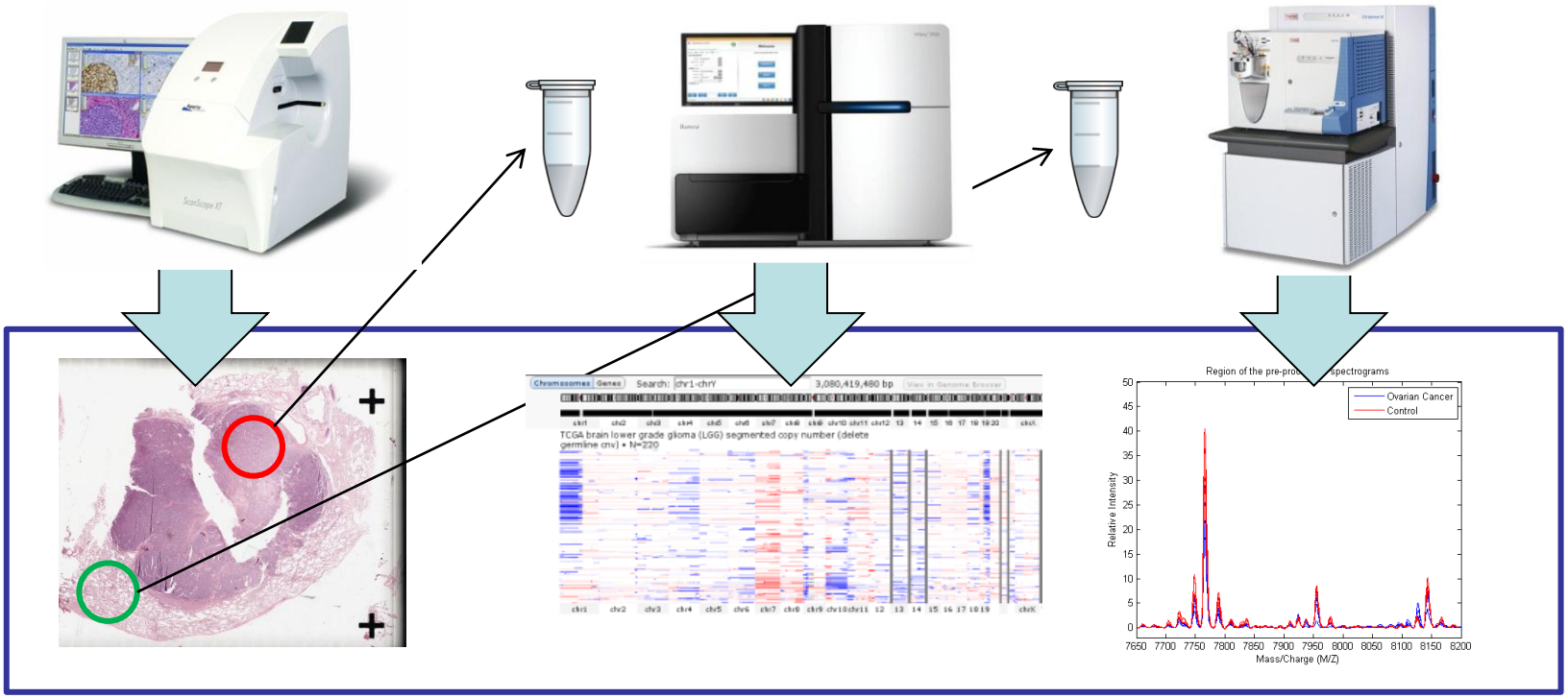
Extending OMERO to Support Non-image Data

Gianluigi Zanetti

CRS4 & Omero

- Currently dealing with production, managing and analysis of data intensive biology outputs
  - NGS, proteomics, … (~400TBytes)
  - Moving toward the confluence of above and digital pathology
- We are using omero (omero.biobank)
  - to provide a consistent computable platform (omero + galaxy) capable of handling both image and non image data
  - to support a scalable error propagation/graph of dependency data model

Our image + non image data use case (=> Next Gen Pathology)

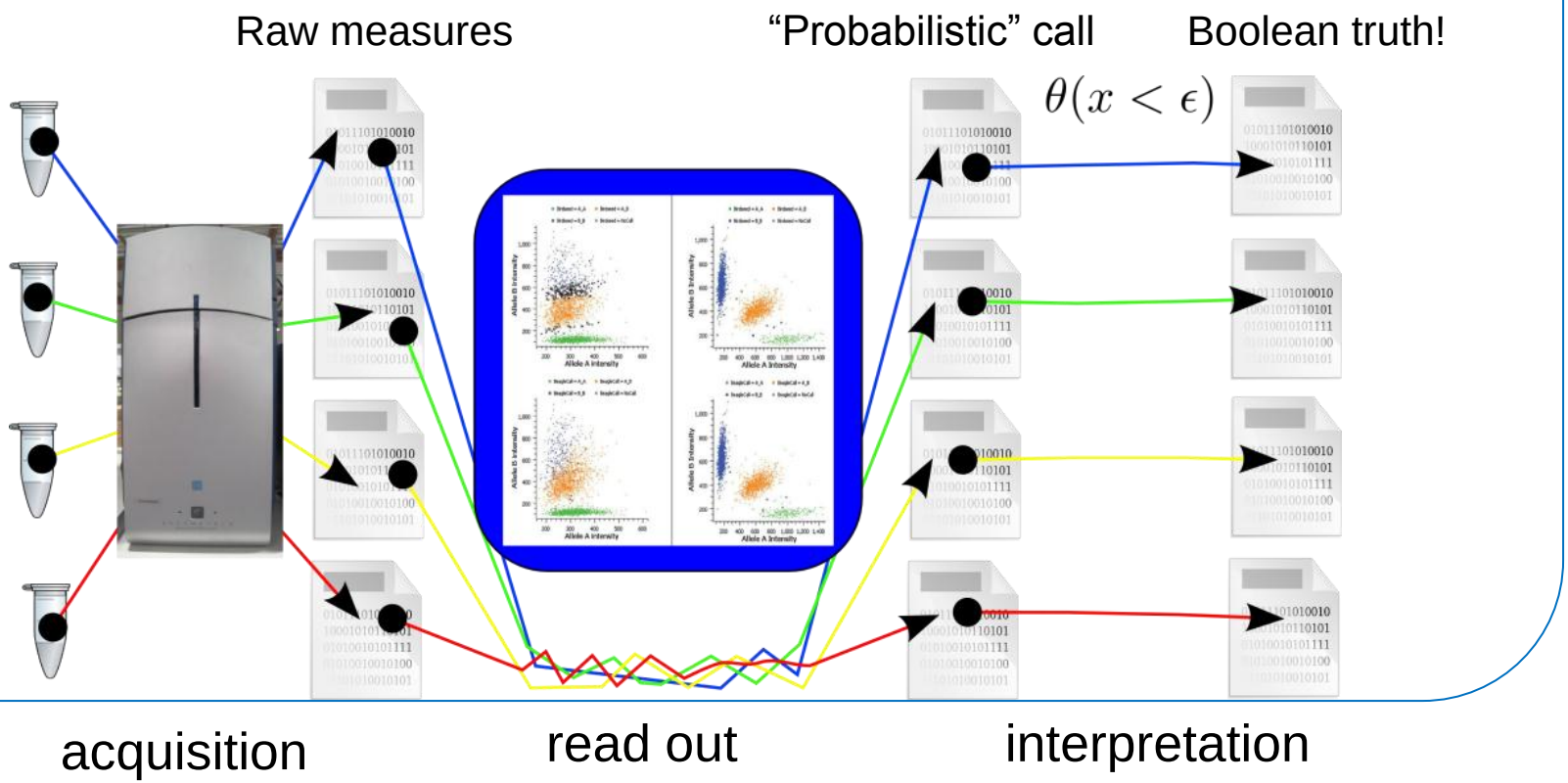


coming soon

We do now

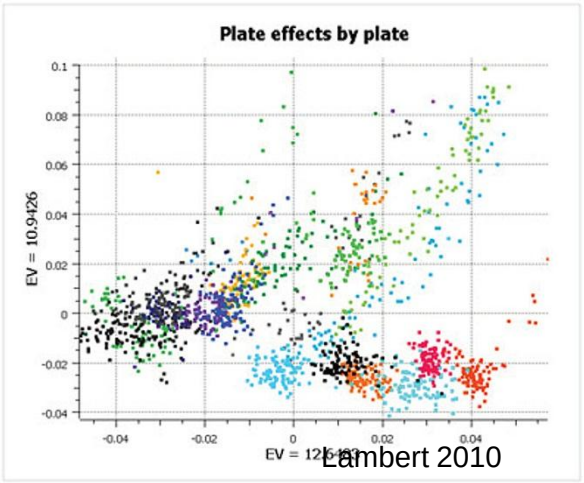
Data Intensive (quantitative) biology needs some way to handle error propagation

Example: birdseed algorithm for SNP calling



Keeping track of errors is not enough

- One needs to track also the **actions** that made an **object** depend on others

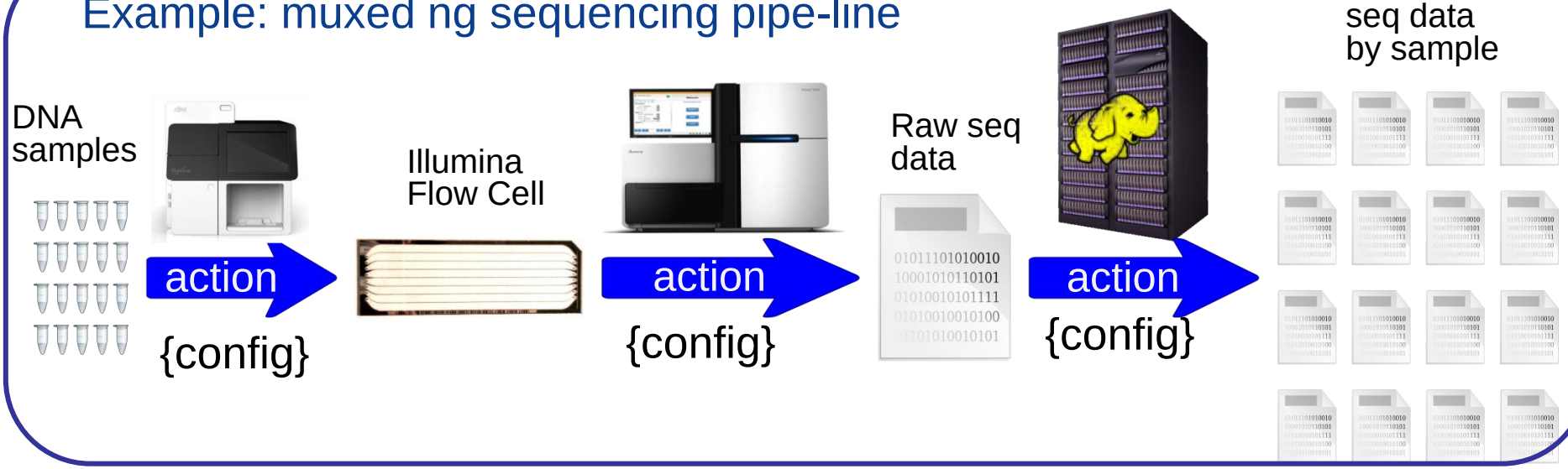


Lambert 2010

omero.biobank

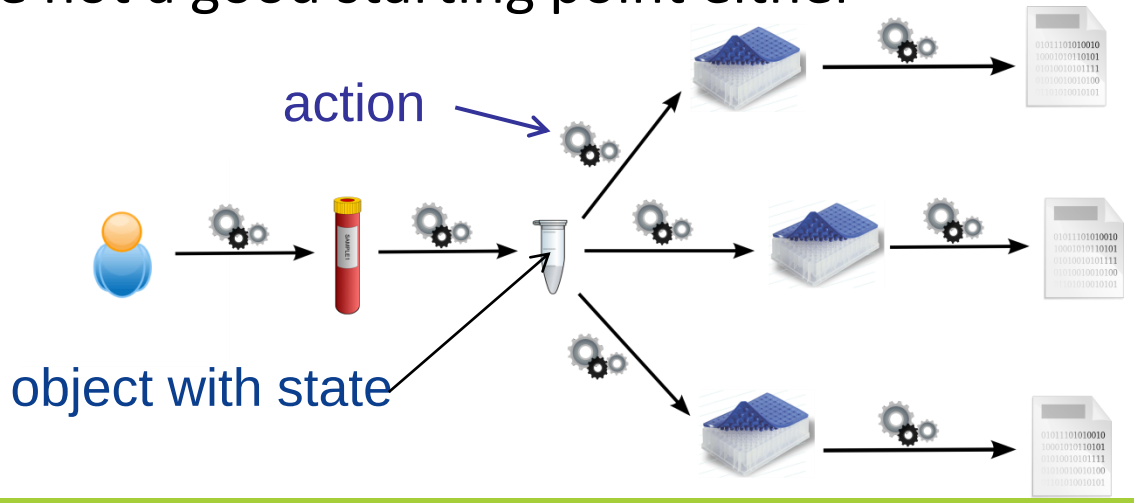
- **Objects** and **actions** as new omero models, “probabilistic” data (e.g., genotyping) mapped to omero.table(s)
- Specialized API
- Supports Galaxy as computational workflow engine

Example: muxed ng sequencing pipe-line



Yet another biodata storage system?

- Currently available systems are designed for “Here, Now, and to Uphold Tradition”.
  - Focused on current state of available data.
  - No dependency tracking, no “actions” that describe how things have been created and how they depend on each other.
  - No concept of maintaining “errors”, e.g., SNP are either NC or AA, AB, BB
- Relational db(s) are not a good starting point either



Why omero anyway?

- Omero is a configurable, distributed, platform that deals with collections of objects
  - Agnostic vs objects models
  - Agnostic vs programming languages (client side)
  - Efficient support for tabular data
  - Very efficient parameterization on models meta class description
- Omero is as expressive language well suited to describe what we want to build.
- And it can handle images too!

omero.biobank roadmap

- Omero.biobank code in `http://github.com/crs4/omero.biobank`

When	What
now	Core system
July	Genotyping data support
September	NGS pipelines automation support
Autumn	Proteomics data support
TBD	Graph component -> omero.graph
TBD	Events components -> omero.events