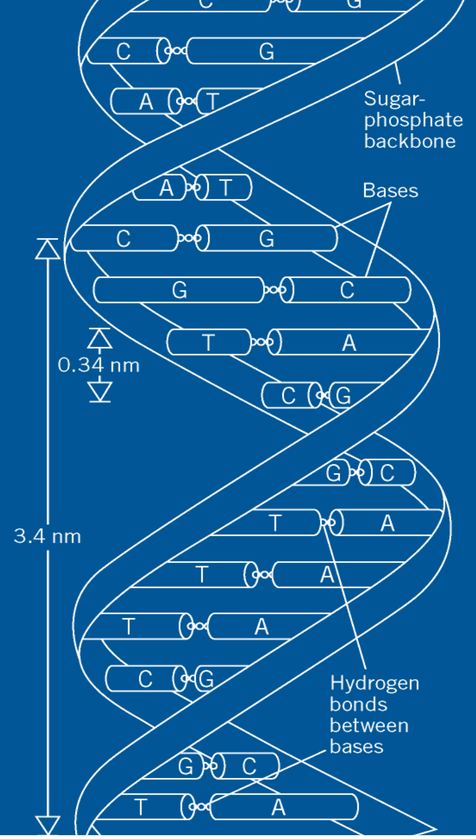


Facilitating the collaborative pursuit of scientific discovery with The FacileData Ecosystem

Steve Lianoglou
Computational Biologist
Cancer Immunology
Genentech

@slianoglou
@lianos
@faciledata

These tools will be open sourced

https://github.com/faciledata

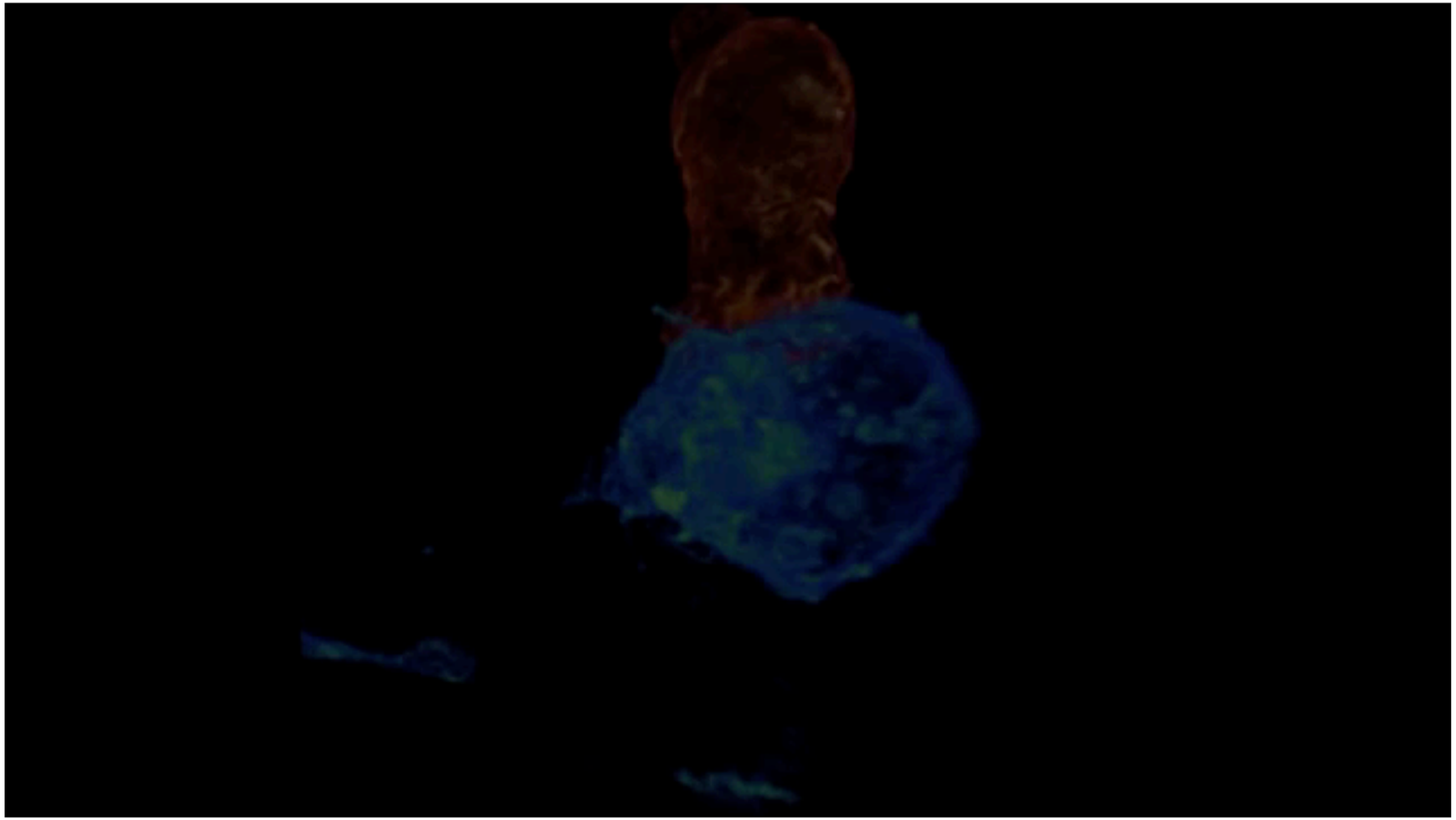
Slides and code used for presentation

https://github.com/faciledata/talks/2017-plotcon

★ star `faciledata/talks` repository to be updated ★

when tools are released

Cancer Immunology: the immune system can combat cancer



Alex Ritter

We can read DNA at scale

Illumina HiSeq

In a single run (4 days)*
12 Genomes
100 Transcriptomes

* https://www.illumina.com/systems/sequencing-platforms/hiseq-3000-4000/specifications.html

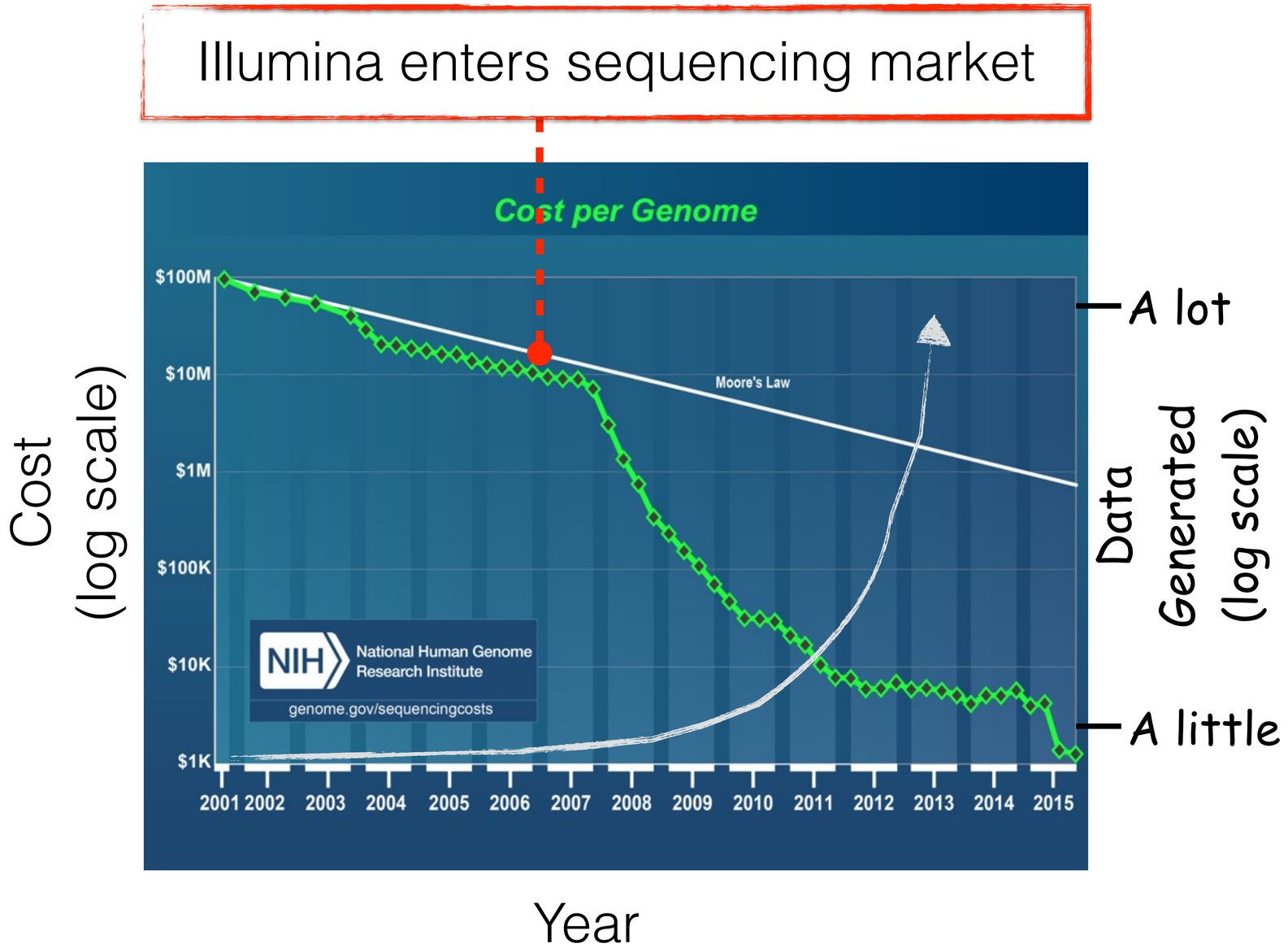
Fast and accurate DNA sequencing revolutionized study of biology

- **Genome sequencing**: identifies alterations in DNA
  - Mutations
  - Insertions, Deletions, Amplifications
  - Rearrangements
- **Transcriptome sequencing**: characterizes the identify and abundance of genes (RNA) expressed in a tissue or cell
  - Understand the mechanisms by which the cell performs its function
  - Understand what makes a tumor different than normal tissue
  - Understand why a tumor responds (or not) to a particular treatment

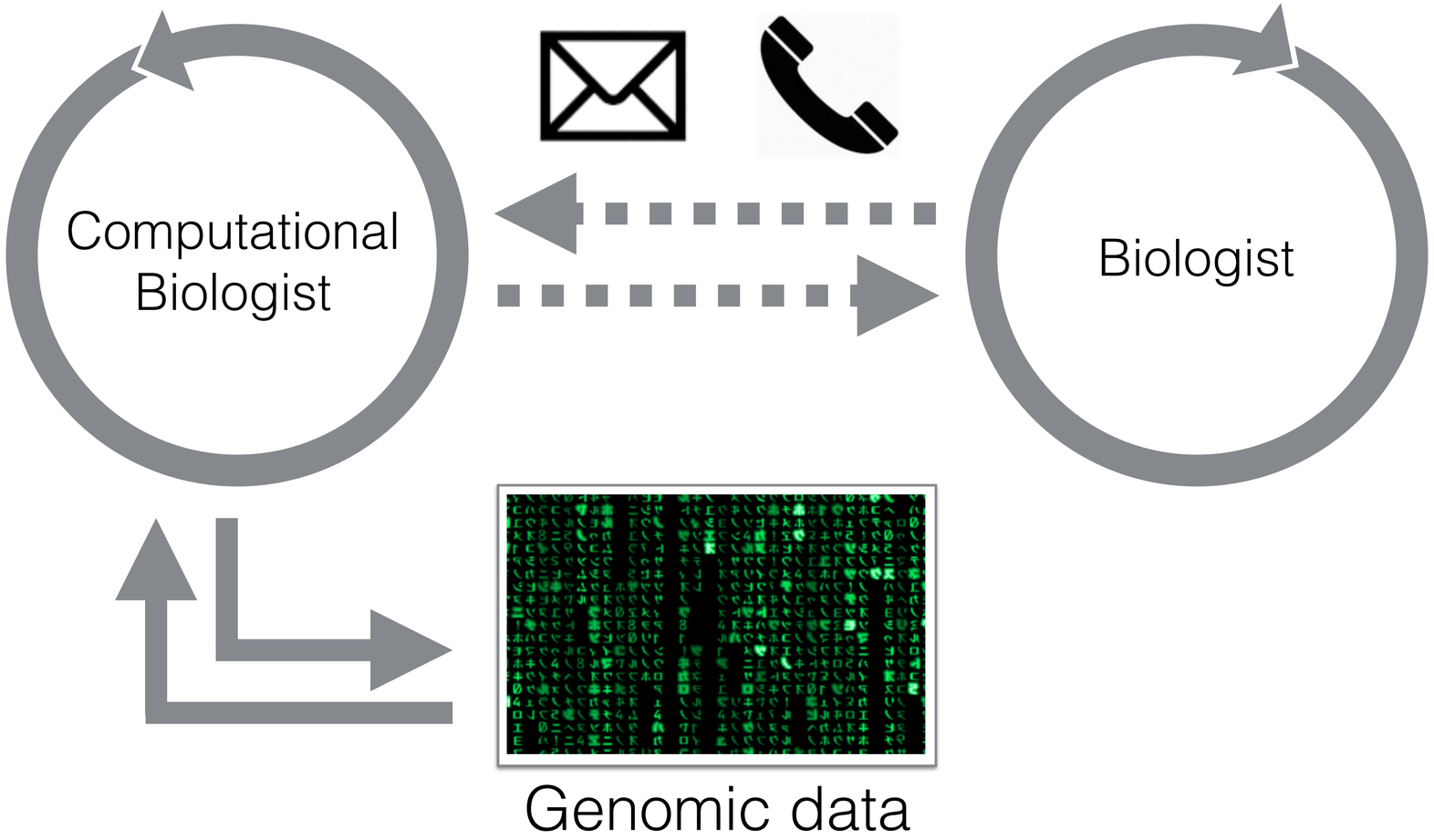
The universe of things the genome can do

The thing the genome is doing right now

Biology has been transformed into a data driven science

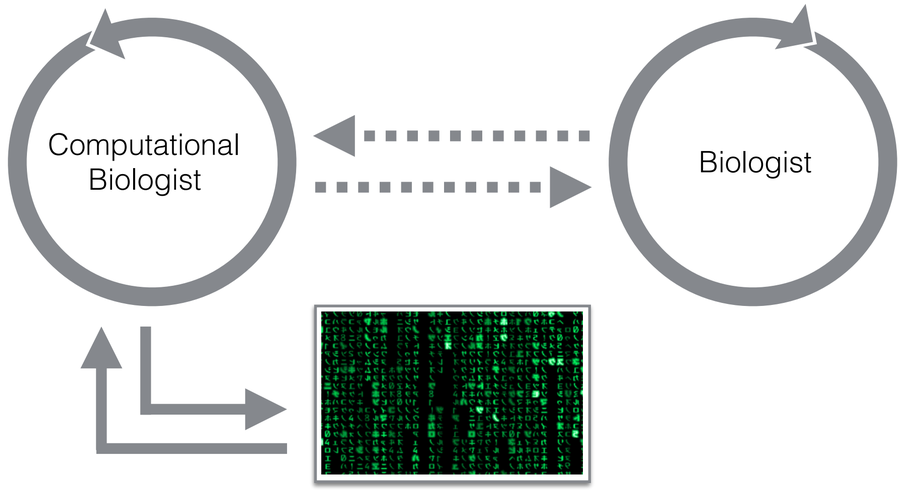


The state of data driven exploration and discovery



Problems with current state of data driven exploration

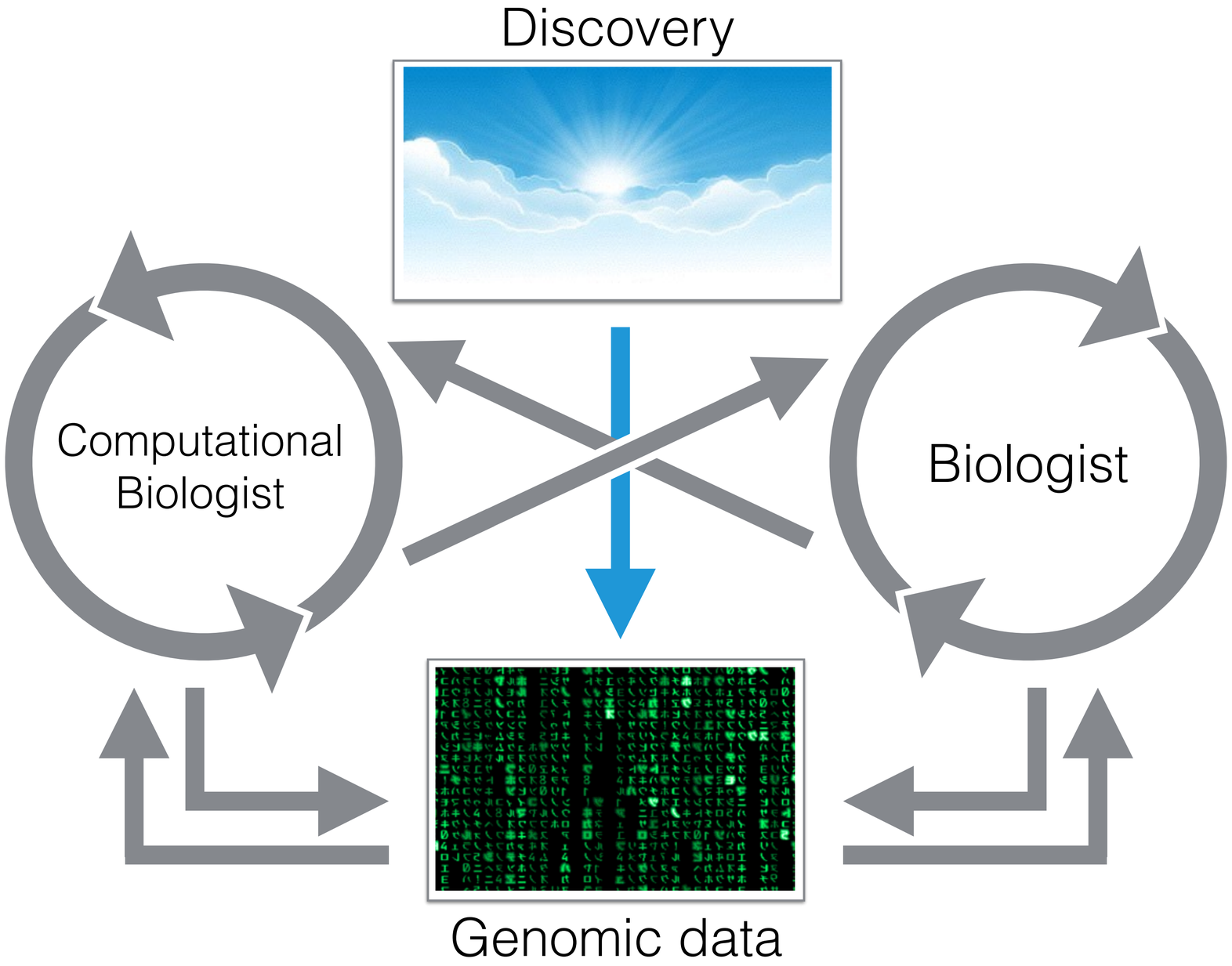
- Biologist has limited ways to meaningfully interact directly with the data.
  - Relies on computational biologist for even most simple of EDA
  - Long turn around times (Colin Ware: "delay creates a universe of ideas lost")
  - Context switches and broken *flow* for all



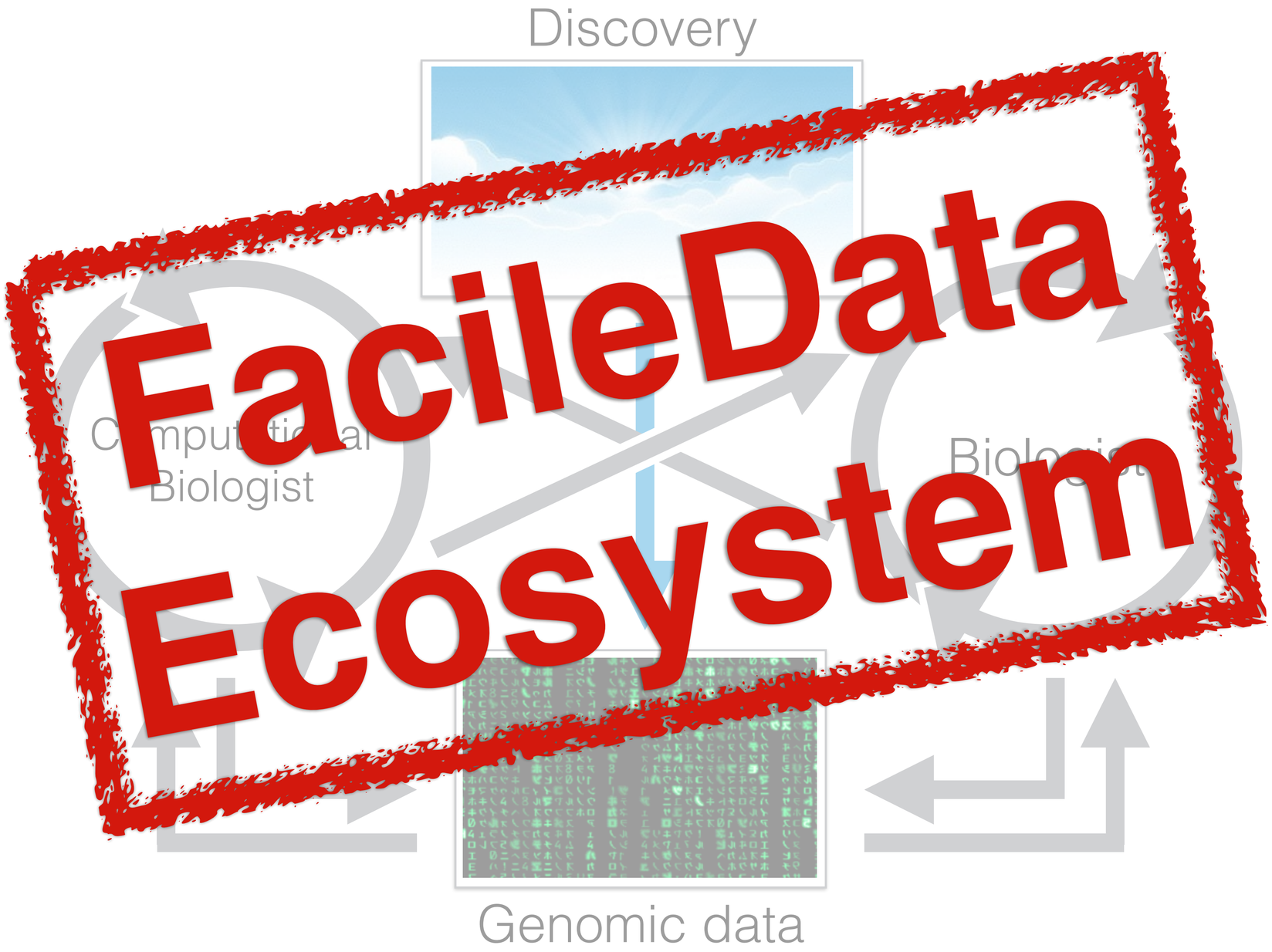
- The units of "knowledge transfer" are emails, tables of statistics, and figures **divorced from the data** (even if using Rmarkdown)
  - Makes it hard to iterate on EDA between collaborators

This is a *service model* not a *collaborative model*

A model for collaborative data driven exploration and discovery



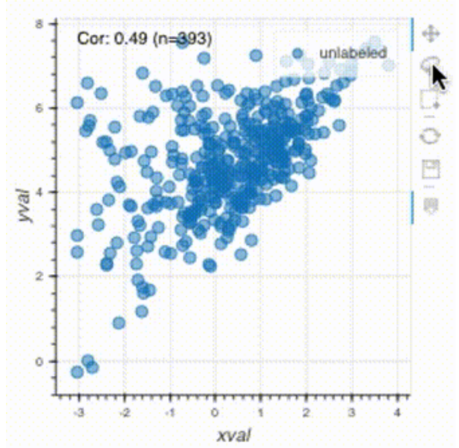
A model for collaborative data driven exploration and discovery



Tools to facilitate collaborative exploration

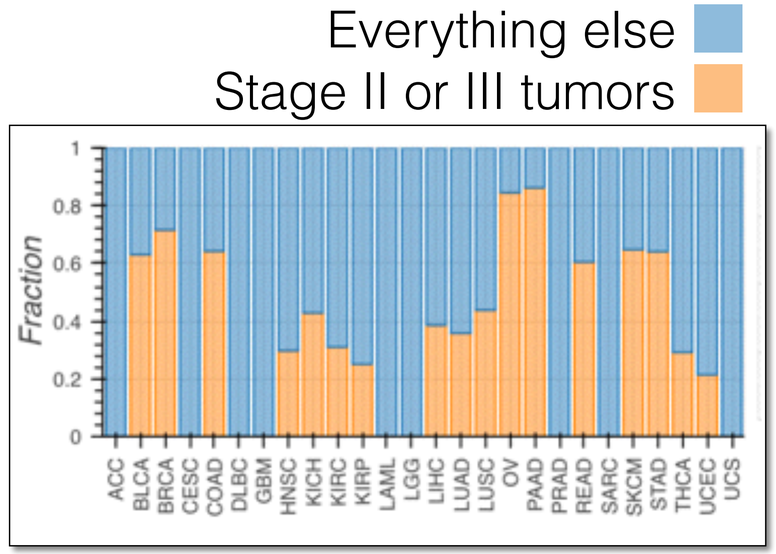
The FacileData Frontend (FacileExplorer)

- Designed to enable **sustained and independent** interactive data exploration by non-informaticians
- Empowers users to compute over data via its GUI
- Provides ability to hand off analyses "in flight", not just results



The FacileData Backend (FacileDataSet)

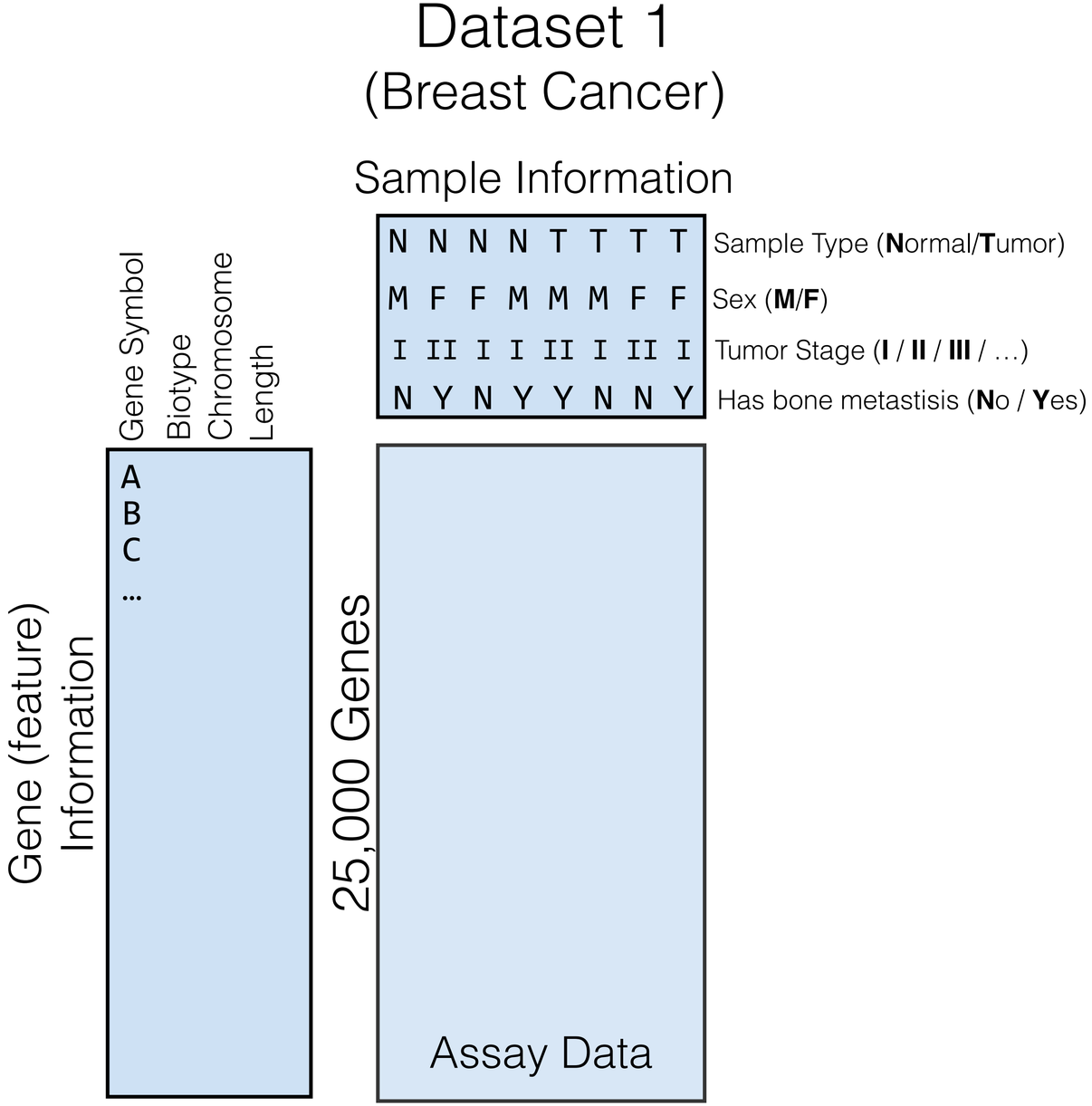
- Consolidates different high-throughput genomics datasets behind a single point of access
- Fast and efficient query and retrieval of **arbitrary data subsets** (features and samples) across datasets.
- Provides a **covariate-centric view** over these data
- Provides a data access API conducive to exploratory data analysis via code or GUI (dplyr/tidy-*ish*)



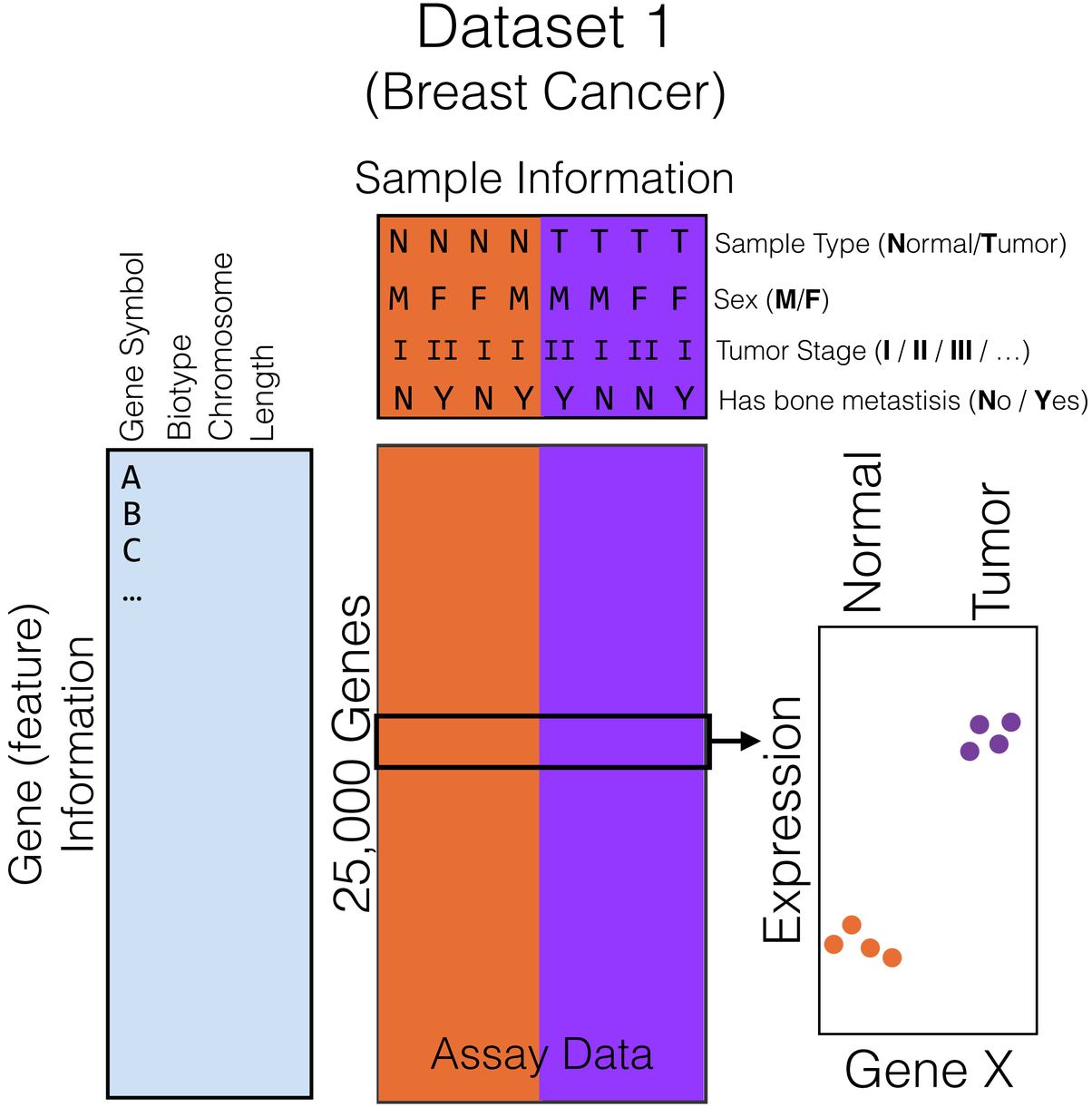
3321 / 8024 TCGA samples

The FacileData Backend (FacileDataSet)

High-throughput genomics data is large and complex

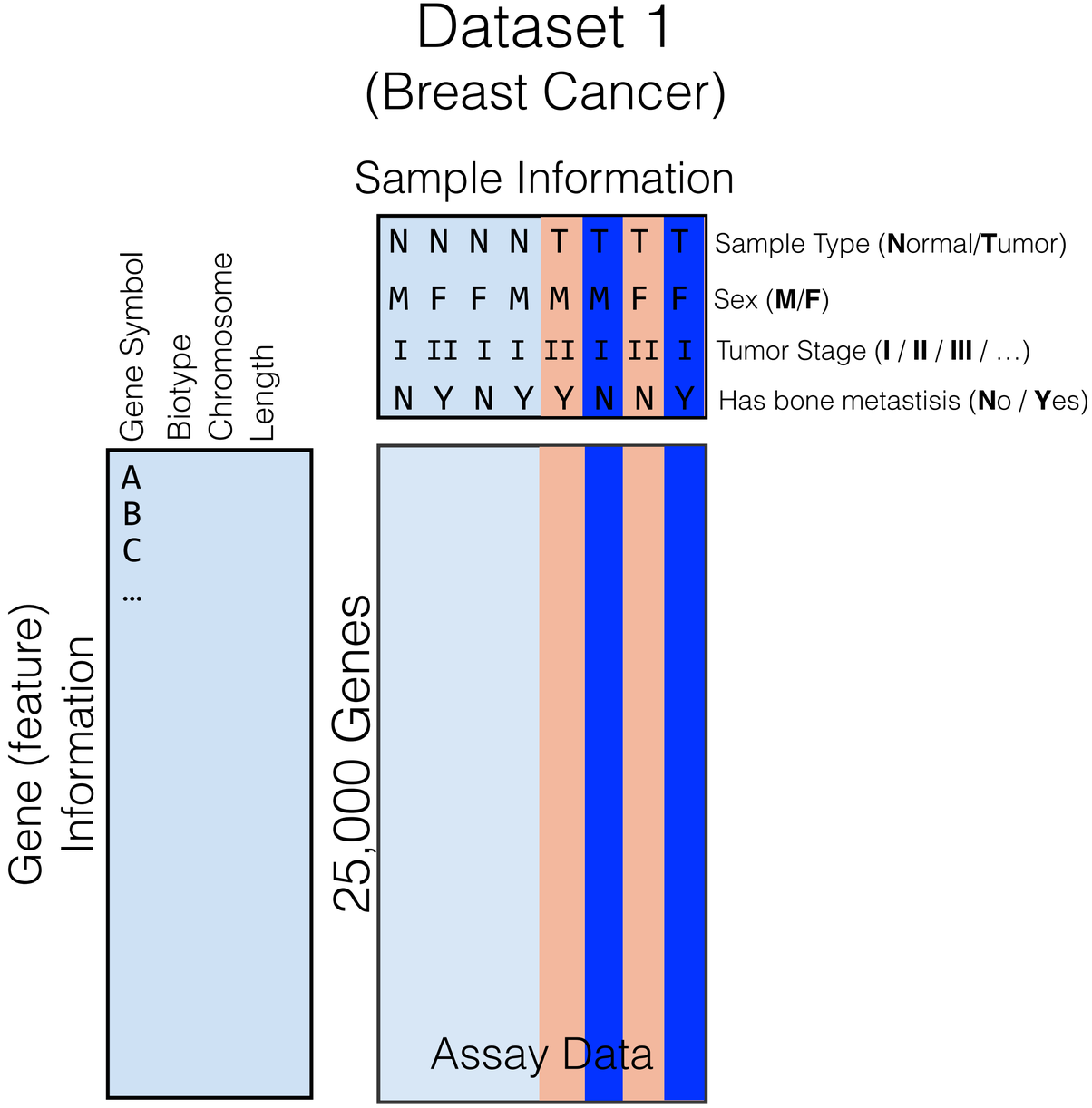


High-throughput genomics data is large and complex



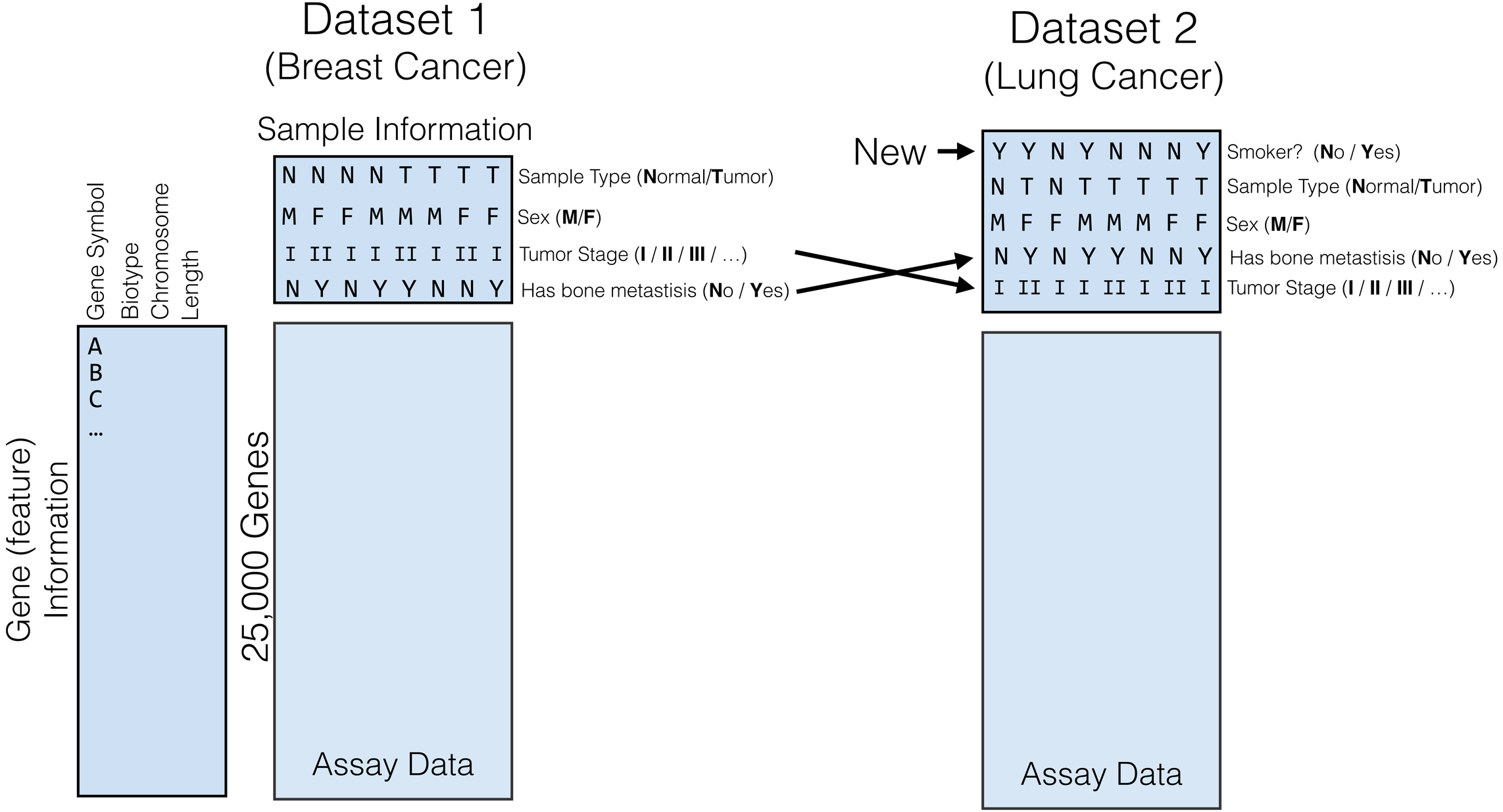
Normals vs Tumors

High-throughput genomics data is large and complex

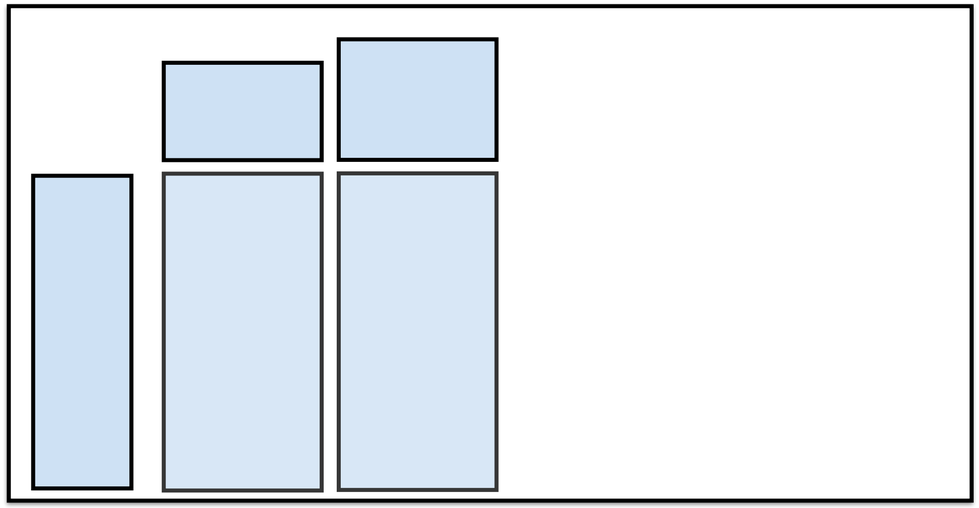


Stage I vs Stage II

High-throughput genomics data is large and complex

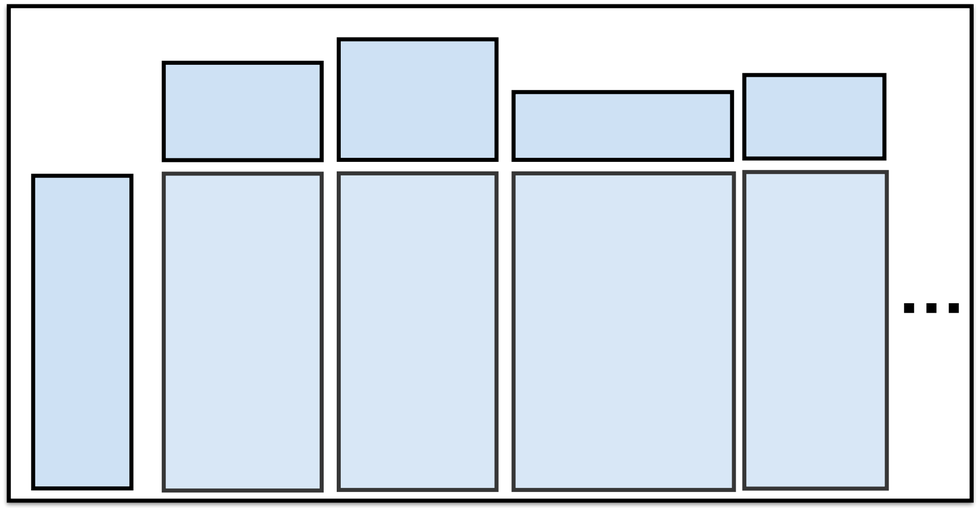


FacileDataSet provides single point of access across multiple datasets



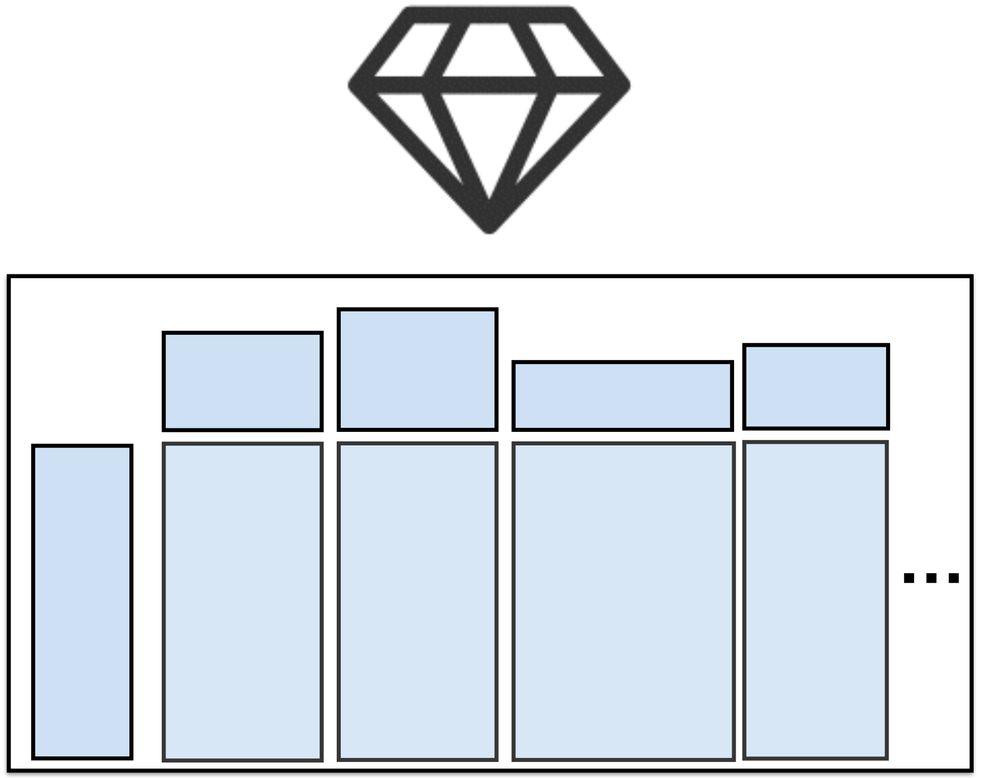
FacileDataSet

FacileDataSet provides single point of access across multiple datasets



FacileDataSet

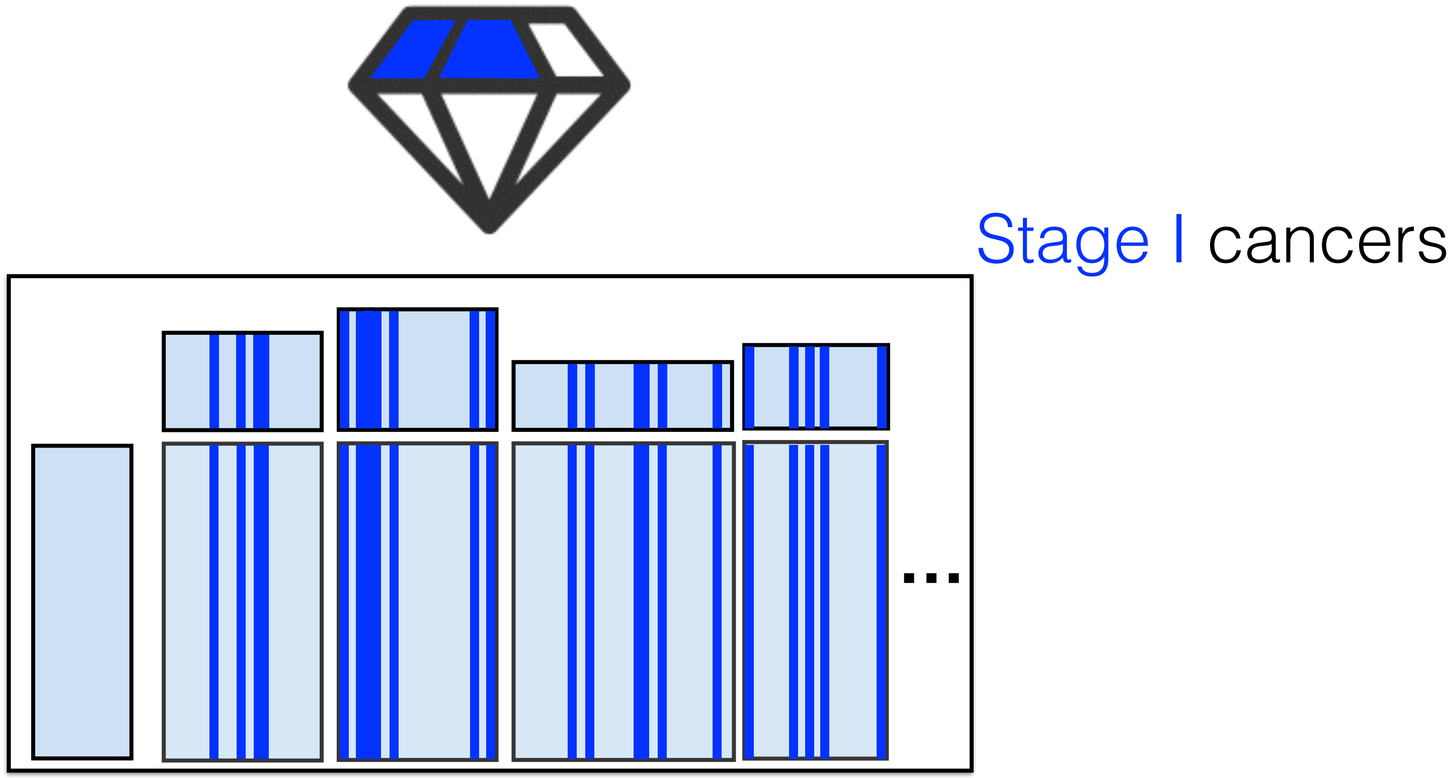
FacileDataSet provides a covariate-centric interface to assay data across multiple datasets



FacileDataSet

Think of data as a multi-faceted diamond

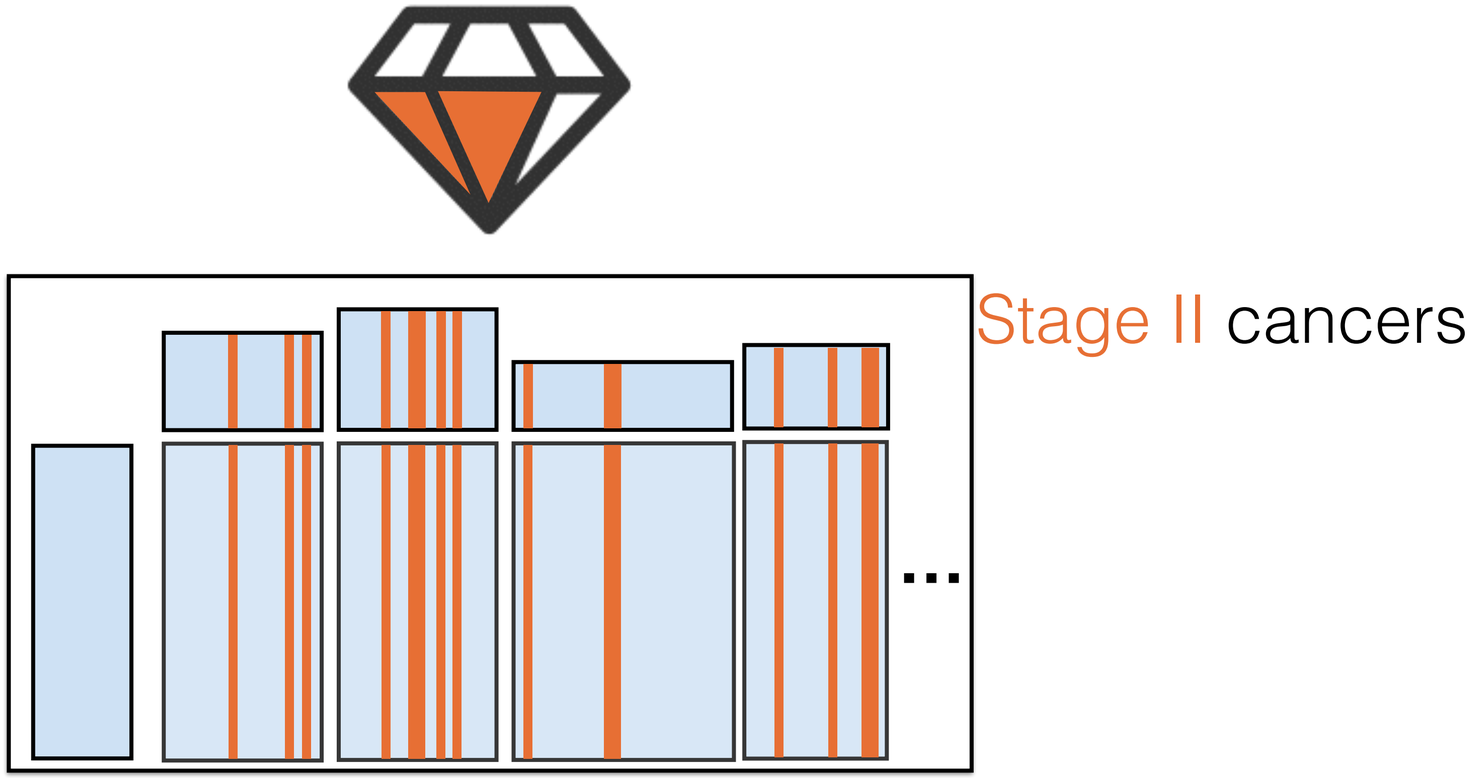
FacileDataSet provides a covariate-centric interface to assay data across multiple datasets



FacileDataSet

Think of data as a multi-faceted diamond
Facets are defined by combinations of covariates

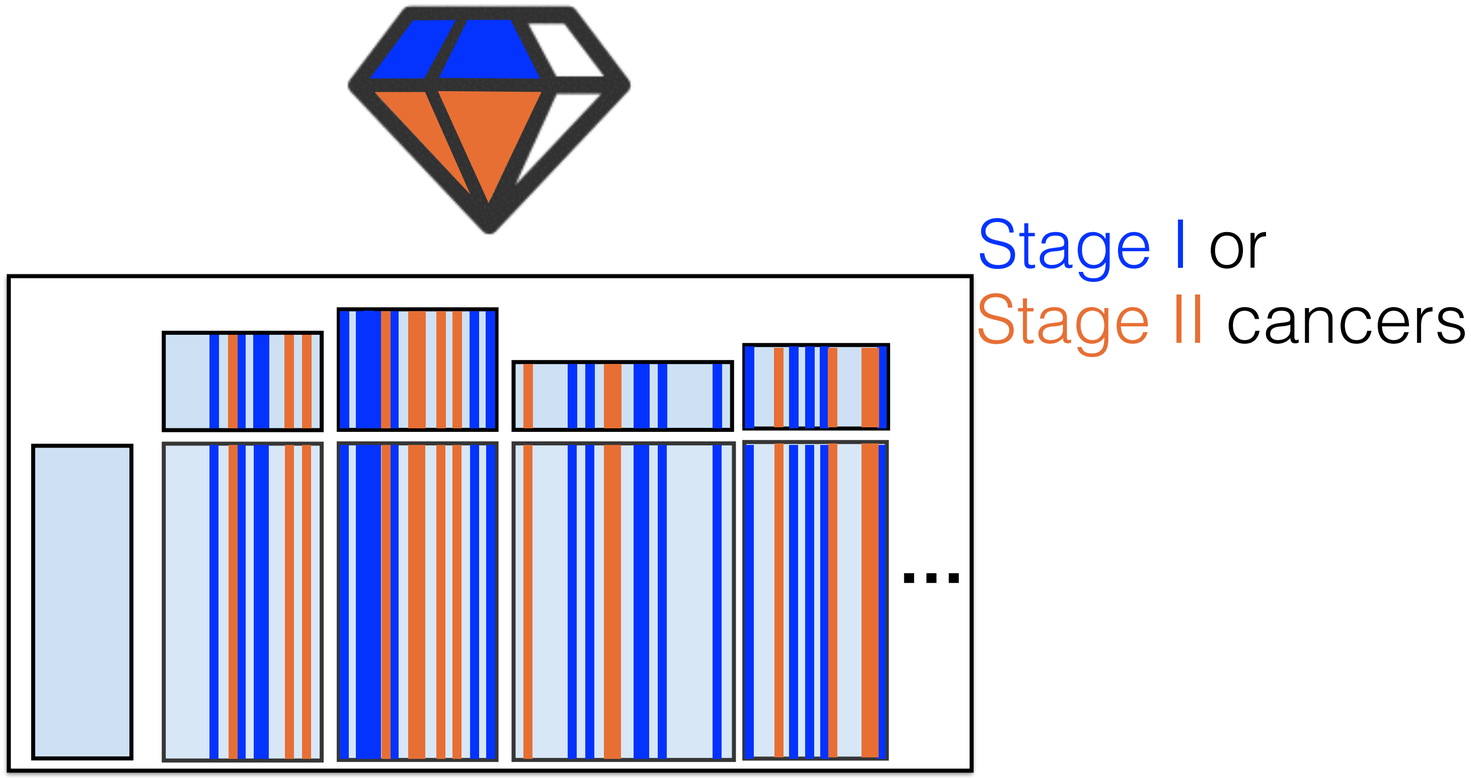
FacileDataSet provides a covariate-centric interface to assay data across multiple datasets



FacileDataSet

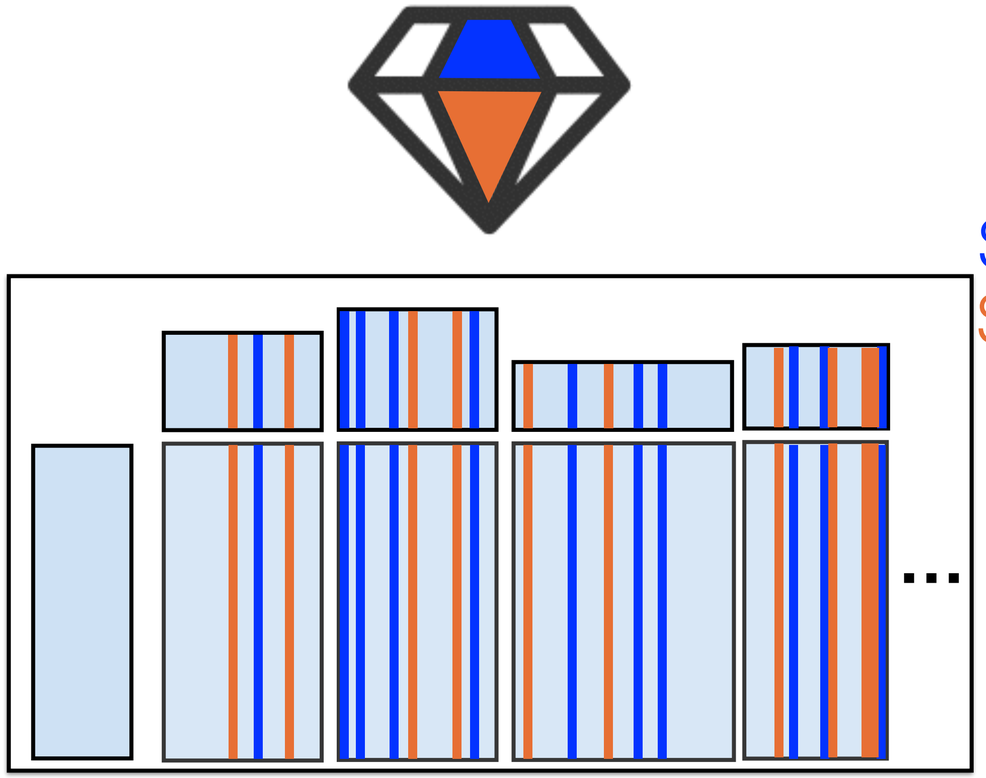
Think of data as a multi-faceted diamond
Facets are defined by combinations of covariates

FacileDataSet provides a covariate-centric interface to assay data across multiple datasets



Think of data as a multi-faceted diamond
Facets are defined by combinations of covariates

FacileDataSet provides a covariate-centric interface to assay data across multiple datasets



Stage I or
Stage II cancers

… that respond to treatment X

FacileDataSet

The FacileData Frontend (FacileExplorer)

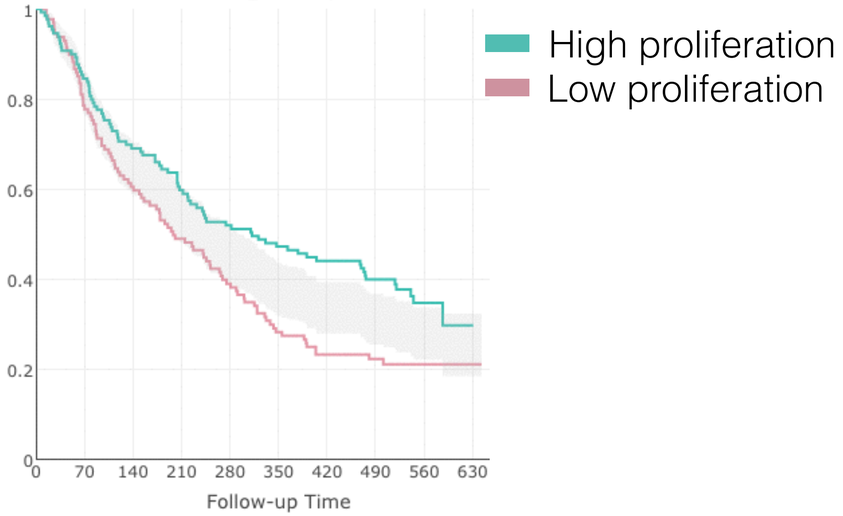
Anatomy of a Facile Survival Analysis

These results are hypothetical

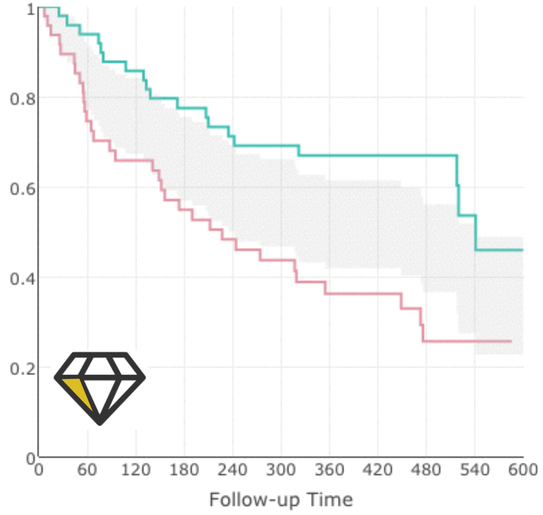
- User decides to analyze lung squamous tumors in TCGA FacileDataSet
- User asks if gene signature for proliferation is prognostic biomarker for overall survival
- User "one-clicks" an *Iterative Faceted Analysis* to identify other facets of data this result generalizes to

TCGA

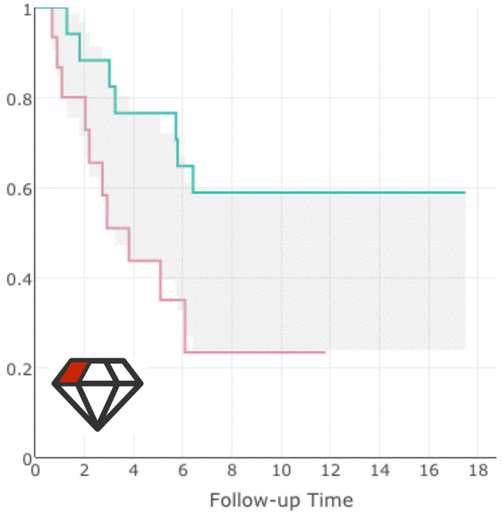
TCGA Lung Squamous



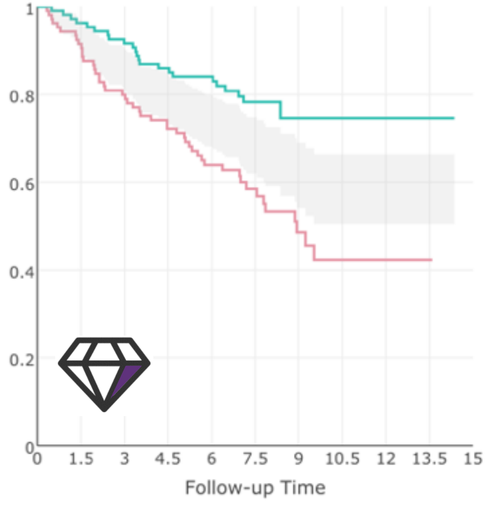
TCGA Lung Adenosarcoma



TCGA: ER+ Breast



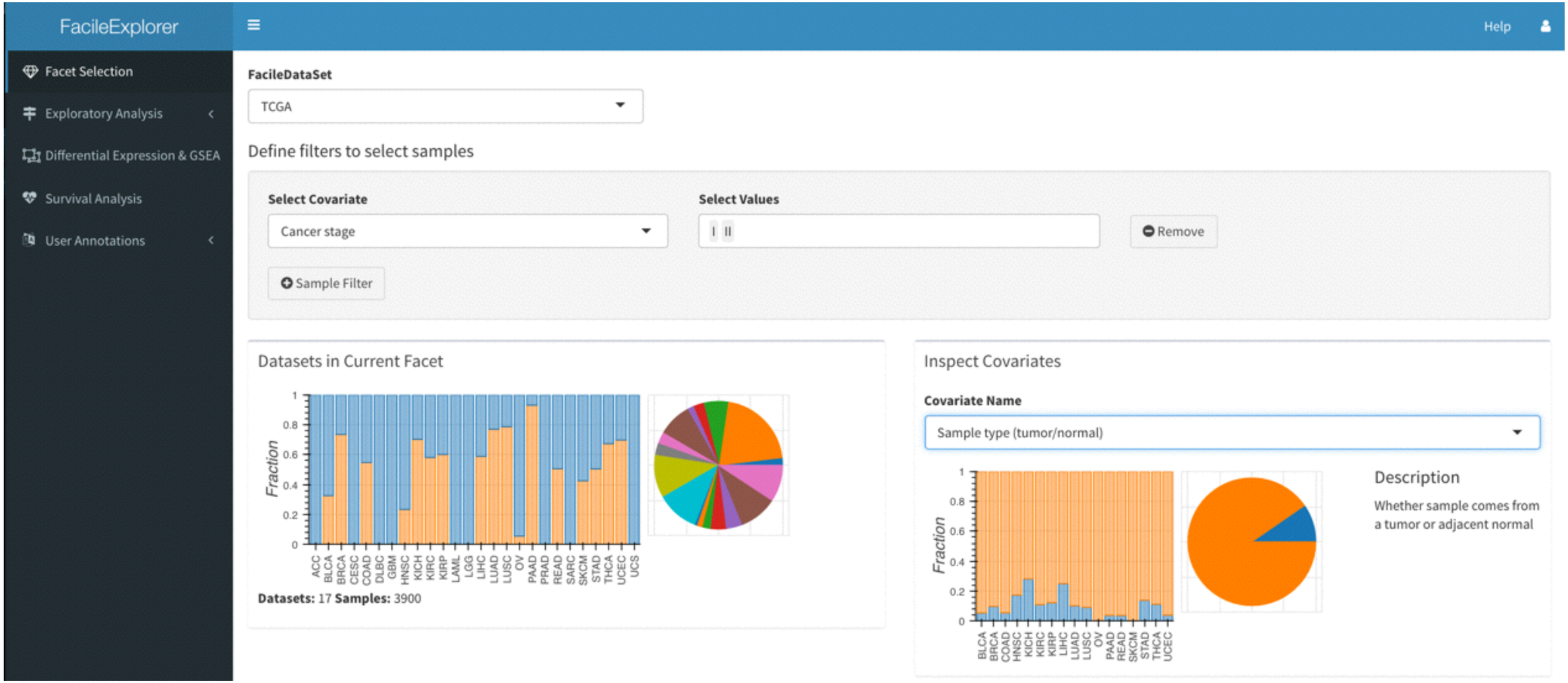
TCGA: Bladder



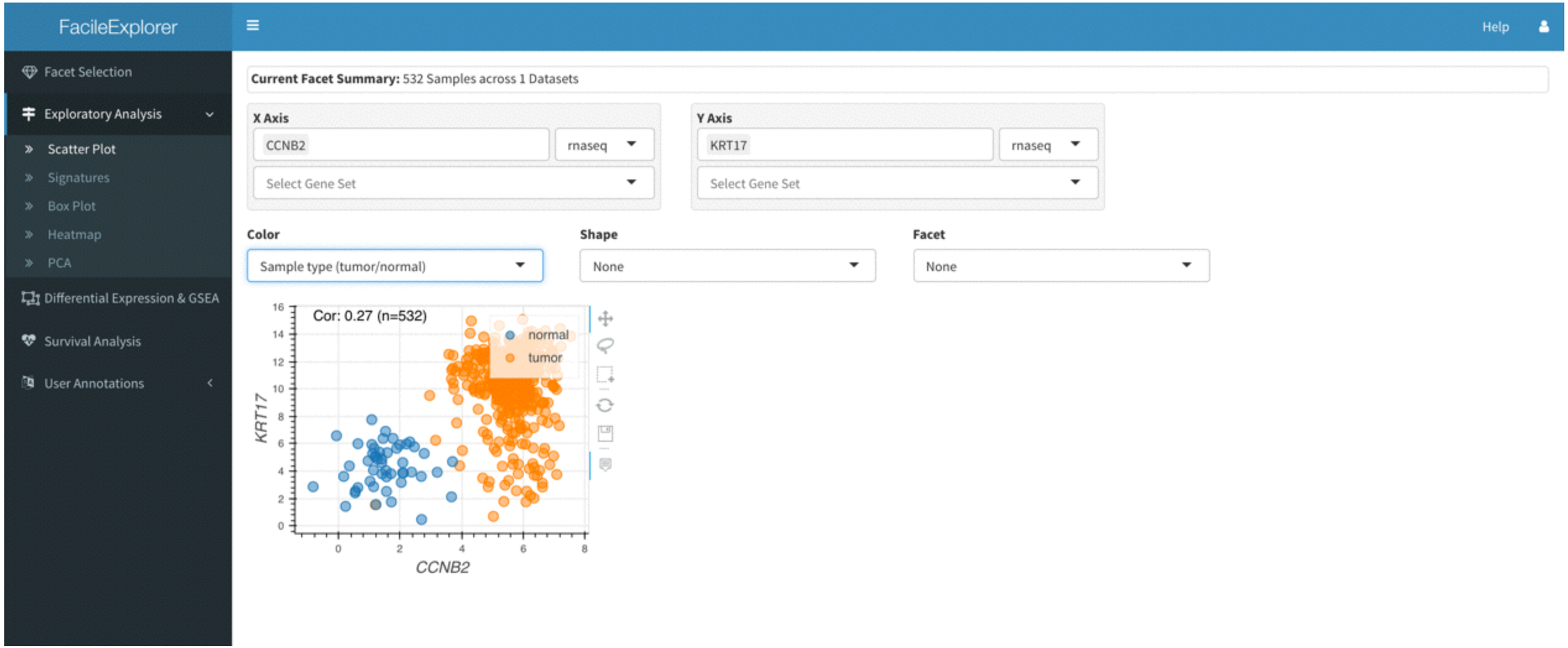
Live demo of FacileExplorer

GUI → Analyst → GUI

Live Demo: Exploring Facets of TCGA



Live Demo: Single gene markers of Proliferation vs P53 response in TCGA Lung Squamous (tumors and normals)



Live Demo: Recover brushed annotations and generate graph via code


```
library(FacileExplorer)
library(FacileTCGADataSet)
tcga <- FacileTCGADataSet()

samples <- tcga %>%
  filter_samples(indication == "LUSC")

features <- tcga %>%
  filter_features(name %in% c("CCNB2", "KRT17")) %>%
  mutate(assay="rnaseq")

dat <- samples %>%
  with_sample_covariates('sample_type') %>%
  with_assay_data(features, normalized=TRUE, spread="name")

ggplot(dat, aes(CCNB2, KRT17, color=sample_type)) +
  geom_point() +
  theme(legend.position="bottom")
```

Live Demo: Recover brushed annotations and generate graph via code


```
library(FacileExplorer)
library(FacileTCGADataSet)
tcga <- FacileTCGADataSet()

samples <- tcga %>%
  filter_samples(indication == "LUSC")

features <- tcga %>%
  filter_features(name %in% c("CCNB2", "KRT17")) %>%
  mutate(assay="rnaseq")

dat <- samples %>%
  with_sample_covariates('sample_type') %>%
  with_assay_data(features, normalized=TRUE, spread="name")

ggplot(dat, aes(CCNB2, KRT17, color=sample_type)) +
  geom_point() +
  theme(legend.position="bottom")
```

Live Demo: Recover brushed annotations and generate graph via code

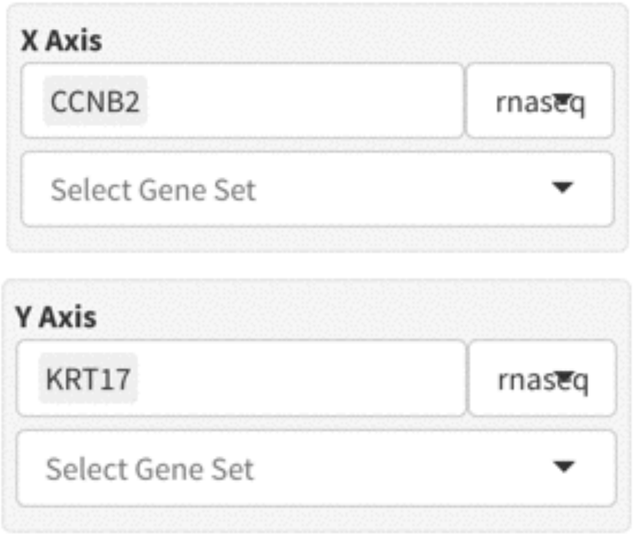

```
library(FacileExplorer)
library(FacileTCGADataSet)
tcga <- FacileTCGADataSet()

samples <- tcga %>%
  filter_samples(indication == "LUSC")

features <- tcga %>%
  filter_features(name %in% c("CCNB2", "KRT17")) %>%
  mutate(assay="rnaseq")

dat <- samples %>%
  with_sample_covariates('sample_type') %>%
  with_assay_data(features, normalized=TRUE, spread="name")

ggplot(dat, aes(CCNB2, KRT17, color=sample_type)) +
  geom_point() +
  theme(legend.position="bottom")
```



Live Demo: Recover brushed annotations and generate graph via code


```
library(FacileExplorer)
library(FacileTCGADataSet)
tcga <- FacileTCGADataSet()

samples <- tcga %>%
  filter_samples(indication ==

features <- tcga %>%
  filter_features(name %in% c(
  mutate(assay="rnaseq")

dat <- samples %>%
  with_sample_covariates('samp
  with_assay_data(features, nc

ggplot(dat, aes(CCNB2, KRT17,
  geom_point() +
  theme(legend.position="botto
```

dataset	sample_id	CCNB2	KRT17	sample_type
LUSC	TCGA-18-3406-01A-01R-0980-07	5.0915656	13.758303	tumor
LUSC	TCGA-18-3407-01A-01R-0980-07	4.7749408	12.631085	tumor
LUSC	TCGA-18-3408-01A-01R-0980-07	5.2399864	9.227850	tumor
LUSC	TCGA-18-3409-01A-01R-0980-07	4.6457598	12.166467	tumor
LUSC	TCGA-18-3410-01A-01R-0980-07	7.1772091	7.319666	tumor
LUSC	TCGA-18-3411-01A-01R-0980-07	6.3701353	12.286020	tumor
LUSC	TCGA-18-3412-01A-01R-0980-07	6.2838198	12.437087	tumor
LUSC	TCGA-18-3414-01A-01R-0980-07	6.1370424	11.362722	tumor
LUSC	TCGA-18-3415-01A-01R-0980-07	6.5267915	12.290913	tumor
LUSC	TCGA-18-3416-01A-01R-0980-07	5.5907905	8.691251	tumor

Live Demo: Recover brushed annotations and generate graph via code

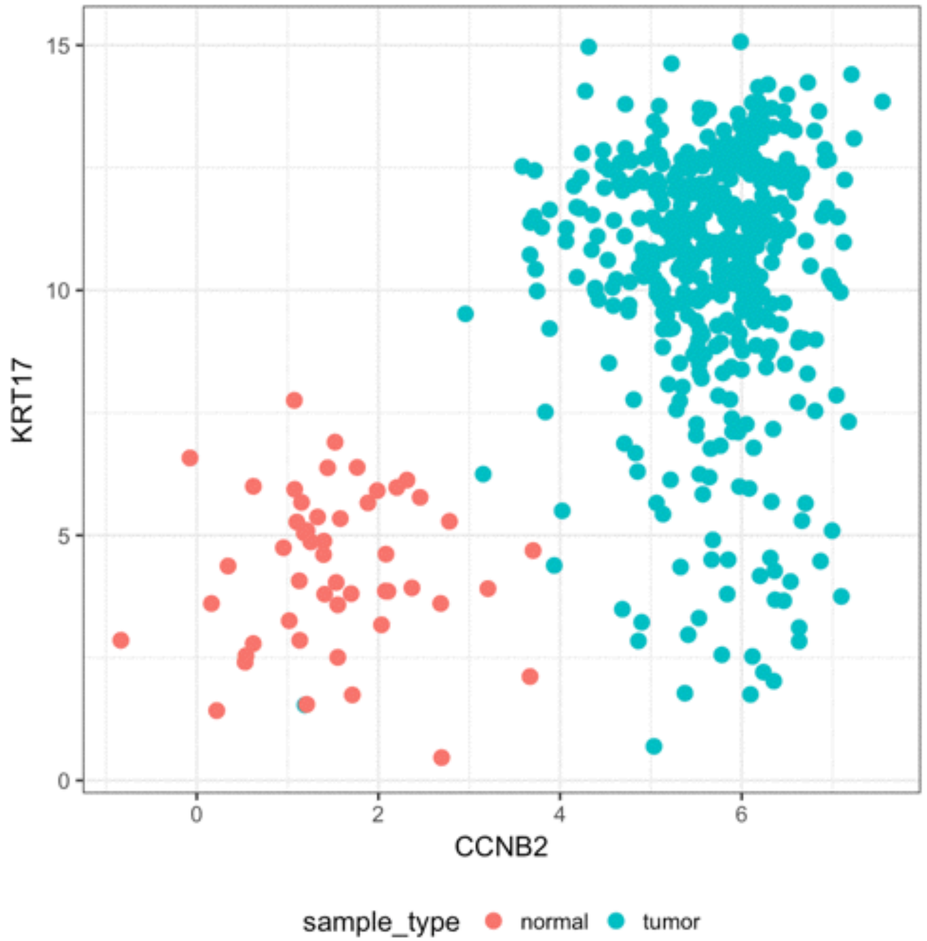

```
library(FacileExplorer)
library(FacileTCGADataSet)
tcga <- FacileTCGADataSet()

samples <- tcga %>%
  filter_samples(indication == "LUSC")

features <- tcga %>%
  filter_features(name %in% c("CCNB2", "KRT17")) %>%
  mutate(assay="rnaseq")

dat <- samples %>%
  with_sample_covariates('sample_type') %>%
  with_assay_data(features, normalized=TRUE, spread="name")

ggplot(dat, aes(CCNB2, KRT17, color=sample_type)) +
  geom_point() +
  theme(legend.position="bottom")
```



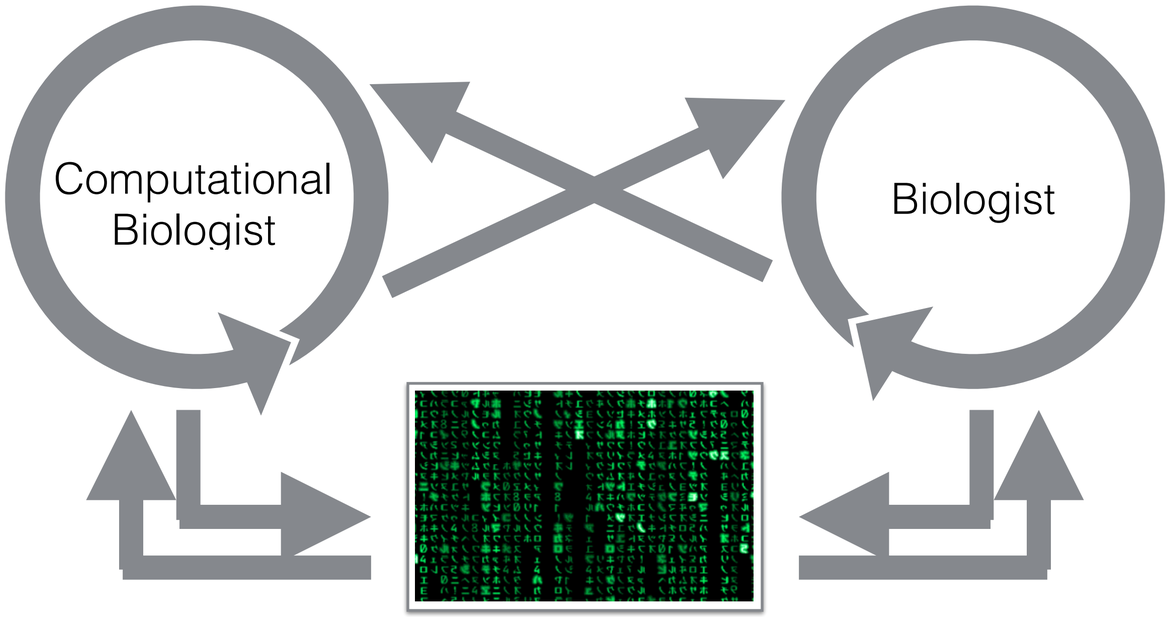
Long Term Goal

- Exploit modular design of analysis components to enable dynamic Rmarkdown report generation within the FacileExplorer.
  - Individual module result (scatter plot, survival, heatmap, etc) can be exported with prose to a growing report (thanks to Michael Lawrence for idea).
  - These reports maintain connection to data backend
- Analyst can consume URL of report in local workspace to re-instantiate the state of analysis for any module embedded in the report (baton passed).

```
state <- restore("http://facilereports/d04be94#prolif-scatter")
```

Summary

- FacileData Ecosystem provides tools to make collaborative science more efficient
  - Empowers sustained, independent, and interactive exploration
  - Enables each scientist to bring their expertise to bear on problem
  - Provides means to "hand off" analyses for further exploration by keeping results connected to their data



Acknowledgements

The FacileData Working Group


Matthew Albert


Marcia Belvin


Richard Bourgon


Haiyin Chen


Sami Mahrus


Vincent Rouilly

Genentech

Michael Lawrence
Luciana Molinero
Yulei Wang
Christine Moussion
Pete Haverty
Gabe Becker
Dorothee Nickles
Ben Shaibe
Matt Buechler
Sascha Rutz
Tess Delfino
Shannon Turley

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Martin Morgan
Yubo Cheng (MultiExperimentDb)
Herve Pages (HDF5 libraries)

support.bioconductor.com
community

rbokeh


Ryan Hafen