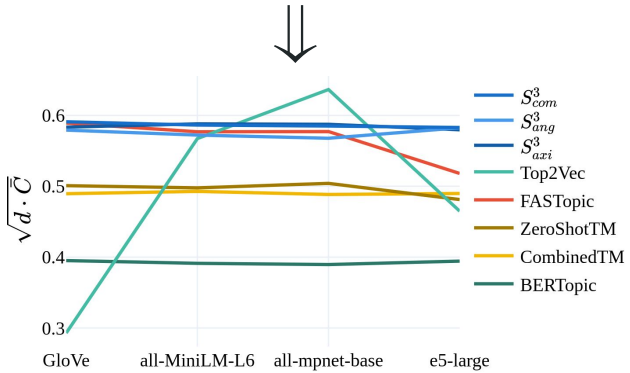
Visualization for Interpretation and Communication in Quantitative Research

Quantitative Research 101

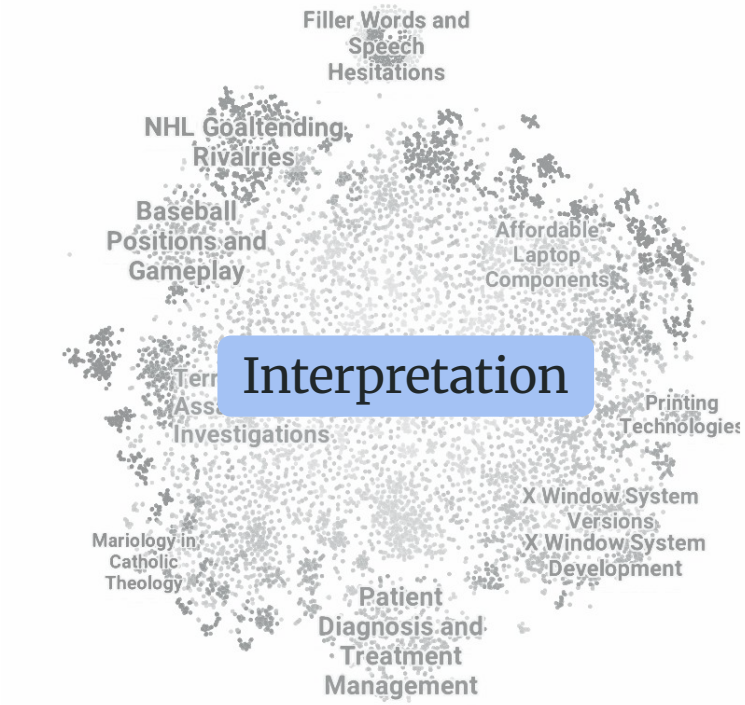
1. You have a clearly delineated and well-thought-out research question.
2. Find or collect some data that is at least tangentially related to your hypothesis.
3. Find out that your data, and by extension the real world, is far more complicated than you had imagined.
4. Denial, anger, bargaining, depression
5. **Acceptance**

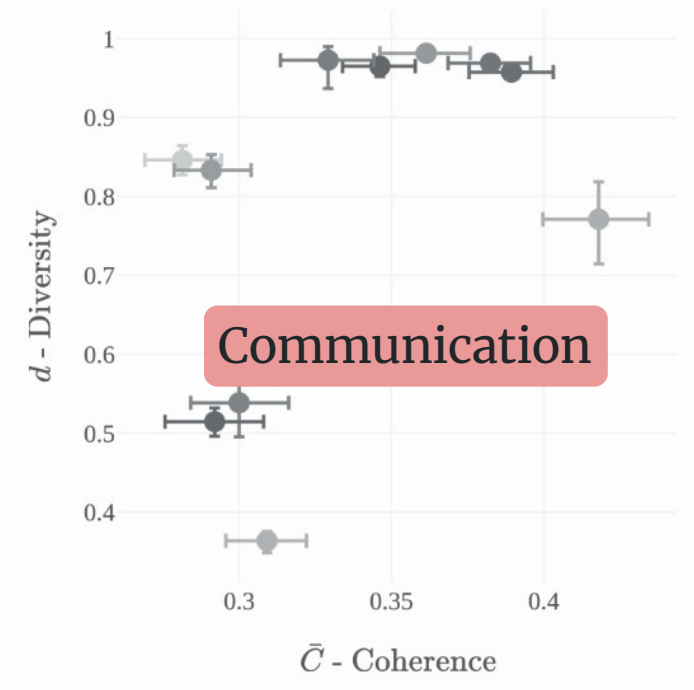
Navigating Complexity ⇒ Use Your Eyes

n_topics int64	topic_descriptions sequence · lengths	runtime_s float64	encoder string · classes	diversity float64	c_npmi float64	wec_ex float64	wec_in float64
10–14 20.2%	6–11 19.6%	0.93–5.23k 96.7%	all-MiniLM… 30.1%	0.27–0.36 3.8%	-0.03–0.03 22.2%	0.19–0.23 27.4%	0.5–0.59 9.4%
10	[["the", "of", "to", "time", "and"…	14.212026	all-MiniLM-L6-v2	0.29	-0.030181	0.222051	0.532827
20	[["in", "with", "the", "image",…	15.018337	all-MiniLM-L6-v2	0.395	-0.00707	0.172834	0.582825
20	[["segmentation", "the", "and", "to",…	14.292851	all-MiniLM-L6-v2	0.395	-0.009121	0.174477	0.578752
20	[["we", "policy", "that", "learning",…	37.529427	all-MiniLM-L6-v2	0.357895	-0.008285	0.179466	0.583949
20	[["the", "of", "and", "to", "in",…	13.924605	all-MiniLM-L6-v2	0.38	-0.006415	0.178178	0.564837
30	[["distributed", "the", "federated",…	11.780045	all-MiniLM-L6-v2	0.390476	-0.006637	0.170969	0.576168



- Humans are bad with numbers
- Enormous tables are the worst thing to look at
- Your visual system is a powerful ally
- Vision is very information-rich and high-dimensional
- …yet we navigate the world with incredible efficiency


Filler Words and Speech Hesitations
NHL Goaltending Rivalries
Baseball Positions and Gameplay
Affordable Laptop Components
Interpretation
Printing Technologies
Investigations
X Window System Versions
X Window System Development
Mariology in Catholic Theology
Patient Diagnosis and Treatment Management


Communication
1
0.9
0.8
0.7
0.6
0.5
0.4
d - Diversity
0.3
0.35
0.4
$\bar{C}$ - Coherence

Interpretation is about Gaining Intuition

Exploratory Data Analysis

Data ⇒ Visualization ⇒ Intuition

You want to gain an understanding of the structure of the data

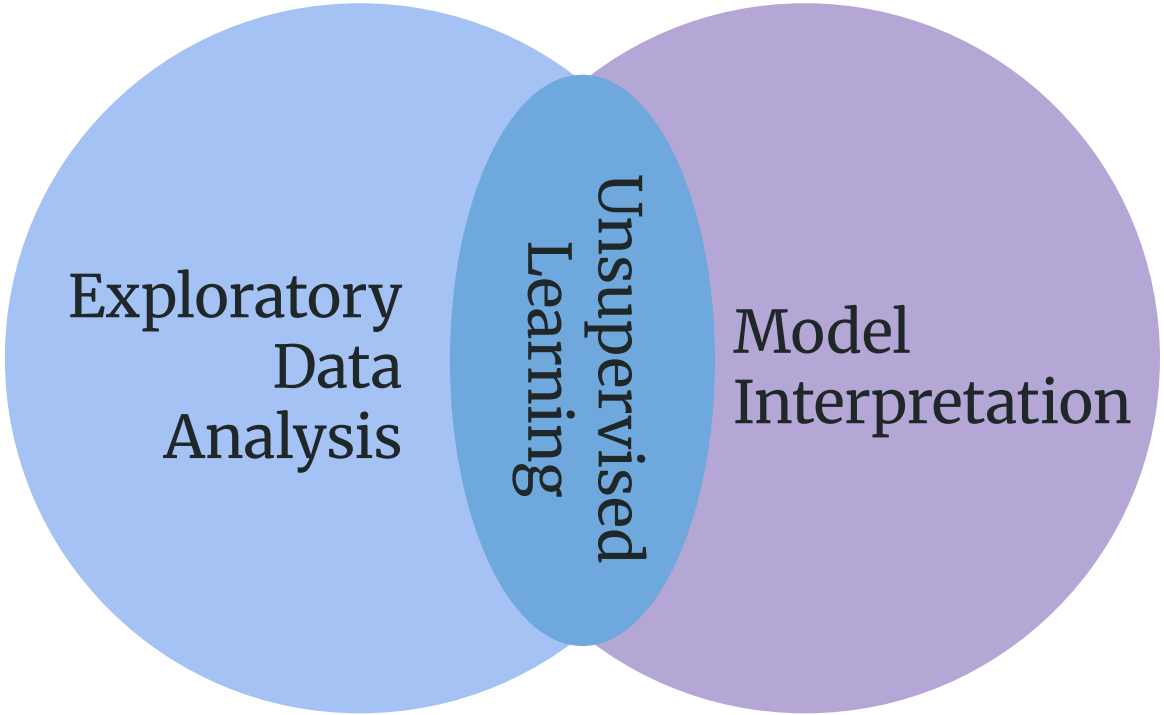
You either visualize the data directly or use a model to discover structure

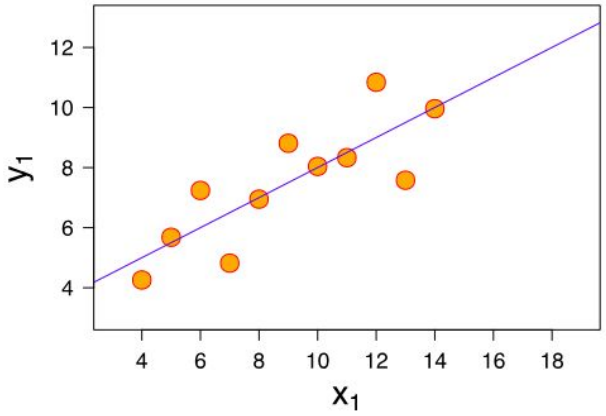
Model Interpretation

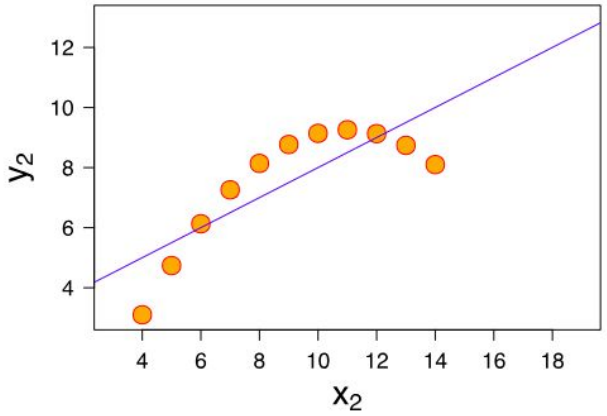
Data ⇒ Model ⇒ Visualization ⇒ Insight

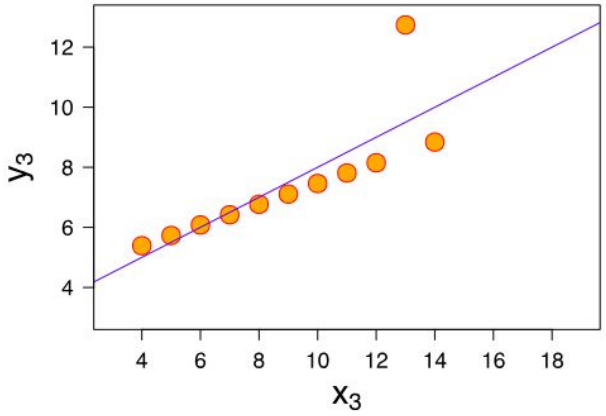
You have a model, and you want to understand its outputs

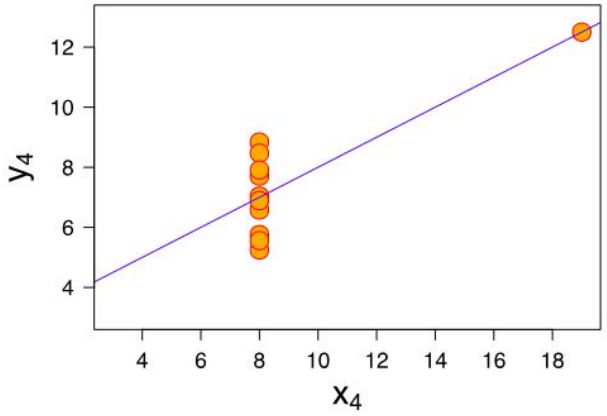
You want to draw meaningful conclusions from the model's parameters

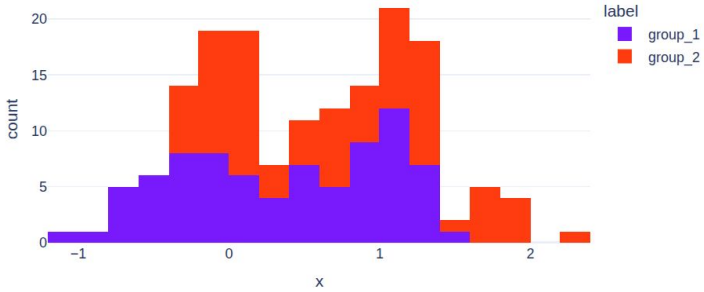

Exploratory Data Analysis
Unsupervised Learning
Model Interpretation

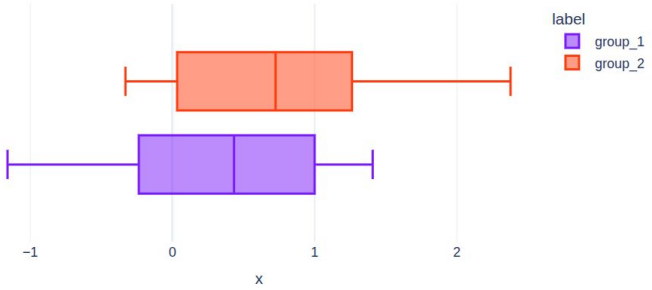

y1
x1
4
6
8
10
12
14
16
18

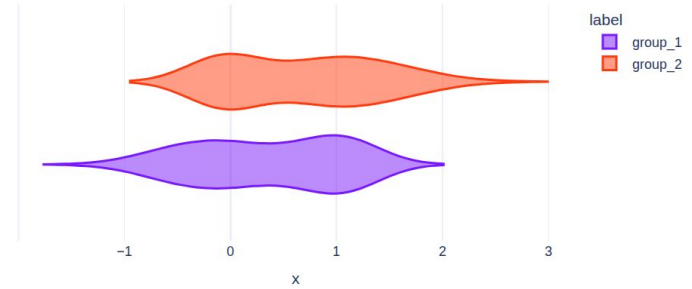

y2
x2
4
6
8
10
12
14
16
18


y3
x3
4
6
8
10
12
14
16
18


y4
x4
4
6
8
10
12
14
16
18

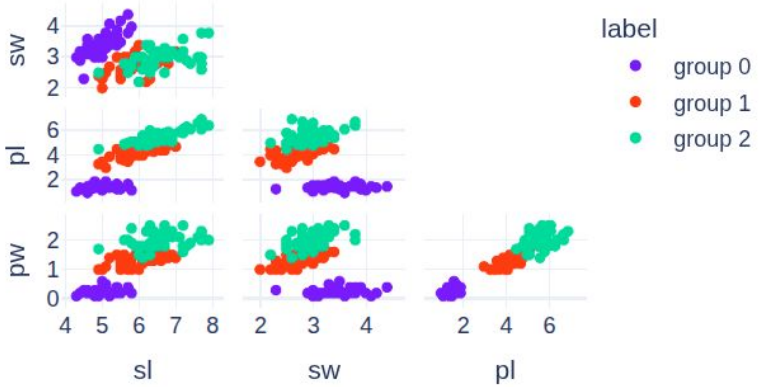






Exploratory Data Analysis

- Investigate **distributions**
- Different tools for different **dimensionality**
- Single dimension (*marginal plots*):
  - Box
  - Histogram
  - Violin
- An alternative for **descriptive statistics**

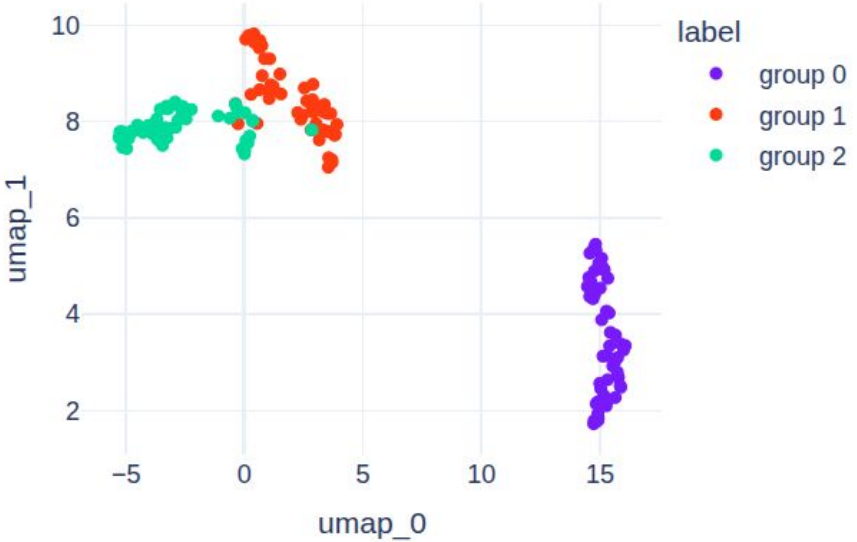


Multiple Dimensions

- Interacting variables ⇒ Increased complexity
- Marginal distributions might not tell you the whole story
- Scatterplots ⇒ Scatterplot matrix
- Bubble charts
- Heatmaps
- ...but unfortunately, data is usually much higher dimensionality than we can visualize

Visualizing High-Dimensional Data

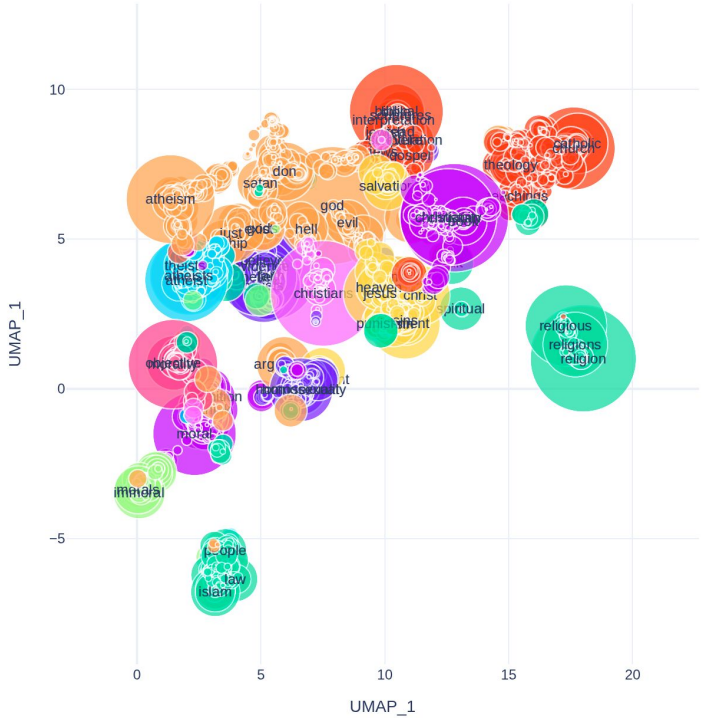
Data ⇒ UMAP ⇒ N dimensions



- Principal Component Analysis
- Manifold methods:
  - UMAP
  - TSNE
- These dimensions are typically rather abstract
- Relative positions mean more than absolute ones
- This is already **unsupervised learning**

Visualizing Semantic Data: Images and Text

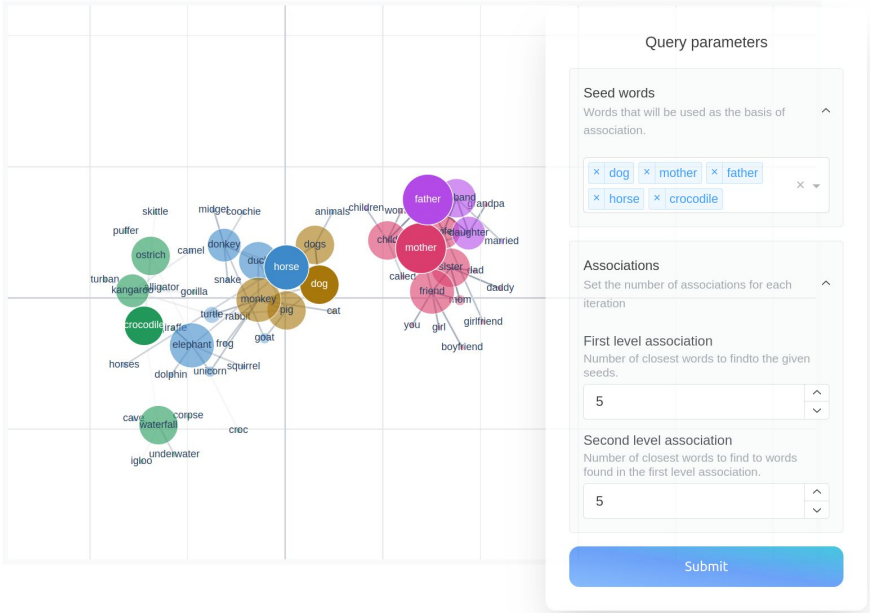
Image|Text ⇒ Embeddings ⇒ UMAP



- Text is unlike most other data:
  - No clear numerical features
  - Context-dependent
  - Has **meaning**
- Images are peculiar too:
  - Colour, Space
  - Very High-dimensional
- **Embedding models** ⇒ High-dimensional representations
- Optional Clustering (this is almost *topic modelling*)

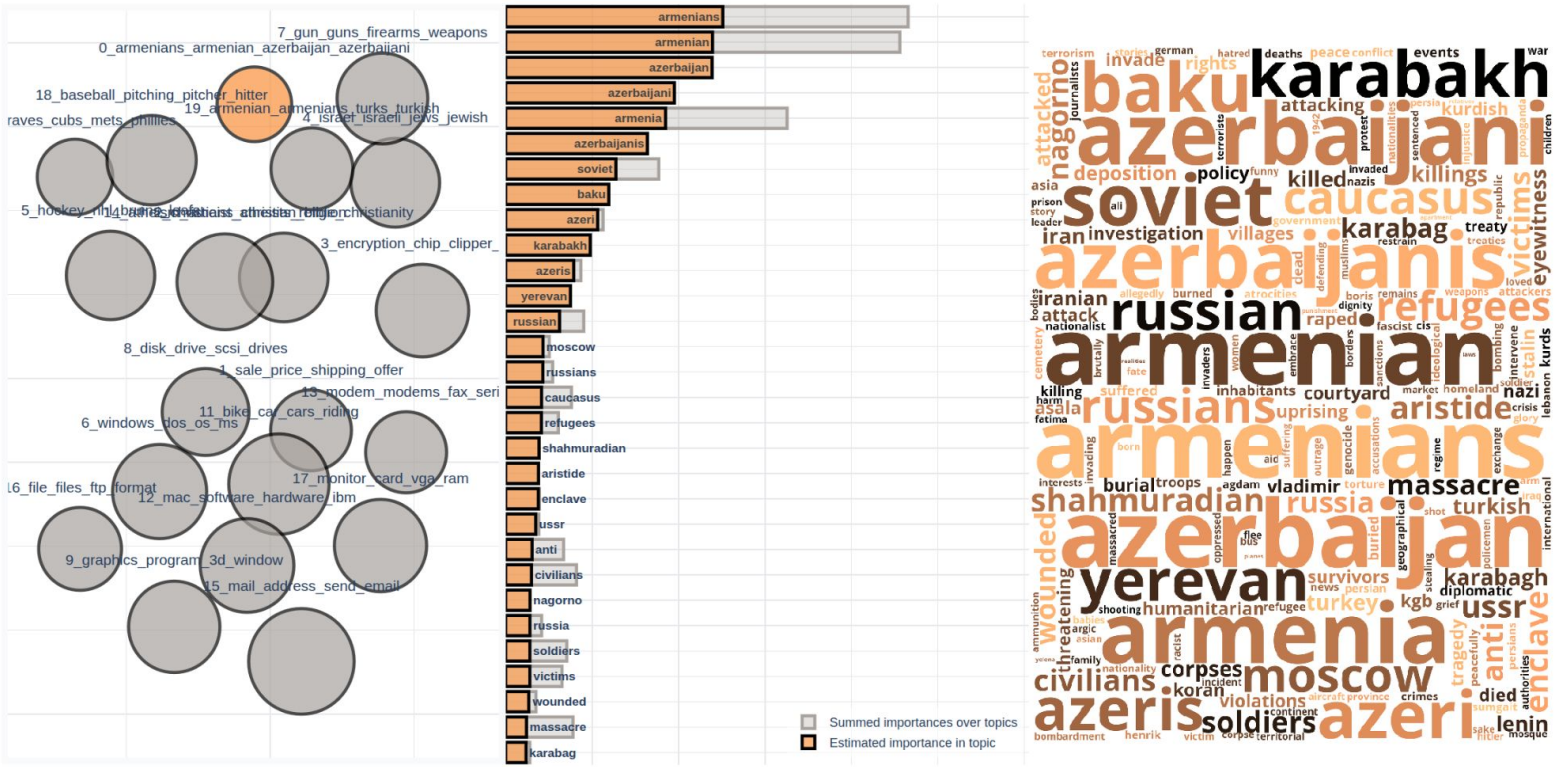
Unsupervised Methods for EDA

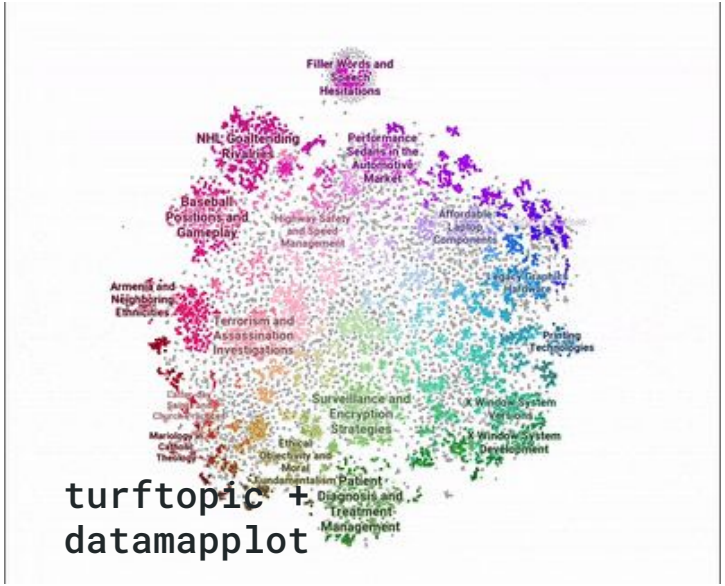
Data ⇒ Unsupervised Learning ⇒ Structure

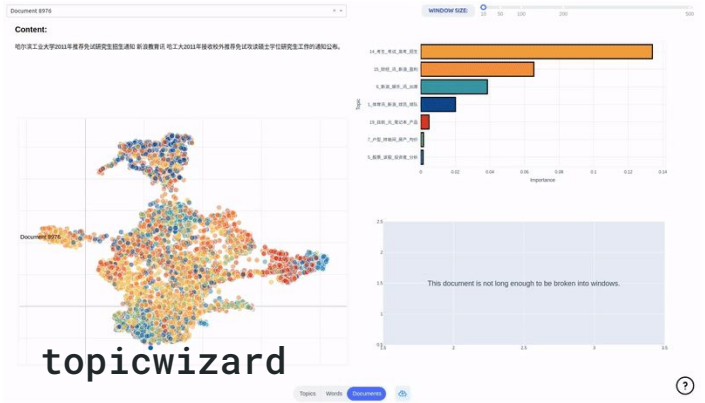


`embedding-explorer`

- No explicit hypothesis
- The data is very complex, and you might have to learn some structure in the data for exploratory analyses
- Topic modelling
  - Learn the structure of a corpus and discover themes
- Semantic graphs
  - Learn associative relations between concepts in a corpus with word embeddings

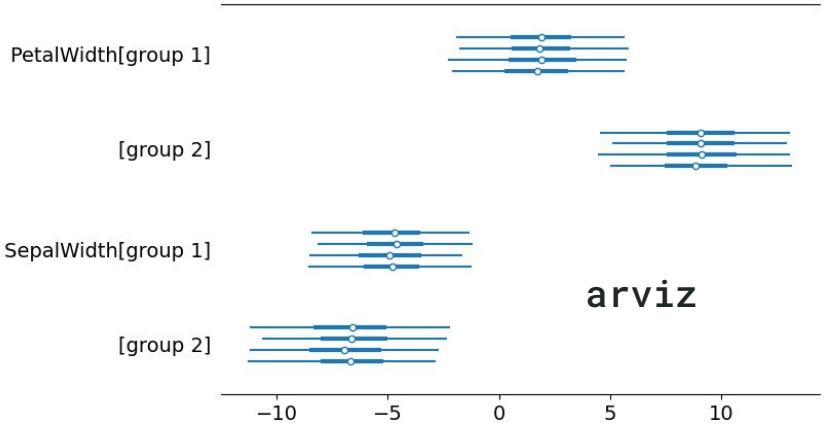

Rename topic...
7_gun_guns_firearms_weapons
0_armenians_armenian_azerbaijan_azerbaijani
18_baseball_pitching_pitcher_hitter
19_armenian_armenians_turks_turkish
raves_cubs_mets_phillies
4_israel_israeli_jews_jewish
5_hockey_nhl_
3_encryption_chip_clipper_
8_disk_drive_scsi_drives
1_sale_price_shipping_offer
13_modem_modems_fax_seri
11_bike_car_cars_riding
6_windows_dos_os_ms
17_monitor_card_vga_ram
6_file_files_ftp_format
12_mac_software_hardware_ibm
9_graphics_program_3d_window
15_mail_address_send_email
armenians
armenian
azerbaijan
azerbaijani
armenia
azerbaijanis
soviet
baku
azeri
karabakh
azeris
yerevan
russian
moscow
russians
caucasus
refugees
shahmuradian
aristide
enclave
ussr
anti
civilians
nagorno
russia
soldiers
victims
wounded
massacre
karabag
Summed importances over topics
Estimated importance in topic
baku
karabakh
azerbaijani
soviet
caucasus
azerbaijan
armenian
russian
refugees
armenians
russians
aristide
shahmuradian
azerbaijan
yerevan
armenia
moscow
azeris
soldiers
azeri
topicwizard
Topics
Words
Documents
Groups

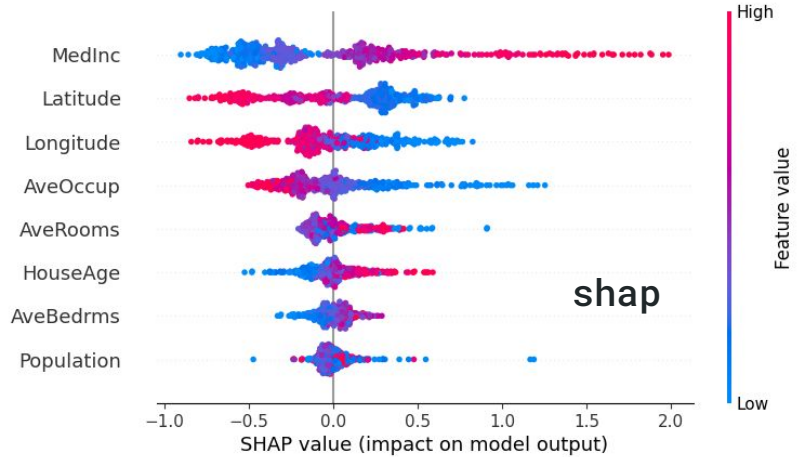




A Comment on Interactivity

- We don't really look at plots on paper most of the time
- We can do things with a cursor:
  - Zoom, Pan, Hover etc.
- You can fit much more in a graph than otherwise would be possible
- **Granular|Hierarchical** insights
- The **Plotly** library is great (Python and R)
- Datamapplot for large datasets

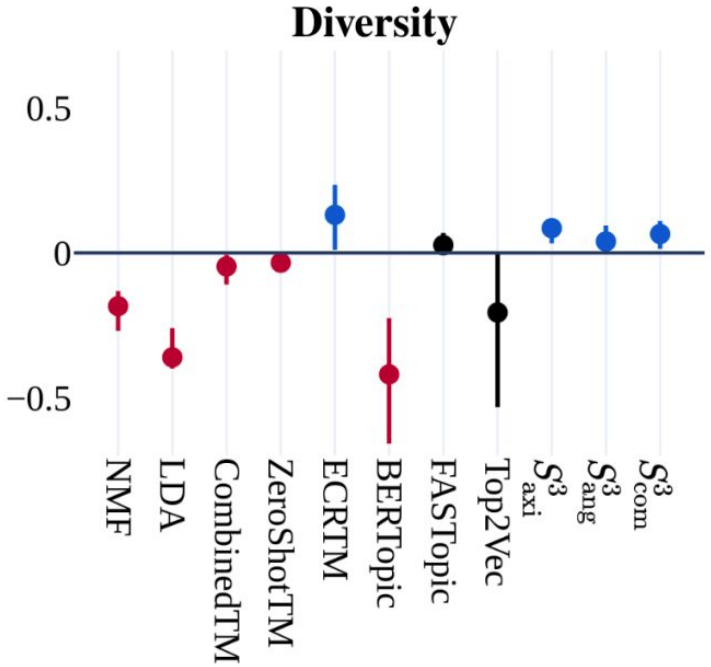




Model Interpretation

Data ⇒ Model ⇒ Visualization ⇒ Insight

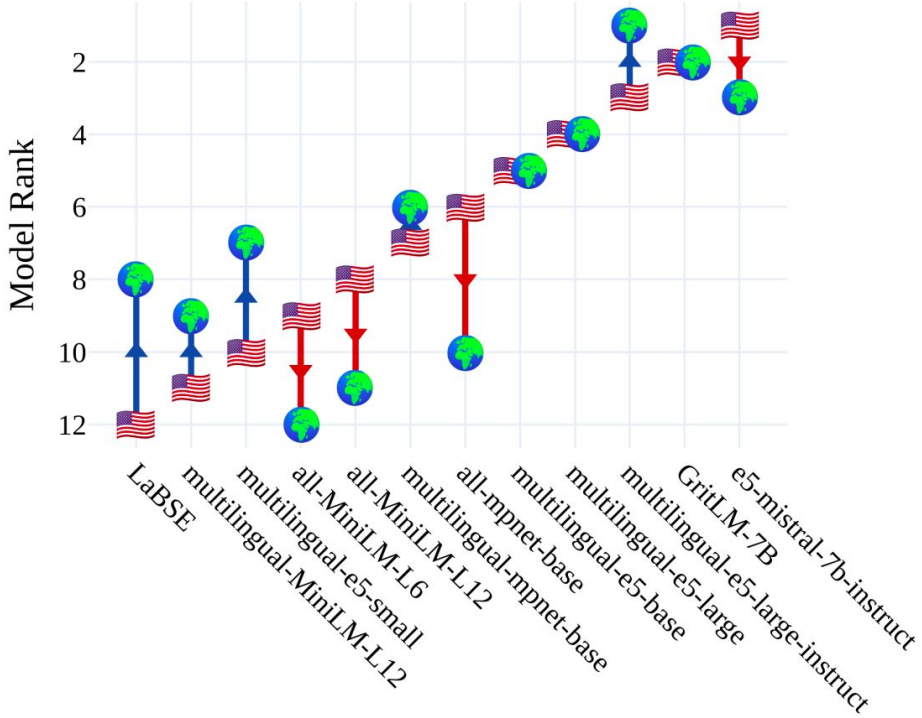
- You have an explicit hypothesis
- … which you formulate into a model
- Hard problems ⇒ Complex models
- Model parameters still need interpretation
- Visualization can give you a richer and more precise understanding
- Effect sizes, model diagnostics etc.
- White-box | Black-box

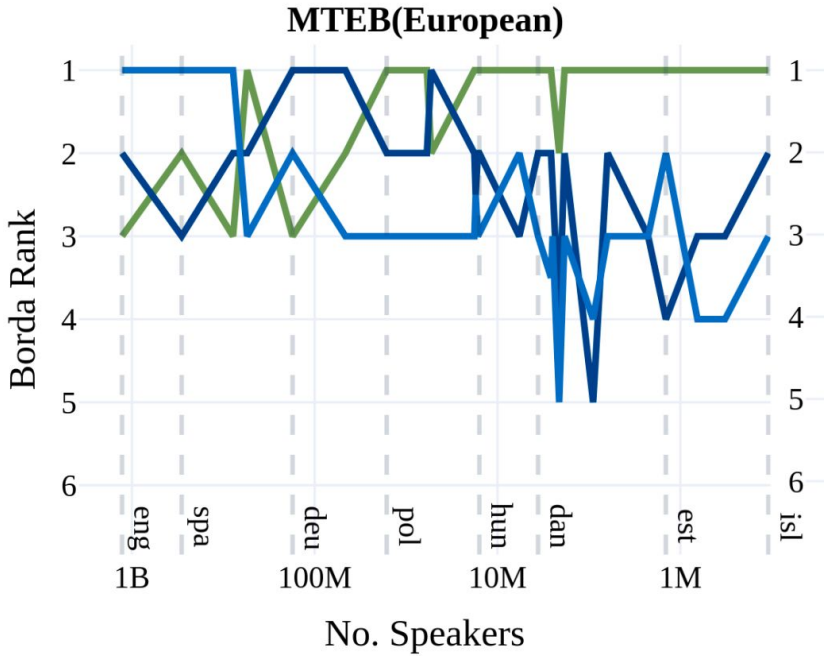


S^3 - Semantic Signal Separation

Visualization is Communication

- Having insights and results is not enough
- We need to communicate them
- Good reports
  - are easy to understand
  - empower the reader
- A simple plot can be **more honest** than a table, because you don't hide behind a plethora of numbers.
- *Show readers what you want them to see, but be honest and transparent.*





MMTEB: Massive Multilingual Text Embedding Benchmark

Thank You for Listening!