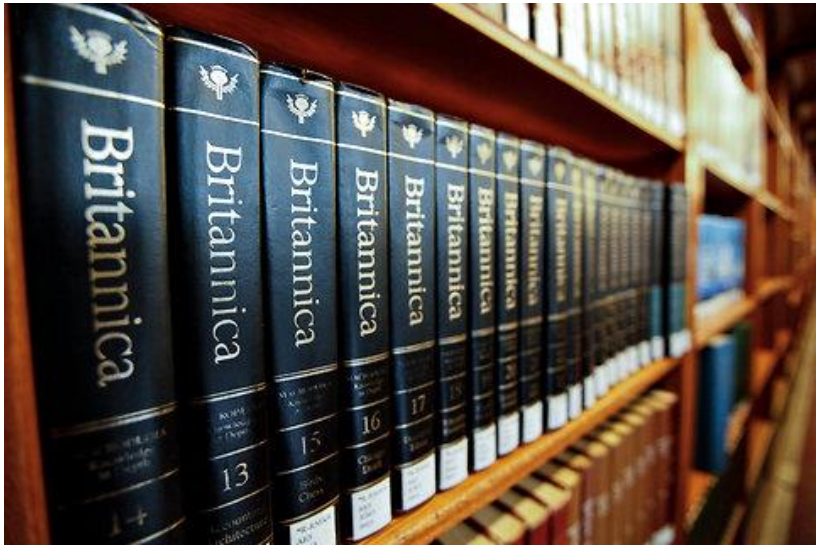
WHAT ABOUT DATA?

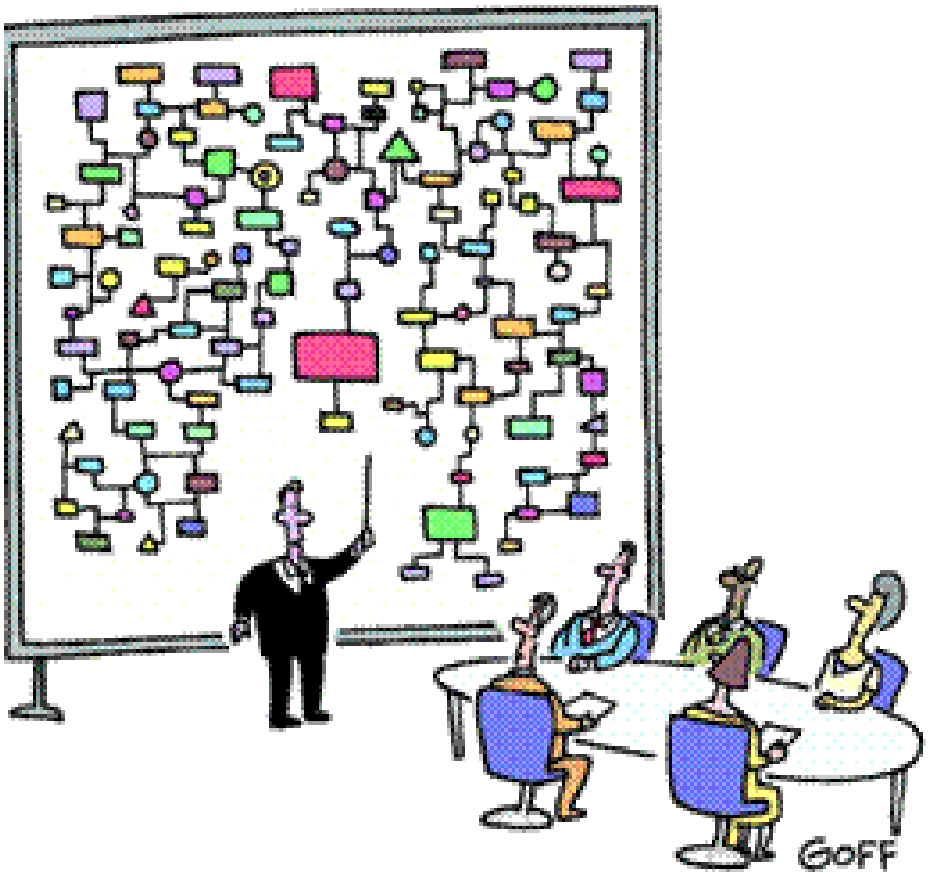
@Chris_Evelo
Department of Bioinformatics – BiGCaT
Maastricht University

Open data versus closed data



Typically we want to:

- Find studies
- Process data
- Integrate
- Evaluate
- Combine with yet other data

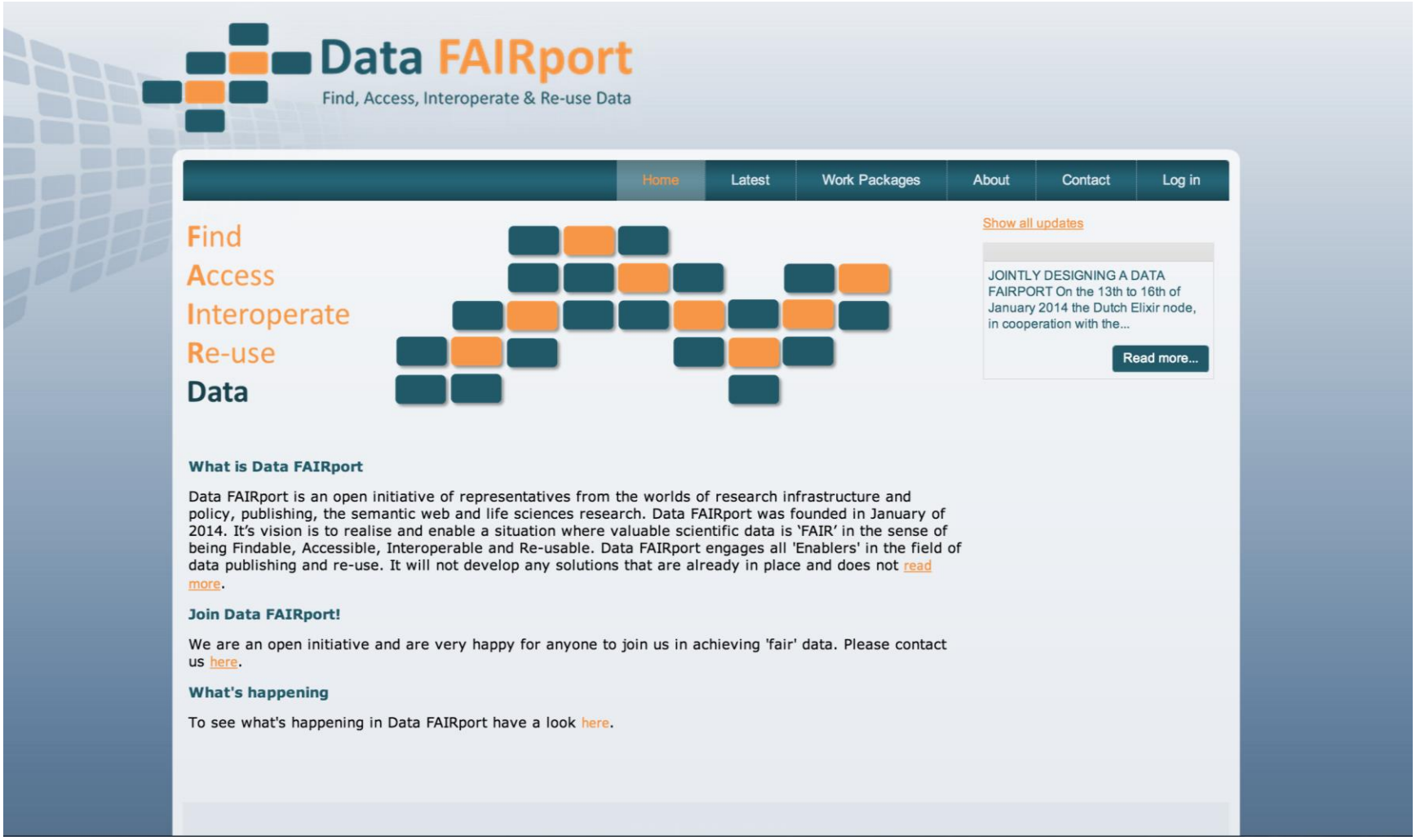


"And that's why we need a computer."

FAIR data

- Findable
- Accessible
- Interoperable
- Reusable

Find it through a FAIRport



But first we need to capture it

NATURE GENETICS | COMMENTARY OPEN

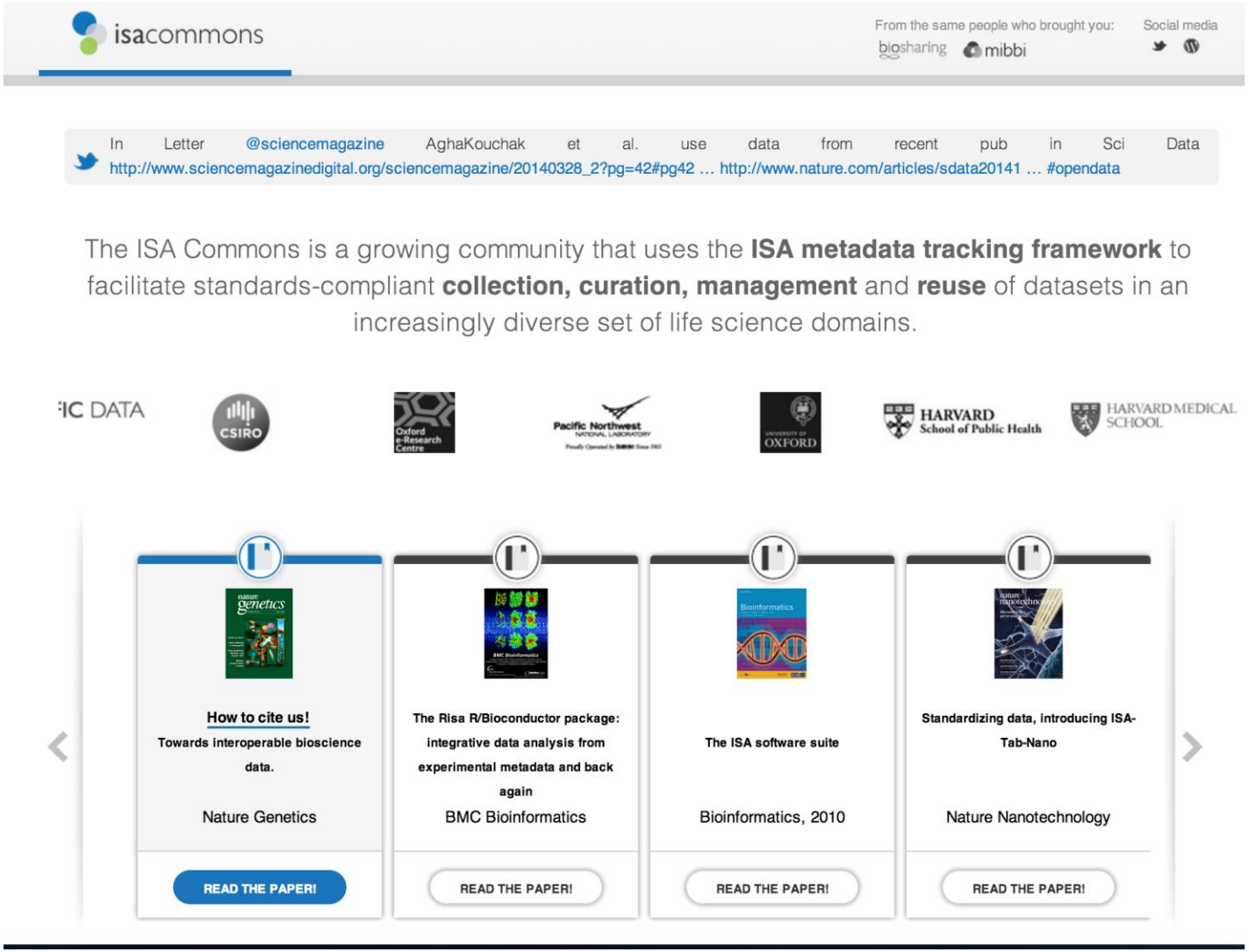
Toward interoperable bioscience data

Susanna-Assunta Sansone, Philippe Rocca-Serra, Dawn Field, Eamonn Maguire, Chris Taylor, Oliver Hofmann, Hong Fang, Steffen Neumann, Weida Tong, Linda Amaral-Zettler, Kimberly Begley, Tim Booth, Lydie Bougueleret, Gully Burns, Brad Chapman, Tim Clark, Lee-Ann Coleman, Jay Copeland, Sudeshna Das, Antoine de Daruvar, Paula de Matos, Ian Dix, Scott Edmunds, Chris T Evelo, Mark J Forster + ***et al.***

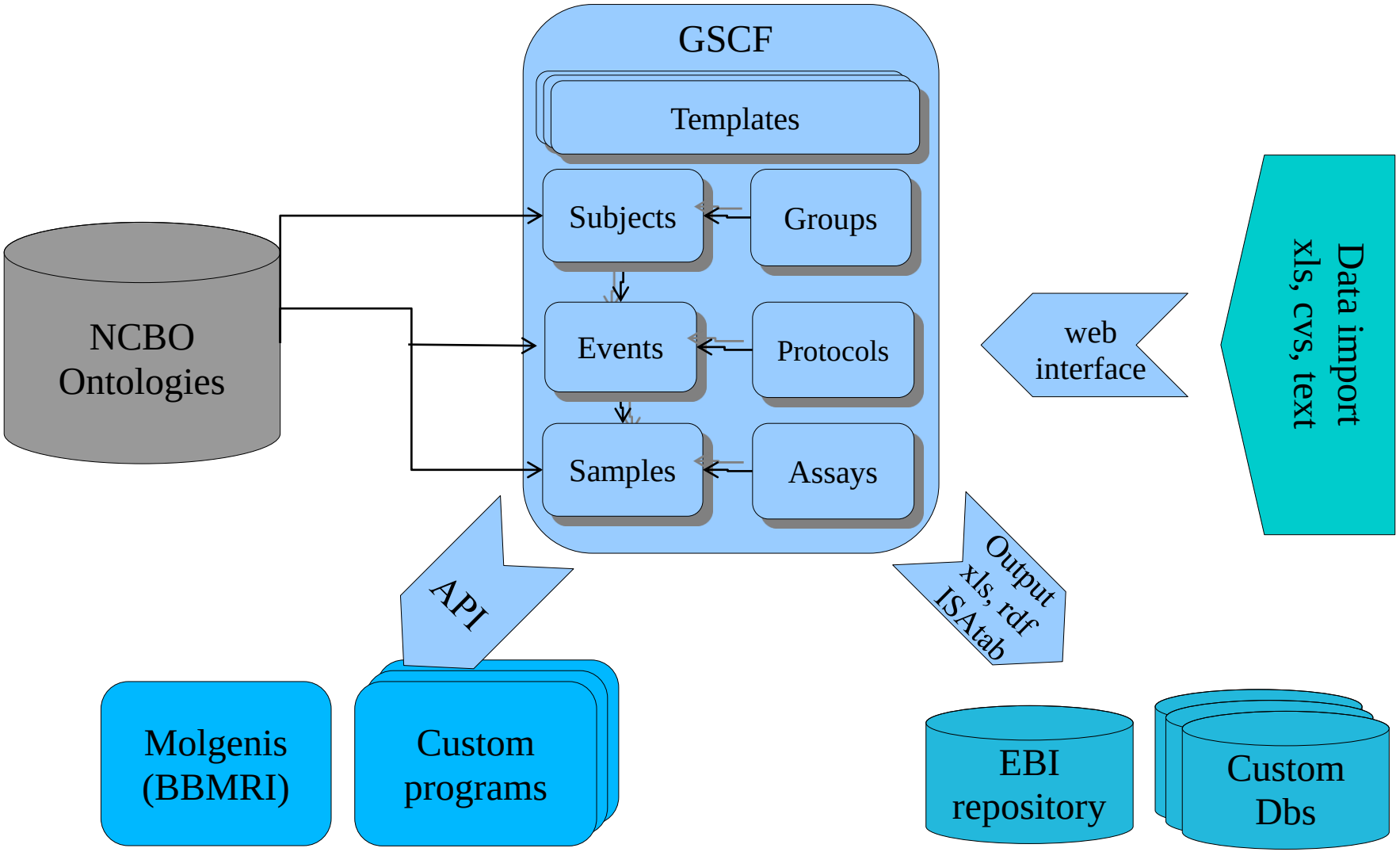
Affiliations | **Corresponding author**

Nature Genetics **44**, 121–126 (2012) | doi:10.1038/ng.1054
Published online 27 January 2012

A common standard



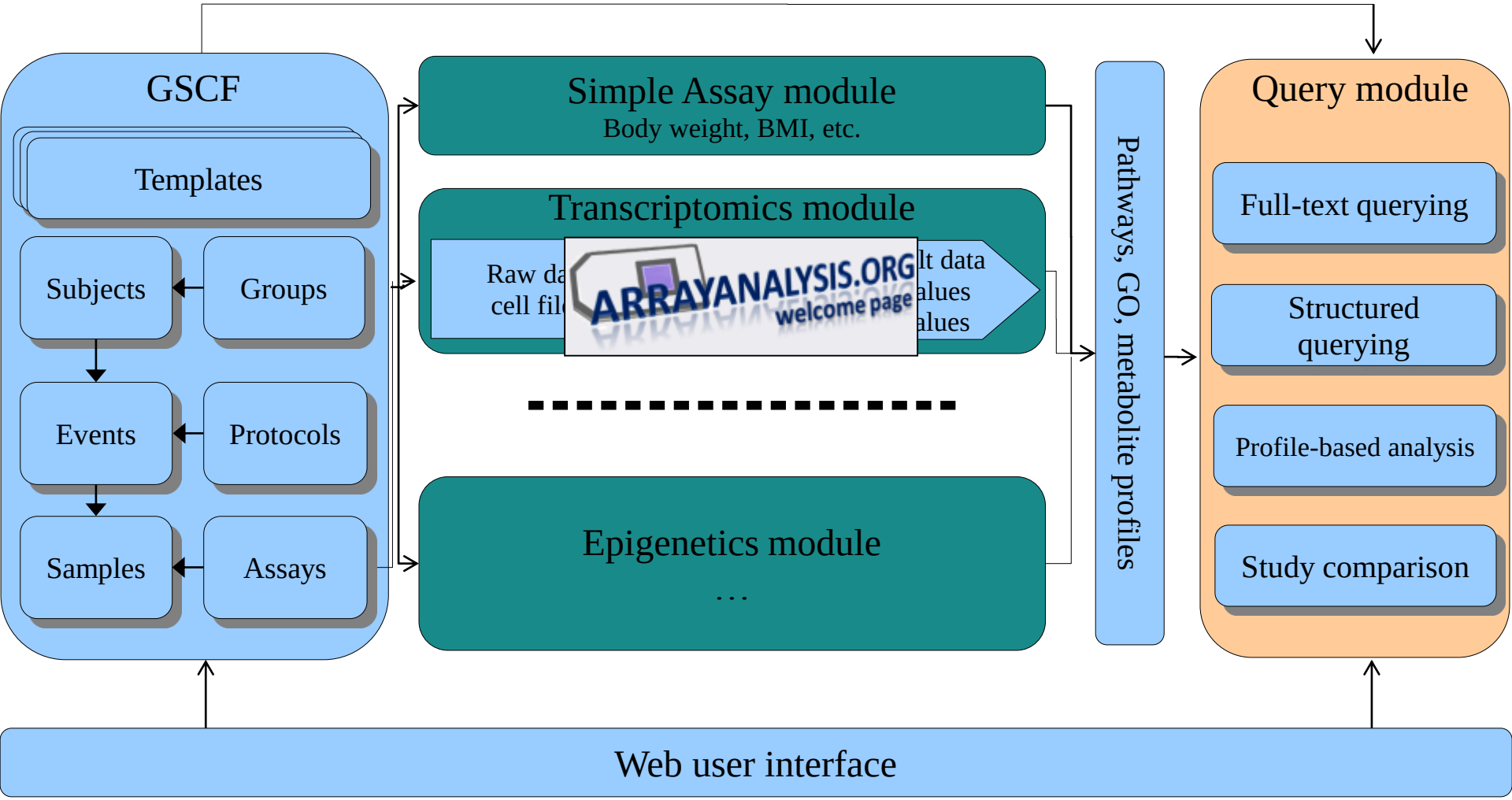
Generic Study Capture Framework
Data input / output



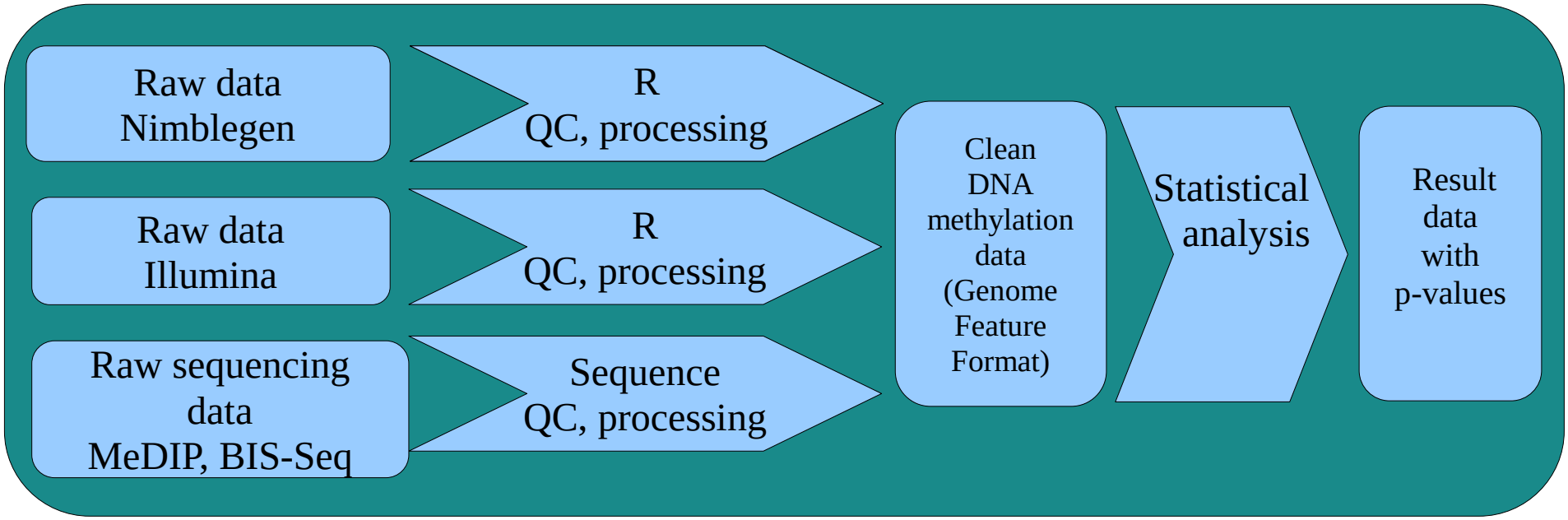
The study description points to:

The raw data	• And is the raw data provenance
The data processing protocols	• Which together is the processed data provenance
The processed data	• Typically two levels: clean and interpreted
The integrated, evaluated data	• For which we need knowledge based models

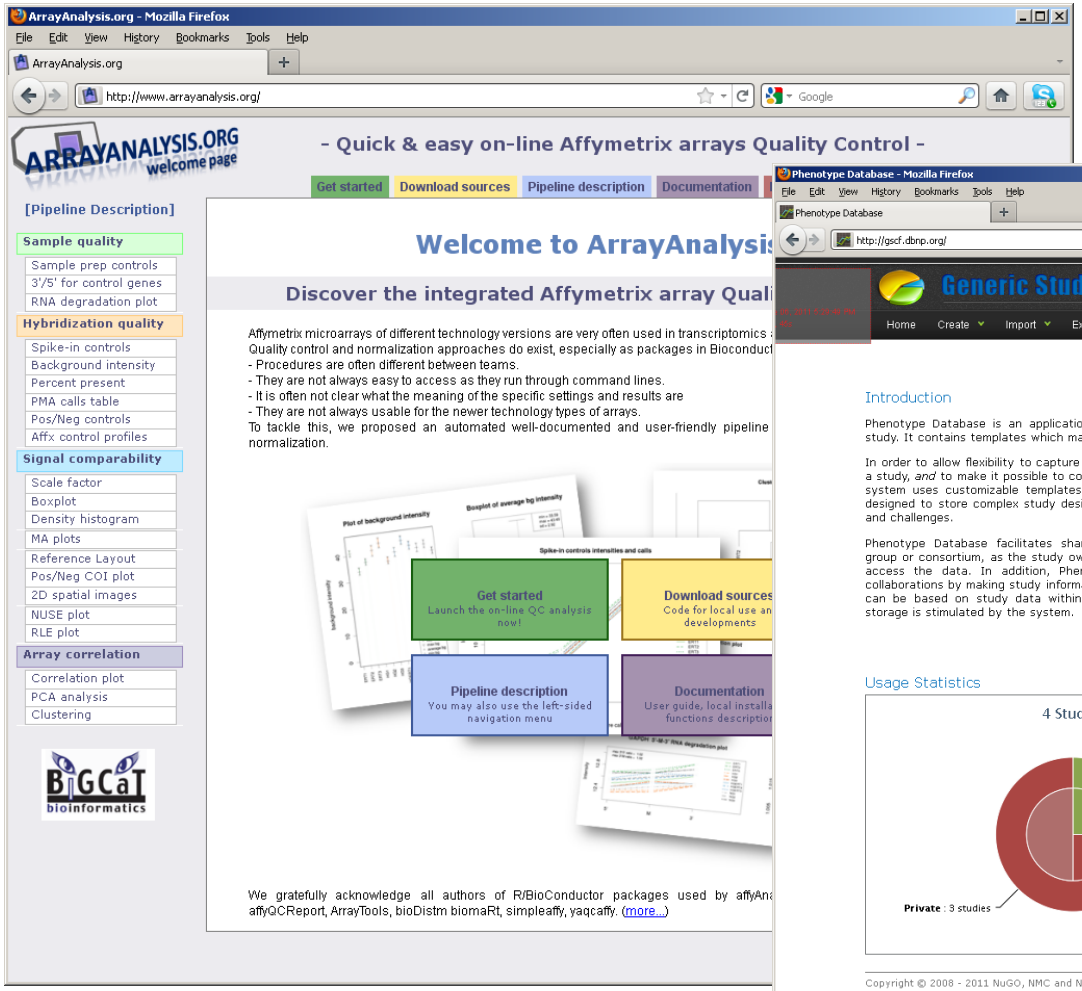
dbNP Architecture



Epigenetics DNA Methylation Pipeline

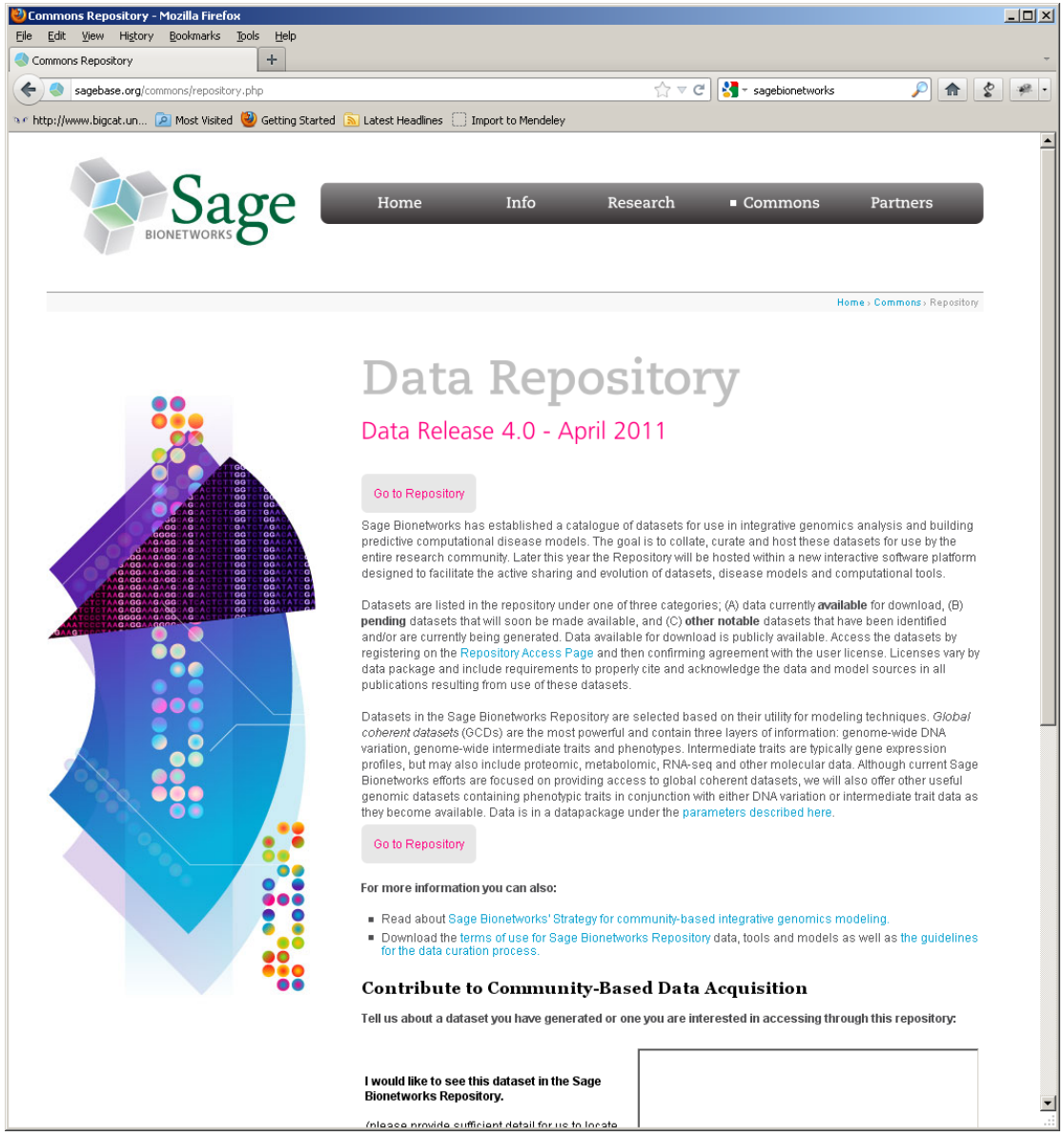


Process the data…

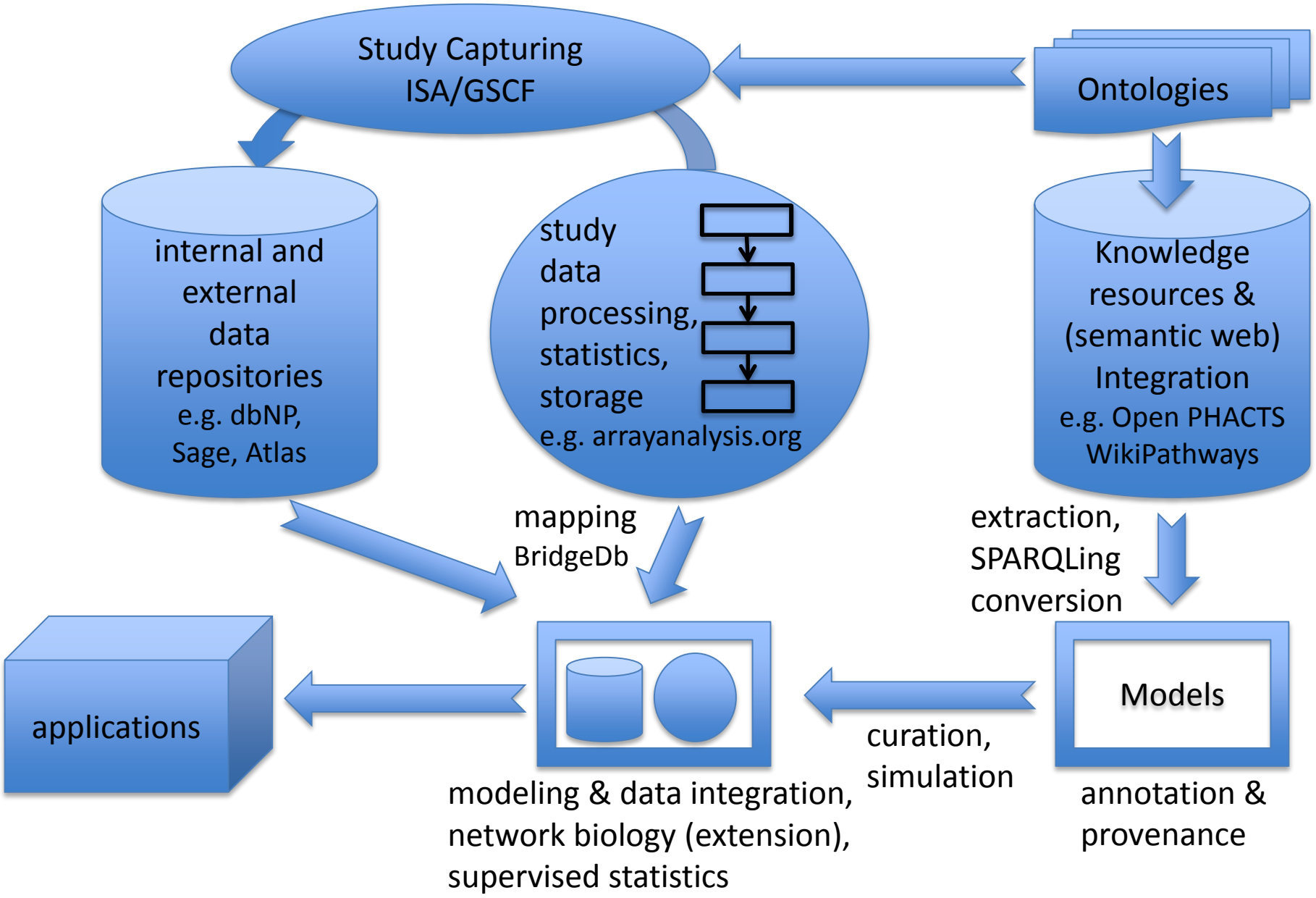


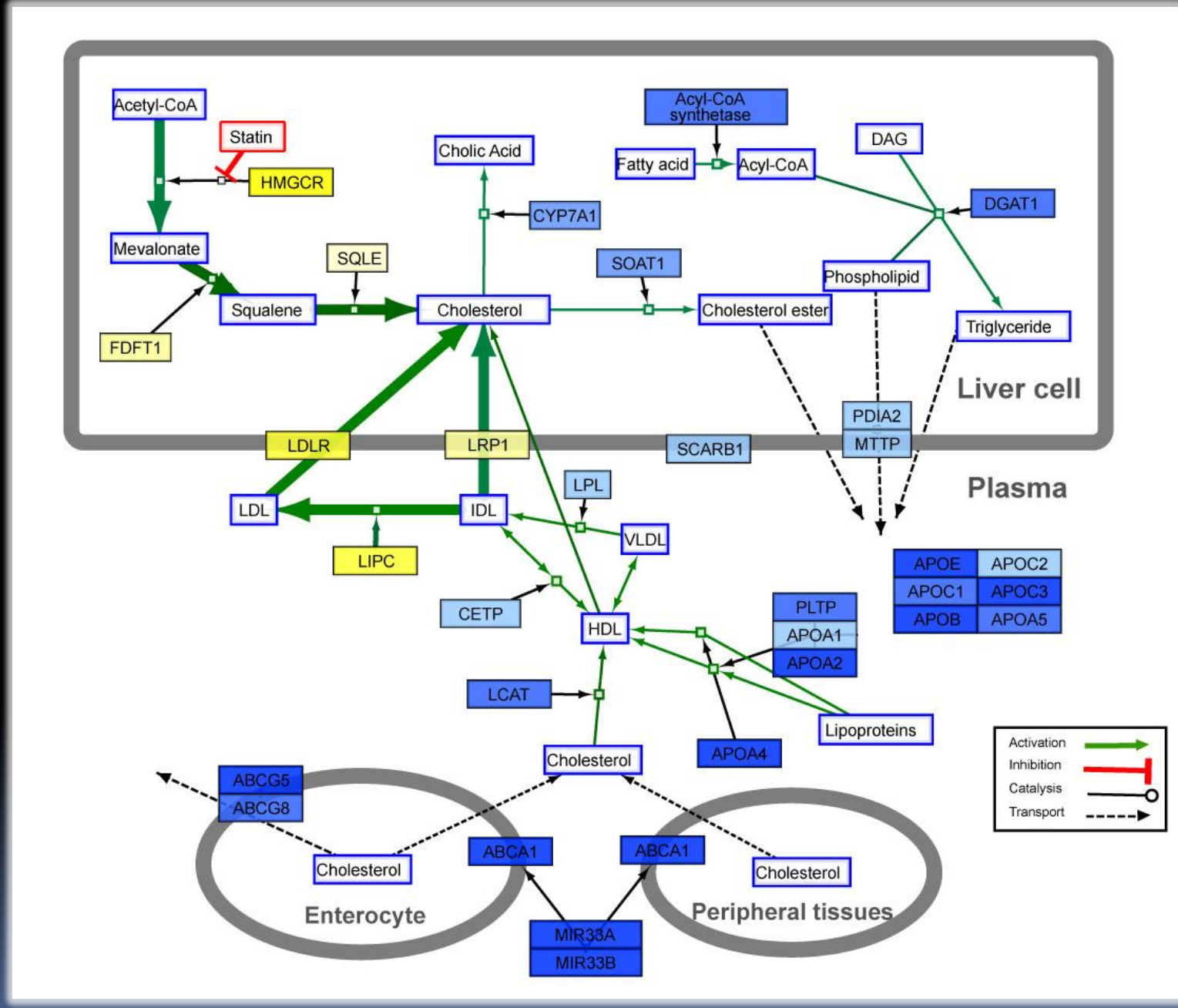
Fairports: connecting to other data

We both need Study Capturing



BiGCaT approach to Bioinformatics for Integrative Systems Biology



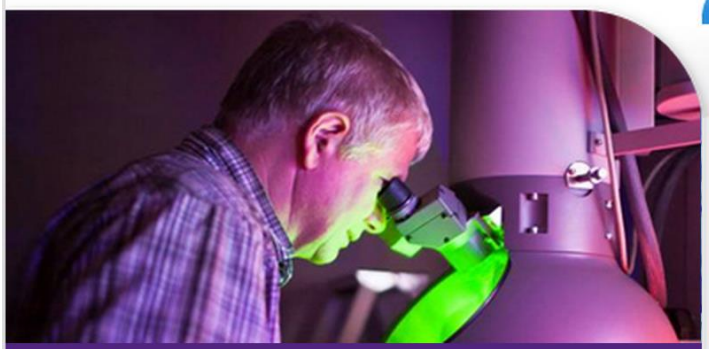


http://www.wikipathways.org/index.php/Pathway:WP430

DTL

DUTCH TECHCENTRE FOR LIFE SCIENCES

Search here...



Programmes | Activities | Expert groups | ELIXIR-NL | About DTL

NEWS

- Pan-European Imaging infrastructure gains momentum
- DTL Focus meeting: NGS Data Storage and Sharing
- CGtag: complete genomics toolkit and annotation in a cloud-based Galaxy
- Proposal on Open Discovery and Exchange of (big) data in Life sciences funded for 2 years
- News overview

EVENTS

- 08.04.2014 | Netherlands Bioinformatics Conference 2014
- 09.04.2014 | Join ELIXIR-NL session at NBIC 2014
- 14.04.2014 | DTL Focus meeting: Proteomics bioinformatics
- 15.04.2014 | DTL Focus Meeting: 'NGS Production Pipelines'
- Events overview

DUTCH TECHCENTRE FOR LIFE SCIENCES (DTL)

DTL Headlines:

- ZonMw and NWO-ALW award 35 projects access to DTL technology hotels
- Netherlands joins international ELIXIR Consortium on biological data

The Dutch Techcentre for Lifesciences (DTL) focuses on the high-end technologies that drive next generation life science research in the clinical & health, nutrition, agro-genomics and industrial microbiology sectors. DTL **aims** to establish a world-class research infrastructure enabling the Netherlands to remain at the forefront of biology research and innovation.

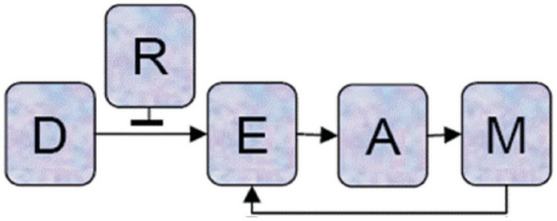
DTL is deeply embedded in the Dutch life sciences community and has been **founded** with strong involvement from universities, university medical centres, research institutes, science policy makers & funders and industry. Their joint **mission** is to create a platform in DTL where they can bundle their efforts to establish an integrated research infrastructure that is cross-technology, cross-sector, cost effective and that stimulates international collaboration.

DTL focuses its efforts to develop, combine and apply technologies in three closely interlinked **DTL|programmes**:

- DTL|Data
- DTL|Technologies
- DTL|Learning

Over 100 **expert groups** collaborate within these programmes to share and provide access to their advanced research facilities and expertise.

As part of its data activities, DTL hosts the Dutch node of **ELIXIR**, the European initiative that builds a sustainable international infrastructure for biological information.



Announcing the 2013 DREAM8.5 Challenges

We are pleased to announce three new DREAM8.5 challenges. Best performers in all DREAM8.5 Challenges will be invited to present at the 2014 DREAM conference (date and location to be determined). We are also working to establish publishing partners for each of these challenges. The DREAM8.5 Challenges are now open for registration, and will begin active problem-solving in late 2013 or early 2014.

Click on a link below to read the Challenge detail and register for a DREAM8.5 Challenge.

Alzheimer's Disease Big Data DREAM Challenge #1

In the first of what will be a series of Alzheimer's Disease (AD) Big Data Challenges, participants will utilize data from the Alzheimer's Disease Neuroimaging Initiative (ADNI). Data will consist of cognitive, imaging, biological, and whole genome sequencing data on cohorts of volunteers, who range from cognitively normal, mild cognitive impairment and dementia. Participants will analyze the data to solve two sub-challenges: (1) Build a model that best predicts change over time in AD cognitive scores using all available test and adjacent data, and (2) Build a model that best predicts discordance between biomarkers suggestive of amyloid perturbations and lack of cognitive impairment. These models will be used to better understand the biomolecular mechanisms leading to Alzheimer's disease, and ultimately to develop new therapies. We expect to announce a publishing partner for this Challenge shortly.

ICGC-TCGA-DREAM Somatic Mutation Calling Challenge

Working with technology partners Google and Annai, we will provide 9 terabytes of raw human sequence data derived from pairs of normal and tumor tissue (from prostate and pancreas). Approved participants will analyze the data to solve two sub-challenges: (1) build a model that accurately predicts cancer mutations that alter a single nucleotide in the genome (single nucleotide variants, SNVs) (2) Build a model that accurately predicts cancer mutations that alter the order of a large stretch of the genome (i.e. a structural variation, SV), such as a rearrangement, inversion or copy-number aberration. Improving the algorithms that correctly identify these variations is important because these variations provide key genetic data which can be used by predictive models to guide personalized cancer therapies. *Nature Publishing Group* enthusiastically welcomes the opportunity to consider for publication work that achieves best performance in this Challenge.

The Rheumatoid Arthritis Responder Challenge

Participants will have access to whole genome genotype data (2.2 million SNPs) and clinical data collected from 2,000 individuals with Rheumatoid Arthritis who have been treated with anti-TNF therapy. Up to one third of these patients fail to enter clinical remission. Participants will use these data to solve two sub-challenges: (1) Build a model that best predicts treatment response as measured by the change in disease activity score (DAS28) in response to anti-TNF therapy, and (2) Build a model that best predicts poor responders as defined by specific criteria (yet to be specified). The winning model will be the one that can predict the largest portion of this sample subset (~10% of the population) with a positive predictive value greater than a predetermined