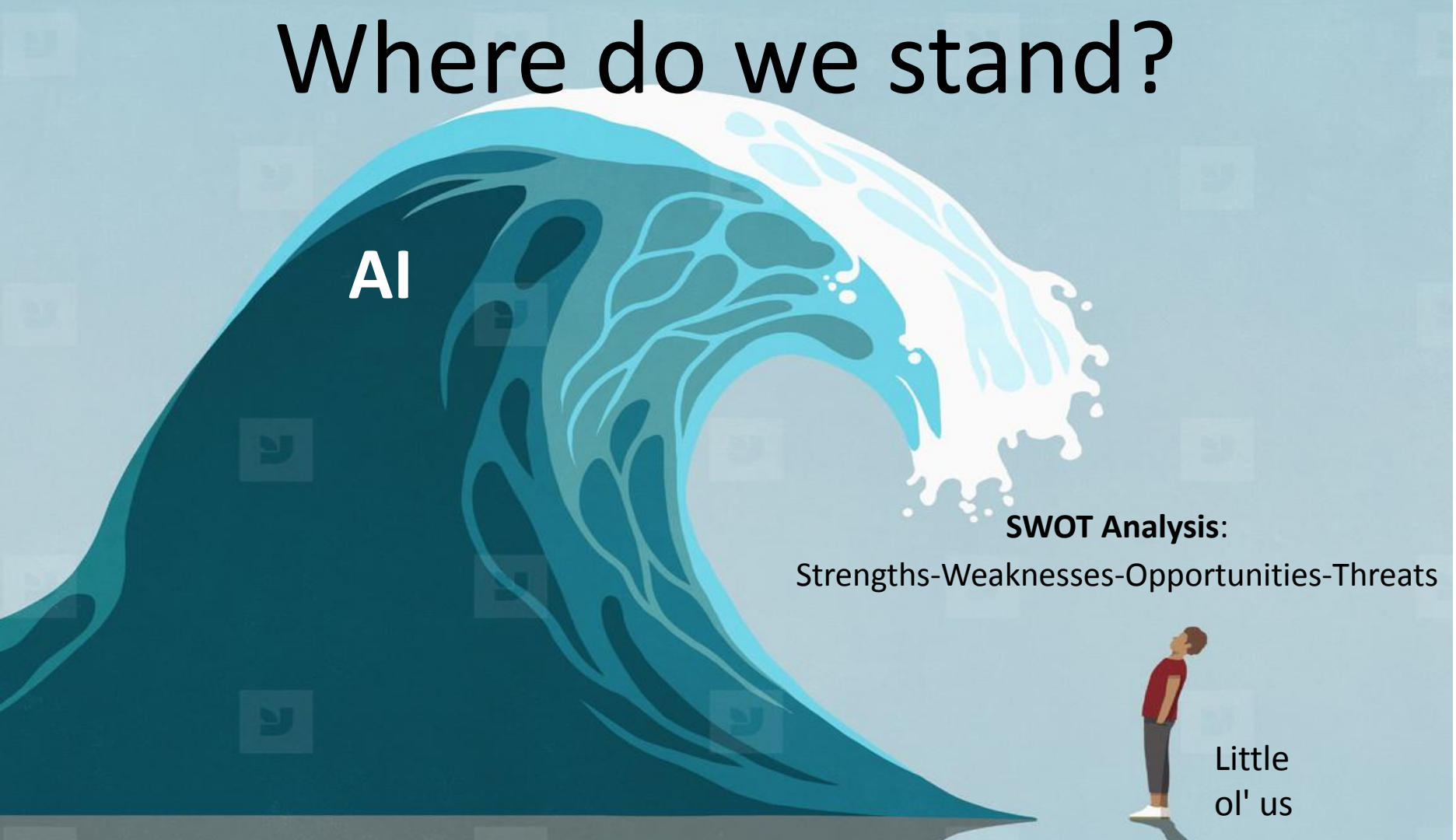
The Next "New Normal" for Language Professionals

Mike Dillinger, PhD
Free-range Knowledge Architect

MemoQfest, Budapest, 2024

Where do we stand?

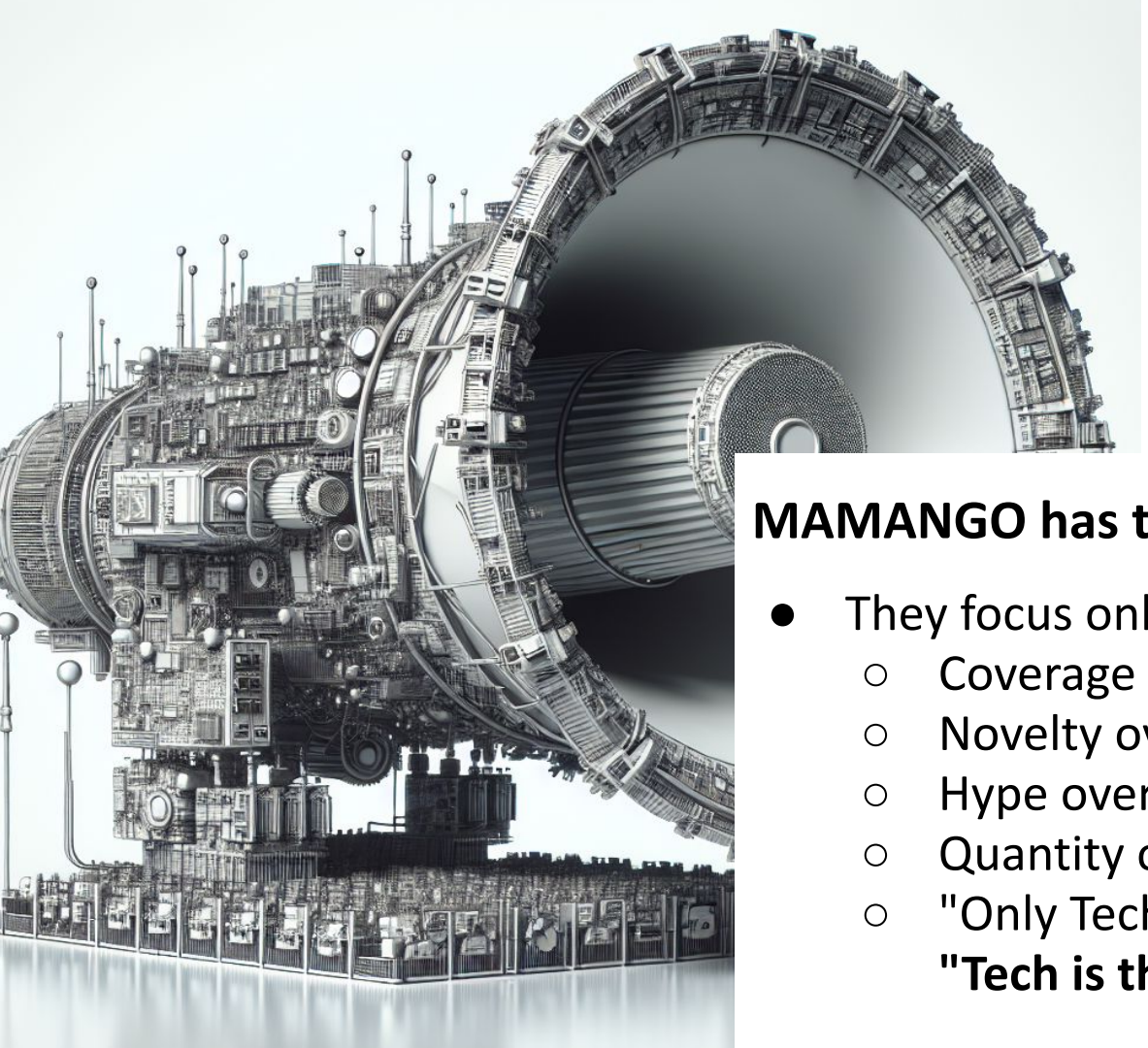


Threat #1. It's a jungle out there!

MAMANGO are 8,000 lb gorillas

Meta, Amazon, Microsoft, Apple, Nvidia, Google, OpenAI

Apex predators who want to eat our lunch!



Threat #2. VERY Big Megaphones

MAMANGO has the budget for brainwashing

- They focus only on *their use cases* and values
  - Coverage over accuracy
  - Novelty over reliability
  - Hype over truth
  - Quantity over quality
  - "Only Tech can do this";
    "Tech is the best solution"

Our Weaknesses

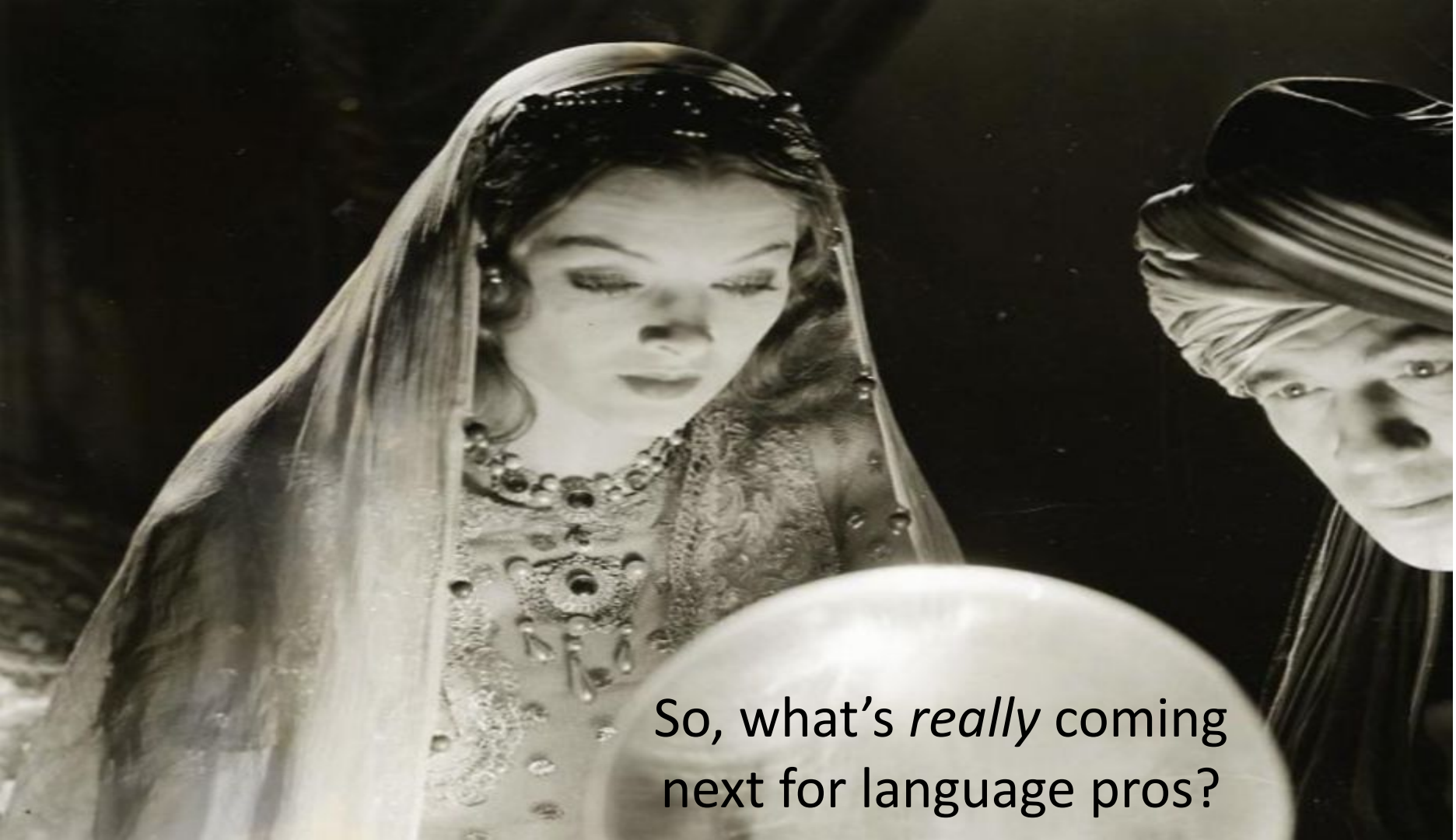
- Fragmented organization; fragmented funding; little public awareness
- Puny human-scale delivery speed
- Tiny human-scale coverage of vocabulary, domains
- Translations that are only as valuable as the source content is
- We have no control over the source
- Etc., etc., etc.

This time it's for real!!!

AI will eliminate
the need for human translators!

Promises! Predictions!

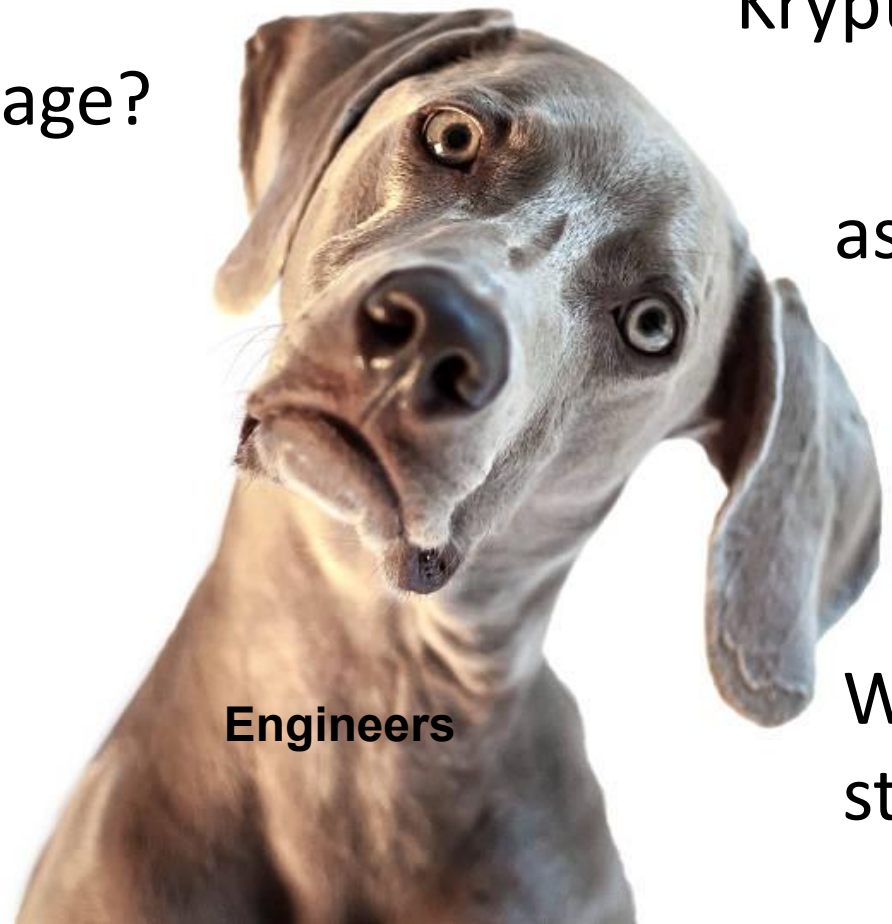
Real Solutions?



The best way to **predict** the future

is to **build** the one you want!

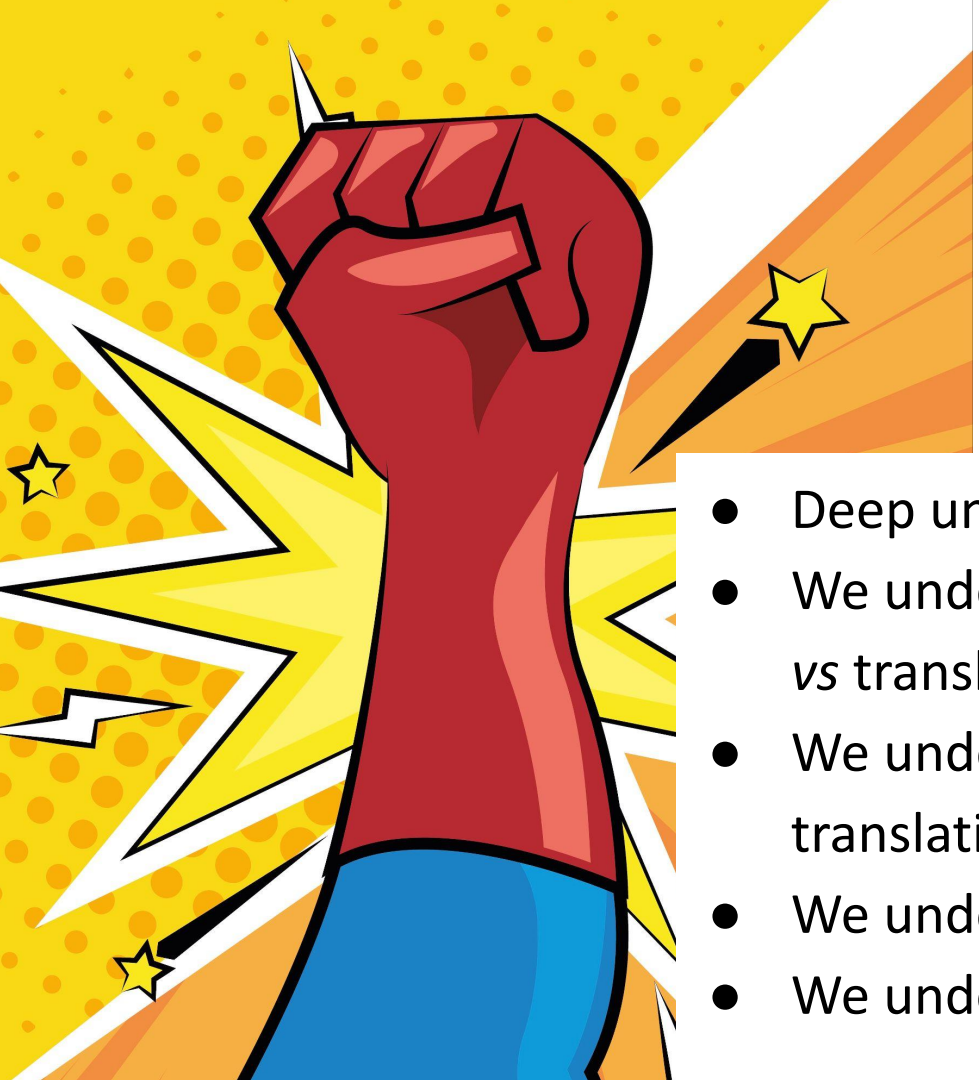
Text?
Language?



Kryptonite for Engineers!

"Unstructured" data
as in "random stuff"?!

We see structure,
structure everywhere!



Superpowers we can build on

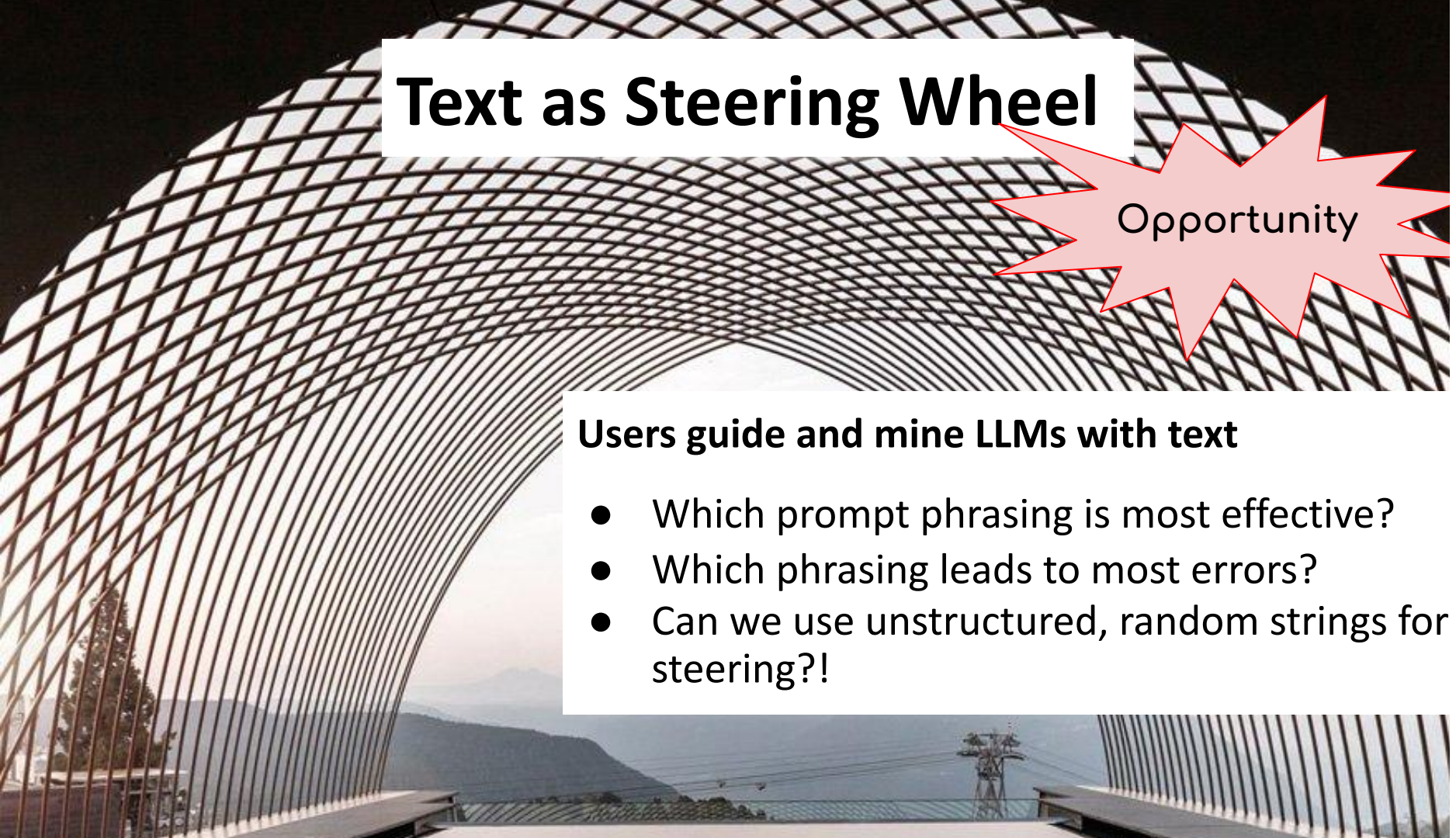
- Deep understanding of text & language
- We understand localization of content *vs* translation of strings
- We understand the value of non-routine translations
- We understand reuse vs regeneration
- We understand translation vs re-creation

Superpowers we can build on (2)

- We have domain knowledge and expertise
- We excel at audience awareness: connotations, offensiveness, appropriateness
- We understand author personas
- We show extreme adaptability in the face of chaotic processes


What’s next in AI?
More text; More need for language professionals

Text as Steering Wheel



Users guide and mine LLMs with text

- Which prompt phrasing is most effective?
- Which phrasing leads to most errors?
- Can we use unstructured, random strings for steering?!

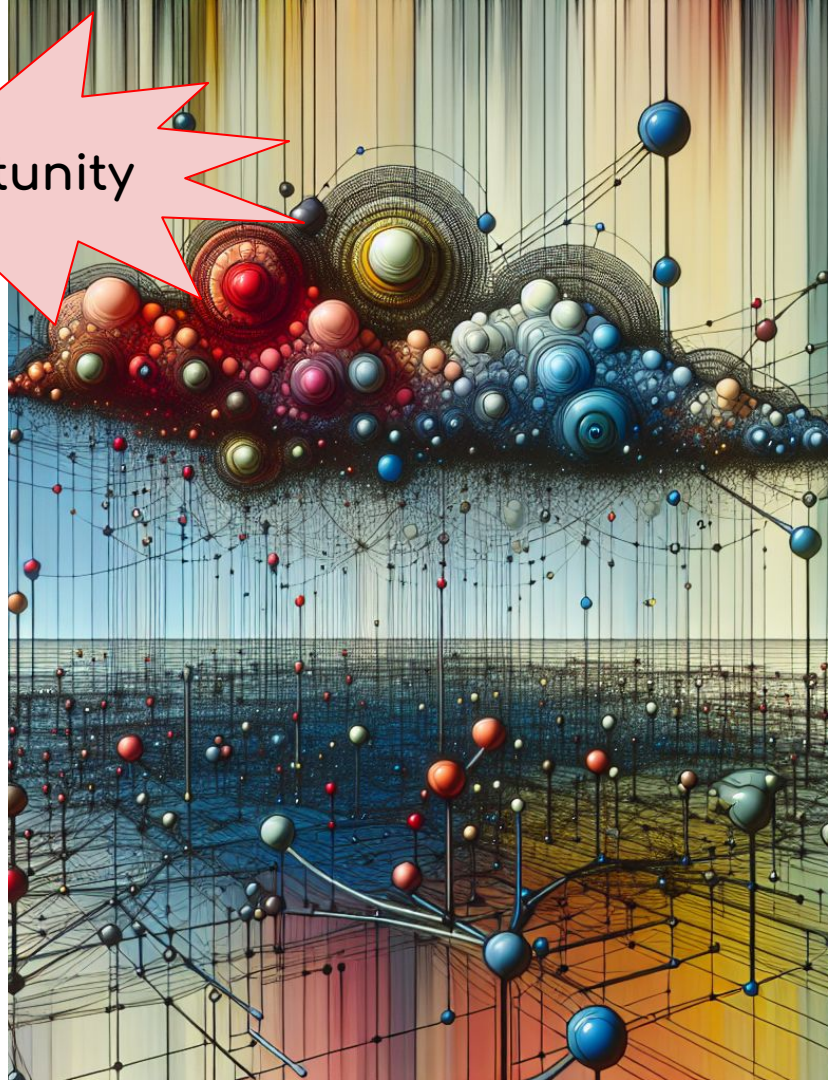
Opportunity

Text as Pivot

for multimodal models:

audio, video, images, …

- What if the text is wrong?
- What if the text is ambiguous?
- What if the text is offensive?
- What if the text has unknown terms?

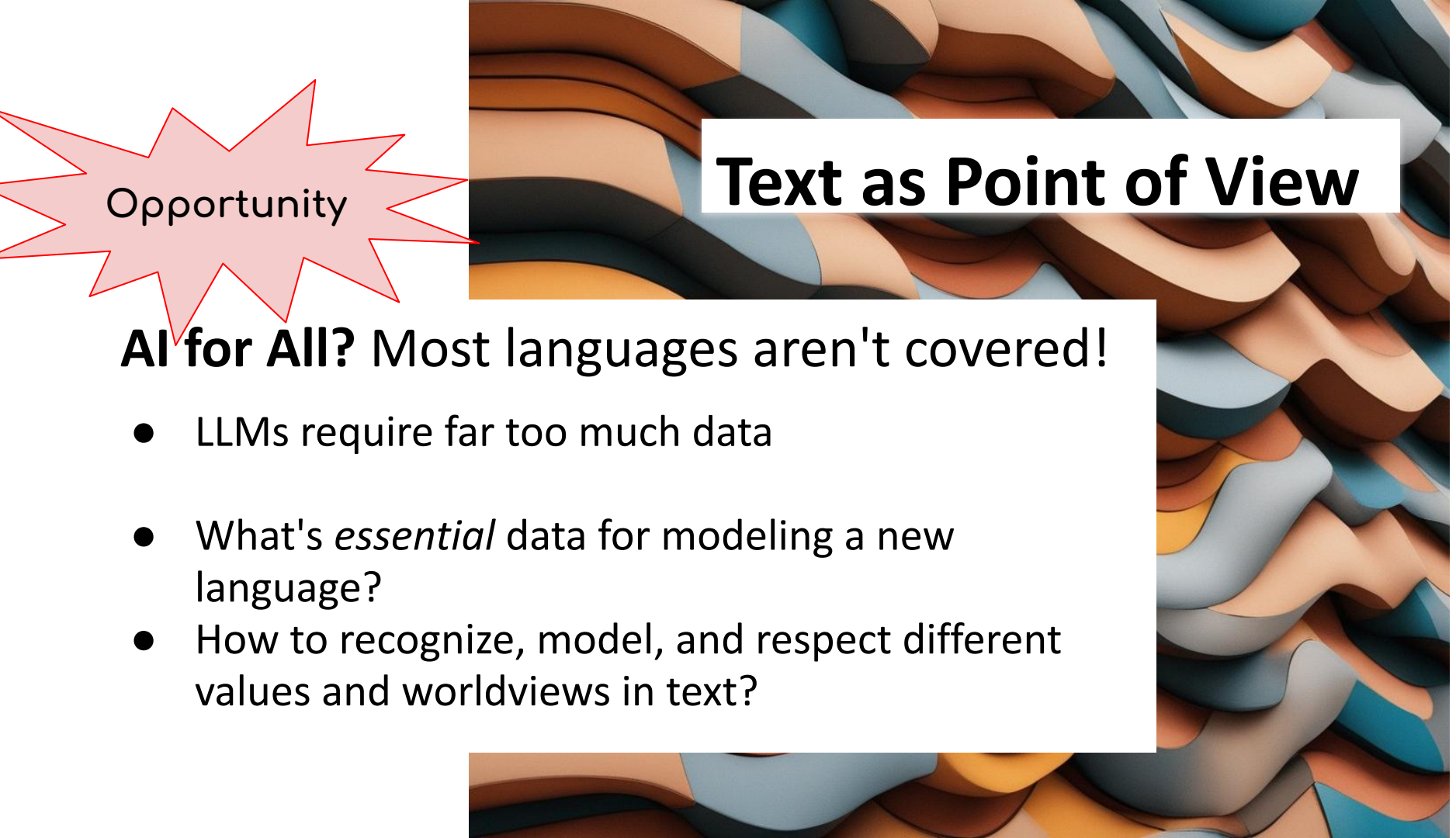


Text as Interface

LLMs are the ultimate interface with humans

- The best tool for communicating with AIs

- **How to model audience appropriateness?**
  - Make appealing, engaging text
  - Make culturally relevant, understandable text
  - Avoid offensive, off-putting text



Opportunity

Text as Point of View

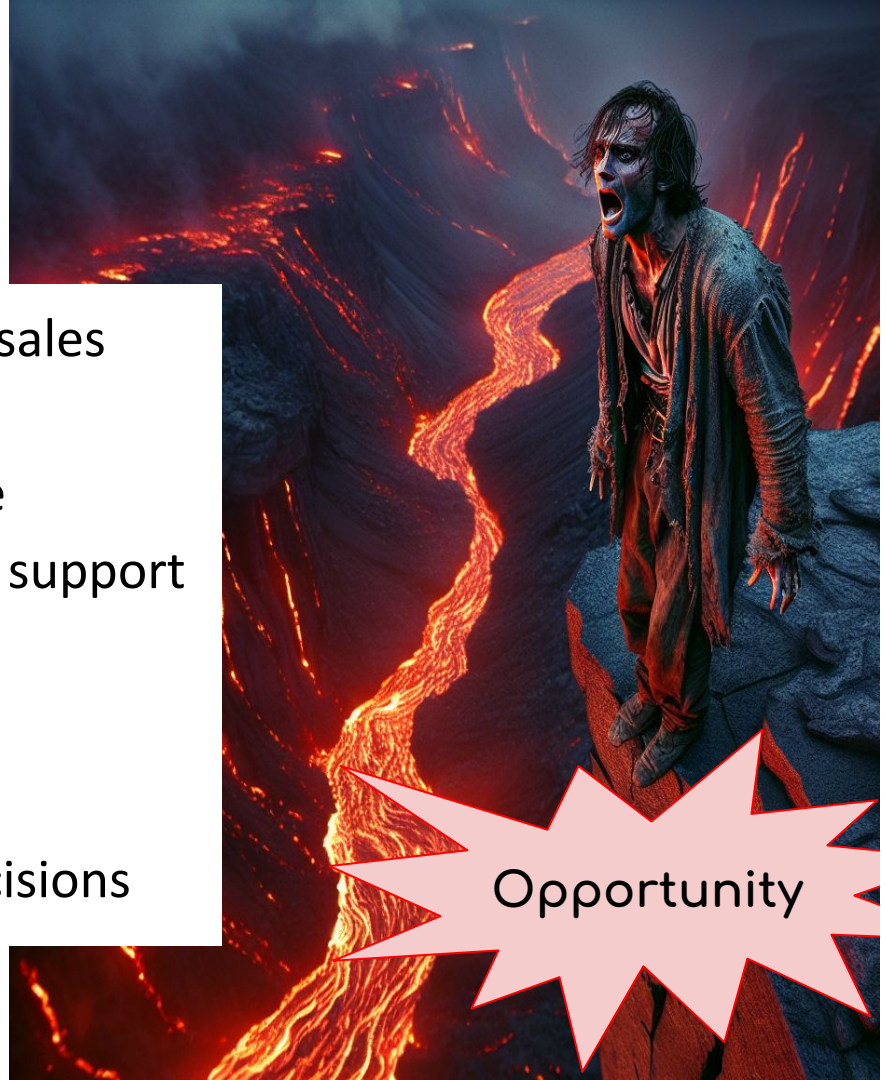
AI for All? Most languages aren't covered!

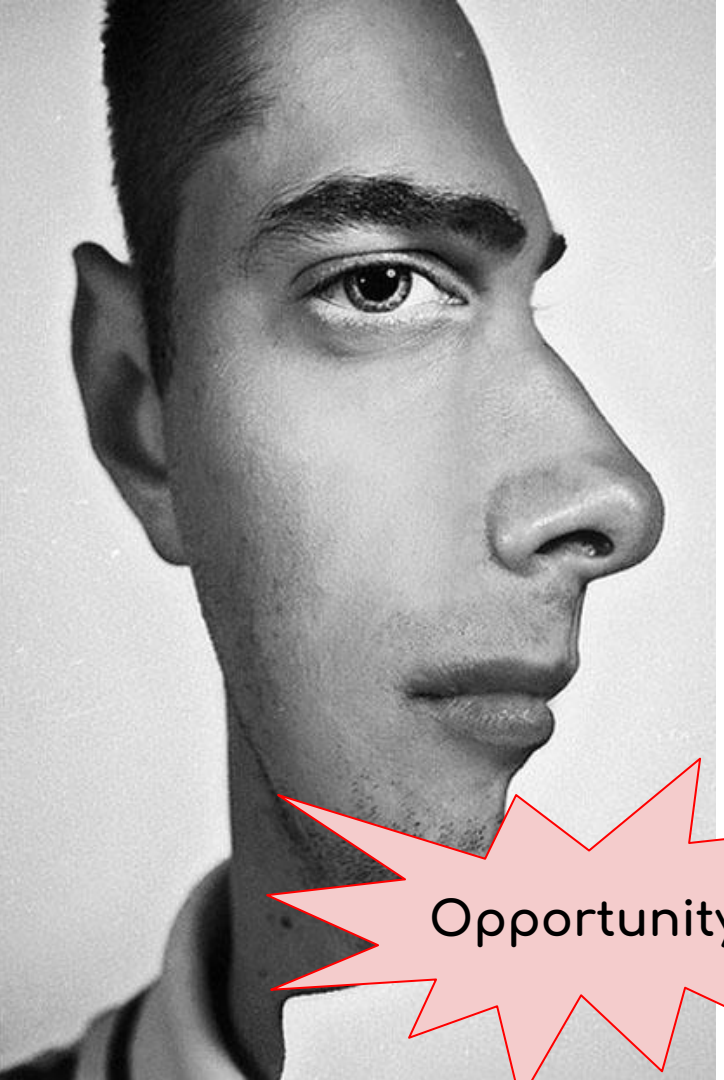
- LLMs require far too much data
- What's *essential* data for modeling a new language?
- How to recognize, model, and respect different values and worldviews in text?

Text as Risk Mitigation

Why do we use text in business?

- Avoid risk of sluggish or no growth in sales
- Avoid risks to reputation and brand
- Avoid legal risks regarding compliance
- Avoid risk of higher development and support costs
- Avoid risk of longer time to market
- Avoid risk of poor product-market fit
- Avoid risk of ill-informed business decisions





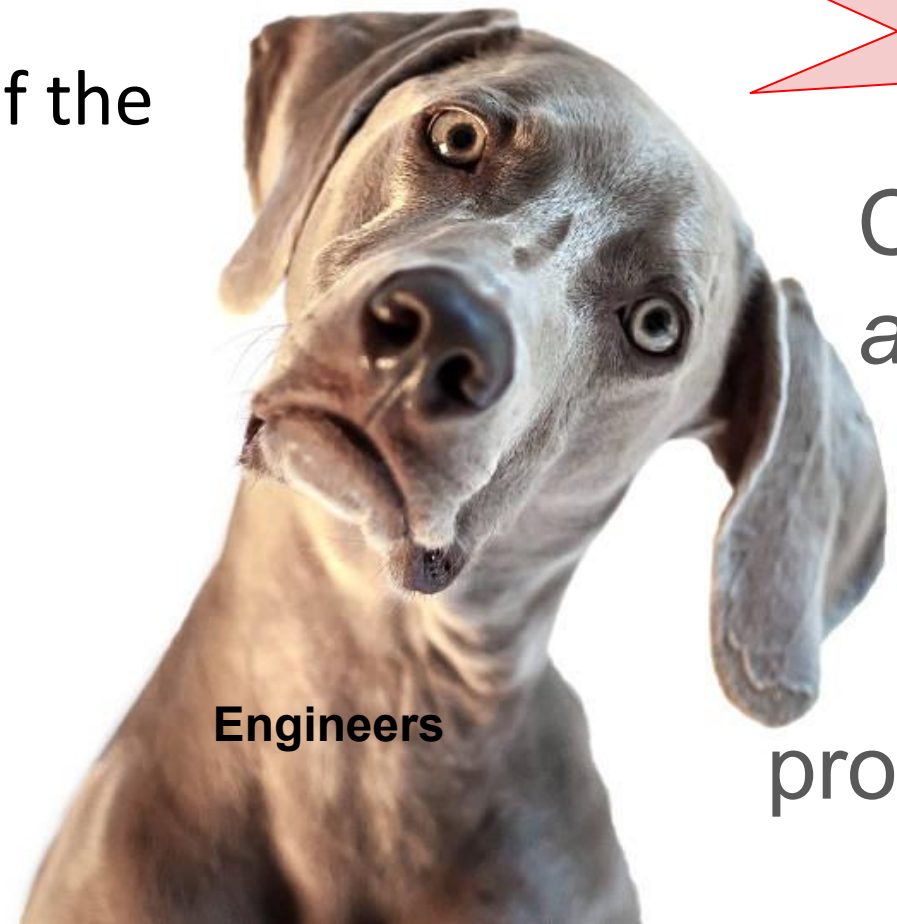
Text as Mirror of the Soul

Text expresses personality (or brand) **directly and indirectly**

- Phrasing reflects values, assumptions, and training
- Phrasing reflects emotions, goals, and intentions

- How to recognize, model, and respect different individual perspectives?
- How to identify and suppress hate speech, prejudices, anti-social points of view?

Huh?
Mirror of the what?!



Opportunities with high-impact text

Can orgs do *any* of these things well without language professionals?

But most of the remaining work has still language workers delivering

Text as Data

But Data is the New Mud

- Too much data is a problem
- Too little data is a problem
- Bad data is a problem
- The wrong data is a problem
- Versioning the data is a problem
- Extracting value from data is a problem
- Data is hard to control, transport, store, verify, correct, appraise, …

We're actually making life more *difficult* for our clients when we create more data and more content for them to manage.

How can we make things *simpler* for clients?

And provide a higher-value product?

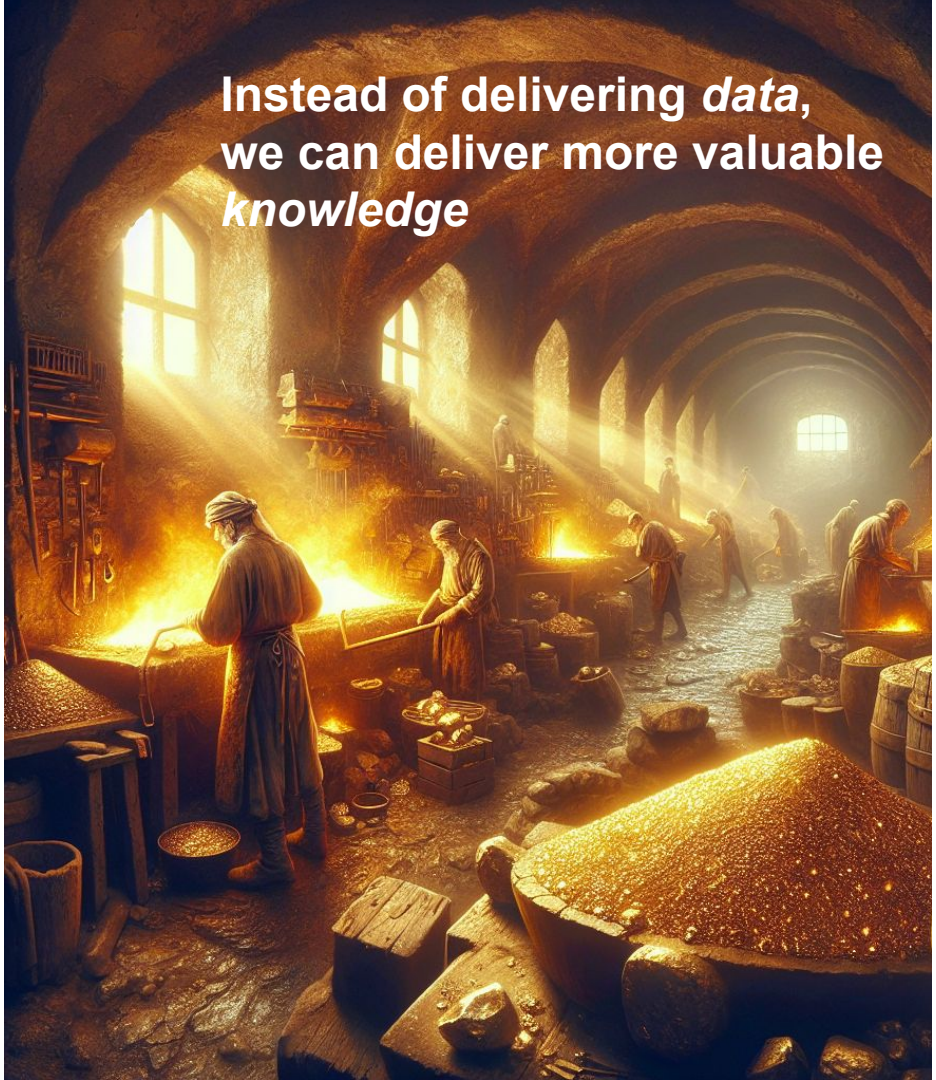
Text as Ore

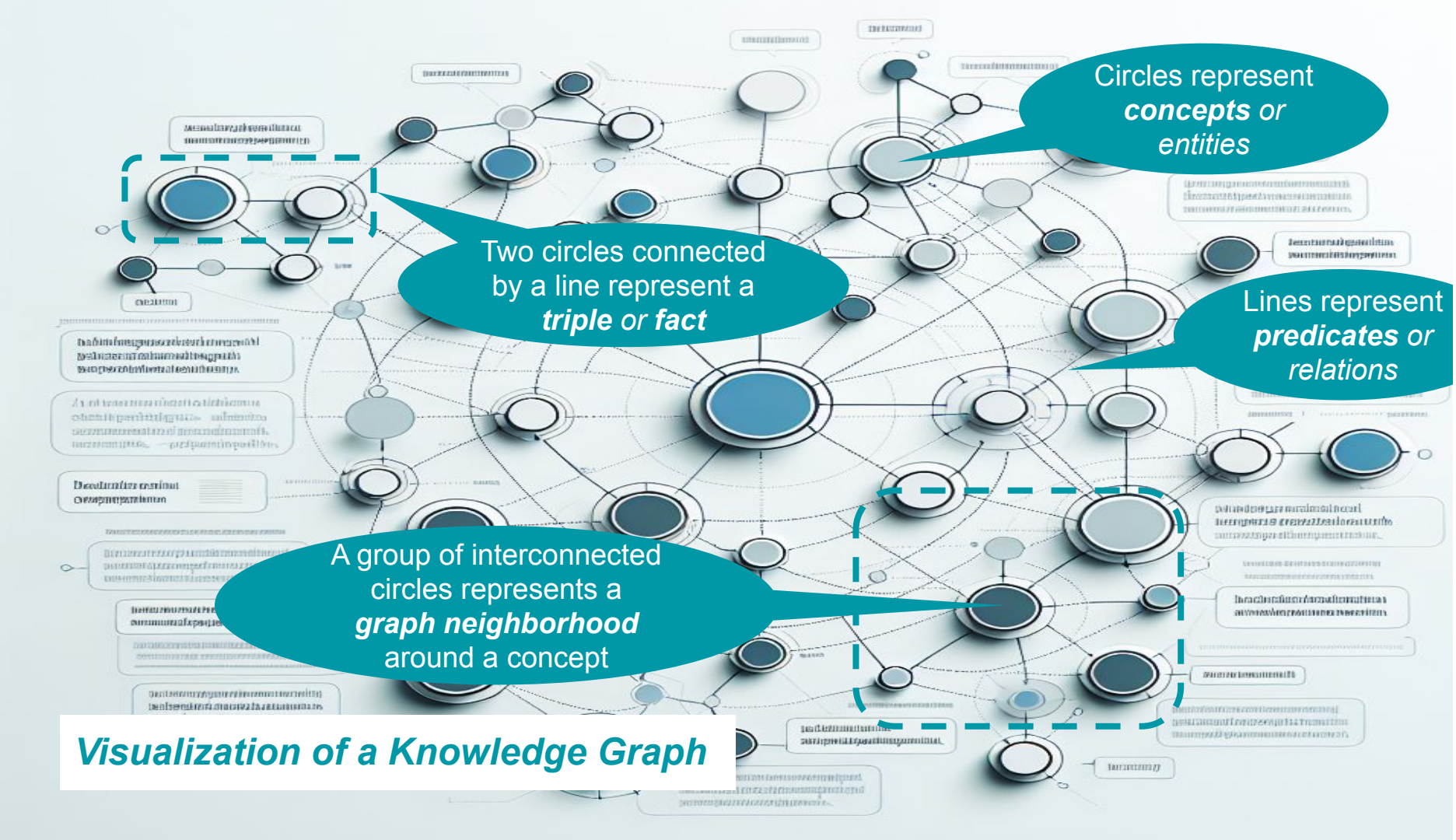
for Ingots of Knowledge

Text is the most important source of reliable knowledge, which we can store in **knowledge graphs**

We can add value to text data by "translating" it into

knowledge graphs





Visualization of a Knowledge Graph

Knowledge Graphs

Terminology work
in a new format

Knowledge graphs are collections of facts with explicit, computer-interpretable meanings

- **Facts** are formatted as concept-relation-concept triples
- **Concepts** are unique *collections* of facts (i.e., definitions) and have many labels: *dog, pooch, mutt, cur, fleabag, chien, cachorro…*
- Concepts contain facts about their components and characteristics to describe their meaning (so we can *unpack* it and manipulate it)
- Concepts are related in many different ways
  (taxonomies and ontologies are *parts* of a knowledge graph)

Knowledge Graphs, stored as a list of "triples"

Concept	*Predicate*	*Concept*
knowledge graphs	**subcategoryOf**	collections of facts
dictionaries	**subcategoryOf**	collections of facts
encyclopedias	**subcategoryOf**	collections of facts
knowledge graphs	**hasPart**	facts
facts	**hasPart**	concept nodes
facts	**hasPart**	predicates
facts	**hasPart**	features
taxonomies	**subcategoryOf**	knowledge graphs
ontologies	**subcategoryOf**	knowledge graph
AI systems	**hasPart**	knowledge graph
knowledge graphs	**usedFor**	Store conc
knowledge graphs	**usedFor**	Organ
knowledge graphs	**hasAttribute**	dom
knowledge graphs	**hasAttribute**	mach
knowledge graphs	**hasAttribute**	recursive
…	**…**	…

In the computer, these are all IDs, not strings

Each label represents *all* of the triples in this concept's graph neighborhood – not just a single string

Why bother?

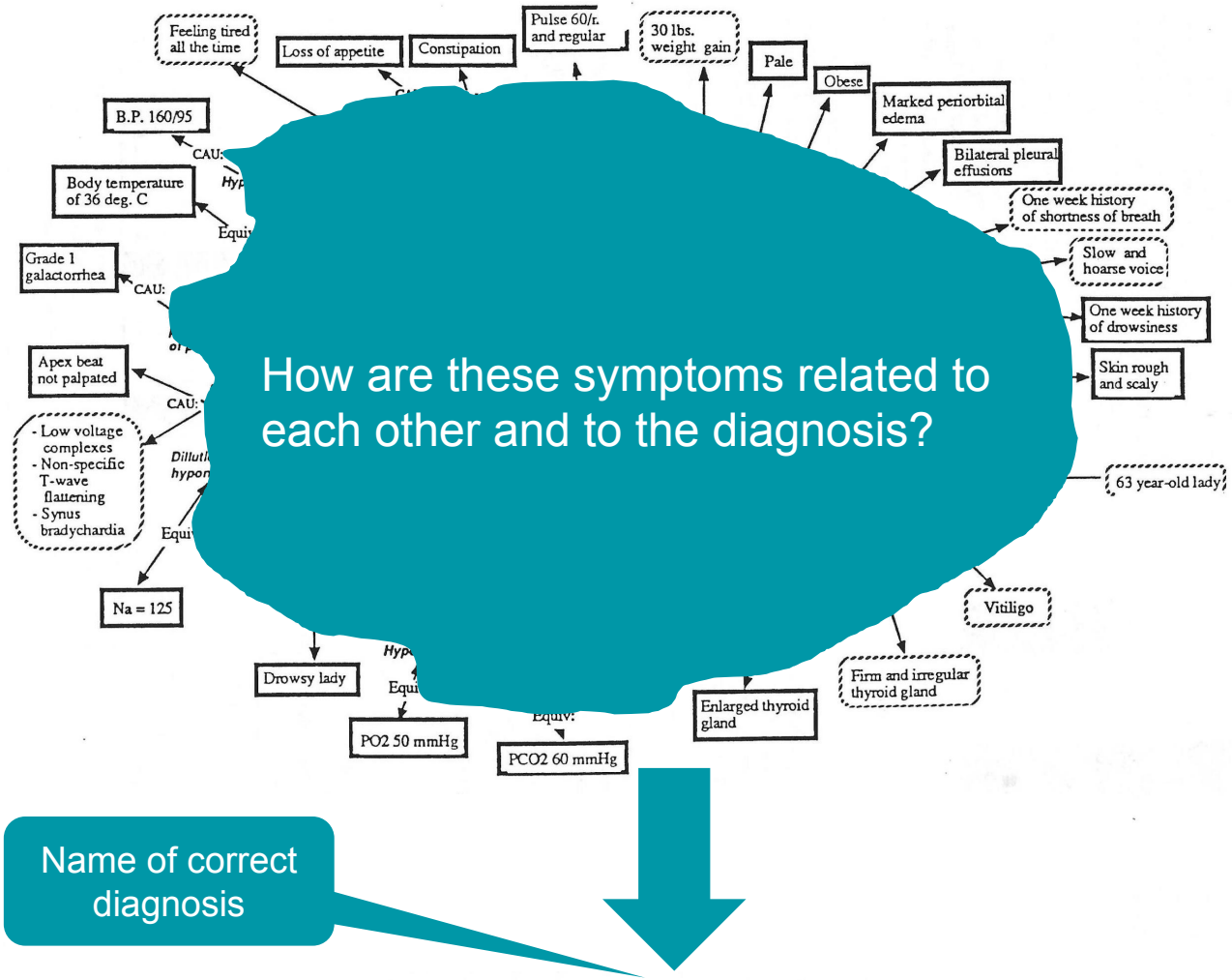


FIGURE 1. A reference model for the Hashimoto's thyroiditis with myxedema pre-coma.

Joseph & Patel, 1990

Today's Medical AI: correlations between a list of **unrelated** (!) symptoms and a diagnosis.

How can we **explain and verify** the diagnosis without knowledge of the internal mechanisms?

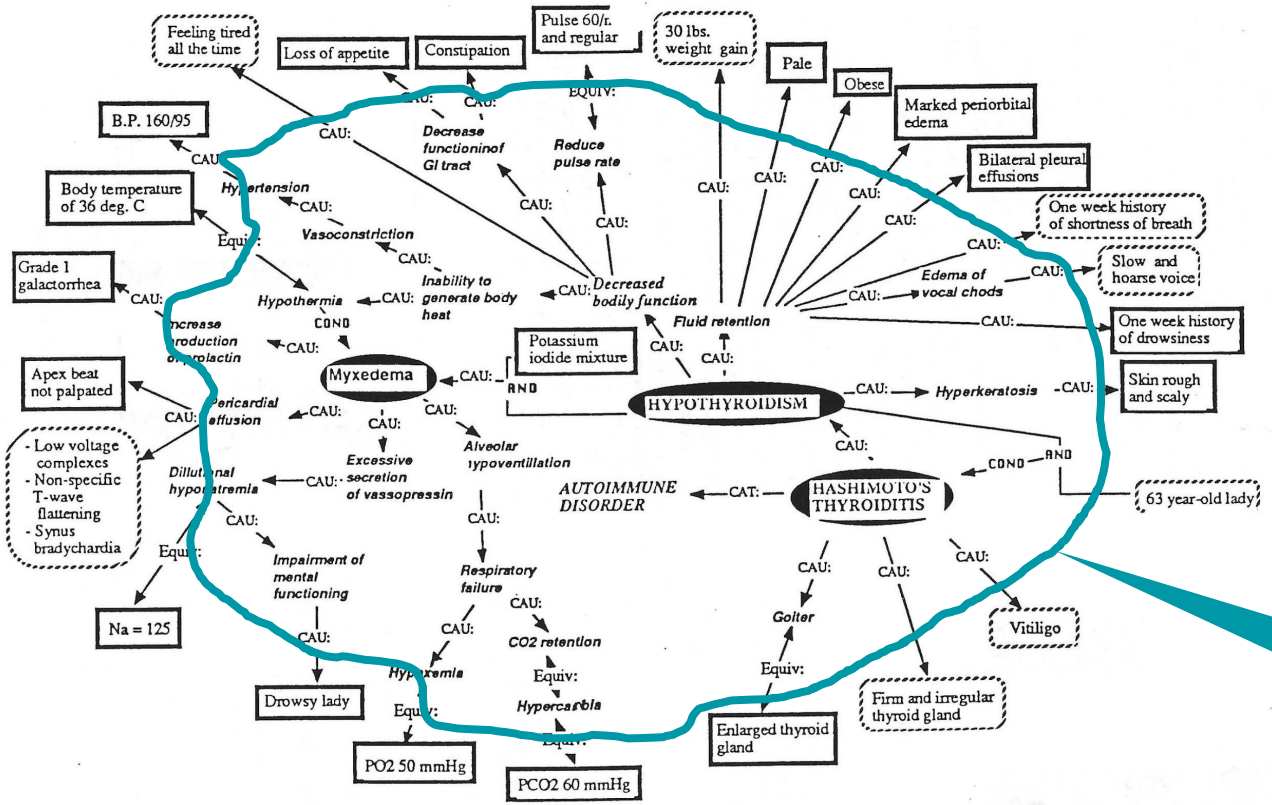


FIGURE 1. A reference model for the Hashimoto's thyroiditis with myxedema pre-coma.

Joseph & Patel, 1990

Tomorrow's AI: explicit, curated, *verifiable* relations between causes and symptoms.

A **knowledge graph** captures an explicit model of the internal mechanisms.

Who has the skills to build this from text?

Knowledge Graphs *to power next-gen tools*

Knowledge graphs for

- Indexing and matching segments by meaning, not by spelling
- Synonym spotting; meaning overlap
- Smaller, more concentrated contexts
- Meaning-based quality management
- Single-source updates and corrections
- Contradiction and consistency detection
- Automated reasoning

Knowledge Graphs *are very different from* LLMs

Knowledge Graphs		LLMs
Facts	What do they model?	Sequences of letter strings
Language-independent concepts and relations	Components	Language-specific strings
Factual accuracy, domain knowledge, coherent reasoning	Strengths	Linguistic diversity, linguistic patterns, human-machine communication
Semantics	Focus	Syntax
In computers	Where is meaning?	In human users

Knowledge Graphs *are very different from* LLMs (2)

Knowledge Graphs		LLMs
Humans, for now	**Who creates them?**	Algorithms
Written materials, human expertise	**From what?**	Written materials
Expert review, Emerging tools	**How?**	Counting frequencies of strings in different contexts
Very high	**Reliability**	Very variable
High / Slow	**Cost / Speed**	High / Fast

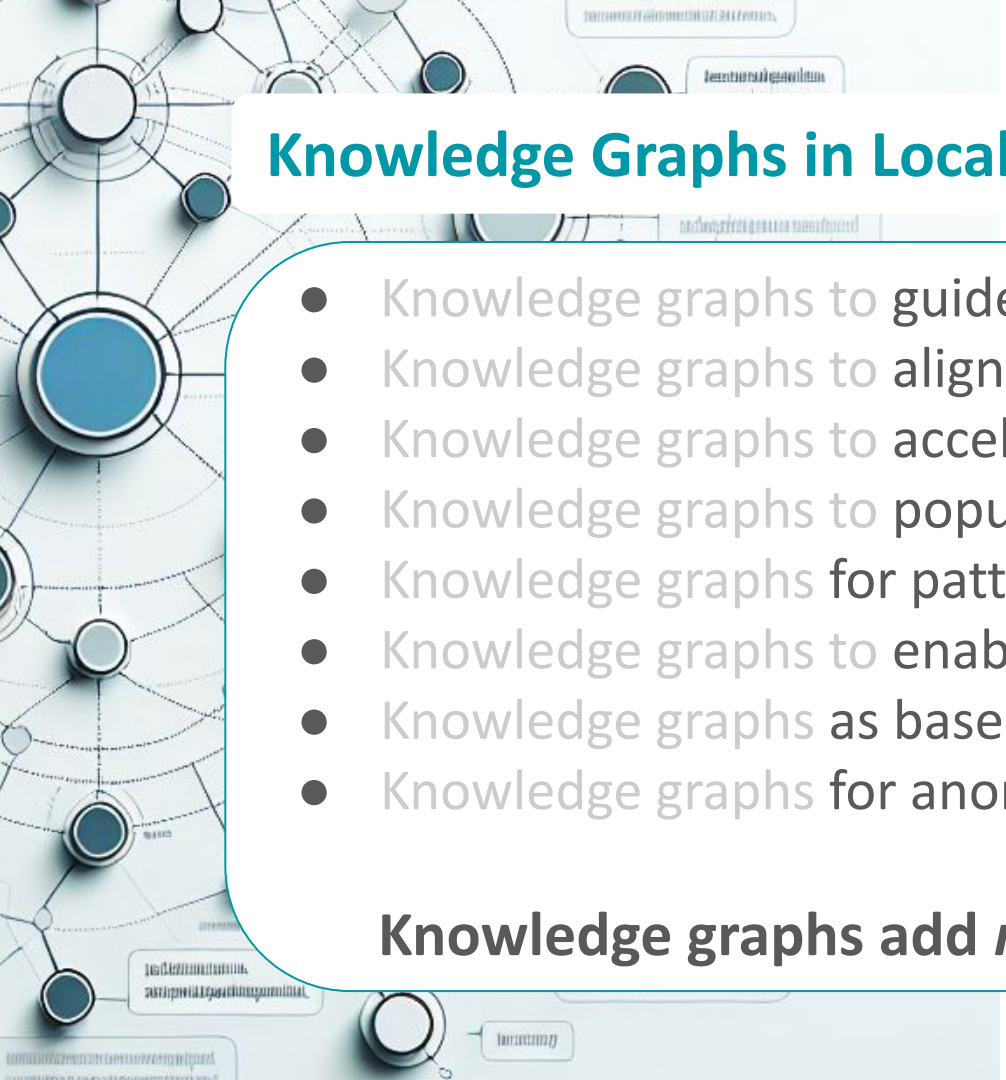
LLMs are *almost* ready…

Opportunities

Practical Issues of Localizing with LLMs

- Scarce linguistic resources and weak language coverage
- Problems of Word Alignment for training
- Speed (slow) & Cost (high)
- *Content* Localization?!
- Client- or project-specific style and guidelines
- Idiosyncratic document formats during training / [illegible] production
- Unpredictable quality issues like hallucinations
- Unpredictable results from re-use of prompts

Prompts are for **coaxing** unruly AIs, not for giving them orders.

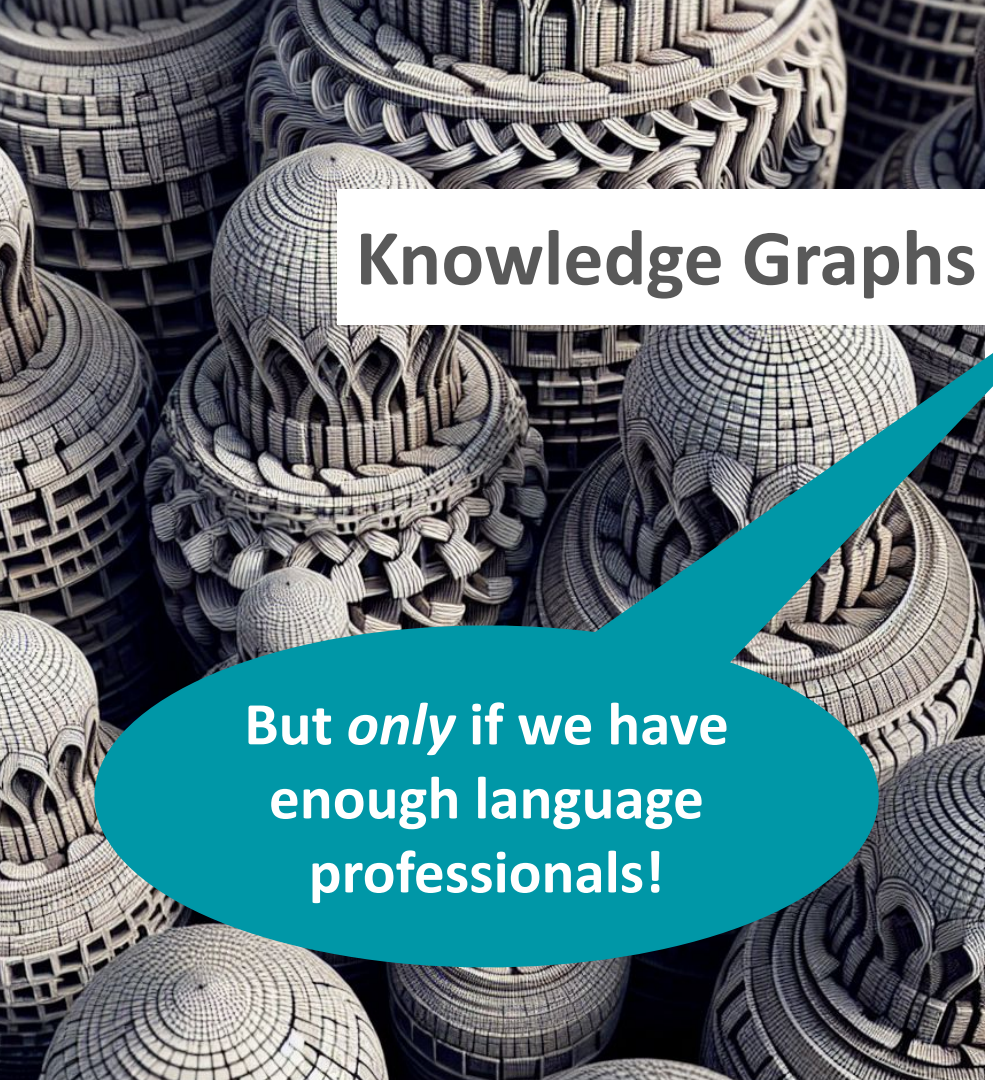


Knowledge Graphs in Localization

Opportunities

- Knowledge graphs to guide training of LLMs
- Knowledge graphs to align data sources
- Knowledge graphs to accelerate fine tuning of LLMs
- Knowledge graphs to populate prompts in RAG, like in AGT
- Knowledge graphs for patterns in guardrails to filter output
- Knowledge graphs to enable knowledge-based validation
- Knowledge graphs as baselines for evaluation
- Knowledge graphs for anomaly detection based on meaning

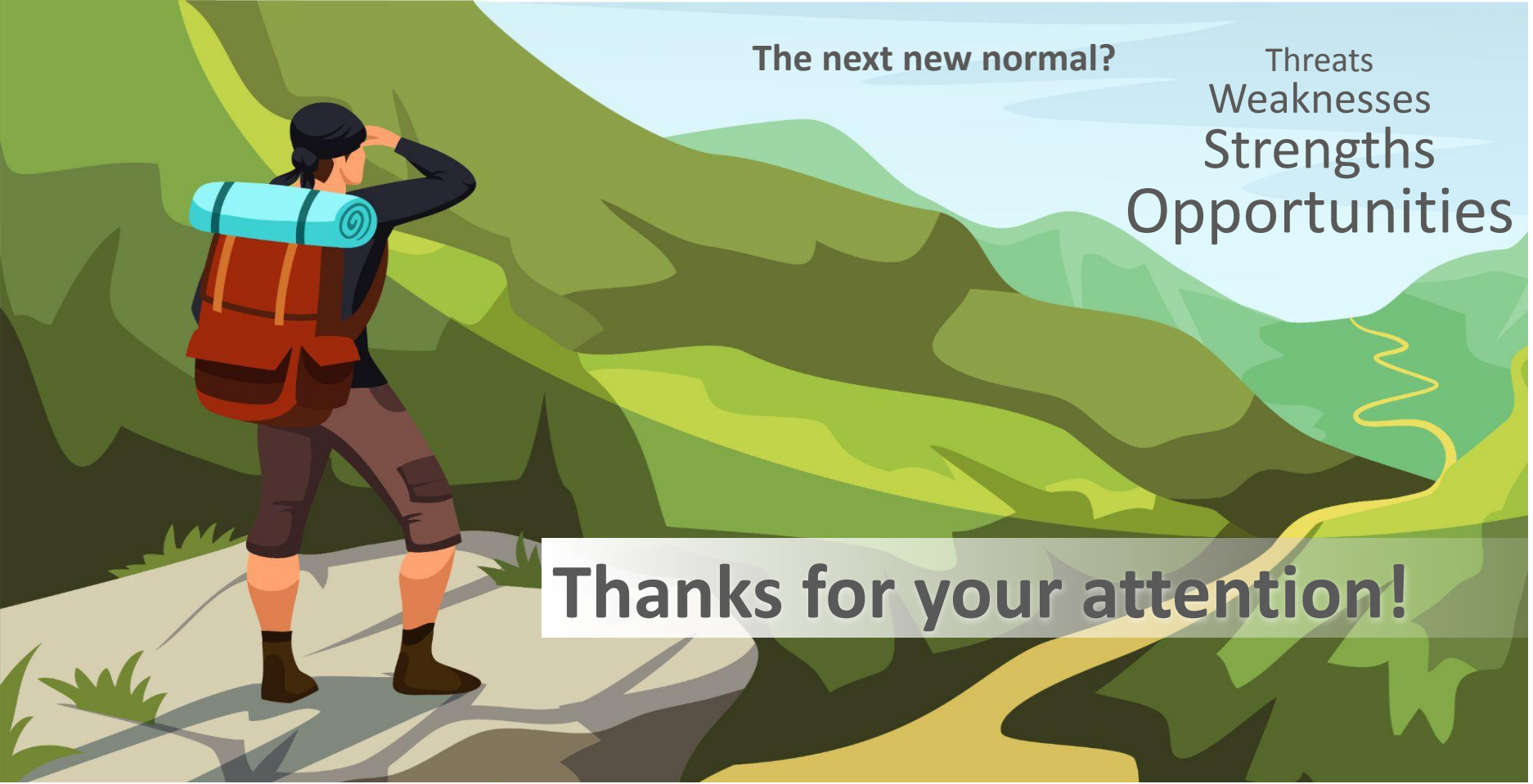
Knowledge graphs add *meaning* to AI models



Knowledge Graphs Everywhere

But *only* if we have enough language professionals!

- Built fact by fact
- Craftsmanship counts!
- To share reliable knowledge, not to generate throwaway text
- To support the creation of new knowledge
- For domains common and rare
- For access through languages large and small
- Built and used with AI tooling



Mike Dillinger, PhD

Read more at: www.linkedin.com/in/mikedillinger