

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>

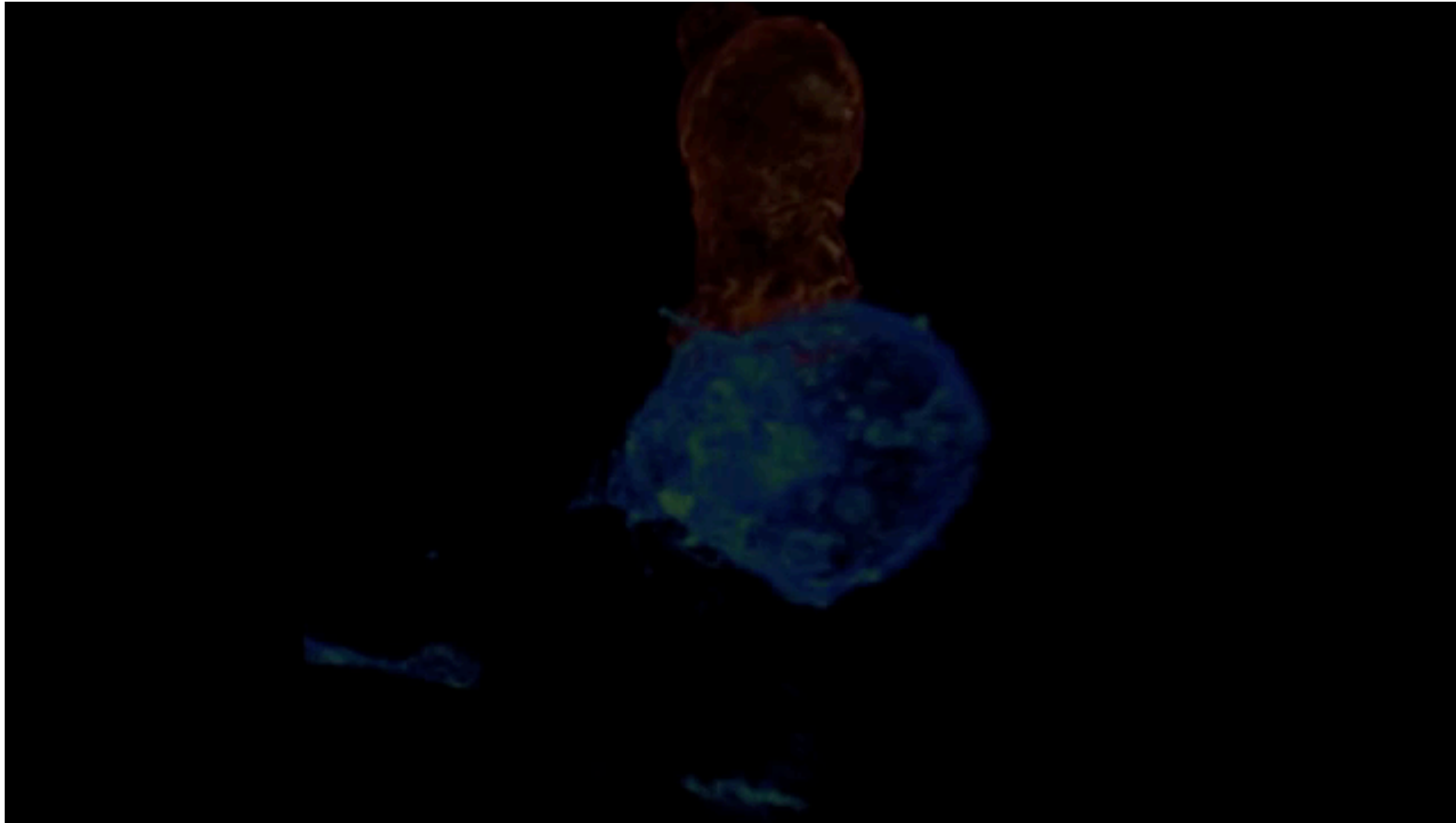
Repository with these slides
and presentation materials

<https://github.com/faciledata/talks>

★ star the `faciledata/talks` repository to be updated ★
when tools are released

Cancer Immunology: the immune system can combat cancer

3



Alex Ritter

We can read DNA at scale

4



Illumina HiSeq

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

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

- **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

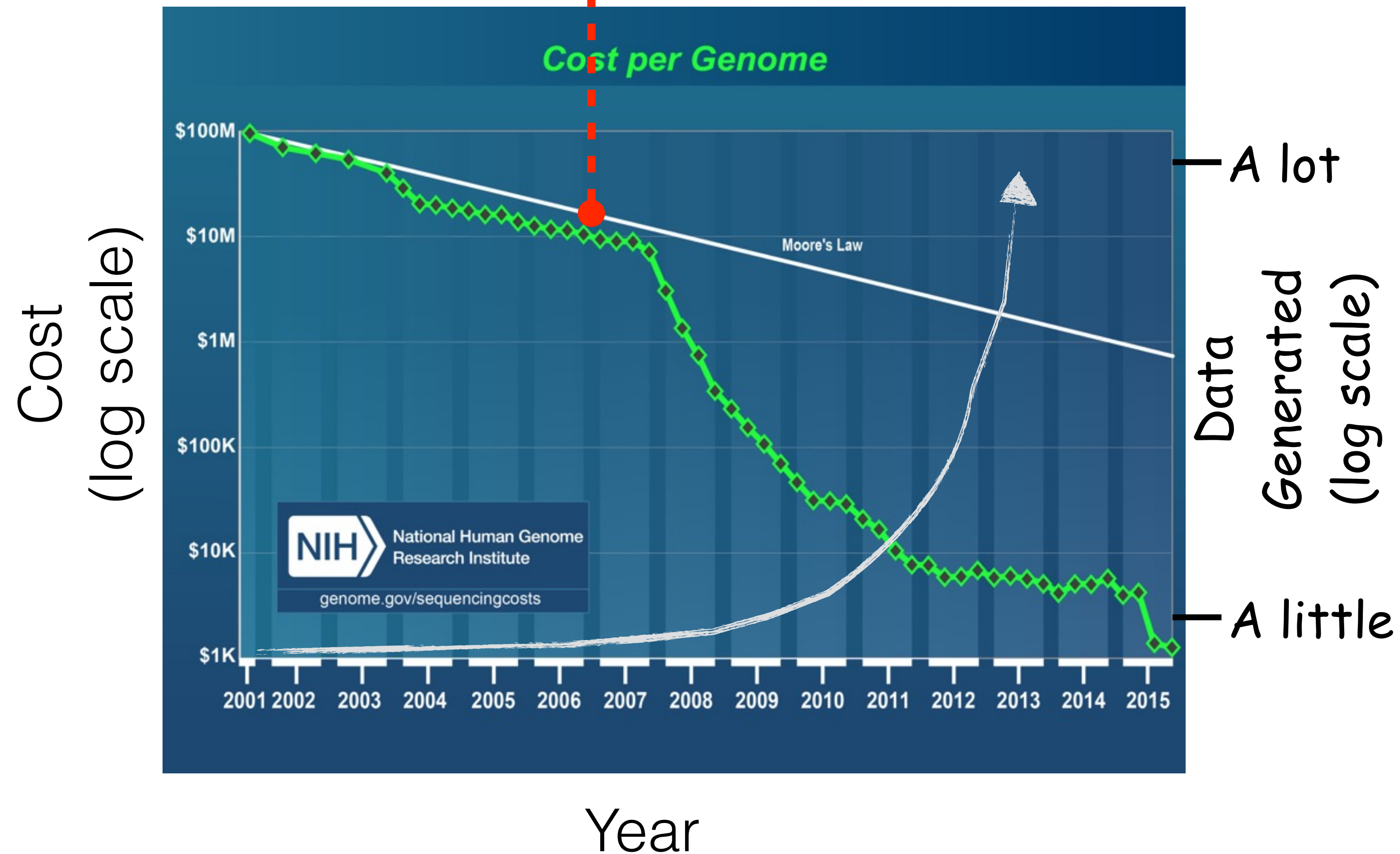
The universe of things the genome can do

The thing the genome is doing right now

Biology has been transformed into a data driven science

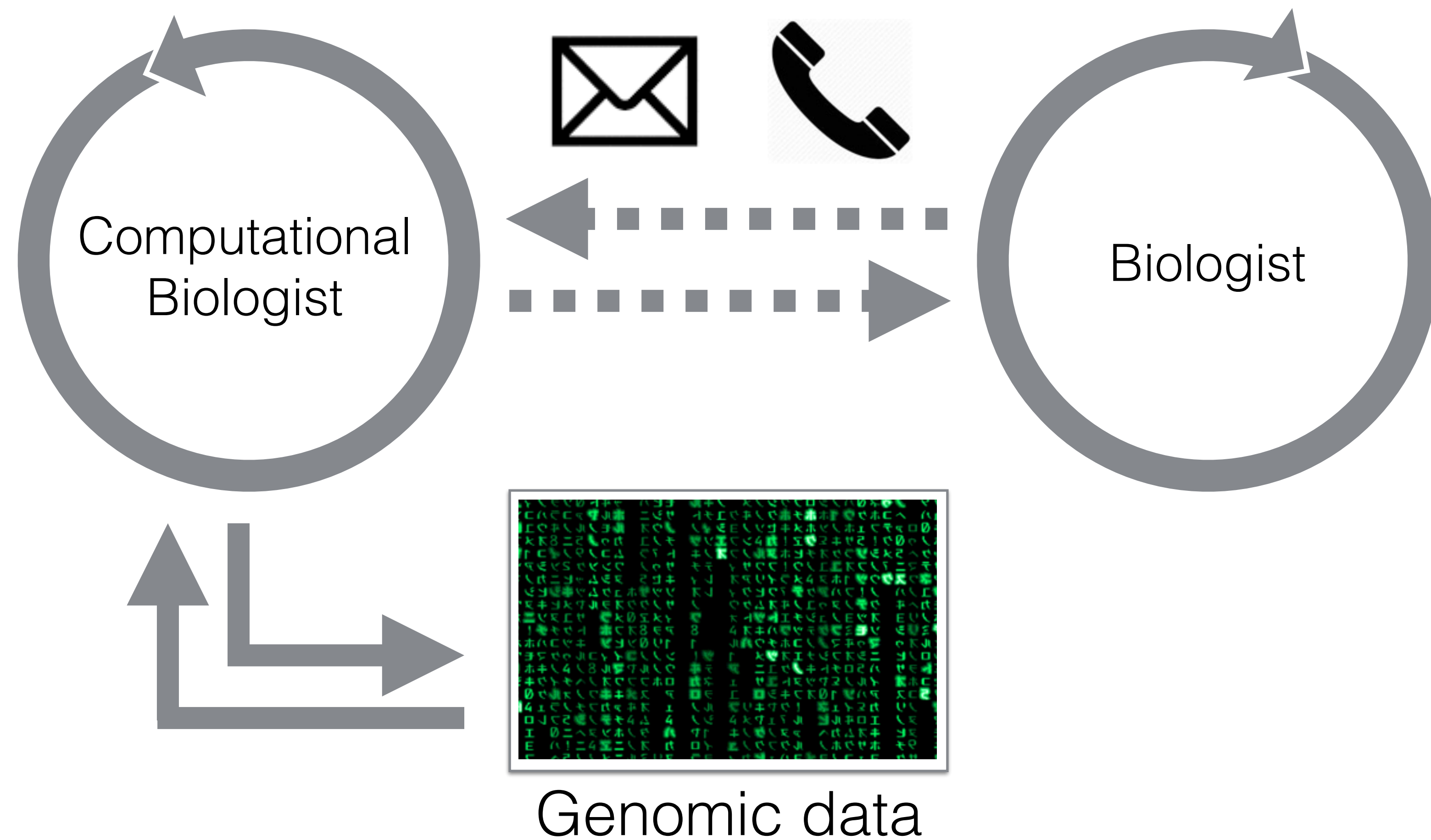
6

Illumina enters sequencing market



The state of data driven exploration and discovery

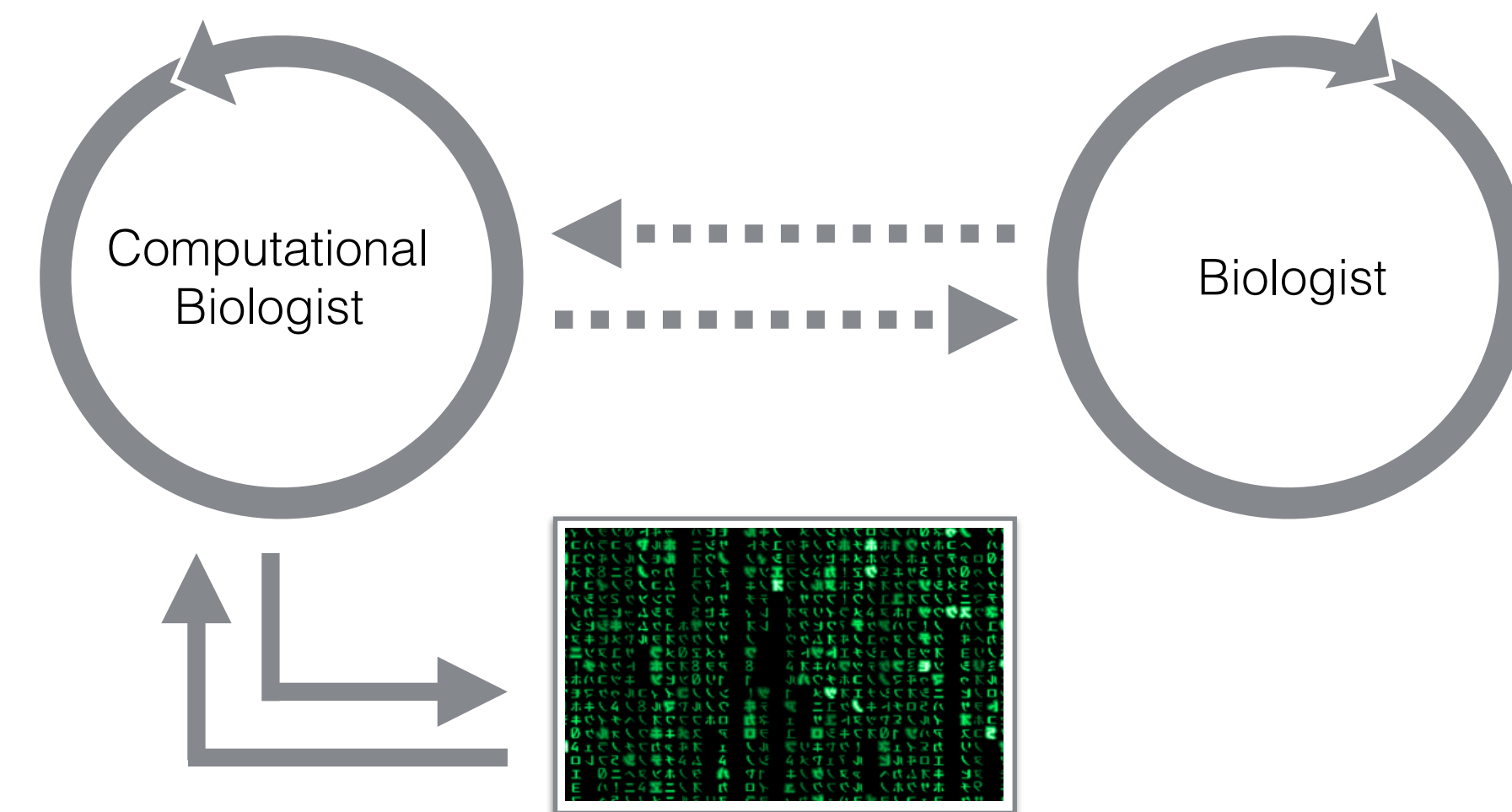
7



Problems with current state of data driven exploration

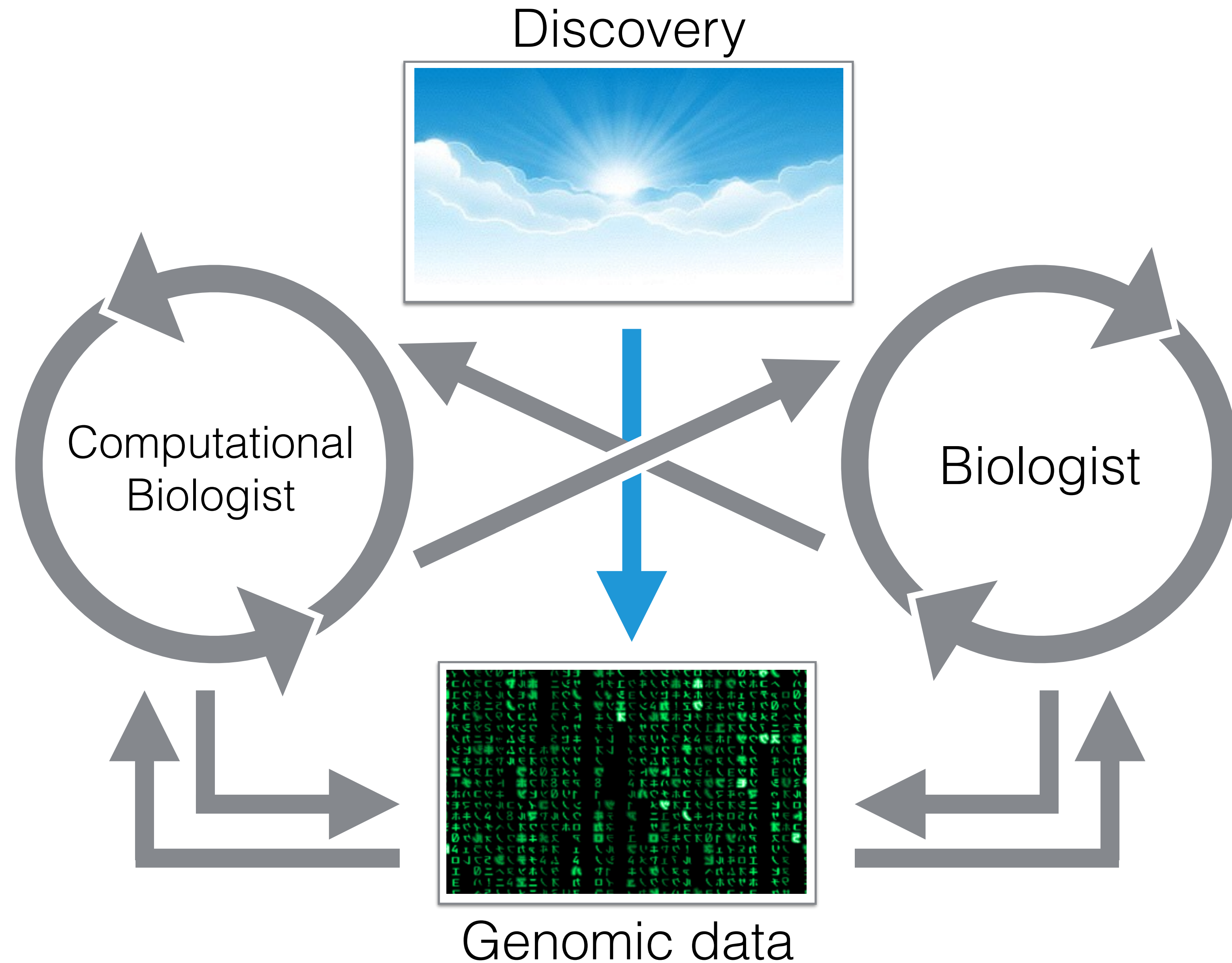
8

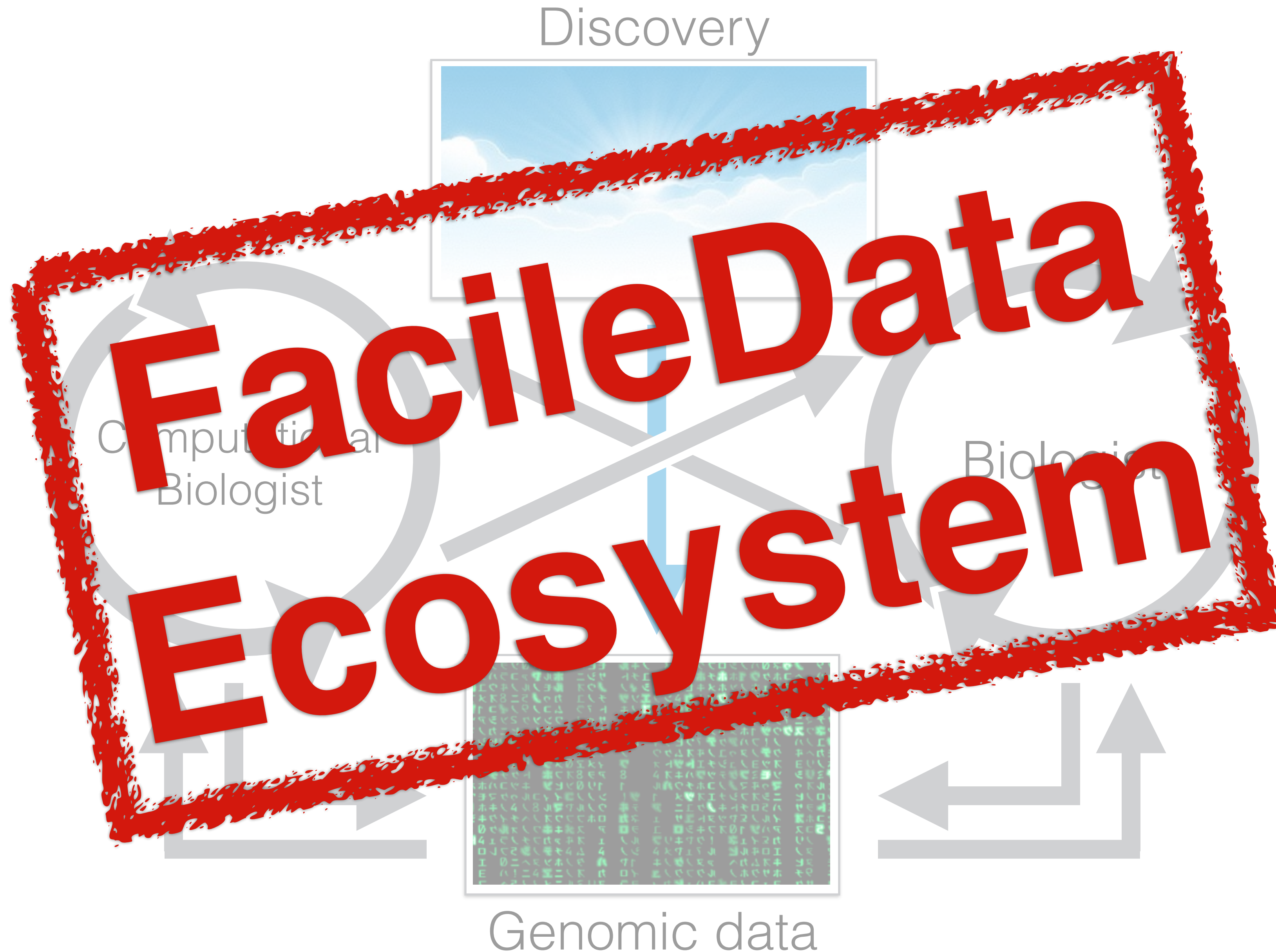
- Biologist has limited ways to meaningfully interact directly with the data.
- Relies on computational biologist for even most simple of EDA
- Long turn around times (Colin Ware: "delay creates a universe of ideas lost")
- Context switches and broken *flow* for all
- The units of "knowledge transfer"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



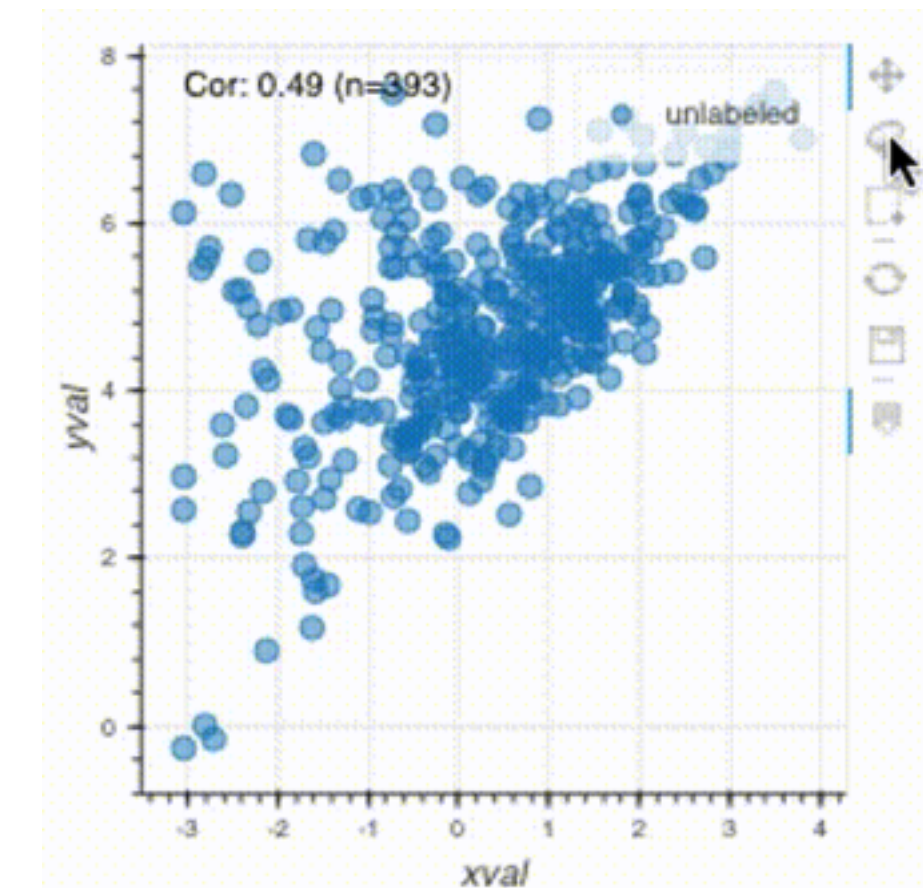


Tools to facilitate collaborative exploration

11

The FacileData Frontend (FacileExplorer)

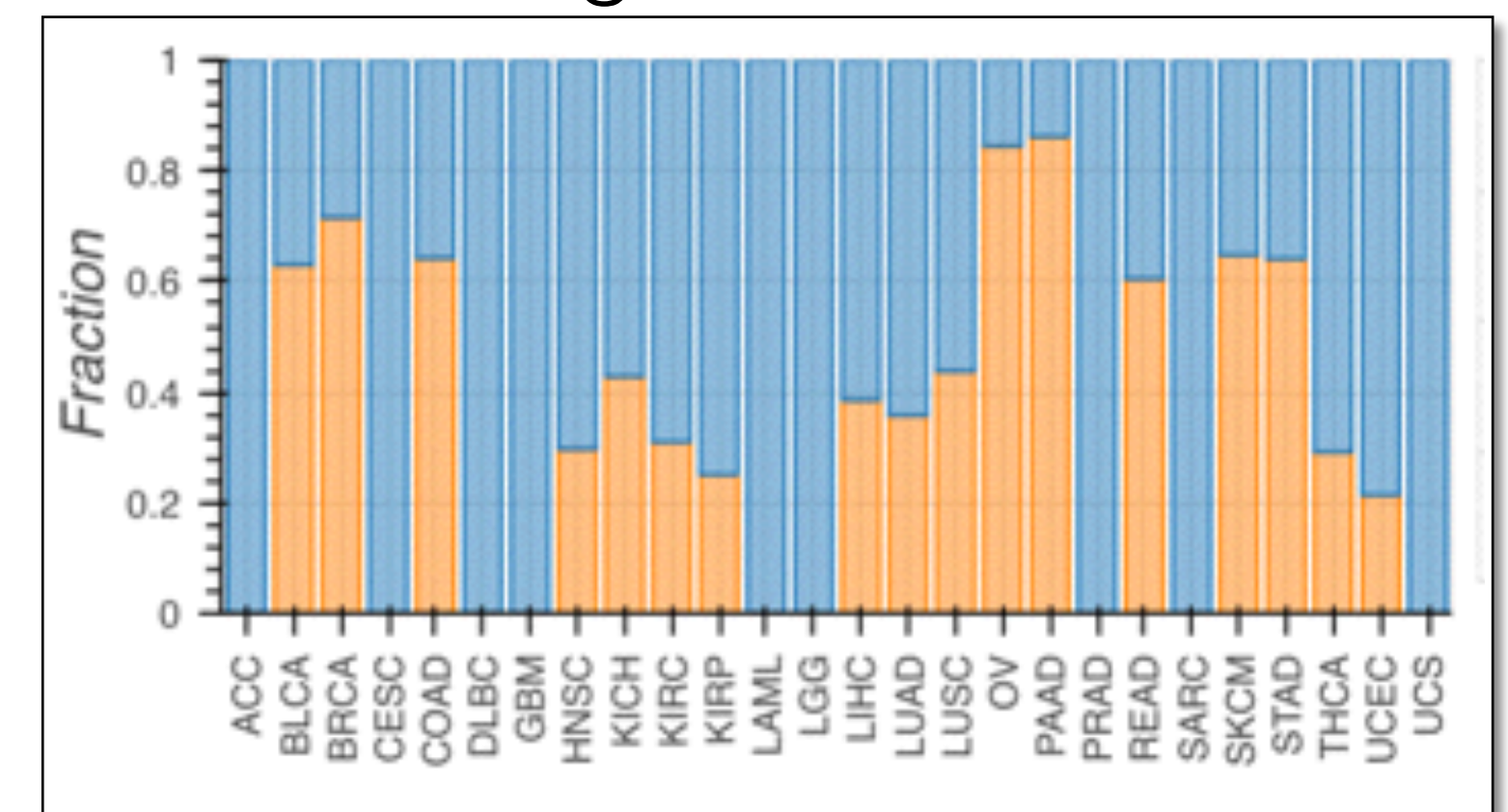
- Designed to enable **sustained and independent** interactive data exploration by non-informaticians
- Empowers users to compute over data via its GUI
- Provides ability to hand off analyses "in flight", not just results



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*)

Everything else ■
Stage II or III tumors ■



3321 / 8024 TCGA samples



The FacileData Backend (FacileDataSet)

High-throughput genomics data is large and complex

Dataset 1 (Breast Cancer)

Sample Information

N	N	N	N	T	T	T	T
M	F	F	M	M	M	F	F
I	II	I	I	II	I	II	I
N	Y	N	Y	Y	N	N	Y

Sample Type (**N**ormal/**T**umor)

Sex (**M**/**F**)

Tumor Stage (**I** / **II** / **III** / ...)

Has bone metastasis (**N**o / **Y**es)

Gene (feature)
Information

Gene Symbol
Biotype
Chromosome
Length

A
B
C
:

25,000 Genes

Assay Data

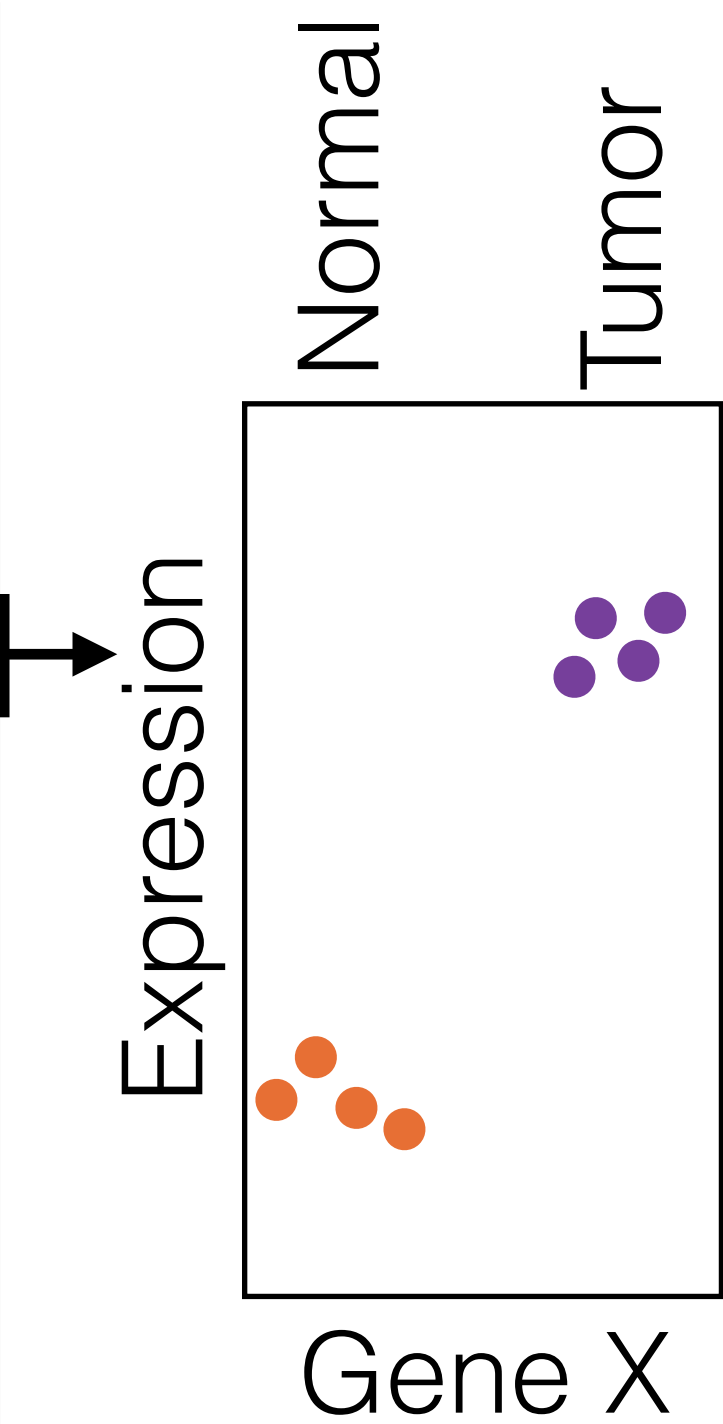
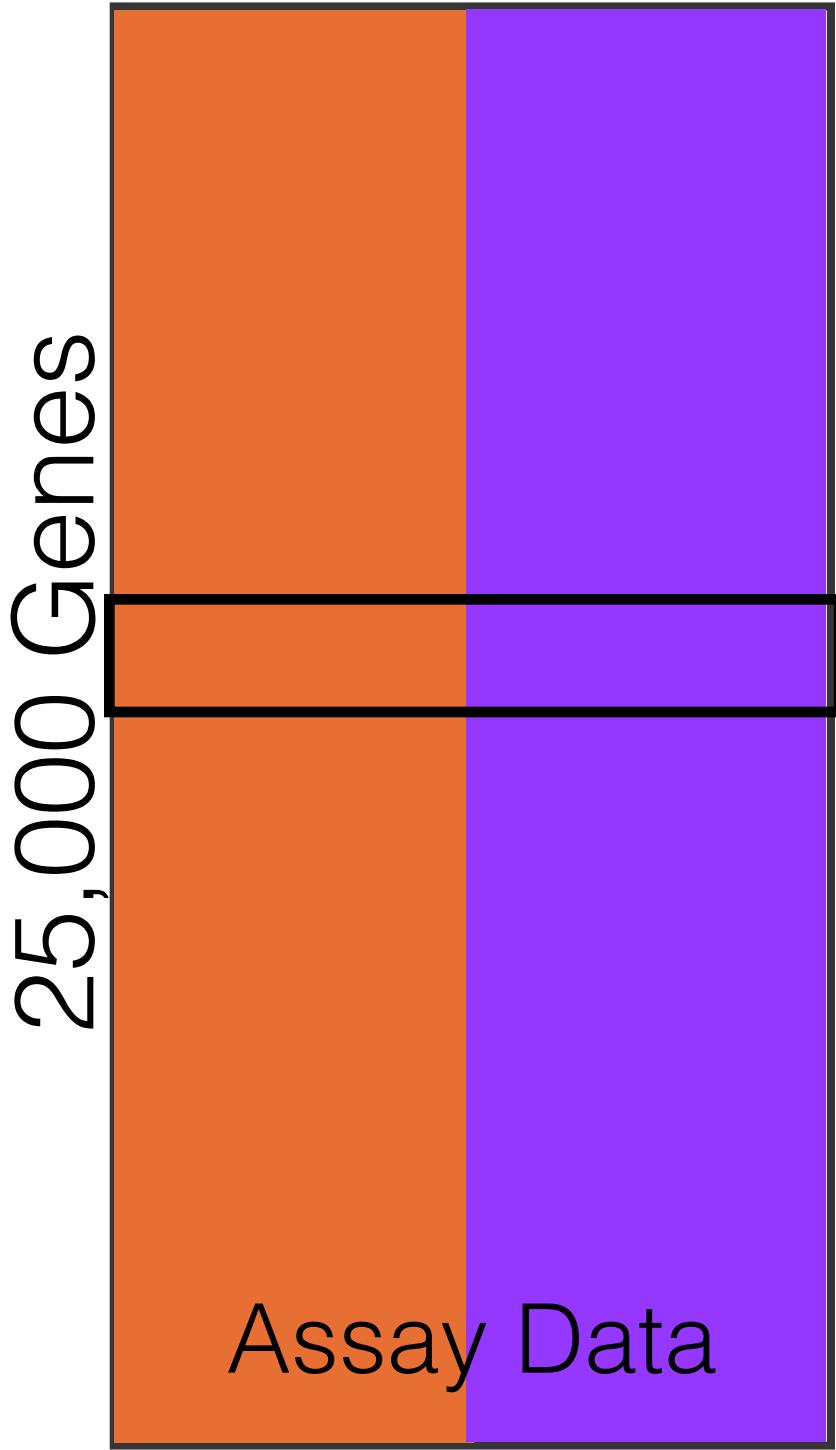
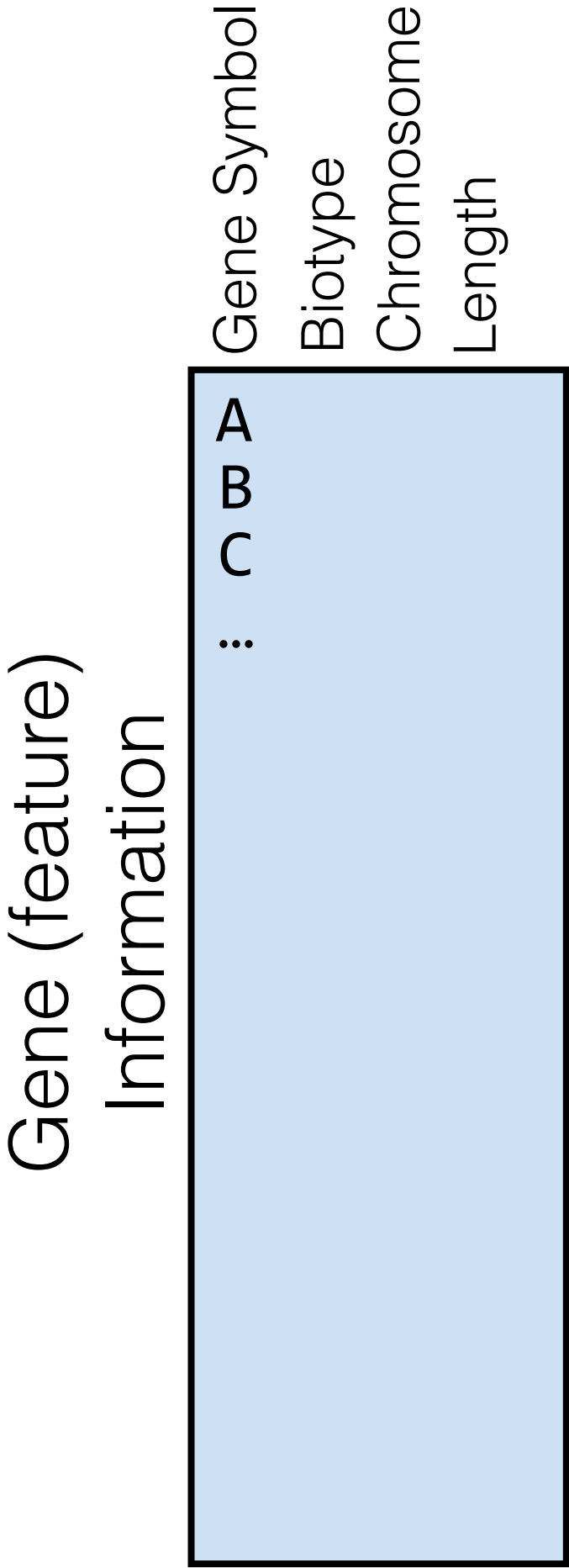
High-throughput genomics data is large and complex

Dataset 1 (Breast Cancer)

Sample Information

N	N	N	N	T	T	T	T	Sample Type (N ormal/ T umor)
M	F	F	M	M	M	F	F	Sex (M /F)
I	II	I	I	II	I	II	I	Tumor Stage (I / II / III / ...)
N	Y	N	Y	Y	N	N	Y	Has bone metastasis (N o / Y es)

Normals vs Tumors



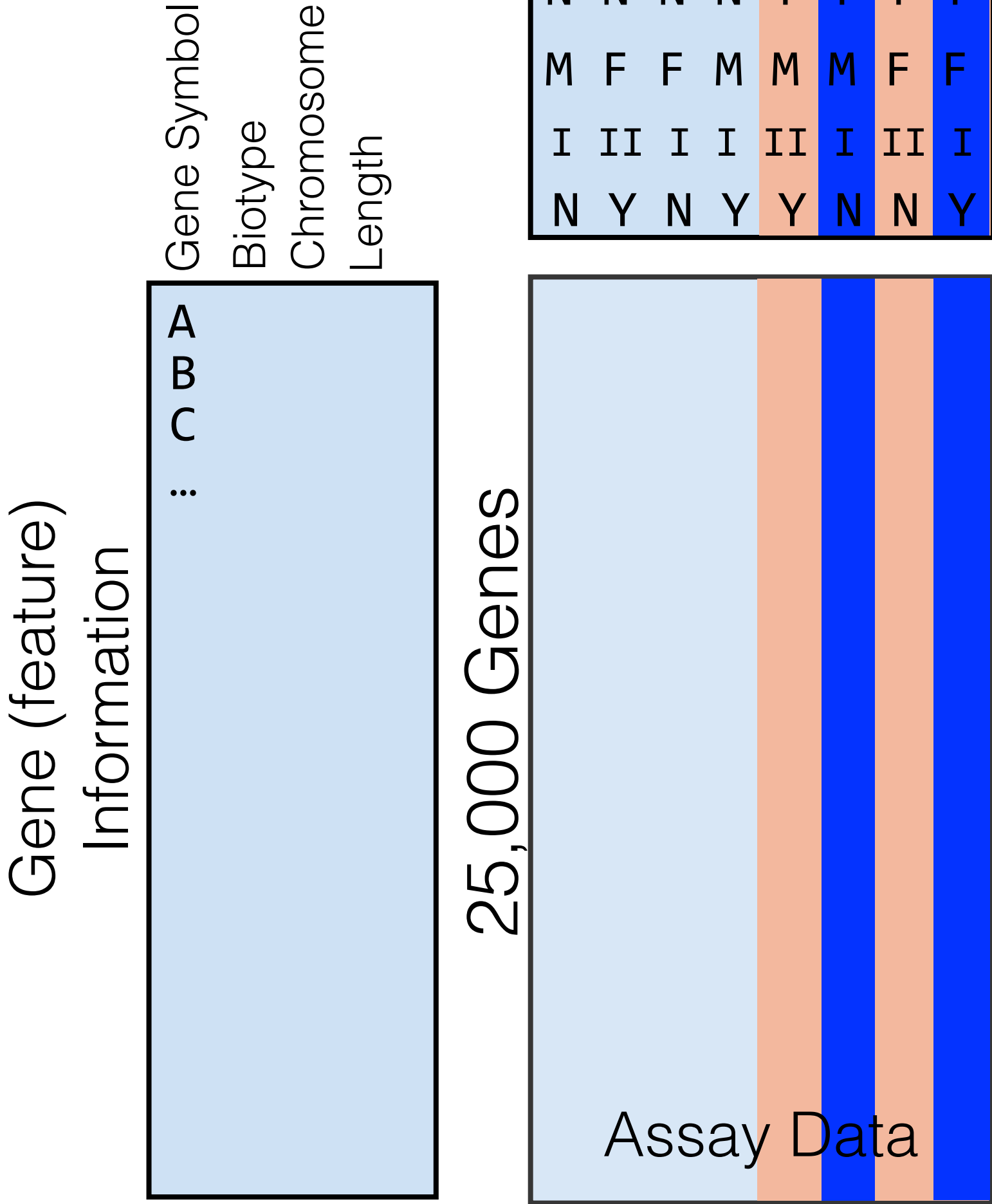
High-throughput genomics data is large and complex

Dataset 1 (Breast Cancer)

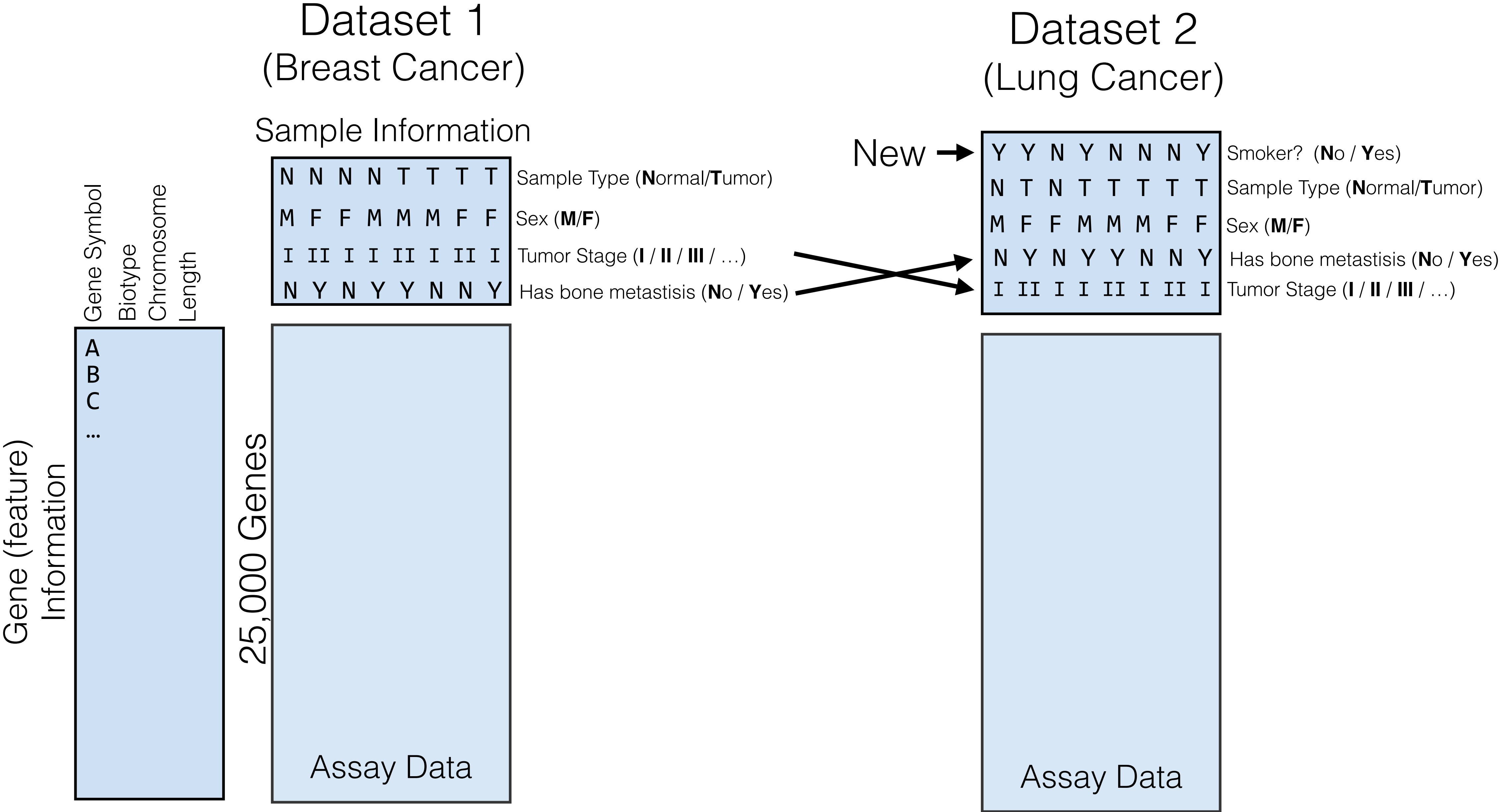
Sample Information

N	N	N	N	T	T	T	T	Sample Type (N ormal/ T umor)
M	F	F	M	M	M	F	F	Sex (M /F)
I	II	I	I	II	I	II	I	Tumor Stage (I / II / III / ...)
N	Y	N	Y	Y	N	N	Y	Has bone metastasis (N o / Y es)

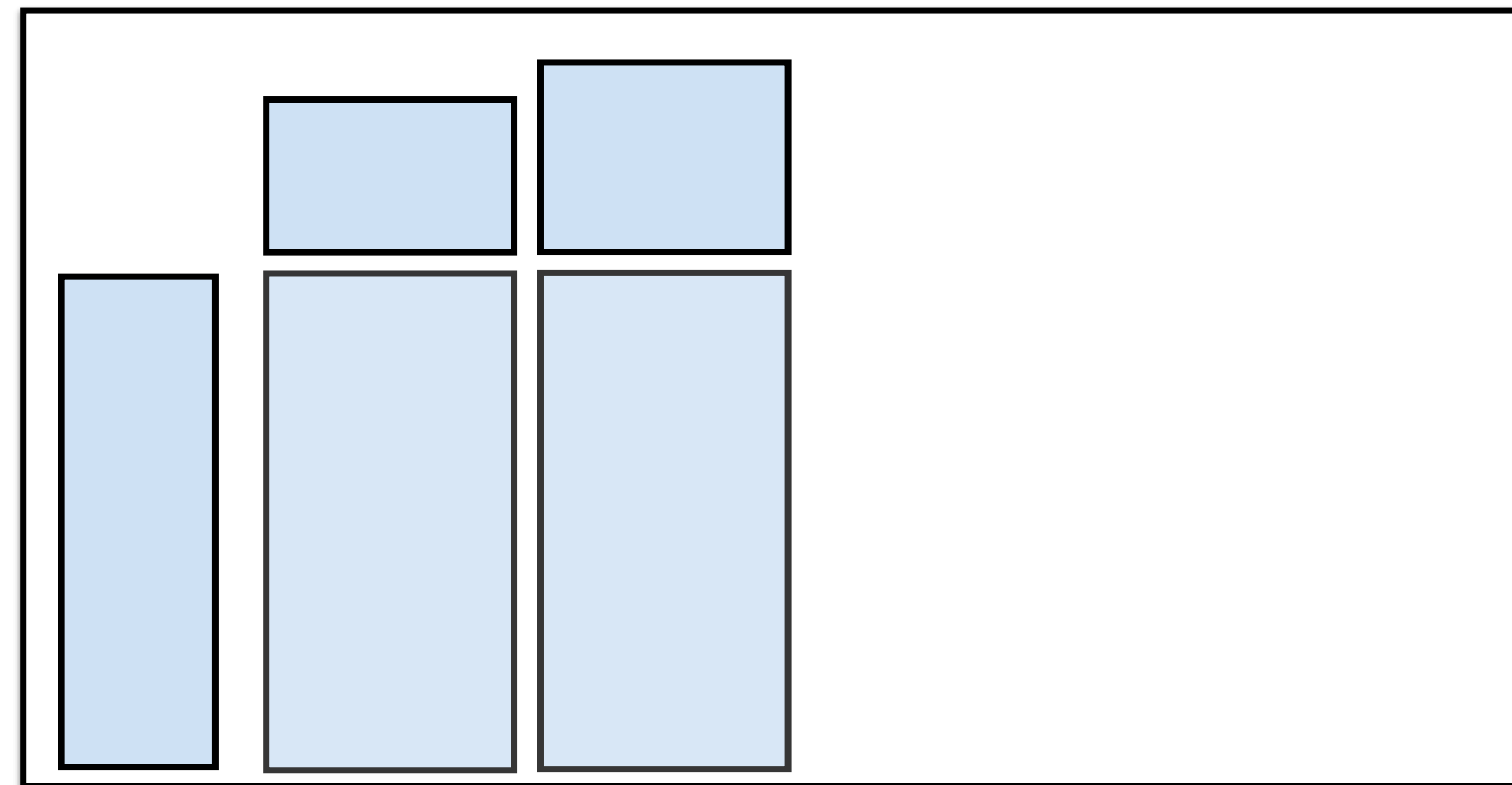
Stage I vs Stage II



High-throughput genomics data is large and complex



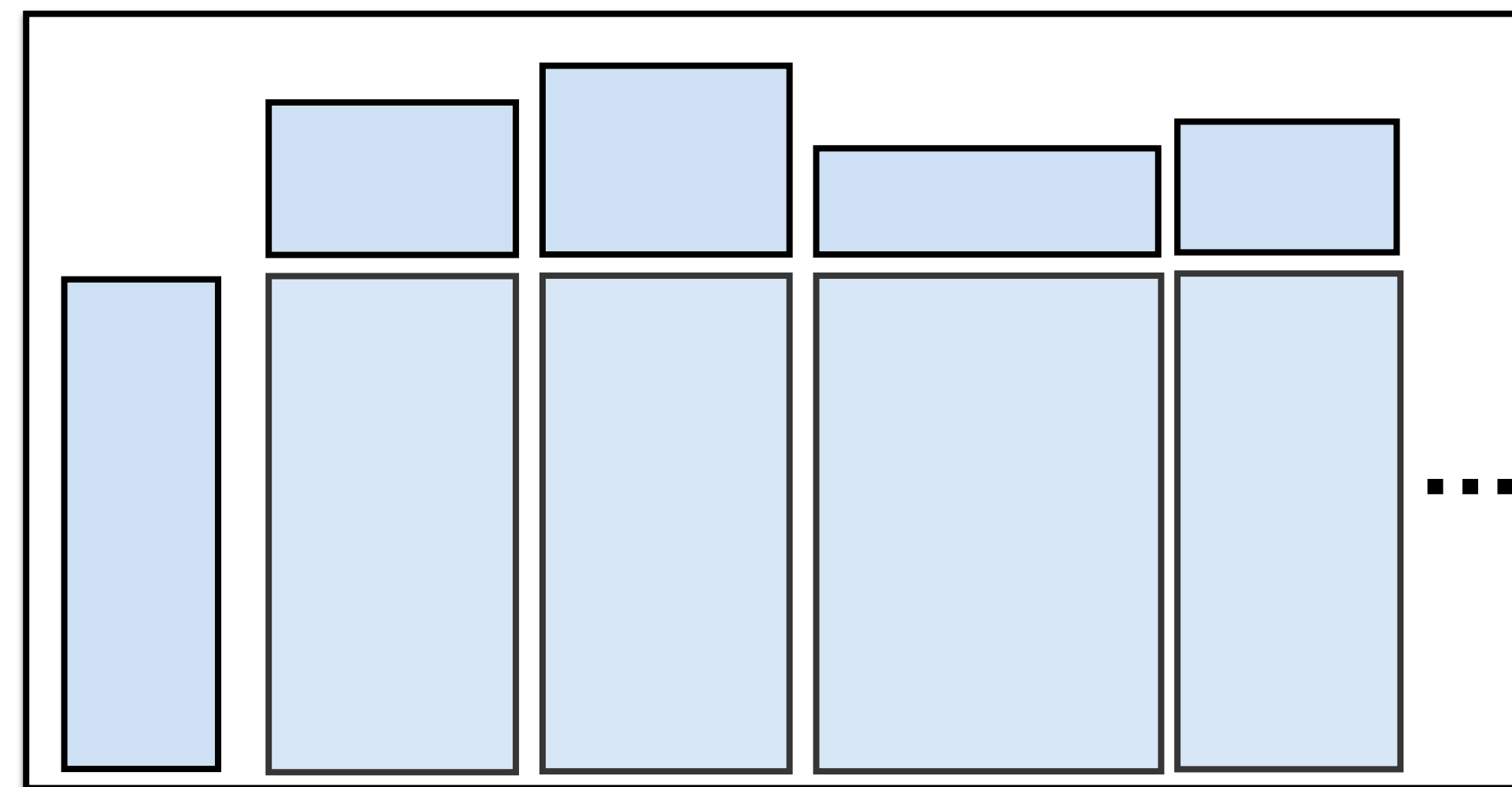
FacileDataSet provides single point of access across multiple datasets



FacileDataSet

FacileDataSet provides single point of access across multiple datasets

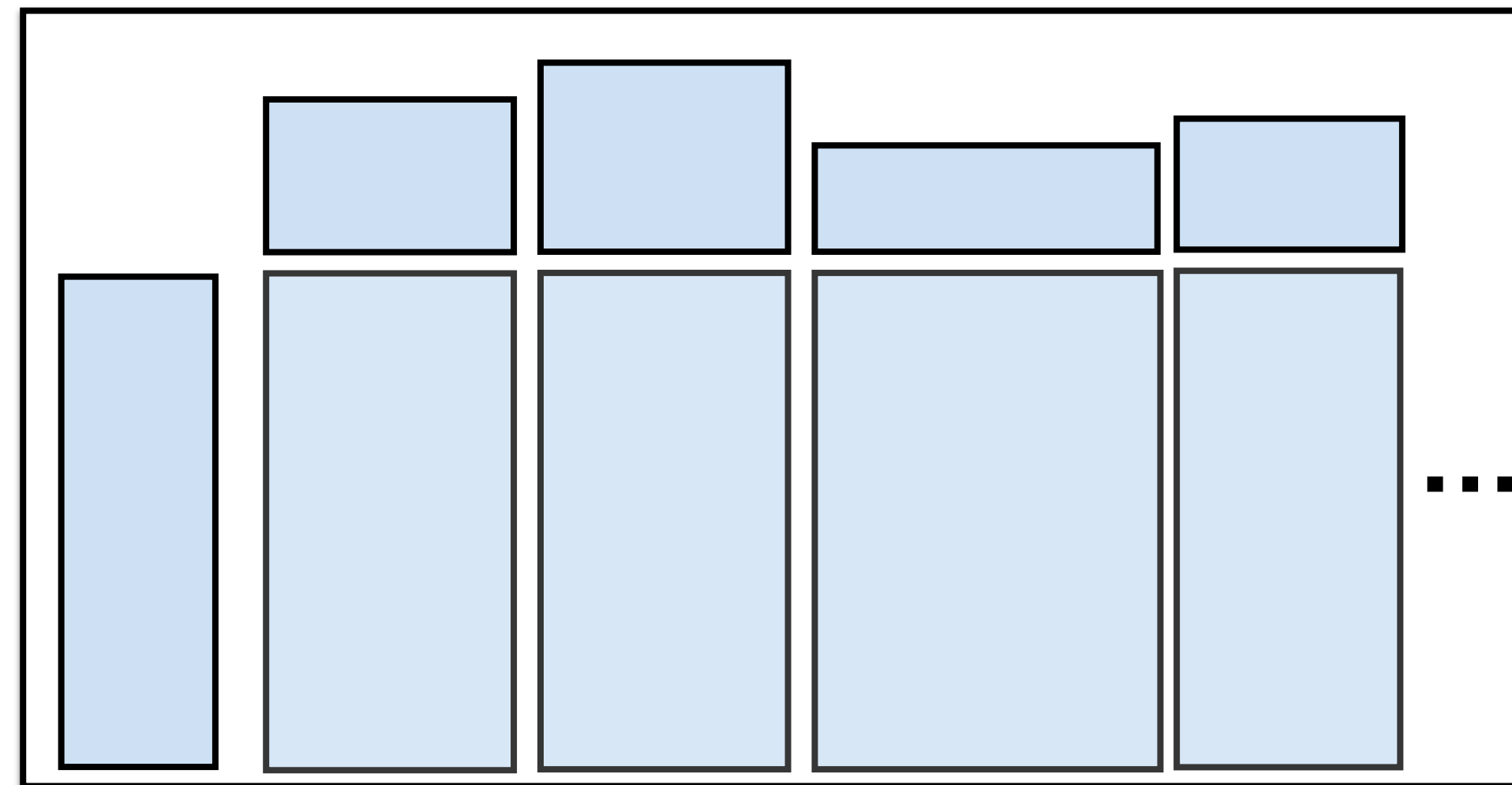
18



FacileDataSet

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

19

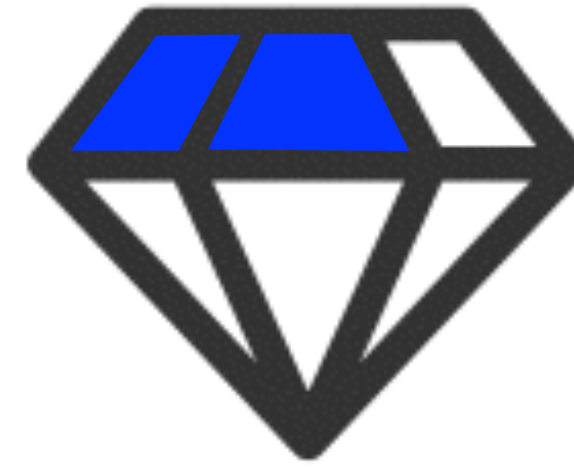


FacileDataSet

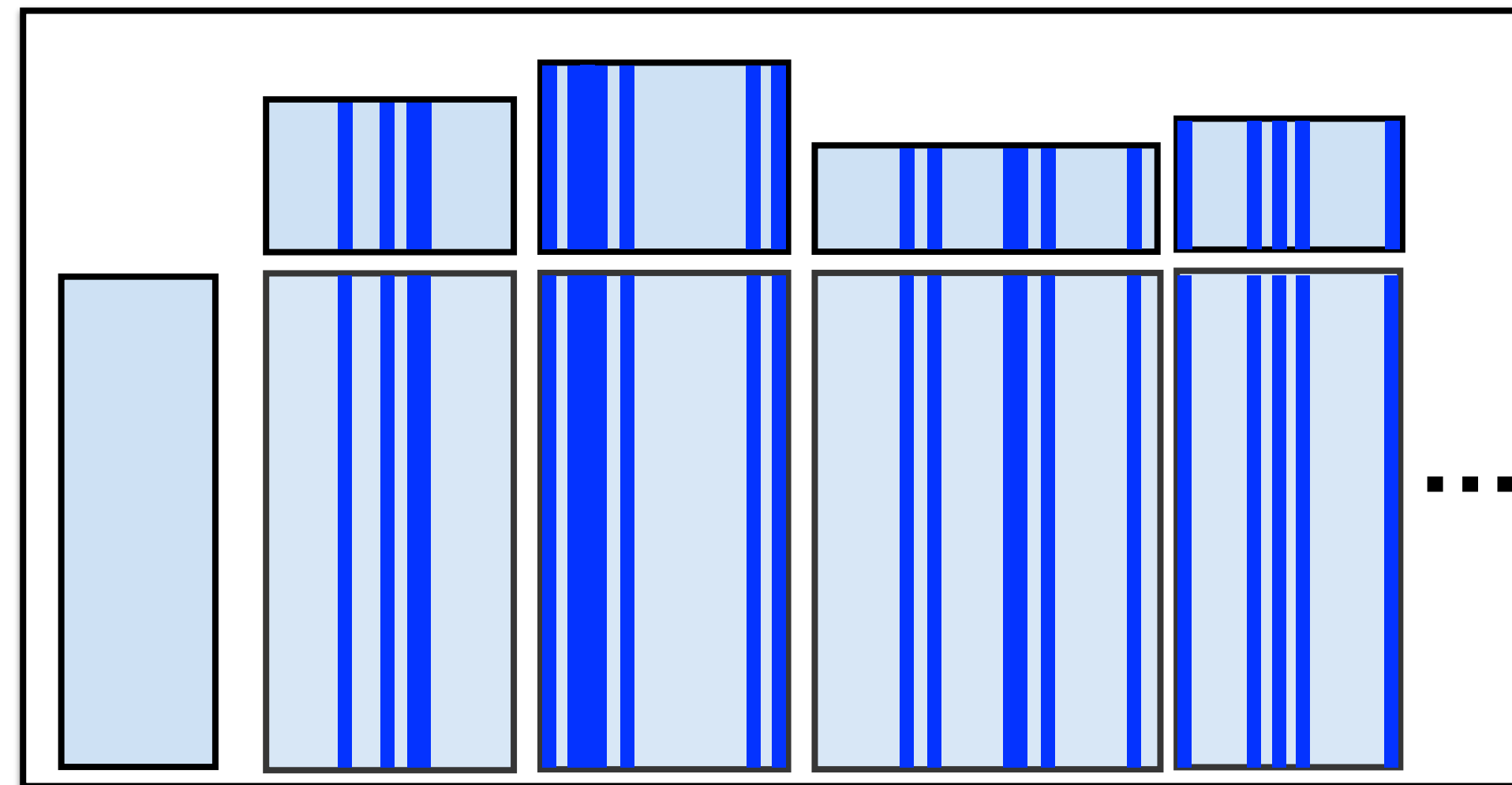
Think of data as a multi-faceted diamond

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

20



Stage I cancers

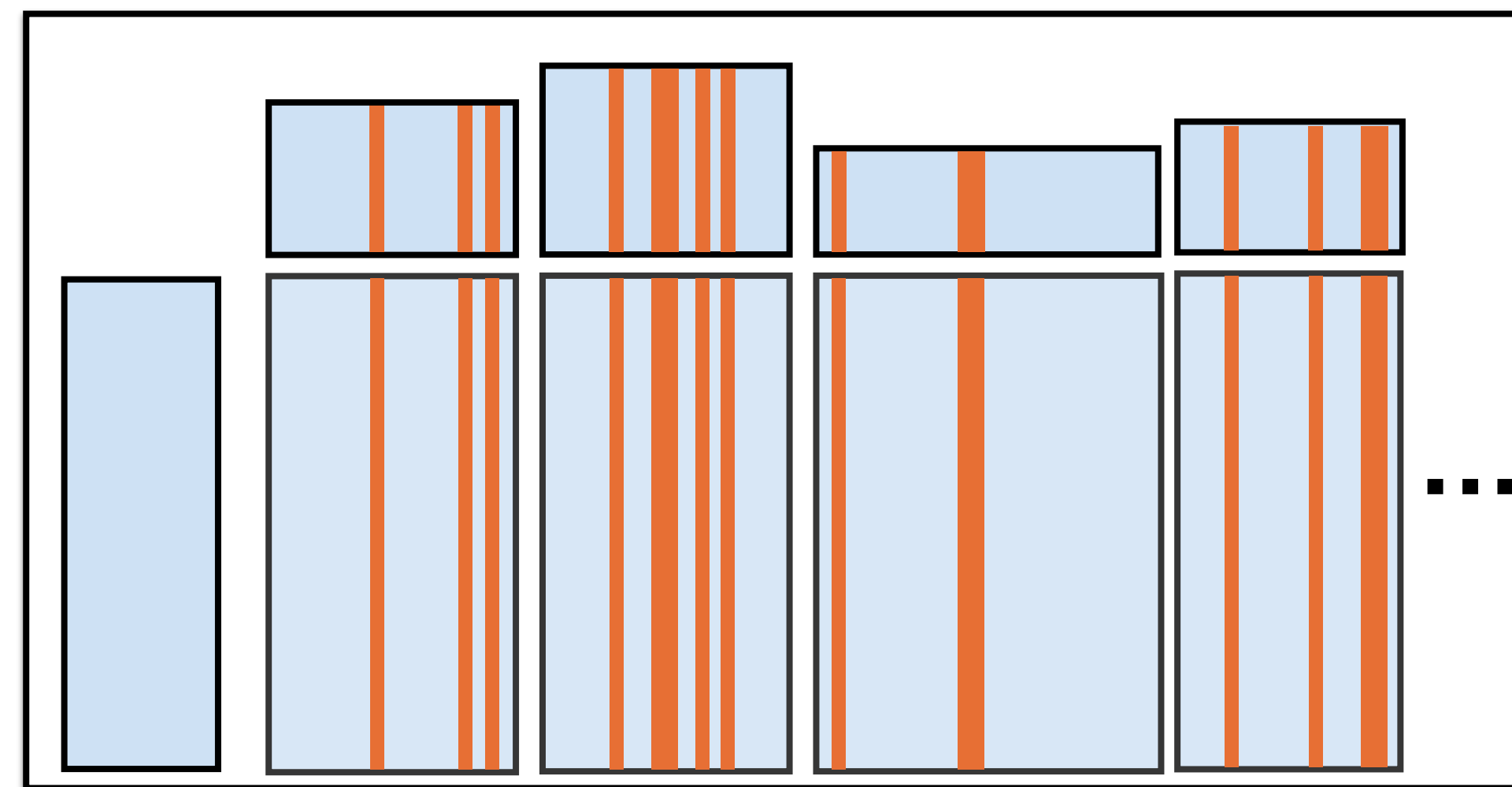


FacileDataSet

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

21



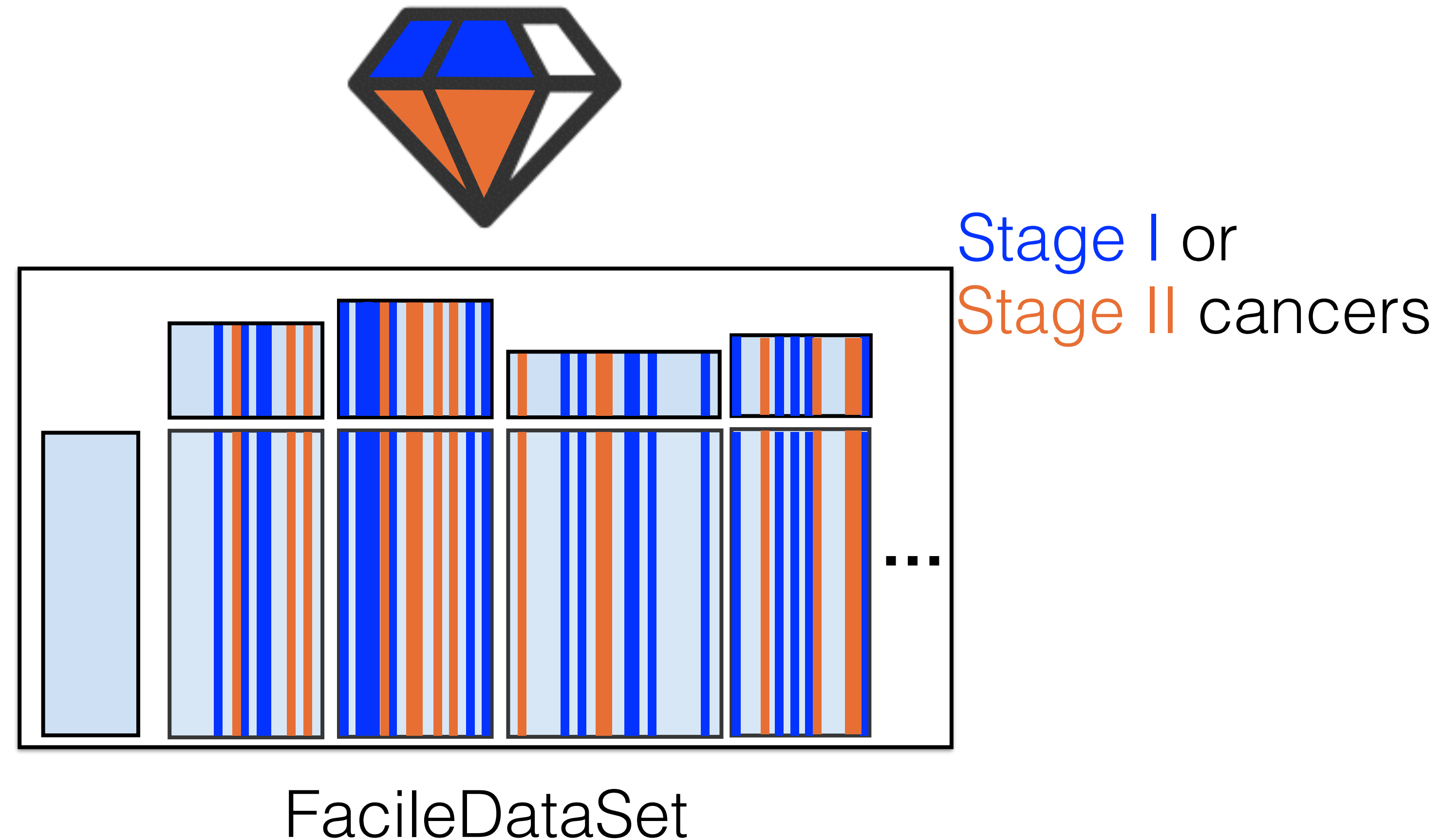
Stage II cancers

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

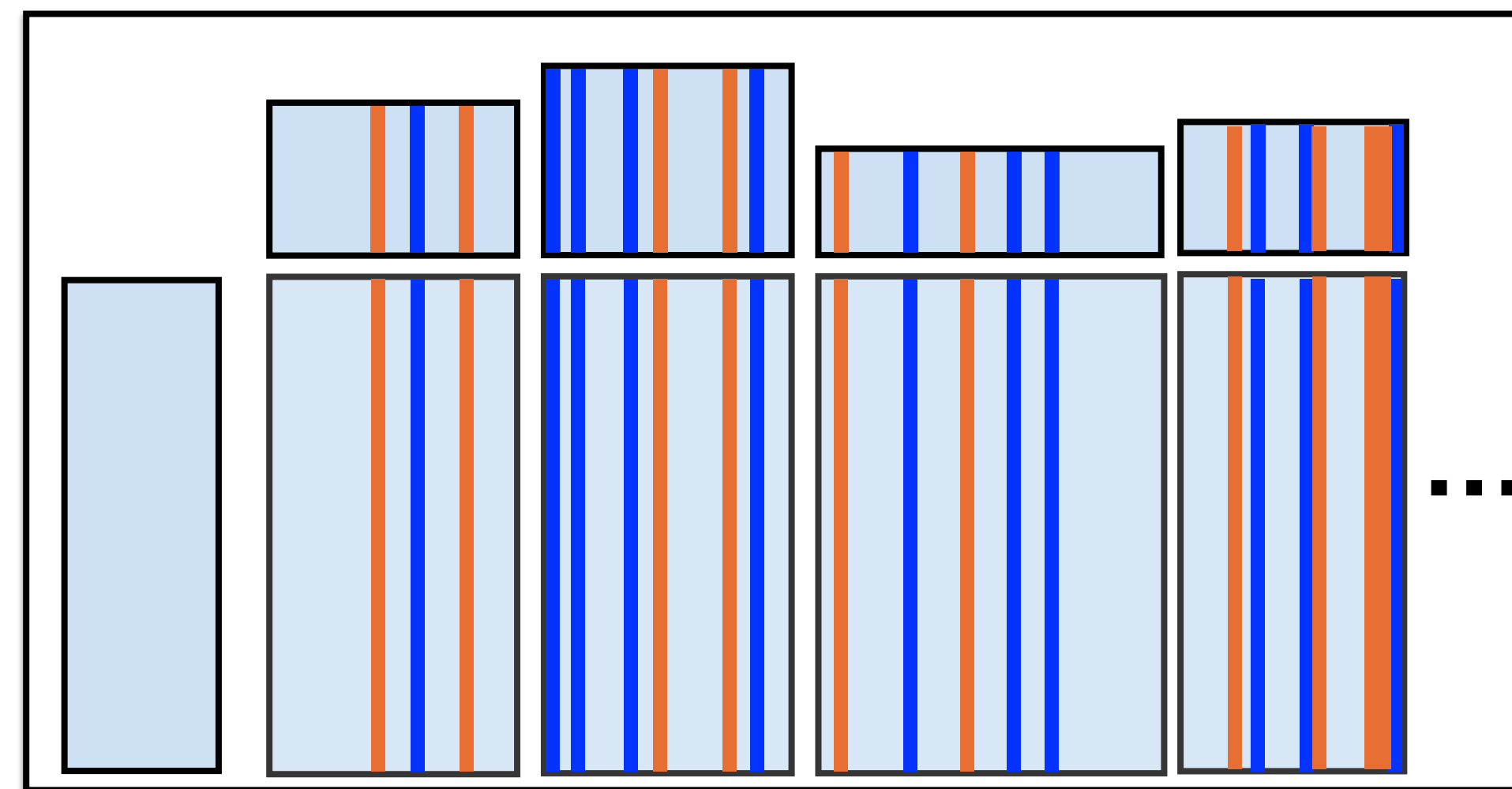
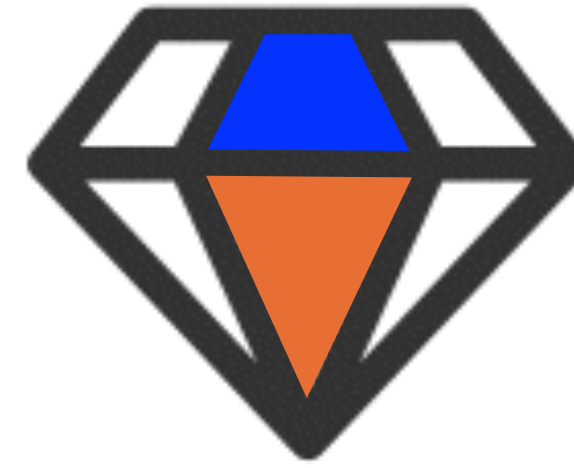
22



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

23



Stage I or
Stage II cancers
... that respond to
treatment X

FacileDataSet



The FacileData Frontend (FacileExplorer)

Anatomy of a Facile Survival Analysis

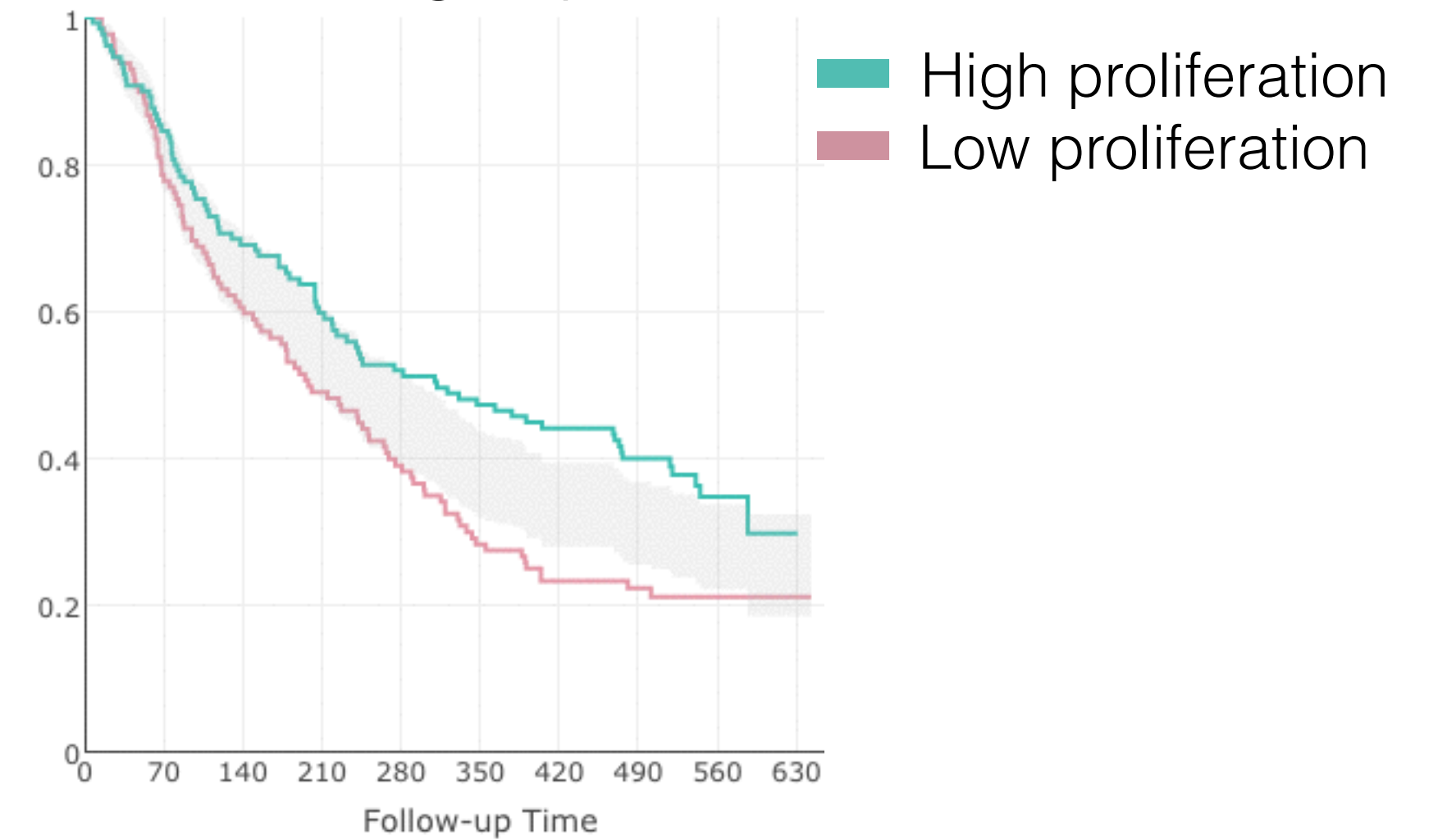
These results
are hypothetical

25

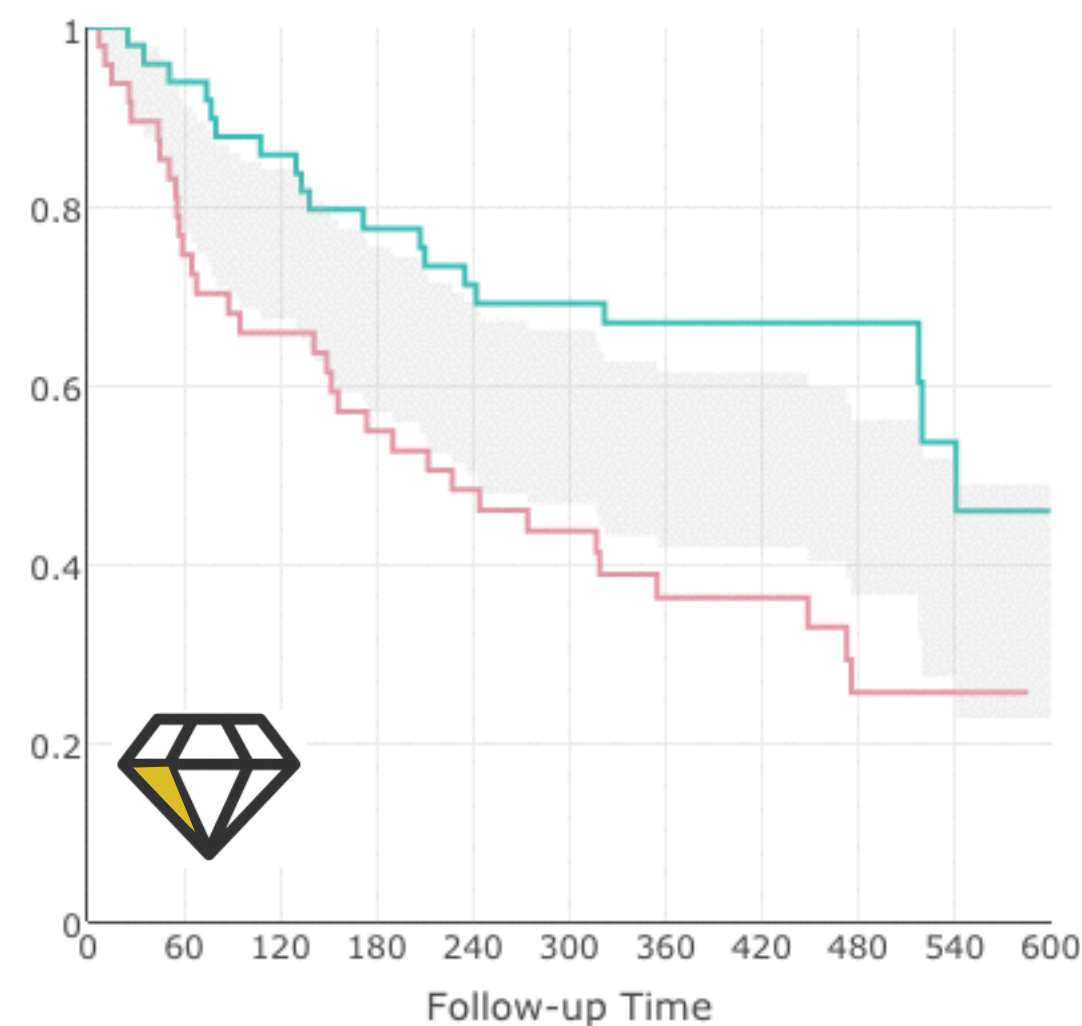
- User decides to analyze lung squamous tumors in TCGA FacileDataSet
- User asks if gene signature for proliferation is prognostic biomarker for overall survival
- User "one-clicks" an *Iterative Faceted Analysis* to identify other facets of data this result generalizes to



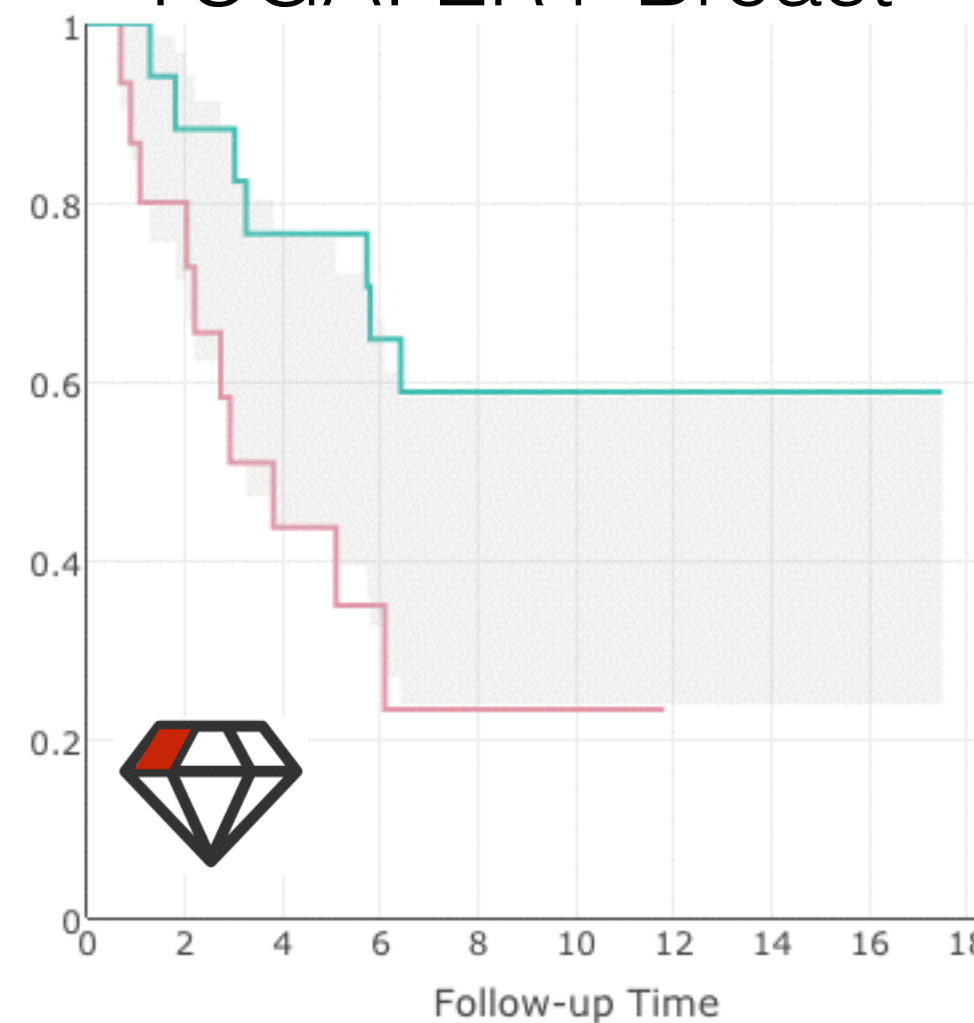
TCGA Lung Squamous



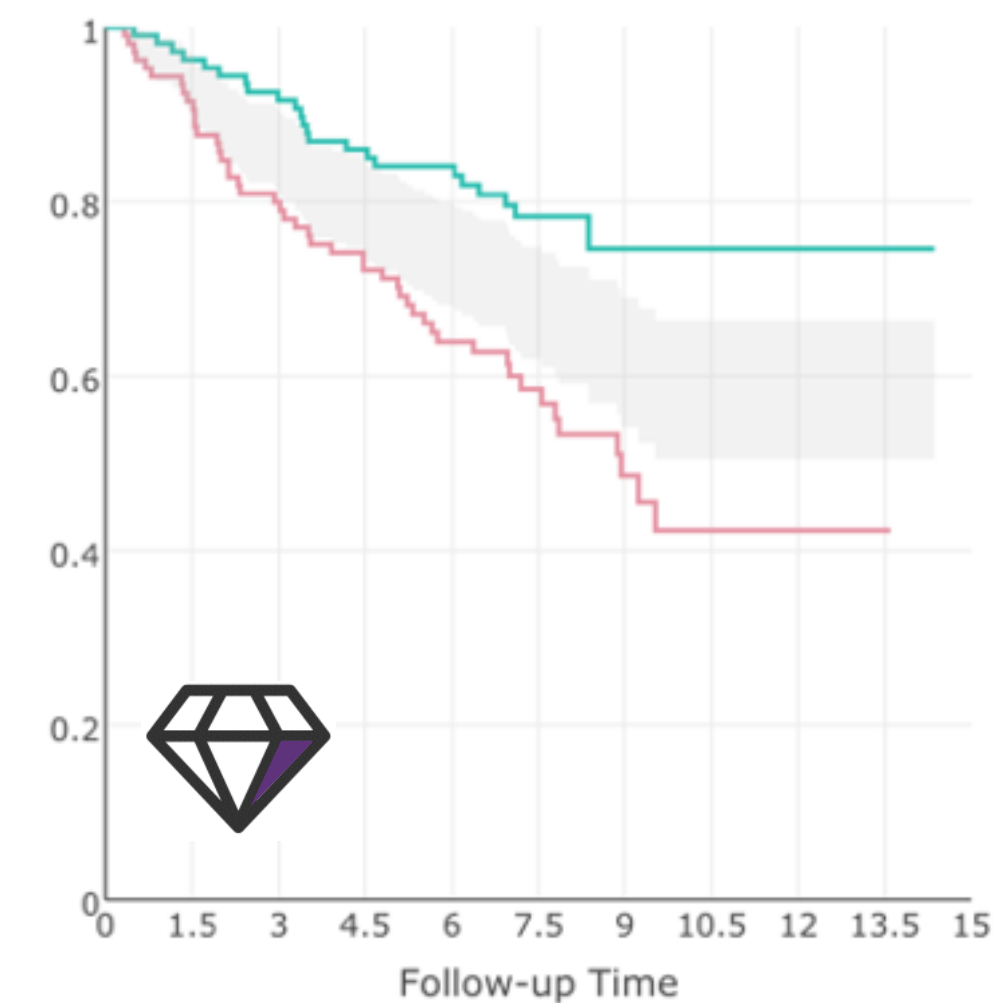
TCGA Lung Adenosarcoma



TCGA: ER+ Breast



TCGA: Bladder





Live demo of FacileExplorer

GUI → Analyst → GUI

Live Demo: Exploring Facets of TCGA

27

FacileExplorer

Facet Selection

Exploratory Analysis

Differential Expression & GSEA

Survival Analysis

User Annotations

FacileDataSet

TCGA

Define filters to select samples

Select Covariate

Cancer stage

Select Values

I II

Remove

Sample Filter

Datasets in Current Facet

Datasets: 17 Samples: 3900

Inspect Covariates

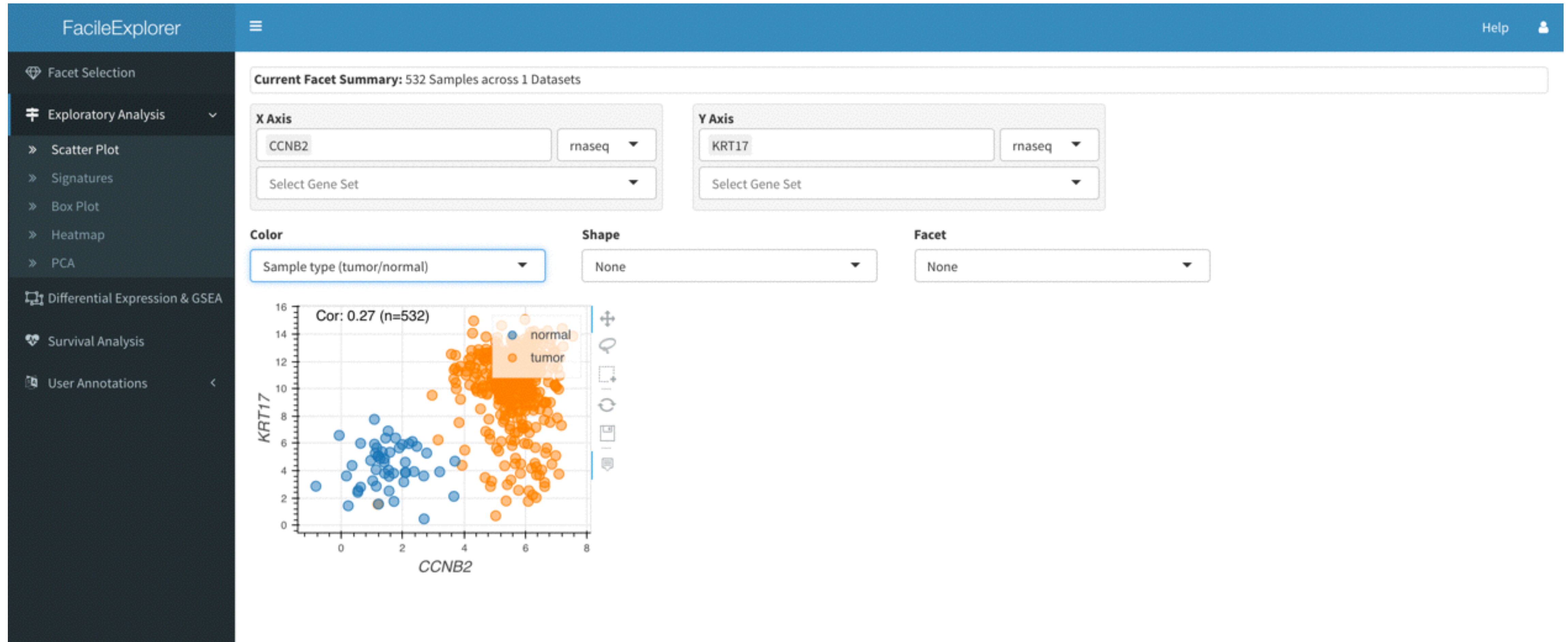
Covariate Name

Sample type (tumor/normal)

Description
Whether sample comes from a tumor or adjacent normal

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

28




```
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

30

```
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")
```

The image shows a screenshot of the Facile Explorer web interface. It features two main sections: 'Select Covariate' and 'Select Values'. Under 'Select Covariate', there is a dropdown menu currently showing 'Cancer Indication'. Under 'Select Values', there is a text input field containing 'LUSC'. To the right of the input field is a 'Remove' button with a minus icon. Below these sections is a '+ Sample Filter' button.

Live Demo: Recover brushed annotations and generate graph via code

31

```
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")
```

The screenshot shows the Facile Explorer interface for configuring a scatter plot. It consists of two main sections: 'X Axis' and 'Y Axis'. Each section has a text input field for a gene name, a dropdown menu for 'Select Gene Set', and a button for the assay type 'rnaseq'.

X Axis

- Gene Name: CCNB2
- Select Gene Set: [Dropdown arrow]
- Assay: rnaseq

Y Axis

- Gene Name: KRT17
- Select Gene Set: [Dropdown arrow]
- Assay: rnaseq


```
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, no_impute=T)

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

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

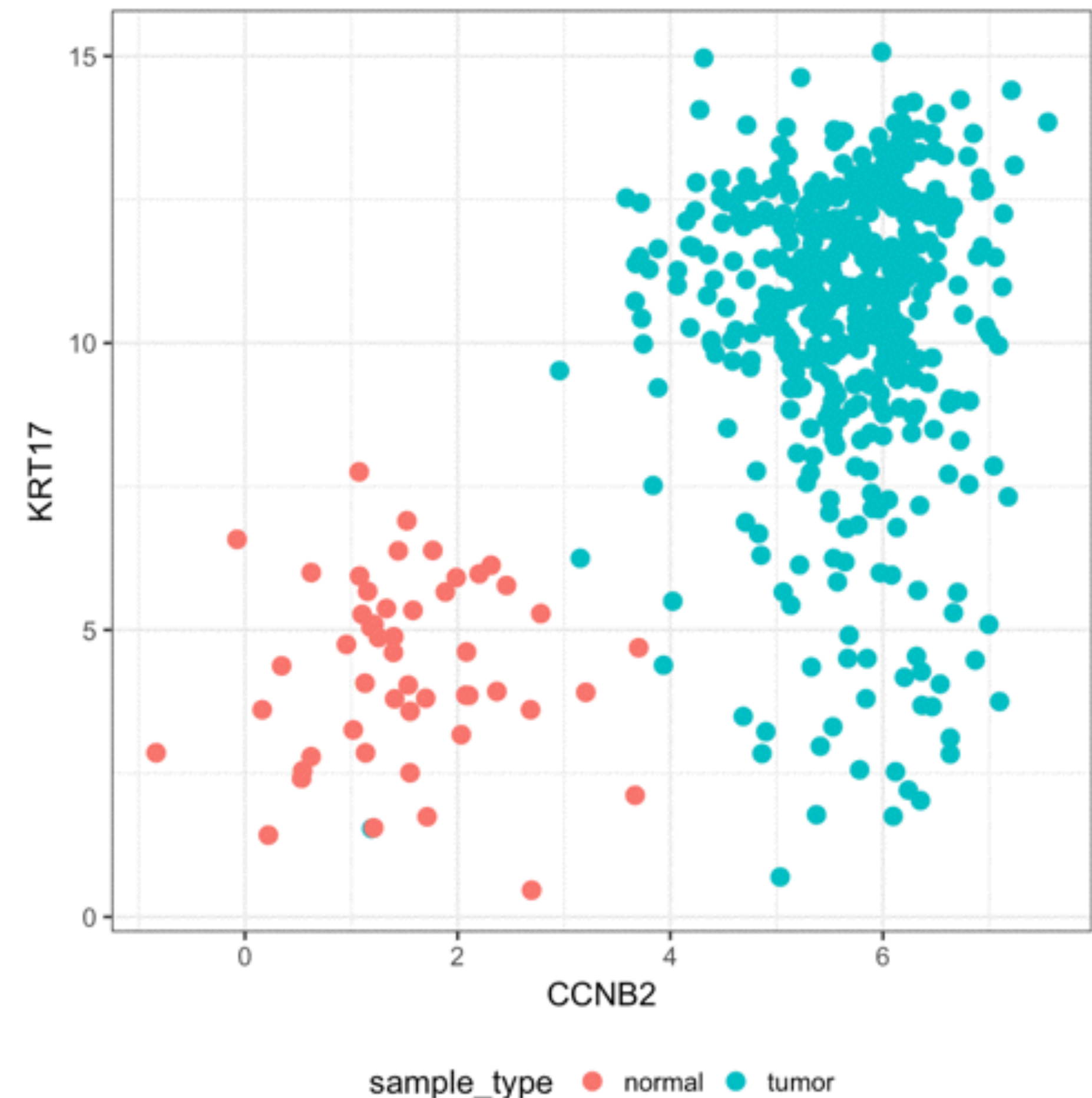

```
library(FacileExplorer)
library(FacileTCGADataset)
tcga <- FacileTCGADataset()

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")
```



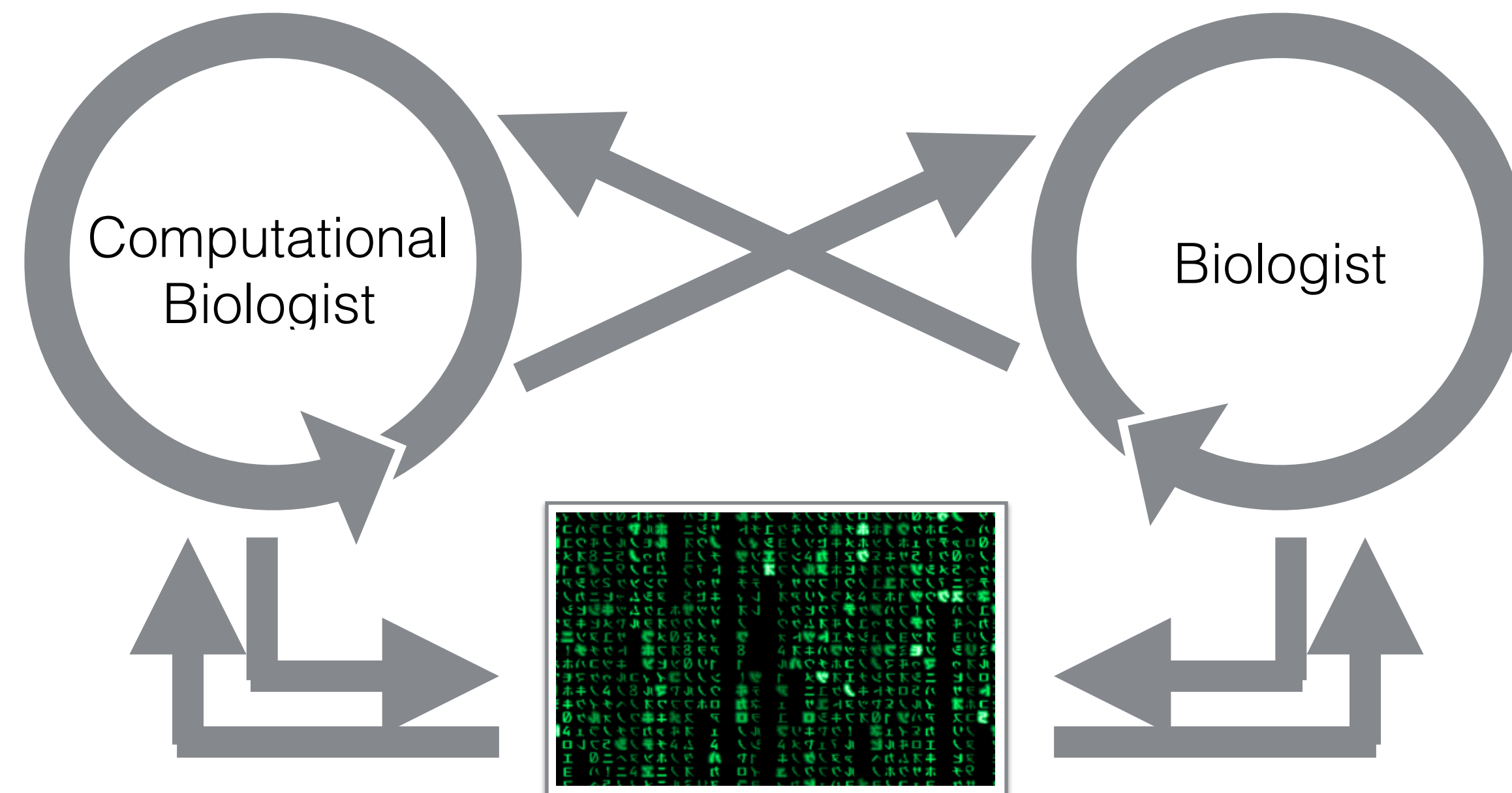
Long Term Goal

34

- 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://facilerports/d04be94#prolif-scatter")
```

- 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

36

The FacileData Working Group



Matthew
Albert



Marcia
Belvin



Richard
Bourgon



Haiyin
Chen



Sami
Mahrus



Vincent
Rouilly

Genentech

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



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

support.bioconductor.com
community



Ryan
Hafen