Reconnaissance and Recommendation: Wayfinding Through Data With Visualization

Tamara Munzner
Department of Computer Science
University of British Columbia

Visualization in Data Science 2023 keynote
23 Oct 2023, Melbourne Australia

http://www.cs.ubc.ca/~tmm/talks.html#vds23

UBC

DESIGNING *for* PEOPLE

UBC Data Science Institute

CAIDA

@tamara@vis.social

@tamaramunzner

Extended analogy

- wayfinding through the world with road trips

- wayfinding through data with visualization

http://www.cs.ubc.ca/~tmm/talks.html#vds23

https://unsplash.com/photos/2oYHfuRe4OU

Questions in road trips

- where are we?
- what's here?
- are we there yet? are we lost?

https://unsplash.com/photos/2oYHfuRe4OU

http://www.cs.ubc.ca/~tmm/talks.html#vds23

Questions in road trips - and visualization in data science!

- with each VDS project, addressing more questions

- where are we?
  - Data Reconnaissance & Task Wrangling
- what's here?
  - Automatic Encodings through Recommendation
- are we there yet? are we lost?
  - Visual Assessment of ML Training Completion & Quality

http://www.cs.ubc.ca/~tmm/talks.html#vds23

https://unsplash.com/photos/2oYHfuRe4OU

Uncovering Data Landscapes through

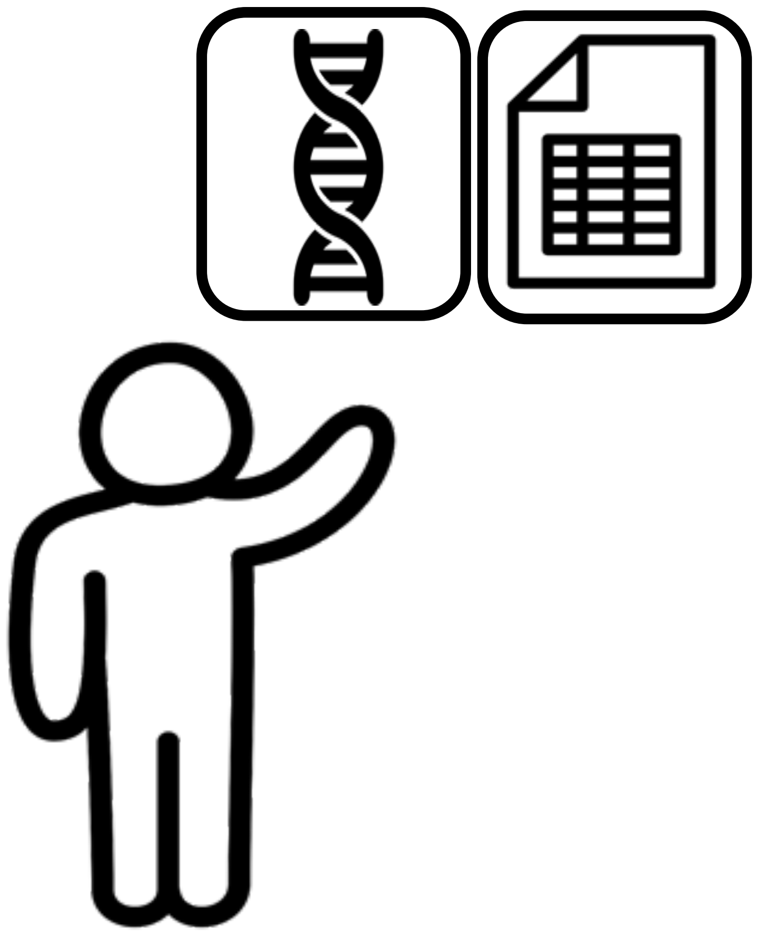
Data Reconnaissance & Task Wrangling

Anamaria Crisan
@amcrisan
UBC/Tableau

Tamara Munzner
@tamaramunzner
@tamara@vis.social
UBC

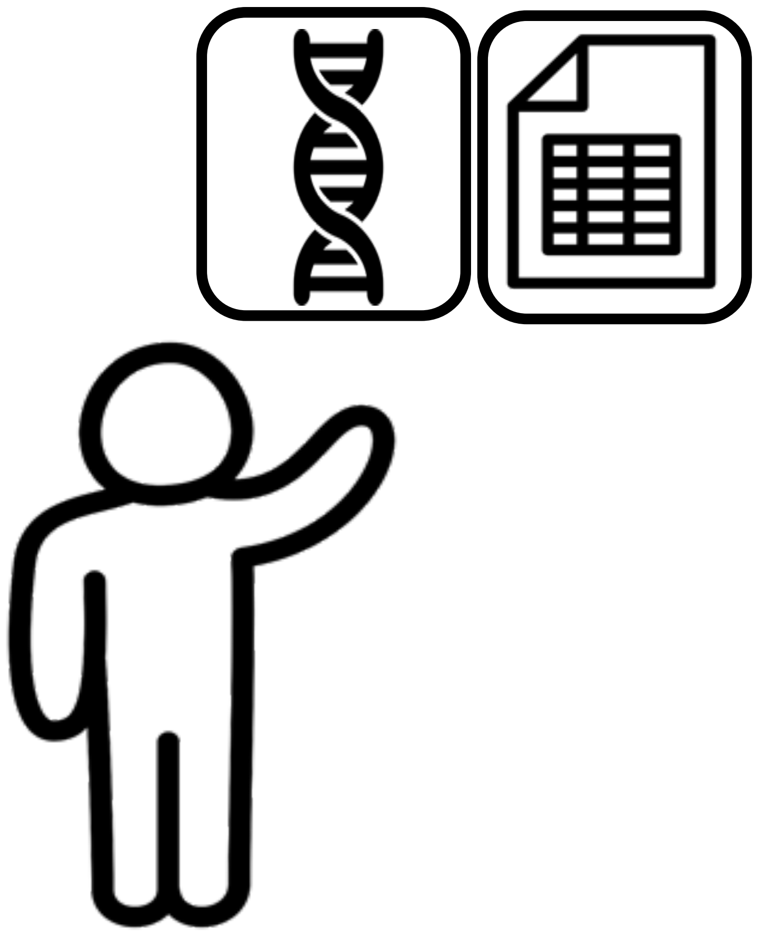
https://amcrisan.github.io/assets/files/papers/Data_Recon_and_Task_Wrangling.pdf

Uncovering Data Landscapes through Data Reconnaissance and Task Wrangling
Crisan, Munzner. Proc. IEEE VIS 2019, pp. 46-50.



Where are we?

Domain experts need help uncovering and reasoning about heterogeneous data landscapes

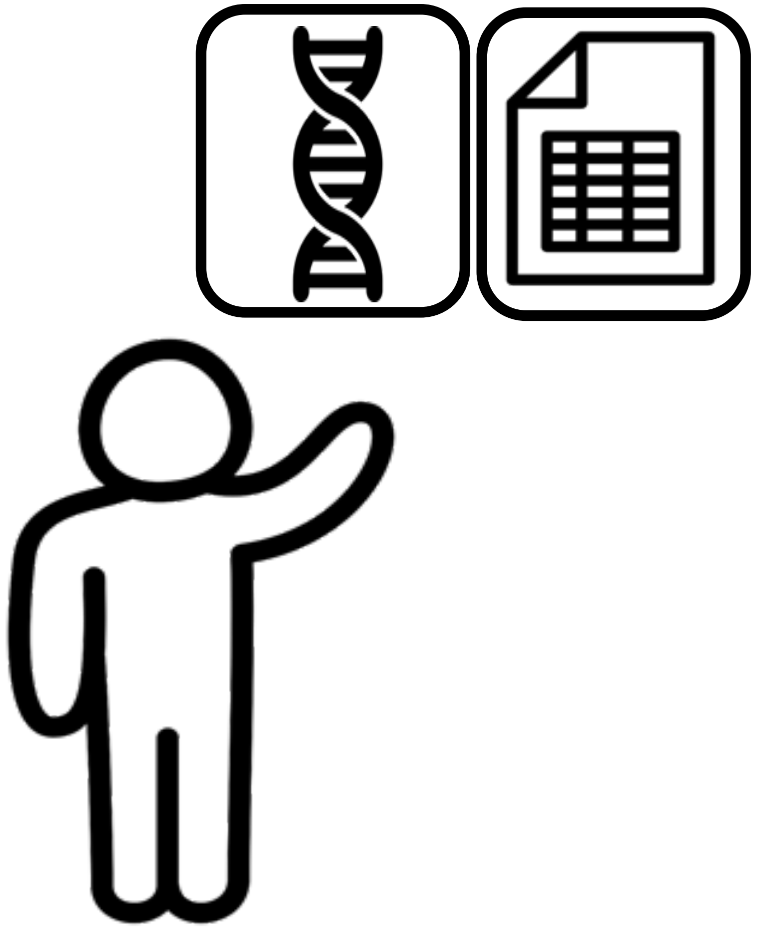


Data landscape

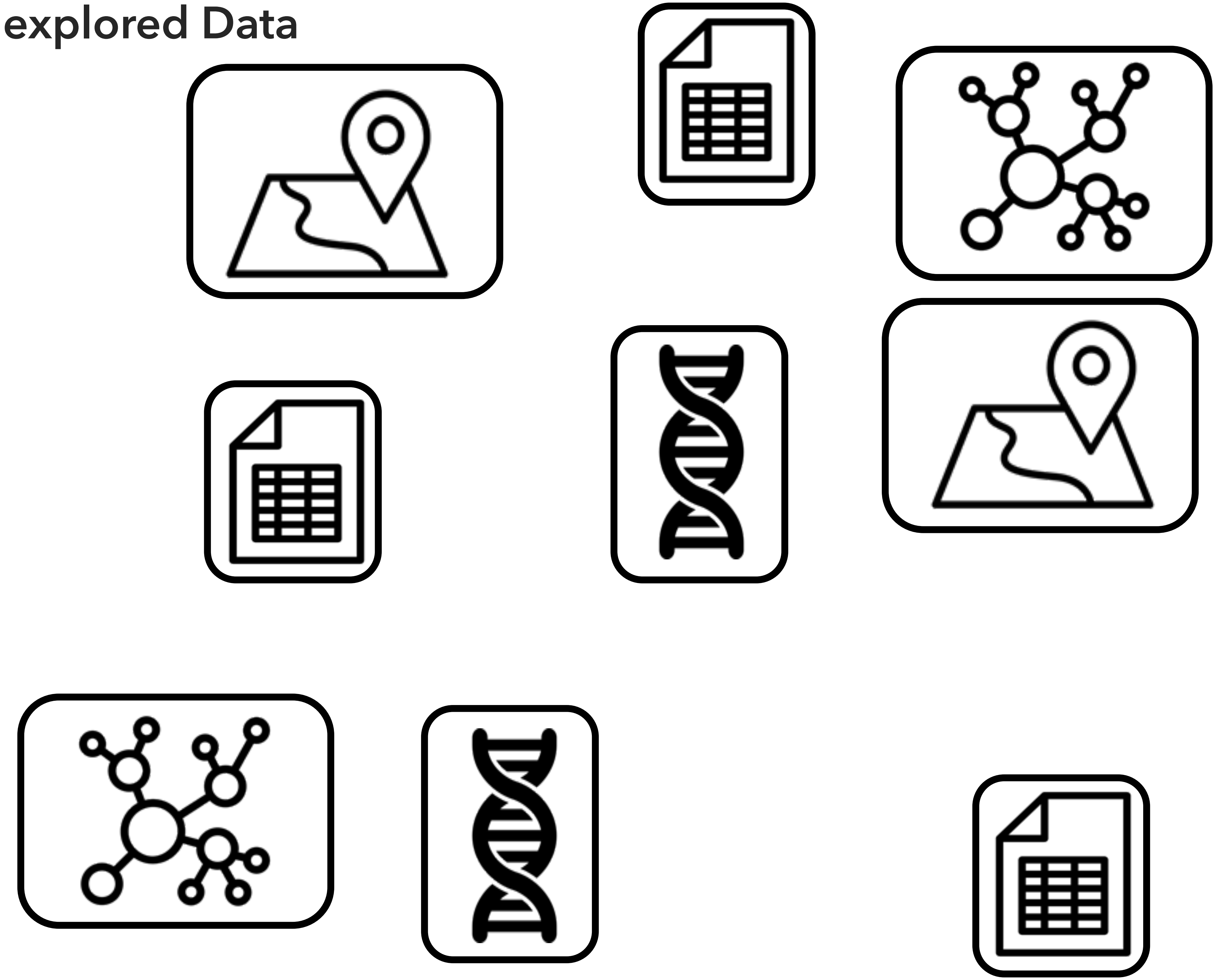
the very large space of existing heterogeneous and multidimensional datasets that are not yet understood by a specific person

Data landscape : collection of heterogeneous datasets

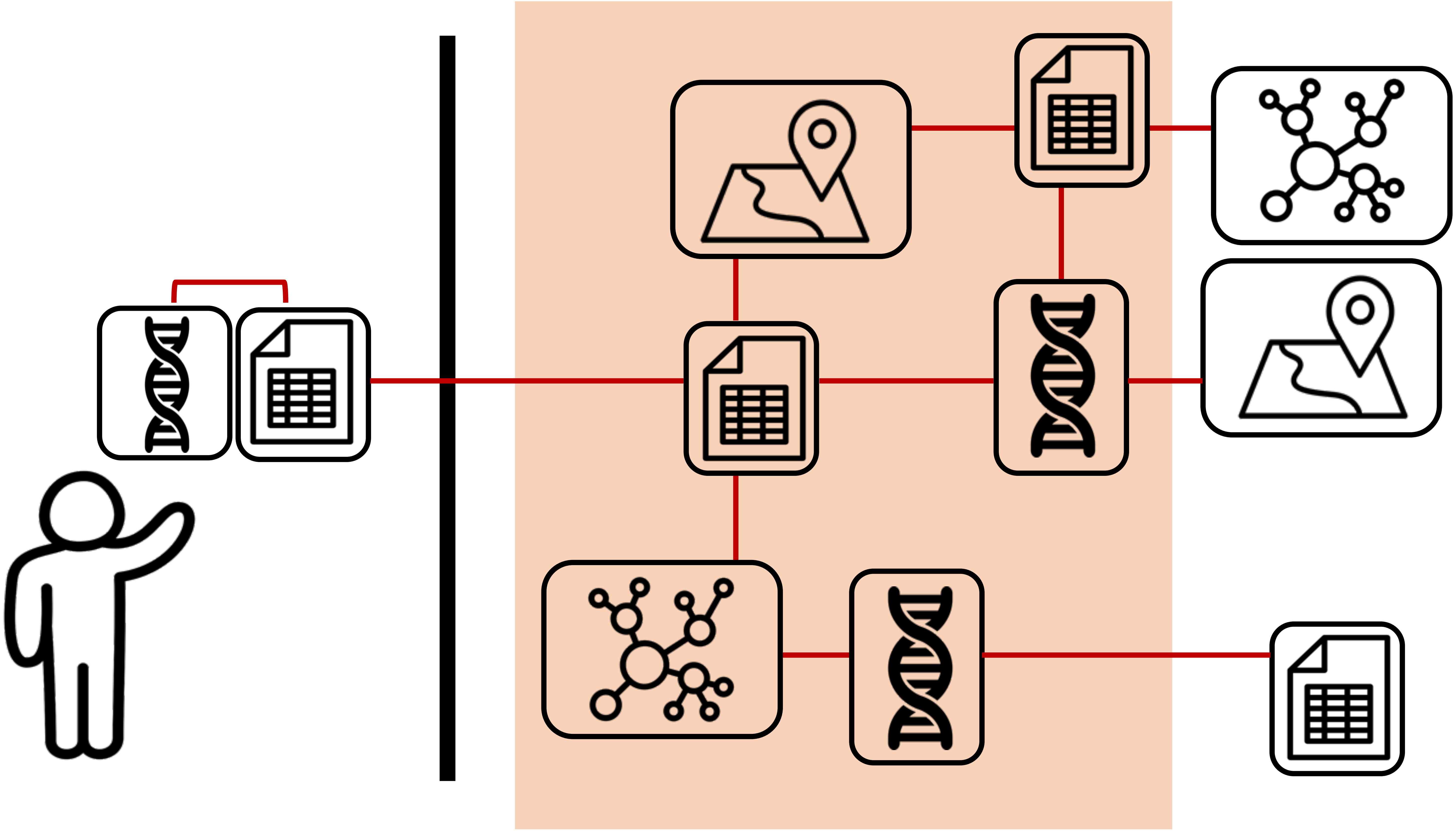
Domain Expert's
Currently Available Data



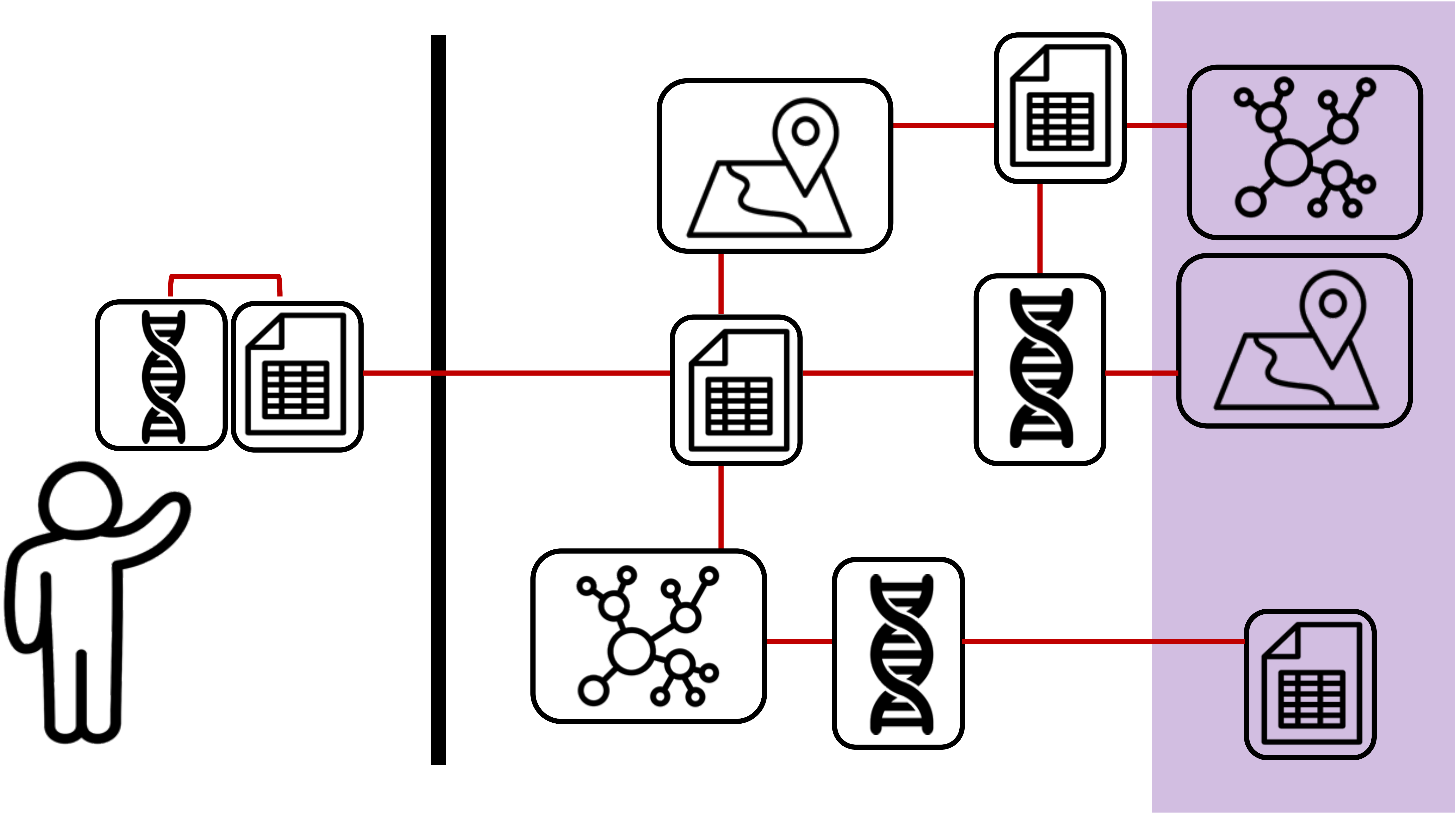
Unexplored Data



Experts may not have access to all of data



Experts may have not yet uncovered some data



New idea :

Operational definitions for
data reconnaissance and task wrangling

Two interrelated processes uncover data landscapes:

Data Reconnaissance

the process of uncovering an unfamiliar data landscape, including datasets that are known, available, unavailable, & unknown

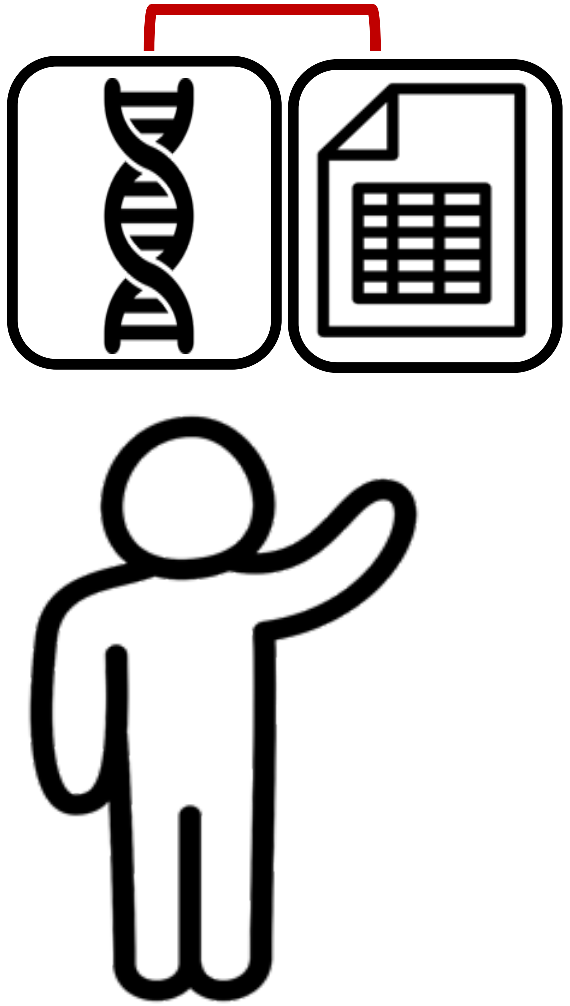
Task Wrangling

the process of progressively forming a crisper notion of tasks and assessing whether available and known datasets are suitable

Two interrelated processes uncover data landscapes:

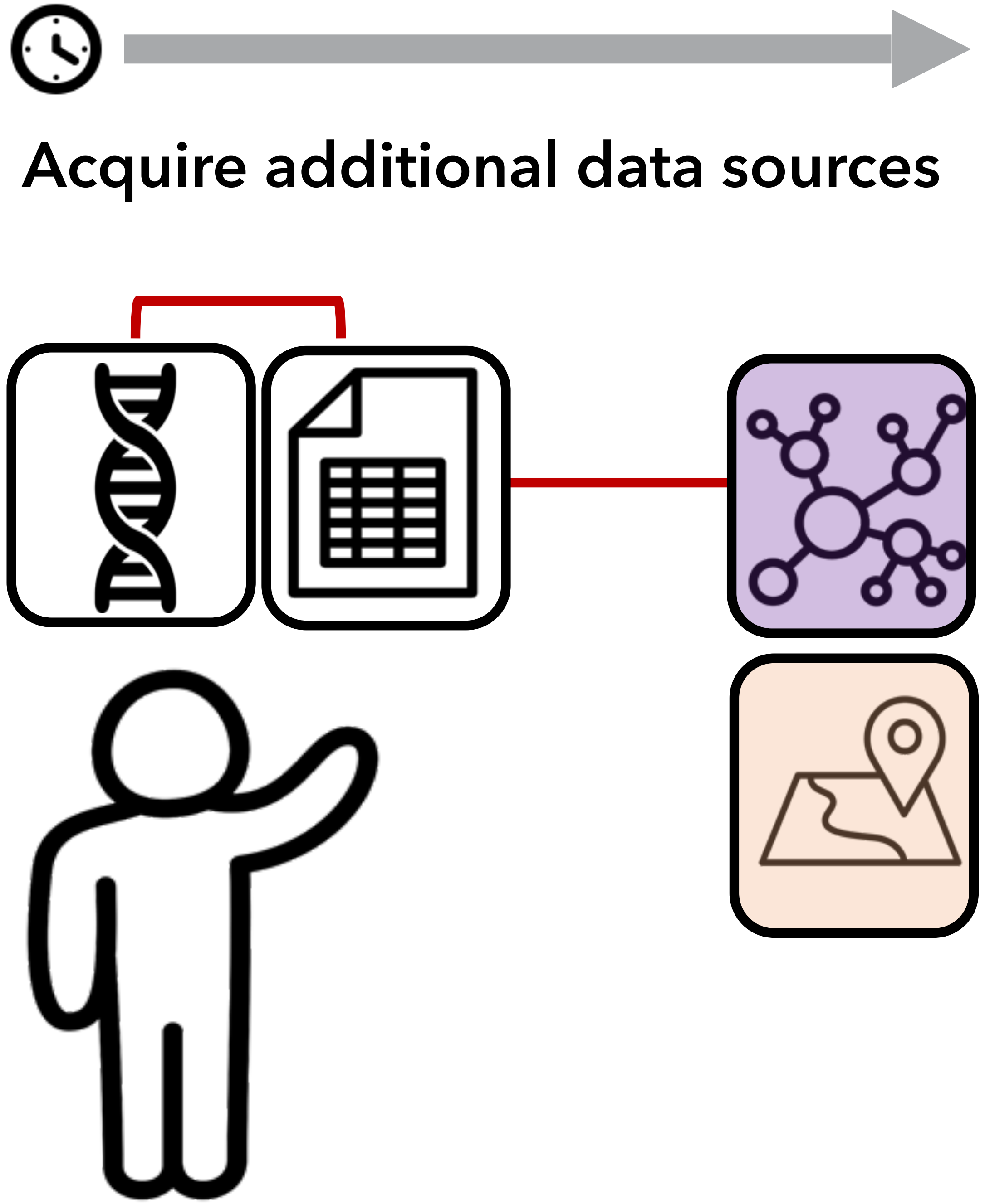
Data Reconnaissance

Some Data



Two interrelated processes uncover data landscapes:

Data Reconnaissance



Analysis & visualization of **available** data sources supports acquisition of **new** data:

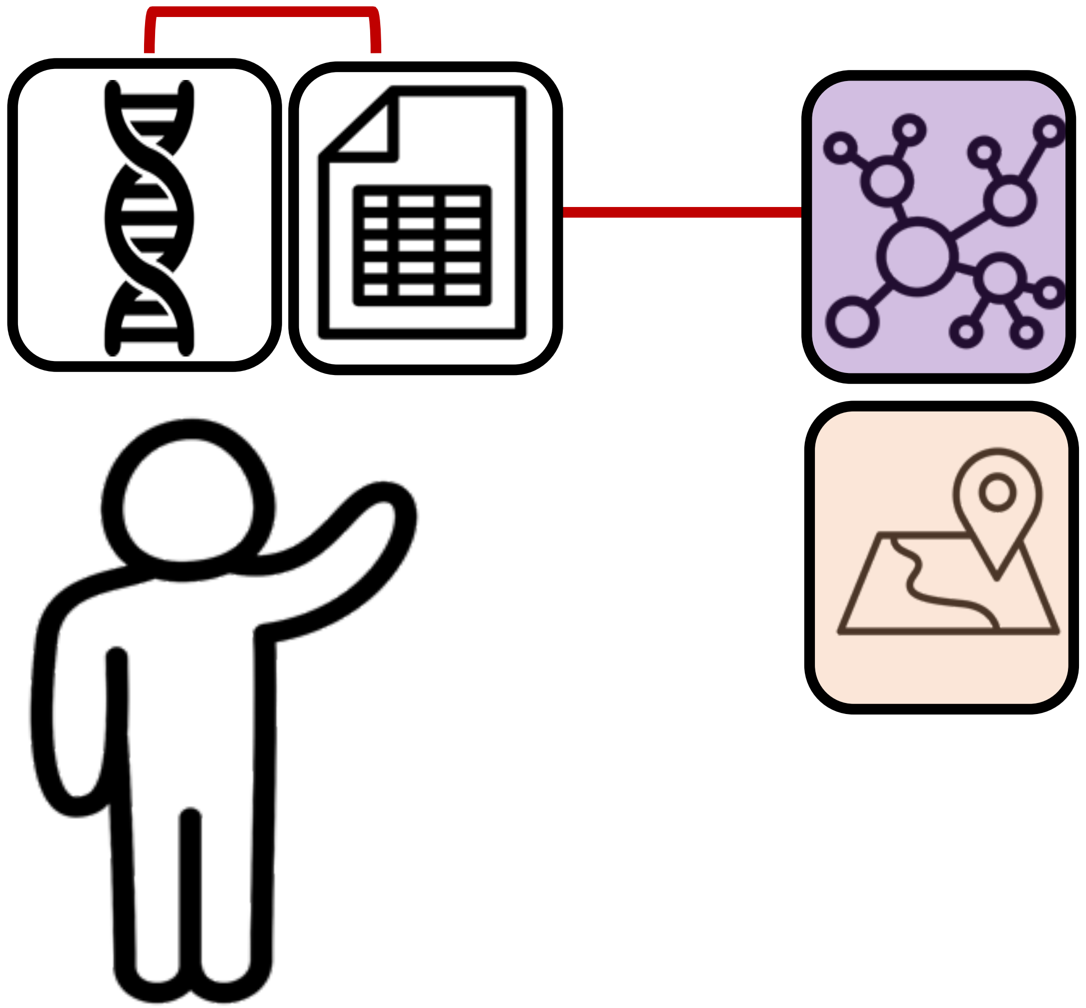
Acquire new dataset

Acquire available, but previously restricted, dataset

Two interrelated processes uncover data landscapes:

Data Reconnaissance

Acquire additional data sources



Analysis & visualization of **available** data sources supports acquisition of **new** data:

Acquire new dataset

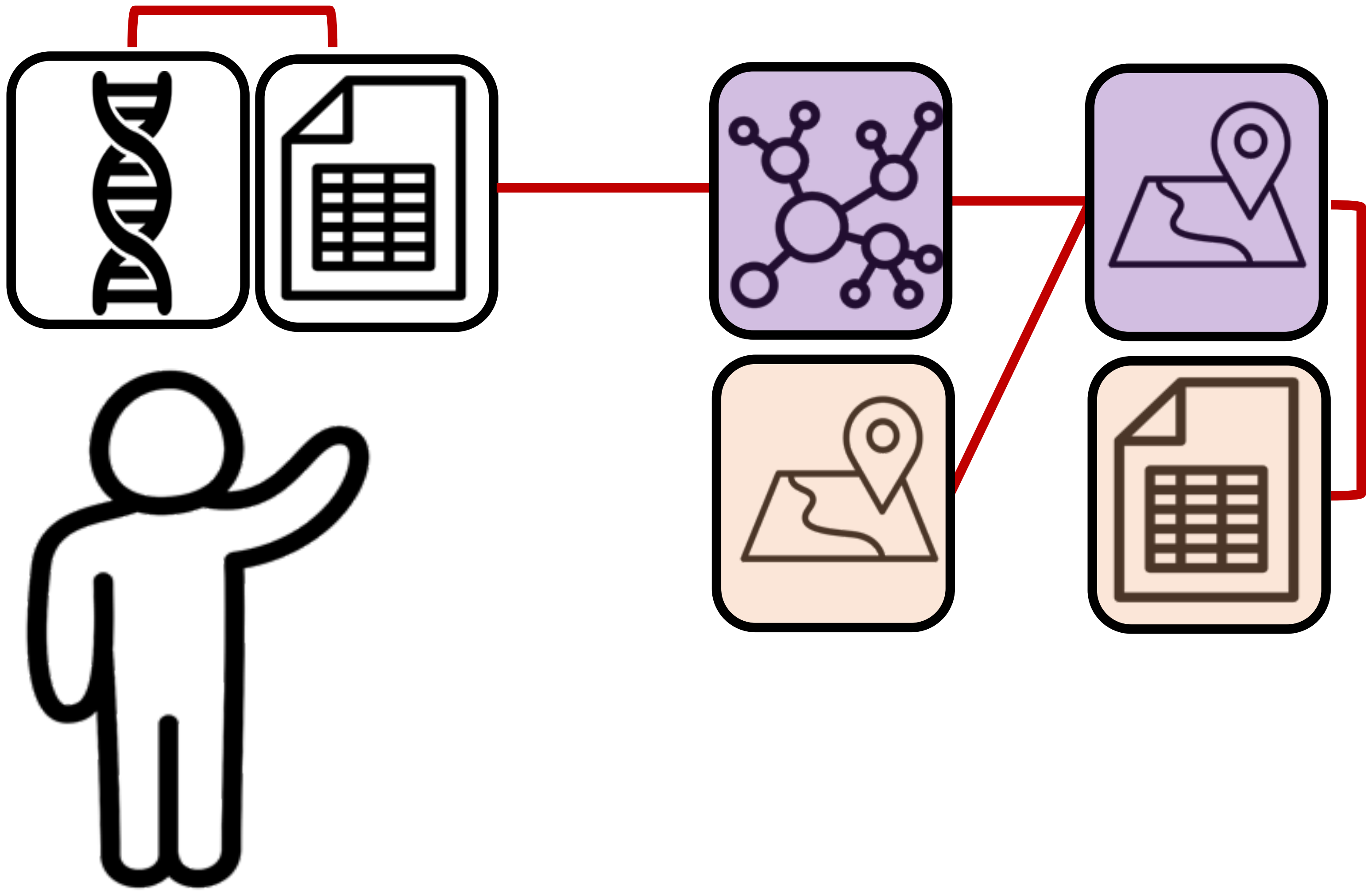
Acquire available, but previously restricted, dataset

Crisan & Munzner.
On Regulatory and Organizational Constraints in Visualization Design and Evaluation.
***Proc BELIV* 2016.**

Two interrelated processes uncover data landscapes:

Data Reconnaissance

Arrive at a finalized data set



Finalized dataset can be analyzed & visualized in depth

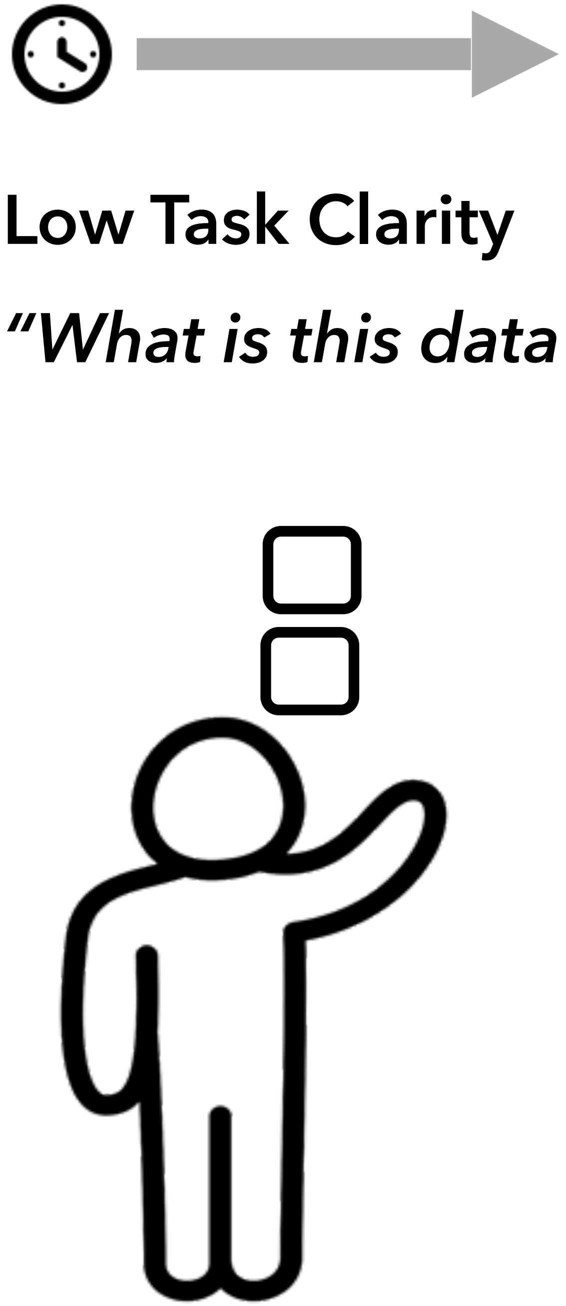
Two interrelated processes uncover data landscapes:

Data Reconnaissance

Task Wrangling

Low Task Clarity

"What is this data?"



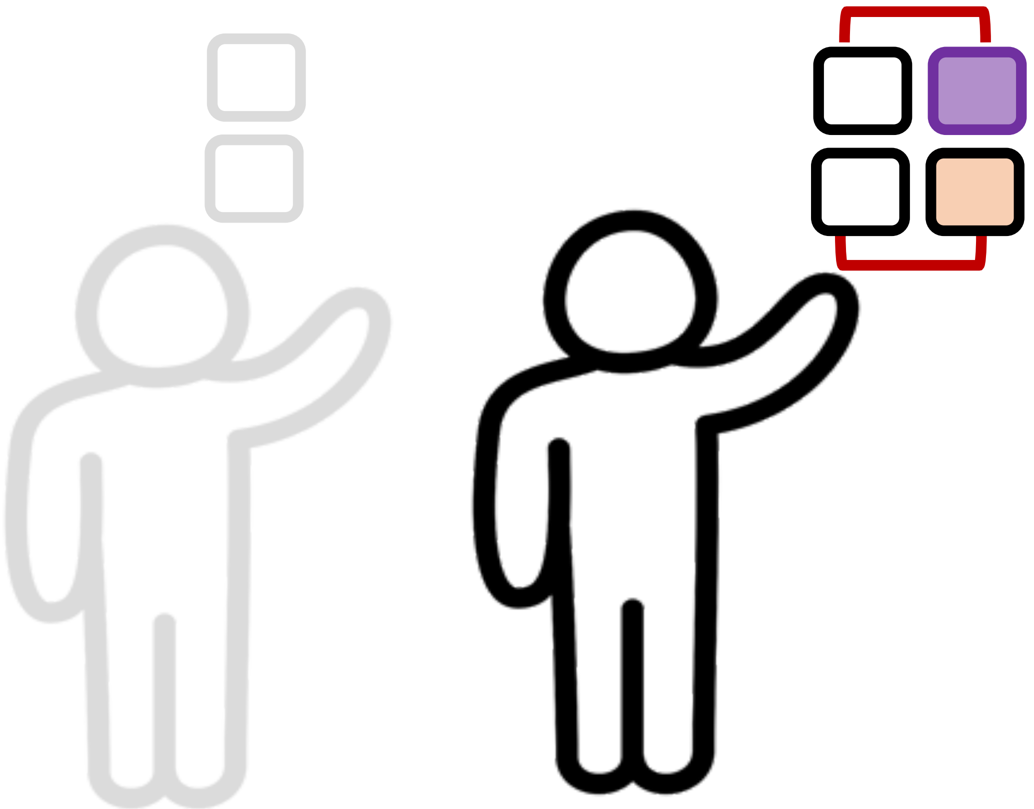
Two interrelated processes uncover data landscapes:

Data Reconnaissance

Task Wrangling

Evolving Task Clarity

" I think might want to see geographic patterns"



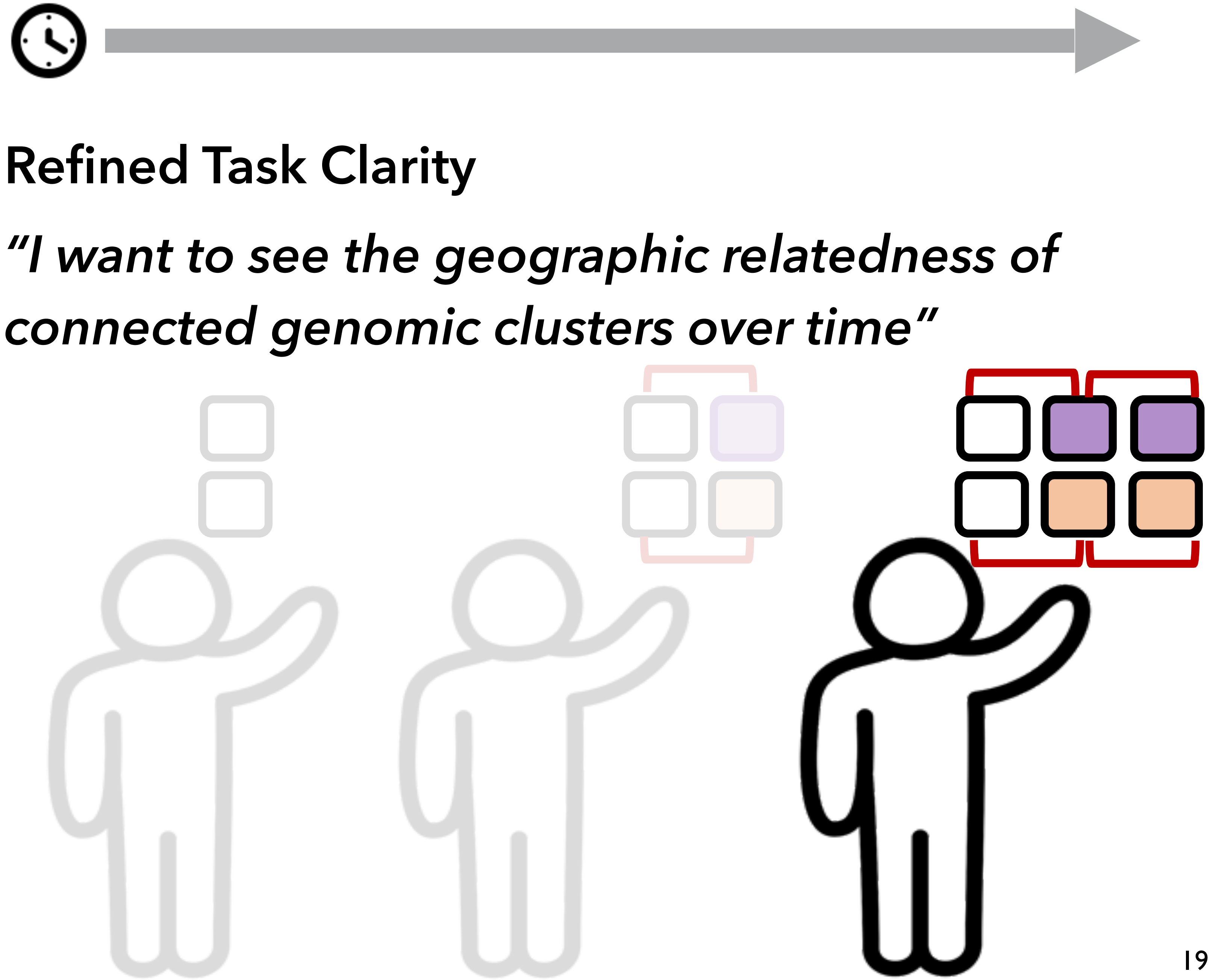
Two interrelated processes uncover data landscapes:

Data Reconnaissance

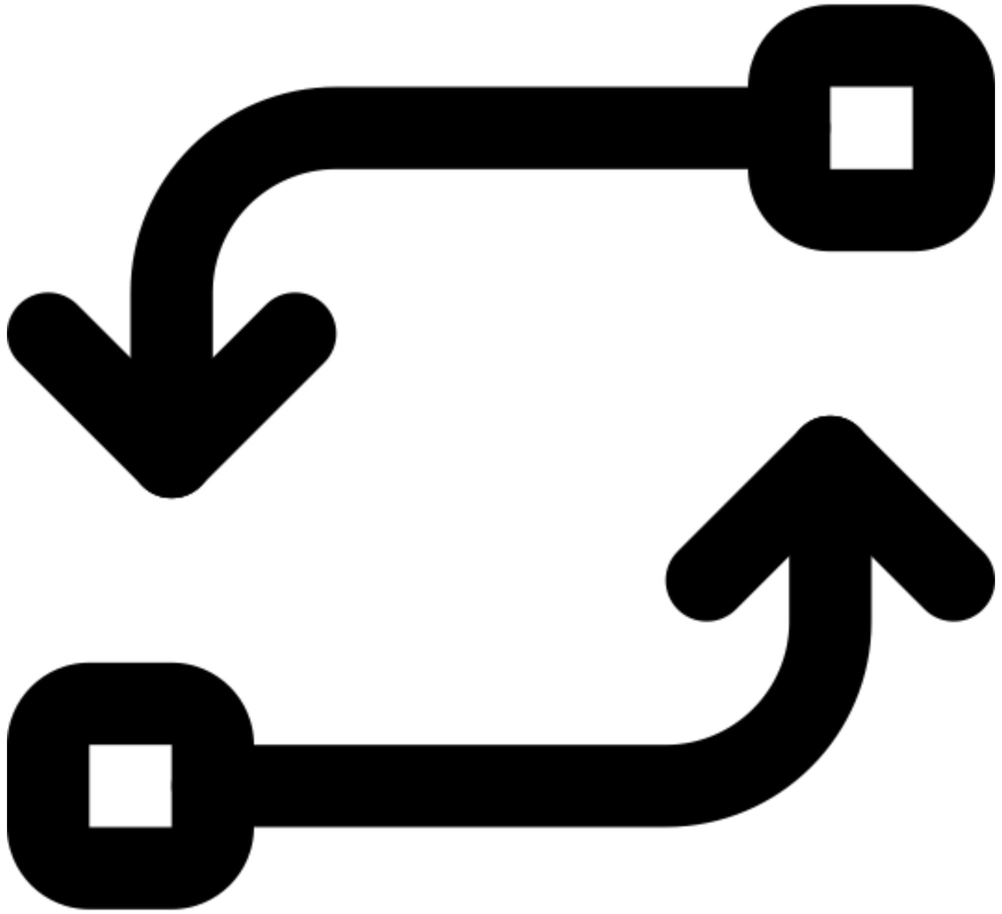
Task Wrangling

Refined Task Clarity

"I want to see the geographic relatedness of connected genomic clusters over time"



Processes influence each other over time



Data Reconnaissance

Data access inspires & refines tasks

Task Wrangling

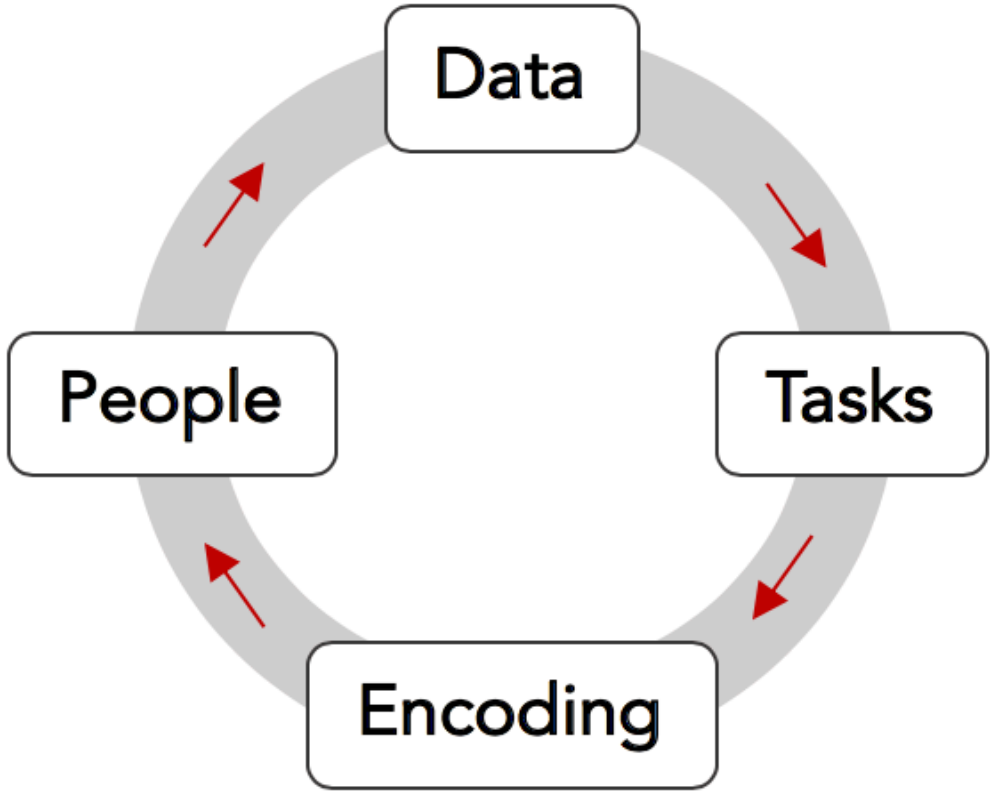
Refined tasks guide the pursuit of data

New idea :

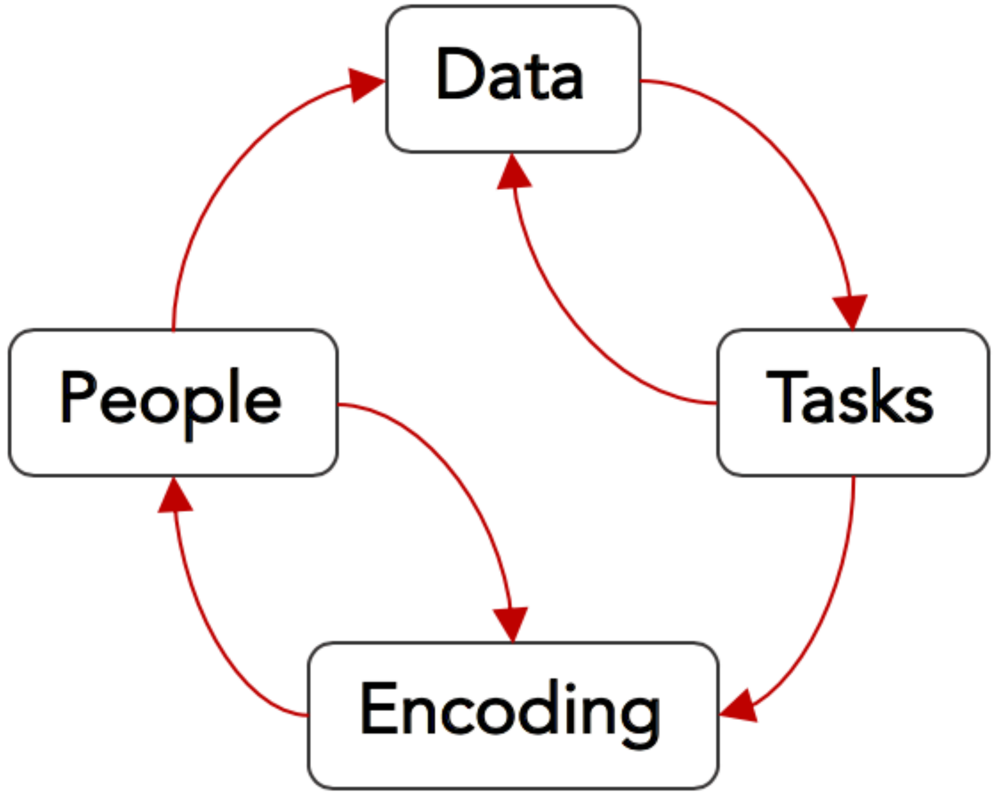
A conceptual framework for data reconnaissance and task wrangling

Existing methods can be slow

Human Centered Design

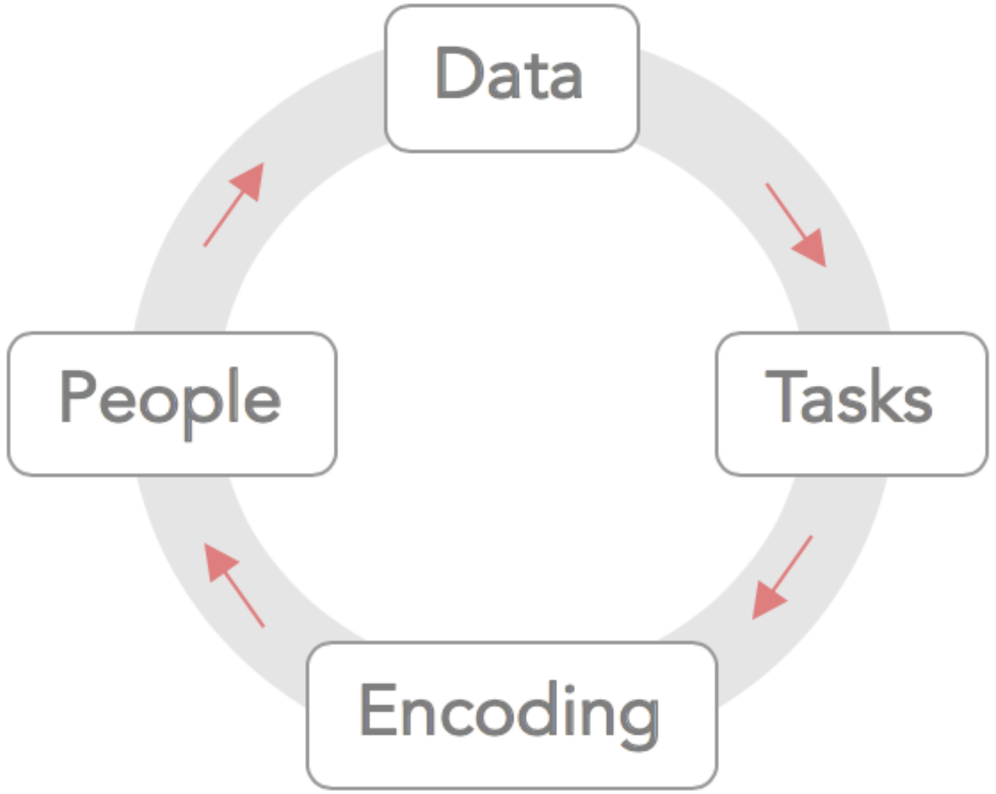


Design Study Methodology

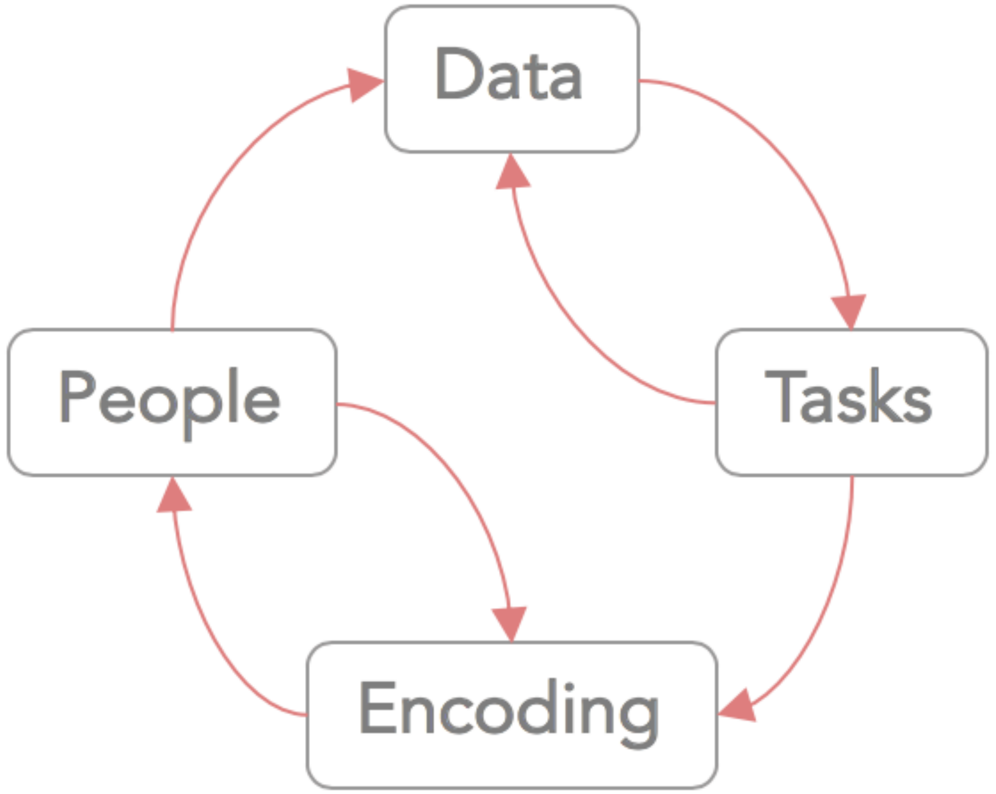


Uncover data & tasks faster with shortcuts

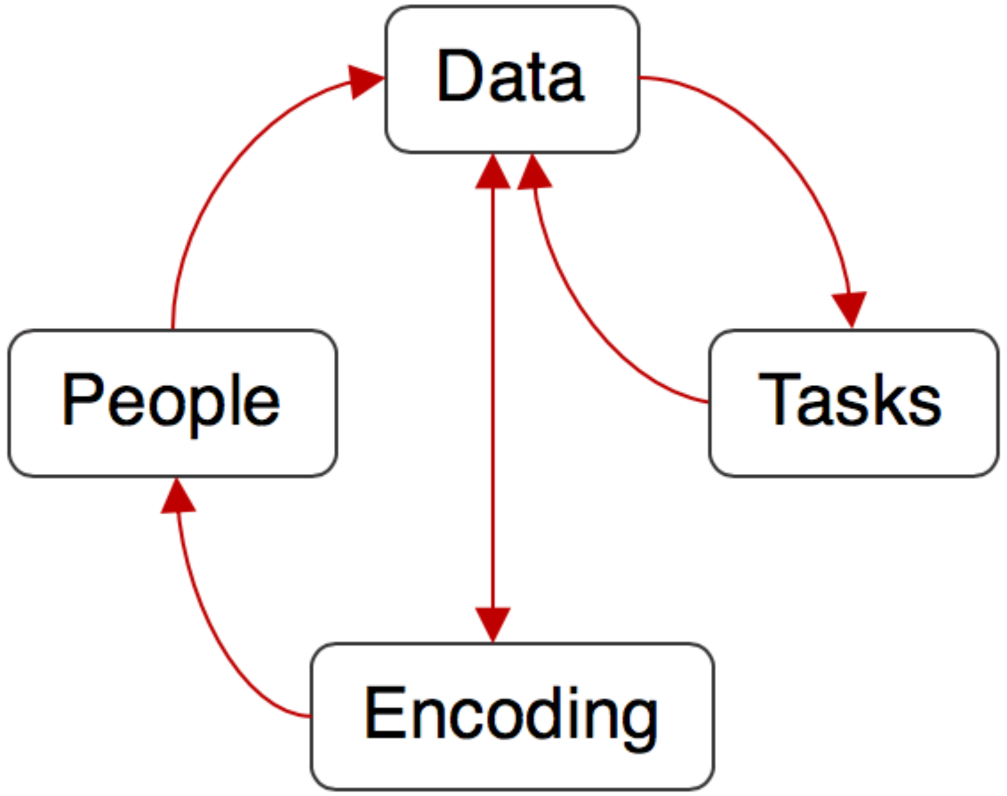
Human Centered Design



Design Study Methodology

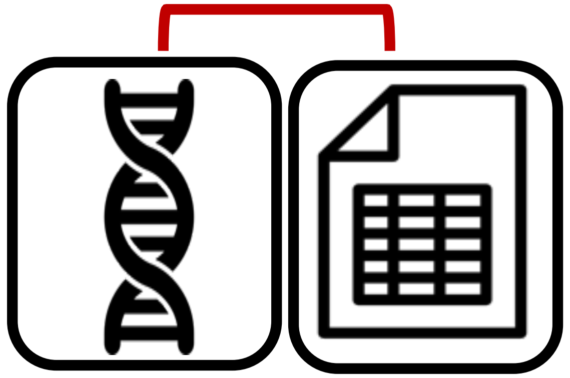


Our speedup

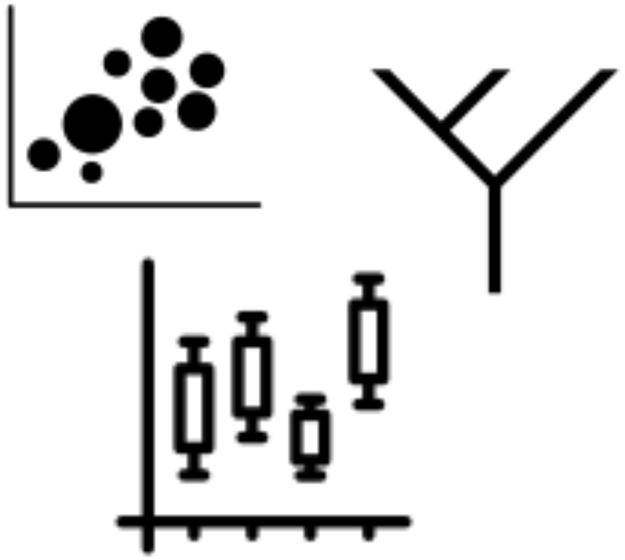


Steps in our conceptual framework

Acquire

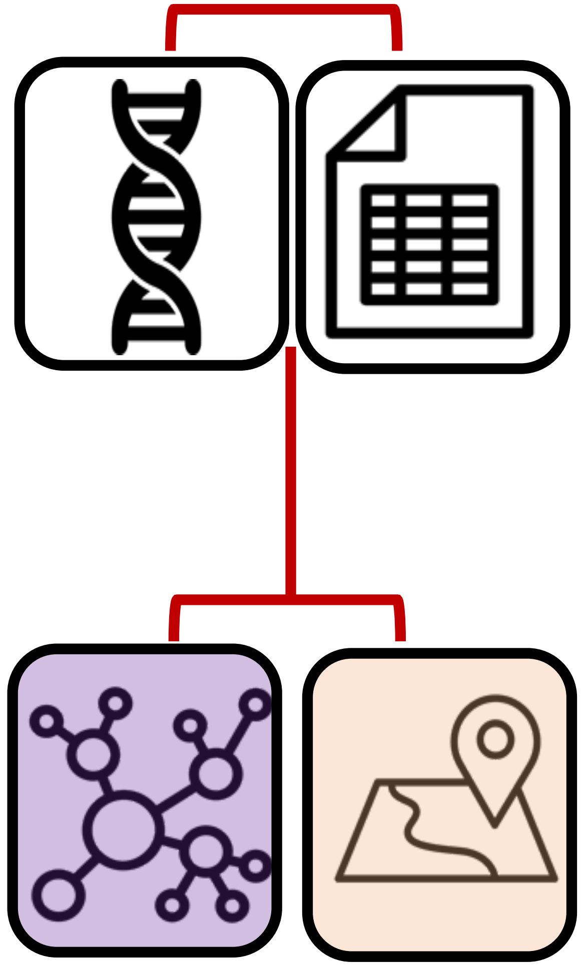


View

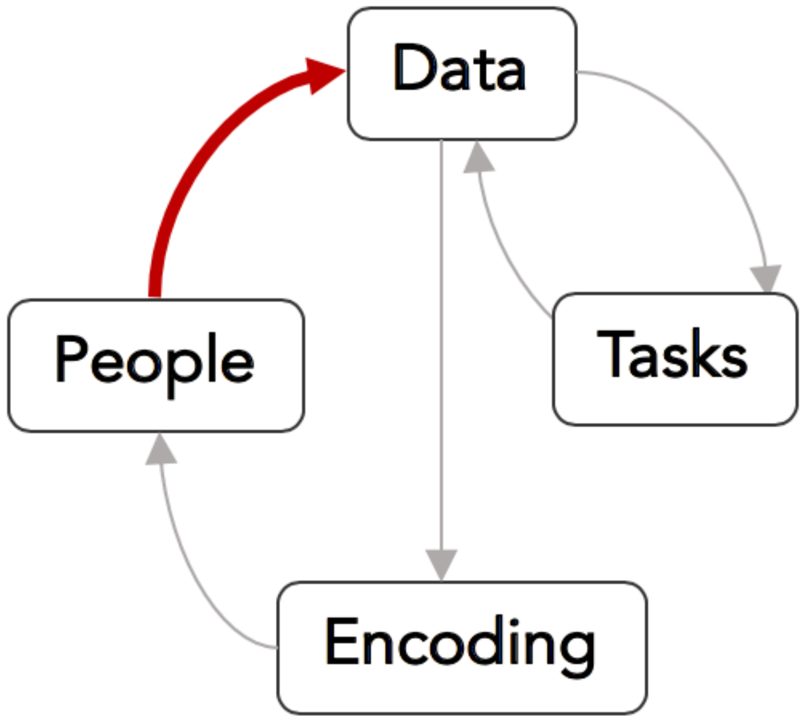


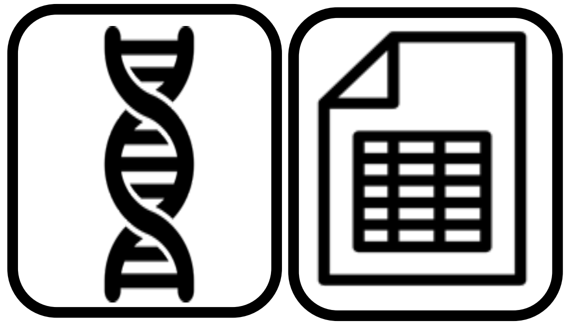
Assess

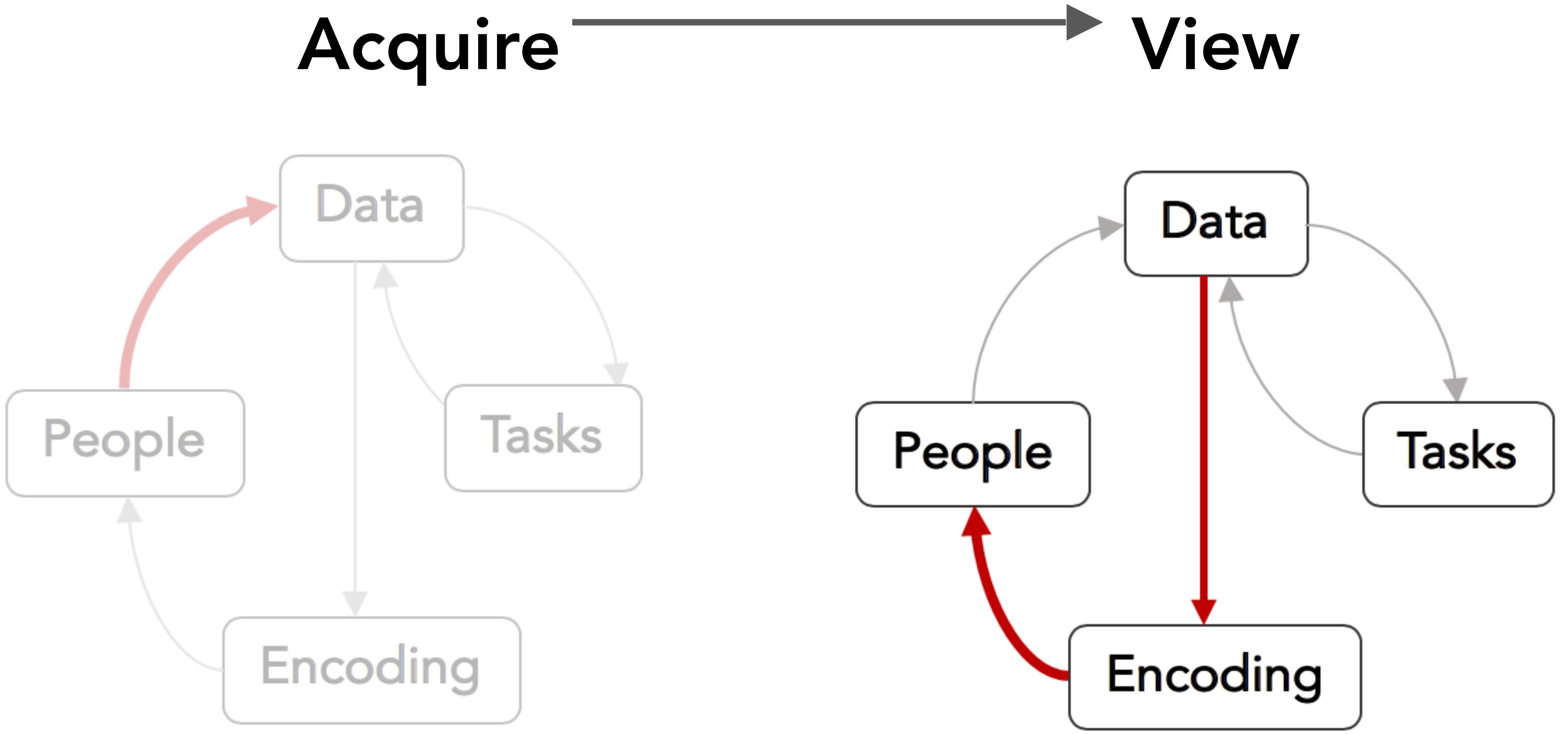
Pursue



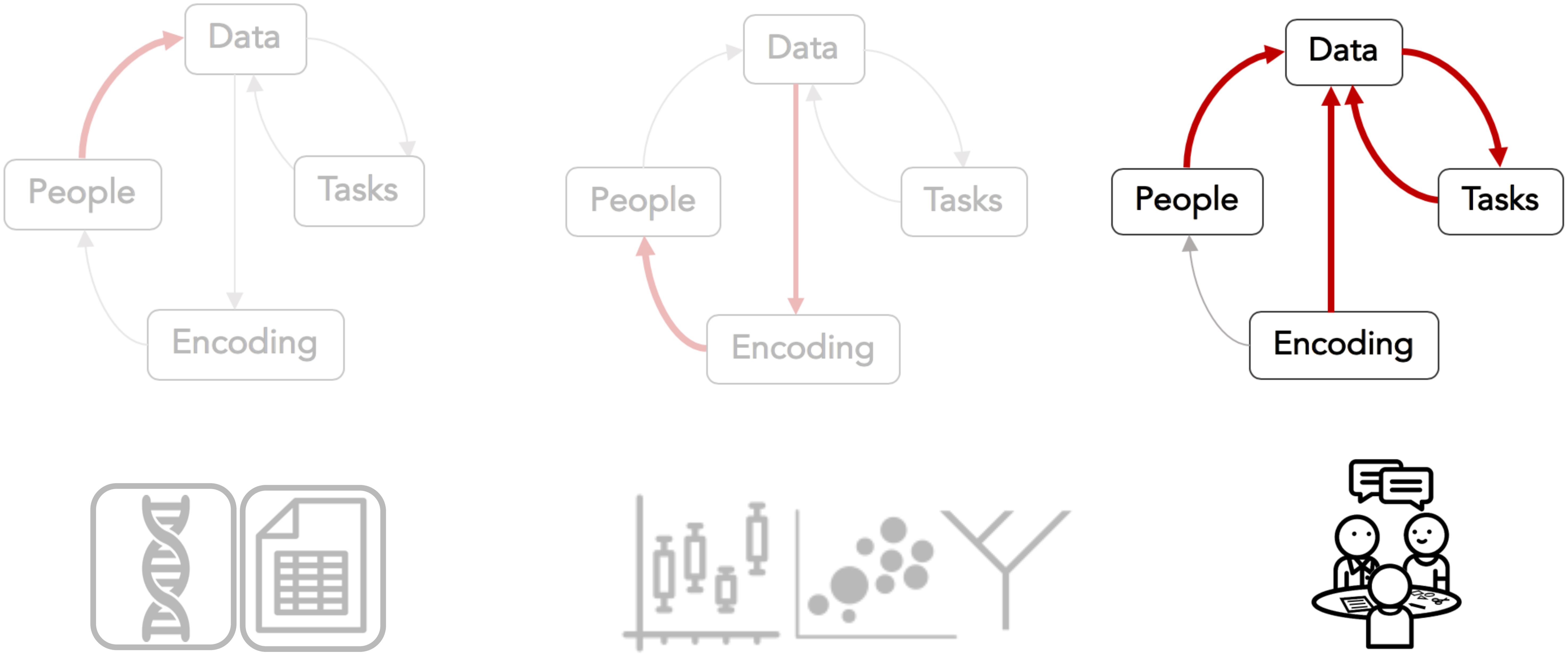
Acquire

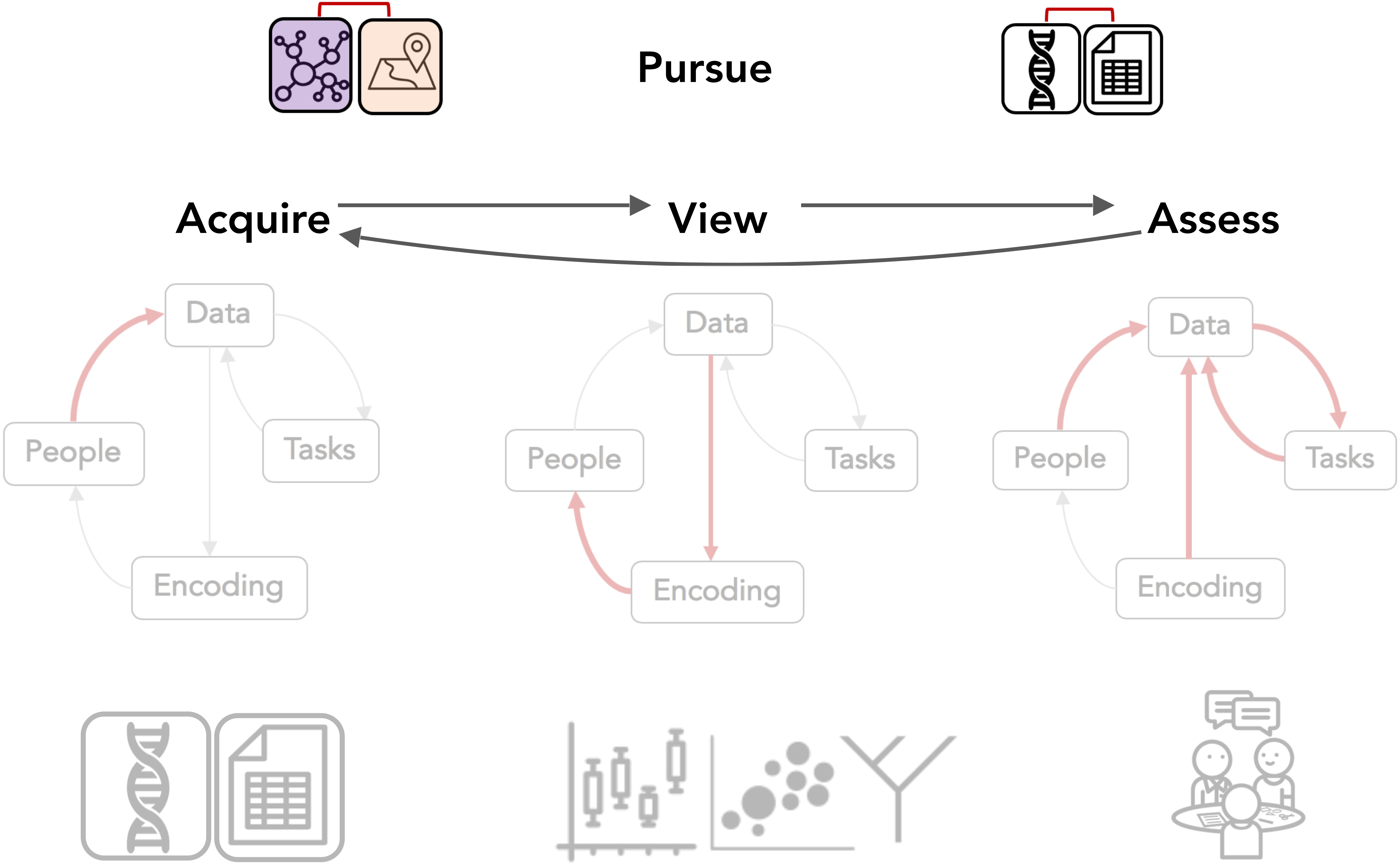




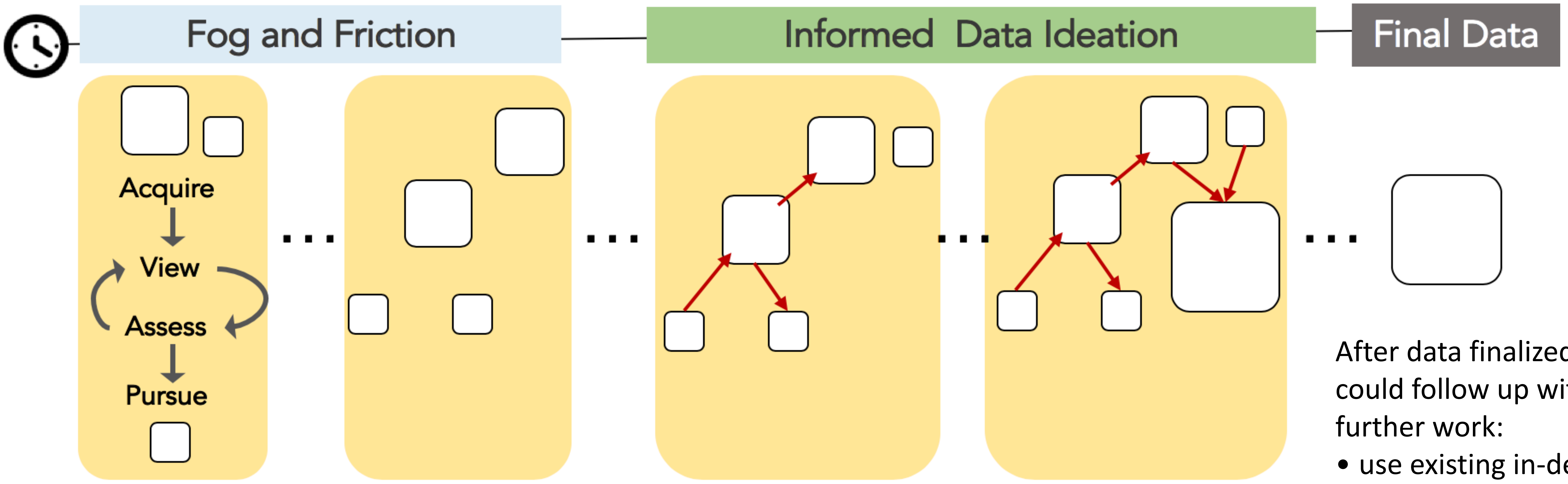

Acquire
View
Data
People
Tasks
Encoding
Data
People
Tasks
Encoding

Acquire → View → Assess




Pursue
Acquire
View
Assess
Data
People
Tasks
Encoding
Data
People
Tasks
Encoding
Data
People
Tasks
Encoding

From unknown landscape to the final dataset



After data finalized, could follow up with further work:

- use existing in-depth EVA tool
- create bespoke system through design study

Where do we go from here?

Building systems suitable for
data reconnaissance and task wrangling

Questions in road trips - and visualization in data science!

- where are we?
  - Uncovering Data Landscapes through Data Reconnaissance & Task Wrangling
- what's here?
  - Automatic Encodings through Recommendation

http://www.cs.ubc.ca/~tmm/talks.html#vds23

GEViTRec:

Data Reconnaissance Through Recommendation Using a Domain-Specific Visualization Prevalence Design Space

Anamaria Crisan
@amcrisan
UBC/Tableau

Shannah Fisher
UBC/USask

Jenn Gardy
@jennifergardy
UBC/BCCDC/
Gates Foundation

Tamara Munzner
@tamaramunzner
@tamara@vis.social
UBC

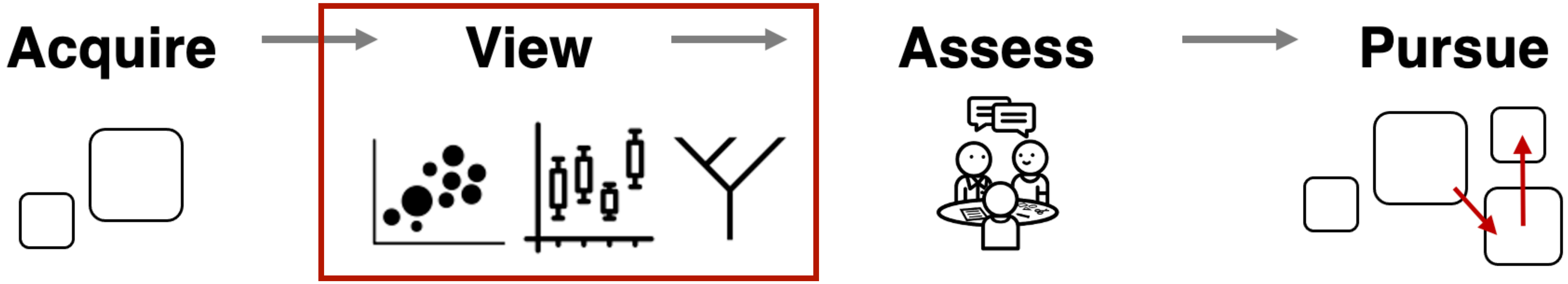
https://www.cs.ubc.ca/group/infovis/pubs/2021/gevitrec/

GEViTRec: Data Reconnaissance Through Recommendation Using a Domain-Specific Visualization Prevalence Design Space.
Crisan, Fisher, Gardy, Munzner. IEEE TVCG 28(12):4855-4872, 2022.

Data Reconnaissance

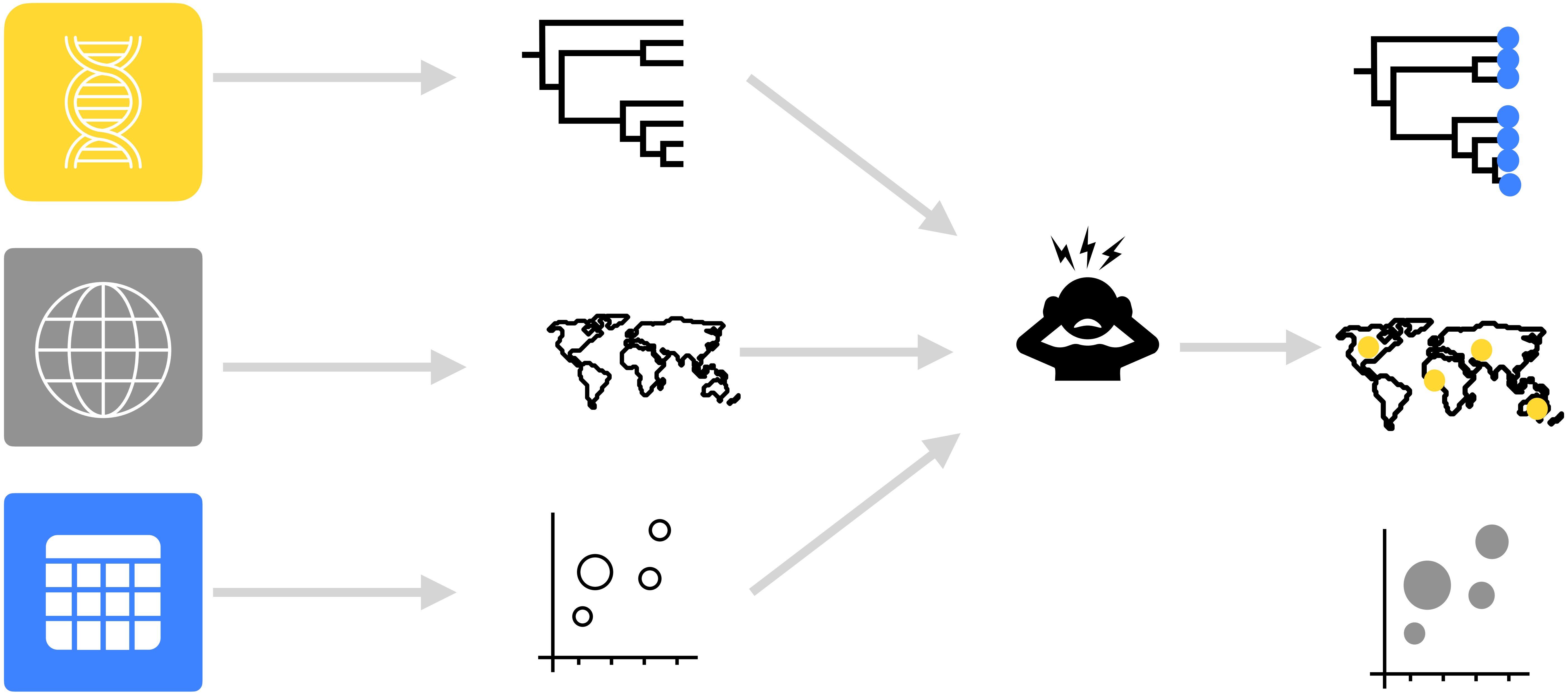
the process of exploring an unfamiliar data landscape; the very large space of existing heterogeneous and multidimensional datasets that are not yet understood by a specific person

biggest need:
accelerate this part

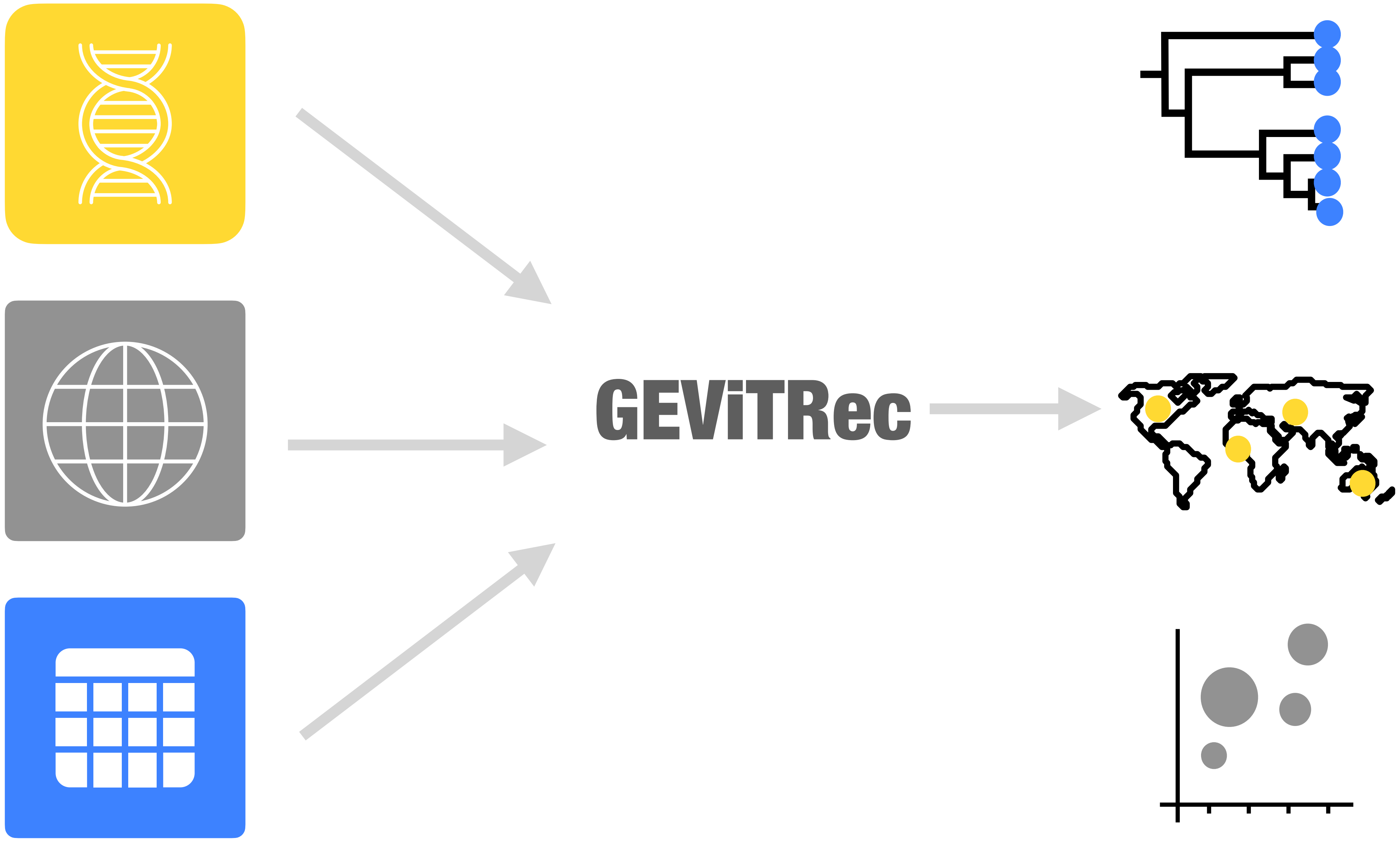


what's here?

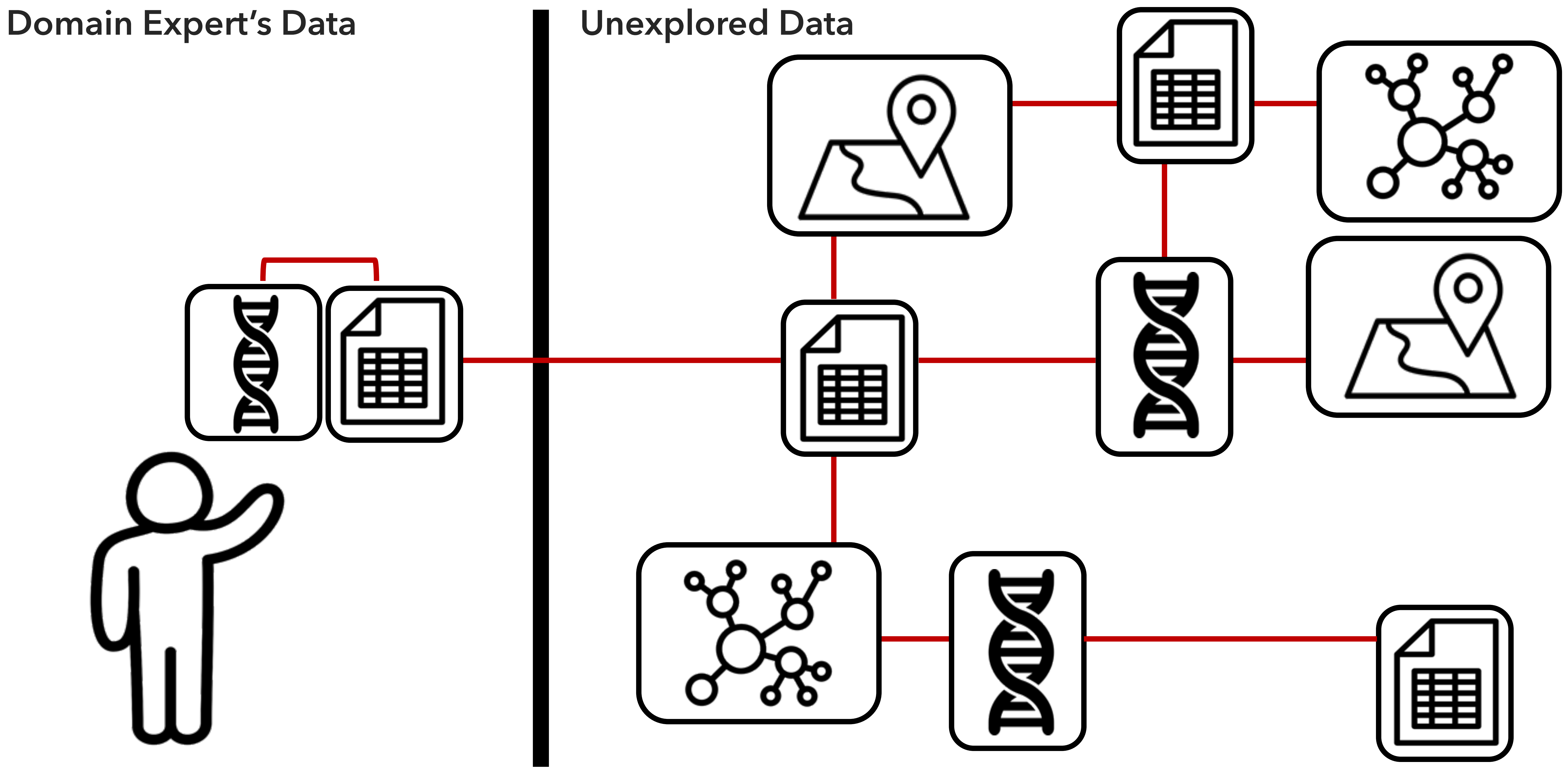
Manually Constructing Chart Combinations



Automatically Constructing Visually Coherent Chart Combinations

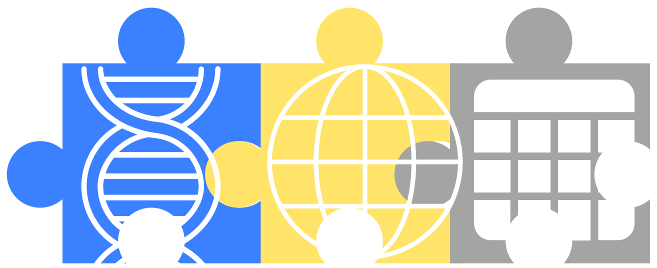


How to connect datasets? Identify shared attributes!



How to show connections for data recon?

Visually Coherent Chart Combinations



that prioritize visual coordination of **shared information** between charts with respect to **layout and consistency among visual channels** (position, color)

Static charts avoid interactive view coordination complexities and costs

Fast to view
Easy to disseminate

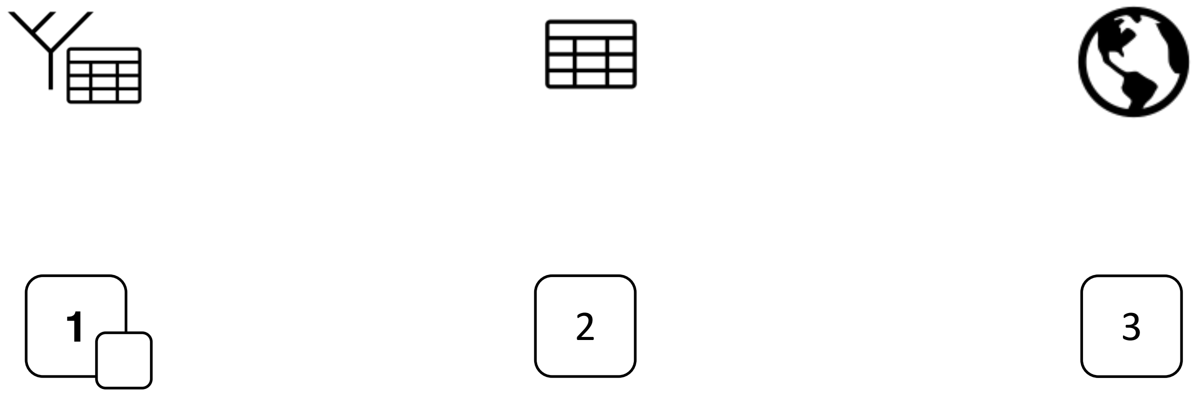
New Idea: Visually Coherent Chart Combinations Through Gradual Binding

- Automatically coordinating static charts is not trivial
  - Cannot change encoding after chart rendered into box of pixels!
- **Declarative approach of gradual binding**
  - Initially generate partial specification using template
  - Modify specification in discrete stages, to enforce consistency of channels (color, position) according to desired combination
  - Pass final specification to rendering library
  - Simply concatenate resulting boxes of pixels to display

GEViTRec algorithm: Overview

Data Type

Data Source



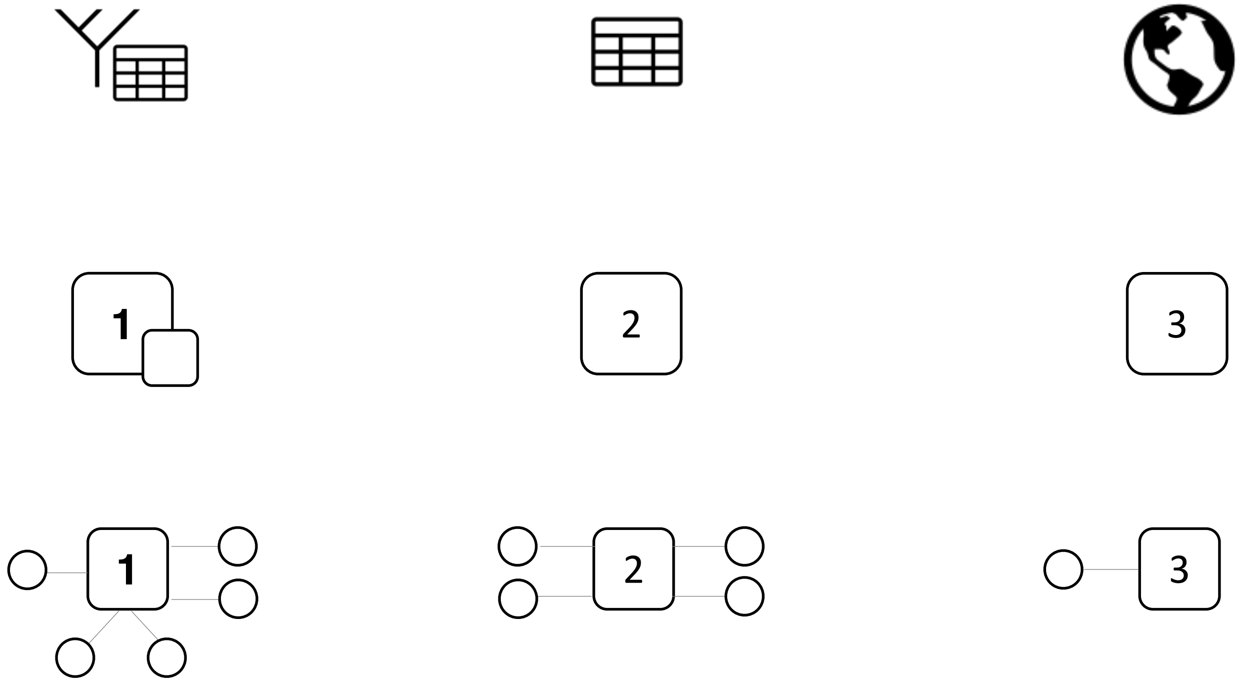
- **Example: three analysis datasets**
  - 1 : Tree data w/ associated tabular data
  - 2 : Tabular Data
  - 3: Spatial Data

GEViTRec algorithm: Overview

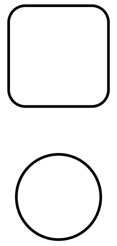
Data Type

Data Source

Exploded Fields



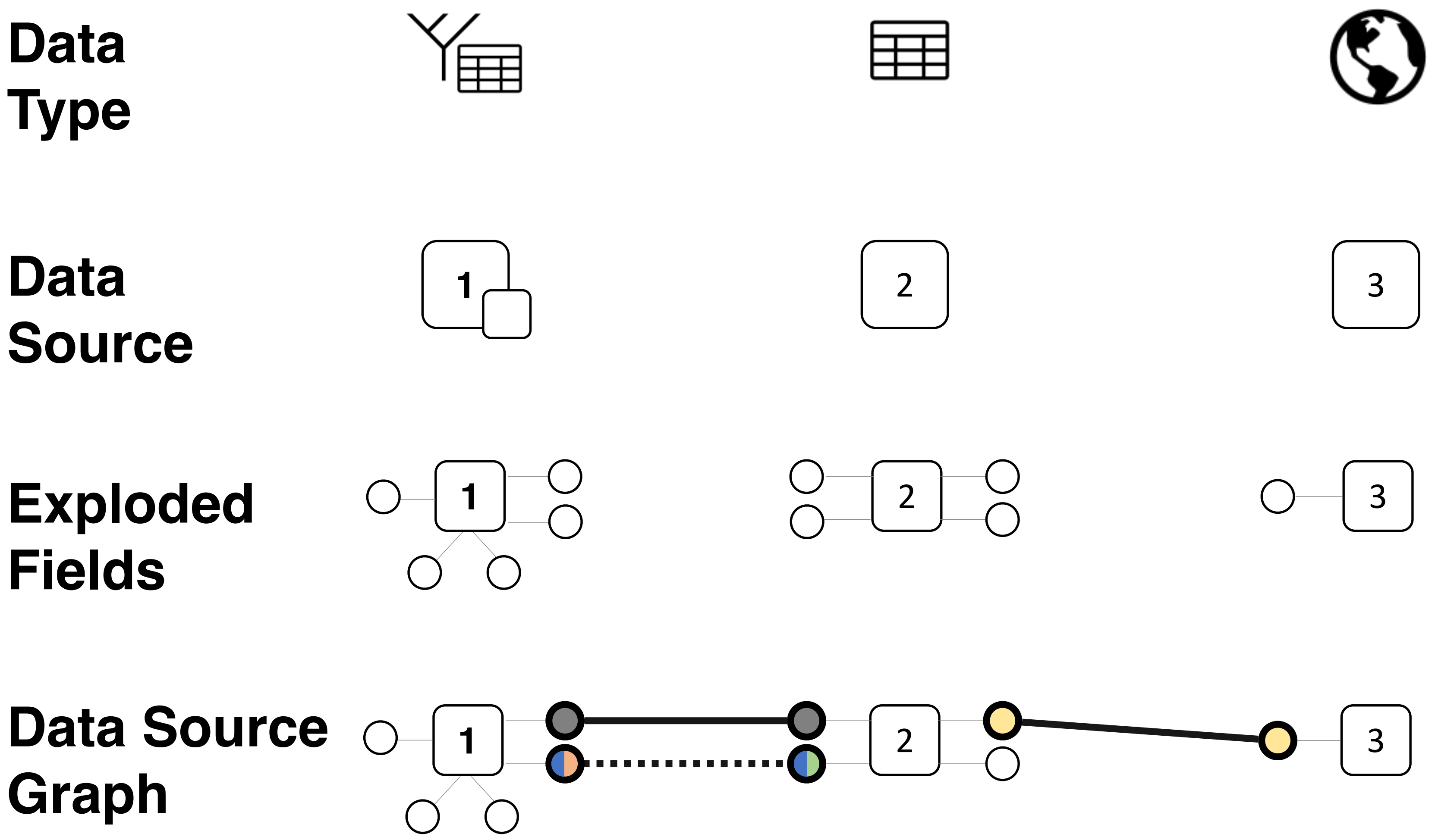
- **‘Explode’ attribute fields extracted from data sources**



= data source

= attribute field

GEViTRec algorithm: Overview

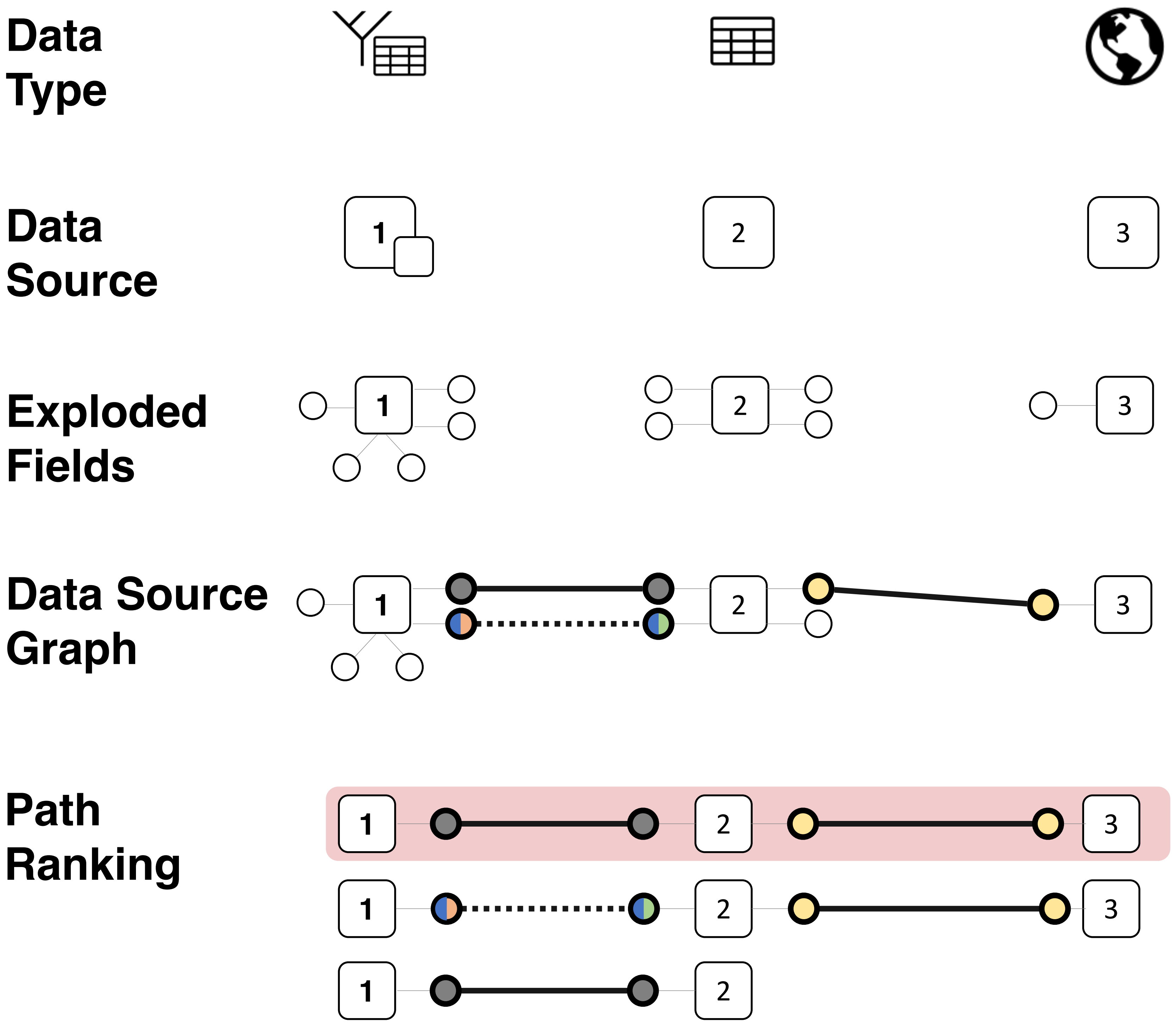


- **Analyze to identify commonalities and create dataset linkages through categorical attribute fields**

Fields A	Fields B	Jaccard Index	Linkage Type
		1	Exact
		$0<j<1$	Partial
		0	None

- **Input data is now modelled as a graph!**

GEViTRec algorithm: Overview

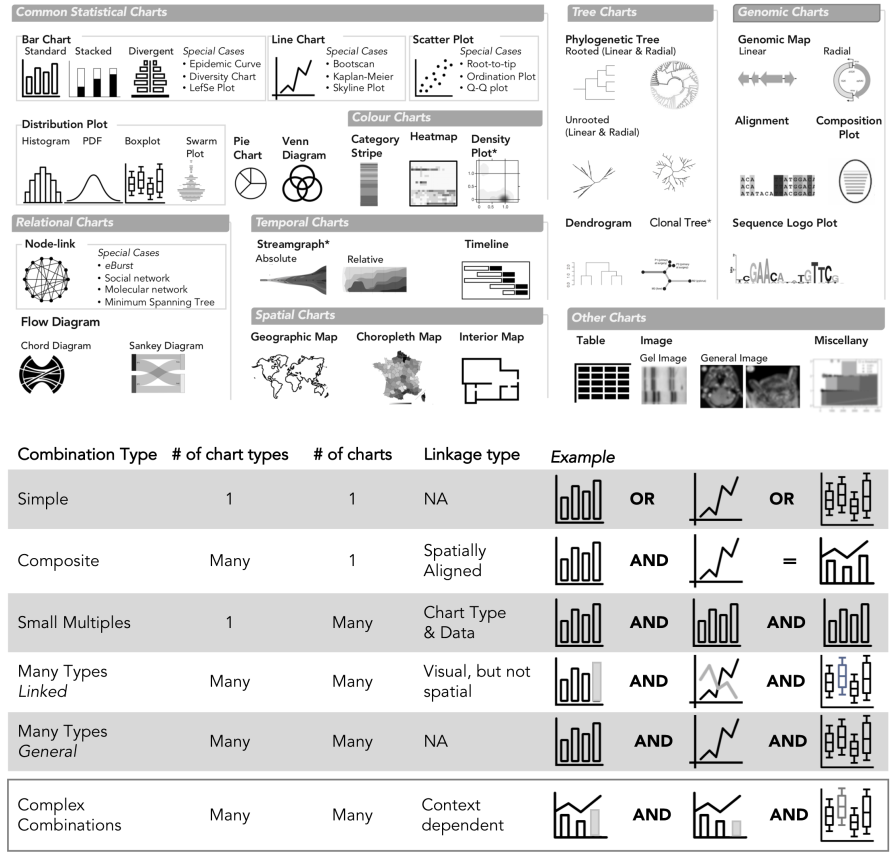


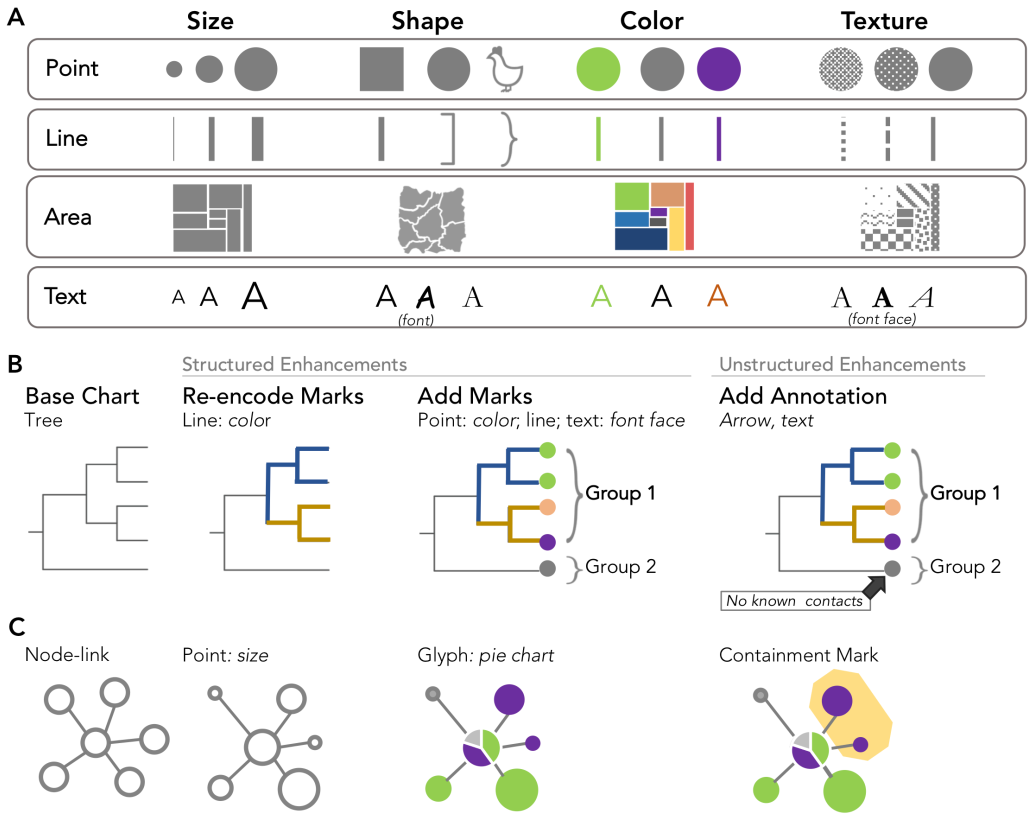
- **Traverse graph: enumerate & rank paths** linking all pairs of data, using three metrics
  - **Strength** of linkages
  - **Diversity** of data types
  - **Relevance** to domain
    - **New idea:** using **domain prevalence design space** in visualization recommendation

Domain Prevalence Design Space:

Captures full scope of visual encodings used by defineable set of experts, includes quantitative estimate for prevalence of each strategy within that domain

Domain-level answer to question of what's here?

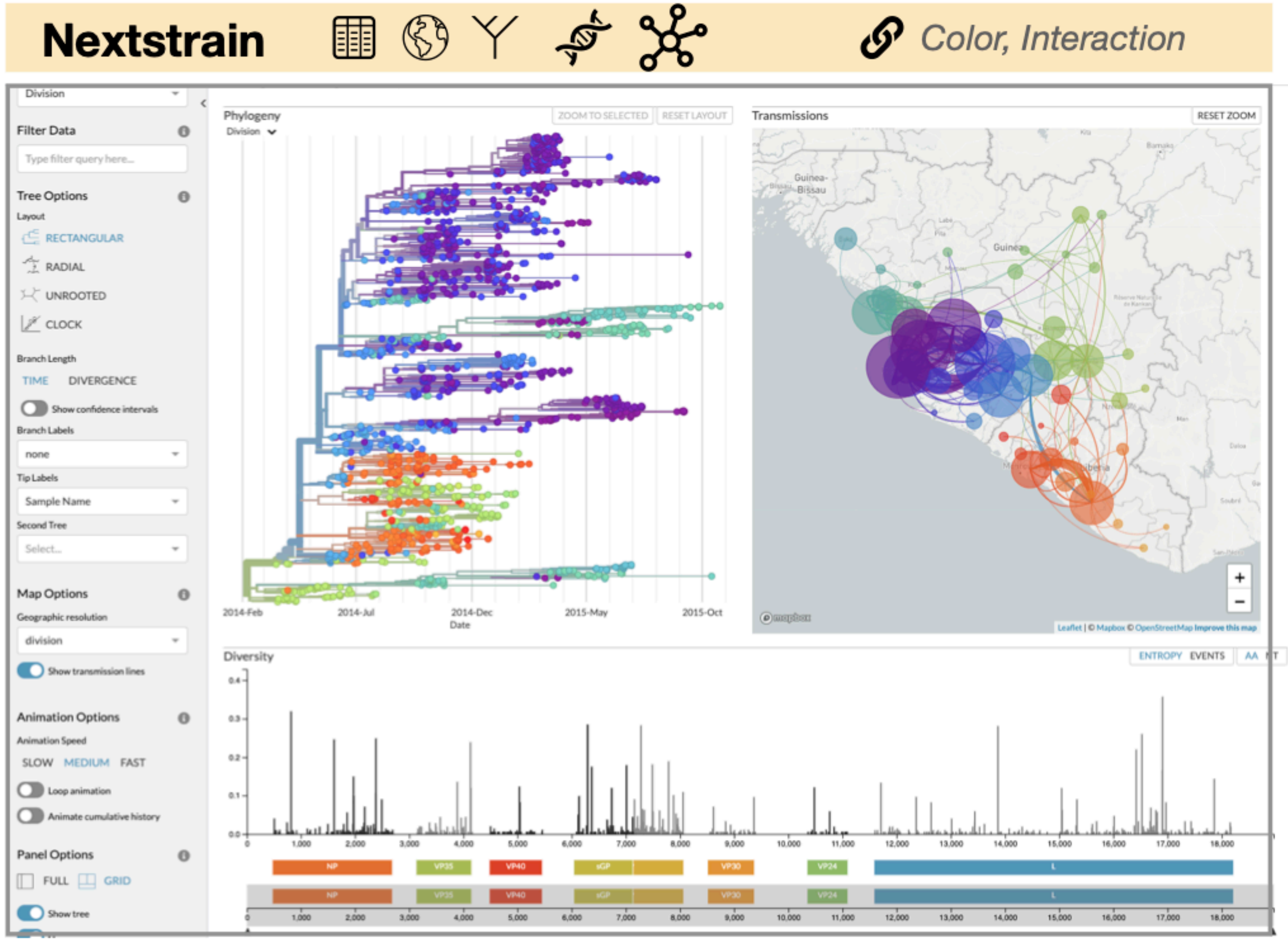




A Crisan, JL Gardy, T Munzner.
A systematic method for surveying data visualizations and a resulting genomic epidemiology visualization typology: GEViT.
Bioinformatics 35(10):1668–1676, 2019.

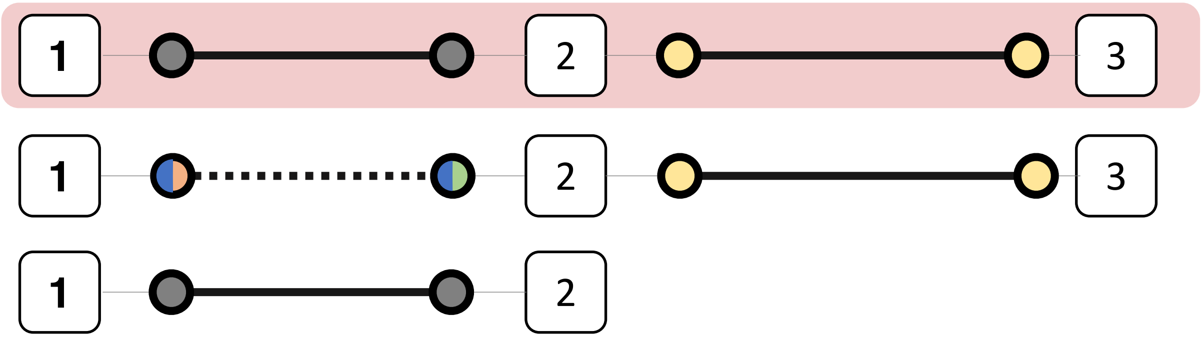
https://doi.org/10.1093/bioinformatics/bty832

Domain Context:
Genomic Epidemiology



GEViTRec algorithm: Overview

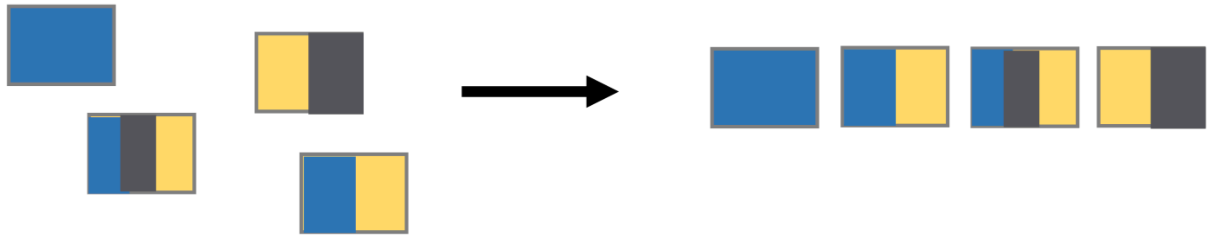
Path Ranking



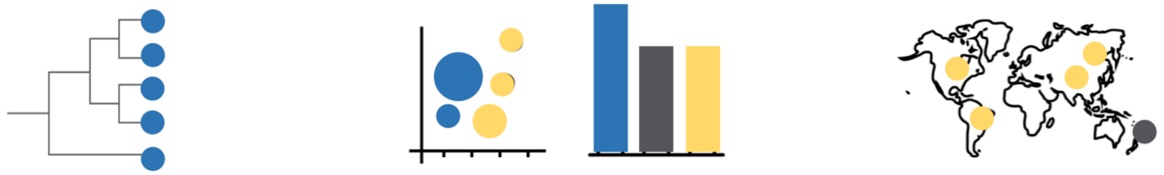
Singleton Specification

```
chart_spec("phylogenetic tree",
           data = DS_1,
               metadata = DS_1_Tab,
               color = var_A)

 chart_spec(”geographic map",
            data = DS_3,
                color = var_A)
```

Align & Combine



Layout & Display



- For each high-ranked path, automatically generate initial partial specifications for each singleton chart
- Align & combine specifications into multi-chart combination
- Use existing packages for layout
- Combine rendered boxes of pixels into final display

Automatically Constructing Visually Coherent Chart Combinations

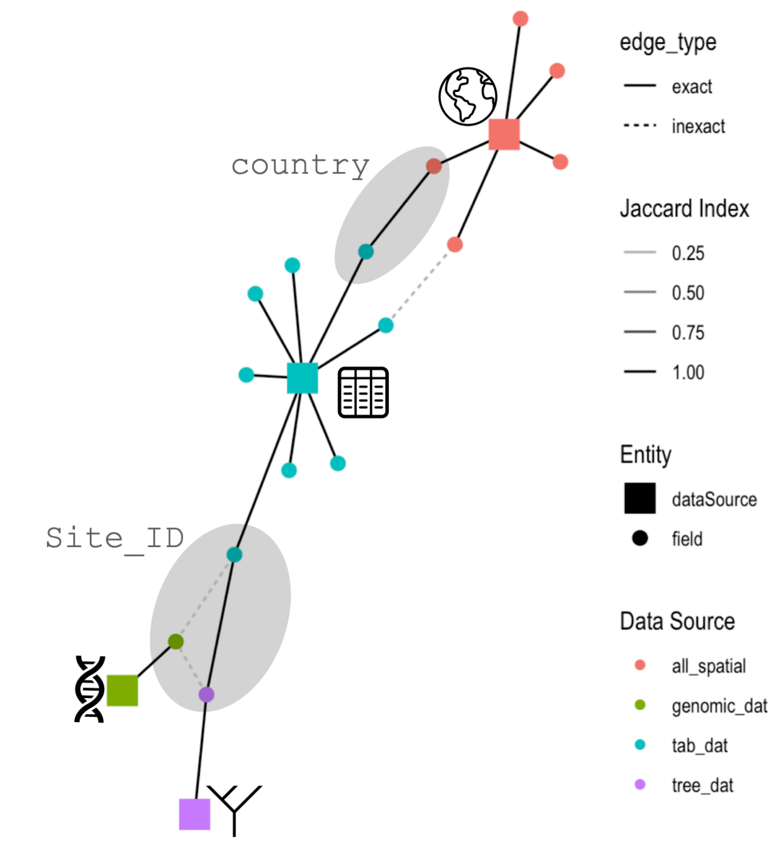
- **GEViTRec runs in R Markdown notebooks**
- **Example: 2013-2016 Ebola outbreak data**

A) GEViTRec Code

Input Data *Tabular* *Geospatial* *Genetic* *Phylogenetic*

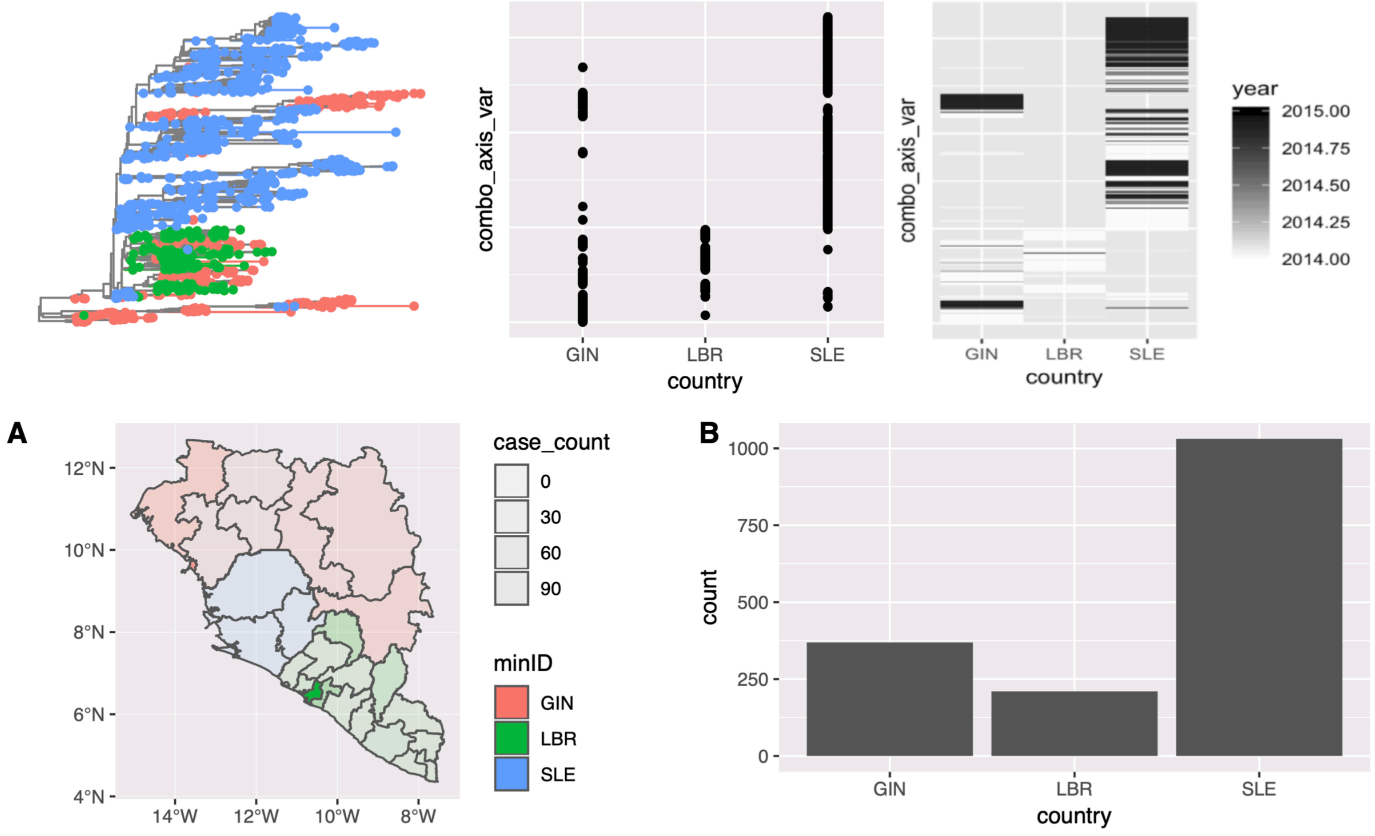
```
# Analyze different
# data types automatically
harmon_obj<-data_linkage(tab_dat,
tree_dat,genomic_dat,all_spatial)

# Create specifications
component_specs<-get_spec_list(harmon_obj)

#plot the result one view at a time
plot_view(component_specs,view_num=1)
```

B) Data Source Graph



C) Top Ranked View

Positional, Color, Field



Automatically Constructing Visually Coherent Chart Combinations

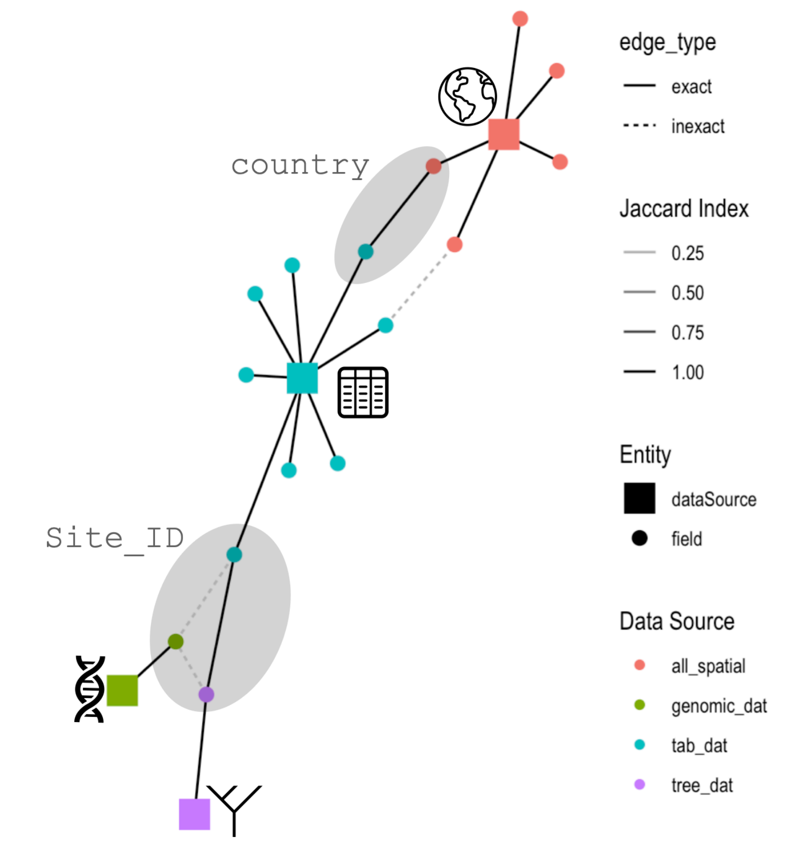
- **GEViTRec runs in R Markdown notebooks**
- **Example: 2013-2016 Ebola outbreak data**

A) GEViTRec Code

Input Data *Tabular* *Geospatial* *Genetic* *Phylogenetic*

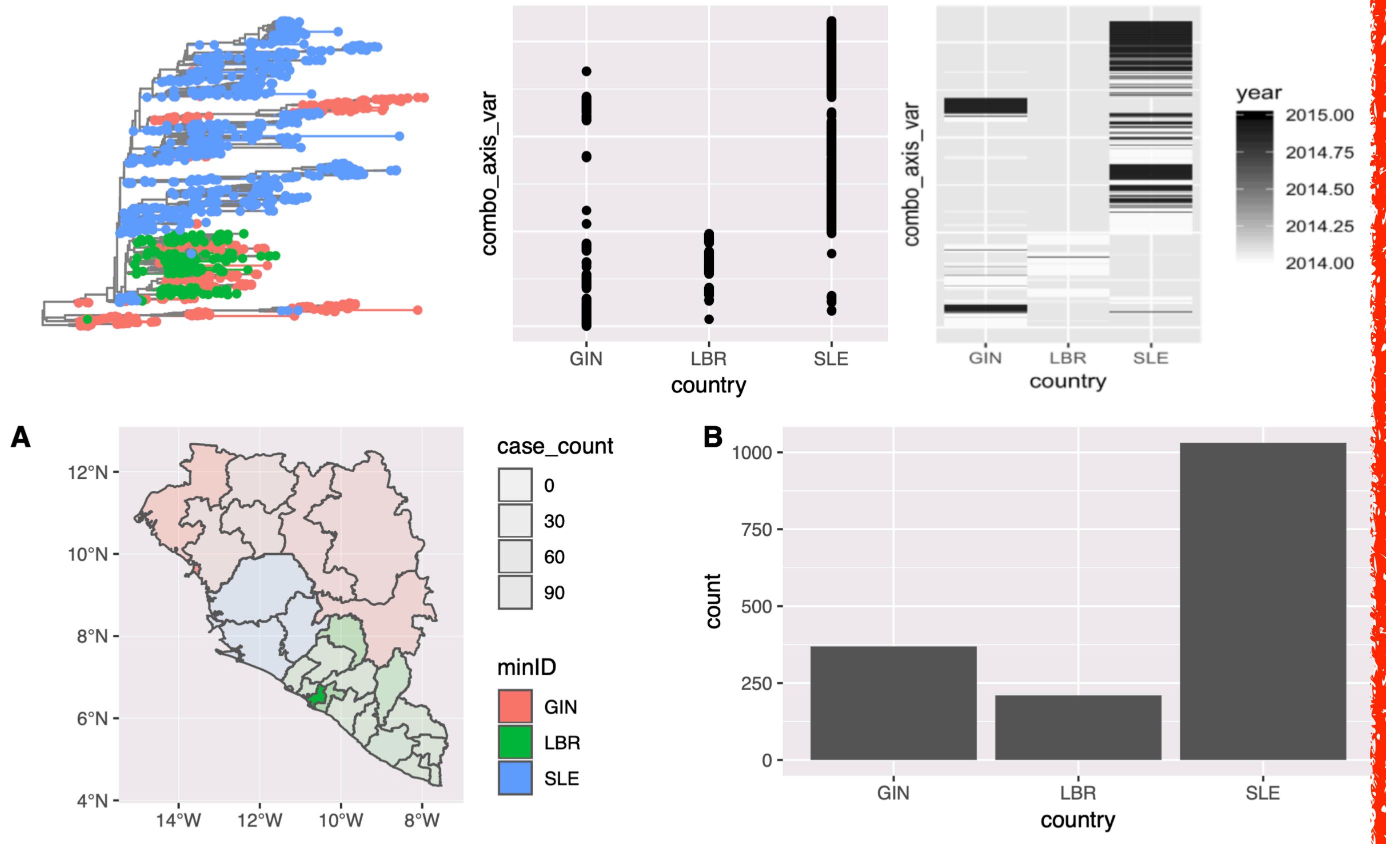
```
# Analyze different
# data types automatically
harmon_obj<-data_linkage(tab_dat,
tree_dat,genomic_dat,all_spatial)

# Create specifications
component_specs<-get_spec_list(harmon_obj)

#plot the result one view at a time
plot_view(component_specs,view_num=1)
```

B) Data Source Graph



C) Top Ranked View

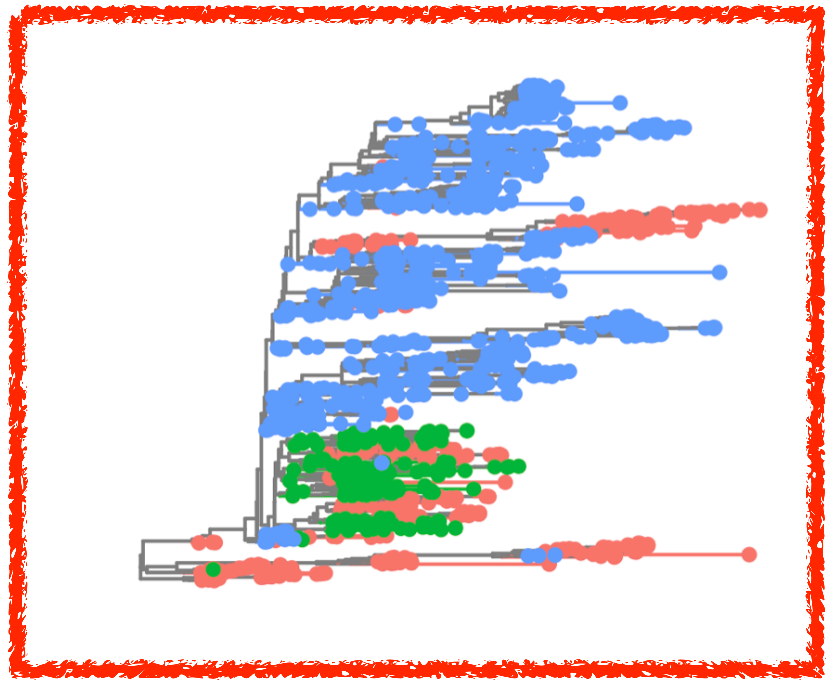
Positional, Color, Field

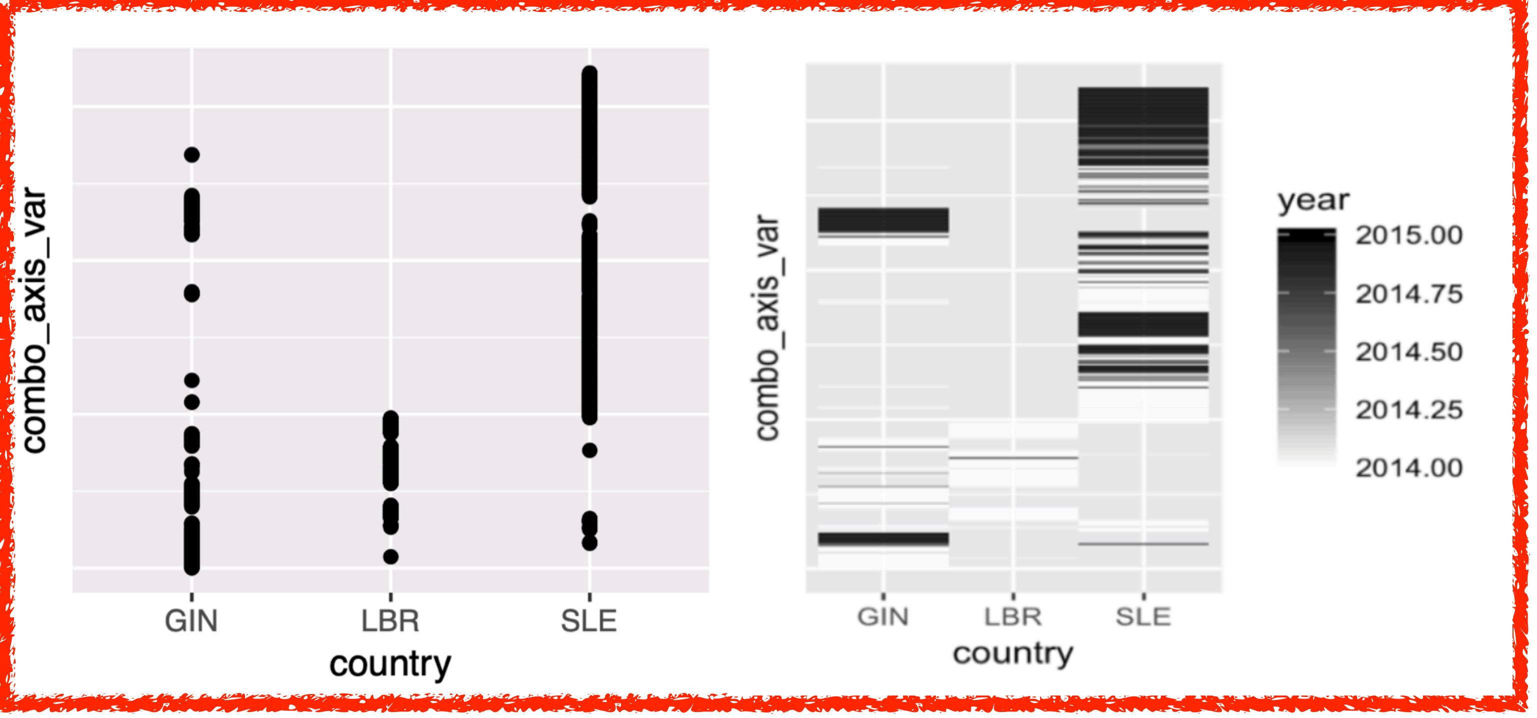


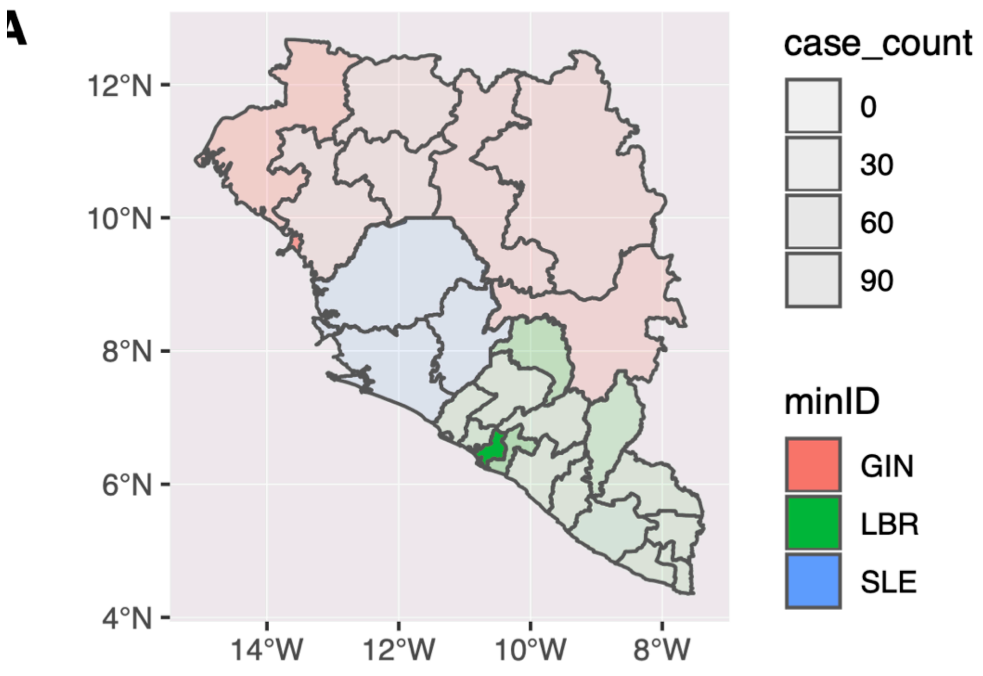
Automatically Constructing Visually Coherent Chart Combinations

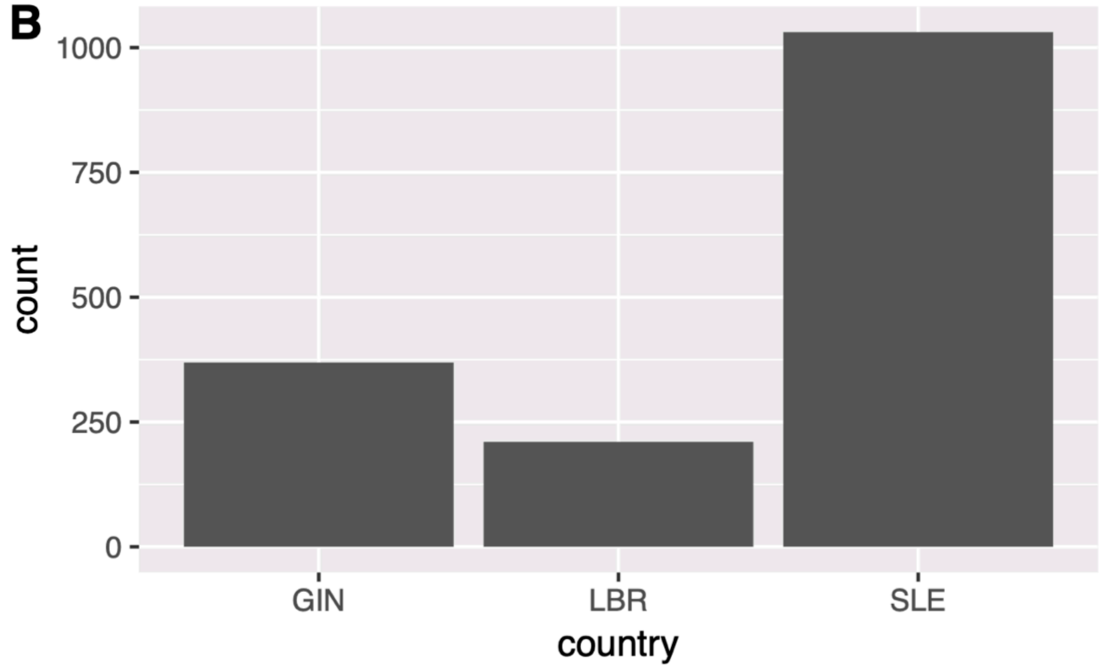
C) Top Ranked View

Positional, Color, Field





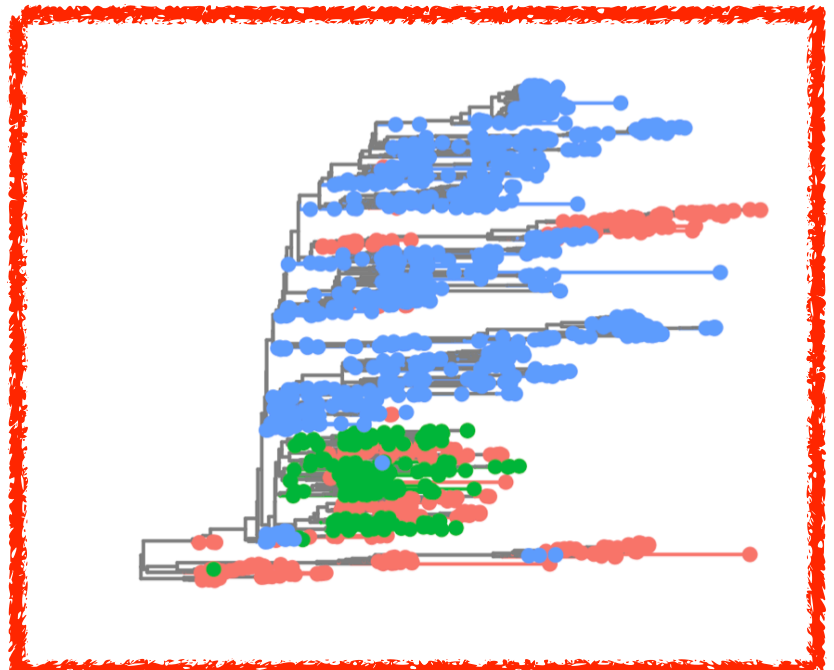


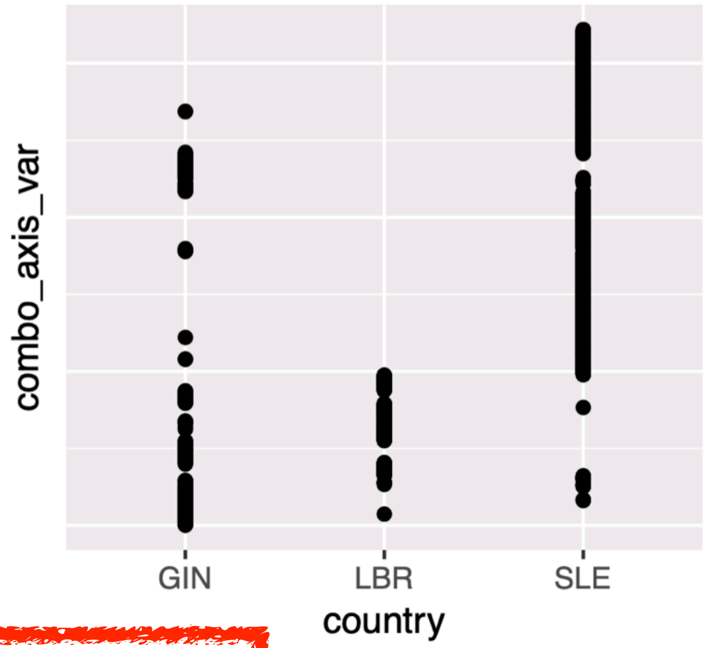


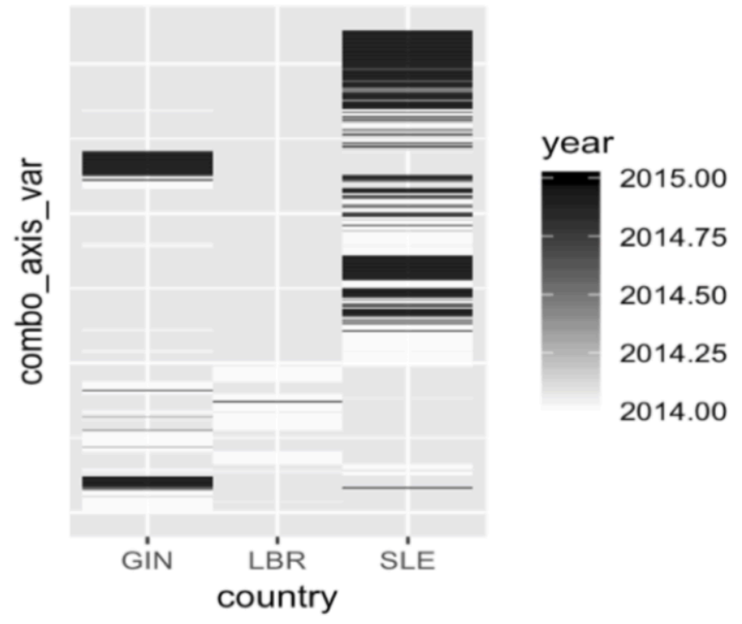
Automatically Constructing Visually Coherent Chart Combinations

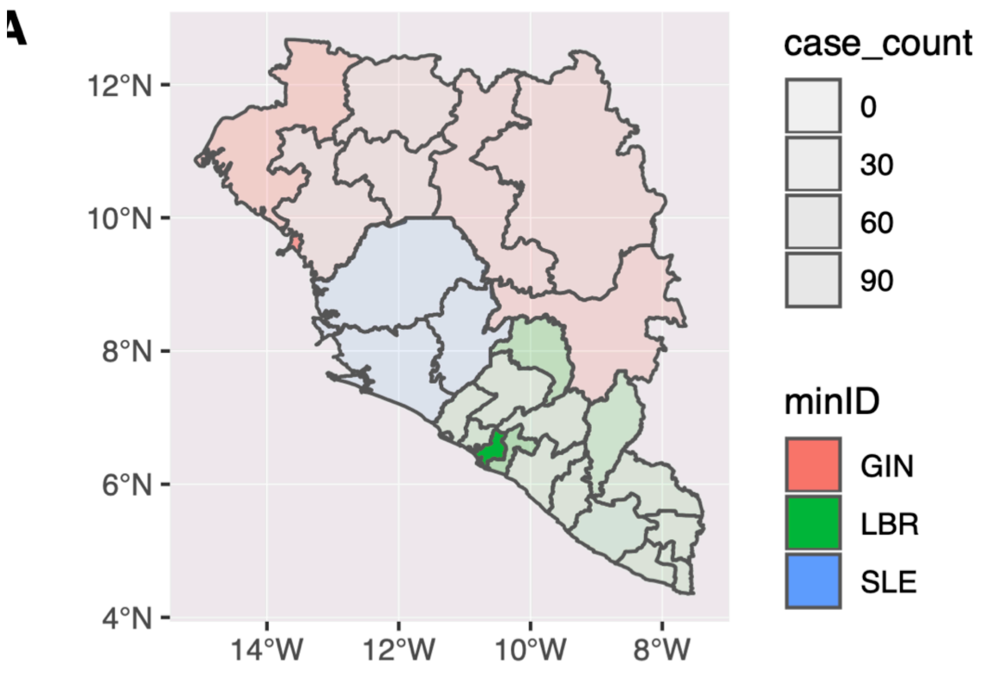
C) Top Ranked View

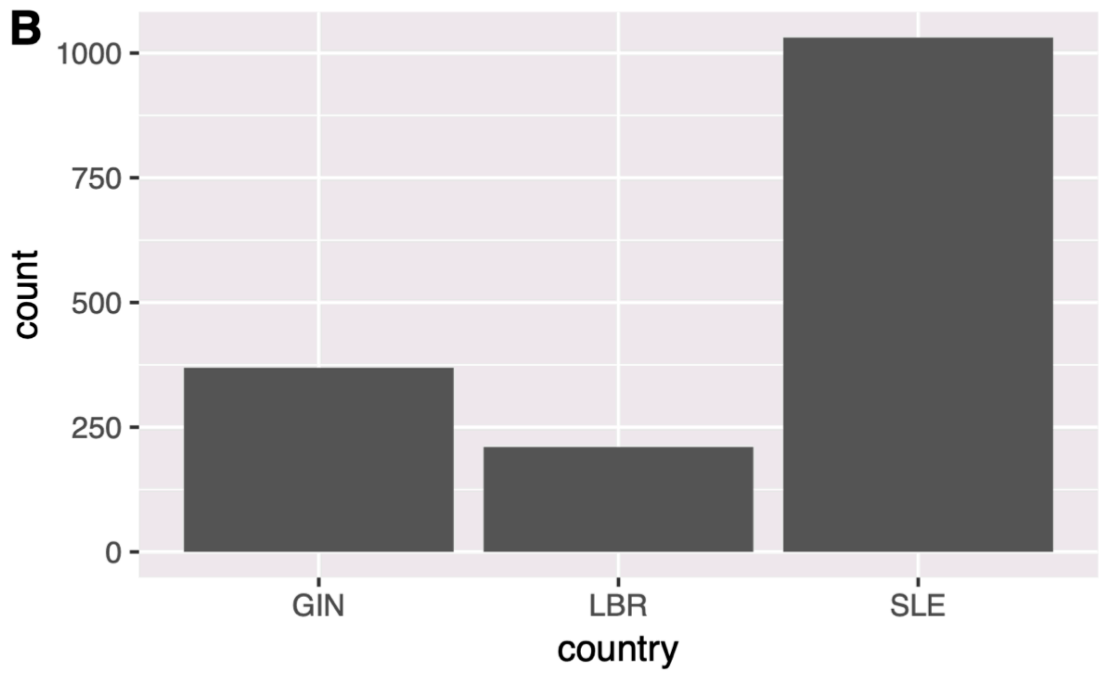
Positional, Color, Field







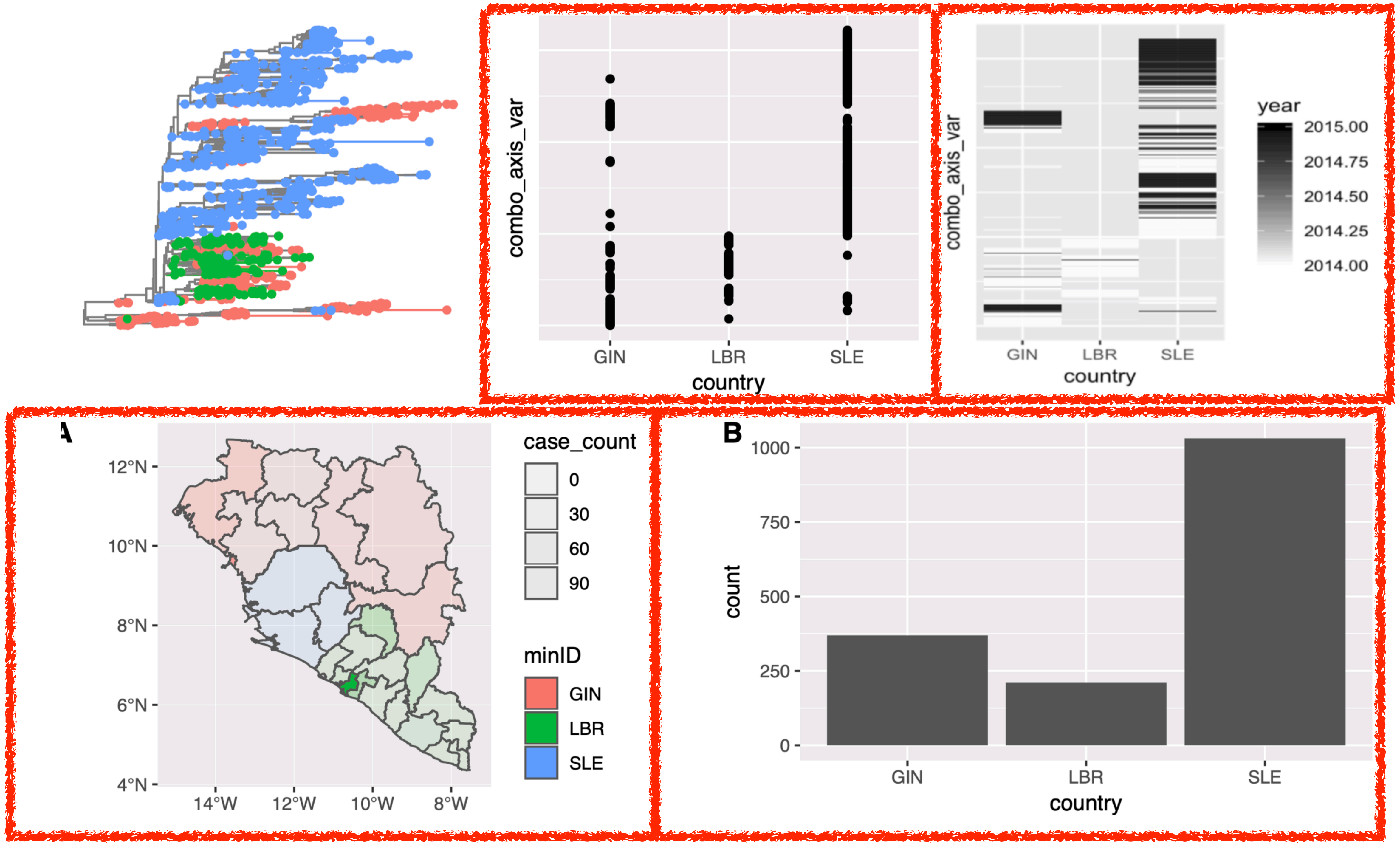


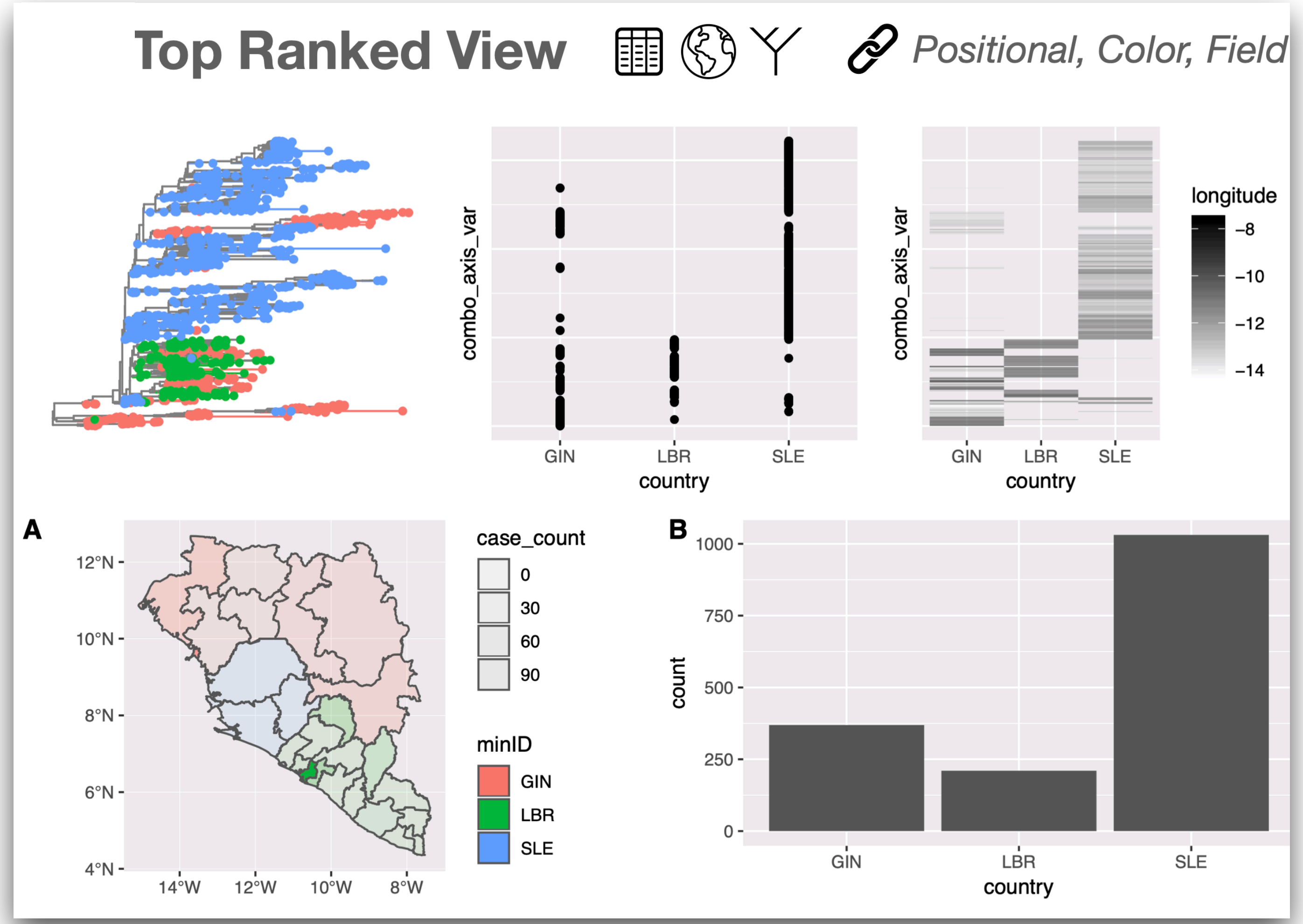


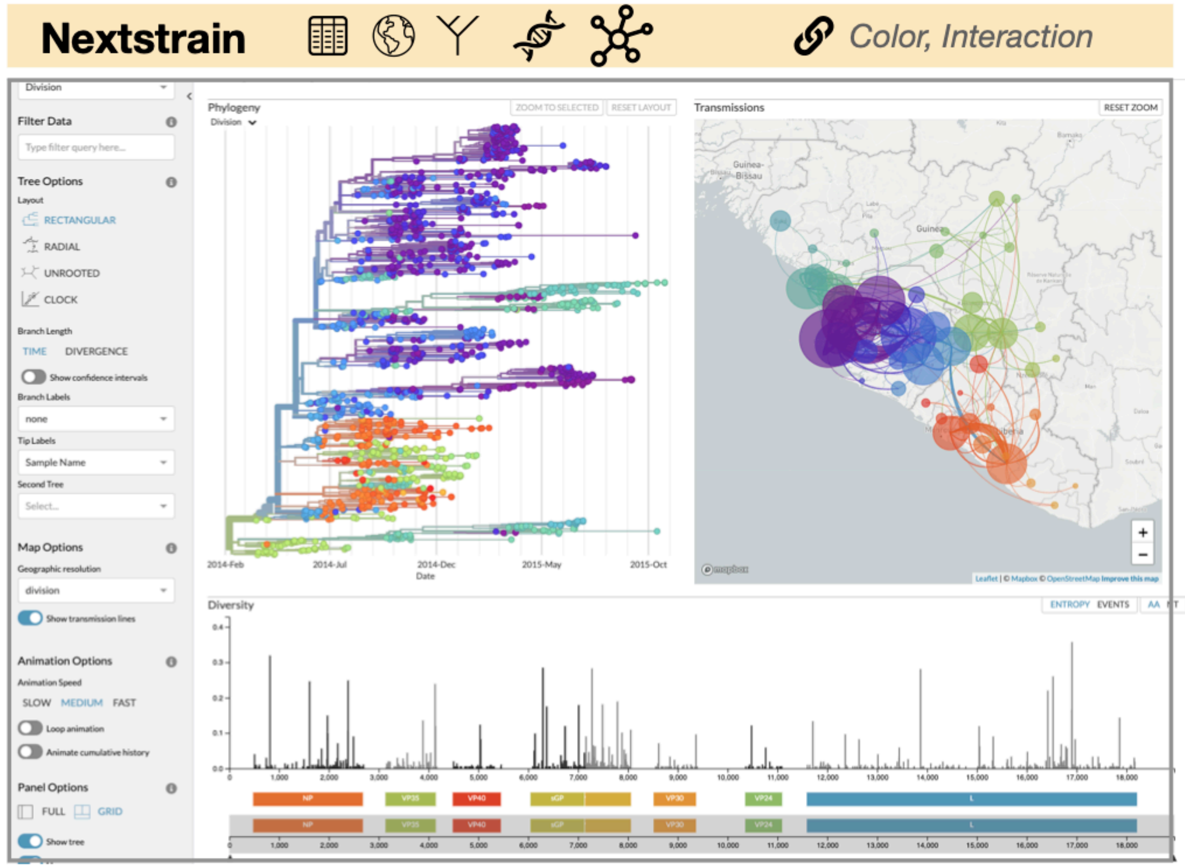
Automatically Constructing Visually Coherent Chart Combinations

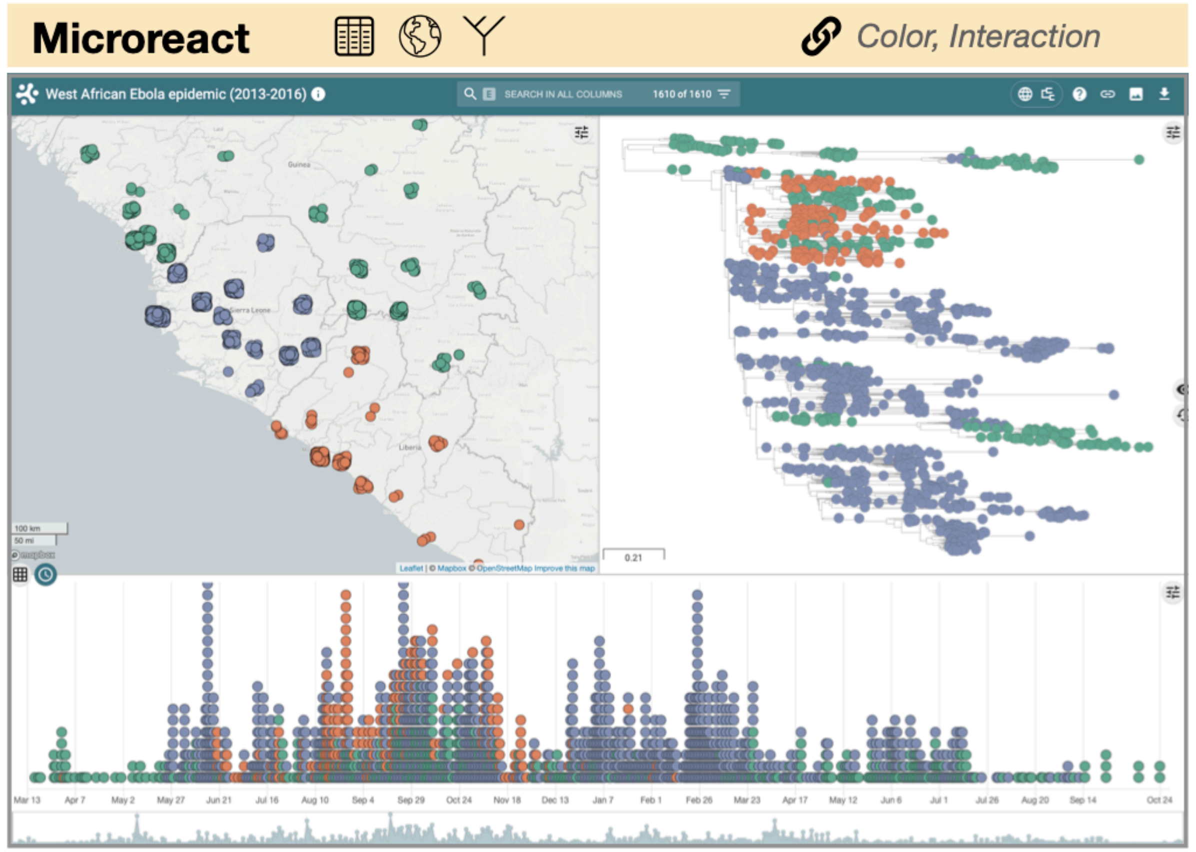
C) Top Ranked View

Positional, Color, Field



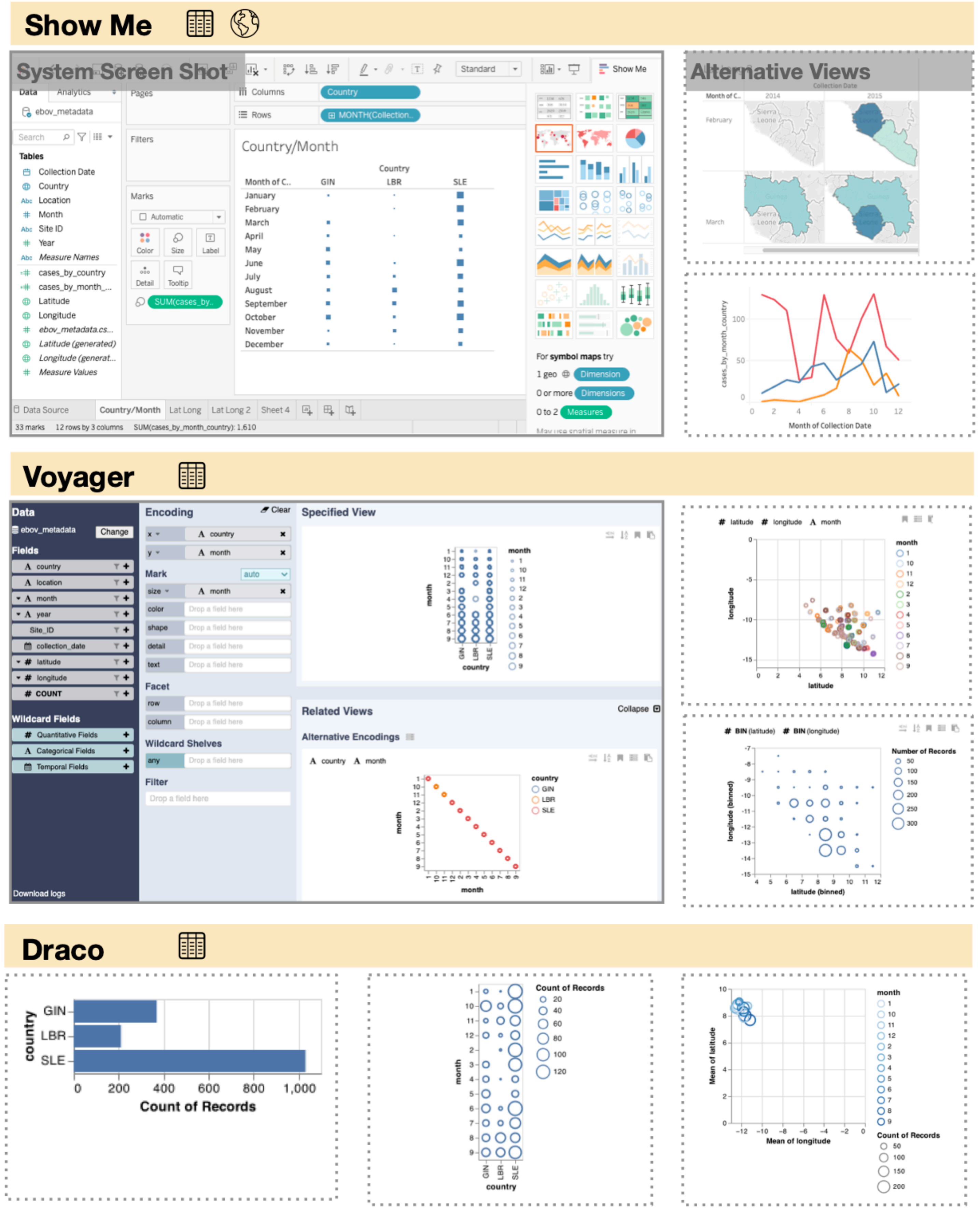


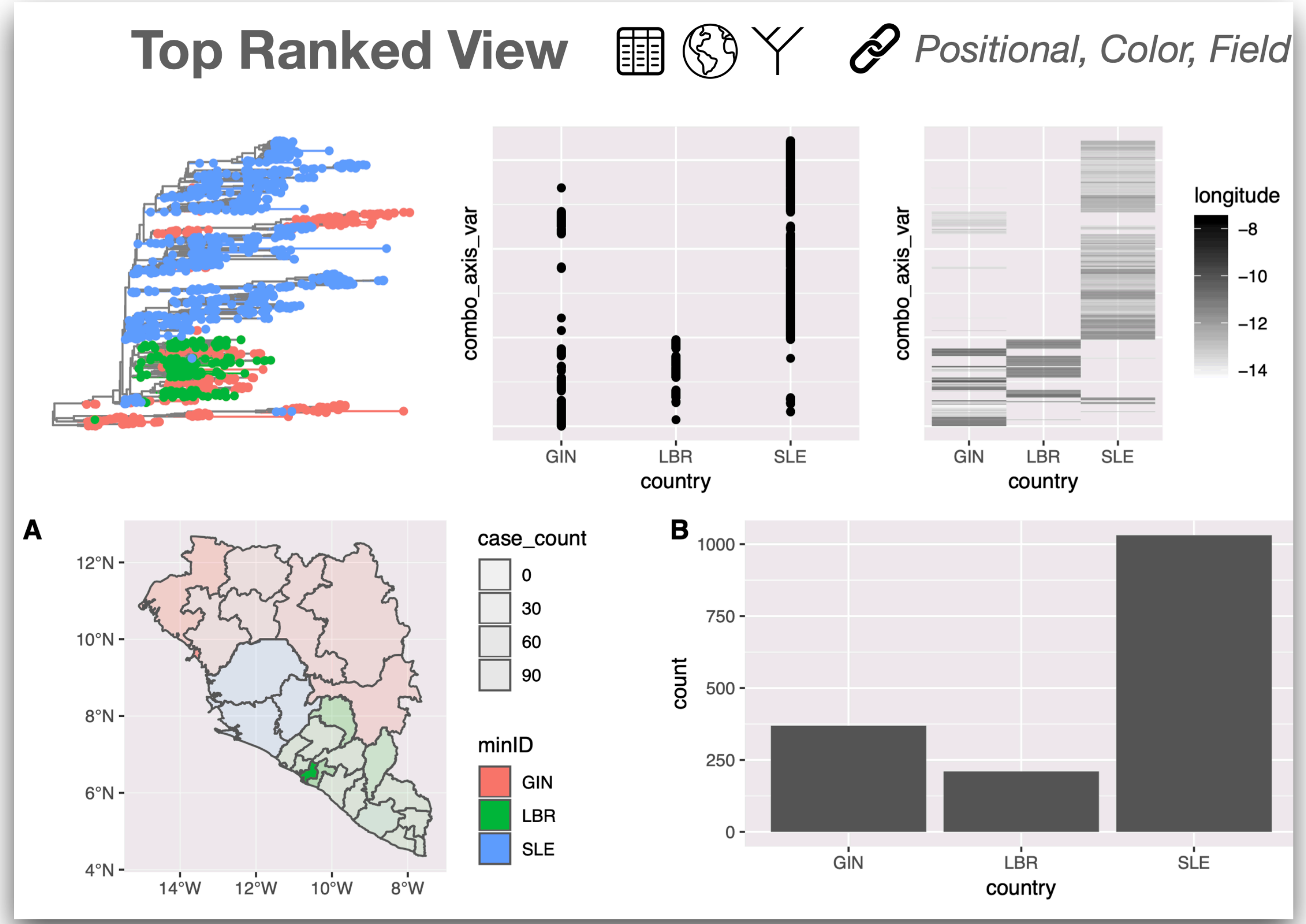




- Comparing to existing ***bespoke*** tools:
  - **Slow:**
    - Require extensive **manual curation**
    - Are **less adaptive** to changing data
  - **Aligned**: Have **better alignment** between chart types
  - **Heterogeneity support**: handle multiple types of data

Data Legend *Tabular* *Geospatial* *Phylogenetic* *Genetic* *Network*





- Comparing to existing ***recommendation*** tools:
  - **Fast: easy to use**
  - **Unaligned**
    - Suggest **one chart** at a time
    - Require **manual curation** for alignment
  - **Heterogeneity support limited**

Data Legend *Tabular* *Geospatial* *Phylogenetic* *Genetic* *Network*

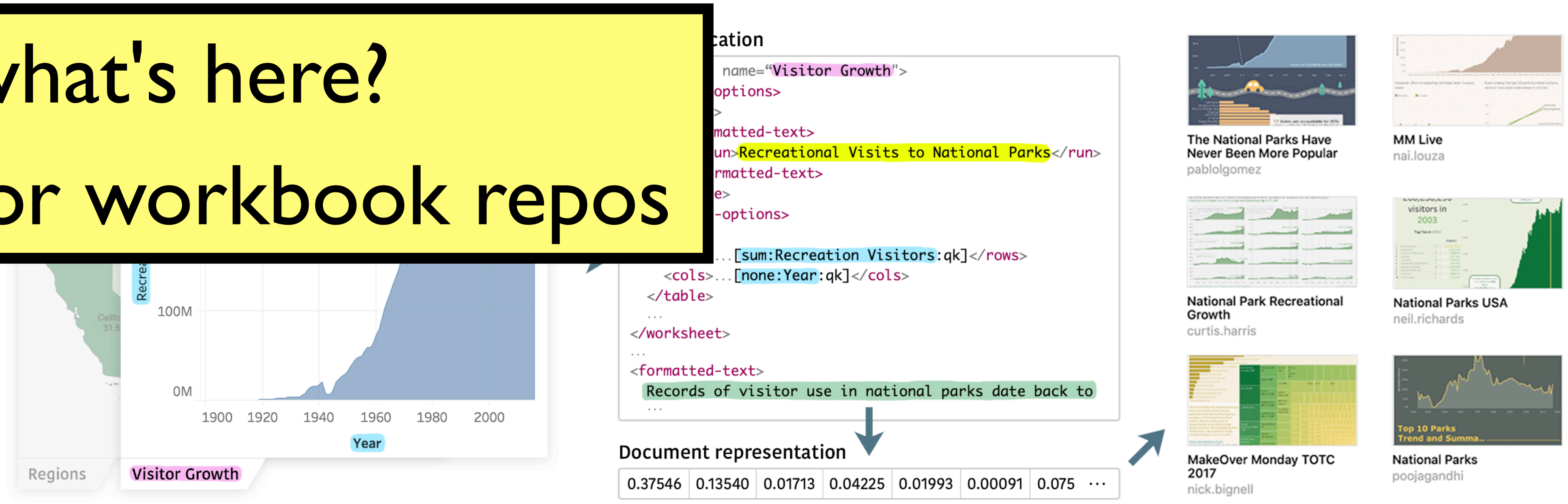
GEViTRec lowers burden to quickly visualize data

- **Speeds up** the process of data reconnaissance - **where are we?**

- Automatically shows us **what's here?**
  - Identifies connections among datasets
  - Exploits domain prevalence design space
  - Constructs visually coherent chart combinations through gradual binding

what's here?

for workbook repos



VizCommender:

Computing Text-Based Similarity in Visualization Repositories for Content-Based Recommendations

Michael Oppermann
UBC/
Virtual Identity

Robert Kincaid
Tableau

Tamara Munzner
UBC

https://www.cs.ubc.ca/group/infovis/pubs/2020/vizcommender/

VizCommender: Computing Text-Based Similarity in Visualization Repositories for Content-Based Recommendations

Oppermann, Kincaid Munzner. IEEE TVCG 27(2): 495-505, 2021 (Proc. VIS 2020).

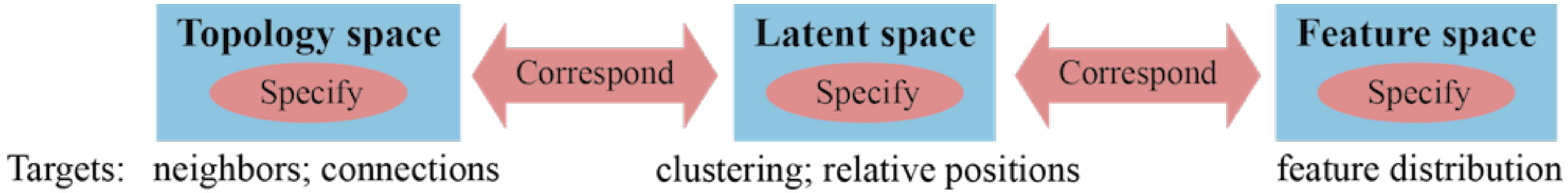
Questions in road trips - and visualization in data science!

- where are we?
  - Data Reconnaissance & Task Wrangling
- what's here?
  - Automatic Encodings through Recommendation

    to shed light on data landscapes
- are we there yet? are we lost?
  - Visual Assessment of ML Training Completion & Quality

https://unsplash.com/photos/2oYHfuRe4OU

http://www.cs.ubc.ca/~tmm/talks.html#vds23

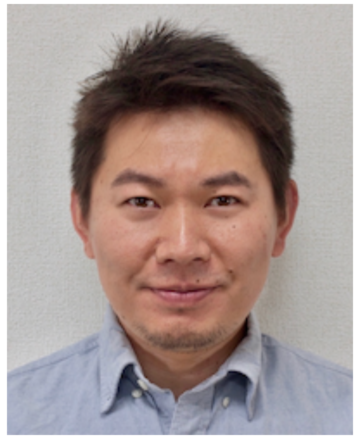


Visualizing Graph Neural Networks with CorGIE:

Corresponding a Graph to Its Embedding

Zipeng Liu
UBC/Beihang

Yang Wang
Uber/Facebook



Jürgen Bernard
UBC/Zurich

Tamara Munzner
UBC

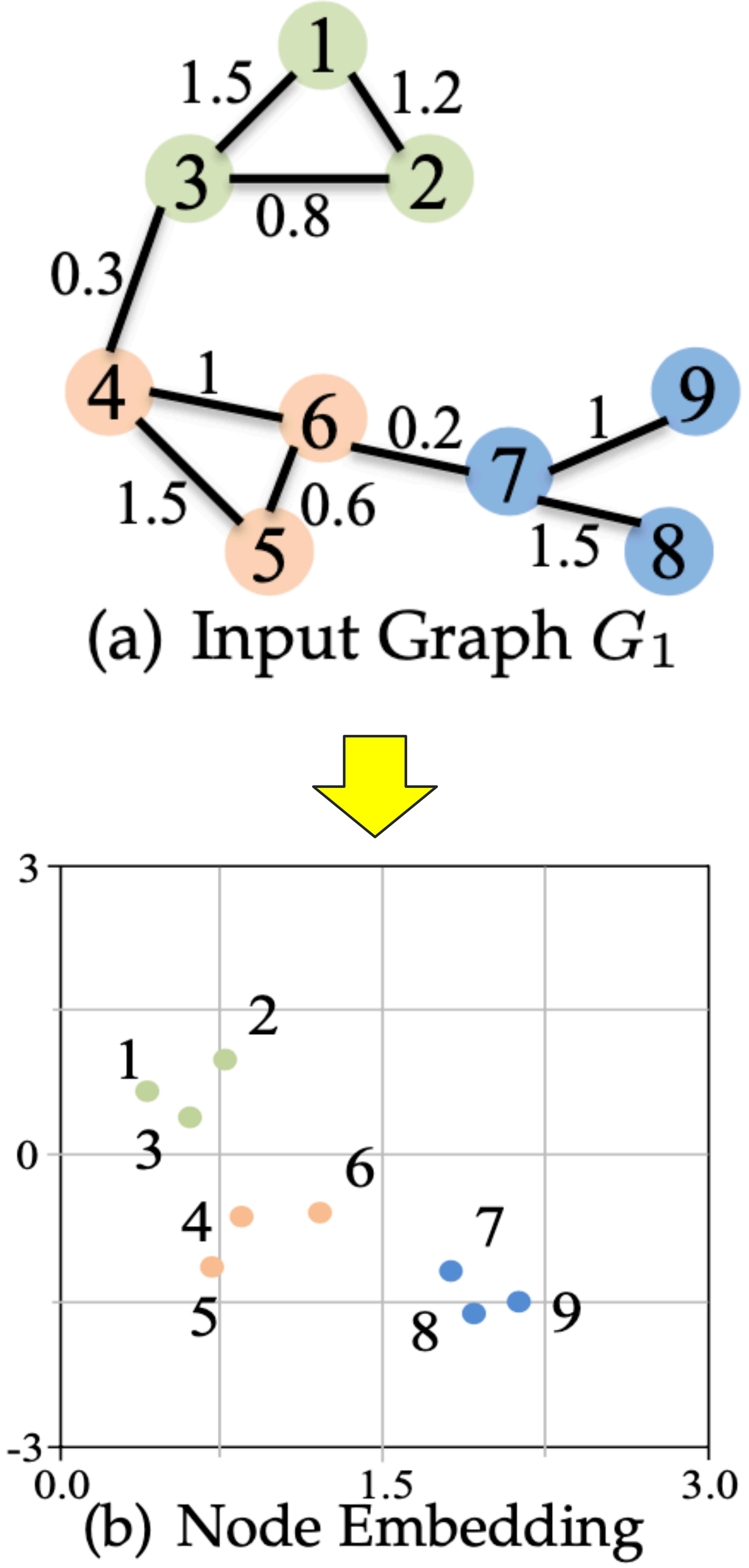
https://arxiv.org/abs/2106.12839

Visualizing Graph Neural Networks with CorGIE: Corresponding a Graph to Its Embedding.
Liu, Wang, Bernard, Munzner. IEEE TVCG 28(6):2500-2516, 2022.

Graph neural network (GNN)

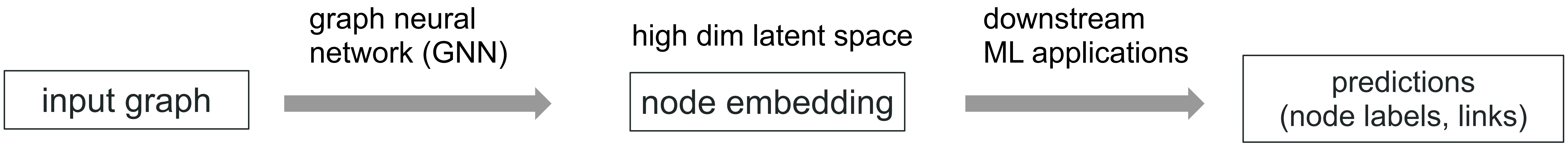
- machine learning (ML) models for graphs
  - like CNN for images
  - like Transformer for text

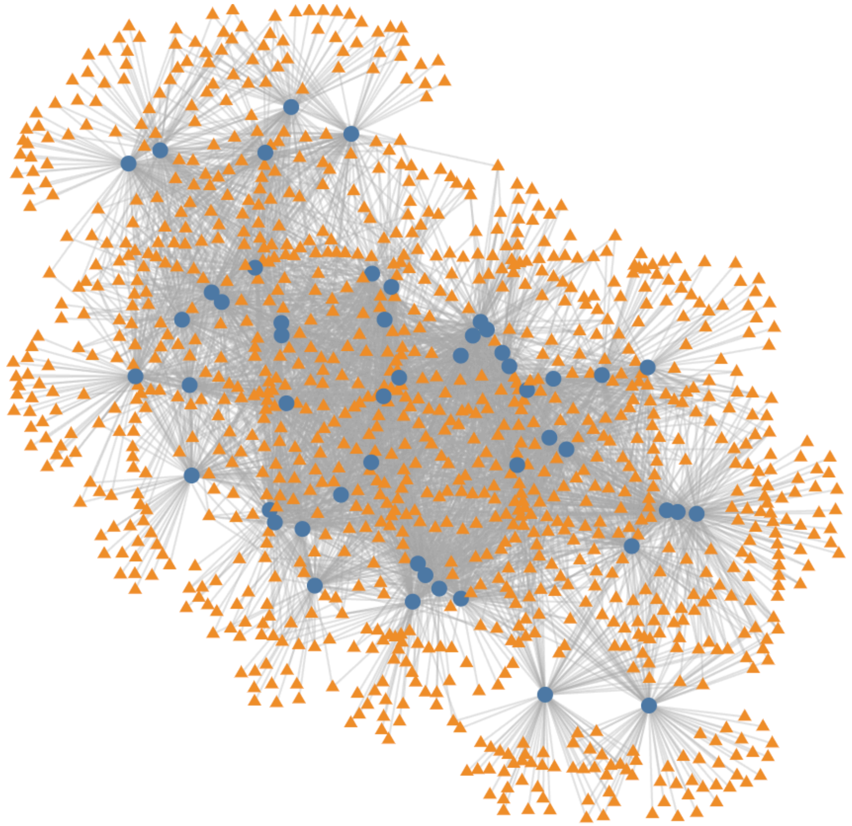
- many real-world graph-related applications
  - node classification
    - examples: fraud detection, disease classification
  - link prediction
    - examples: product recommendation, protein interactions



[Cai et al. TKDE'18]

Graph neural network (GNN)





movie – user graph

Node 0	-1.98	0.74	-0.51	1.19	-1.20	0.97	-1.66	0.90	0.06	-1.89	-1.33	-0.77	0.37	1.60	-0.13	1.47
Node 1	-0.21	0.11	-0.08	0.17	-0.16	0.14	-0.21	0.15	-0.03	-0.20	-0.18	-0.15	0.09	0.15	-0.01	0.16
Node 2	-0.23	0.12	-0.09	0.20	-0.17	0.16	-0.23	0.17	-0.04	-0.22	-0.21	-0.17	0.10	0.16	-0.02	0.18
Node 3	-0.27	0.15	-0.11	0.23	-0.20	0.19	-0.27	0.21	-0.05	-0.26	-0.25	-0.21	0.13	0.18	-0.02	0.21
Node 4	-0.30	0.17	-0.12	0.27	-0.23	0.23	-0.31	0.24	-0.07	-0.29	-0.29	-0.25	0.15	0.20	-0.03	0.23
Node 5	-0.19	0.09	-0.06	0.14	-0.13	0.11	-0.17	0.11	-0.01	-0.18	-0.15	-0.11	0.06	0.15	-0.01	0.15
Node 6	-0.28	0.16	-0.11	0.26	-0.22	0.22	-0.30	0.23	-0.07	-0.28	-0.28	-0.24	0.14	0.19	-0.03	0.22
Node 7	-0.30	0.17	-0.12	0.27	-0.23	0.22	-0.31	0.24	-0.07	-0.29	-0.28	-0.25	0.15	0.20	-0.03	0.23
Node 8	-0.23	0.12	-0.08	0.18	-0.16	0.14	-0.21	0.15	-0.02	-0.22	-0.19	-0.15	0.09	0.17	-0.01	0.18
Node 9	-0.31	0.18	-0.12	0.28	-0.24	0.24	-0.33	0.25	-0.08	-0.30	-0.30	-0.26	0.16	0.20	-0.03	0.24
Node 10	-0.33	0.19	-0.13	0.30	-0.26	0.25	-0.35	0.27	-0.08	-0.32	-0.32	-0.28	0.17	0.22	-0.03	0.26
Node 11	-0.21	0.11	-0.07	0.17	-0.16	0.14	-0.20	0.15	-0.03	-0.20	-0.18	-0.14	0.09	0.16	-0.01	0.17
Node 12	-0.20	0.10	-0.07	0.16	-0.15	0.13	-0.19	0.14	-0.03	-0.19	-0.17	-0.13	0.08	0.15	-0.01	0.16
Node 13	-0.26	0.14	-0.10	0.23	-0.20	0.19	-0.26	0.20	-0.05	-0.25	-0.24	-0.20	0.12	0.18	-0.02	0.20
Node 14	-0.19	0.08	-0.06	0.13	-0.13	0.11	-0.17	0.11	-0.01	-0.18	-0.15	-0.10	0.06	0.14	-0.01	0.15
Node 15	-0.16	0.06	-0.04	0.09	-0.10	0.07	-0.13	0.07	0.01	-0.14	-0.11	-0.06	0.03	0.13	-0.00	0.12

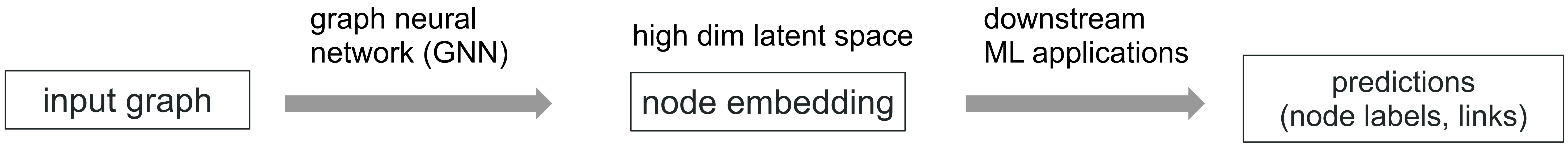
a vector for each node

node 0: Alice

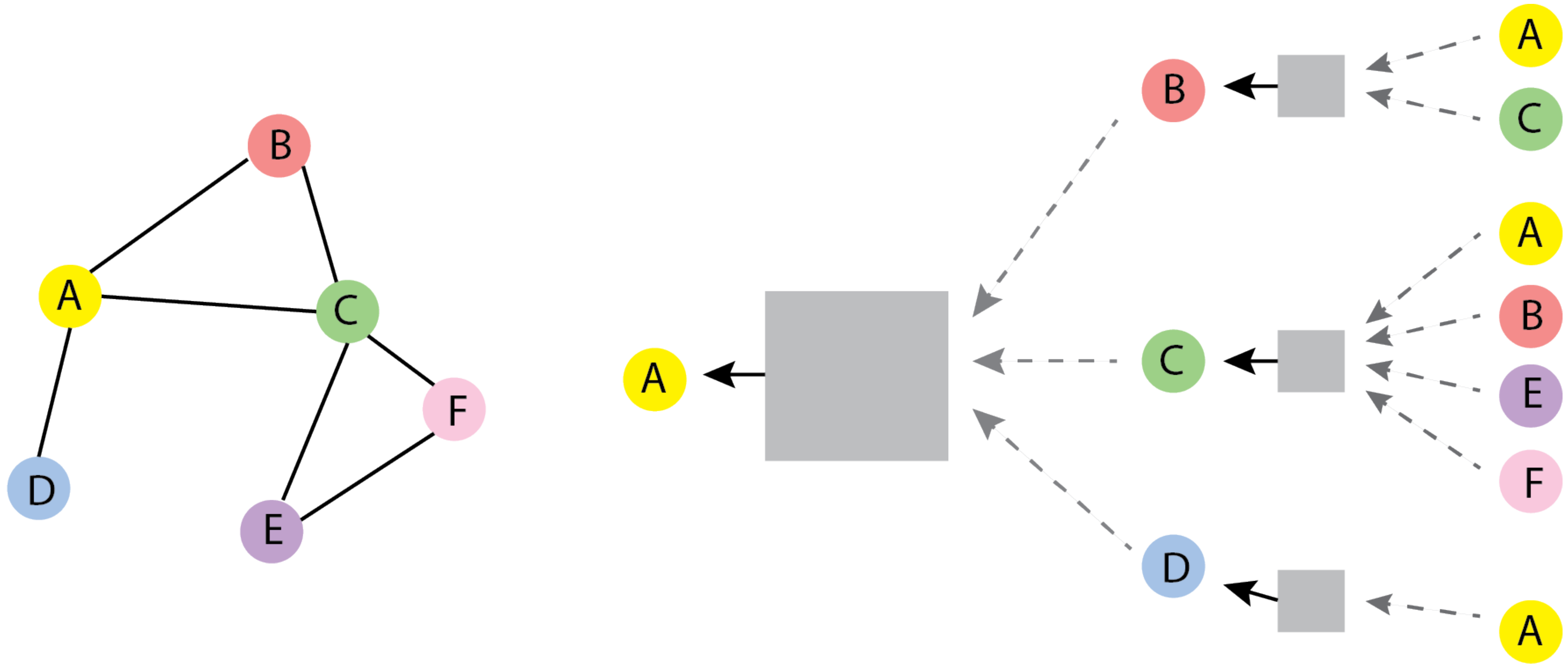
movie
recommendation

node 12: Lord of the Rings

Graph neural network (GNN)



node features are aggregated / passed through **topological neighborhood**



Remake from https://snap-stanford.github.io/cs224w-notes/machine-learning-with-networks/graph-neural-networks

Evaluating GNN quality

Two big-picture questions

- Are we there yet? *Should we train / tune more?*
- Are we lost? *Does it behave as we expect?*

Evaluate GNN: CorGIE idea

high dim latent space

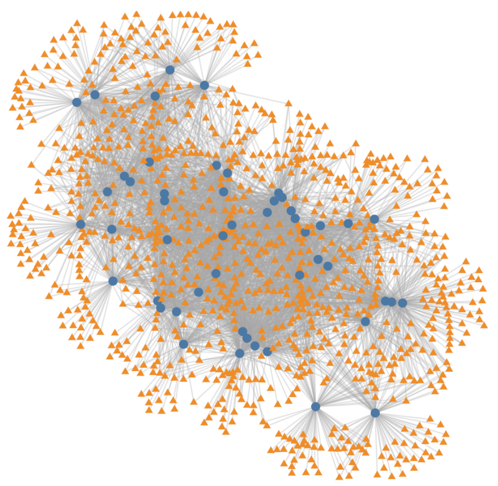
downstream
ML applications

input graph → graph neural network (GNN) → node embedding → predictions (node labels, links)

shared topo neighbors
similar node features

explore correspondences

nearby positions



where are we?
what's here?

Evaluate GNN: CorGIE idea

input graph → graph neural network (GNN) → high dim latent space: node embedding → downstream ML applications → predictions (node labels, links)

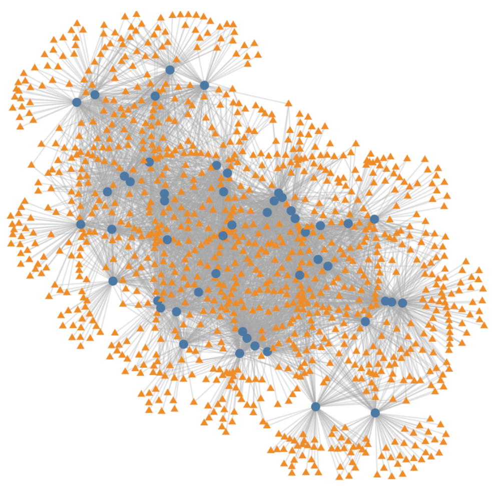
shared topo neighbors
similar node features

explore correspondences

nearby positions

Examples of correspondences:

Check [similar topology? Similar node features?] ← – – – Pick [a cluster]



where are we?
what's here?

Evaluate GNN: CorGIE idea

input graph → graph neural network (GNN) → node embedding (high dim latent space) → downstream ML applications → predictions (node labels, links)

shared topo neighbors
similar node features

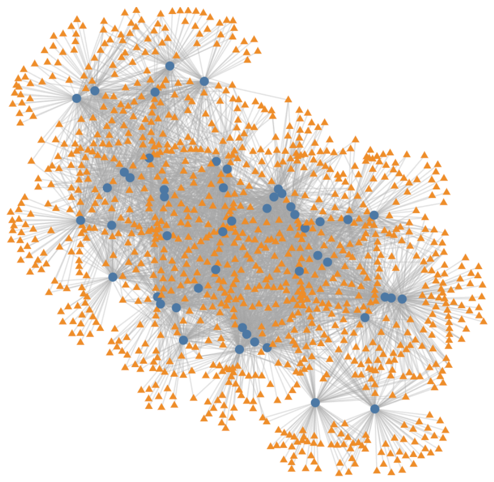
explore correspondences

nearby positions

Examples of correspondences:

Check [similar topology? Similar node features?] ← – – – Pick [a cluster]

Check [different topology? Different node features?] ← – – – Pick [two far-away clusters]



where are we?
what's here?

Evaluate GNN: CorGIE idea

input graph

graph neural network (GNN)

high dim latent space

node embedding

downstream
ML applications

predictions
(node labels, links)

shared topo neighbors
similar node features

explore
correspondences

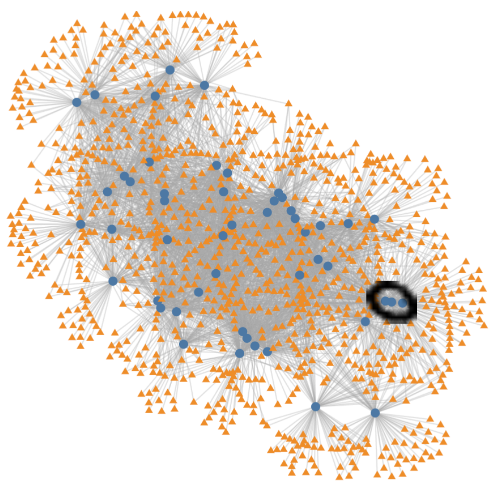
nearby positions

Examples of correspondences:

Check [similar topology? Similar node features?] ← – – – Pick [a cluster]

Check [different topology? Different node features?] ← – – – Pick [two far-away clusters]

Pick [two nodes sharing many topo neighbors] – – – → Check [how close the nodes are compared to others?]



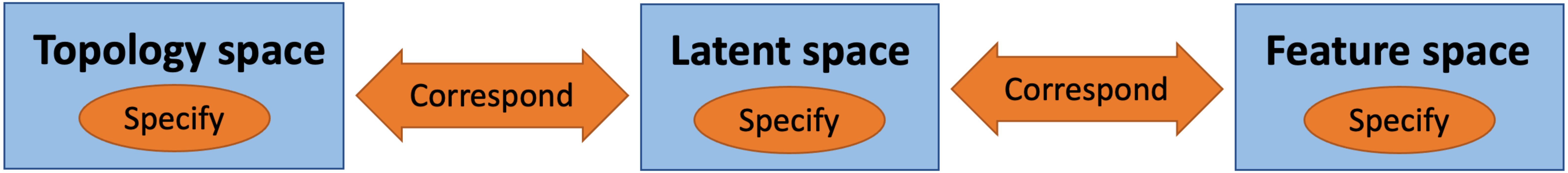
where are we?
what's here?

Data and tasks

Topology space | **Latent space** | **Feature space**

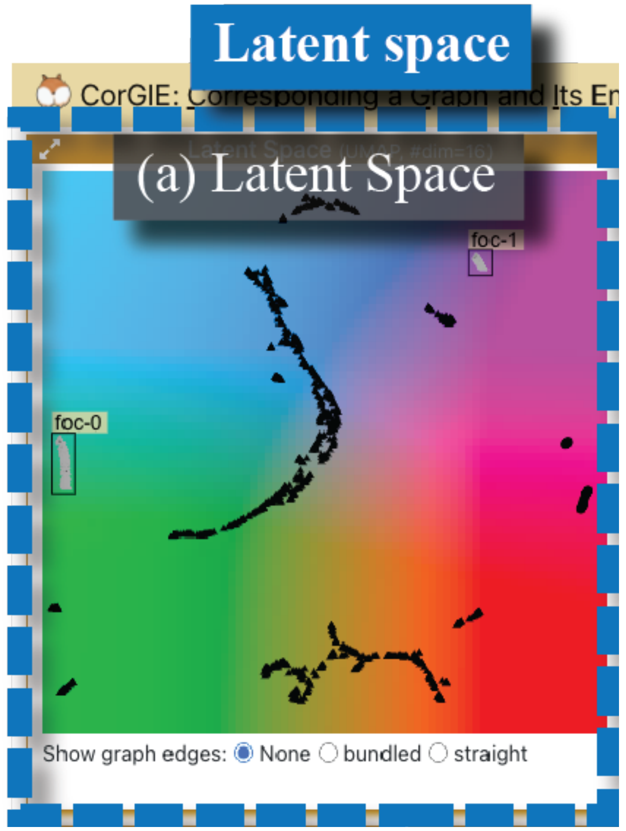
- three data spaces

Data and tasks

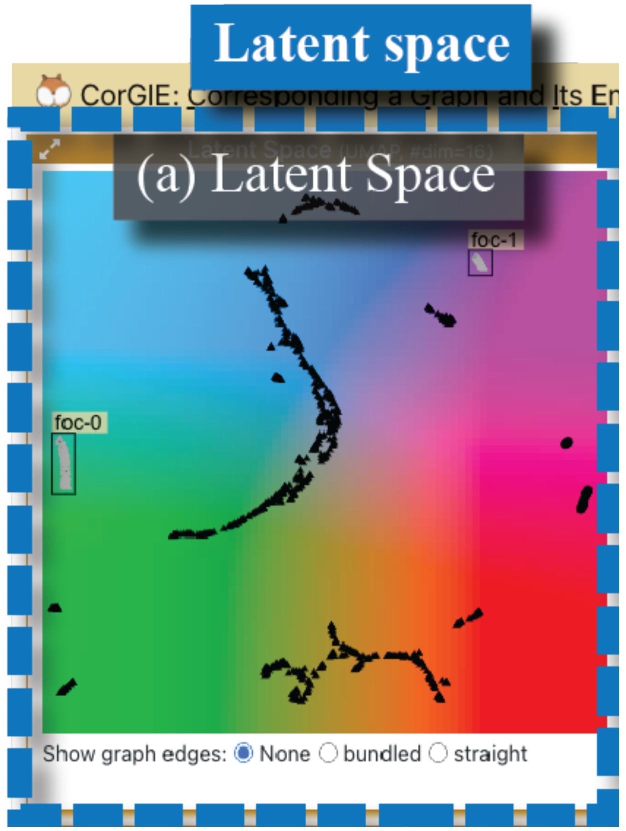


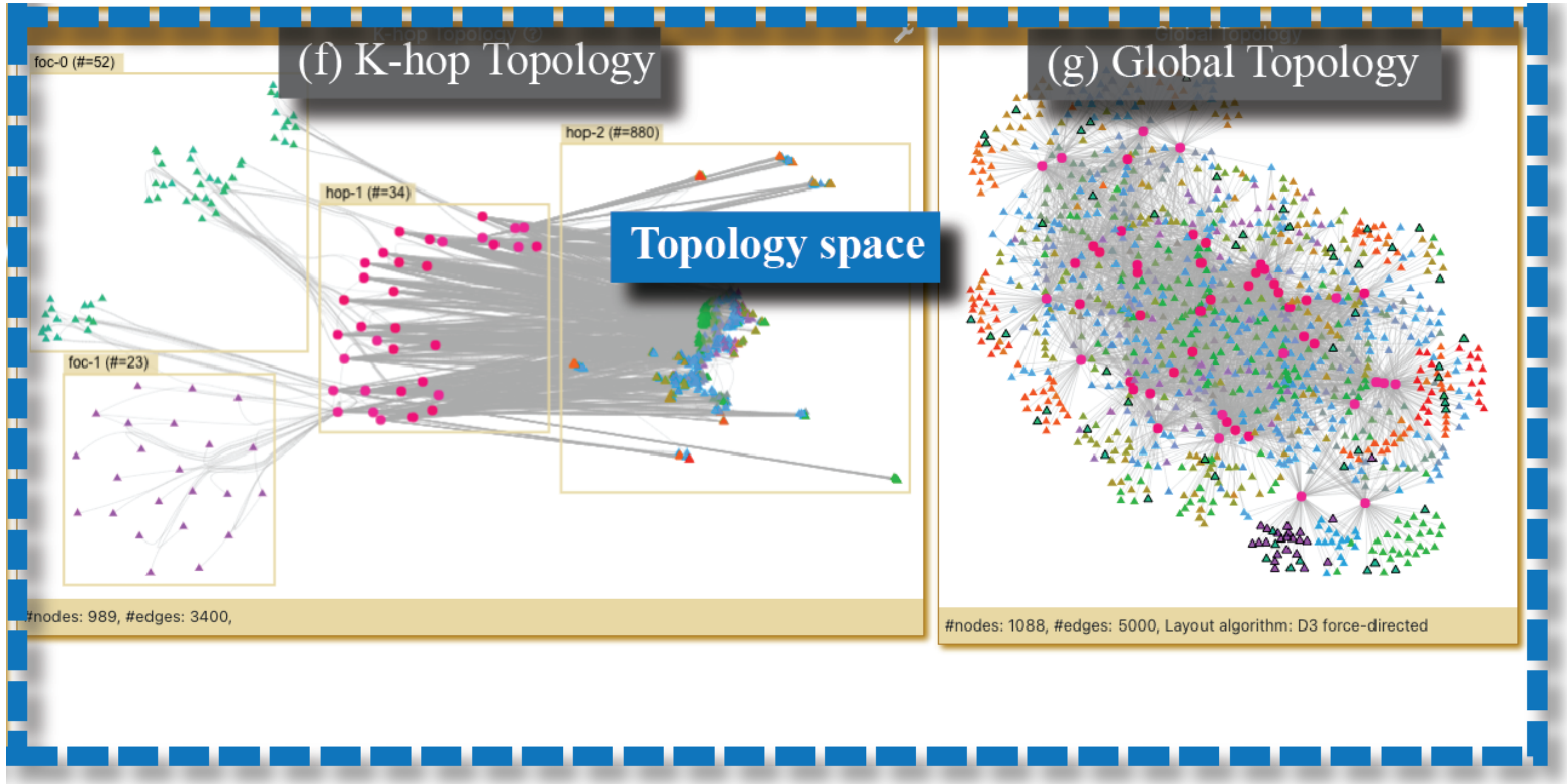
- three data spaces
- tasks
  - specify
  - correspond

CorGIE multi-view interactive interface

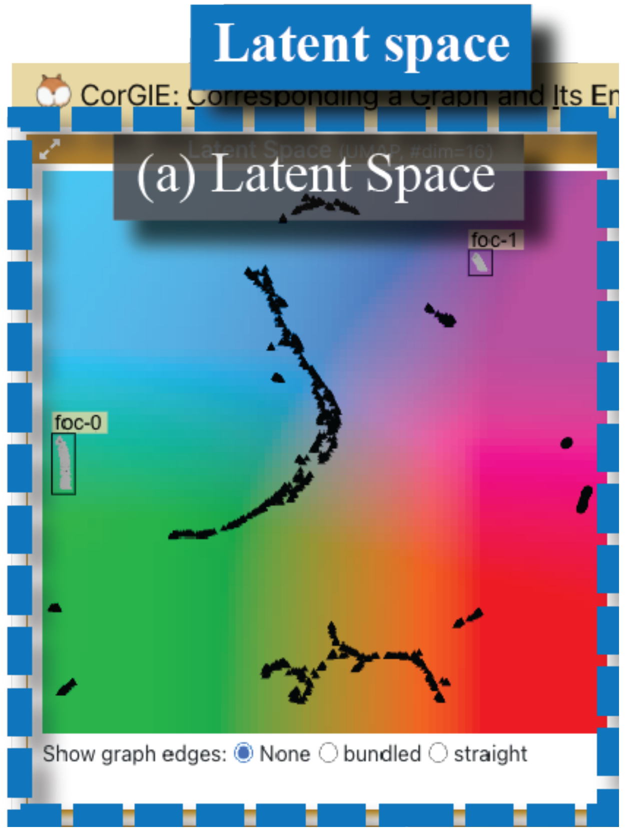


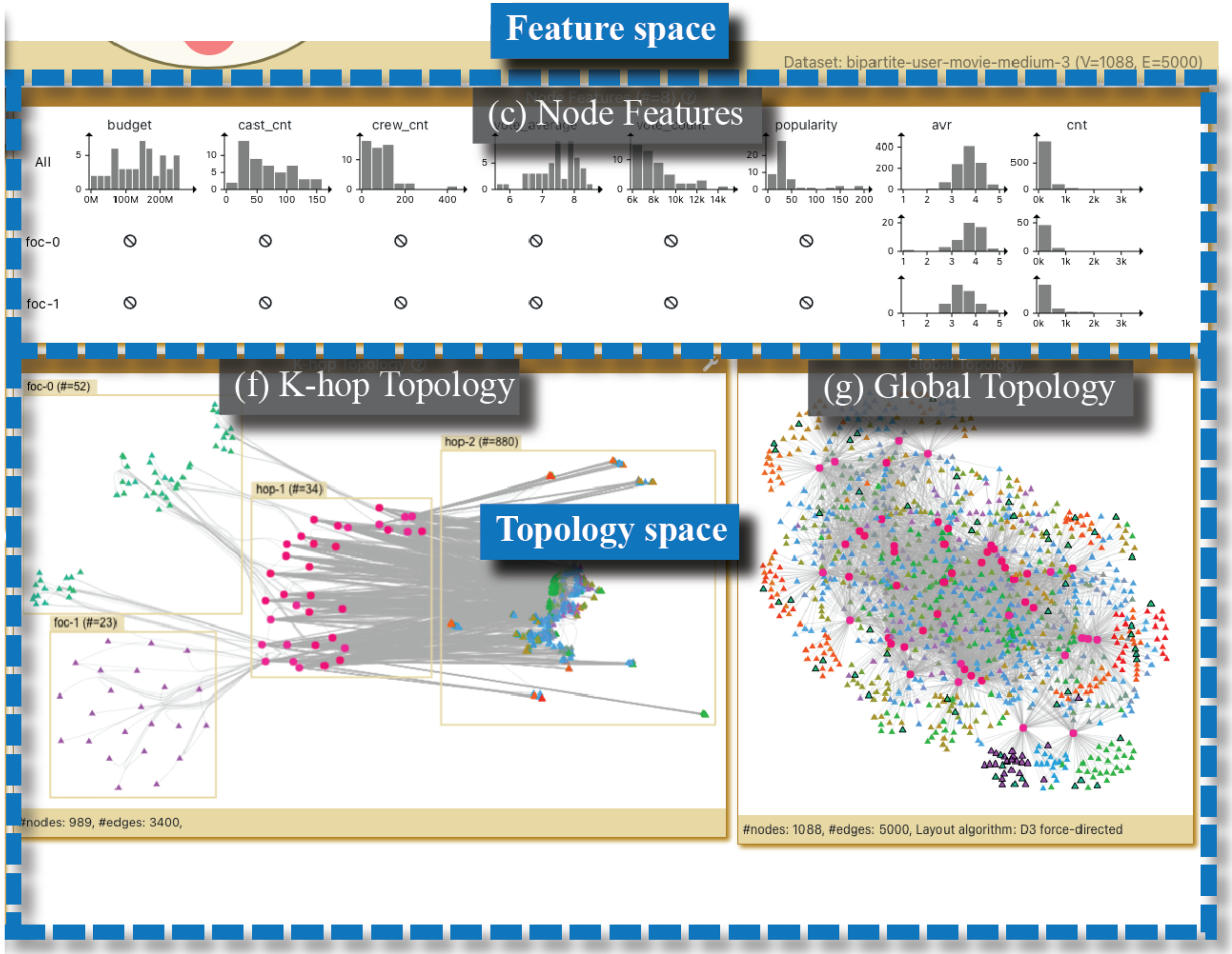
CorGIE multi-view interactive interface



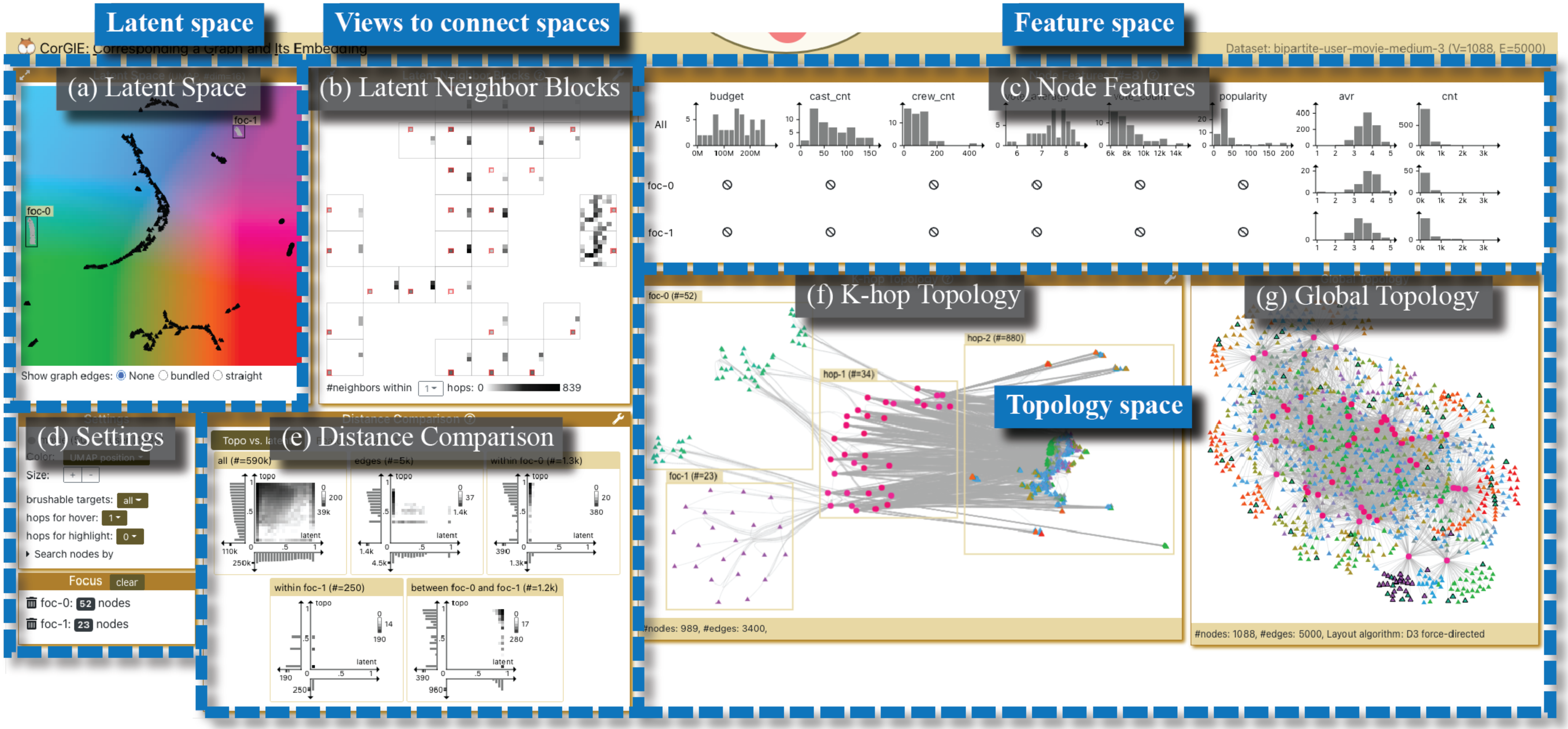


CorGIE multi-view interactive interface





CorGIE multi-view interactive interface



CorGIE: Visual Assessment of ML Training Completion & Quality

- Addresses **where are we?**
  - Visually explore correspondences between input graph and node embedding to show **what's here?**
- Addresses are **we there yet?**
  - Has the GNN training process captured all expected data about k-hop neighborhoods in the input graph, or should we keep going with train/tune?
- Addresses **are we lost?**
  - Are the GNN predictions high quality or low quality?

Questions in road trips - and visualization in data science!

- one VDS project for each question

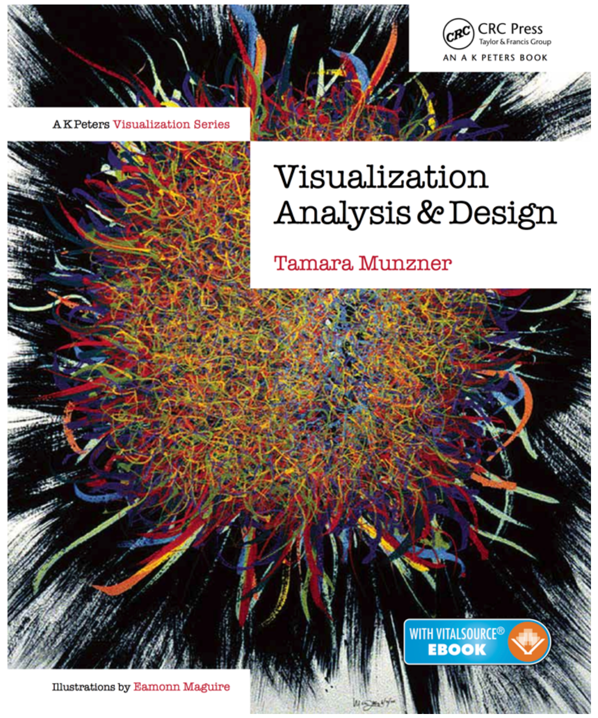
- where are we?
  - Data Reconnaissance & Task Wrangling
- what's here?
  - Automatic Encodings through Recommendation
- are we there yet? are we lost?
  - Visual Assessment of ML Training Completion

http://www.cs.ubc.ca/~tmm/talks.html#vds23

https://unsplash.com/photos/2oYHfuRe4OU

More information

- this talk
  http://www.cs.ubc.ca/~tmm/talks.html#vds23
- full courses, papers, videos, software, talks
  http://www.cs.ubc.ca/group/infovis
  http://www.cs.ubc.ca/~tmm
- book
  http://www.cs.ubc.ca/~tmm/vadbook
- VIS23 book table from CRC/Routledge
  – physical table
  – virtual bookshop: https://bit.ly/IEEEVIS23



Visualization Analysis and Design. Munzner.
CRC Press, AK Peters Visualization Series, 2014.

@tamara@vis.social

@tamaramunzner