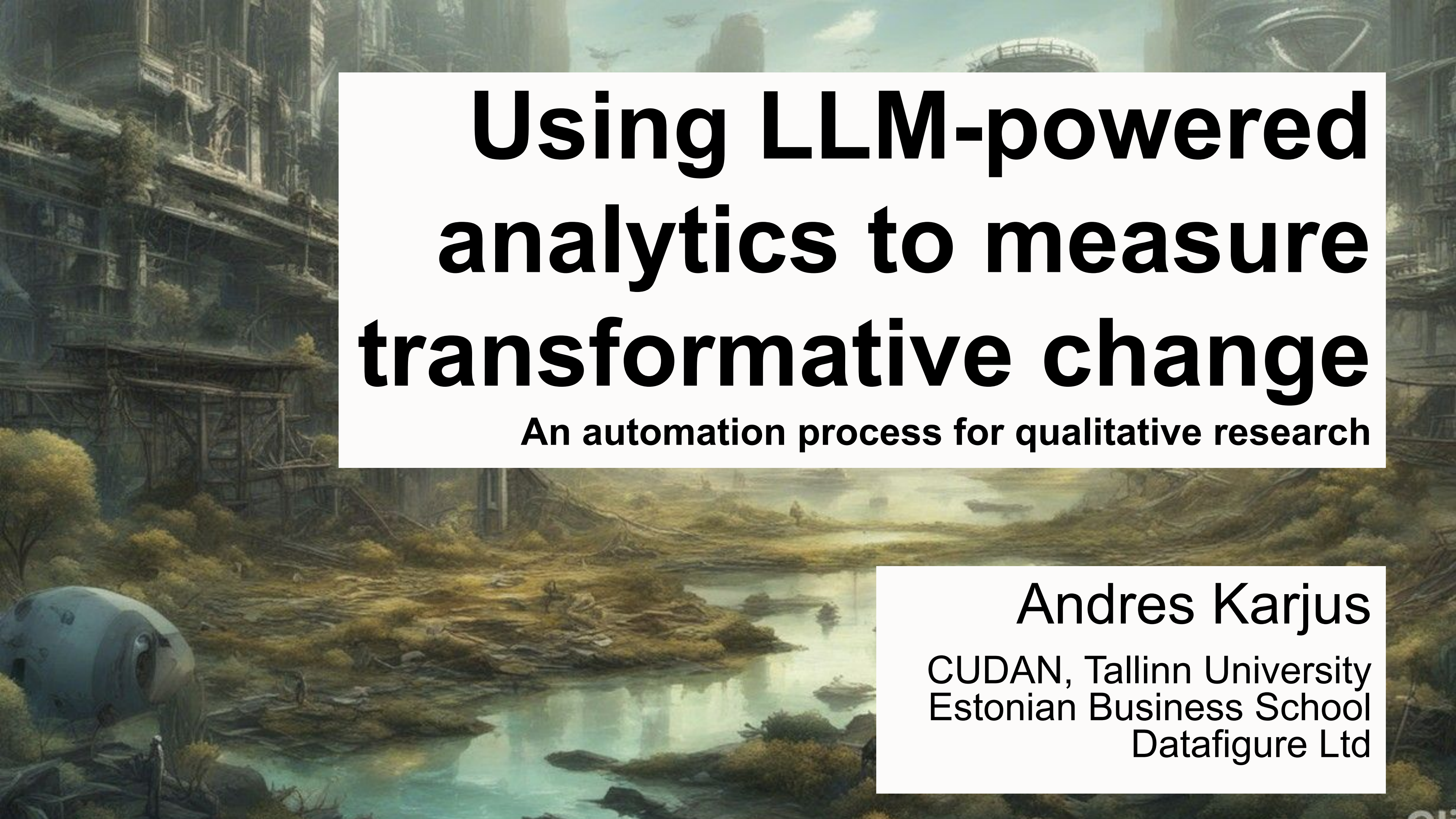


Using LLM-powered analytics to measure transformative change

An automation process for qualitative research

Andres Karjus

CUDAN, Tallinn University
Estonian Business School
Datafigure Ltd

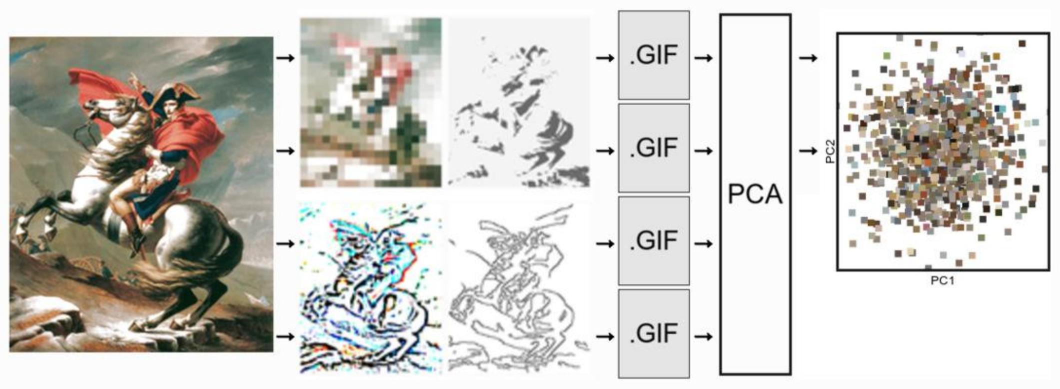
Piet Mondrian Windmill in the Gein 1906-1907 original size: 99 x 126 cm image source: wikiart.org (public domain)

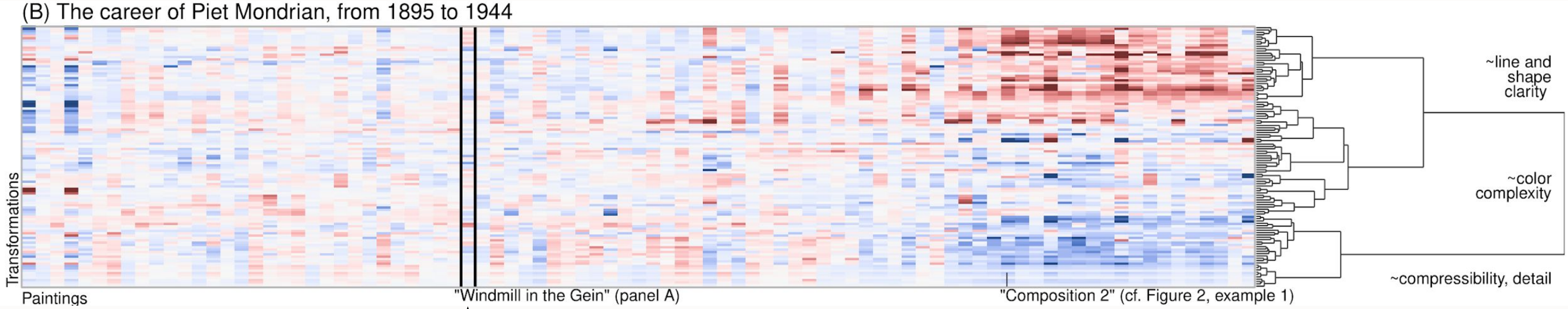
Compression ensembles quantify aesthetic complexity and the evolution of visual art

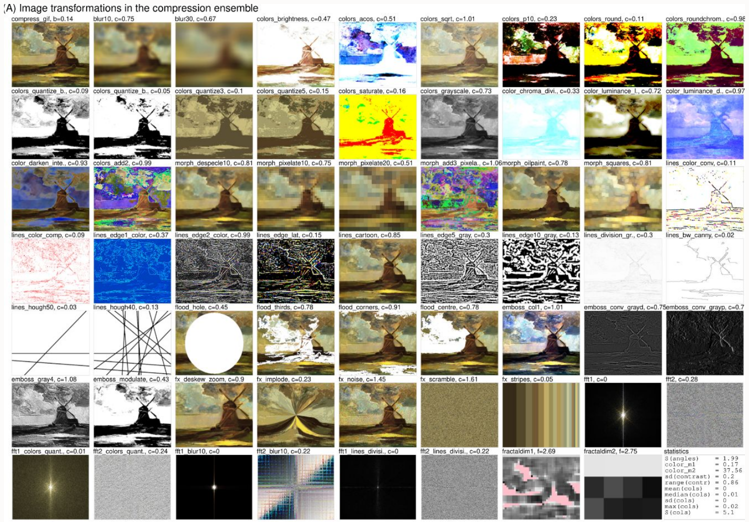
Andres Karjus, Mar Canet Solà, Tillmann Ohm, Sebastian E. Ahnert & Maximilian Schich

EPJ Data Science **12**, Article number: 21 (2023) | Cite this article

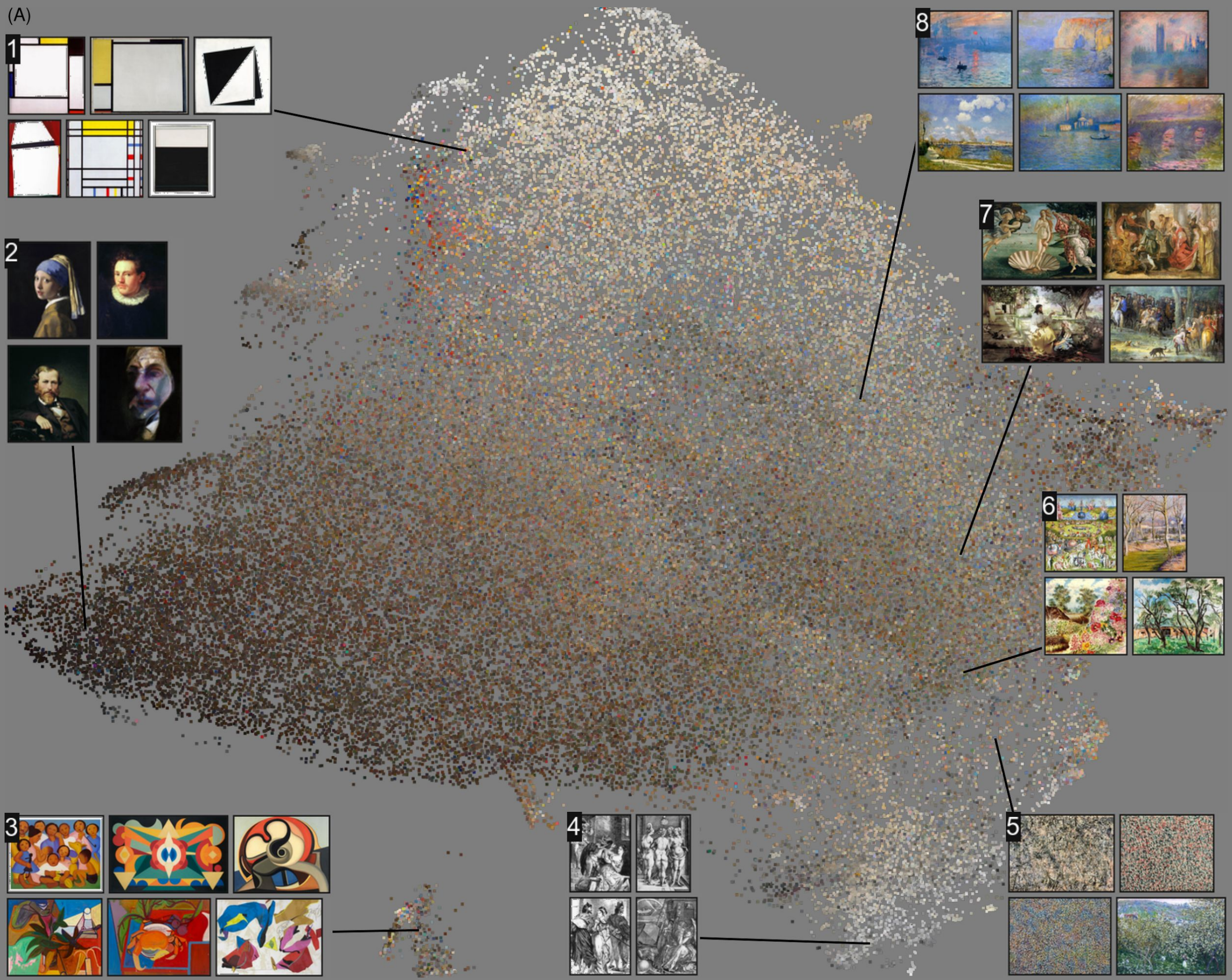
comparison framework for images. Instead of attempting to find a single algorithm or metric to best match human judgments or perform on downstream tasks, we argue for an ensemble approach of concurrently using multiple measures. This con-




(B) The career of Piet Mondrian, from 1895 to 1944
Transformations
Paintings
"Windmill in the Gein" (panel A)
"Composition 2" (cf. Figure 2, example 1)
~line and shape clarity
~color complexity
~compressibility, detail


(A) Image transformations in the compression ensemble

Dataset: 74k 2D artworks from Wikiart/art500k (Mao et al 2017)



Important: object similarity is incidental. It's a measure of aesthetic complexity, but similar subjects are often depicted using a similar complexity profile.

Society and societal change is also complex

- Instead of trying to find one estimator/index, use them all. No aggregation, no arbitrary weights.
  An ensemble. S-curves --> S-surfaces
- Instead of 1 measure, a vector (space) of numeric measures
  (scale if necessary; decorrelate via PCA if needed)
- Quantitative measures (GDP, number of vehicles)
- But qualitative measures…? Ideas are not numeric :(
- Qualitative data (text, images) is… hard. So far it has necessitated either human annotation (expensive!) or task-specific, often complex ML solutions.

Enter large language models

- Recent developments in multilingual generative LLM tech open up unprecedented avenues of analytics
(yes ChatGPT is a nice toy but here we're interested in LLMs as generative classifiers)
- Already currently available LLMs perform at near-human level in many textual tasks (*above* crowdsourcing quality in many tasks)
- Zero-shot learning via instruction. No need for complex computational pipelines, training or tuning for every single task.

- *Tag this text as supportive or against the idea of expanding the nuclear energy sector: "While the need for green energy is apparent, more nuclear plants is the last thing this country needs."*
- *Against.*

Machine-assisted mixed methods: augmenting humanities and social sciences with artificial intelligence

- MAMM: a framework that casts the application of instructable machines (e.g. zero-shot LLMs) as annotators/classifiers in a mixed methods framework
- Quantit*izing* design
  (aka quantitative-qualitative, integration through data transformation)
- Unitization > qual analysis/coding > quant/stats
- Change, transitions, niches, regimes: can then be measured in such *quantitized* units.

Machine-assisted mixed methods: augmenting humanities and social sciences with artificial intelligence

16 tasks
9 languages
10+ disciplines

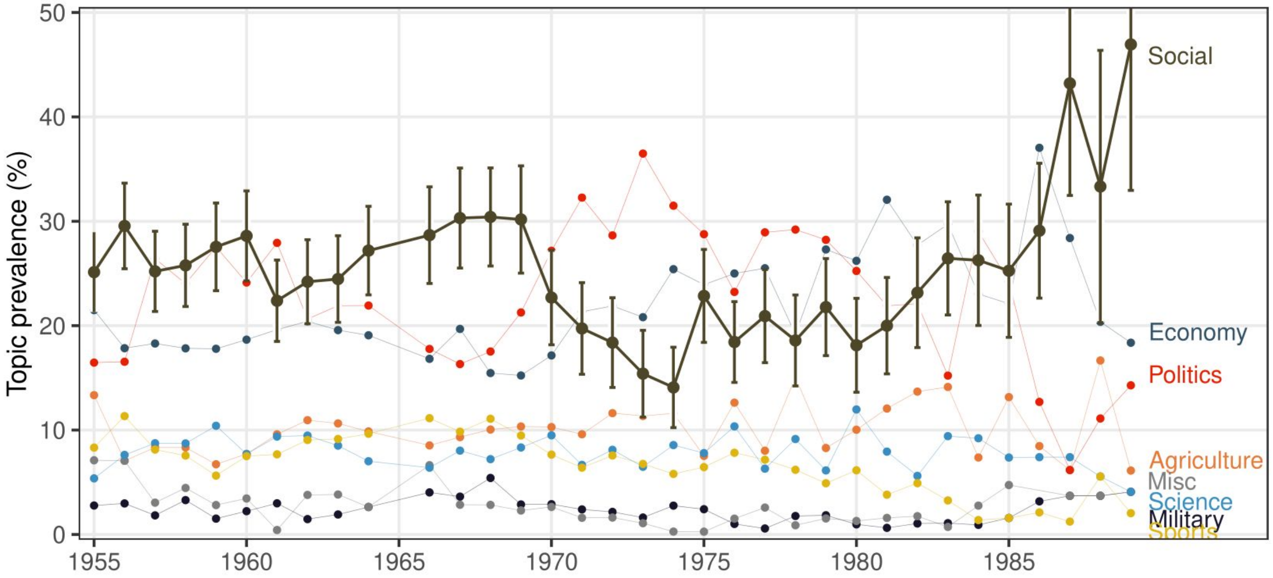
Table 1: Summary of case studies in this contribution, replicating and emulating various humanities and social sciences research tasks. The Acc column displays raw accuracy of the best-performing LMM at the task (compared to human-annotated ground truth; results marked with ρ are Spearman's rho values instead of accuracy). The Adj column shows the kappa or baseline-chance adjusted agreement where this is applicable. Open-ended results are marked with an asterisk*.

Task	Language	Acc	Adj	Data domain	Complexities
Topic prediction	Russian	0.88	0.85	Cultural history, media	Historical, abbreviations
Event cause detection	Estonian	0.88	0.83	Maritime history	Historical, abbreviations
Interview analytics	English	1	1	Discourse/content analysis	
Relevance filtering	English	0.92	0.82	Text mining, history, media	Low quality OCR
Text&idea reuse	Eng, Rus	1	1	History of ideas	Multilingual
Usage feature analysis	Eng (18th c)	0.94	0.89	Linguistics, culture	Historical
Semantic change	English	$^{\rho}0.81$		Linguistics, NLP	Historical
Semantic change	German	$^{\rho}0.75$		Linguistics, NLP	Historical
Semantic change	Latin	$^{\rho}0.1$		Linguistics, NLP	Historical
Semantic variation	English	$^{\rho}0.6$		Sociolinguistics	Social media text, emoji
Stance: relevance	Estonian	0.95	0.91	Media analytics	
Stance: polarity	Estonian	0.95	0.92	Media analytics	
Lit. genre detection	English	0.8	0.73	Literature	Books mix genres
Translation analytics, censorship detection	Eng, Italian, Japanese	0.96	0.95	Translation studies, culture	Multilingual
Novel sense inference	Eng, Est, Turkish	~1		Lexicography, linguistics	Minimal context
Data augmentation	Finnish	0.72		Media studies	Minimal context
Visual analytics	-	*	*	Film & art, cultural analytics	Multi-modal
Social network inference	English	*	*	Network science, literature	Many characters, ambig. references

Task	Language	Acc
Topic prediction	Russian	0.88
Event cause detection	Estonian	0.88
Interview analytics	English	1
Relevance filtering	English	0.92
Text&idea reuse	Eng, Rus	1
Usage feature analysis	Eng (18th c)	0.94

gpt-3.5 accuracy:
0.88 (kappa 0.85)

(A) Predicted topics in Soviet newsreels 1955–1989



A little replication

ELSEVIER

Global Environmental Change
Volume 72, January 2022, 102447

Deep Transitions: Towards a comprehensive framework for mapping major continuities and ruptures in industrial modernity

Laur Kanger [a b], Peeter Tinits [a], Anna-Kati Pahker [a], Kati Orru [a], Amaresh Kumar Tiwari [c], Silver Sillak [a d], Artjoms Šeļa [e f], Kristiina Vaik [g]

Relevance filtering with OCR correction

LLM. The authors (Kanger et al. 2022) kindly provided a human-annotated test set of 99 excerpts for this exercise. The experiments here include both GPT-3.5 and GPT-4, and also evaluate the effect of adding an OCR-correction step before the classification step. While most many of the corpus texts are fairly readable, they also contain examples such as this:

principally to easing in » u ¿ allan consolidated bonds nine Issues Siorln « fallí and on'y two lssues ßalnl » 8 The lïttei Included the 3 . per cent 1942 in which laigf pa'cek were bou.ht The Syd Ii , banks lollnqulshed a small pait of recent rlim Arünstnaturalleacilon in t . limited S , r of issues the main body of Indu- irai continued to find keen support.

Does it work?

model	accuracy	adjusted (kappa)
gpt-3.5-turbo (raw text)	0.79	0.49
gpt-4 (raw text)	0.9	0.77
gpt-3.5-turbo (cleaned)	0.82	0.56
gpt-4 (cleaned)	**0.92**	0.82

Just text?

Figure 6: Examples of visual analytics using a multimodal LLM (Microsoft Bing AI, August 2023 version). See the Appendix for the full prompts and outputs which have been truncated here.
(A) Prompt: Explain this joke. — Output: The joke is based on a wordplay that involves removing the letter B from the word "bananas" and getting the word "ananas", which is the French word for "pineapple".
(B) This is a still from a movie. Analyze it like a film scholar. — The image depicts a scene of violence and conflict, as indicated by the presence of guns, masks, and broken glass. The scene is likely part of an action or thriller genre, which typically features such elements to create tension and excitement for the audience. The scene also suggests a power imbalance between the four men, as three of them are shooting at one, who is falling backwards. The falling man is in a vulnerable and defeated position, while the other three are in a dominant and aggressive stance. This could imply that the falling man is either the protagonist or an ally of the protagonist, who is facing a threat or a challenge from the antagonists. The scene is shot from a low angle, which enhances the sense of drama and danger. /.../ The image is slightly blurry, which suggests fast movement and adds to the feeling of chaos and urgency. The setting of the scene is a gray stone building with columns, which could indicate a public or official place, such as a government office, a bank, or a museum. /.../ The debris and broken glass scattered around the scene add to the sense of destruction and disorder caused by the shootout.
(C) Make a nested list of all prominent objects and characters in this painting. — [painting [sky [clouds]] [buildings [roofs] [walls] [windows] [doors]] [street [cobblestones]] [people [man in black suit and top hat] [woman in blue dress and white hat] [boy in brown shirt and pants] [girl in pink dress and bonnet]] [animals [horse] [cart] [chickens] [rabbits]]].

Conclusions

- Instead of aggregated indices, consider multidimensional ensembles
- In research design and analytics, there is no substitute for expert human judgment, reasoning and calibration.
- However, human time is a bottleneck. With suitable instructable machines, expert capacities can be augmented and scaled up to enable research on much larger volumes of data.
Humans don't scale, machines do.
- The quantit*izing* mixed methods approach provides a flexible and replicable framework for working with qualitative data types.
- The MAMM is an augmentation of the QMM with machine learning.

Thanks!

follow:
@andreskarjus (twitter/x, linkedin, mastodon)
andreskarjus.github.io

3.4 Social network inference from literary texts

This short section showcases the potential of using LLMs as information retrieval engines. Figure 2.A depicts a character network manually constructed from "Les Misérables" by Victor Hugo, often used as a textbook example in network science and related fields. Figure 2.B is a network of interacting characters inferred automatically from the full text of the same book, by feeding each chapter into GPT-3.5 with the prompt to list pairs of characters who directly converse in the chapter. The result may well have some errors — some anomalous pairs like street names and unspecific characters ("people" etc.) were filtered out post-hoc. Better results may well be achieved by better prompts and using more capable models like GPT-4. Still, the result is also much richer than the smaller manual version, including non-plot characters discussed by Hugo in tangential sections of the book. This limited exercise shows that LLMs can be used for information retrieval tasks like this in H&SS contexts, while preprocessing with specialized models (named entity recognition, syntactic parsing, etc.) is no longer strictly required (cf. Elson et al. 2010).