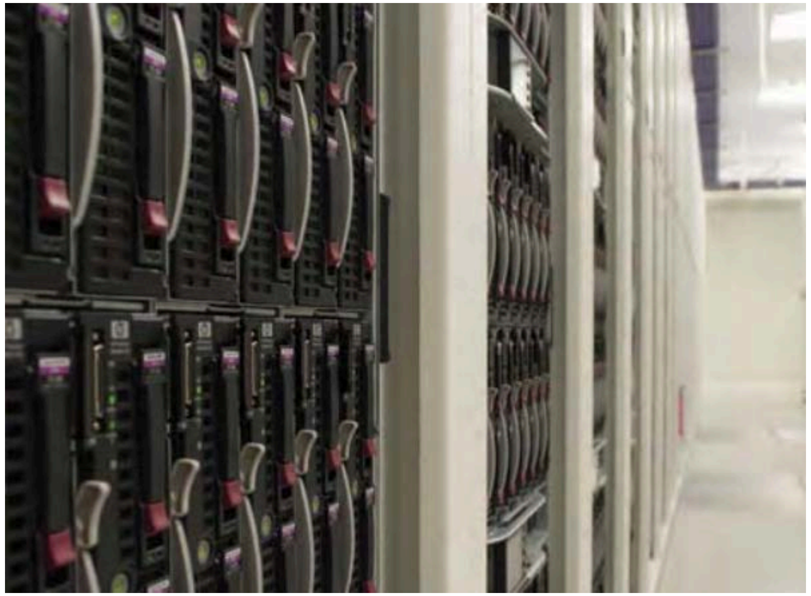
Introduction to Big Data

Bin Chen
Sep 27 2019

Session 1: Big Data in translational bioinformatics

Session 2: Big Data in R

EMBL-EBI

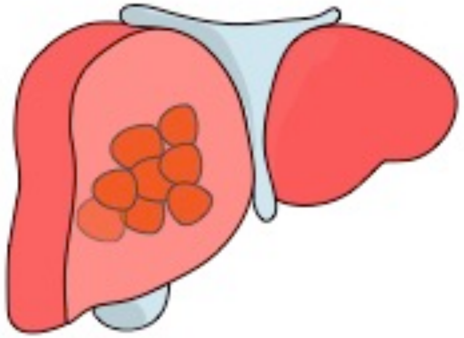


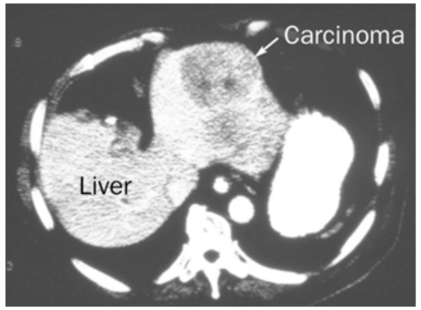
25 petabytes	2014
60 petabytes	2015

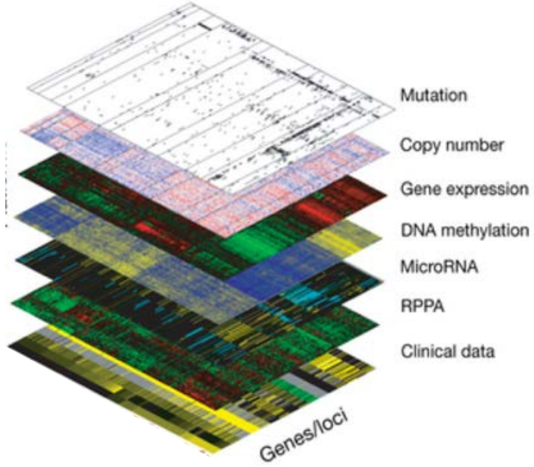
http://bit.ly/1OyTuqZ
http://bit.ly/2o3QJdy

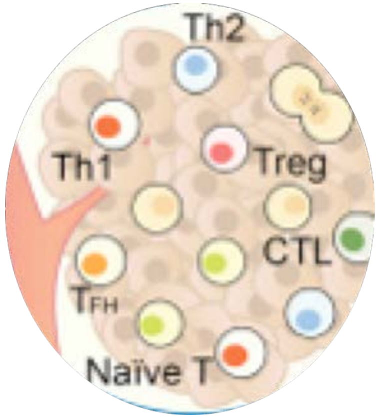
60 petabytes = × 30,000

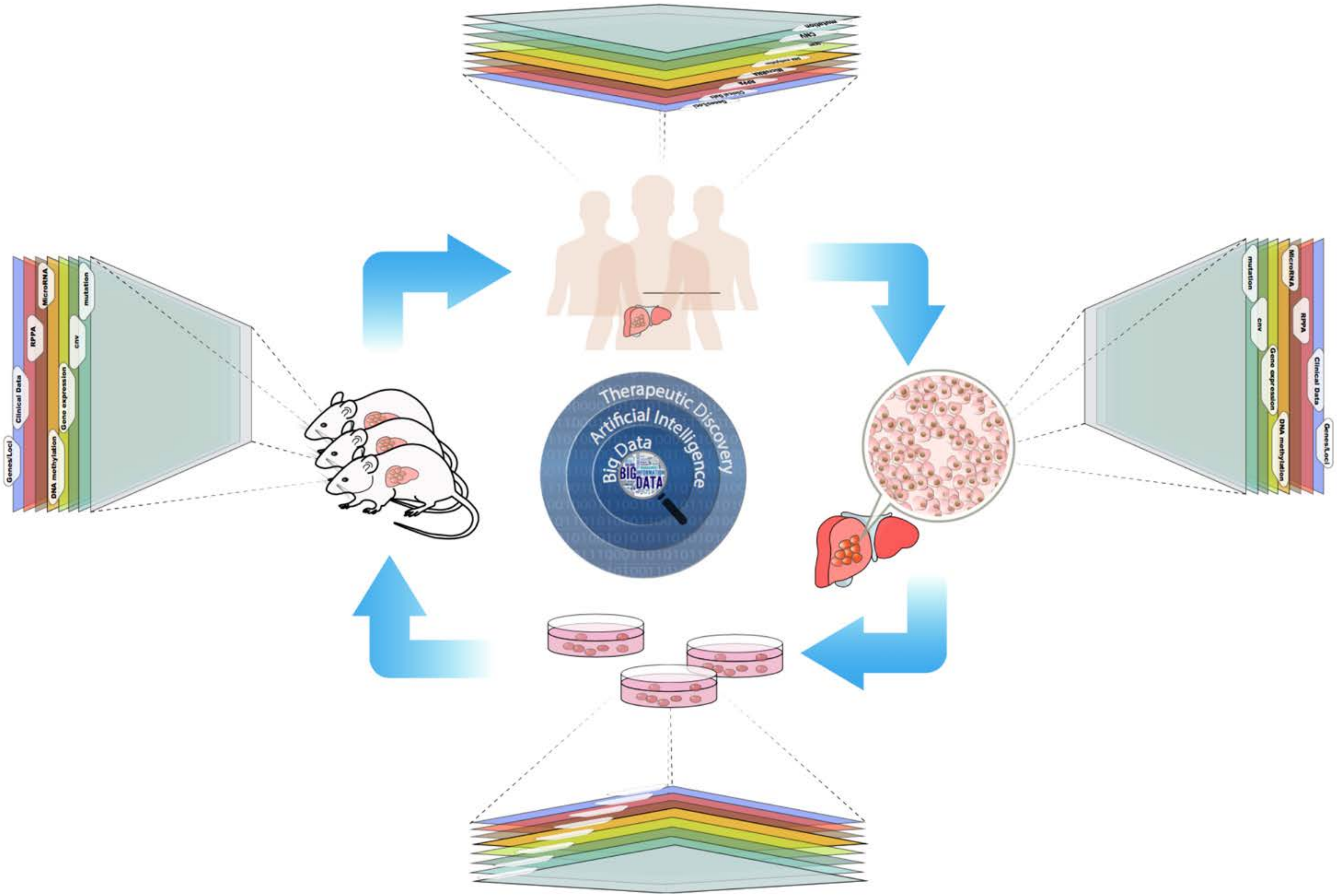
120 000 datasets!

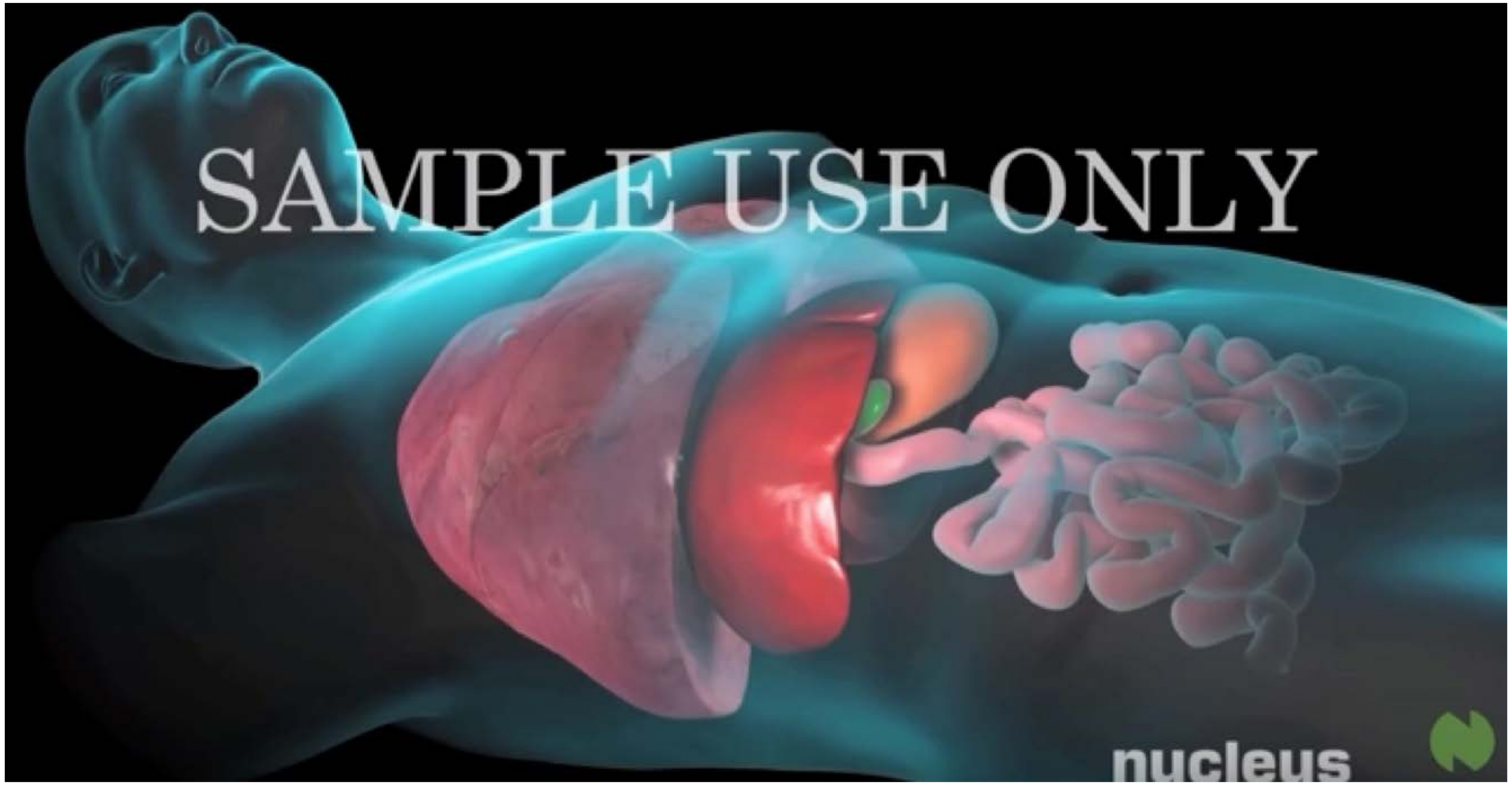



Carcinoma
Liver


Mutation
Copy number
Gene expression
DNA methylation
MicroRNA
RPPA
Clinical data
Genes/loci


Th2
Th1
Treg
CTL
TFH
Naïve T

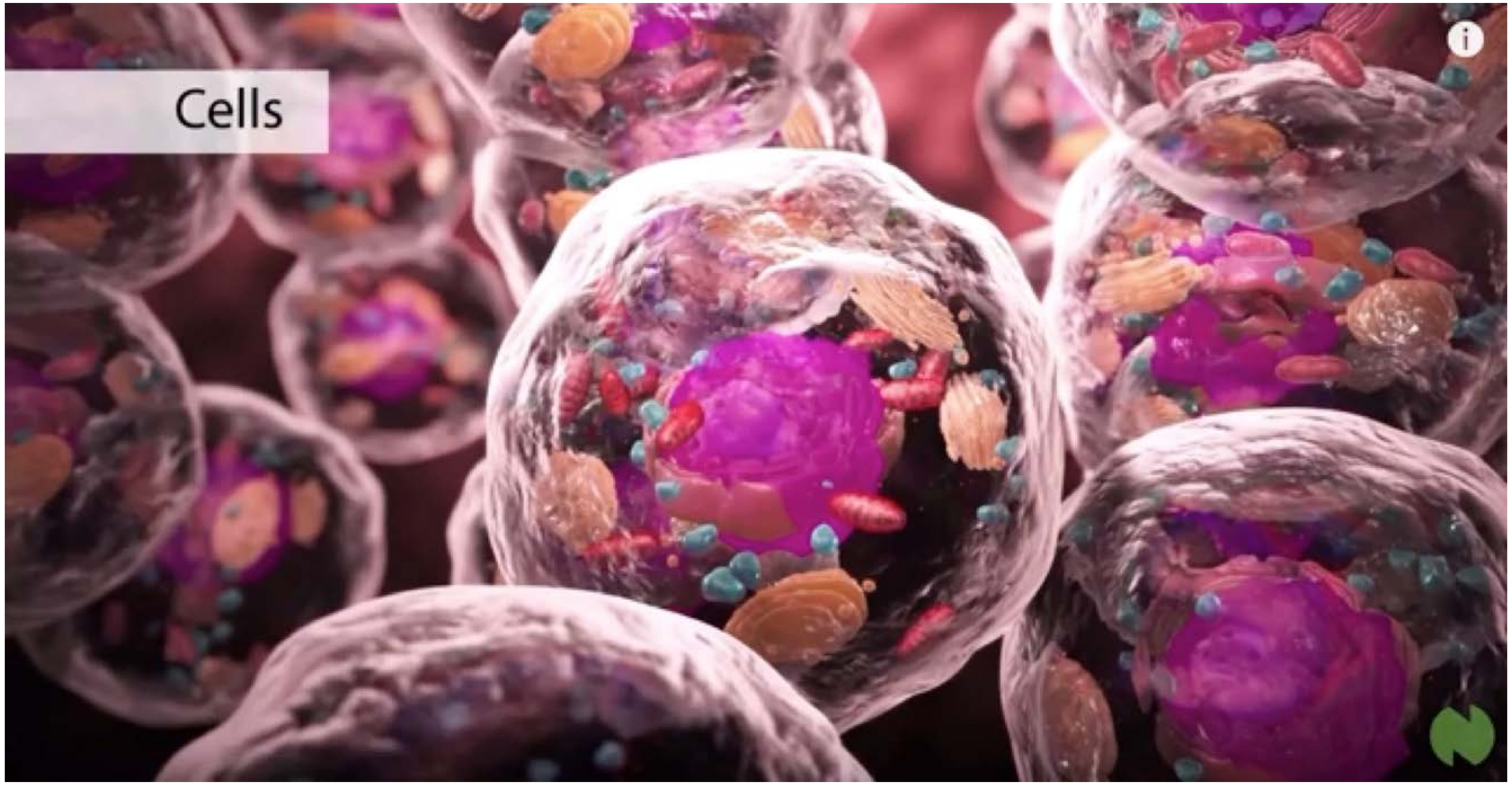

Therapeutic Discovery
Artificial Intelligence
Big Data
BIG DATA
Genes/Loci
Clinical Data
RPPA
MicroRNA
DNA methylation
Gene expression
cnv
mutation
mutation
cnv
Gene expression
DNA methylation
MicroRNA
RPPA
Clinical Data
Genes/Loci
mutation
CNV



https://www.youtube.com/watch?v=mnHPx5XEvfQ

https://www.youtube.com/watch?v=CMRKKl9XSDU

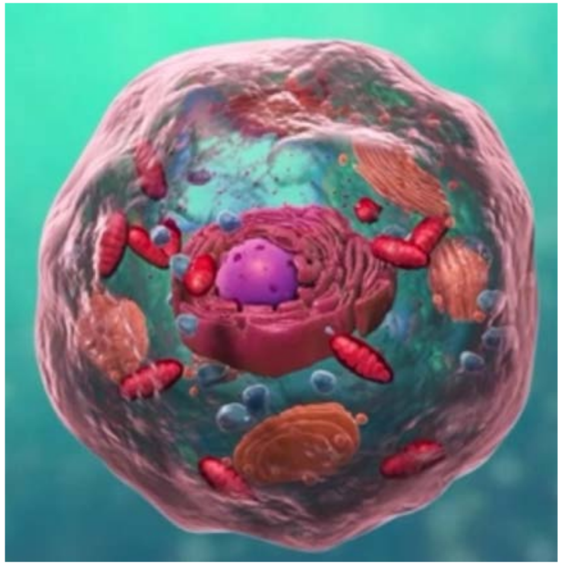
https://www.youtube.com/watch?v=CK78IXTRH0s

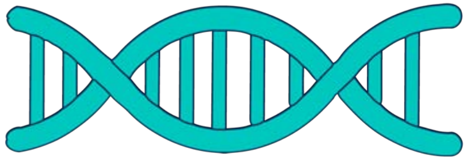


https://www.youtube.com/watch?v=URUJD5NEXC8

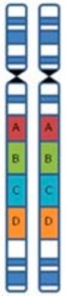
https://www.youtube.com/watch?v=gG7uCskUOrA

- Ignore spatial info
- Ignore dynamic
- Ignore cell-cell variation





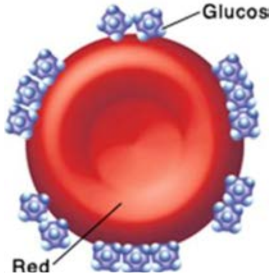
DNA



Chromosome
(copy number)

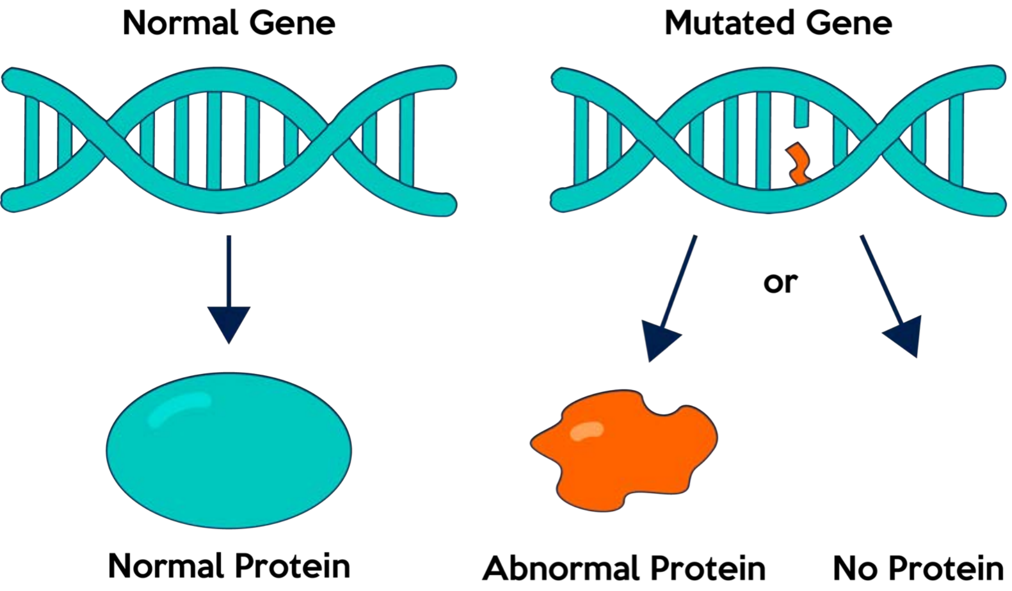
mRNA (gene)

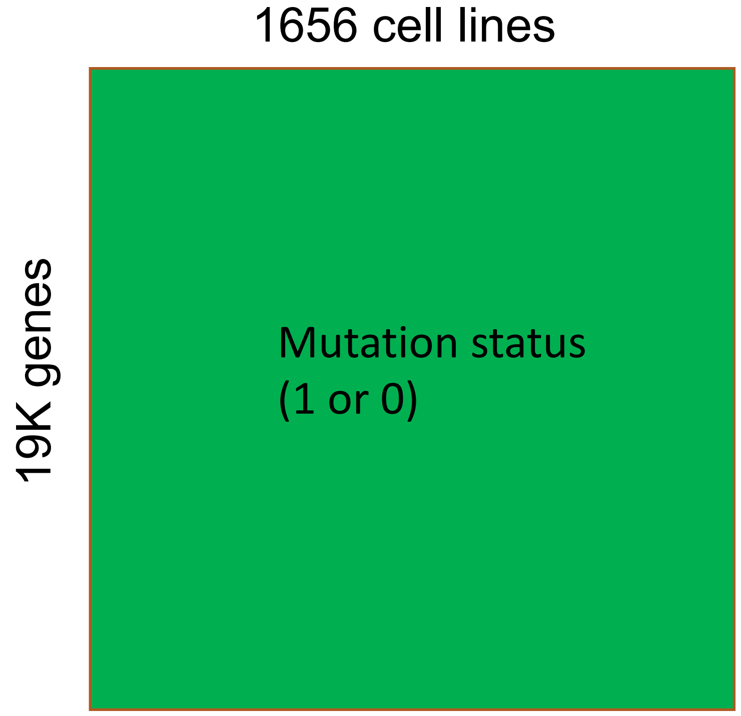
Protein



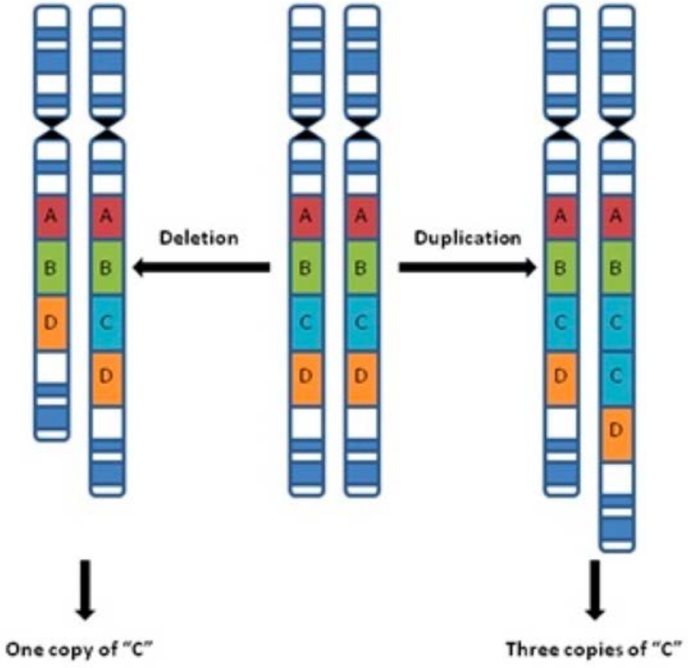
metabolite

……


Normal Gene
Mutated Gene
or
Normal Protein
Abnormal Protein
No Protein


1656 cell lines
19K genes
Mutation status
(1 or 0)

Genomics



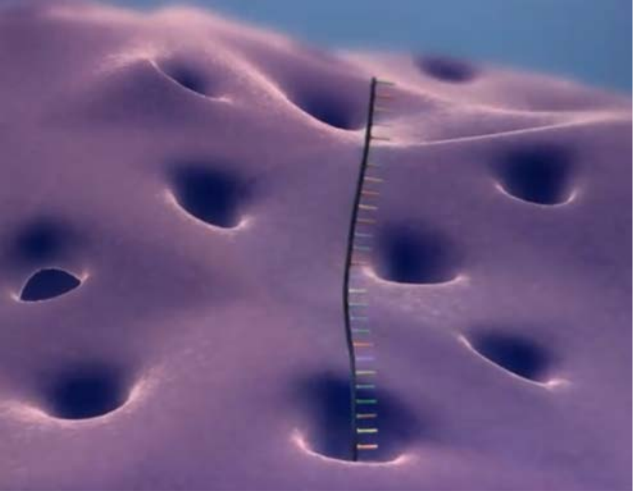
1657 cell lines

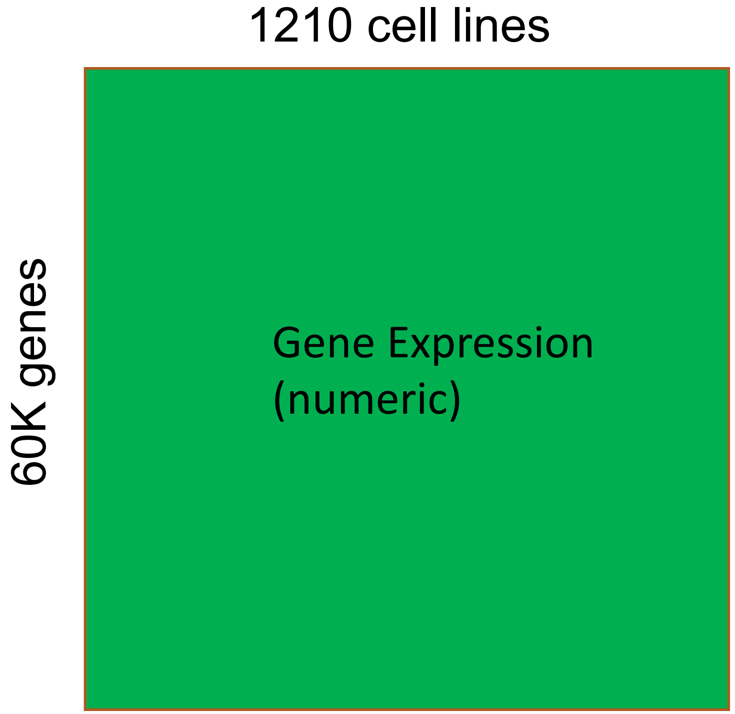
28K genes

Copy number
(-2, -1, 0, 1, 2)

Genomics

EpiGenomics



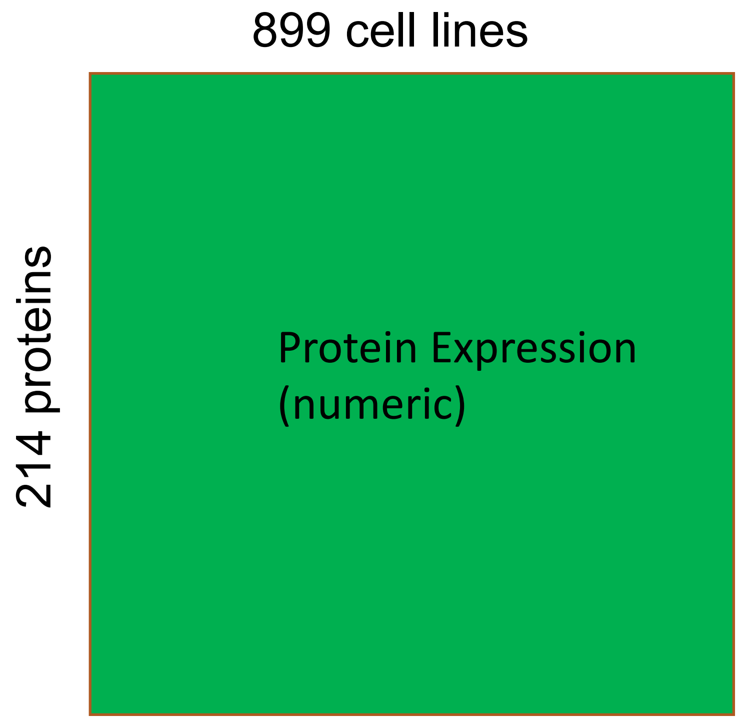


1210 cell lines

60K genes

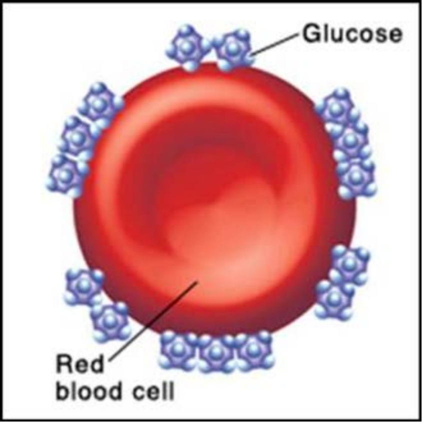
Gene Expression
(numeric)

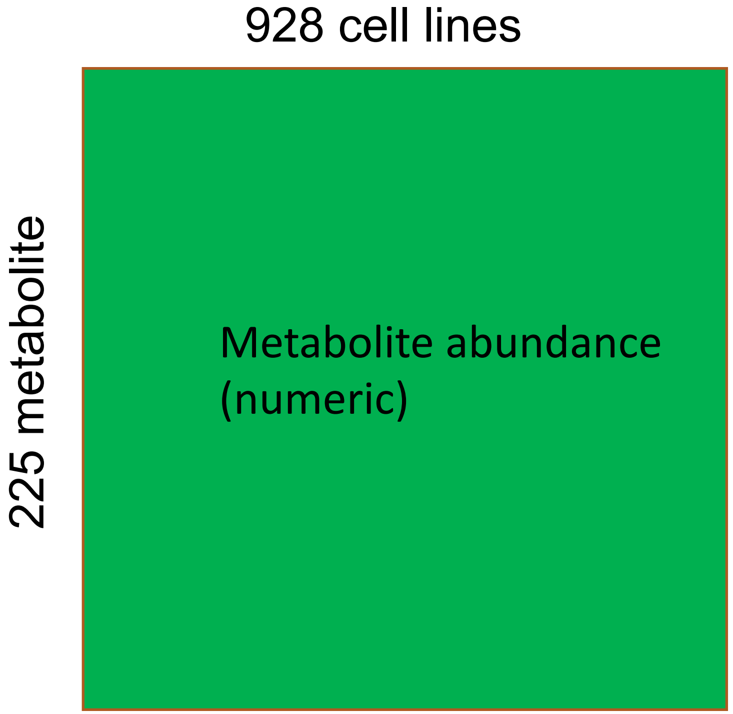
Functional Genomics


899 cell lines

214 proteins

Protein Expression
(numeric)



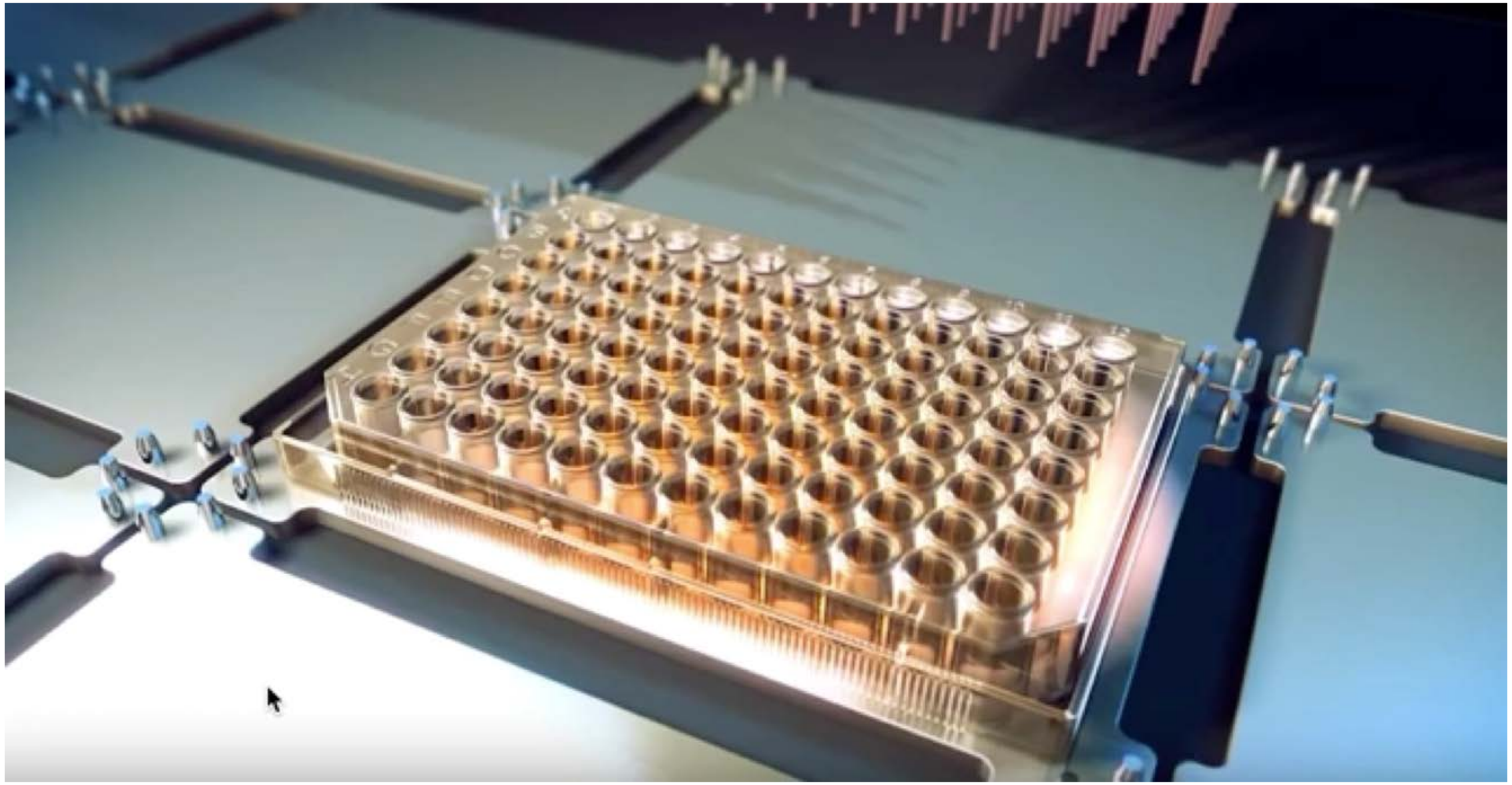


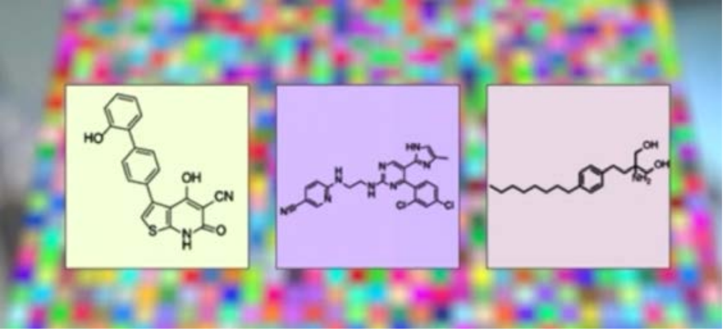
928 cell lines

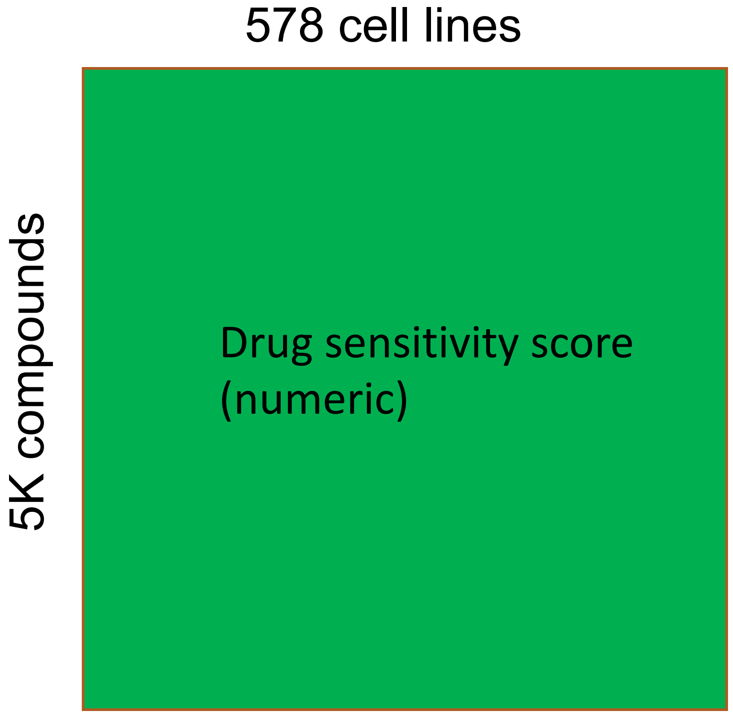
225 metabolite

Metabolite abundance (numeric)

Imaging





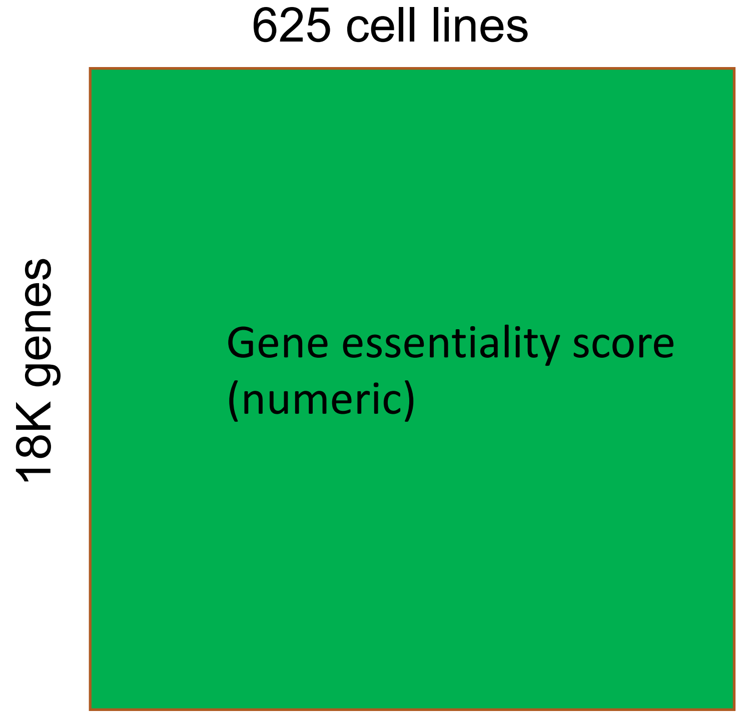


578 cell lines

5K compounds

Drug sensitivity score (numeric)

Response data

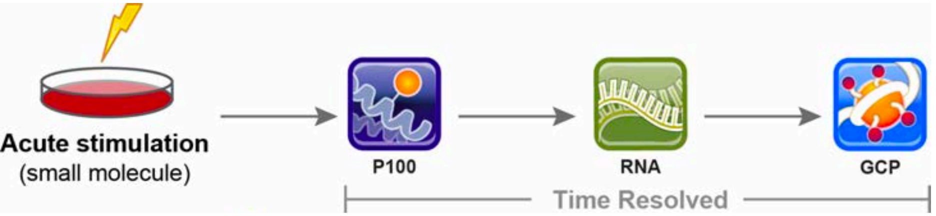


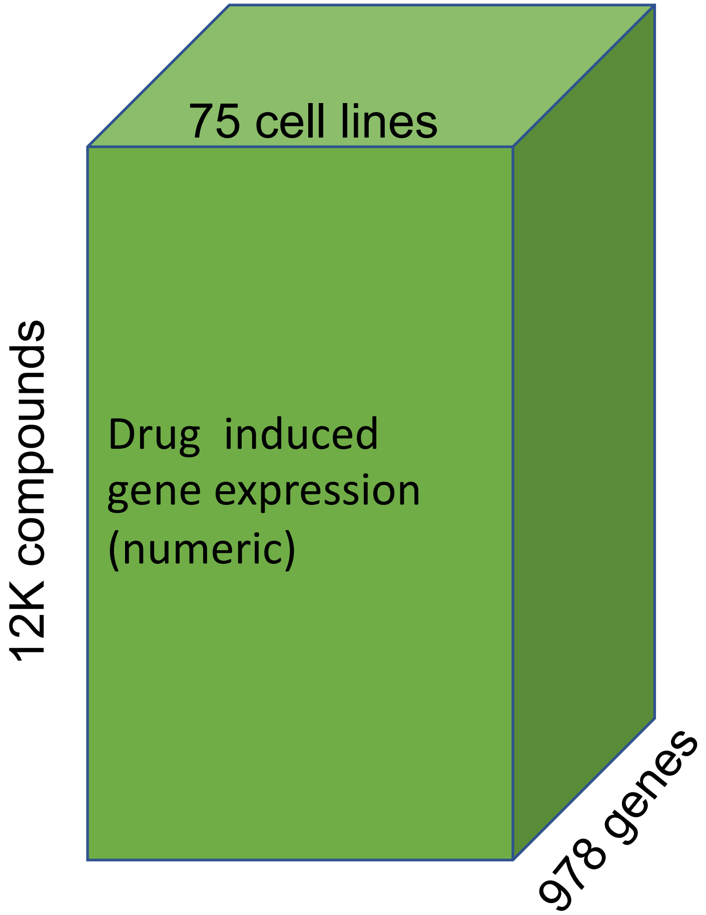
625 cell lines

18K genes

Gene essentiality score (numeric)

Response data





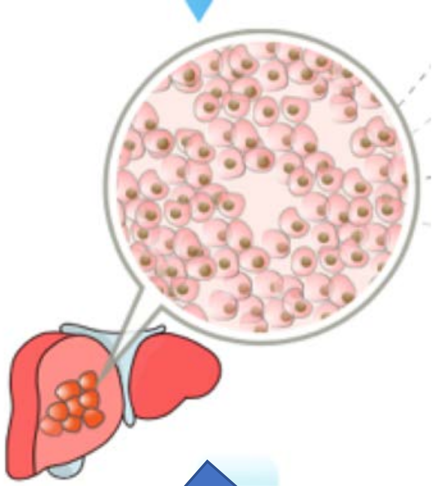
75 cell lines

12K compounds

Drug induced gene expression (numeric)

978 genes

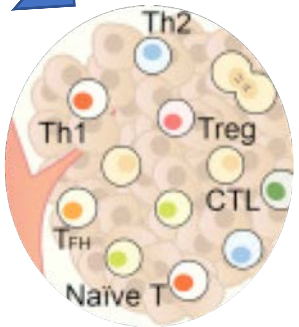
Response data

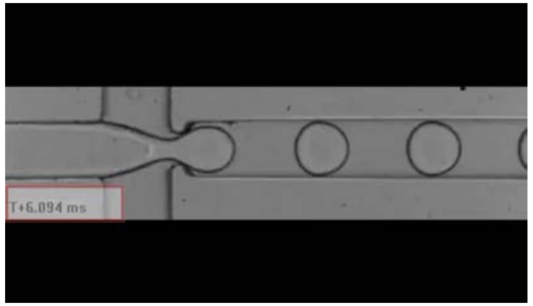


20K patient tissues

60K genes

Gene Expression
(numeric)



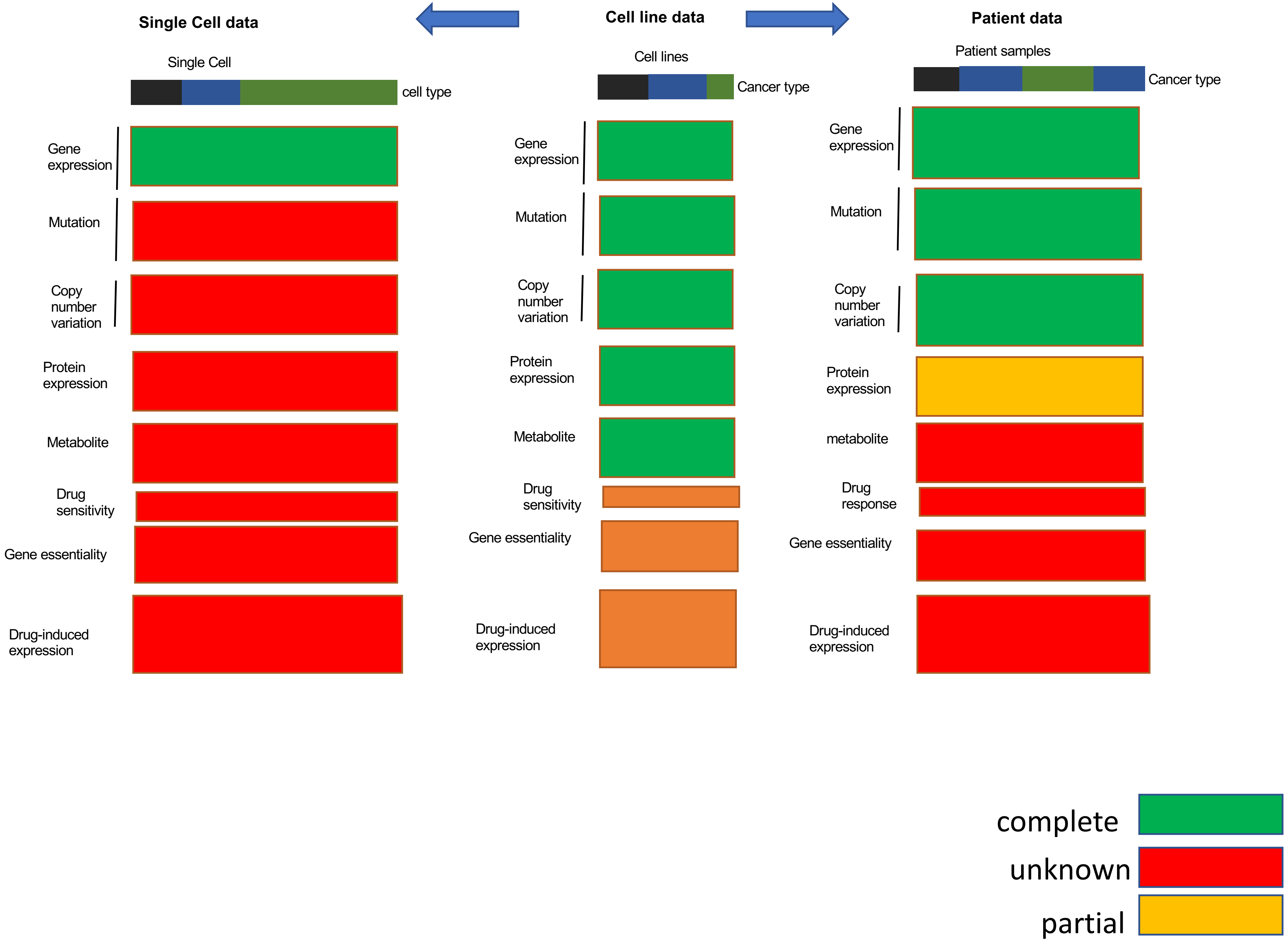


Millions of single cells

60K
genes

Gene Expression
(numeric)

Single cells



Single Cell data		Cell line data		Patient data	
Single Cell	cell type	Cell lines	Cancer type	Patient samples	Cancer type
Gene expression	🟩	Gene expression	🟩	Gene expression	🟩
Mutation	🟥	Mutation	🟩	Mutation	🟩
Copy number variation	🟥	Copy number variation	🟩	Copy number variation	🟩
Protein expression	🟥	Protein expression	🟩	Protein expression	🟨
Metabolite	🟥	Metabolite	🟩	metabolite	🟥
Drug sensitivity	🟥	Drug sensitivity	🟧	Drug response	🟥
Gene essentiality	🟥	Gene essentiality	🟧	Gene essentiality	🟥
Drug-induced expression	🟥	Drug-induced expression	🟧	Drug-induced expression	🟥

complete 🟩

unknown 🟥

partial 🟨

Network (knowledge graph)

Unstructured data

Session 1: Big Data in translational bioinformatics

Session 2: Big Data in R

R

- Install R
- Install RStudio
- Install R package
- Install Bioconductor package

Data Type

- Numeric
- Character
- Logical
- Factor

Data Type

- Vector
- Data.frame
- Matrics
- Arrays
- List
- RData

Advanced data Type

- Image

- Unstructured text

- Class

Big files

- Data.table
- H5

Subsetting

- A = a[1:3]

Basic Operators and Calculations

- AND
- OR
- Not

- A = 1 + 10

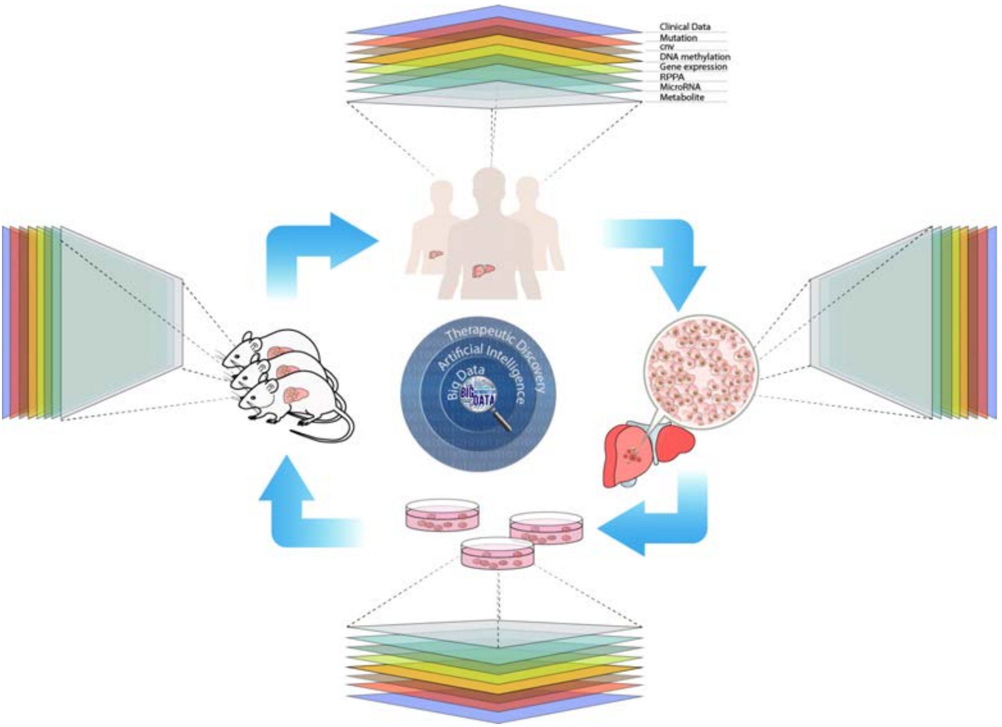
Data input and output

Data summary

Data properties

- Matrix (small-samples, big features)
- Sparse
- Noisy (High-throughput)
- Batch effect
- Processed data
- Context dependent
- Time series (some matrices)

Connecting open data points to facilitate translational research



Data (grey: to be processed)

Cell lines

- Gene expression (1210 cell lines X 19k genes)
- Mutation (1656 cell lines X 19K genes)
- Copy number (1657 cell lines X 28K genes)
- Protein expression (899 cell lines X 214 proteins)
- Metabolite abundance (928 cell lines X 225 metabolites)

- CRISPR (625 cell lines X 18K genes)
- RNAi (712 cell lines X 17K genes)
- Drug sensitivity (578 cell lines X 5K cmpds)
- Drug expression profile (75 cell lines X 12K cmpds X 978 genes)

Animals

- Ad-hoc

Patient tissues

- Bulk disease (18K samples X 60K transcripts)
- Bulk normal (7K samples X 60K transcripts)
- Single cell

Patient EMR (Spectrum Health)

- Medication, lab test, bill, outcome, disease condition

Tool/Model

- Query
- Correlation analysis
- Clustering
- Predictive models

Workshop structure

- Data manipulation and visualization
- Basic statistical analysis
- Machine learning
- RNA-Seq
- Single cell RNA-Seq
- Cheminformatics/pharmacogenomics
- Structure-based drug design
- R markdown/R package/Shiny

Lab session