Intelligence Artificielle: Apports en génomique

Prospective en Science des Données, IA et Biologie
2/12/2020

Chloé-Agathe Azencott
Centre for Computational Biology, Mines ParisTech
`chloe-agathe.azencott@mines-paristech.fr`
`@cazencott`

~~Intelligence Artificielle~~ Machine Learning: Apports en génomique

Prospective en Science des Données, IA et Biologie
2/12/2020

Chloé-Agathe Azencott
Centre for Computational Biology, Mines ParisTech
chloe-agathe.azencott@mines-paristech.fr
@cazencott

Inserm
Institut national de la santé et de la recherche médicale

institut Curie

PR[AI]RIE
PaRis Artificial Intelligence Research InstitutE

MINES ParisTech | PSL

Supervised Machine Learning

- **Data:**
  - n **samples** (from the same distribution) X
    - Tabular data
    - Images, sequences, graphs
  - their n **labels** y
    - A single categorical/qualitative or continuous/quantitative value

Supervised Machine Learning

- **Data:**
  - n **samples** (from the same distribution) X
    - Tabular data
    - Images, sequences, graphs
  - their n **labels** y
    - A single categorical/qualitative or continuous/quantitative value
- **Questions:**
  - Predict y from X (**regression**/**classification**)

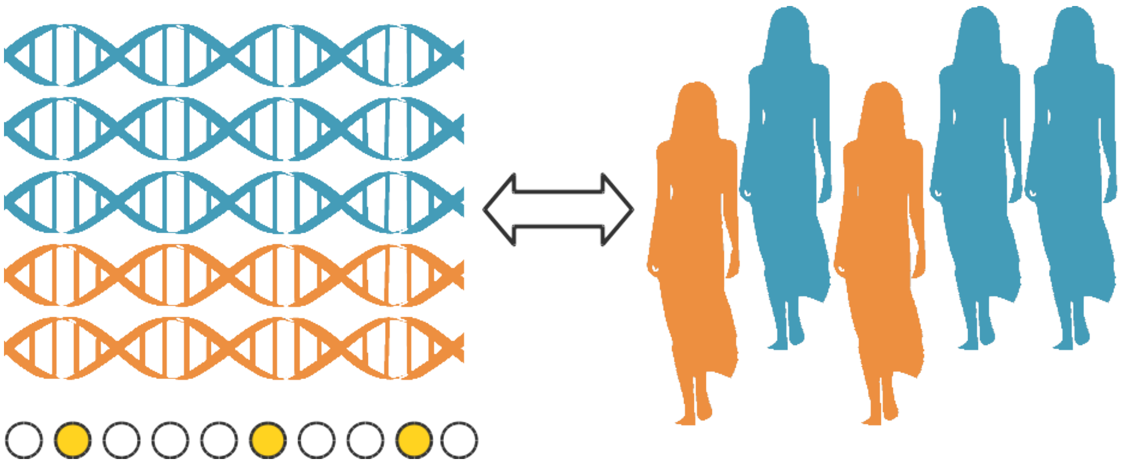
    General idea: find a model that **minimizes** (more or less accurately) a **loss** on the training data (+ some constraints)
  - Understand which features from X make it possible to predict y (**feature selection** & **interpretable models**)

Supervised Machine Learning

- **Data:**
  - n **samples** (from the same distribution) X
    - Tabular data
    - Images, sequences, graphs
  - their n **labels** y
    - A single categorical/qualitative or continuous/quantitative value
- **Questions: DESCRIBE and UNDERSTAND**
  - Predict y from X (**regression**/**classification**)

    General idea: find a model that **minimizes** (more or less accurately) a **loss** on the training data (+ some constraints)
  - Understand which features from X make it possible to predict y (**feature selection** & **interpretable models**)

Example 1: Biomarker discovery

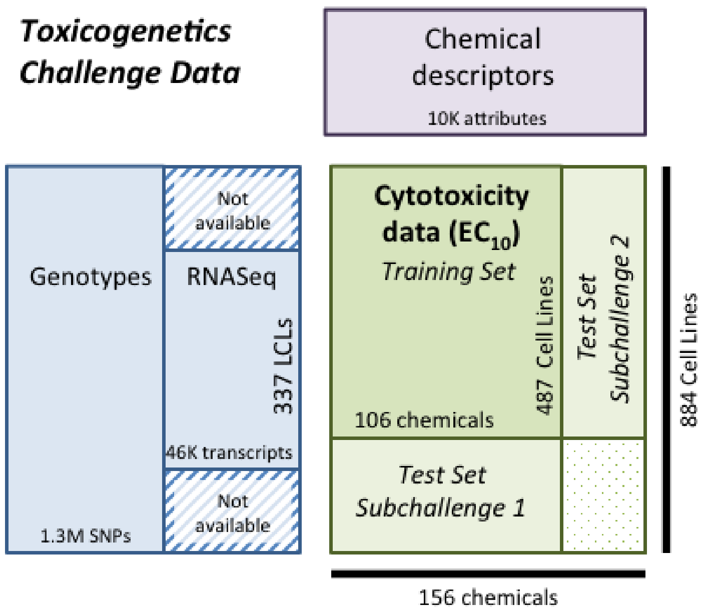


Which SNPs (or other **genomic measurements**) explain the phenotype?

Chloé-Agathe Azencott. **Machine learning tools for biomarker discovery**, Sorbonne Université, HDR dissertation, tel-02354924 (2020).

Lotfi Slim, Clément Chatelain, Chloé-Agathe Azencott, **Nonlinear post-selection inference for genome-wide association studies**, BioRxiv (2020).

Example 2: Chemogenomics



Which SNPs (or other **genomic measurements**) explain the **response-to-treatment** phenotype?

Federica Eduati et al. **Prediction of human population responses to toxic compounds by a collaborative competition**, Nature Biotechnology (2015).

Example 3: DNA sequencing

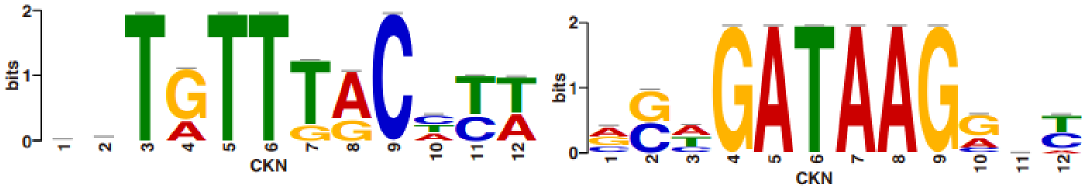
Predict **base identity** from **changes in electric current** measured by Oxford Nanopore long read sequencers

Ryan R. Wick et al. **Performance of neural network basecalling tools for Oxford Nanopore sequencing**, Genome Biology (2019).

Variant calling: predict **variant** from **sequence alignments converted to image data**

Ryan Poplin et al. **A universal SNP and small-indel variant caller using deep neural networks**, Nature Biotechnology (2018).

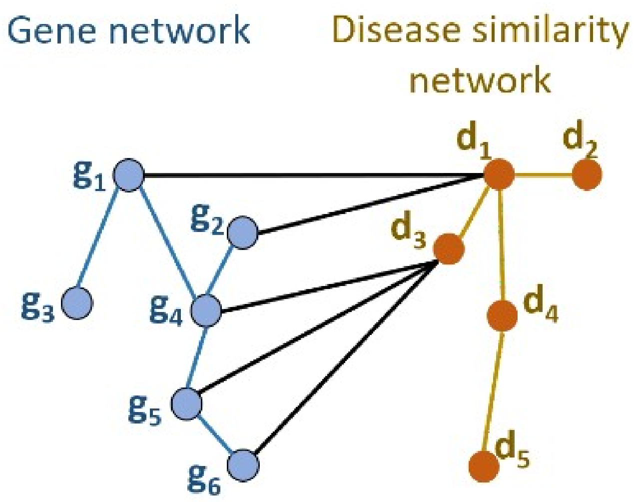
Example 4: TFBS Prediction



Predict whether a **DNA sequence** binds a given transcription factor.

Dexiong Chen et al. **Biological sequence modeling with convolutional kernel networks**, Bioinformatics (2019).

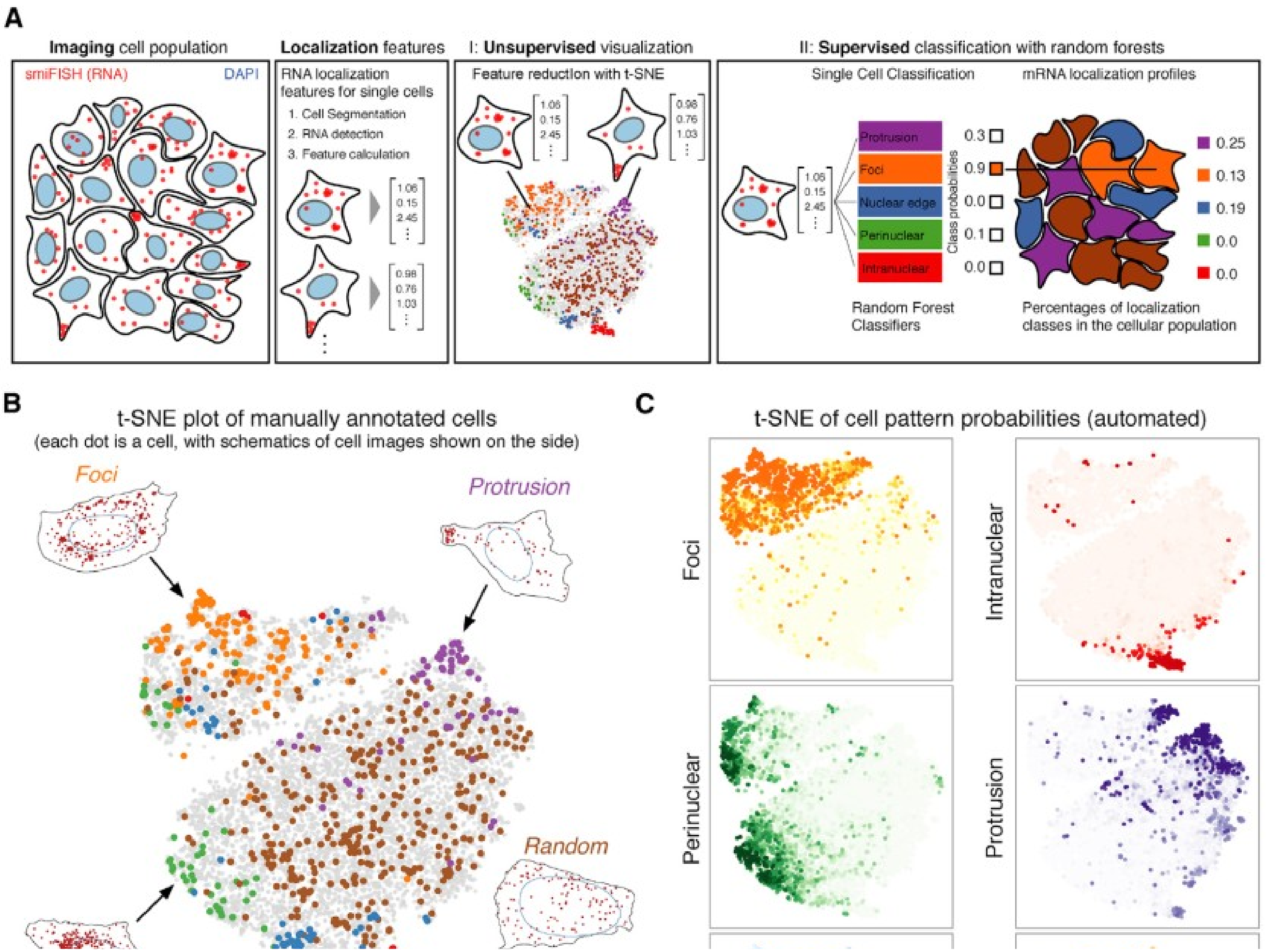
Example 5: Disease-gene prediction



Which **nodes of a gene network** are associated with which disease?

Sezin Kircali Ata et al. **Recent Advances in Network-based Methods for Disease Gene Prediction**, Briefings in Bioinformatics (2020).

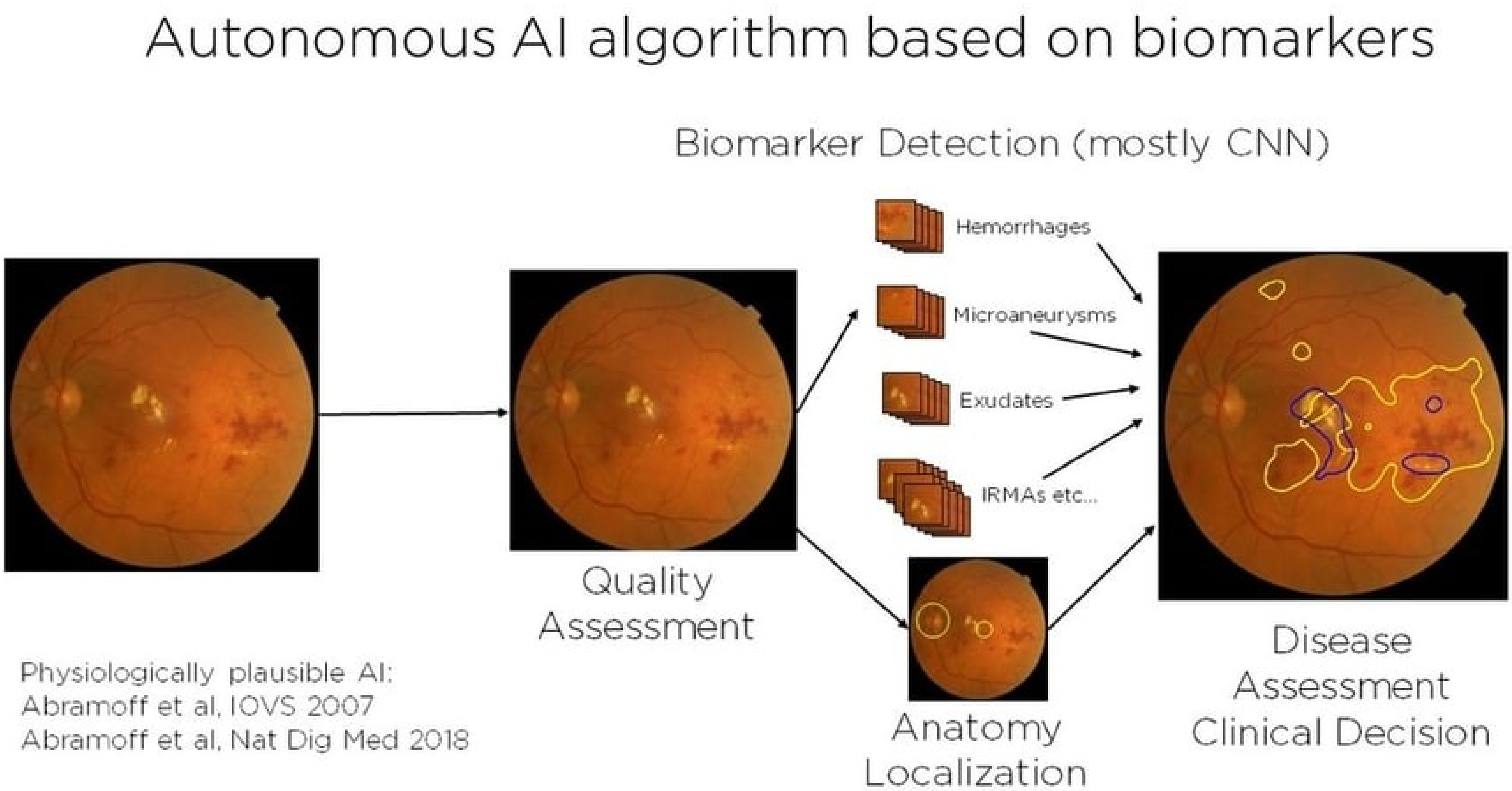
Example 6: Spatial transcriptomics



Automated classification of **mRNA localization patterns**

Racha Chouaib et al. **A Dual Protein-mRNA Localization Screen Reveals Compartmentalized Translation and Widespread Co-translational RNA Targeting**, Developmental Cell (2020).

Why haven't we cured cancer yet?



Idx-DR: automatic detection of diabetic retinopathy, FDA approved in 2018

What works well

- **Nature of the data**
  - **Images** (modeling well understood)

What works well

- **Nature of the data**
  - **Images** (modeling well understood)

    Challenge: we don't understand genomes nearly as well.

What works well

- **Nature of the data**
  - **Images** (modeling well understood)

    Challenge: we don't understand genomes nearly as well.
  - **Very large data sets** to train models on

    ImageNet > 14 million images

    transfer learning makes it possible to start from a neural network trained on natural images to learn from medical images

What works well

- **Nature of the data**
  - **Images** (modeling well understood)

    Challenge: we don’t understand genomes nearly as well.
  - **Very large data sets** to train models on

    ImageNet > 14 million images

    transfer learning makes it possible to start from a neural network trained on natural images to learn from medical images

    Challenge: Our data sets are small.

What works well

- **Nature of the data**
  - **Images** (modeling well understood)

    Challenge: we don't understand genomes nearly as well.
  - **Very large data sets** to train models on

    ImageNet > 14 million images

    transfer learning makes it possible to start from a neural network trained on natural images to learn from medical images

    Challenge: Our data sets are small.
- **Nature of the question**
  - Humans can perform the task.

What works well

- **Nature of the data**
  - **Images** (modeling well understood)

    Challenge: we don’t understand genomes nearly as well.
  - **Very large data sets** to train models on

    ImageNet > 14 million images

    transfer learning makes it possible to start from a neural network trained on natural images to learn from medical images

    Challenge: Our data sets are small.
- **Nature of the question**
  - Humans can perform the task.

    Challenge: Humans cannot perform the task.

    That’s why they’re interesting :-)

Relevant current ML challenges

- Learning from **small data sets**
  - **few-shot learning**
- Learning from **several data sets**
  - **federated learning** / **differential privacy** / **domain adaptation**
- **Describing** & **understanding**
  - **interpretability**
- **Trusting** what is learned
  - **verification** / **certification**
- Learning from **heterogeneous data types** (sequences, genomic measurements, images, graphs and more)
  - **multi-modality** / **multi-view learning**

Acknowledgements

- Talks by **Ewan Birney**, **Gabriele Schweikert**, **Jean-Philippe Vert**
- 2020 report of the **PHG foundation** on Artificial Intelligence for genomic medicine

  https://www.phgfoundation.org/documents/artificial-intelligence-for-genomic-medicine.pdf
- **CBIO** & **U900**, **PrAIrie**, **MLFPM**