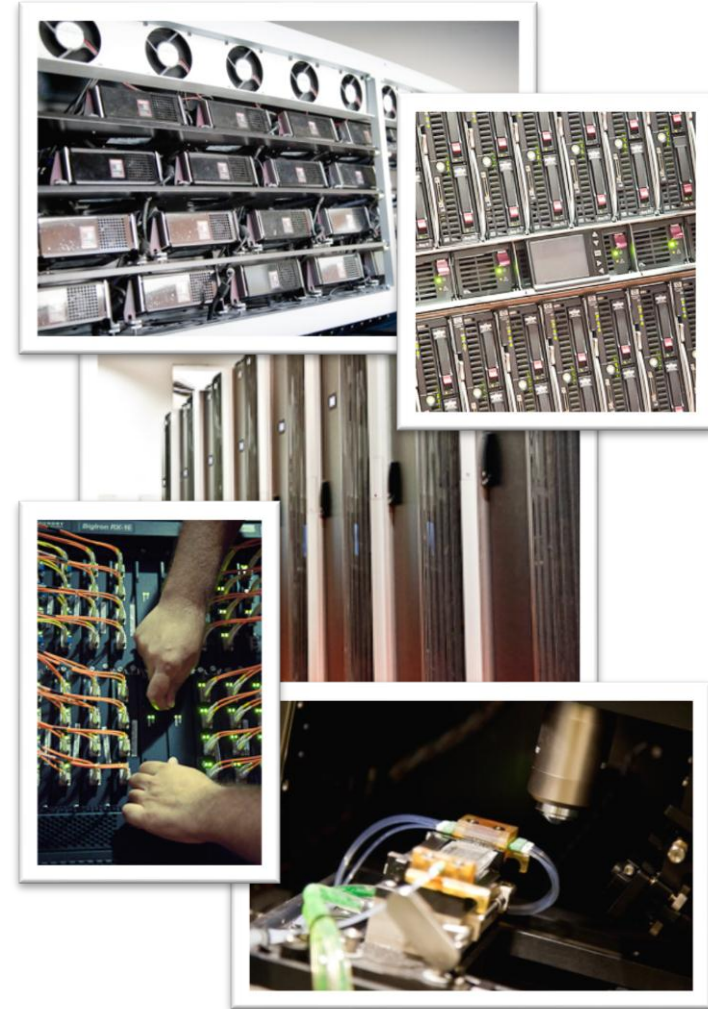
Key Strengths

- **Research and Development on enabling technologies**
 - Direct experience in the application context with primary focus on: Energy & Environmental sciences; Information society; Biomedical sciences
- **Strong Collaboration Network**
 - PON, EU FP7, Wellcome Trust, NIH, APL, ...
- **Technology Transfer to Industry**
 - ENI, INPECO, IBM, NICE, GEXCEL,
- **Technology Transfer to Regional Institutions**
 - Health Care, Environment, Cultural Heritage, ...
- **Big data/open data scalable infrastructure and technological applications**
 - CNR, City of Cagliari, ...



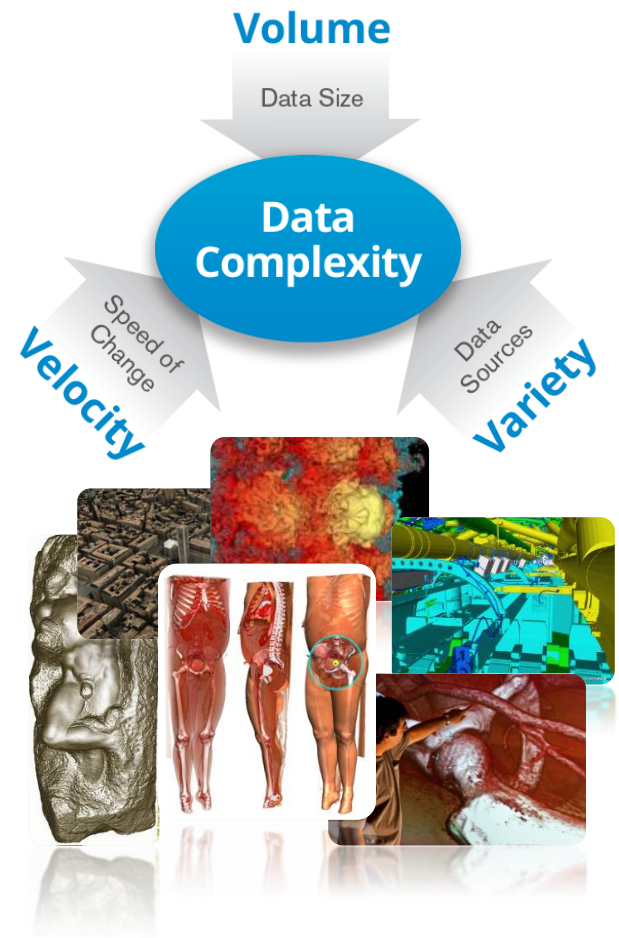
Research facilities

- **State of the art computational facilities**
 - In Italy top five, wide variety of hybrid configurations (GPU, FPGA, ...)
- **High Speed connections**
 - Multiple 10GbE towards Sardinian Regional Research Network (RTR) and National Research Network (GARR)
- **High throughput analysis facility**
 - The largest Next Generation sequencing facility in Italy, with a cumulative output of up to 5.4 TBytes of raw sequencing data every month
 - CRS4 is the computational & bioinformatics backbone to large scale biomedical projects in Sardinia



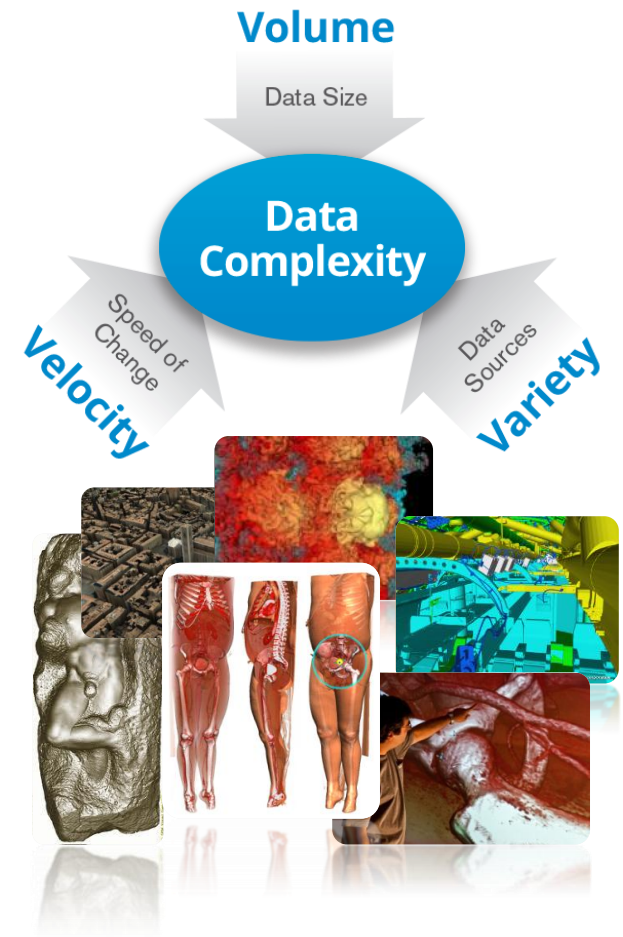
The Big Data picture

- Explosion of data in all areas of science, engineering, health and business applications, driven by hardware and information processing technology
 - **Acquisition/Generation:** 3D imaging, remote sensing, ubiquitous sensing devices, ..., computer simulations...
 - **Data management:** on a different scale
 - **Data analysis:** from data to information
- Need for novel tools, techniques, and expertise!
 - CRS4 focuses on scalable technologies

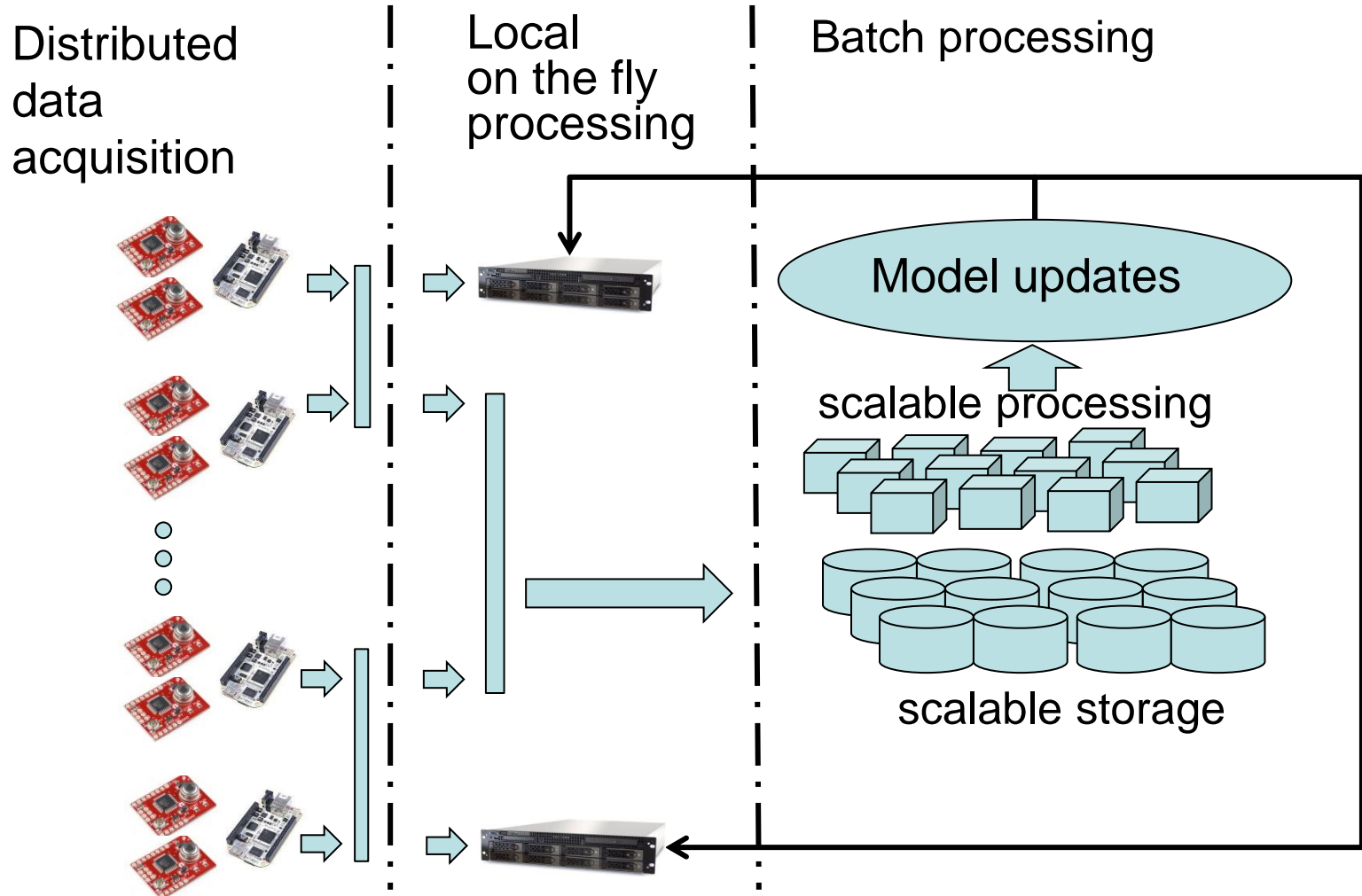


Some Big Data applications @ CRS4

- **Smart grids**
 - Forecasting renewable energy production, consumption, smart grid control.
- **Cultural Heritage**
 - Rendering complex 3D models very high res scanning of historical artifacts.
- **Data-intensive biology**
 - Analysis of population level, heterogeneous, data sets



Lambda-architecture



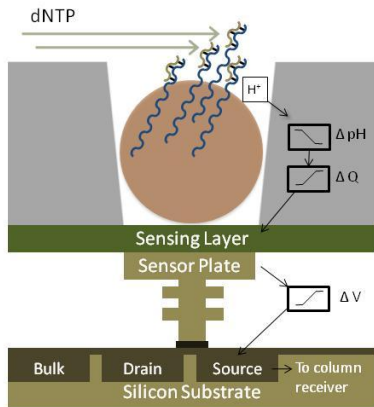
Data intensive biology @ CRS4

- **Data management and analysis support to a very precise characterization of the Sardinian population (with CNR-IRBG)**
 - based on samples coming from population wide studies on auto-immune diseases and aging 25,000 microarray datasets (SNPs/CNV and expression),
 - about 3000 Whole genome Re-sequencing, about 700 samples RNA-seq and
 - hundreds of Exome-seq samples coming from population wide studies on auto-immune diseases and aging.
 - **Asset that qualifies Sardinia at the European scale** (BBMRI, H2020) made “live” by CRS4 data management and analysis support
- **TIGET quality control pipeline for gene therapy**
 - Lentiviral Hematopoietic Stem Cell Gene Therapy of Metachromatic Leukodystrophy and Wiskott-Aldrich Syndrom
 - CRS4 **provides a key enabling technology** by speeding up analysis x100

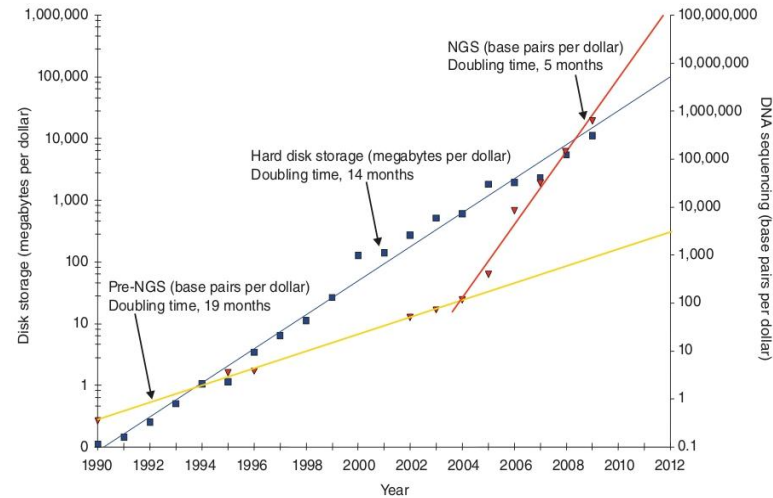


Massively parallel biochemistry

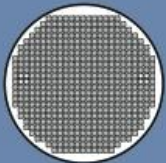
-- digitalization via semiconductor tech



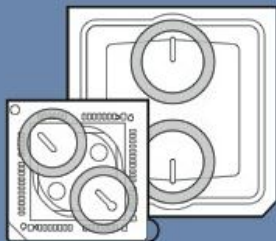
life technologies™ | ion torrent
Semiconductor Sequencing for Life™



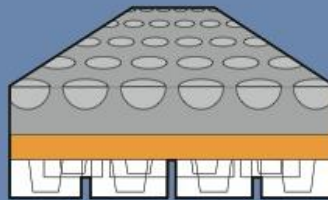
WAFER
Semiconductor Manufacturing



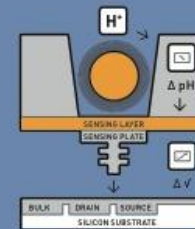
PGM / PROTON CHIP
Semiconductor Packaging



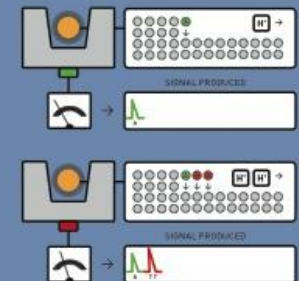
MILLIONS OF SENSORS
Semiconductor Design



SINGLE SENSOR
Chemical to Digital Sequence



SEQUENCING

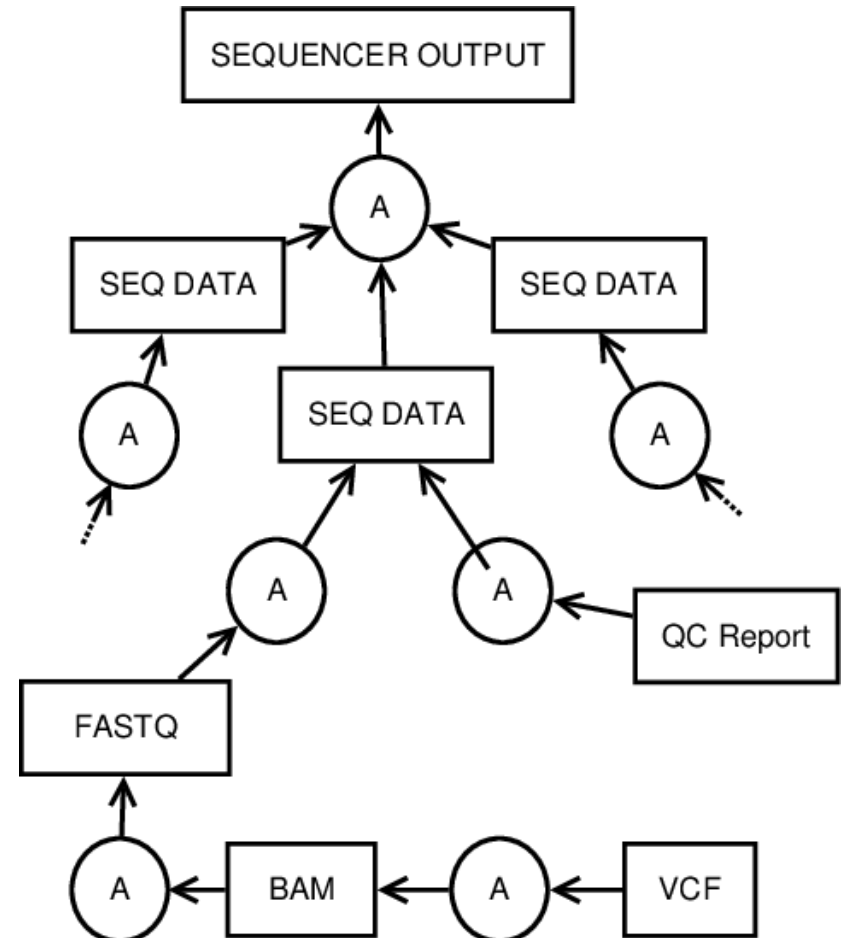


Data management and analysis challenges

- **Scaling computational throughput**
 - Cope with data size ($\sim 20\text{TByte/week}$)
 - Cope with turn-around times ($\sim 8\text{ hrs}$)
- **Tracking processes and managing large datasets**
 - Medium-size (228 samples) cancer study generates about 3900 derived datasets and a complex dependency graph structure.
- **Maximize biologists access to data and tools**
 - Typical analysis workflows require 30 or more steps

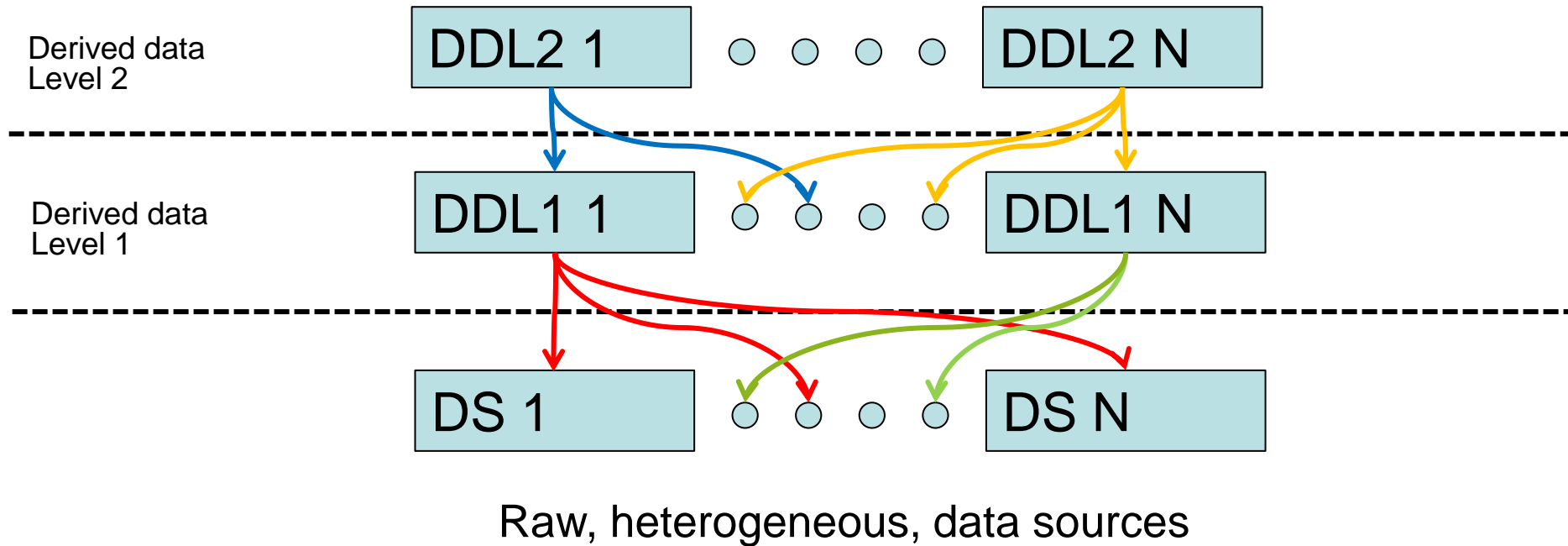
“Actionable” data provenance graph

- Full tracing of the graph of operations performed
 - Rectangles are objects (artifacts)
 - Circles are fully defined “actions”
- Usage examples:
 - Automatically schedule missing processing steps
 - Automatic quality/consistency checking
 - Confirm previous results using a different analysis approach
 - Large scale ($\sim 10^3$ genomes) comparisons
 - ...



part of the traceability graph produced by a rare-disease data analysis workflow

General need for computable data provenance as companion to open data



Conclusions & take home messages

- **The more sensors/heterogeneous data you have, the more big-data it becomes.**
- **Production of derived data should be directly bound to computable full data provenance information.**
- **We are currently working with the city of Cagliari on a open-data/big-data strategy for the metropolitan region of Cagliari**
 - next time I hope to have more relevant examples 😊.

Thank you for your time!

gianluigi.zanetti@crs4.it