(ab)using omero

Gianluigi Zanetti

(gianluigi.zanetti@crs4.it)

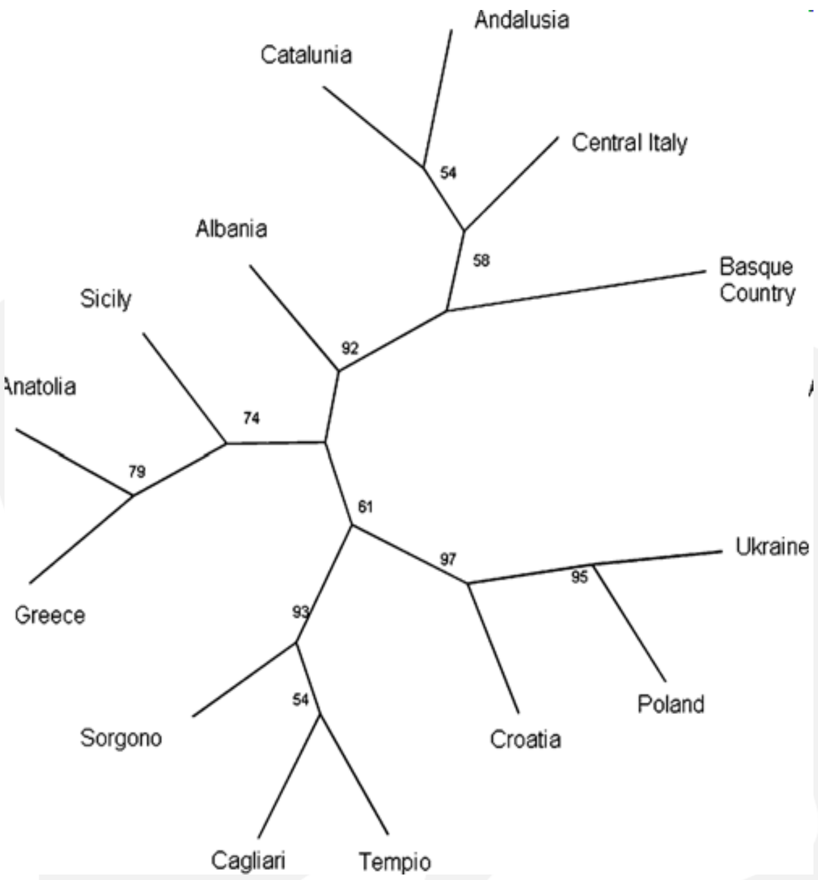
CRS4

- Center for Research, Development, and Advanced Studies in Sardinia
- Interdisciplinary research center focused on computational sciences
- Located in the POLARIS Science and Technology Park (Pula, Sardinia, Italy)
- Operational since 1992
- RTD staff of ~180 people

CRS4
POLARIS Edificio 1
C.P. 25
09010 Pula (CA), ITALY
www.crs4.it

Sardinia and its genetic isolated founder populations

- Sardinia is characterized by genetic isolated founder populations
    - An island that is big enough and where it is difficult to travel
- CRS4 cooperates with CNR-IRGB in two large scale studies
    - on longevity
    - and auto-immune diseases



Progenia: search for genetic influences on longevity

- Joint project CNR-IRGB/NIH
- Population level studies
  - 6,148 individuals - aged 14-102y, 95% are known to have all grandparents born in Sardinia
- Highly characterized samples
  - Traits (> 150) ranging from anthropometric measurements to personality facets, repeated measurements on the scale of a decade
  - High resolution genotyping, (soon) deep sequencing
- Cast of '00, see *

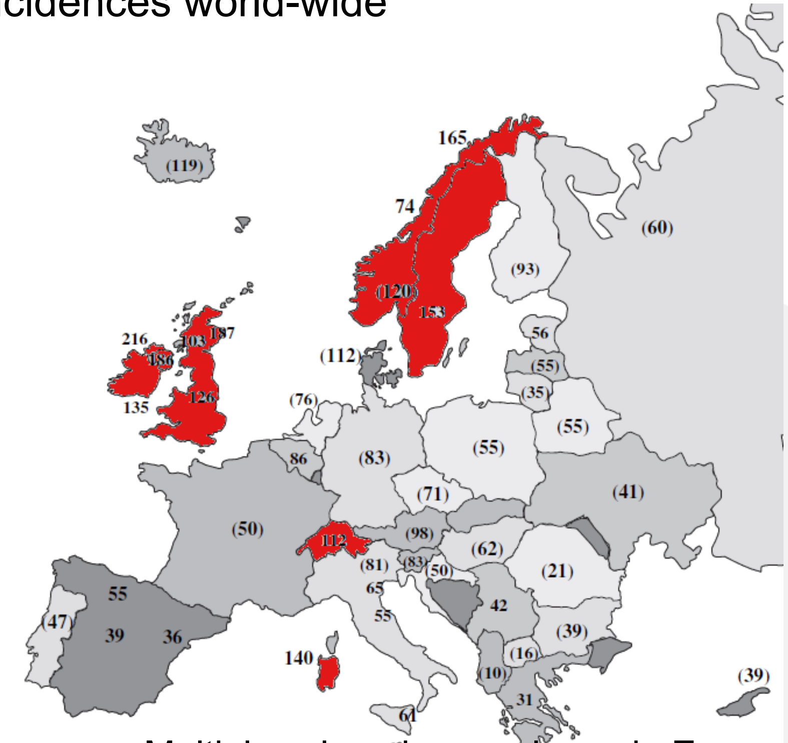
Arzana
Eliní
Ilbono
Lanusei
Stintino
Porto Torres
Sassari
Olmedo
Ossi
Ploaghe
Alghero
Ittiri
E25
Ozieri
Buddusò
Budoni
Siniscola
Villanova Monteleone
Irgoli
Orosei
Nuoro
Dorgali
Bosa
Macomer
Oliena
Orgosolo
Cuglieri
Ghilarza
Sardegna
Fonni
San Vero Milis
Samugheo
Baunei
Cabras
Arzana
Tortolì
Oristano
Lanusei
Terralba
Tertenia
E25
Sanluri
Arbus
San Gavino Monreale
Villacidro
Villaputzu
Villasor
Dolianova
Iglesias
Assemini
Sinnai
Gonnesa
Elmas
Portoscuso
Carbonia
Quartu Sant'Elena
Cagliari
Carloforte
Villasimius
Sant'Antioco
Sarroch
Pula

* http://sardinia.nia.nih.gov/Project_Team/project_team.html

Auto-immune diseases

- Collaboration with CNR-IRGB
- Population level study
  - 4,000 affected individuals and 2,000 healthy volunteers
  - High resolution genotyping
  - Order of 10^2 reseq.
- Cast of ‘00, see *

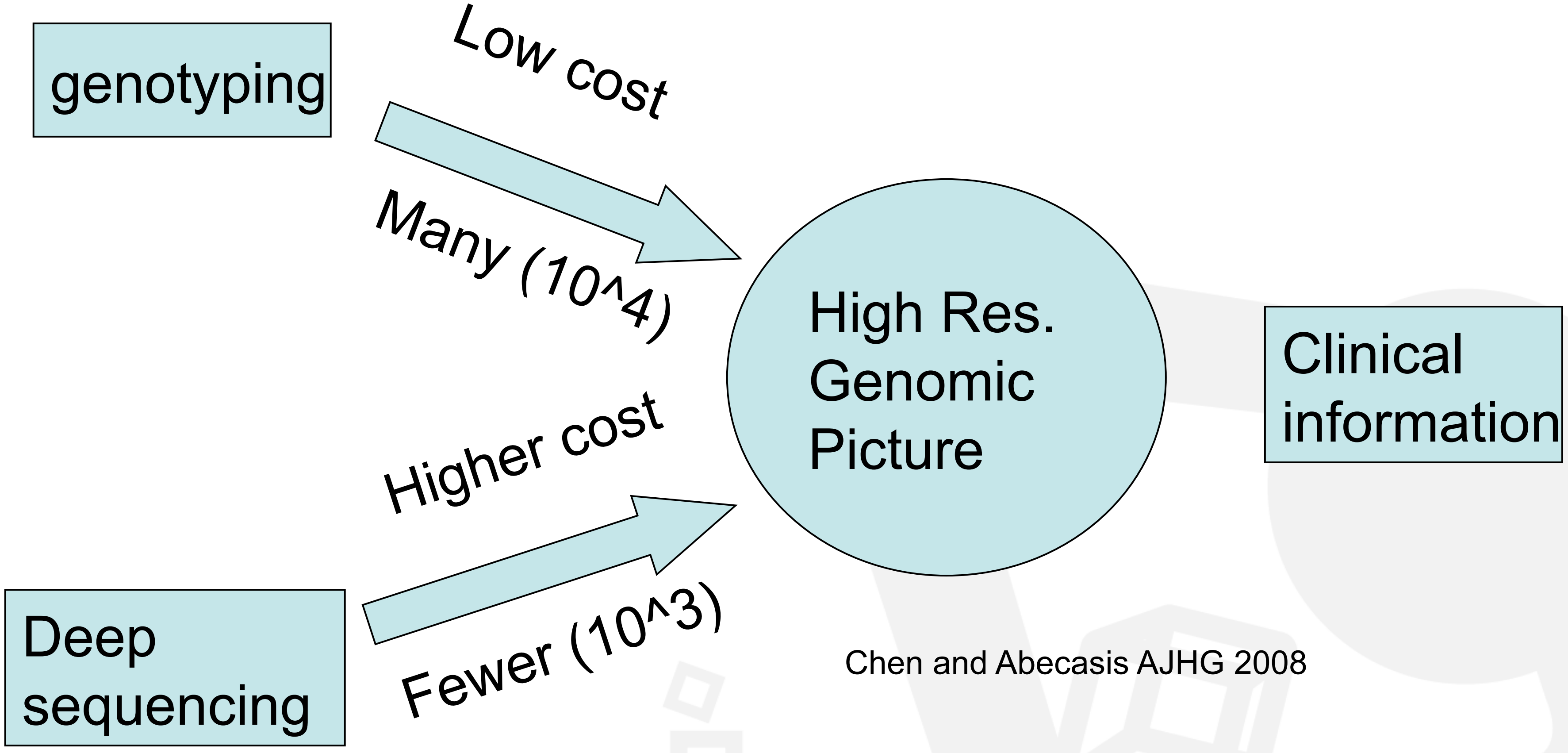
Auto-immune diseases such as type-1diabetes and MS have in Sardinia one of the highest incidences world-wide



Multiple sclerosis prevalence in Europe (from Pugliatti et al. Eu. J. Neur. 2006)

* Sanna S., Pitzalis M., Zoledziewska M., et al. Nat Genet. 42, 495 - 497 (2010).

Scale of the problem (1/4)

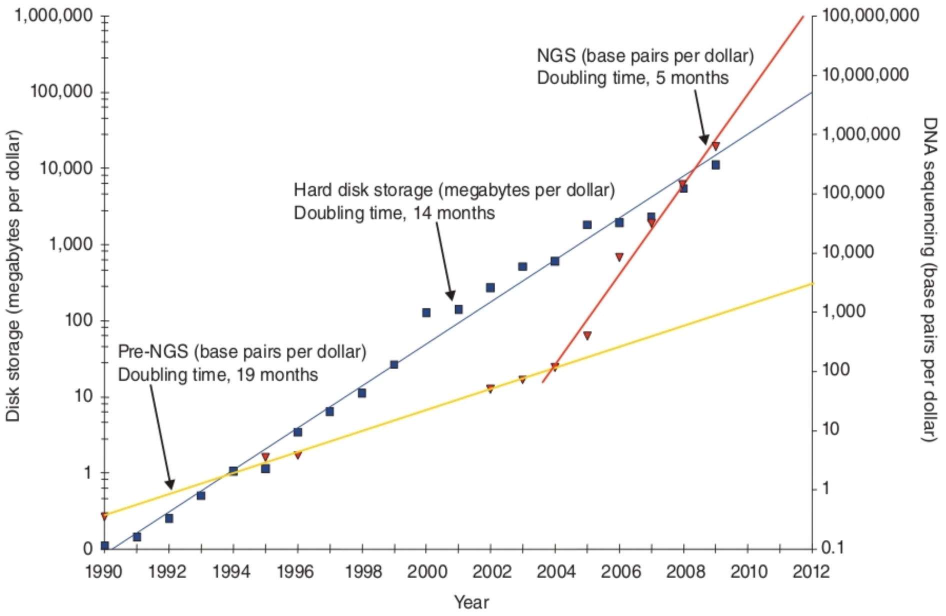


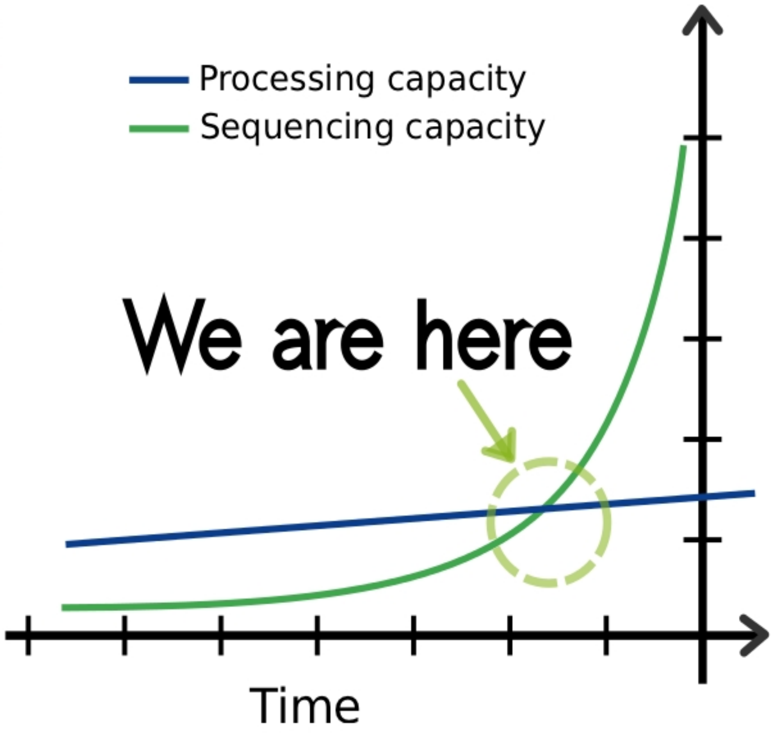
Chen and Abecasis AJHG 2008

Scale of the problem (2/4)

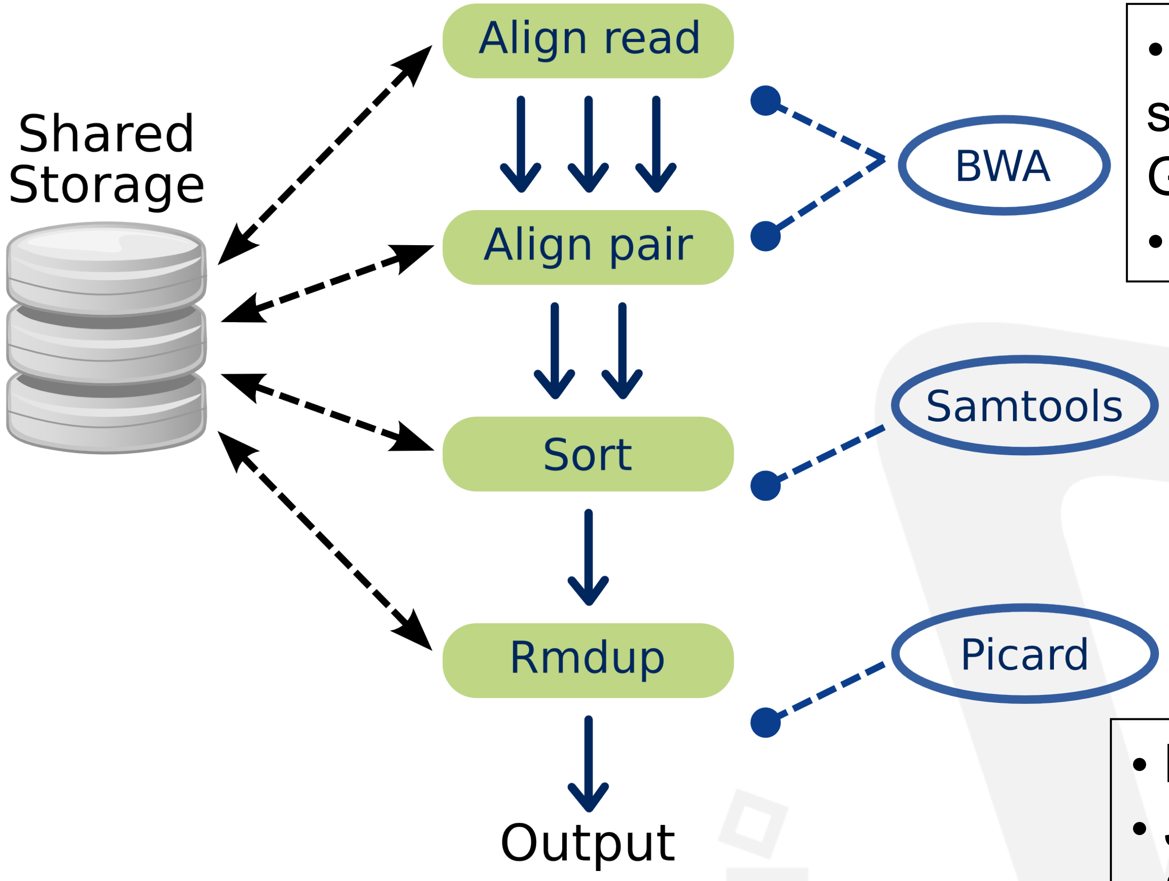
- Transition from independent ‘labs’ to a single distributed pipe-line
  - Biosamples are geographically distribuited
  - Multiple genomic technologies
  - Multiple clinical data sources
    - Existing biobanks (CNR-IRGB, Progenia)
    - Feed from Regional Health system
  - Comparable with a small hospital
- Non trivial ‘Data intensive’ problem
  - Genotyping dataset order of 10^4 ind.
  - Deep sequencing datasets order of 10^3 ind. (>4TB/week)
- Moving Target
  - More detailed clinical data
  - Soon start adding epigenomic data, e.g., ChIP-seq, RNAseq, …

A software crisis





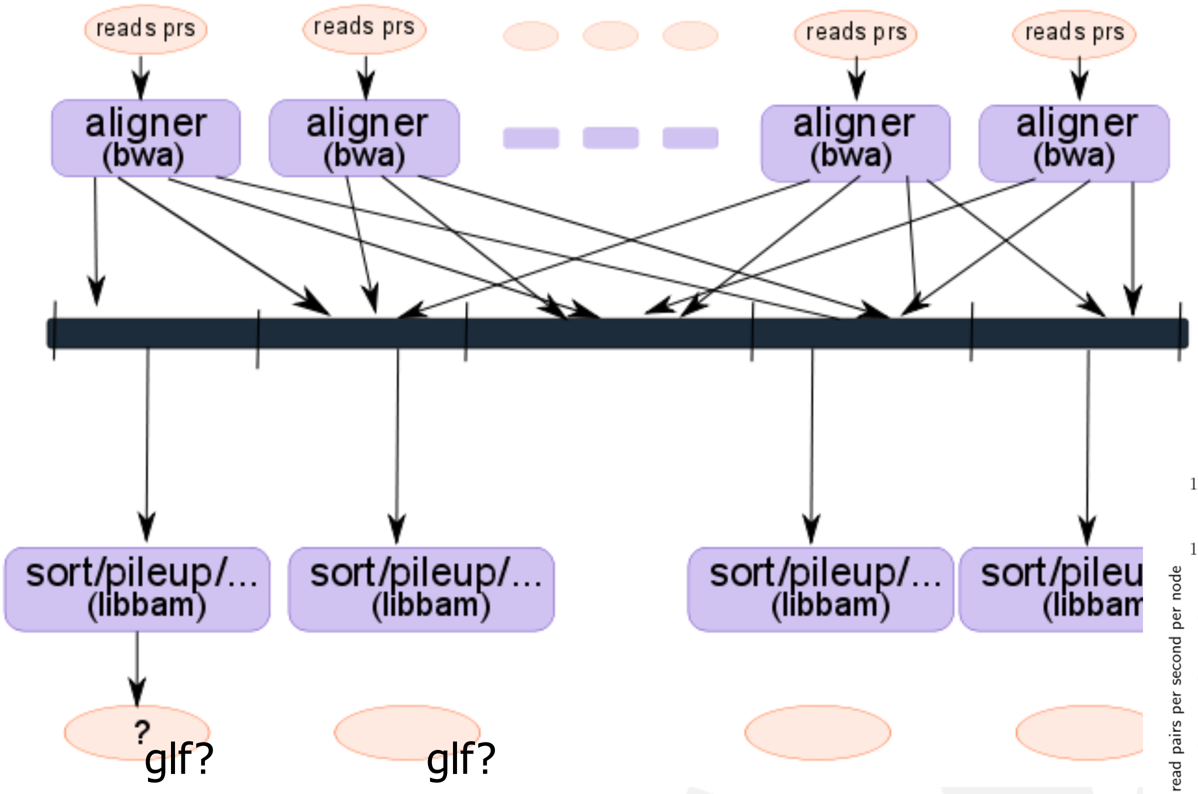
Example: re-sequencing pipeline (old way)

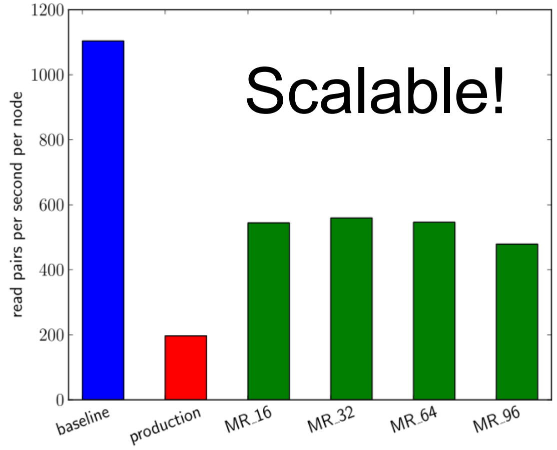


- Direct “transposition” of serial pipeline to GridEngine
- Shared file system

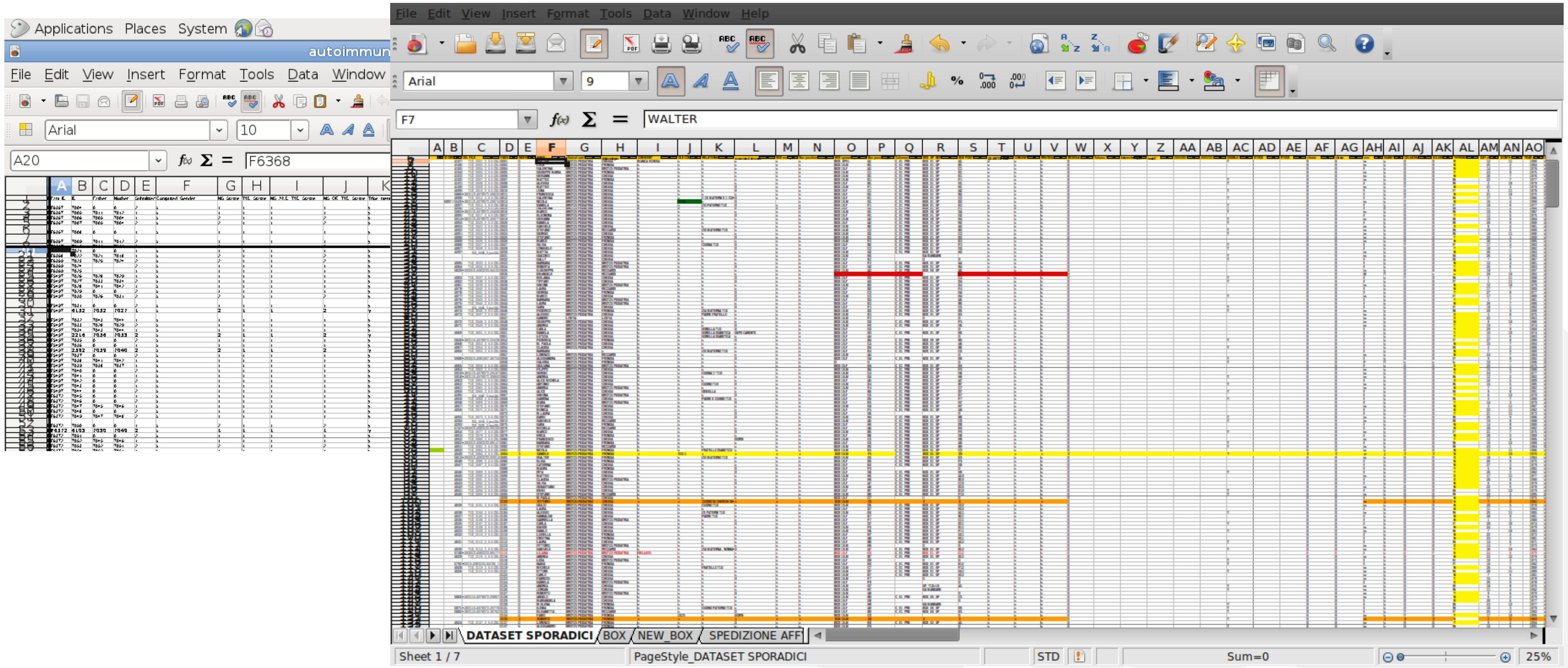
- Low parallelism
- Job failures
- Shared storage failures
- Difficult maintenance

Example: re-sequencing pipeline (new way)





A software crisis: aka excel will break!


Applications Places System
autoimmun
File Edit View Insert Format Tools Data Window
Arial 10
A20
F6368
File Edit View Insert Format Tools Data Window Help
Arial 9
F7
WALTER
DATASET SPORADICI
BOX
NEW_BOX
SPEDIZIONE AFF
Sheet 1 / 7
PageStyle_DATASET SPORADICI
STD
Sum=0
25%

Distributed excel is an inherently brittle technology

Need to capture statics and dynamics

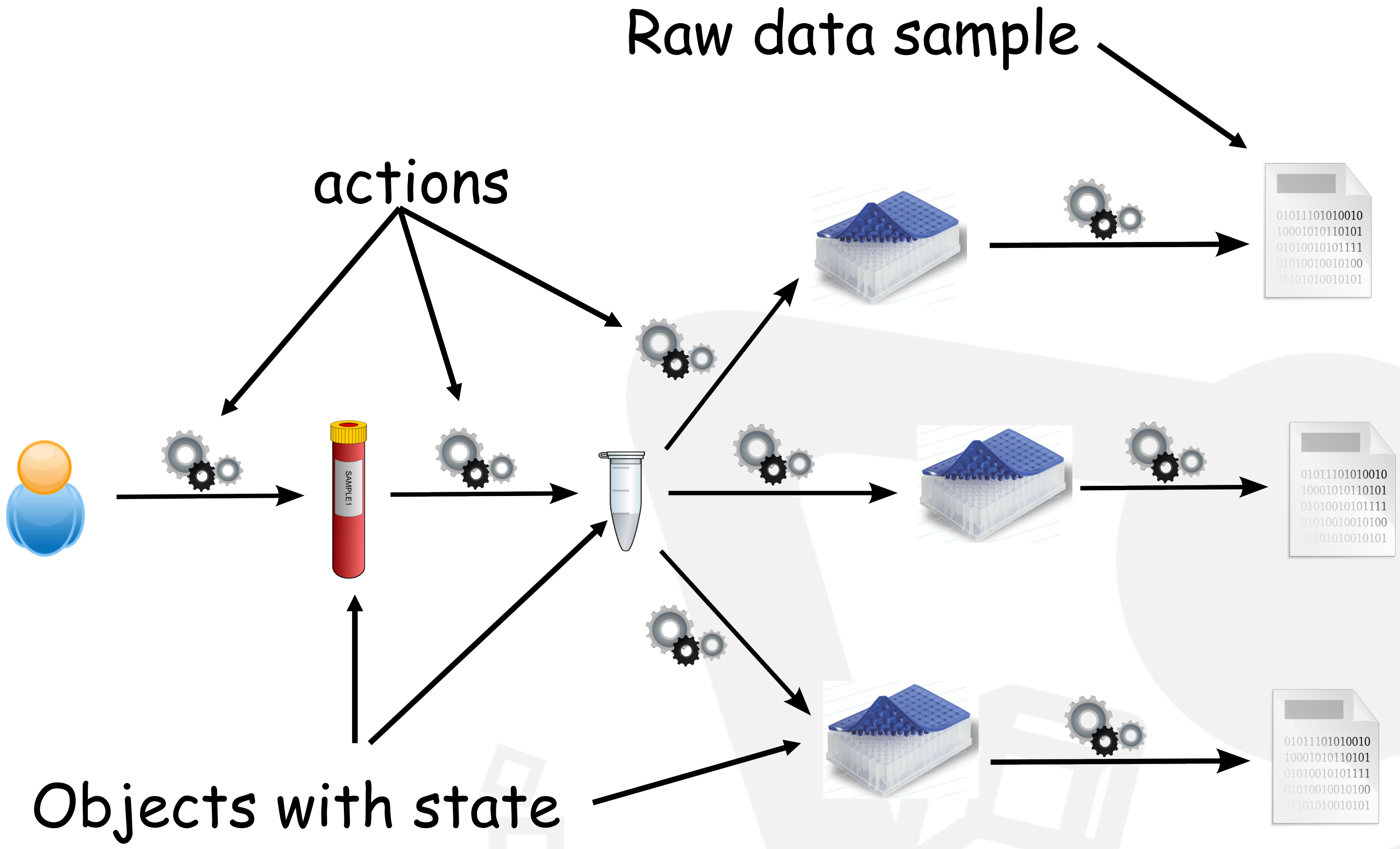
- Electronic Health Record
  - Multiple sources
  - Implementation specific details
- Samples
  - Bio and synthetic
    - Physical location / big dataset
  - Chain of Custody

results

- Operation description
  - ‘Experimental’ and ‘digital’ ops
- Computational driven inference process
  - Uniform access to data

process

Desiderata: Model state and transitions

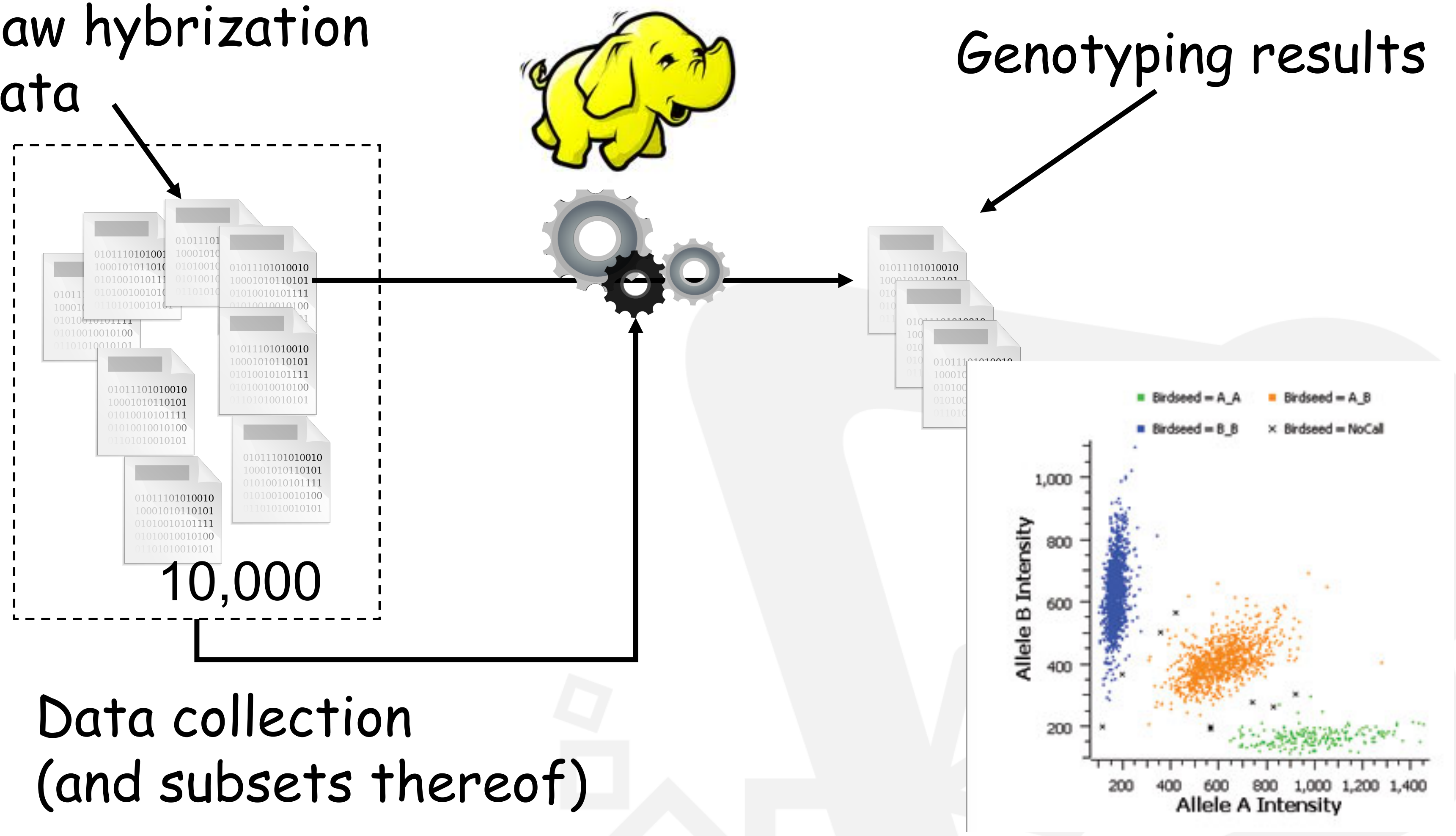


Raw data sample

actions

Objects with state

Desiderata: Keep track of dependencies



Raw hybrization data

10,000

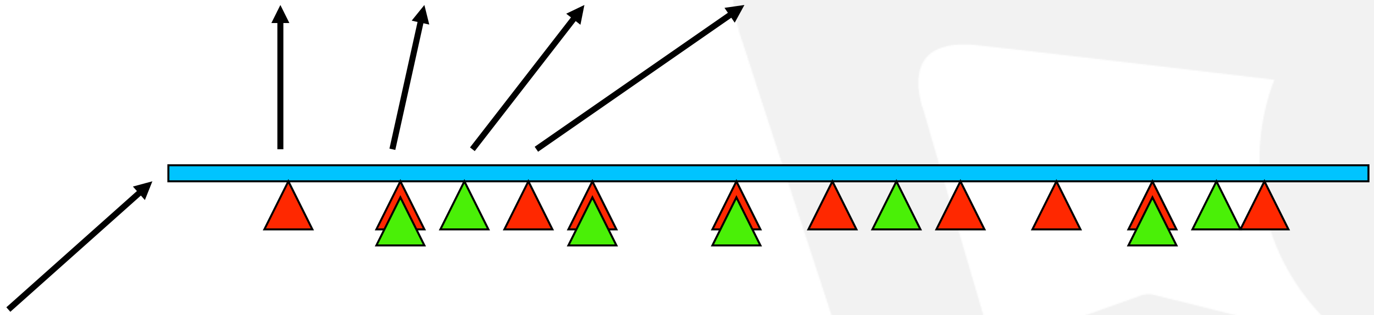
Genotyping results

Data collection (and subsets thereof)

Desiderata : Uniform SNPs mngmt

AA	0.2	0.9		0.3	0.1	…	dataset1
BB	0.8	0.1		0.7	0.9	…	

AA		0.9	0.9		0.1	…	dataset2
BB		0.1	0.1		0.9	…	



Reference genome

Flanking regions

CCGACCTAGTAGGCAA[A/G]TAGACACTGAGGCTGA

Desiderata:
Clean access to data (1/3)

```
def main():
    kb = KB(driver='omero')(…)
    maker, model = 'crs4-bl', 'taqman-foo'
    mset = kb.get_markers_set(maker, model)
    s = gkb.get_gdo_iterator(mset)
    counts = algo.count_homozygotes(s)
    mafs = algo.maf(counts)
    hwe  = algo.hwe(counts)
```

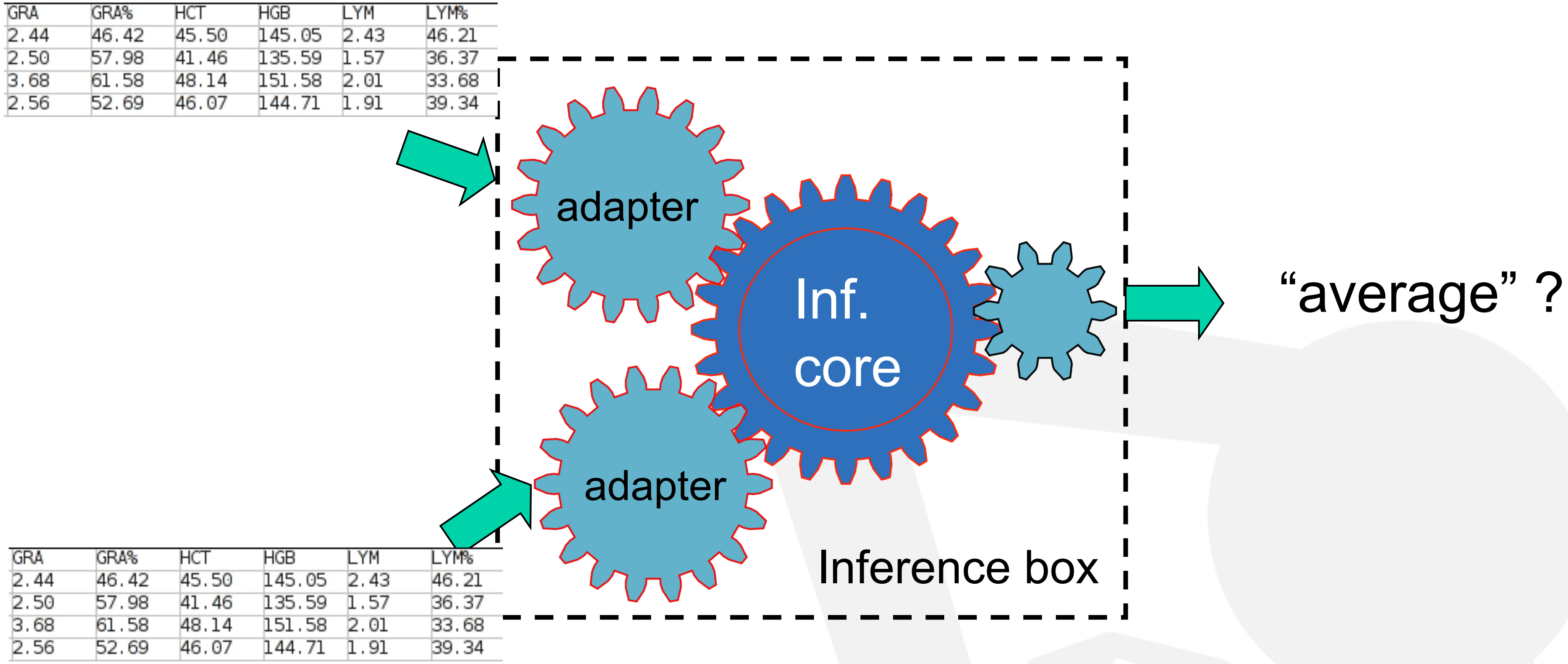
Clean access to data (2/3)

```
def main():
   kb = KB(driver='omero')(…)
   …
   enrolled = kb.get_enrolled(study)
   #--
   for e in enrolled:
      dsets = kb.get_gdos(e.individual)
      support, mean, sigma = compare_dsets
(dsets)
```

Clean access to data (3/3) (well, if you like numpy)

```
def compare_dsets(gdos):
  support, mapping = algo.find_shared_support(gdos)
  ad = [np.vstack([g['probs'][:,i], g['confs'][:,i]])
        for (g, i) in it.izip(gdos, mapping)]
  map(lambda _ : np.reshape(_, (1,) + _.shape), all_data)
  all_data = np.vstack(ad)
  v  = all_data[:,0:2,:].sum(axis=0)
  v2 = (all_data[:,0:2,:]**2).sum(axis=0)
  N = all_data.shape[0]
  mean = v/N
  # FIXME: I know, this is not the variance...
  var = v2/N - mean**2
  return (support, mean, np.sqrt(var))
```

Computable data semantics is crucial



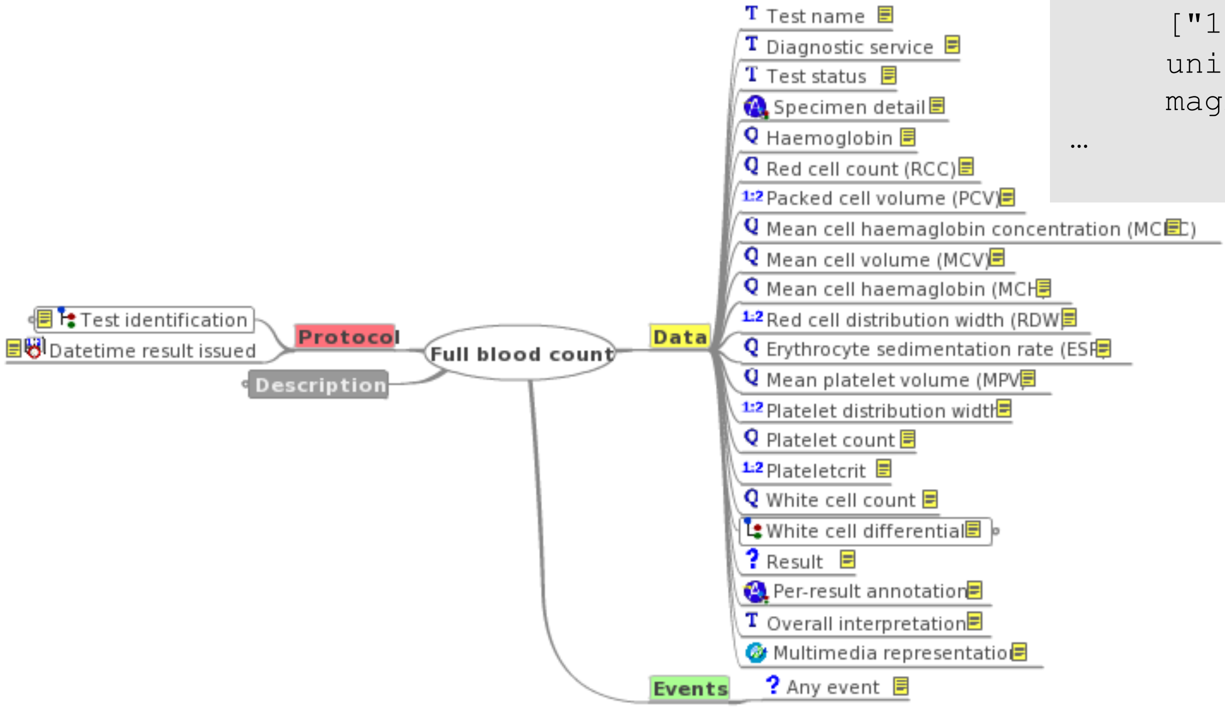
GRA	GRA%	HCT	HGB	LYM	LYM%
2.44	46.42	45.50	145.05	2.43	46.21
2.50	57.98	41.46	135.59	1.57	36.37
3.68	61.58	48.14	151.58	2.01	33.68
2.56	52.69	46.07	144.71	1.91	39.34

GRA	GRA%	HCT	HGB	LYM	LYM%
2.44	46.42	45.50	145.05	2.43	46.21
2.50	57.98	41.46	135.59	1.57	36.37
3.68	61.58	48.14	151.58	2.01	33.68
2.56	52.69	46.07	144.71	1.91	39.34

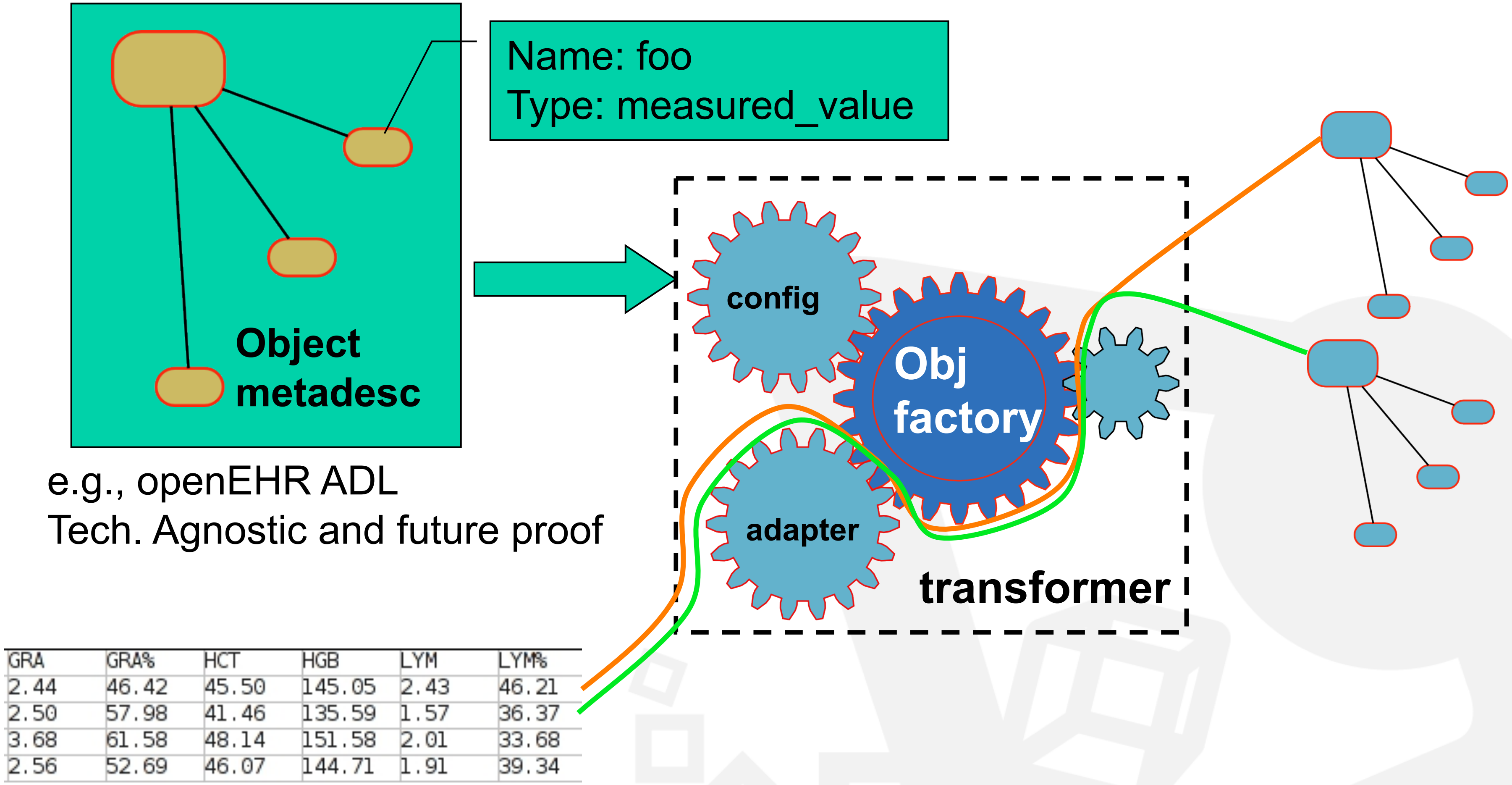
"Semantic understanding belongs to the humans that wrote the adapter"

Computable clinical semantics

```
… ADL SNIPPET
ELEMENT[at0078.7]
 occurrences matches {0..1} matches {
 -- Mean cell haemaglobin conc. (MCHC)
 value matches {
   C_DV_QUANTITY <
       property = <[openehr::119]>
       list = <
     ["1"] = <
     units = <"gm/l">
     magnitude = <|>=0.0|>
…
```



Keep biomedical & computational semantics



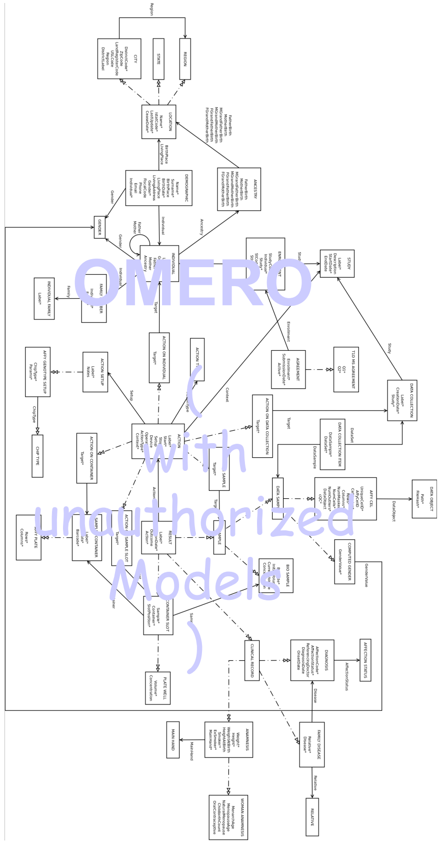
e.g., openEHR ADL
Tech. Agnostic and future proof

GRA	GRA%	HCT	HGB	LYM	LYM%
2.44	46.42	45.50	145.05	2.43	46.21
2.50	57.98	41.46	135.59	1.57	36.37
3.68	61.58	48.14	151.58	2.01	33.68
2.56	52.69	46.07	144.71	1.91	39.34

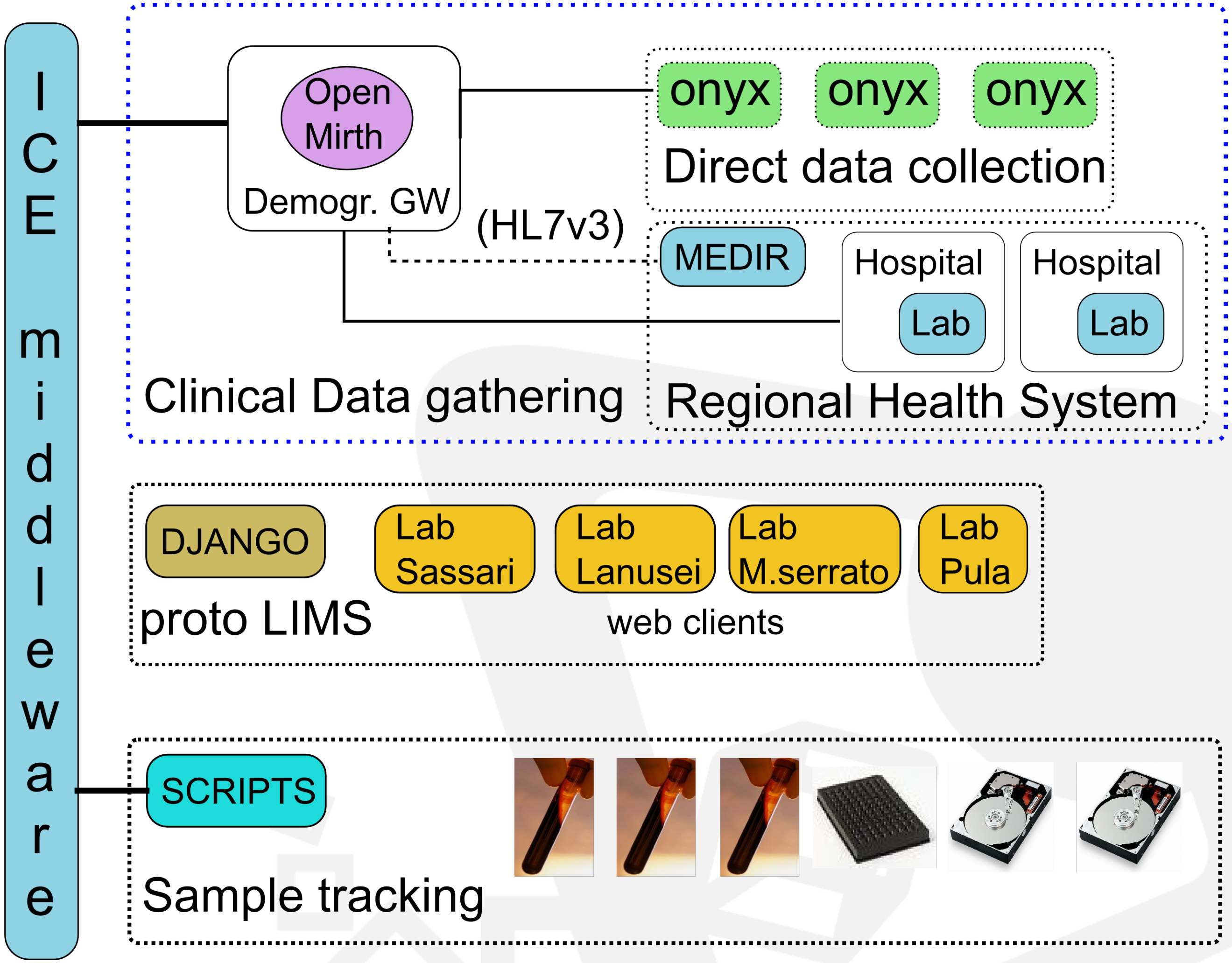
Why we like omero

- Omero is agnostic
  - Configurable, distributed, platform that deals with collections of objects
  - Agnostic vs objects models
  - Agnostic vs programming languages (client side)
- Omero can grow
  - Meta class description of objects
    - Automatically map openEHR archetypes to models
  - Omero Tables!
  - Minimal down-time for model set extension
    - Keep db, install delta (probably an hack)

Misusing Omero



OMERO
(
with
unauthorized
Models
)



ICE middleware

Open Mirth

Demogr. GW

(HL7v3)

onyx onyx onyx

Direct data collection

MEDIR

Hospital Lab

Hospital Lab

Clinical Data gathering

Regional Health System

DJANGO

proto LIMS

Lab Sassari

Lab Lanusei

Lab M.serrato

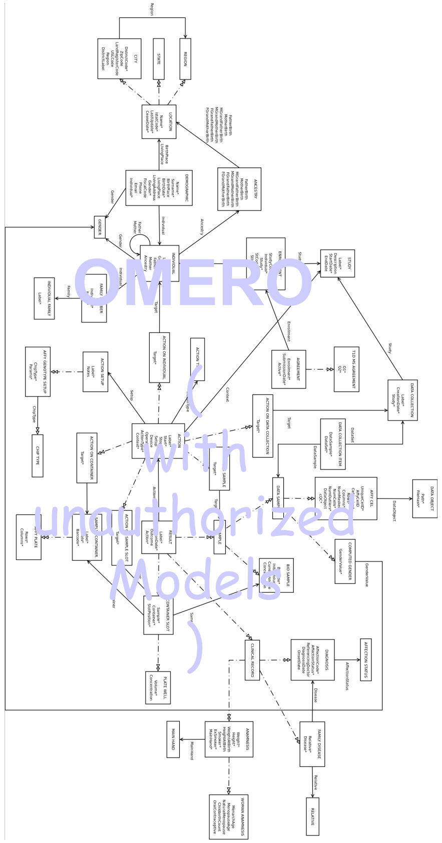
Lab Pula

web clients

SCRIPTS

Sample tracking

More Omero misuses

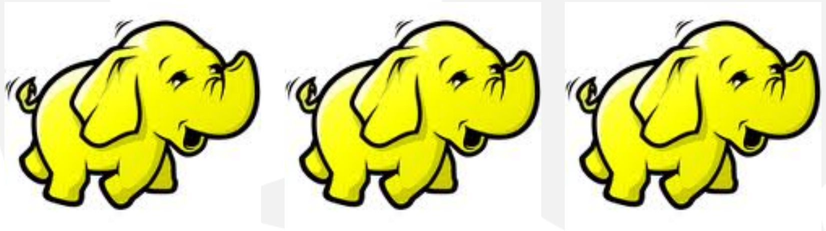


OMERO

(with unauthorized Models)

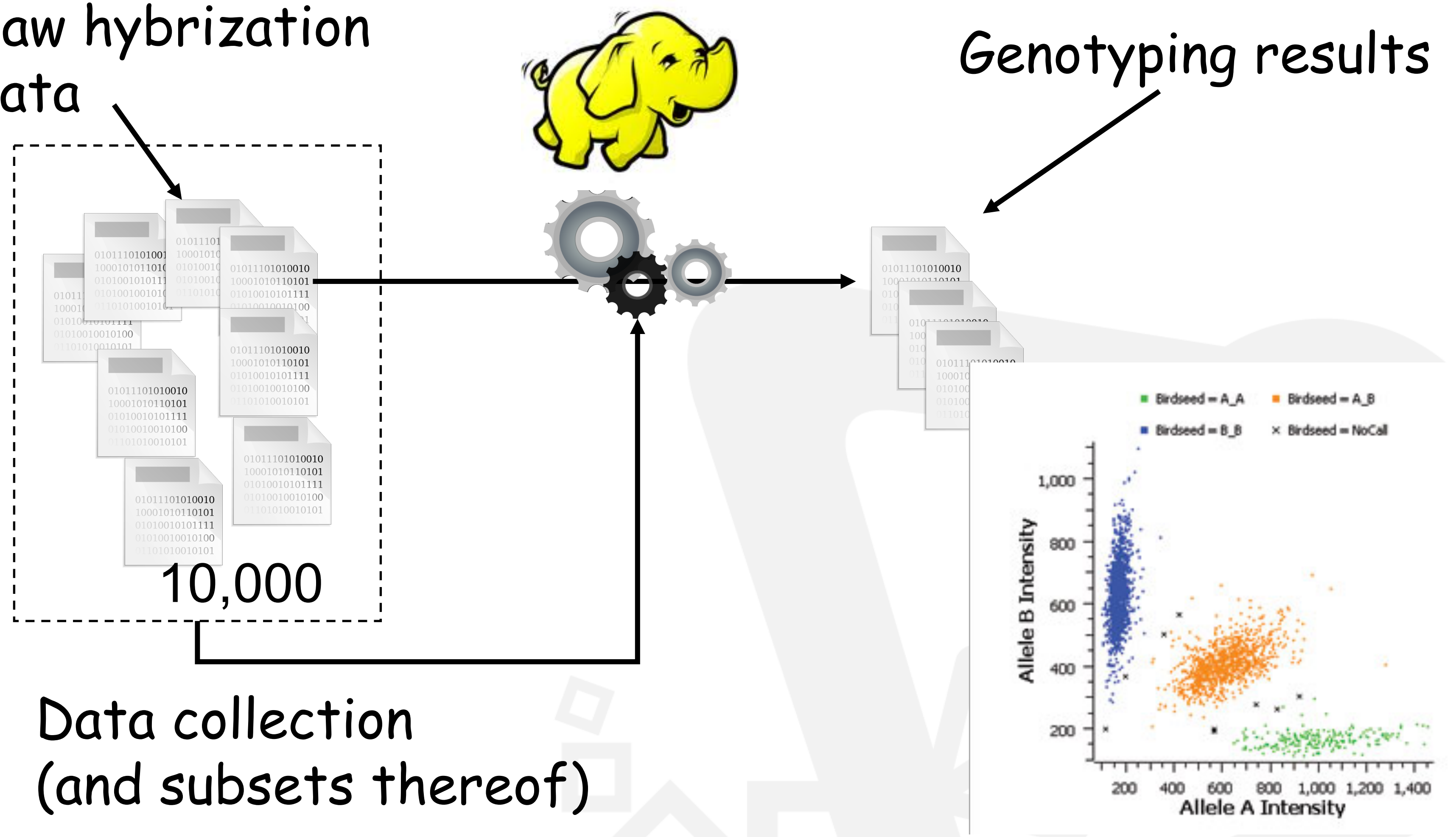
ICE middleware

Meta job dispatcher



Hadoop jobs

Desiderata:
Keep track of dependencies

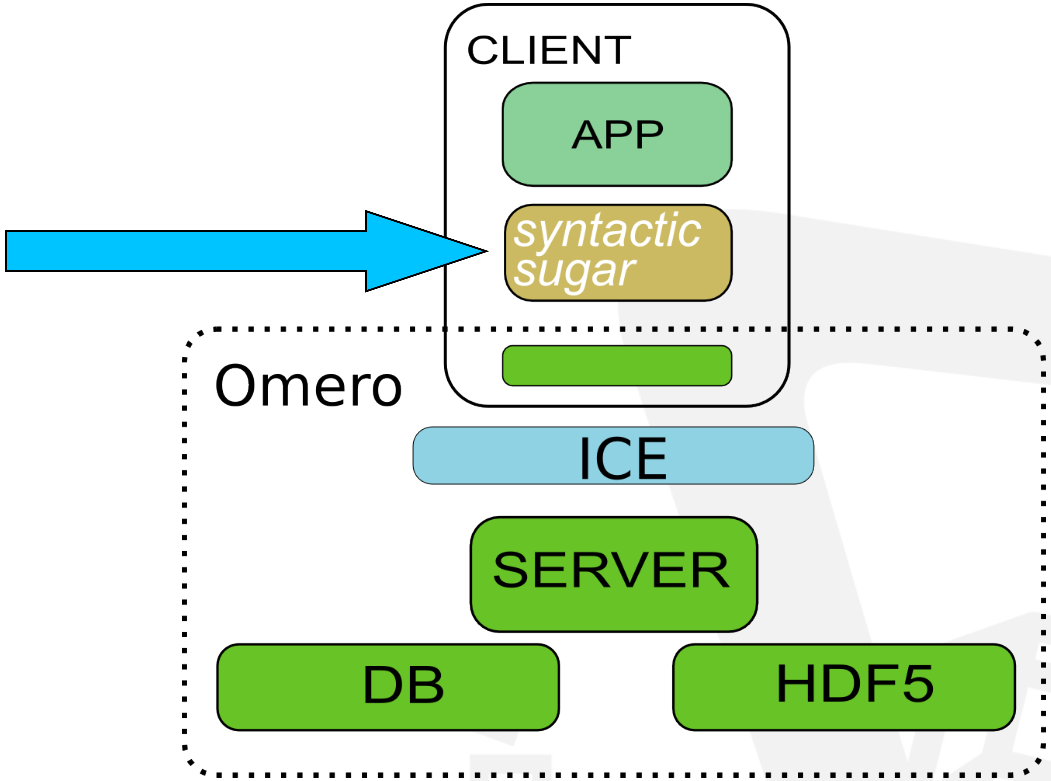


Raw hybrization data

Genotyping results

Data collection
(and subsets thereof)

An omero specialization



New models and related sugar

```
<type id="ome.model.vl.Device">
    <properties>
      <required name="vid" type="string" unique="true"/>
      <required name="label" type="string" unique="true"/>
      <required name="maker" type="string"/>
      <required name="model" type="string"/>
      <required name="release" type="string"/>
      <optional name="physicalLocation" type="string"/>
    </properties>
</type>
```

```
class Device(wp.OmeroWrapper):
  OME_TABLE = 'Device'
  __fields__ = [('vid', wp.VID, wp.REQUIRED),
                ('label', wp.STRING, wp.REQUIRED),
                ('maker', wp.STRING, wp.REQUIRED),
                ('model', wp.STRING, wp.REQUIRED),
                ('release', wp.STRING, wp.REQUIRED),
                ('physicalLocation', wp.STRING, wp.OPTIONAL)]
```

Wrap-up

- Omero is much more than bioimages handling
- Adding sugar has been enough for our needs
  - There could be some ad-hoc improvements…
- We are hiring….

CRS4
POLARIS Edificio 1
C.P. 25
09010 Pula (CA), ITALY
www.crs4.it