Data integration in systems genetics & aging research

Alexis Rapin

Laboratory of
Integrative Systems
Physiology

epfl.ch/labs/auwerx-lab

Laboratory of Integrative Systems Physiology

Team

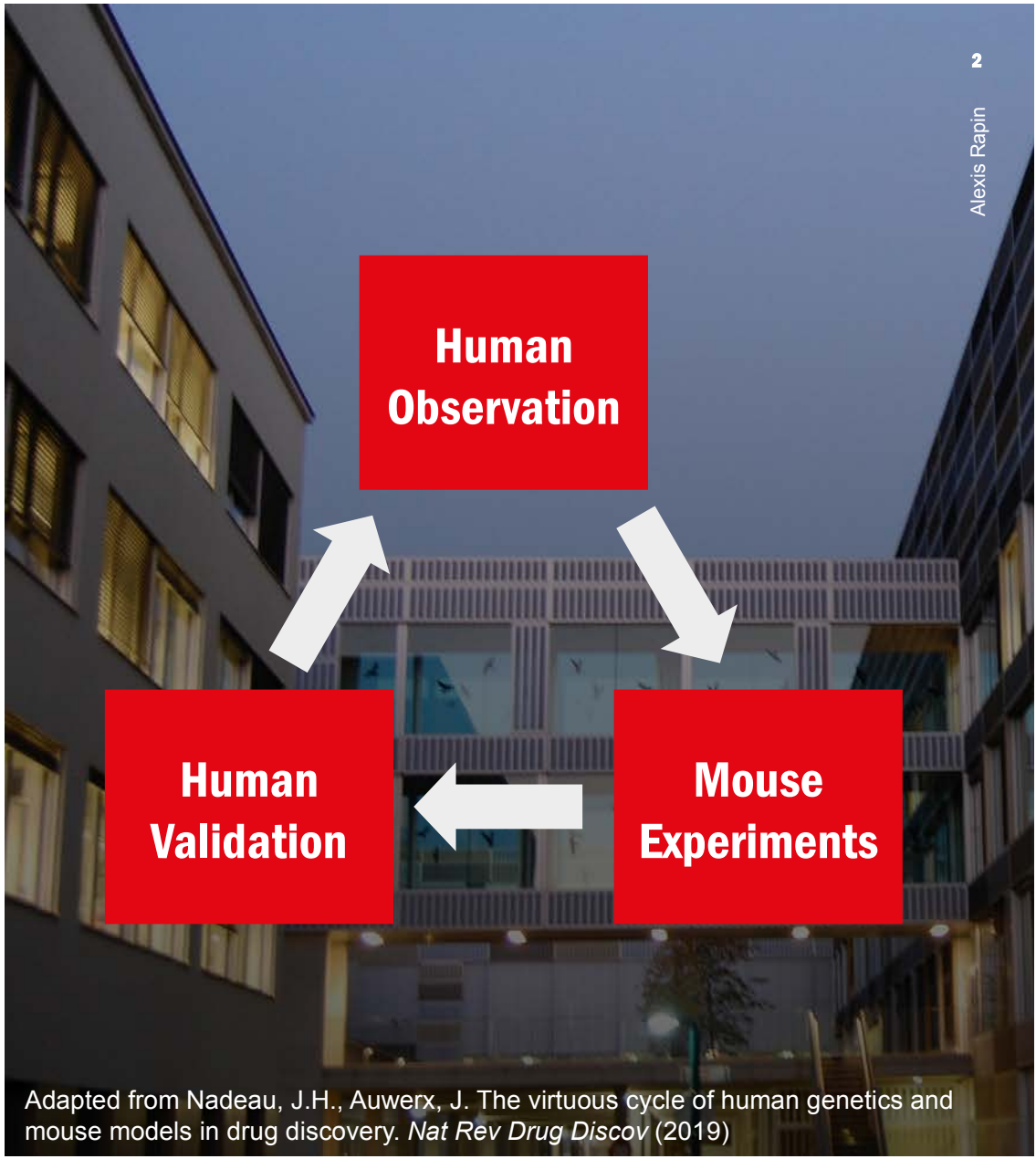
- Head: Johan Auwerx
- ~20 molecular biologists
- ~10 bioinformaticians

Research focus

- Aging and metabolic disorders
- Mitochondrial metabolism

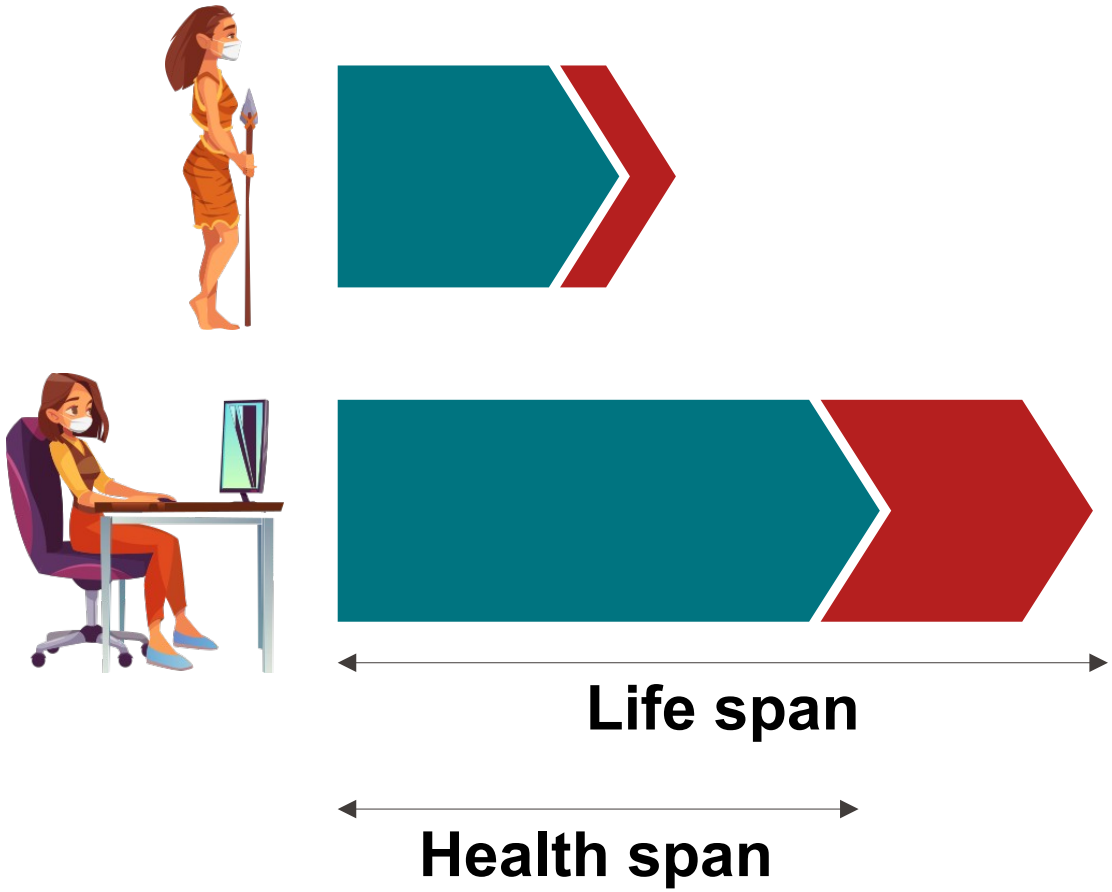
Key methods

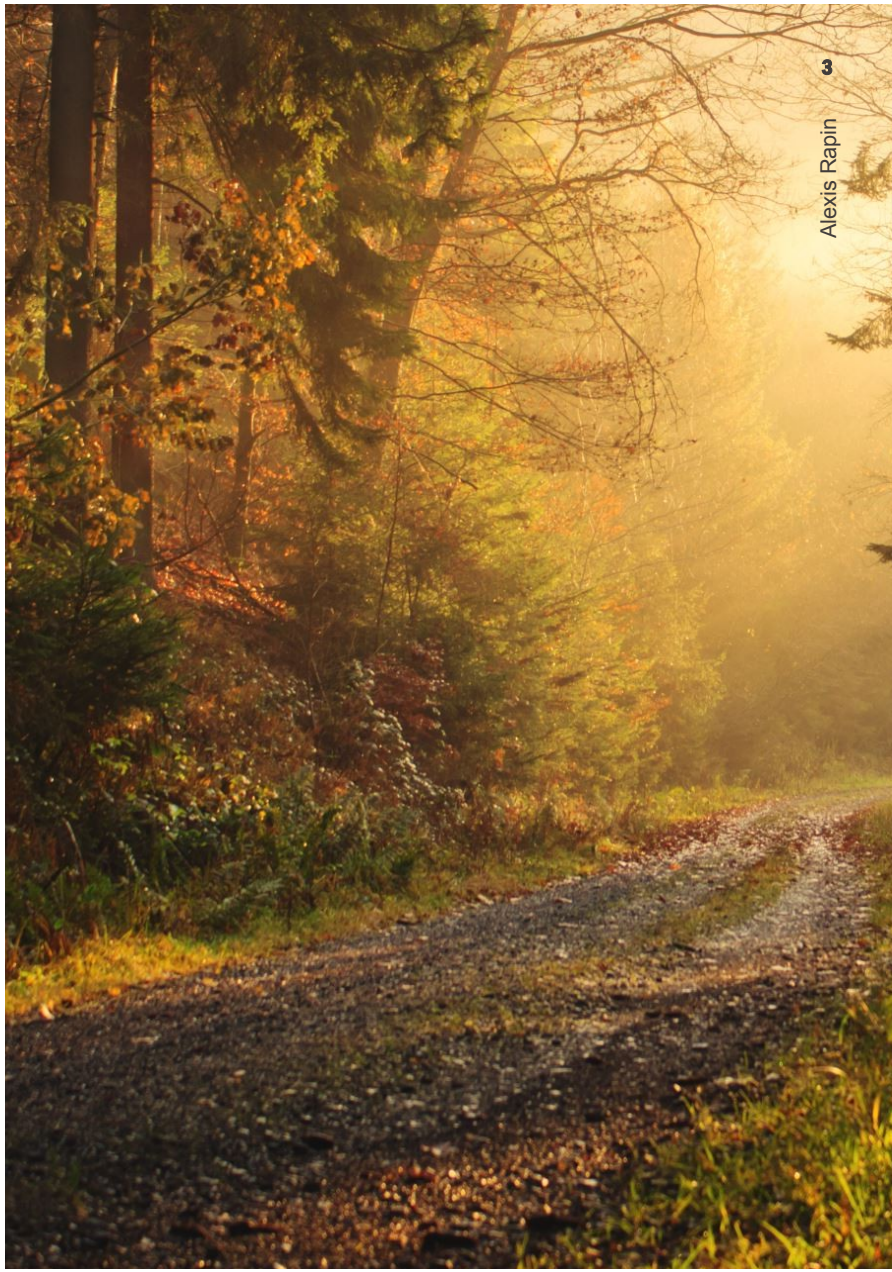
- Models of genetic diversity
- Omics (genomics, proteomics, …)
- Phenotyping
- Drugs/compounds screening



Adapted from Nadeau, J.H., Auwerx, J. The virtuous cycle of human genetics and mouse models in drug discovery. *Nat Rev Drug Discov* (2019)

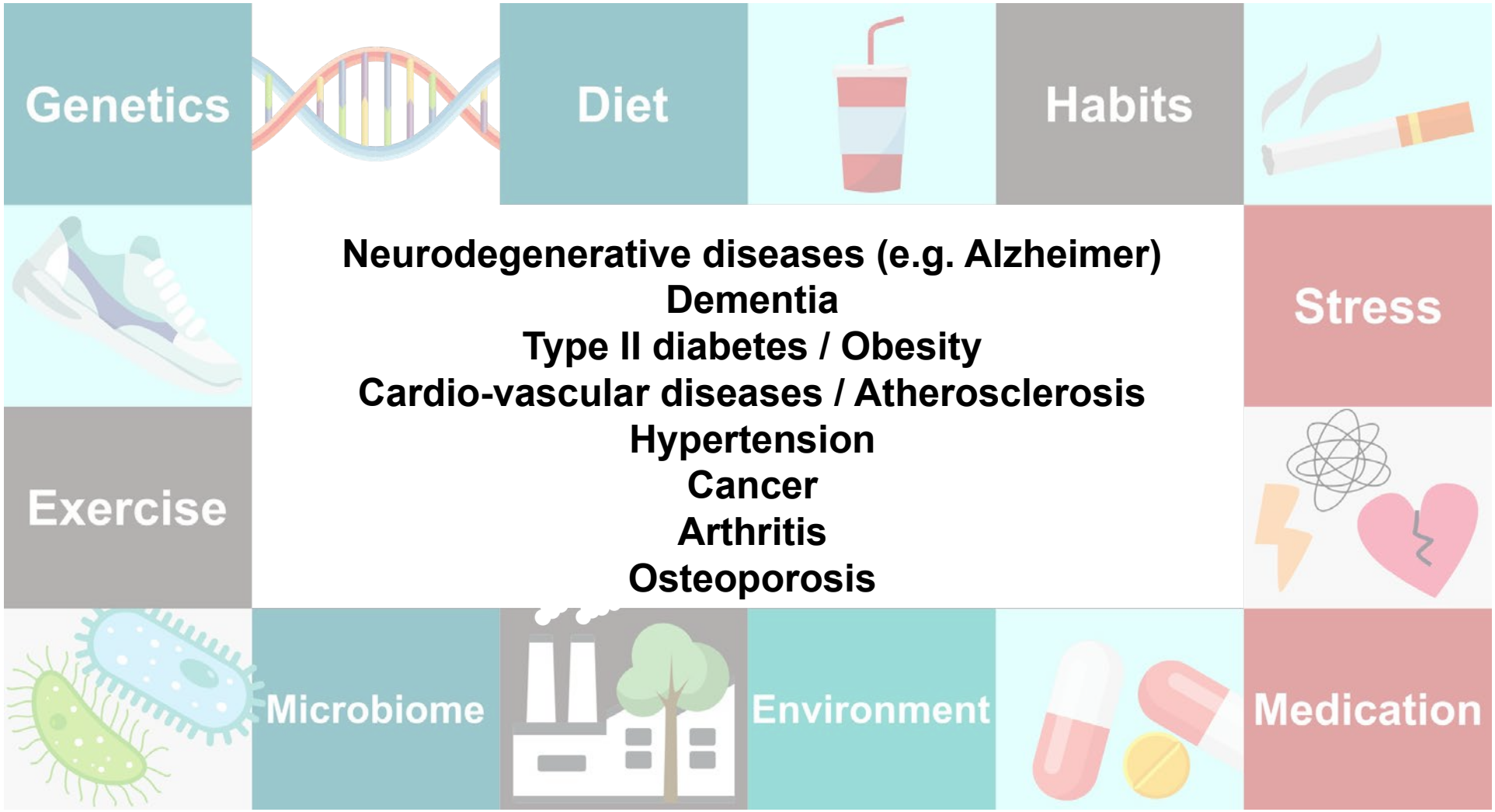
More people will suffer from age-related diseases



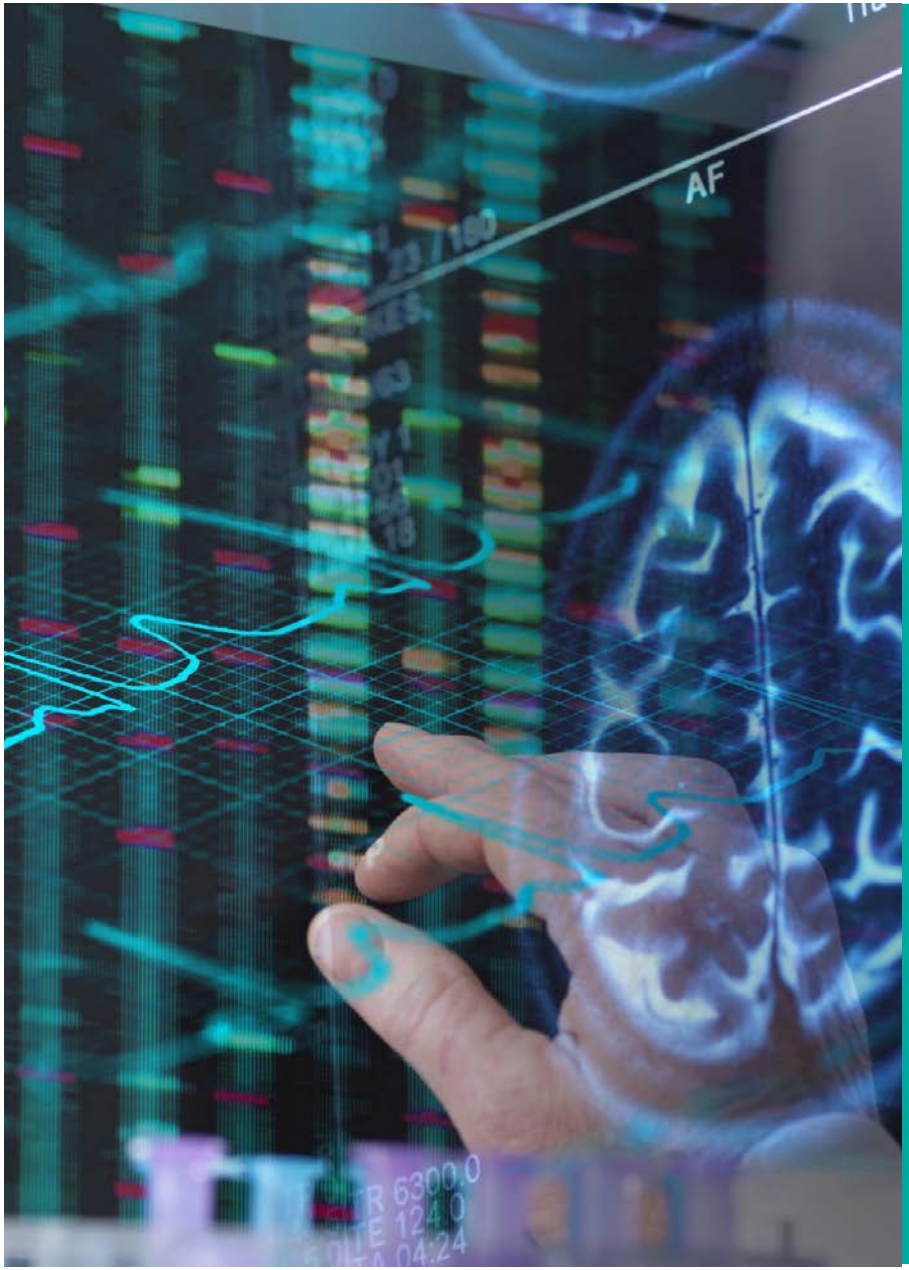


Alexis Rapin

The roots of age-related diseases are complex



We do not all age the same way

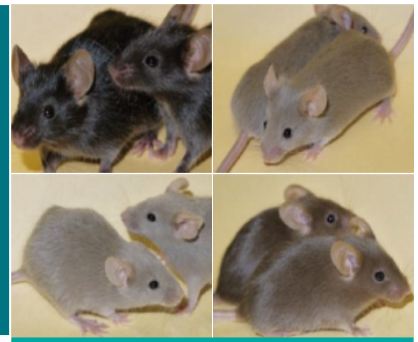


Precision medicine

A good treatment for **you** may not be a good treatment for **me**

Genetic diversity models:
The experimental side of precision medicine

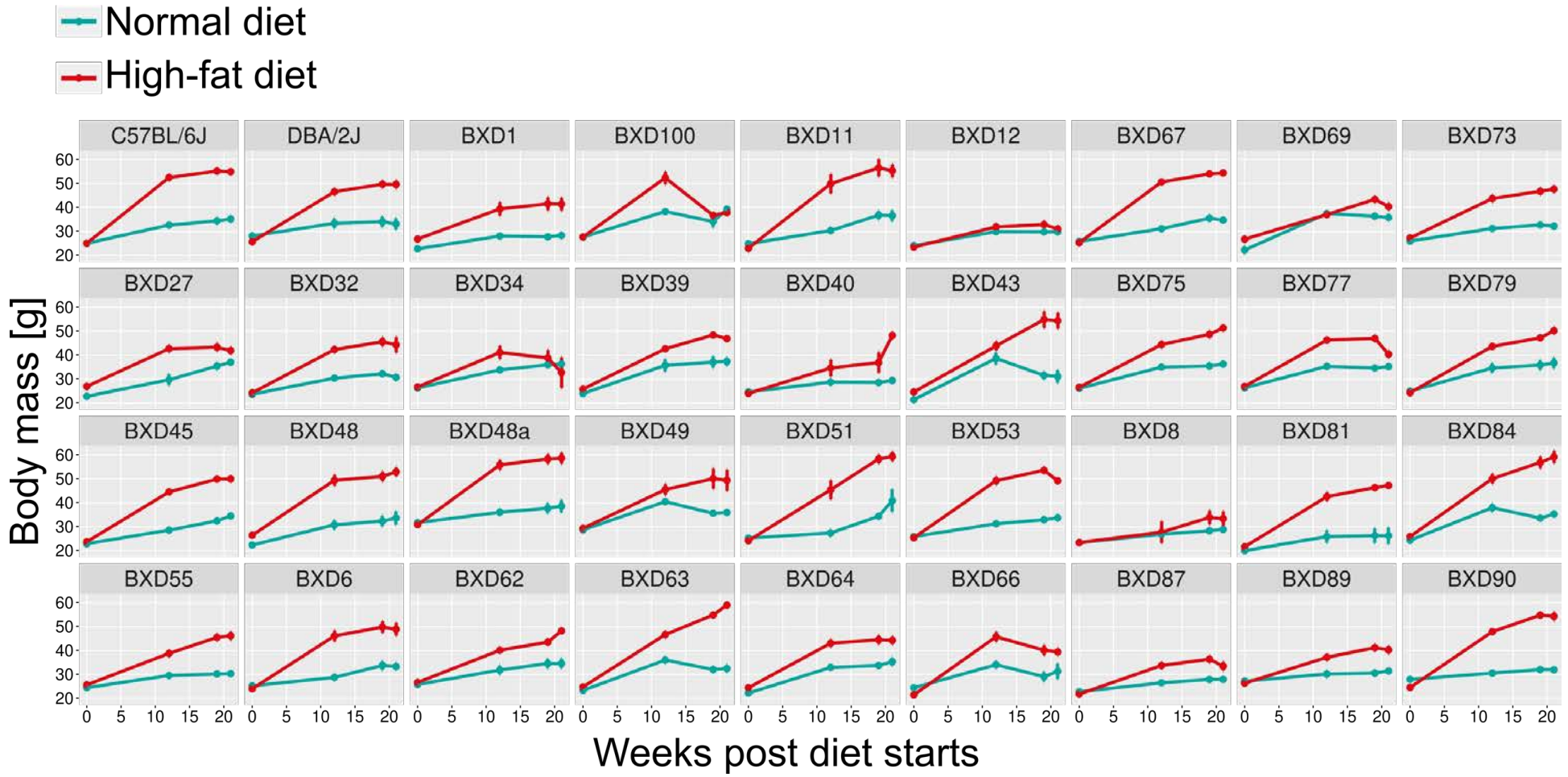
controlled
Genetics
Genetic Diversity Panels



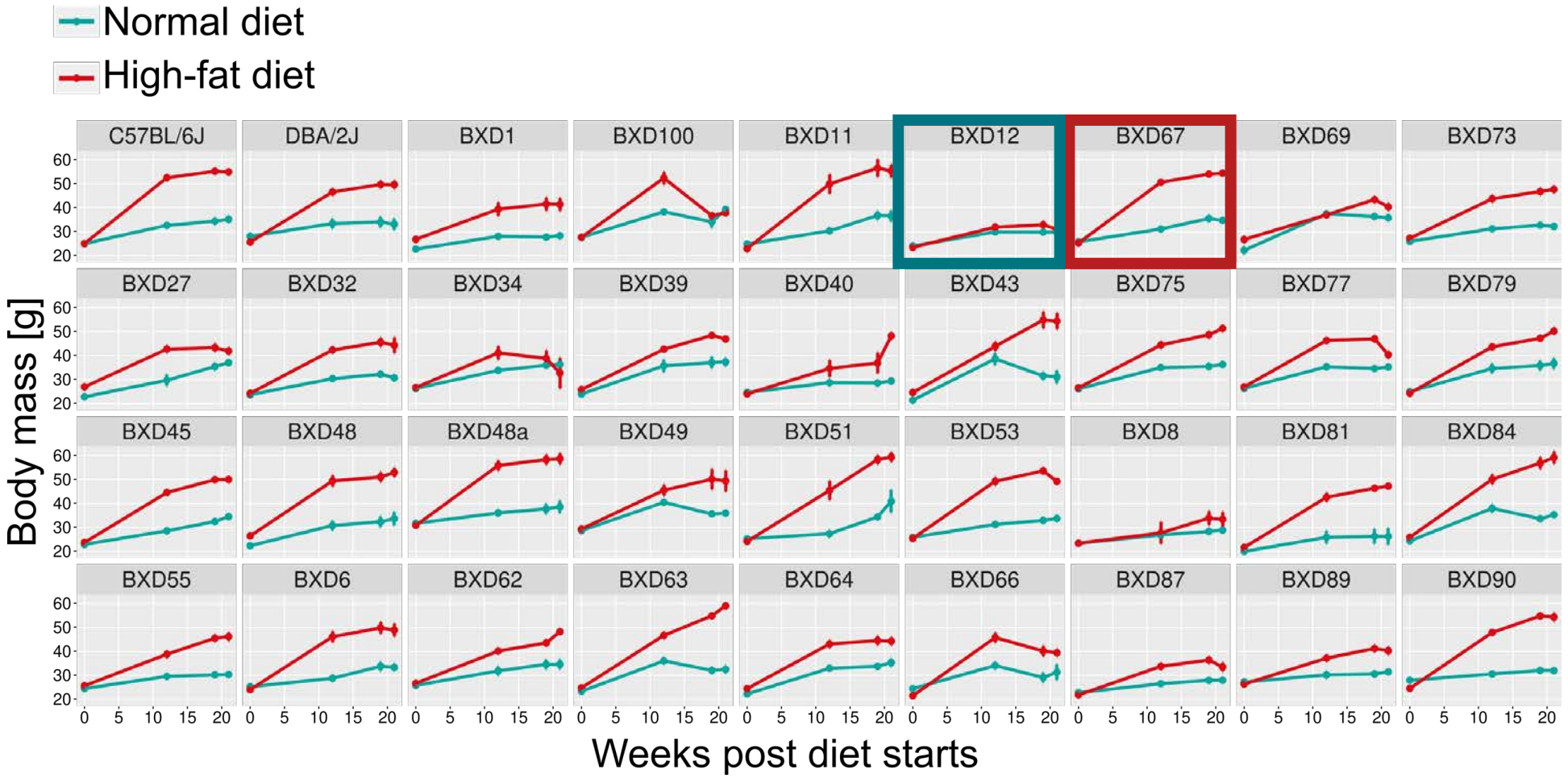
controlled
Environment

controlled
Diet

Genetic diversity model populations allow to link trait variations to genetic variations

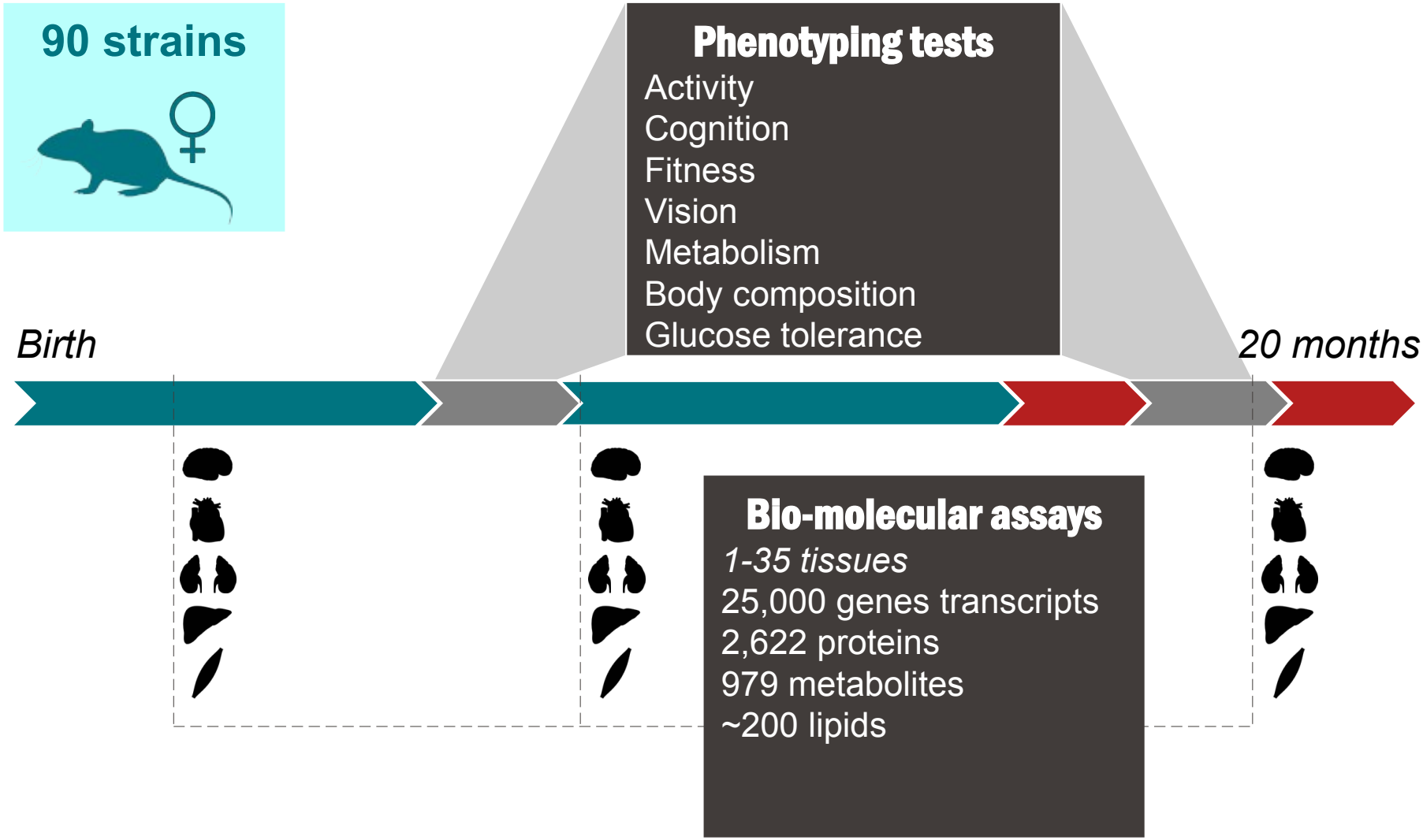


Genetic diversity model populations allow to link trait variations to genetic variations



Alexis Rapin

Observations are made at the organism and the bio-molecular levels



Macroscopic phenotypes are shaped by the microscopic bio-molecular machinery and the environment

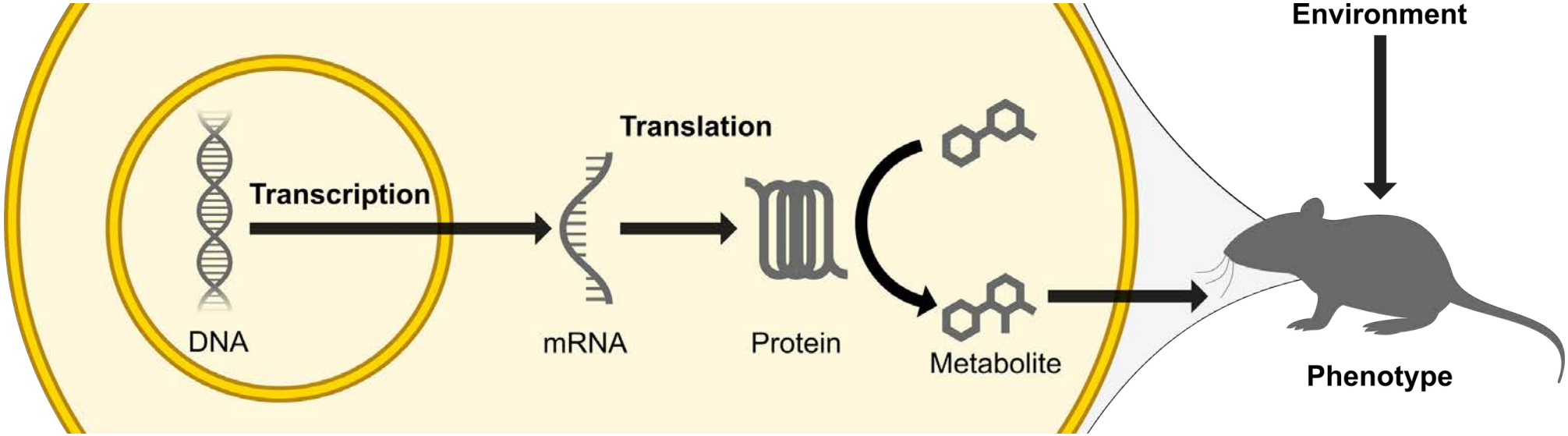
Genome
97 strains
6,800 markers

Transcriptome
34 tissues
25,000 transcripts

Proteome
1 tissue
2,622 proteins

Metabolome
2 tissues
979 metabolites

Phenome
5,000 phenotypes



Adapted from Li, H. et *al.* An Integrated Systems Genetics and Omics Toolkit to Probe Gene Function. *Cell Syst* 6, 90–120.e4 (2018).

Datasets are heterogeneous and sparse

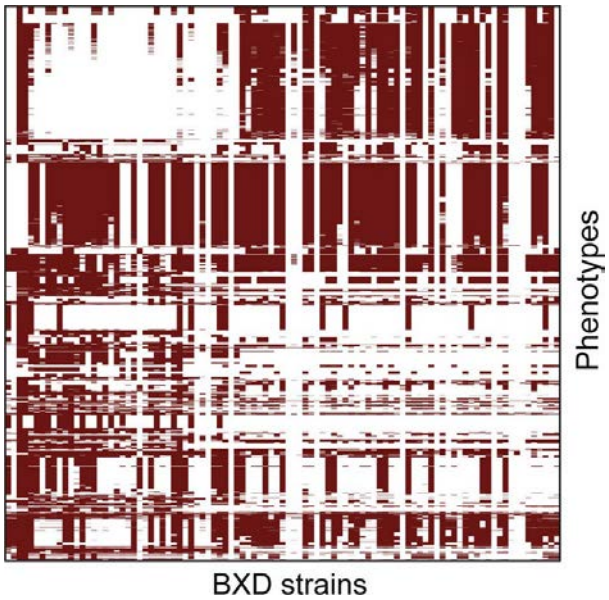
Genome
97 strains
6,800 markers

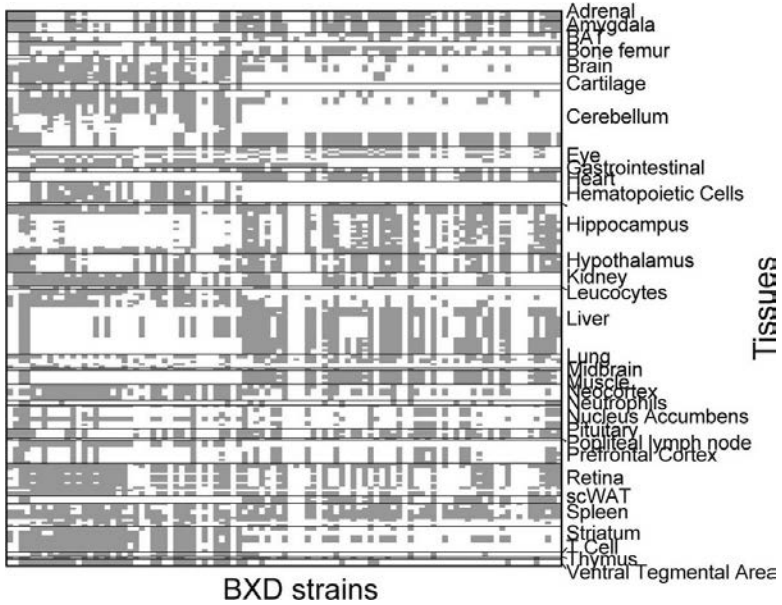
Transcriptome
34 tissues
25,000 transcripts

Proteome
1 tissue
2,622 proteins

Metabolome
2 tissues
979 metabolites

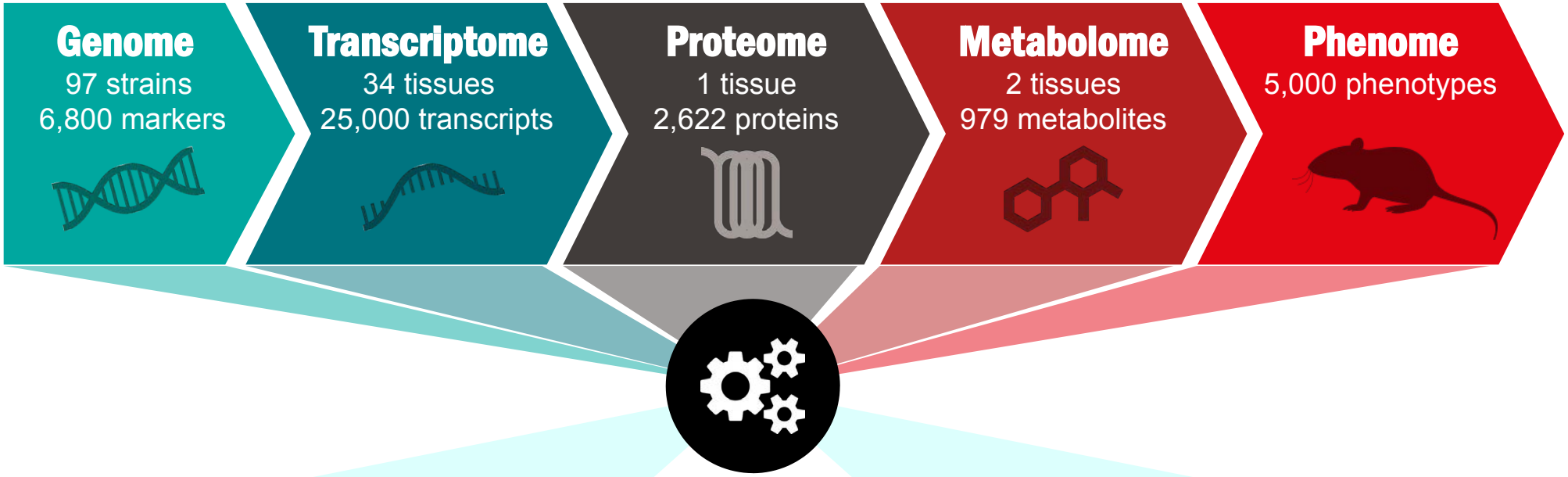
Phenome
5,000 phenotypes





Adapted from Li, H. et *al.* An Integrated Systems Genetics and Omics Toolkit to Probe Gene Function. *Cell Syst* 6, 90–120.e4 (2018).

Important patterns may be observed only by combining multiple layers of biological data



No disease

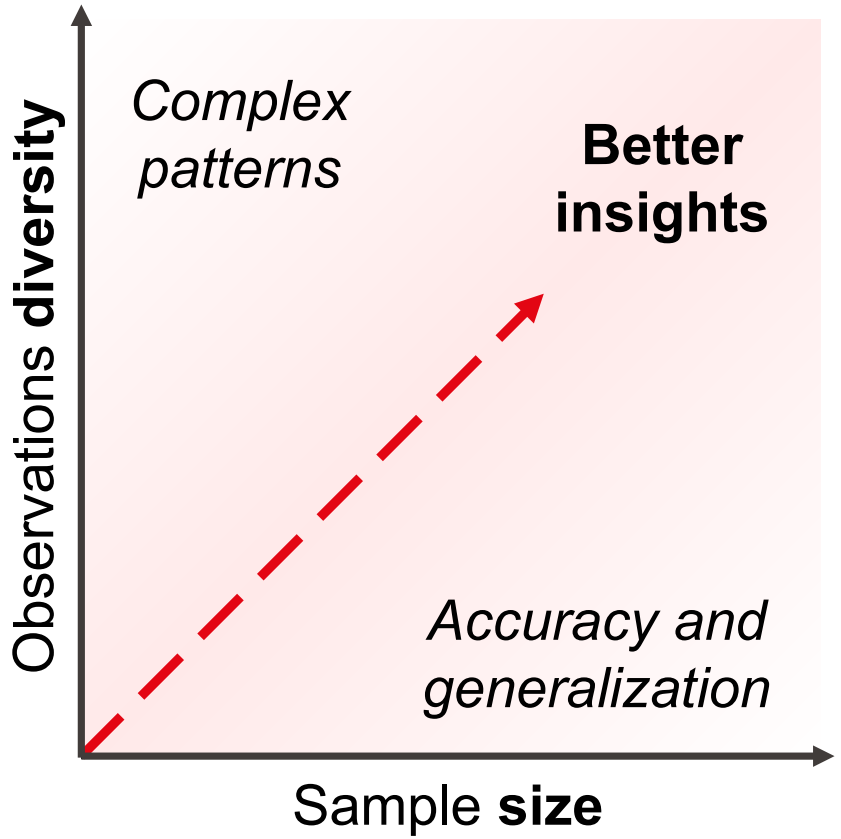
GeneX associated with disease
AND Low expression in Liver
AND High expression in Brain
AND Normal blood pressure

Disease

GeneY associated with disease
AND High expression in Liver
AND High blood pressure

Adapted from:
- Li, H. et *al.* An Integrated Systems Genetics and Omics Toolkit to Probe Gene Function. *Cell Syst* 6, 90–120.e4 (2018).
- Zitnik, M. et *al.* Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Inf Fusion* 50, 71-92 (2019).

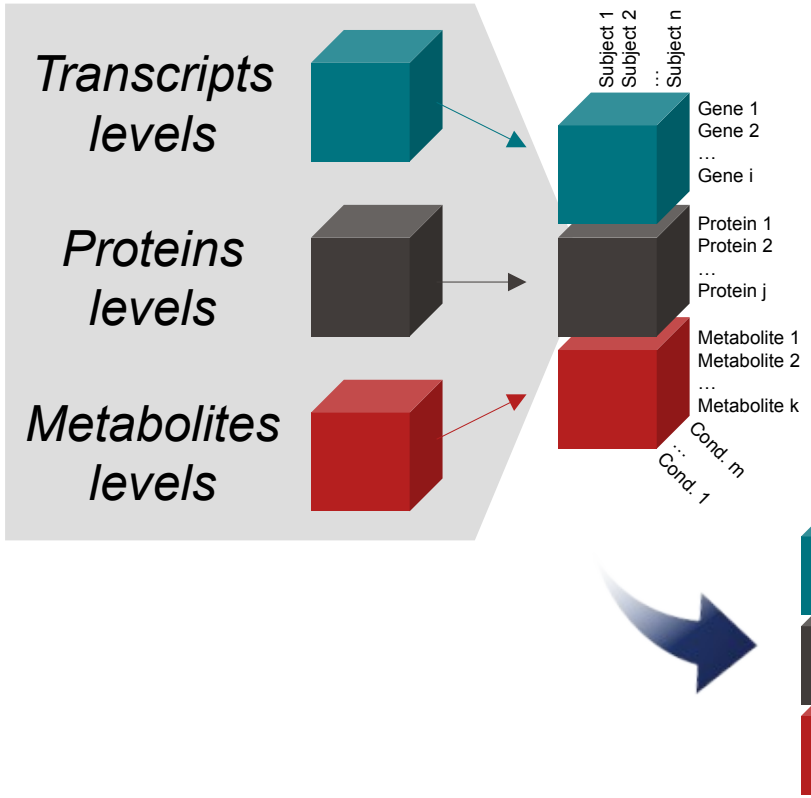
The more the better



Integration can diversify and enlarge datasets

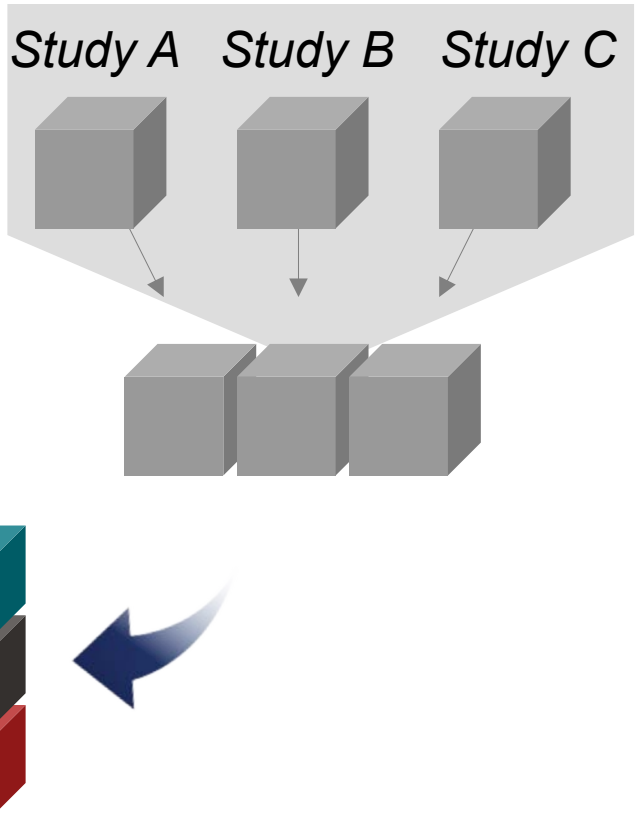
Vertical (intra-study) integration

Increasing observations **diversity**



Horyzontal (inter-studies) integration

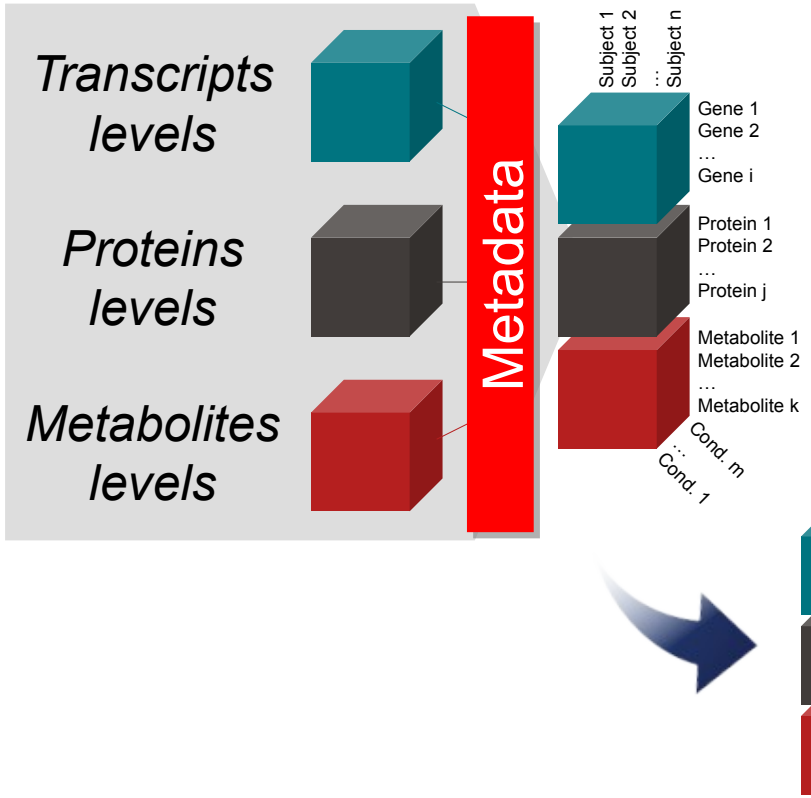
Increasing dataset sample **size**



Integration can diversify and enlarge datasets

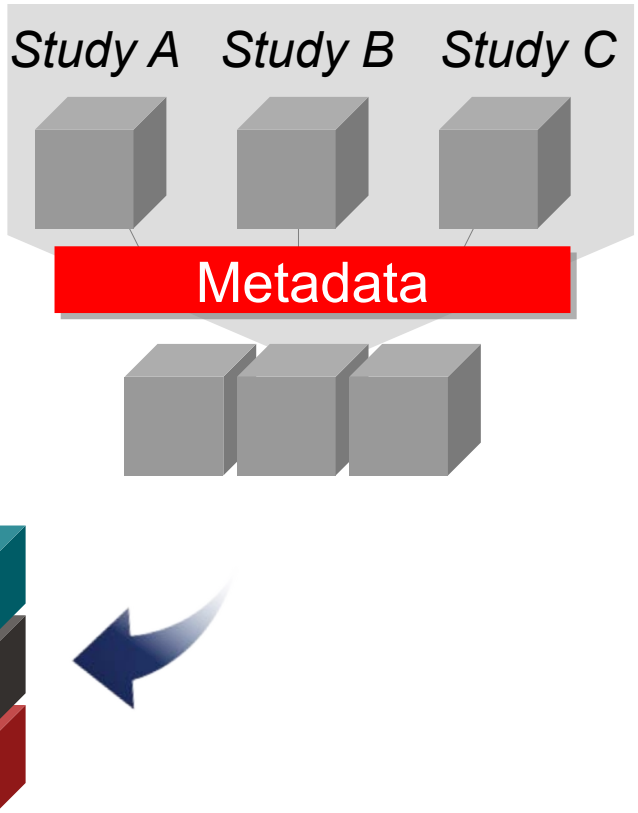
Vertical (intra-study) integration

Increasing observations **diversity**



Horyzontal (inter-studies) integration

Increasing dataset sample **size**

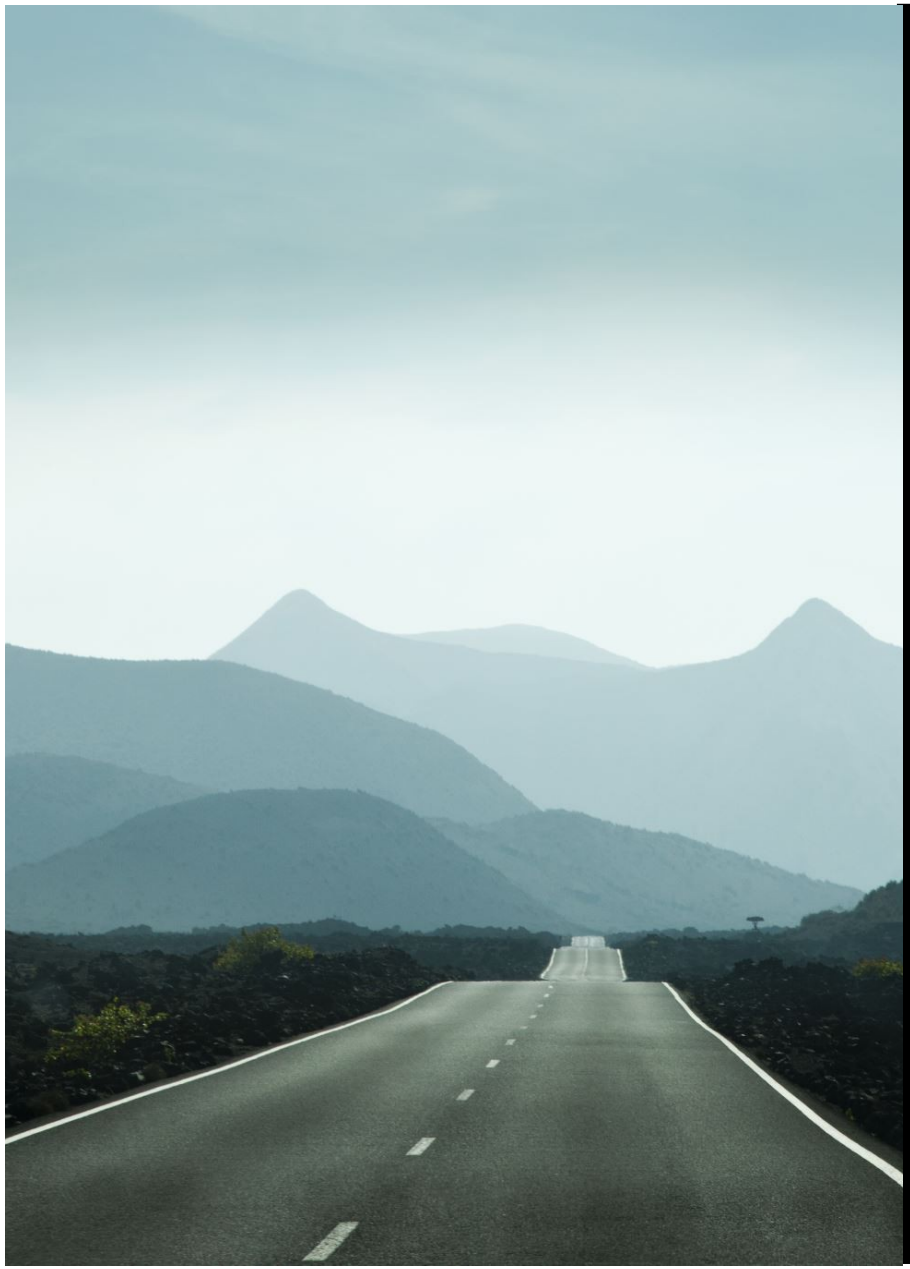


The FAIR principles ease datasets integration

Findable **A**ccessible **I**nteroperable **R**eusable

> ***Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process.***

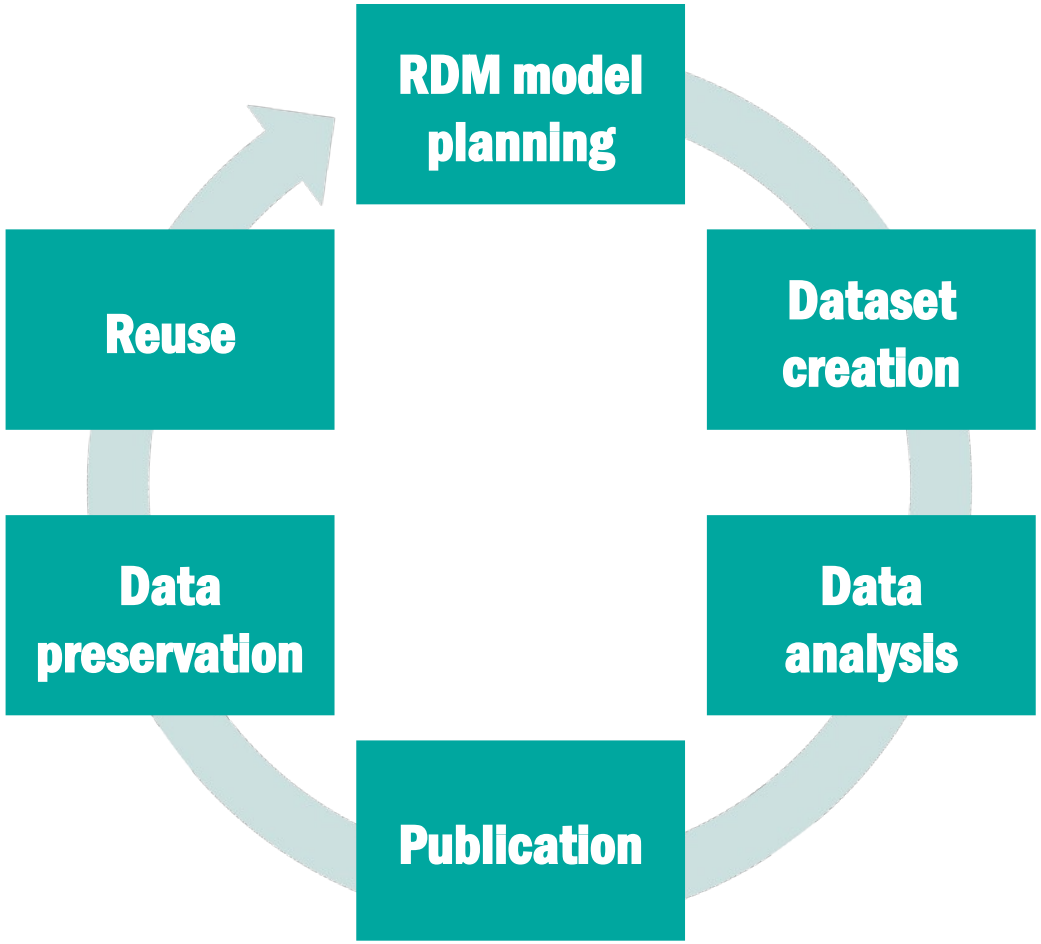
Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.*
The FAIR Guiding Principles for scientific data management and stewardship.
***Sci Data* 3, 160018 (2016).**



Streamlining research datasets with the FAIR principles

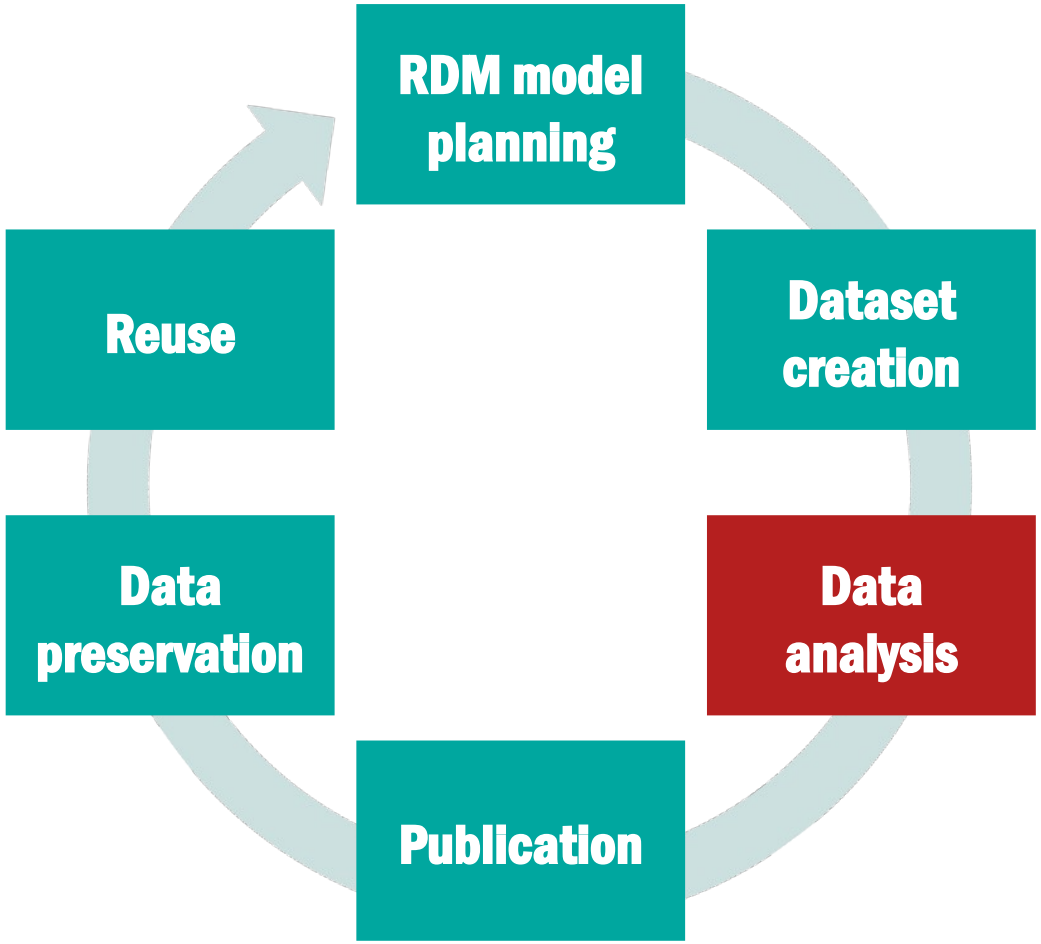
Easing datasets **integration** and **reuse**

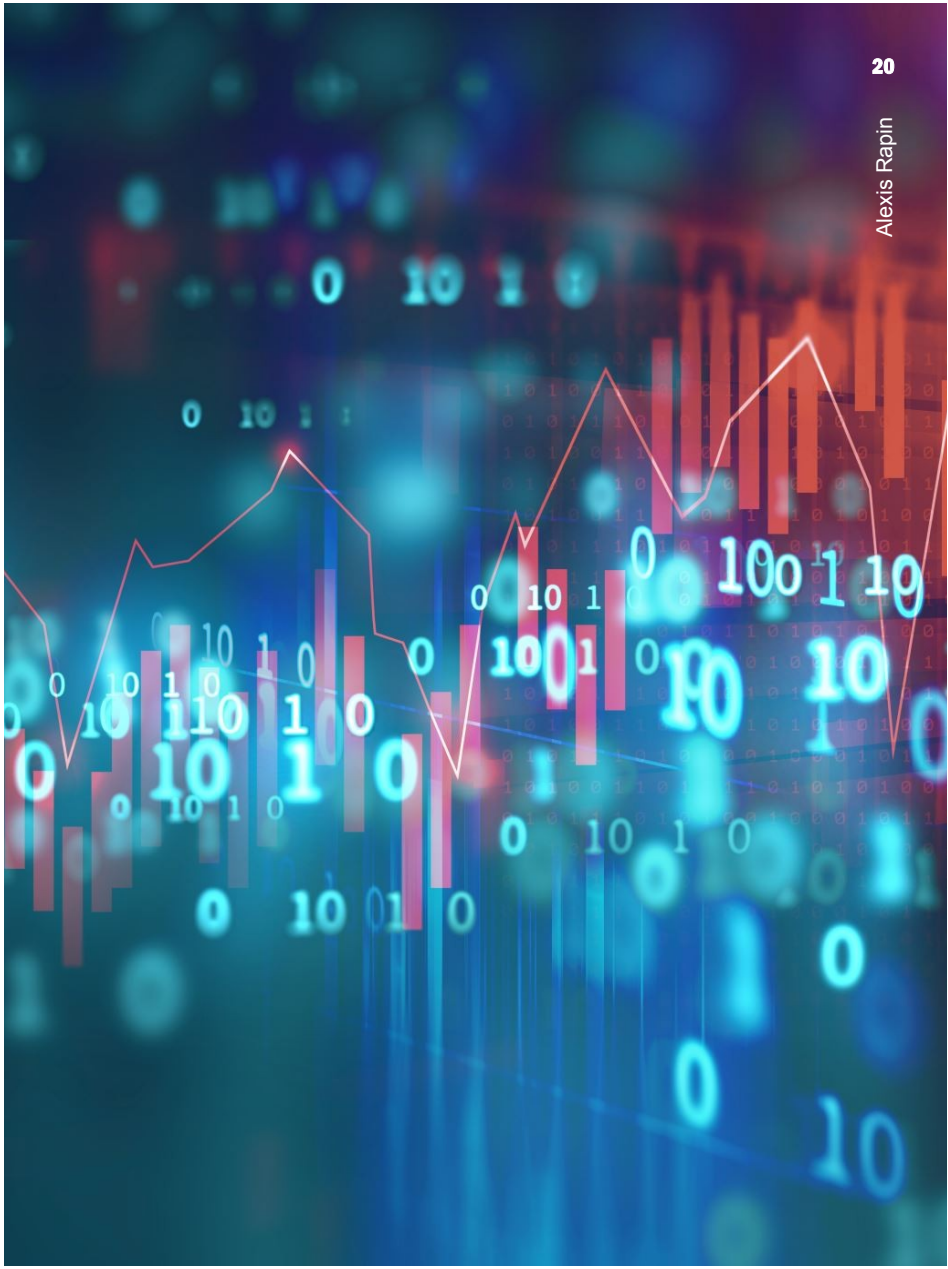
Planning the data life cycle



Alexis Rapin

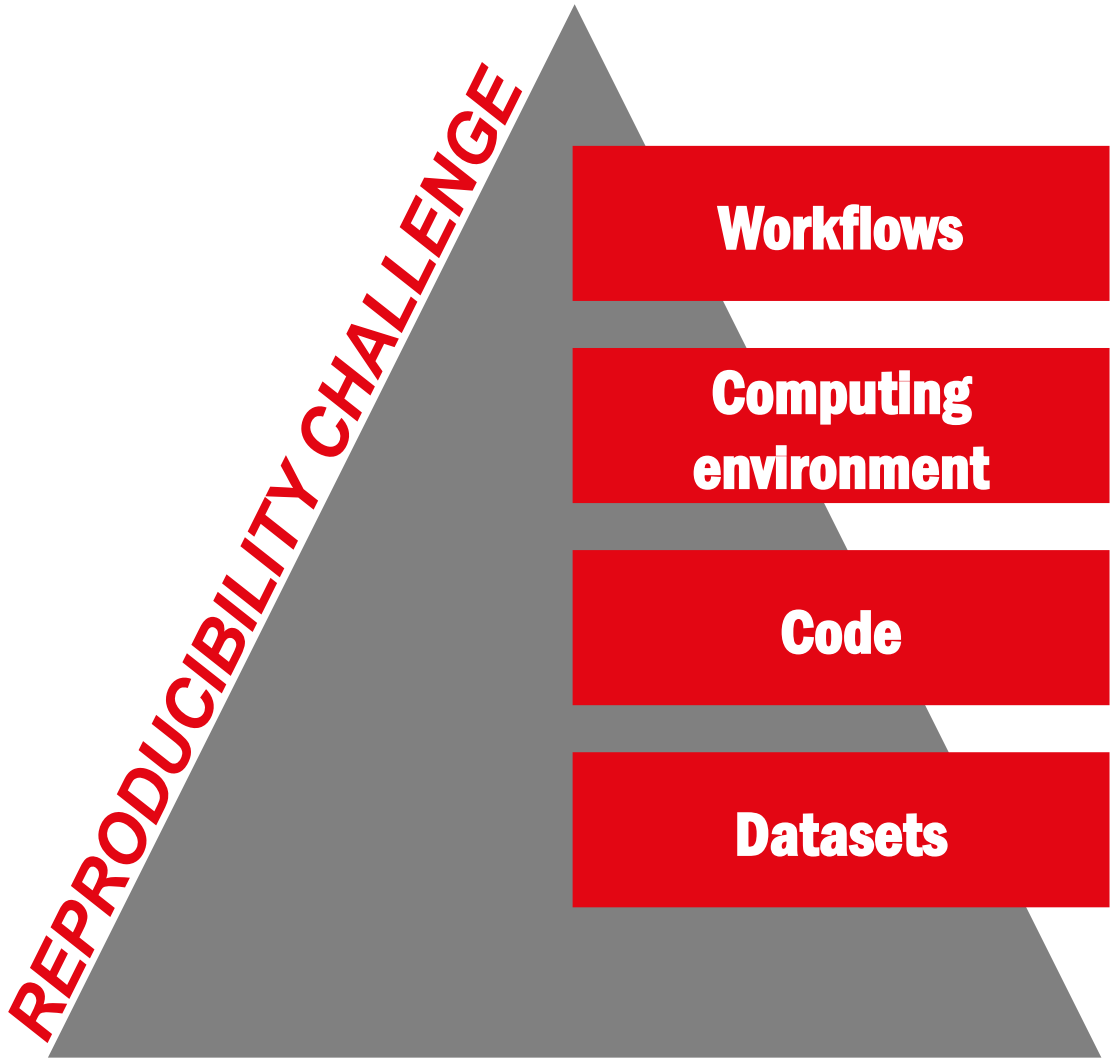
Data analysis





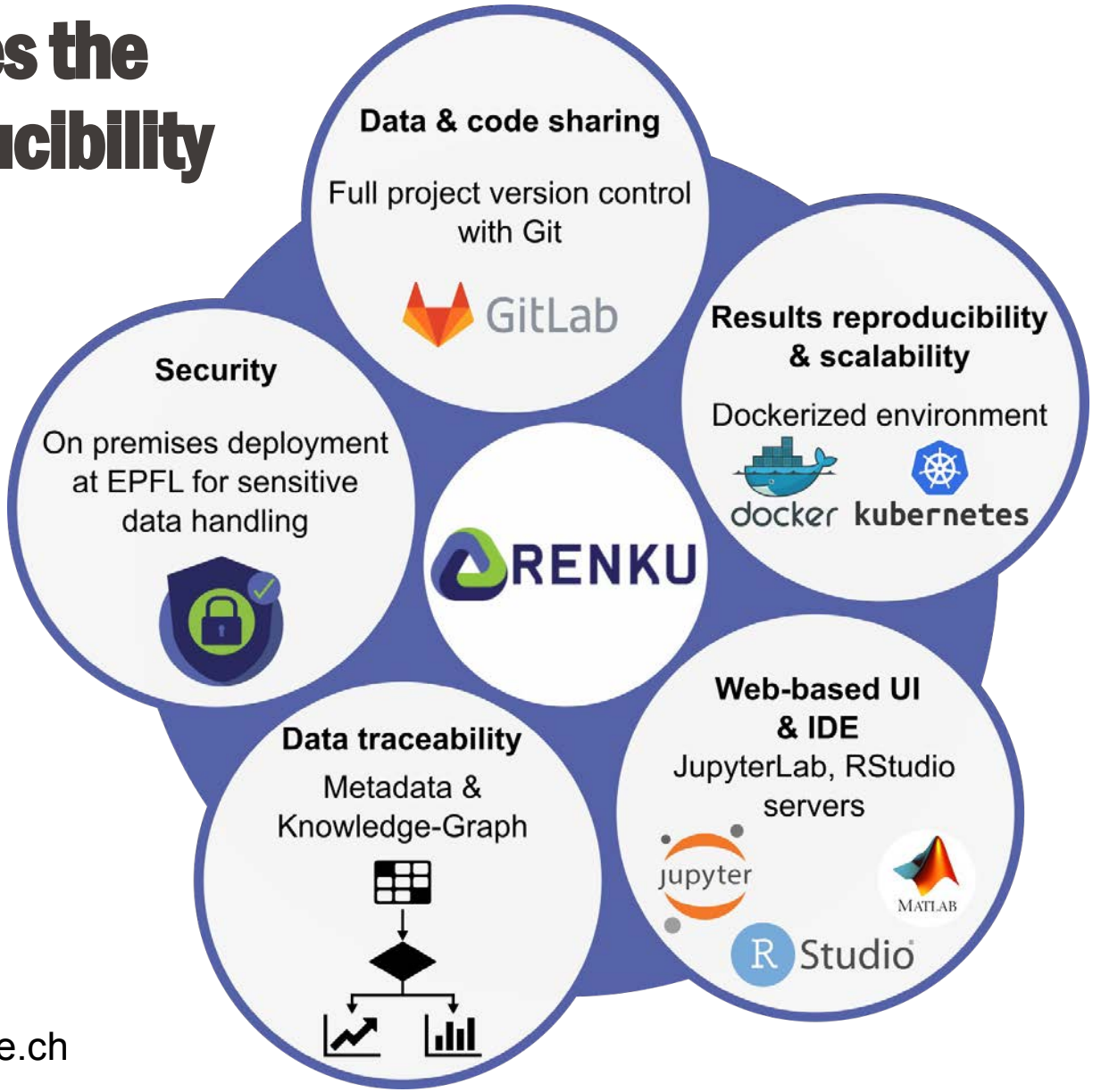
Alexis Rapin

"But, it works on my machine!"



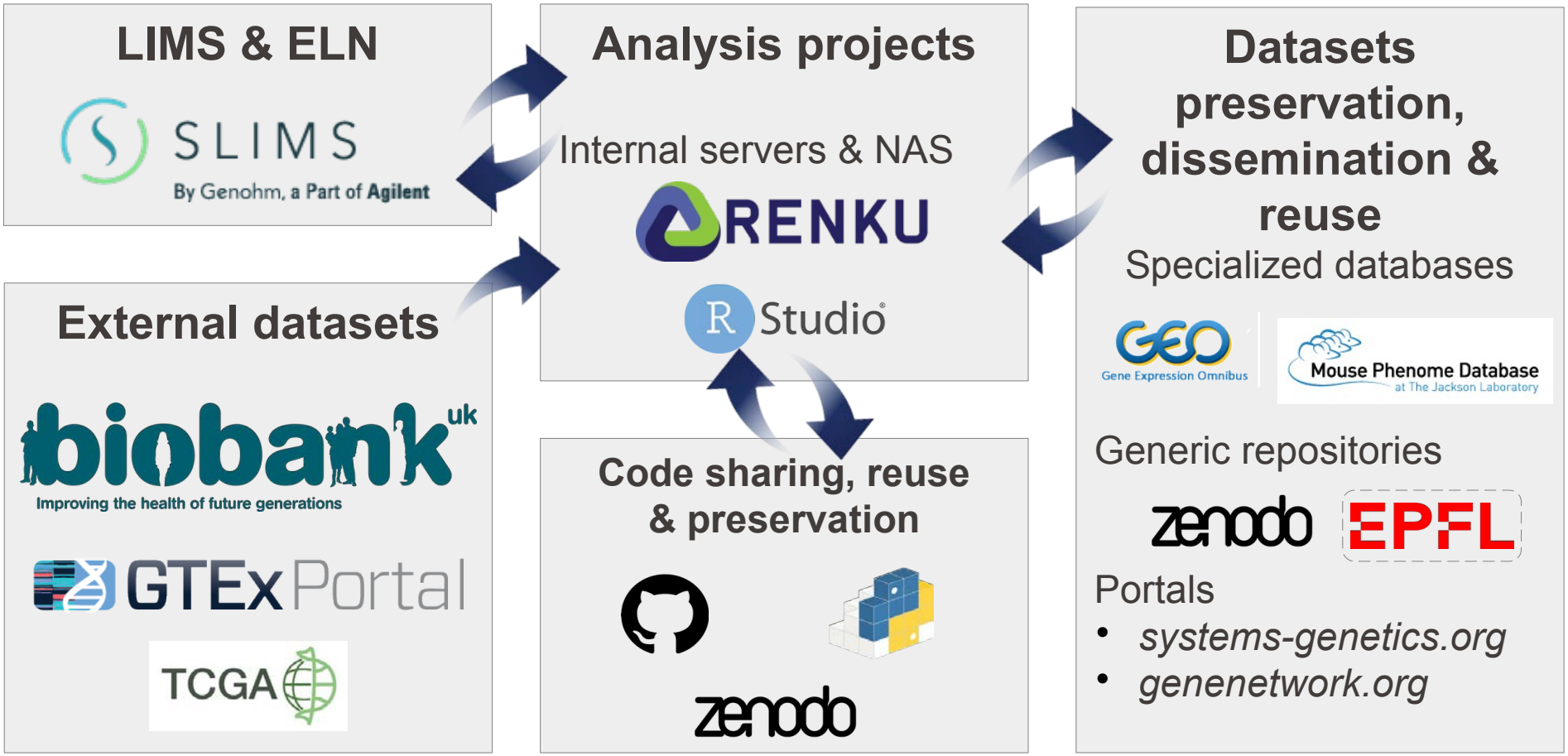
Alexis Rapin

Renku addresses the analysis reproducibility challenge



More at https://datascience.ch

Our data analysis ecosystem (not exhaustive)



Alexis Rapin

FAIRification enables vertical integration - use case

57 strains
2 diets

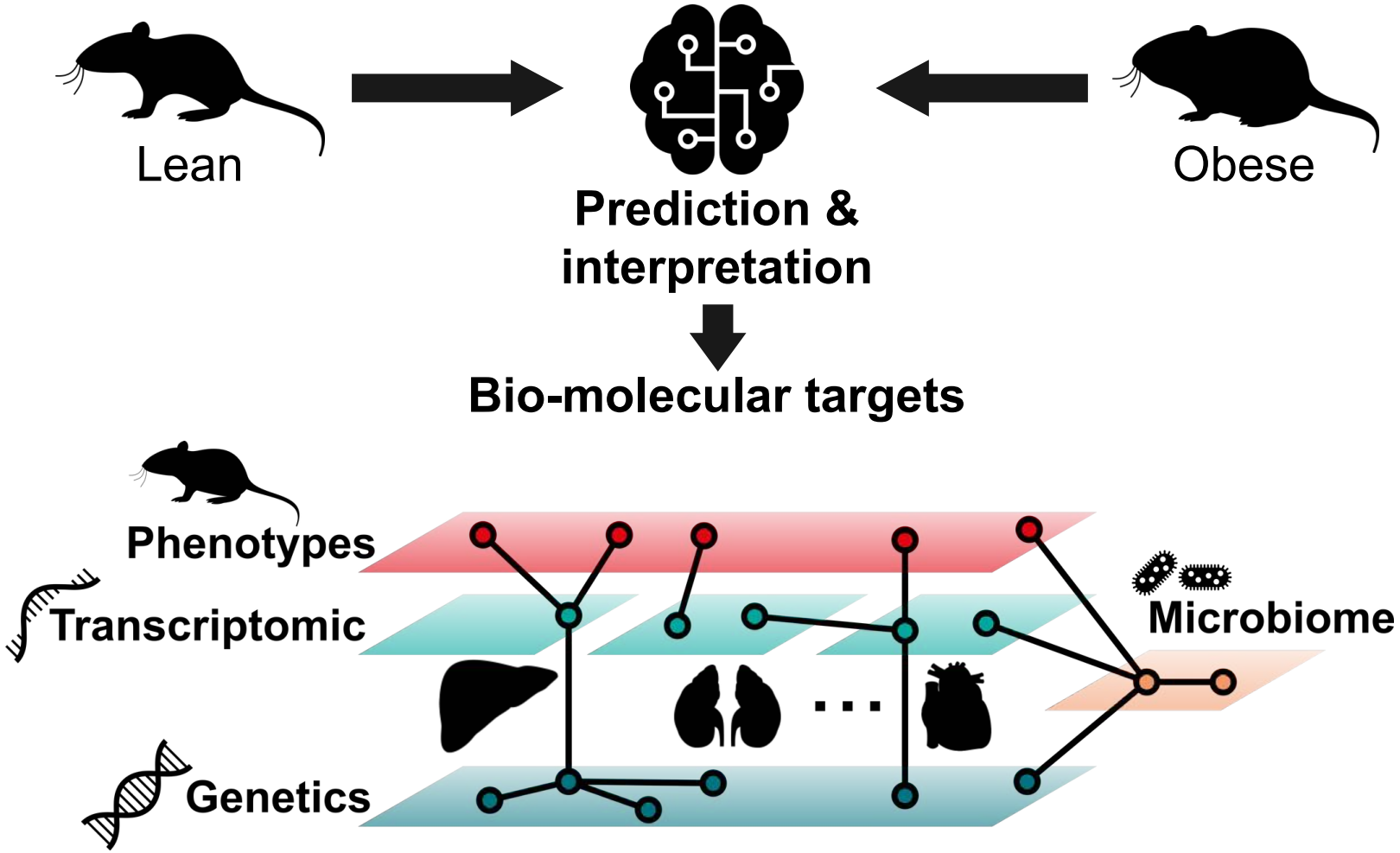
Data type	Tissue	Dimension	Location
Phenome	NA	117 measurements x 536 subjects	MPD:1023
Transcriptome	Liver	35,556 genes x 81 samples	GEO:GSE60149
Transcriptome	Quadriceps	35,556 genes x 79 samples	GEO:GSE60151
Transcriptome	Heart	41,155 genes x 81 samples	GEO:GSE60489
Transcriptome	scWAT	65,770 genes x 80 samples	GEO:GSE79016
Transcriptome	Colon	35,556 genes x 81 samples	Internal NAS
Transcriptome	Ileum	35,556 genes x 81 samples	Internal NAS
Microbiome	Ceacum	5,215 ASVs x 89 samples	Internal NAS
...	...	...	...

Metadata
- Standardization
- Enrichment
- Cross-referencing

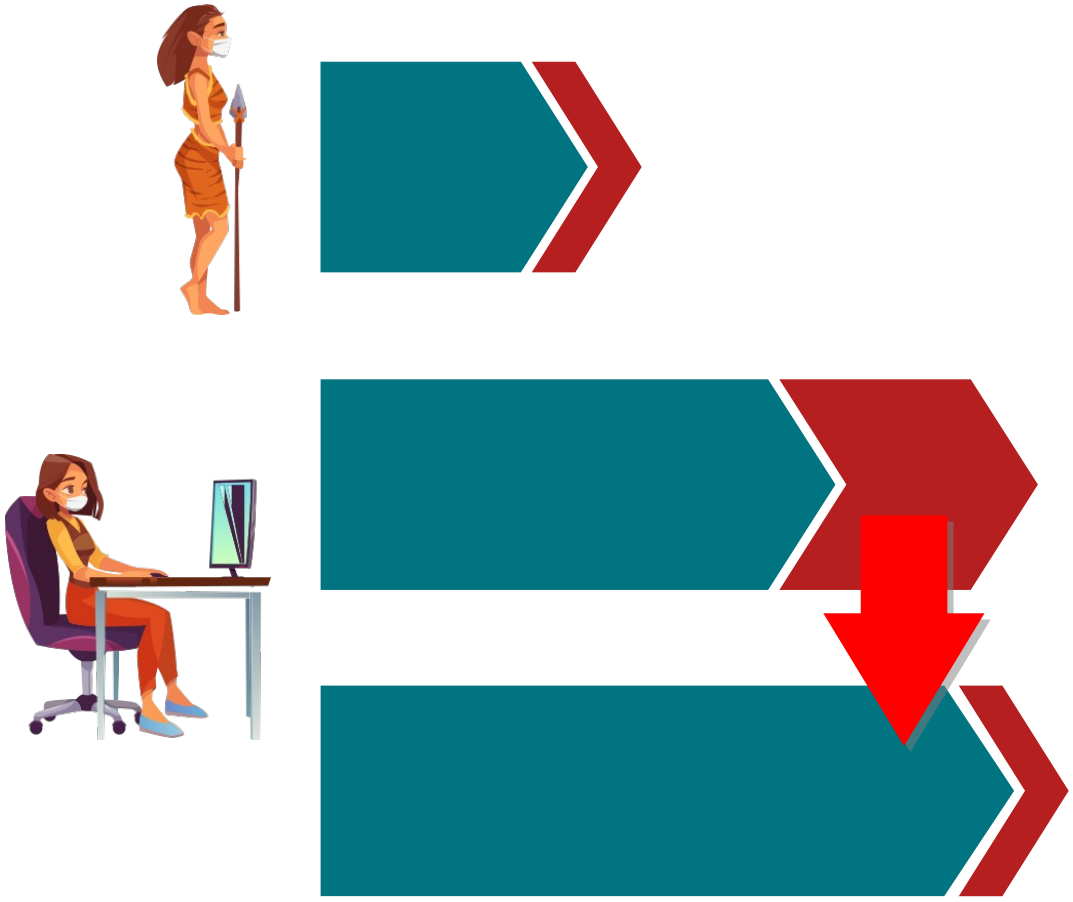
Release / update

Integration

Advanced ML will enable the discovery of new mechanisms that may be exploited to fight age-related diseases



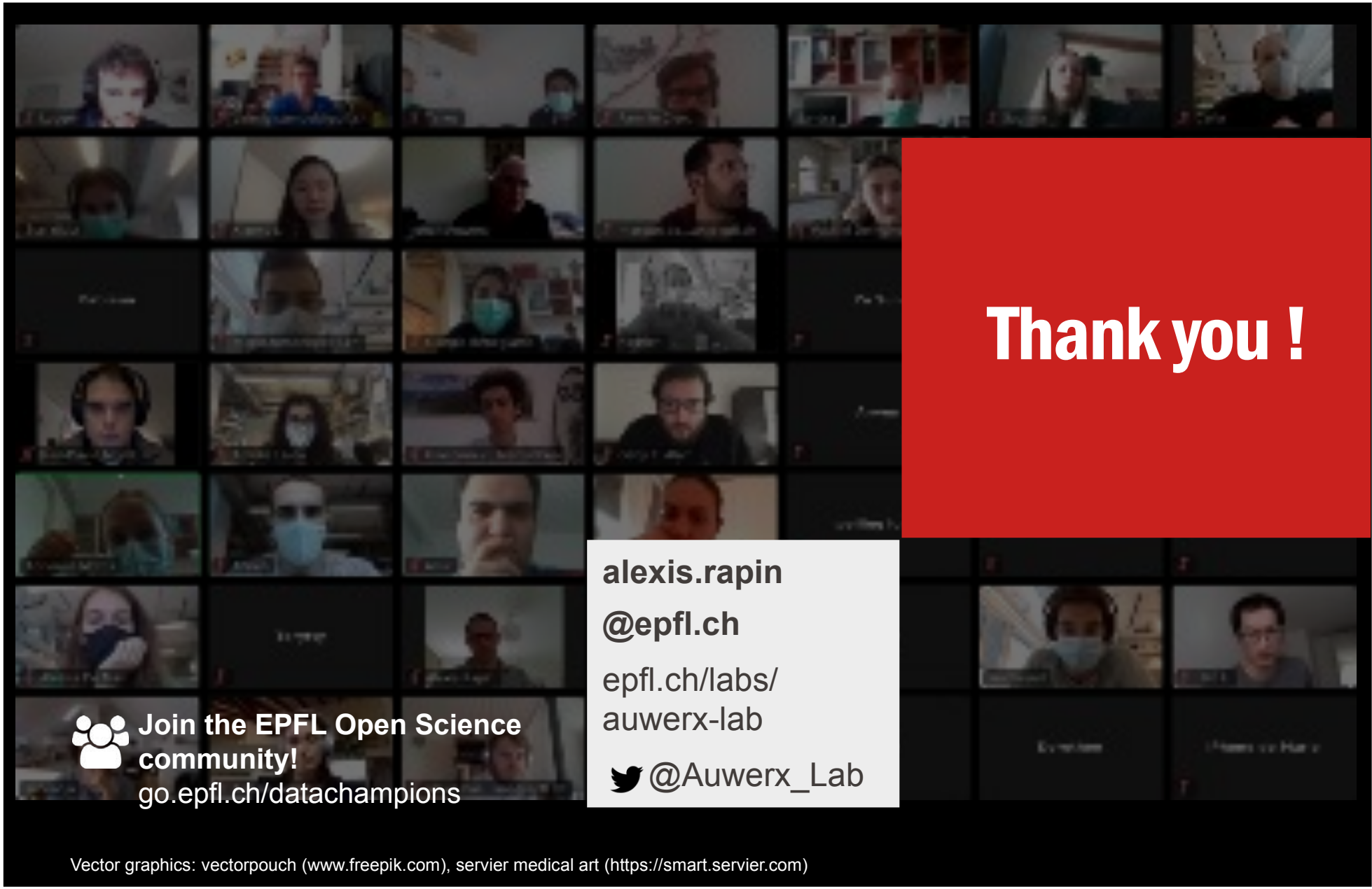
As aging research accelerates, we may live healthier, longer



Alexis Rapin

EPFL

SDSC



Thank you !

alexis.rapin

@epfl.ch

epfl.ch/labs/
auwerx-lab

@Auwerx_Lab

Join the EPFL Open Science community!
go.epfl.ch/datachampions

Vector graphics: vectorpouch (www.freepik.com), servier medical art (https://smart.servier.com)