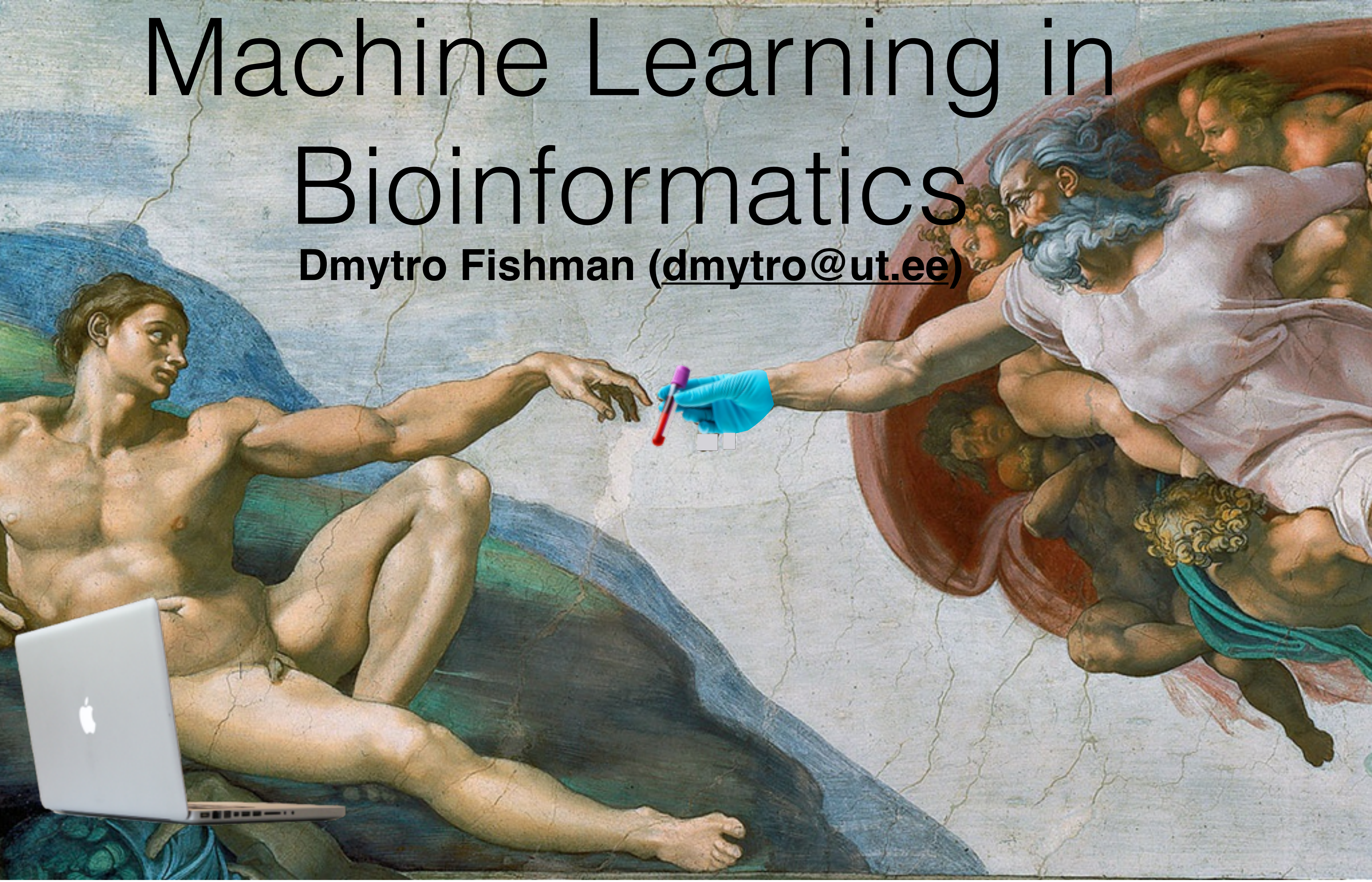
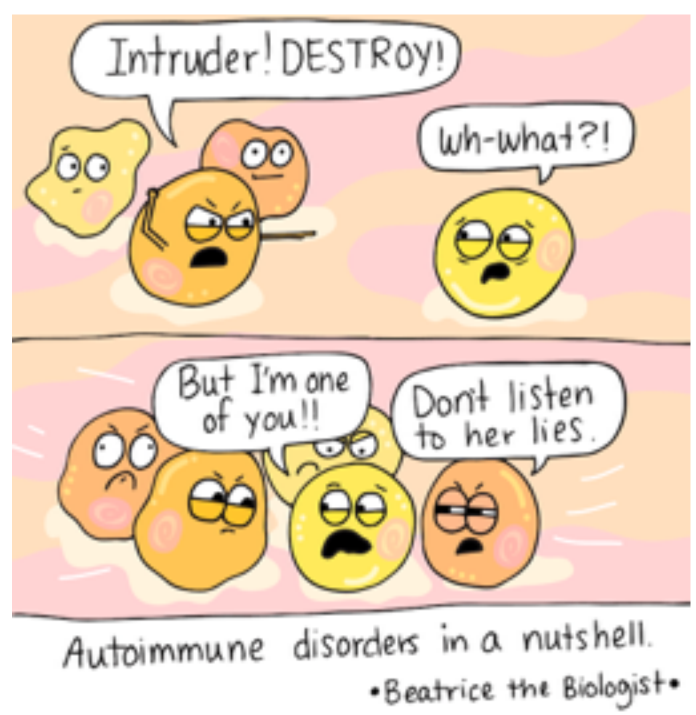
Machine Learning in Bioinformatics

Dmytro Fishman (dmytro@ut.ee)



BIIT

Dmytro Fishman

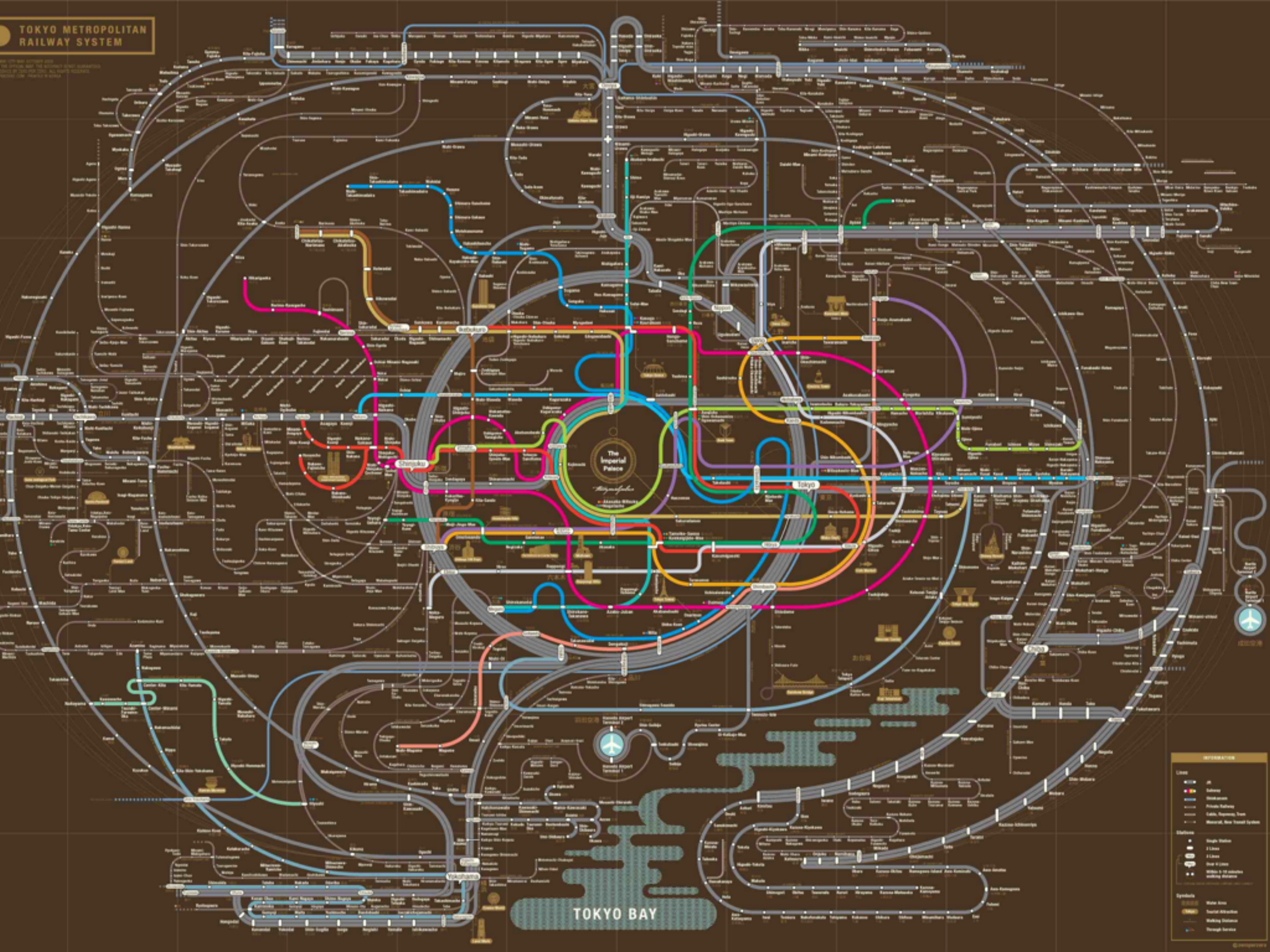


Oct 21 - 23, Tartu

E-mail: **dmytro@ut.ee**

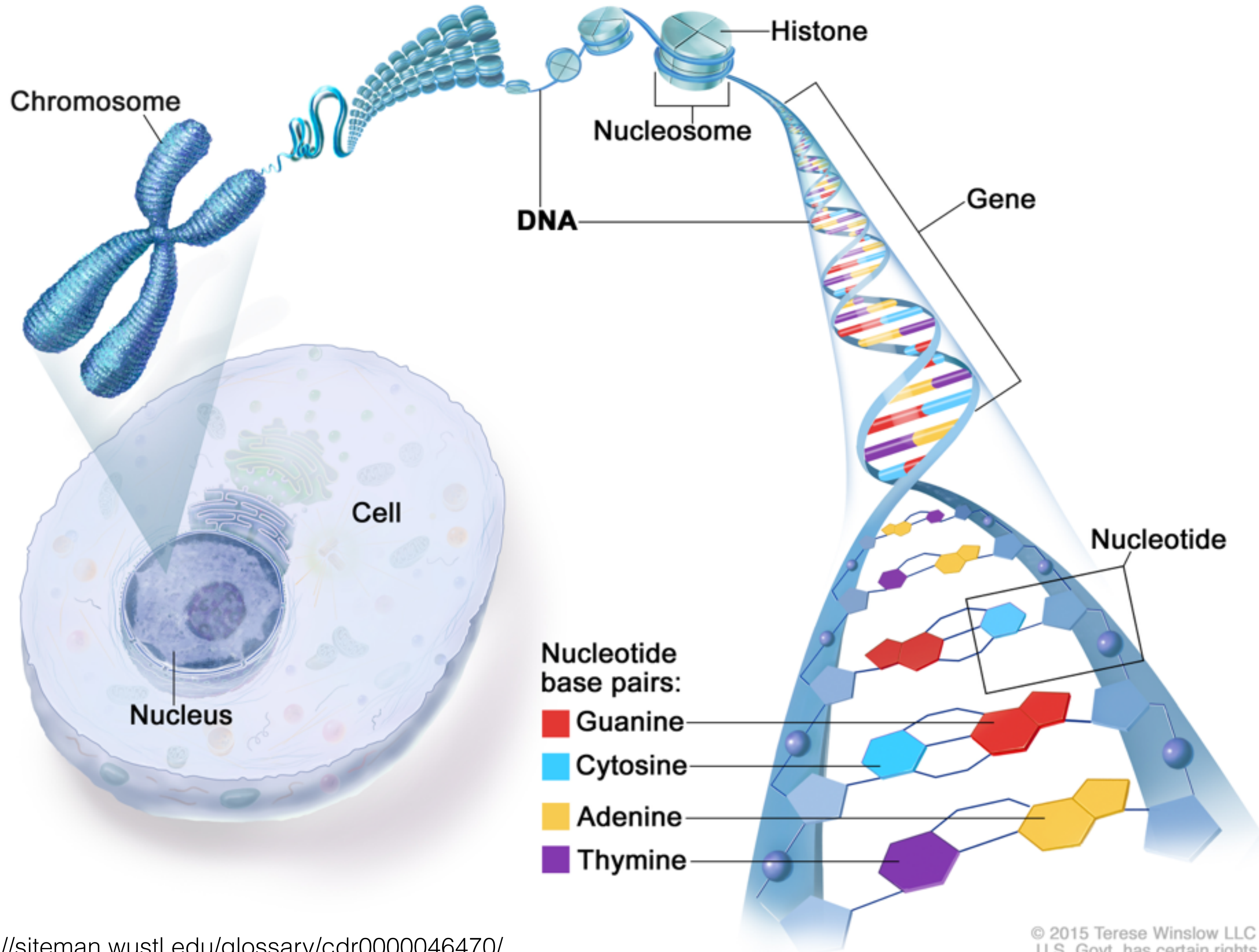
Machine Learning in Bioinformatics

Machine Learning in **Bioinformatics**



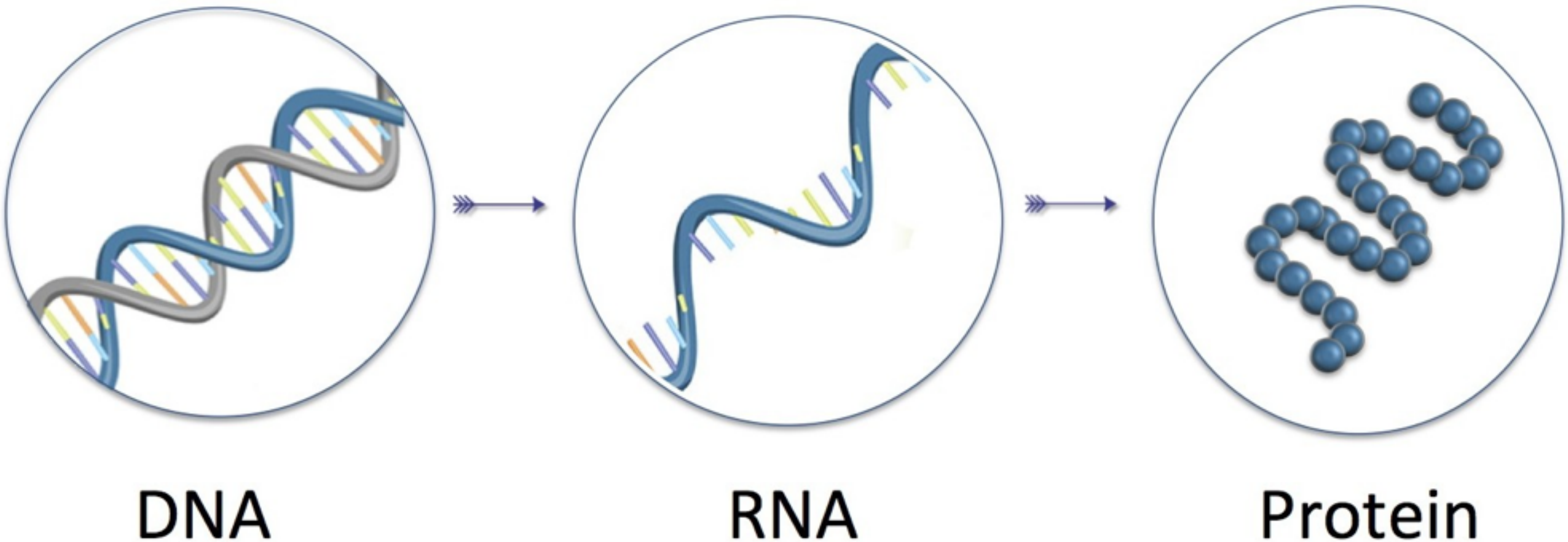
TOKYO METROPOLITAN
RAILWAY SYSTEM
The Imperial Palace
Tokyo
Shinjuku
TOKYO BAY
INFORMATION

DNA Structure



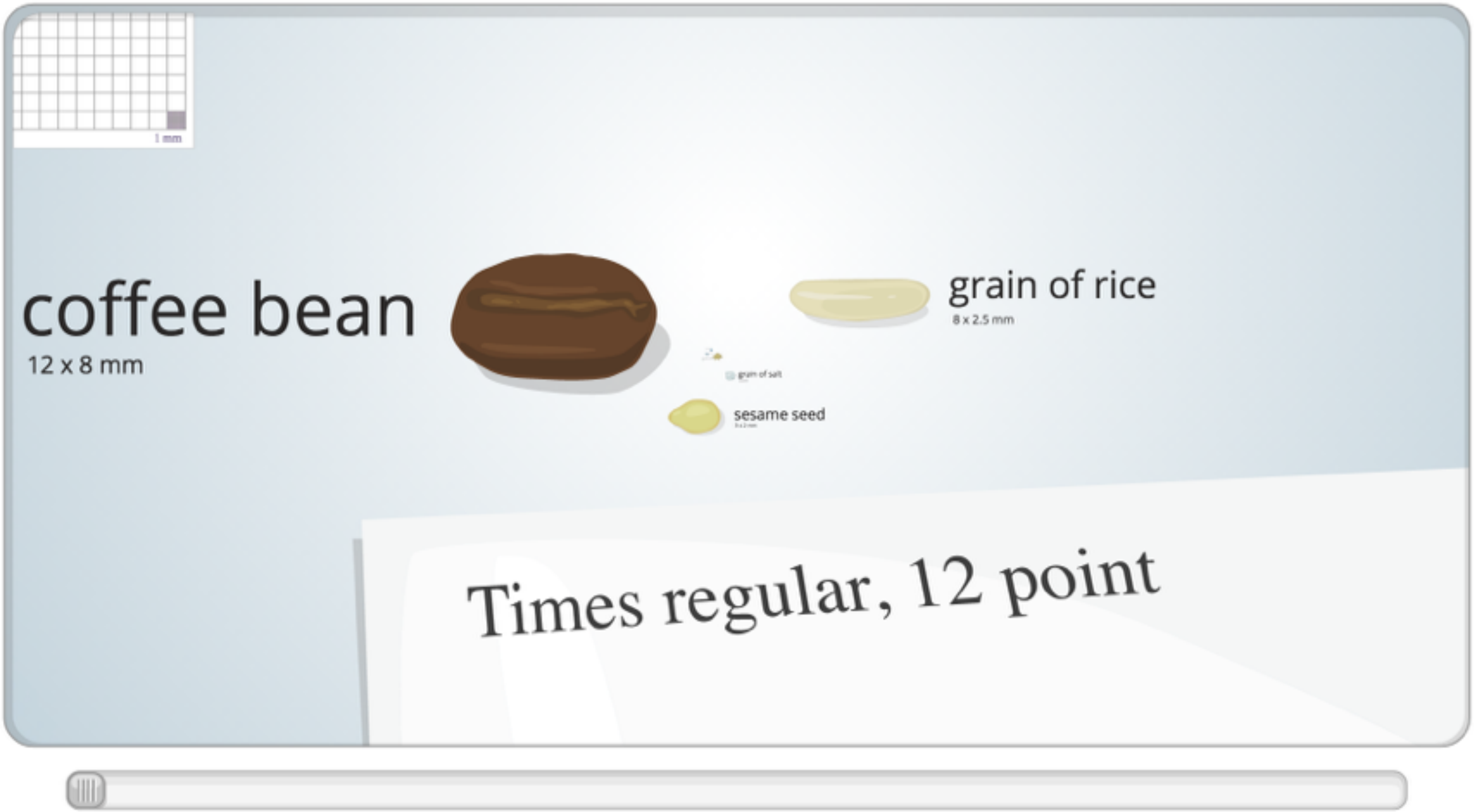
https://siteman.wustl.edu/glossary/cdr0000046470/

Central dogma



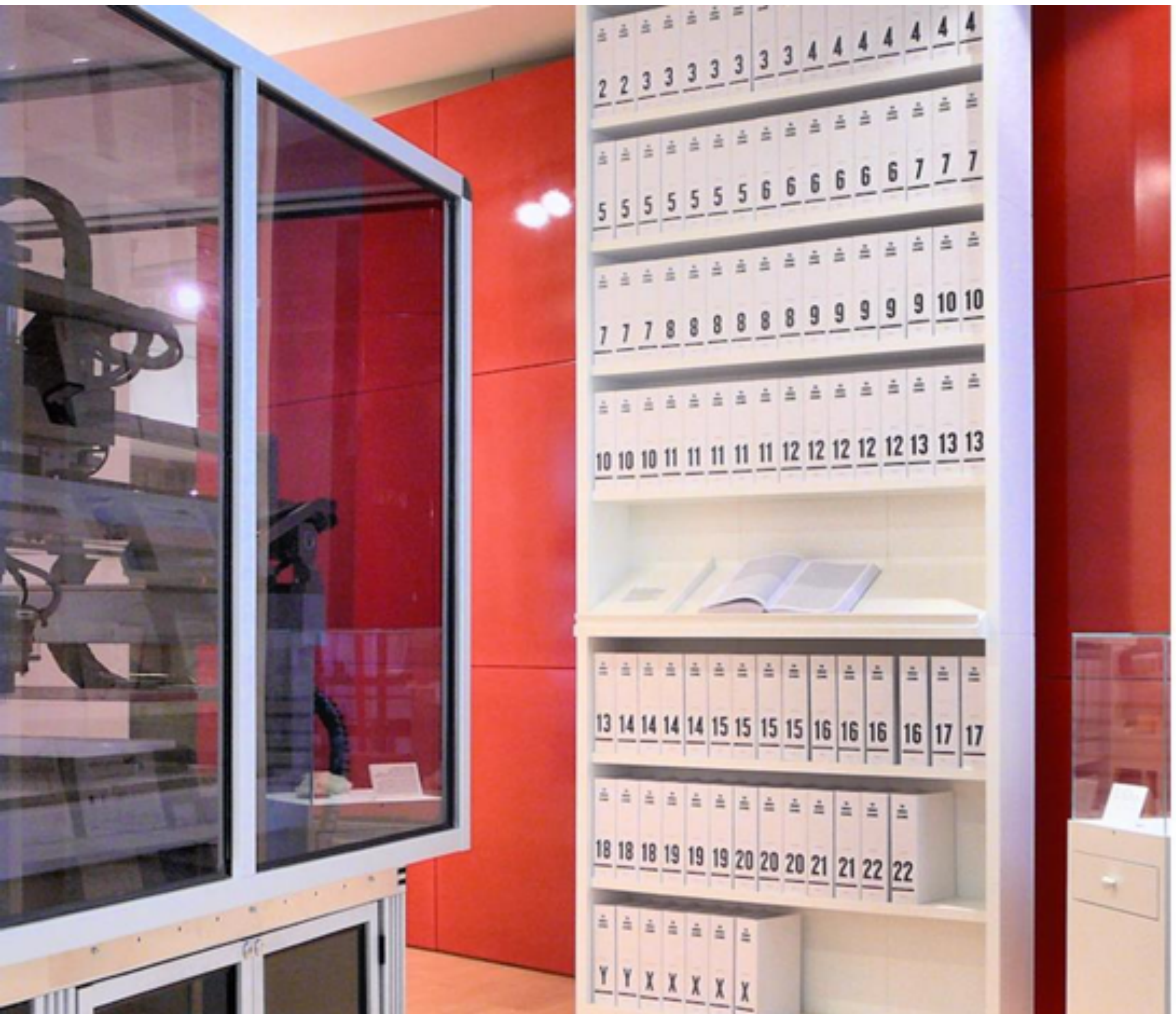


Microlevel

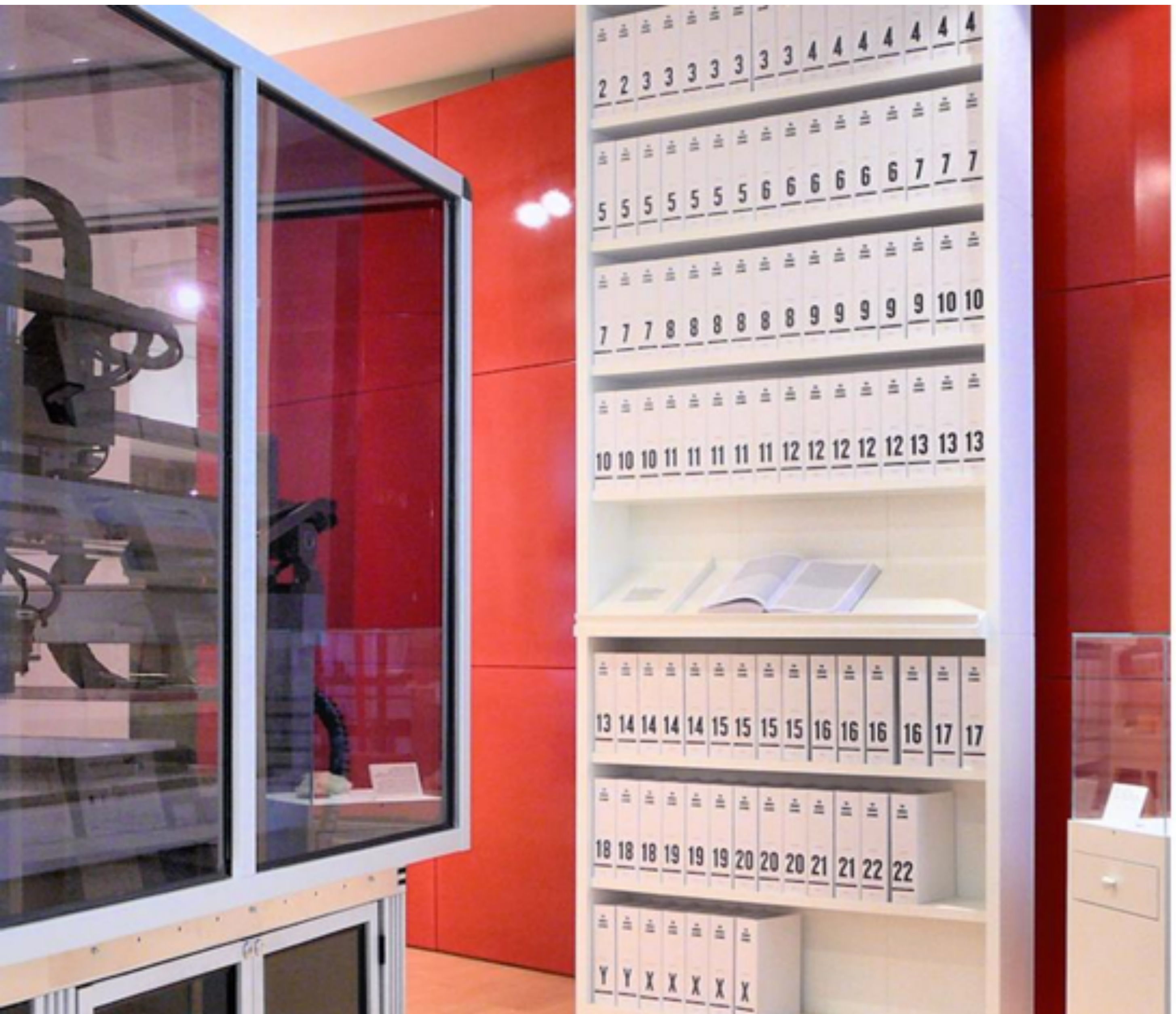


Meter	Centimeter	Millimeter	Micrometer	Nanometer	Angstrom	Picometer
10^{0} m	10^{-2} m	10^{-3} m	10^{-6} m	10^{-9} m	10^{-10} m	10^{-12} m
1 m	0.01 m	0.001 m	0.000001 m	0.000000001 m	0.0000000001 m	0.000000000001 m
	1/100 m	1/1,000 m	1/1,000,000 m	1/1,000,000,000 m	1/10,000,000,000 m	1/1,000,000,000,000 m
	hundreth of a meter	thousandth of a meter	millionth of a meter	billionth of a meter	ten billionth of a meter	trillionth of a meter

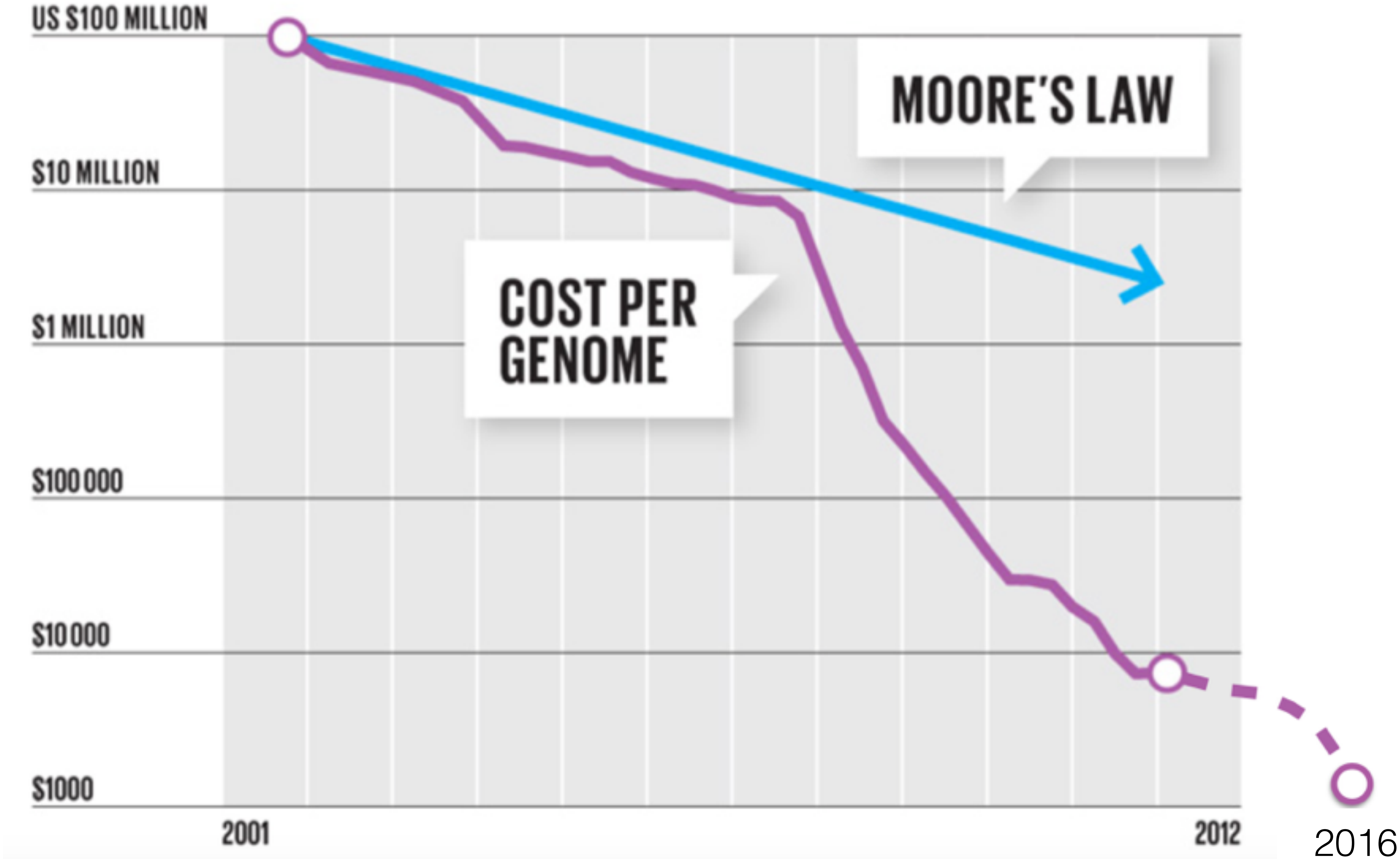
http://learn.genetics.utah.edu/content/cells/scale/



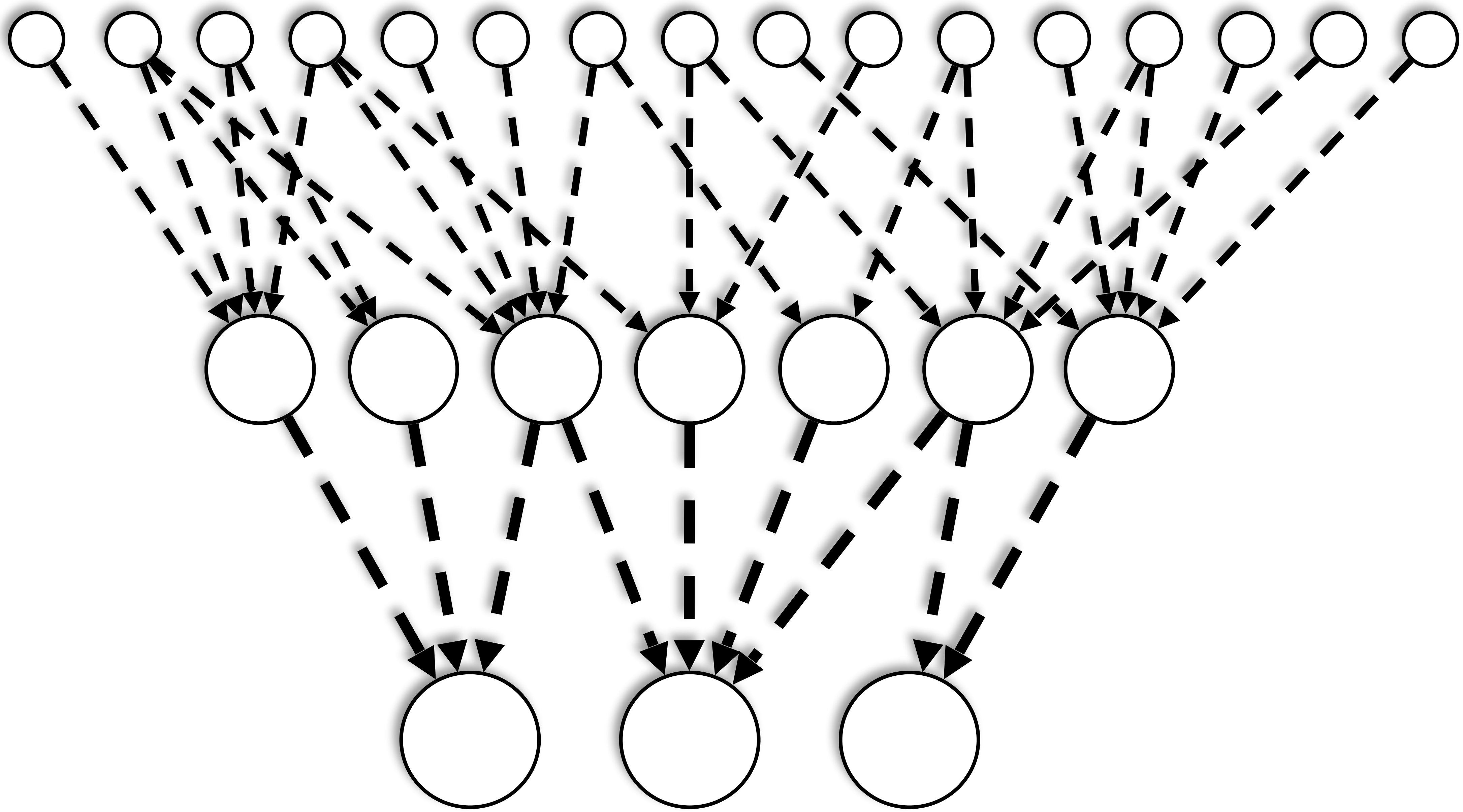
What these **books** contain?



This is a single
human genome
~4GB


US $100 MILLION
$10 MILLION
$1 MILLION
$100 000
$10 000
$1000
MOORE'S LAW
COST PER GENOME
2001
2012
2016

Labs that produce data



Research groups that can analyse data

What is Bioinformatics?

"Bioinformatics is an interdisciplinary field that develops methods and software tools for extracting knowledge from biological data."

https://en.wikipedia.org/wiki/Bioinformatics

What is Bioinformatics?

"**Bioinformatics** is

- an **interdisciplinary** field
- develops **methods** and **software tools**
- for extracting knowledge from **biological data**."

What is Bioinformatics?

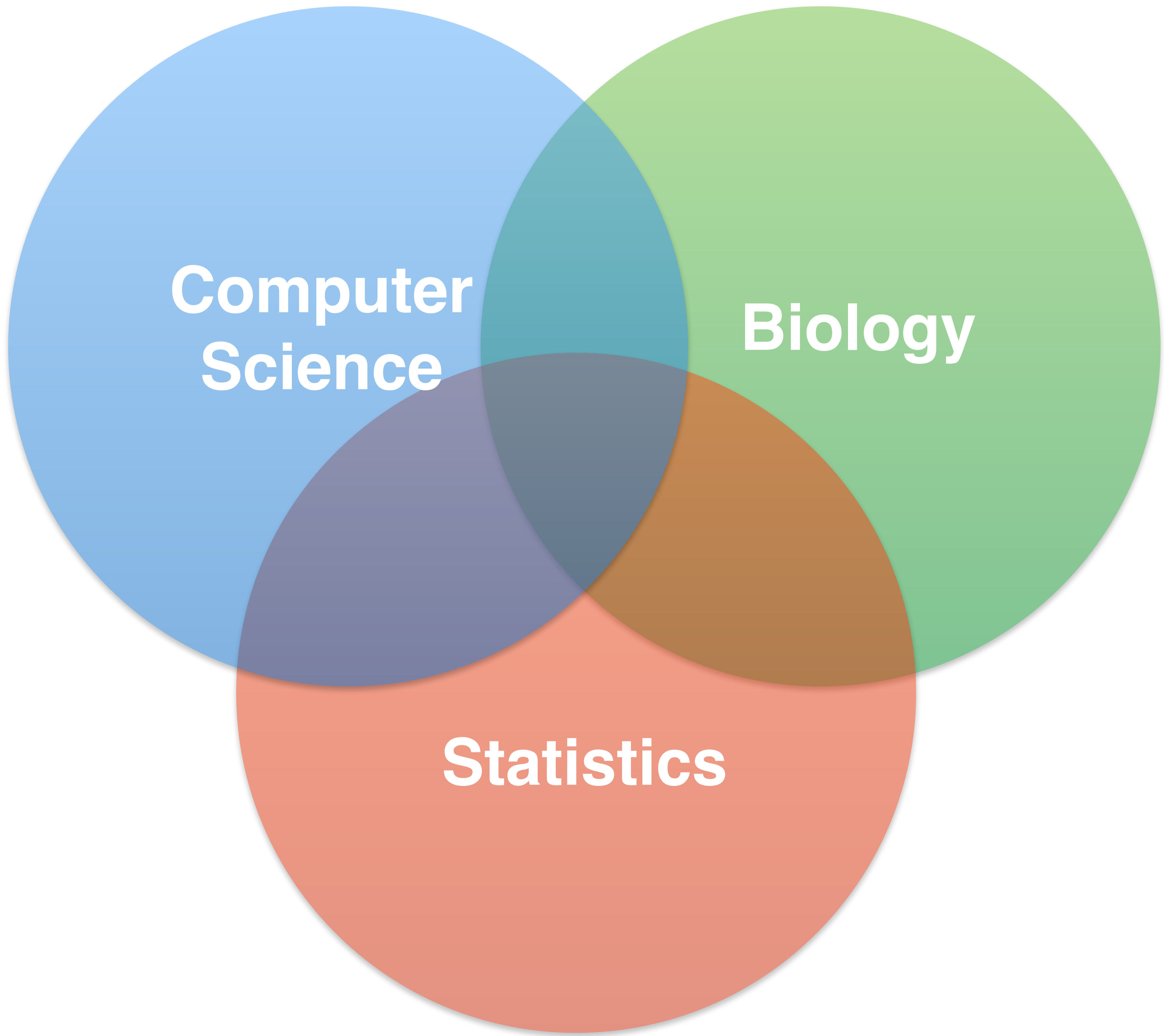
"**Bioinformatics** is

- an **interdisciplinary** field
- develops **methods** and **software tools**
- for extracting knowledge from **biological data**."

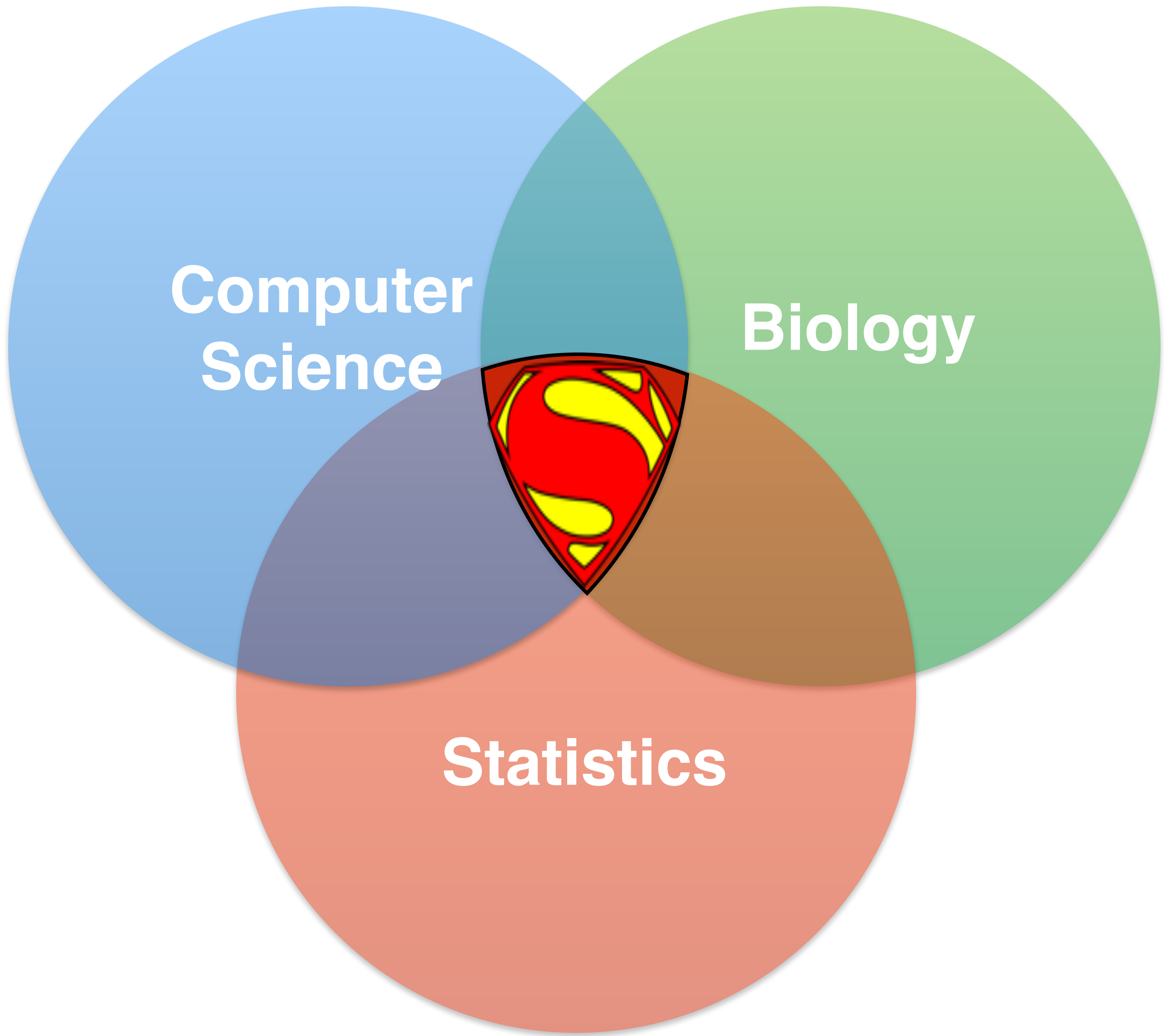
✚

Presenting this knowledge to **non-experts**

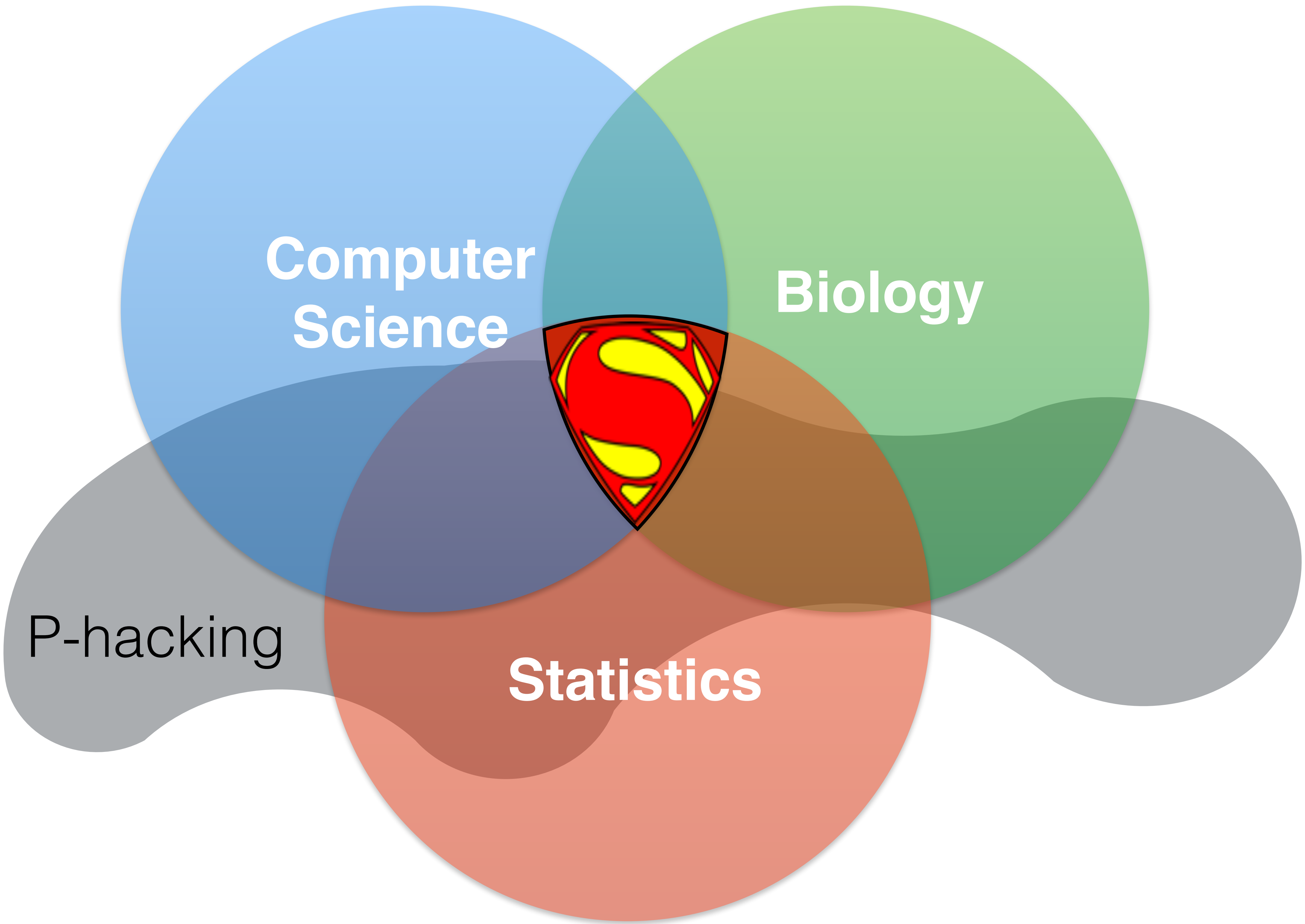
Bioinformatics



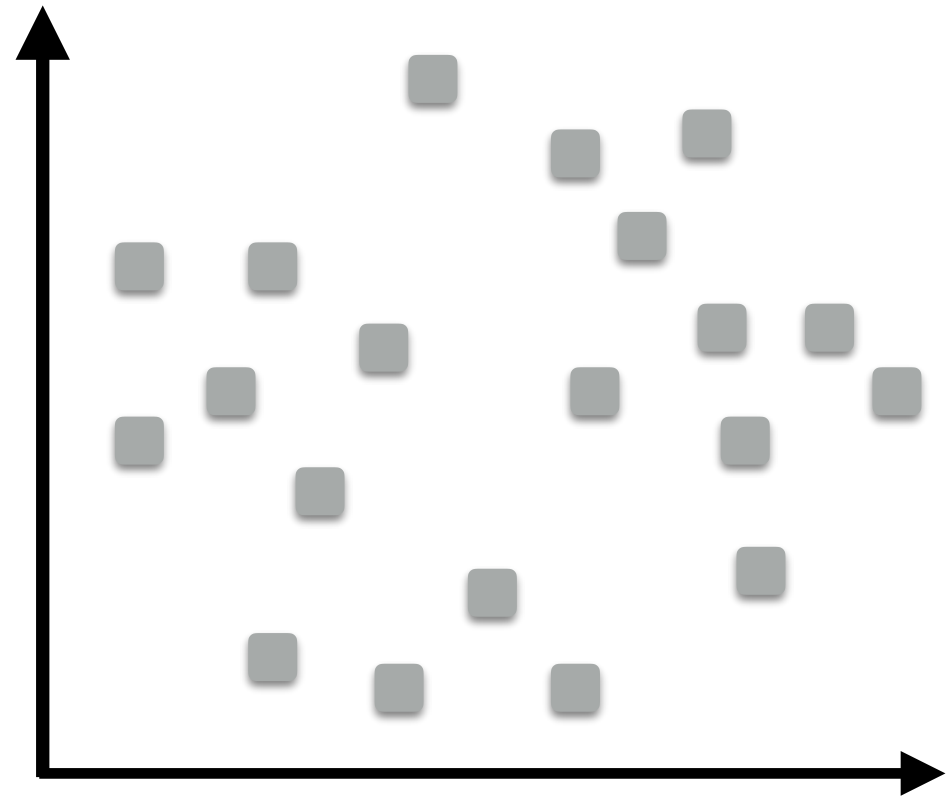
Bioinformatics

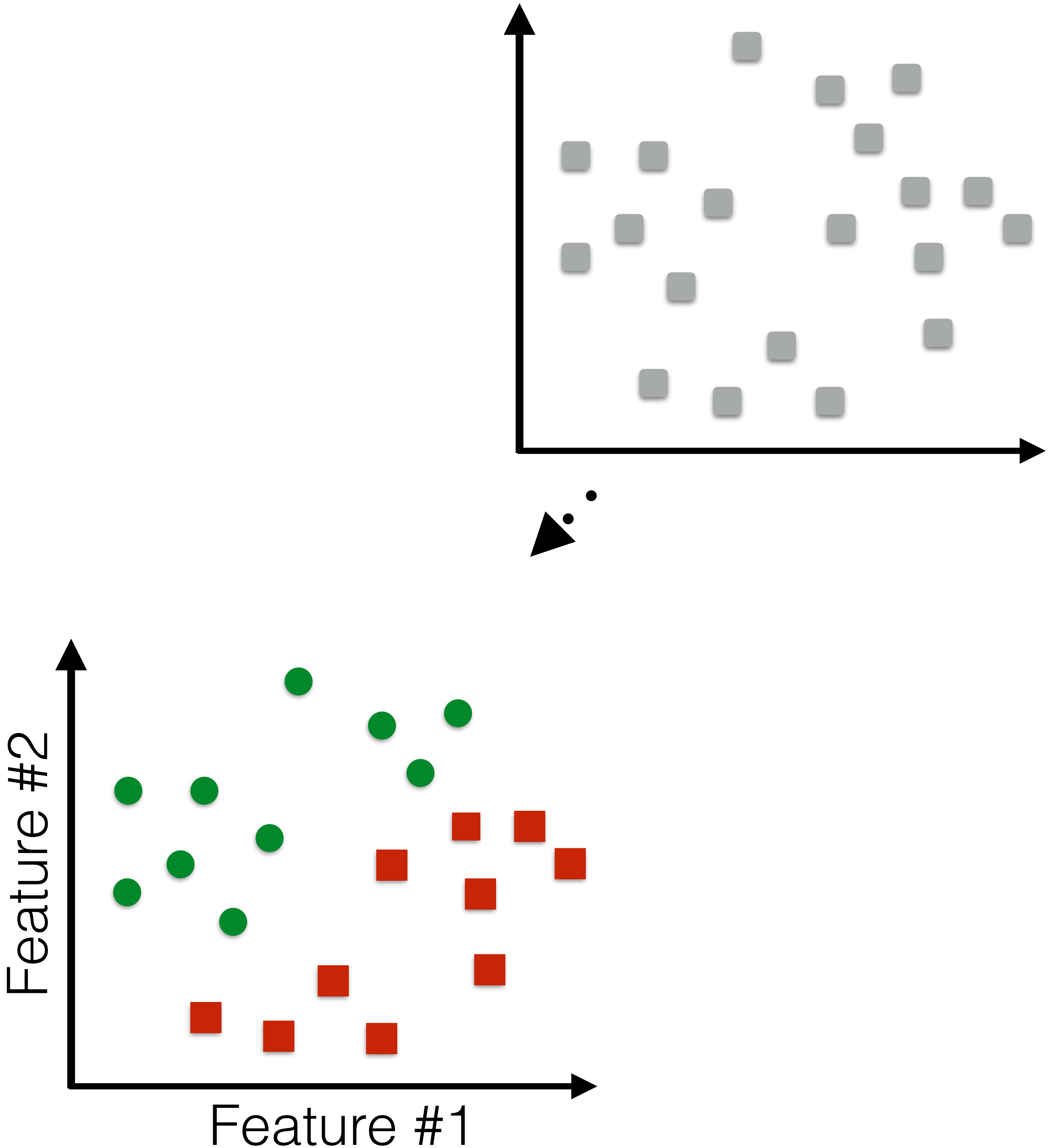


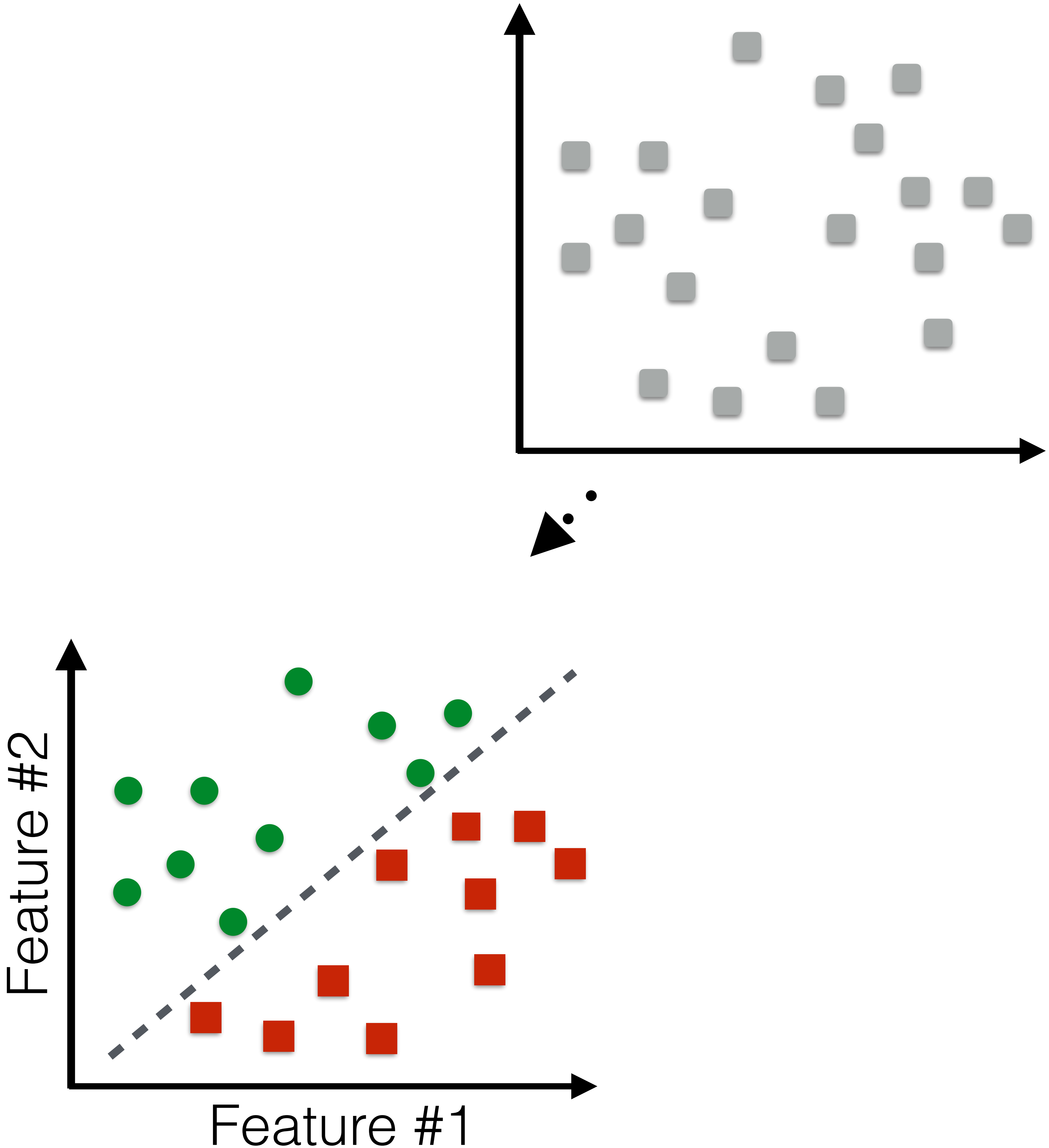
Bioinformatics

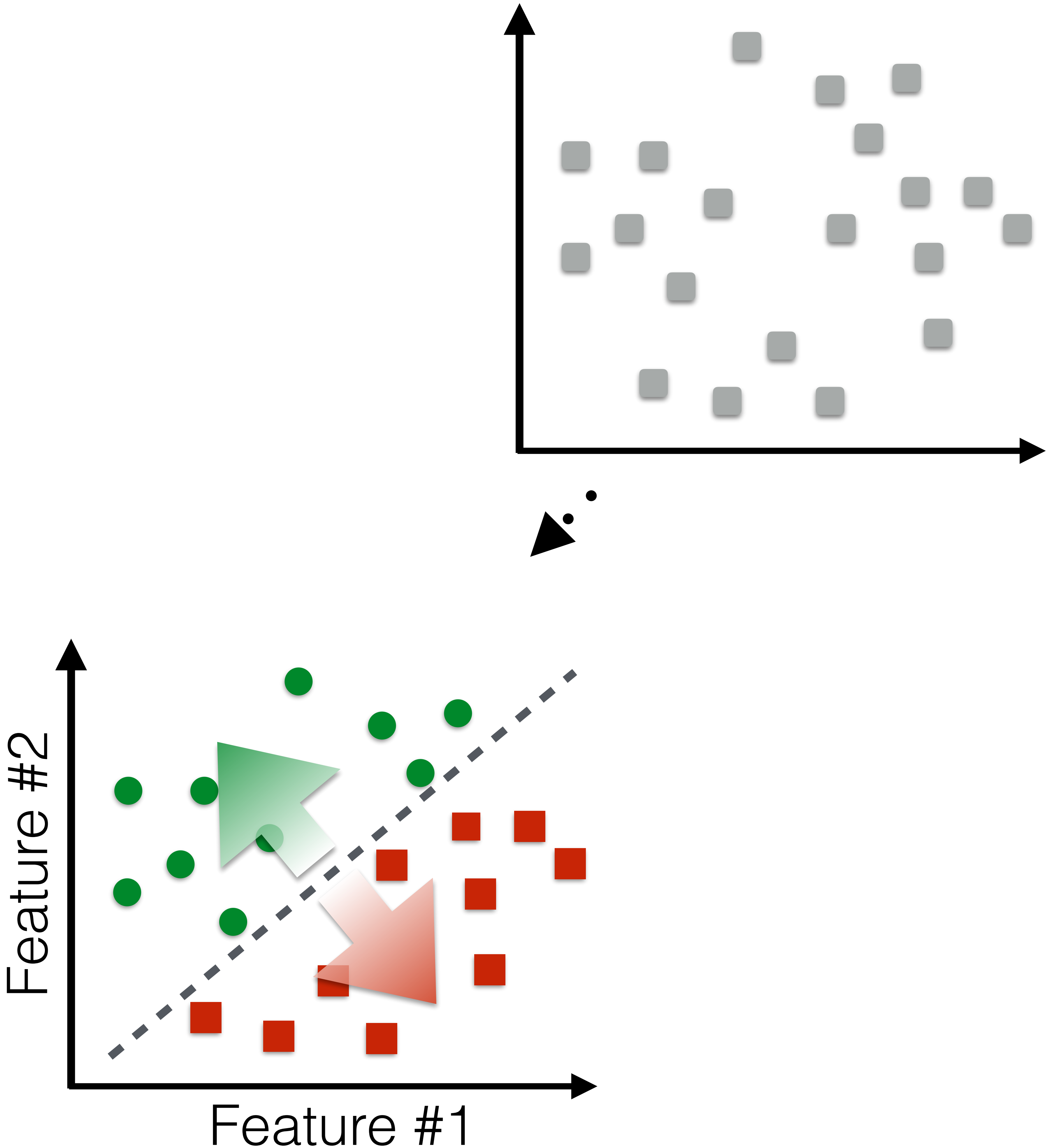


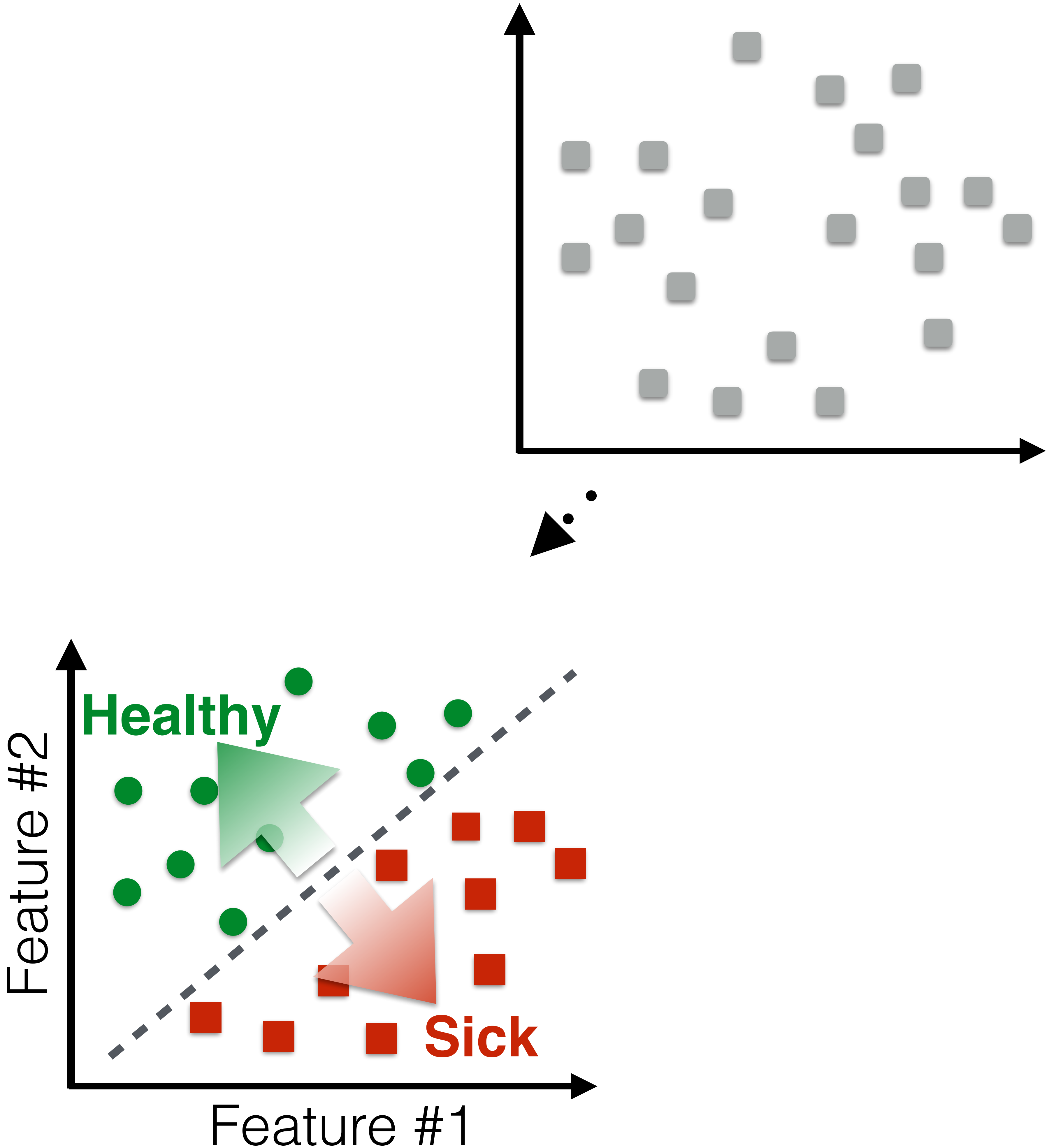
Machine Learning in Bioinformatics

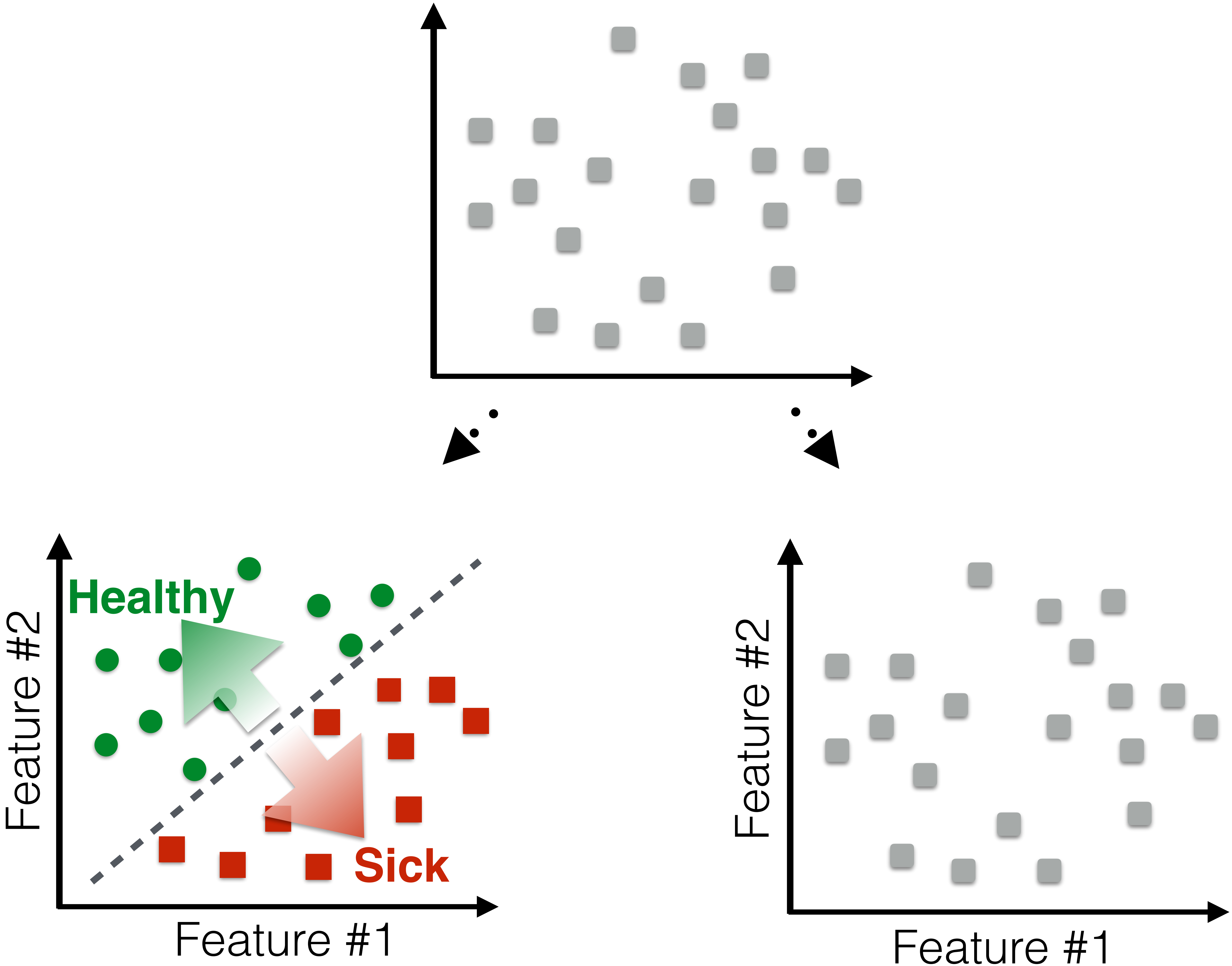



Feature #2
Feature #1

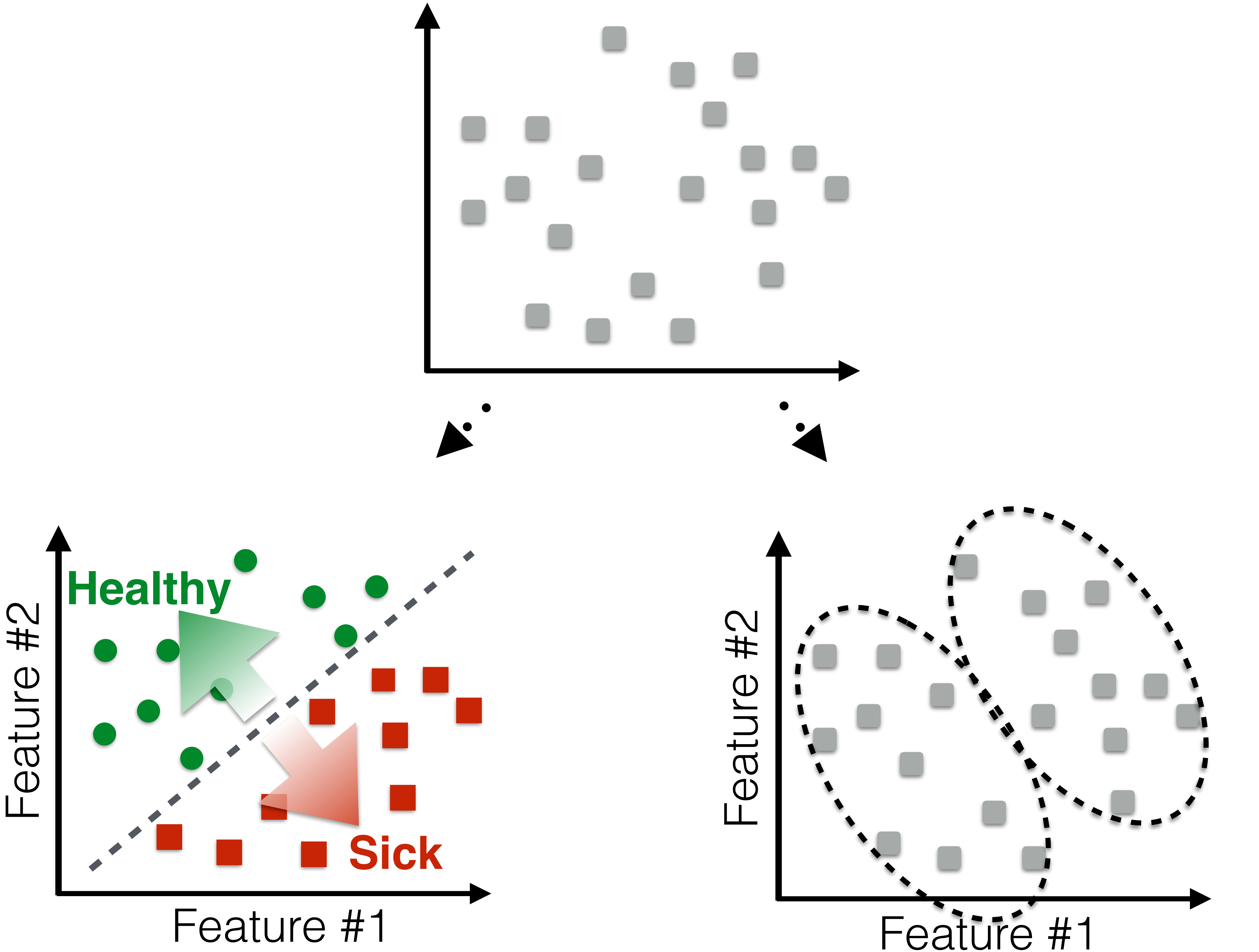

Feature #2
Feature #1

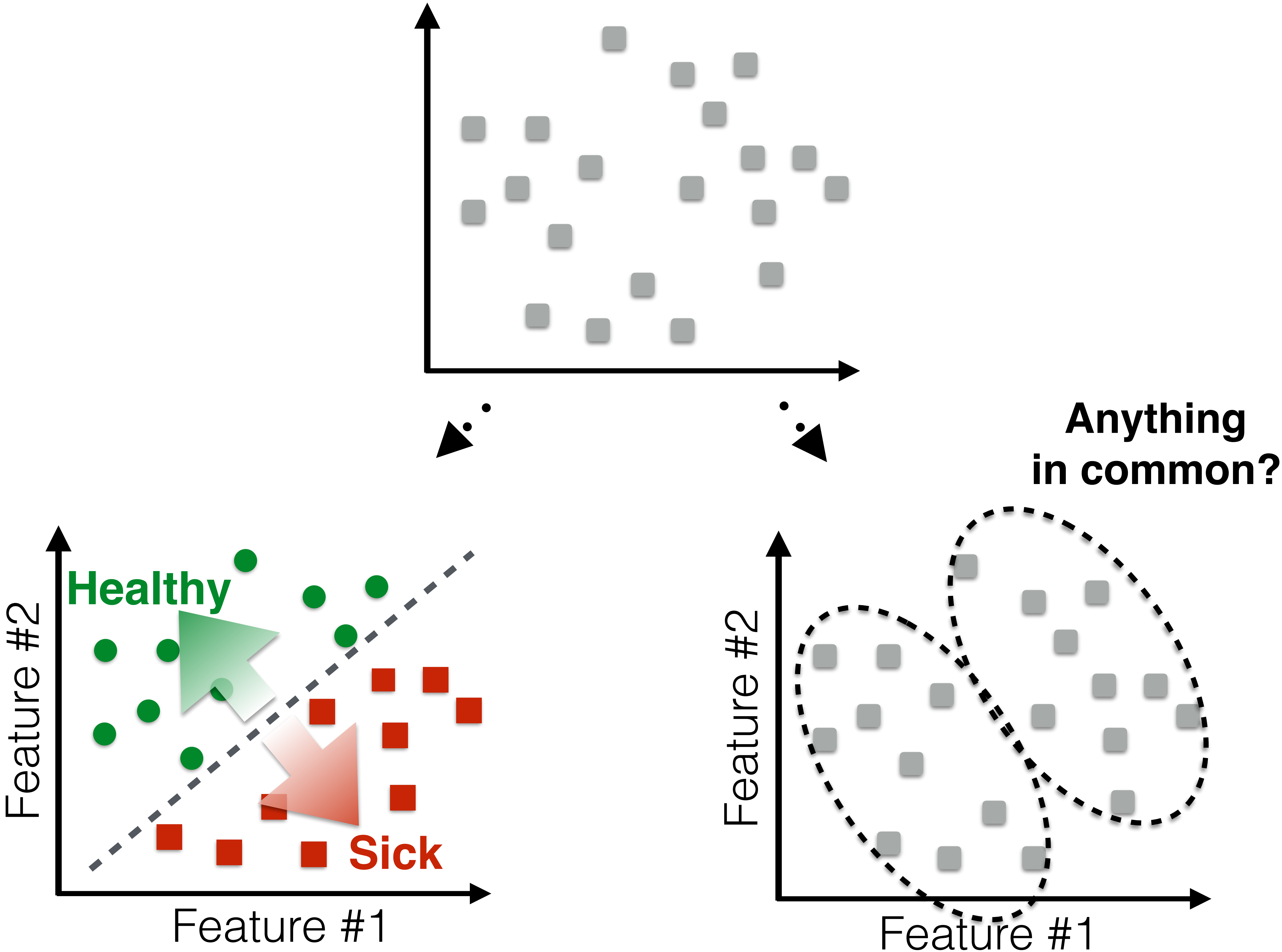

Feature #2
Feature #1

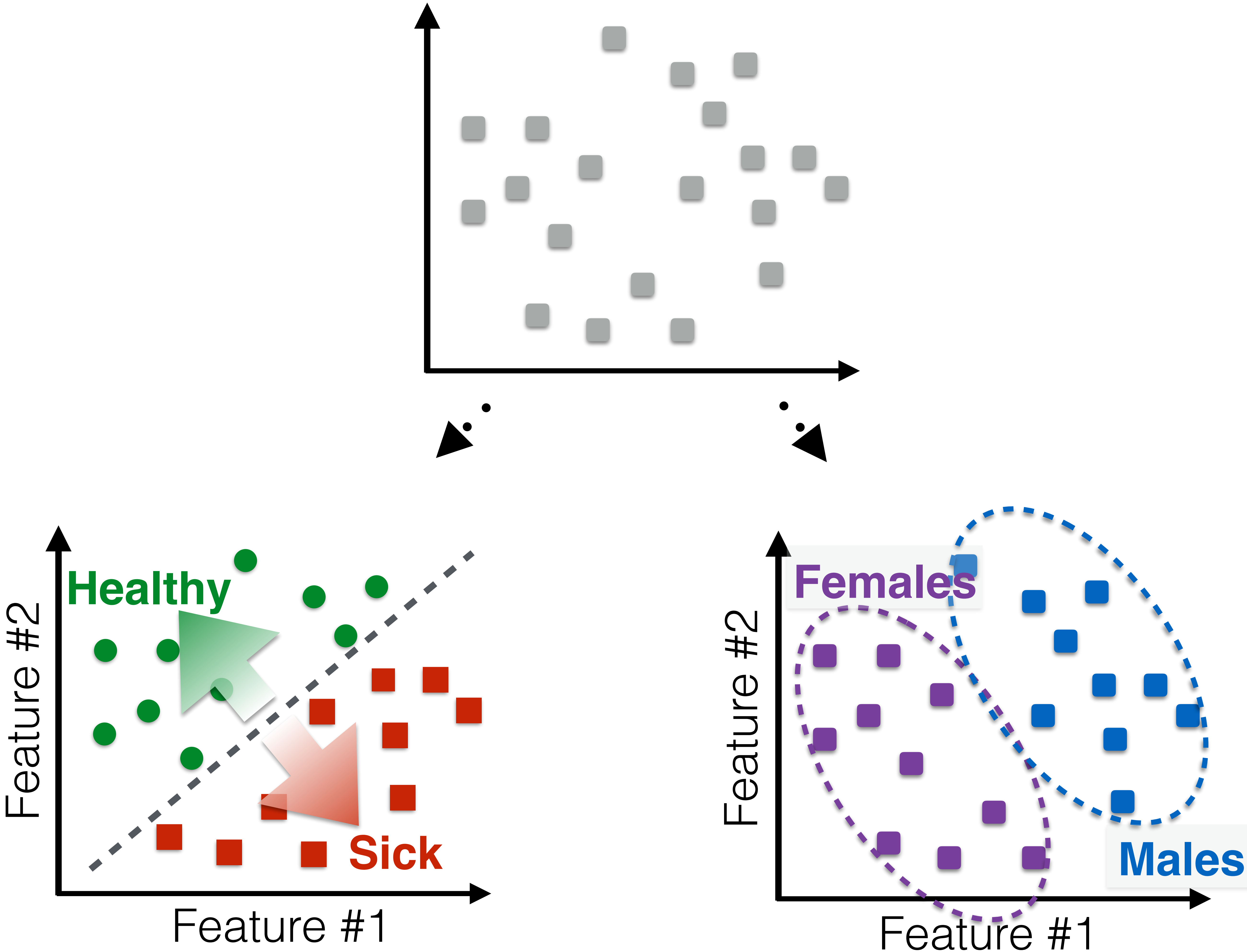

Healthy
Sick
Feature #2
Feature #1

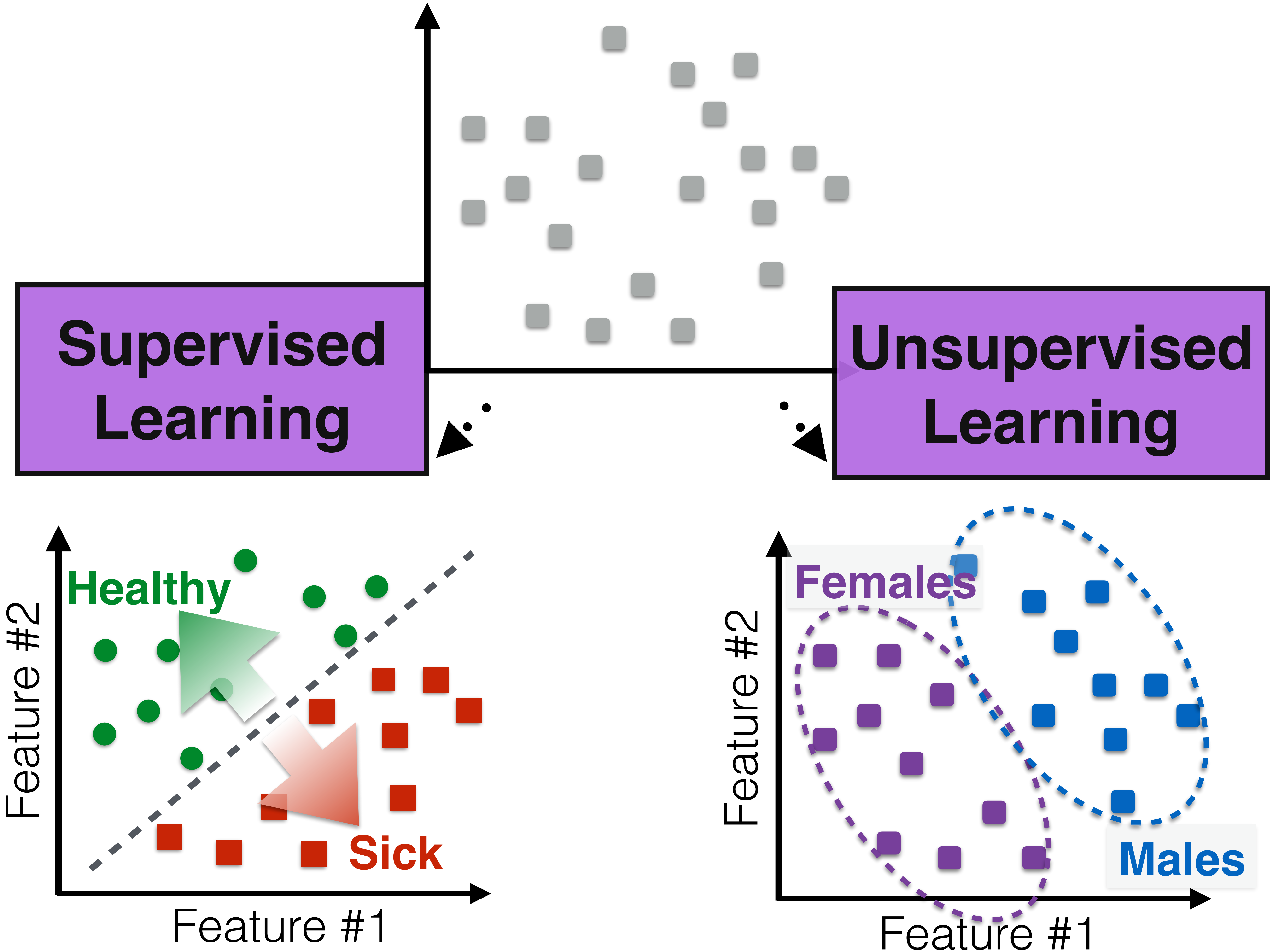


Healthy
Sick
Feature #2
Feature #1
Feature #2
Feature #1

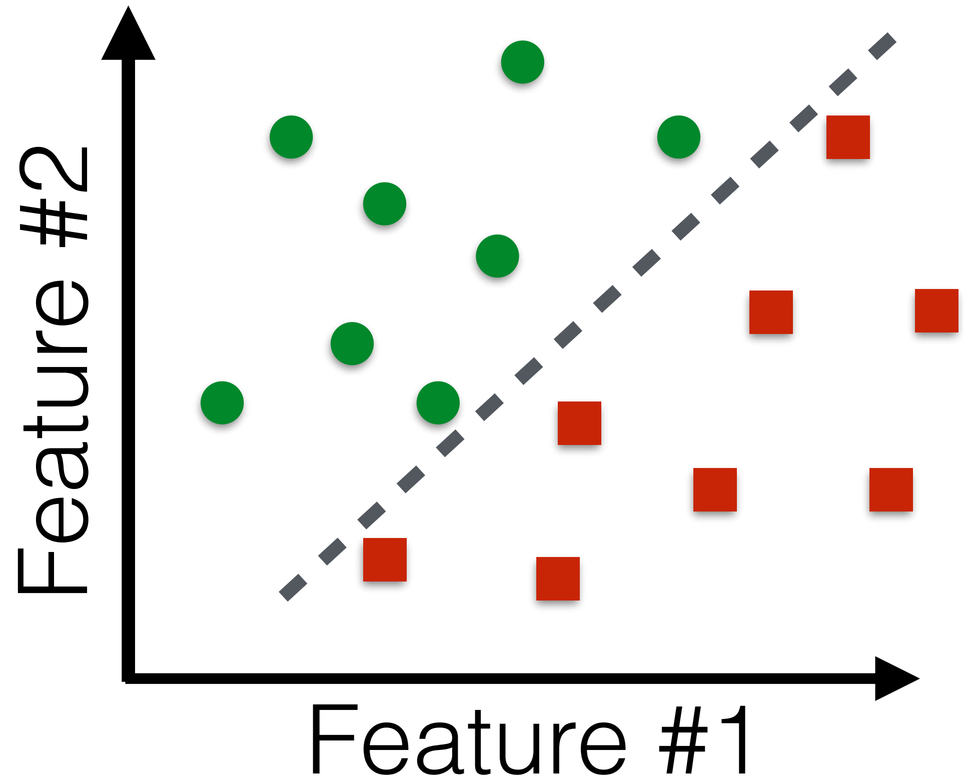

Healthy
Sick
Feature #2
Feature #1
Feature #2
Feature #1


Healthy
Sick
Feature #2
Feature #1
Anything
in common?
Feature #2
Feature #1

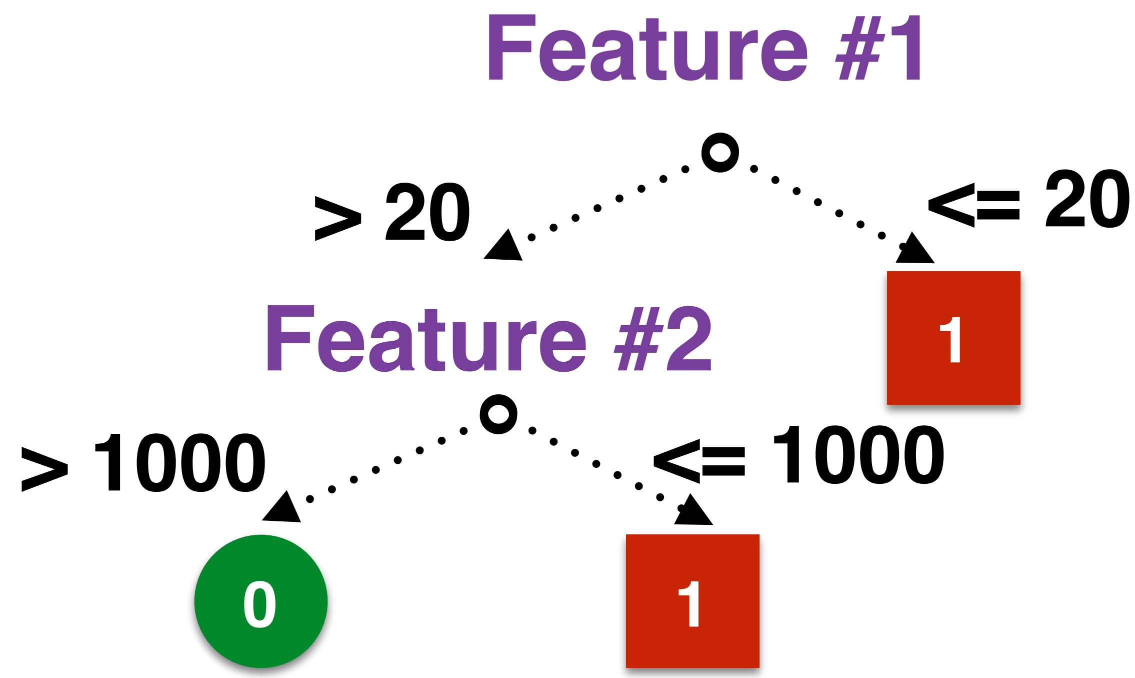

Healthy
Sick
Feature #2
Feature #1
Females
Males
Feature #2
Feature #1


Supervised Learning
Unsupervised Learning
Healthy
Sick
Feature #2
Feature #1
Females
Males
Feature #2
Feature #1

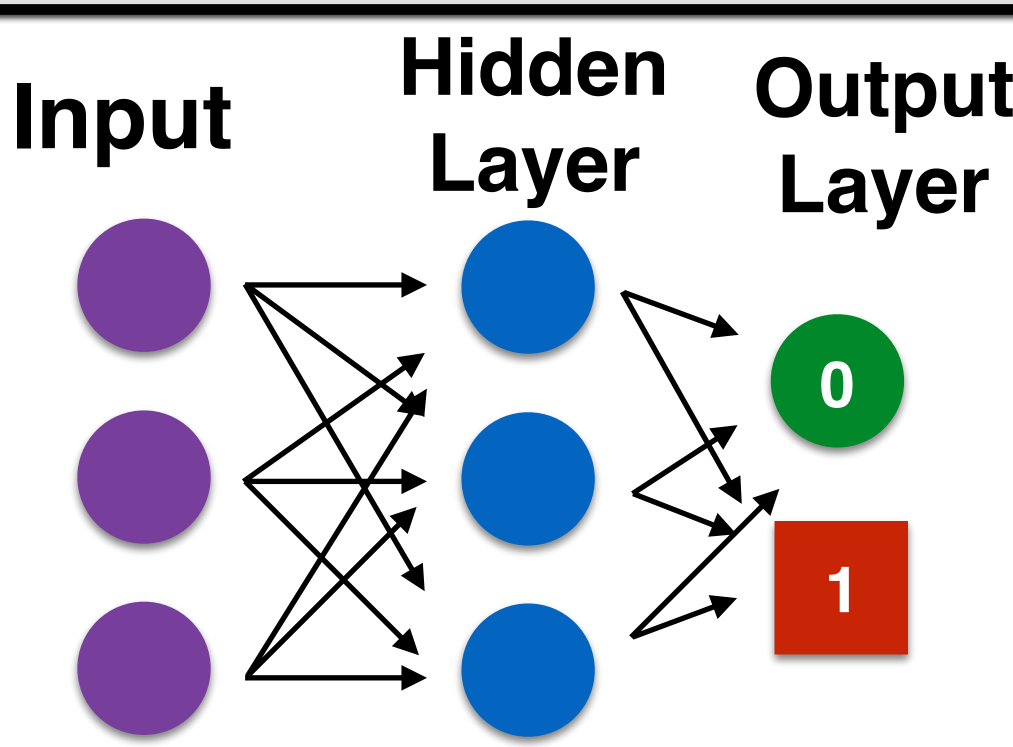
Linear Classifiers



Decision Trees



Neural Networks



Others

k-Nearest Neighbors

Self-Organizing Map

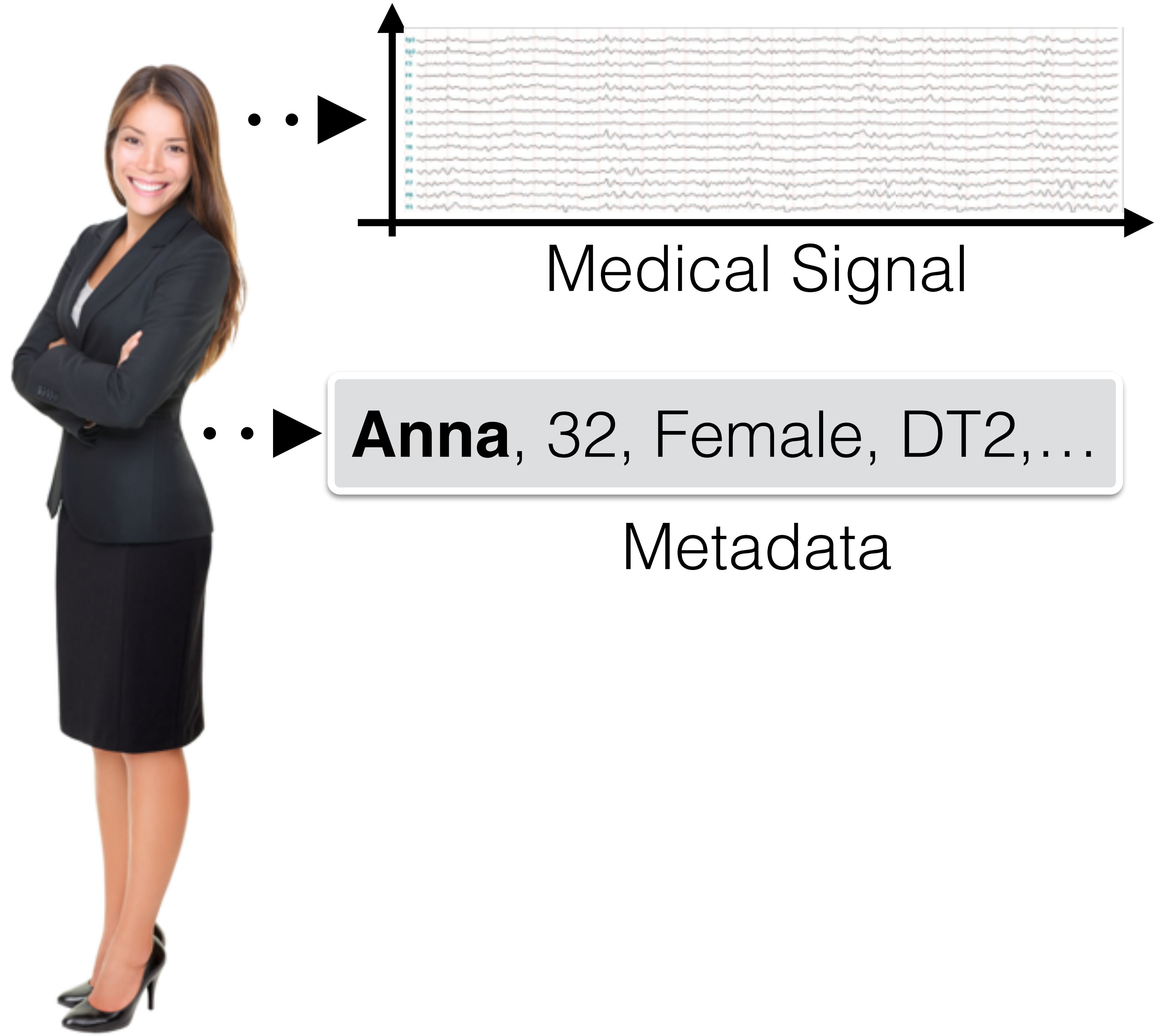
Many more

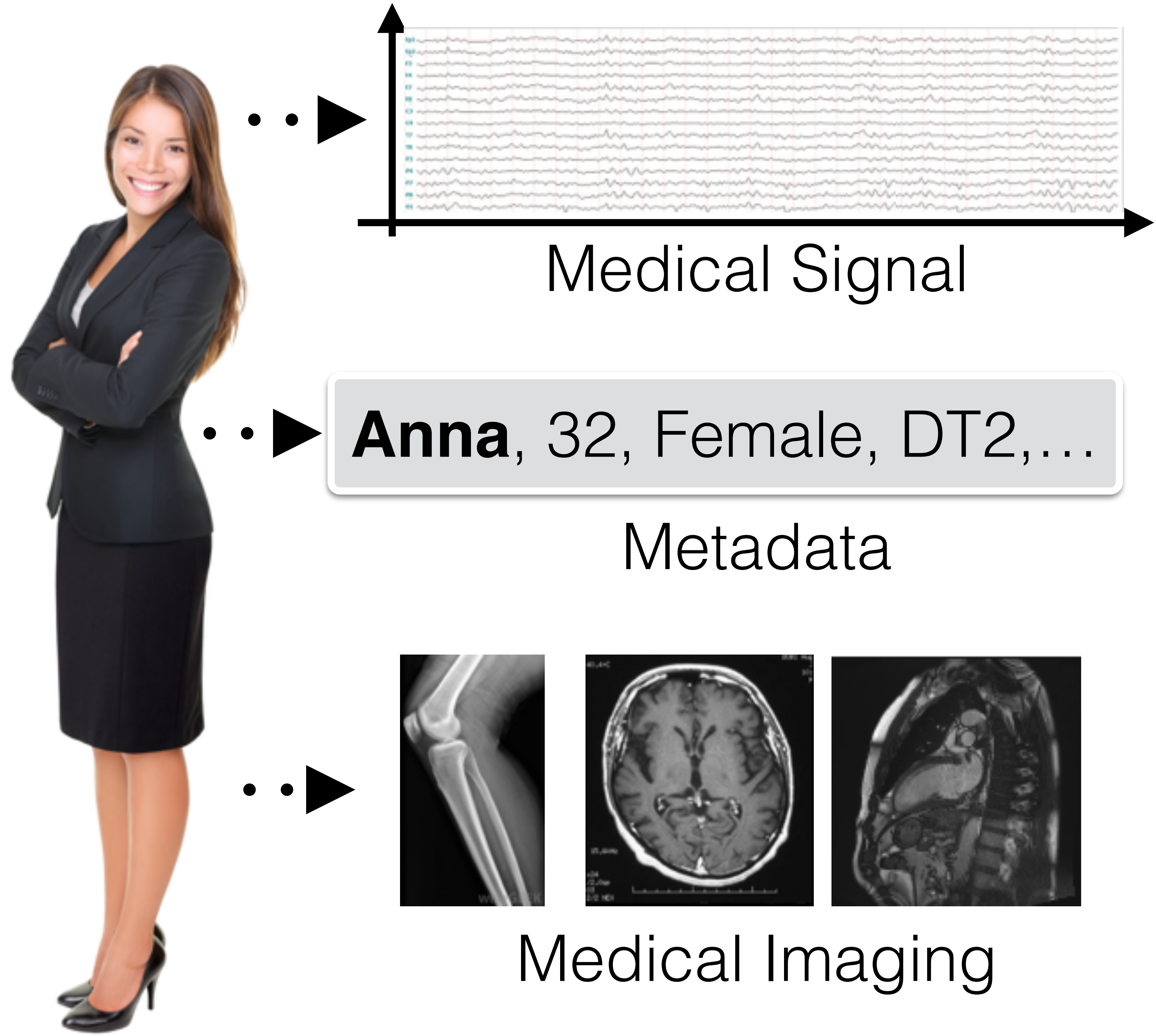
https://topepo.github.io/caret

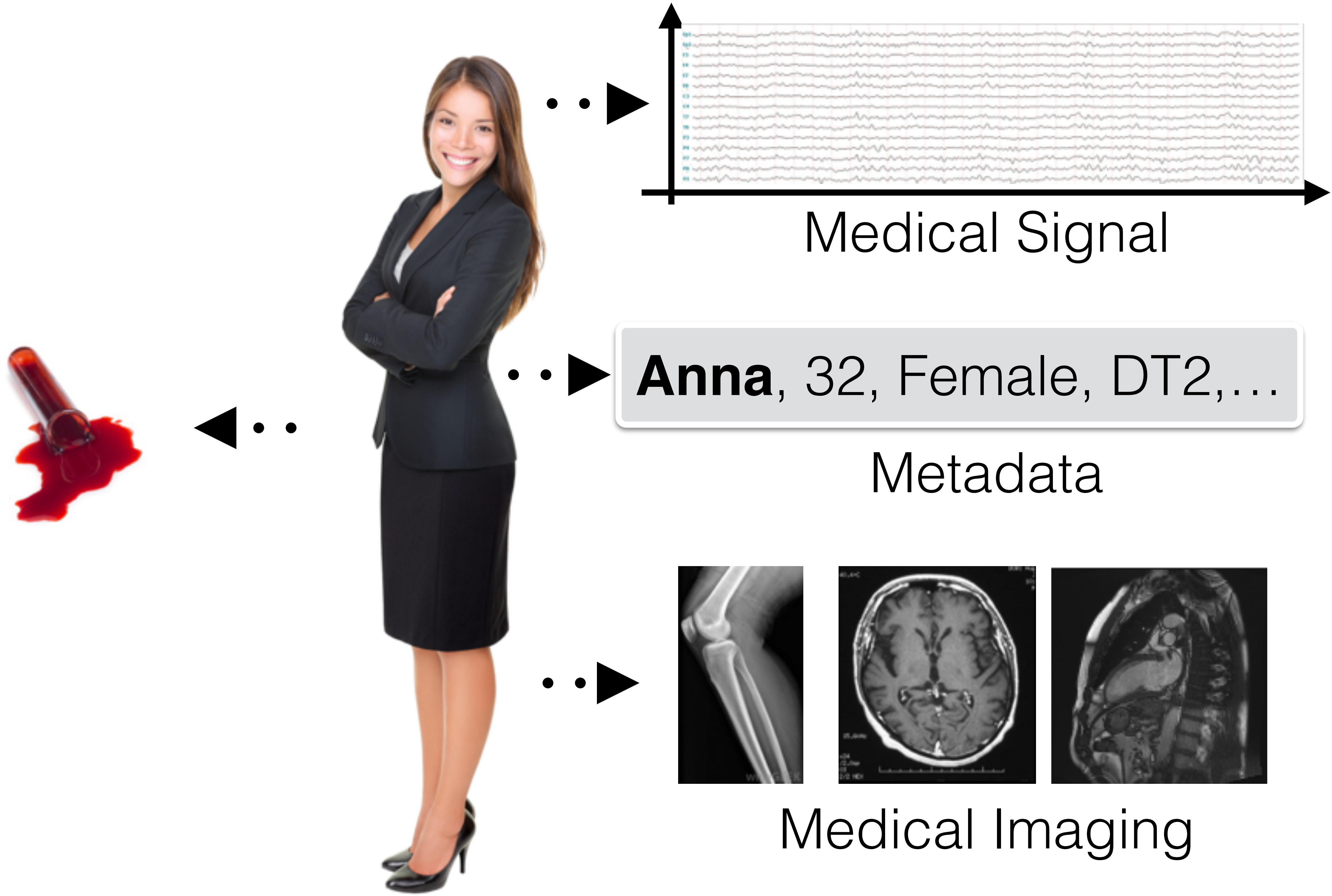
Machine Learning in Bioinformatics

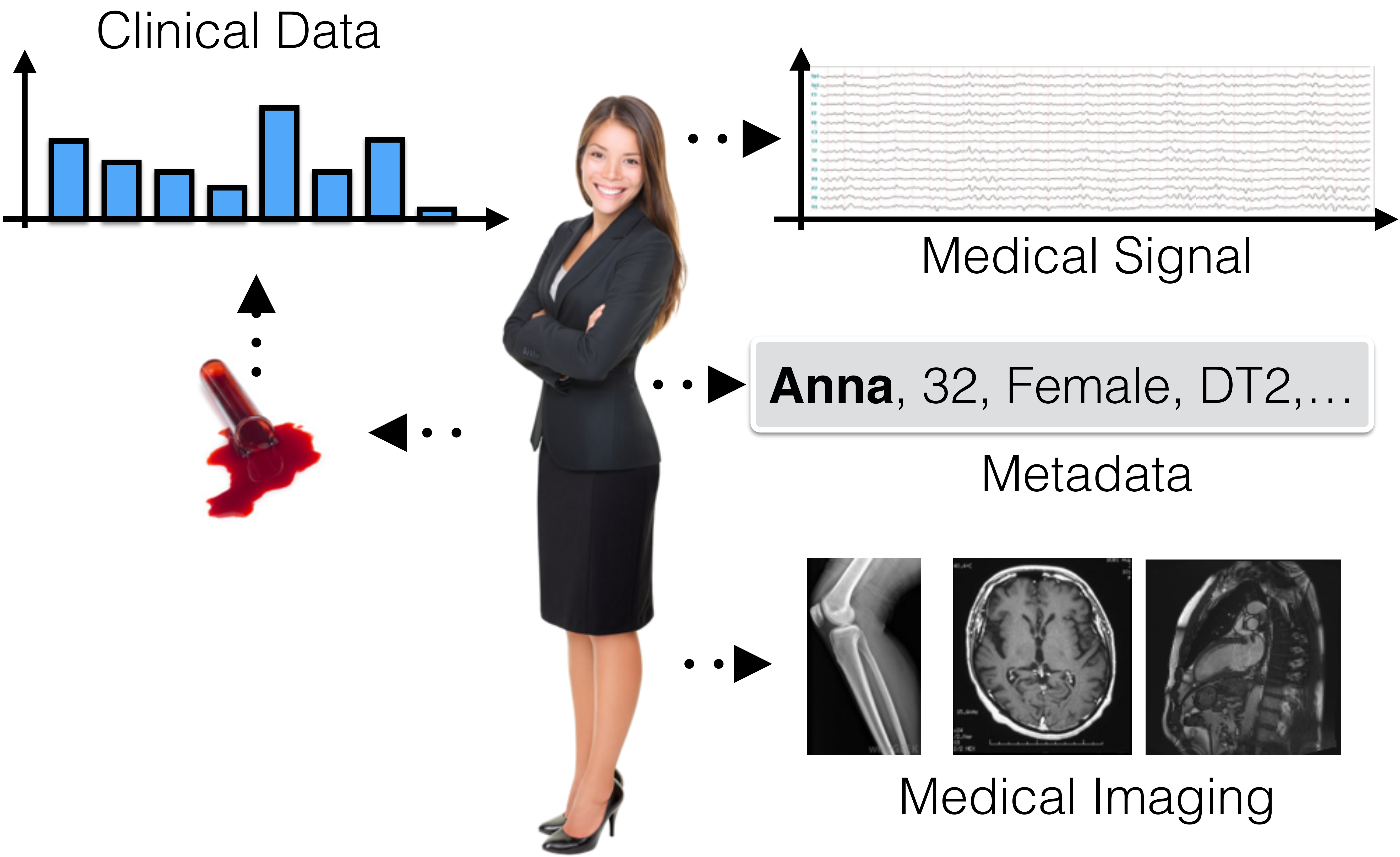
Anna, 32, Female, DT2,…

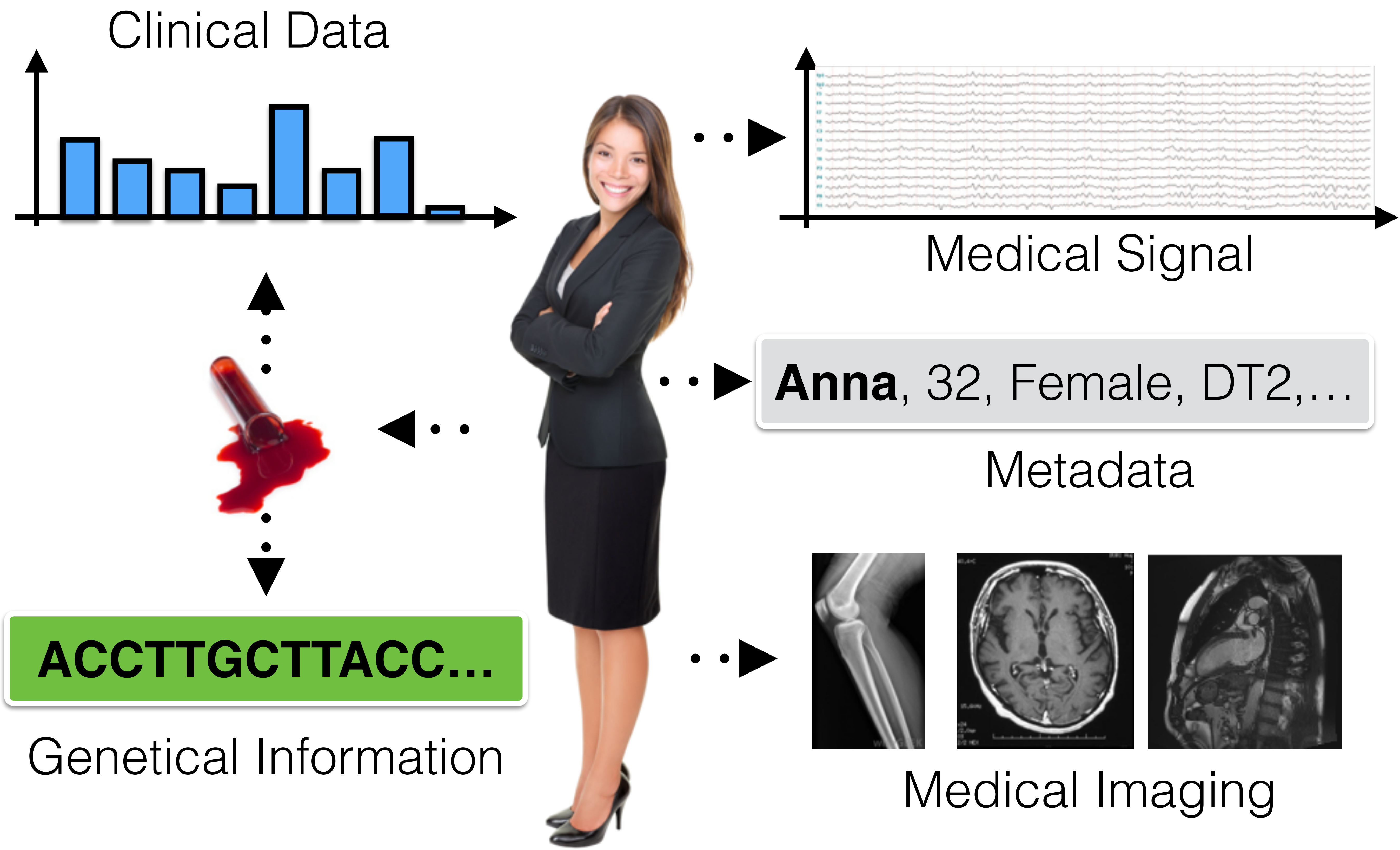
Metadata


Medical Signal
Anna, 32, Female, DT2,...
Metadata

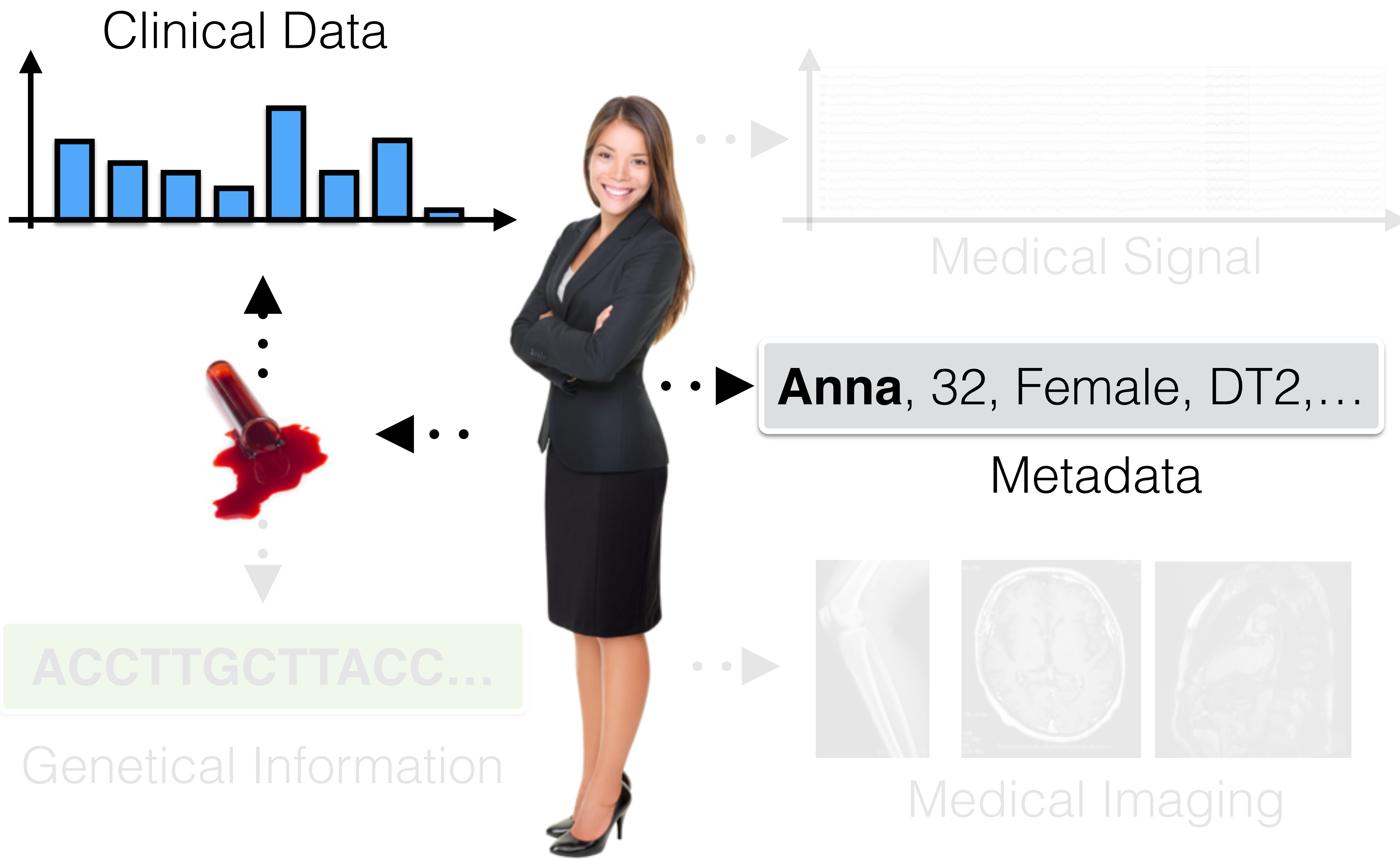

Medical Signal
Anna, 32, Female, DT2,…
Metadata
Medical Imaging


Medical Signal
Anna, 32, Female, DT2,...
Metadata
Medical Imaging

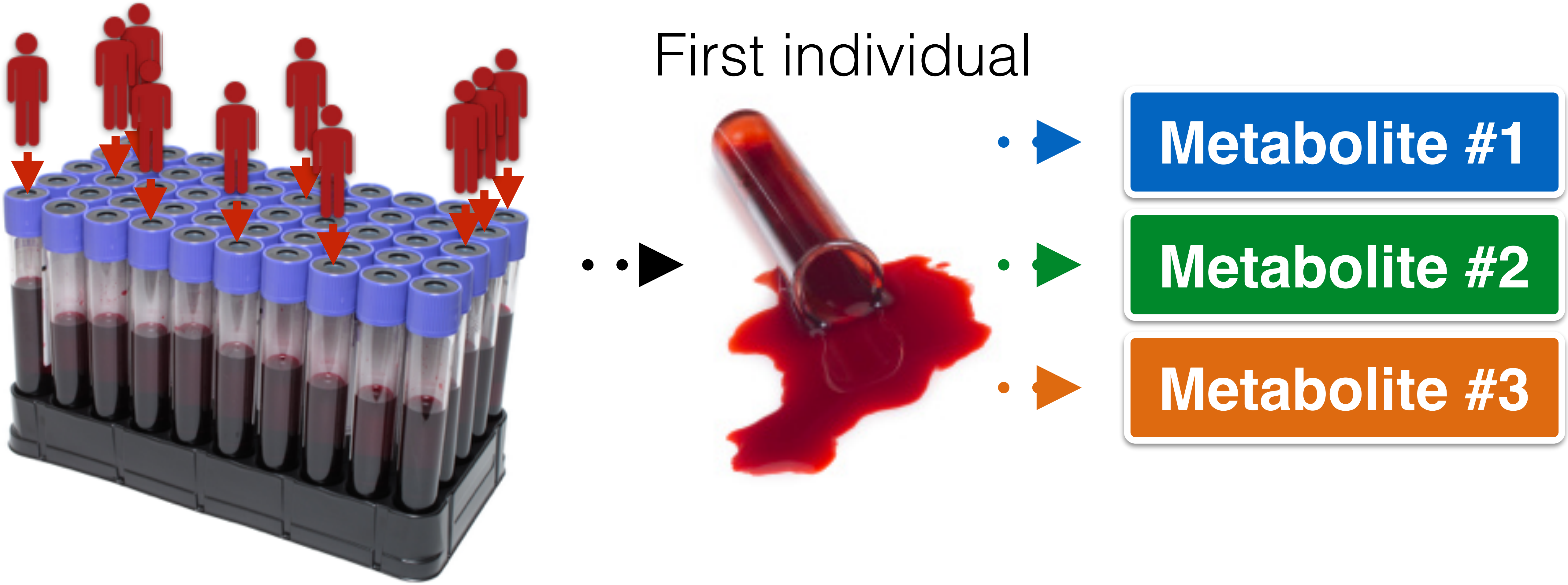

Clinical Data
Medical Signal
Anna, 32, Female, DT2,…
Metadata
Medical Imaging

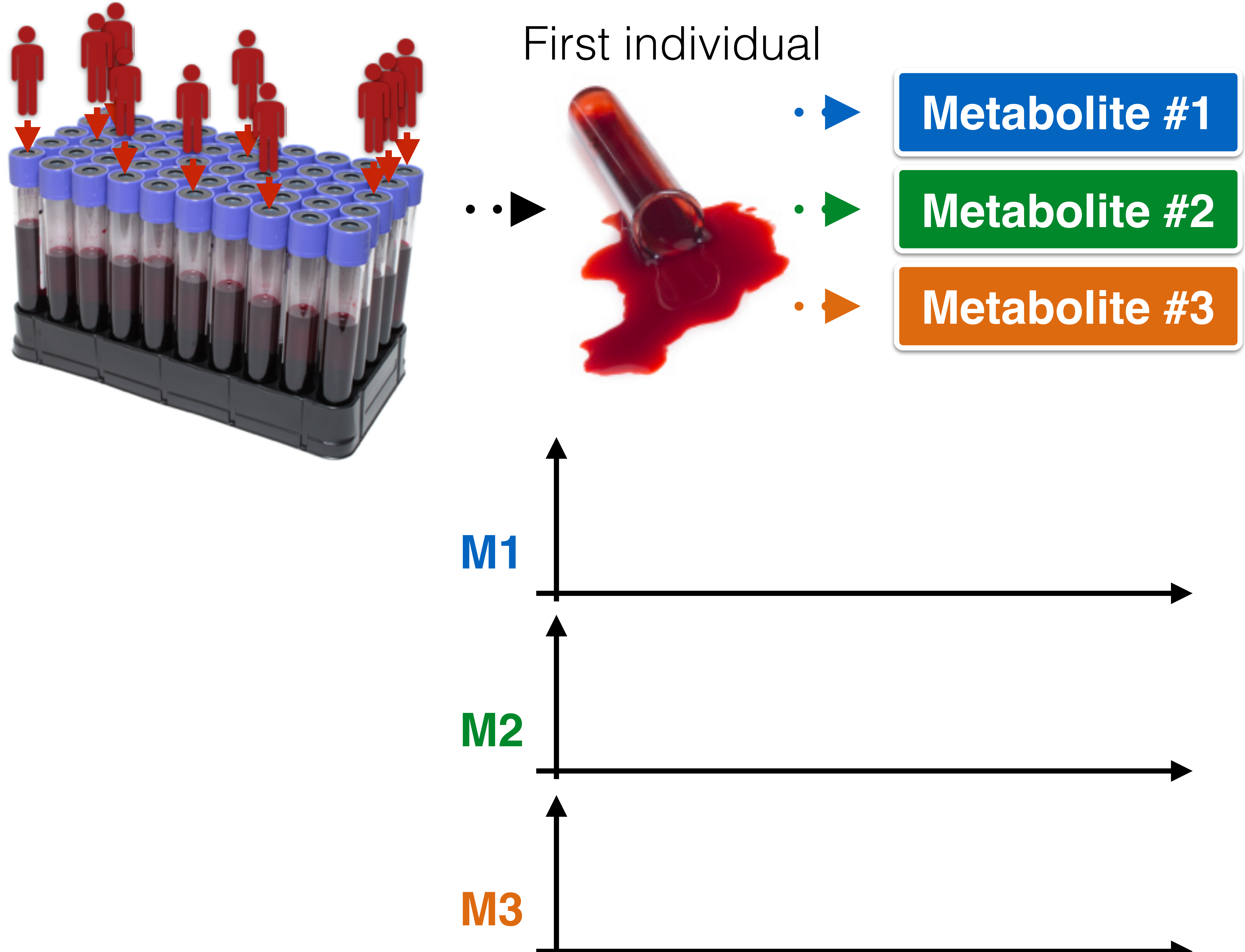


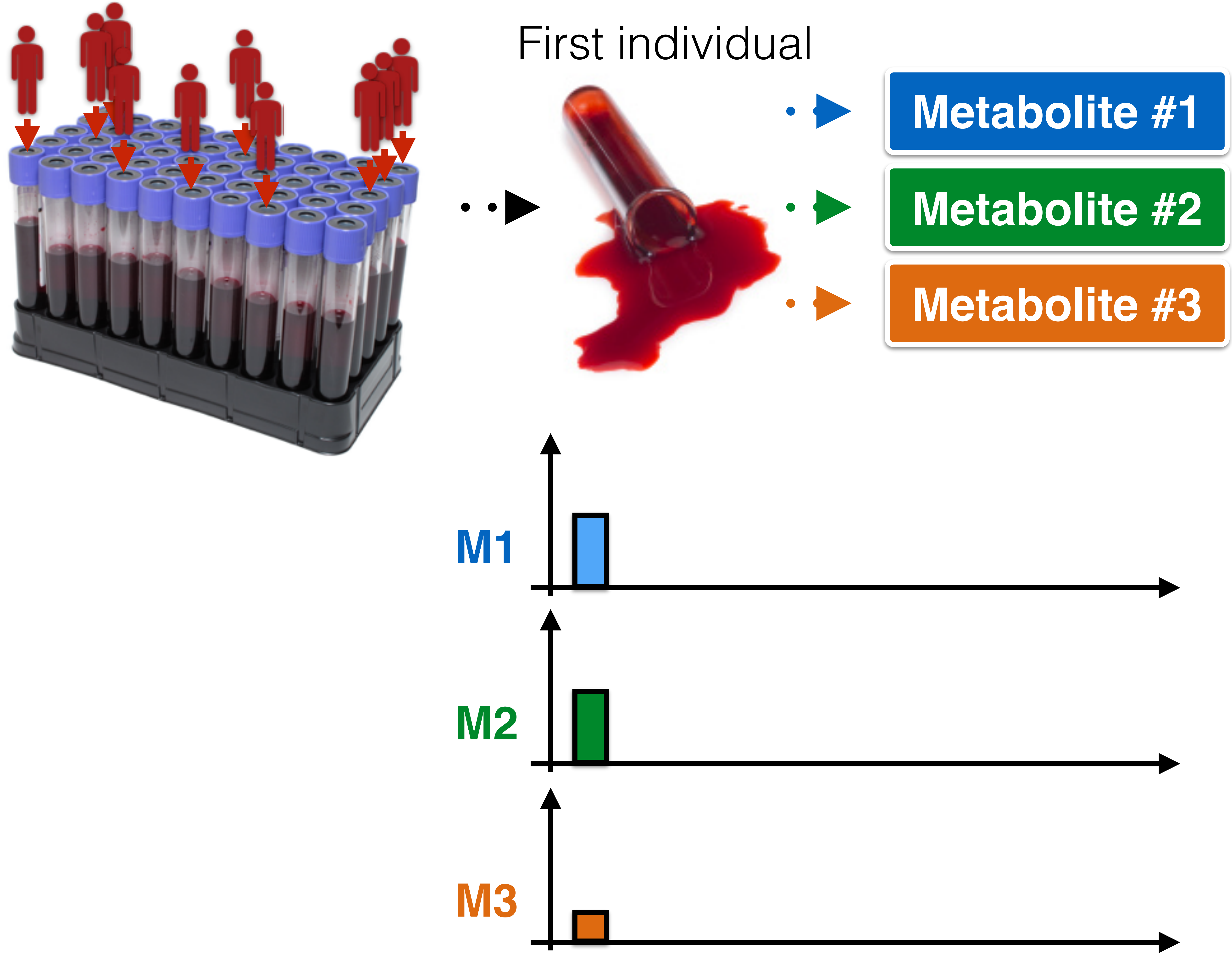
Clinical Data
Medical Signal
Anna, 32, Female, DT2,…
Metadata
ACCTTGCTTACC…
Genetical Information
Medical Imaging

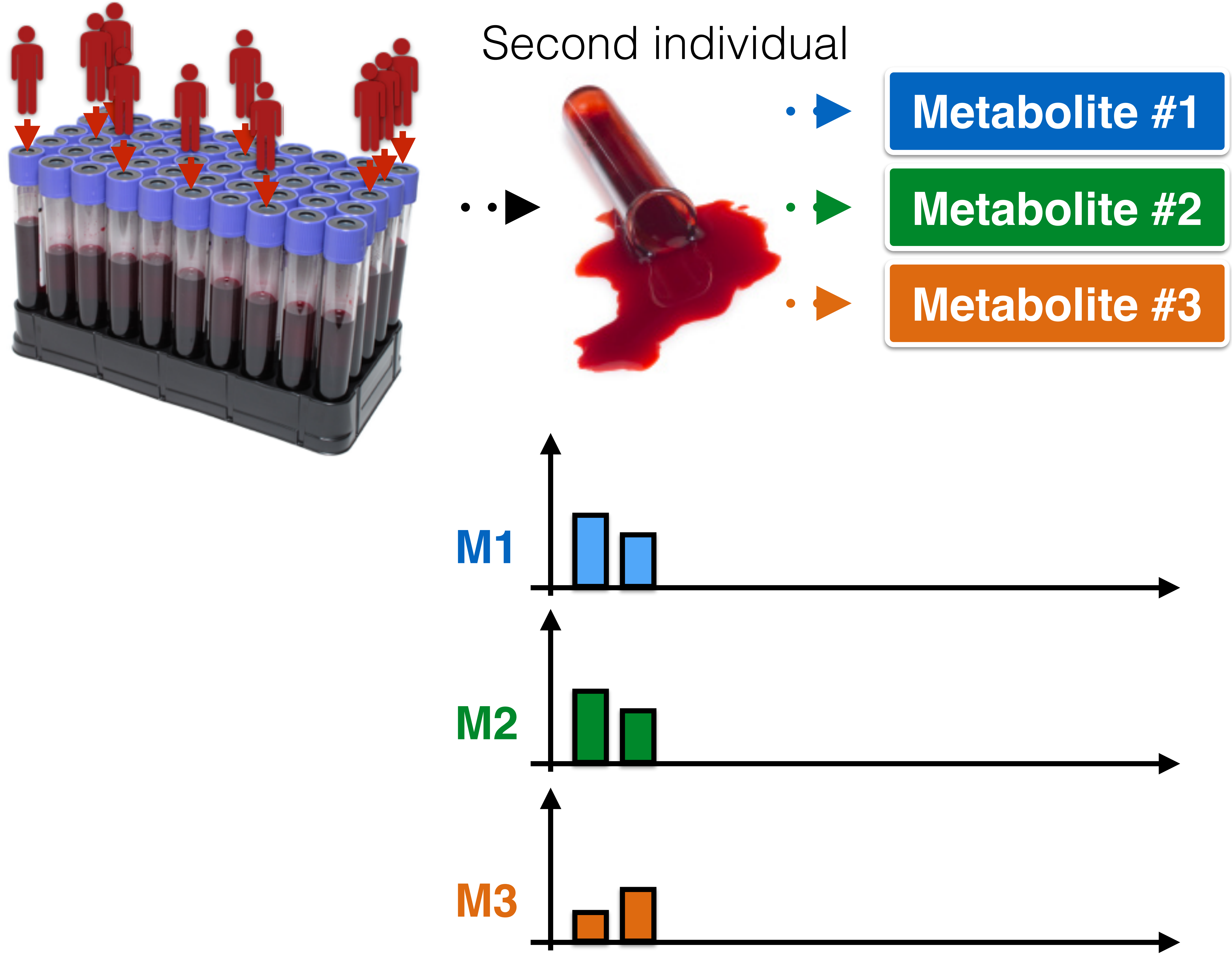

Clinical Data
Medical Signal
Anna, 32, Female, DT2,…
Metadata
ACCTTGCTTACC...
Genetical Information
Medical Imaging

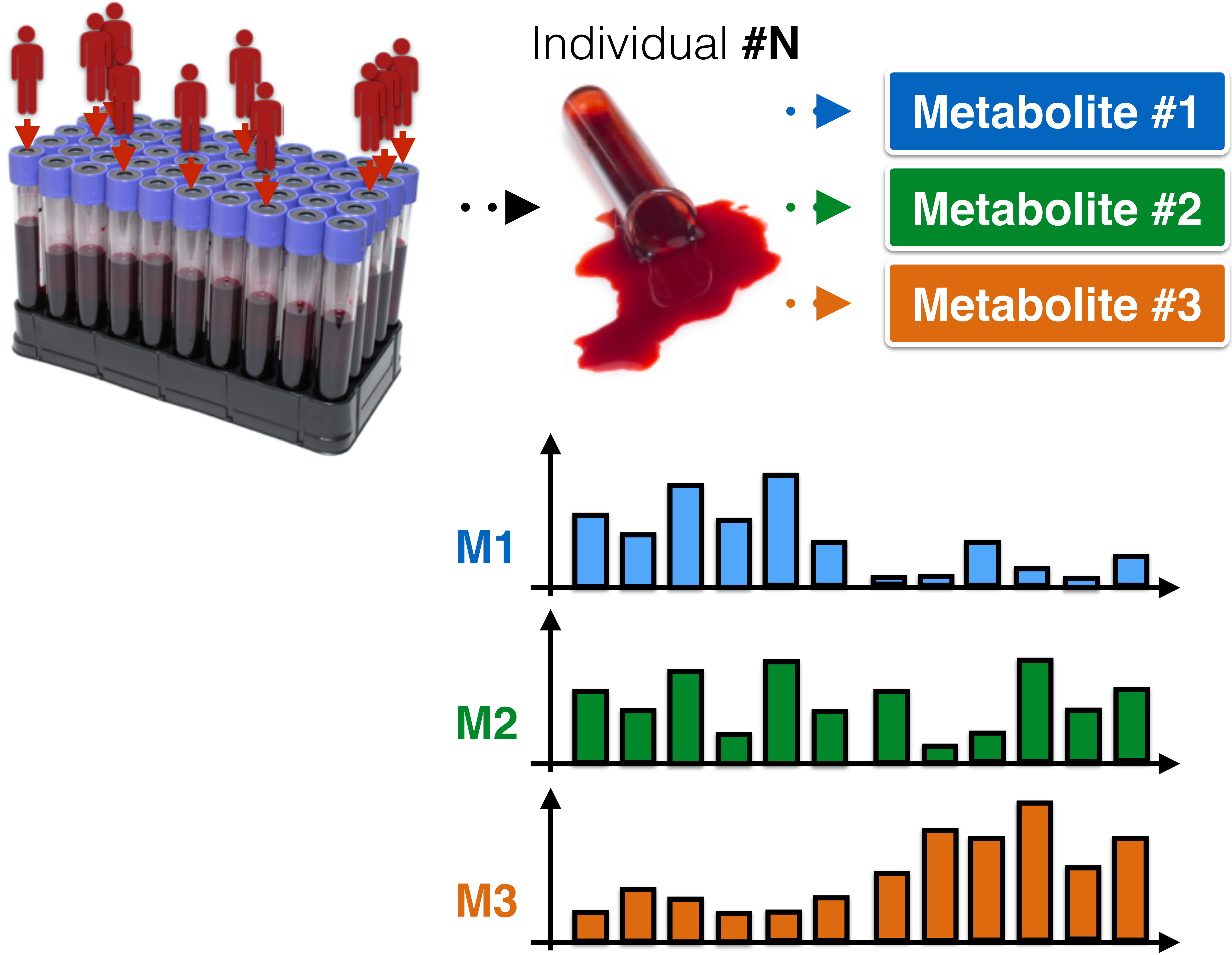
First individual



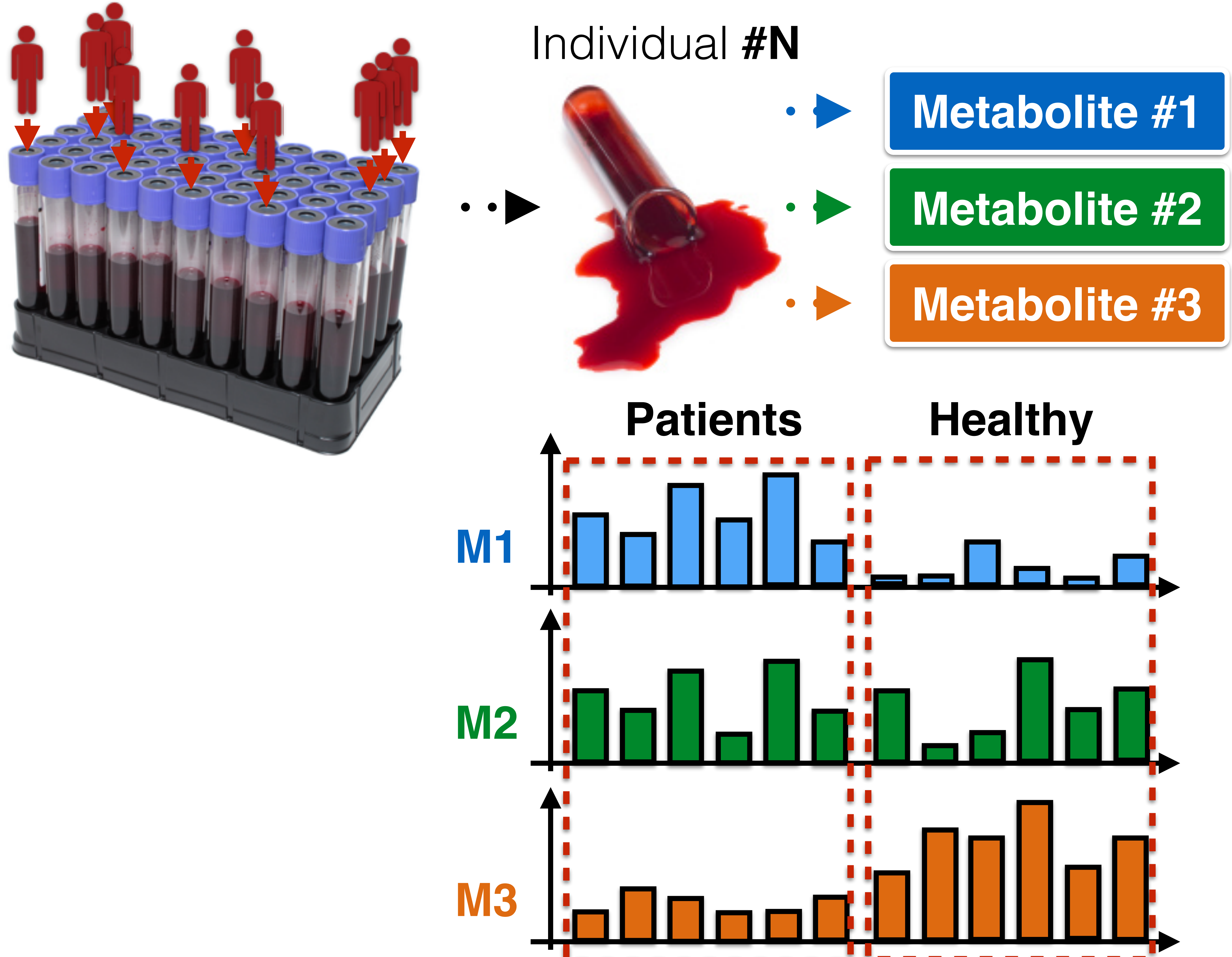

First individual
Metabolite #1
Metabolite #2
Metabolite #3
M1
M2
M3

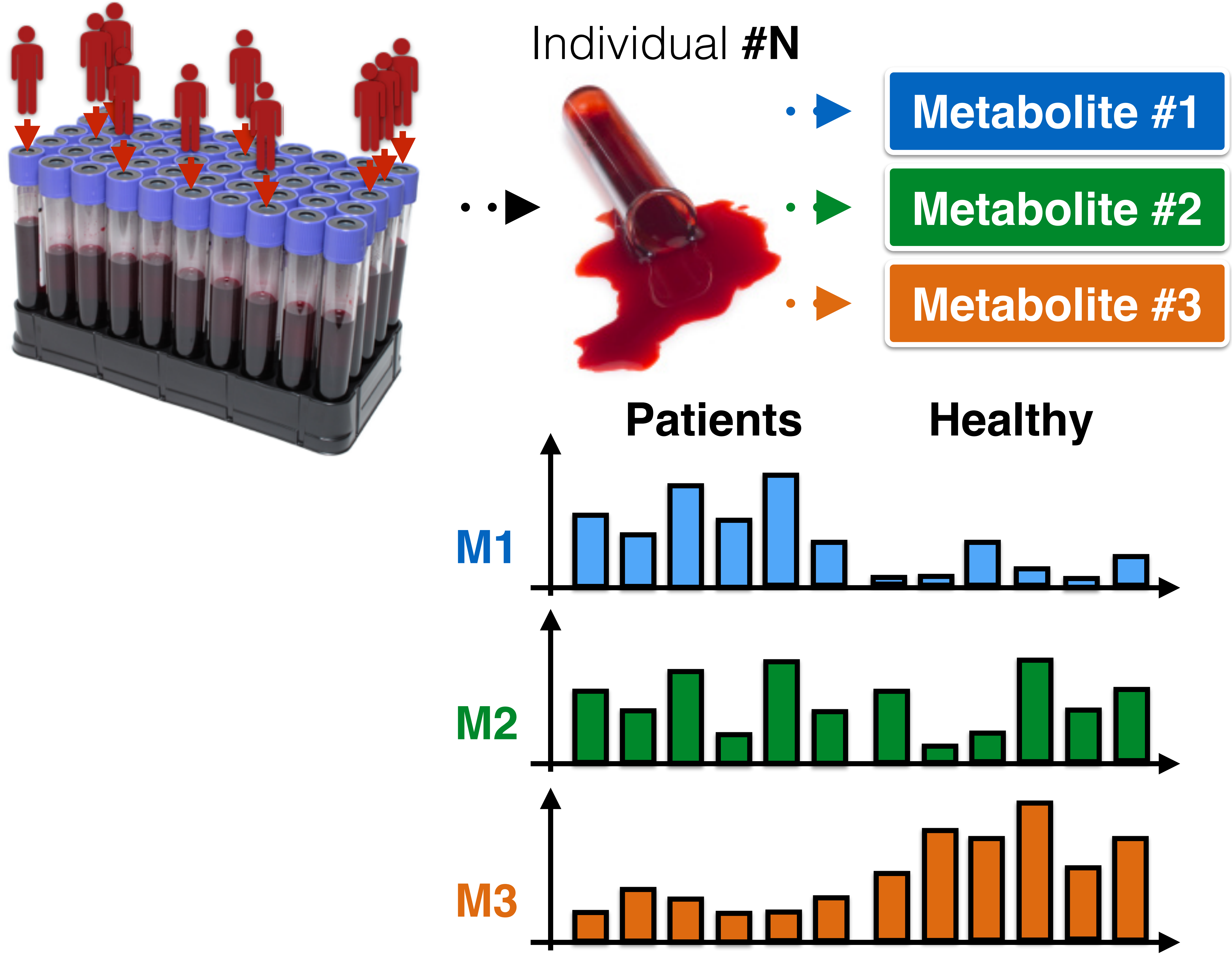

First individual
Metabolite #1
Metabolite #2
Metabolite #3
M1
M2
M3


Second individual
Metabolite #1
Metabolite #2
Metabolite #3
M1
M2
M3

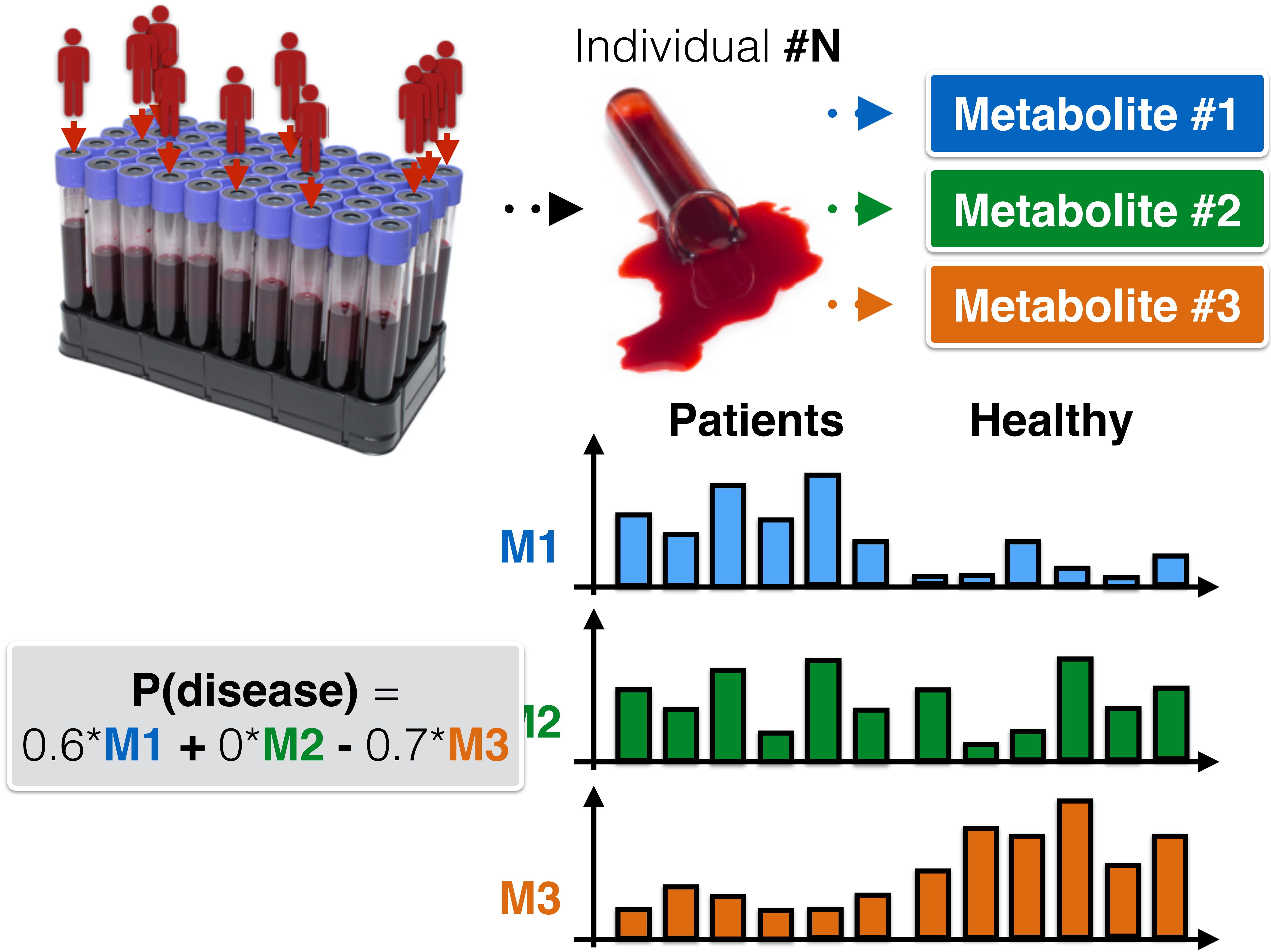


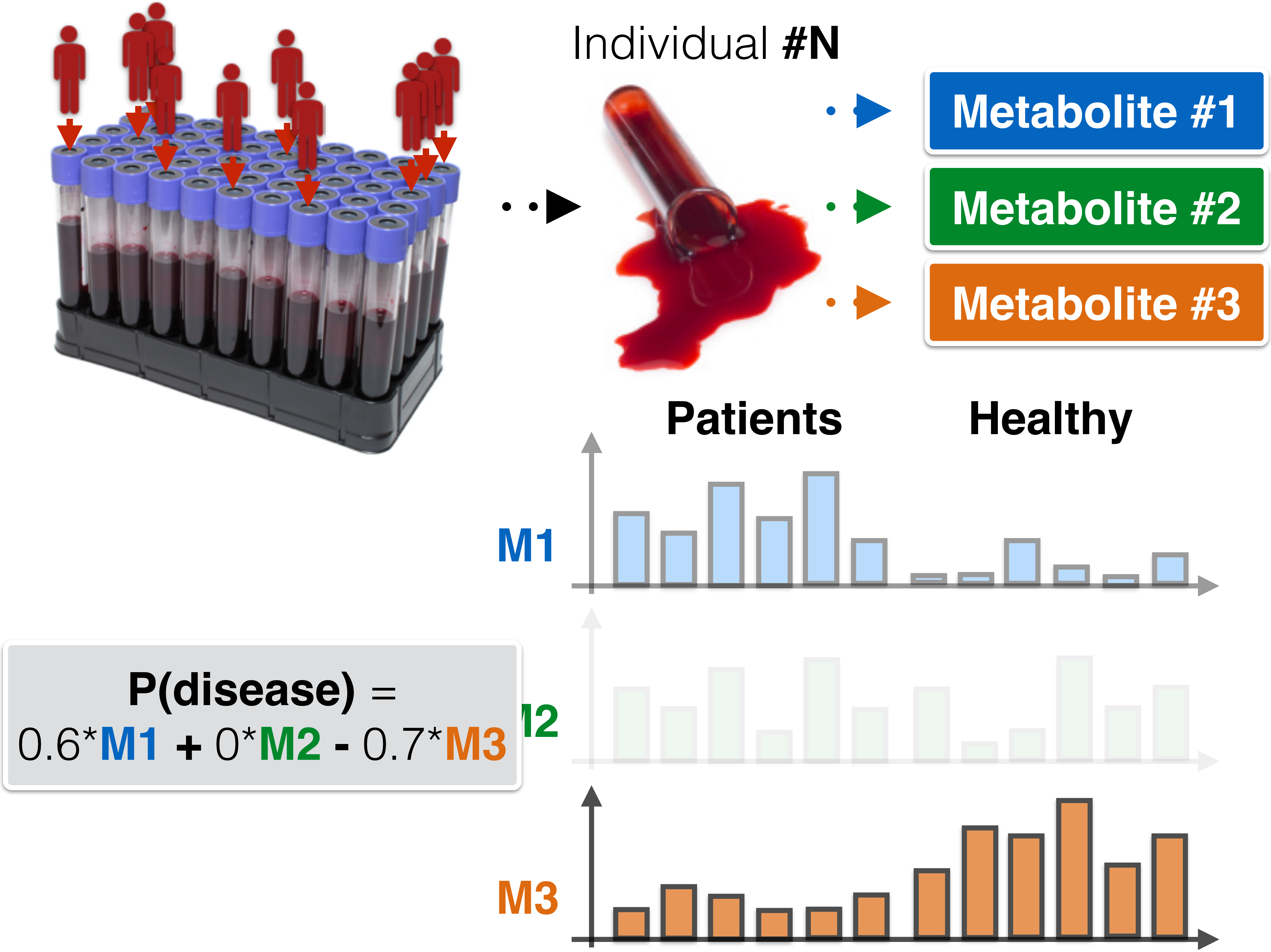
Individual #N
Metabolite #1
Metabolite #2
Metabolite #3
M1
M2
M3

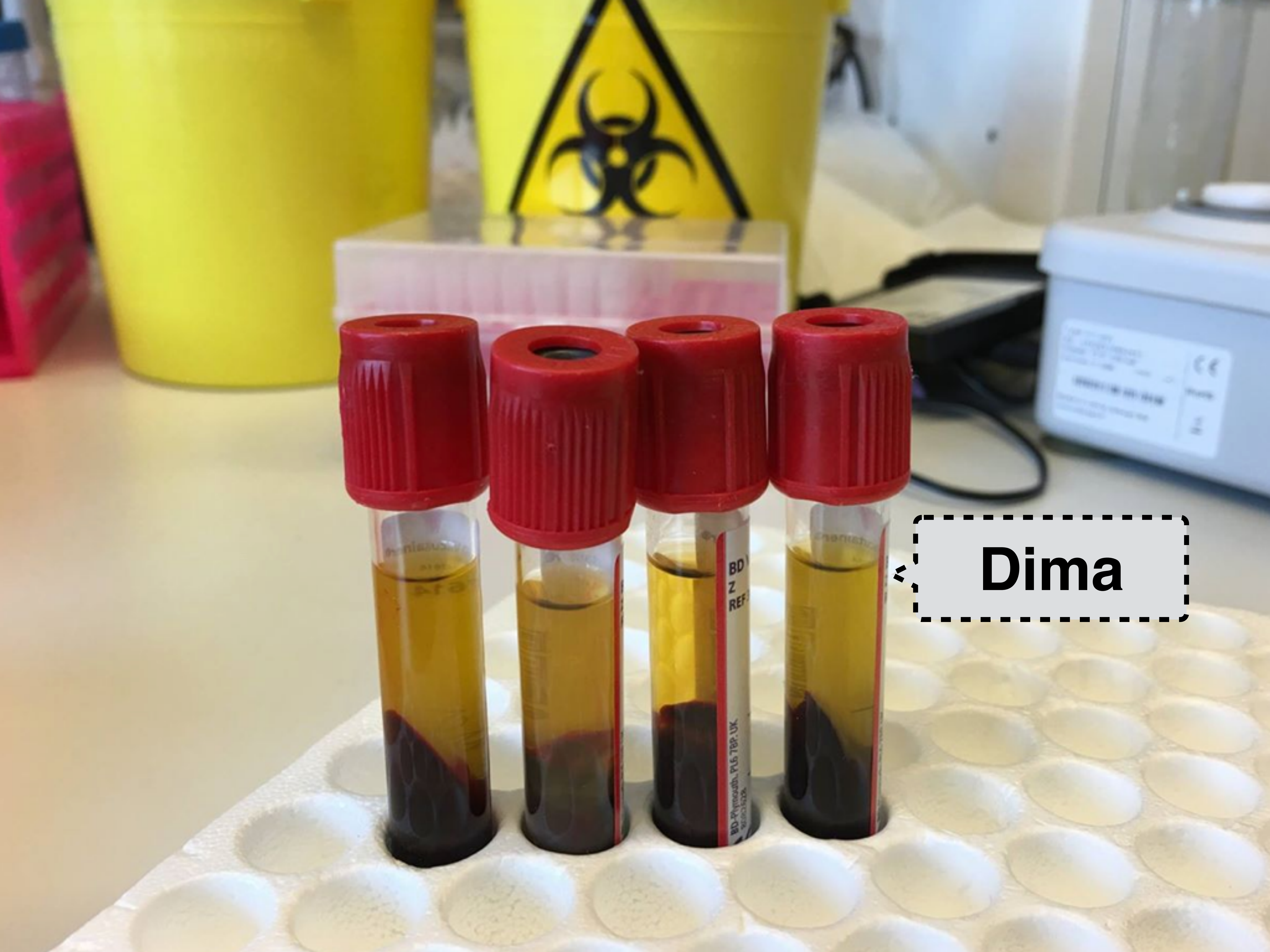

Individual #N
Metabolite #1
Metabolite #2
Metabolite #3
Patients
Healthy
M1
M2
M3

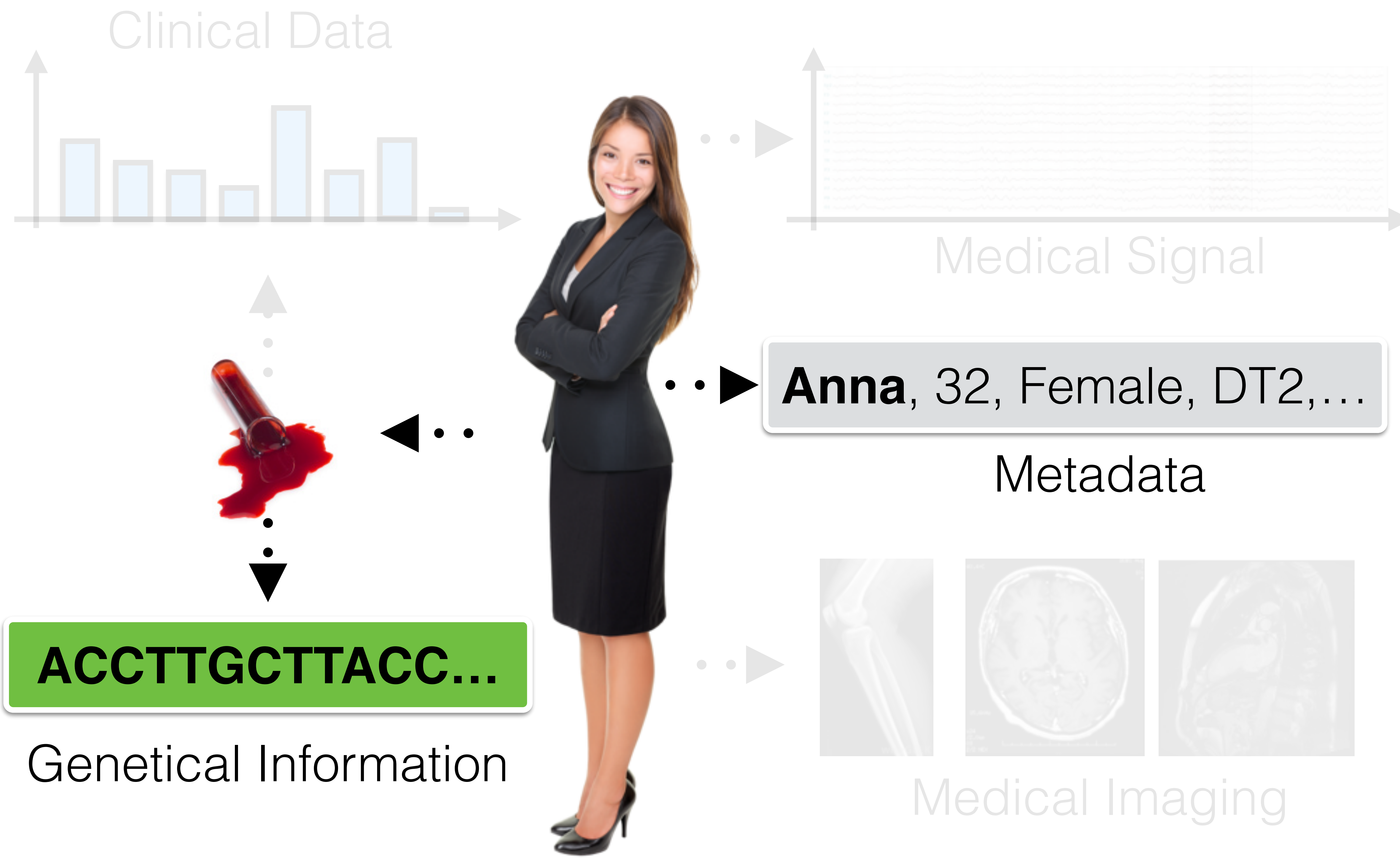


Individual #N
Metabolite #1
Metabolite #2
Metabolite #3
Patients
Healthy
M1
M2
M3


Individual #N
Metabolite #1
Metabolite #2
Metabolite #3
Patients
Healthy
M1
M3
P(disease) =
0.6*M1 + 0*M2 - 0.7*M3


Individual #N
Metabolite #1
Metabolite #2
Metabolite #3
Patients
Healthy
M1
M3
P(disease) =
0.6*M1 + 0*M2 - 0.7*M3


Dima


Clinical Data
Medical Signal
Anna, 32, Female, DT2,…
Metadata
ACCTTGCTTACC…
Genetical Information
Medical Imaging

…AACCTTTACATTAAACTGGGACCCGG… Individual **#1**

…AACCTTTACATTATACTGGGAGCCGG… Individual **#2**

…

…CTAAACCACATTATACTGGGACCCGG… Individual **#N**

...AACCTTTACATTAAACTGGGACCCGG... Individual **#1**

...AACCTTTACATTATACTGGGAGCCGG... Individual **#2**

...

...CTAAACCACATTATACTGGGACCCGG... Individual **#N**

Human genomes are 99.5% similar

...AACCTTTACATTAAACTGGGACCCGG... Individual **#1**

...AACCTTTACATTATACTGGGAGCCGG... Individual **#2**

...

...CTAAACCACATTATACTGGGACCCGG... Individual **#N**

Mutations

...AACCTTTACATTAAACTGGGACCCGG... Individual **#1**

...AACCTTTACATTATACTGGGAGCCGG... Individual **#2**

...

...CTAAACCACATTATACTGGGACCCGG... Individual **#N**

Deletions

...AACCTTTACATTAAACTGGGACCCGG... Individual **#1**

...AACCTTTACATTATACTGGGAGCCGG... Individual **#2**

...

...CTAAACCACATTATACTGGGACCCGG... Individual **#N**

Insertions

...AACCTTTACATTAAACTGGGACCCGG... Individual **#1**

...AACCTTTACATTATACTGGGAGCCGG... Individual **#2**

...

...CTAAACCACATTATACTGGGACCCGG... Individual **#N**

Compute pairwise distance

...AACCTTTACATTAAACTGGGACCCGG... Individual **#1**

...AACCTTTACATTATACTGGGAGCCGG... Individual **#2**

...

...CTAAACCACATTATACTGGGACCCGG... Individual **#N**

Compute pairwise distance

	#1	**#2**	**...**	**#N**
#1	0	29	…	34
#2	29	0	…	56
...	…	…	…	…
#N	34	56	…	0

Distance Matrix

...AACCTTTACATTAAACTGGGACCCGG... Individual **#1**

...AACCTTTACATTATACTGGGAGCCGG... Individual **#2**

...

...CTAAACCACATTATACTGGGACCCGG... Individual **#N**

	#1	**#2**	**...**	**#N**
#1	0	29	...	34
#2	29	0	...	56
...	...	...	...	...
#N	34	56	...	0

Distance Matrix

...AACCTTTACATTAAACTGGGACCCGG... Individual **#1**

...AACCTTTACATTATACTGGGAGCCGG... Individual **#2**

...

...CTAAACCACATTATACTGGGACCCGG... Individual **#N**

	#1	**#2**	**...**	**#N**
#1	0	29	...	34
#2	29	0	...	56
...	...	...	...	...
#N	34	56	...	0

Distance Matrix

PCA

Principal component analysis of the combined autosomal genotypic data of individuals from Russia and seven European countries (Finnland, Estonia, Latvia, Poland, Czech Republic, Germany [5] and Italia [22]). Khrunin et al.

...AACCTTTACATTAAACTGGGACCCGG... Individual **#1**

...AACCTTTACATTATACTGGGAGCCGG... Individual **#2**

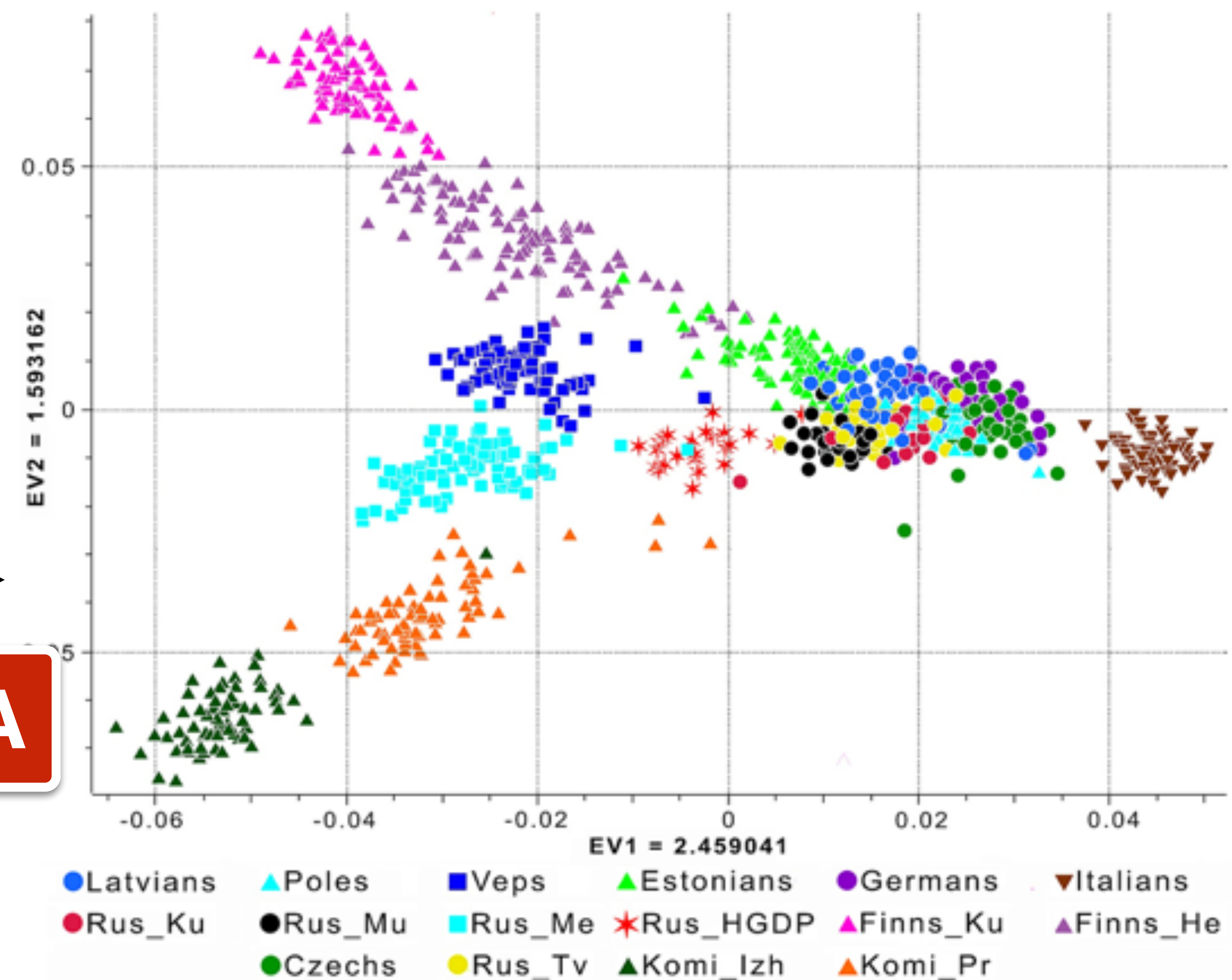
...

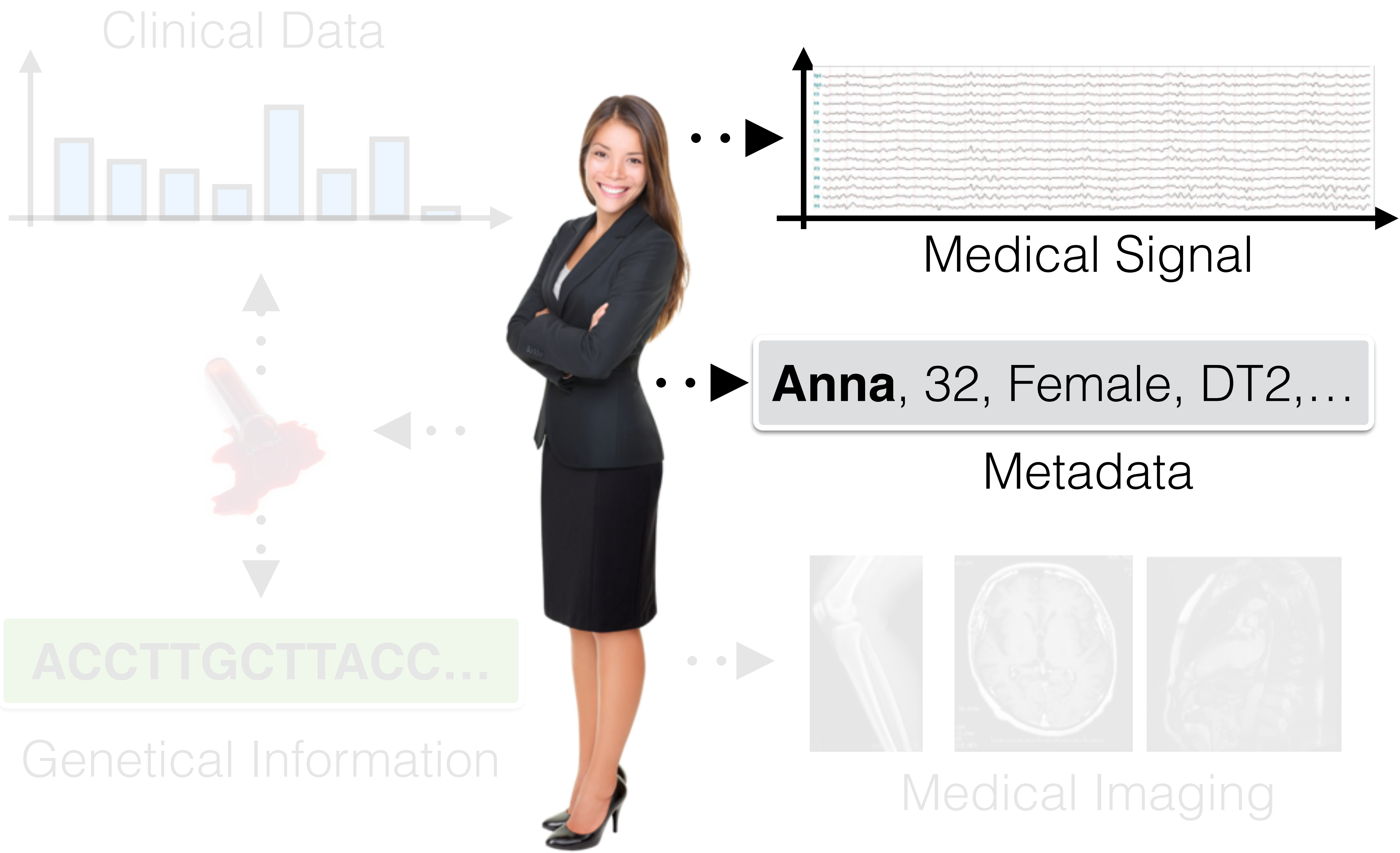
...CTAAACCACATTATACTGGGACCCGG... Individual **#N**

	#1	**#2**	**...**	**#N**
#1	0	29	...	34
#2	29	0	...	56
...	...	...	...	...
#N	34	56	...	0

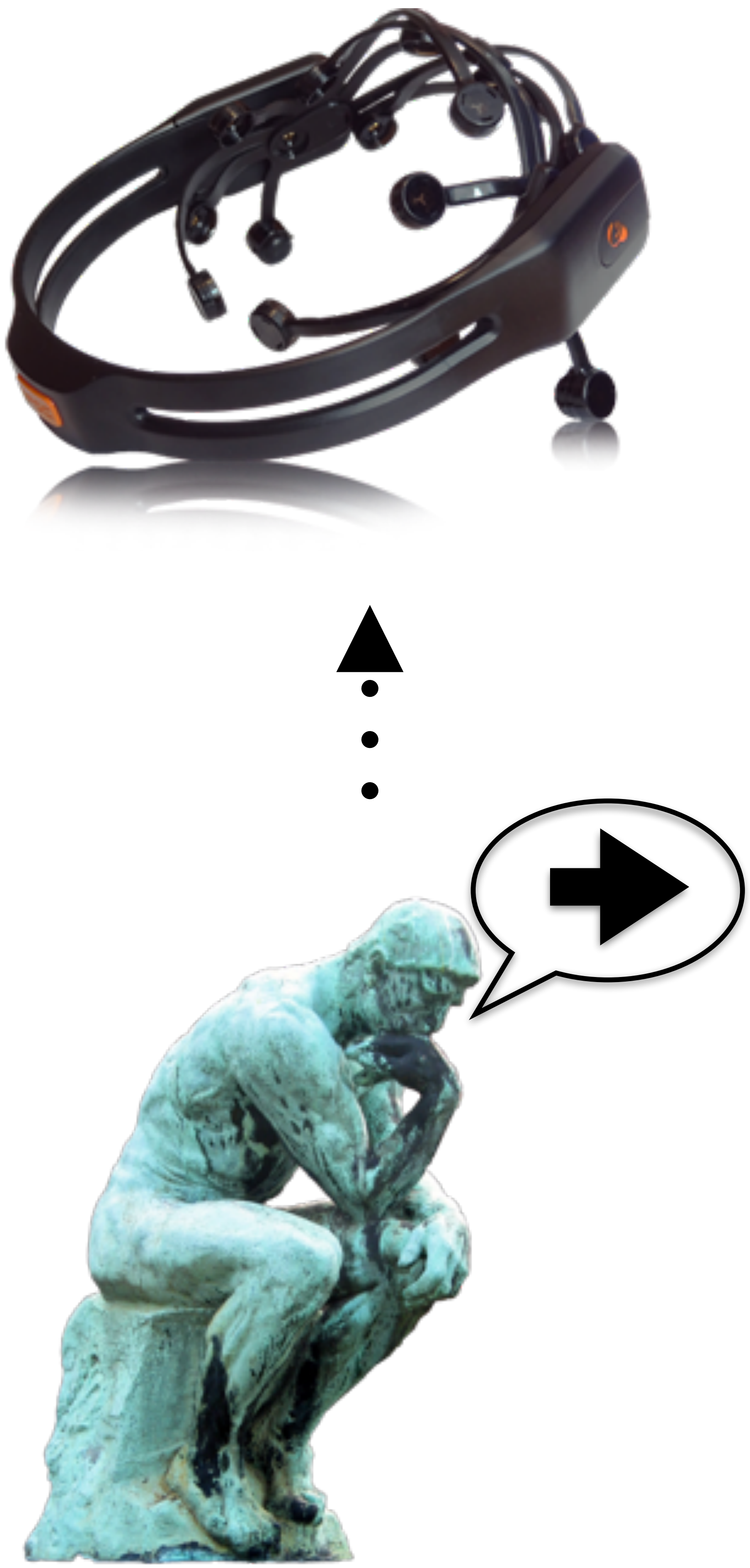
Distance Matrix

PCA



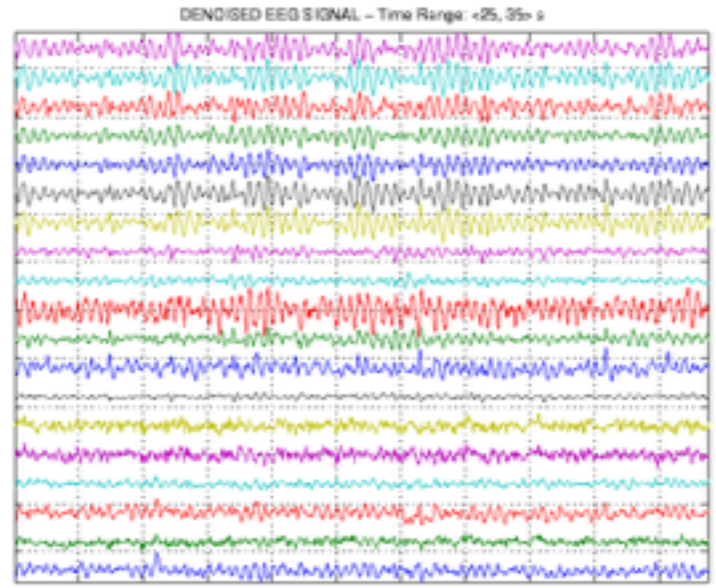

Clinical Data
Medical Signal
Anna, 32, Female, DT2,...
Metadata
ACCTTGCTTACC...
Genetical Information
Medical Imaging

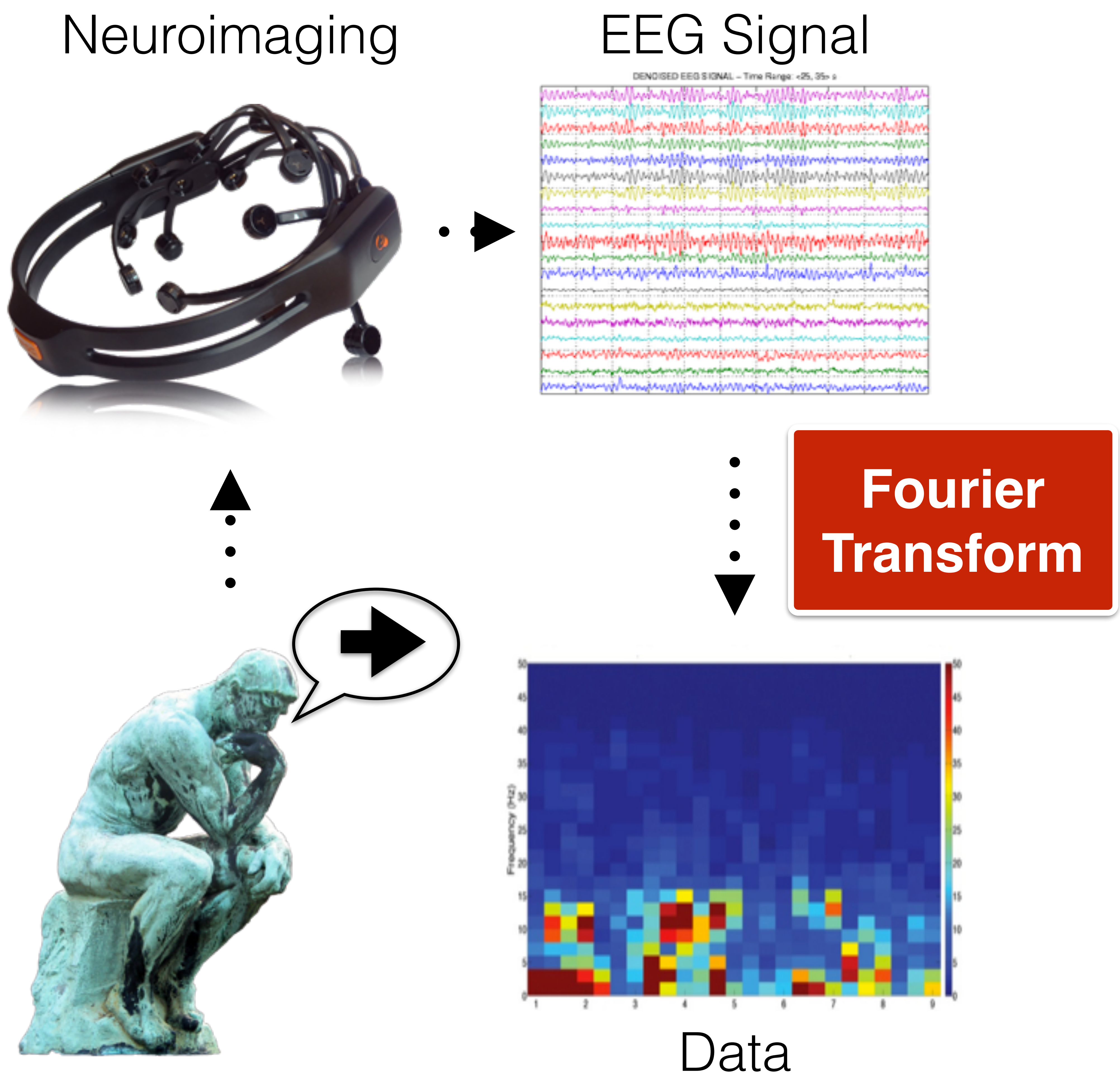
Neuroimaging



Neuroimaging

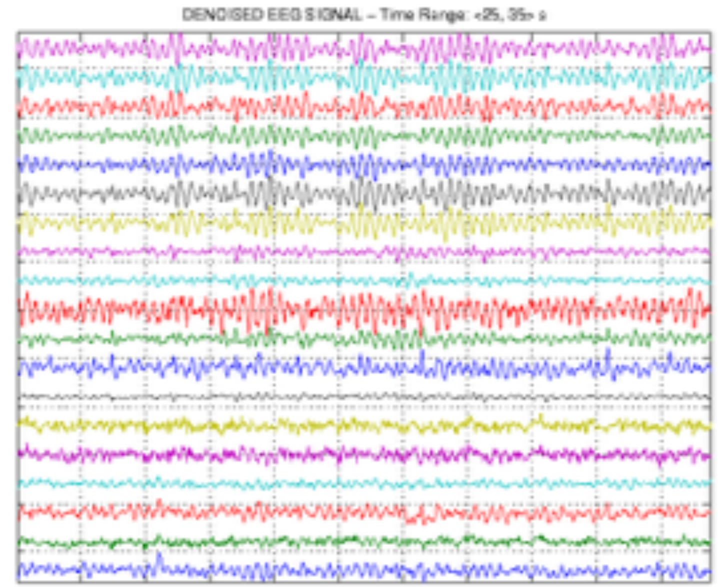
EEG Signal

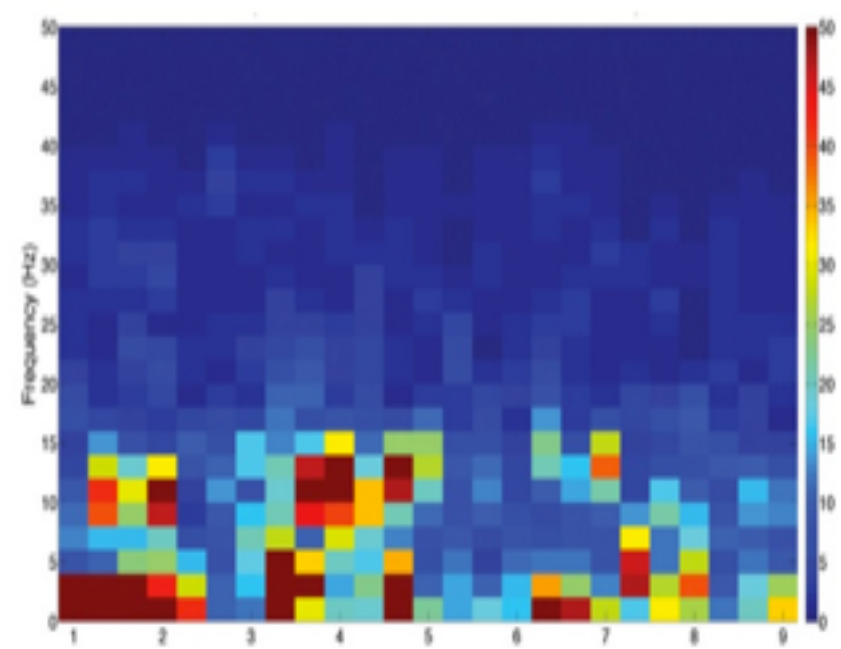



Neuroimaging
EEG Signal
Fourier Transform
Frequency (Hz)
Data

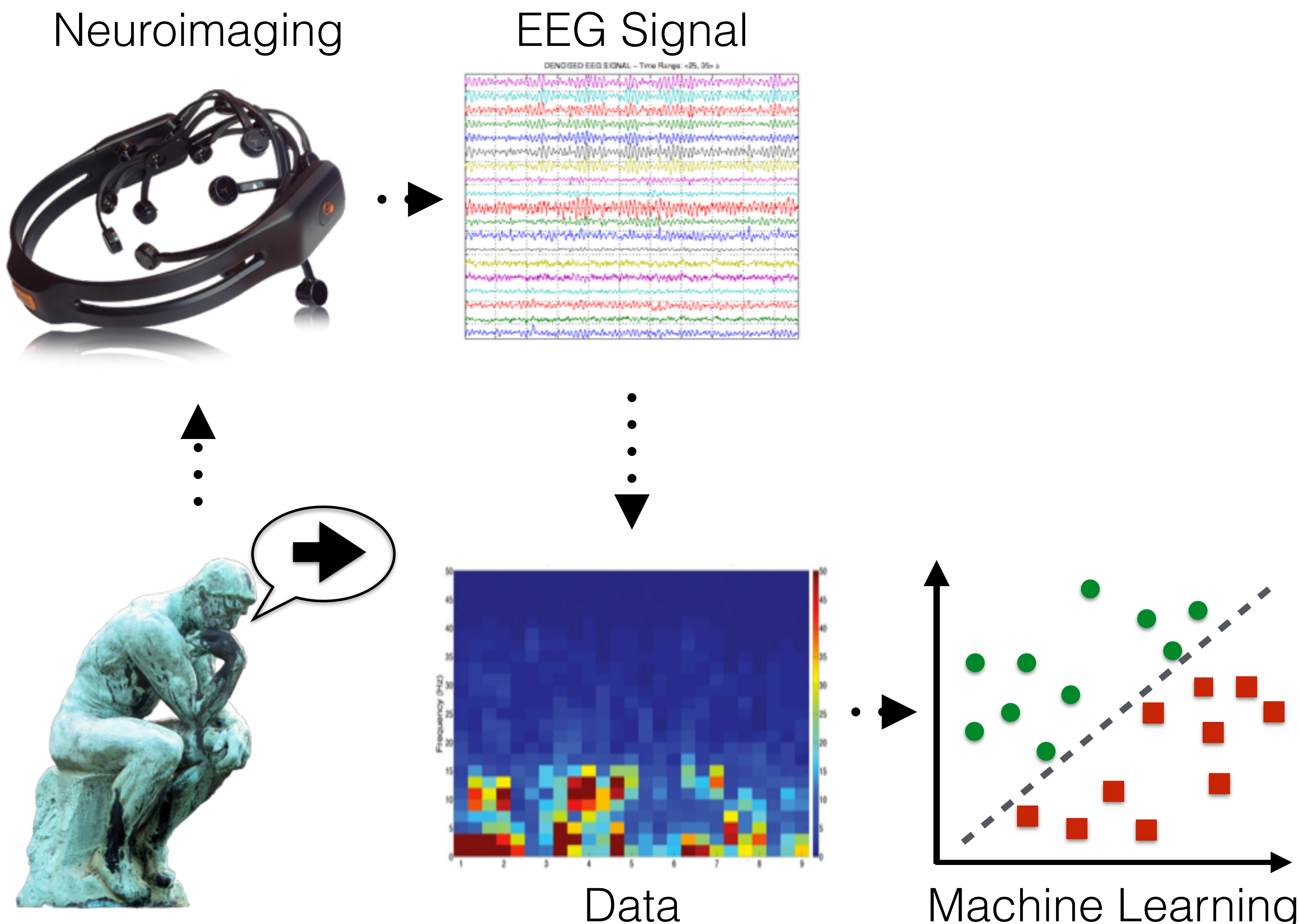
Neuroimaging

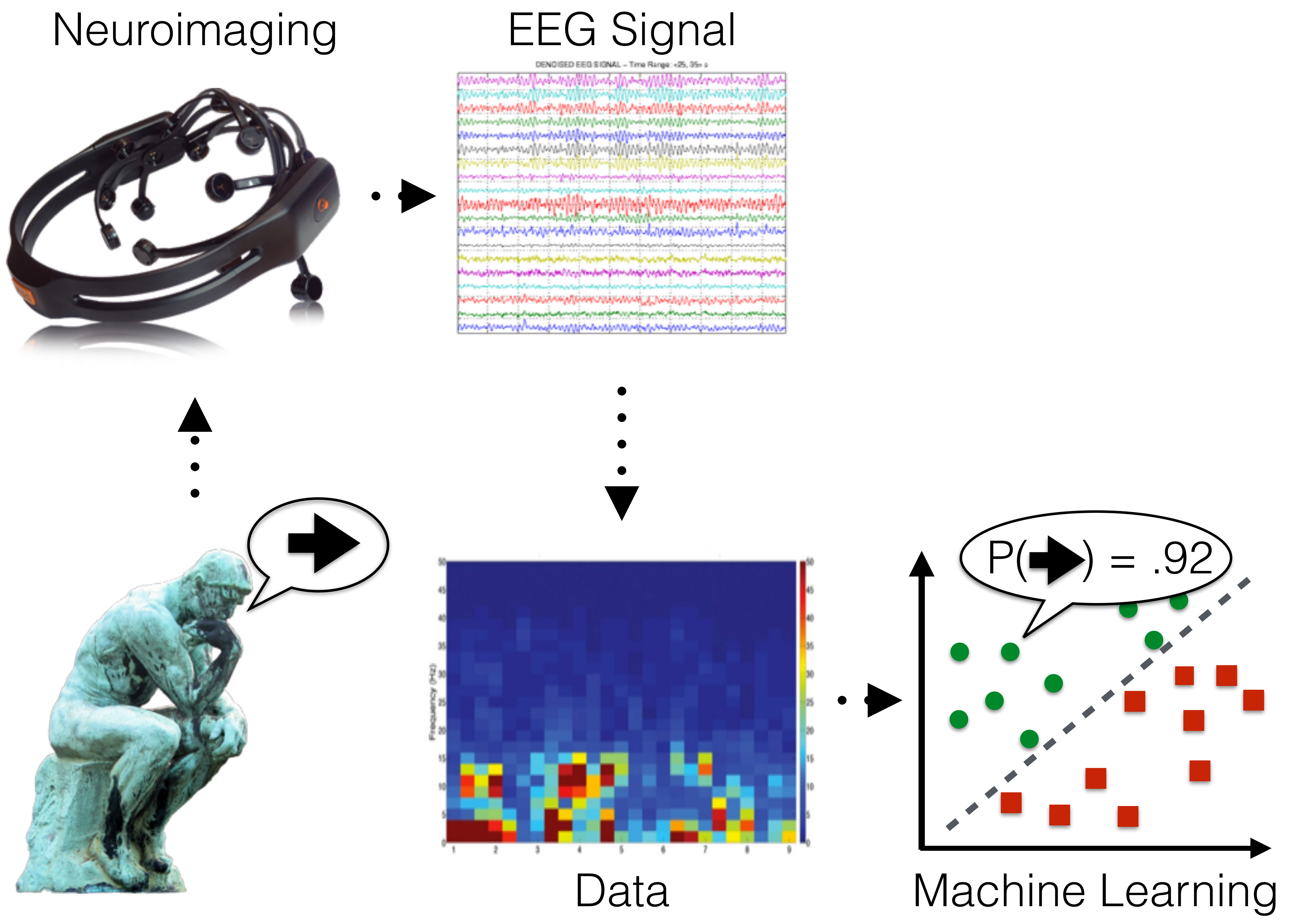
EEG Signal

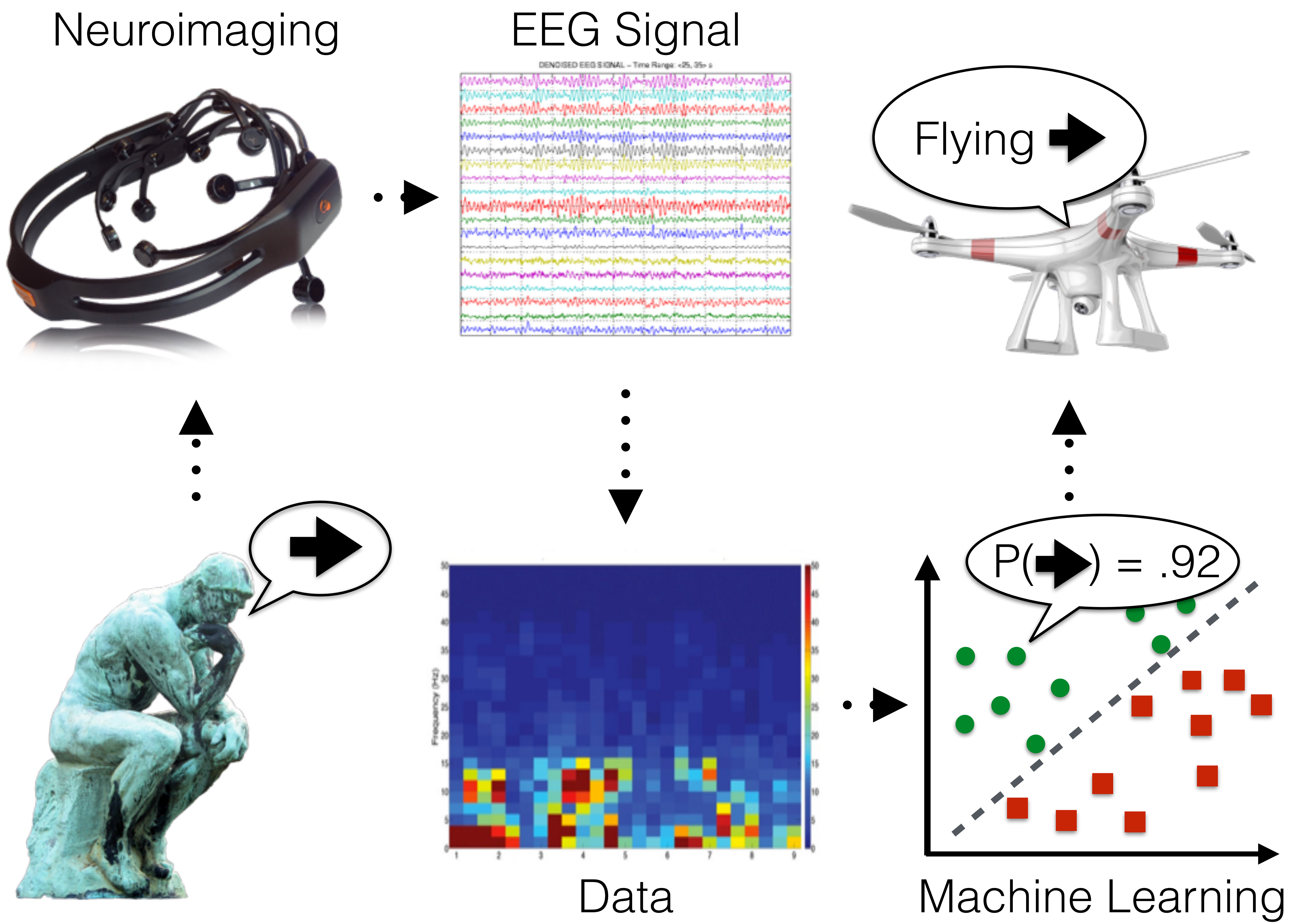




Data

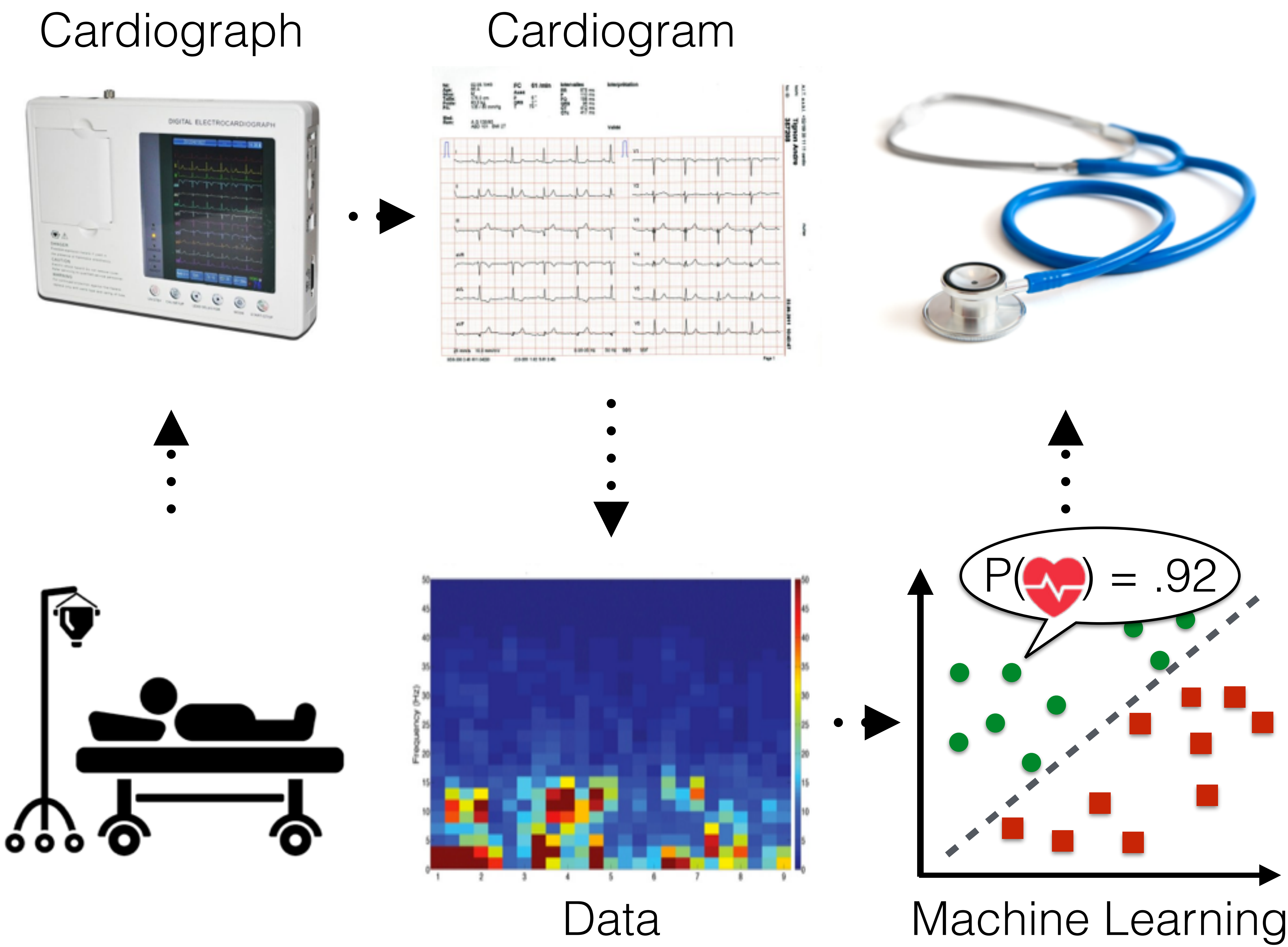

Neuroimaging
EEG Signal
Data
Machine Learning

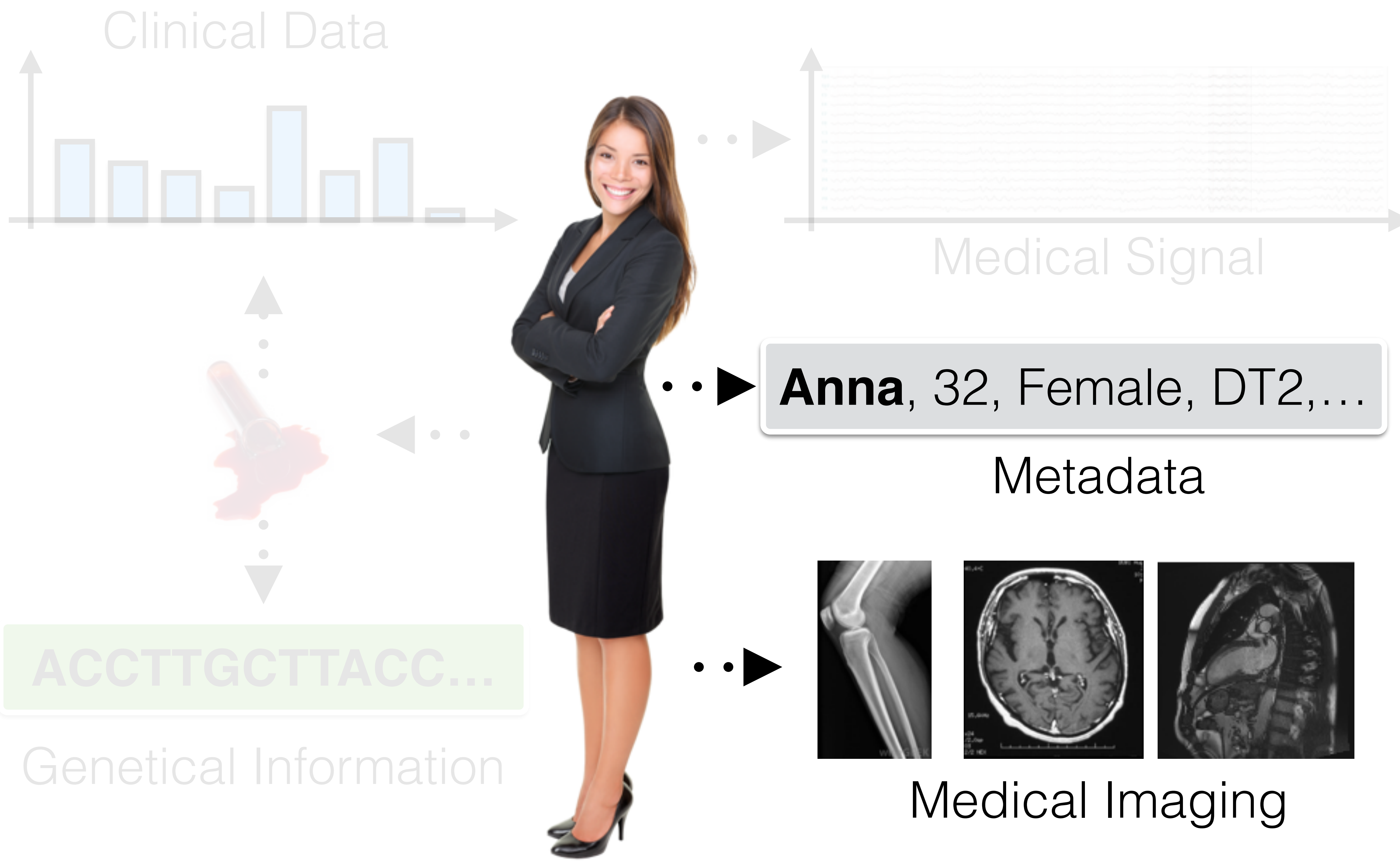

Neuroimaging
EEG Signal
P(➡) = .92
Data
Machine Learning

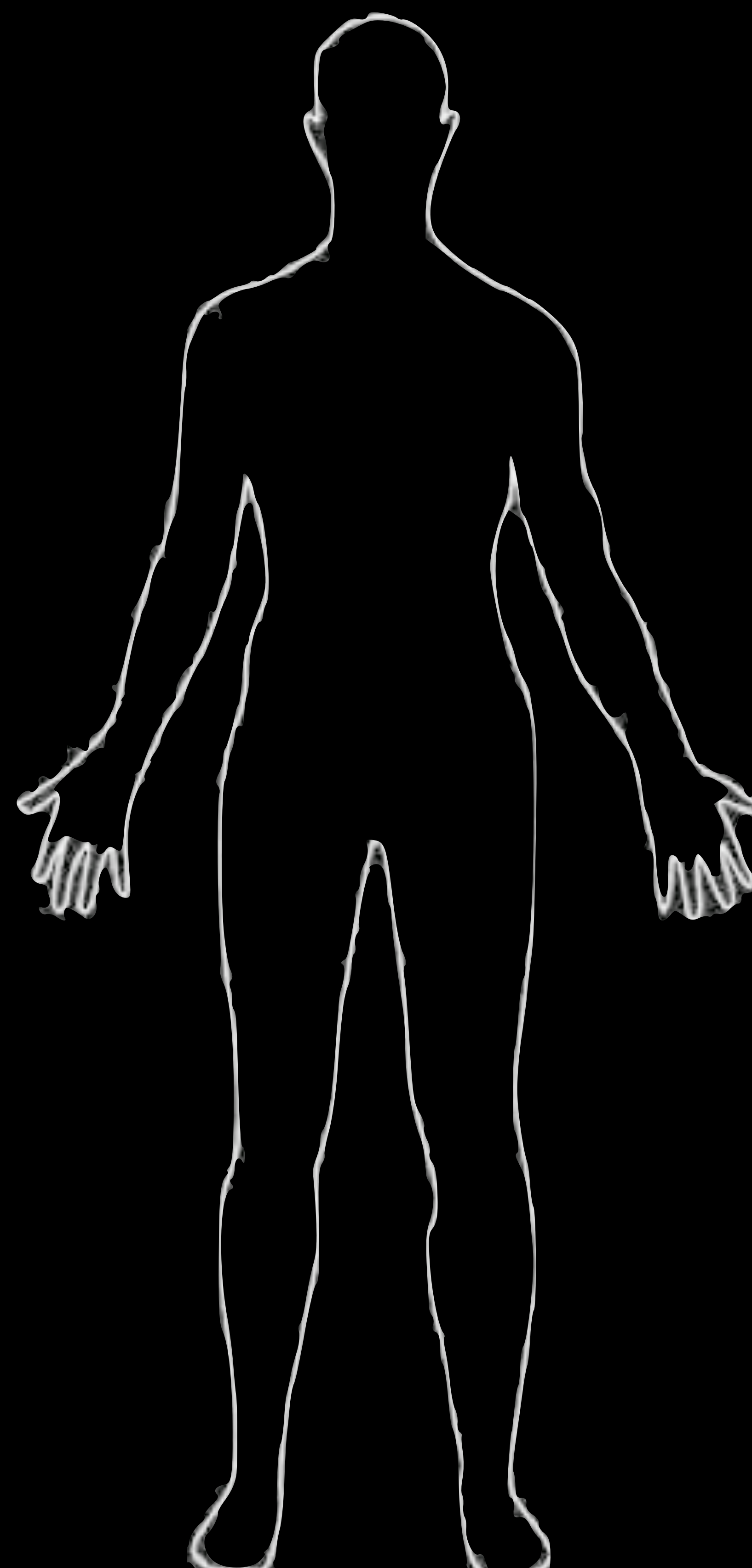

Neuroimaging
EEG Signal
Flying
P( ) = .92
Data
Machine Learning

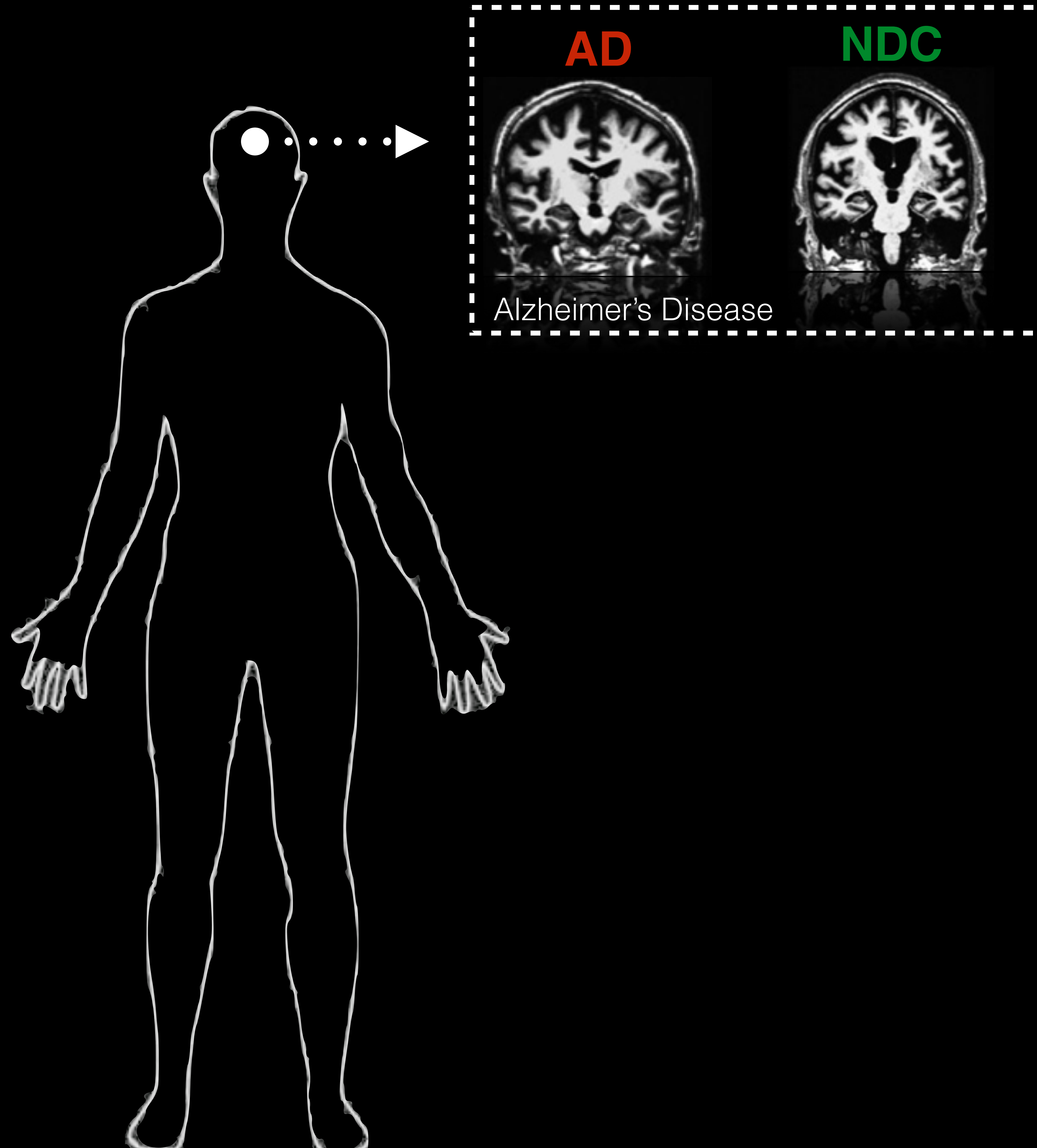
Soft Introduction to Brain-Computer Interfaces

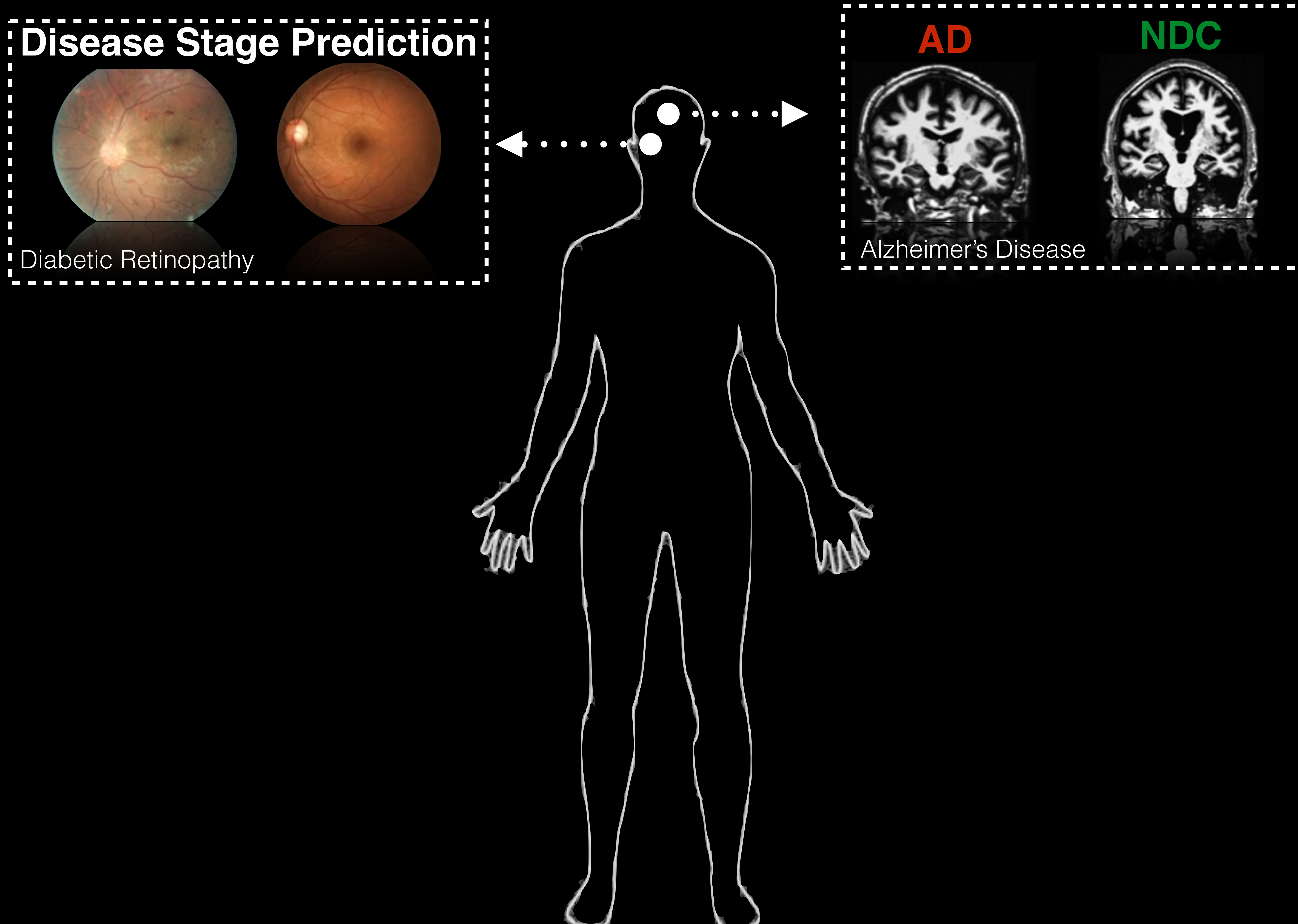
by Ilya Kuzovkin (www.ikuz.eu)

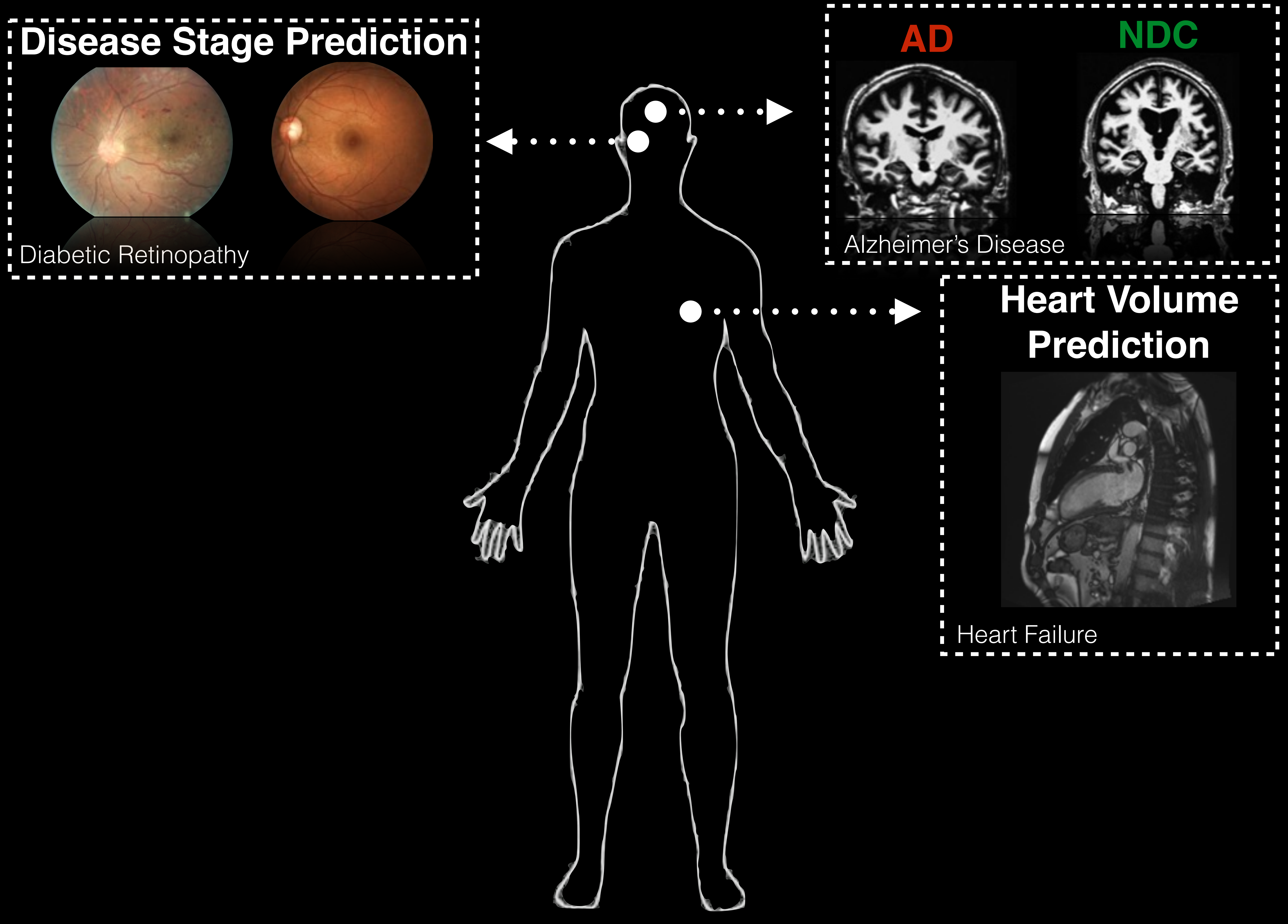

Cardiograph
Cardiogram
Data
P( ) = .92
Machine Learning

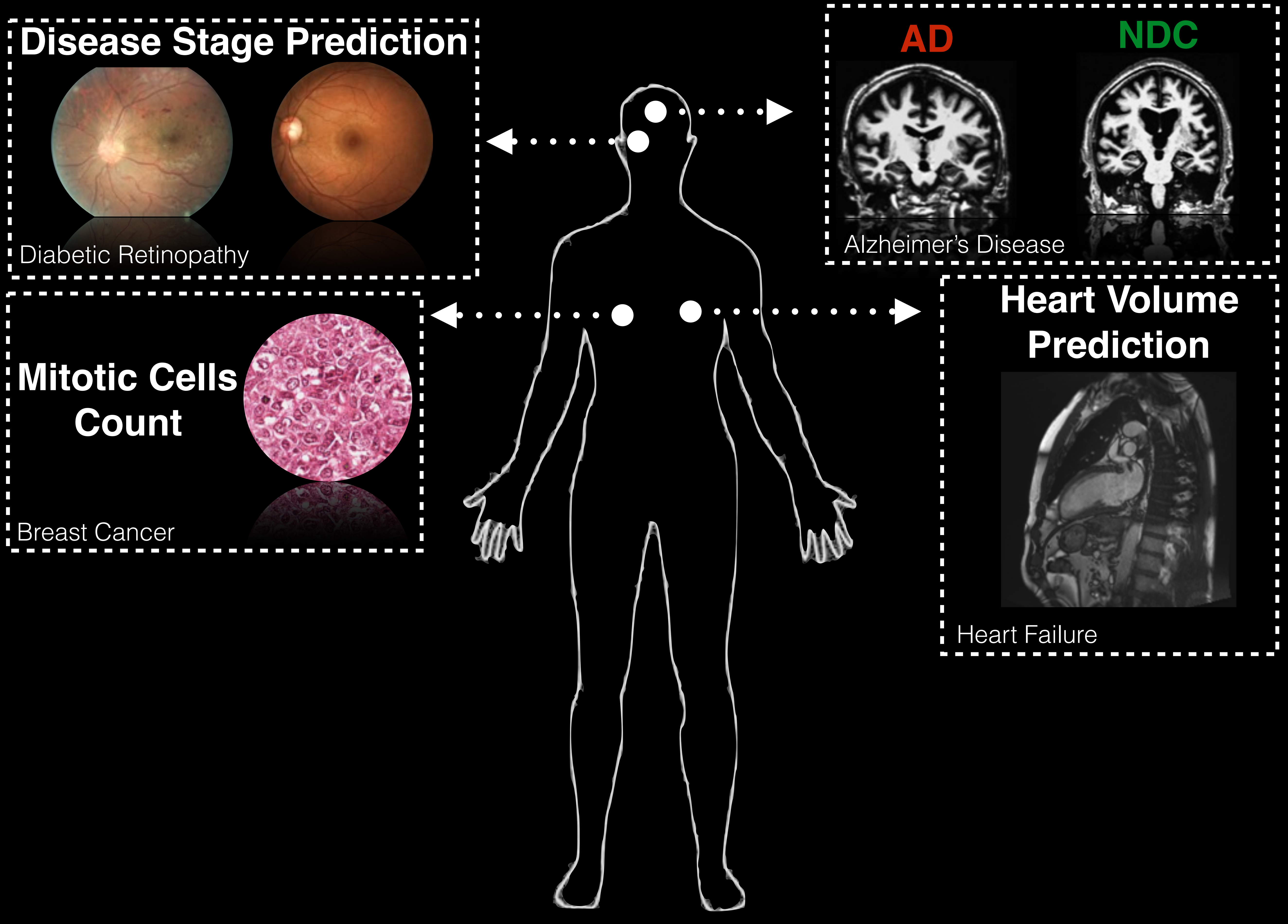

Clinical Data
Medical Signal
Anna, 32, Female, DT2,...
Metadata
ACCTTGCTTACC...
Genetical Information
Medical Imaging

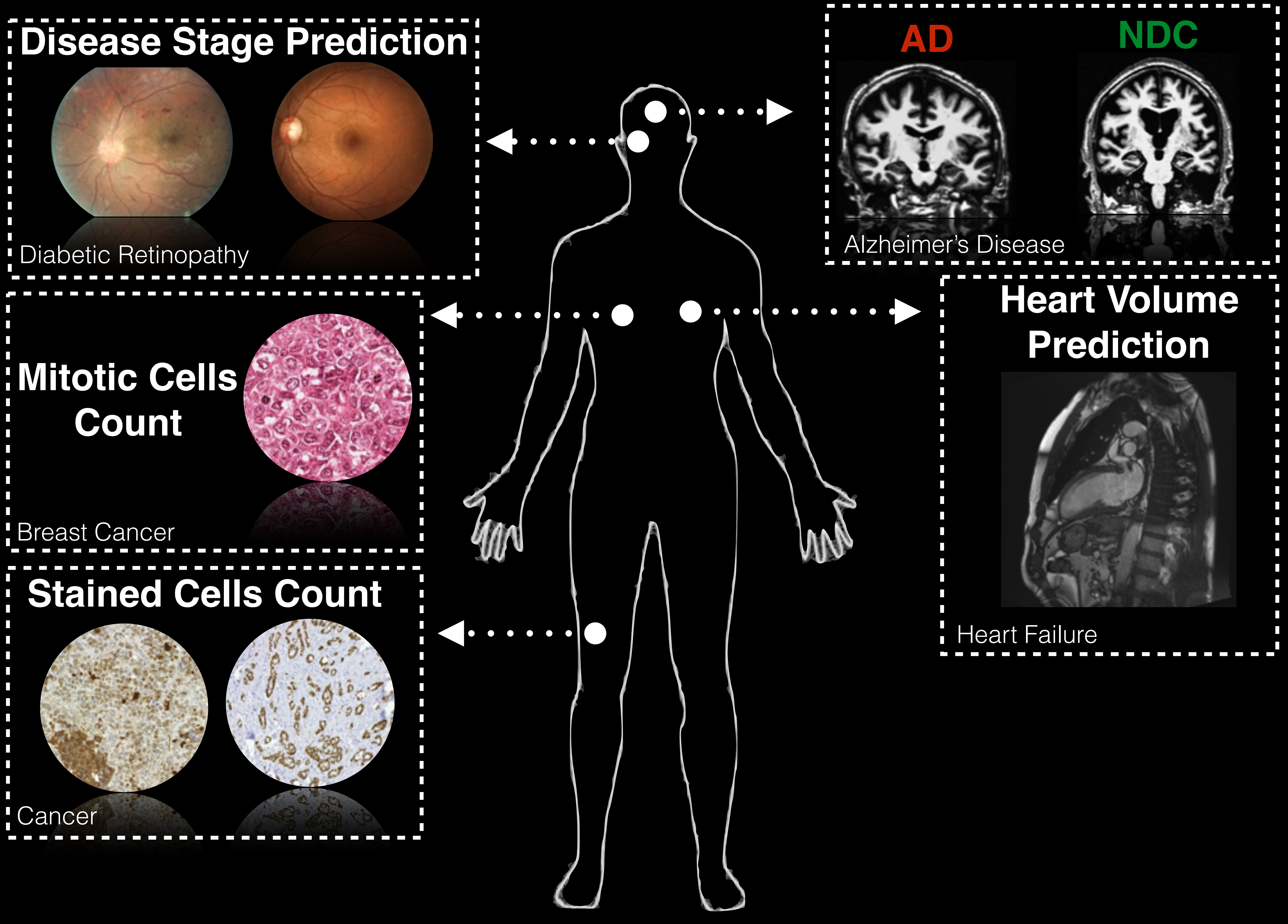


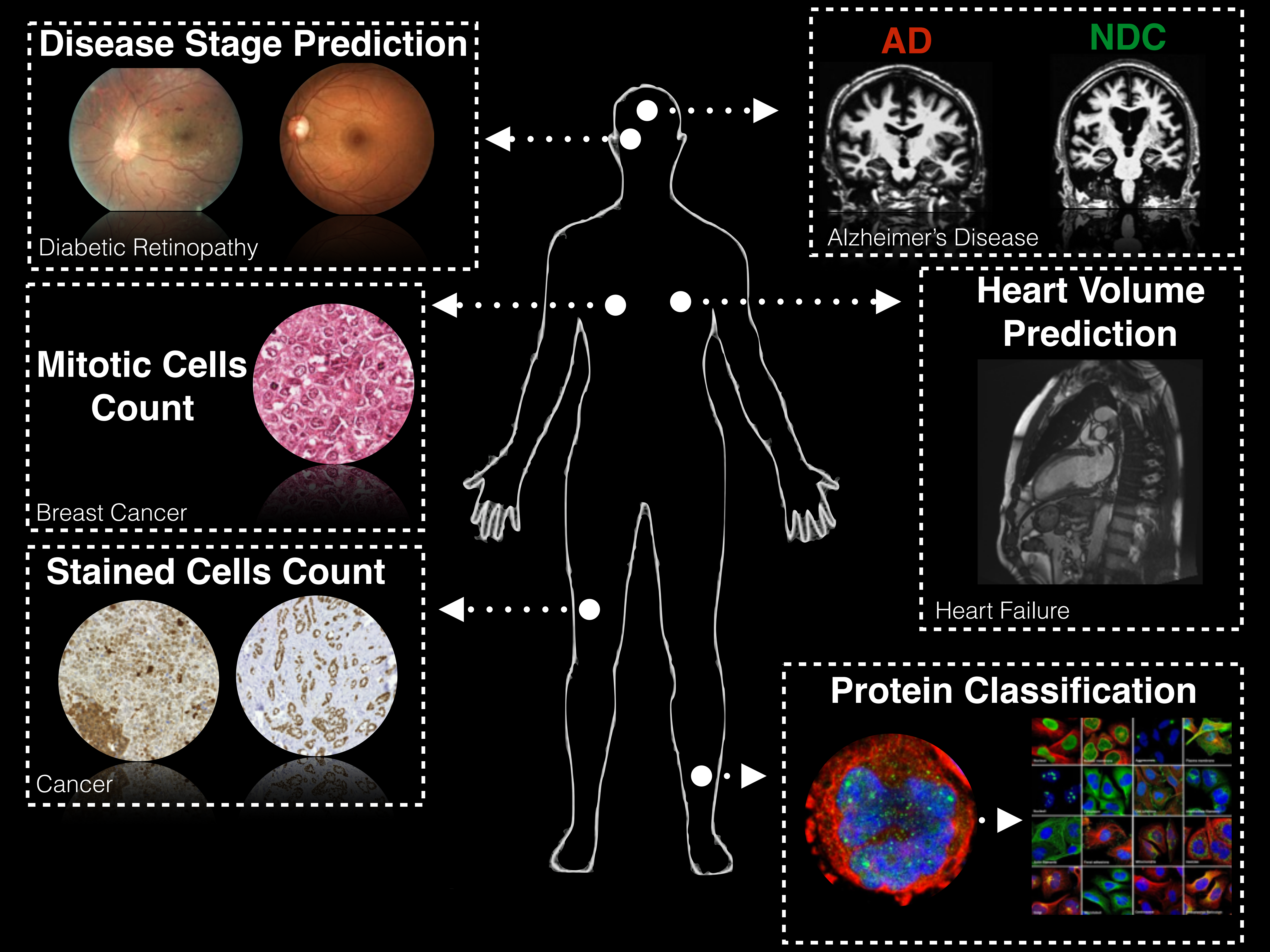

AD
NDC
Alzheimer's Disease

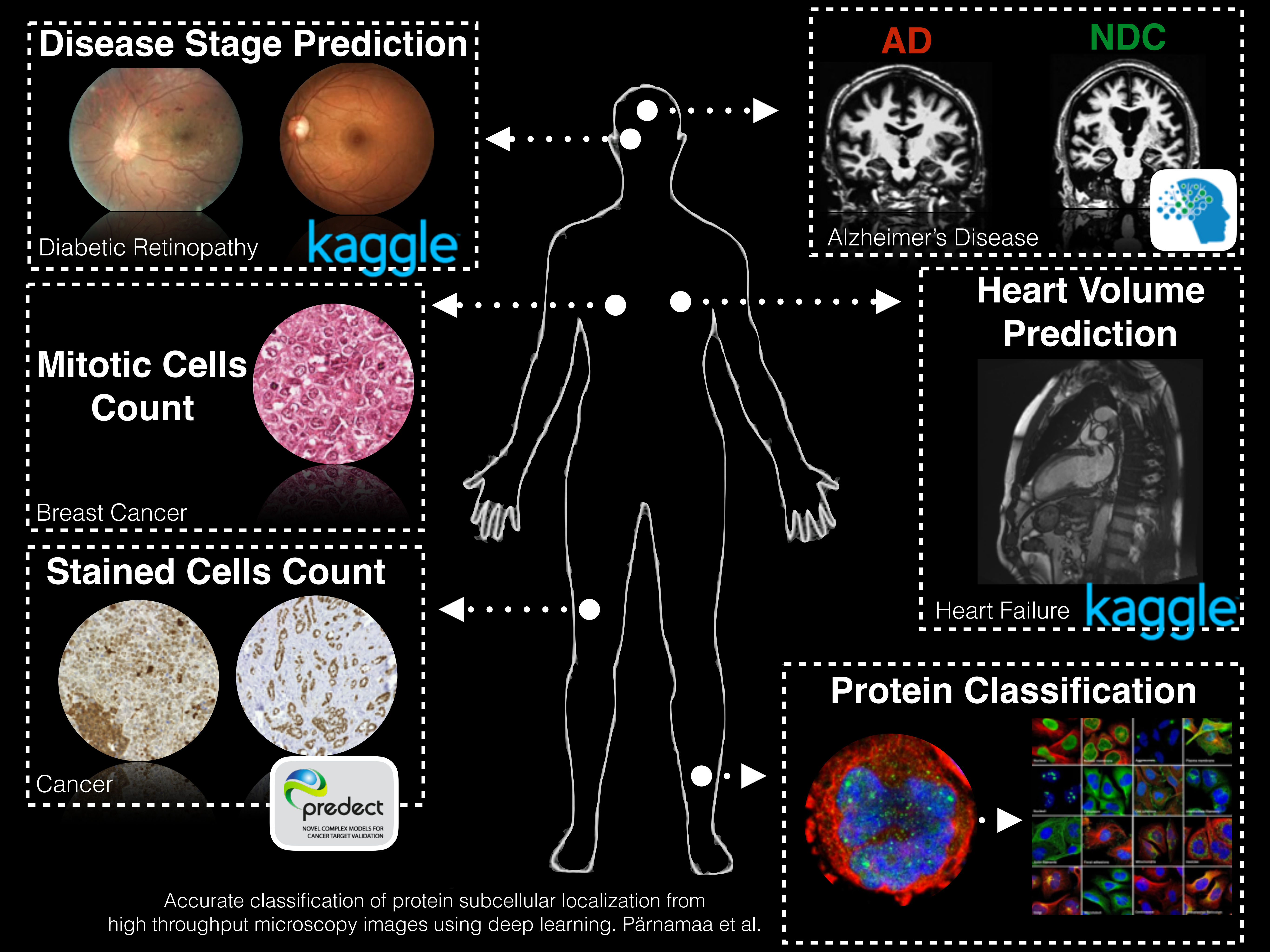

Disease Stage Prediction
Diabetic Retinopathy
AD
NDC
Alzheimer's Disease

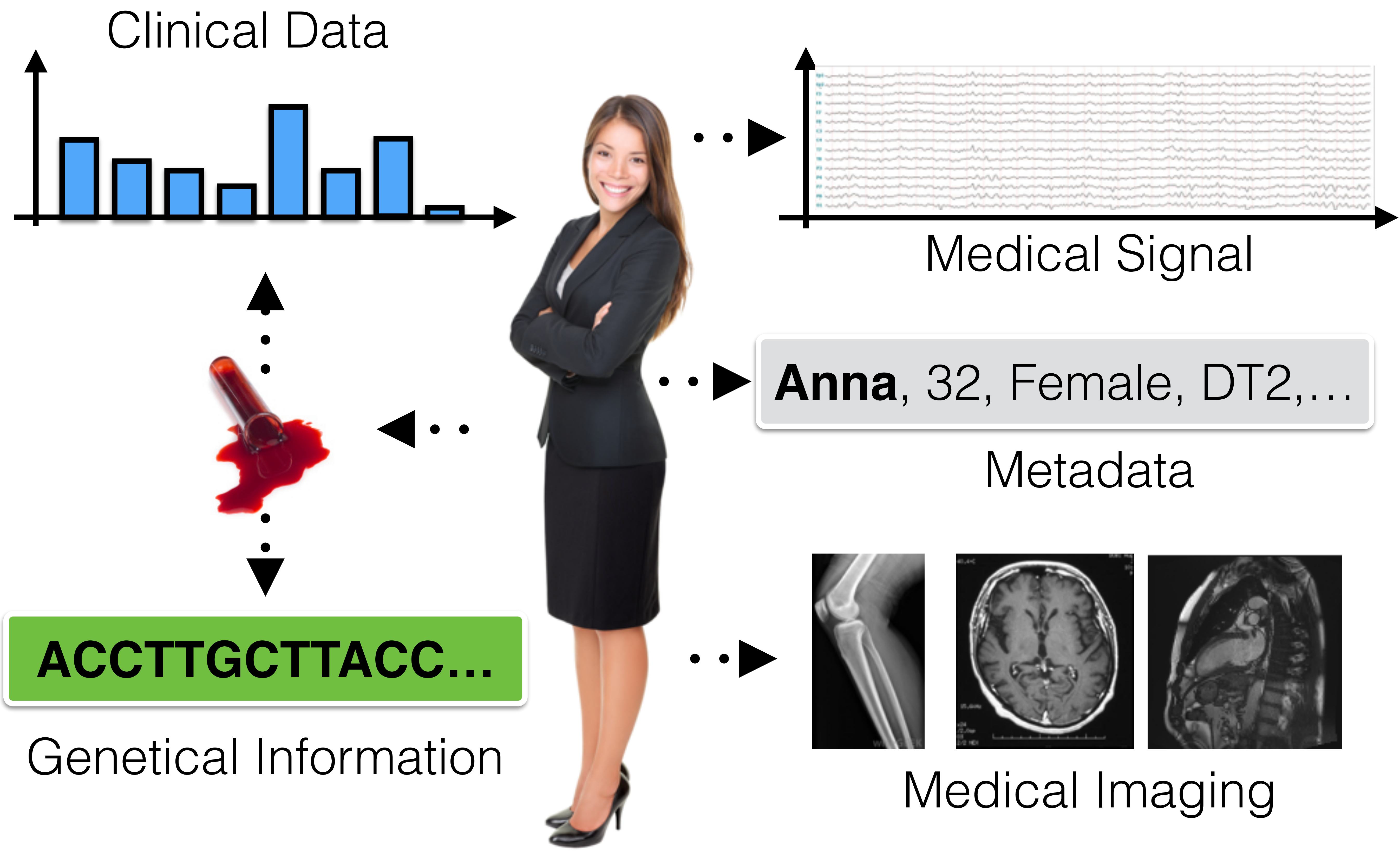

Disease Stage Prediction
Diabetic Retinopathy
AD
NDC
Alzheimer's Disease
Heart Volume Prediction
Heart Failure


Disease Stage Prediction
Diabetic Retinopathy
AD
NDC
Alzheimer's Disease
Mitotic Cells Count
Breast Cancer
Heart Volume Prediction
Heart Failure


Disease Stage Prediction
Diabetic Retinopathy
AD
NDC
Alzheimer's Disease
Mitotic Cells Count
Breast Cancer
Heart Volume Prediction
Heart Failure
Stained Cells Count
Cancer


Disease Stage Prediction
Diabetic Retinopathy
AD
NDC
Alzheimer's Disease
Mitotic Cells Count
Breast Cancer
Heart Volume Prediction
Heart Failure
Stained Cells Count
Cancer
Protein Classification


Disease Stage Prediction
Diabetic Retinopathy
kaggle
AD
NDC
Alzheimer's Disease
Mitotic Cells Count
Breast Cancer
Heart Volume Prediction
Heart Failure
kaggle
Stained Cells Count
Cancer
predect
NOVEL COMPLEX MODELS FOR CANCER TARGET VALIDATION
Protein Classification
Accurate classification of protein subcellular localization from high throughput microscopy images using deep learning. Pärnamaa et al.


Clinical Data
Medical Signal
Anna, 32, Female, DT2,...
Metadata
ACCTTGCTTACC...
Genetical Information
Medical Imaging

National Health System

peterson@nhs.eu

●●●●●●●●●●

Log In

Having trouble logging in?

Dr. Peterson

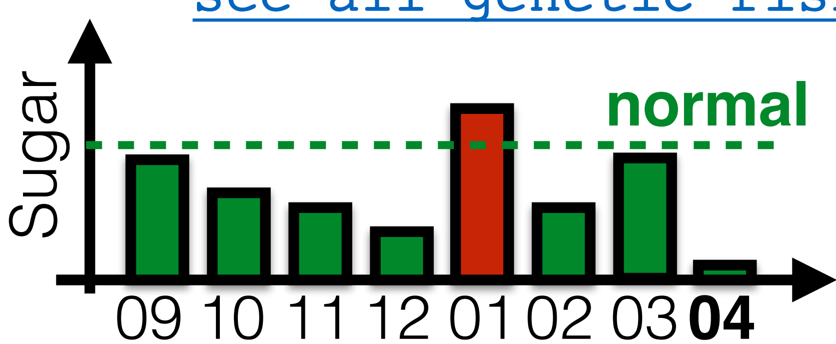
Logout

Anna, 32, Female

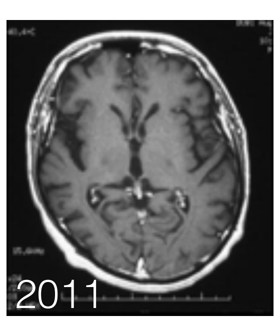
see full profile

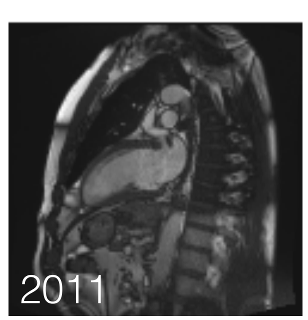
Susceptible to DT2

see all genetic risks



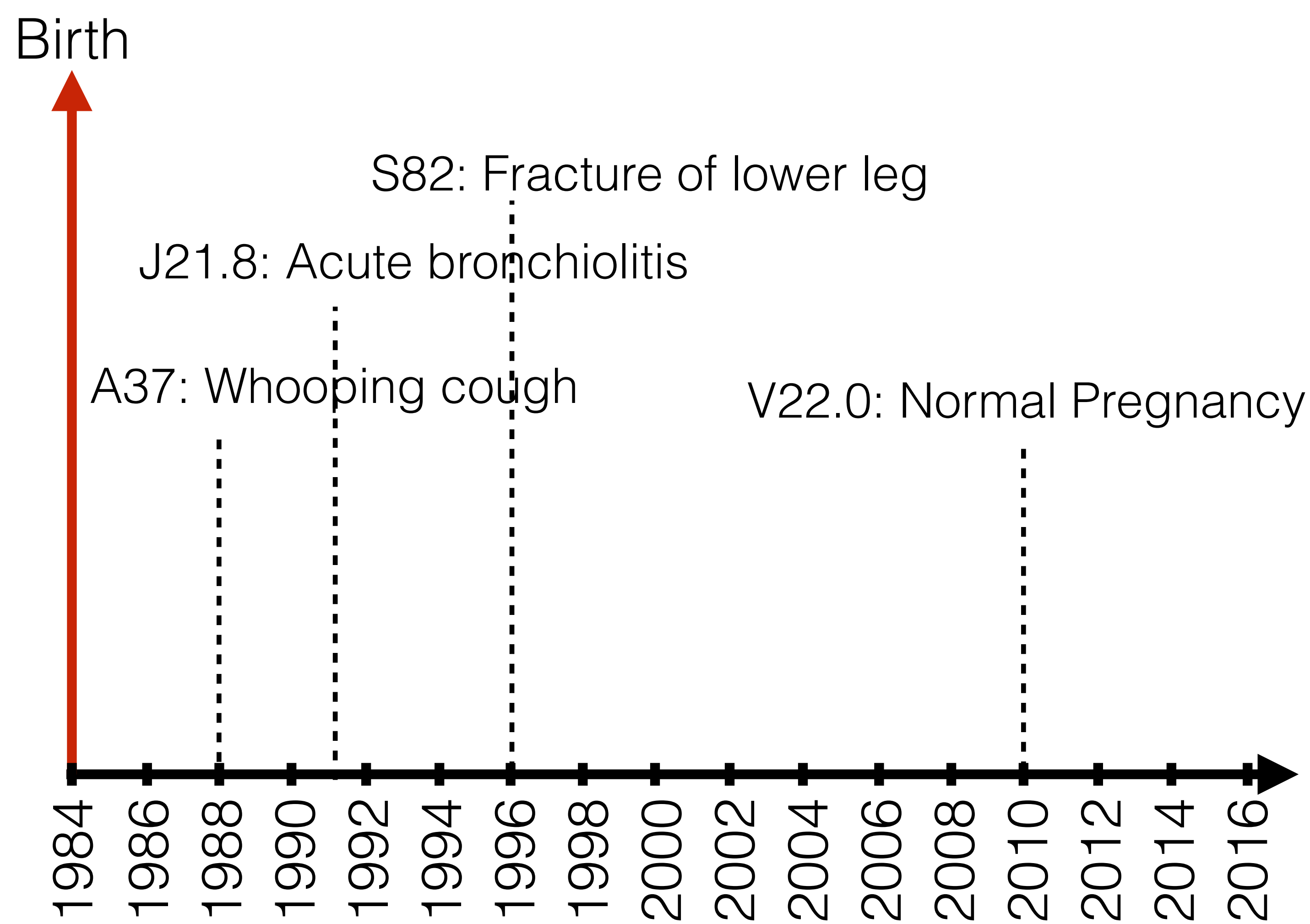
see other tests results





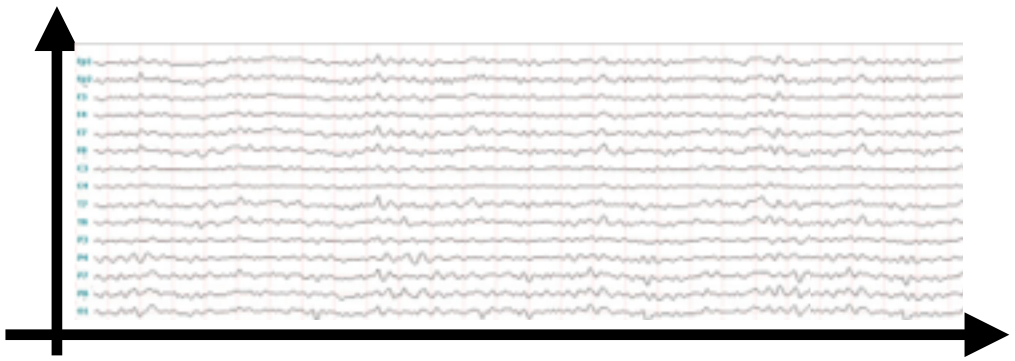
see all recent scans

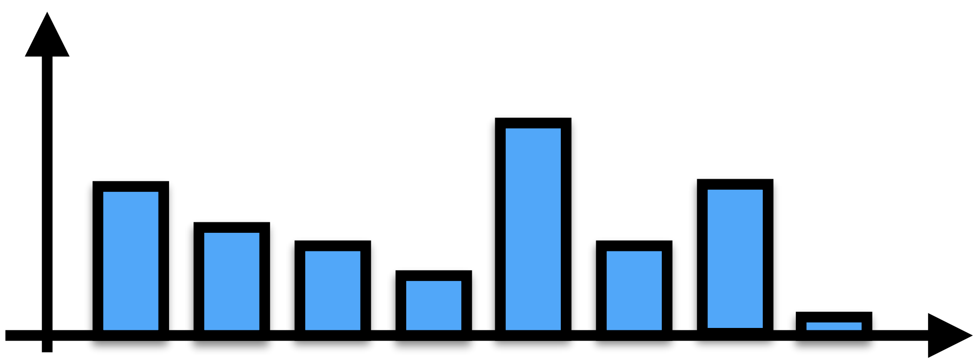
Enter symptoms here…



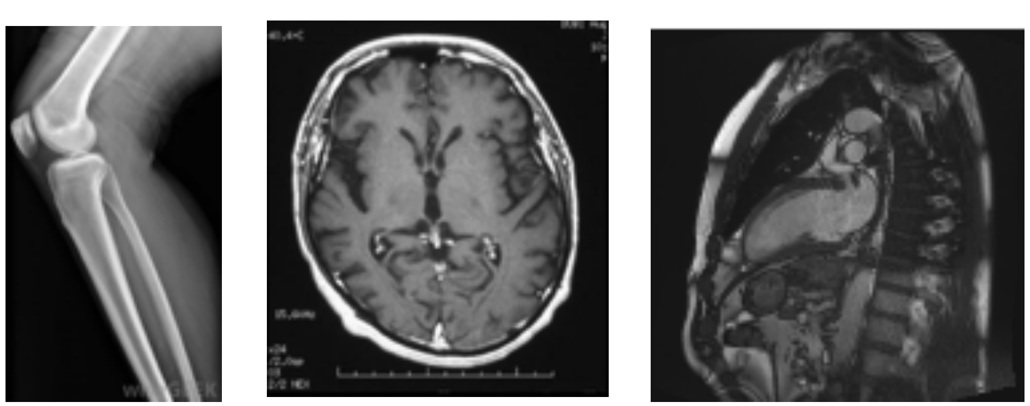
related medical reports

Anna, 32, Female, DT2,…





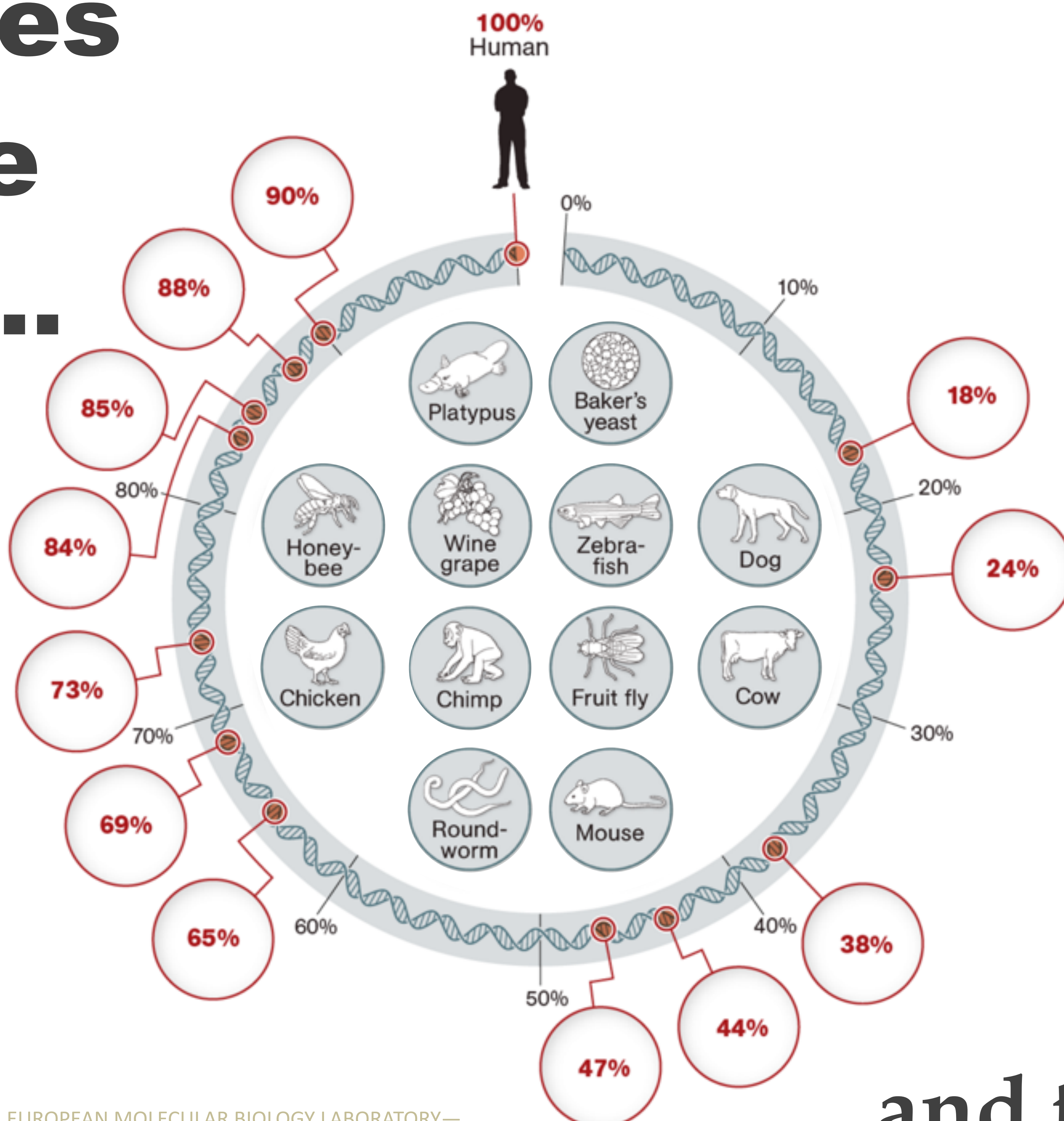
ACCTTGCTTACC…



Data Integration for better decision making is the **key challenge**

Quiz

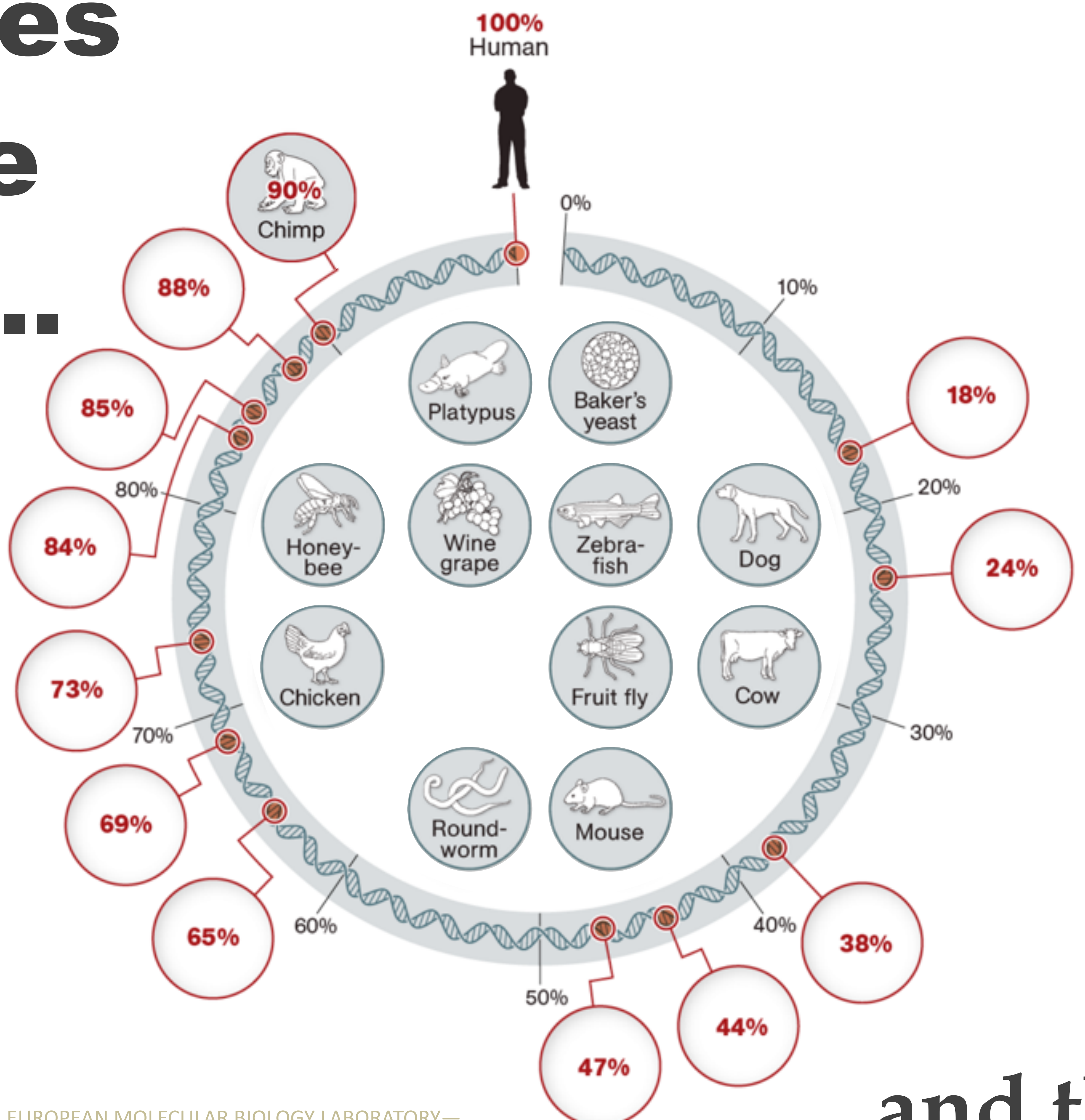
Genes are us...



SOURCE: JAVIER HERRERO, EUROPEAN MOLECULAR BIOLOGY LABORATORY—EUROPEAN BIOINFORMAT ICS INSTITUTE

and them.

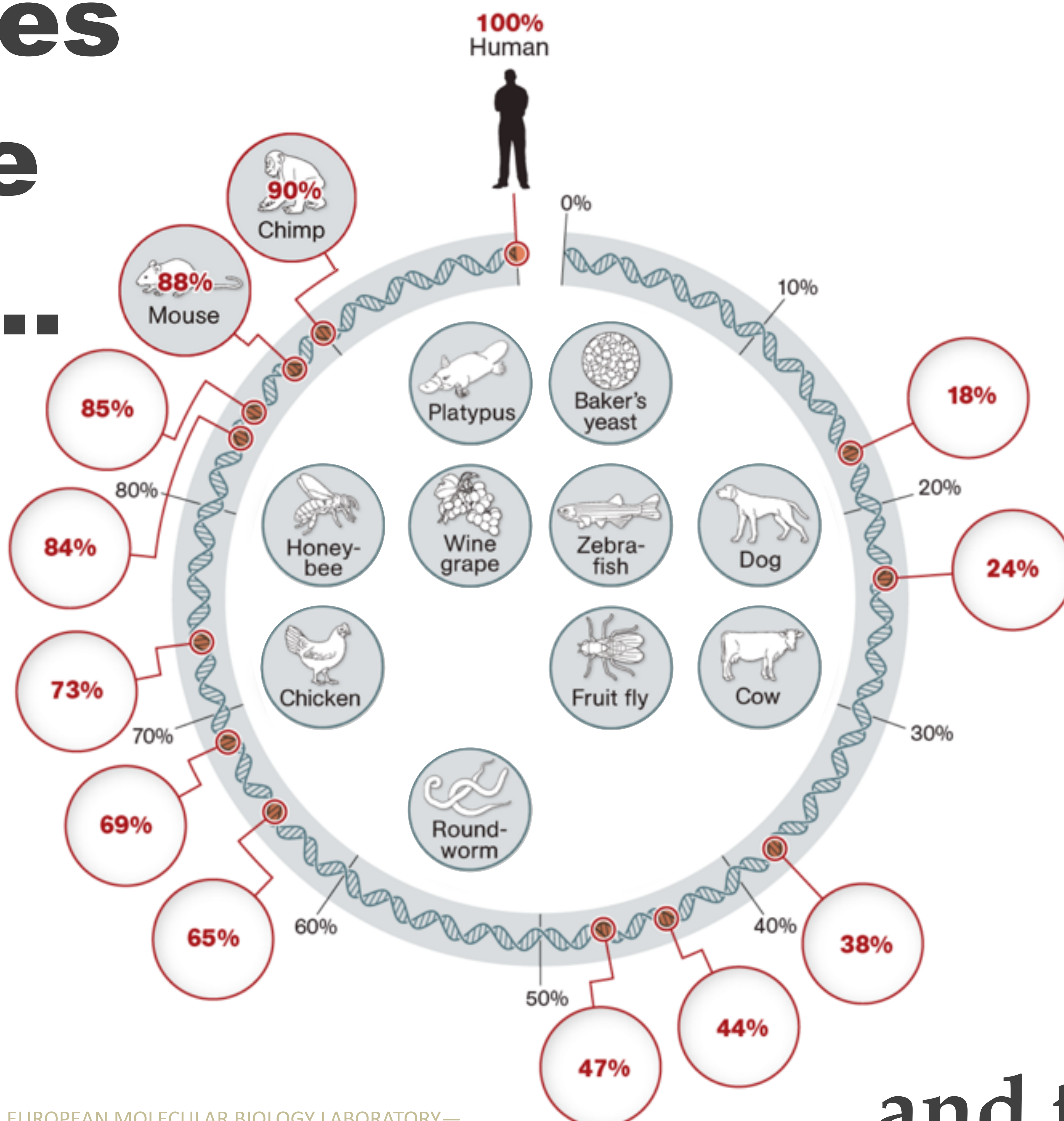
Genes are us…



SOURCE: JAVIER HERRERO, EUROPEAN MOLECULAR BIOLOGY LABORATORY—EUROPEAN BIOINFORMAT ICS INSTITUTE

and them.

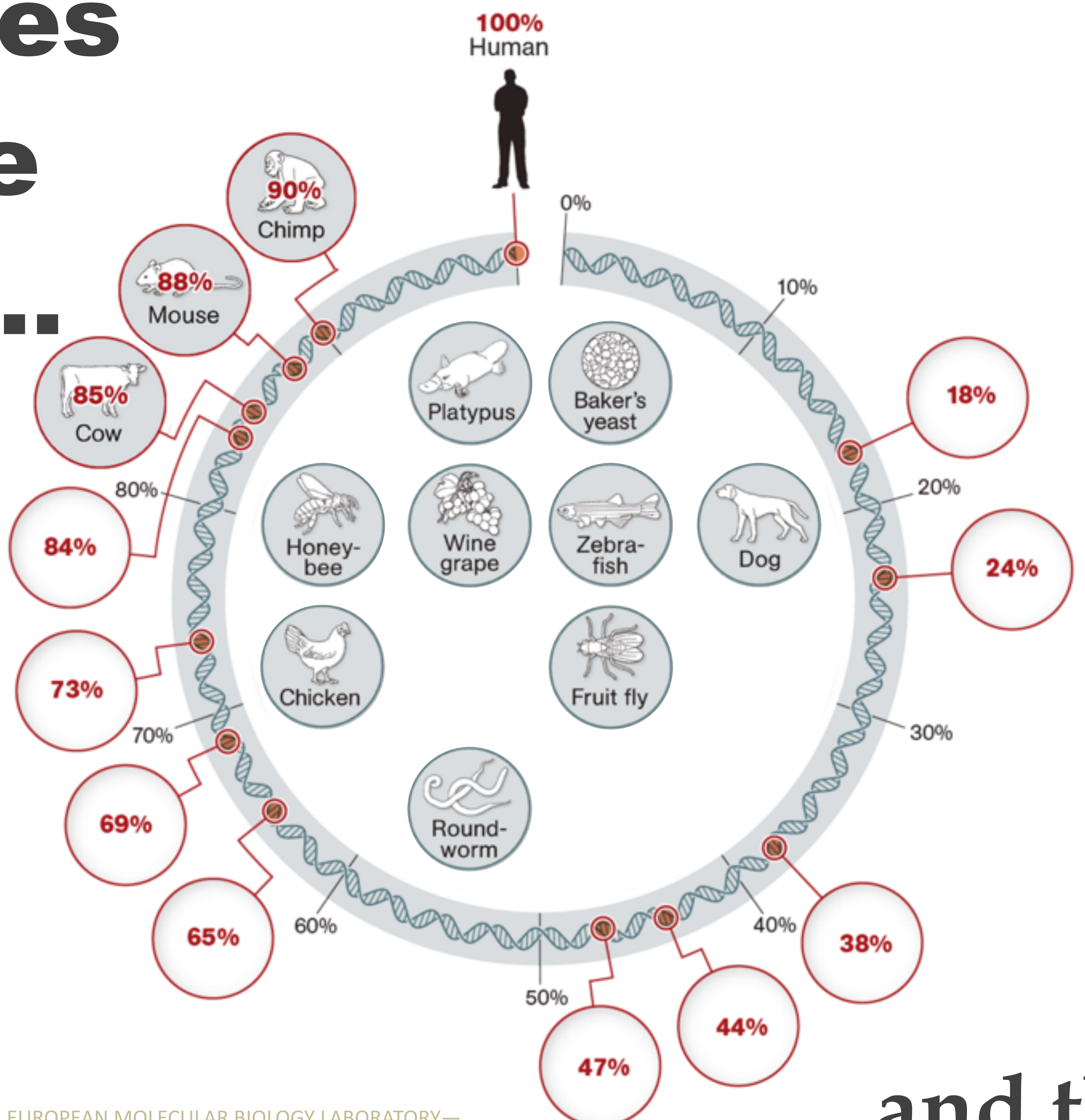
Genes are us...



SOURCE: JAVIER HERRERO, EUROPEAN MOLECULAR BIOLOGY LABORATORY—
EUROPEAN BIOINFORMAT ICS INSTITUTE

and them.

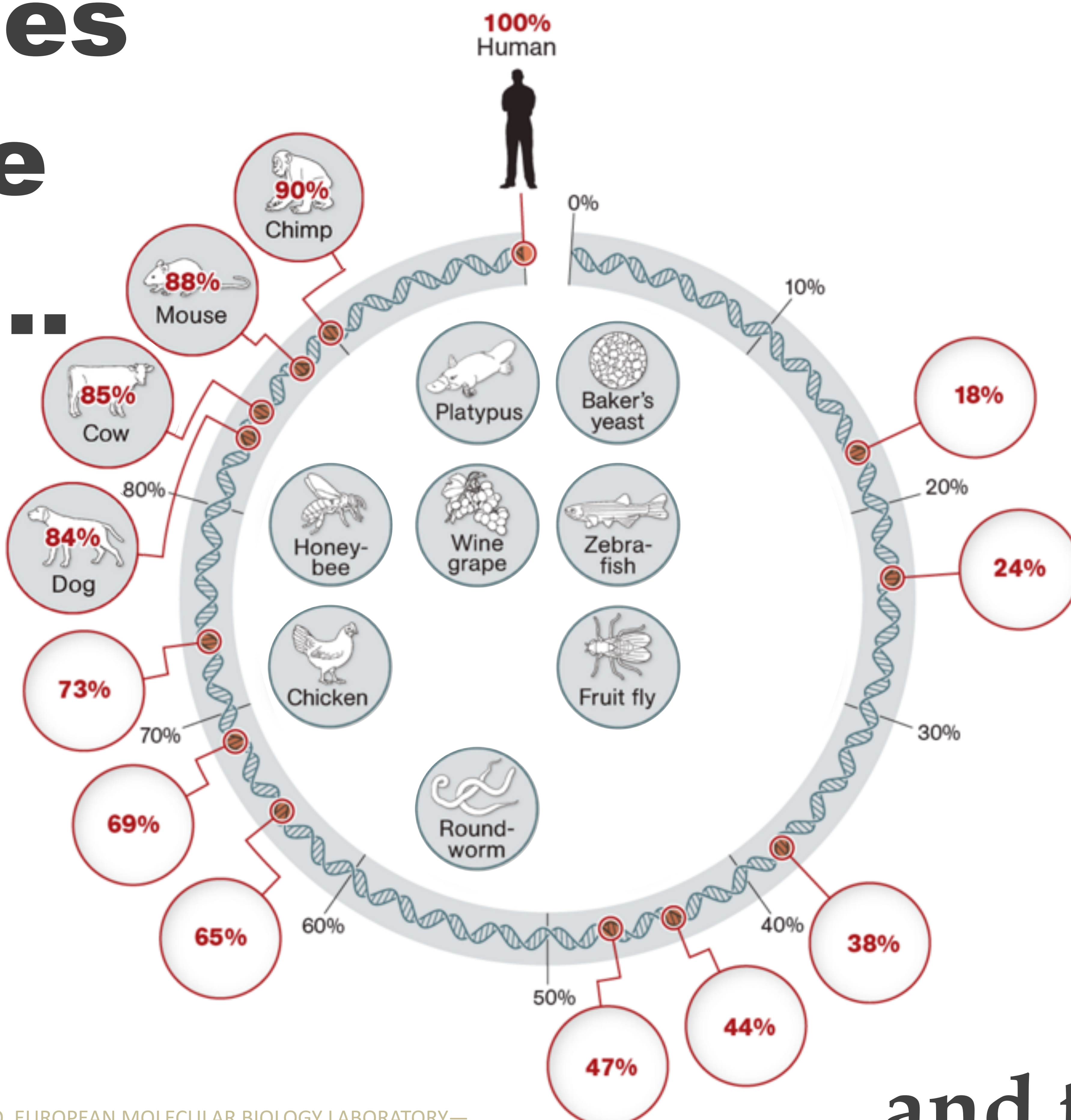
Genes are us...



SOURCE: JAVIER HERRERO, EUROPEAN MOLECULAR BIOLOGY LABORATORY— EUROPEAN BIOINFORMAT ICS INSTITUTE

and them.

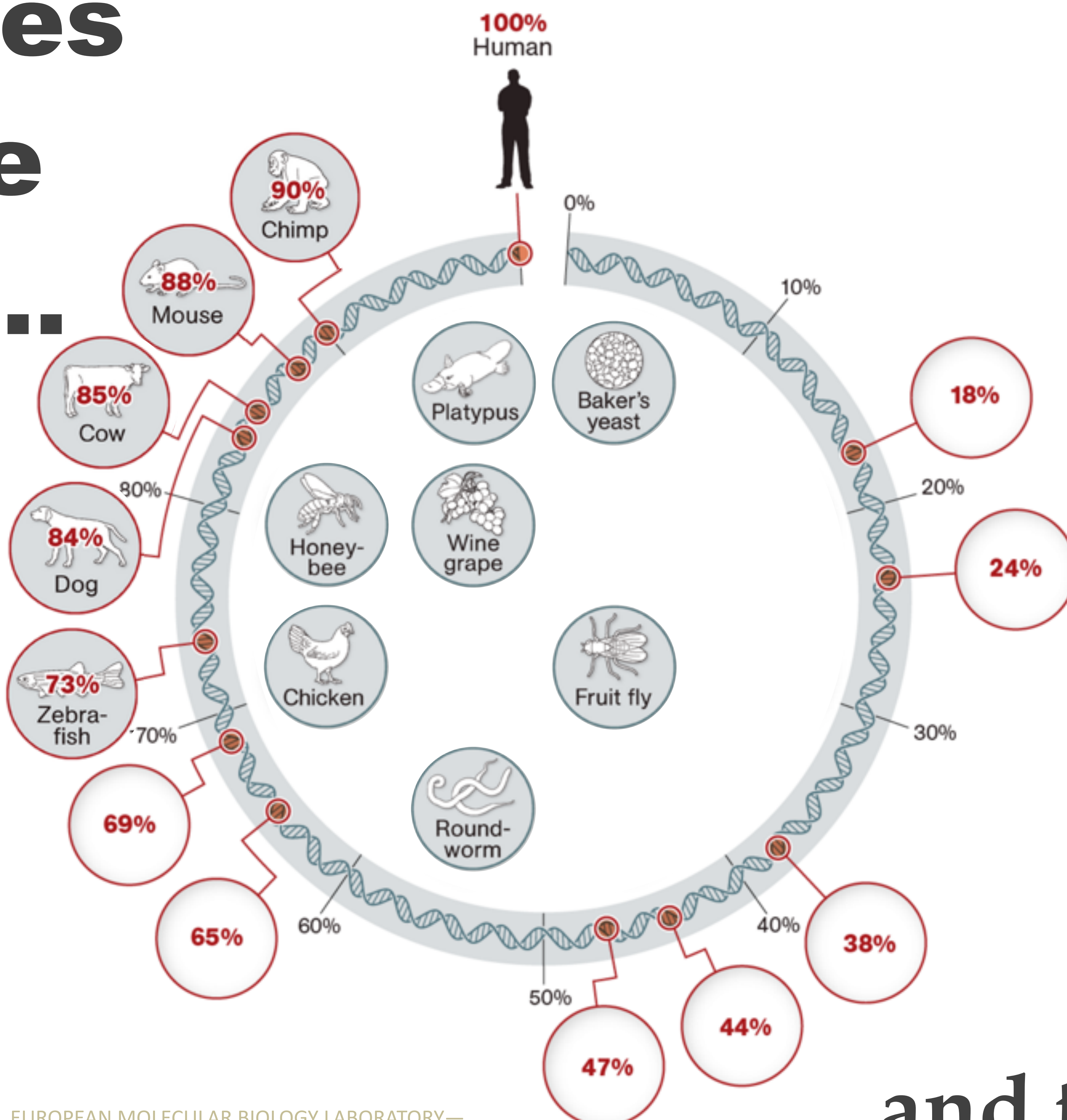
Genes are us…



and them.

SOURCE: JAVIER HERRERO, EUROPEAN MOLECULAR BIOLOGY LABORATORY—EUROPEAN BIOINFORMAT ICS INSTITUTE

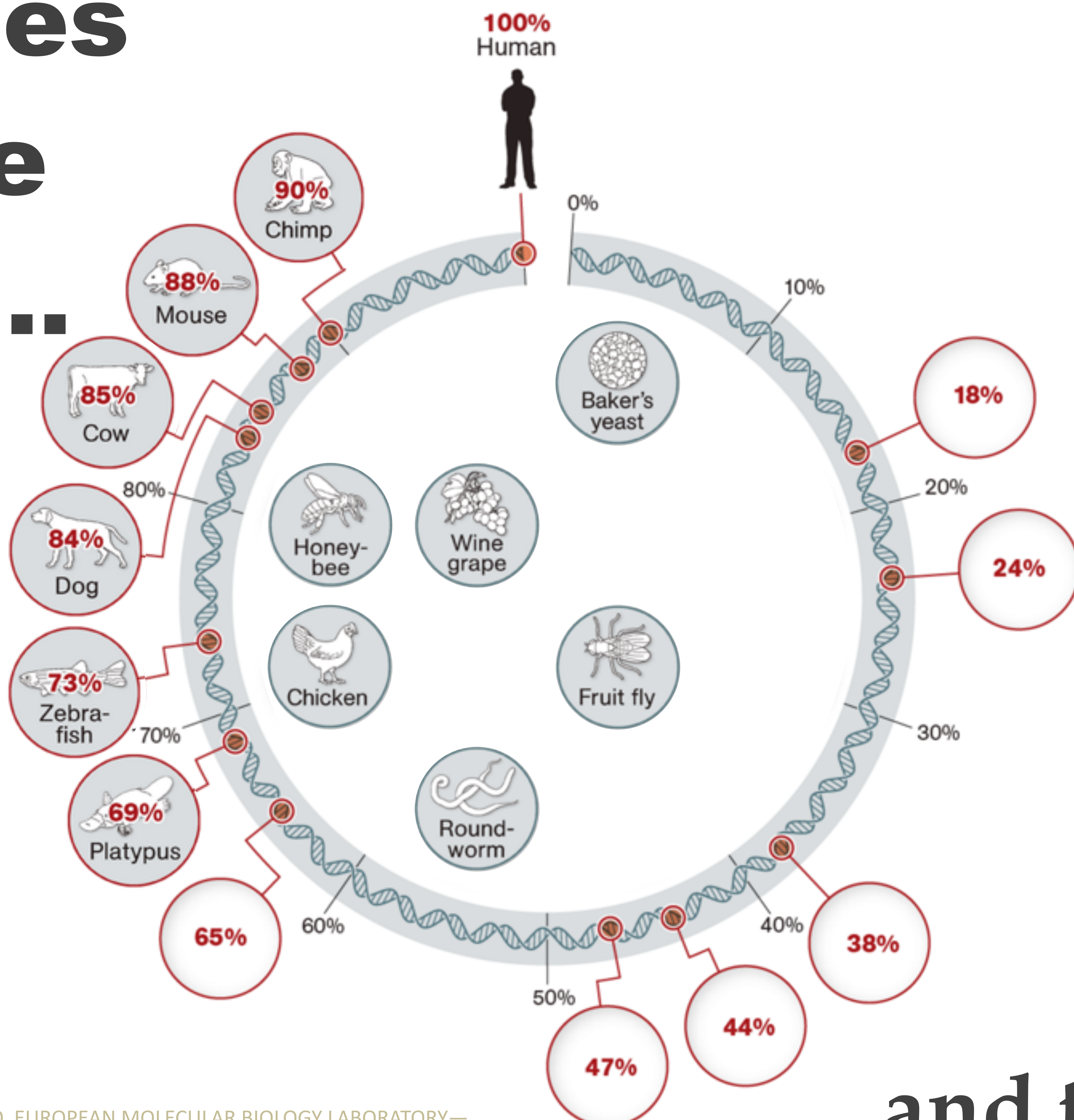
Genes are us...



and them.

SOURCE: JAVIER HERRERO, EUROPEAN MOLECULAR BIOLOGY LABORATORY—EUROPEAN BIOINFORMAT ICS INSTITUTE

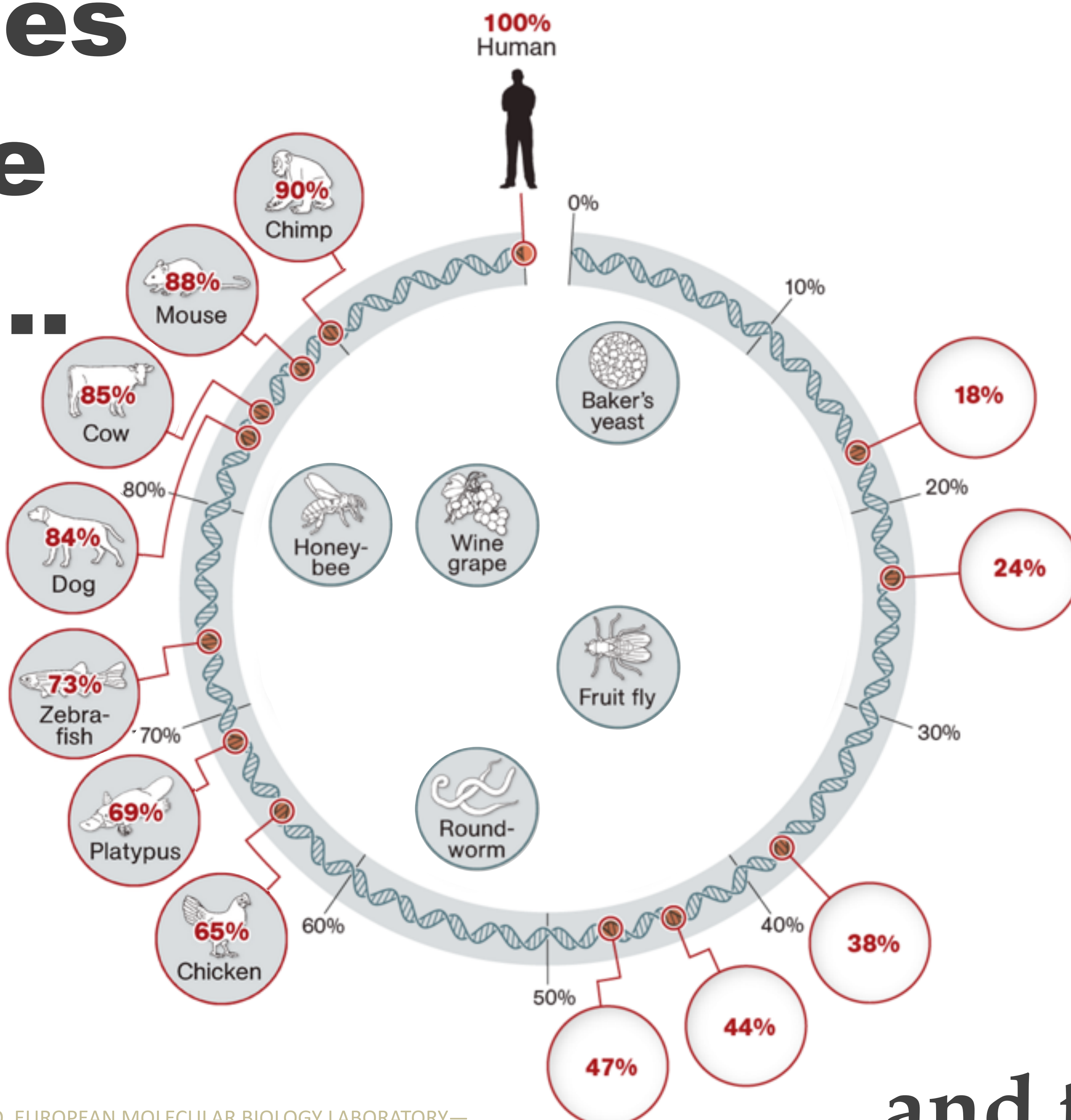
Genes are us...



and them.

SOURCE: JAVIER HERRERO, EUROPEAN MOLECULAR BIOLOGY LABORATORY—
EUROPEAN BIOINFORMAT ICS INSTITUTE

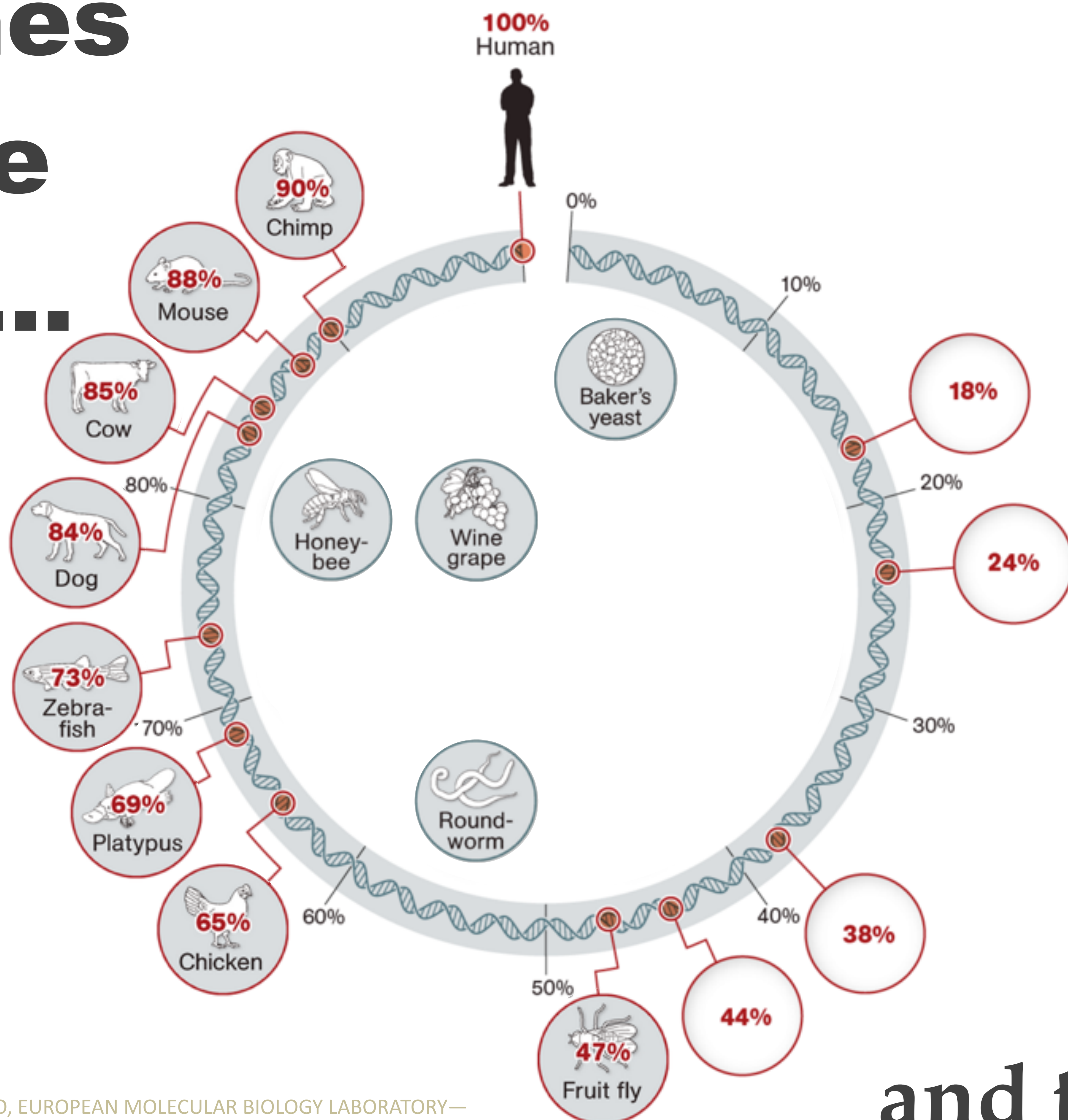
Genes are us...



SOURCE: JAVIER HERRERO, EUROPEAN MOLECULAR BIOLOGY LABORATORY—
EUROPEAN BIOINFORMAT ICS INSTITUTE

and them.

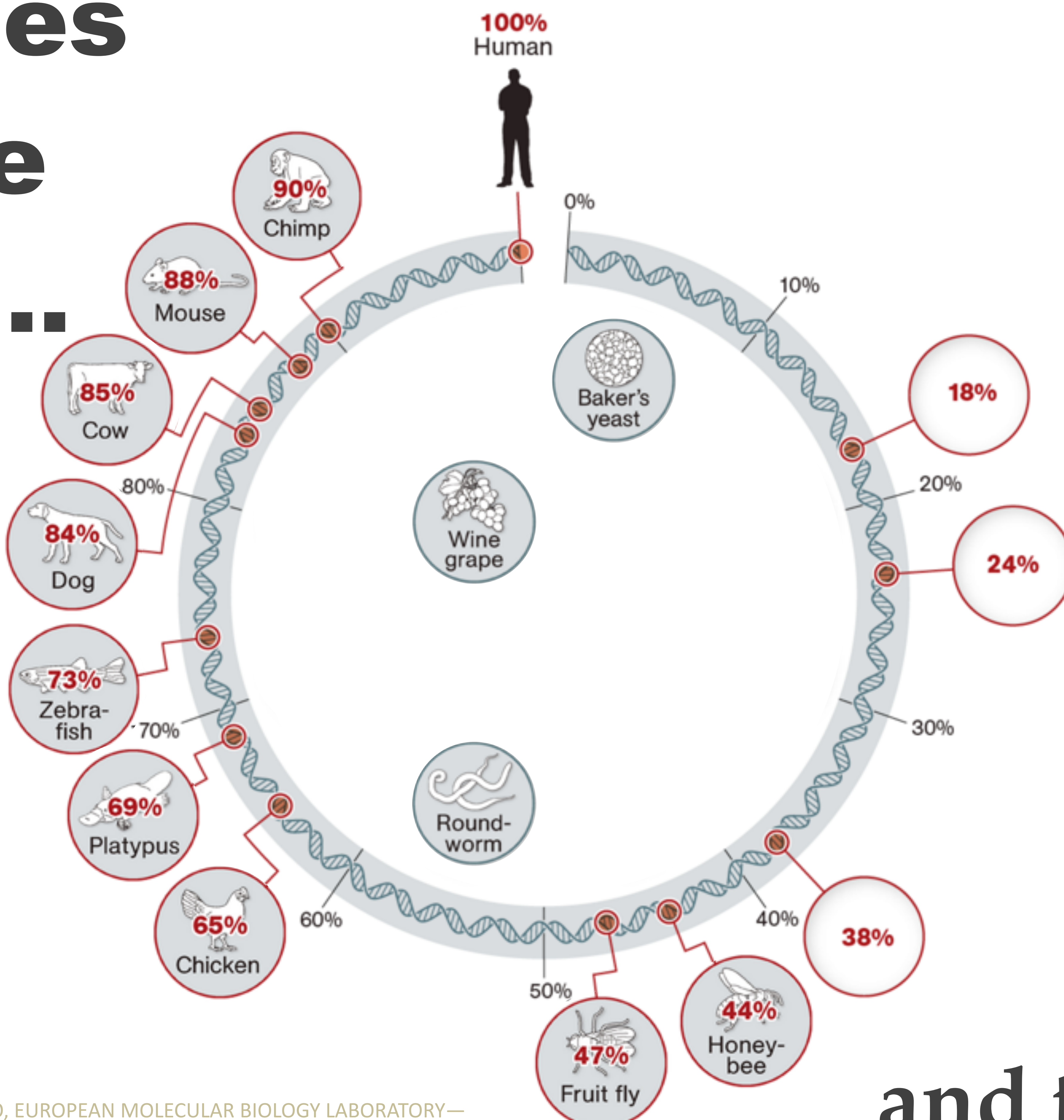
Genes are us…



SOURCE: JAVIER HERRERO, EUROPEAN MOLECULAR BIOLOGY LABORATORY—
EUROPEAN BIOINFORMAT ICS INSTITUTE

and them.

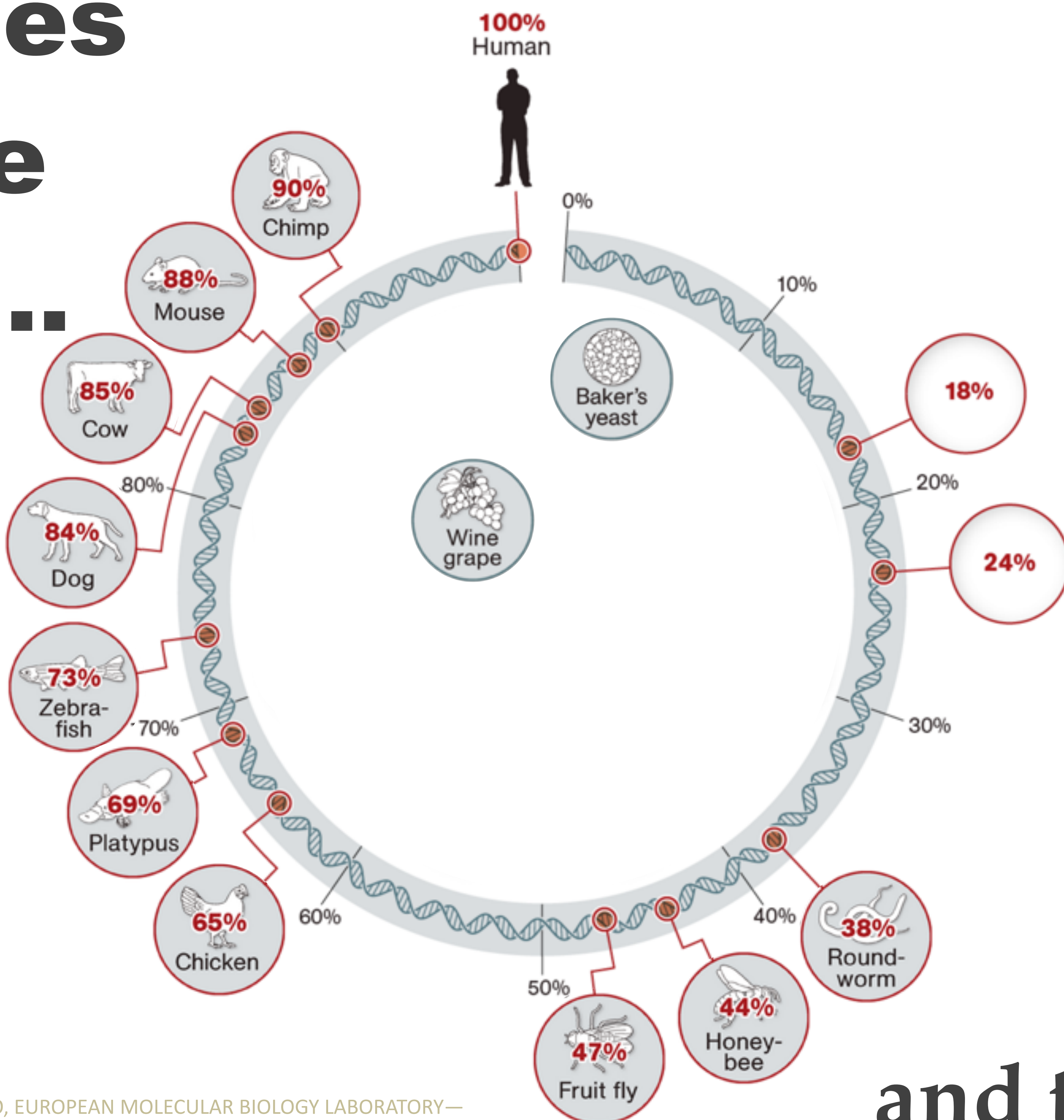
Genes are us...



SOURCE: JAVIER HERRERO, EUROPEAN MOLECULAR BIOLOGY LABORATORY—
EUROPEAN BIOINFORMAT ICS INSTITUTE

and them.

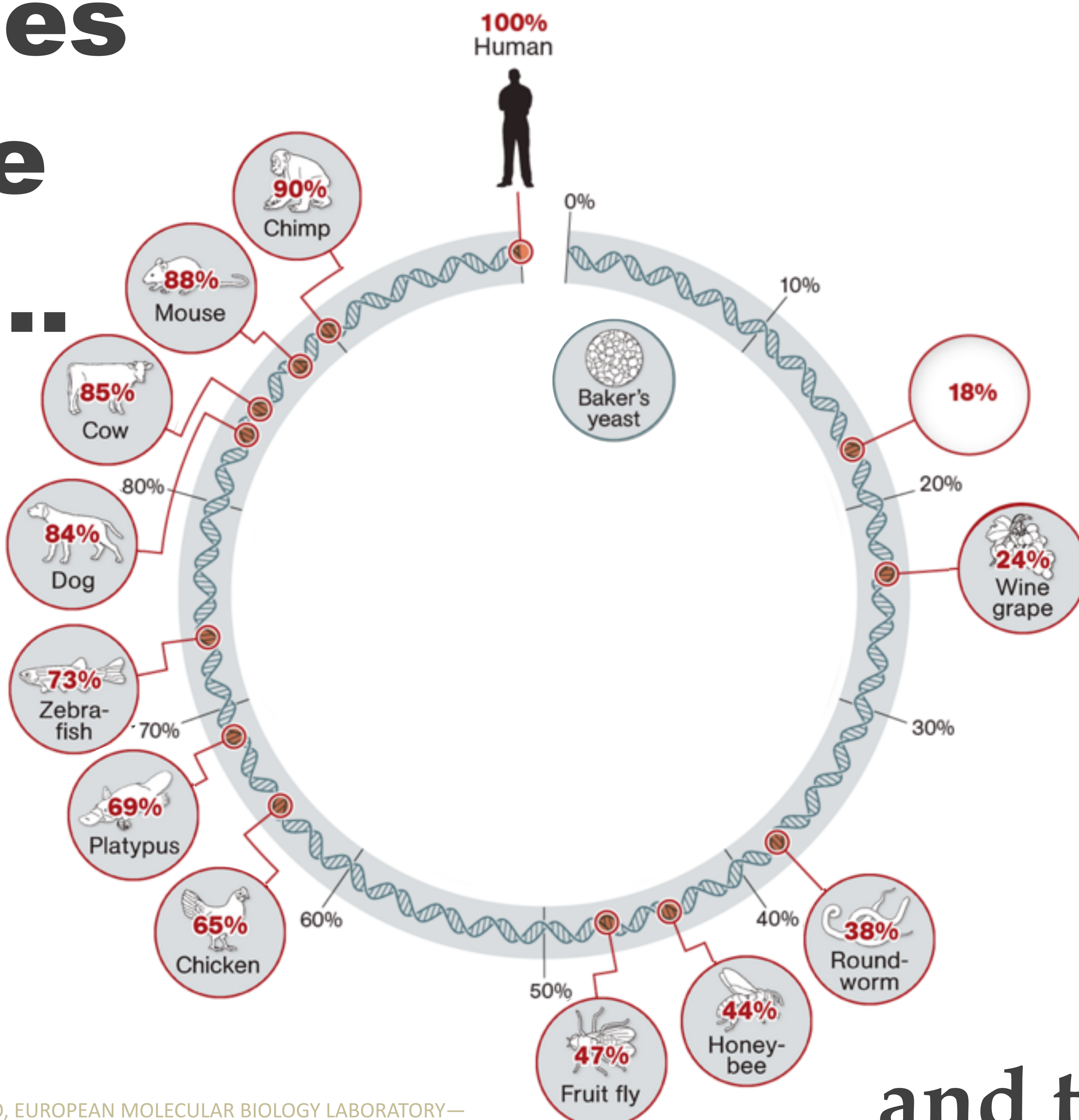
Genes are us…



SOURCE: JAVIER HERRERO, EUROPEAN MOLECULAR BIOLOGY LABORATORY—
EUROPEAN BIOINFORMAT ICS INSTITUTE

and them.

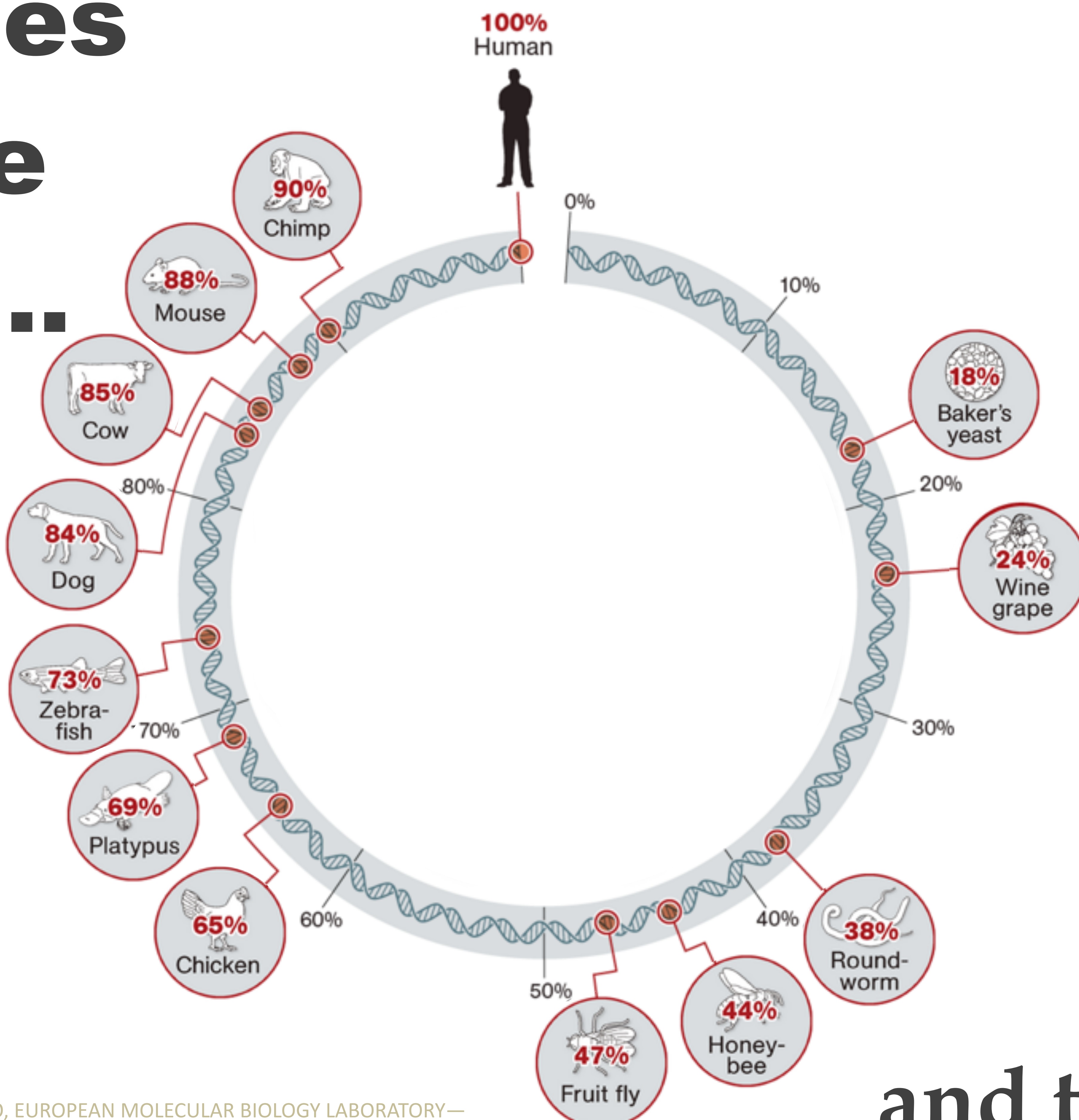
Genes are us…



SOURCE: JAVIER HERRERO, EUROPEAN MOLECULAR BIOLOGY LABORATORY—
EUROPEAN BIOINFORMAT ICS INSTITUTE

and them.

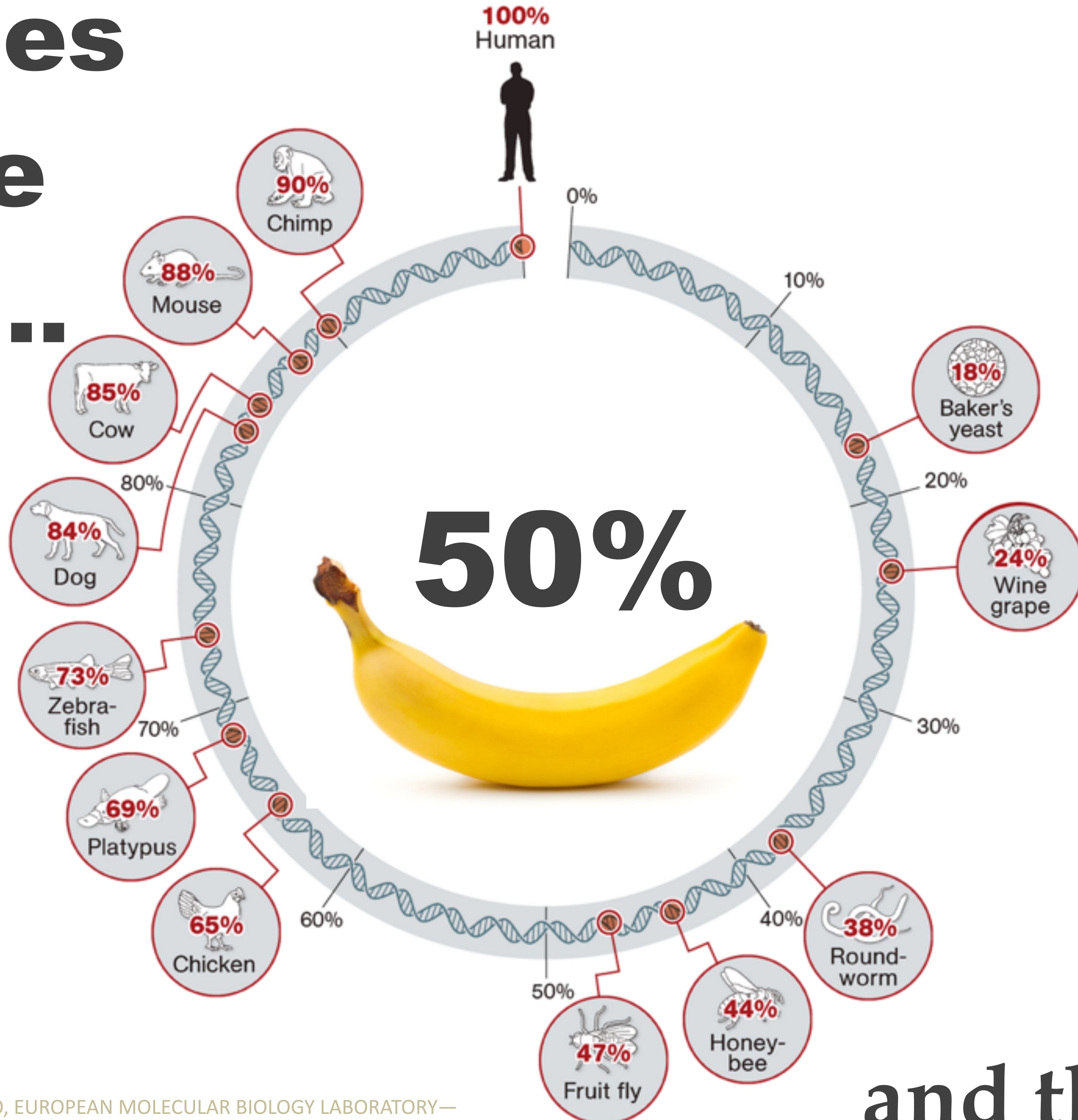
Genes are us…



SOURCE: JAVIER HERRERO, EUROPEAN MOLECULAR BIOLOGY LABORATORY—
EUROPEAN BIOINFORMAT ICS INSTITUTE

and them.

Genes are us...



and them.

SOURCE: JAVIER HERRERO, EUROPEAN MOLECULAR BIOLOGY LABORATORY—
EUROPEAN BIOINFORMAT ICS INSTITUTE


UPEER
MEETUP

dmytro@ut.ee

BIIT

QureTEC

Rocket science is for kids
Bioinformatics is for scientists

CLC bio