

# The ISB Cancer Genomics Cloud

Workshop at ISB

August 2<sup>nd</sup> 2016



# ISB-CGC Team Members



**Ilya Shmulevich**  
Varsha Dhankani  
David Gibbs  
Abigail Hahn  
Kelly Iverson  
Phyliss Lee  
Kalle Leinonen  
Michael Miller  
Suzanne Paquette  
Sheila Reynolds  
Zack Rodebaugh

**Jonathan Bingham**  
Matt Bookman  
Nicole Deflaux  
Larry Quesada

**David Pot**  
Mark Backus  
Ross Casanova  
Madelyn Reyes

# Our Goals for this Workshop

## for YOU

- to understand what the ISB-CGC platform provides
- to know how to find and use the data and tools that suit your needs
- to know your way around the Google Cloud Platform

## for US

- to better understand your use-cases, research goals, and needs
- to get feedback & suggestions
  - new data or metadata, sources, and/or formats?
  - new features?
  - new tools?
  - other ideas?

# What the ISB-CGC is...

Open platform integrated with the Google Cloud, providing Data, Tools, and Code Samples for a broad range of users

- Data as a Service (DaaS)
  - TCGA clinical and molecular data (multiple formats and technologies)
  - Genome- and Platform-Reference data (*eg* GENCODE, miRTarBase, Kaviar, *etc*)
  - Additional open-access data sets (*eg* CCLE)
- Tools & Applications (SaaS)
  - web app allows interactive and custom visualizations of the TCGA data
  - R and Python examples to get you started implementing your own custom analyses
  - programmatic API endpoints to supplement Google APIs
  - framework to help you use the new Google Pipelines API to run and manage tens of thousands of compute tasks

# What the ISB-CGC is...

Open platform integrated with the Google Cloud, providing Data, Tools, and Code Samples for a broad range of users

- Data as a Service (DaaS)
  - TCGA clinical and molecular data (multiple formats and technologies)
  - Genome- and Platform-Reference data (*eg* GENCODE, miRTarBase, Kaviar, *etc*)
  - Additional open-access data sets (*eg* CCLE)
- Tools & Applications (SaaS)
  - web app allows interactive and custom visualizations of the TCGA data
  - R and Python examples to get you started implementing your own custom analyses
  - programmatic API endpoints to supplement Google APIs
  - framework to help you use the new Google Pipelines API to run and manage tens of thousands of compute tasks



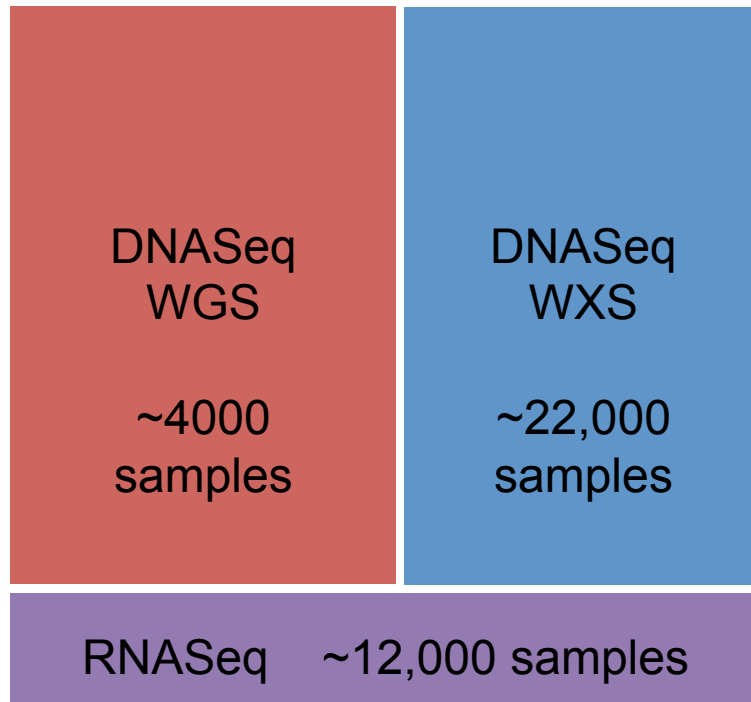
You are the builder

# What the ISB-CGC is not ...

- a repository from which to *download* large data sets
  - bring *your* tools and data to the cloud, to work with *our* tools and data!
- a “walled garden” with a *single* entrance
  - only “fence” is around the controlled-access data, but once you have a key, you can choose the path you prefer (GUI, CLI, API)
- a repository of “approved” or “recommended” pipelines for performing standard tasks
  - best left to the research community (*eg* DREAM challenges, GA4GH, PCAWG, tools developers, and you!)

# TCGA Size & Complexity

>1 PB of sequence data  
(controlled access)



~400,000 files of  
heterogeneous data  
(mostly open-access)

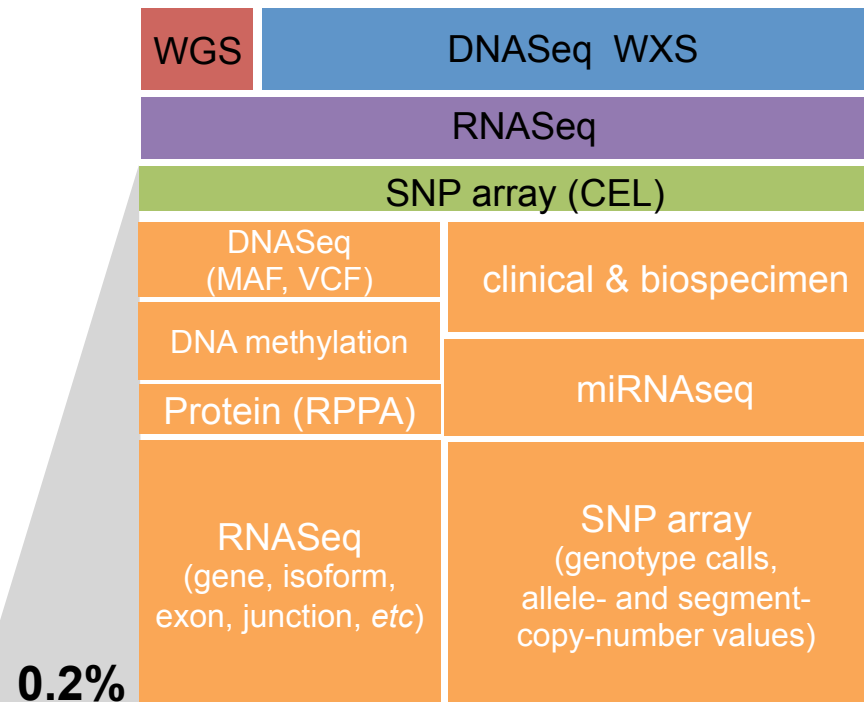


Table Details: Clinical\_data

Schema			
ParticipantBarcode	STRING	NULLABLE	Describe this field...
Study	STRING	NULLABLE	Describe this field...
Project	STRING	NULLABLE	Describe this field...
ParticipantUUID	STRING	NULLABLE	Describe this field...
TSSCode	STRING	NULLABLE	Describe this field...
age_at_initial_pathologic_diagnosis	INTEGER	NULLABLE	Describe this field...
anatomic_neoplasm_subdivision	STRING	NULLABLE	Describe this field...
batch_number	INTEGER	NULLABLE	Describe this field...
bcr	STRING	NULLABLE	Describe this field...
clinical_M	STRING	NULLABLE	Describe this field...
clinical_N	STRING	NULLABLE	Describe this field...

Table Details: Somatic\_Mutation\_calls

Schema			
ParticipantBarcode	STRING	NULLABLE	Describe this field...
Tumor_SampleBarcode	STRING	NULLABLE	Describe this field...
Tumor_AliquotBarcode	STRING	NULLABLE	Describe this field...
Tumor_SampleTypeLetterCode	STRING	NULLABLE	Describe this field...
Normal_SampleBarcode	STRING	NULLABLE	Describe this field...
Normal_AliquotBarcode	STRING	NULLABLE	Describe this field...
Normal_SampleTypeLetterCode	STRING	NULLABLE	Describe this field...
Study	STRING	NULLABLE	Describe this field...
Annotation_Transcript	STRING	NULLABLE	Describe this field...
CCLE_ONCOMAP_Total_Mutations_In_Gene	INTEGER	NULLABLE	Describe this field...
COSMIC_Total_Alterations_In_Gene	INTEGER	NULLABLE	Describe this field...
Center	STRING	NULLABLE	Describe this field...
Chromosome	STRING	NULLABLE	Describe this field...
DNAREpairGenes_Role	STRING	NULLABLE	Describe this field...
DbSNP_RS	STRING	NULLABLE	Describe this field...
DbSNP_Val_Status	STRING	NULLABLE	Describe this field...
DrugBank	STRING	NULLABLE	Describe this field...
End_Position	INTEGER	NULLABLE	Describe this field...
Entrez_Gene_Id	INTEGER	NULLABLE	Describe this field...
GC_Content	FLOAT	NULLABLE	Describe this field...
GENCODE_Transcript_Name	STRING	NULLABLE	Describe this field...
GENCODE_Transcript_Status	STRING	NULLABLE	Describe this field...
GENCODE_Transcript_Type	STRING	NULLABLE	Describe this field...
GO_Biological_Process	STRING	NULLABLE	Describe this field...
GO_Cellular_Component	STRING	NULLABLE	Describe this field...
GO_Molecular_Function	STRING	NULLABLE	Describe this field...
Gene_Type	STRING	NULLABLE	Describe this field...
Genome_Change	STRING	NULLABLE	Describe this field...

Table Details: DNA\_Methylation\_betas

Schema			
ParticipantBarcode	STRING	NULLABLE	Describe this field...
SampleBarcode	STRING	NULLABLE	Describe this field...
SampleTypeLetterCode	STRING	NULLABLE	Refer: <a href="https://tcga-data.nci.nih.gov/datareports/codeTablesReport.htm">https://tcga-data.nci.nih.gov/datareports/codeTablesReport.htm</a>
AliquotBarcode	STRING	NULLABLE	The Aliquot ID is an identifier/barcode of TCGA data. Refer: <a href="https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode">https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode</a>
Platform	STRING	NULLABLE	Refer: <a href="https://tcga-data.nci.nih.gov/datareports/codeTablesReport.htm">https://tcga-data.nci.nih.gov/datareports/codeTablesReport.htm</a>
Study	STRING	NULLABLE	TCGA disease type
Probe_Id	STRING	NULLABLE	Illumina's CpG loci IDs. Refer: <a href="http://www.illumina.com/content/dam/illumina-marketing/documents/products/technote/technote_cpg_loci_identification.pdf">http://www.illumina.com/content/dam/illumina-marketing/documents/products/technote/technote_cpg_loci_identification.pdf</a>
Beta_Value	FLOAT	NULLABLE	The beta value (β) is used to estimate the methylation level of the CpG locus using the ratio of intensities between methylated and unmethylated DNA.

Table Details: Biospecimen\_data

Schema			
ParticipantBarcode	STRING	NULLABLE	Describe this field...
SampleBarcode	STRING	NULLABLE	Describe this field...
SampleTypeLetterCode	STRING	NULLABLE	Describe this field...
SampleType	STRING	NULLABLE	Describe this field...
Study	STRING	NULLABLE	Describe this field...
Project	STRING	NULLABLE	Describe this field...
SampleTypeCode	STRING	NULLABLE	Describe this field...
avg_percent_lymphocyte_infiltration	FLOAT	NULLABLE	Describe this field...
avg_percent_monocyte_infiltration	FLOAT	NULLABLE	Describe this field...
avg_percent_necrosis	FLOAT	NULLABLE	Describe this field...
avg_percent_neutrophil_infiltration	FLOAT	NULLABLE	Describe this field...
avg_percent_normal_cells	FLOAT	NULLABLE	Describe this field...
avg_percent_stromal_cells	FLOAT	NULLABLE	Describe this field...
avg_percent_tumor_cells	FLOAT	NULLABLE	Describe this field...
avg_percent_tumor_nuclei	FLOAT	NULLABLE	Describe this field...
batch_number	INTEGER	NULLABLE	Describe this field...
bcr	STRING	NULLABLE	Describe this field...
days_to_collection	FLOAT	NULLABLE	Describe this field...
days_to_sample_procurement	FLOAT	NULLABLE	Describe this field...

Table Details: Copy\_Number\_segments

Schema			
max_percent_lymphocyte_infiltration	max_percent_lymphocyte_infiltration	max_percent_lymphocyte_infiltration	max_percent_lymphocyte_infiltration
max_percent_monocyte_infiltration	max_percent_monocyte_infiltration	max_percent_monocyte_infiltration	max_percent_monocyte_infiltration
max_percent_necrosis	max_percent_necrosis	max_percent_necrosis	max_percent_necrosis
max_percent_neutrophil_infiltration	max_percent_neutrophil_infiltration	max_percent_neutrophil_infiltration	max_percent_neutrophil_infiltration
max_percent_normal_cells	max_percent_normal_cells	max_percent_normal_cells	max_percent_normal_cells
max_percent_stromal_cells	max_percent_stromal_cells	max_percent_stromal_cells	max_percent_stromal_cells
max_percent_tumor_cells	max_percent_tumor_cells	max_percent_tumor_cells	max_percent_tumor_cells
max_percent_tumor_nuclei	max_percent_tumor_nuclei	max_percent_tumor_nuclei	max_percent_tumor_nuclei

Table Details: Annotations

Schema			
annotationId	INTEGER	NULLABLE	Describe this field...
annotationCategoryId	INTEGER	NULLABLE	Describe this field...
annotationCategoryName	STRING	NULLABLE	Describe this field...
annotationClassification	STRING	NULLABLE	Describe this field...
annotationNoteText	STRING	NULLABLE	Describe this field...
Study	STRING	NULLABLE	Describe this field...
ItemTypeName	STRING	NULLABLE	Describe this field...
ItemBarcode	STRING	NULLABLE	Describe this field...
AliquotBarcode	STRING	NULLABLE	Describe this field...
ParticipantBarcode	STRING	NULLABLE	Describe this field...
SampleBarcode	STRING	NULLABLE	Describe this field...
dateAdded	STRING	NULLABLE	Describe this field...
dateCreated	STRING	NULLABLE	Describe this field...
dateEdited	STRING	NULLABLE	Describe this field...

Table Details: mRNA\_UNC\_HiSeq\_RSEM

Schema			
ParticipantBarcode	STRING	NULLABLE	Describe this field...
SampleBarcode	STRING	NULLABLE	Describe this field...
AliquotBarcode	STRING	NULLABLE	Describe this field...
Study	STRING	NULLABLE	Describe this field...
SampleTypeLetterCode	STRING	NULLABLE	Describe this field...
Platform	STRING	NULLABLE	Describe this field...
original_gene_symbol	STRING	NULLABLE	Describe this field...
HGNC_gene_symbol	STRING	NULLABLE	Describe this field...
gene_id	INTEGER	NULLABLE	Describe this field...
normalized_count	FLOAT	NULLABLE	Describe this field...

Table Details: mRNA\_BCGSC\_HiSeq\_RPKM

Schema			
ParticipantBarcode	STRING	NULLABLE	Describe this field...
SampleBarcode	STRING	NULLABLE	Describe this field...
SampleTypeLetterCode	STRING	NULLABLE	Describe this field...
AliquotBarcode	STRING	NULLABLE	Describe this field...
Study	STRING	NULLABLE	Describe this field...
Platform	STRING	NULLABLE	Describe this field...
original_gene_symbol	STRING	NULLABLE	Describe this field...
HGNC_gene_symbol	STRING	NULLABLE	Describe this field...
gene_id	INTEGER	NULLABLE	Describe this field...
normalized_count	FLOAT	NULLABLE	Describe this field...

Table Details: miRNA\_expression

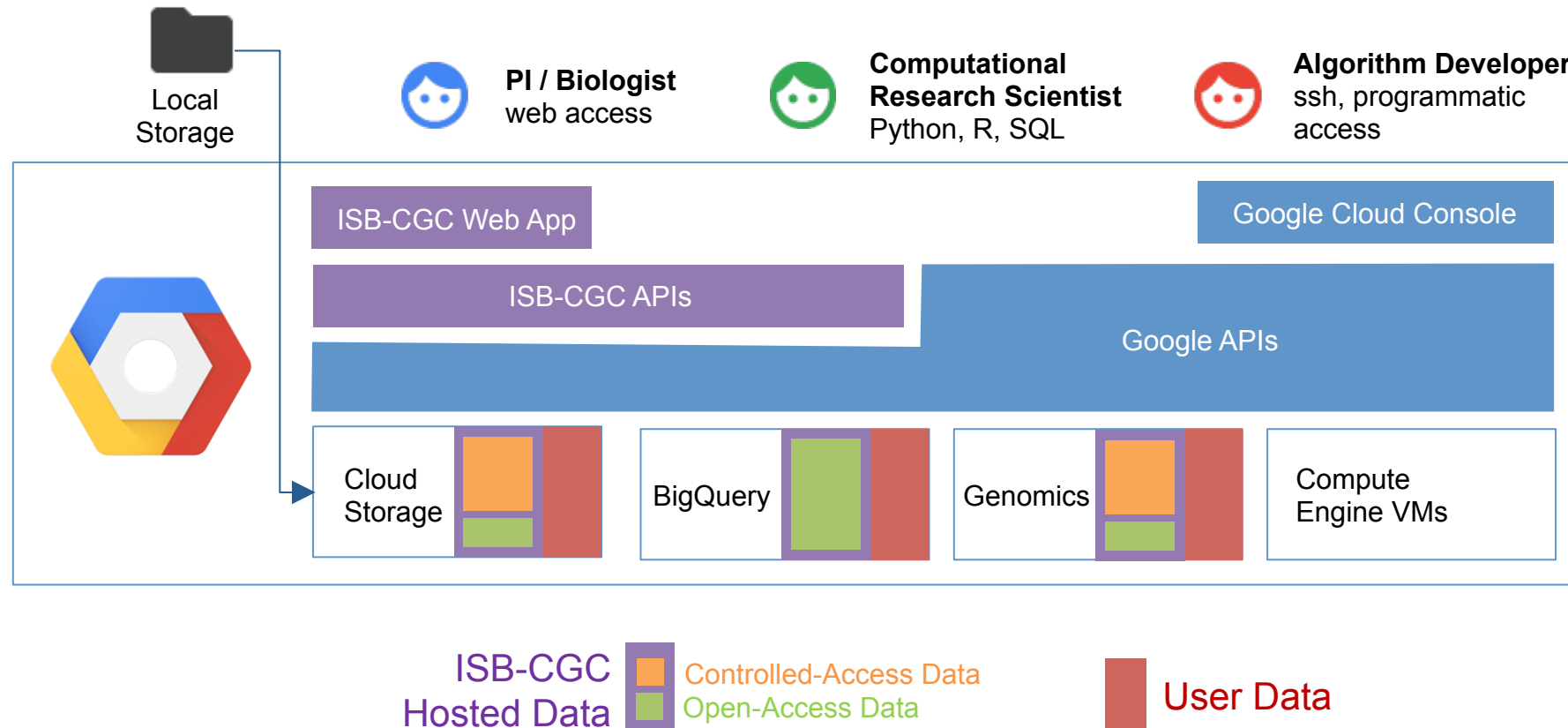
Schema			
ParticipantBarcode	STRING	NULLABLE	Describe this field...
SampleBarcode	STRING	NULLABLE	Describe this field...
AliquotBarcode	STRING	NULLABLE	Describe this field...
SampleTypeLetterCode	STRING	NULLABLE	Describe this field...
Study	STRING	NULLABLE	Describe this field...
Platform	STRING	NULLABLE	Describe this field...
mirna_id	STRING	NULLABLE	Describe this field...
mirna_accession	STRING	NULLABLE	Describe this field...
normalized_count	FLOAT	NULLABLE	Describe this field...





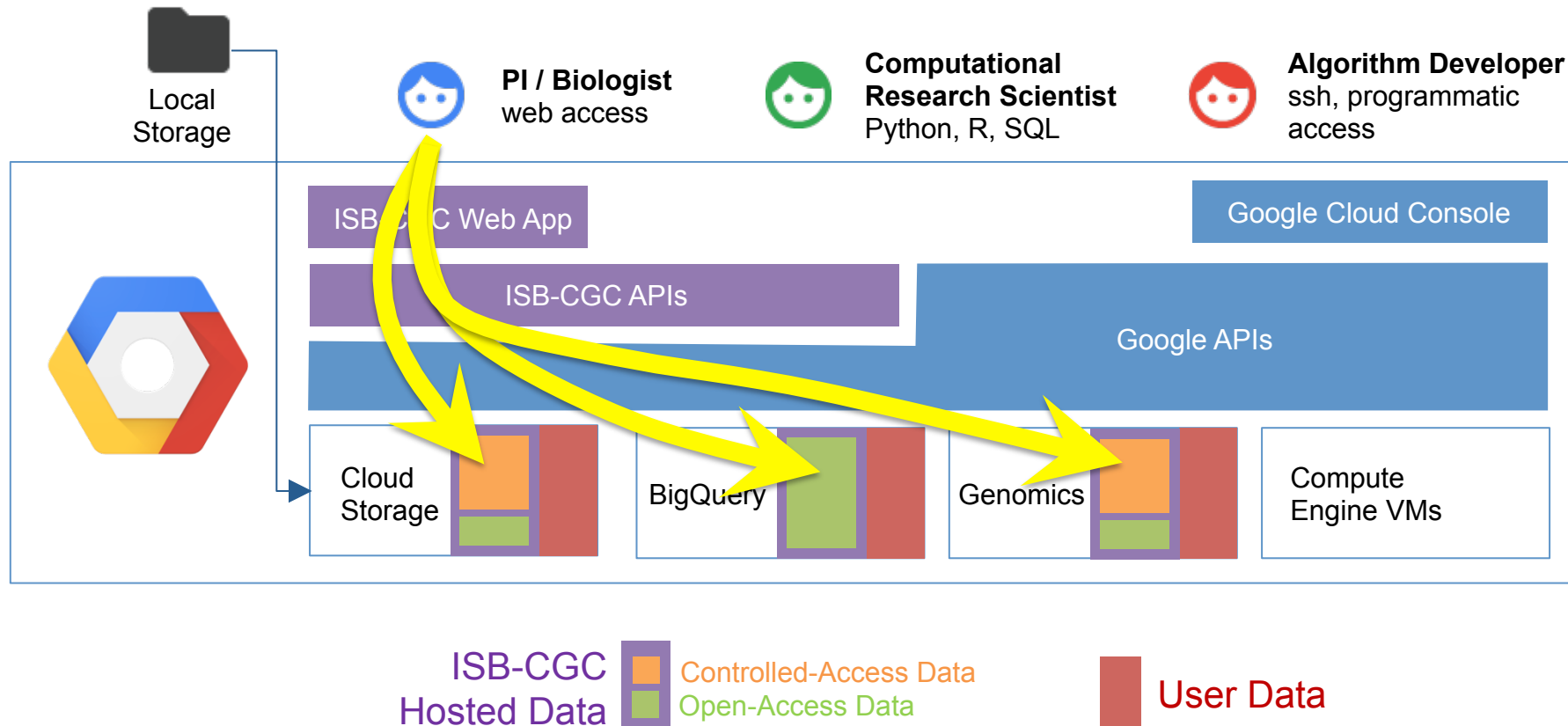
# ISB-CGC Approach

- build an open platform that can grow and evolve to satisfy a broad range of users and use-cases
- leverage the best existing tools and technologies, as they are released
- collaborate with the research community in areas of data standards, containers and workflows, *etc*
- provide a range of examples and tutorials to get newcomers up and running quickly!

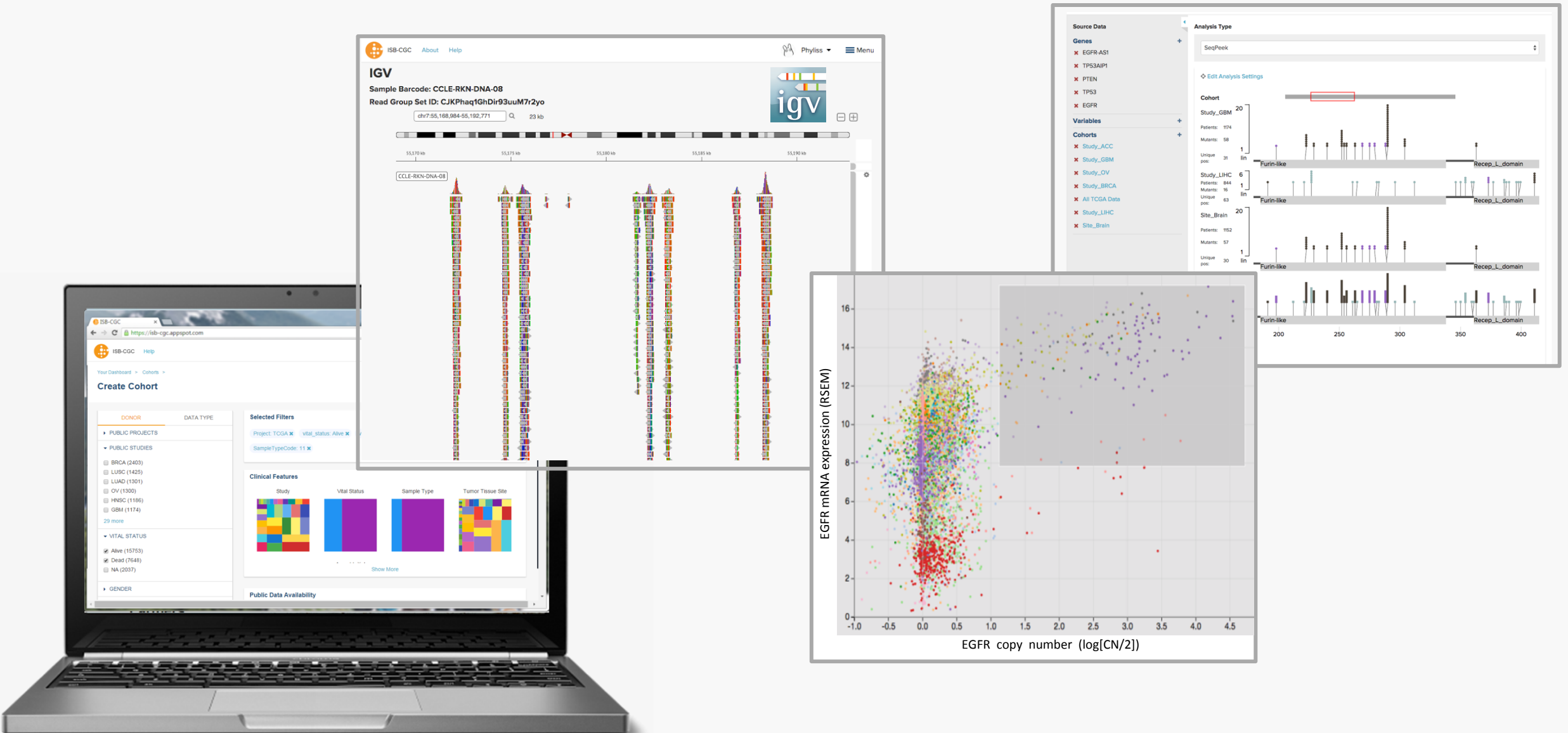


# ISB-CGC Approach

- build an open platform that can grow and evolve to satisfy a broad range of users and use-cases
- leverage the best existing tools and technologies, as they are released
- collaborate with the research community in areas of data standards, containers and workflows, *etc*
- provide a range of examples and tutorials to get newcomers up and running quickly!

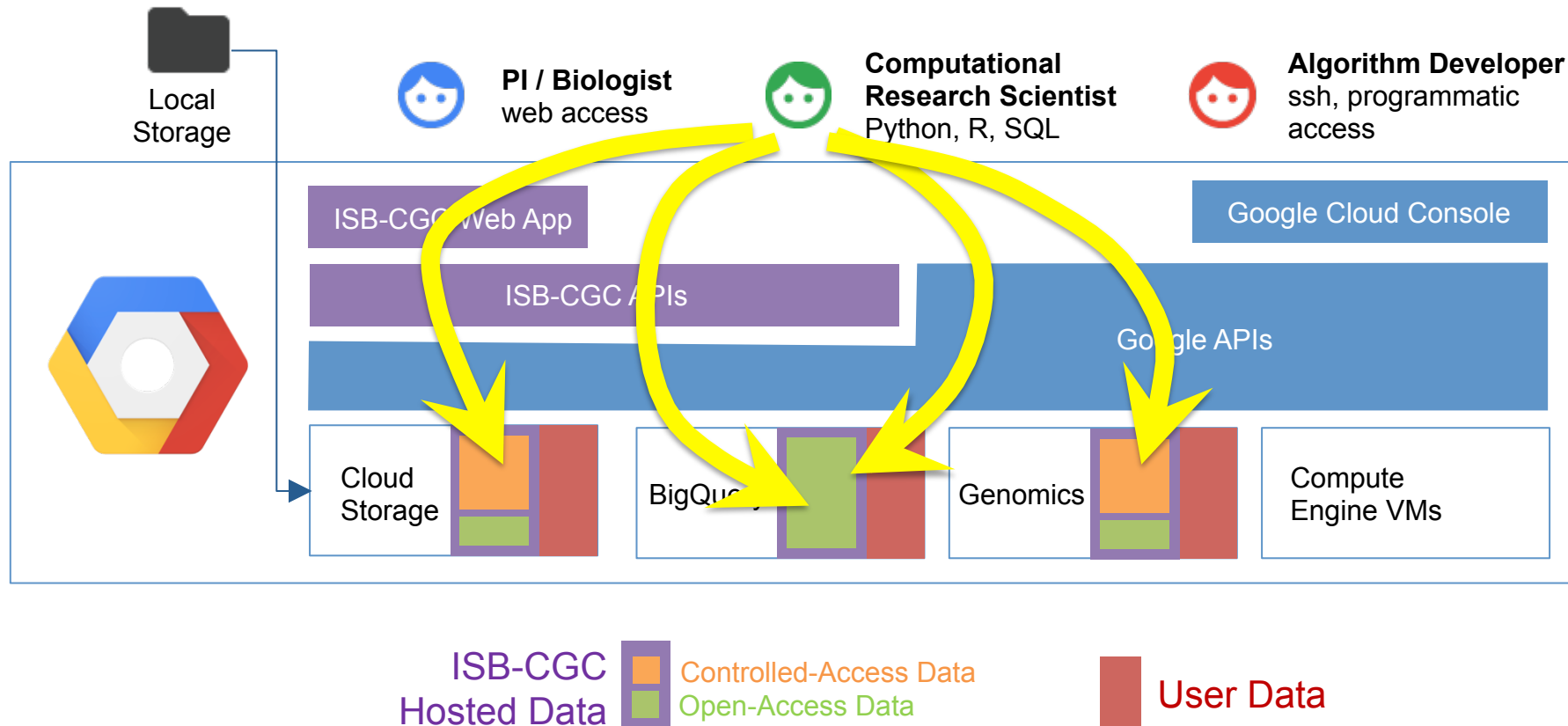


# Web access for the PI / Biologist:



# ISB-CGC Approach

- build an open platform that can grow and evolve to satisfy a broad range of users and use-cases
- leverage the best existing tools and technologies, as they are released
- collaborate with the research community in areas of data standards, containers and workflows, *etc*
- provide a range of examples and tutorials to get newcomers up and running quickly!



# Python, R, and SQL for the Computational Scientist:

IP[y]: IPython  
Interactive Computing



### Copy Number segments (Broad)

The goal of this notebook is to introduce you to the Copy Number (CN) segments Broad table.

This table contains all available TCGA Level-3 copy number data produced by the Broad Genome Wide SNP6 array, as of October 2015. (Actual archive dates range from April 2011 to October 2015). The data was downloaded from the DCC, and data extracted from all files matching the criteria. Each of these segmentation files has six columns: Sample, Chromosome, Start, End, and Segment\_Mean. During ETL the sample identifier contained in the segmentation files was mapped to the SDRF file in the associated image-tab archive.

In order to work with BigQuery, you need to import the python bigquery module (get the name(s) of the table(s) you are going to be working with:

```
import gcp.bigquery as bq
cn_BQtable = bq.Table('tcbg-cgc:tcga_201510_alpha.Copy_Number_Segments')
```

From now on, we will refer to this table using this variable (\$cn\_BQtable), but we will use the table name each time.

Let's start by taking a look at the table schema:

```
%%bigquery schema --table $cn_BQtable
```

name	type	mode	description
ParticipantBarcode	STRING		
SampleBarcode	STRING		
SampleTypeLetterCode	STRING		
AliquotBarcode	STRING		
Study	STRING		
Platform	STRING		
Chromosome	STRING		
Start	INTEGER		
End	INTEGER		
Num_Probes	INTEGER		
Segment_Mean	FLOAT		

Unlike most other molecular data types in which measurements are available for a single sample, this data is produced using a data-driven approach for each aliquot and its size and positions of these segments can vary widely from one sample to another.

Now we'll use matplotlib to create some simple visualizations.

```
import numpy as np
import matplotlib.pyplot as plt
```

For the segment means, let's invert the log-transform and then bin the values to see what the distribution is.

```
%%sql --module getCNhist
SELECT
  1/n_bin,
  COUNT(*) AS n
FROM (
  SELECT
    Segment_Mean,
    (2.*POW(2,Segment_Mean)) AS 1/n_bin,
    INTEGER(((2.*POW(2,Segment_Mean))+0.50)/1.0) AS 1/n_bin
  FROM
    cn_BQtable
  WHERE
    ((End-Start+1)>1000 AND SampleTypeLetterCode="TP") )
GROUP BY
  1/n_bin
HAVING
  ( n > 2000 )
ORDER BY
  1/n_bin ASC
```

```
CNhist = bq.Query(getCNhist,t=cn_BQtable).results().to_dataframe()
bar_width=0.80
plt.bar(CNhist['1/n_bin']*.1,CNhist['n'],bar_width,alpha=0.8);
plt.xticks(CNhist['1/n_bin']*10,CNhist['1/n_bin']);
plt.title('Histogram of Average Copy-Number');
plt.xlabel('# of segments');
plt.ylabel('integer copy-number');
```

```
bin
ORDER BY
bin ASC
```

```
%%sql --module getSLhist_1k_amp
SELECT
  bin,
  COUNT(*) AS n
FROM (
  SELECT
    (END-Start+1) AS segLength,
    INTEGER((END-Start+1)/1000) AS bin
  FROM
    cn_BQtable
  WHERE
    (END-Start+1)<1000000 AND SampleTypeLetterCode="TP" AND Segment_Mean>0.5
  GROUP BY
    bin
ORDER BY
  bin ASC
```

```
SLhistdel = bq.Query(getSLhist_1k_del,t=cn_BQtable).results().to_dataframe()
SLhistamp = bq.Query(getSLhist_1k_amp,t=cn_BQtable).results().to_dataframe()
```

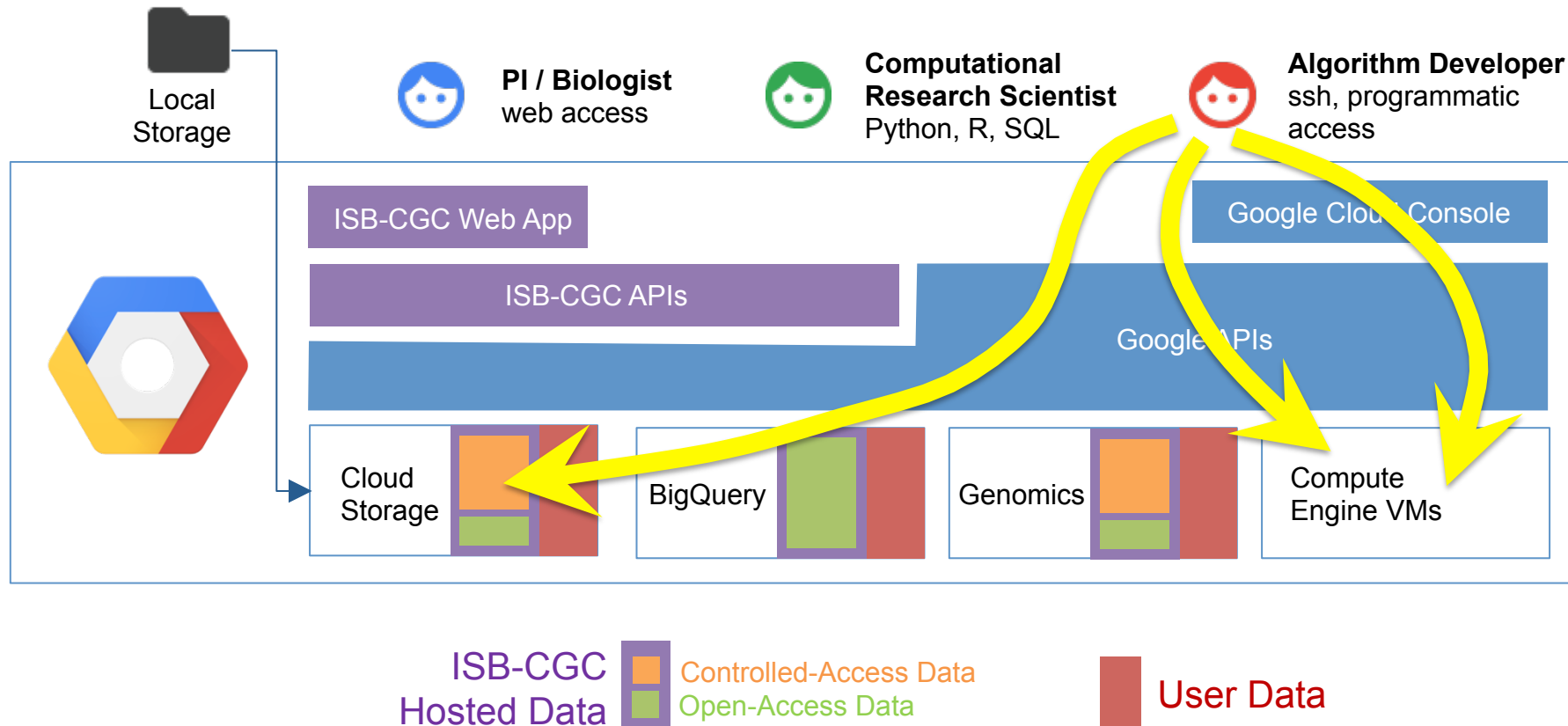
```
plt.plot(SLhistdel['bin'],SLhistdel['n'],'ro');
plt.plot(SLhistamp['bin'],SLhistamp['n'],'bo');
plt.plot(SLhistdel['bin'],SLhistdel['n'],'go',alpha=0.3);
plt.xscale('log');
plt.ylabel('# of segments');
plt.xlabel('Segment length (kb)');
plt.title('Distribution of Segment Lengths');
```

And now we'll take a look at histograms of the average copy-number for these three genes. TP53 (in green) shows a significant number of partial deletions (CN<0), while MYC (in blue) shows some partial amplifications -- more frequently than EGFR, while EGFR (pale red) shows a few extreme amplifications (log2(CN/2) > 2). The final figure shows the same histograms on a semi-log plot to bring up the rarer events.

```
binwidth = 0.2
binvals = np.arange(-2+(binwidth/2.), 6-(binwidth/2.), binwidth)
plt.hist(tp53CN['avgCN'],bins=binvals,normed=False,color='green',alpha=0.9,label='TP53');
plt.hist(mycCN['avgCN'],bins=binvals,normed=False,color='blue',alpha=0.7,label='MYC');
plt.hist(egfrCN['avgCN'],bins=binvals,normed=False,color='red',alpha=0.5,label='EGFR');
```

# ISB-CGC Approach

- build an open platform that can grow and evolve to satisfy a broad range of users and use-cases
- leverage the best existing tools and technologies, as they are released
- collaborate with the research community in areas of data standards, containers and workflows, *etc*
- provide a range of examples and tutorials to get newcomers up and running quickly!



# A generalizable workflow using Docker

