CytoData 2018
The Challenge

CytoData

CytoData Symposium 2018

September 21-25

Gold Sponsor

Silver Sponsor

Bronze Sponsors

Prizes and Hackathon Sponsors

Table of contents - Part I

1. Welcome and Schedule

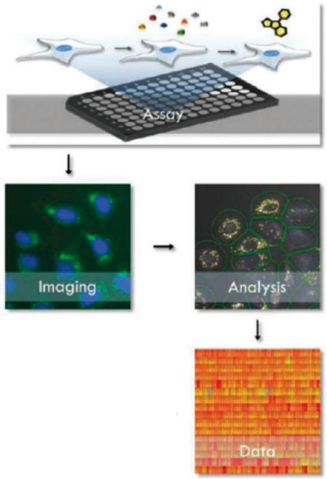
Day 1:

- ❑ 9:00 am - 10:00 am challenge introduction
- ❑ 10:00 am - 11:00 am hands on session
- ❑ 11:00 am - 5:00 pm hacking (with lunch break)
- ❑ 5:00 pm - 5:30 pm discussion and team presentations
- ❑ 6:00 pm pizza at Area 4

Day 2

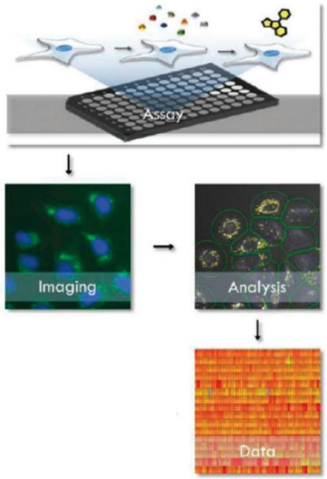
- ❑ 9:00 am - 4:00 pm Hacking
- ❑ 4:00 pm - 4:30 pm final submission and evaluation
- ❑ 4:30 pm - 5:00 pm team presentations and winner announcement

2. Problem Description - Cross Data set matching

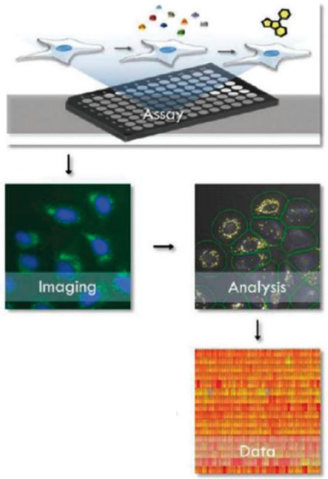


Batch 1

2. Problem Description - Cross Data set matching

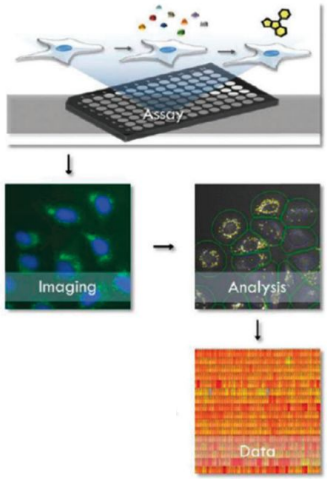


Batch 1

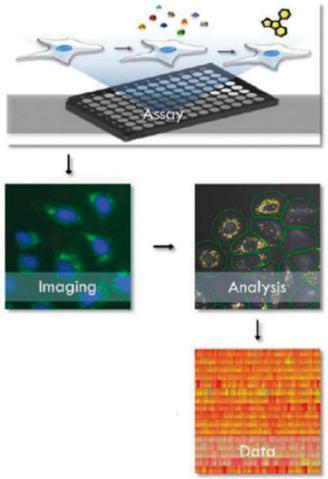


Batch 2

2. Problem Description - Cross Data set matching



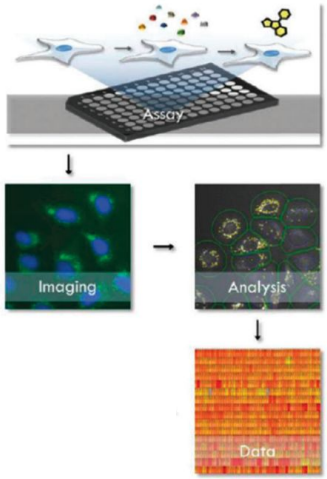
Batch 1



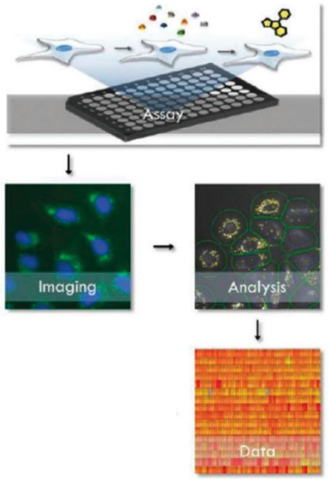
Batch 2 - differences

- microscope
- cells
- perturbations
- time
- Plate layout
- Summer - winter?

2. Problem Description - Cross Data set matching



Batch 1



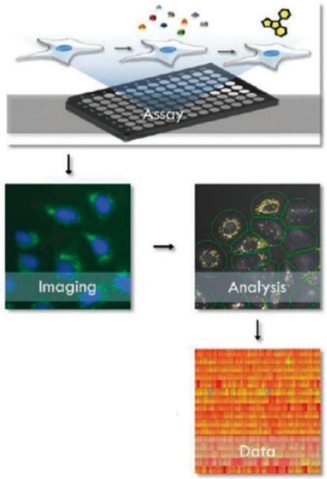
Batch 2 - differences

- microscope
- cells
- perturbations
- time
- Plate layout
- Summer - winter?

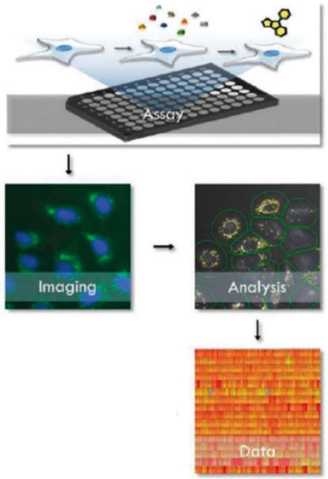
- Imaging data is subject to **batch effects/undesired artifacts** (biology!)
- two batches of microscopy images with the same treatments, but acquired under different technical conditions show differences in the quantitative measurements
- These differences are not due to meaningful biological variations

- Question: how can we remove these difference using **computational methods?**

2. Problem Description - Cross Data set matching



Batch 1



Batch 2 - differences

- microscope
- cells
- perturbations
- time
- Plate layout
- Summer - winter?

- Analyze the profiles of **two different batches** of data and design **computational methods** to correct batch effects .
- A successful method will be able to **align the information** content of both batches
- **Goal**: **transform** data so that profiles of the same treatment have similar measurements without distorting the relationships among other treatments

2. Problem Description - Cross Data set matching

a Detection of batch effects

1. Determine well-averaged feature vectors

Sample No.	Feature No. 1	2	3	4	5	6	7
S1	−0.77	0.49	−0.71	−0.99	−0.97	−0.18	−0.94
S2	−0.47	0.75	−0.17	−0.98	−0.72	0.31	−0.82
S3	−0.15	−0.45	−0.43	0.37	0.05	−0.12	0.39
S4	−0.87	−1.17	−0.56	−2.36	−1.45	0.23	−0.74

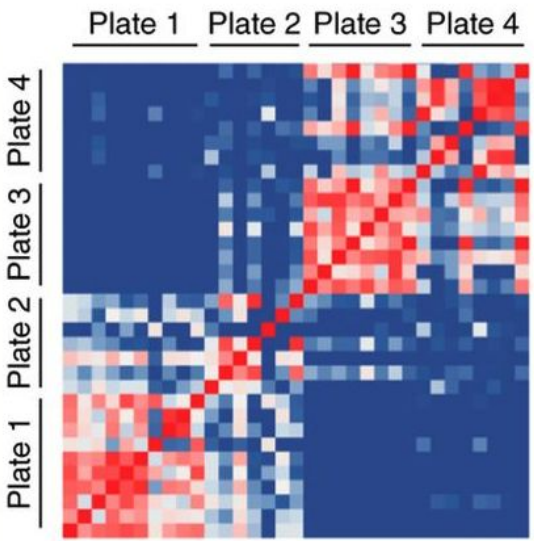
2. Calculate distance or correlation between samples

$$(S1,S2) = \sqrt{(S1_1 - S2_1)^2 + (S1_2 - S2_2)^2 + \cdots + (S1_n - S2_n)^2}$$

3. Create matrix with pairwise distances

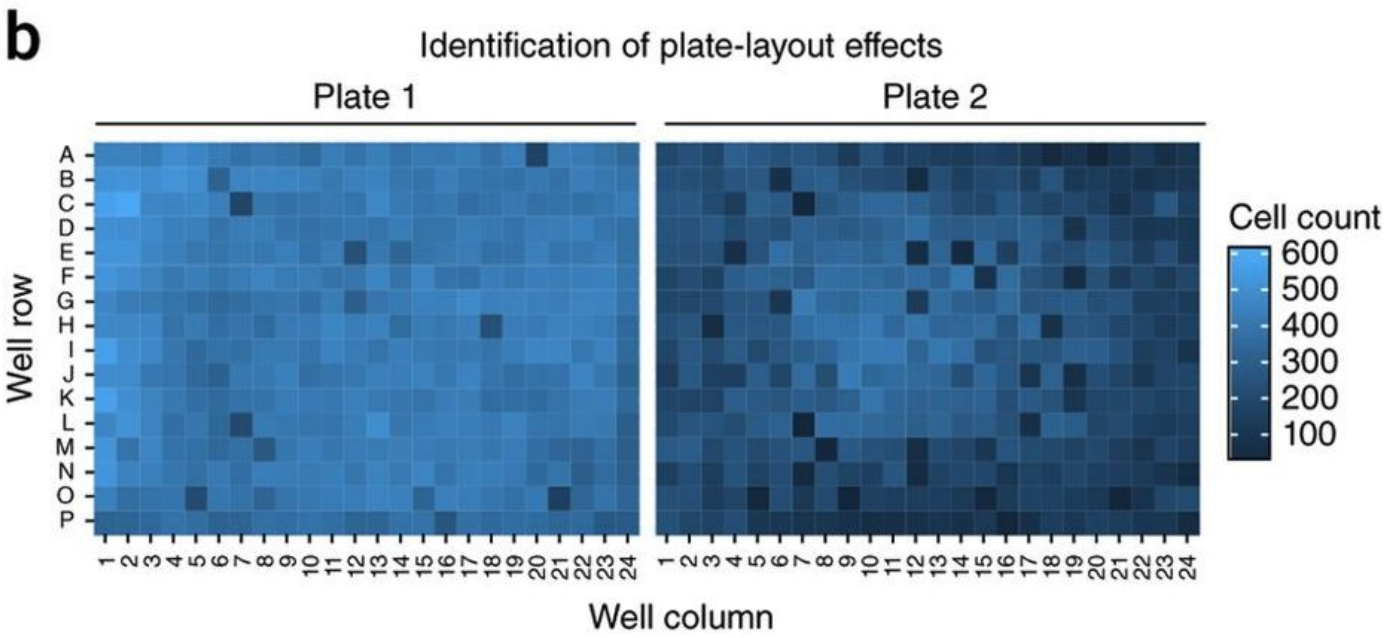
	S1	S2	S3	S4
S1	2.27	2.77	0.93	0.00
S2	2.58	2.87	0.00	0.93
S3	3.70	0.00	2.87	2.77
S4	0.00	3.70	2.58	2.27

4. Plot distance matrix for controls across screen to reveal batch effects



Caicedo et al. 2017

2. Problem Description - Cross Data set matching



Caicedo et al. 2017

2. Problem Description - Cross Data set matching

Caicedo et al. 2017

High-throughput Imaging using Cell Painting

	Dataset	Perturbation	Treatments	Cell line	Plates	Images	Size (GB)
Day 1	**BBBC037** LUAD	genetic	596	A549	16	55,296	802
	BBBC043 TA-ORF	genetic	196	U2OS	5	11,520	246
Day 2	**BBBC022** bioactives	chemical	1,600	U2OS	20	69,120	278
	BBBC036 CDRP	chemical	2,500	U2OS	55	126,720	561

Profiling data - available formats

- CellProfiler data
  - Single cell data as SQL DB
  - **Aggregated features on replicate level**
    - All features ~1,700
    - Feature selected
    - Feature selected and normalized

- DeepProfiler data
  - UNet-based nucleus segmentation
  - Inception-ResNet V2 features (single cells)
  - 1,520 features per channel (5 channels)
  - **Aggregated features on replicate level**

Profiling data - available formats

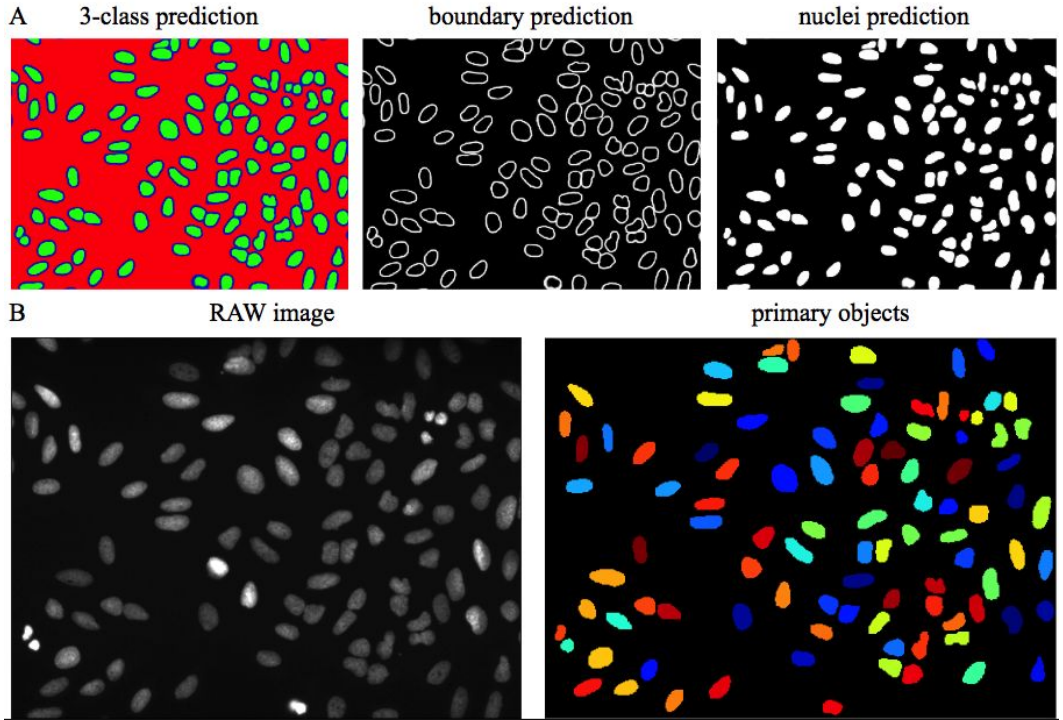
- CellProfiler data
  - Single cell data as SQL DB
  - **Aggregated features on replicate level**
    - All features ~1,700
    - Feature selected
    - Feature selected and normalized

- DeepProfiler data
  - UNet-based nucleus segmentation
  - Inception-ResNet V2 features (single cells)
  - 1,520 features per channel (5 channels)
  - **Aggregated features on replicate level**

Profiling data - available formats

DeepProfiler
image-based profiling using deep learning

Step 1: segment nuclei using pre-trained UNet



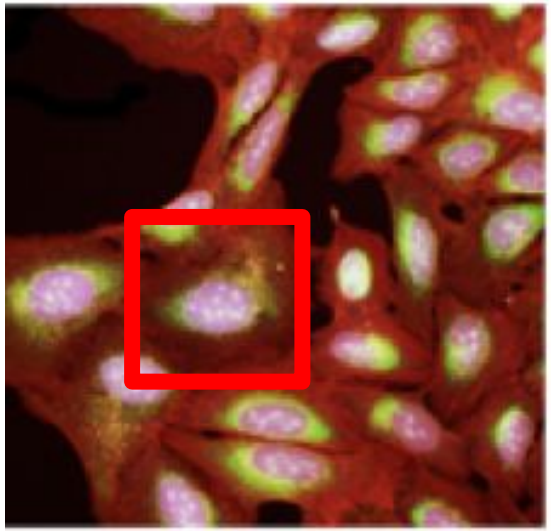
Profiling data - available formats

DeepProfiler
image-based profiling using deep learning

Step 1: segment nuclei using pre-trained UNet

Step 2: extract feature using a pre-trained Inception ResNet vs2

- Bounding box 128x128
- Feature extracted in 5 channels independently
- 1,520 features per channel
- Data is provided aggregated to replicate level



Challenge - Access to data sets

- BBBC datasets are hosted on AWS
- AWS Public Dataset Program
- Publicly accessible
  - image data and single cell information:
  - `s3://cytodata/datasets/`

- Aggregated features
  - `s3://cytodata/evaluation/`
  - Splitted in **test and training** sets

- Data sets are accessible for the hackathon and go online during October

Challenge - Access to data sets

- BBBC datasets are hosted on AWS
- AWS Public Dataset Program
- Publicly accessible
  - image data and single cell information:
  - `s3://cytodata/datasets/`

- Aggregated features
  - `s3://cytodata/evaluation/`
  - Splitted in **test and training** sets

- Data sets are accessible for the hackathon and go online during October

Aggregated features for day 1

```
├── LUAD-BBBC043-Caicedo
│   ├── profiles_cp
│   │   ├── bbbc043_test.csv
│   │   └── bbbc043_train.csv
│   └── profiles_dp
│       ├── bbbc043_test.csv
│       └── bbbc043_train.csv
└── TA-ORF-BBBC037-Rohban
    ├── profiles_cp
    │   ├── bbbc037_test.csv
    │   └── bbbc037_train.csv
    └── profiles_dp
        ├── bbbc037_test.csv
        └── bbbc037_train.csv
```

Challenge - Access to data sets

https://github.com/cytodata/cytodata-hackathon-2018/blob/master/cytodata-toolkit/datasets.csv

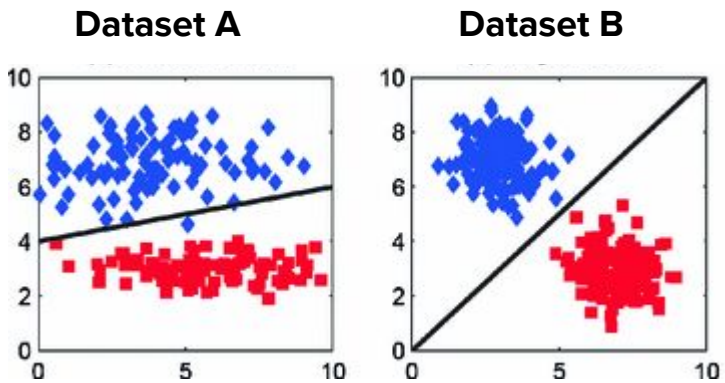
10 lines (9 sloc) | 1012 Bytes

Raw | Blame | History

Search this file...

	Dataset	Partition	Features	Link
2	BBBC037	Test	CellProfiler	https://s3.amazonaws.com/cytodata/evaluation/TA-ORF-BBBC037-Rohban/profiles_cp/bbbc037_test.csv
3	BBBC037	Train	CellProfiler	https://s3.amazonaws.com/cytodata/evaluation/TA-ORF-BBBC037-Rohban/profiles_cp/bbbc037_train.csv
4	BBBC043	Test	CellProfiler	https://s3.amazonaws.com/cytodata/evaluation/LUAD-BBBC043-Caicedo/profiles_cp/bbbc043_test.csv
5	BBBC043	Train	CellProfiler	https://s3.amazonaws.com/cytodata/evaluation/LUAD-BBBC043-Caicedo/profiles_cp/bbbc043_train.csv
6	BBBC037	Test	DeepLearning	https://s3.amazonaws.com/cytodata/evaluation/TA-ORF-BBBC037-Rohban/profiles_dp/bbbc037_test.csv
7	BBBC037	Train	DeepLearning	https://s3.amazonaws.com/cytodata/evaluation/TA-ORF-BBBC037-Rohban/profiles_dp/bbbc037_train.csv
8	BBBC043	Test	DeepLearning	https://s3.amazonaws.com/cytodata/evaluation/LUAD-BBBC043-Caicedo/profiles_dp/bbbc043_test.csv
9	BBBC043	Train	DeepLearning	https://s3.amazonaws.com/cytodata/evaluation/LUAD-BBBC043-Caicedo/profiles_dp/bbbc043_train.csv

How to solve this problem?



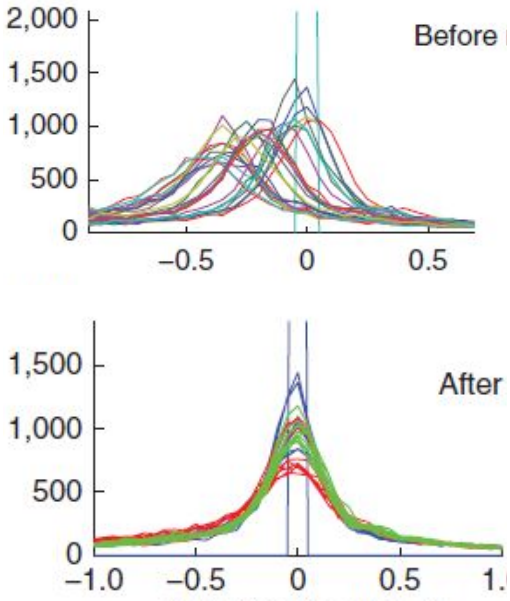
Treatments

Controls

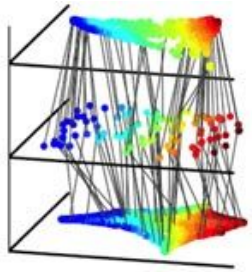
Transform feature space by aligning known common data points

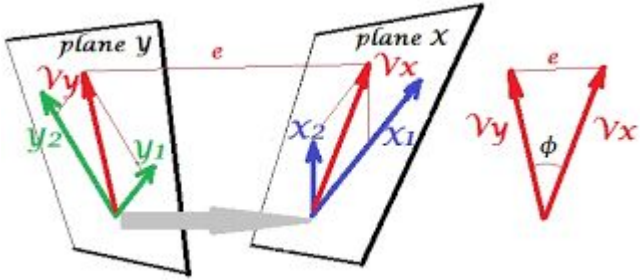
How to solve this problem?

Feature normalization



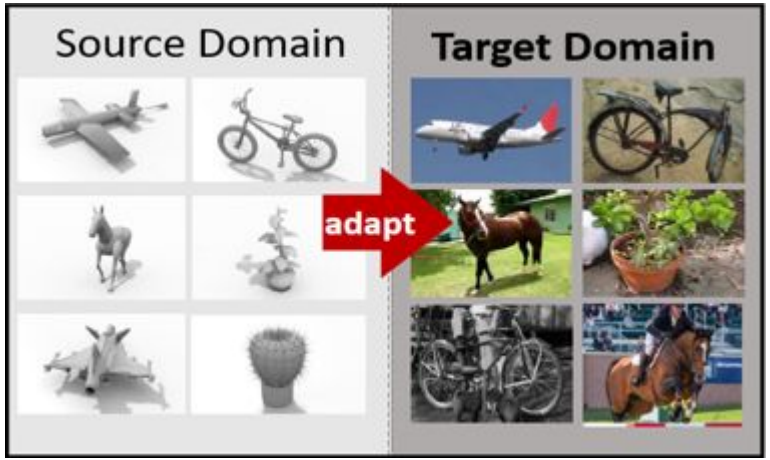
Subspace / manifold alignment



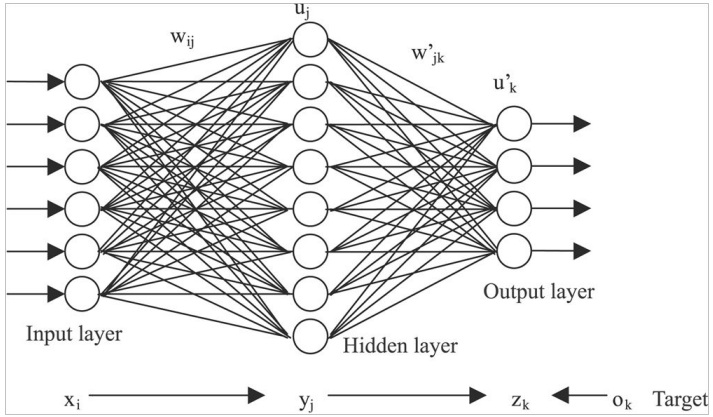


How to solve this problem?

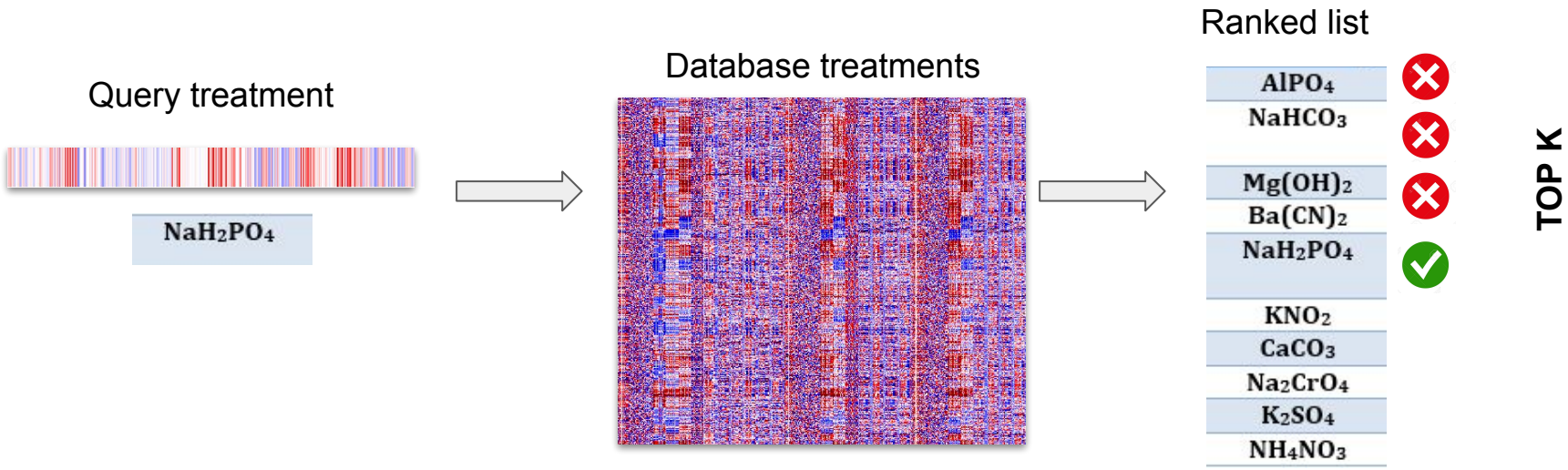
Domain adaptation



Feature encoding with NN



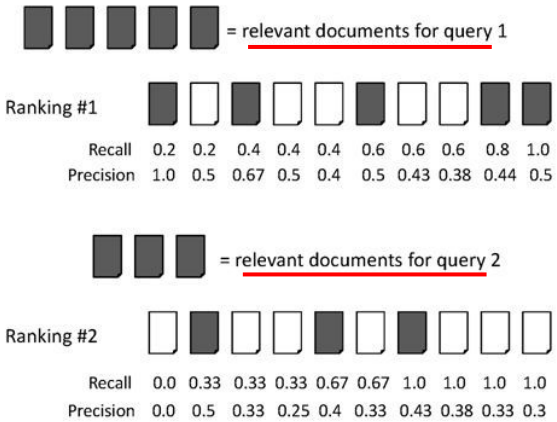
Evaluation metric - Treatment Matching



Treatment matching score:
Fraction of compounds that find the correct treatment in the top K results

Evaluation metric - Mean Average Precision

Consider the problem of document (treatment) retrieval



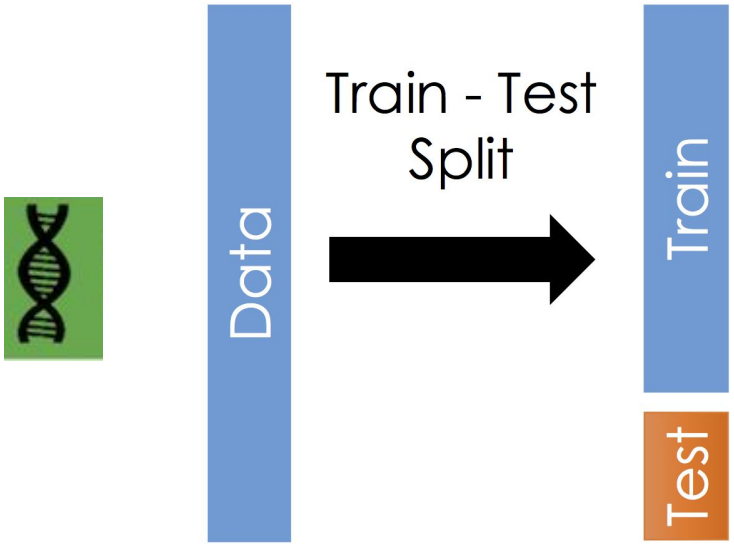
$$average\ precision\ query\ 1 = (1.0 + 0.67 + 0.5 + 0.44 + 0.5)/5 = 0.62$$
$$average\ precision\ query\ 2 = (0.5 + 0.4 + 0.43)/3 = 0.44$$

$$mean\ average\ precision = (0.62 + 0.44)/2 = 0.53$$

Relevant treatments: Defined by ground truth biological connections (gene pathway or Mechanism of Action (MoA))

Dataset for Day 1

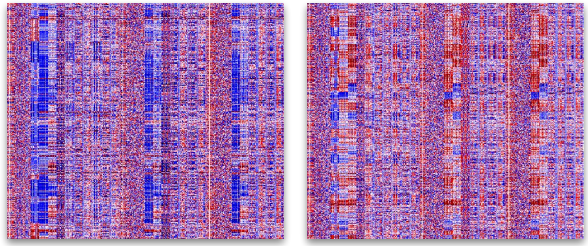
- Genetic perturbations
- Each database has training treatments and testing treatments
- Replicate level (well) profiles



- Learn models on **training** data
- Apply models to train **AND** test data

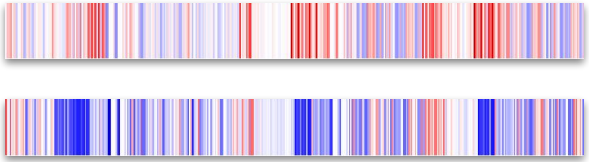
Submission for Day 1

Replicate level data



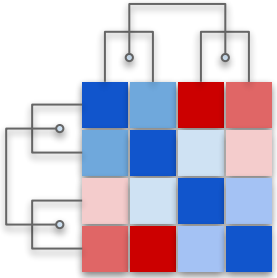
Aggregate

Treatment level profiles



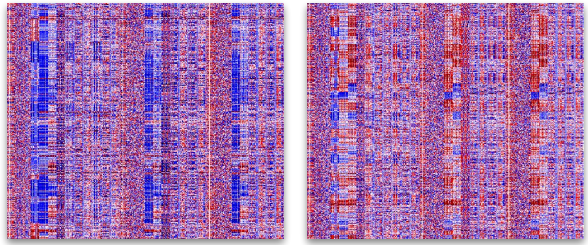
Similarity

Connectivity matrix



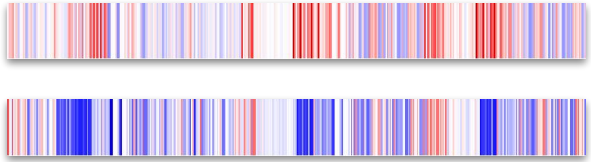
Submission for Day 1

Replicate level data



Aggregate

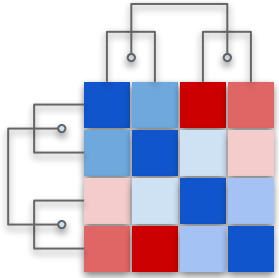
Treatment level profiles



Similarity

Submit this

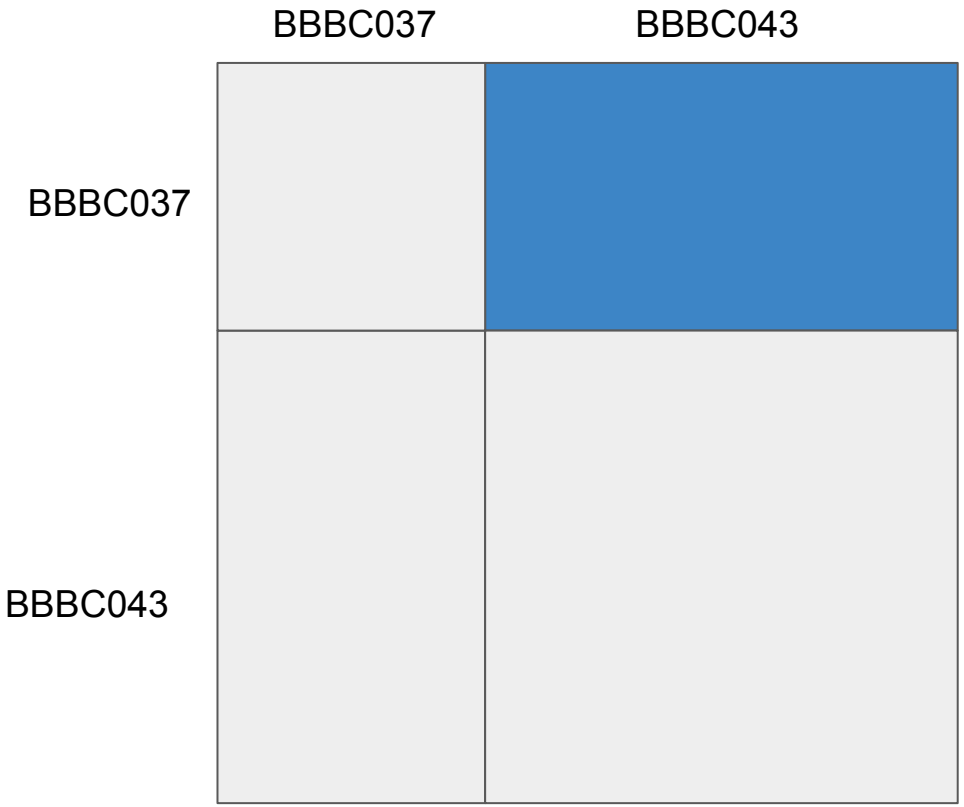
Connectivity matrix



Submission for Day 1 - connectivity or similarity scores

	BBBC037	BBBC043
BBBC037		
BBBC043		

Submission for Day 1 - connectivity or similarity scores



Submit This!

Example code is available!

Table of Contents - Part II

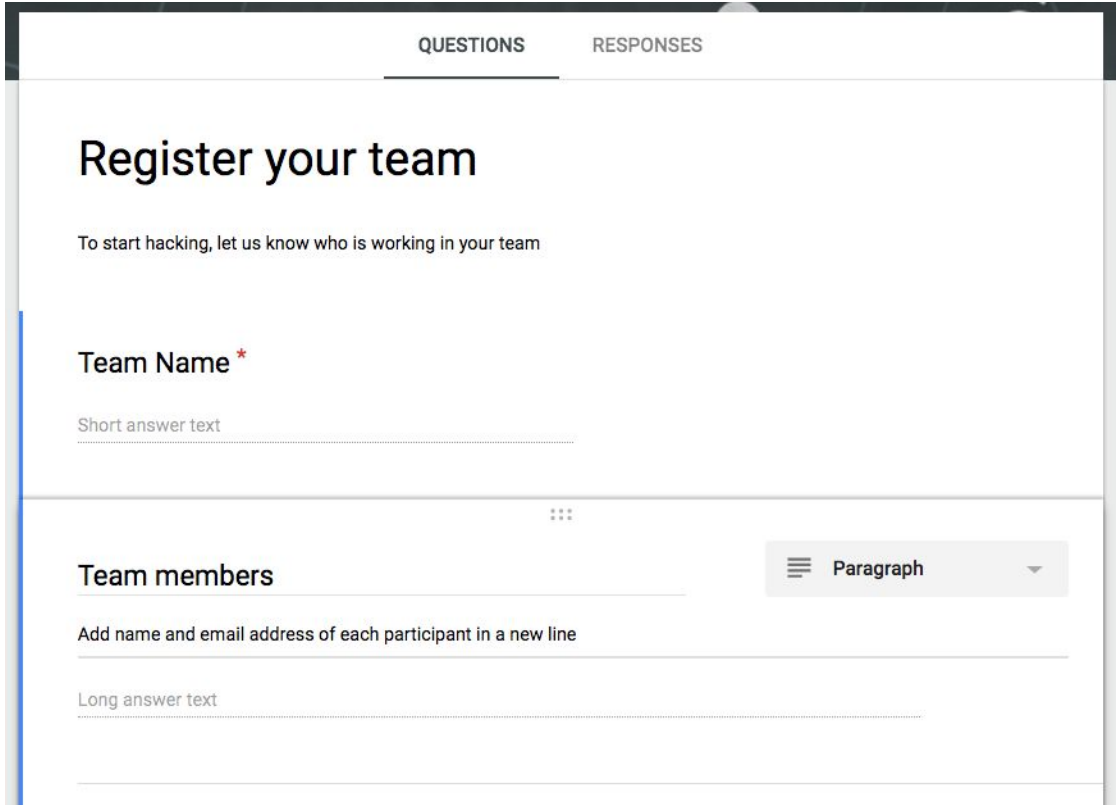
1. Resources
    a. GitHub Repo https://github.com/cytodata/cytodata-hackathon-2018
    b. Live demos (R and Python)
        i. R notebook
        ii. Python notebook
    c. Submissions: upload system
    d. AWS infrastructure
    e. Slack
2. Teams
    a. Instructions to form a team [3-6 people from different institutions]
    b. Register teams here
    c. AWS access / assign resources [Christopher, @cfriedri]

Team up

We suggest to create team s of up to 6 people from different institutes!

Please sign up your team using this google form:

https://goo.gl/forms/yMzMKzec0MOxNNjq2



AWS infrastructure - EC2 instances

Each team has access to one EC2 instance

```
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 396.37                 Driver Version: 396.37                    |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla K80           On   | 00000000:00:1E.0 Off |                    0 |
| N/A   37C    P8    31W / 149W |      0MiB / 11441MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
```

- EC2 instance p2.xlarge
  - 61 GB Memory
  - 1 GPU
  - 4x CPU cores
- Access via **ssh** and **jupyter notebook**
- AMI predefined for several deep learning environment

Let us know if you need more resources!

Ready. Set. Go!

CytoData 2018 - Challenge

CytoData