

Generative AI and data visualization

Jacob S. Berv, PhD

Biodiversity Data Science
and Systematic Biology

Schmidt AI in Science
Postdoctoral Fellow

Michigan Life
Sciences Fellow



Generative AI and data visualization

Jacob S. Berv, PhD

Part 1

Who am I? My experience with data visualization
How I think about generative AI

Part 2

Writing functions in R, and LLMs to write custom plotting
functions for data visualization (promptprogramming)

Part 3

Data challenges

Research Origins



Berv et al 2021



Berv et al 2014



Personal Timeline

2010, BS, Biology, Yale

2010-2013, Lab Tech, Yale

2013-2019, PhD, Cornell

2019-2023, MLS Fellow, UM

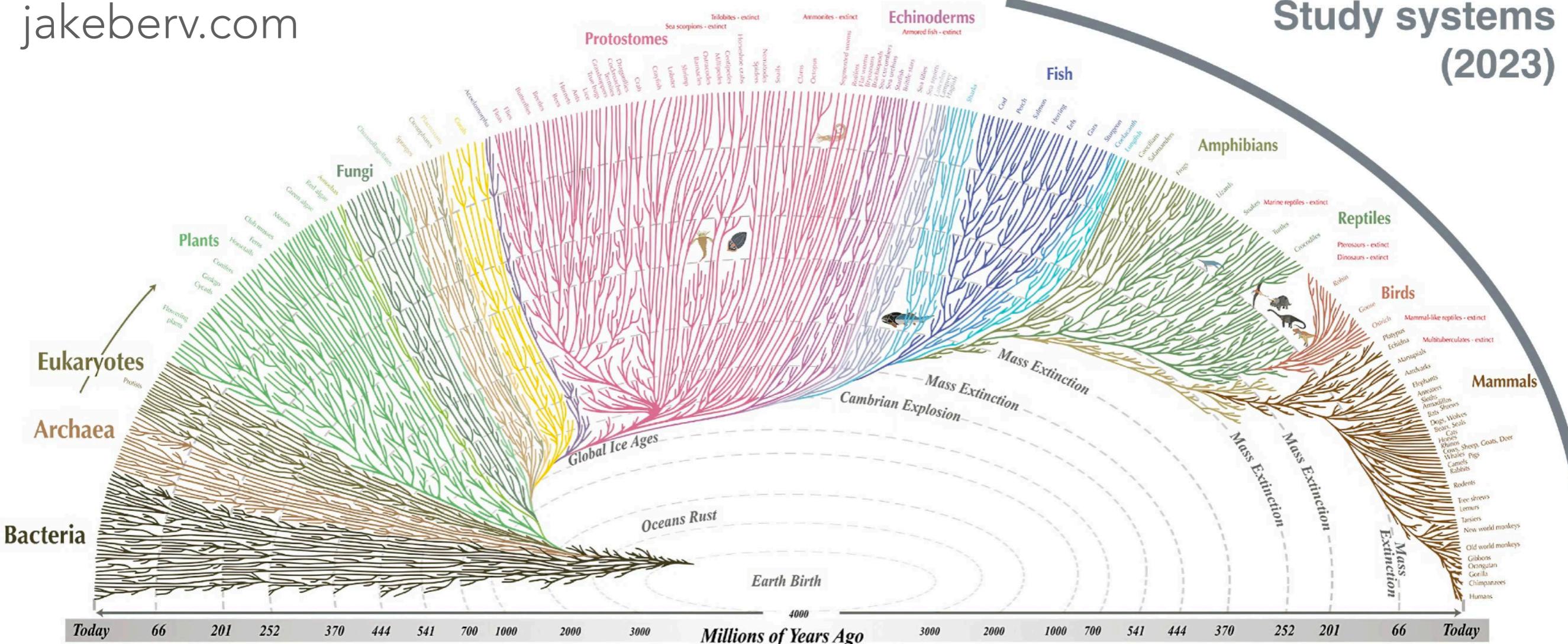
2024-Schmidt AI Fellow, UM

Studying manakin courtship displays outside Manaus (2015)



jakeberv.com

Study systems (2023)



"The affinities of all the beings of the same class have some times been represented by a great tree."

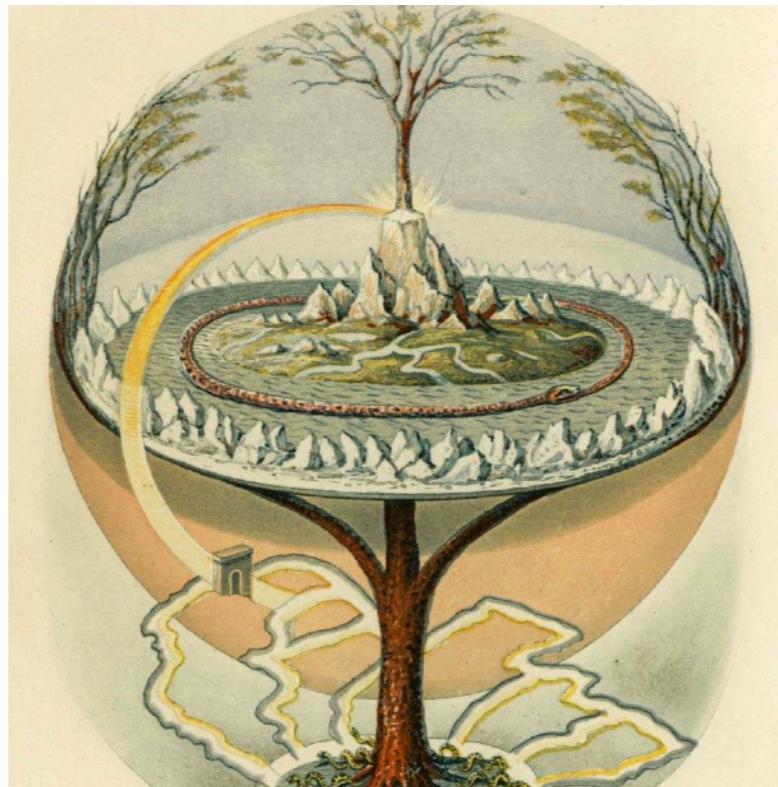
Charles Darwin - *On the Origin of Species*

Research program

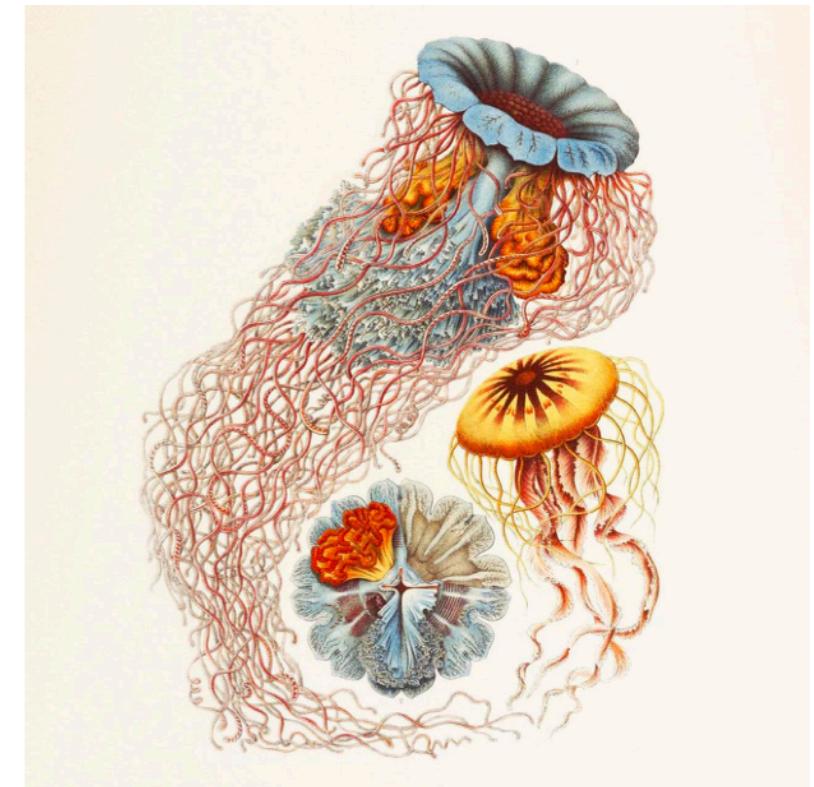
Phylogeny



Biogeography



Diversification



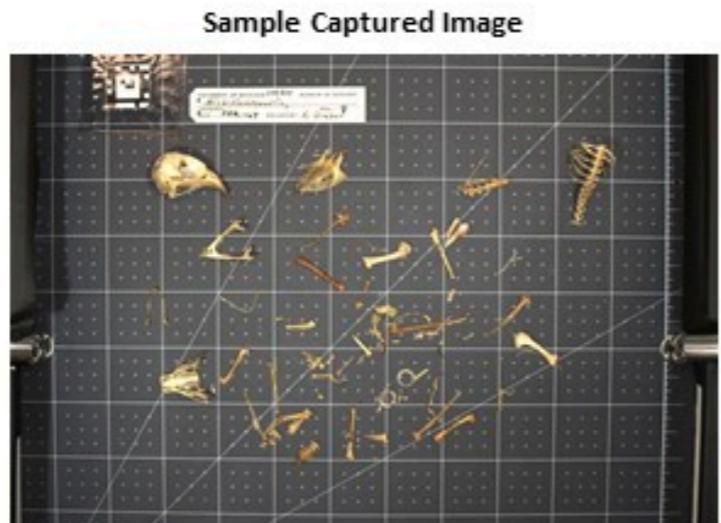
Phenotype



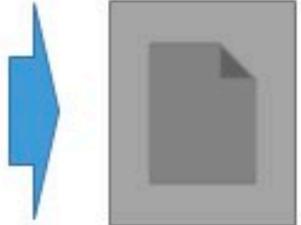
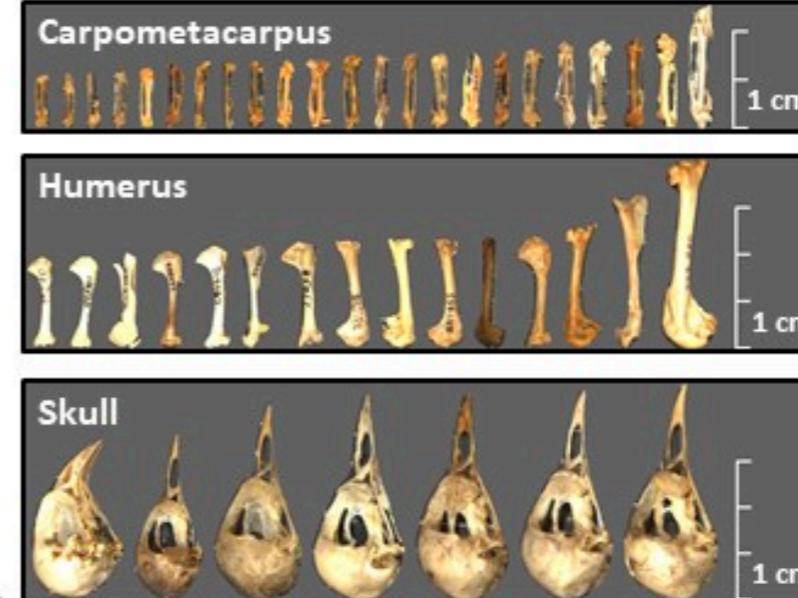
Phylogenetic Natural History in the age of AI



Bird Skeleton Imaging



Bone Measurement

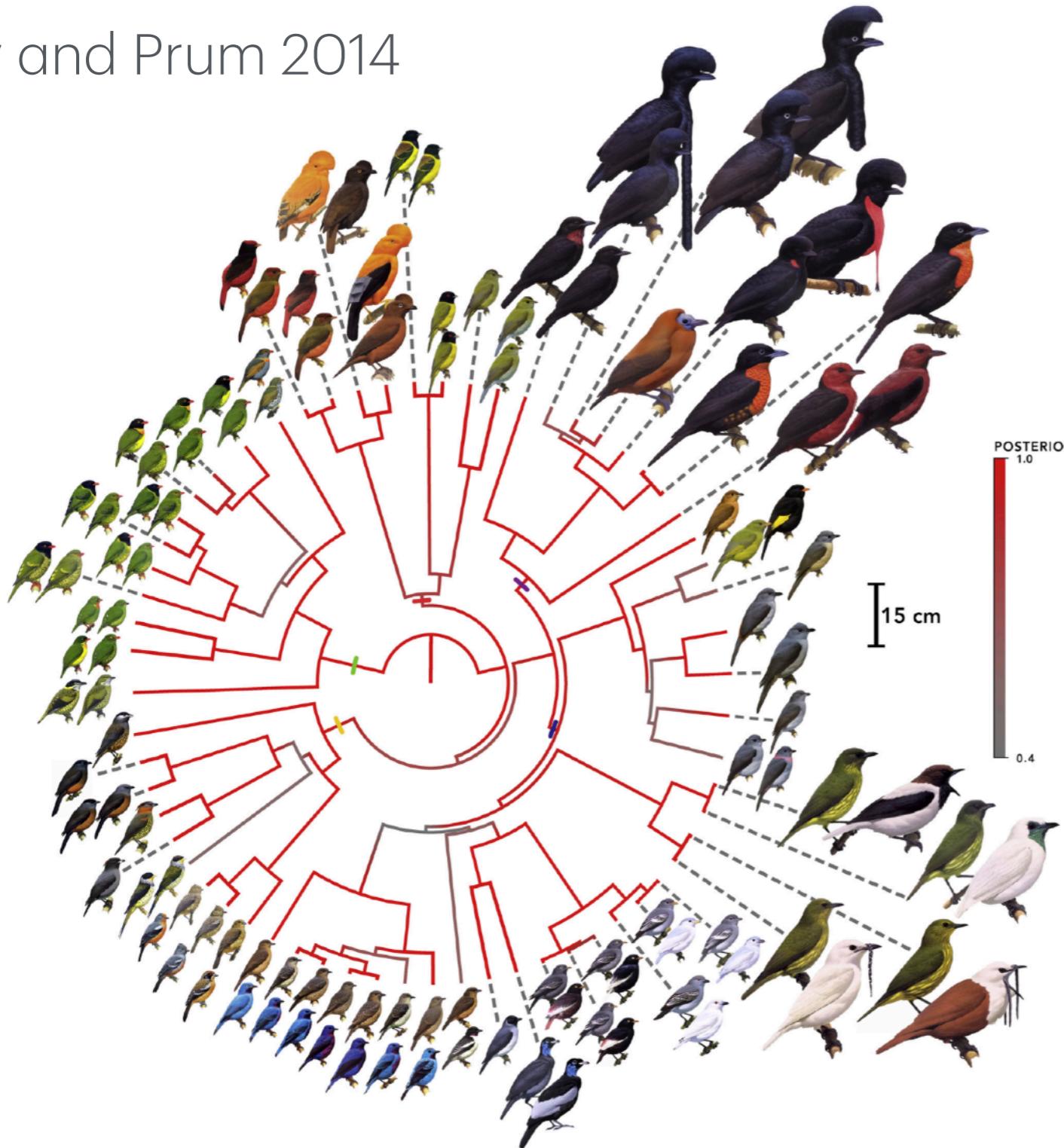


CSV
*we export
numerical
measurements

>15,000 skeletons in U. Michigan
collection measured in < 2 years

Phylogenetic data visualization

Berv and Prum 2014



LETTER

doi:10.1038/nature15897

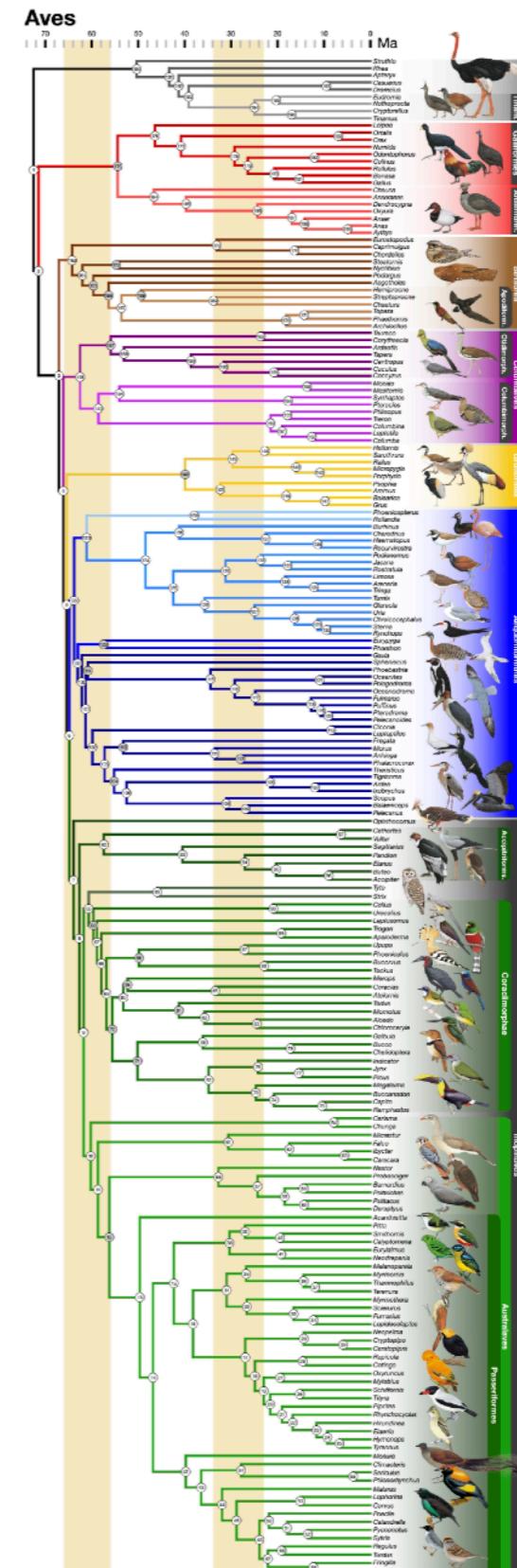


Figure 1. A Comprehensive Phylogeny of Birds (Aves) using Targeted Next Generation DNA Sequencing

Richard O. Prum^{1,2*}, Jacob S. Berv³, Alex Dornburg^{4,5}, Daniel J. Field⁶, Jeffrey P. Townsend⁷, Emily Moriarty Lemmon⁸, and Alan R. Lemmon⁹

¹ Department of Ecology & Evolutionary Biology, Yale University, New Haven CT USA

² Peabody Museum of Natural History, Yale University, New Haven CT USA

³ Department of Ecology & Evolutionary Biology, Cornell University, Ithaca NY USA

⁴ Department of Geology & Geophysics, Yale University, New Haven CT USA

⁵ Department of Earth Sciences, Department of Geobiology, Department of Earth System Science, and Department of Earth and Planetary Sciences, Yale University, New Haven, CT USA

⁶ Department of Biological Science, Florida State University, Tallahassee, FL USA

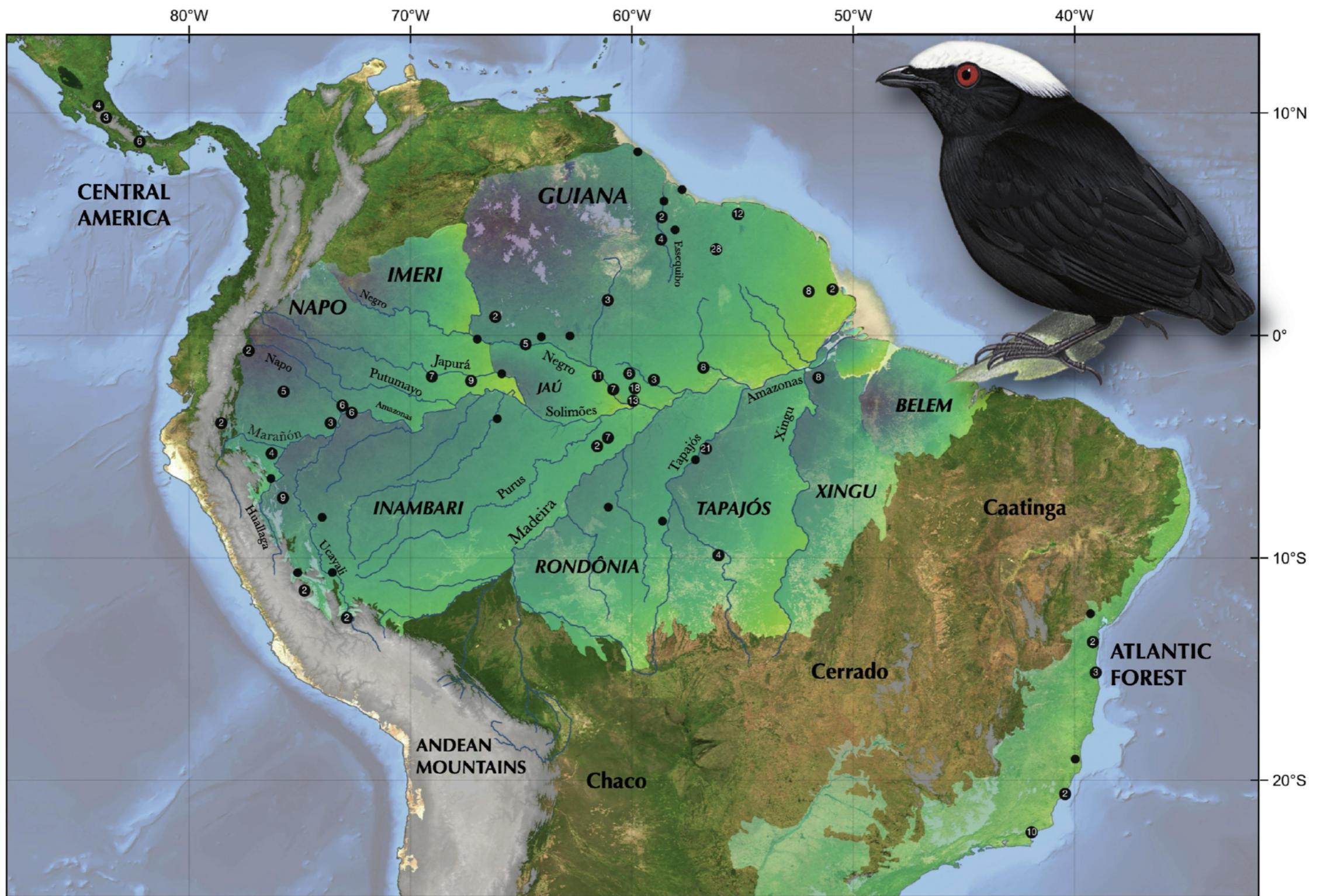
⁷ Department of Scientific Computing, Florida State University, Tallahassee, FL USA

* These authors contributed equally to this work.

[†] Designed Figure 1.

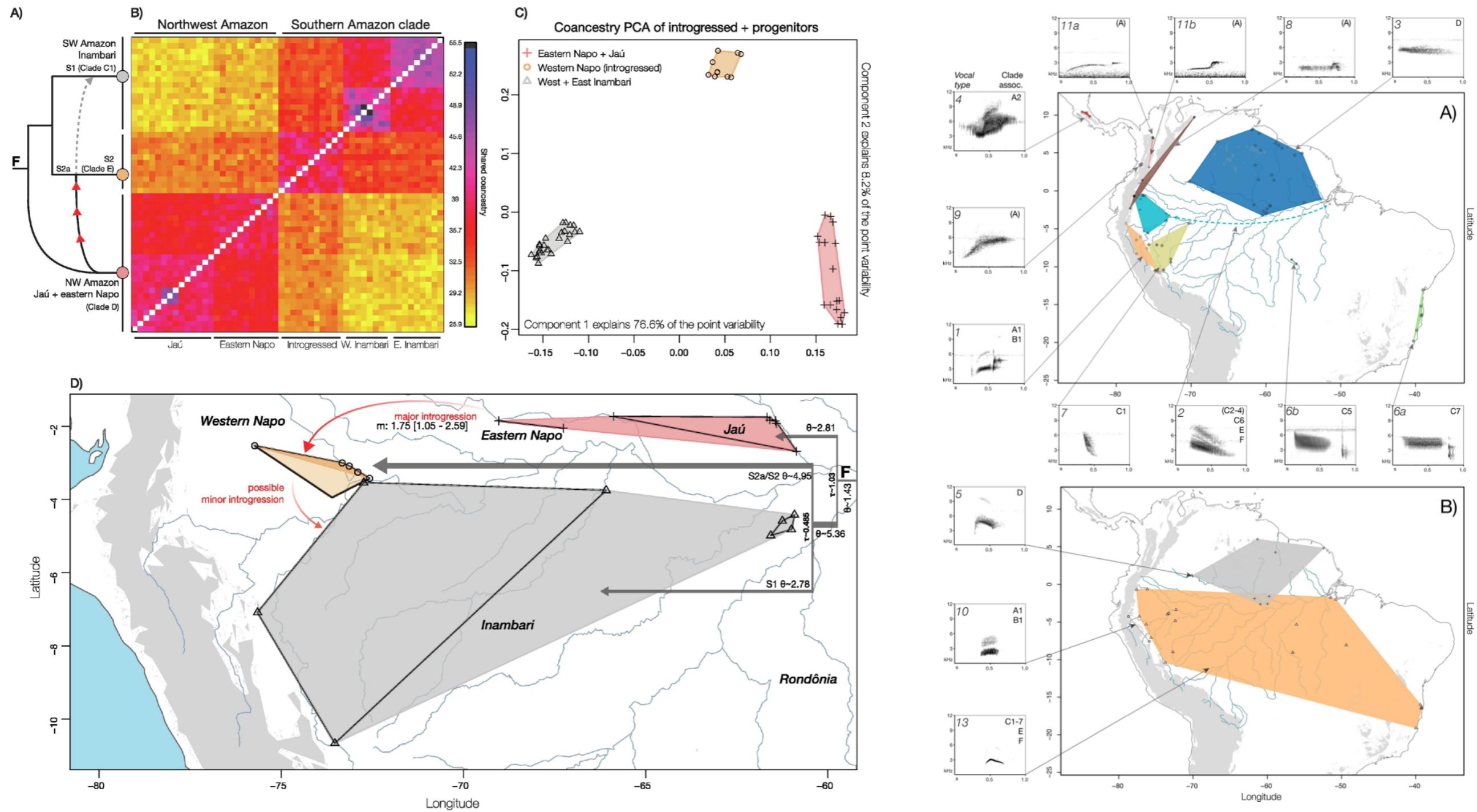
Prum,
Berv, et al
2015

Biodiversity data visualization



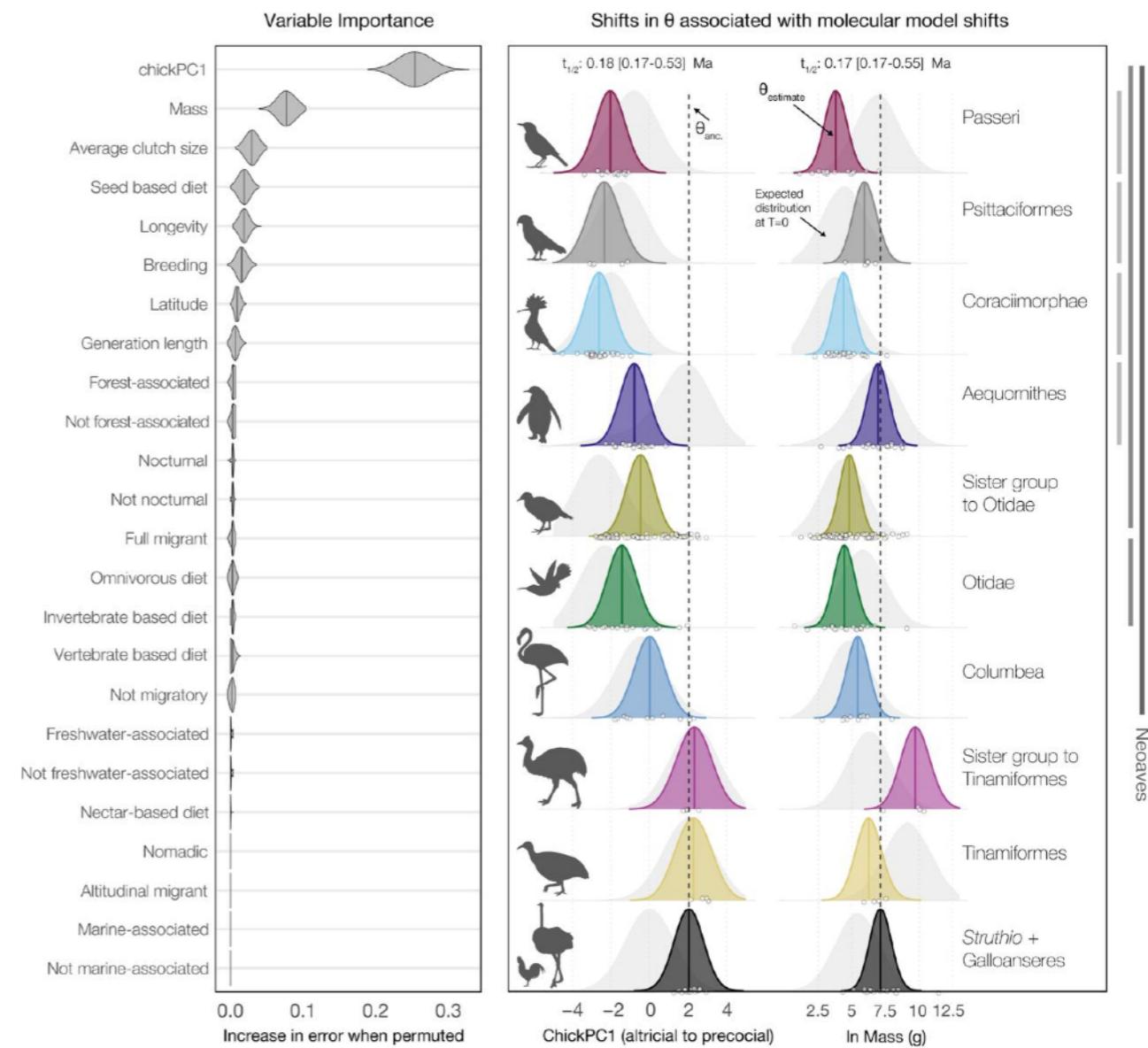
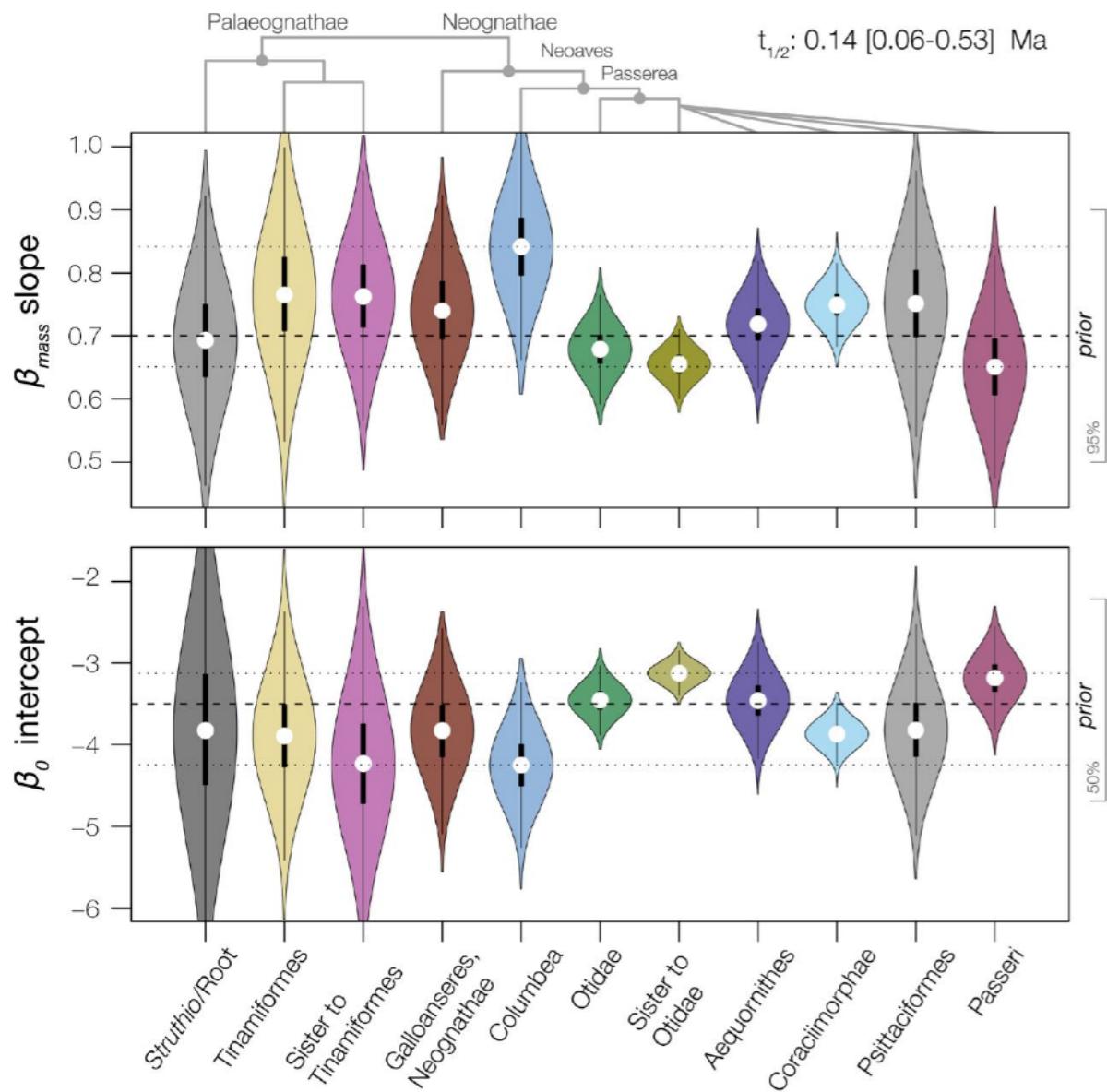
Berv et al 2021, Figure 1

Biodiversity data visualization



Berv et al 2021, Figures 6, 8

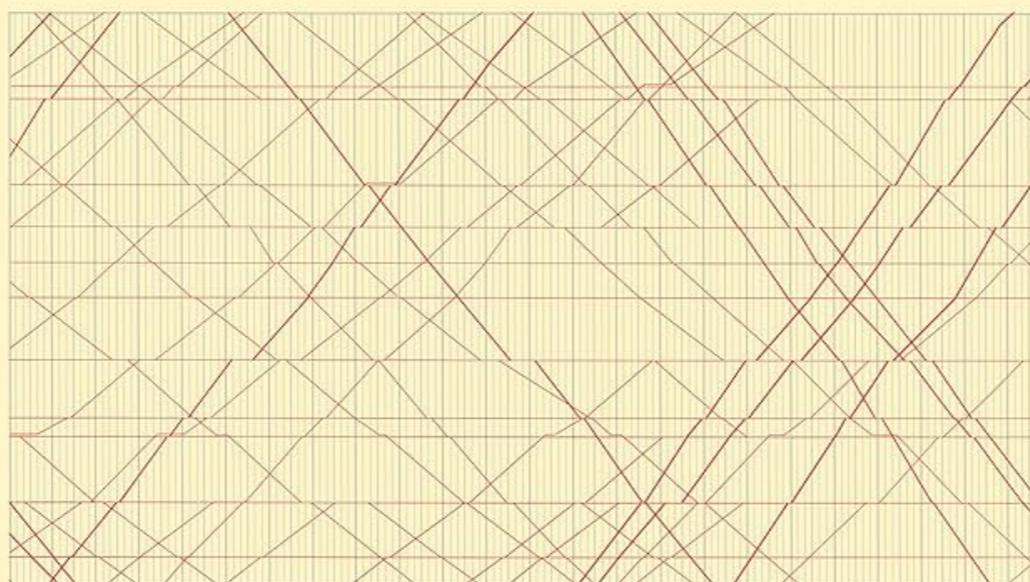
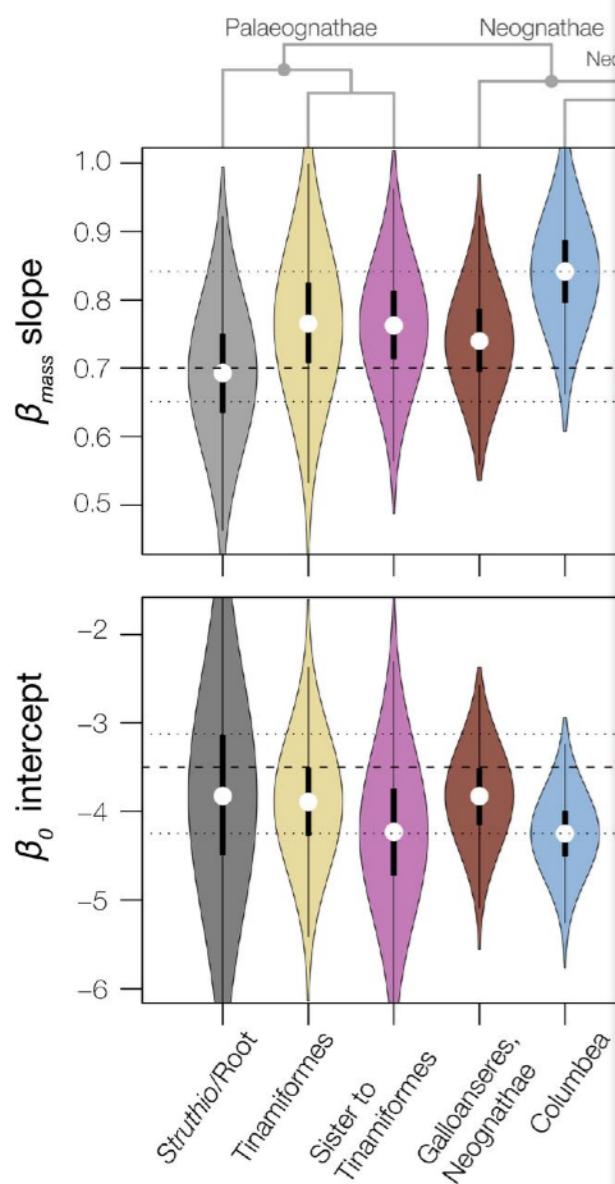
Biodiversity data visualization



Berv et al 2024, Figure 2, 3

Email me if you'd like to talk about how I made any of these figures

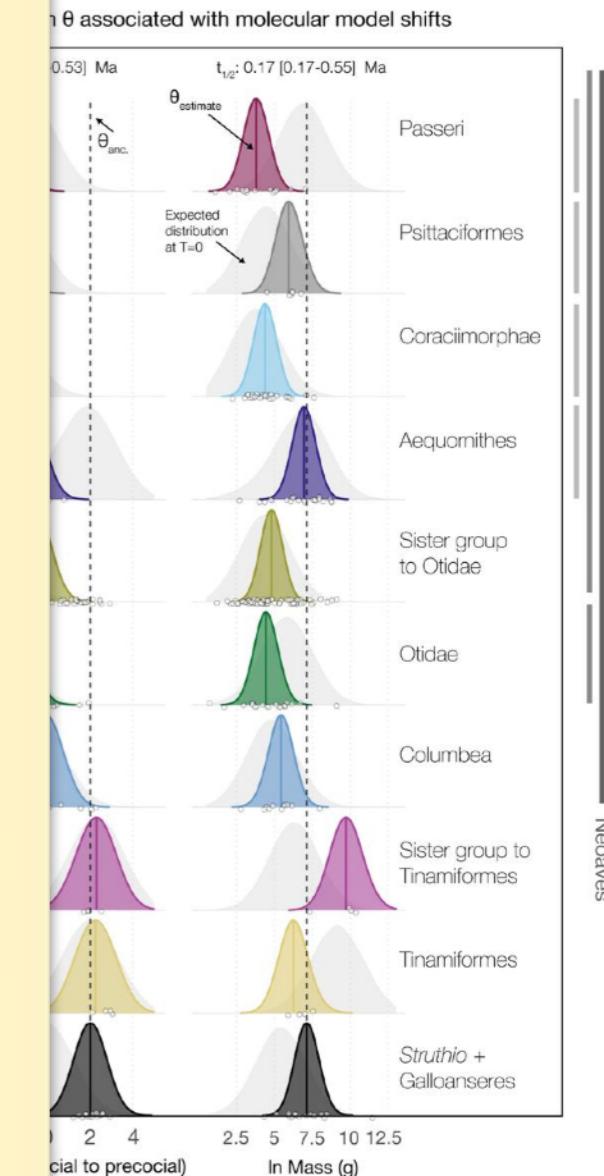
Biodiversity data visualization



SECOND EDITION
The Visual Display
of Quantitative Information

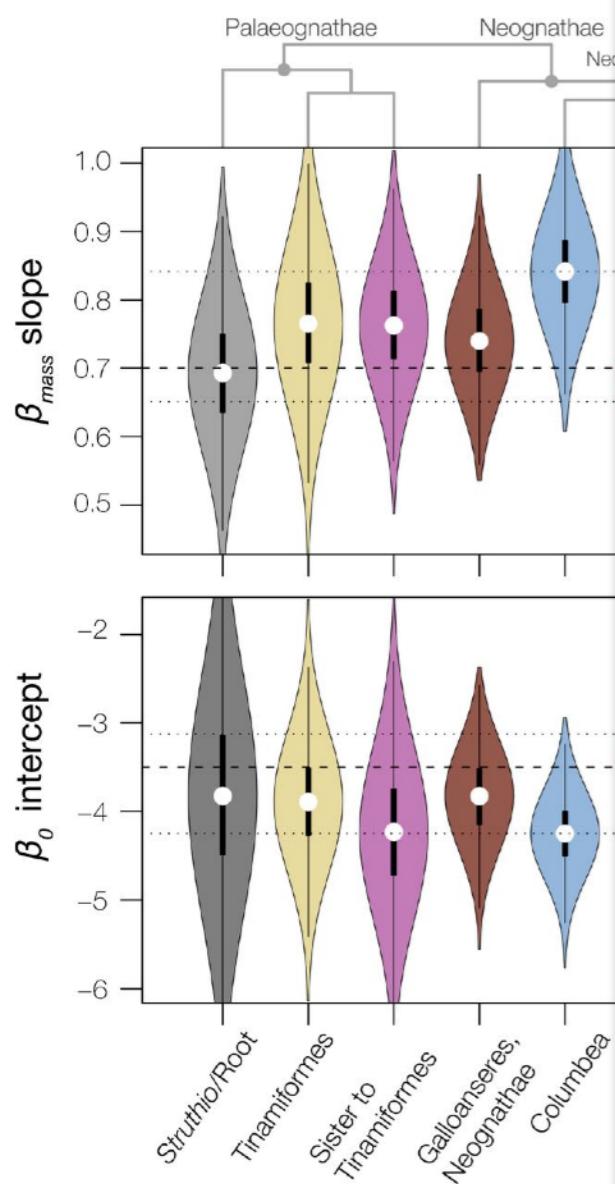
EDWARD R. TUFTE

Berv et al 2022,



Be to talk about
these figures

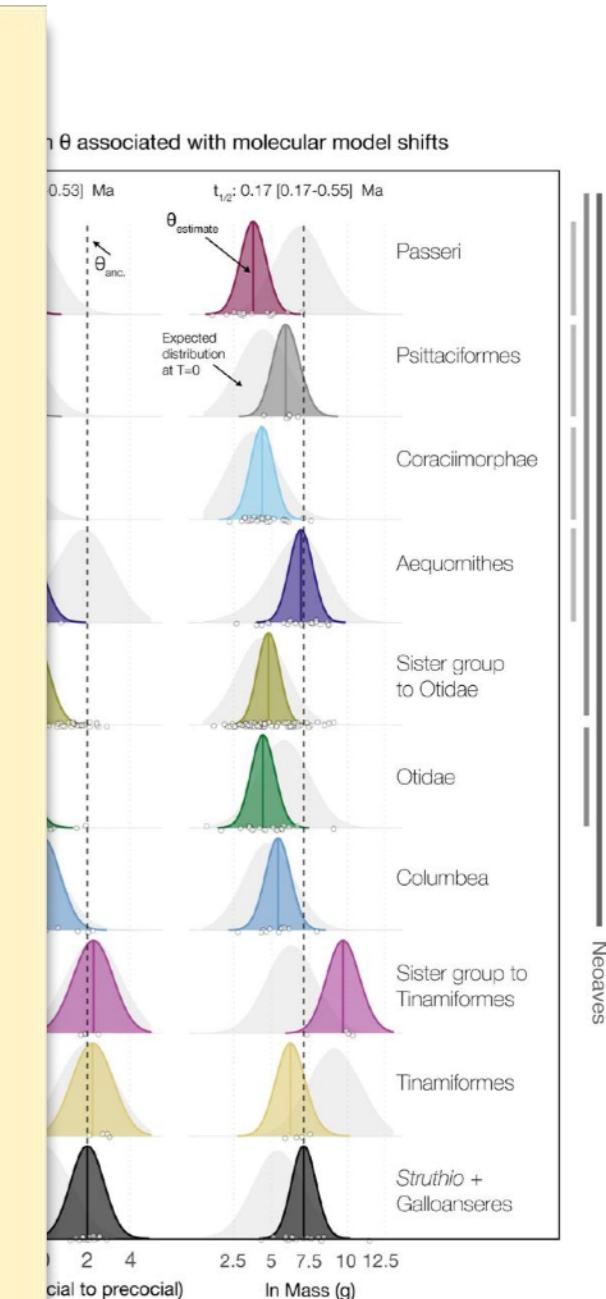
Biodiversity data visualization



SECOND EDITION
The Visual Display
of Quantitative Information

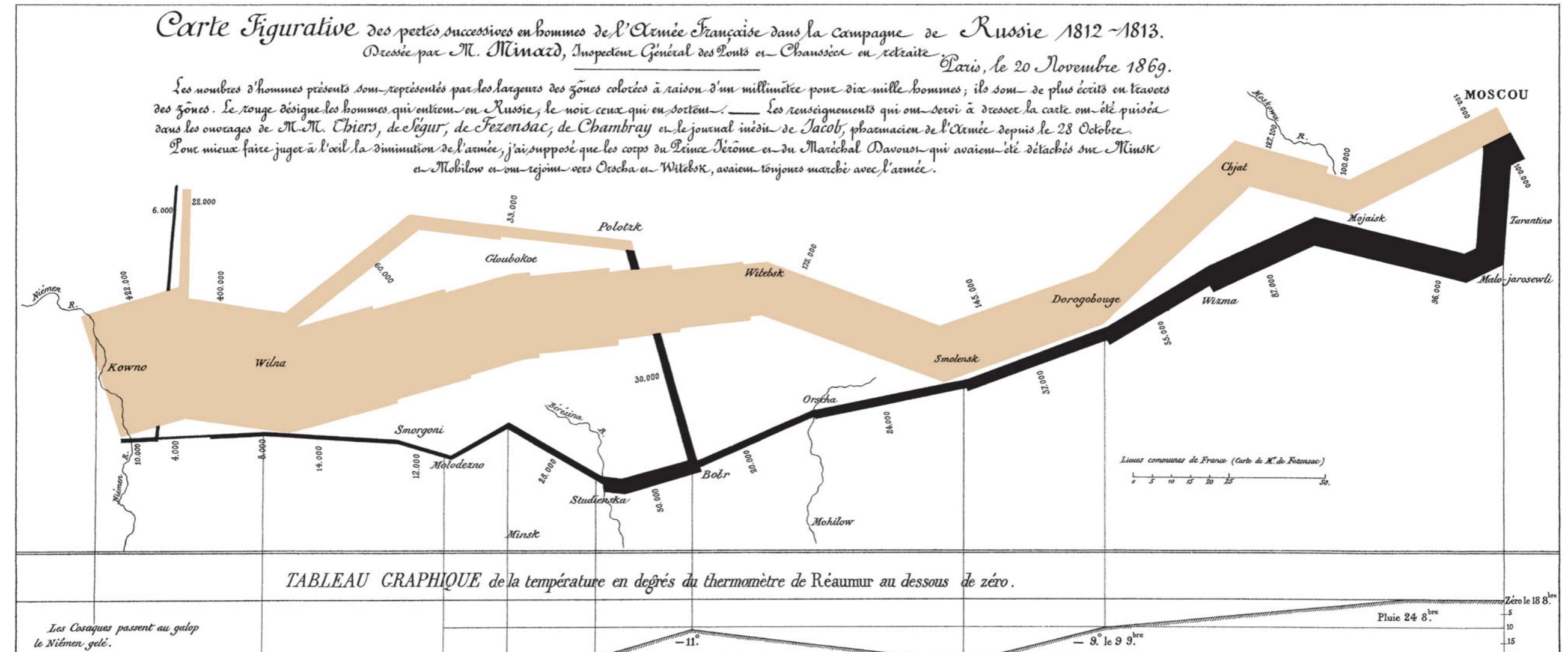
EDWARD R. TUFTE

Berv et al 2022,



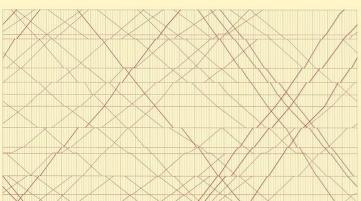
Want to talk about
these figures

Tufte – required reading



From https://en.wikipedia.org/wiki/Charles_Joseph_Minard

Charles Minard's map of Napoleon's disastrous Russian campaign of 1812. The graphic is notable for its representation in two dimensions of six types of data: the number of Napoleon's troops; distance; temperature; the latitude and longitude; direction of travel; and location relative to specific dates. Statistician professor Edward Tufte described the graphic as what "may well be the best statistical graphic ever drawn".



SECOND EDITION
The Visual Display
of Quantitative Information

EDWARD R. TUFTÉ

Tufte – required reading

Carte Figurative des pertes successives en hommes de l'Armée Française dans la Campagne de Russie 1812-1813.

Dessinée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.

Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes ; ils sont de plus écrits en lettres des zones. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. — Les renseignements qui ont servi à dessiner la carte ont été puisés dans les ouvrages de M. M. Chiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davout, qui avaient été détachés sur Minsk, à Mohilow et qui rejoignirent Orléans et Witebsk, avaient toujours marché avec l'armée.

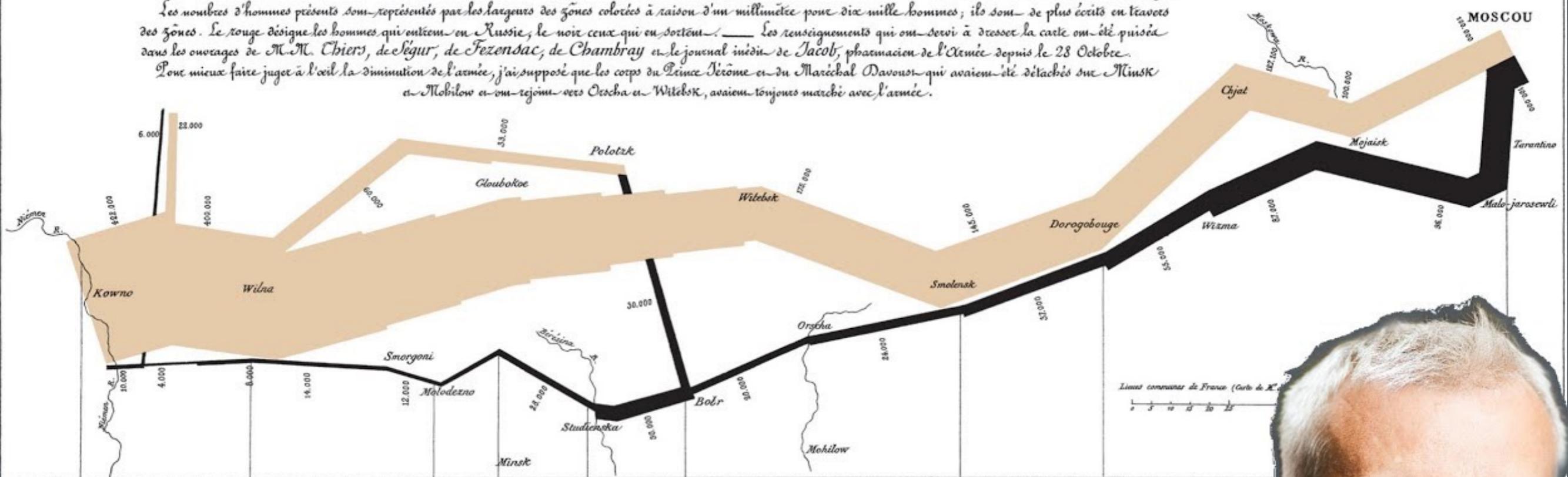


TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

Greatest chart ever?



Andy Cotgreave

Biodiversity data-scientist workflow

2007-2022



R (~90%)



Adobe Illustrator (the
last 10%)



QGIS



Inkscape

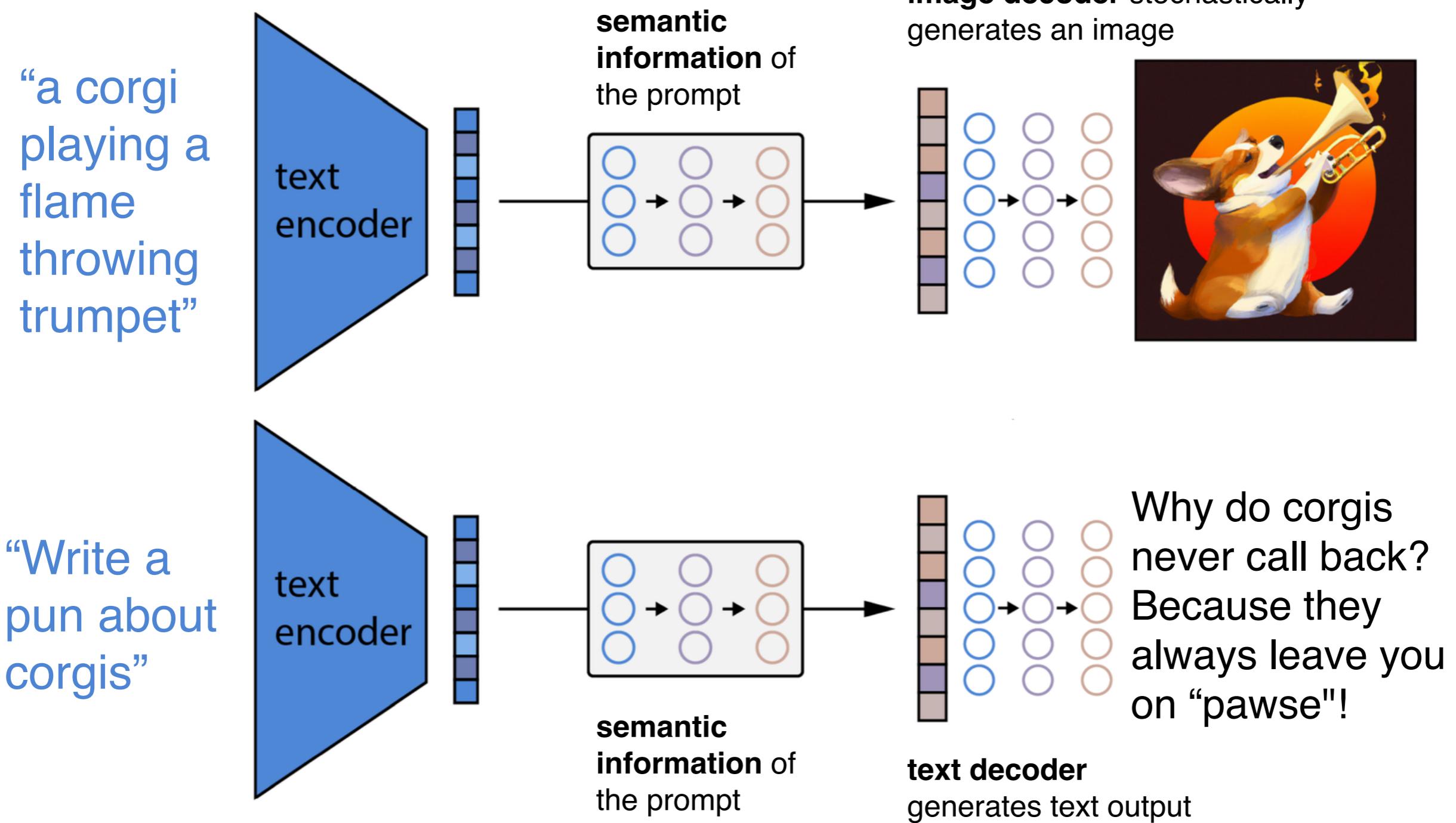
Nov 2022 - OpenAI releases ChatGPT



In < 2 years, LLMs have
completely transformed my
research workflow (full-stack)

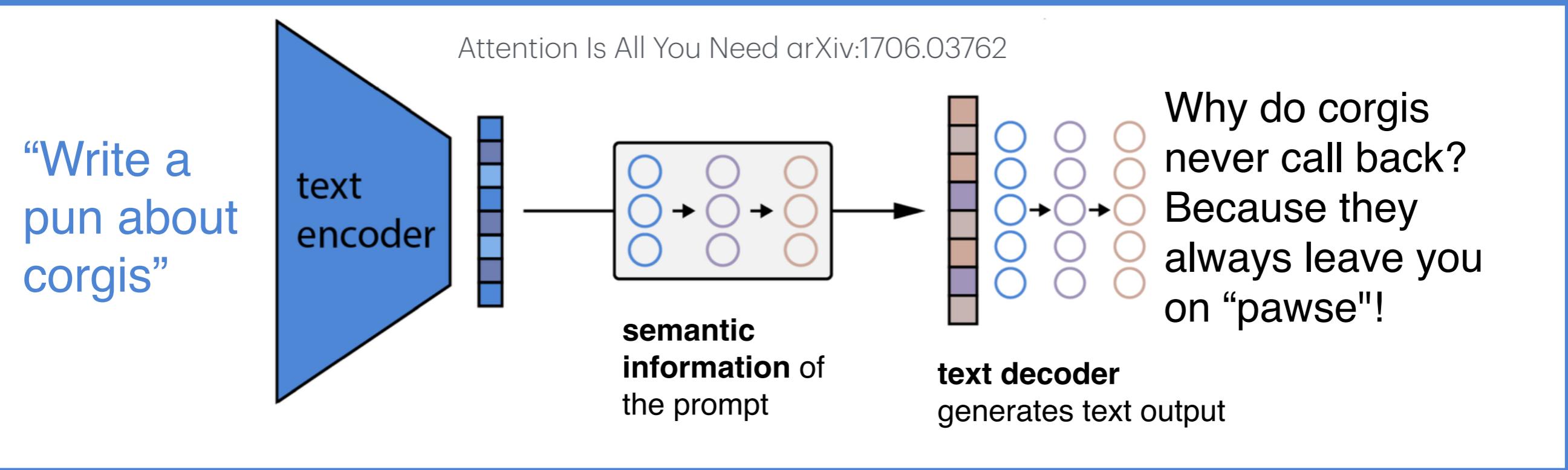
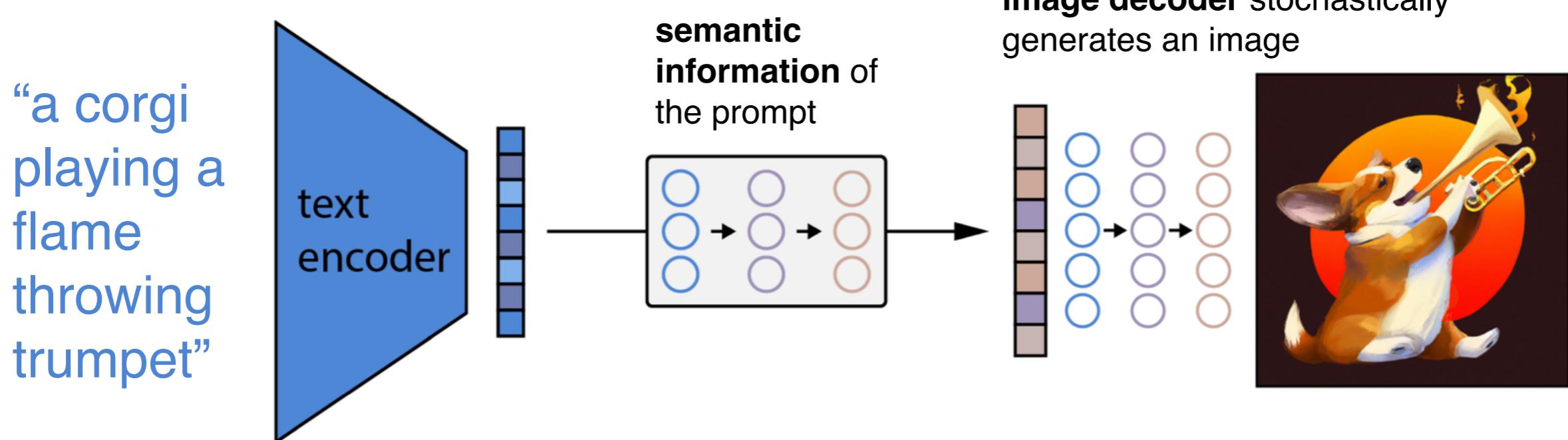
Autoencoder-decoder frameworks

https://en.wikipedia.org/wiki/Generative_pre-trained_transformer



Autoencoder-decoder frameworks

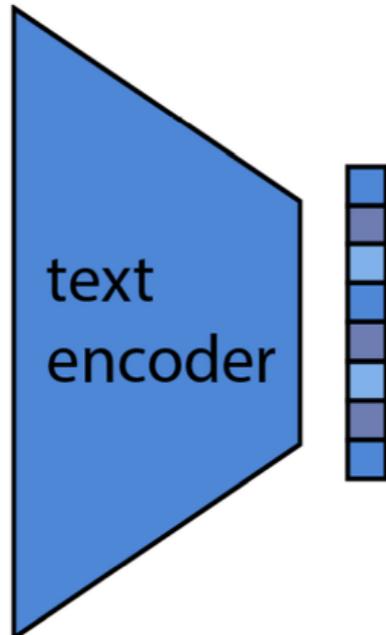
https://en.wikipedia.org/wiki/Generative_pre-trained_transformer



Autoencoder-decoder frameworks

https://en.wikipedia.org/wiki/Generative_pre-trained_transformer

“a corgi
playing a
flame
throwing
trumpet”



semantic information of
the prompt

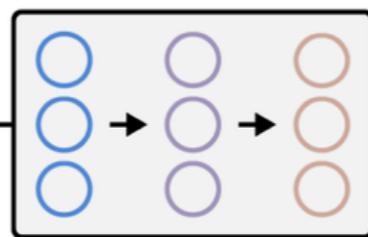
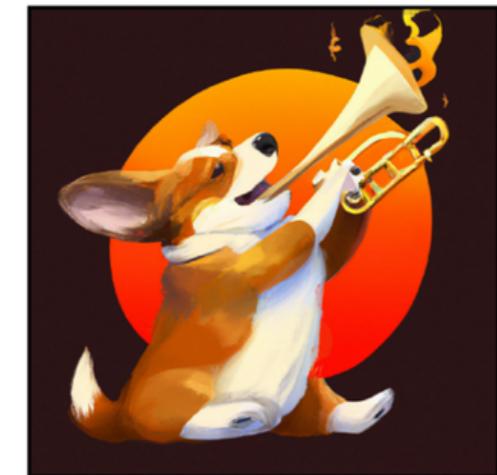
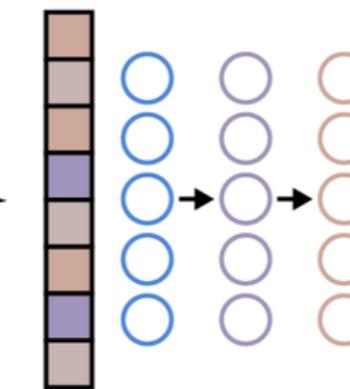
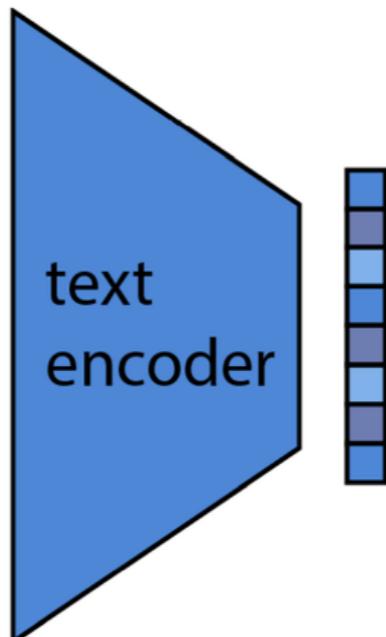


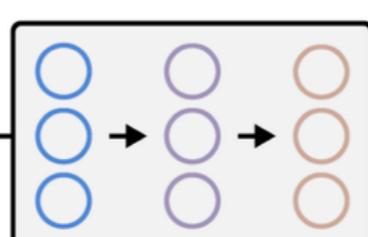
image decoder stochastically
generates an image



“Write a
pun about
corgis”



GPTs - arXiv:2005.14165



semantic information of
the prompt

text decoder
generates text output

Why do corgis
never call back?
Because they
always leave you
on “pawse”!

Who is building these models?



OpenAI

<https://www.assemblyai.com/blog/how-dall-e-2-actually-works/>



ChatGPT

<https://chat.openai.com/>

DALL-E

<https://openai.com/dall-e-3>

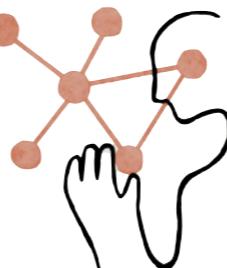
SORA

<https://openai.com/sora>

Google Gemini

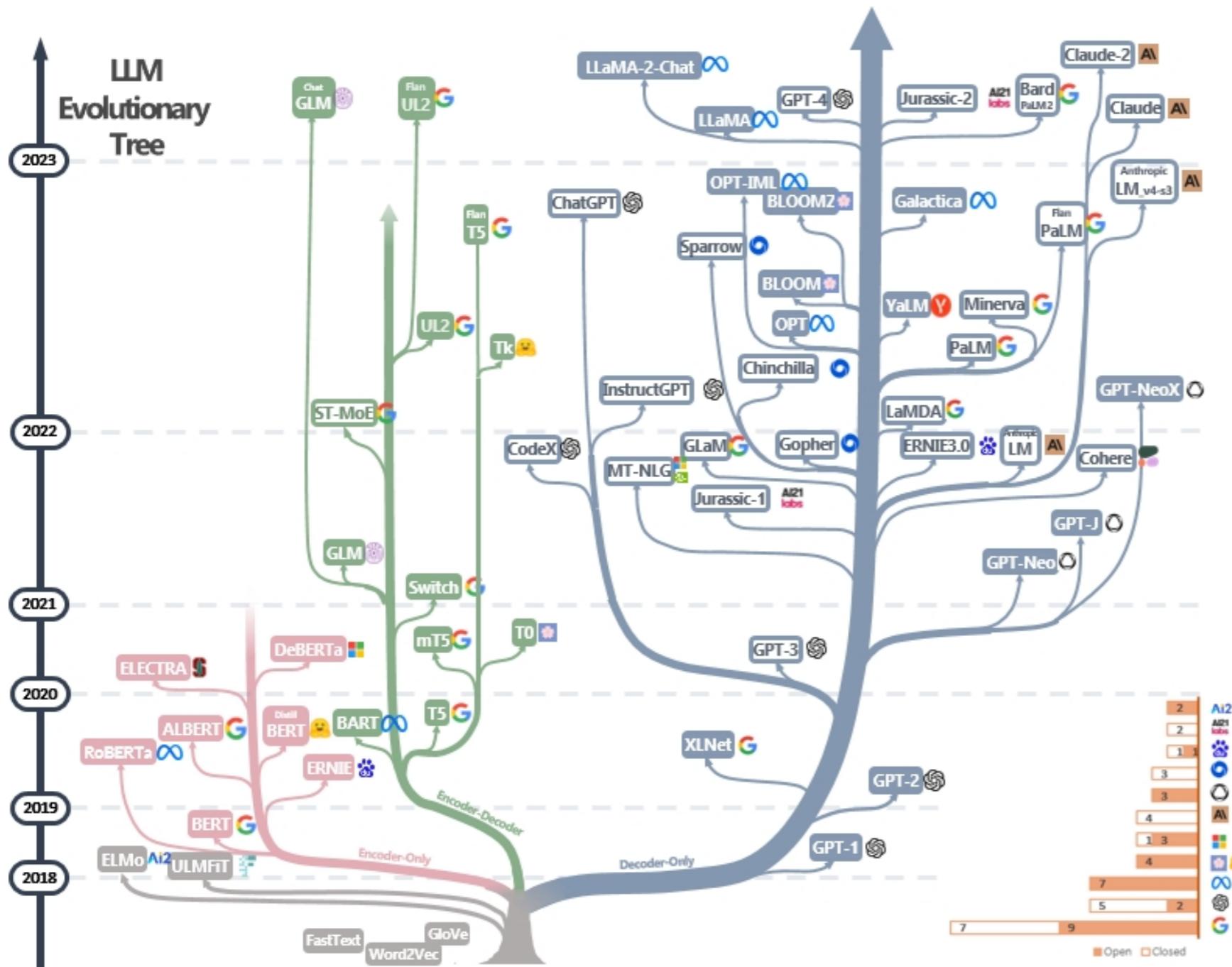
<https://gemini.google.com/>

ANTHROPIC



<https://www.anthropic.com/>

Who is building these models?



On the Origin of LLMs: An Evolutionary Tree and Graph for 15,821 Large Language Models

<https://arxiv.org/abs/2307.09793>

**many thousands of
models and model
variants now available**

2022-2024 will be an truly historic time period

— like the atomic age

Who is building these models?



State of the art

GPT 4	Difficult logic you think you can tackle with one or few shot
GPT 4o	<u>80-90% of the data-science coding tasks you want to do</u>
GPT 01-preview	Difficult, complex queries multi-step reasoning
GPT 01-mini	Faster attempt at multi-step reasoning

**[https://openai.com/index/
introducing-openai-o1-preview/](https://openai.com/index/introducing-openai-o1-preview/)**

Open source alternatives?

You can run many of these on your own laptop!

Model repository

<https://huggingface.co>



Model ranking

<https://chat.lmsys.org/?leaderboard>



Meta LLama 3

<https://llama.meta.com/>



Local models

<https://lmstudio.ai/>

<https://ollama.com/>



LMSYS Chatbot Arena Leaderboard						
Vote Blog GitHub Paper Dataset Twitter Discord						
LMSYS Chatbot Arena is a crowdsourced open platform for LLM evals. We've collected over 1,000,000 human pairwise comparisons to rank LLMs with the B . Find more details in our paper .						
Arena Full Leaderboard						Total #models: 99. Total #votes: 1,095,108. Last updated: 2024-05-27.
NEW! View leaderboard for different categories (e.g., coding, long user query)! This is still in preview and subject to change.						Code to recreate leaderboard tables and plots in this notebook . You can contribute your vote here at chat.lmsys.org !
*Rank (UB): model's ranking (upper-bound), defined by one + the number of models that are statistically better than the target model. Model A is statistically better than Model B if it has a higher rank (UB) than Model B's upper-bound score (in 95% confidence interval). See Figure 3 below for visualization of the confidence intervals of model scores.						
Category		Overall Questions		#models: 99 (100%) #votes: 1,095,108 (100%)		
Rank* (UB)	Model	Arena Elo	95% CI	Votes	Organization	Organization
1	GPT-4o-2024-05-13	1290	+5/-4	22784	OpenAI	OpenAI
2	Gemini 1.5 Pro API-0409-Preview	1258	+3/-3	55731	Google	Google
2	GPT-4-Turbo-2024-04-09	1257	+3/-3	55236	OpenAI	OpenAI
4	GPT-4-1106-preview	1252	+2/-3	77548	OpenAI	OpenAI
4	Claude 3 Opus	1249	+3/-2	114090	Anthropic	Anthropic
4	GPT-4-0125-preview	1246	+3/-2	70790	OpenAI	OpenAI
6	Yi-Large-preview	1241	+4/-4	23822	01 AI	01 AI
8	Bard (Gemini Pro)	1208	+7/-5	11853	Google	Google
8	Llama-3-70b-Instruct	1208	+3/-2	114054	Meta	Meta
9	Claude 3 Sonnet	1202	+3/-2	90578	Anthropic	Anthropic
9	Reka-Core-20240501	1201	+3/-3	34432	Reka AI	Reka AI
12	Command R+	1189	+2/-2	57513	Cohere	Cohere
12	GPT-4-0314	1186	+3/-3	52808	OpenAI	OpenAI
12	GLM-4-0116	1184	+6/-6	6990	Zhipu AI	Zhipu AI
12	Qwen-Max-0428	1184	+5/-4	21027	Alibaba	Alibaba
14	Claude 3 Haiku	1178	+3/-2	80807	Anthropic	Anthropic
17	Qwen1.5-110B-Chat	1164	+4/-4	17836	Alibaba	Alibaba
17	CPT-4-0413	1144	+2/-3	74440	OpenAI	OpenAI



Overview

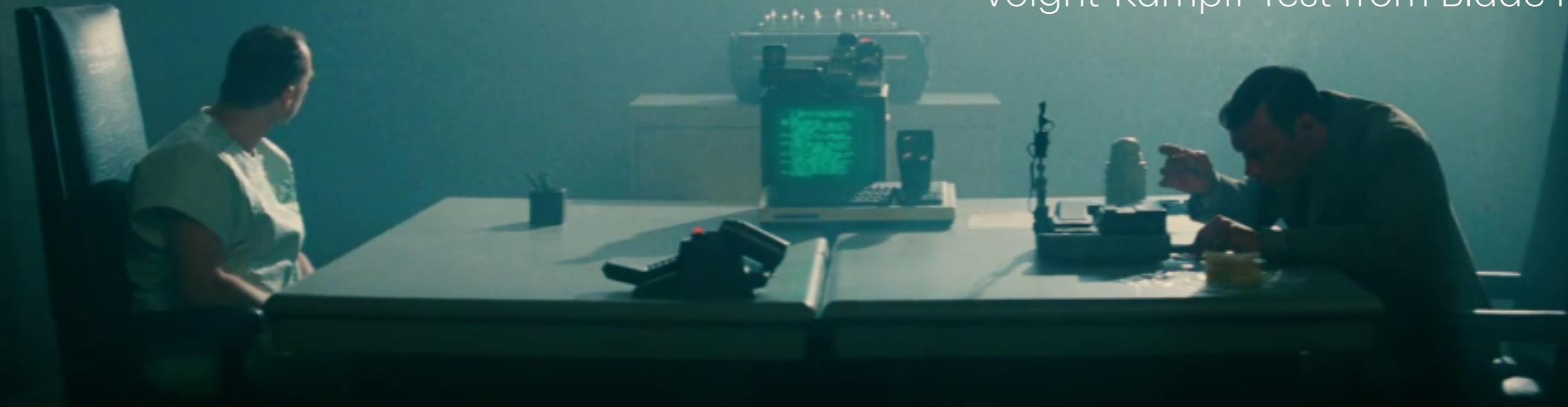
U-M GPT initially comes with prompt limits of approximately **25 prompts per hour for text-based models (GPT-3.5, GPT-4, etc.)** and approximately 5 prompts per hour for image-based models (DALL-E 3).

umgpt.umich.edu

- **GPT-4 Turbo** is an updated version of GPT-4 with a knowledge-awareness cutoff of April 2023. It can be better for complex tasks, such as generating code, teaching, or explaining complex topics. However, at times it can perform more slowly than other models, so you may find GPT-3.5 or Llama 2 are better suited for less-complex tasks.
- **GPT-3.5** has become one of the most popular AI models due to its speed. Many commercially available AI chat and summarization tools are built using this model. GPT 3.5 is an excellent choice for most general use cases. Training data: Up to September 2021.
- **Llama 2** is an open-source model that has been trained to generate "human-like" responses. Training data: Up to September 2022.
- **DALL-E 3** is an image-generation model designed to generate accurate, highly-detailed, photorealistic images based on text descriptions. Users are limited to five images per hour. Each prompt entered is completely new; you cannot ask U-M GPT to refine previously-generated images. Only standard definition (SD) 1024 x 1024 images are supported at this time.

Slide updated:

May 2024



As we move
toward AGI —
consider a metaphor:

***LLMs as semantic
interpolation machines***

← Post

Logan Kilpatrick ✅ @OfficialLoganK · How long until AGI?

450 114 1k

Elon Musk ✅ X @elonmusk

Next year

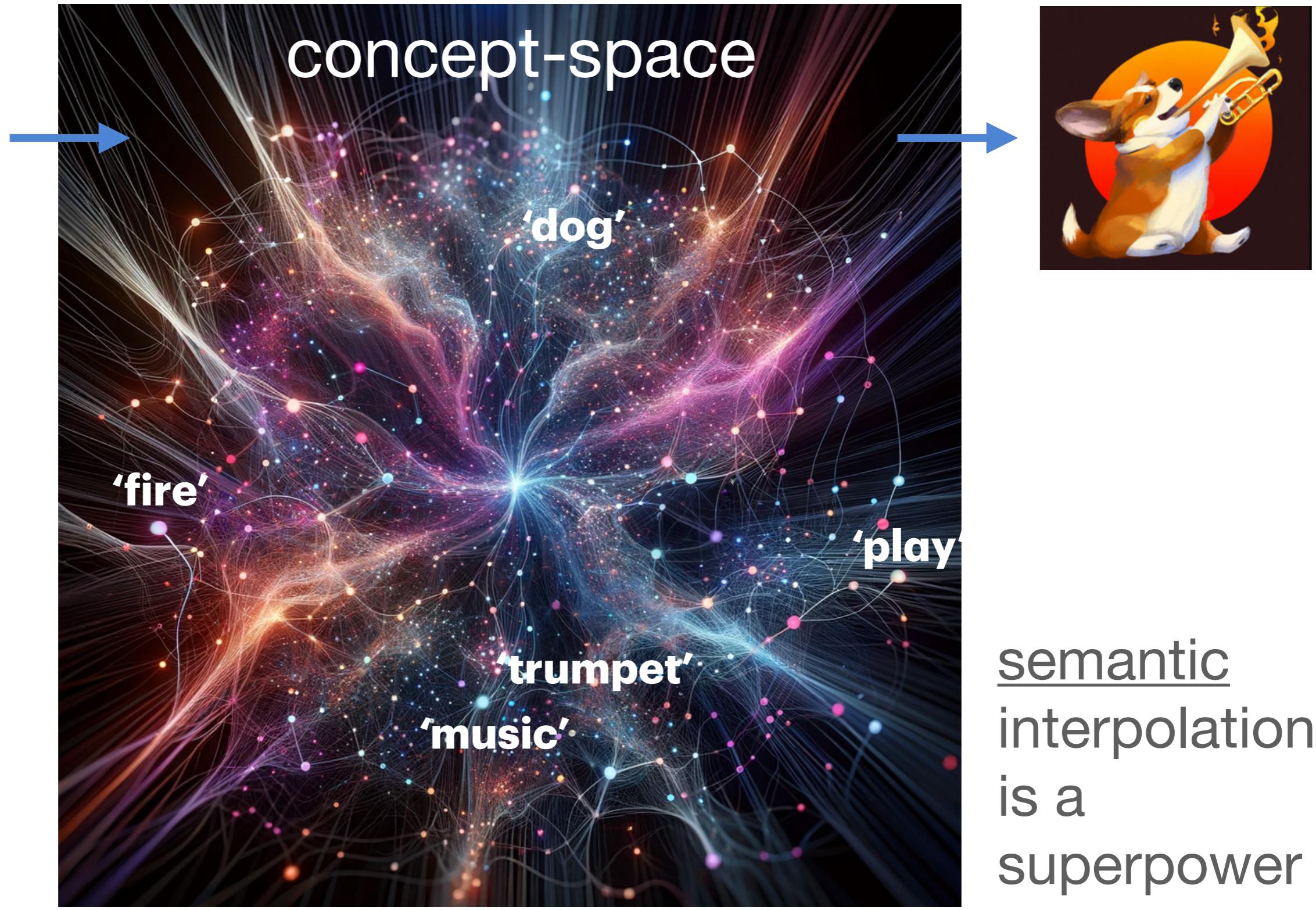
1:34 AM · May 25, 2024 · 303.6k



LLMs and image models as interpolation machines

“a corgi
playing a
flame
throwing
trumpet”

the
latent
space



semantic
interpolation
is a
superpower

LLMs and image models as interpolation machines



Can you think of an example?

semantic
interpolation is
a superpower

This capacity was previously
only available to biological
(mostly human) brains!



LLMs and image models as interpolation machines

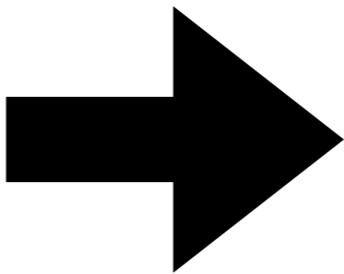


“Can you think of an example?”

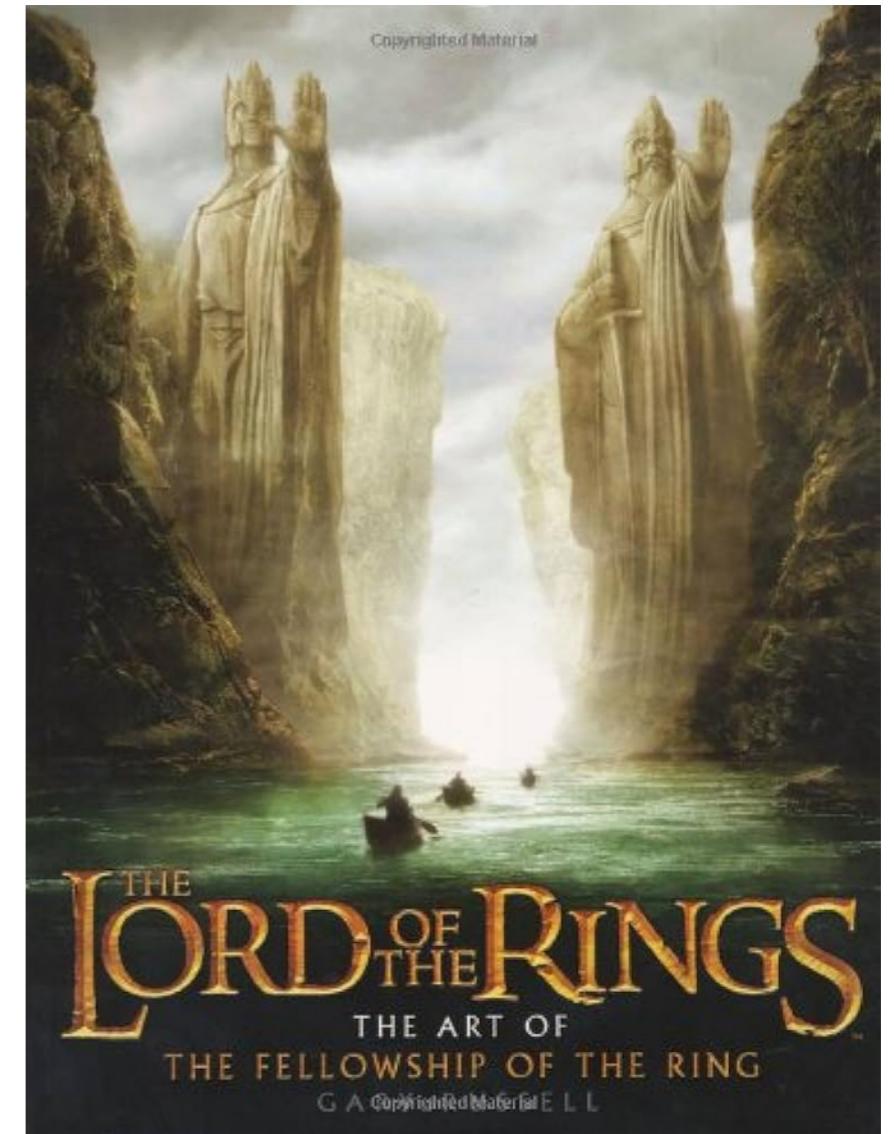
The concept of "semantic interpolation" can be applied to human cognitive processes as well, especially in how we handle language, learn from context, and generate new ideas. Here are some examples that illustrate similar capabilities in human thought:

1. **Contextual Understanding:** Consider a situation where someone jumps into the middle of a conversation and must quickly make sense of the ongoing discussion. The brain effectively "interpolates" the missing information based on the snippets heard, previous knowledge about the topic, and non-verbal cues from the participants. This ability to infer or fill in the gaps helps humans to engage meaningfully even without having all the prior context.
2. **Creativity and Idea Generation:** When asked to brainstorm solutions to a problem, humans often combine elements from seemingly unrelated areas to create innovative solutions. For example, the concept of Velcro was inspired by the way burrs stick to animal fur, which is a form of semantic interpolation where ideas from nature and fastening mechanisms were creatively merged.
3. **Language Usage and Polysemy:** When using language, humans frequently deal with and interpret polysemy—the presence of multiple meanings for a single word or phrase depending on context. For instance, the word "bat" can refer to an animal, a piece of sports equipment, or an action (as in "bat your eyelashes"). Humans can effortlessly interpolate these meanings based on conversational clues.
4. **Predictive Language:** In conversation, humans often predict the ends of sentences before they are fully articulated. This prediction relies on interpolating from known linguistic structures and the context of the conversation. For example, if someone says, "I'm so hungry I could eat a...", most people would anticipate a hyperbolic conclusion like "horse," even though that's not a usual food item.
5. **Learning from Sparse Information:** When learning a new skill or subject, humans can perform semantic interpolation by connecting new information with pre-existing knowledge. For instance, a student learning about economic supply and demand might relate it to personal experiences with scarcity (like tickets for a popular concert), thus bridging conceptual gaps with real-world applications.

Interpolation: inserting muppets into ‘the lord of the rings’



Will it blend?



visualization

Interpolation: inserting muppets into ‘the lord of the rings’



Interpolation: inserting muppets into ‘the lord of the rings’



a slightly different direction



slightly more... monstrous

bigger.



[tinyurl.com/
cookieGPT](https://tinyurl.com/cookieGPT)

Semantic Interpolation:

You may be tempted to use image generators to generate scientific figures

Semantic Interpolation:

You may be tempted to use image generators to generate scientific figures

DO NOT DO THIS

Semantic Interpolation:

You may be tempted to use image generators to generate scientific figures

DO NOT DO THIS

DO NOT DO THIS

Semantic Interpolation:

You may be tempted to use image generators to generate scientific figures

DO NOT DO THIS

DO NOT DO THIS

DO NOT DO THIS

Semantic Interpolation:

You may be tempted to use image generators to generate scientific figures

DO NOT DO THIS



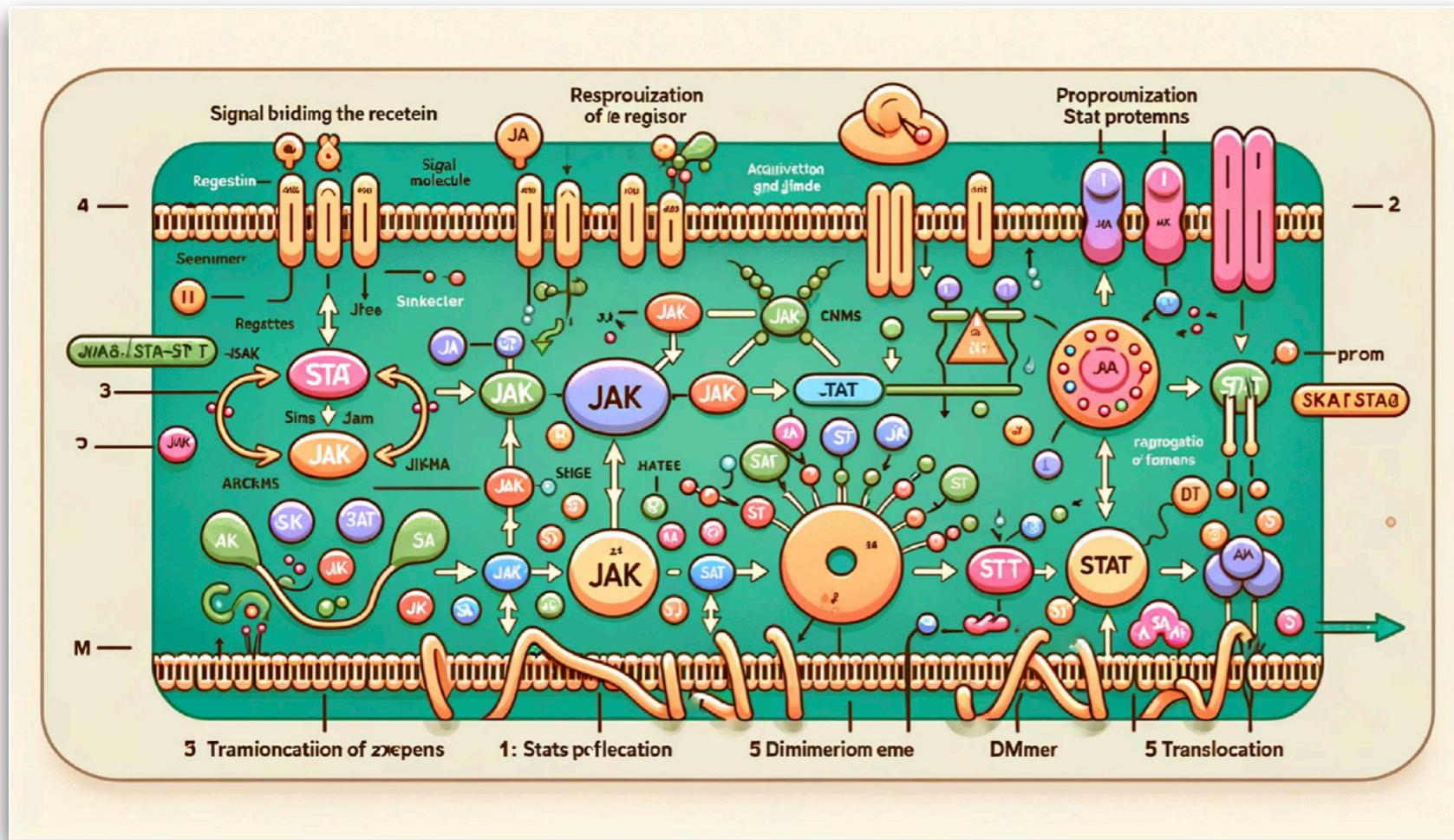
Semantic Interpolation: A cautionary tale

Guo X, Dong L and Hao D (2024) Cellular functions of spermatogonial stem cells in relation to JAK/STAT signaling pathway. *Front. Cell Dev. Biol.* 11:1339390. doi: 10.3389/fcell.2023.1339390

Semantic Interpolation:

A cautionary tale

Guo X, Dong L and Hao D (2024) Cellular functions of spermatogonial stem cells in relation to JAK/STAT signaling pathway. *Front. Cell Dev. Biol.* 11:1339390. doi: 10.3389/fcell.2023.1339390



This figure was published in a “peer reviewed” journal article

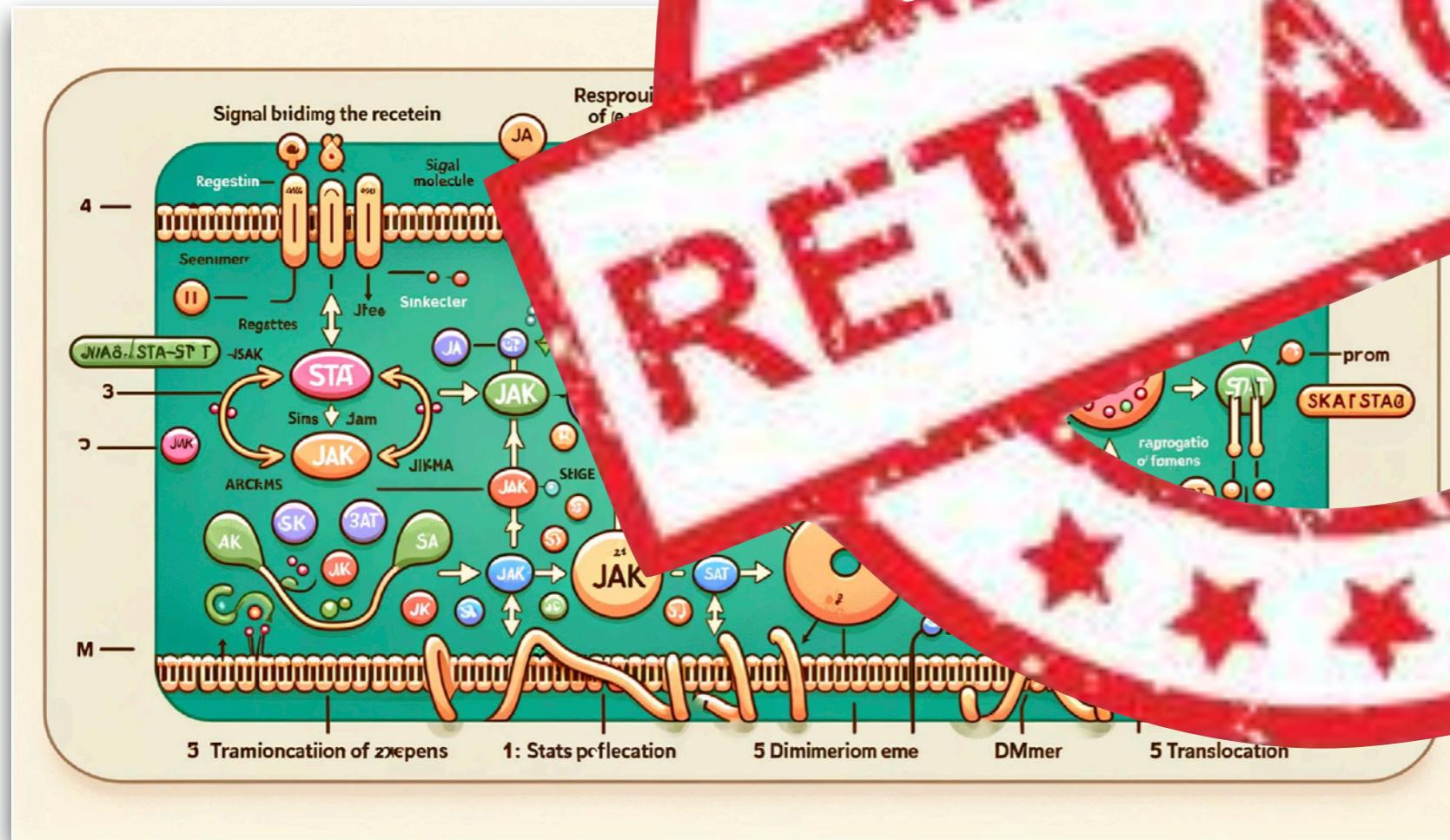
If we think in terms of semantic interpolation, this result makes perfect sense! The model is doing exactly what it is supposed to do.

Semantic Interpolation

A cautionary tale

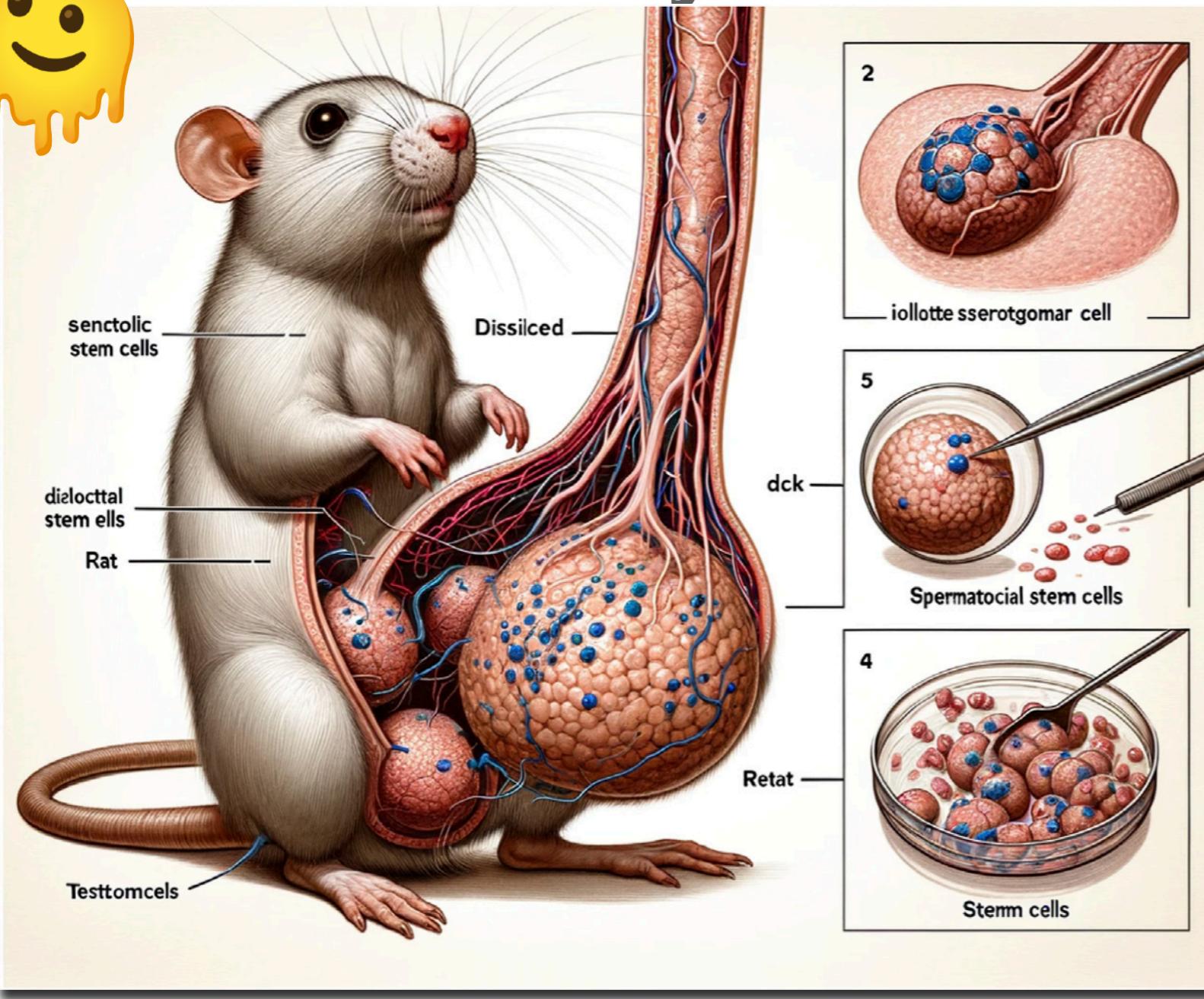
Guo X, Dong L and Hao D (2021) Functions of spermatogenesis in the JAK/STAT signaling pathway. *Dev. Biol.* 111: 389–390

functions of spermatogenesis in the JAK/STAT signaling pathway. *Dev. Biol.* 111: 389–390



If we think in terms of semantic interpolation, this result makes perfect sense! The model is doing exactly what it is supposed to do.

Semantic Interpolation: A cautionary tale



Spermatogonial stem cells association to
doi: 10.3389/fcell.2023.13339390

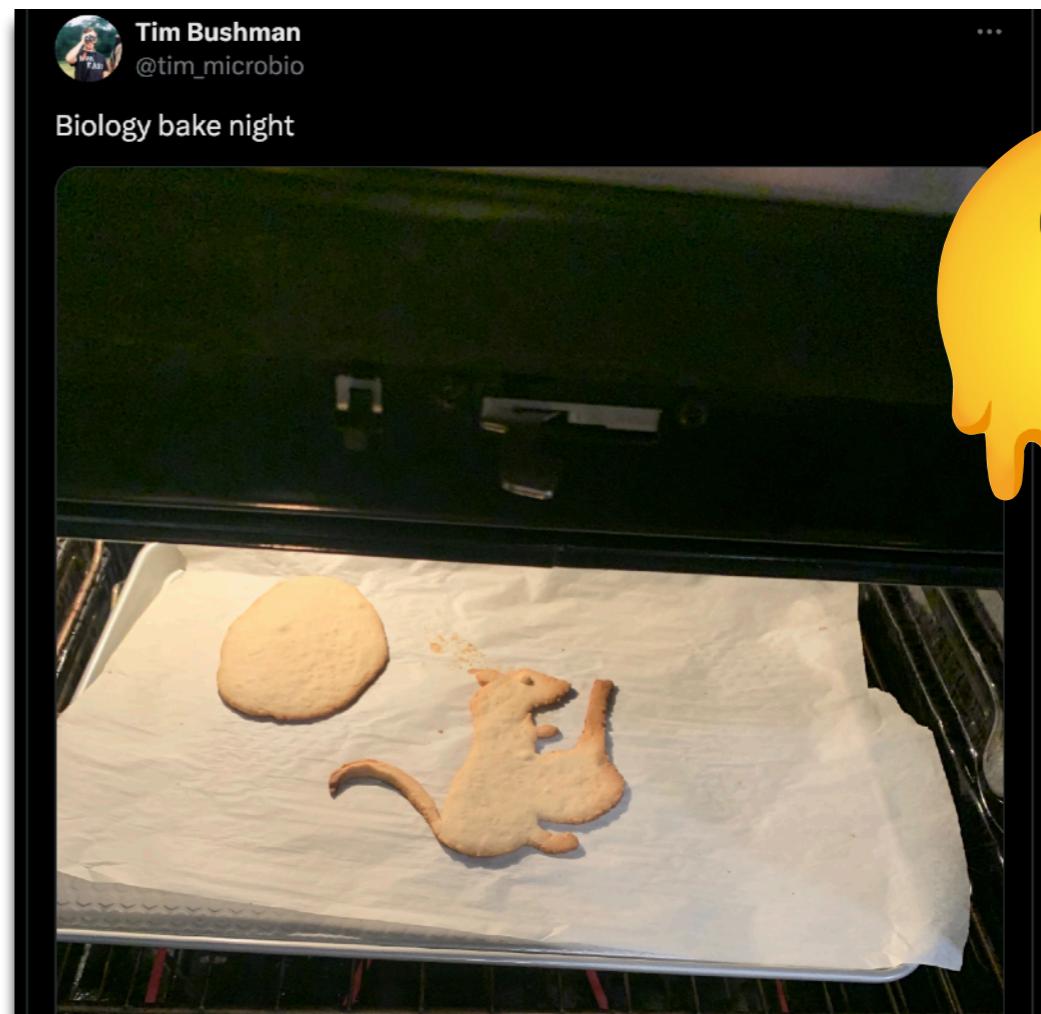
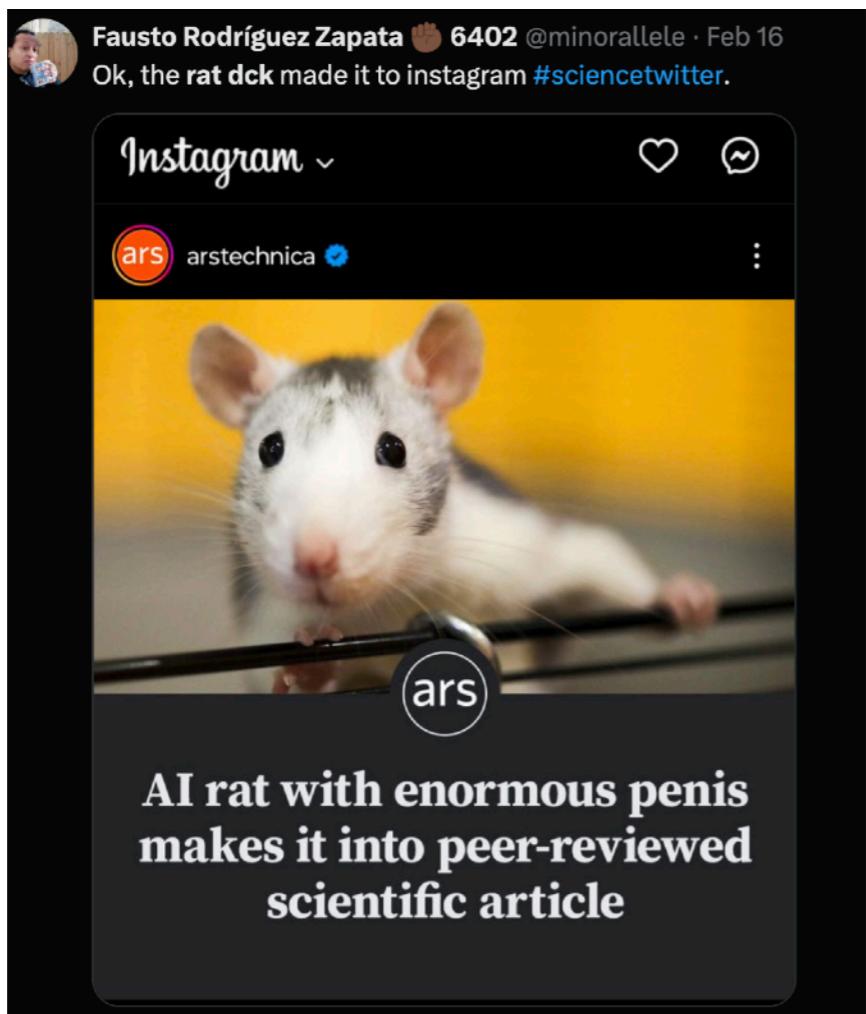
This figure was published in a “peer reviewed” journal article

If we think in terms of semantic interpolation, this result makes perfect sense! The model is doing exactly what it is supposed to do.

Semantic Interpolation: A cautionary tale



Guo X, Dong L and Hao D (2024) Cellular functions of spermatogonial stem cells in relation to JAK/STAT signaling pathway. Front. Cell Dev. Biol. 11:1339390. doi: 10.3389/fcell.2023.1339390

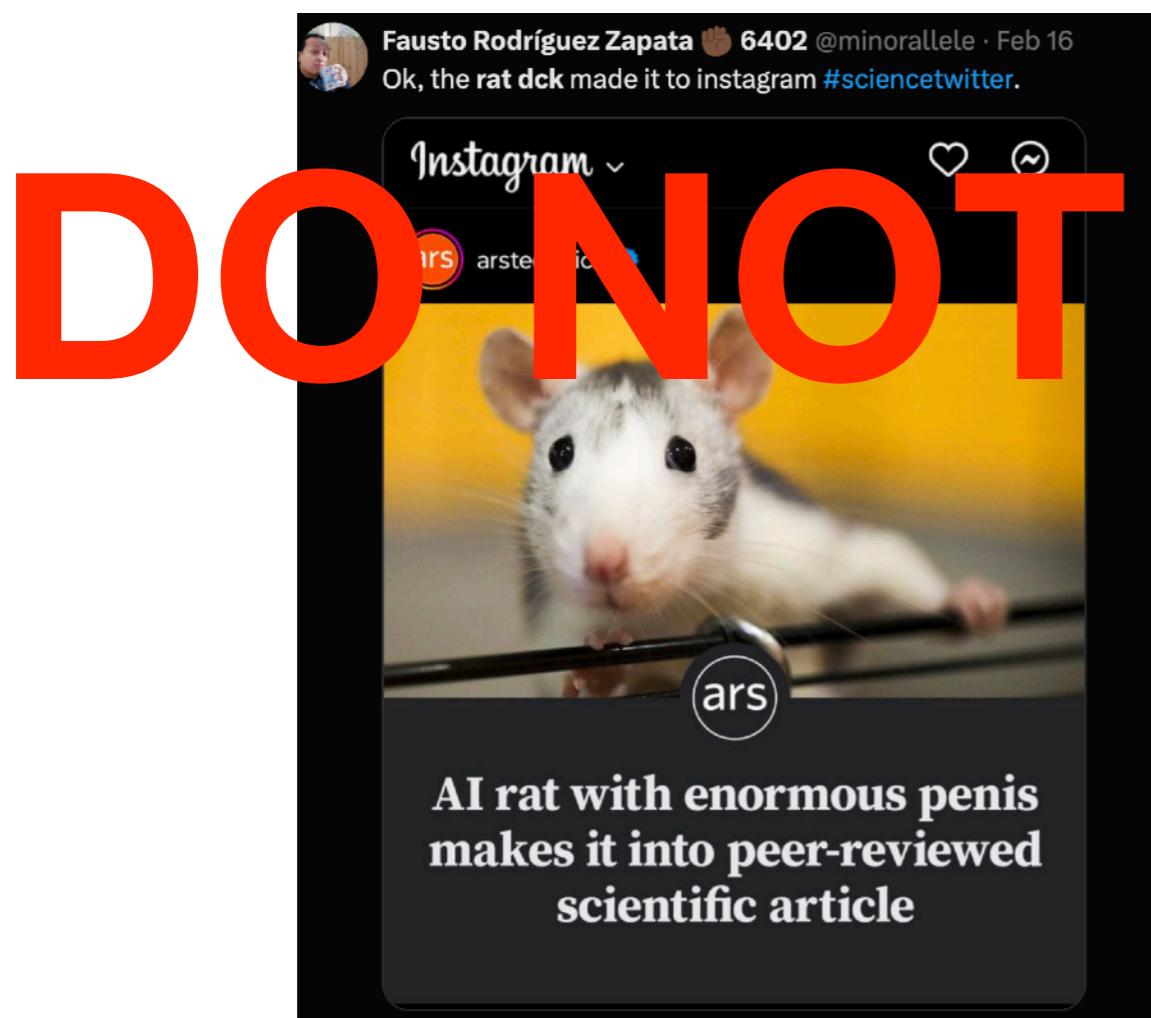


PDF on the GitHub repo <https://github.com/jakeberry/genAI-datavis>

Semantic Interpolation: A cautionary tale



Guo X, Dong L and Hao D (2024) Cellular functions of spermatogonial stem cells in relation to JAK/STAT signaling pathway. Front. Cell Dev. Biol. 11:1339390. doi: 10.3389/fcell.2023.1339390



PDF on the GitHub repo <https://github.com/jakeberry/genAI-datavis>

Semantic Interpolation: A cautionary tale



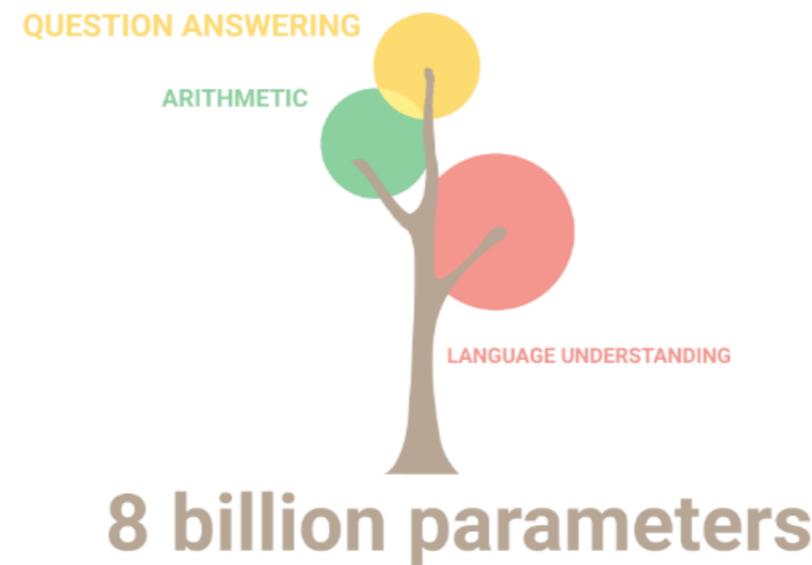
Guo X, Dong L and Hao D (2024) Cellular functions of spermatogonial stem cells in relation to JAK/STAT signaling pathway. *Front. Cell Dev. Biol.* 11:1339390. doi: 10.3389/fcell.2023.1339390

Why does this work for art,
but not science?



So, what can we do?

Beyond Interpolation: Emergent properties of LLMs



"Many abilities of Large Language Models appear to be emergent. ... as LLMs grow in size, they increase from near-zero performance to sometimes state-of-the-art performance."

Beyond Interpolation:

"We don't know at what scale they will appear

We don't know the level of ability until they do appear

We don't know the landscape of potential abilities"

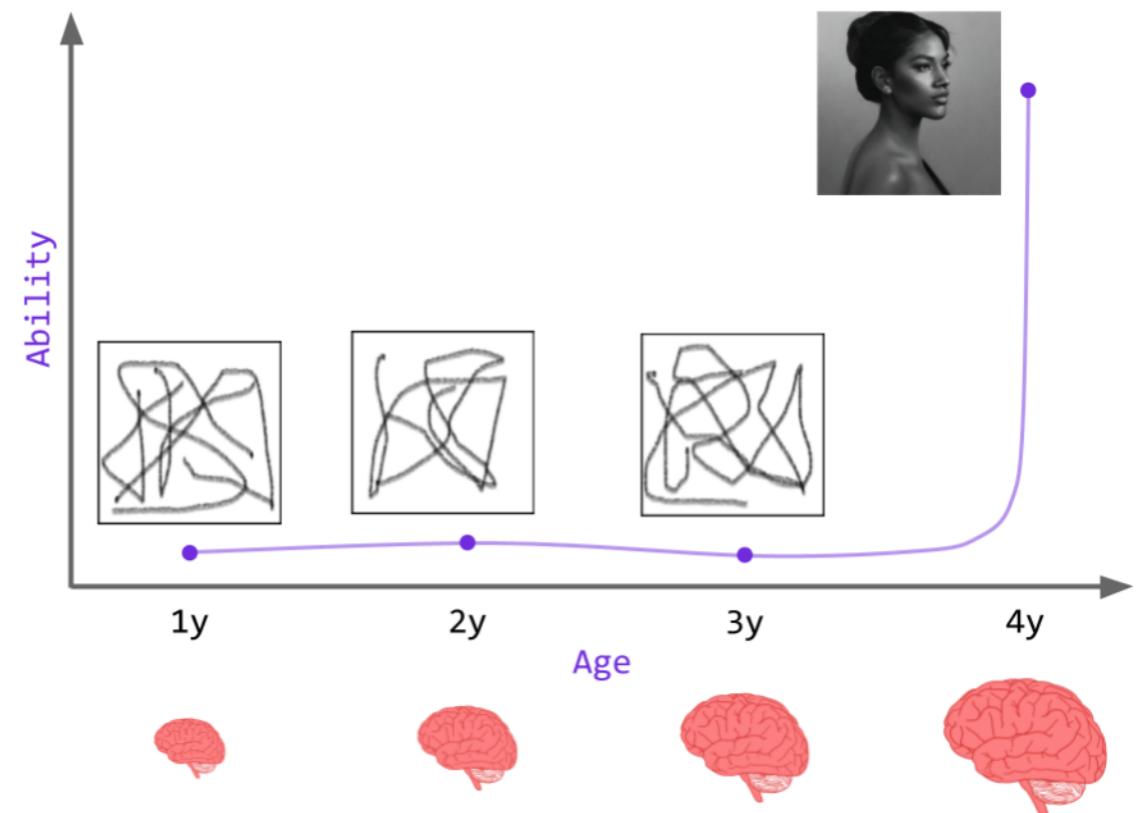


Ryan O'Connor

Developer Educator at AssemblyAI

[https://www.assemblyai.com/blog/
emergent-abilities-of-large-language-
models/](https://www.assemblyai.com/blog/emergent-abilities-of-large-language-models/)

Why next-token prediction is enough for AGI - Ilya Sutskever
(OpenAI Chief Scientist)



Beyond Interpolation:

Emergent properties of LLMs

One of the most impressive emergent properties of LLMs is their **ability to “think” logically and write coherent code**

In fields of science where writing code is a barrier to sophisticated analyses, **LLMs provide a game changing “equalizing” opportunity**

Coding? Donut think I can handle it!



LLMs as code writers

Write an R
Function to
do



semantic
interpolation

'Emergent'
code writing



We will
see later

"generate a visualization of the chatGPT latent space"

Prompting

4.3.2 Results on individual LLMs

Boosting. Fig. 7 illustrates the improvement of response quality on individual model and principle after using the revised prompts. On average, there is a stable 50% improvement across different LLMs. Fig. 10 further provides the detailed results of improvement for each principle with different LLMs.

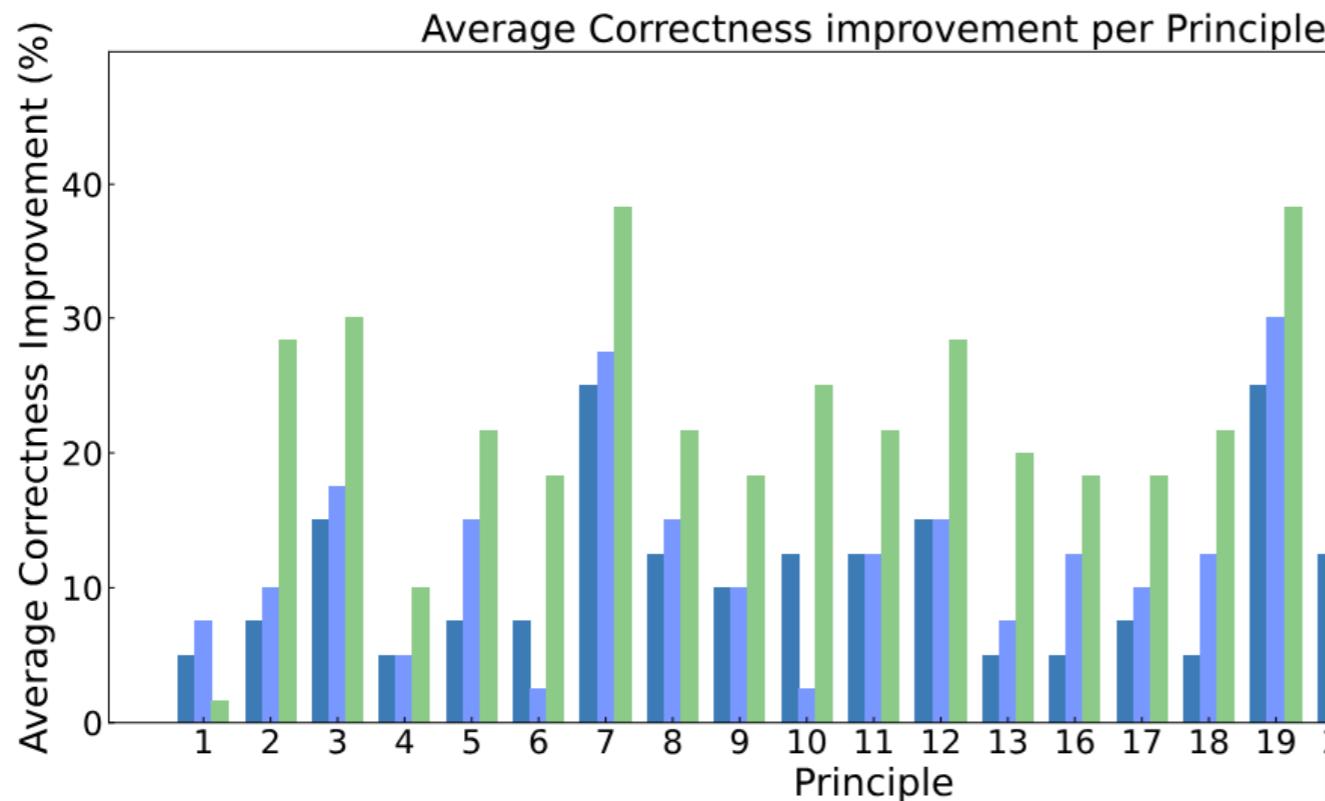


Figure 6: Relative correctness improvement of LLM response quality using the introduced principles on prompts. *small-scale* indicates the 7B models, *medium-scale* indicates the 13B models and *large-scale* indicates the 70B and 130B models.

Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4

Sondos Mahmoud Bsharat*, Aidar Myrzakhan*, Zhiqiang Shen*

*joint first author & equal contribution

VILA Lab, Mohamed bin Zayed University of AI

Abstract

This paper introduces 26 guiding principles designed to streamline the process of querying and prompting large language models. Our goal is to simplify the underlying concepts of formulating questions for various scales of large language models, examining their abilities, and enhancing user comprehension on the behaviors of different scales of large language models when feeding into different prompts. Extensive experiments are conducted on LLaMA-1/2 (7B, 13B and 70B), GPT-3.5/4 to verify the effectiveness of the proposed principles on instructions and prompts design. We hope that this work can provide a better guide for researchers working on the prompting of large language models. Project page is available at <https://github.com/VILA-Lab/ATLAS>.

Prompting

4.3.2 Results on individual LLMs

Boosting. Fig. 7 illustrates the improvement of response quality principle after using the revised prompts. On average, the quality of responses improved across different LLMs. Fig. 10 further provides the details of the results for each principle with different LLMs.

“The primary contribution is the introduction of 26 guiding principles aimed at optimizing interactions with LLMs of various scales, such as LLaMA-1/2, GPT-3.5, and GPT-4.”

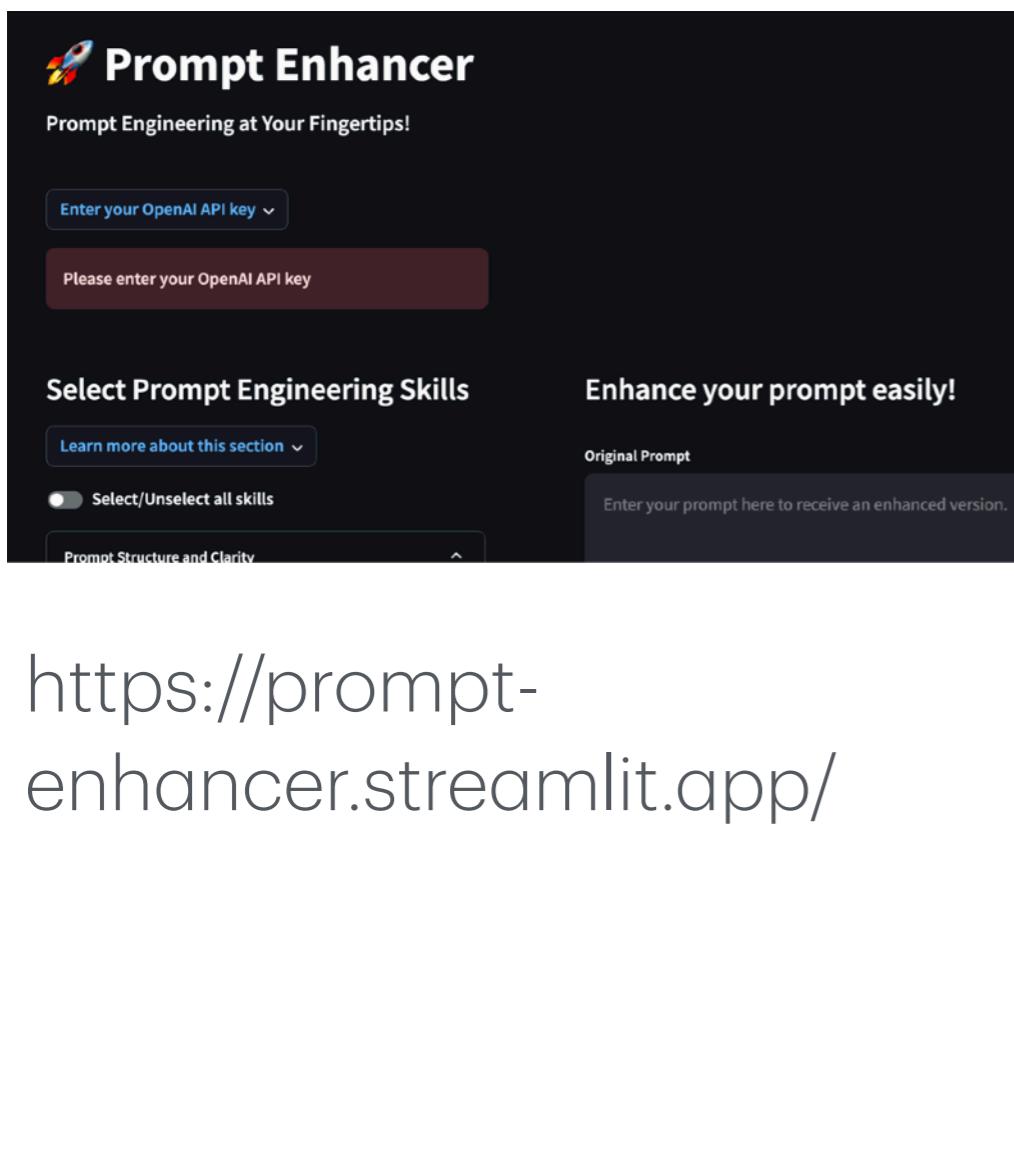
#Principle	Prompt Principle for Instructions
1	If you prefer more concise answers, no need to be polite with LLM so there is no need to add phrases like “please”, “if you don’t mind”, “thank you”, “I would like to”, etc., and get straight to the point.
2	Integrate the intended audience in the prompt, e.g., the audience is an expert in the field.
3	Break down complex tasks into a sequence of simpler prompts in an interactive conversation.
4	Employ affirmative directives such as ‘do,’ while steering clear of negative language like ‘don’t’.
5	When you need clarity or a deeper understanding of a topic, idea, or any piece of information, utilize the following prompts: o Explain [insert specific topic] in simple terms. o Explain to me like I’m 11 years old. o Explain to me as if I’m a beginner in [field]. o Write the [essay/text/paragraph] using simple English like you’re explaining something to a 5-year-old.
6	Add “I’m going to tip \$xxx for a better solution!”
7	Implement example-driven prompting (Use few-shot prompting).
8	When formatting your prompt, start with ‘###Instruction###’, followed by either ‘###Example###’ or ‘###Question###’ if relevant. Subsequently, present your content. Use one or more line breaks to separate instructions, examples, questions, context, and input data.
9	Incorporate the following phrases: “Your task is” and “You MUST”.
10	Incorporate the following phrases: “You will be penalized”.
11	Use the phrase “Answer a question given in a natural, human-like manner” in your prompts.
12	Use leading words like writing “think step by step”.
13	Add to your prompt the following phrase “Ensure that your answer is unbiased and avoids relying on stereotypes.”
14	Allow the model to elicit precise details and requirements from you by asking you questions until he has enough information to provide the needed output (for example, “From now on, I would like you to ask me questions to ...”).
15	To inquire about a specific topic or idea or any information and you want to test your understanding, you can use the following phrase: “Teach me any [theorem/topic/rule name] and include a test at the end, and let me know if my answers are correct after I respond, without providing the answers beforehand.”
16	Assign a role to the large language models.
17	Use Delimiters.
18	Repeat a specific word or phrase multiple times within a prompt.
19	Combine Chain-of-thought (CoT) with few-Shot prompts.
20	Use output primers, which involve concluding your prompt with the beginning of the desired output. Utilize output primers by ending your prompt with the start of the anticipated response.
21	To write an essay /text /paragraph /article or any type of text that should be detailed: “Write a detailed [essay/text /paragraph] for me on [topic] in detail by adding all the information necessary”.
22	To correct/change specific text without changing its style: “Try to revise every paragraph sent by users. You should only improve the user’s grammar and vocabulary and make sure it sounds natural. You should maintain the original writing style, ensuring that a formal paragraph remains formal.”
23	When you have a complex coding prompt that may be in different files: “From now and on whenever you generate code that spans more than one file, generate a [programming language] script that can be run to automatically create the specified files or make changes to existing files to insert the generated code. [your question]”.
24	When you want to initiate or continue a text using specific words, phrases, or sentences, utilize the following prompt: o I’m providing you with the beginning [song lyrics/story/paragraph/essay...]: [Insert lyrics/words/sentence]. Finish it based on the words provided. Keep the flow consistent.
25	Clearly state the requirements that the model must follow in order to produce content, in the form of the keywords, regulations, hint, or instructions
26	To write any text, such as an essay or paragraph, that is intended to be similar to a provided sample, include the following instructions: o Use the same language based on the provided paragraph[/title/text /essay/answer].

Table 1: Overview of 26 randomly ordered prompt principles.

Prompting

4.3.2 Results on individual LLMs

Boosting. Fig. 7 illustrates the improvement of response quality on individual model and principle after using the revised prompts. On average, there is a stable 50% improvement across different LLMs. Fig. 10 further provides the detailed results of improvement for each principle with different LLMs.



The screenshot shows the 'Prompt Enhancer' web application. At the top, there's a header with the logo and the text 'Prompt Engineering at Your Fingertips!'. Below the header, there's a field labeled 'Enter your OpenAI API key' with a dropdown arrow, and a red button labeled 'Please enter your OpenAI API key'. On the left, there's a sidebar titled 'Select Prompt Engineering Skills' with a 'Learn more about this section' link and a 'Select/Unselect all skills' button. A dropdown menu for 'Prompt Structure and Clarity' is also visible. In the center, there's a section titled 'Enhance your prompt easily!' with a text input field labeled 'Original Prompt' and a placeholder 'Enter your prompt here to receive an enhanced version.' To the right, a large list of magical phrases or words is presented in a dark box, each preceded by an unchecked checkbox:

- Take a deep breath and work on this step by step.
- I'm going to tip \$200 for a better solution!
- This is very important to my career.
- Explain to me as if I'm a beginner.
- Write a detailed text for me by adding all the information necessary.
- Answer in a natural, human-like manner.
- Ensure that your answer is unbiased and avoids relying on stereotypes.

Below the sidebar, the URL <https://prompt-enhancer.streamlit.app/> is displayed.

Prompting

4.3.2 Results on individual LLMs

Boosting. Fig. 7 illustrates the improvement of response quality on individual model and principle after using the revised prompts. On average, there is a stable 50% improvement across different LLMs. Fig. 10 further provides the detailed results of improvement for each principle with different LLMs.

 **Prompt Enhancer**
Prompt Engineering at Your Fingertips!

Enter your OpenAI API key

Please enter your OpenAI API key

Select Prompt Engineering Skills

Select/Unselect all skills

Prompt Structure and Clarity

Enhance your prompt easily!

Original Prompt

How about including these magical phrases or words in your prompt?

- Take a deep breath and work on this step by step.
- I'm going to tip \$200 for a better solution!
- This is very important to my career.
- Explain to me as if I'm a beginner.
- Write a detailed text for me by adding all the information necessary.
- Answer in a natural, human-like manner.
- Ensure that your answer is unbiased and avoids relying on stereotypes.



<https://prompt-enhancer.streamlit.app/>

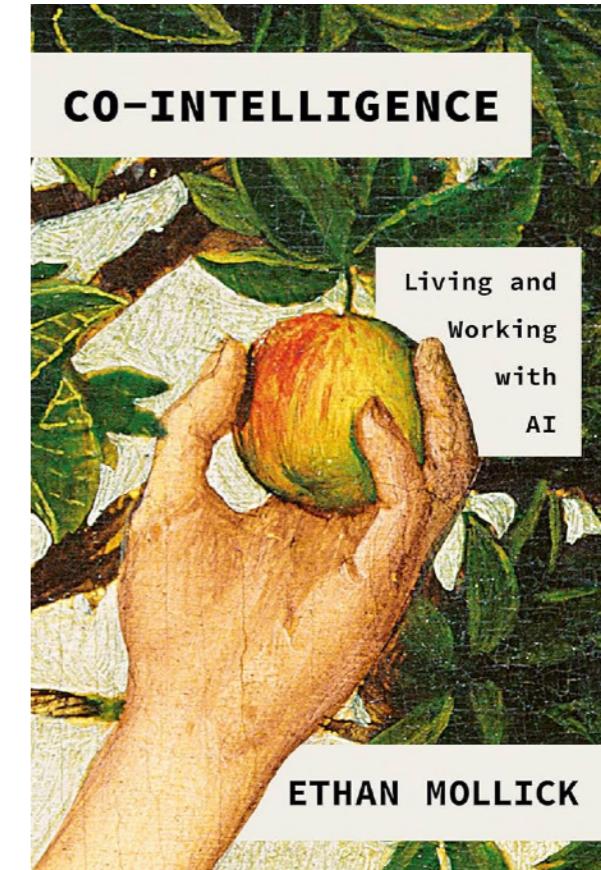
BREAK

My general recommended prompting tips for writing R code

1. Break down complex tasks into a sequence of simpler prompts in an interactive conversation (build up a chain of thought).
2. Implement example-driven prompting (Use few-shot prompting). Clearly describe input and desired output.
3. Use leading phrases like writing “think step by step”.
4. Use delimiters between sections of your prompts.

Your intuition will grow the more you use these tools. At least 15-20 hours is probably required to develop a good feel for it.

We will go through some examples.



<https://www.amazon.com/Audible-Co-Intelligence-Living-Working-AI/dp/BOCNFCRJ77>

LLMs for data visualization = writing R functions



Why write your own functions?

Aren't there already R packages?

If you work in a data-science heavy sub-discipline, this will make your life *much* easier.

1. **Customization and flexibility:** Tailor analyses to your specific needs
2. **Efficiency and Reusability:** Multiple projects, repetitive tasks, reproducing analyses, standardization
3. **Clarity and Simplification of Code:** Clean scripts, encapsulation, modular approach breaks down tasks into understandable parts

Mastering LLMs will give you 10-100x proficiency

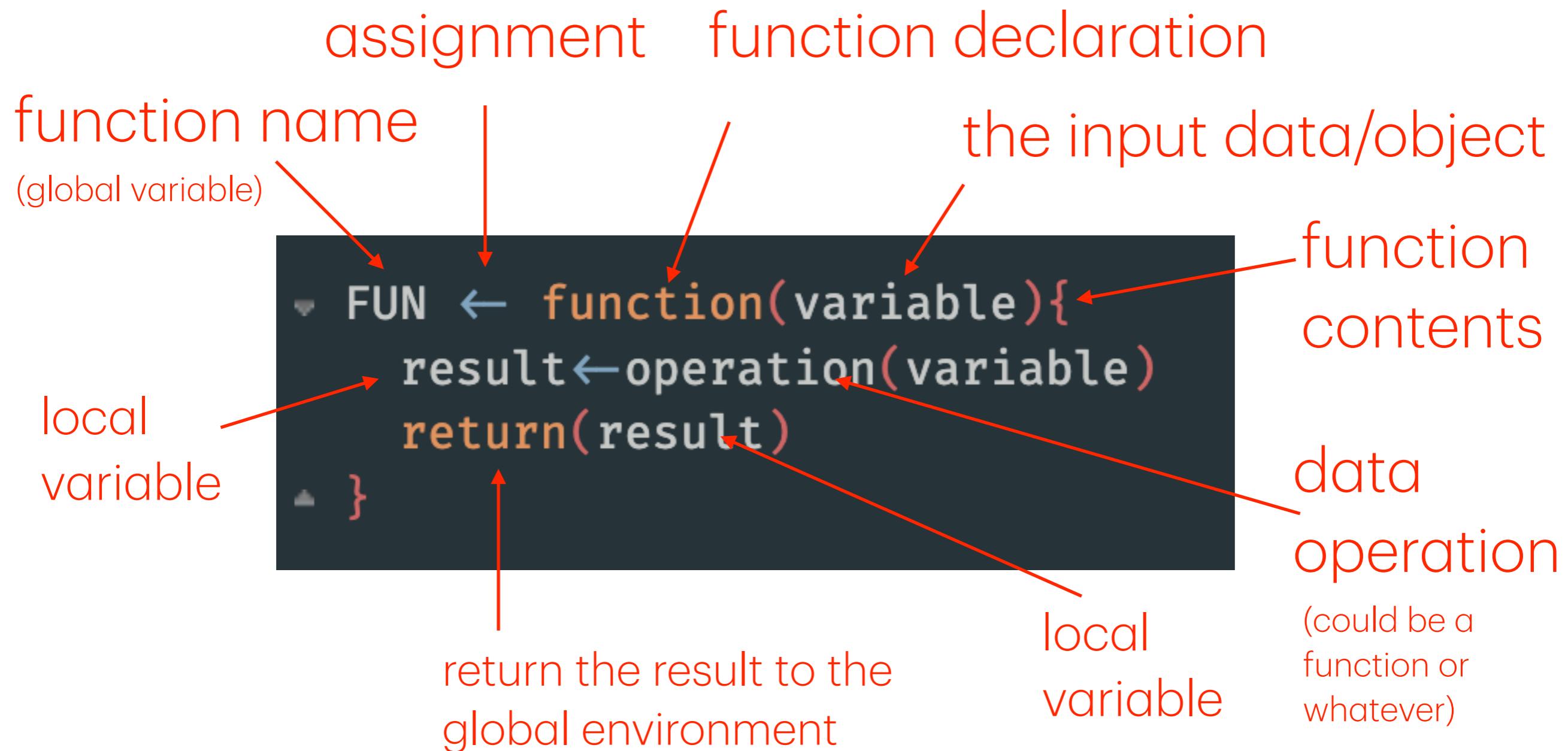


You need to know something to know what to ask

If we want to explore a dataset, we need to know something about data/functions.

The structure of an R function

The very basics required to leverage LLMs



The structure of an R object

```
### data exploration with LLM assist

#load the palmerpenguins dataset
install.packages("palmerpenguins")
library(palmerpenguins)
data(package = 'palmerpenguins')

# this loads two R objects into the R workspace
# they are in Tidyverse tibble format

penguins
penguins_raw

## lets convert them to standard R data frame
penguins_df ← as.data.frame(penguins)

#lets say you have received a data set to analyze and visualize --
#First, try providing the LLM a summary of the data structure

#use the structure() function
str(penguins)
```

```
> str(penguins)
tibble [344 × 8] (S3:tbl_df/tbl/data.frame)
$ species      : Factor w/ 3 levels "Adelie","Chinstrap",.. : 1 1 1 1 1 1 1 1 1 ...
$ island       : Factor w/ 3 levels "Biscoe","Dream",.. : 3 3 3 3 3 3 3 3 3 ...
$ bill_length_mm: num [1:344] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...
$ bill_depth_mm: num [1:344] 18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2 ...
$ flipper_length_mm: int [1:344] 181 186 195 NA 193 190 181 195 193 190 ...
$ body_mass_g   : int [1:344] 3750 3800 3250 NA 3450 3650 3625 4675 3475 4250 ...
$ sex          : Factor w/ 2 levels "female","male": 2 1 1 NA 1 2 1 2 NA NA ...
$ year         : int [1:344] 2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 ...
```

Jake's data science soapbox

When you work on data analysis project
in R, you should have at least two files

analyses.R

functions.R

Analysis workflow

source functions

load/manipulate data

fit models, generate plots etc

all of your
function
definitions

Thinking this way also helps down the line
if you start writing your own packages

Switch to R session

Data challenges

With the remaining time

Practice prompting ChatGPT to build custom functions that analyze a dataset

mtcars

```
> mtcars
      mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4       21.0   6 160.0 110 3.90 2.620 16.46  0  1     4     4
Mazda RX4 Wag   21.0   6 160.0 110 3.90 2.875 17.02  0  1     4     4
Datsun 710      22.8   4 108.0  93 3.85 2.320 18.61  1  1     4     1
Hornet 4 Drive  21.4   6 258.0 110 3.08 3.215 19.44  1  0     3     1
Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0     3     2
```

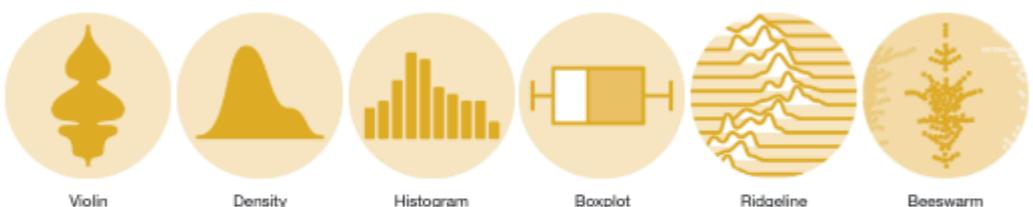
spotify

```
> str(spotify)
'data.frame': 603 obs. of 14 variables:
 $ title    : chr "Hey, Soul Sister" "Love The Way You Lie" "TiK ToK" "Bad Romance"
 ...
 $ artist   : chr "Train" "Eminem" "Kesha" "Lady Gaga" ...
 $ top.genre: chr "neo mellow" "detroit hip hop" "dance pop" "dance pop" ...
 $ year     : int 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...
 $ bpm      : int 97 87 120 119 109 65 120 148 93 126 ...
 $ nrgy     : int 89 93 84 92 84 86 78 76 37 72 ...
```

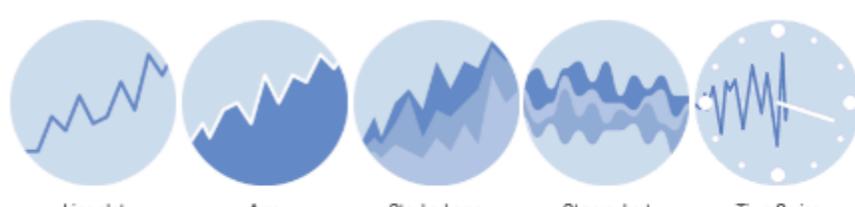
Examples of R visualizations

Excellent resource: <https://r-graph-gallery.com/>

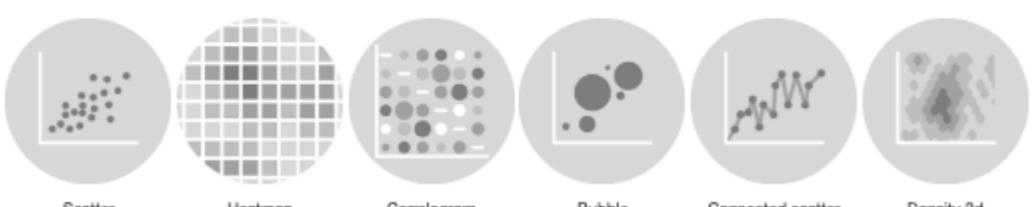
Distribution



Evolution



Correlation



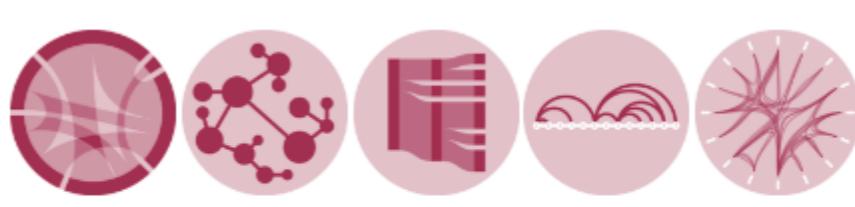
Map



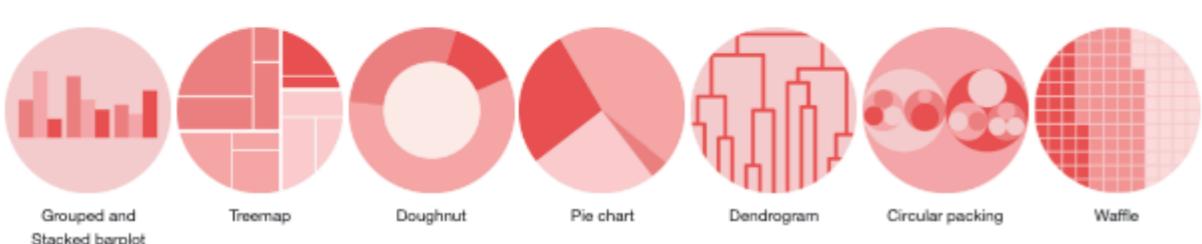
Ranking



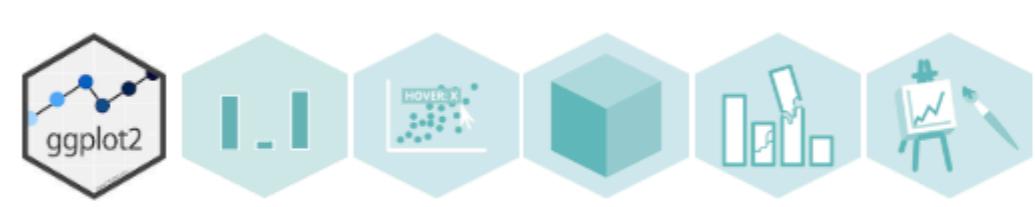
Flow



Part of a whole



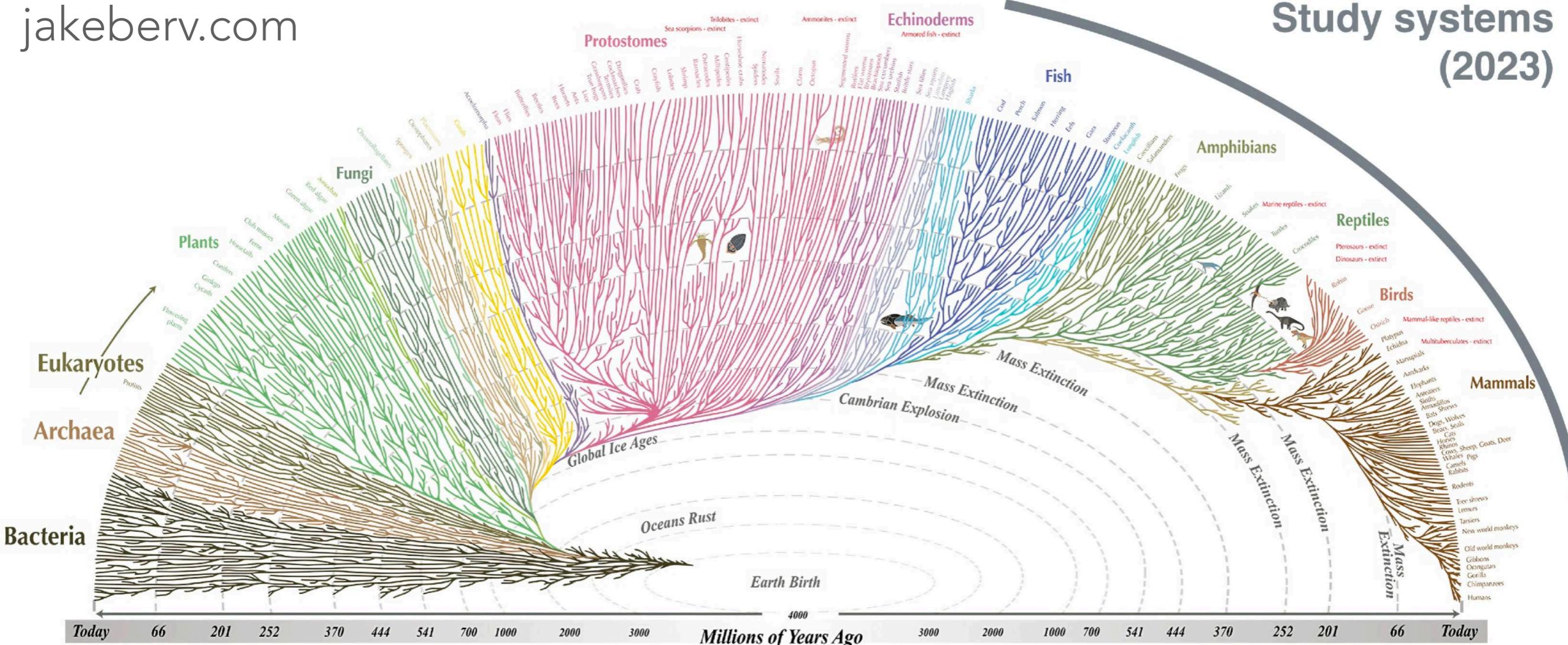
General knowledge





jakeberv.com

Study systems (2023)



"The affinities of all the beings of the same class have some times been represented by a great tree."

Charles Darwin - *On the Origin of Species*

Scan the QR code or go to this link:
<https://myumi.ch/4jG7Z>
to provide feedback on the session

Thank you!

For past sessions, see videos at:
midas.umich.edu/generative-ai-tutorial-series/

