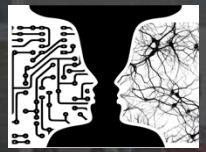
Département d'informatique
et de génie logiciel

Bridging Humans and Data: AI-Supported Interactive Visualizations

Jiayi Hong
01/20/2026

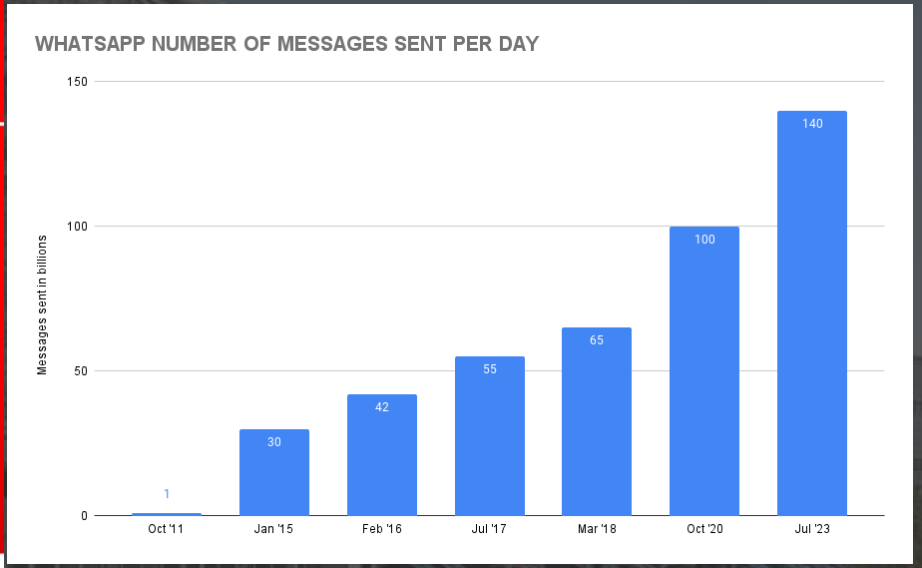


GRAAL

LVSN

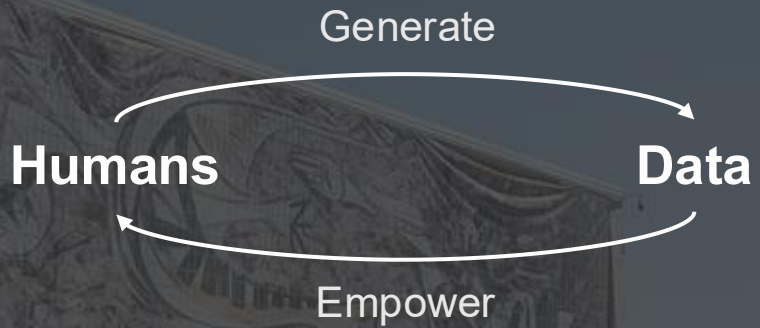

DATA

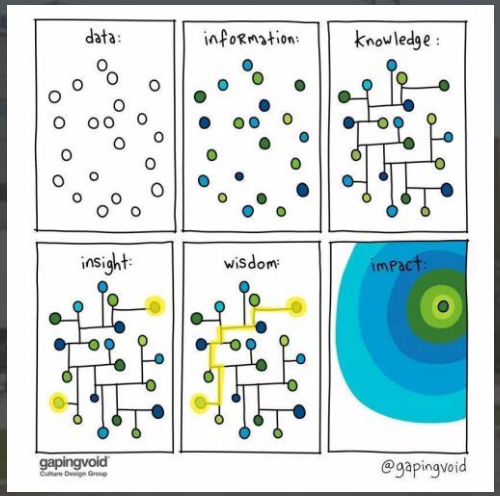
Fun Facts



https://www.sellcell.com/blog/how-many-text-messages-are-sent-a-day-2023-statistics/

Humans and Data





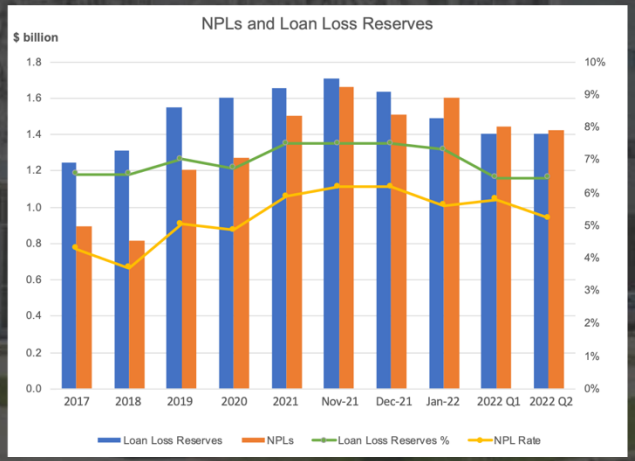

How can we empower **Humans** to effectively
understand and use **Data** within AI Applications?

Humans and Data

Humans → Data

Empower

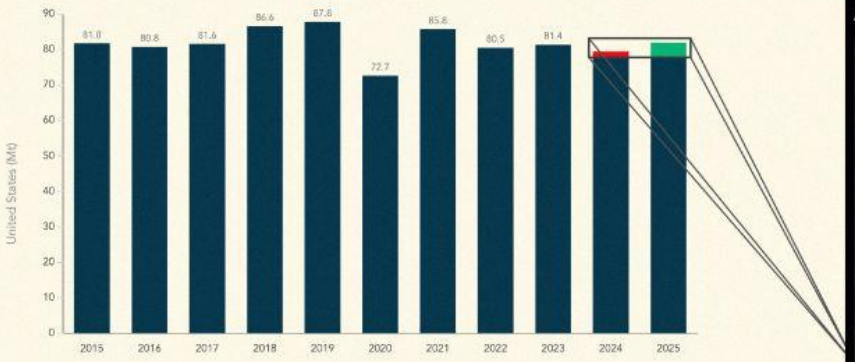
Interactive Visualization



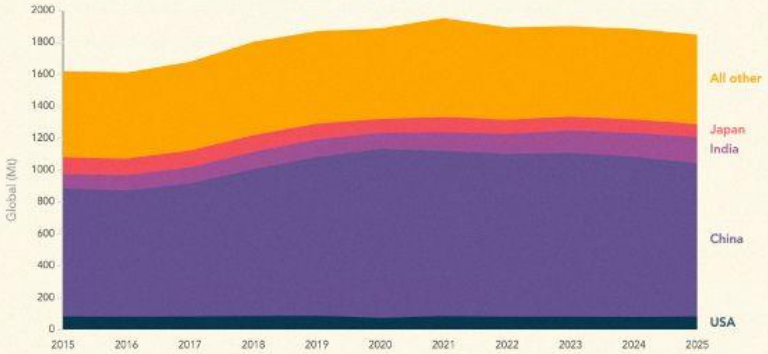
Steel Production, 2015 to 2025

Source: World Steel Association (except for USA's 2024 and 2025 data, which is sourced from the White House tweet below)

United States Steel Production By Year



Global Steel Production By Year



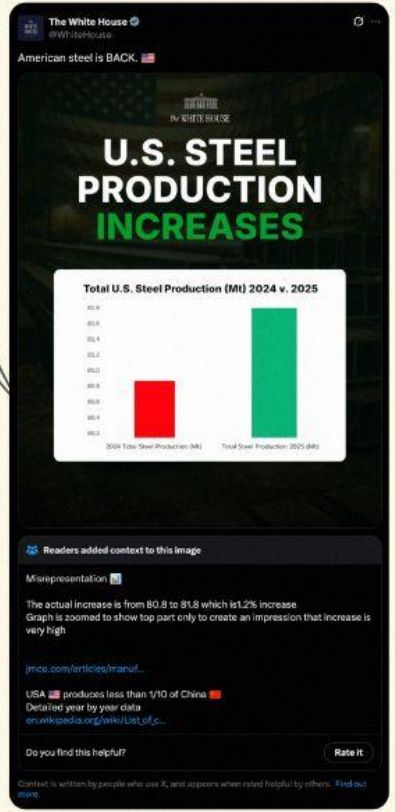


chart by shri (@shrikhalpada.dev)

Humans and Data

Humans

Data

Empower

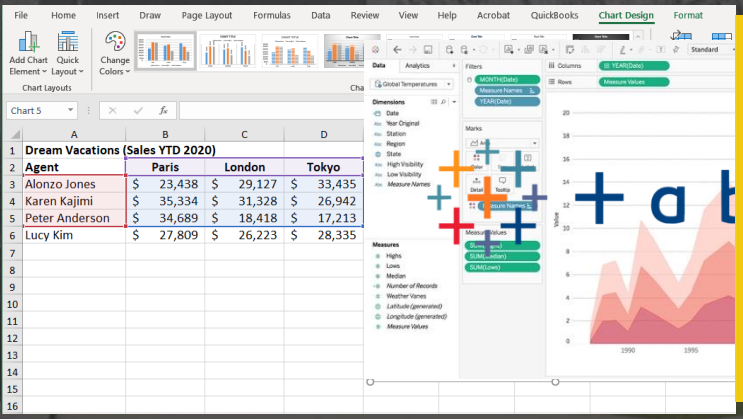
Interactive Visualization

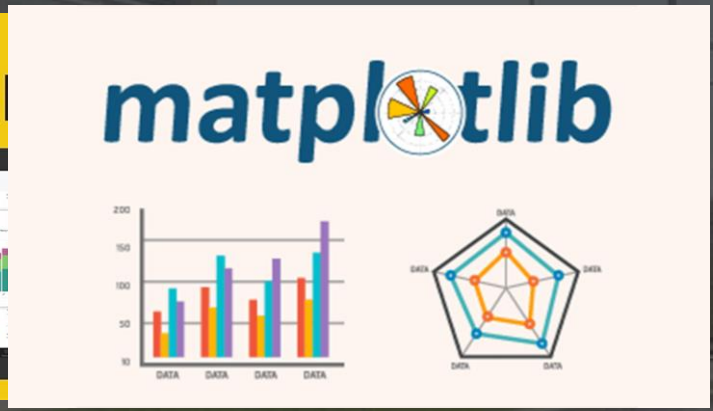
Humans and Data

Humans → Data

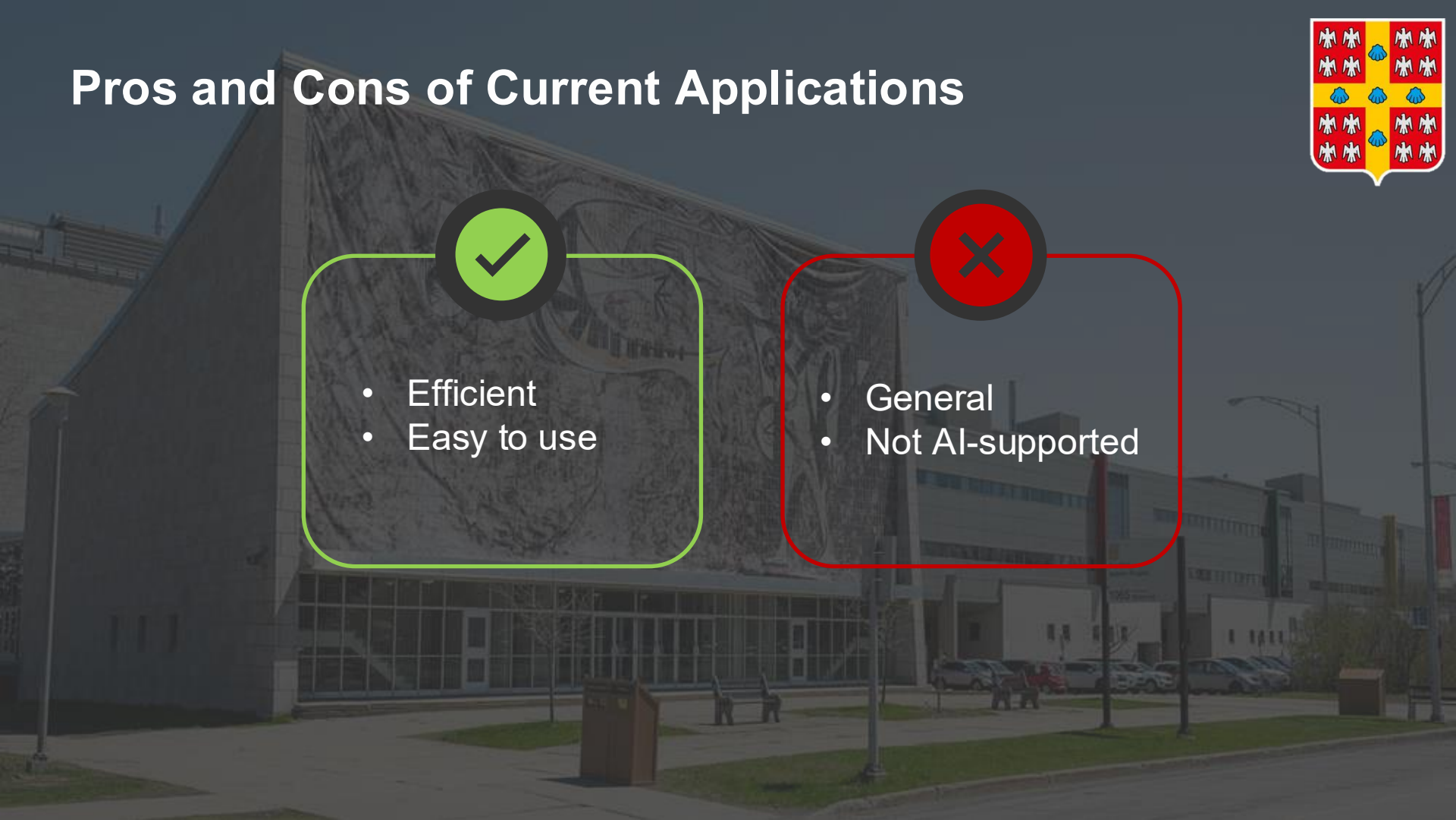
Empower

Interactive Visualization





Pros and Cons of Current Applications



- Efficient
- Easy to use

- General
- Not AI-supported

CLAUDE 3

GEMINI

GPT-4

Human-AI teaming in visualization systems

Received 5 April 2023, accepted 3 May 2023, date of publication 8 May 2023, date of current versi
Digital Object Identifier 10.1109/ACCESS.2023.3274199

RESEARCH ARTICLE

Chat2VIS: Generating Data Visualiza Natural Language Using ChatGPT, C GPT-3 Large Language Models

PAULA MADDIGAN AND TEO SUSNJAK
School of Mathematical and Computational Sciences, Massey University, Auckland 0632, New Zealand
Corresponding author: Teo Susnjak (t.susnjak@massey.ac.nz)

ABSTRACT The field of data visualisation has long aimed to devise soluti directly from natural language text. Research in Natural Language Interfaces the development of such techniques. However, the implementation of w challenging due to the inherent ambiguity of natural language, as well as poorly written user queries which pose problems for existing language m Instead of pursuing the usual path of developing new iterations of langua proposes leveraging the advancements in pre-trained large language model GPT-3 to convert free-form natural language directly into code for approp presents a novel system, Chat2VIS, which takes advantage of the capabili how, with effective prompt engineering, the complex problem of language u efficiently, resulting in simpler and more accurate end-to-end solutions th shows that LLMs together with the proposed prompts offer a reliable appr from natural language queries, even when queries are highly misspecified a also presents a significant reduction in costs for the development of NLI visualisation inference abilities compared to traditional NLP approaches rules and tailored models. This study also presents how LLM prompts ca preserves data security and privacy while being generalisable to different d performance of GPT-3, Codex and ChatGPT across several case studies and prior studies.

ChartQA: A Benchmark for Question Answering with Visual and Logical Reasoning

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty
York University, Canada
Nanyang Technological University, Singapore, Salesfo
{masry20, enamulh}@yorku.ca
{xuanlong001@e.ntu, C190022@e.ntu, srjoty@ntu

Abstract

Charts are very popular for analyzing data. When exploring charts, people often ask a variety of complex reasoning questions that involve several logical and arithmetic operations. They also commonly refer to visual features of a chart in their questions. However, most existing datasets do not focus on such complex reasoning questions as their questions are template-based and answers come from a fixed-vocabulary. In this work, we present a large-scale benchmark covering 9.6K human-written questions as well as 23.1K questions generated from human-written chart summaries. To address the unique challenges in our benchmark involving visual and logical reasoning over charts, we present two transformer-based models that combine visual features and the data table of the chart in a unified way to answer questions. While our models achieve the state-of-the-art results on the previous datasets as well as on our benchmark, the evaluation also reveals several challenges in answering complex reasoning questions.

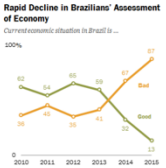


Figure 1: Sample qu

dicting the answer. Thi tasks such as QA on te and tables (Pasupat an input for ChartQA is a that can draw a reader' nent features such as tre 2020, 2021). Also, peo referring to visual attrib in Fig. 1, Q2 refers to and its attribute ('peak

While the task of Ch attentions in recent ye

Is Your Code Generated by ChatGPT Really Correct? Rigorous Evaluation of Large Language Models for Code Generation

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, Lingming Zhang
University of Illinois Urbana-Champaign, Nanjing University
{jiawei6, chunqiu2, lingming}@illinois.edu yuyao6@outlook.com

Abstract

Program synthesis has been long studied with recent approaches focused on directly using the power of Large Language Models (LLMs) to generate code. Programming benchmarks, with curated synthesis problems and test-cases, are used to measure the performance of various LLMs on code synthesis. However, these test-cases can be limited in both quantity and quality for fully assessing the functional correctness of the generated code. Such limitation in the existing benchmarks begs the following question: *In the era of LLMs, is the code generated really correct?* To answer this, we propose EvalPlus – a code synthesis evaluation framework to rigorously benchmark the functional correctness of LLM-synthesized code. EvalPlus augments a given evaluation dataset with large amounts of test-cases newly produced by an automatic test input generator, powered by both LLM- and mutation-based strategies. While EvalPlus is general, we extend the test-cases of the popular HUMANEVAL benchmark by 80× to build HUMANEVAL+. Our extensive evaluation across 26 popular LLMs (*e.g.,* GPT-4 and ChatGPT) demonstrates that HUMANEVAL+ is able to catch significant amounts of previously undetected wrong code synthesized by LLMs, reducing the pass@k by up-to 19.3-28.9%. We also surprisingly found that test insufficiency can lead to mis-ranking. For example, both WizardCoder-CodeLlama and Phind-CodeLlama now outperform ChatGPT on HUMANEVAL+, while none of them could on HUMANEVAL. Our work not only indicates that prior popular code synthesis evaluation results do not accurately reflect the true performance of LLMs for code synthesis, but also opens up a new direction to improve such programming benchmarks through automated testing. We have open-sourced our tools, enhanced datasets as well as all LLM-generated code at https://github.com/evalplus/evalplus to facilitate and accelerate future LLM-for-code research.

Human-AI teaming in visualization systems

How can we **evaluate** these visualizations?

LLMs

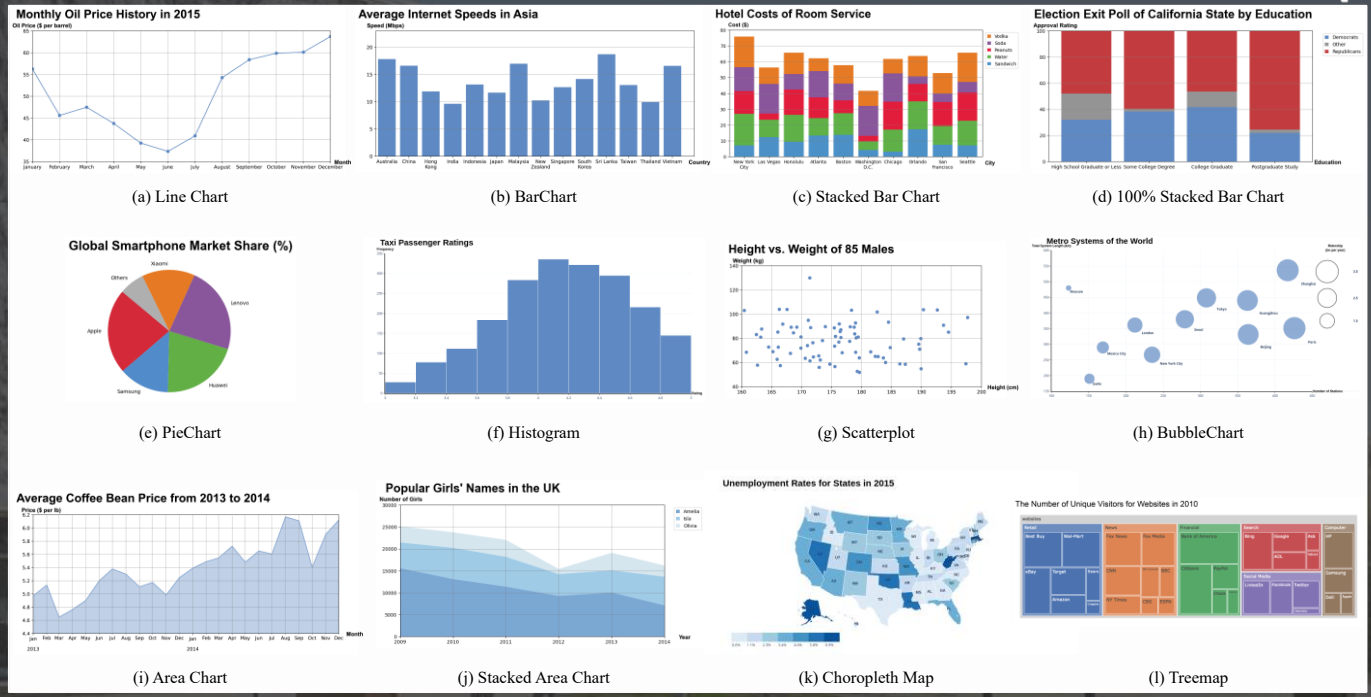
How effectively do LLMs interpret charts?

- What are LLMs' performance compared with human?
- Do LLMs answer chart-related questions based on charts or pre-obtained knowledge?

LLMs' Visualization Literacy

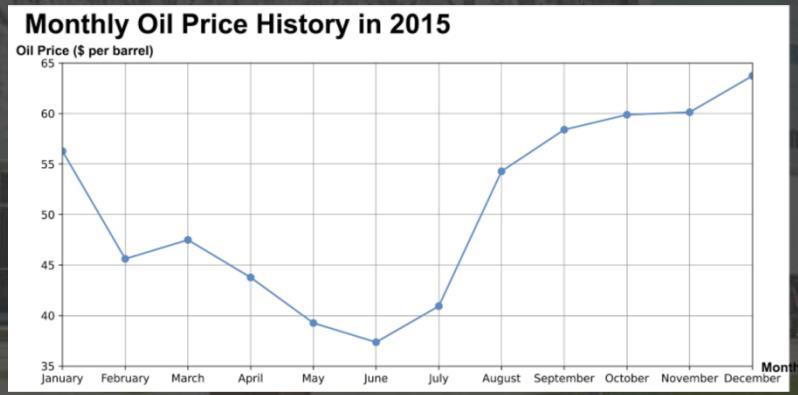
ChatGPT

Gemini



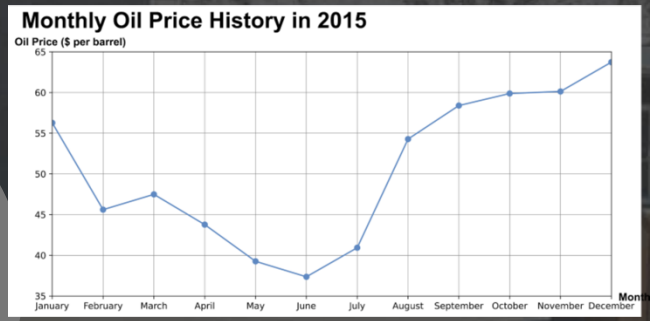
Factors

- **Visualization**
  - Whether LLMs would be influenced by pre-learned knowledge



Factors

- **Visualization**
  - Whether LLMs would be influenced by pre-learned knowledge
  - LLMs' performance in answering open questions
- **Answer choices**



What was the price of a barrel of oil in February 2015?
a. $58.31 b. $45.61 c. $50.28 d. $54.67 e. Omit

Results

Visualization	Task	Stem	VLAT	GPT-4	Gemini	Random
Line Chart	Retrieve Value	What was the price of a barrel of oil in February 2015?	0.95	0.56	0.25	0.25
	Find Extremum	In which month was the price of a barrel of oil the lowest in 2015?	0.97	0.02	0.22	0.25
	Determine Range	What was the price range of a barrel of oil in 2015?	0.56	0.23	0.23	0.25
	Find Correlation/Trend	Over the course of the second half of 2015, the price of a barrel of oil was _____.	0.98	0.87	0.42	0.33
	Make Comparisons	About how much did the price of a barrel of oil rise from June to September in 2015?	0.77	0.87	0.28	0.25
Bar Chart	Retrieve Value	What is the average internet speed in Japan?	0.88	0.40	0.56	0.25
	Find Extremum	In which country is the average internet speed the fastest in Asia?	0.98	0.00	0.01	0.25
	Determine Range	What is the range of the average internet speed in Asia?	0.54	0.29	0.68	0.25
	Make Comparisons	How many countries in Asia is the average Internet speed slower than South Korea?	0.40	0.20	0.13	0.25
Stacked Bar Chart	Retrieve Value (Absolute Value)	What is the cost of peanuts in Las Vegas?	0.38	0.23	0.25	0.25
	Retrieve Value (Relative Value)	About what is the ratio of the cost of a sandwich to the total cost of room service in Seattle?	0.36	0.04	0.23	0.25
	Find Extremum	In which city is the cost of soda the highest?	0.69	0.17	0.05	0.25
	Make Comparisons (Absolute Value)	The cost of water in Boston is higher than that of New York City.	0.59	0.70	0.78	0.50
	Make Comparisons (Relative Value)	The ratio of the cost of peanuts to the cost of water in Las Vegas is higher than that of San Francisco.	0.47	0.94	0.43	0.50
100% Stacked Bar Chart	Retrieve Value (Absolute Value)	What is the approval rating of Republicans among the people who have the education level of Postgraduate Study?	0.49	0.00	0.79	0.25
	Find Extremum (Relative Value)	What is the education level of people in which the Democrats have the lowest approval rating?	0.90	0.00	0.21	0.25
	Make Comparisons (Relative Value)	The approval rating of Republicans for the people who have the education level of Some College Degree is lower than that for the people who have the education level of Postgraduate Study.	0.54	0.15	0.98	0.50
Pie Chart	Retrieve Value (Relative Value)	About what is the global smartphone market share of Huawei?	0.72	0.00	0.36	0.25
	Find Extremum (Relative Value)	In which company is the global smartphone market share the smallest?	0.98	0.00	0.30	0.25
	Make Comparisons (Relative Value)	The global smartphone market share of Lenovo is larger than that of Samsung.	1.00	0.00	0.43	0.50
Histogram	Retrieve Value (Derived Value)	How many people have rated the taxi between 4.0 and 4.2?	0.84	0.38	0.18	0.25
	Find Extremum (Derived Value)	What is the rating that the people have rated the taxi the most?	0.94	0.06	0.13	0.25
	Make Comparisons (Derived Value)	More people have rated the taxi between 4.6 and 4.8 than between 4.2 and 4.4.	0.86	0.89	0.00	0.50
		rson who is 165.1 cm tall?	0.85	0.20	0.49	0.25
		est person among the 85 males?	0.76	0.85	1.00	0.25
		or the 85 males?	0.53	0.86	0.76	0.25
		n who lies outside the others the most?	0.42	0.00	0.00	0.25
		d around the height of 176 cm and the weight of 70 kg.	0.90	0.06	0.50	0.50
		tionship between the height and the weight of the 85 males.	0.52	1.00	0.00	0.50
		e height of 188 cm are all the same.	0.79	1.00	1.00	0.50
		f a pound of coffee beans in June 2013?	0.75	0.28	0.13	0.25
		f a pound of coffee beans at minimum?	0.44	0.57	0.73	0.25
		verage price of a pound of coffee beans between January 2013 and	0.38	0.23	0.13	0.25
		average price of a pound of coffee beans was _____.	0.94	1.00	1.00	0.33
		named 'Amelia' in 2010 in the UK?	0.15	0.36	0.35	0.25
		e number of girls named 'Amelia' to those named 'Isla' in 2014 in the	0.25	0.17	0.30	0.25
		een 2009 and 2014, when was the number of girls named 'Amelia' at	0.97	0.00	0.01	0.25
		sla' was _____ from 2009 to 2012.	0.96	0.00	0.72	0.33
		s named 'Amelia' in 2014 was more than it was in 2013.	0.20	0.51	0.99	0.50
		een 2009 and 2014, the number of girls named 'Isla' was always more	0.24	1.00	1.00	0.50
		metro system in Beijing?	0.41	0.02	0.15	0.25
		s the largest number of stations?	0.69	0.00	0.11	0.25
		length of the metro systems?	0.29	0.25	0.26	0.25
		es lie outside the relationship between the total system length and the	0.53	0.53	0.06	0.25
		s of the world has approximately 200 stations and around a 200 km	0.59	0.38	0.48	0.50
		e metro system increases as the number of stations increases.	0.26	0.93	1.00	0.50
		s more ridership than the metro system in New York City.	0.80	0.00	0.09	0.50
		rate for Indiana (IN) in 2015?	0.24	0.00	0.00	0.25
		loyment rate the highest in 2015?	0.97	0.93	0.72	0.25
	Make Comparison (Approximate Value)	In 2015, the unemployment rate for Arizona (AZ) was higher than that of Oklahoma (OK).	0.92	0.80	0.00	0.50
Treemap	Find Extremum (Relative Value)	For which website was the number of unique visitors the largest in 2010?	0.68	0.01	0.00	0.25
	Make Comparison (Relative Value)	The number of unique visitors for Target was more than that of Ask in 2010.	0.42	0.03	0.53	0.50
	Identify the Hierarchical Structure	Amazon is nested in the Computer category.	0.92	1.00	0.00	0.50

Results

- LLMs are influenced by factual accuracy, or the “real truth.”
- Removing the context could potentially avoid the influence of pre-learned knowledge.
- LLMs are prone to overalignment for open questions.

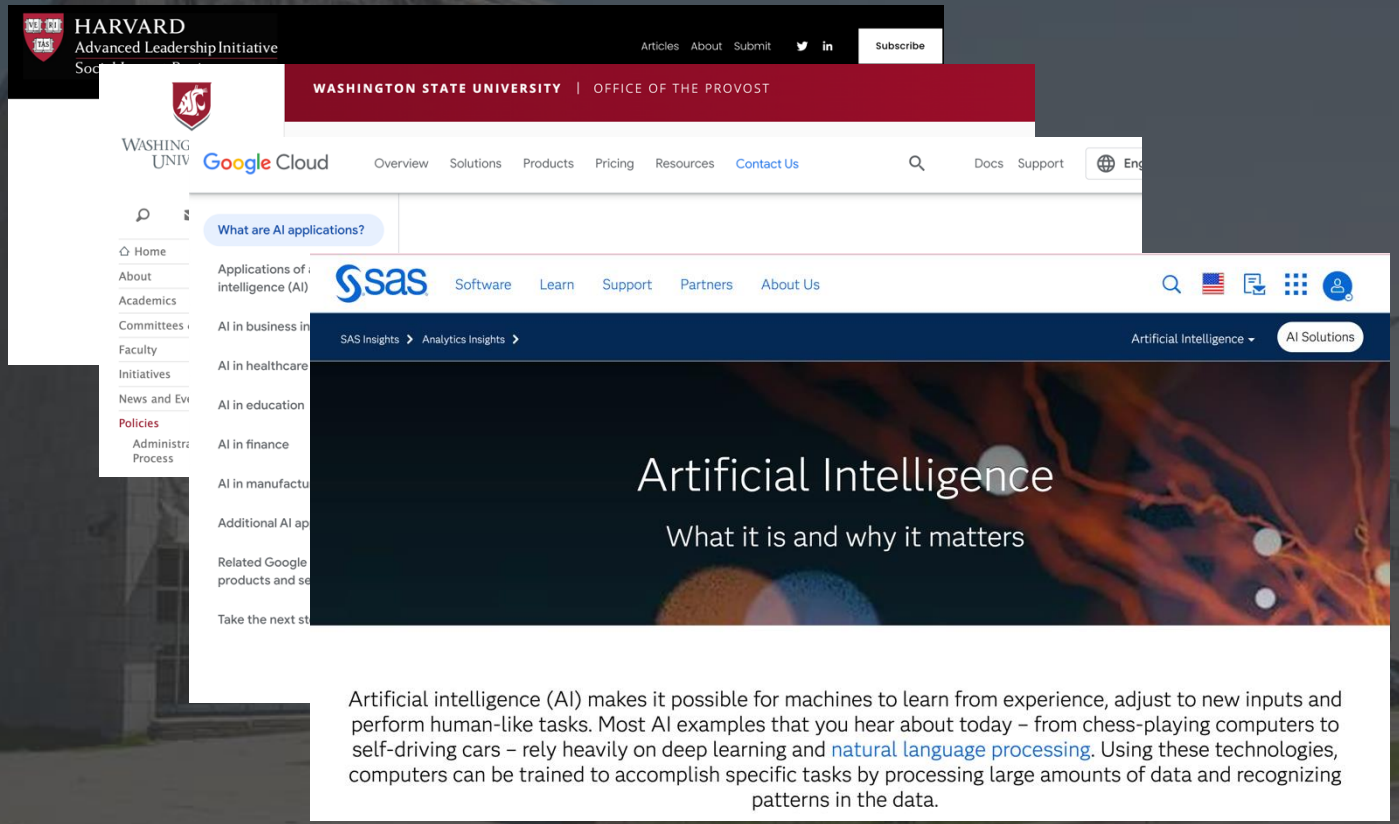
Takeaways

LLMs are not yet capable of fully replacing humans in interpreting visualizations or answering related questions.

However, they still hold potential for use in alternative applications.

(**Hong** et al., IEEE TVCG, 2025)

The Importance of AI Applications

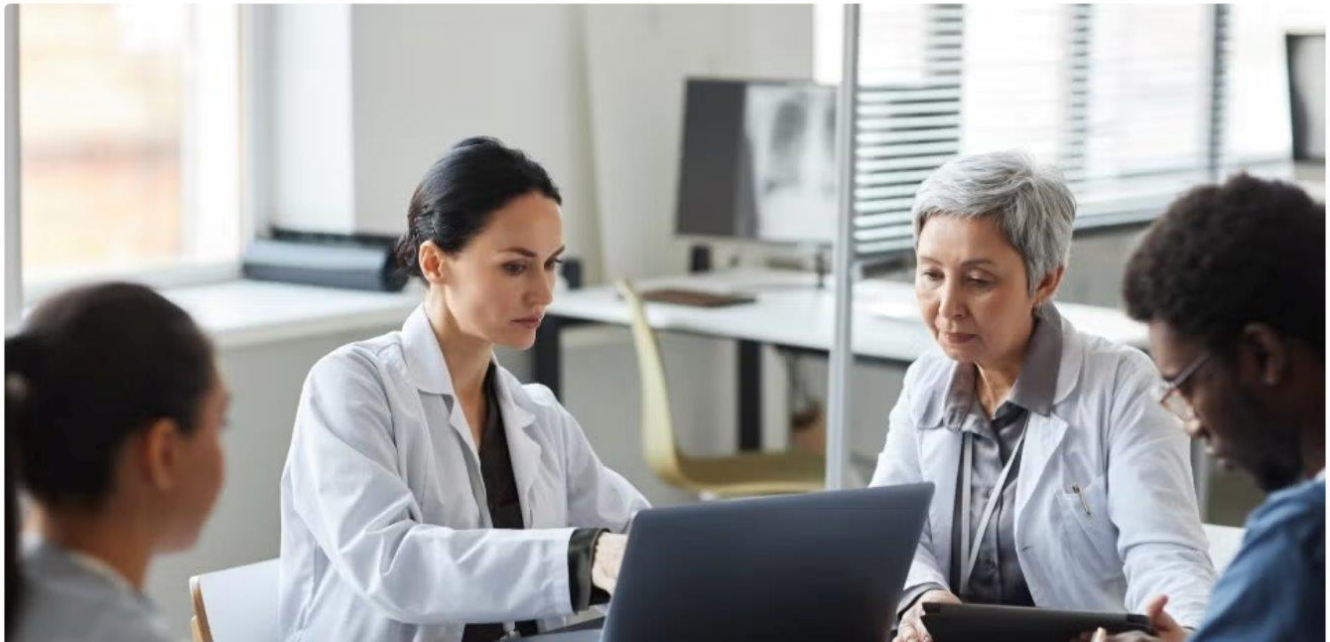


RESEARCH | 9 MAY 2025

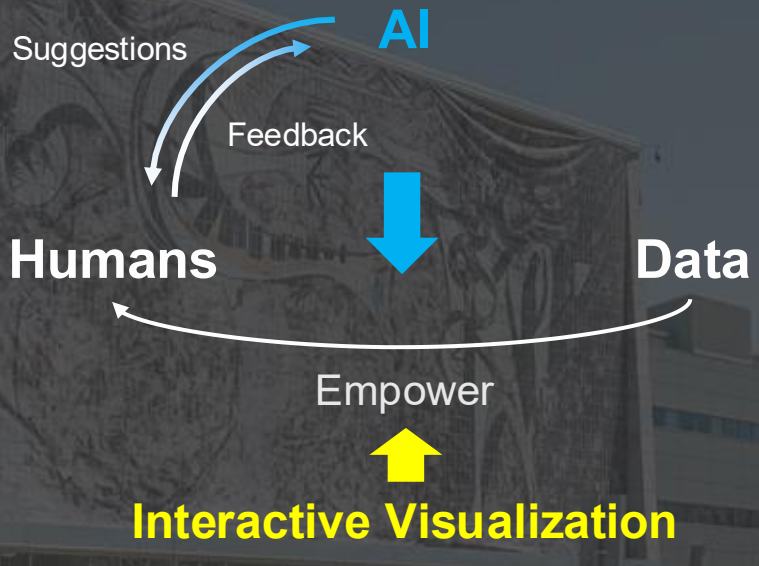
Generative AI Can Help Doctors Diagnose Patients — But Is it Biased?

A new study by Professor Damon Centola tested if AI tools could help improve medical care without increasing bias.

By Hailey Reissman

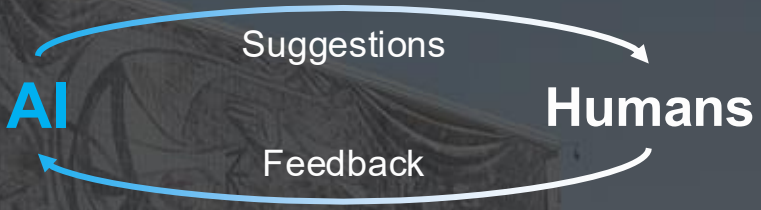


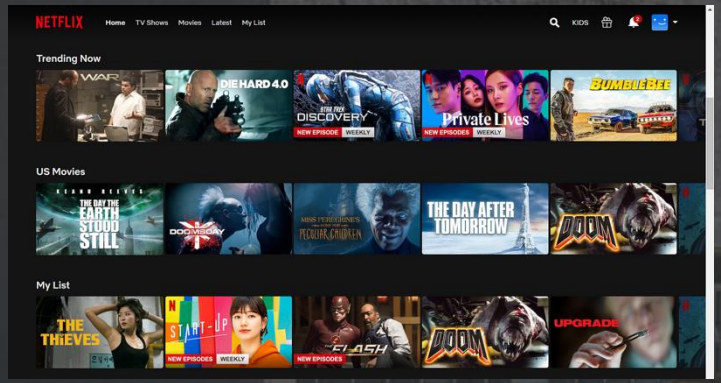
Role of AI

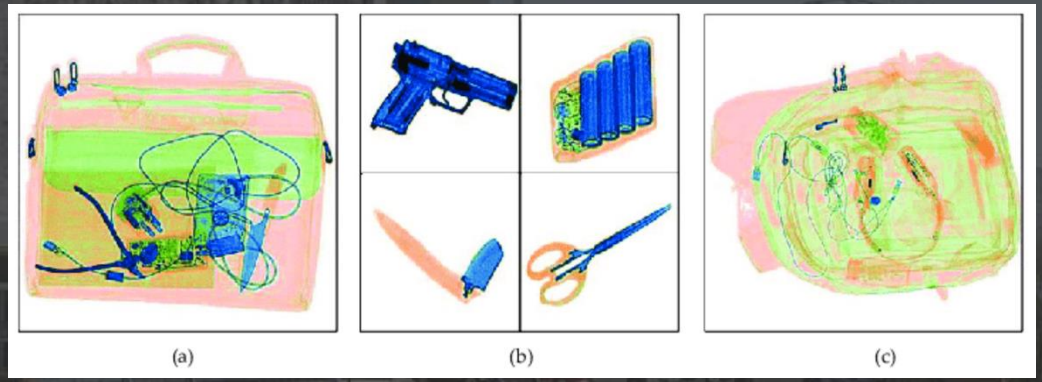

AI
Suggestions
Feedback
Humans
Data
Empower
Interactive Visualization

Human-AI Teaming

- Black Box
- Can make mistakes



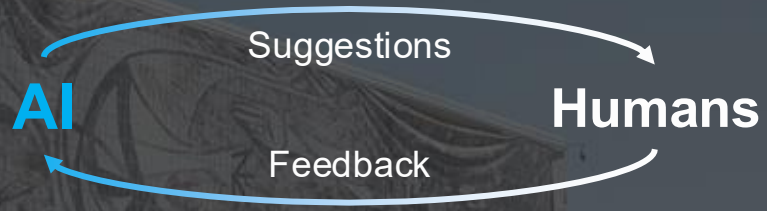




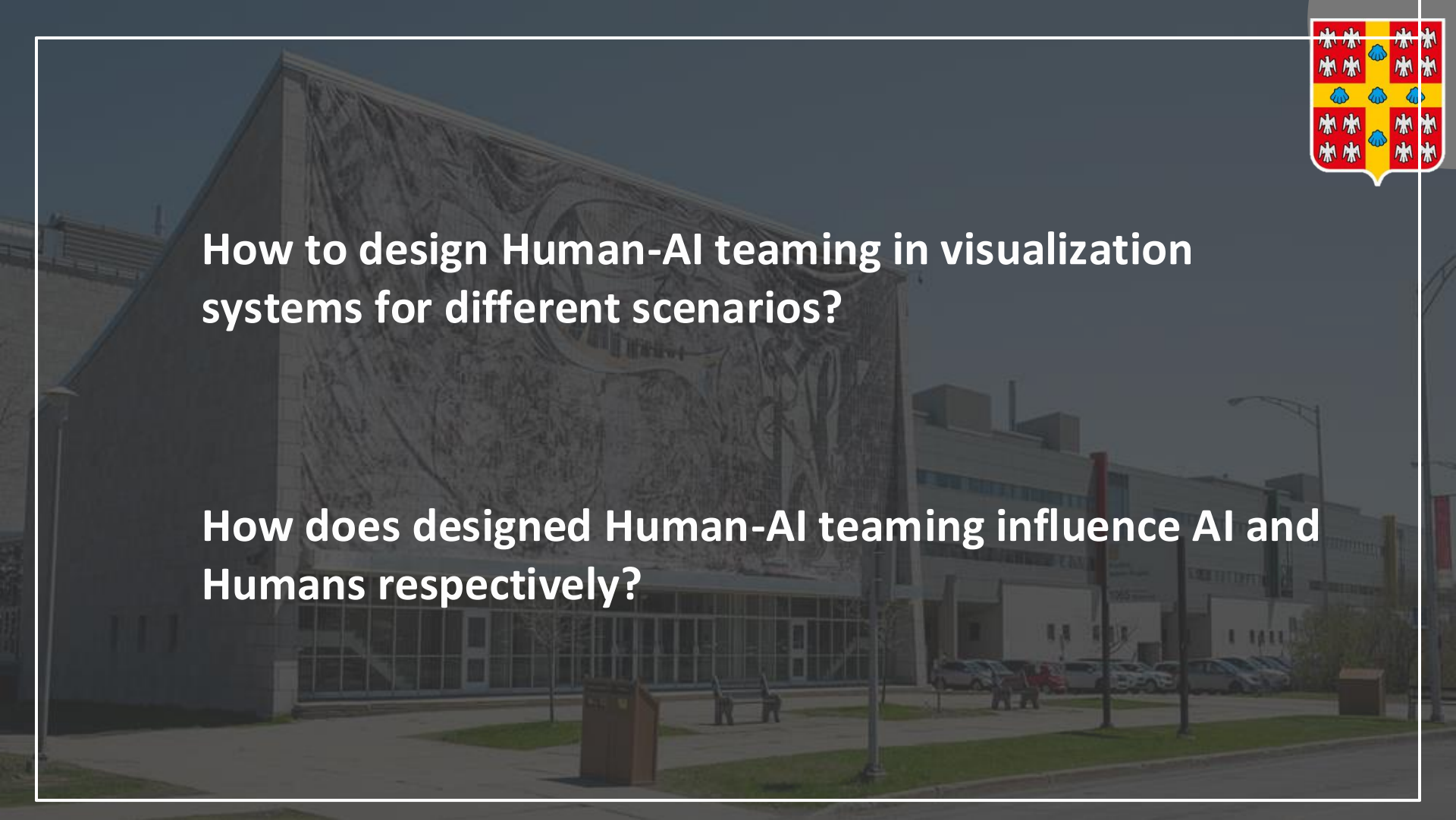
[Zhu et al. 2020]

Human-AI Teaming

- Black Box
- Can make mistakes



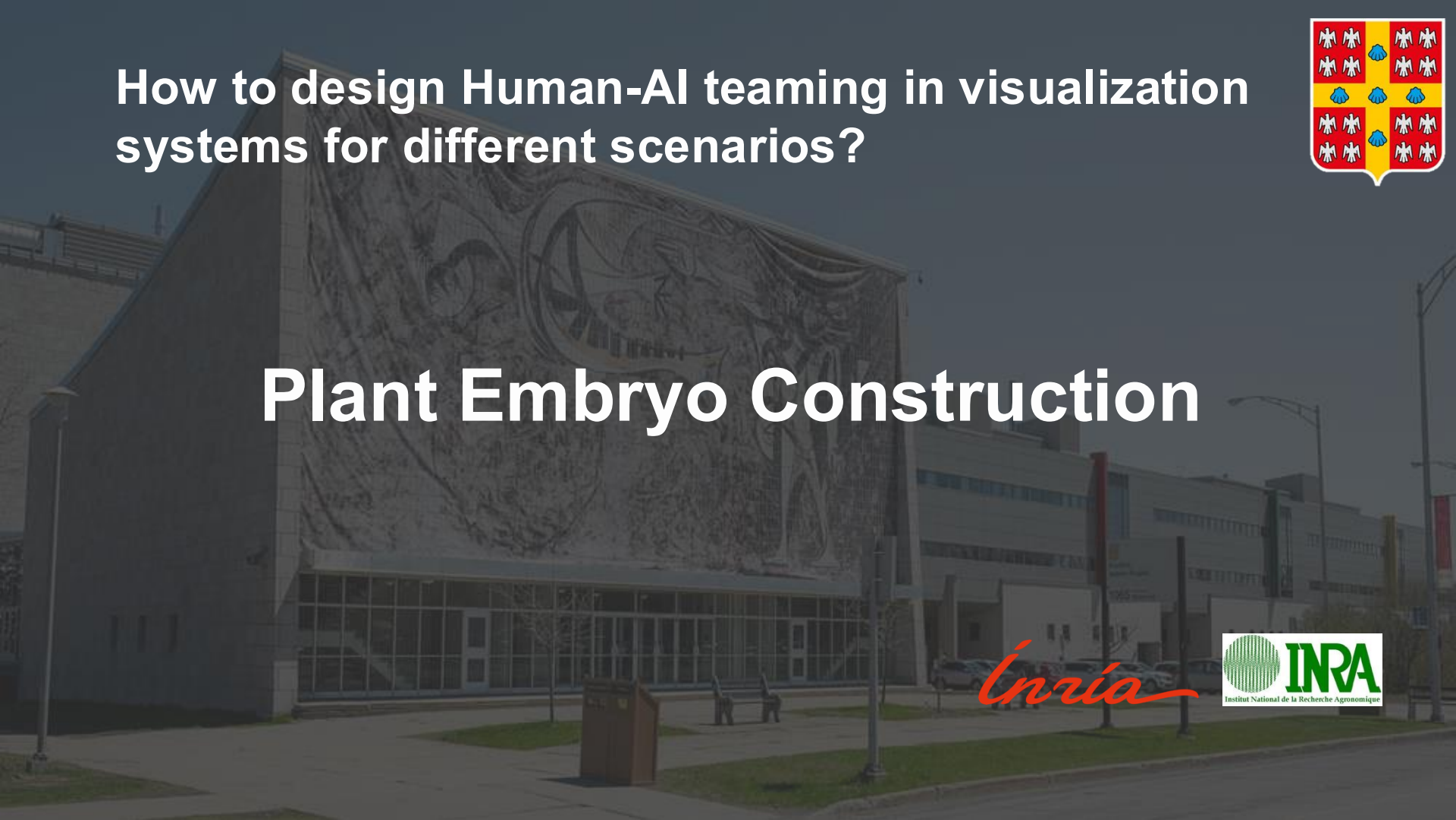
Different Design of Interactive Visualization



How to design Human-AI teaming in visualization systems for different scenarios?

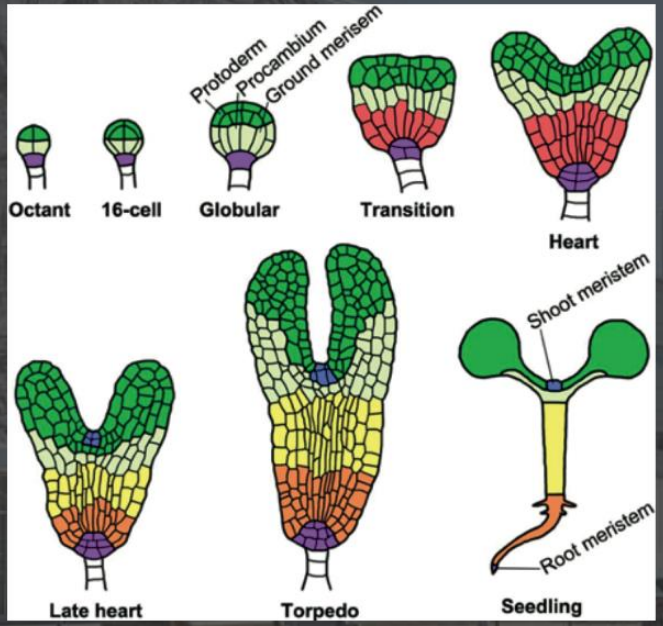
How does designed Human-AI teaming influence AI and Humans respectively?

How to design Human-AI teaming in visualization systems for different scenarios?



Plant Embryo Construction

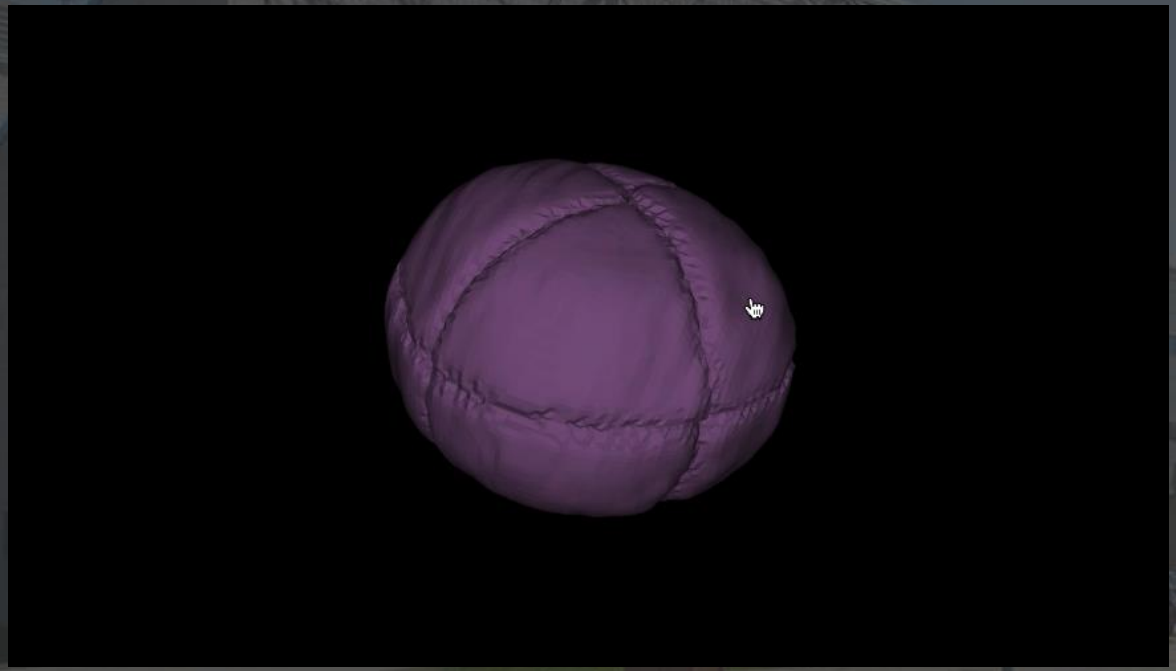
Plant Embryos

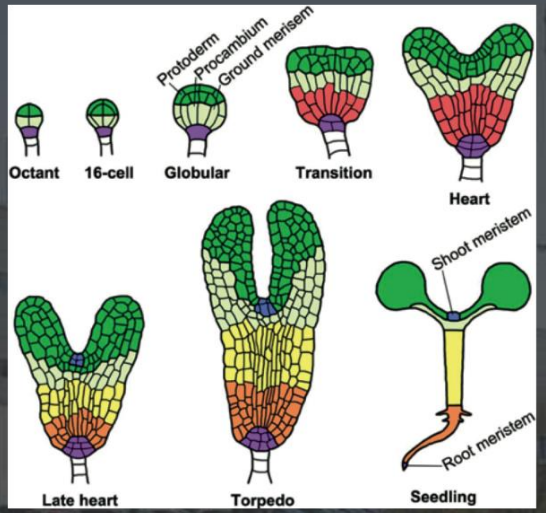


[Kim et al. 2017]

Plant Cell Lineage

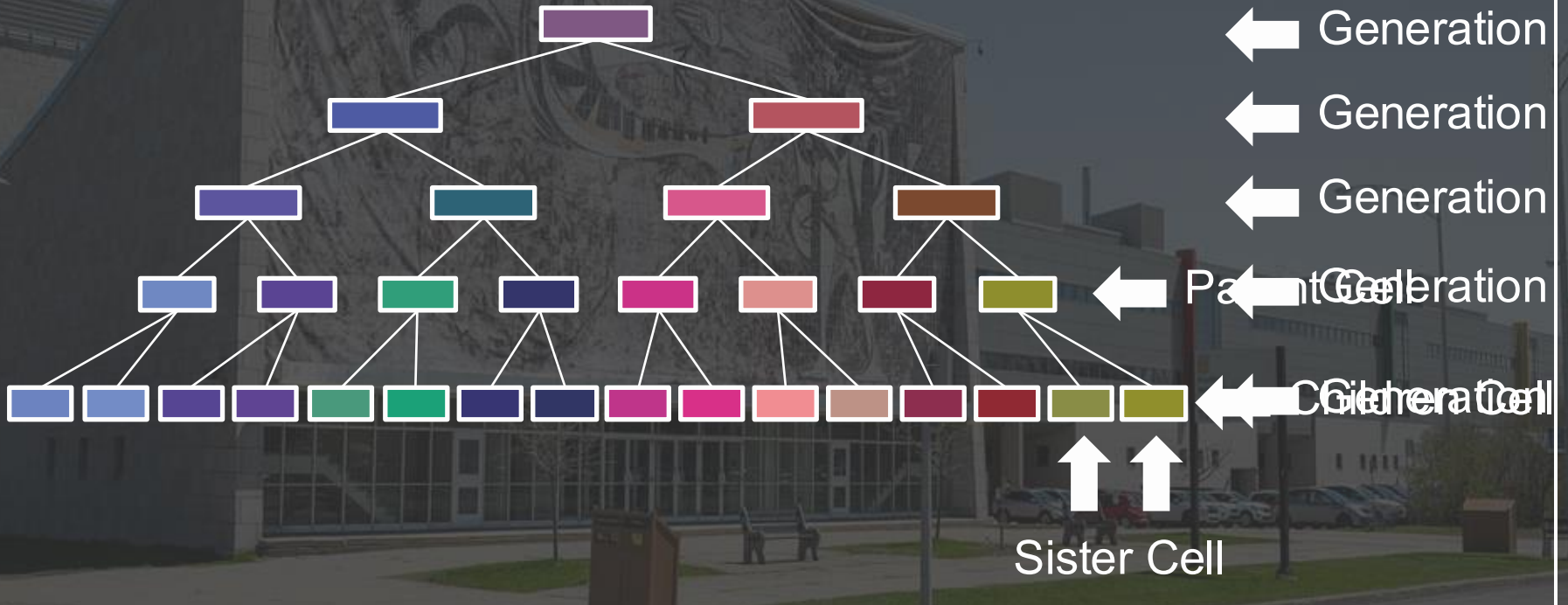
The development history of plant embryos



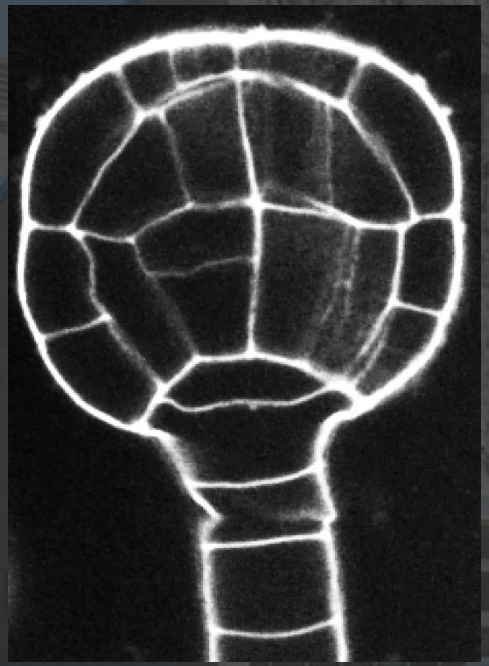


[Kim et al. 2017]

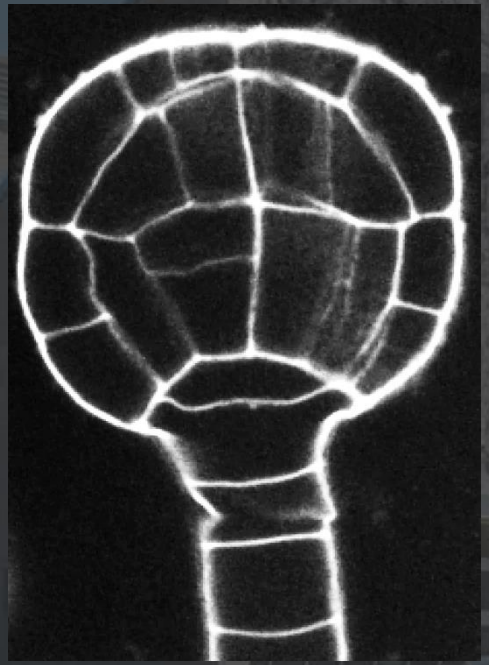
Plant Cell Lineage

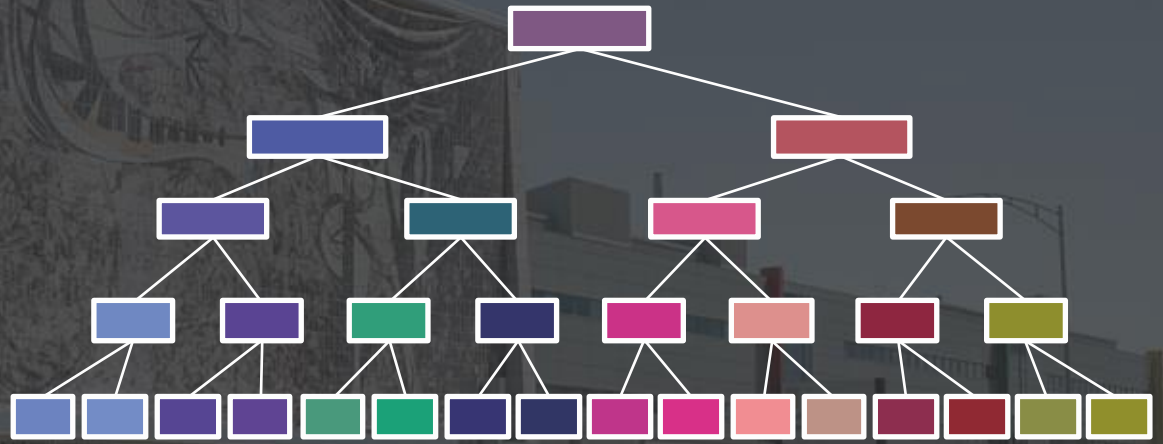


Biologists' Task

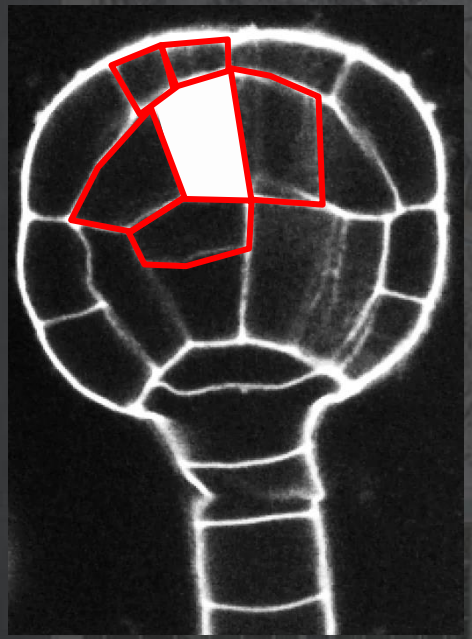


Biologists' Task



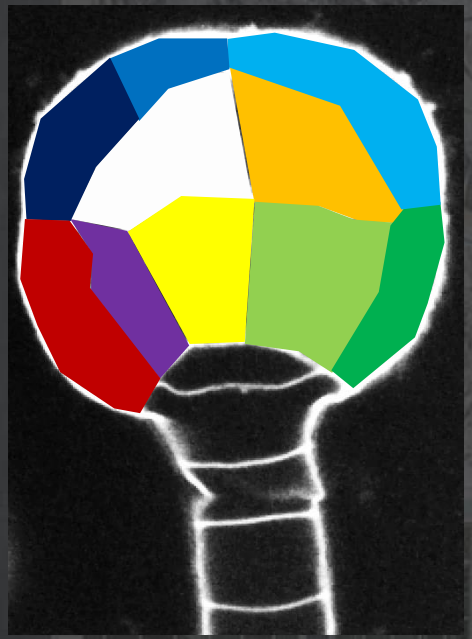


Current Approach



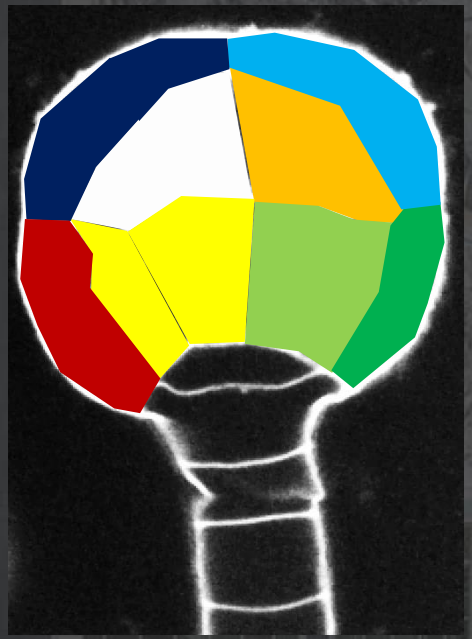
Biologists need to find the right sister cell for every cell in an embryo.

Current Approach



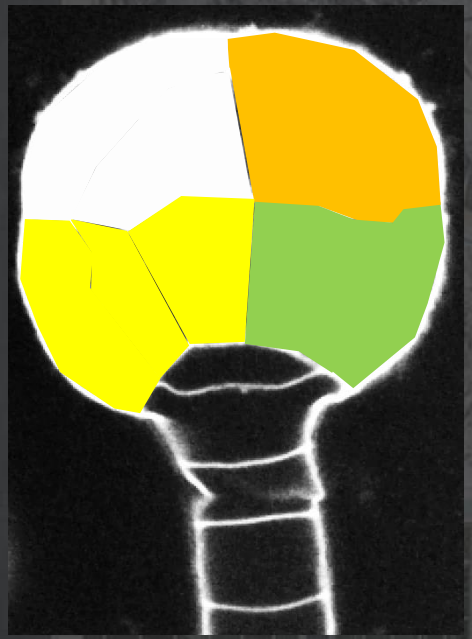
Once decided, they would merge cells and continue assigning the remaining cells.

Current Approach



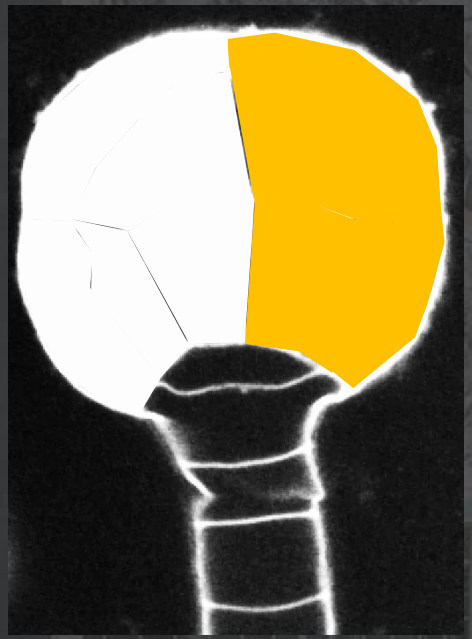
Biologists will continue this process to the new generation until there is only one cell left.

Current Approach



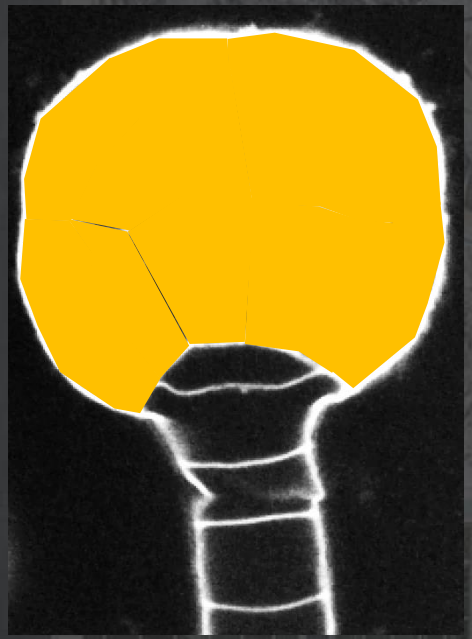
Biologists will continue this process to the new generation until there is only one cell left.

Current Approach



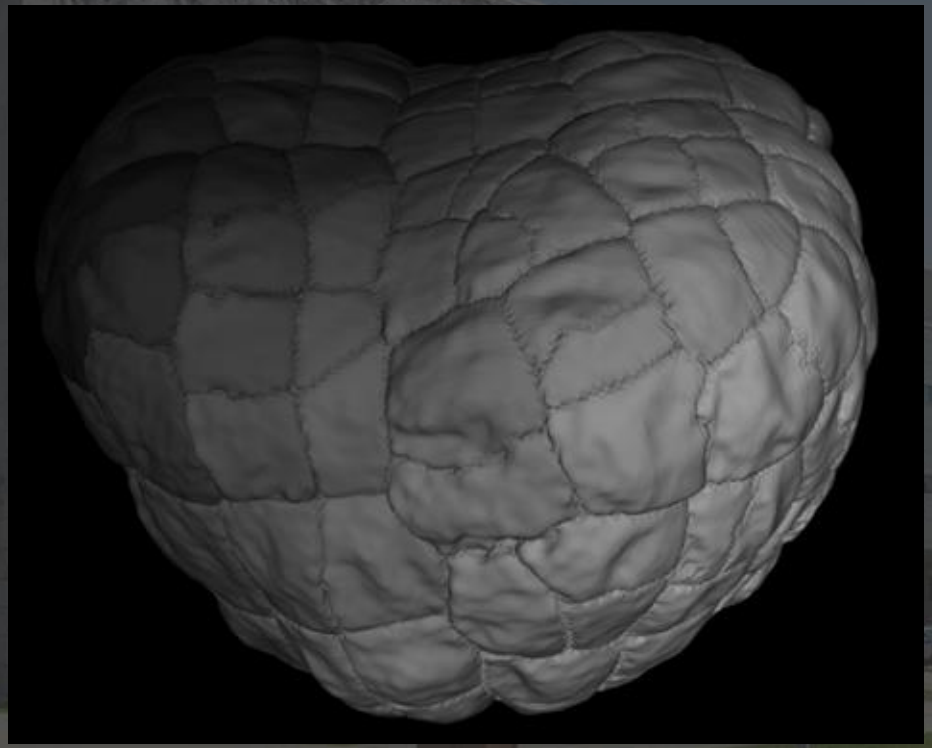
Biologists will continue this process to the new generation until there is only one cell left.

Current Approach



Biologists will continue this process to the new generation until there is only one cell left.

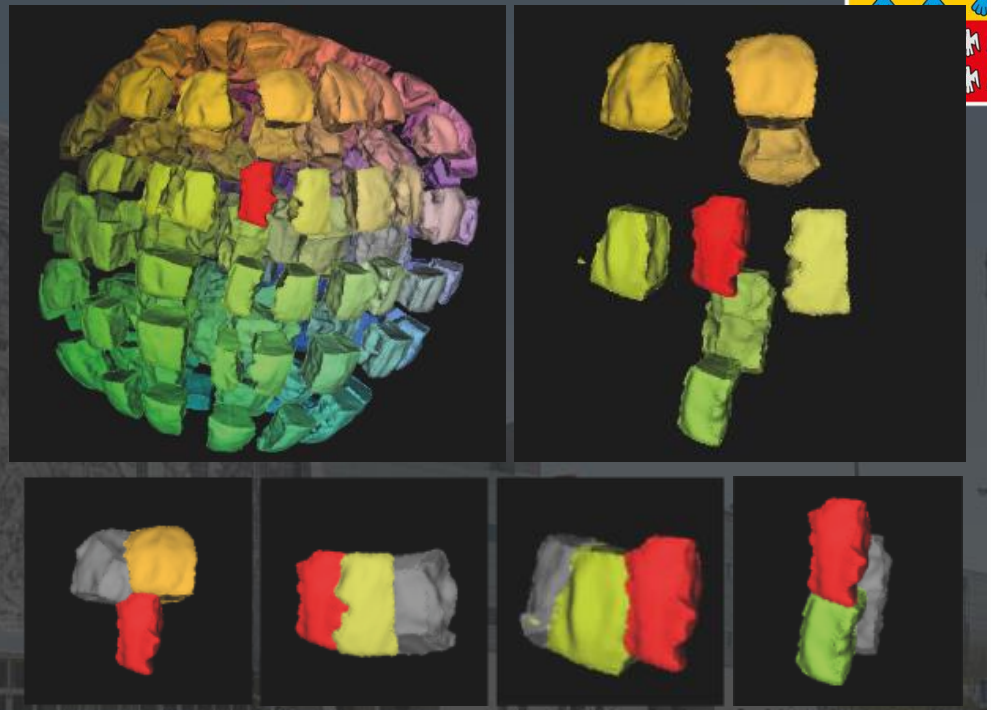
Why do biologists seek help?



8 Generations?
Or more?

Machine Learning

Neighboring cells:
Sisters (1) / Non-sisters (0)



(**Hong** et al., in Graphics Interface, 2021)

Binary Classification Problem

Datasets

93 Embryos

Data

Machine Learning

Assignments

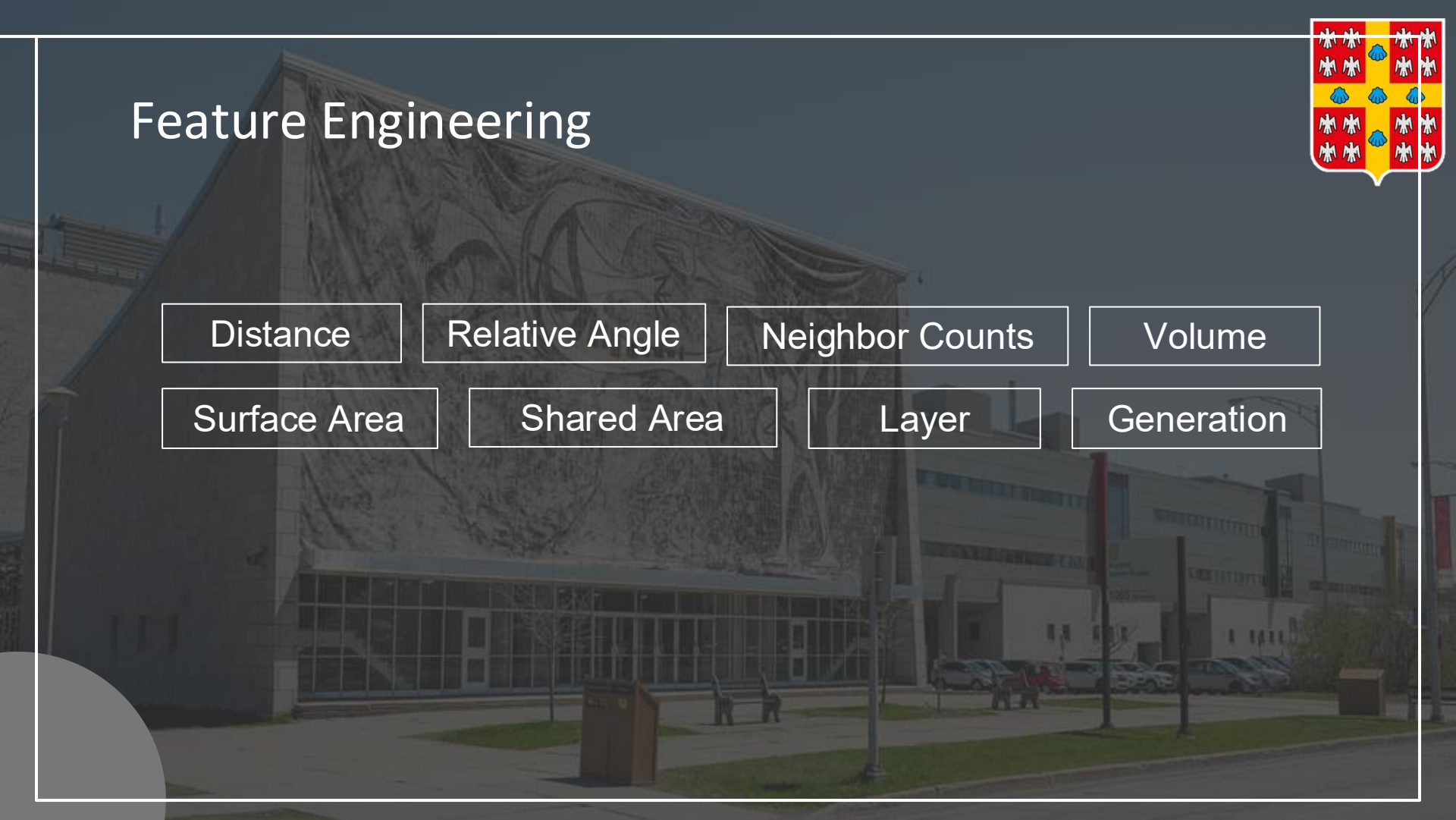
*"I know what **features** I usually use to make decisions!!"*

Data

Machine Learning

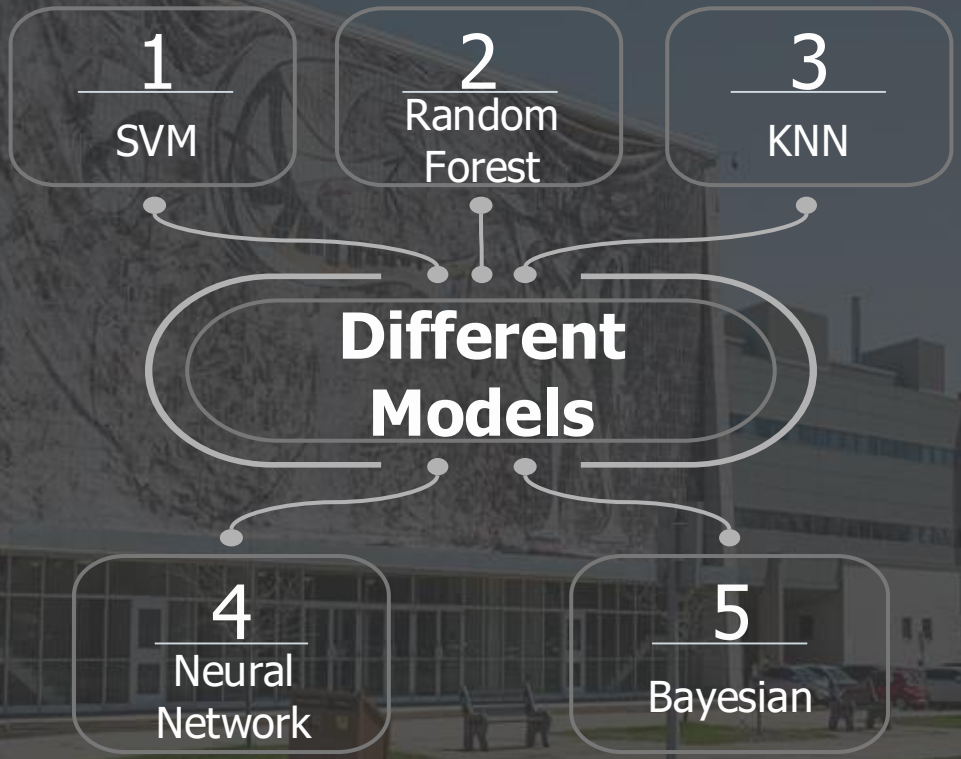
Assignments

Feature Engineering



- Distance
- Relative Angle
- Neighbor Counts
- Volume
- Surface Area
- Shared Area
- Layer
- Generation

Model Training



Model Training

1 SVM

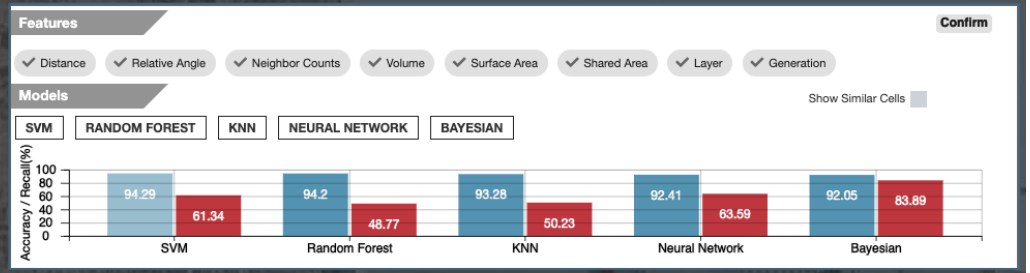
2 Random Forest

3 KNN

Different Models

4 Neural Network

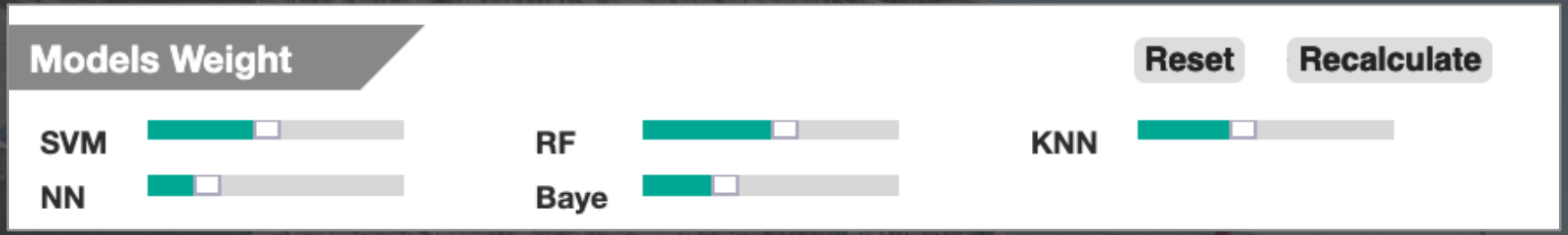
5 Bayesian

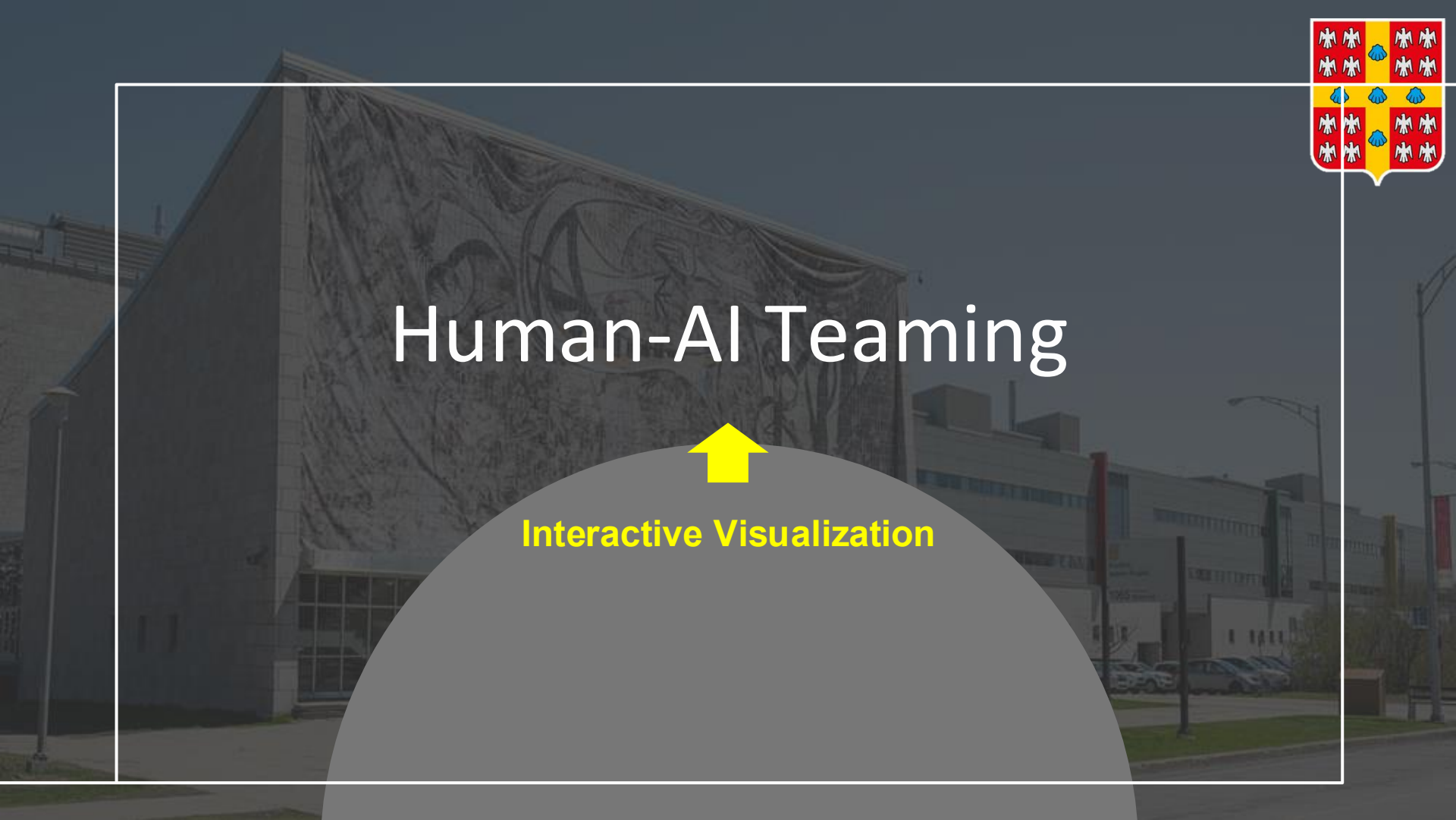


Accuracy

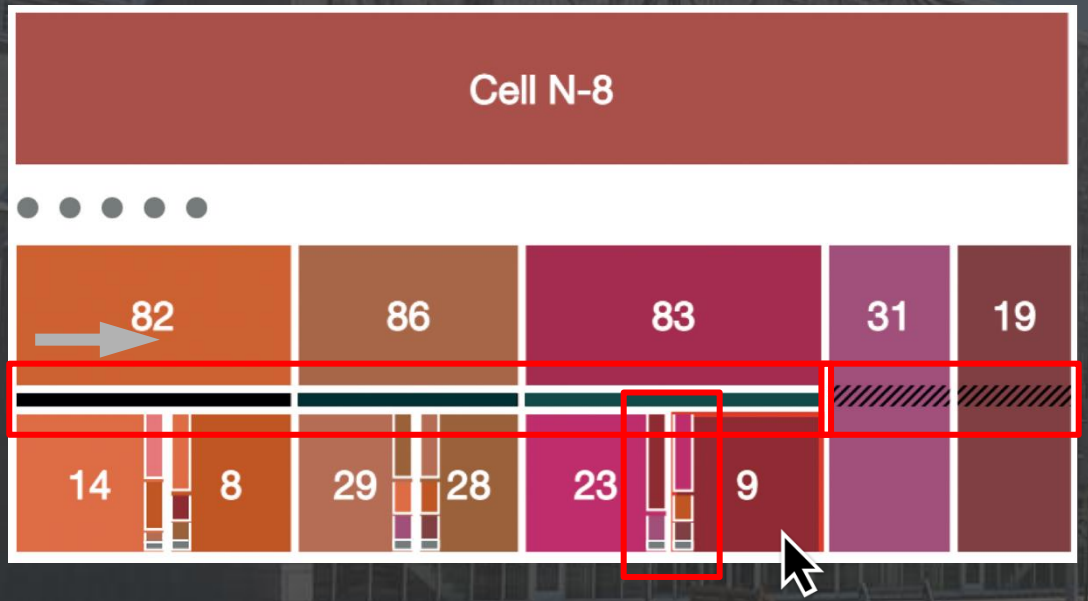
Recall

Models Weight



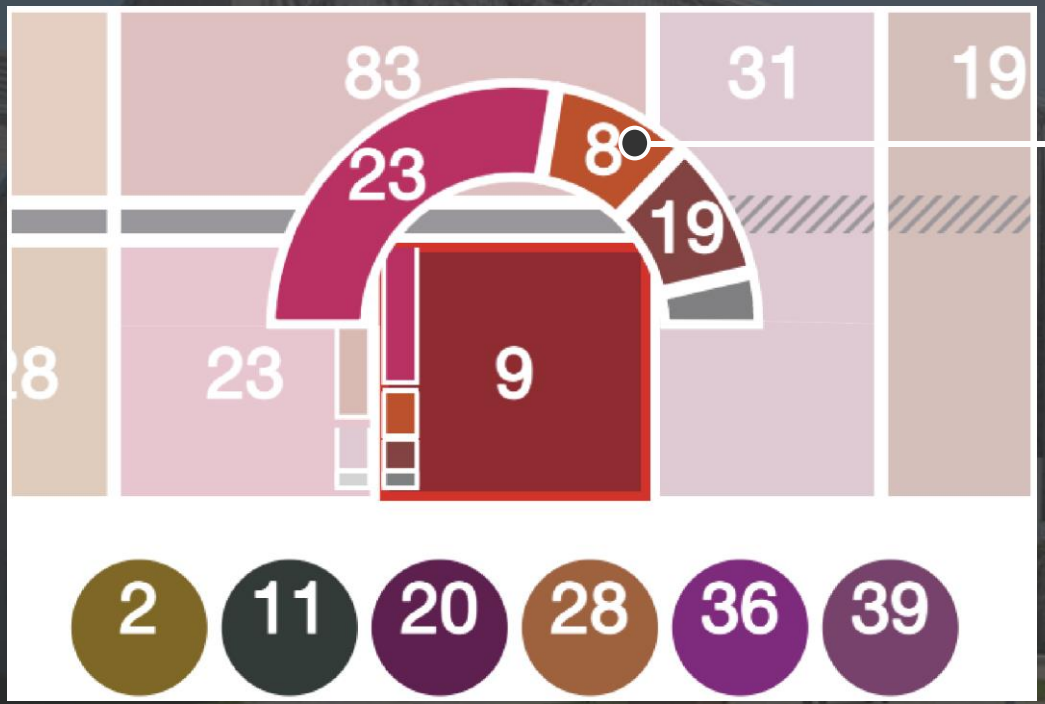

Human-AI Teaming
Interactive Visualization

Prediction Visualization



Combining the prediction results from five models, we visualized them with stacked bar charts on each node.

Prediction Visualization

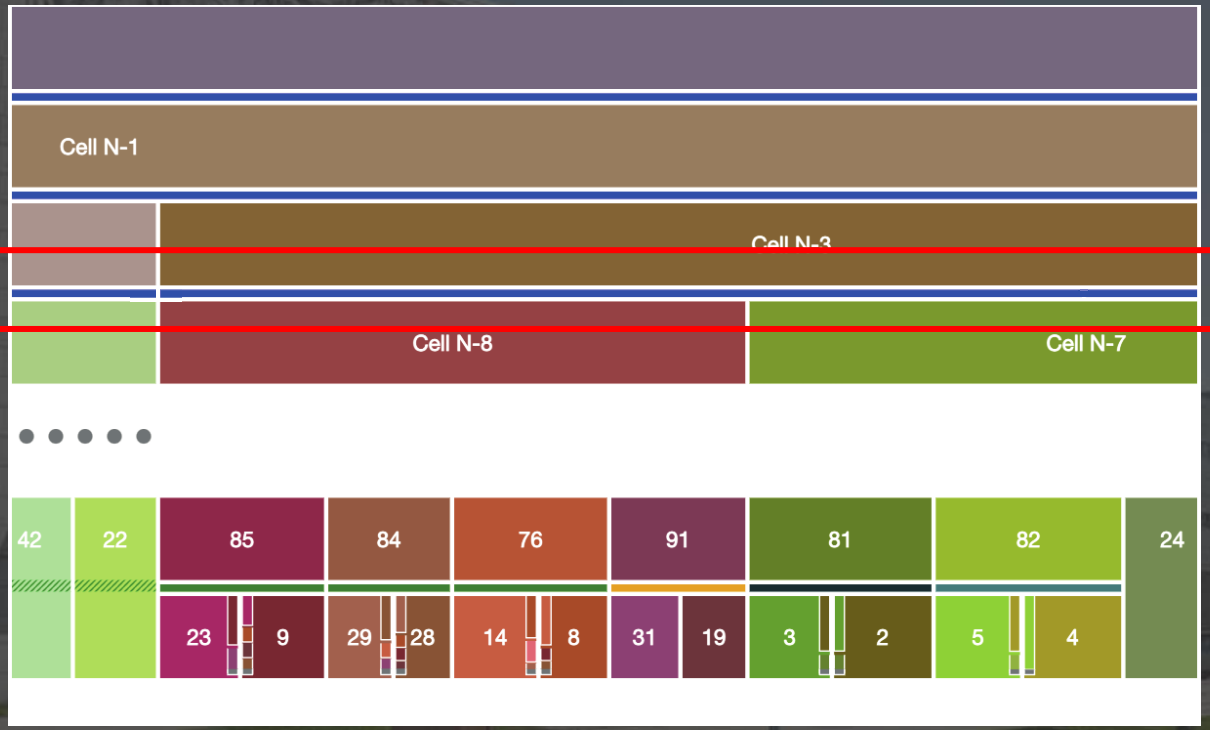


All predicted sisters

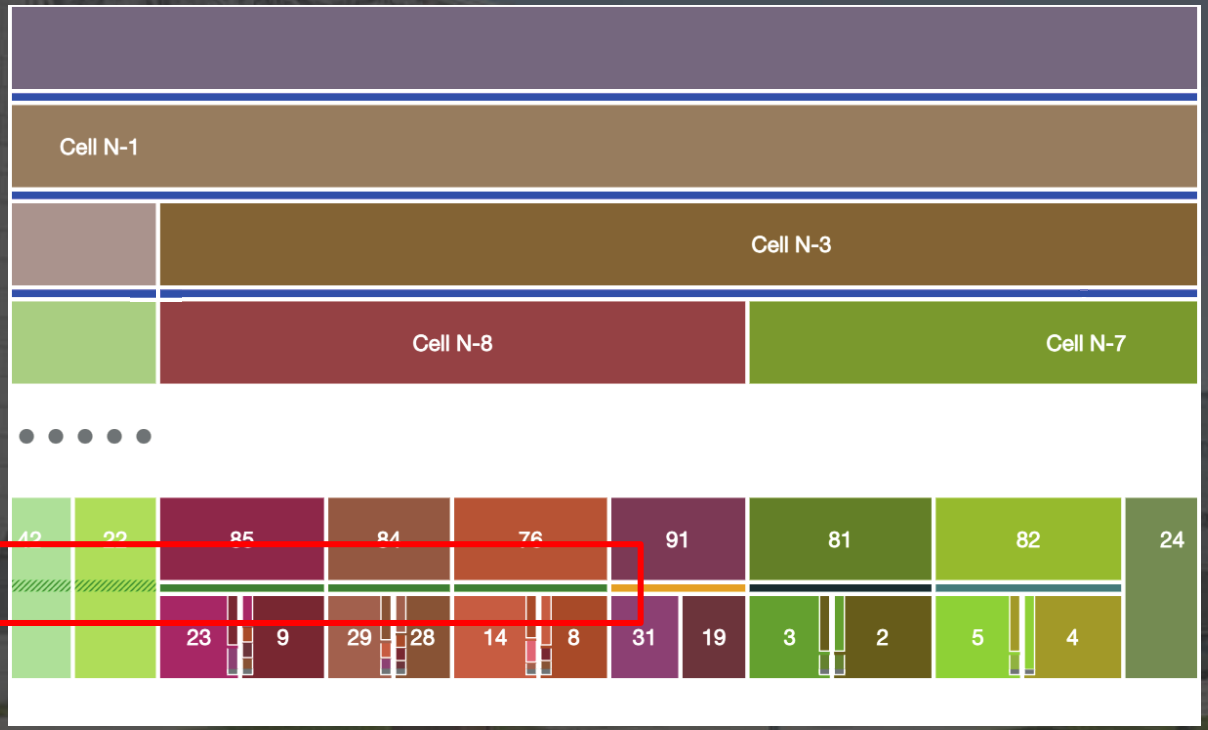
Vertical thumbnail of all predictions

All the other neighboring cells

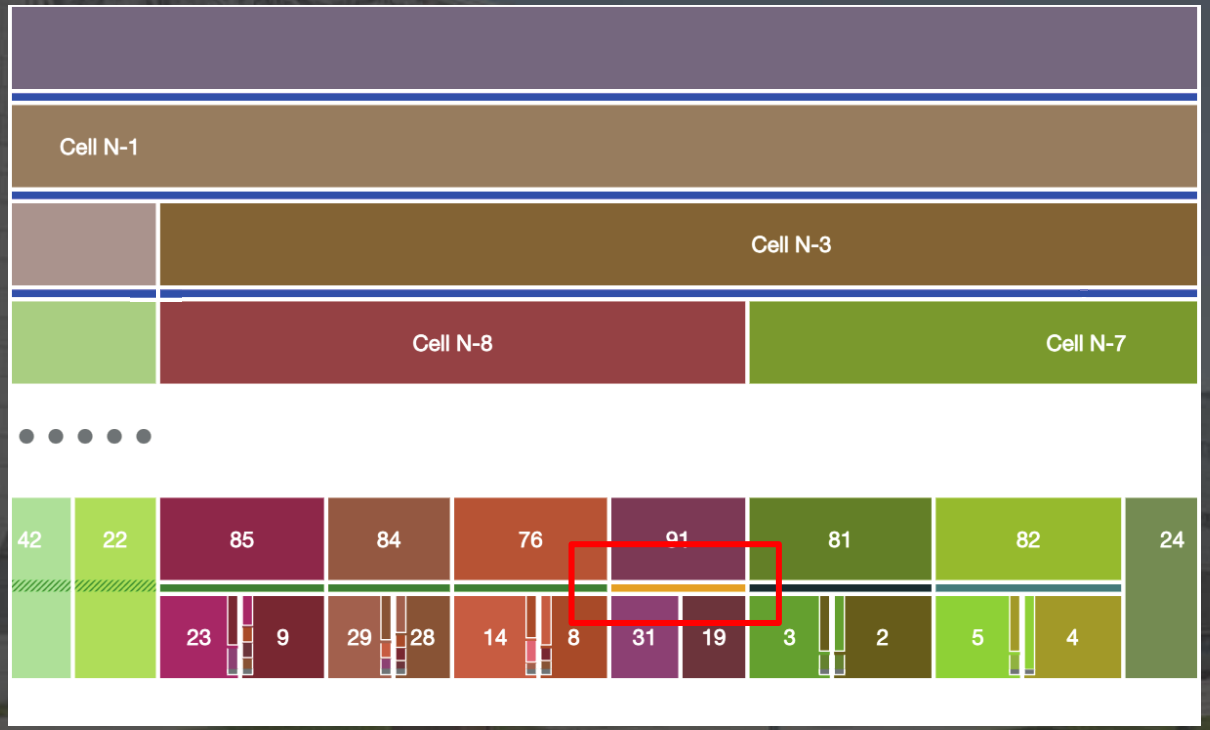
Differentiate Manual Assignments



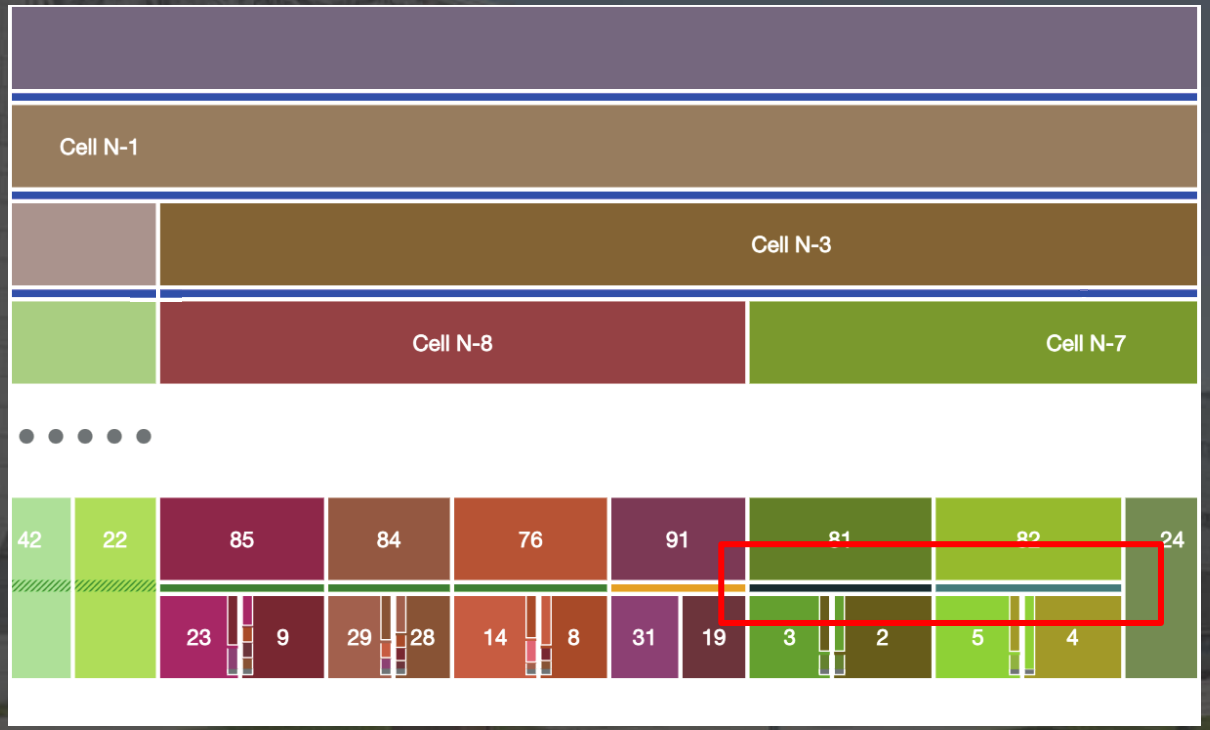
Differentiate Manual Assignments

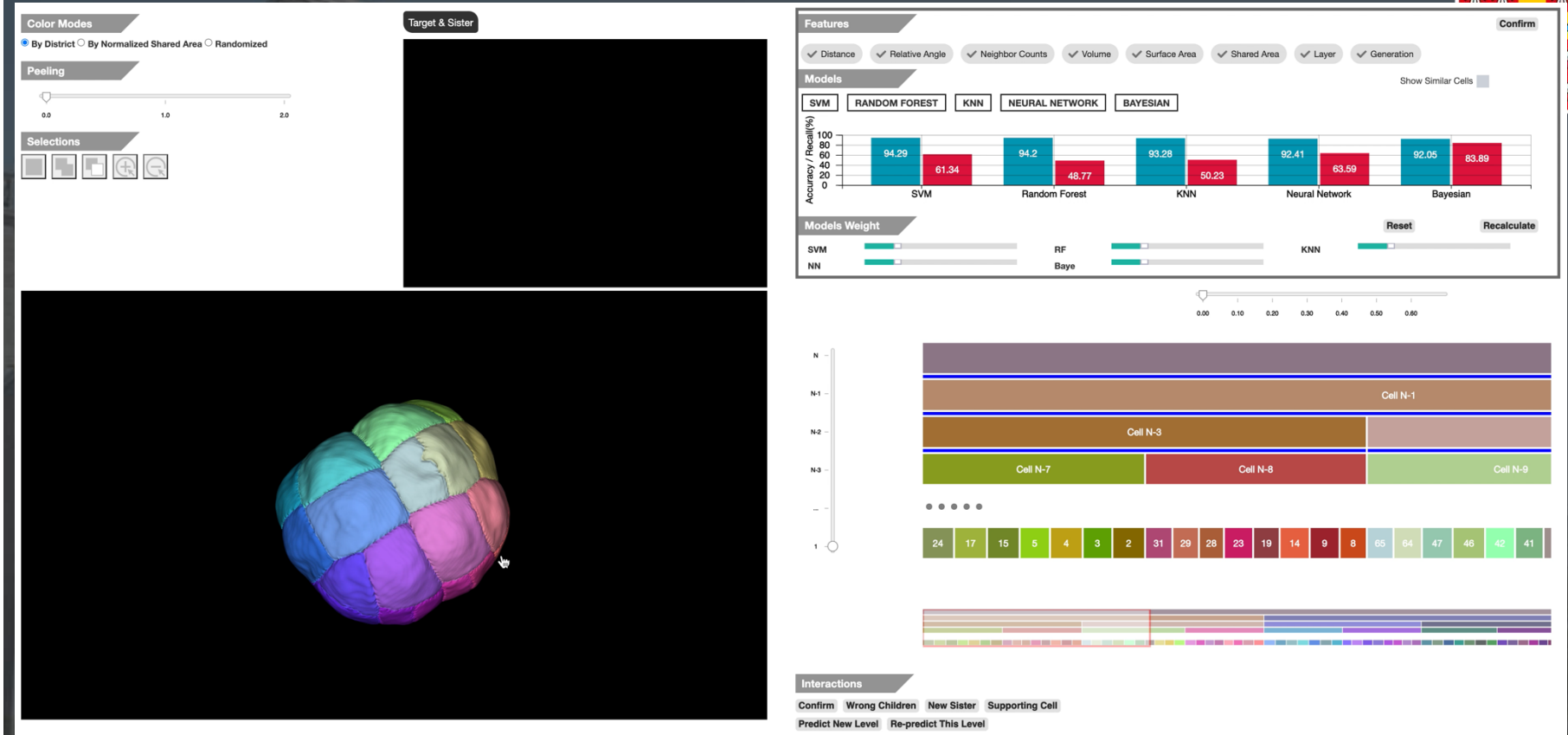


Differentiate Manual Assignments



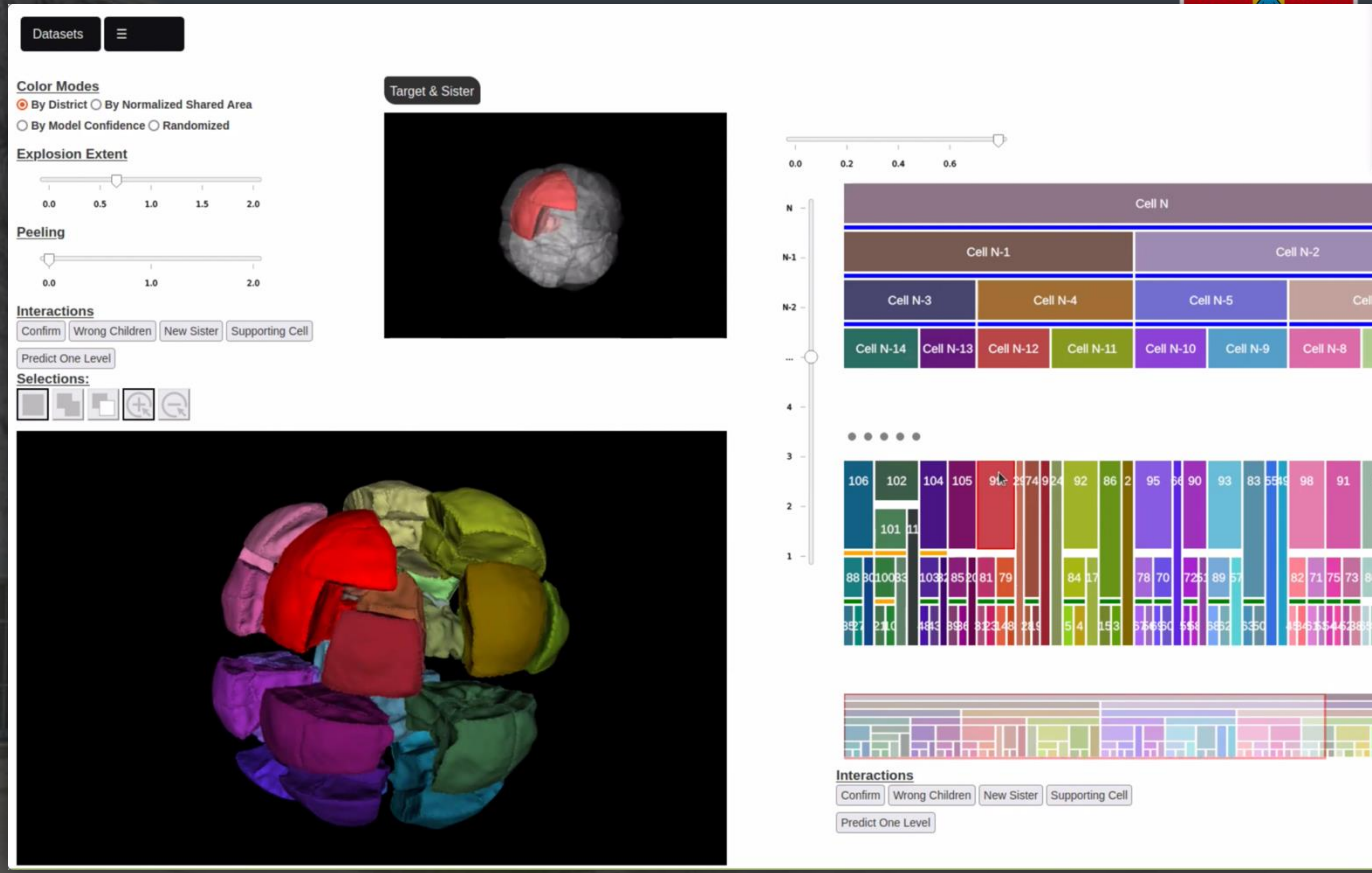
Differentiate Manual Assignments

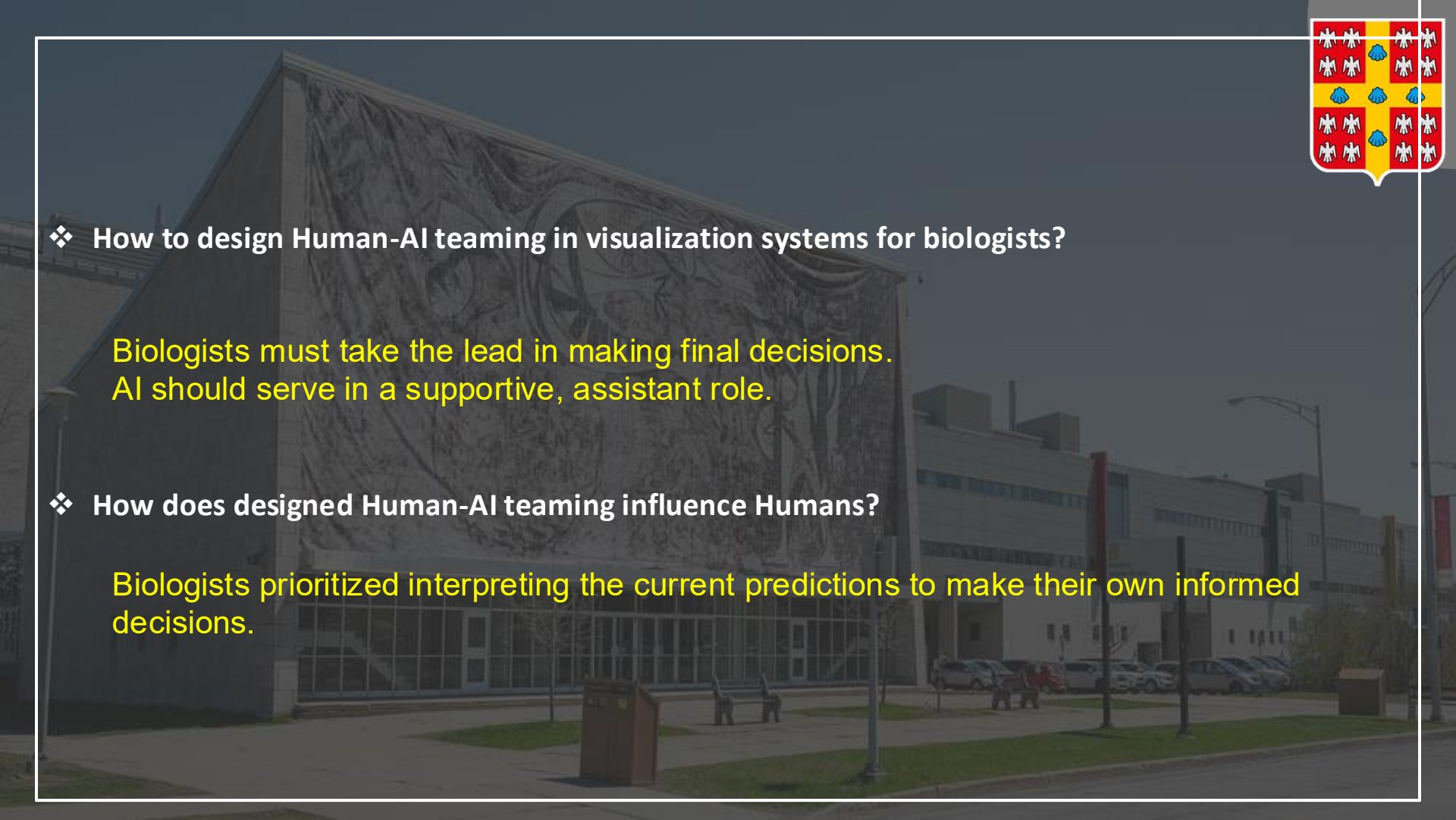




(**Hong** et al., Computer Graphics Forum, 2022
Hong et al., IEEE TVCG, 2024)

“I like the way you present the predictions and the way you allow us to correct assignments!....
It made me feel like I was working with my colleague."





- **How to design Human-AI teaming in visualization systems for biologists?**

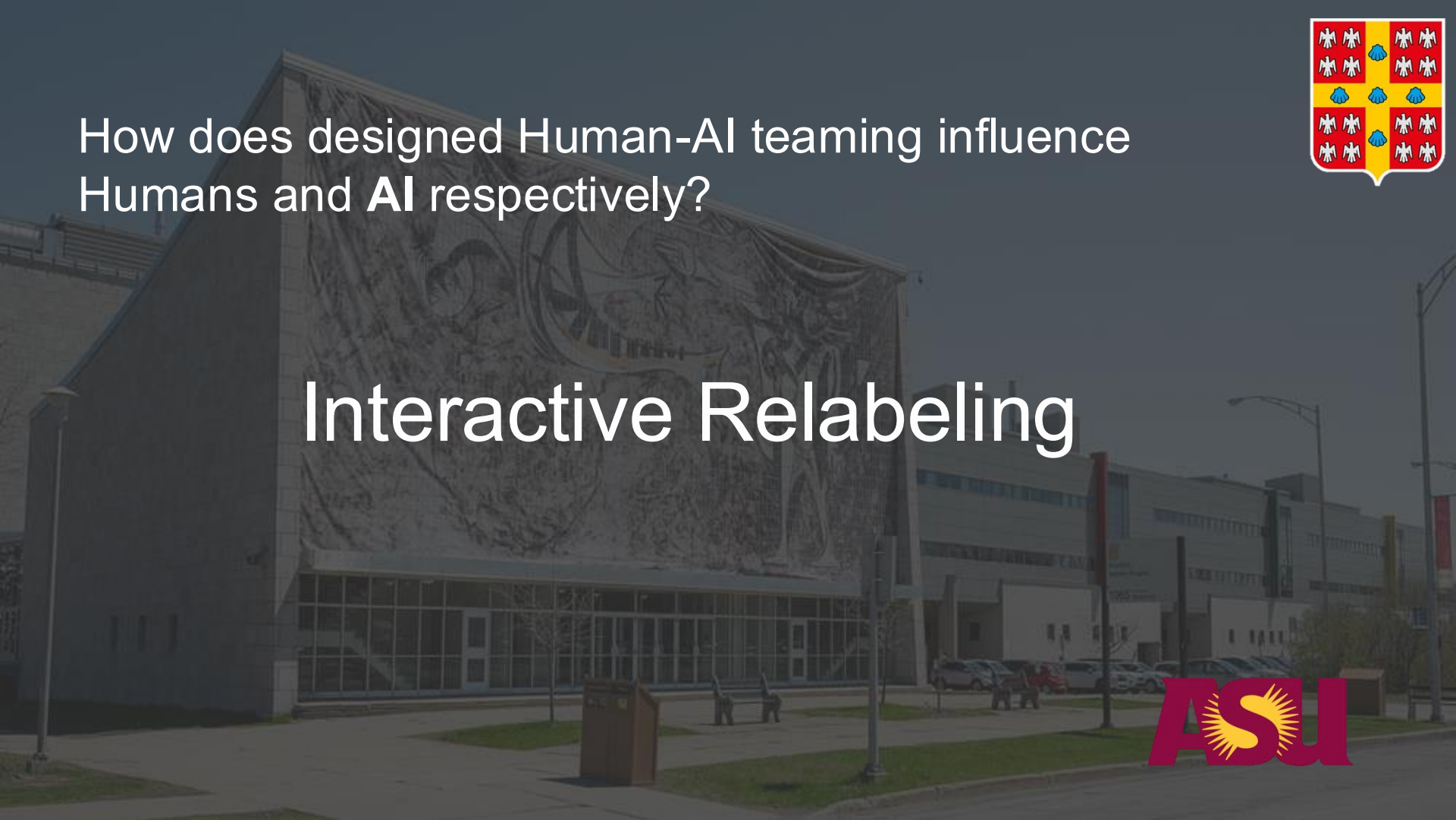
  Biologists must take the lead in making final decisions.
  AI should serve in a supportive, assistant role.

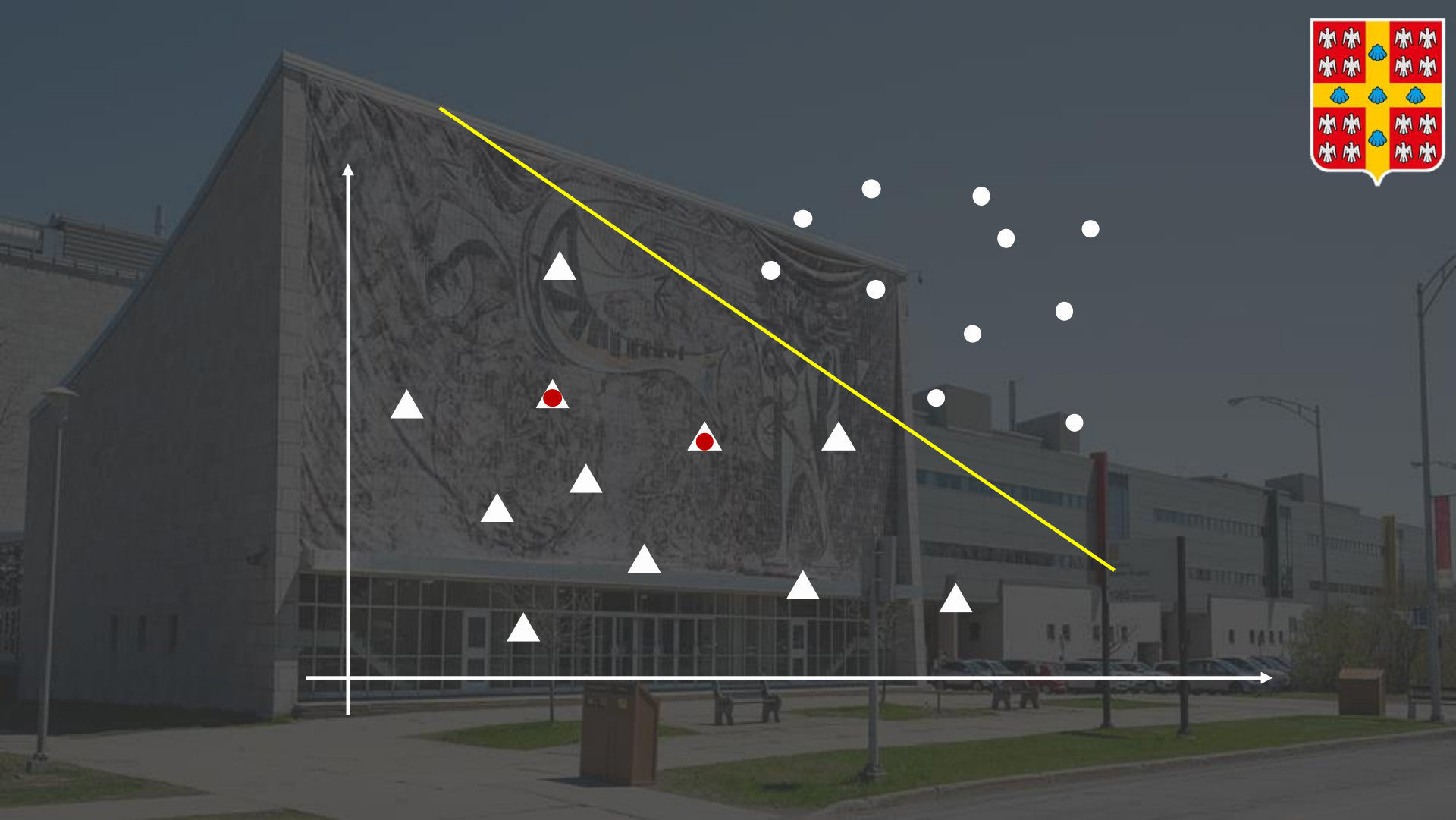
- **How does designed Human-AI teaming influence Humans?**

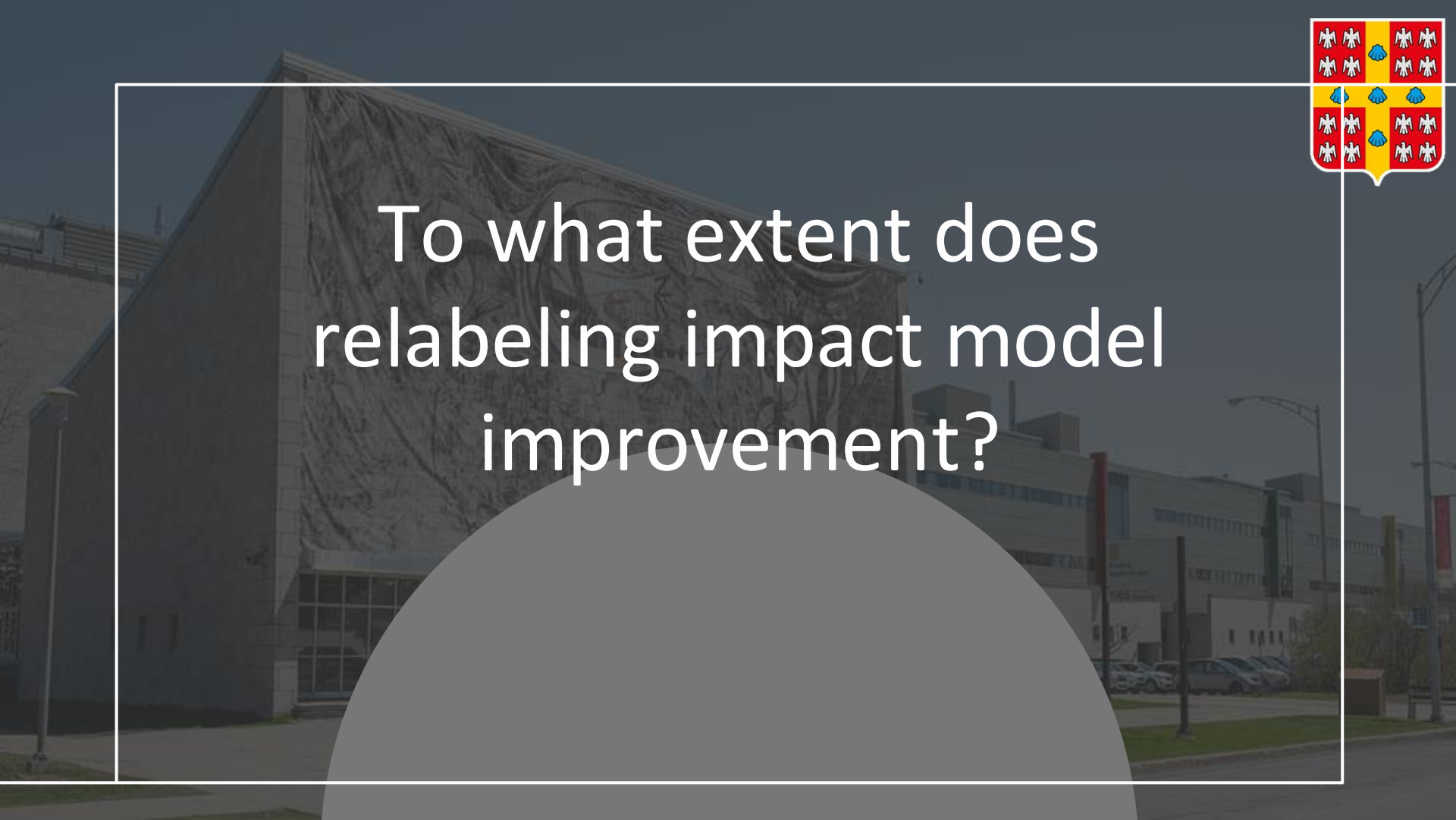
  Biologists prioritized interpreting the current predictions to make their own informed decisions.

How does designed Human-AI teaming influence Humans and **AI** respectively?

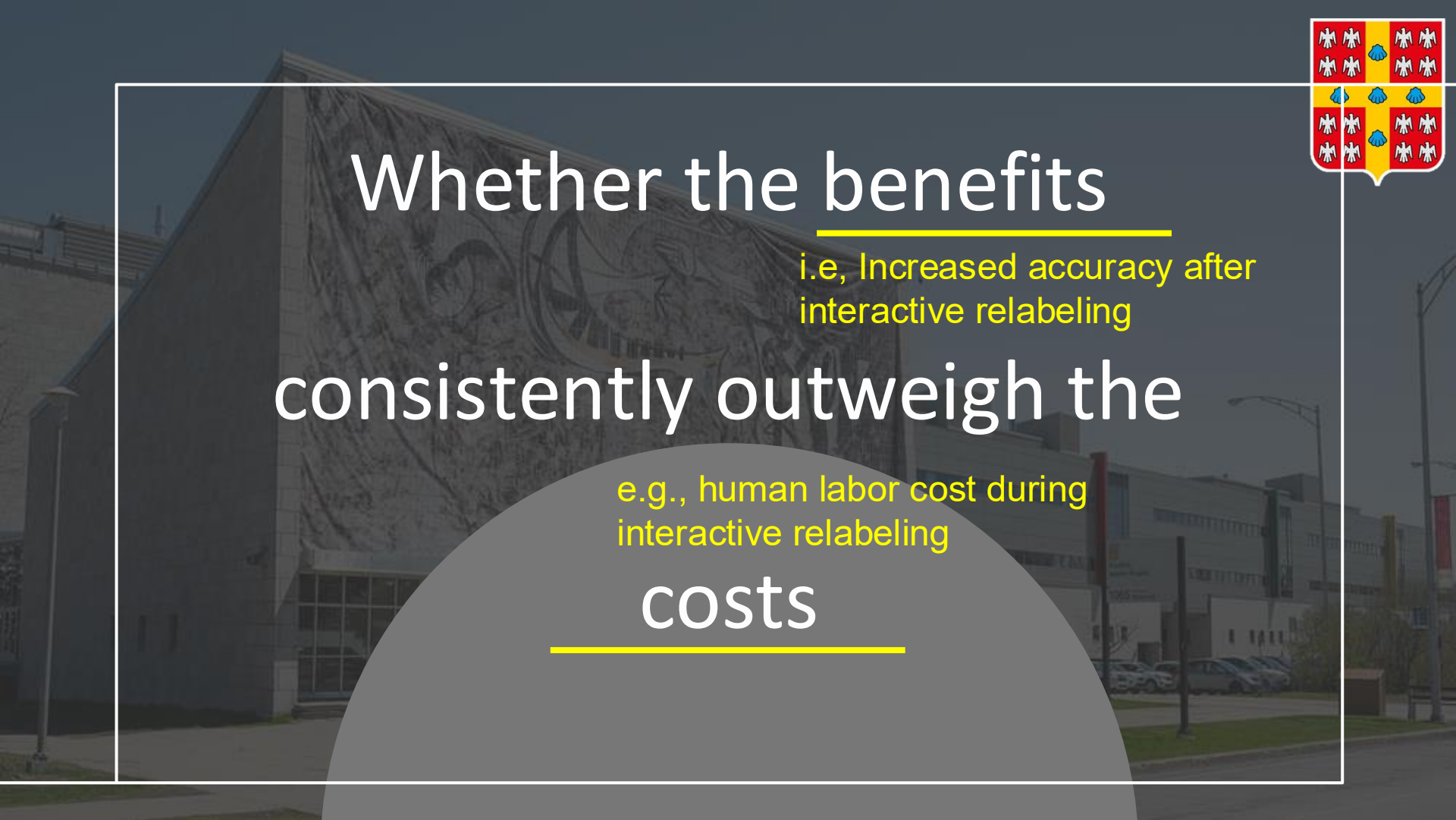
Interactive Relabeling







To what extent does relabeling impact model improvement?



Whether the benefits

i.e, Increased accuracy after interactive relabeling

consistently outweigh the

e.g., human labor cost during interactive relabeling

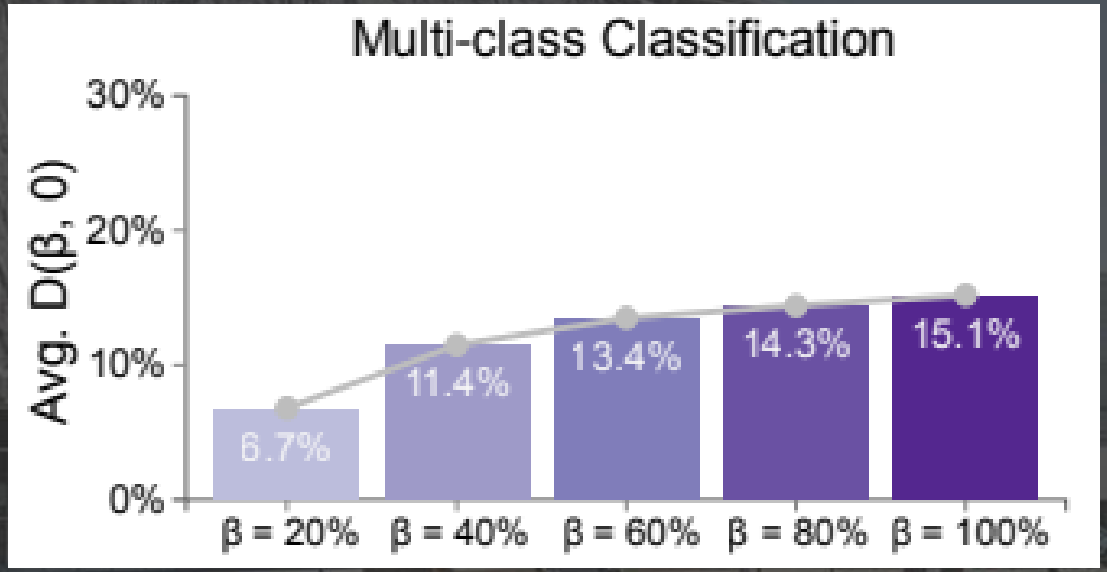
costs

Upper-bound Case

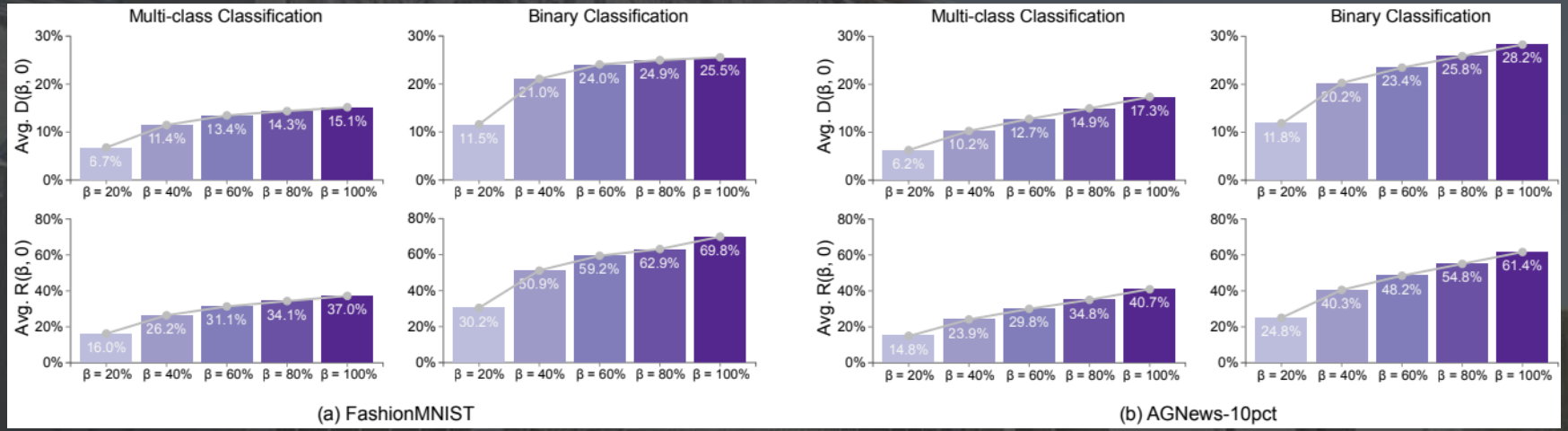
- A perfect visualization system to detect mislabels
- Mislabels will always be correctly relabeled
- Costs depend only on the number/proportion of mislabels

Simulation

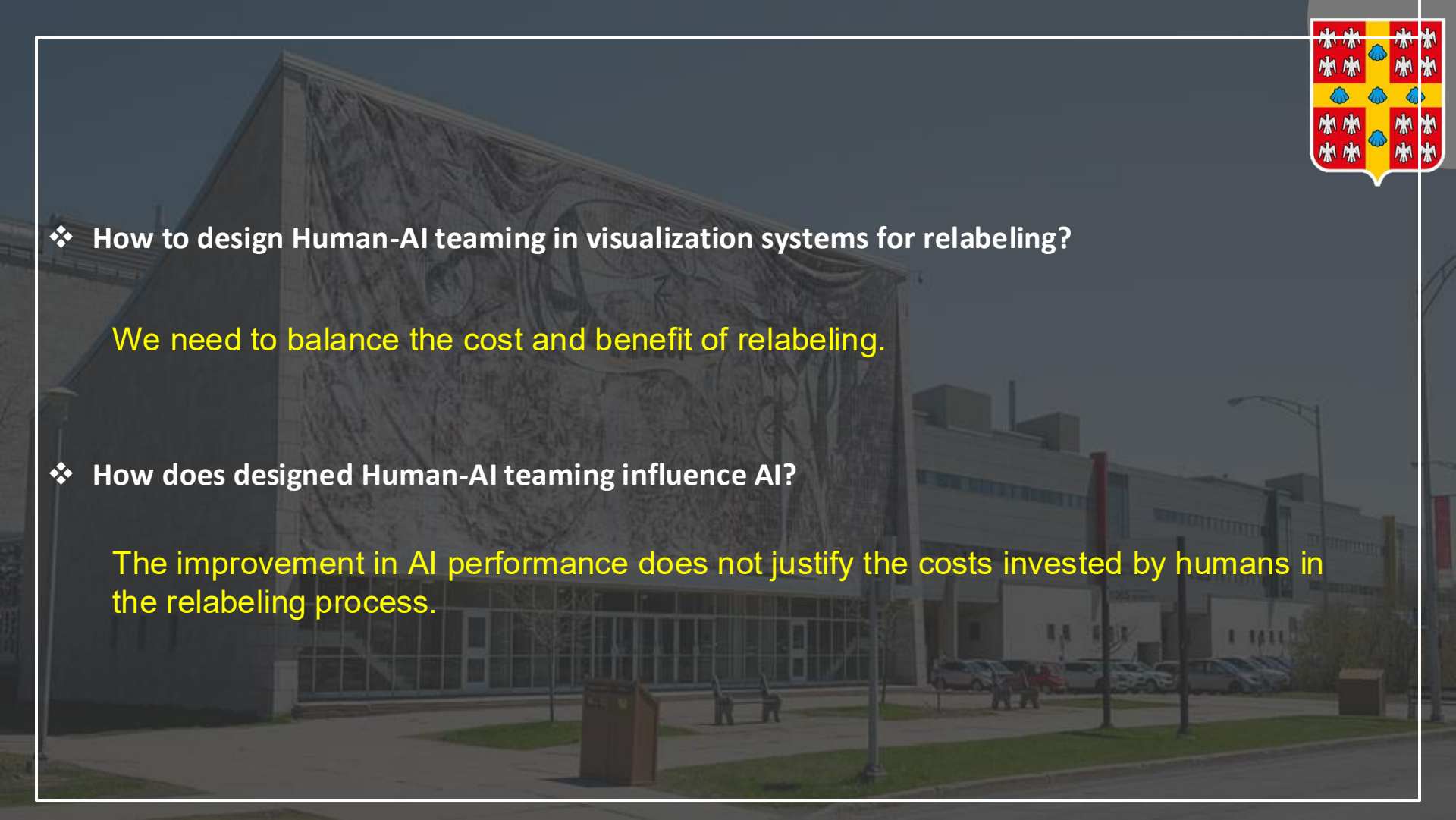
Results | FashionMINST



Results



Wang et al., IEEE TVCG, 2024



- **How to design Human-AI teaming in visualization systems for relabeling?**

  We need to balance the cost and benefit of relabeling.

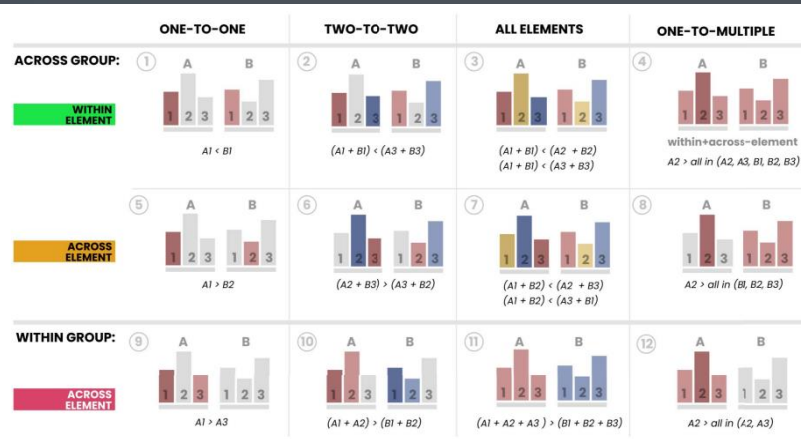
- **How does designed Human-AI teaming influence AI?**

  The improvement in AI performance does not justify the costs invested by humans in the relabeling process.


Visualization of Different scales?

5.3.1 One-to-One Comparisons (1:1)

Viewers can compare two individual bars in their takeaways. We refer to them as "one-to-one" comparisons, as shown in the leftmost column in Figure 5, comparison types C1, C5, and C9. For ***across group - within element*** operations, one-to-one comparison means that the viewer compares one element in one group to the same element in another group, such as comparing element 1 in group A (which we will refer to as A1) to element 1 in group B (which we will refer to as B1) . For ***across group - across element*** operations, the viewer compares one element in one group to another element in a different group, such as comparing A1 to B2 . For ***within group - across element*** operations, one-to-one comparison means that the viewer compares one element in one group to another element in the same group, such as comparing A1 to A3 .



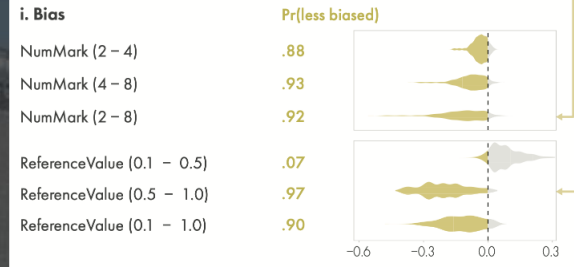
g. 5. Twelve categories of comparisons in two by three bar charts in the adjacent arrangement.

[Xiong et al. 2022]

A

nderstand the differences in visual channels for the reproduction , we report various effects on each of the precision, bias, and error measures. We then derive ranks for the visual channels.

We base our inference on the first distribution of the mixture model and the posterior distributions (marginal, conditional, and predictive distributions). Marginal posterior distributions summarize all the known information for one parameter; conditional posterior distributions tell us the expected value of one parameter in a specific situation; and posterior predictive distributions provide unobserved data conditioning on the observed data and the fitted model.

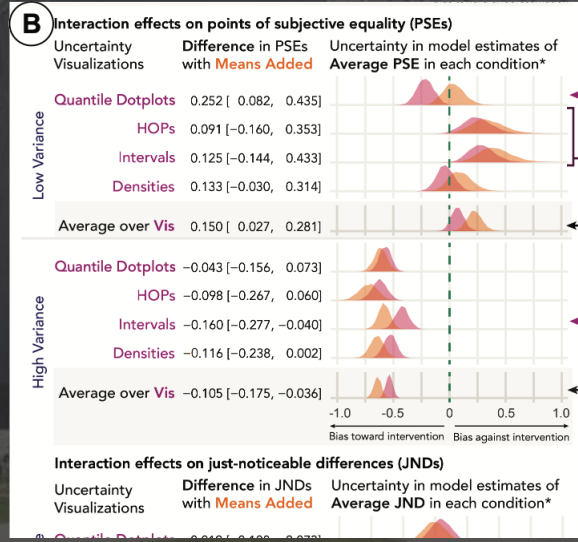


Number of marks. The average participant is very likely to be *less* biased in the reproduction, when the number of marks is *small*. For an average visual channel and an average reference value, the estimated probability that the average participant is less biased in a chart with 2 marks than with 8 marks is .92. That is, for the same reference value, we expect 92% of responses with 2 marks to exhibit less bias than the responses with 8 marks.

Reference value. The average participant is very likely to *overestimate* a small reference value and seriously *underestimate* a large reference value, and are least biased with a reference value around .4 or .5 (median). For an average visual channel and an average number of marks, the estimated probability that the participant is less biased in a chart with the median reference value (.5) than the minimum value (.1) is .93 (this is 1-.07). Similarly, the estimated probability that an average participant is less biased in a chart with the median reference value (.5) than the maximum value (1.0) is .97.

Interaction effects. The effects of *NumMark* and *ReferenceValue* interact, and each interacts with *VisualChannel*. For most of the visual channels but position (line), response bias increases when the number of marks is large and a reference value deviates from the median further. Overall, angle is the visual channel where response bias is most sensitive to either a change in the number of marks or the reference value; position (line) is where bias is sensitive to the reference value, but robust to the number of marks for large reference values.

B



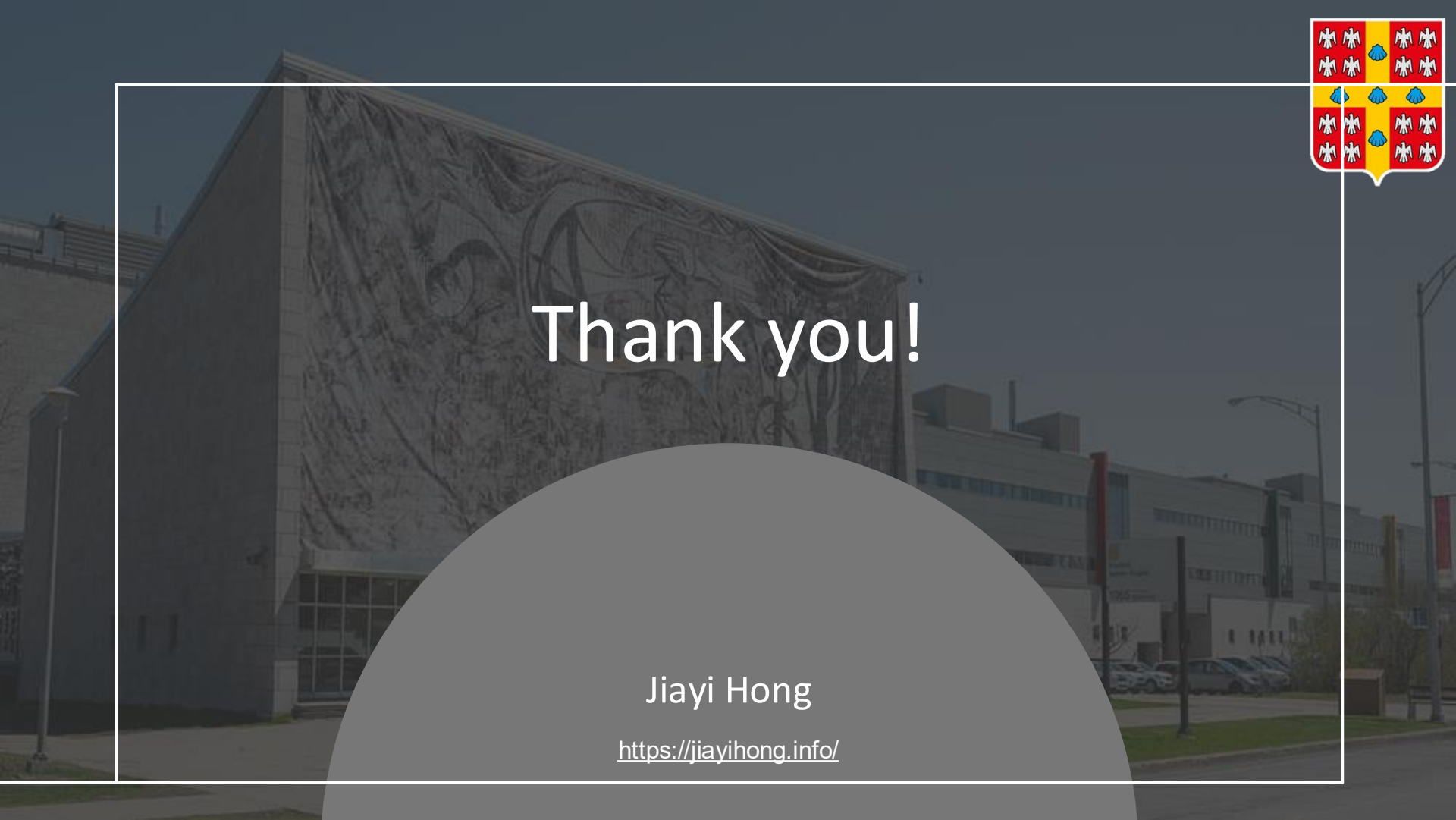
4.2 Intervention Decisions

4.2.1 Points of Subjective Equality

For each uncertainty visualization, **adding means** at **low variance** increases PSEs. This results in different effects depending on whether the visualization with **no means** has a PSE below or above utility-optimal. Recall that a PSE of zero is utility-optimal, a negative PSE indicates intervening too often, and a positive PSE indicates not intervening often enough. Users of **quantile dotplots** with **no means** have negative PSEs which become unbiased when we **add means**. Users of **HOPs and intervals** with **no means** have positive PSEs, biases which increase when we **add means**. Users of **densities** with **no means** have PSEs near zero and become more biased when we **add means**. Only the effect for quantile dotplots is reliable. When we **average over uncertainty visualizations**, at **low variance** the average user may have a PSE 0.6 percentage points above utility-optimal with **no means**, and **adding means** increases this mild bias by about 1.7 percentage points in terms of the probability of winning.

At **high variance**, **adding means** decreases PSEs. Since PSEs for all uncertainty visualizations with **no means** are below optimal, **adding means** increases biases in all conditions, however, the effect is only reliable for **intervals**. When we **average over uncertainty visualizations**, at **high variance** the average user has a negative PSE 9.5 percentage points below utility-optimal with **no means**, and **adding means** increases this bias by about 2.1 percentage points.

How could we design intuitive embellishments within papers for better readability?



Thank you!

Jiayi Hong

https://jiayihong.info/