UNIVERSITÉ DE MONTPELLIER

AI applied to medicine

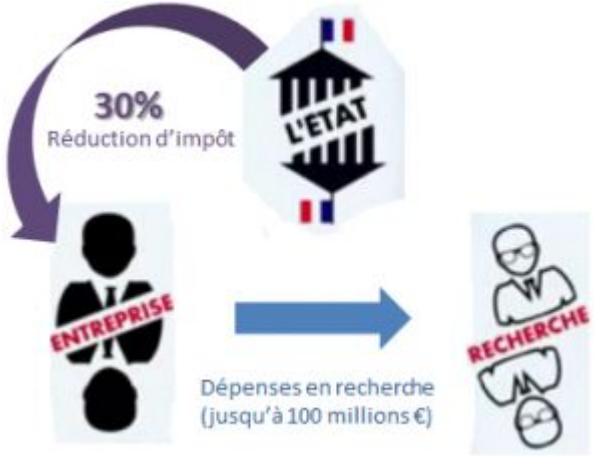
Kévin Yauy
Physician-Scientist Fellow, CHU Montpellier
MD in Medical Genetics
PhD in Bioinformatics and Machine Learning

Summary

1. A short brief for future PhD candidates
2. An introduction to artificial intelligence
3. Applied AI to the genomic medicine example

UNIVERSITÉ DE MONTPELLIER

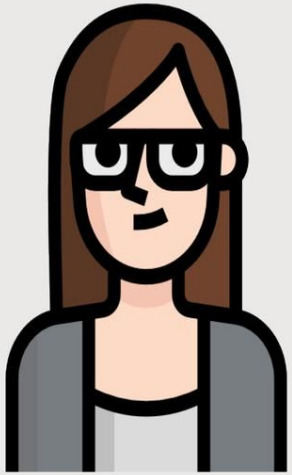
A short introduction to CIFRE thesis before…



Exemple

Dispositif jeune docteur

Un jeune docteur, rémunéré 35 k€ brut annuel représente pour l'entreprise :



Avec 100% de son temps passé en R&D

35 k€ x 1,41 x 2 x 2 x 0,3 = 59,2 k€ de CIR

Le jeune docteur ne coûte donc rien et rapporte même 10 k€ à l'entreprise (car elle aura déboursé 35 k€ x 1,41 de cotisations patronales soit 49,4 k€).

Avec 80% de son temps passé en R&D

35 k€ x 1,41 x 2 x 2 x 0,3 x 0,8 = 47,4 k€ de CIR

Le jeune docteur ne coûte donc quasiment rien à l'entreprise

UNIVERSITÉ DE MONTPELLIER

CIFRE in Montpellier?

Who benefit from CIFRE ?

- Small Company : 45 % Cifre
- Big Company: 38 %

La ville abrite aujourd'hui un formidable vivier de « 32 000 étudiants, 4 000 scientifiques, **300 entreprises en santé** et plus de 16 500 emplois », salué par la présidente de la région Occitanie Carole Delga, qui a ouvert ces premières assises.

Actu › Occitanie › Hérault › Montpellier

Montpellier. Recherche : SeqOne Genomics, la pépite médicale lève 20 M€

Fleuron de la recherche médicale, SeqOne Genomics développe des solutions d'analyse génomique pour la médecine personnalisée dont elle veut devenir un leader mondial.

CIFRE Thesis, hands on

1. Seek for a research lab (me in Grenoble)
   and a company (SeqOne Genomics in Montpellier)

2. Administrative and scientific development (1 an)
   a. Write a research project for CIFRE
   b. Company and university settle a collaboration contract

3. Only a yearly report (mainly based on CSI report)

Finally, “la CIFRE”

Lauréat du Prix Sabatier d’Espeyran de la Ville de Montpellier & l’Académie des Sciences et Lettres de Montpellier 2021

=> An incredible journey it has been in Grenoble!

- Learnt a new job : Data Scientist, PhD!
- Improve programming skills and scientific writing
- Two patent-filed applications
- Co-authors of six publications, including 2 first authors, 2 second authors, and 1 third author
- Discover the industrial side of research and company management
- Alps Finalist of 3-Minutes Thesis
- Co-leader of a bilingual MOOC in Genomic Medicine and Bioinformatics

Artificial Intelligence - Quescussion start

Only answers with questions that could some this problem

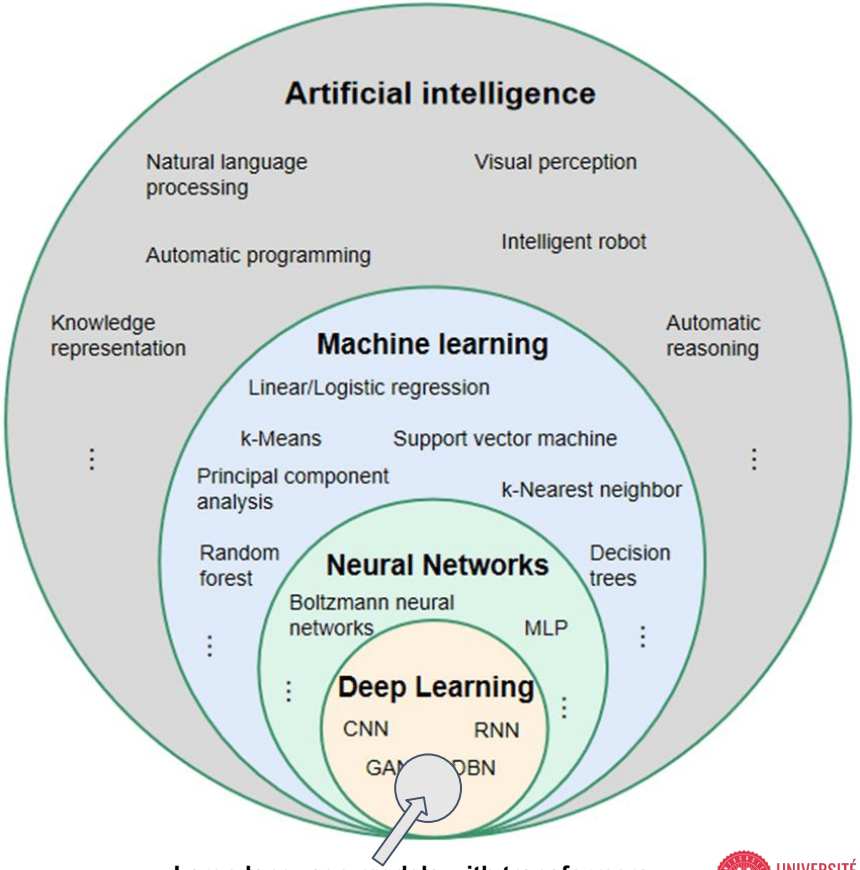
What AI can or can’t do today ?

What are we talking about with AI ?

AI is splitted in different concepts to mimic cognitive human functions

A multitude of methods have emerged starting from
Machine Learning to Neural Networks
to make this happen

Today -> Transformers (and GPT) are most advanced models

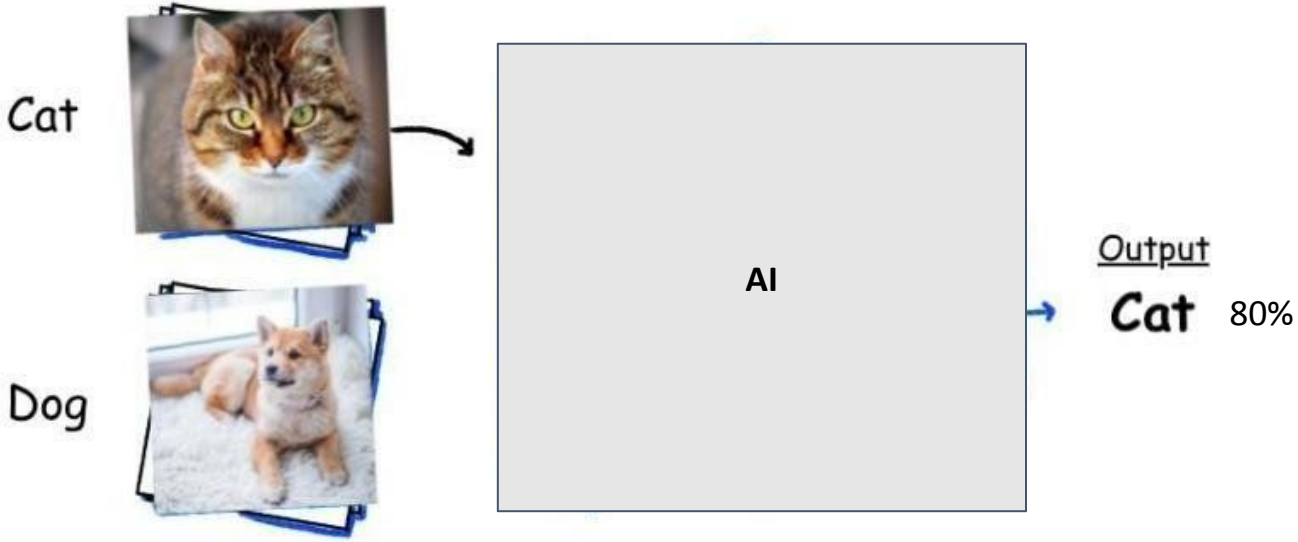


Large language models with transformers

UNIVERSITÉ DE MONTPELLIER

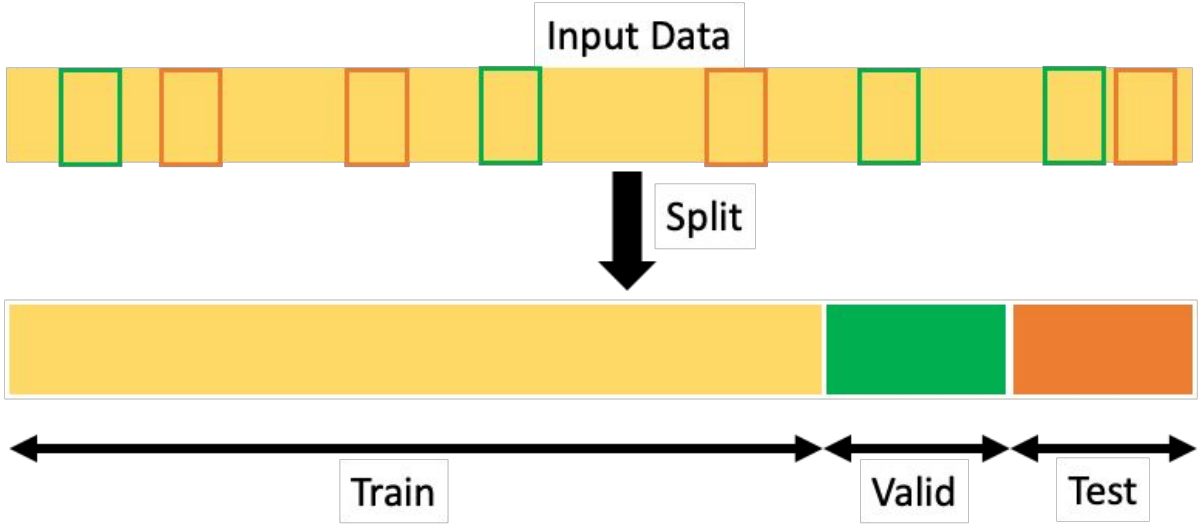
The quest to **classify** objects

The main goal for decades and AI was to train machine to say a YES or NO
Provide probabilities from a list of possibilities



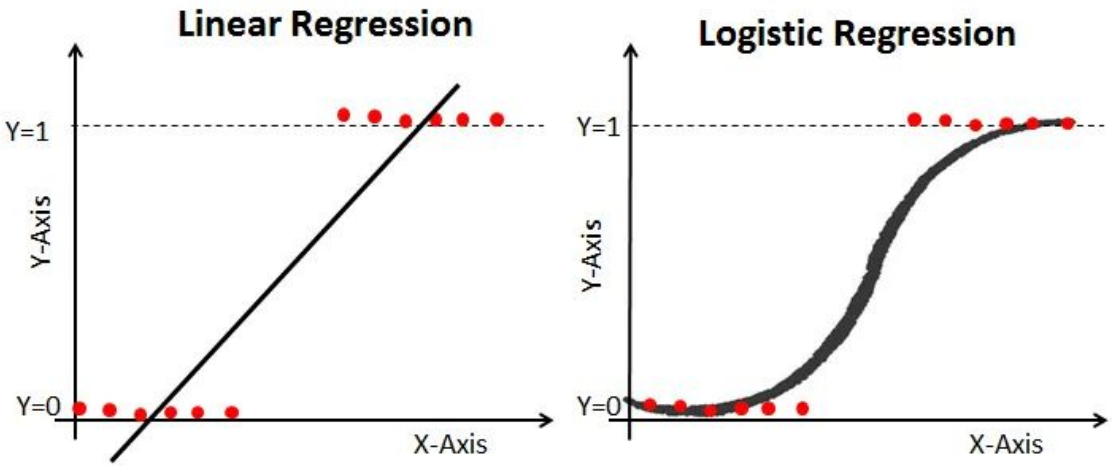
A common methodology

Train, validate and test your model with splitted datasets



Optimization with limited computing and data

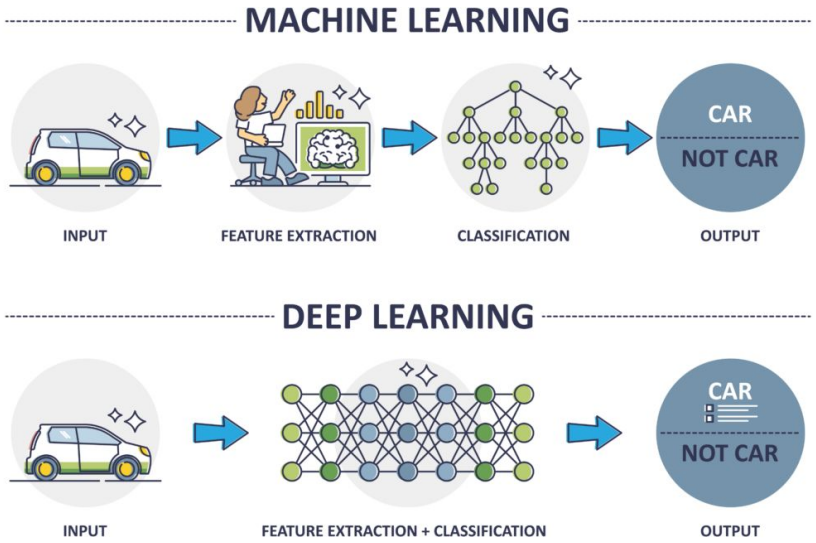
The beginning of machine learning was using mathematics and limited amount of data/computing to made optimized prediction



Look for the best mathematical model to fit your data

Neural network and high computing

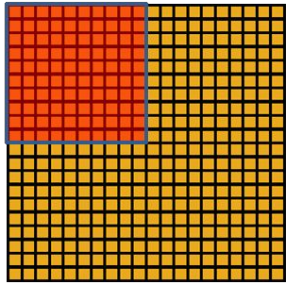
With a lot of data and computing, neural networks find accurate patterns to predict and classify data



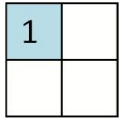
UNIVERSITÉ DE MONTPELLIER

Neural network and convolution

With a lot of data and computing, neural networks find accurate patterns with aggregation of layers with summarized informations



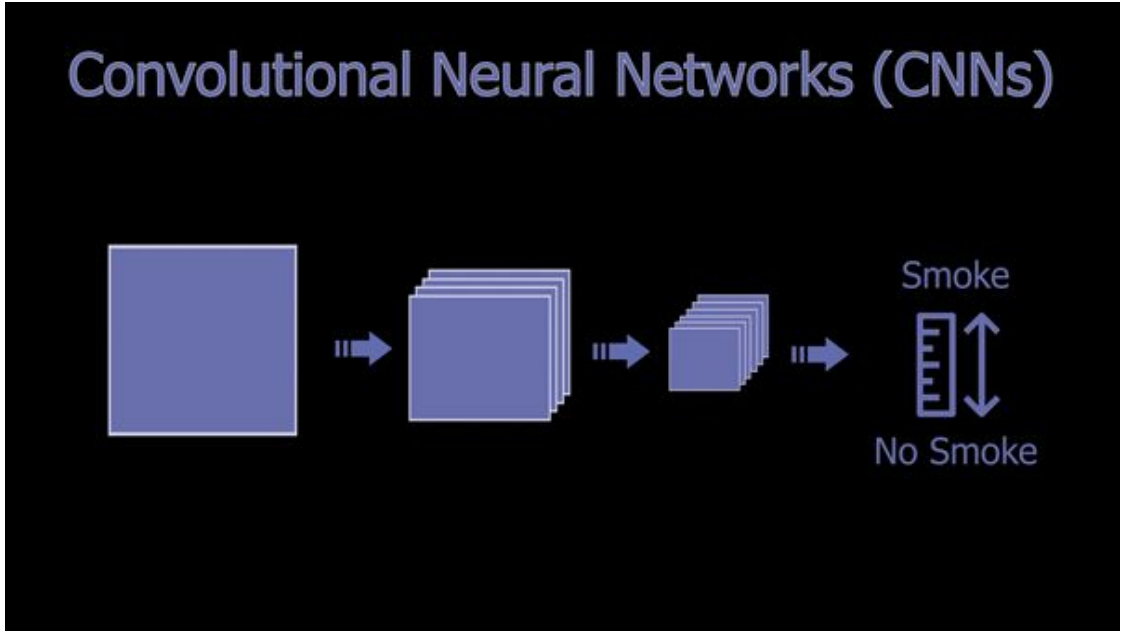
Convolved feature



Pooled feature

Neural network and classification

With a lot of data and computing, neural networks find accurate patterns with aggregation of layers with summarized informations

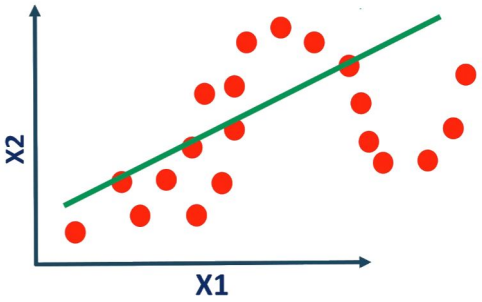


Neural network and training

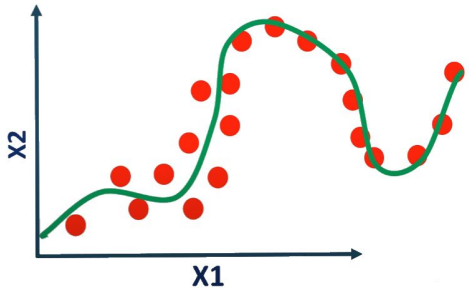
With a lot of data and computing, neural networks find accurate patterns with aggregation of layers with summarized informations <u>after a lot of iteration</u>

Rerun the training, keep what makes improve the results, x times !

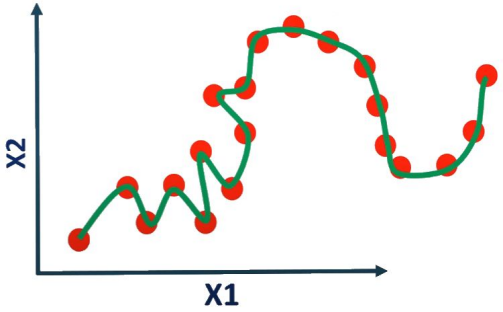
Epoch 1



Epoch 10



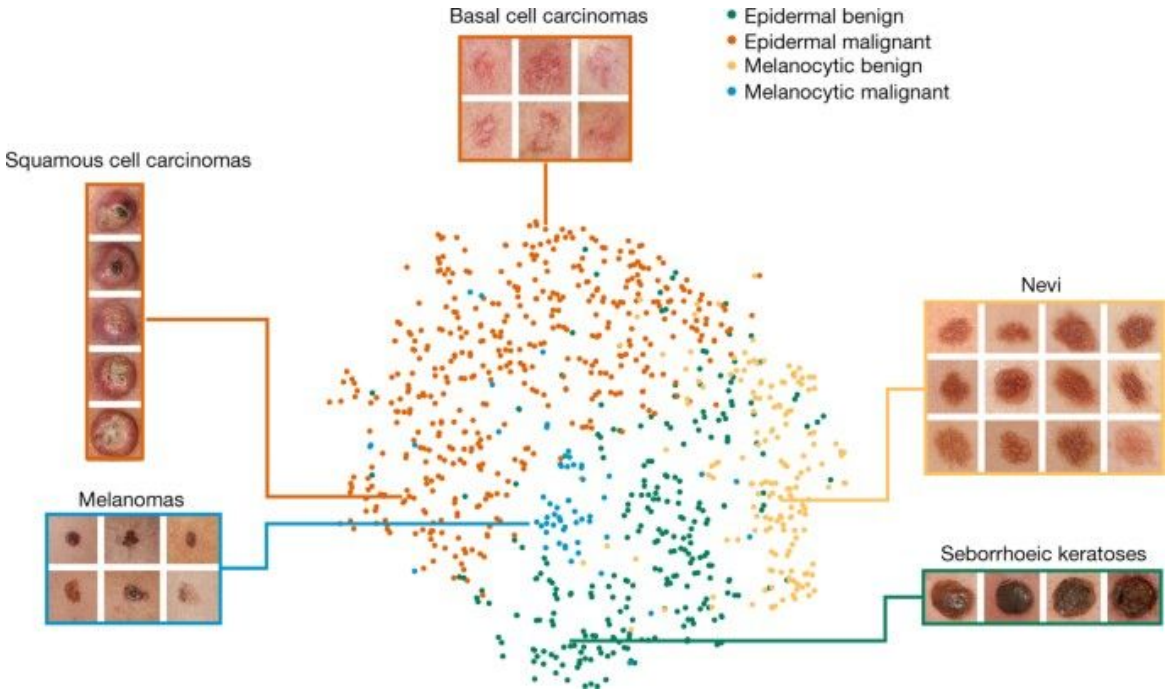
Epoch 50



overfitting

Deep learning in medicine

To do so, we ask human to classify object and we need a lot of them



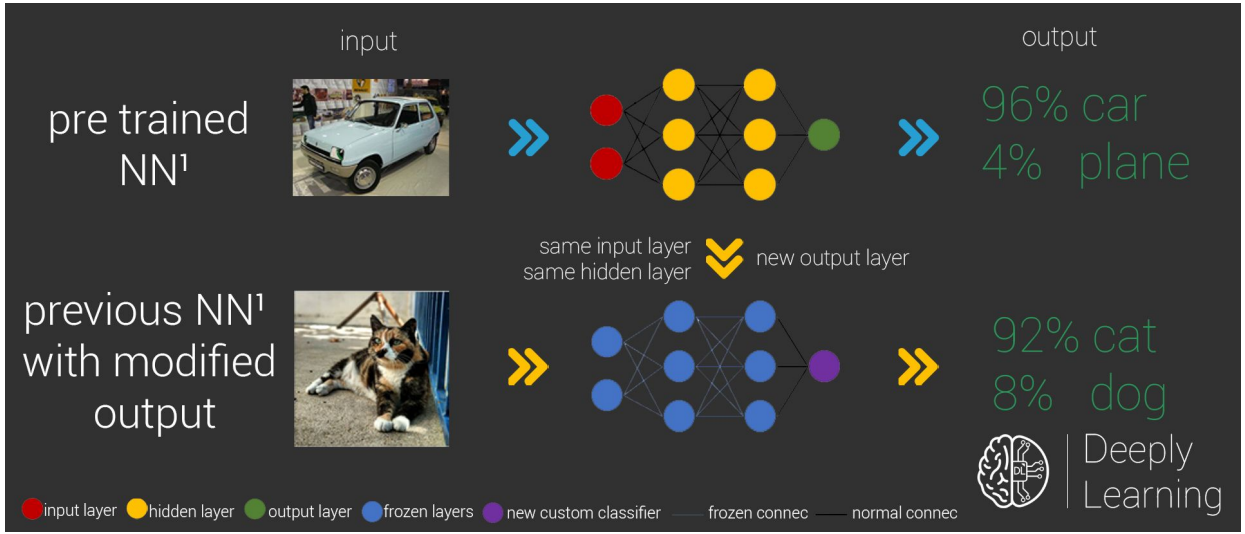
Esteva *et al*. used 129,450 clinical images of skin disease to train a deep convolutional neural network to classify skin lesions

(Nature 2017)

UNIVERSITÉ DE MONTPELLIER

Transfer learning

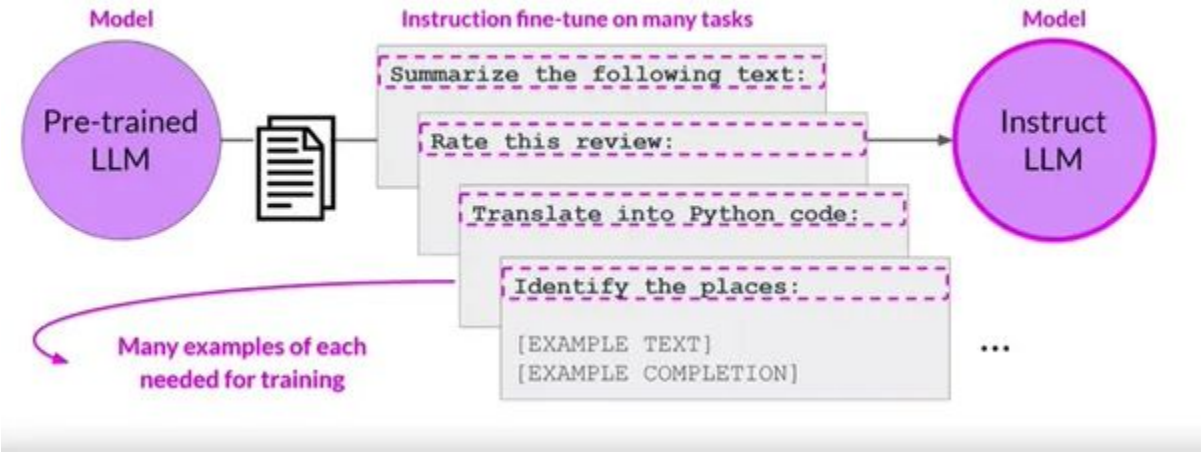
Thanks to transfer learning, applications of AI have spreaded



UNIVERSITÉ DE MONTPELLIER

Fine tuning

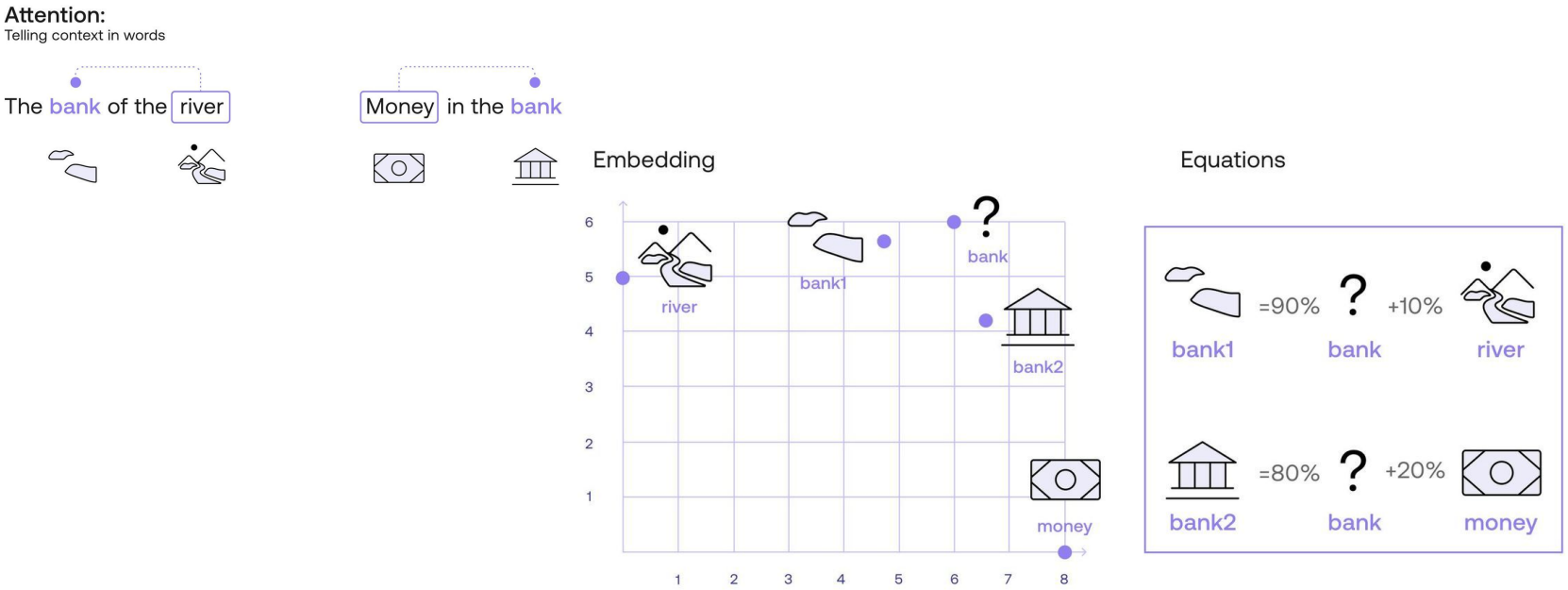
Thanks to fine tuning, applications of AI continuously improve



UNIVERSITÉ DE MONTPELLIER

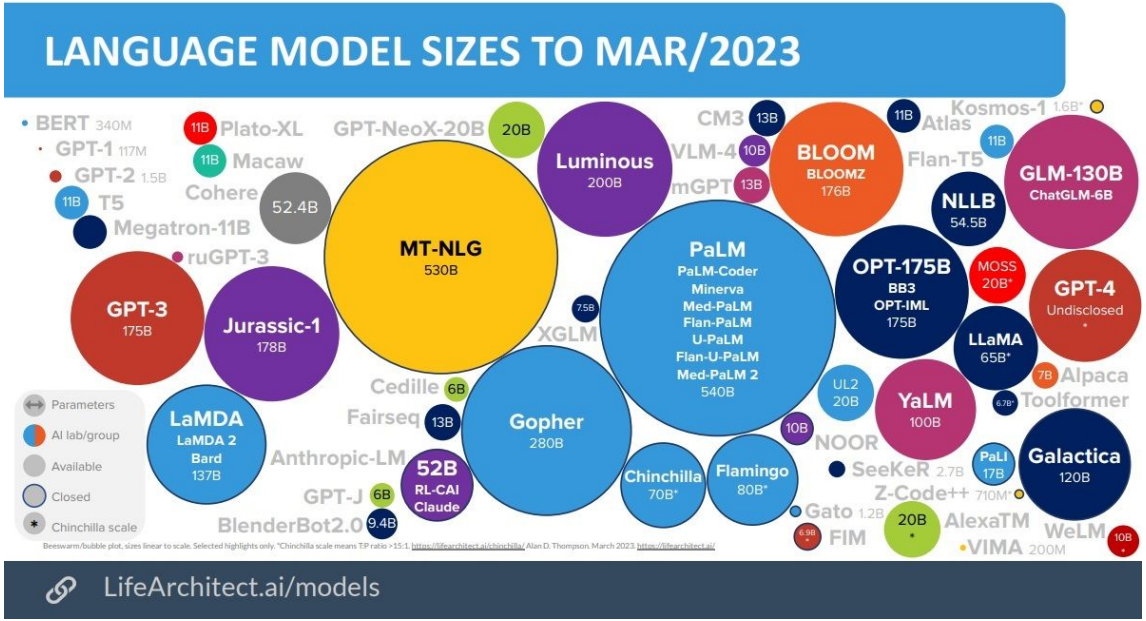
From the era of classifying AI to generative AI

Able to manage an overall context, Transformer model provided the generative era



Large Language Model and GPT

Able to manage an overall context, Transformer model provided the generative era



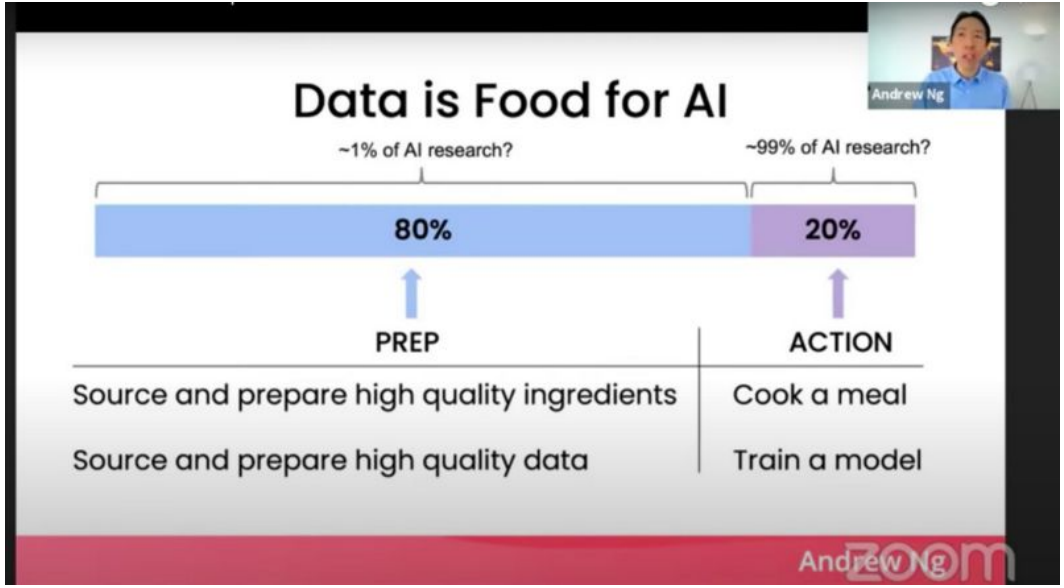
Cohere LLM University

UNIVERSITÉ DE MONTPELLIER

My guess as a Data Scientist

It’s all about data quality and not data quantity

From structured data will shine creativity and new applications



UNIVERSITÉ DE MONTPELLIER

My guess as a Data Scientist

Choose the right model for your data

Few data/sparse data

Machine learning

Texts

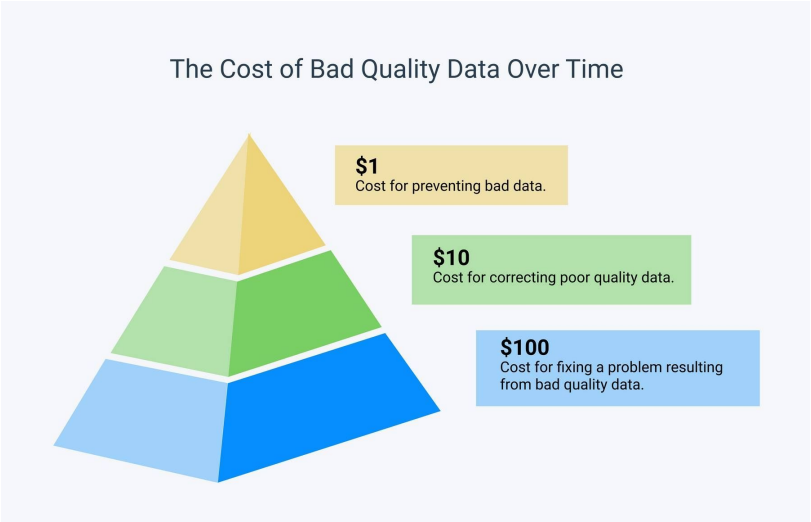
LLM

Lot of data / task specific

Deep learning

My wish as a Data Scientist

Learn minimum skills on how to structure your data if you want AI and/or a data scientist to help you !



UNIVERSITÉ DE MONTPELLIER

UNIVERSITÉ DE MONTPELLIER

AI applied to medicine
The genomic medicine example

Kévin Yauy
Physician-Scientist Fellow, CHU Montpellier
MD in Medical Genetics
PhD in Bioinformatics and Machine Learning

CHU MONTPELLIER CENTRE HOSPITALIER UNIVERSITAIRE

Université de Montpellier FACULTÉ de MÉDECINE Montpellier-Nîmes

Summary

1. Introduction to genomic medicine
2. Application of AI in genomic medicine
3. LLM for (genomic) medicine education

1. Genomic **medicine ?**

Call for ideas

What is the impact of a genetic diagnosis ?

Let me know !

Rare disease diagnostic

The diagnostic odyssea

- About **3 Million** people affected in France
- Still **~50%** patients with no diagnosis

The diagnostic challenge

- There are more than **10000** rare diseases
- A geneticist only has **1** brain

Linkeropathies

Ciliopathies

Kabuki Syndrome

Familial mediterranean fever

22q11.2 Syndrome

Williams-Beuren Syndrome

Impact in patient care

Diagnosis:
- Society issue, disability
- Loneliness, guilty
- Medical aid if genetic condition

Prognosis :
- Known associated complications
- Provide degree of severity

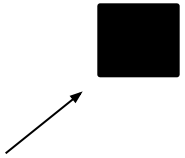
Therapeutics :
- Target therapies
- Pharmacogenomics

Family :
- Shared informations to your family
- Pre-symptomatic diagnosis

Reproduction :
- Prenatal diagnosis
- Pre-implantation diagnosis
- Carrier screening

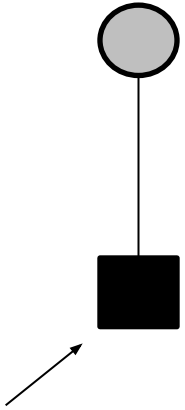
Cas index

- Enfant 3 ans 1/2
  - Maîtresse scolaire vous informe des difficultés
    - Difficultés de concentration
    - Difficultés de compréhension
    - Perturbateur dans la classe
  - 2 crises d’épilepsie fébrile
  - Marche 19 mois
  - Hyperactif
  - WISC4 = QIT 55



Informations familiale

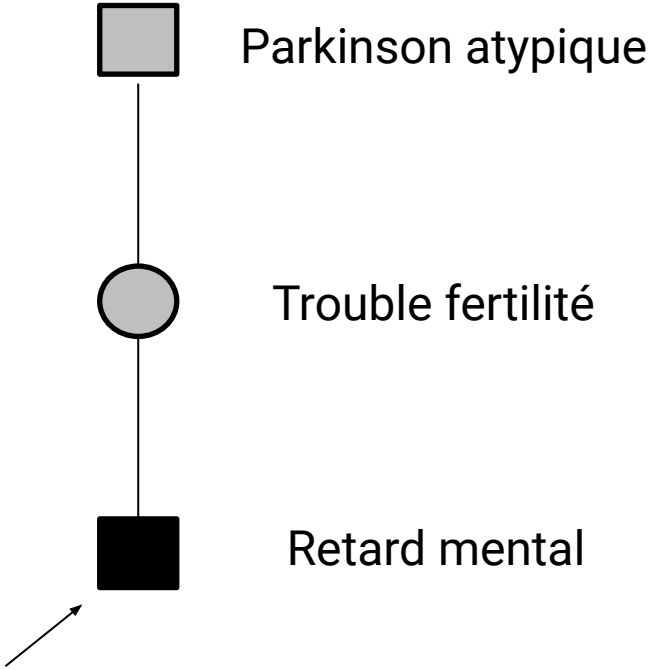
- Maman
  - seul enfant
  - Difficulté procréation



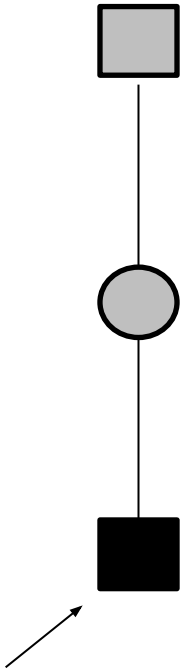
Informations familiale

- Maman
  - seul enfant
  - Difficulté procréation

- Grand père maternel ataxie et Parkinson atypique



Les vrais Diagnostics

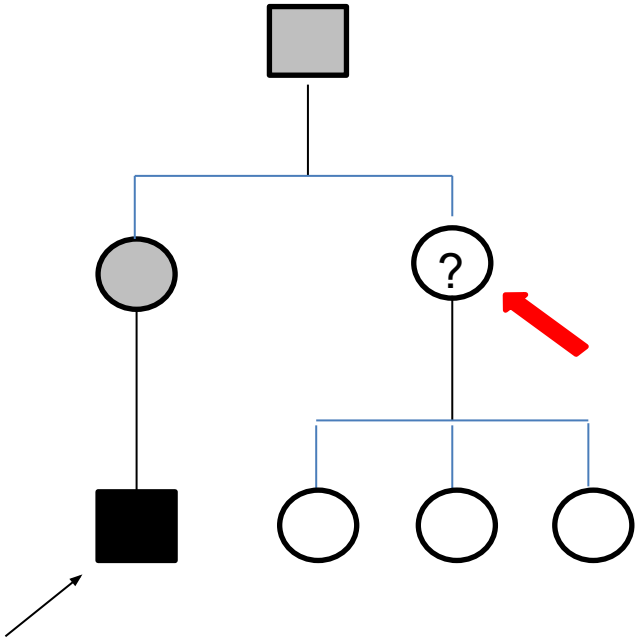


Fragile X Tremor ataxia syndrome
FXTAS

Insuffisance ovarienne précoce
FXPOF/FXPOI

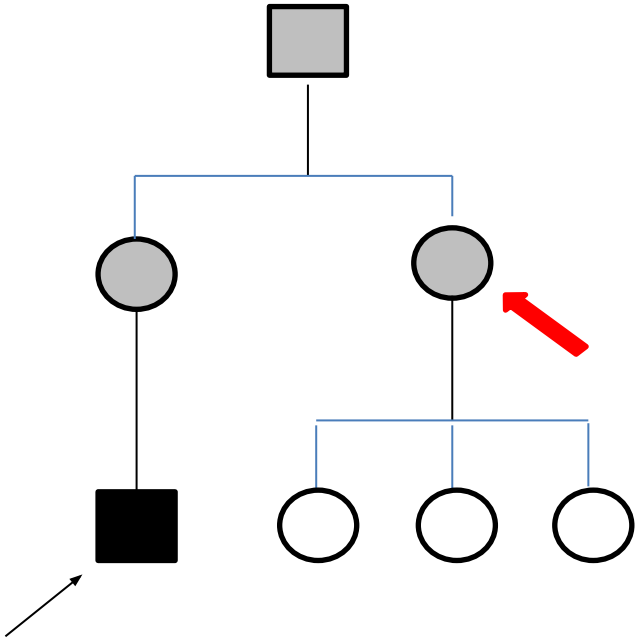
Syndrome de l'X Fragile

Quel conseil génétique donner ?



A Indemne
B Doit faire le teste
C Pas de risque pour ses filles
D Conductrice obligatoire

Quel conseil donner ?

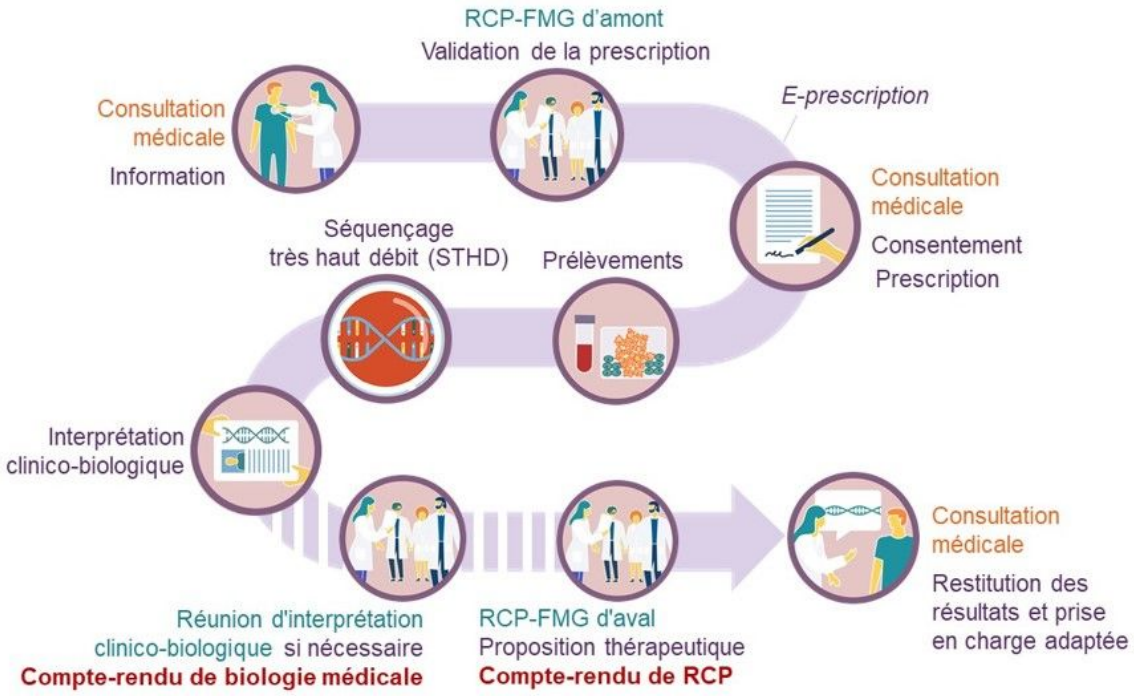


Conductrice obligatoire

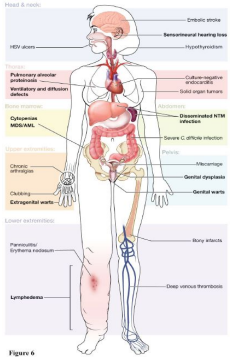
2. Application of AI in genomic medicine

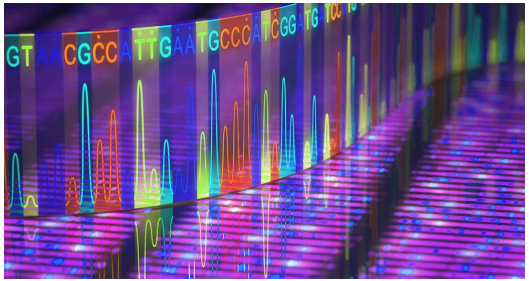
Genome sequencing era in France

Yes, we prescribe complete genome sequencing in routine, since 2019 !



Diagnostic tools for the doctors





1. Recognition of "gestalt" or
2. association of symptoms : "phenotype"

3. DNA analysis/sequencing : "genotype"

Physicians select a targeted sequencing analysis according to their knowledge

Diagnostic tools for the doctors in the genome era

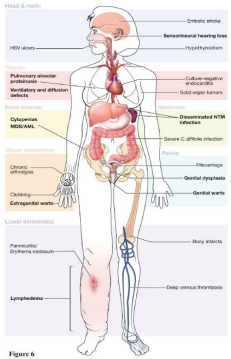
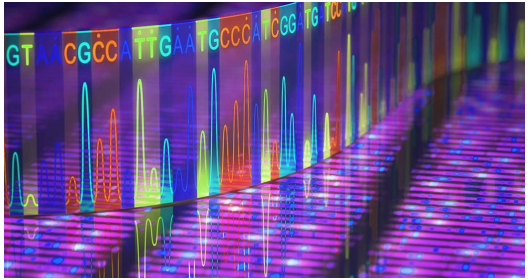


Figure 6



1. Recognition of "gestalt" or
2. association of symptoms : "phenotype"

3. DNA analysis/sequencing : "genotype" - millions of variants

=> How AI can help medical geneticists in their clinical practice ? Three examples

A.

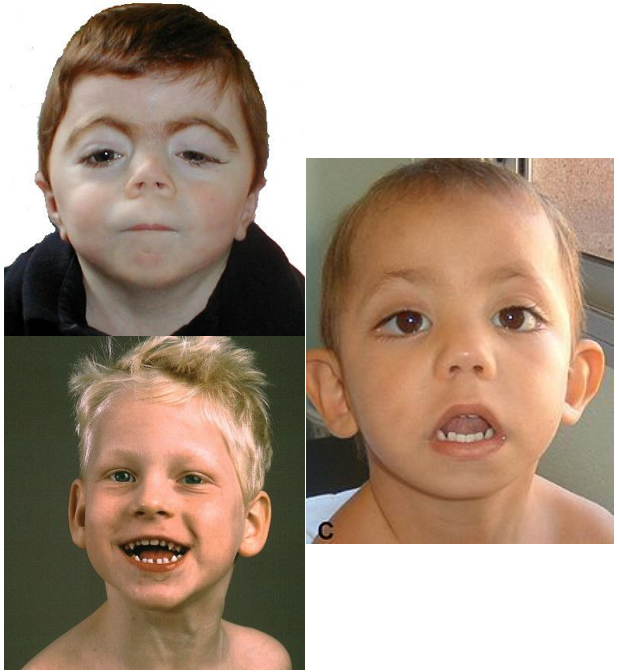
Facial recognition of "gestalt"

(Deep Learning)

Facial recognition of "gestalt"

- One of the main skills before sequencing (200-300 gestalts to)

- Limitation :
- **“Subjective” skills**
- Hard to learn



Syndrome Kabuki

Deep learning & gestalts

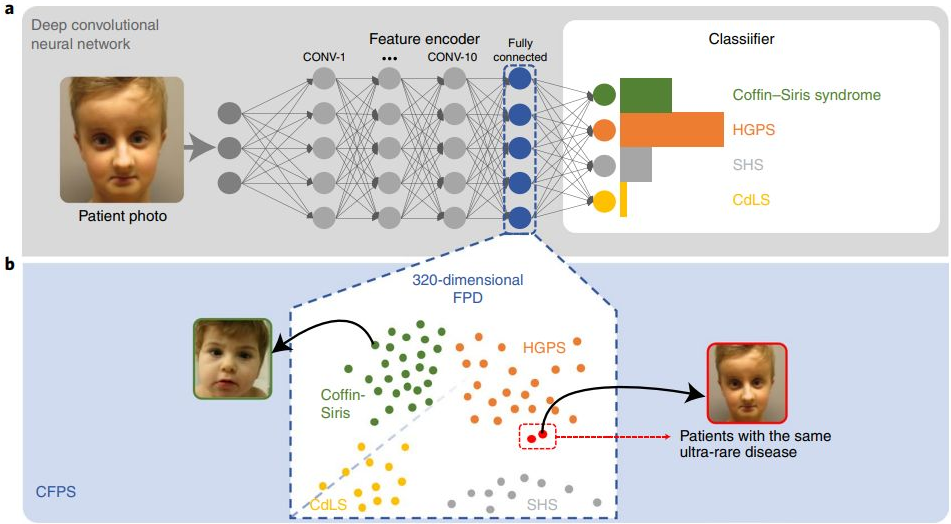
nature genetics

TECHNICAL REPORT

https://doi.org/10.1038/s41588-021-01010-x

Check for updates

GestaltMatcher facilitates rare disease matching using facial phenotype descriptors



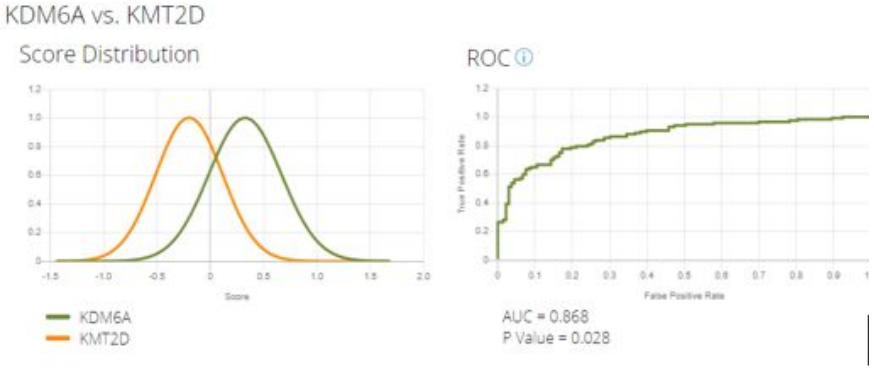
Deep learning & gestalts

Deep learning & gestalts

GestaltMatcher

Find new gestalts:
=> Kabuki 1 et 2 syndrome : no distinguishable gestalts described

AI can find patterns with interpretable results



Rouxel, Yauy* et al. EJHG. (2022)*

B.

Association of **Symptomes** (Machine Learning)

Physicians needs computer helps

Clinical geneticists were early adopters of software as clinical decision support

"For precision medicine human and artificial intelligence need to join efforts." (Peter Krawitz)

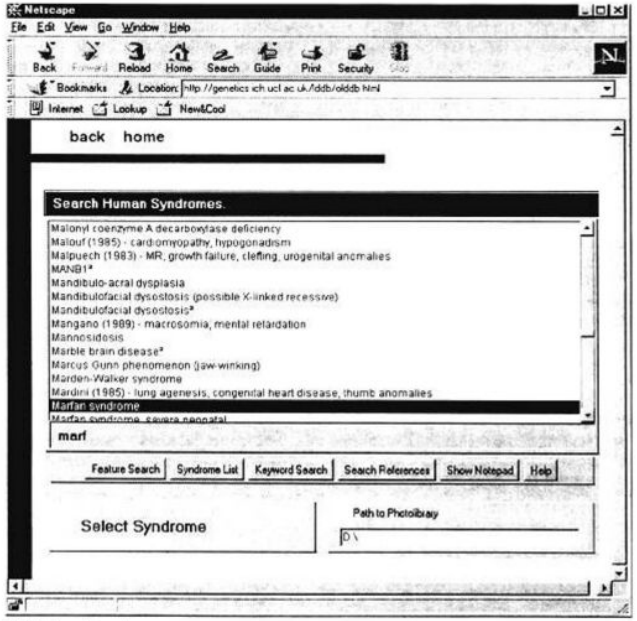


Fig. 1 Syndrome list search screen.

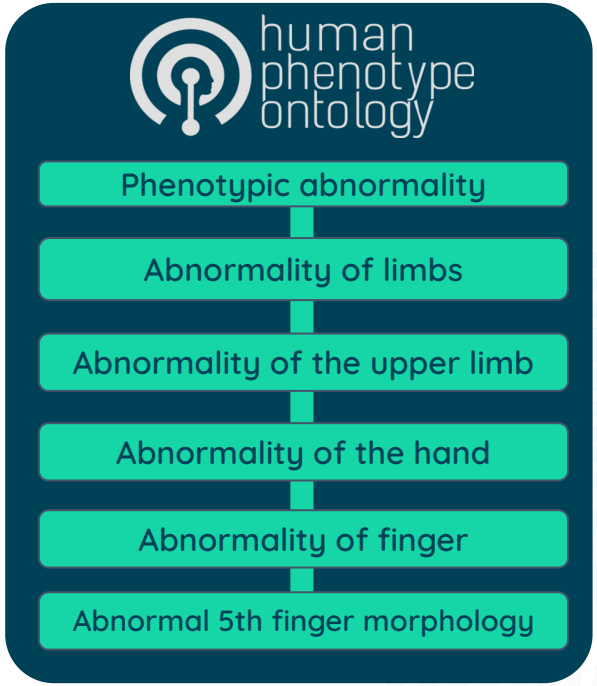
Need for computational phenotype analysis

Phenotyping :

Physicians identification of characteristics deviating from normal morphology, physiology, and behavior

Ontology :

Standardized symptom terms linked according to the human development architecture



A common language between human and machine is necessary for computer support

Computational phenotype analysis :

Identification of diagnostic hypothesis, clinically relevant groups of patients, ...

Bottleneck : “fuzzy” phenotypic profiles

Phenotyping is reported to be “fuzzy”

- Heterogeneity in phenotyping
- Variable expressivity in diseases
- Undescribed associations

=> Studies about phenotyping practices in clinical sequencing are lacking

Haendel et al. NEJM 2018

Wiedemann–Steiner Syndrome Profile	Patient 1 Profile (3-year-old girl)	Patient 2 Profile (14-year-old boy)
DIGITS		
Short toe HP:0001831	○ None	◒ Long toe HP:0010511
Short middle phalanx of finger HP:0005819	◒ Cone-shaped epiphysis of the phalanges of the hand HP:0010230	○ None
DEVELOPMENT		
Developmental delay in speech and language HP:0000750	● Developmental delay in speech and language HP:0000750	● Developmental delay in speech and language HP:0000750
Intellectual disability HP:0001249	◒ Global developmental delay HP:0001263	◒ Global developmental delay HP:0001263
SKELETAL		
Microcephaly HP:0000252	● Microcephaly HP:0000252	◒ Macrocephaly HP:0000256
Short stature HP:0004322	◒ Proportionate short stature HP:0003508	○ None
FACIAL		
Thin upper lip HP:0000219	◒ Thick upper lip HP:0000215	○ None
Hypertelorism HP:0000316	● Hypertelorism HP:0000316	○ None
Blepharophimosis HP:0000581	○ None	● Blepharophimosis HP:0000581
Epicanthus HP:0000286	○ None	◒ Epicanthus inversus HP:0000537

Heterogeneous Phenotyping

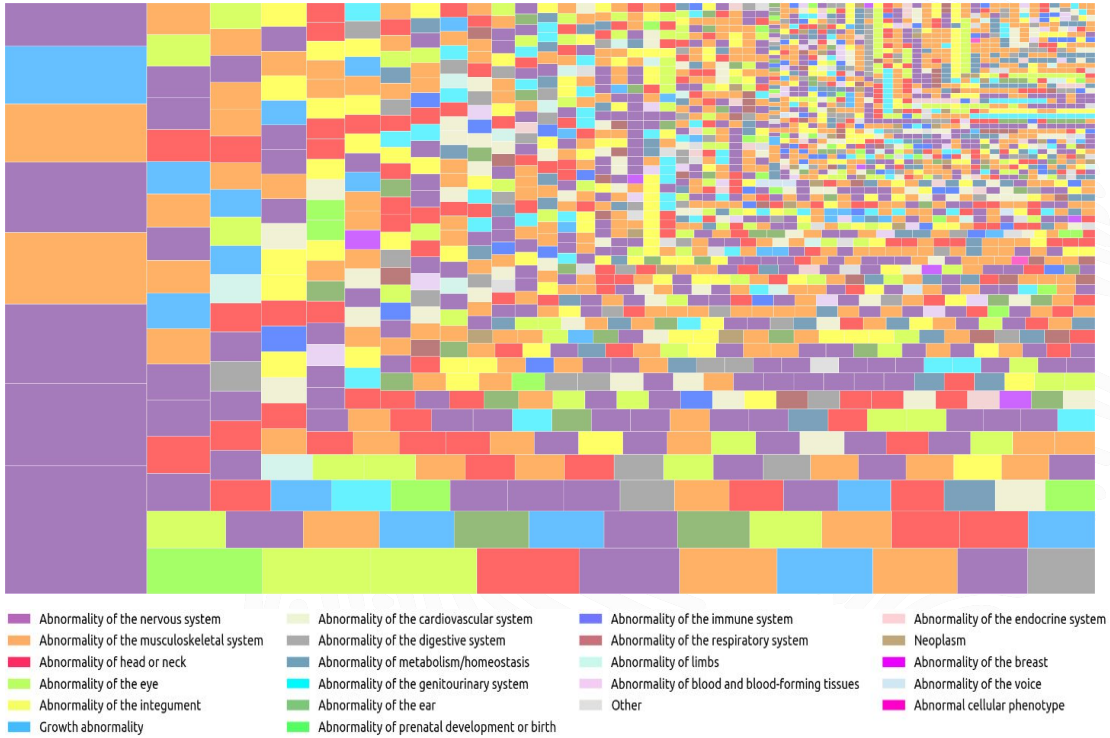
- Phenotypes described in a cohort of **1686 patients**
- **47%** of HPO terms declared only once

Heterogeneity because :

Clinical examination variability?

Physicians phenotyping diversity?

PhenoGenius consortium
Peng *et al.*, NAR Genom Bioinform (2021)
Seo *et al.*, Clin. Genet (2020)
Trujillano *et al.*, *EJHG* (2017)



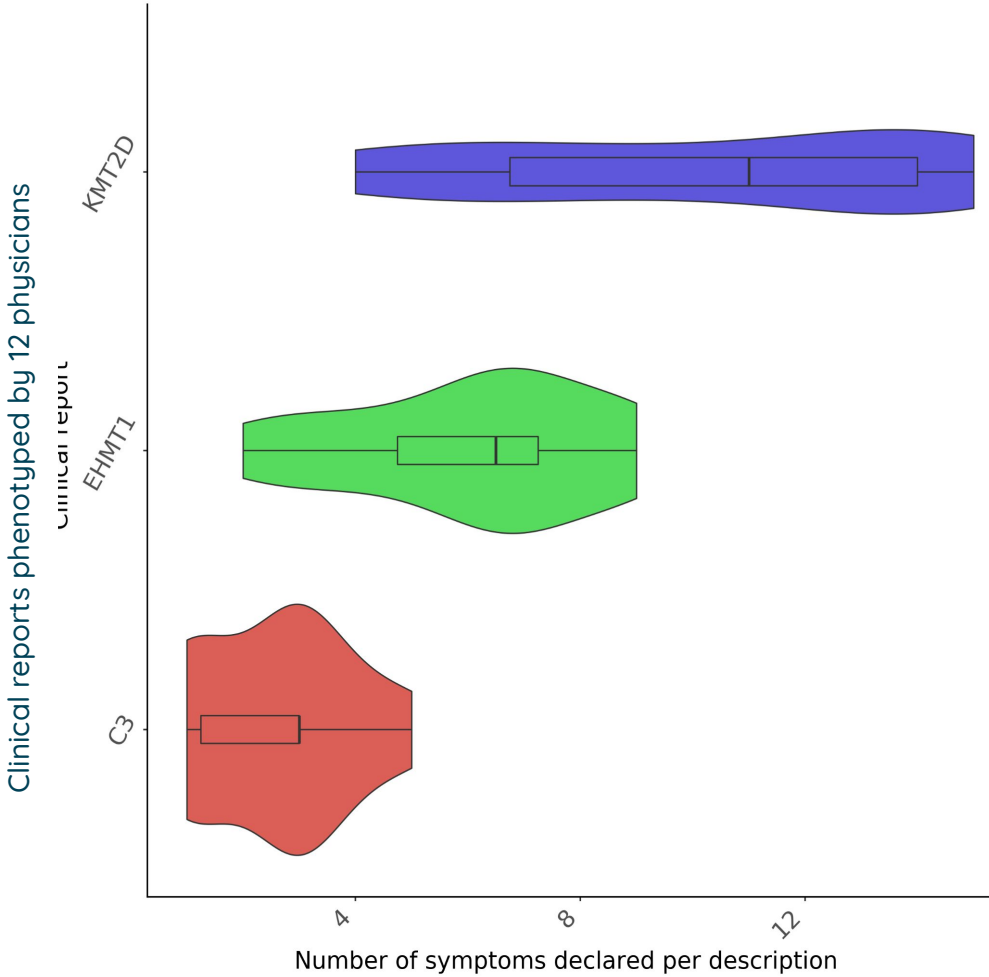
Heterogeneity within reports

Three clinical reports described by 12 physicians

EHMT1 Example

- 29 different terms
- 17 used by 2 or more physicians
- none mentioned by all

Physicians phenotyping diversity explains the observed heterogeneity



AFGC
Association Francophone de Génétique Clinique

Phenotyping unknown associations

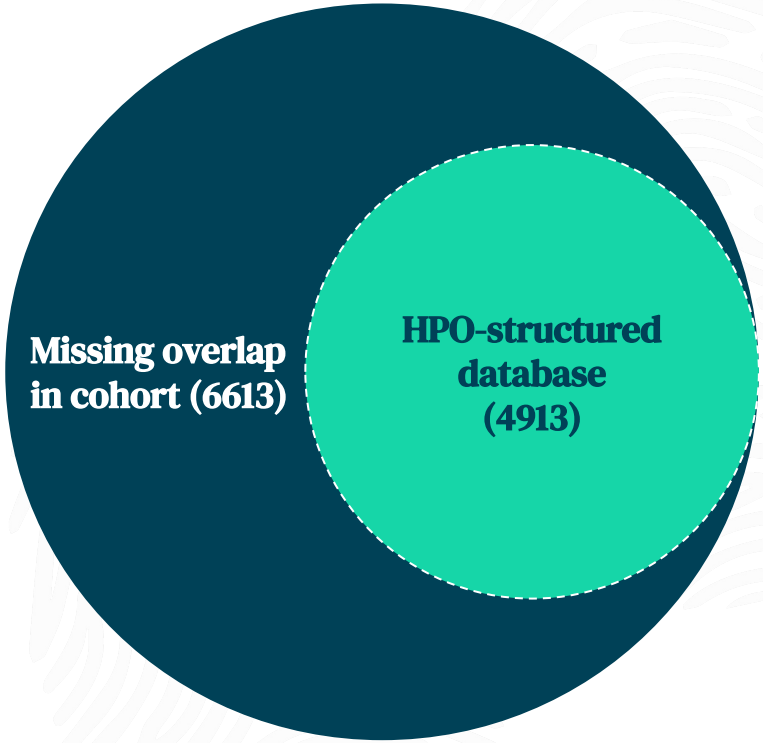
Cohort
11,526 unique symptom-gene associations

HPO-structured database
734,931 unique symptom-gene associations

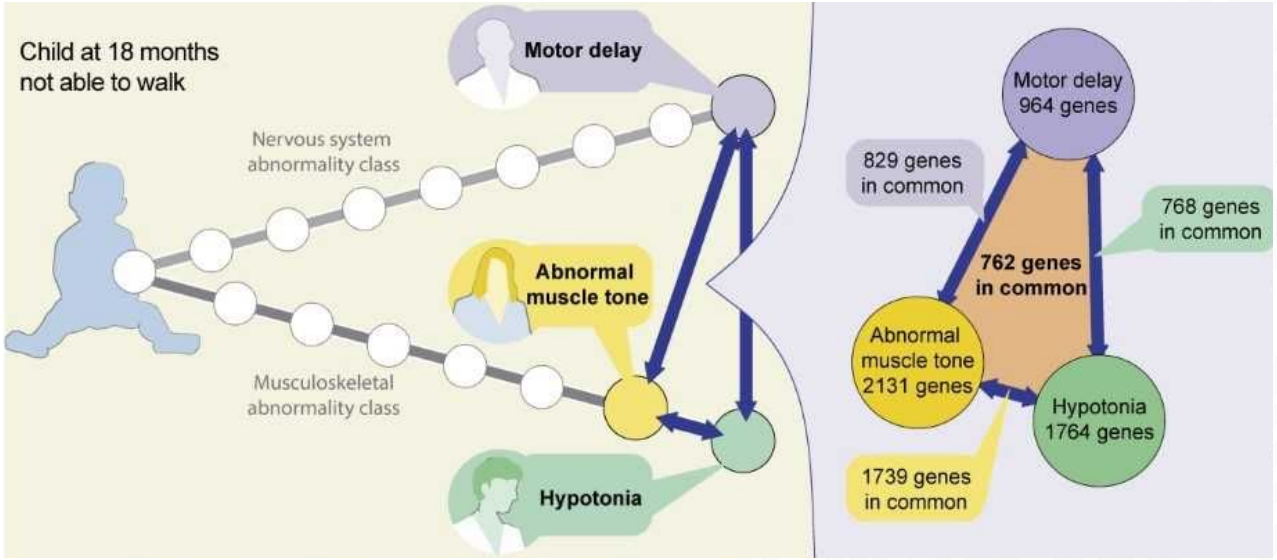
57% symptom-gene association were missing

=> Unexploited information for computational phenotype analysis

=> How to handle physicians' heterogeneous phenotyping ?



Multiple ways to describe patients



Physicians acquire intricate cognitive frameworks to solve diagnostic problems

→ Link between symptoms are very different from Human Phenotype Ontology structure

Inductive reasoning

Microcephaly (HP:0000252)

Delayed ability to walk (HP:0031936)

→

Disease /Phenotypic patterns I know

→

Intellectual deficiency (HP:0031936)

Shin. J Med Educ 2019

Inductive reasoning… through modeling

- Superhero
- Action
- US
- Robert Downey Jr

- Child
- Anime
- UK
- Animals

NETFLIX

Recommendation algorithms are really efficient

Find symptom-gene associations

	Unstructured data			Structured databases			
	Text-matching via elasticsearch *			Merge all HPO-gene association available			HP:0000006 Probability
Gene	OMIM	PubMed	NCBI National Center for Biotechnology Information	The Jackson Laboratory	orphanet	G2P	VOTE
BRCA1	1	1	0	0			0.5

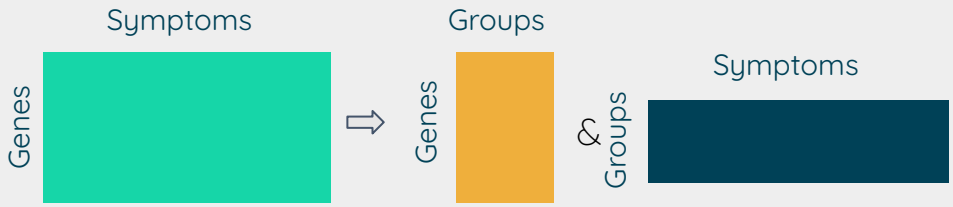
$$
\begin{array}{c}
 \\
\text{Gene 1} \\
\text{Gene 2} \\
\dots \\
\text{Gene 4531}
\end{array}
\begin{array}{c}
\begin{array}{ccccccc} \text{HP01} & \text{HP02} & \text{HP03} & \dots & \dots & \dots & \text{HP15785} \end{array} \\
\begin{pmatrix}
0.23 & 0.41 & 0 & \dots & \dots & \dots & 0 \\
0 & 0.21 & 0.32 & \dots & \dots & \dots & 0.42 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots \\
0.11 & 0.27 & 0.42 & \dots & \dots & \dots & 0.42
\end{pmatrix}
\end{array}
$$

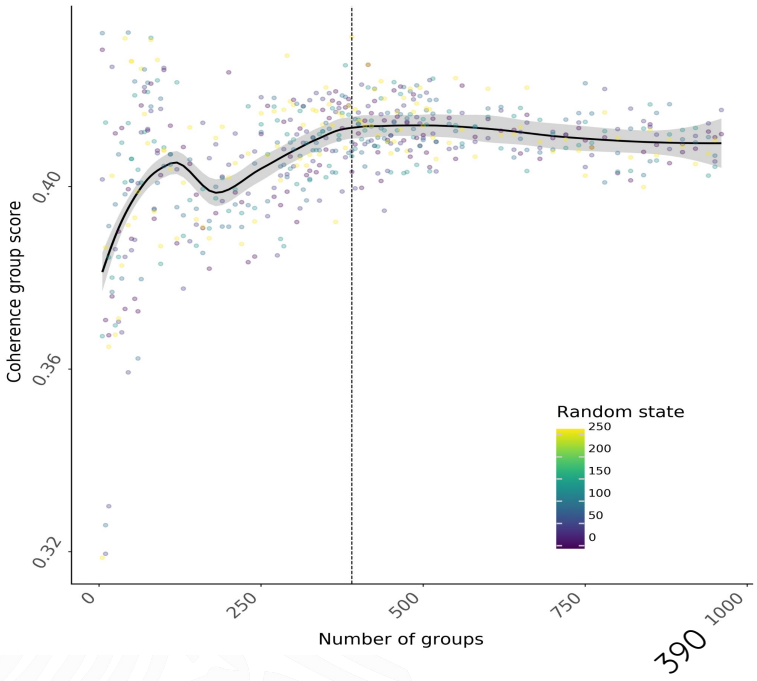
Building a gene-symptom association matrix with ~16000 symptoms

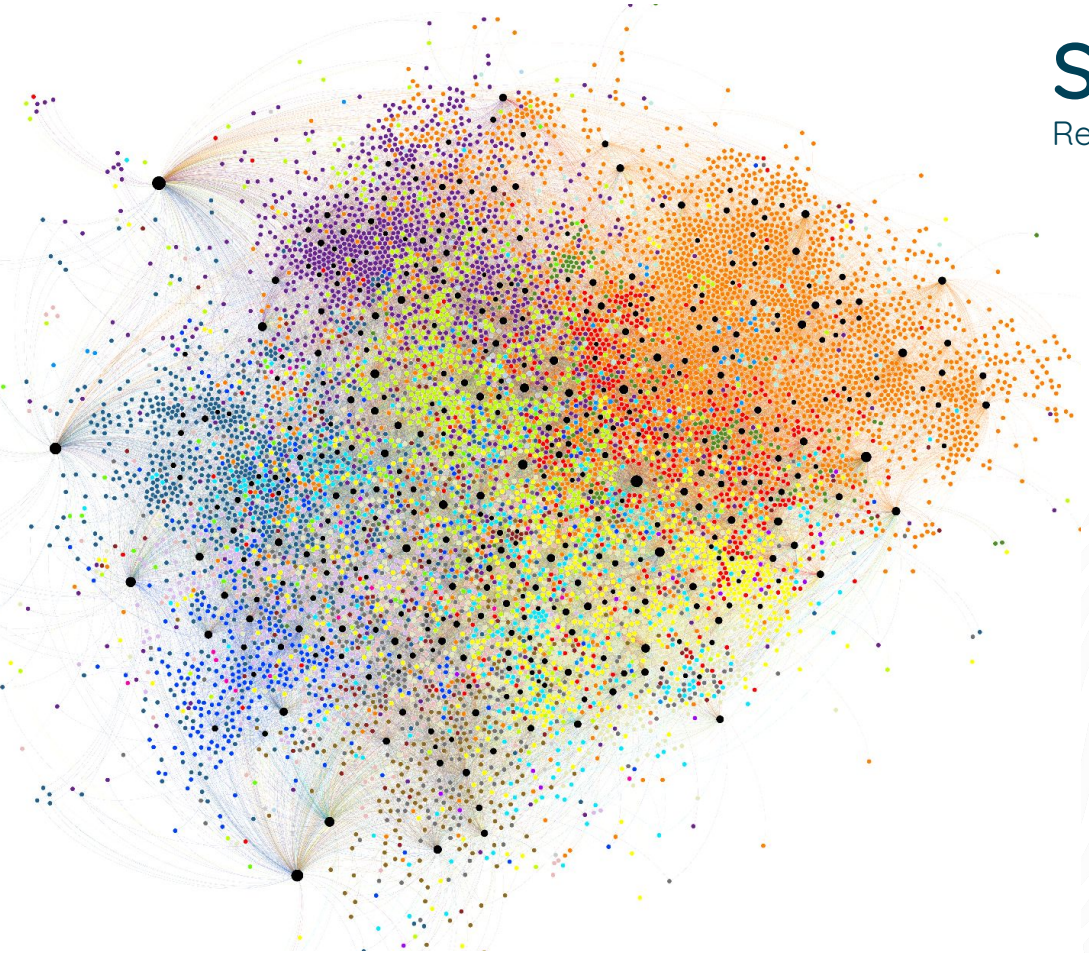
Extract meaningful groups of symptoms

Non-Negative Matrix Factorization

Dimensionally reduces 16,600 symptoms to 390 groups of symptoms according to their current genetic relatedness







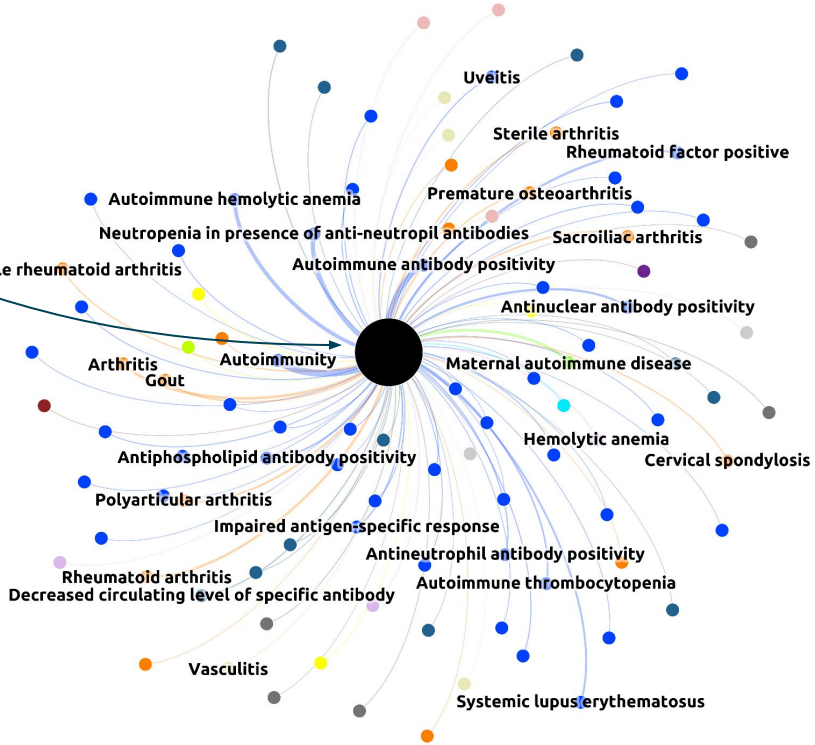
Symptom-gene graph

Retrieving Symptom-gene associations with graphs

- **390 groups** of symptoms (n=43,308, 10% of symptom-group associations)
- 5,971,755 pairs of symptoms
- **3,222,053 additional** NMF-based symptom-gene associations
- **“only” 2% cohort symptom-gene association missing !**

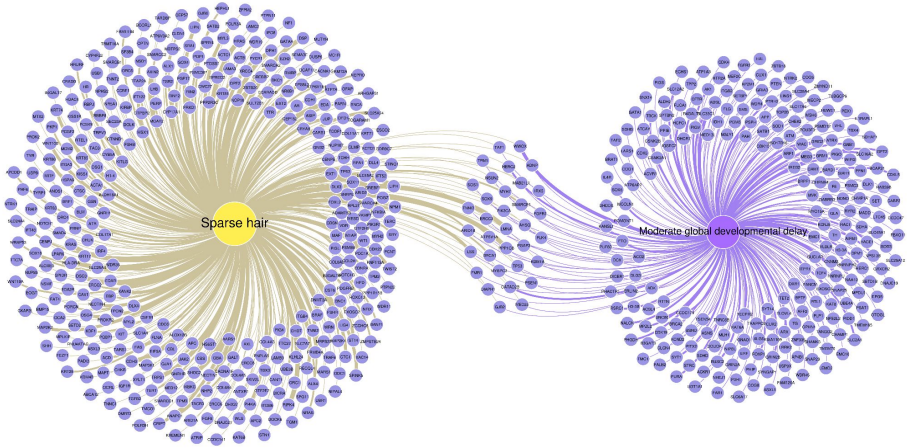
An example: Autoimmunity group

- <u>Main symptom :</u> Autoimmunity
- 99 HPOs included
- 14 different HPO classes
- 545 genes associated to the group

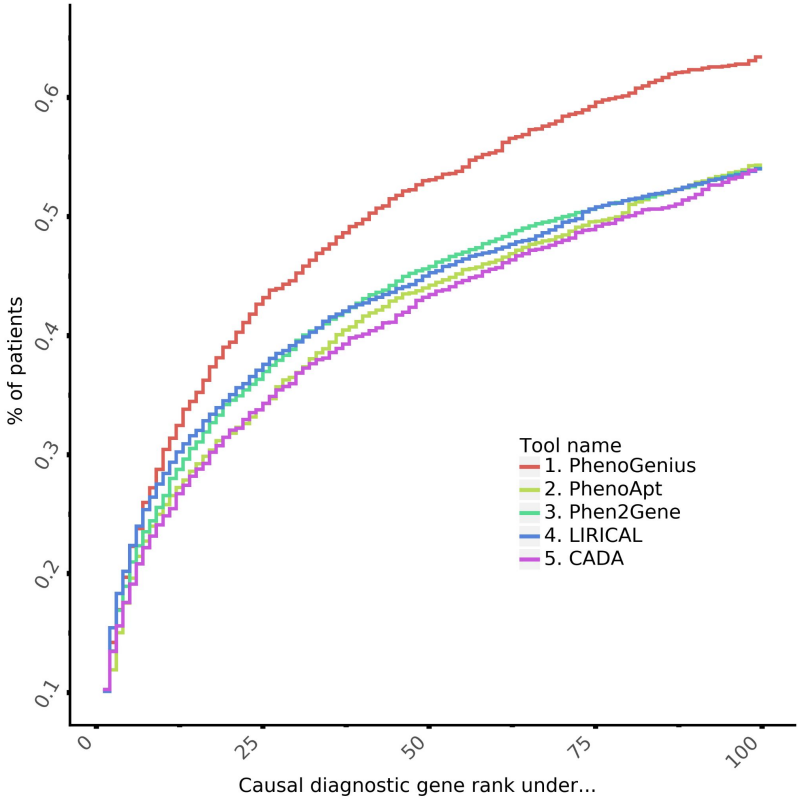


Q. Is it clinically relevant ?

Diagnostic gene ranking experiments on 1,686 patients



Improved gene prioritization with symptom interactions



Using symptom interactions improved the diagnostic performance in gene prioritization by 42 %.

Median rank of diagnostic gene

- **80 with Phen2Gene**
- **41 with PhenoGenius**

Diagnostic gene rank benchmark:

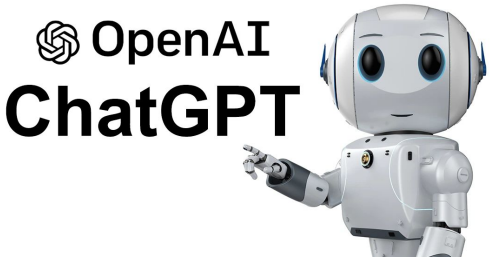
4 state of the arts software with different methodologies :

PhenoApt (Chen *et al. AJHG* 2022)
Phen2Gene (Zhao *et al. NAR* GB 2020)
LIRICAL (Robinson *et al. AJHG* 2020)
CADA (Peng *et al. NAR GB* 2021)

Conclusion

Such a great era for new scientists !

Be part of it !



Thanks for your attention !

Questions ?
kevin.yauy@chu-montpellier.fr