PyCon PL
2019-09-13

Biohack

Jacek Śmietański
Kraków

Epam Systems: data science for industry, Jagiellonian University: bioinformatics, lectures, classes

Bioinformatics

Image source: https://blog.f1000.com/wp-content/uploads/2017/01/Header-Image-crop-1024x559.jpg

data

challenges

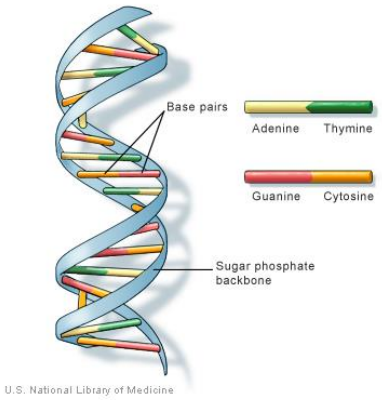
applications

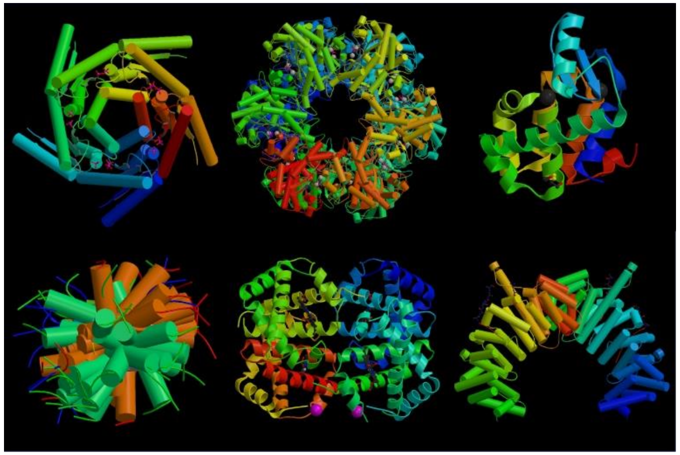
Raw data

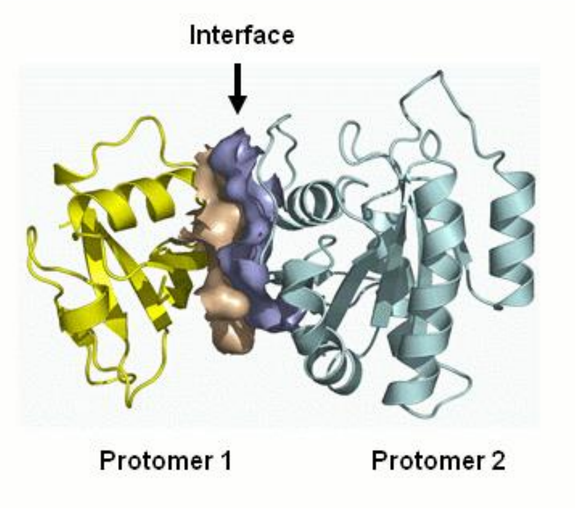
DNA sequence

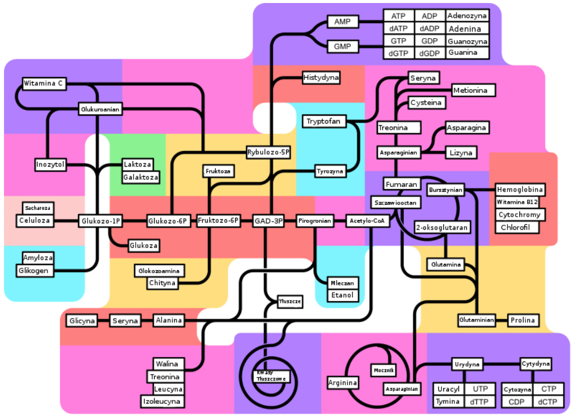
```
> AY169899.1 Morelia viridis strain ABTC66386 cytochrome b gene,
partial cds; mitochondrial gene for mitochondrial product

TTCGGCTCAATATTATTAACATGTTTAGCCCTACAAGTACTACCGGC
TTCTTCTTAGCCGTCCACTACACAGCAAACATCAACCTAGCATTCT
CATCCATTATCCATATCACTCGAGATGTCCCATACGGCTGAATAATA
CAAAACCTACACGCCATCGGAGCATCCATATTCTTCATTTGCATTTA
CATCCACATCGCACGAGGACTATACTACGGATCCTACCTCAACAAA
GAGACTTGAATATCCGGTATCACCCTACTCATCACATTAATAGCAAC
CGCCTTCTTTG
```

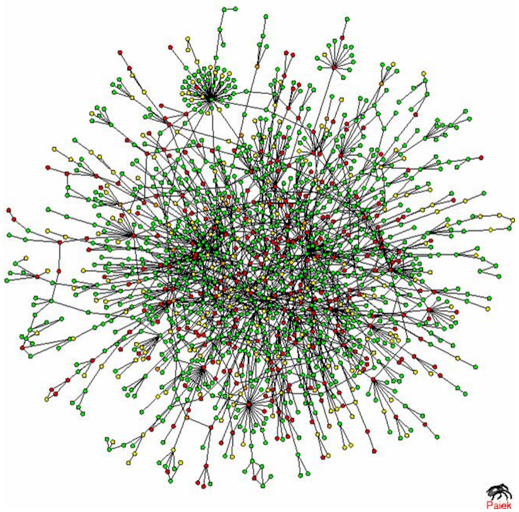
Data visualisation

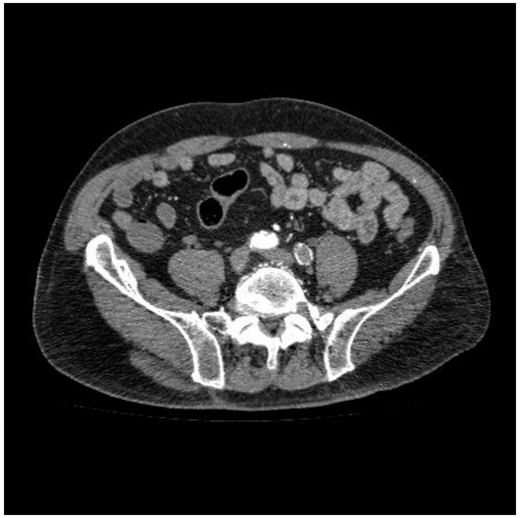




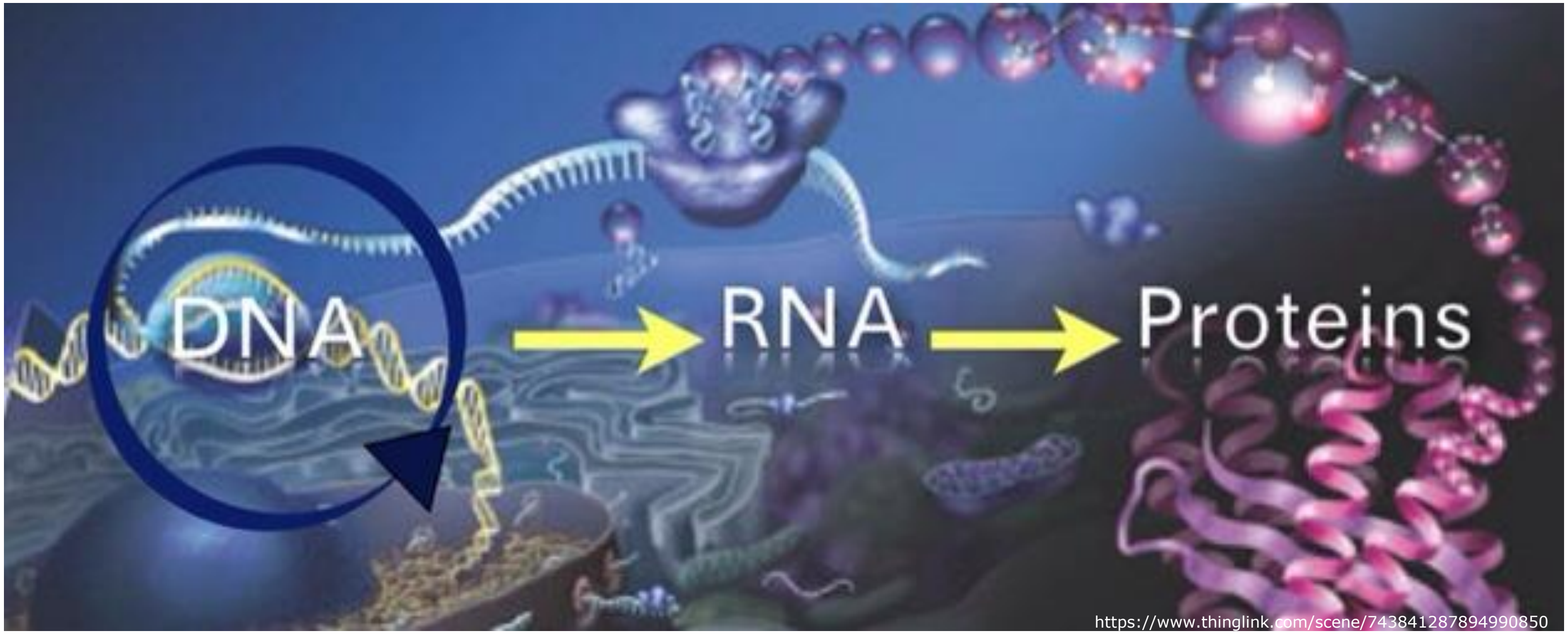








Information flow in cell



https://www.thinglink.com/scene/743841287894990850

Sequence → Structure → Function

Data availability

- Over 1500 bioinformatics databases listed by Nucleic Acid Research journal
- Main databases connected in network (meta-database)

NCBI Entrez:

https://www.ncbi.nlm.nih.gov/search

GenBank

Swiss Prot

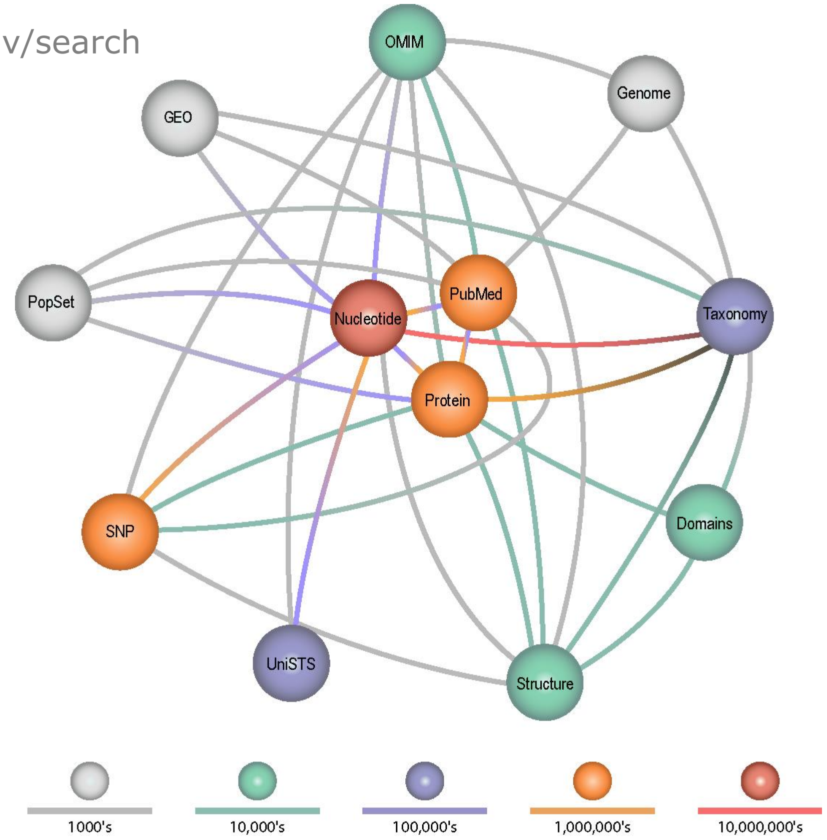
Protein Data Bank

PubMed

OMIM

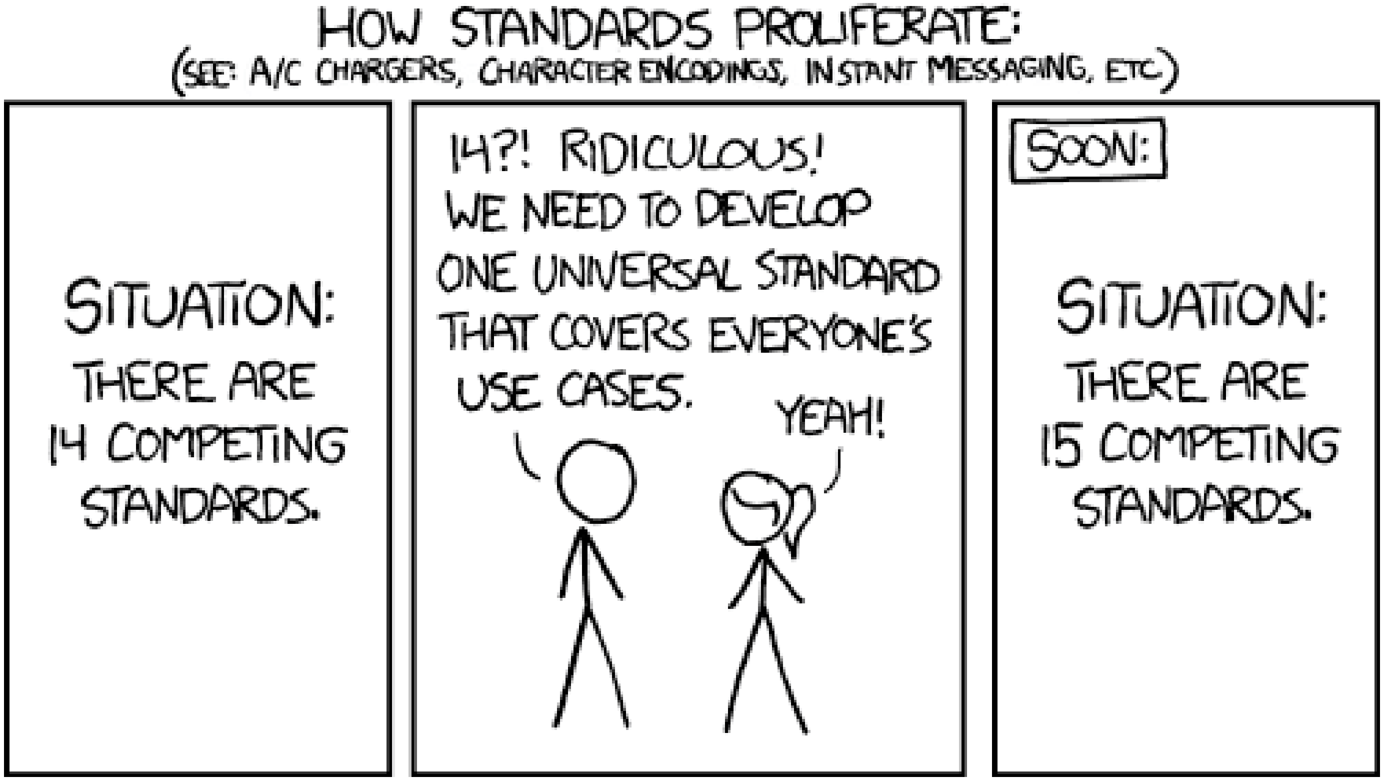
SNP

(...)



Data issues

- Corectness and reliability
- Inconsistency of data formats
- Outdated databases


HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

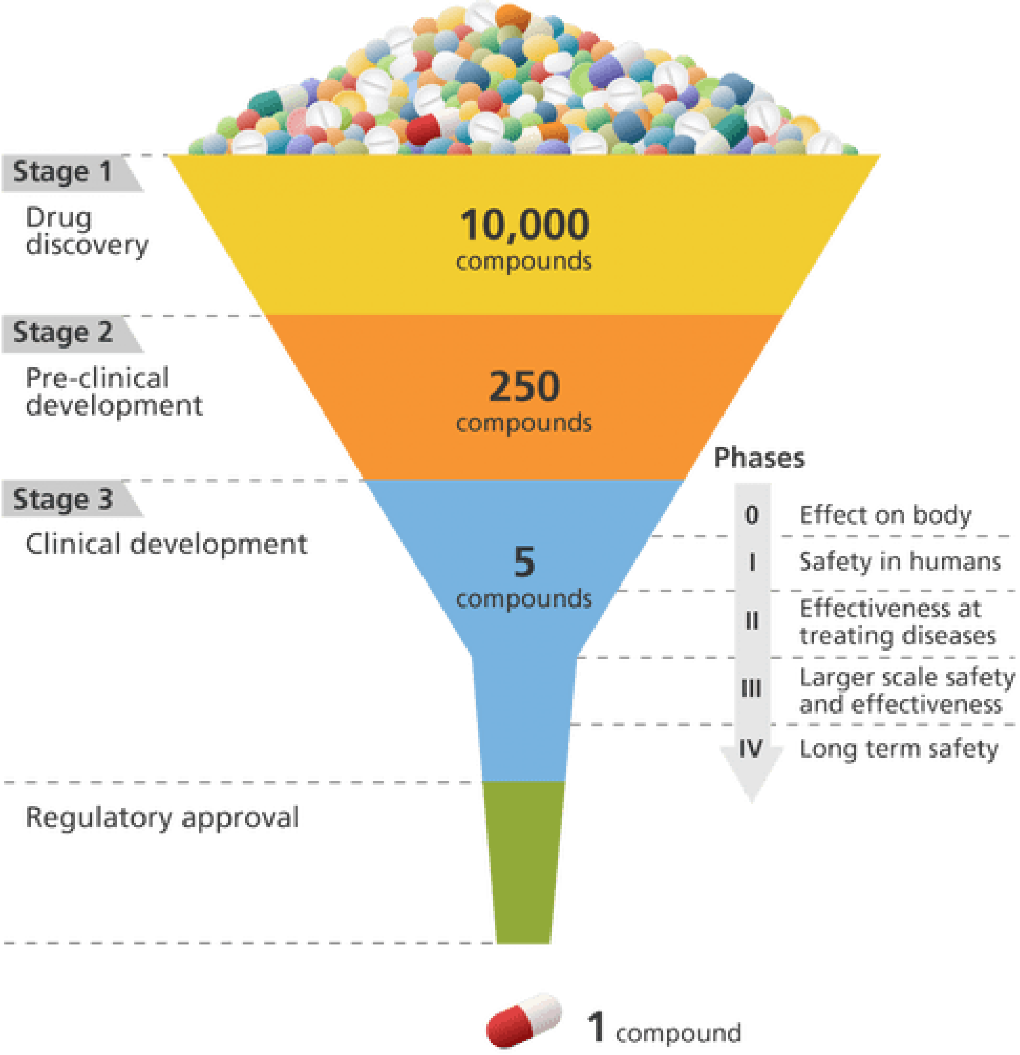
Challenge: next generation sequencing (NGS)

https://singularityhub.com/wp-content/uploads/2019/01/3d-illustration-dna-sequencing_shutterstock_430949605-1068x601.jpg

- sequence assembly
- data storage
- variations

Application example: drug design

- target identification
- structure determination
- lead search
- pharmacokinetics
- response prediction (ADME & toxicology)



https://i1.wp.com/theskepticalchemist.com/wp-content/uploads/2018/09/drug_design_yourgenome.png

Biohack

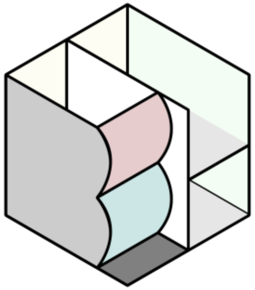
Boinformatics hackathons

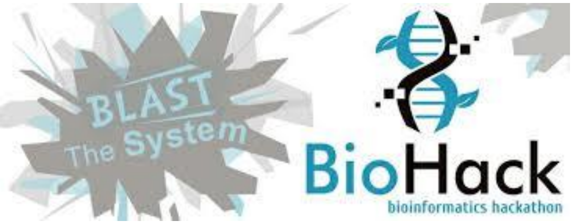
Hosted worldwide by several organisations.

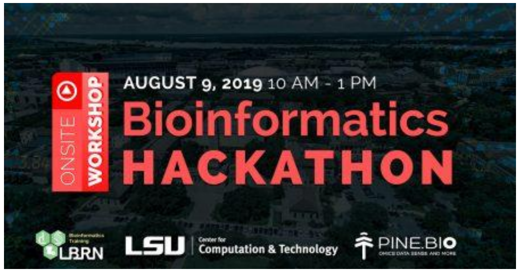
Eg.: http://biohackathon.org (annual event since 2008)

https://biohackathons.github.io

https://www.biohackathon-europe.org






ONSITE WORKSHOP
AUGUST 9, 2019 10 AM - 1 PM
Bioinformatics
HACKATHON
LBRN
LSU Center for Computation & Technology
PINE.BIO

JGI JOINT GENOME INSTITUTE
NVIDIA
OAK RIDGE National Laboratory
LEADERSHIP COMPUTING FACILITY
Join Our
Bioinformatics
Hackathon
May 6-10

Biohackathons in Poland

Biohack (past editions: 2018, 2019)
http://biohack.com.pl/

Bioninja (upcoming, 4-6.10.2019, registration still active ☺)
http://www.bioninjachallenge.eu

- onsite (Łódź)
- free participation
- 4-5 people interdisciplinaty teams
  (if you are alone, orga team creates team for you)
- task are selected randomly
- 24 h for coding
- integration part, board games, pizza, etc...

Tasks 2018

1. The analysis of bacteriophage genomes considering their lytic/lysogenic potential
2. Implementation of balance tree algorithm for abundance difference analysis in metagenomic research
3. Analysis of GWAS data using machine learning algorithms
4. Development of algorithm for translation of common names of disease entities into ICD-10 units (NLP approach preferred)
5. Development of local tool for annotation of bacteriophage genomes
6. Development of method for ordered storage of NGS-obtained data

http://biohack.com.pl/biohack-exemplary-tasks/

Tasks 2019

1. Immunogenicity prediction of viral neoantigens in murine hosts.
2. Feature space engineering for optimal classification of virome dataset.
3. Predicting response to cancer immunotherapy

http://biohack.com.pl/biohack-ii-exemplary-tasks/

Our challenge

Predicting response to cancer immunotherapy

Treatment method:
the patient's immune system is stimulated to destroy tumor cells.

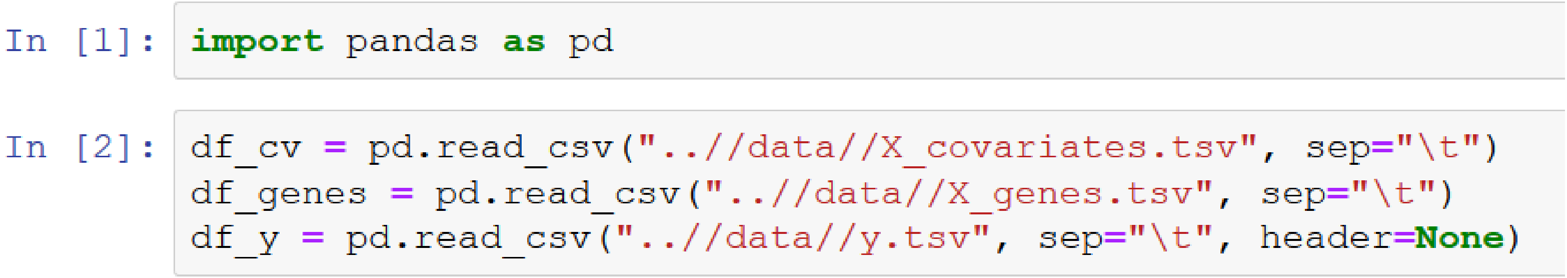
Only 20-40% of patients respond to immunotherapy.

Making prediction which patients will be responders is one of the challenges in cancer research.

Provided data

A dataset related to patients who underwent cancer immunotherapy:

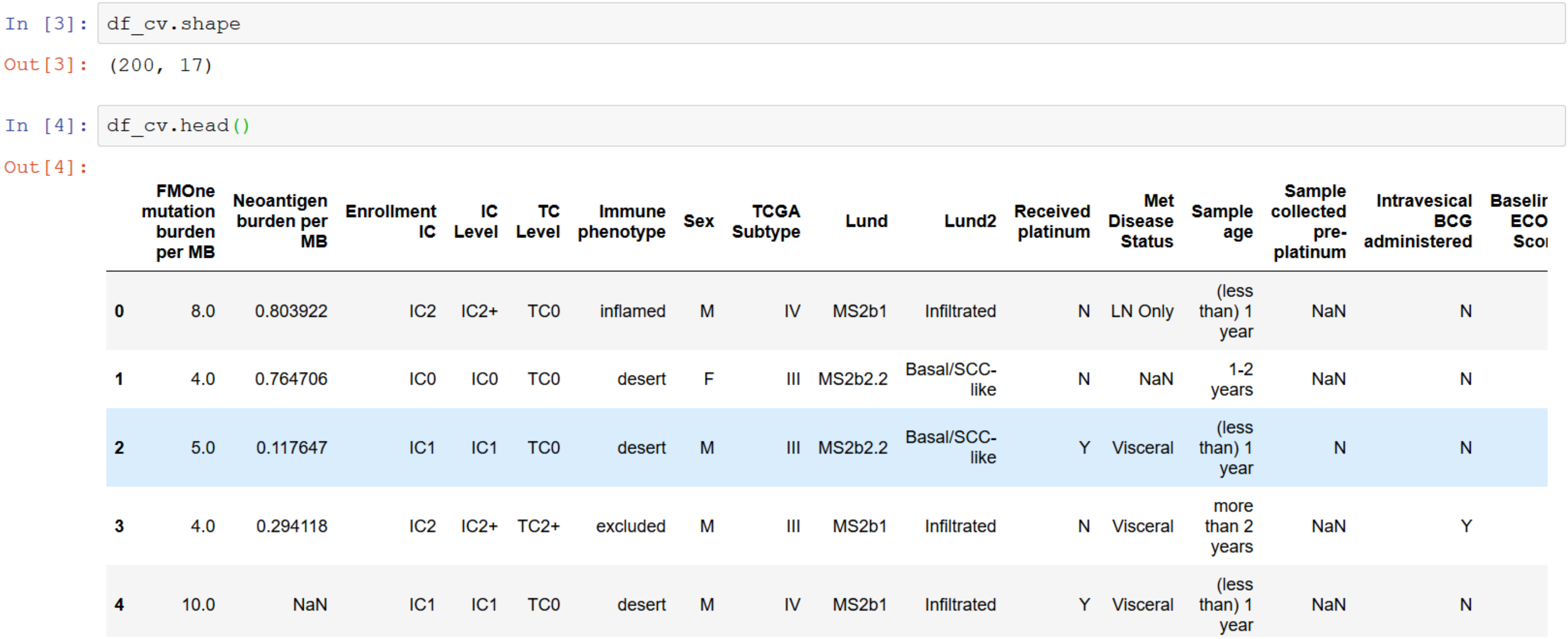
- patients’ clinical information
- RNA sequencing data
- response status



```
In [1]: import pandas as pd

In [2]: df_cv = pd.read_csv("..//data//X_covariates.tsv", sep="\t")
        df_genes = pd.read_csv("..//data//X_genes.tsv", sep="\t")
        df_y = pd.read_csv("..//data//y.tsv", sep="\t", header=None)
```

Clinical information

17 parameters regarding patient interview, diagnosis results and treatment history.



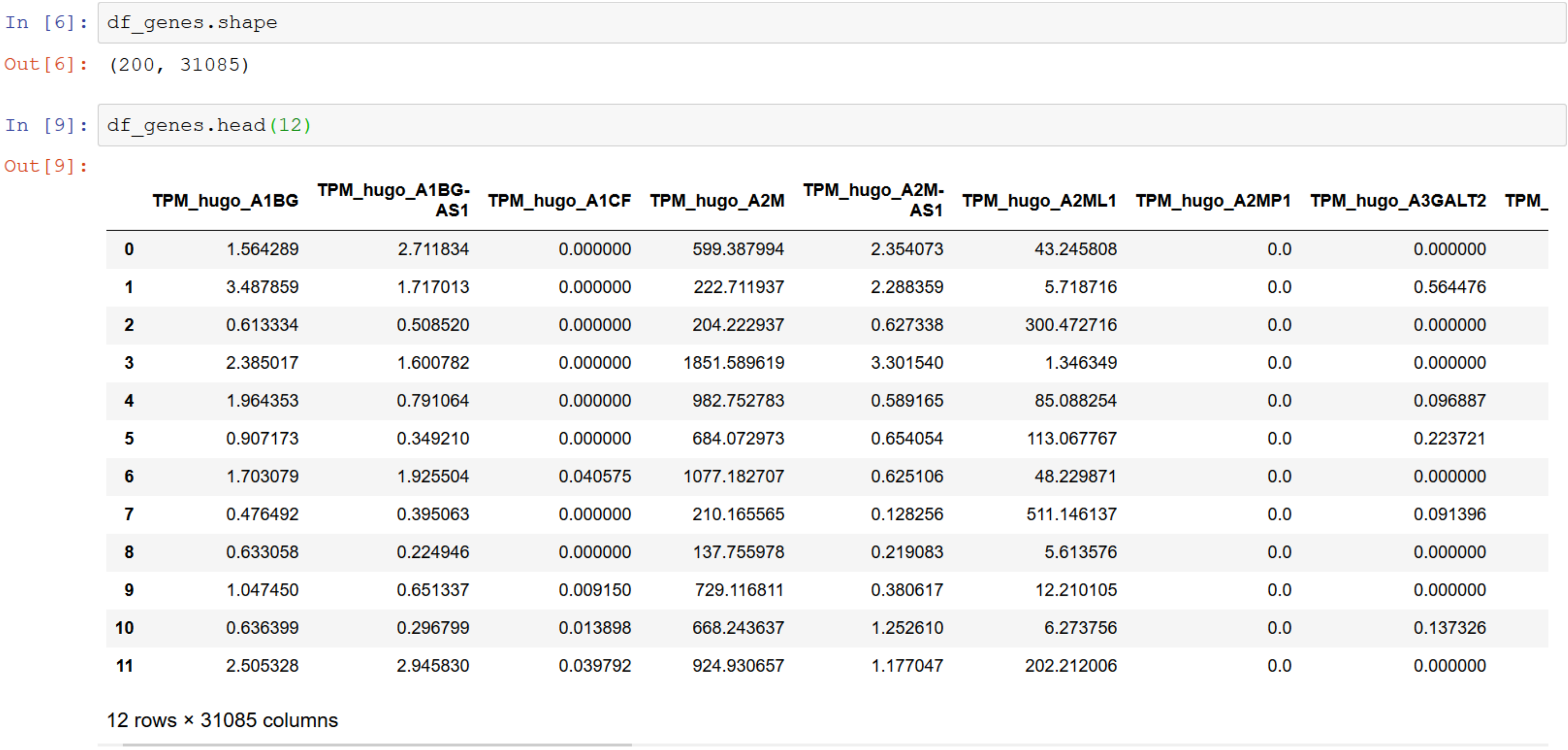
```
In [3]: df_cv.shape
Out[3]: (200, 17)
```

```
In [4]: df_cv.head()
Out[4]:
```

	FMOne mutation burden per MB	Neoantigen burden per MB	Enrollment IC	IC Level	TC Level	Immune phenotype	Sex	TCGA Subtype	Lund	Lund2	Received platinum	Met Disease Status	Sample age	Sample collected pre-platinum	Intravesical BCG administered	Baselir ECO Sco
0	8.0	0.803922	IC2	IC2+	TC0	inflamed	M	IV	MS2b1	Infiltrated	N	LN Only	(less than) 1 year	NaN	N	
1	4.0	0.764706	IC0	IC0	TC0	desert	F	III	MS2b2.2	Basal/SCC-like	N	NaN	1-2 years	NaN	N	
2	5.0	0.117647	IC1	IC1	TC0	desert	M	III	MS2b2.2	Basal/SCC-like	Y	Visceral	(less than) 1 year	N	N	
3	4.0	0.294118	IC2	IC2+	TC2+	excluded	M	III	MS2b1	Infiltrated	N	Visceral	more than 2 years	NaN	Y	
4	10.0	NaN	IC1	IC1	TC0	desert	M	IV	MS2b1	Infiltrated	Y	Visceral	(less than) 1 year	NaN	N	

RNA sequencing data

Expression level for 31085 genes.



```
In [6]: df_genes.shape
Out[6]: (200, 31085)
```

```
In [9]: df_genes.head(12)
Out[9]:
```

	TPM_hugo_A1BG	TPM_hugo_A1BG-AS1	TPM_hugo_A1CF	TPM_hugo_A2M	TPM_hugo_A2M-AS1	TPM_hugo_A2ML1	TPM_hugo_A2MP1	TPM_hugo_A3GALT2	TPM_
0	1.564289	2.711834	0.000000	599.387994	2.354073	43.245808	0.0	0.000000	
1	3.487859	1.717013	0.000000	222.711937	2.288359	5.718716	0.0	0.564476	
2	0.613334	0.508520	0.000000	204.222937	0.627338	300.472716	0.0	0.000000	
3	2.385017	1.600782	0.000000	1851.589619	3.301540	1.346349	0.0	0.000000	
4	1.964353	0.791064	0.000000	982.752783	0.589165	85.088254	0.0	0.096887	
5	0.907173	0.349210	0.000000	684.072973	0.654054	113.067767	0.0	0.223721	
6	1.703079	1.925504	0.040575	1077.182707	0.625106	48.229871	0.0	0.000000	
7	0.476492	0.395063	0.000000	210.165565	0.128256	511.146137	0.0	0.091396	
8	0.633058	0.224946	0.000000	137.755978	0.219083	5.613576	0.0	0.000000	
9	1.047450	0.651337	0.009150	729.116811	0.380617	12.210105	0.0	0.000000	
10	0.636399	0.296799	0.013898	668.243637	1.252610	6.273756	0.0	0.137326	
11	2.505328	2.945830	0.039792	924.930657	1.177047	202.212006	0.0	0.000000	

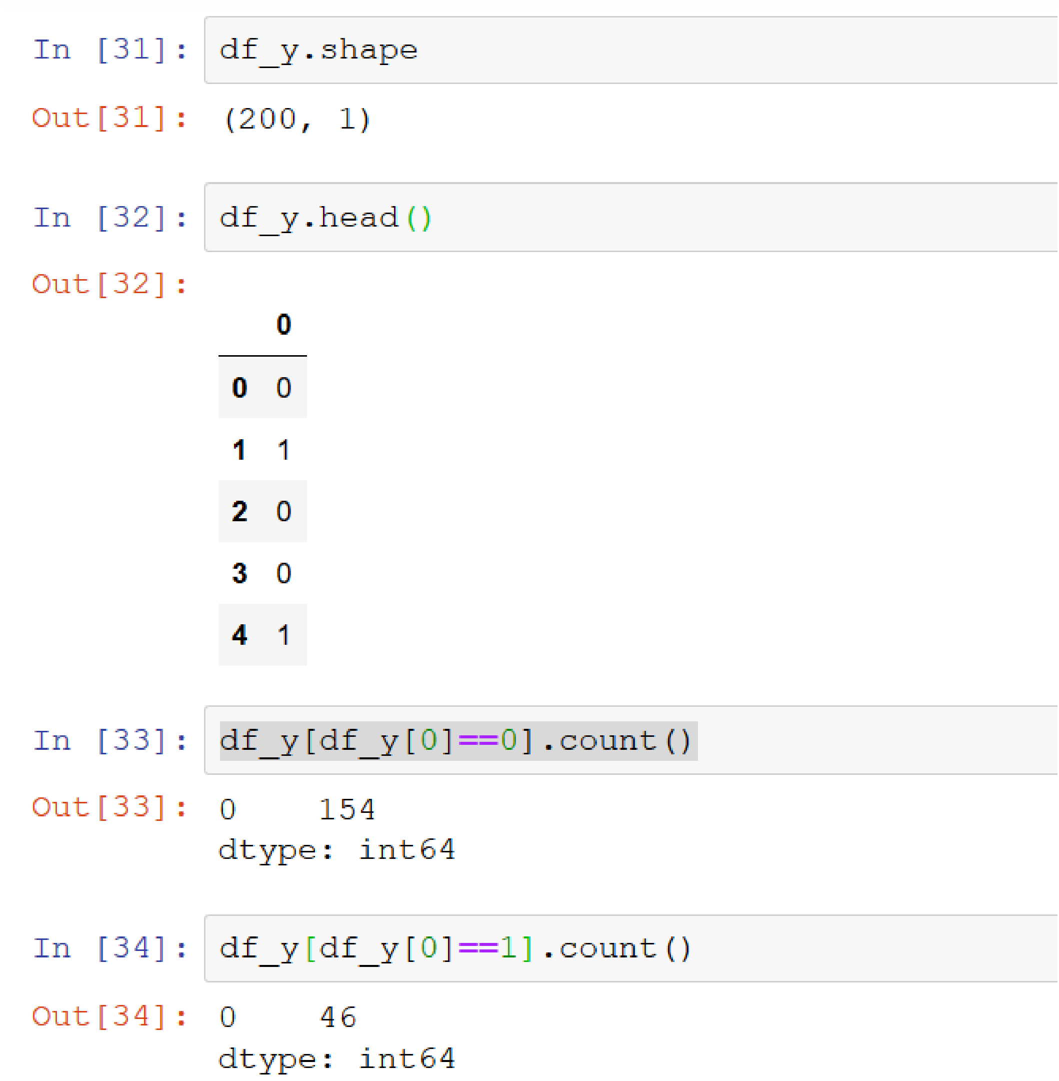
12 rows × 31085 columns

Response status

Binary.

„1” – response for treatment
was observed

„0” – no response for treatment

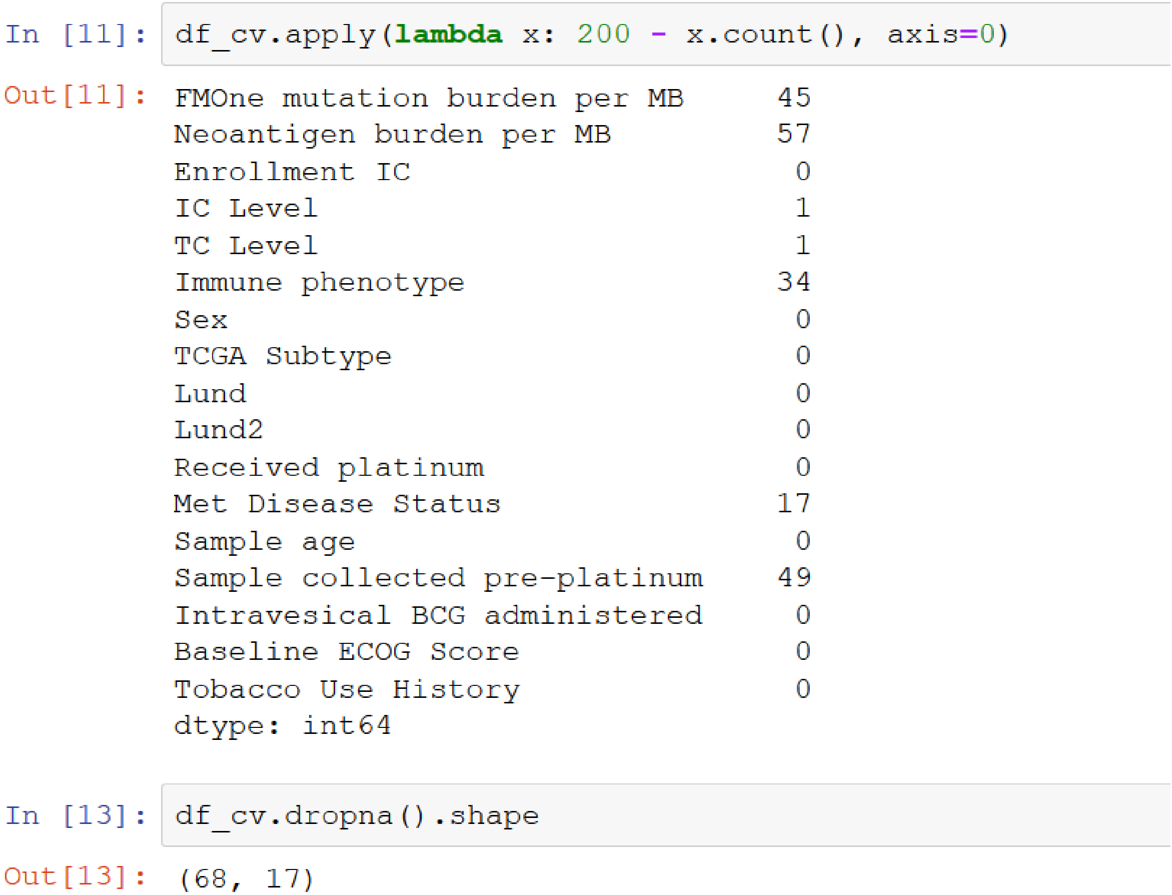


```
In [31]: df_y.shape
Out[31]: (200, 1)

In [32]: df_y.head()
Out[32]:
      0
   0  0
   1  1
   2  0
   3  0
   4  1

In [33]: df_y[df_y[0]==0].count()
Out[33]: 0    154
         dtype: int64

In [34]: df_y[df_y[0]==1].count()
Out[34]: 0    46
         dtype: int64
```

Our team

RSD
doktoranci
UMED
Boost
Biotech
24 H
OF REAL
FORMATICS
BioHack
CERTIFICATE
Marcin Lubocki
Jacek Śmietański
Jan Havranek

Dealing with incomplete data



```
In [11]: df_cv.apply(lambda x: 200 - x.count(), axis=0)

Out[11]: FMOne mutation burden per MB      45
         Neoantigen burden per MB         57
         Enrollment IC                     0
         IC Level                          1
         TC Level                          1
         Immune phenotype                 34
         Sex                               0
         TCGA Subtype                      0
         Lund                              0
         Lund2                             0
         Received platinum                 0
         Met Disease Status               17
         Sample age                        0
         Sample collected pre-platinum    49
         Intravesical BCG administered     0
         Baseline ECOG Score               0
         Tobacco Use History               0
         dtype: int64
```

```
In [13]: df_cv.dropna().shape

Out[13]: (68, 17)
```

Remove incomplete records?

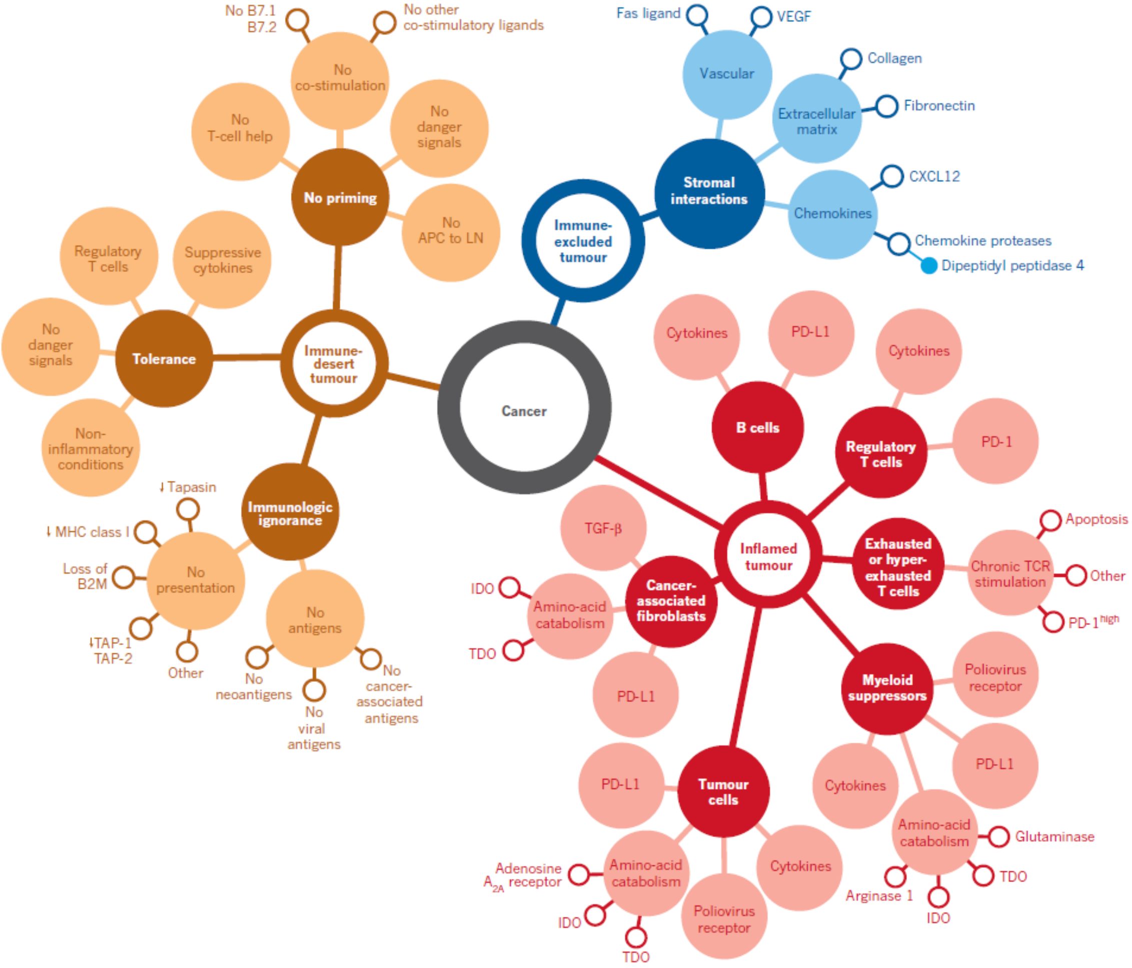
Fill with artificial value?

Try to estimate?

Features selection

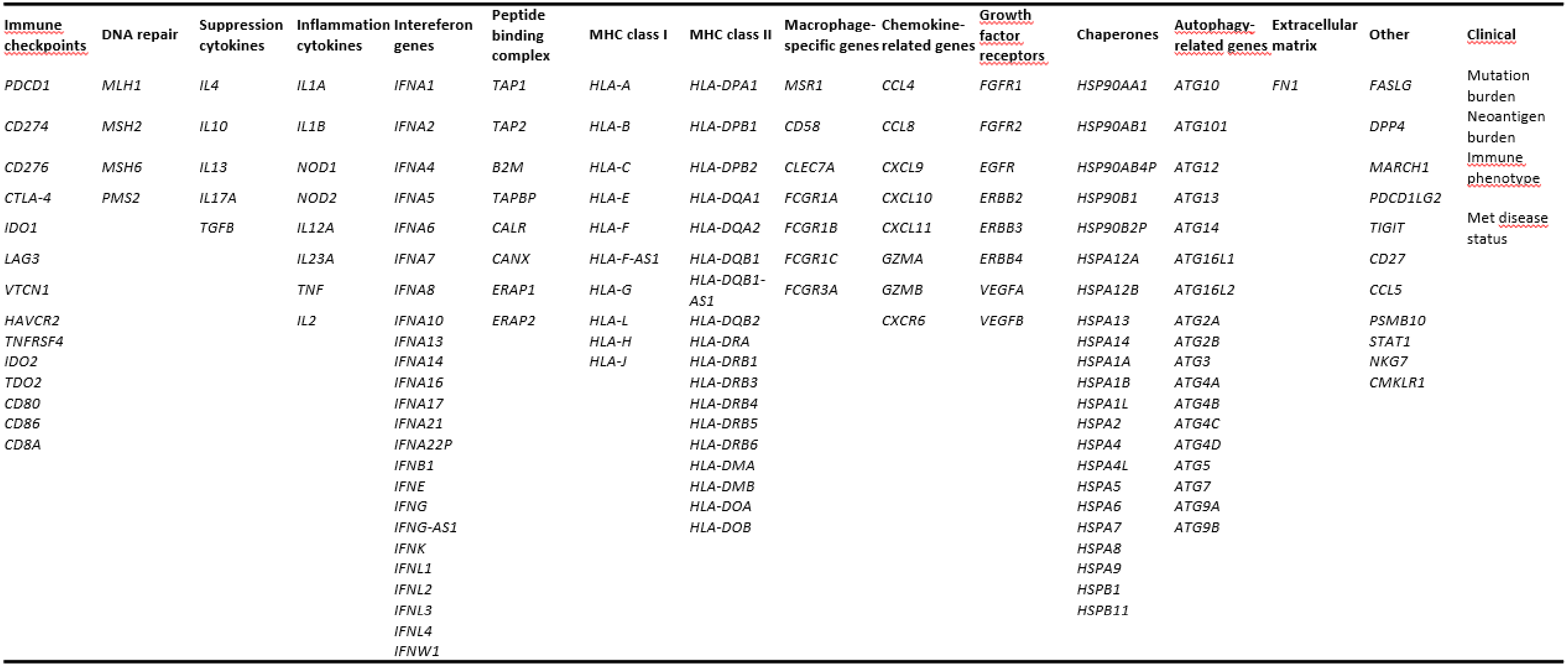
Use clinical data? Genes? Both?

Are some columns more important?



Chen et al., 2017

Literature research



Immune checkpoints	DNA repair	Suppression cytokines	Inflammation cytokines	Intereferon genes	Peptide binding complex	MHC class I	MHC class II	Macrophage-specific genes	Chemokine-related genes	Growth factor receptors	Chaperones	Autophagy-related genes	Extracellular matrix	Other	Clinical
PDCD1	*MLH1*	*IL4*	*IL1A*	*IFNA1*	*TAP1*	*HLA-A*	*HLA-DPA1*	*MSR1*	*CCL4*	*FGFR1*	*HSP90AA1*	*ATG10*	*FN1*	*FASLG*	Mutation burden
CD274	*MSH2*	*IL10*	*IL1B*	*IFNA2*	*TAP2*	*HLA-B*	*HLA-DPB1*	*CD58*	*CCL8*	*FGFR2*	*HSP90AB1*	*ATG101*		*DPP4*	Neoantigen burden
CD276	*MSH6*	*IL13*	*NOD1*	*IFNA4*	*B2M*	*HLA-C*	*HLA-DPB2*	*CLEC7A*	*CXCL9*	*EGFR*	*HSP90AB4P*	*ATG12*		*MARCH1*	Immune phenotype
CTLA-4	*PMS2*	*IL17A*	*NOD2*	*IFNA5*	*TAPBP*	*HLA-E*	*HLA-DQA1*	*FCGR1A*	*CXCL10*	*ERBB2*	*HSP90B1*	*ATG13*		*PDCD1LG2*	
IDO1		*TGFB*	*IL12A*	*IFNA6*	*CALR*	*HLA-F*	*HLA-DQA2*	*FCGR1B*	*CXCL11*	*ERBB3*	*HSP90B2P*	*ATG14*		*TIGIT*	Met disease status
LAG3			*IL23A*	*IFNA7*	*CANX*	*HLA-F-AS1*	*HLA-DQB1*	*FCGR1C*	*GZMA*	*ERBB4*	*HSPA12A*	*ATG16L1*		*CD27*	
VTCN1			*TNF*	*IFNA8*	*ERAP1*	*HLA-G*	*HLA-DQB1-AS1*	*FCGR3A*	*GZMB*	*VEGFA*	*HSPA12B*	*ATG16L2*		*CCL5*	
HAVCR2			*IL2*	*IFNA10*	*ERAP2*	*HLA-L*	*HLA-DQB2*		*CXCR6*	*VEGFB*	*HSPA13*	*ATG2A*		*PSMB10*	
TNFRSF4				*IFNA13*		*HLA-H*	*HLA-DRA*				*HSPA14*	*ATG2B*		*STAT1*	
IDO2				*IFNA14*		*HLA-J*	*HLA-DRB1*				*HSPA1A*	*ATG3*		*NKG7*	
TDO2				*IFNA16*			*HLA-DRB3*				*HSPA1B*	*ATG4A*		*CMKLR1*	
CD80				*IFNA17*			*HLA-DRB4*				*HSPA1L*	*ATG4B*			
CD86				*IFNA21*			*HLA-DRB5*				*HSPA2*	*ATG4C*			
CD8A				*IFNA22P*			*HLA-DRB6*				*HSPA4*	*ATG4D*			
				IFNB1			*HLA-DMA*				*HSPA4L*	*ATG5*			
				IFNE			*HLA-DMB*				*HSPA5*	*ATG7*			
				IFNG			*HLA-DOA*				*HSPA6*	*ATG9A*			
				IFNG-AS1			*HLA-DOB*				*HSPA7*	*ATG9B*			
				IFNK							*HSPA8*				
				IFNL1							*HSPA9*				
				IFNL2							*HSPB1*				
				IFNL3							*HSPB11*				
				IFNL4											
				IFNW1											

166 genes

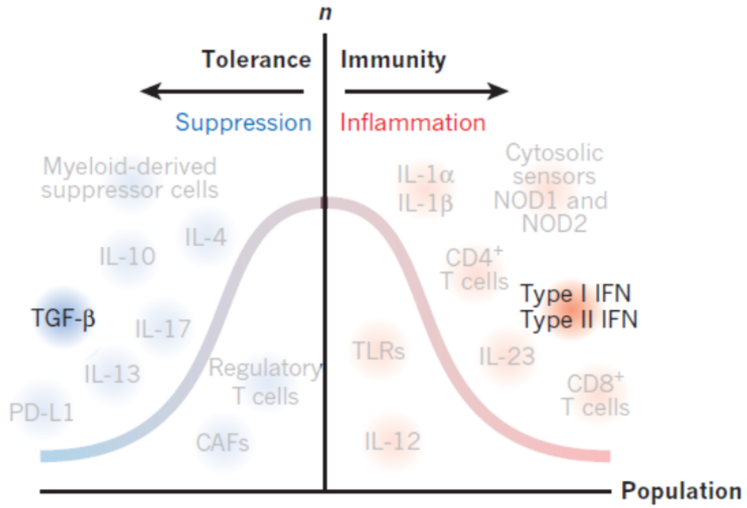
117 directly involved in immune regulation

4 clinical features

Final selection

Statistics revealed best markers

2 genes
IFNG
TGFB1



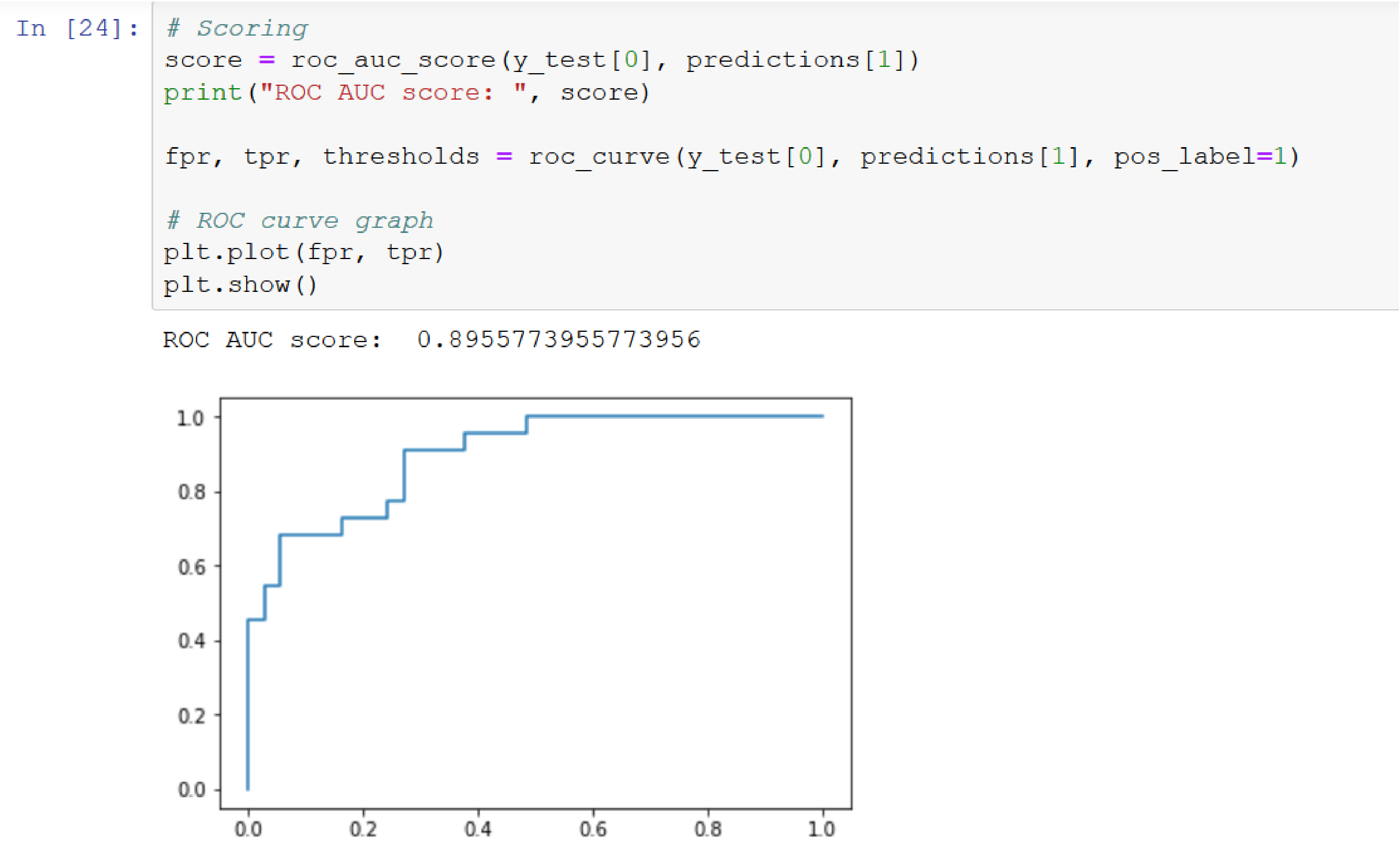
Chen et al., 2017

3 clinical features
MTB
Neoantigen burden
Met disease status

Basic neural network
10x cross-validated

Average accuracy: 0.76
Average ROC: 0.81

Area under ROC



```
In [24]: # Scoring
         score = roc_auc_score(y_test[0], predictions[1])
         print("ROC AUC score: ", score)

         fpr, tpr, thresholds = roc_curve(y_test[0], predictions[1], pos_label=1)

         # ROC curve graph
         plt.plot(fpr, tpr)
         plt.show()
```

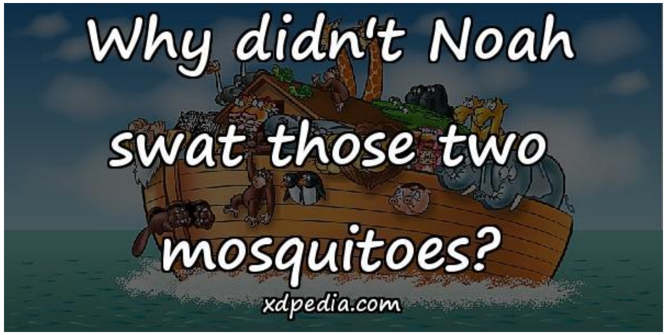
```
ROC AUC score:  0.8955773955773956
```

BioHack

bioinformatics hackathon

Thank you for your attention

jacek.smietanski@ii.uj.edu.pl



Thank you for your attention

jacek.smietanski@ii.uj.edu.pl