Data integration and multi-omics

Thomas Stoeger
thomas.stoeger@northwestern.edu

November 24th, 2020

How this lecture will work:

How this lecture will work:

Q/A on handout

How this lecture will work:

Q/A on handout

mini-presentation

How this lecture will work:

Q/A on handout

mini-presentation

group coding exercise

How this lecture will work:

Q/A on handout

get overview
know where to find details

mini-presentation

group coding exercise

How this lecture will work:

Q/A on handout

get overview
know where to find details

mini-presentation

solidify main points

group coding exercise

How this lecture will work:

Q/A on handout	get overview know where to find details
mini-presentation	solidify main points
group coding exercise	experience

How this lecture will work:

Q/A on handout	get overview know where to find details
mini-presentation	solidify main points
group coding exercise	experience

Q/A on handout

get overview
know where to find details

Q/A on handout

get overview
know where to find details

Handout (7min to read)

https://github.com/tstoeger/course_multi_omics/blob/main/multi-omics.pdf

Q/A on handout

get overview
know where to find details

Handout (7min to read)

https://github.com/tstoeger/course_multi_omics/blob/main/multi-omics.pdf

Questions (~5min)

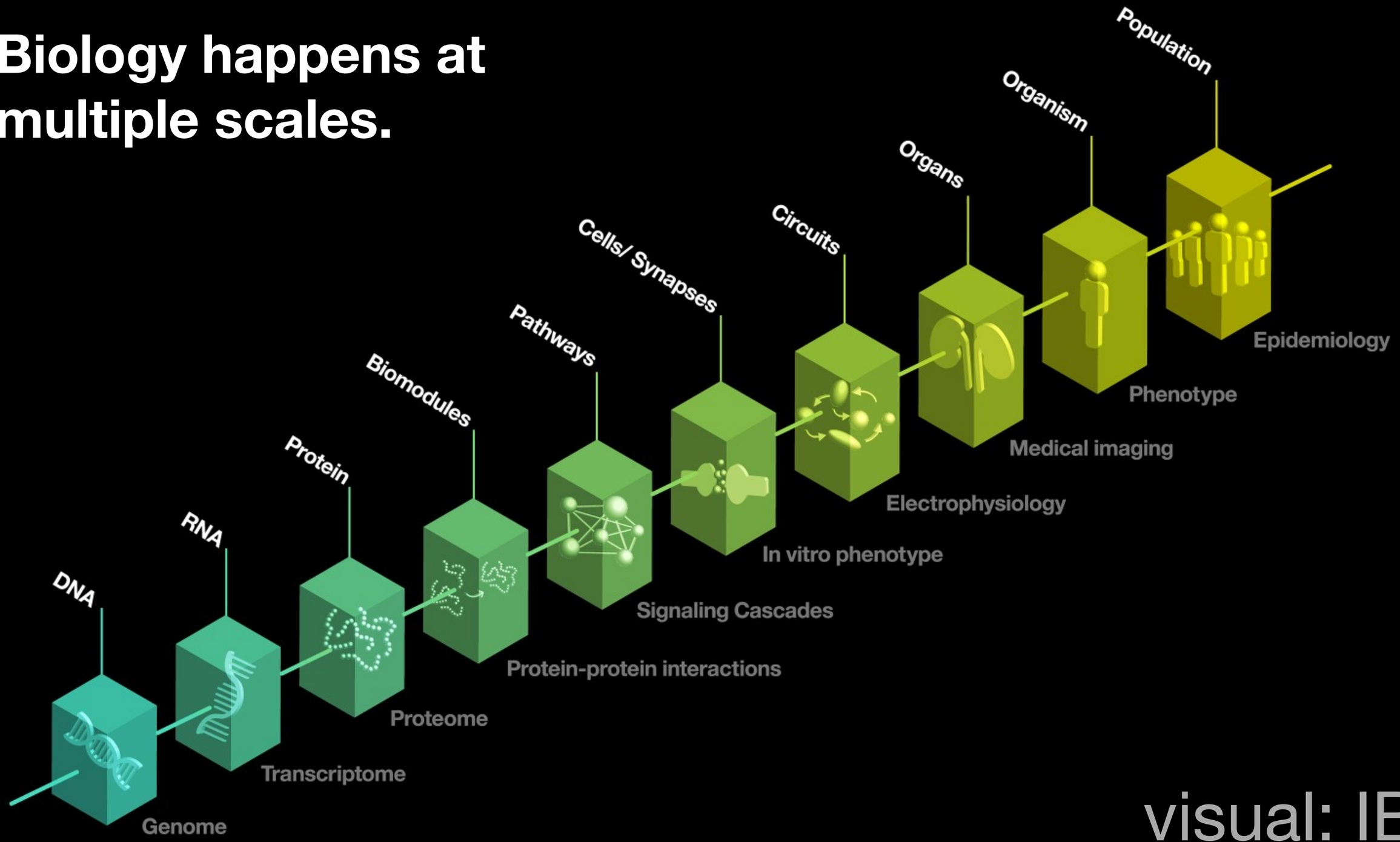
Ask in person.

Ask anonymously on:
https://padlet.com/thomasstoeger/a4mtvpym671dwgnl

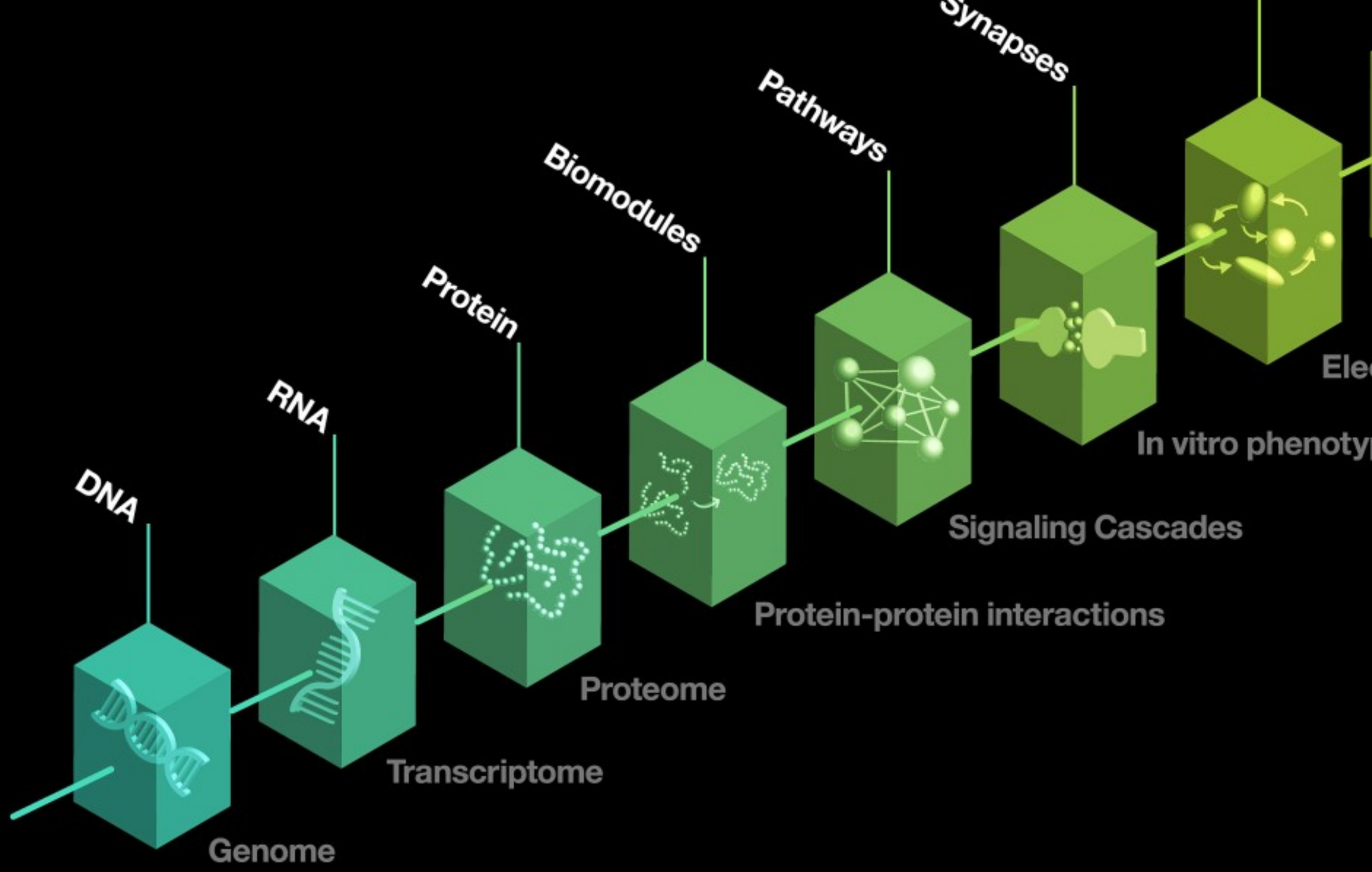
mini-presentation

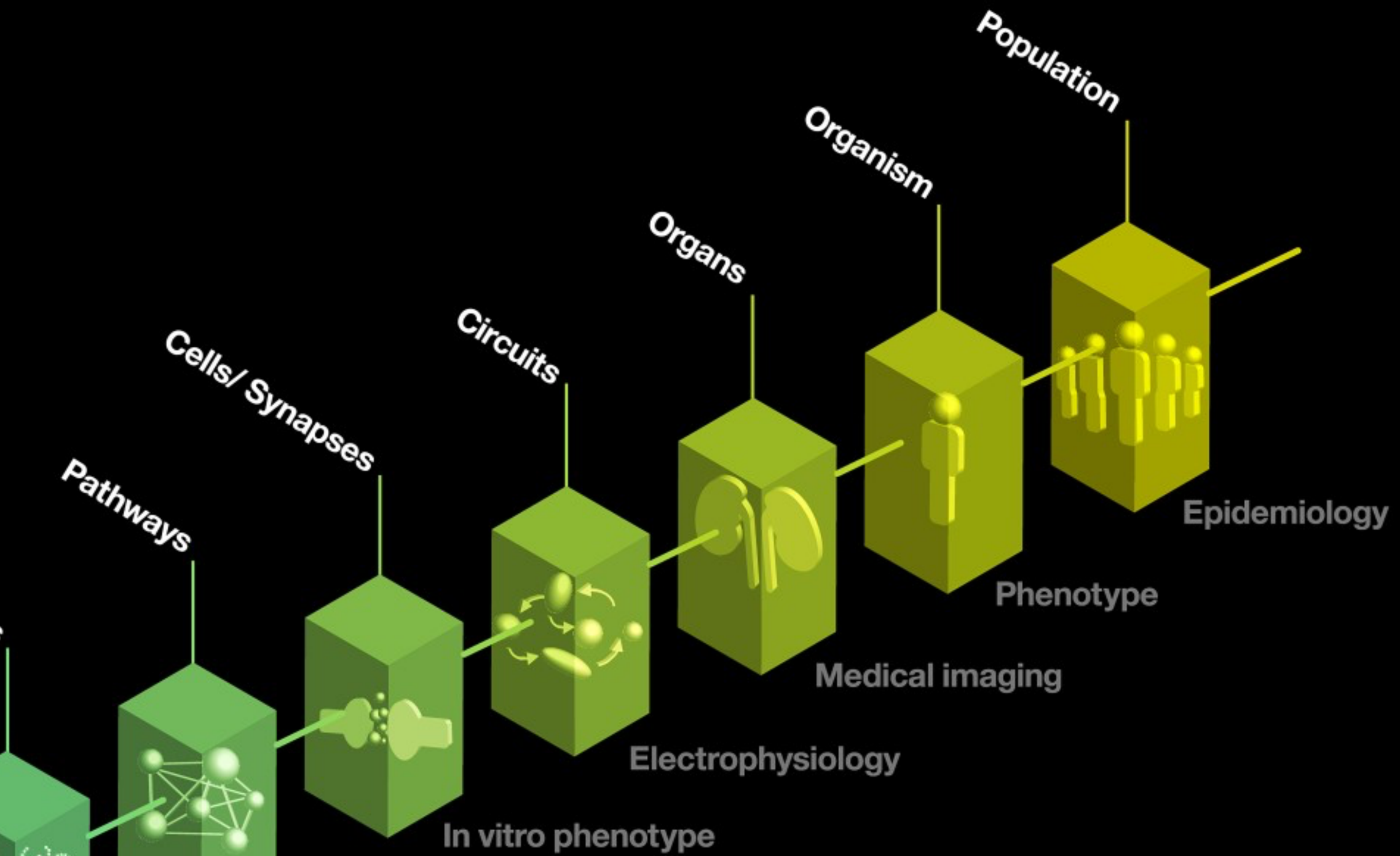
solidify main points

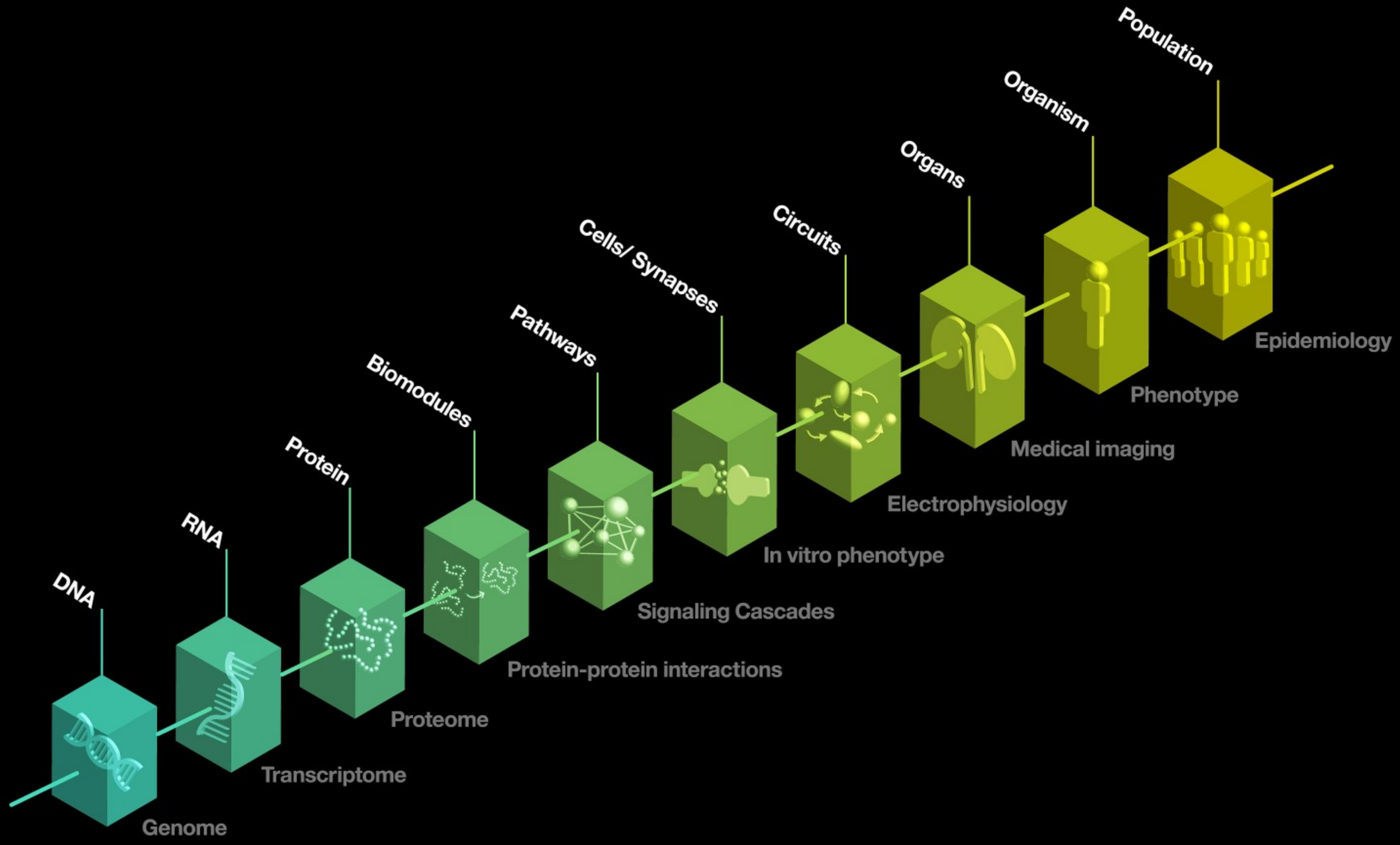
Biology happens at multiple scales.



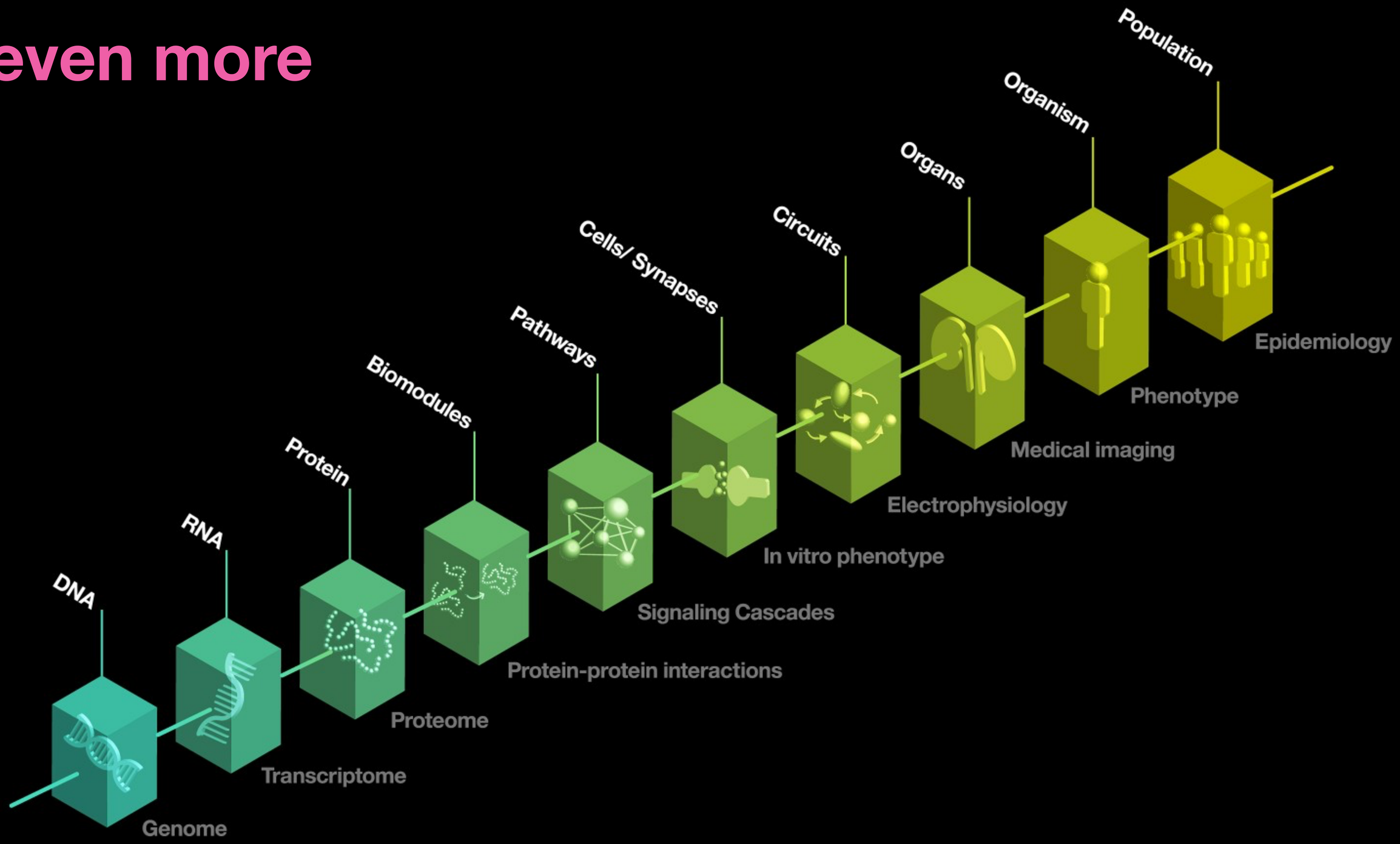
visual: IBM


DNA
RNA
Protein
Biomodules
Pathways
Synapses
Genome
Transcriptome
Proteome
Protein-protein interactions
Signaling Cascades
In vitro phenoty
Ele


Population
Organism
Organs
Circuits
Cells/ Synapses
Pathways
Epidemiology
Phenotype
Medical imaging
Electrophysiology
In vitro phenotype

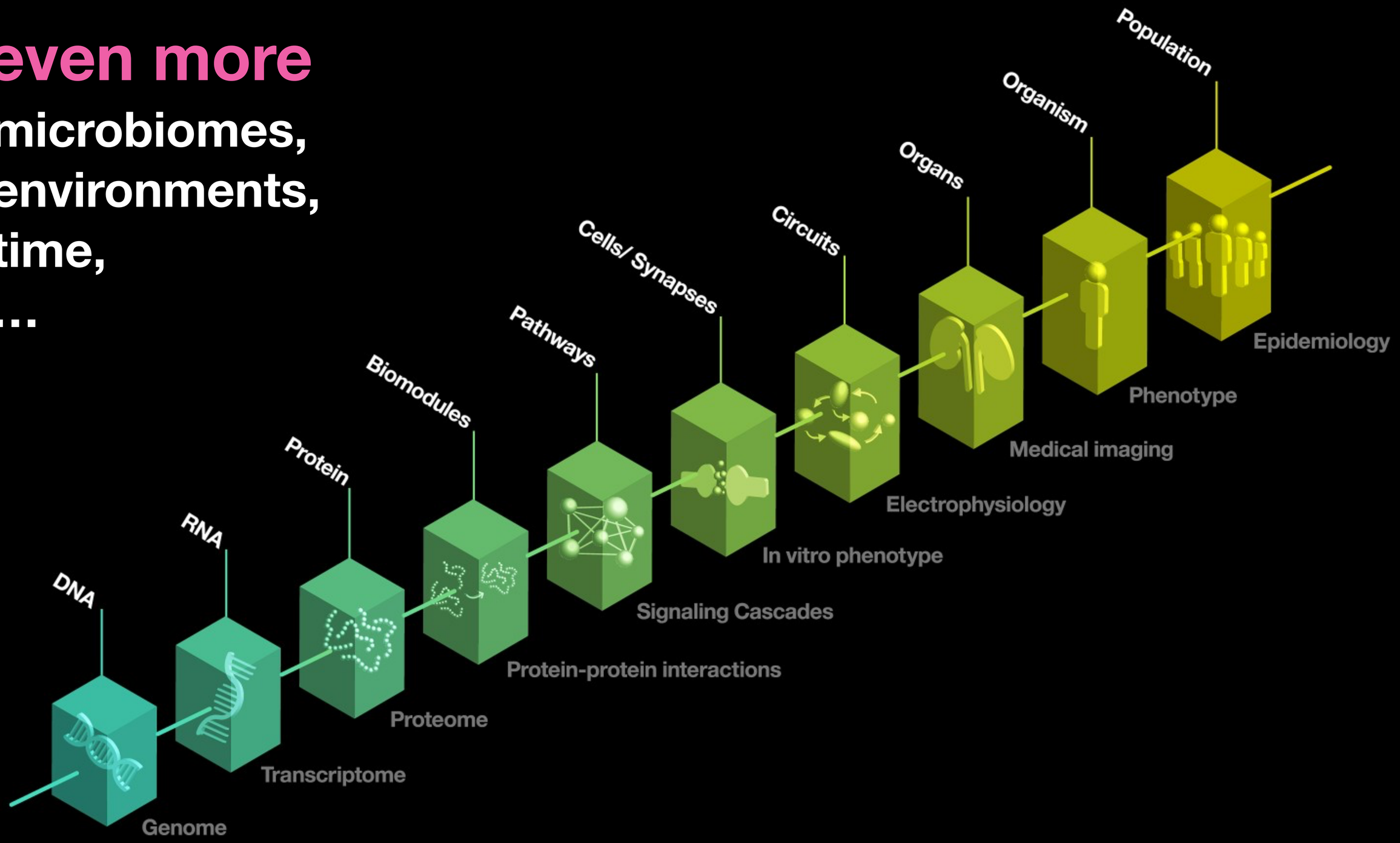

DNA
RNA
Protein
Biomodules
Pathways
Cells/ Synapses
Circuits
Organs
Organism
Population
Genome
Transcriptome
Proteome
Protein-protein interactions
Signaling Cascades
In vitro phenotype
Electrophysiology
Medical imaging
Phenotype
Epidemiology

even more


DNA
RNA
Protein
Biomodules
Pathways
Cells/ Synapses
Circuits
Organs
Organism
Population
Genome
Transcriptome
Proteome
Protein-protein interactions
Signaling Cascades
In vitro phenotype
Electrophysiology
Medical imaging
Phenotype
Epidemiology

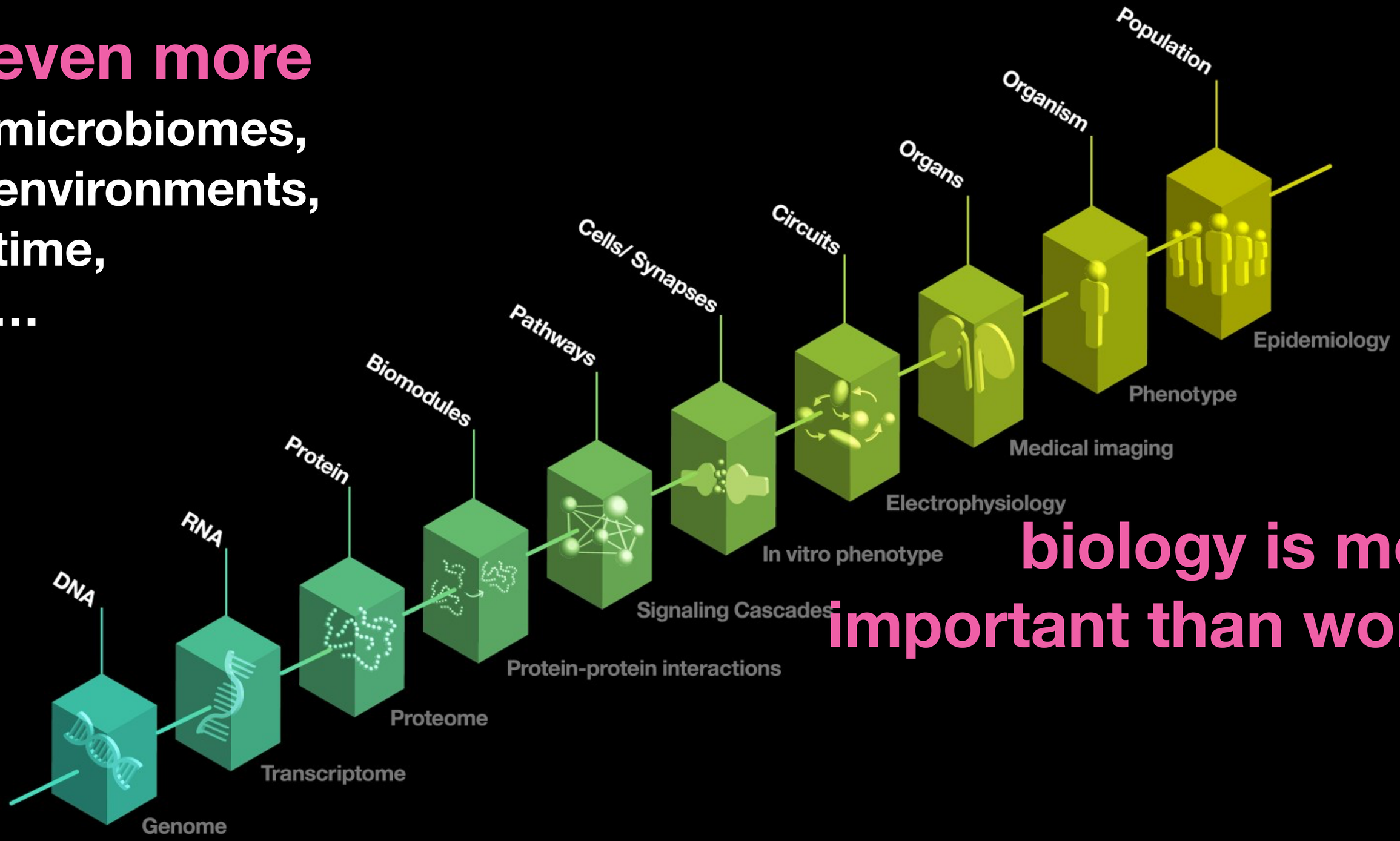
even more

microbiomes,
environments,
time,
...



even more

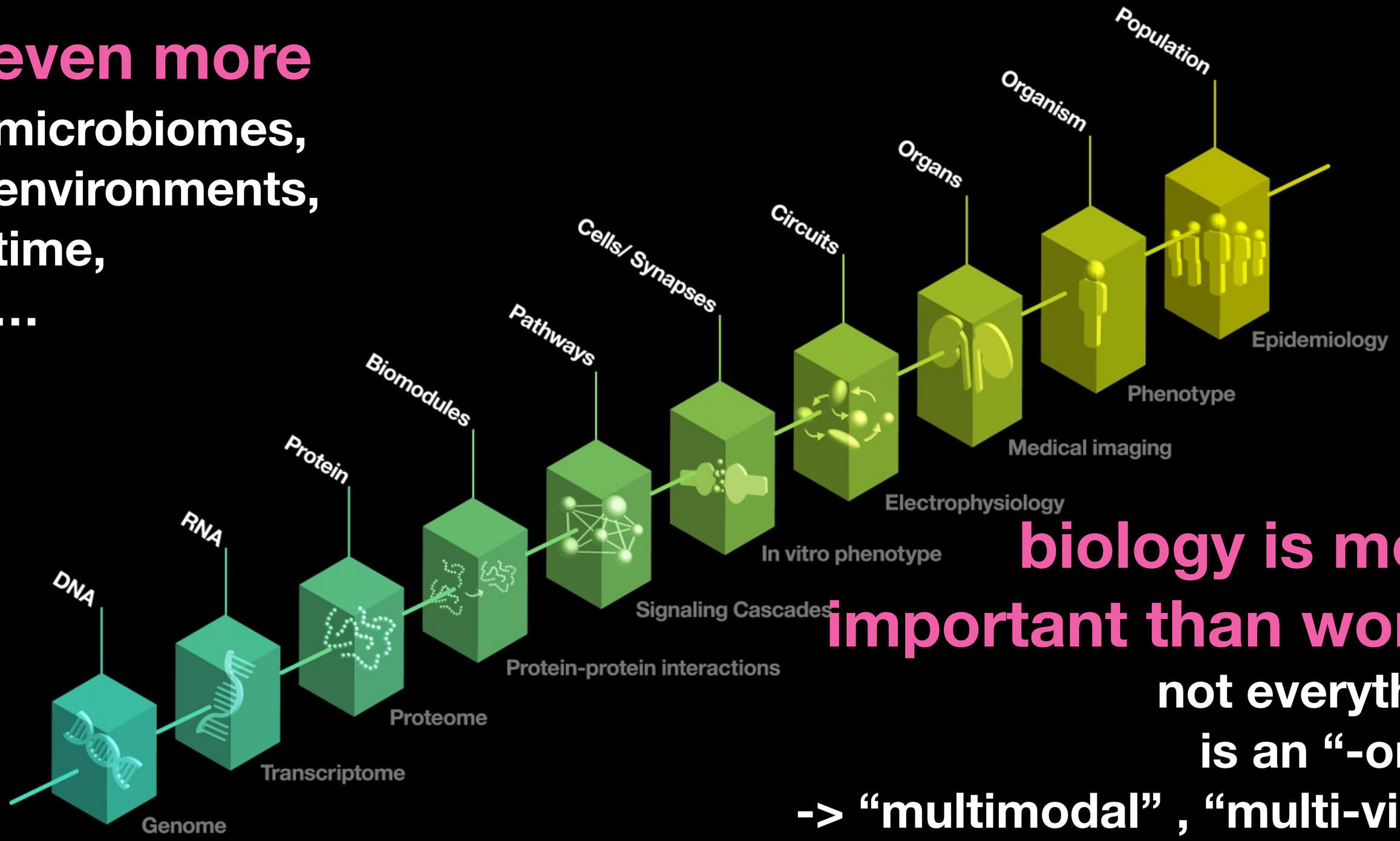
microbiomes,
environments,
time,
...



biology is more important than words

even more

microbiomes,
environments,
time,
...

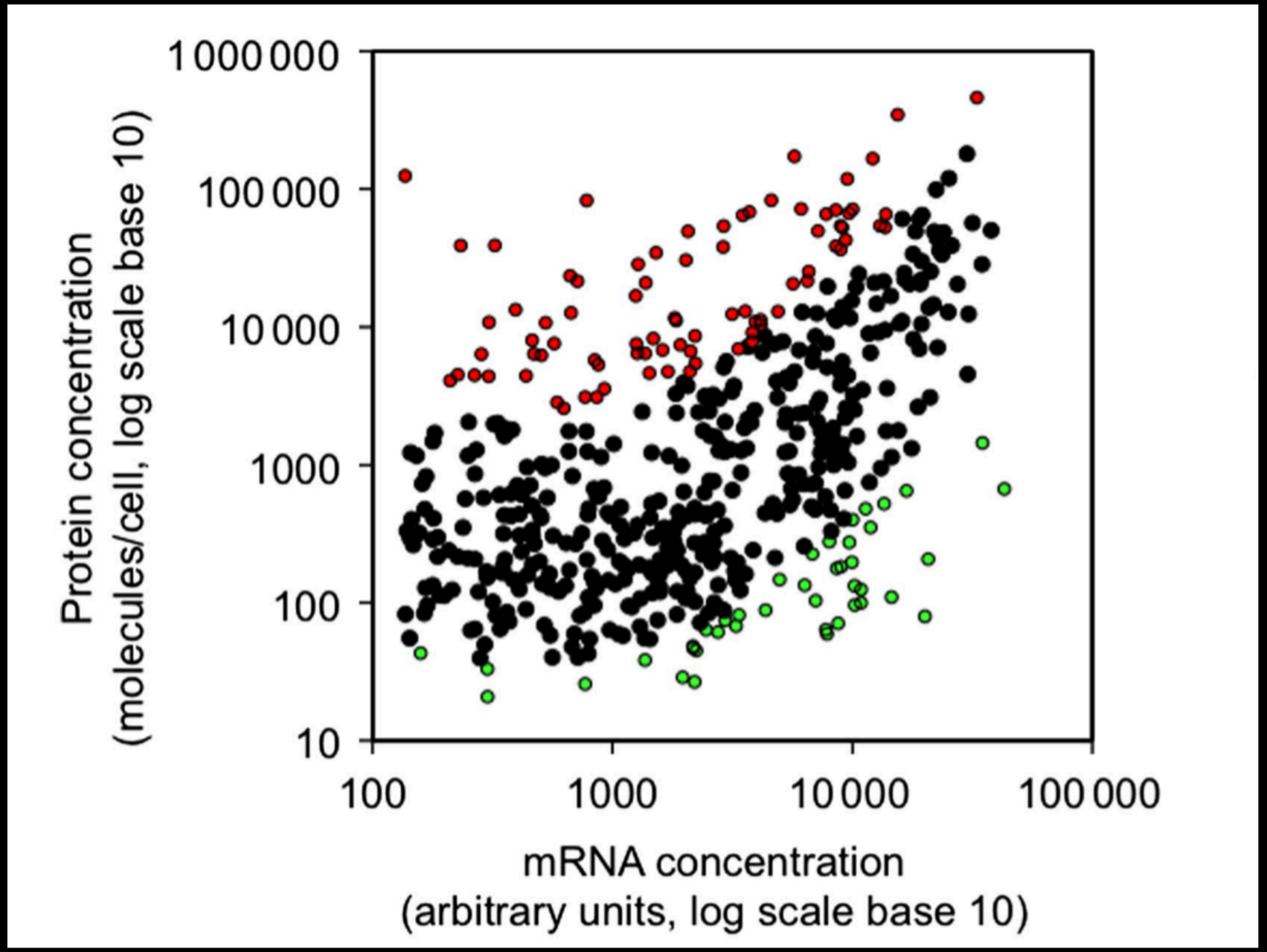


biology is more important than words

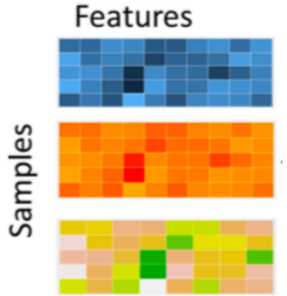
not everything
is an “-ome”
-> “multimodal” , “multi-view”

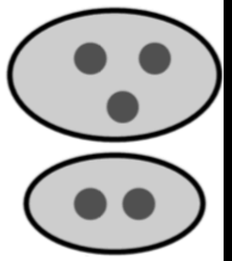
many combinations

Be careful!


1 000 000
100 000
10 000
1000
100
10
Protein concentration
(molecules/cell, log scale base 10)
100
1000
10 000
100 000
mRNA concentration
(arbitrary units, log scale base 10)

There are many different strategies.





Pavlidis et al., 2001

Rappoport and Shamir, 2018

Stuart and Satija, 2019

Figure 1. Overview of multi-omics clustering approaches.

There are many different strategies.

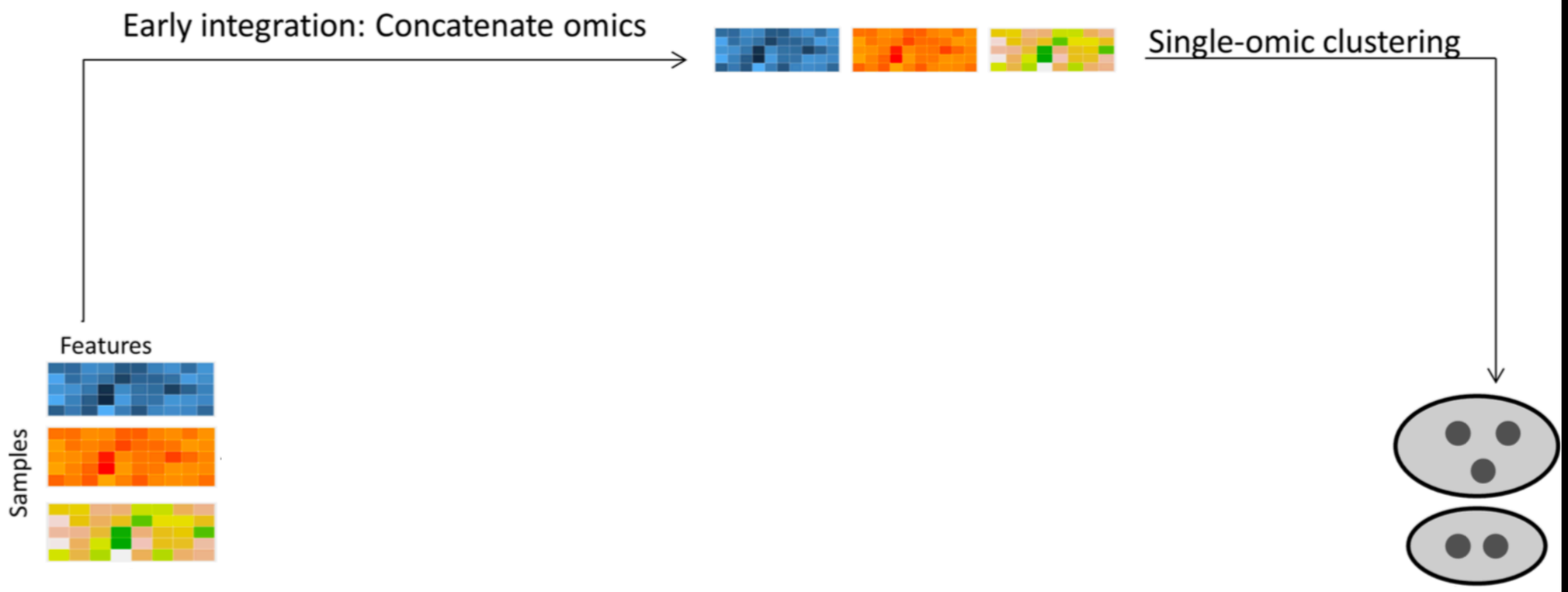


Figure 1. Overview of multi-omics clustering approaches.

Pavlidis et al., 2001

Rappoport and Shamir, 2018

Stuart and Satija, 2019

There are many different strategies.

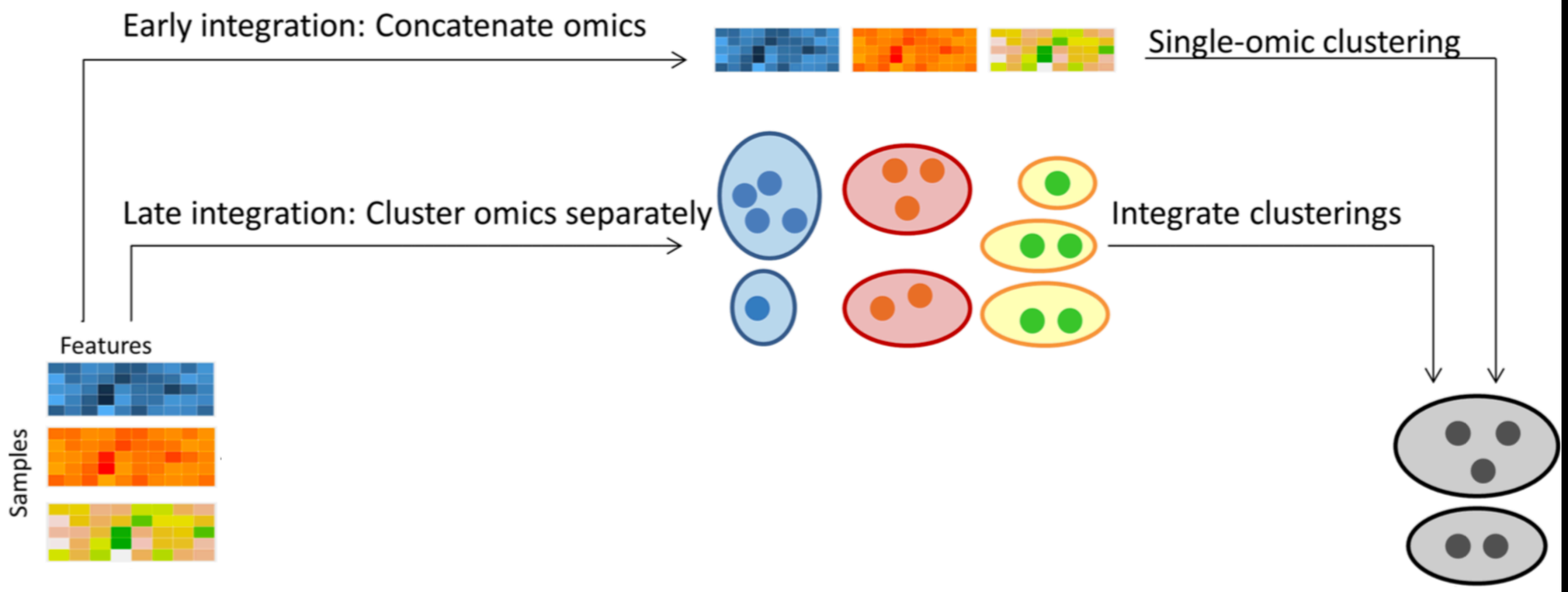


Figure 1. Overview of multi-omics clustering approaches.

Pavlidis et al., 2001

Rappoport and Shamir, 2018

Stuart and Satija, 2019

There are many different strategies.

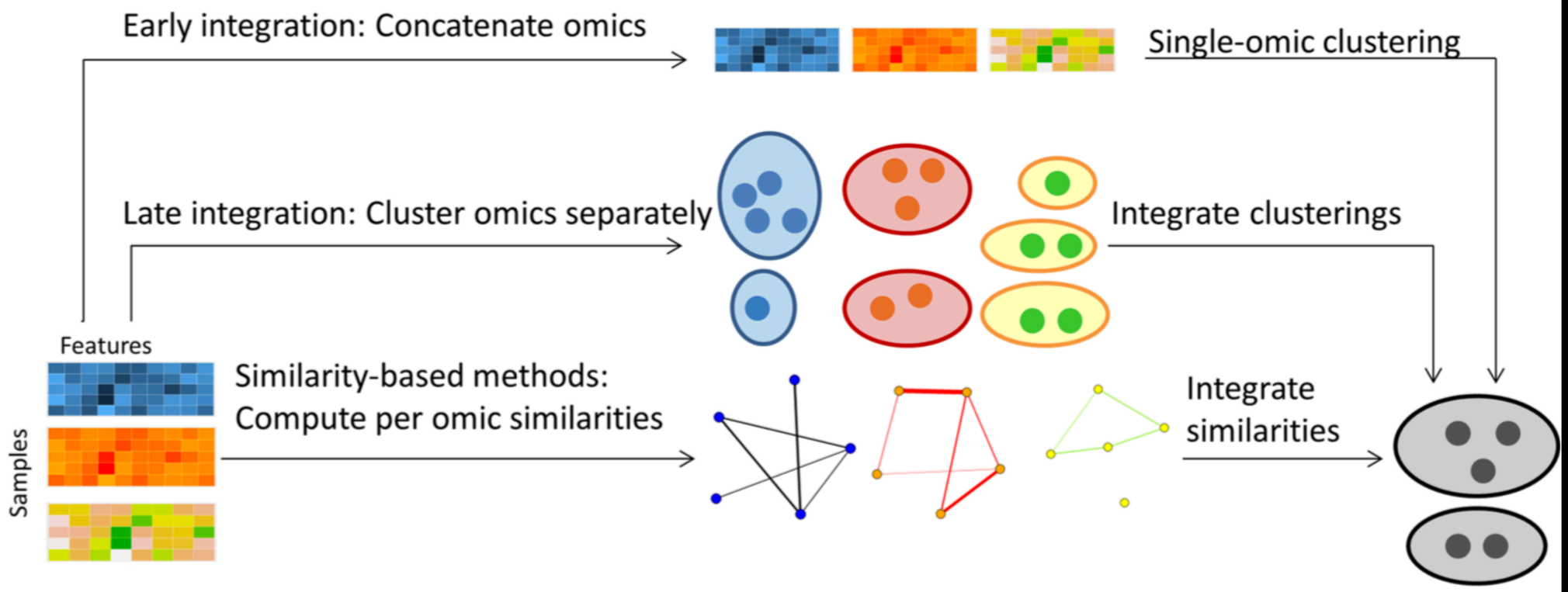


Figure 1. Overview of multi-omics clustering approaches.

Pavlidis et al., 2001

Rappoport and Shamir, 2018

Stuart and Satija, 2019

There are many different strategies.

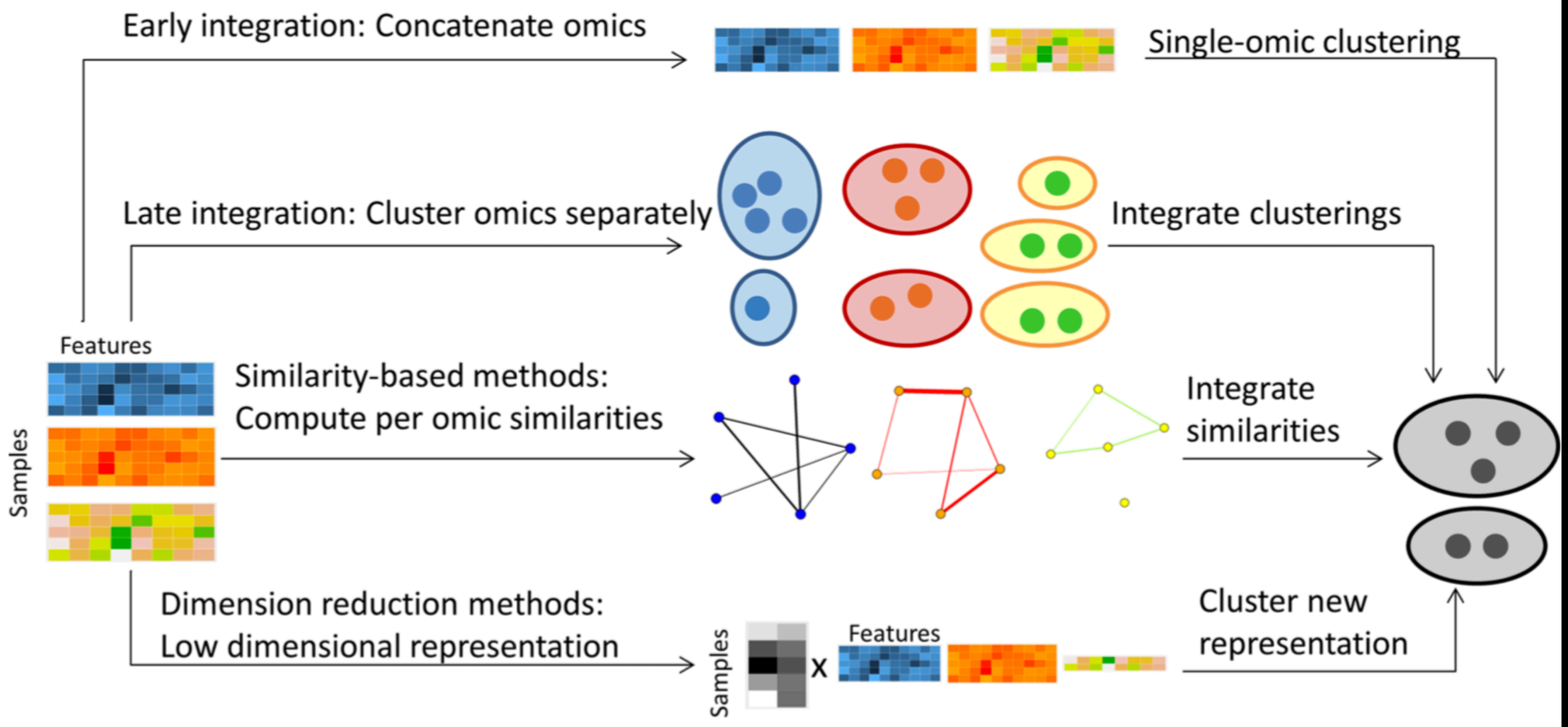


Figure 1. Overview of multi-omics clustering approaches.

Pavlidis et al., 2001

Rappoport and Shamir, 2018

Stuart and Satija, 2019

Organize data by adhering to good practices.

THE AMERICAN STATISTICIAN
2018, VOL. 72, NO. 1, 2–10
https://doi.org/10.1080/00031305.2017.1375989

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Data Organization in Spreadsheets

Karl W. Broman[a] and Kara H. Woo[b]

[a]Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, WI; [b]Information School, University of Washington, Seattle, WA

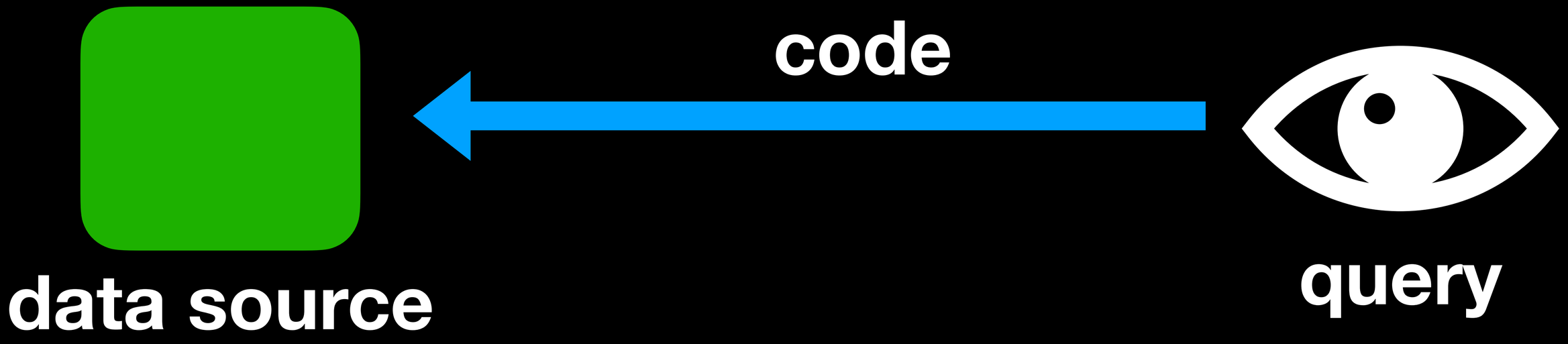
ABSTRACT

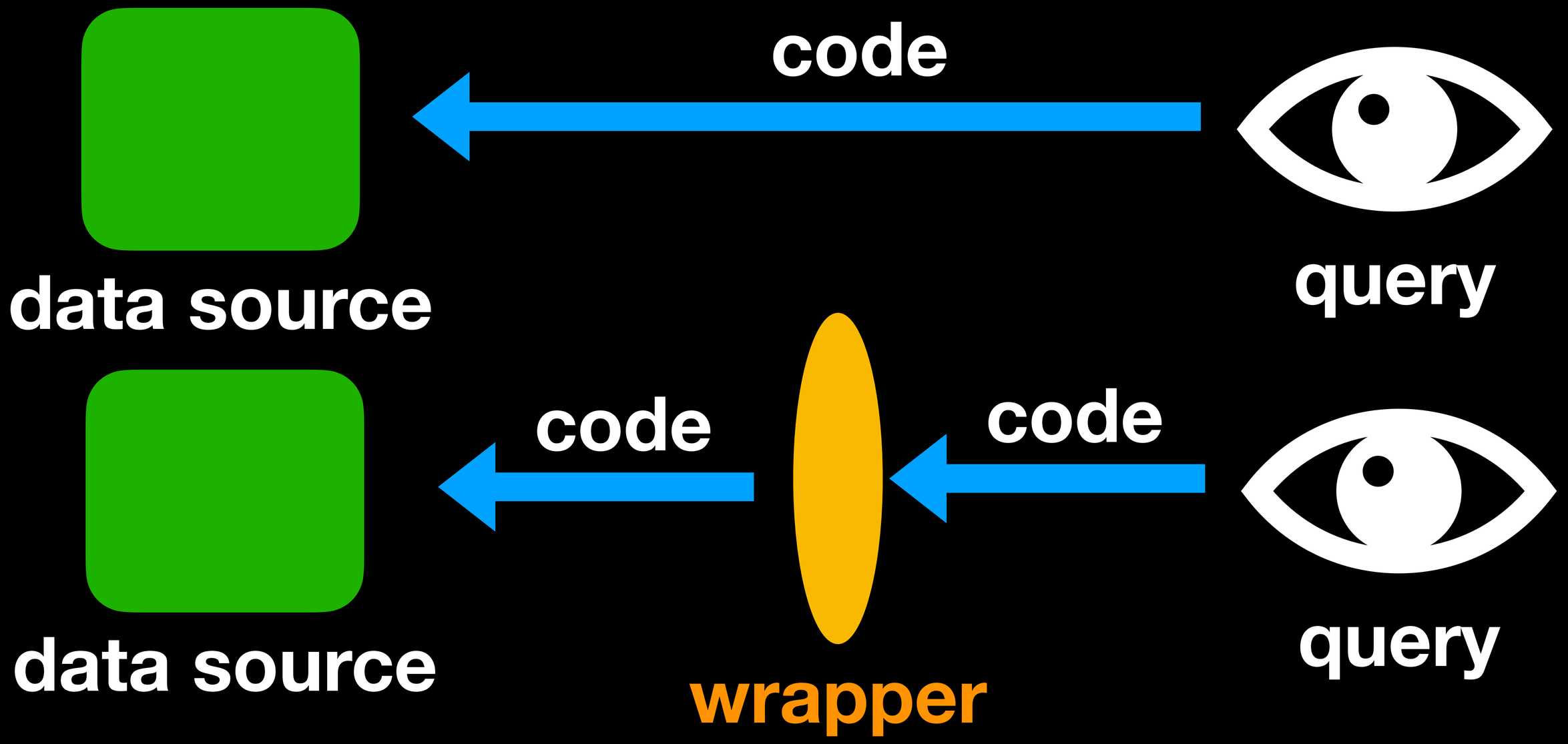
Spreadsheets are widely used software tools for data entry, storage, analysis, and visualization. Focusing on the data entry and storage aspects, this article offers practical recommendations for organizing spreadsheet data to reduce errors and ease later analyses. The basic principles are: be consistent, write dates like YYYY-MM-DD, do not leave any cells empty, put just one thing in a cell, organize the data as a single rectangle (with subjects as rows and variables as columns, and with a single header row), create a data dictionary, do not include calculations in the raw data files, do not use font color or highlighting as data, choose good names for things, make backups, use data validation to avoid data entry errors, and save the data in plain text files.

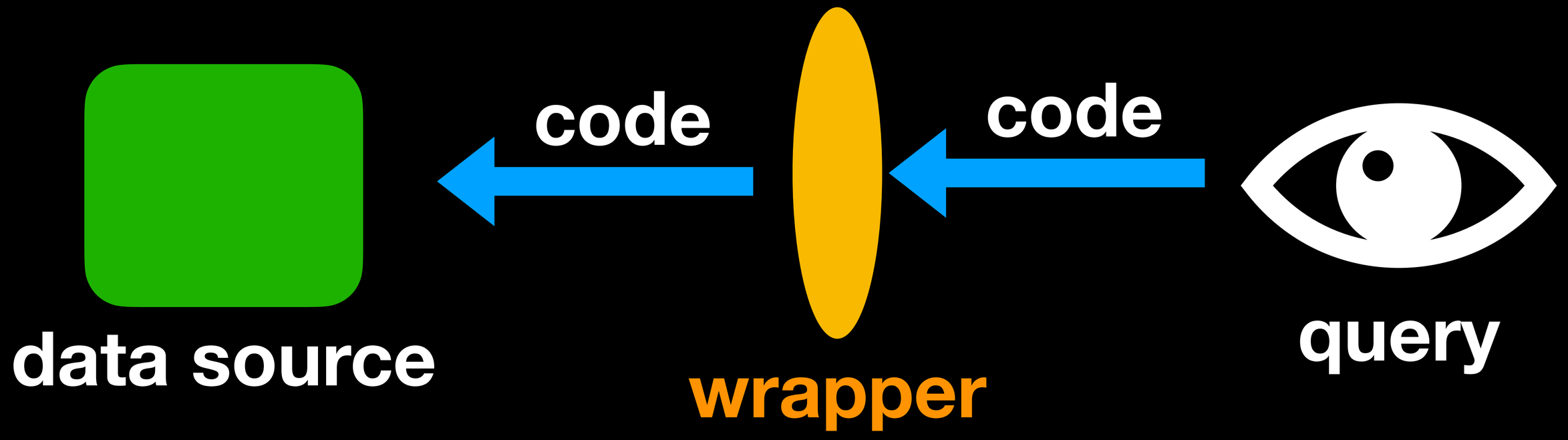
ARTICLE HISTORY
Received June 2017
Revised August 2017

KEYWORDS
Data management; Data organization; Microsoft Excel; Spreadsheets

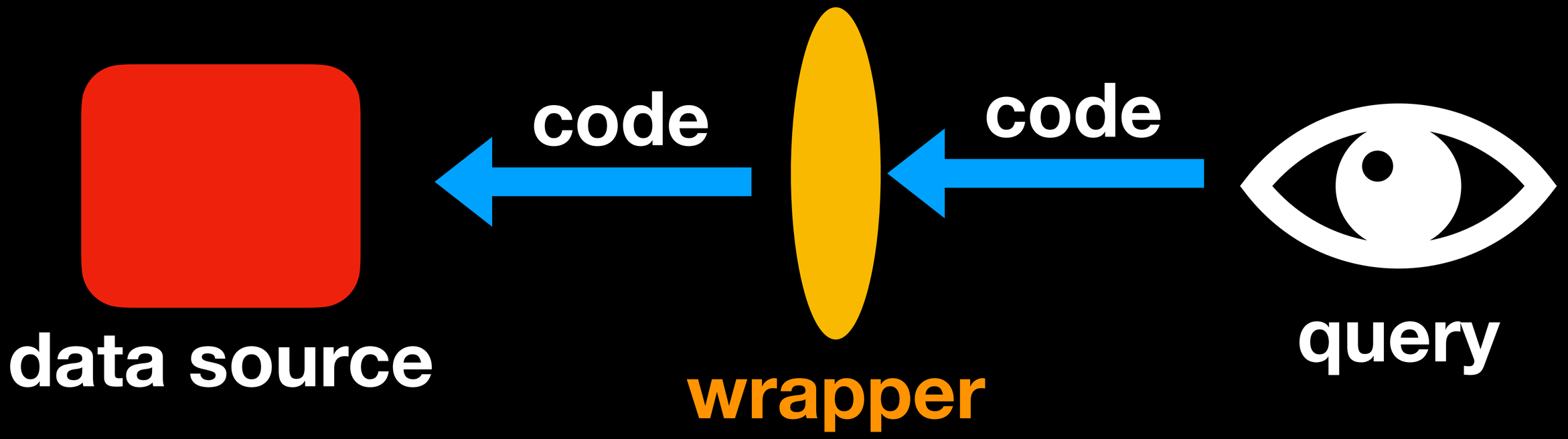
Wrappers, something very useful to learn for everyone working with data.


code
data source
query


code
data source
query
code
code
data source
wrapper
query

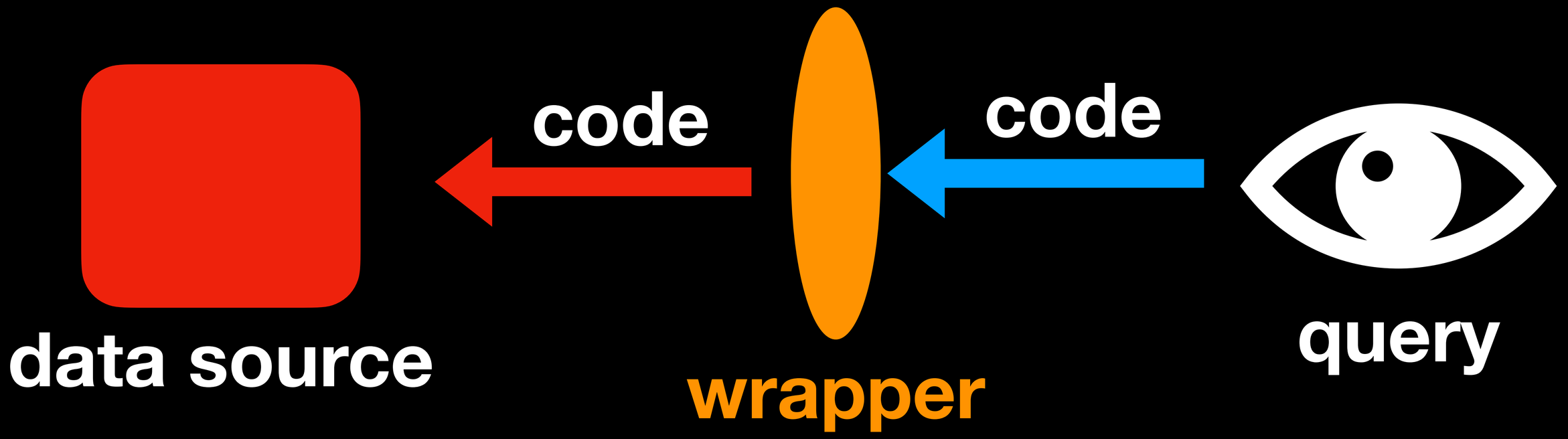

code
code
data source
wrapper
query

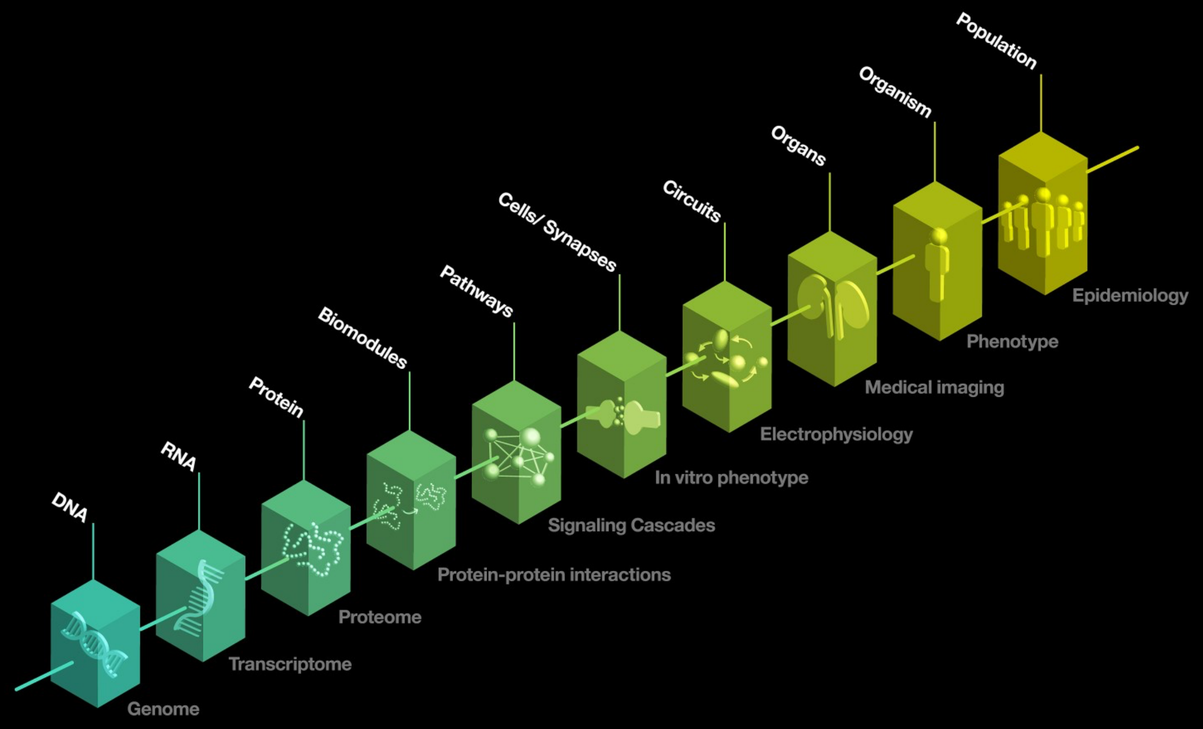
other / new
differently
formatted

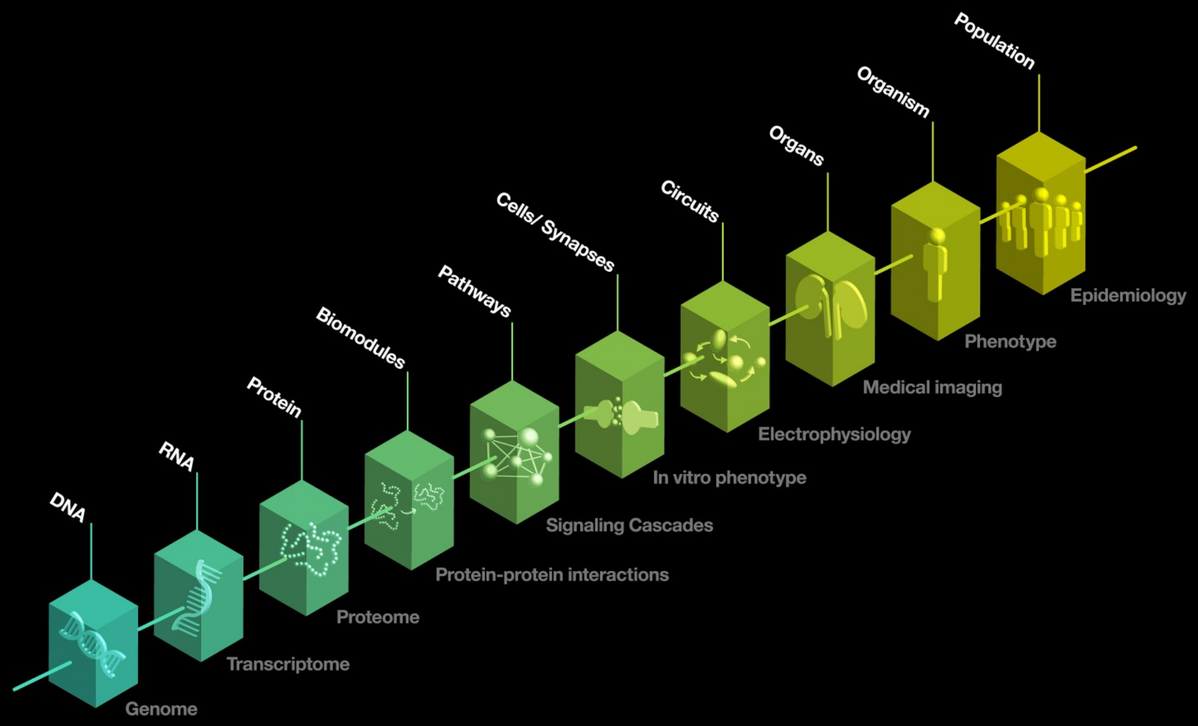


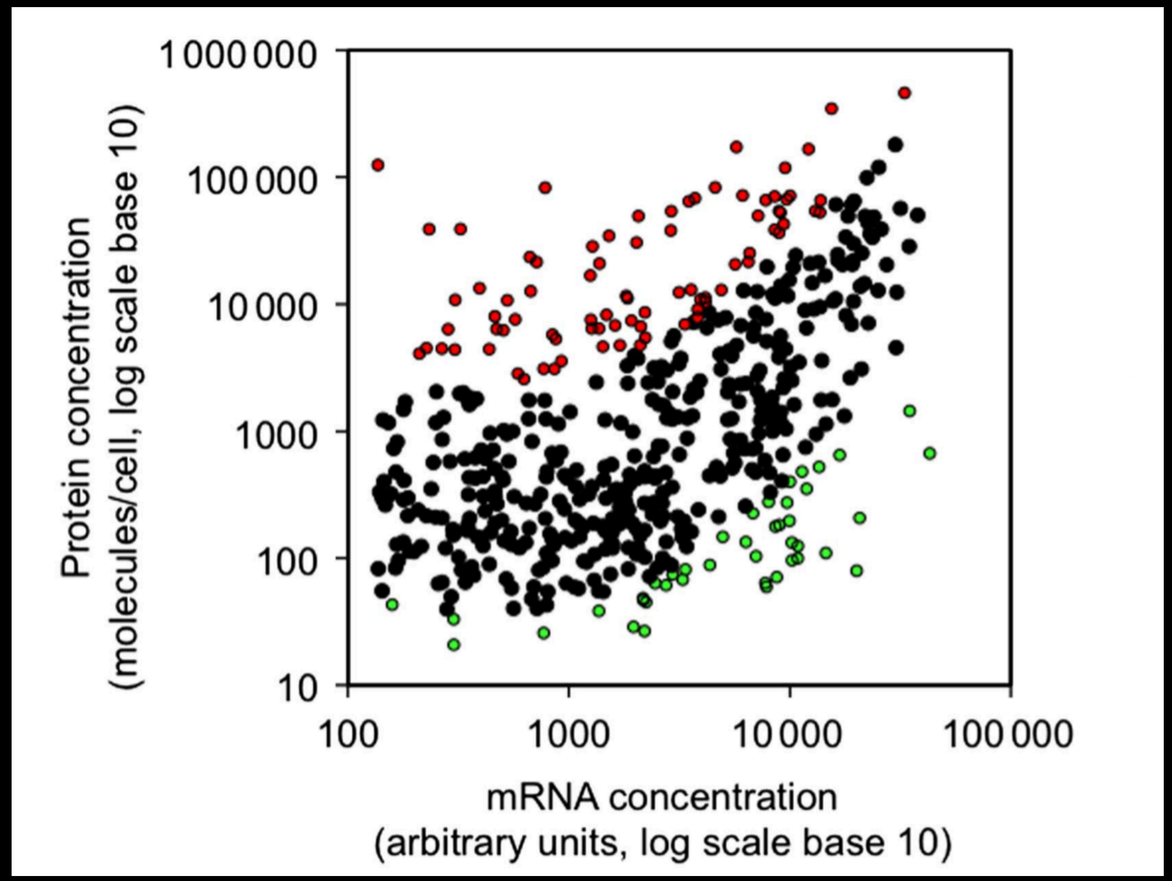
other / new
differently
formatted

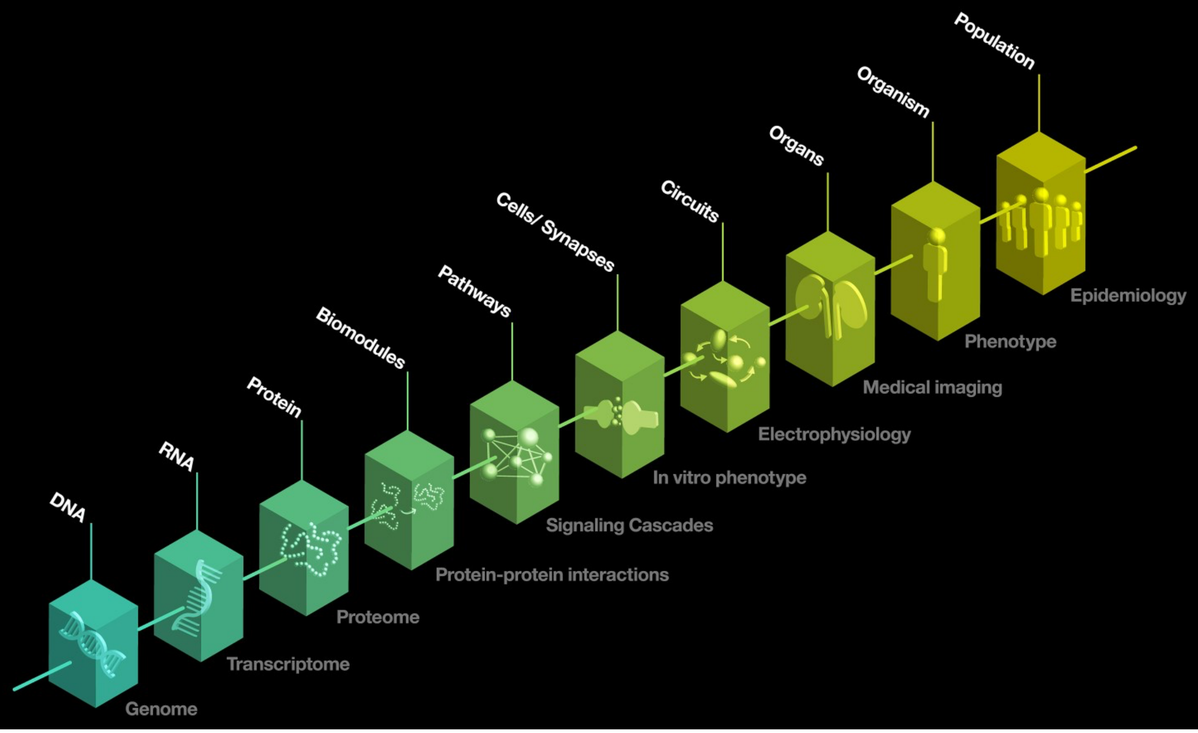
optimize
speed if
frequently used

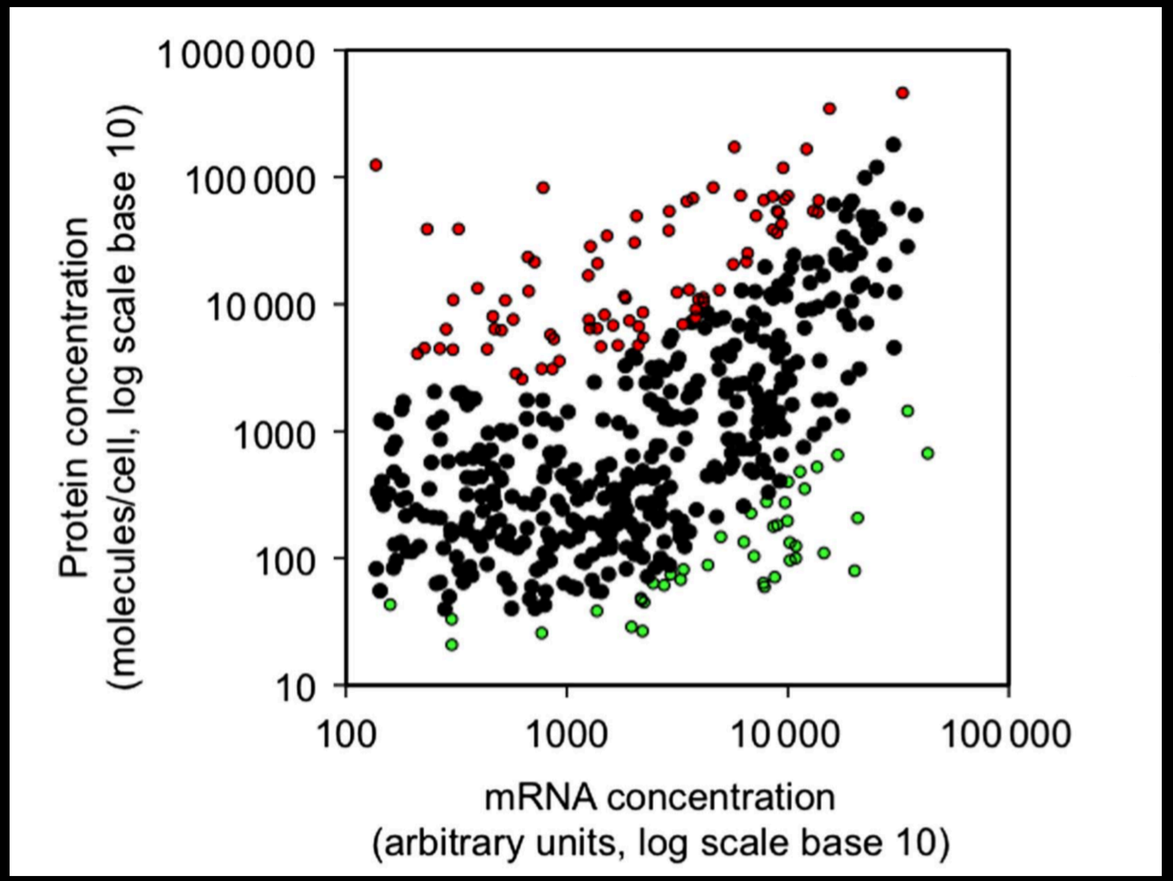



DNA
RNA
Protein
Biomodules
Pathways
Cells/ Synapses
Circuits
Organs
Organism
Population
Genome
Transcriptome
Proteome
Protein-protein interactions
Signaling Cascades
In vitro phenotype
Electrophysiology
Medical imaging
Phenotype
Epidemiology


DNA
RNA
Protein
Biomodules
Pathways
Cells/ Synapses
Circuits
Organs
Organism
Population
Genome
Transcriptome
Proteome
Protein-protein interactions
Signaling Cascades
In vitro phenotype
Electrophysiology
Medical imaging
Phenotype
Epidemiology


Protein concentration
(molecules/cell, log scale base 10)
1 000 000
100 000
10 000
1000
100
10
100
1000
10 000
100 000
mRNA concentration
(arbitrary units, log scale base 10)





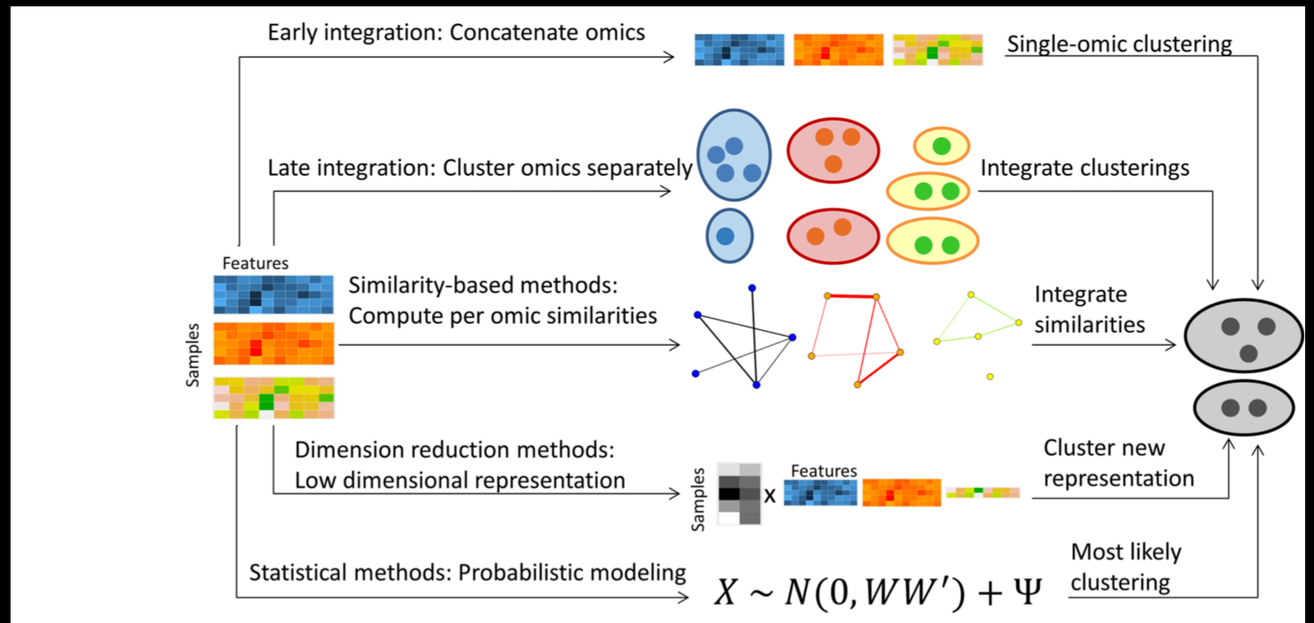
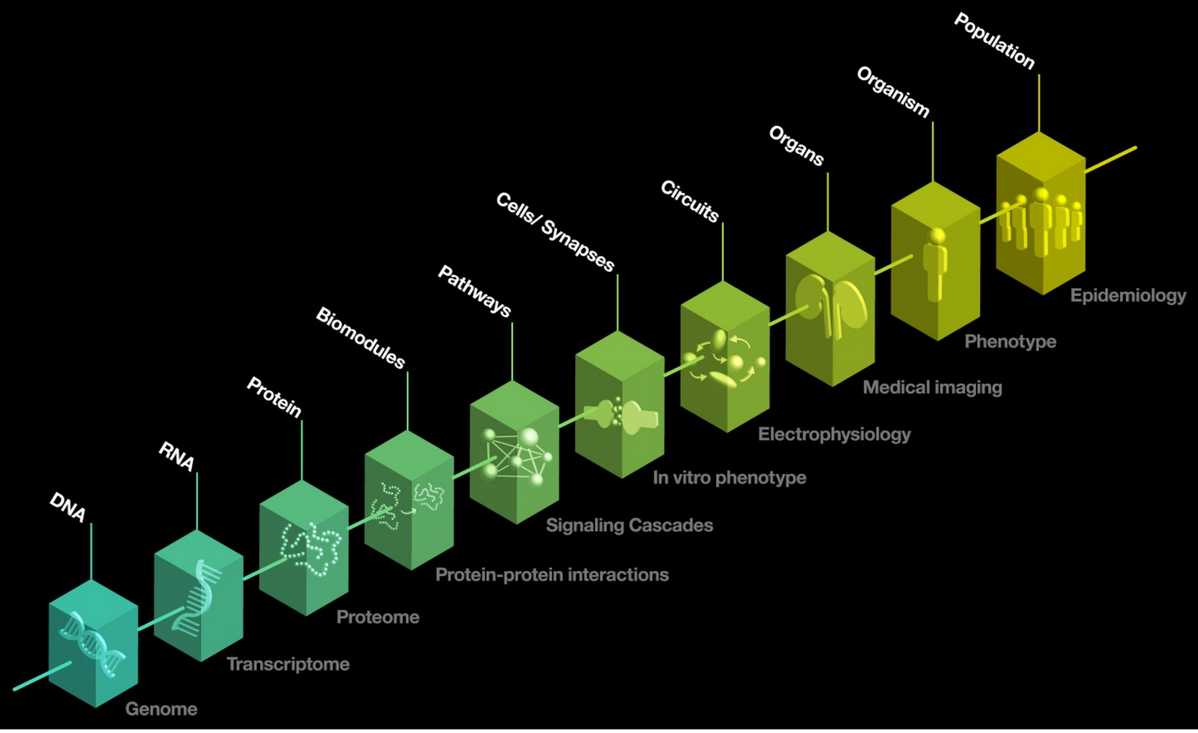
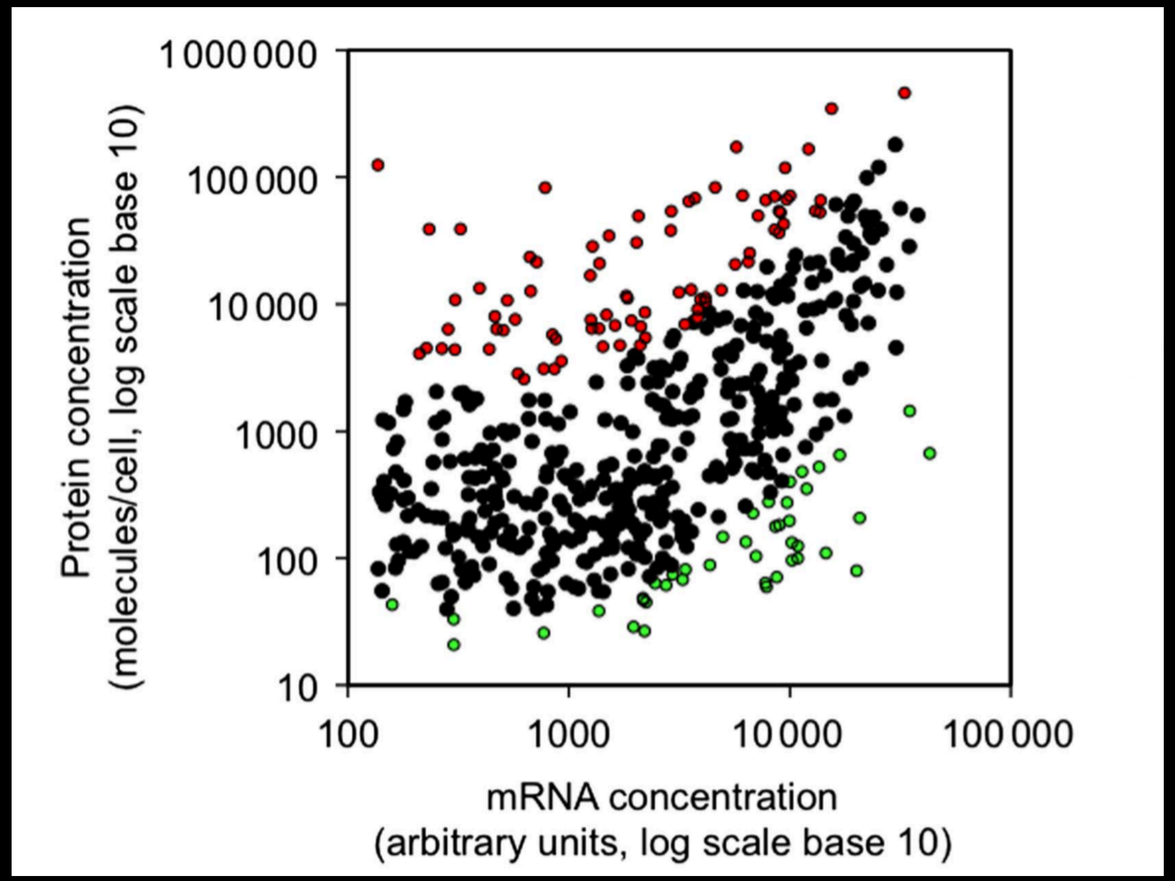


Figure 1. Overview of multi-omics clustering approaches.





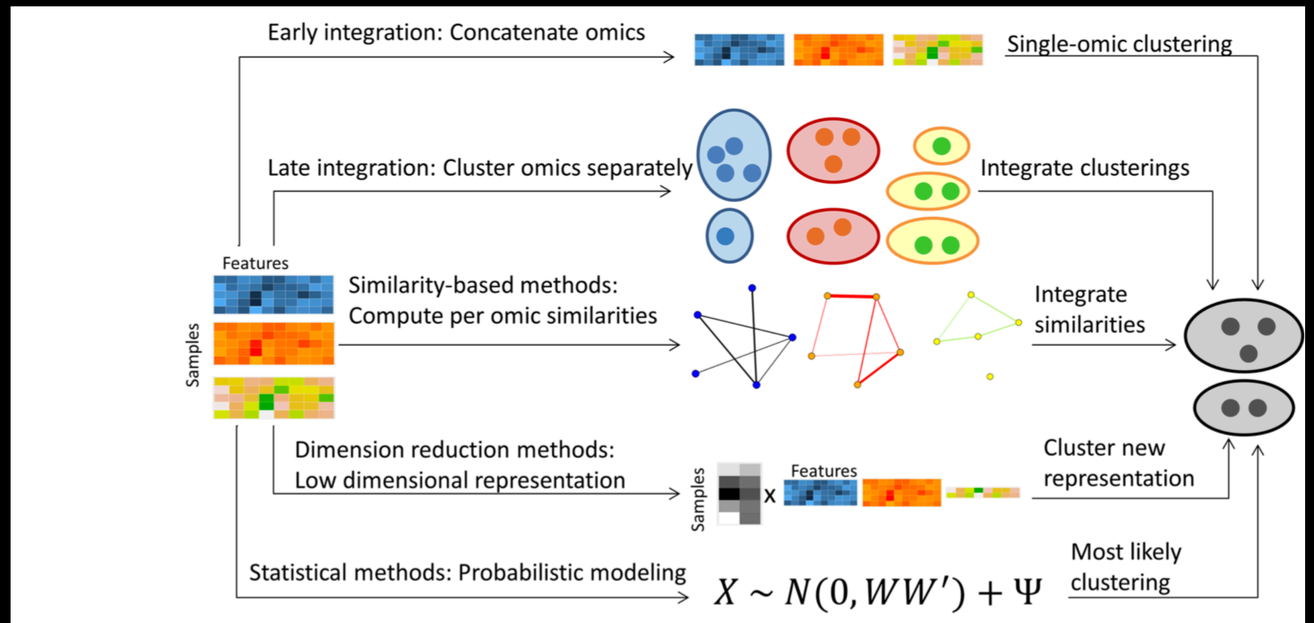
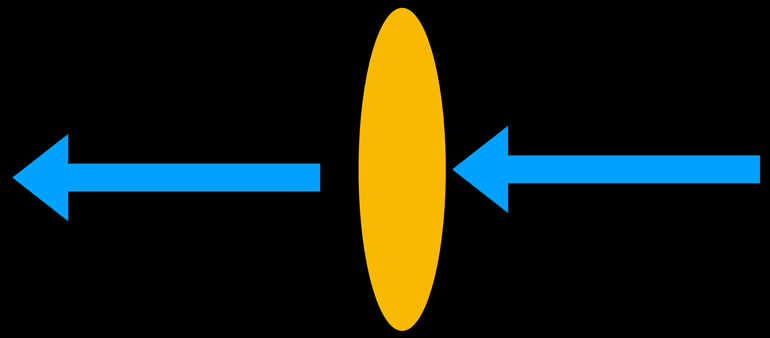


Figure 1. Overview of multi-omics clustering approaches.



wrapper

group coding exercise

experience

30 minutes

30 minutes

3 people per breakout group

30 minutes

3 people per breakout group

free choice of tools

30 minutes

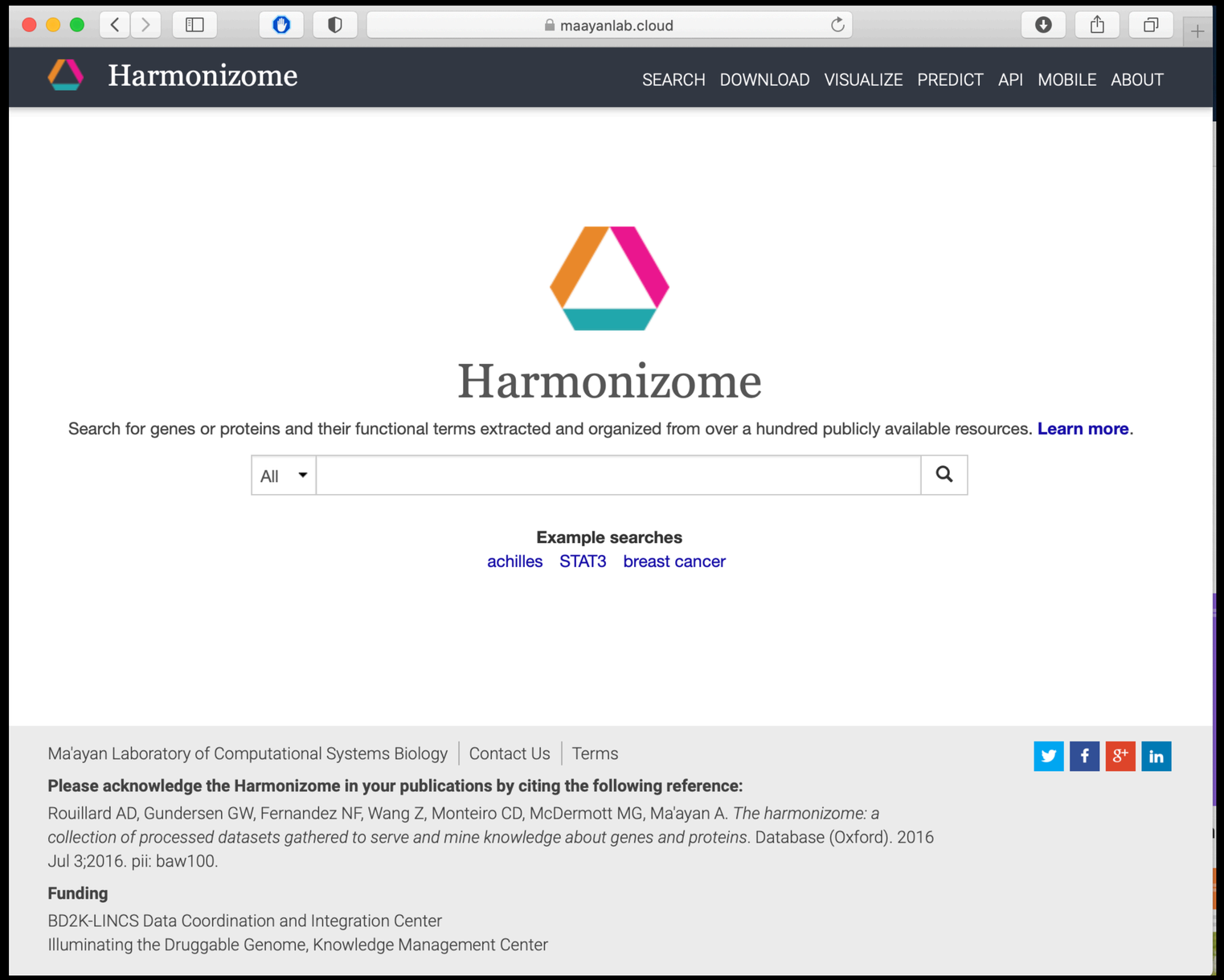
3 people per breakout group

free choice of tools

Are similarities between genes conserved across scales?

Data Source: https://maayanlab.cloud/Harmonizome/

subset: https://northwestern.box.com/s/dvrxld7ioe6jgm2srrs7rj8mzkqpkhsa (see handout)



Provides precomputed similarities between genes.

Are similarities between genes conserved across scales?

Note: while you might answer this within 30 minutes, identifying possible challenges is even more valuable.

Suggested: a) take two datasets that intrigue you – possibly small to limit download time b) work / talk with your fellow team members

Data Source: https://maayanlab.cloud/Harmonizome/

subset: https://northwestern.box.com/s/dvrxld7ioe6jgm2srrs7rj8mzkqpkhsa (see handout)

Template code: https://github.com/tstoeger/course_multi_omics/blob/main/how_well_do_different_views_on_biology_correlate.ipynb

Ask anonymously on: https://padlet.com/thomasstoeger/a4mtvpym671dwgnl