LECTURE 1

GEOSPATIAL REPRESENTATION LEARNING

Introduction to Geospatial Representations

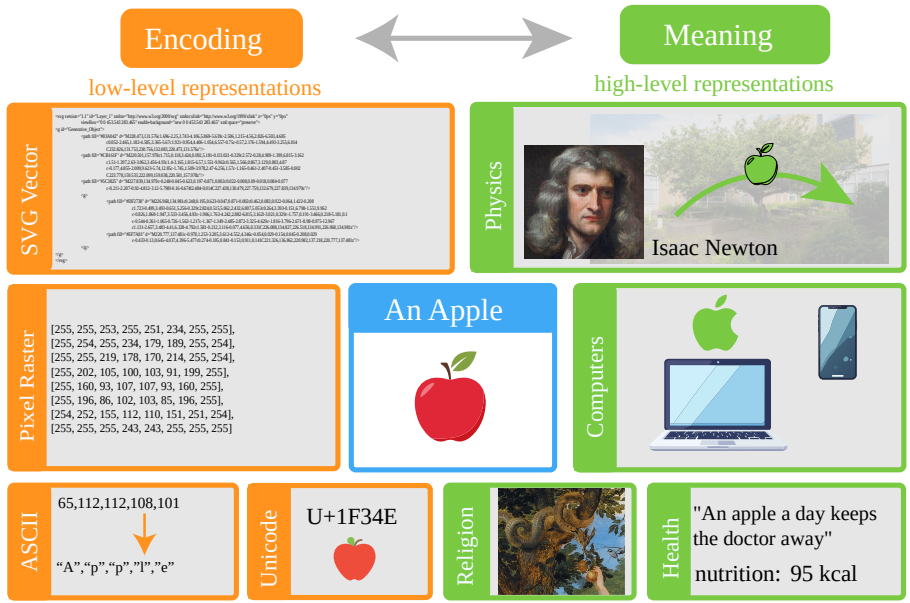
PRESS → OR SPACE.

UNIVERSITÄT BONN

Representations of our World

Describe an Apple.

Meaning and Encoding of Representations



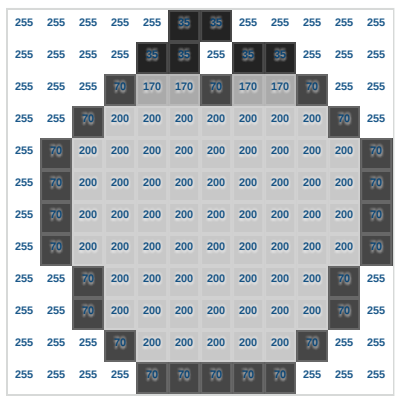
The same object can be encoded in many ways. A good encoding representation makes useful structure visible and computation easier.

But representation is not only encoding.
Representation is also meaning.

Choosing Right Encoding Representations

Examples

Images



Math

```
XLVII + LXXVIII = ?
```

47 + 78 = 125

Places

COORDINATES

50°43'57.73" N,
7°06'16.63" E

PLACE

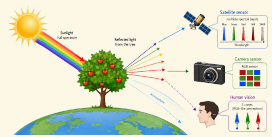
Hofgarten, Bonn, Germany

Lecture Outline

Representations

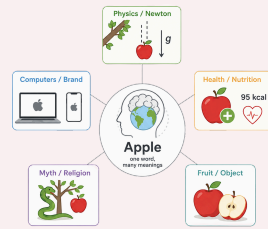
What are representations?

Vision



How vision encodes the world

Language



How text becomes meaning

Spatial

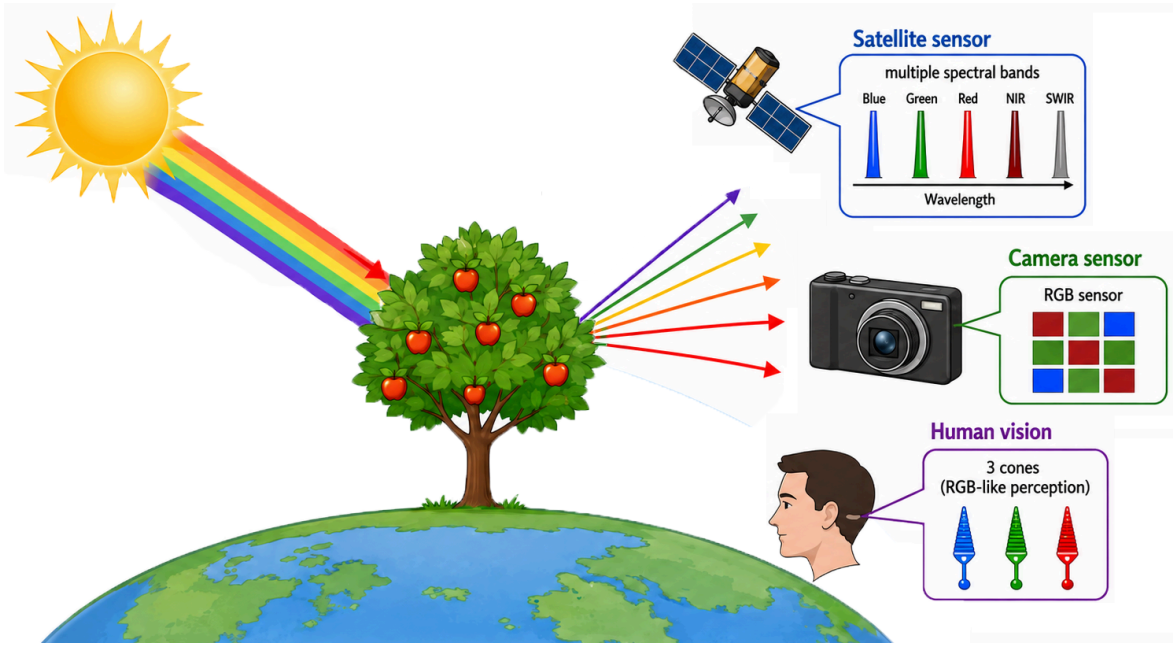
How places carry spatial meaning

Die Stadt in meinem Kopf. (2016). *Süddeutsche Zeitung*.

Visual Representations

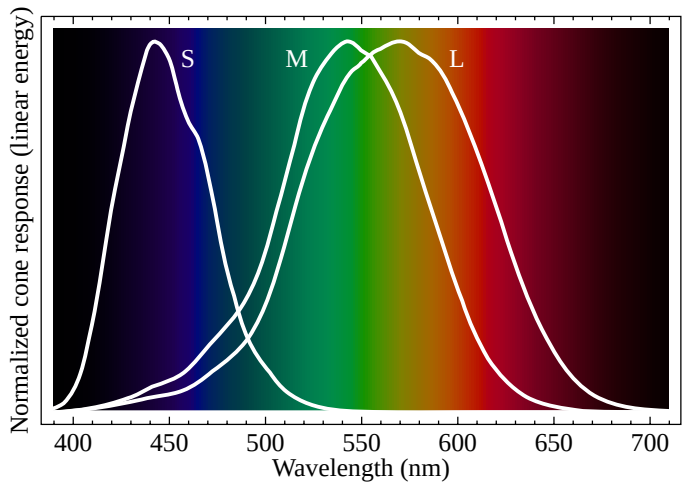
We perceive our world in representations

Example: Vision

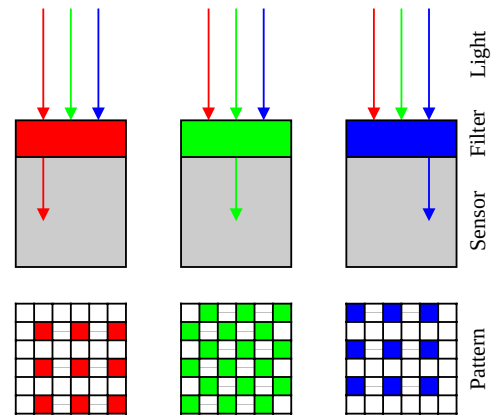


Engineering suitable Reprepresentations.

Sensitivity of cone cells (S, M, L) in human eyes

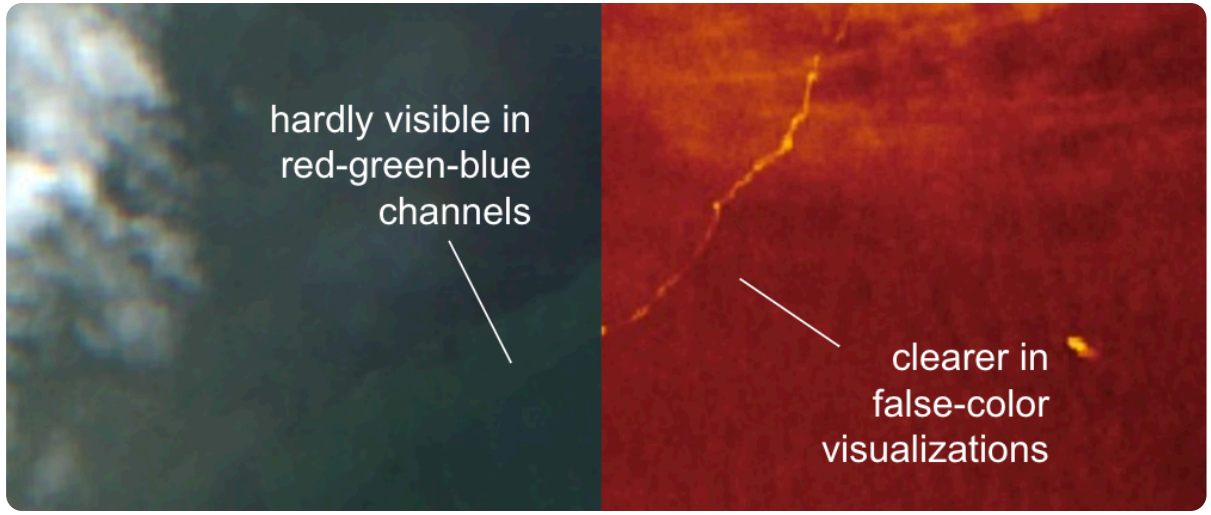


Bayer Filter on Cameras



Cameras are built to mimic our eyes: More green pixels in CCD camera mimics the color sensitivity in our eyes.

Spectral information can make patterns visible.



Example: plastic litter and marine debris (Sentinel-2)

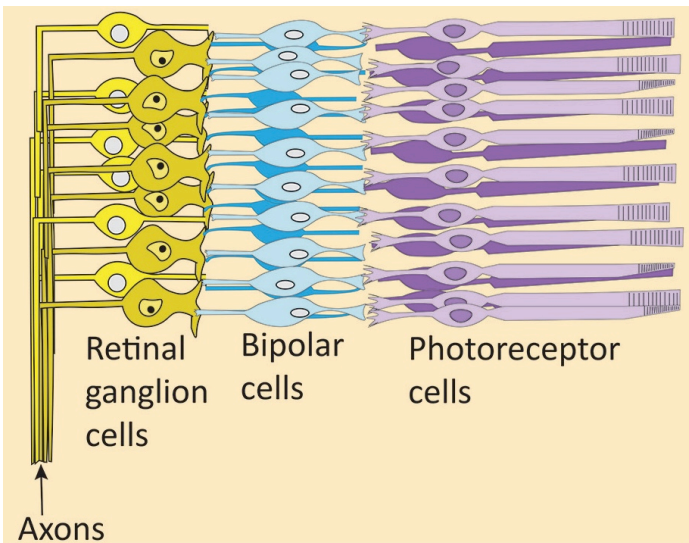
The representation depends on the Problem

A useful representation depends on the problem we are trying to solve.

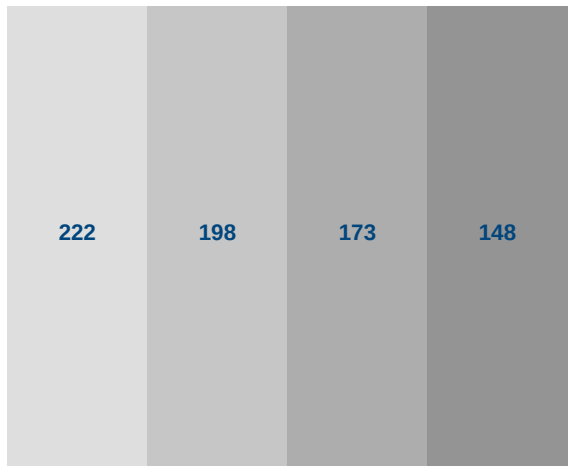
Bees see ultraviolet patterns that humans miss, because their visual system is tuned to different tasks.

Our eyes have edge detectors

Retinal cells implement an edge filter through lateral inhibition.



Mach Bands - Optical Illusion (Mach, 1865)



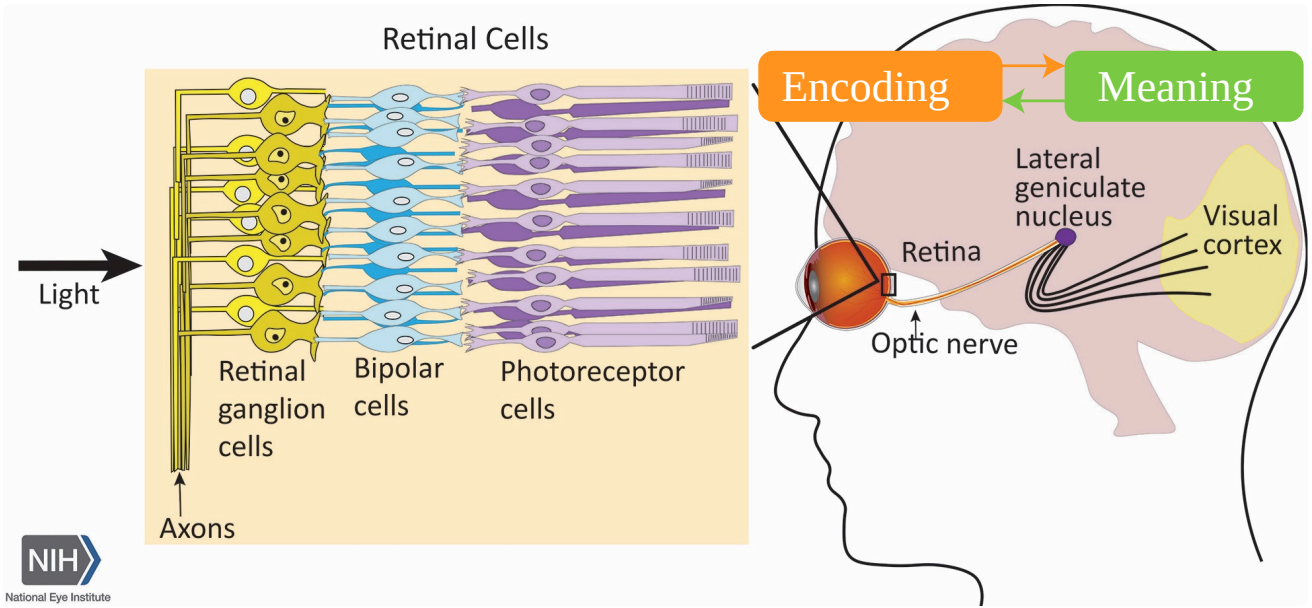
Our eyes are not objective, they capture an edge-enhanced representation of the world.

Shapley, R. M., & Tolhurst, D. J. (1973). Edge detectors in human vision. *Journal of Physiology*, 229, 165-183.

Selective Vision: The Monkey Business Illusion

Simons, D. J. (2010). Monkeying around with the gorillas in our midst: Familiarity with an inattentional-blindness task does not improve the detection of unexpected events. *i-Perception*. https://doi.org/10.1068/i0386

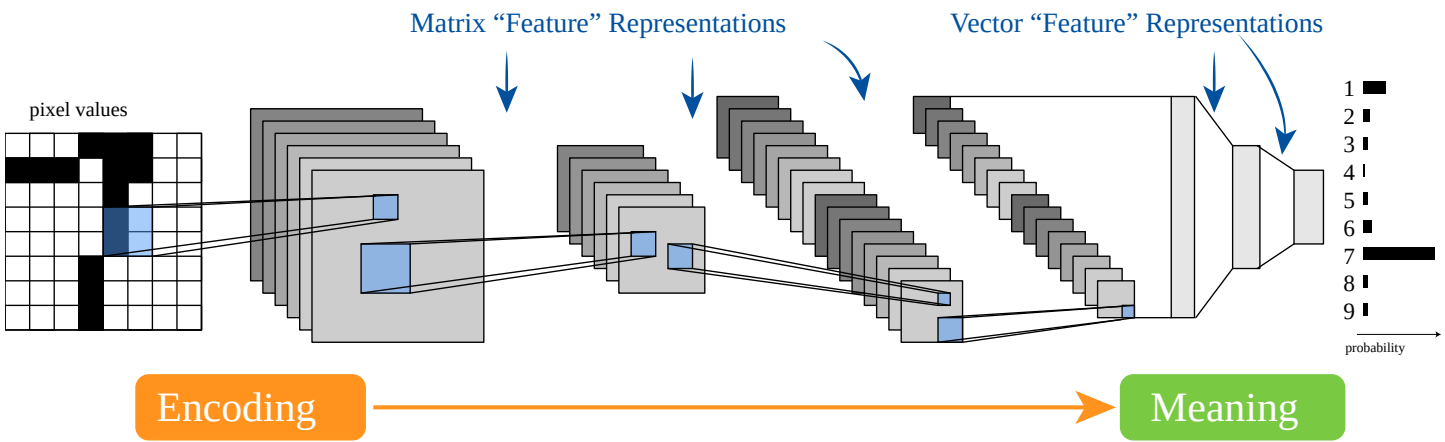
Our mind selects what we perceive



Yang, J., et al. (2025). Differential impact of retinal lesions on visual responses of LGN X and Y cells. *The Journal of Neuroscience*. DOI: 10.1523/JNEUROSCI.0436-25.2025.

From Encoding to Meaning

Feature Learning in Neural Networks



Excurse: Dimensionality Reduction

We often need to visualize high-dimensional data on a lower-dimensional space, as here with 3D data on a 2D screen.

Principal Component Analysis (PCA) linearly rotates high-dimensional data points so that a 2D projection best captures the variance (i.e., spread) of the data

Rotate to PCA projection | Reset | ☑ Show projections

t-distributed Stochastic Neighborhood Embedding (t-SNE) moves data points onto a 2D plane and tries to preserve the local neighborhood relation between points.

Run t-SNE projection | Reset | ☑ Show trajectories

Feature Representations in LeNet (2000)

Supervised Learning: From Labels to Loss

Learning from labeled images

- **Input:** a 28×28 image of a handwritten digit
- **Known truth:** the human-provided label `7`
- **Prediction:** probabilities over the ten digit classes
- **Learning:** iteratively adjust the model to minimize the prediction error
- **Goal:** make the probability for the true label `7` as high as possible

`image pixels → CNN representation → class probabilities → label`

> **How do we measure whether the prediction is correct or incorrect?**

Classification example: Cross-Entropy Loss

$$\mathcal{L}_{CE}(y, \hat{y}) = -\sum_{c=1}^{C} y_c \log(\hat{y}_c)$$

For a one-hot target label:

$$\mathcal{L}_{CE} = -\log(\hat{y}_{\text{correct}})$$

Correct class: `7`

- If `p(7) = 0.90` → low loss
- If `p(7) = 0.10` → high loss

> **Cross entropy rewards high probability on the correct class and penalizes confidence in wrong classes.**

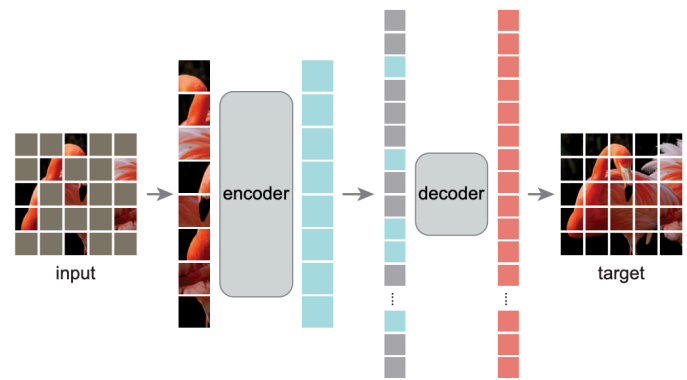
Self-supervised Learning

Contrastive Learning

Identify the corresponding object among two different views of the same object.

Masked Autoencoding

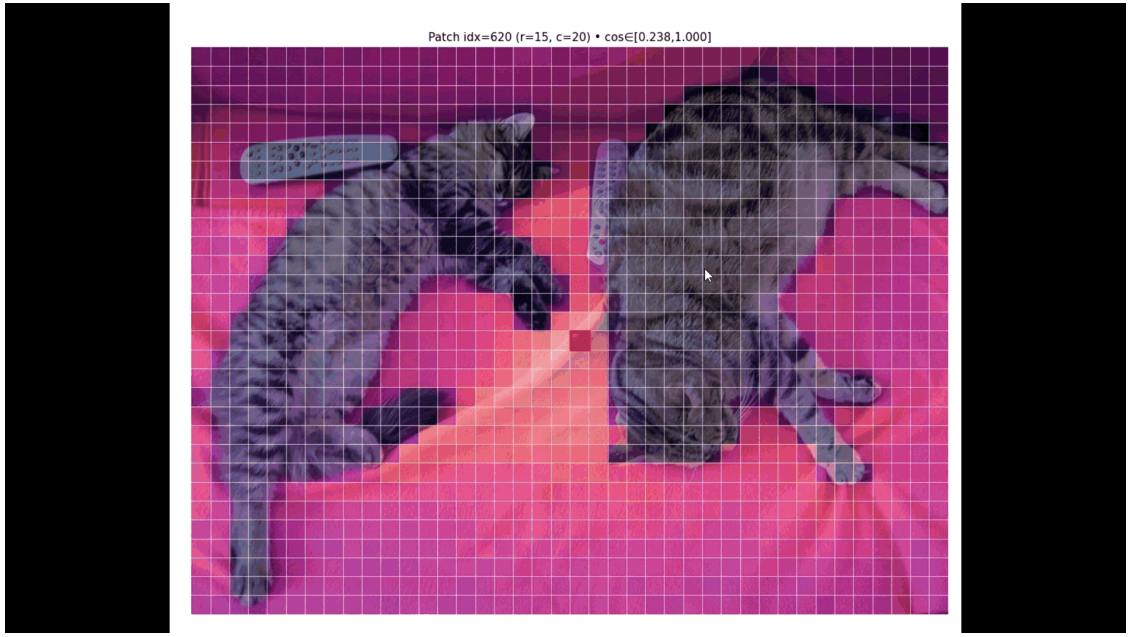
Train the model to correctly inpaint a masked portion of an image.



Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). **Unsupervised Learning of Visual Features by Contrasting Cluster Assignments** . NeurIPS. Code: **facebookresearch/swav**.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). **Masked autoencoders are scalable vision learners** . CVPR, 16000-16009.

High-level Visual Features (Dinov3, 2025)



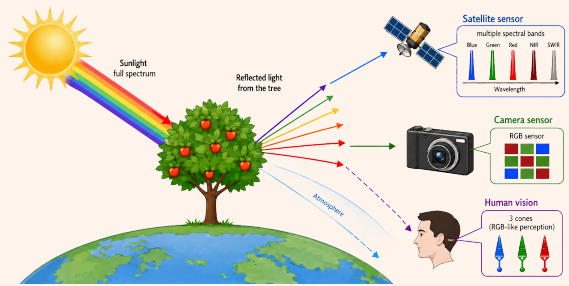
Demo: **devMuniz02/DINOv3-Interactive-Patch-Cosine-Similarity**
DINOv3 model: **Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al. (2025). DINOv3. *arXiv:2508.10104*.**

High-level Visual Features (SAM 2, 2024)

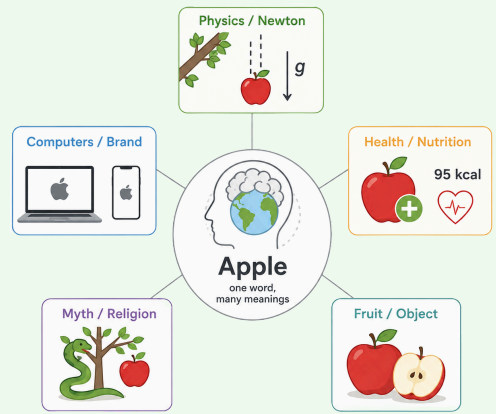
Try the Segment Anything Model (SAM) 2 demo

From Physical to Societal Representations

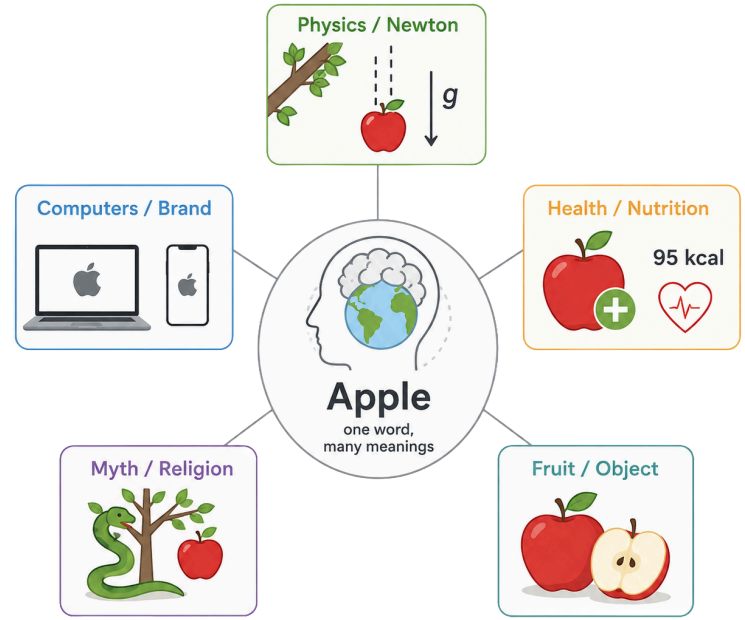
Observing the Physical World with Vision



Understanding Societies by Language



Language Representations



From Text Encoding to Language Models

Words

"Apple"

Text is a human-readable representation of concepts and meaning.

→

Token IDs

[23182]

A tokenizer maps text pieces to numerical IDs that a computer can process.

→

Embeddings

```
[0.12,
-0.41,
... ]
```

Embeddings turn tokens into dense vectors to encode semantic relationships.

→

Language Models

context + prediction

Language models use context to predict likely next tokens.

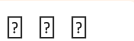
Text and language captures the human world: machines move from low-level encodings of words and tokens to high-level representations associated with meaning in context.

Character Encodings

Human writing systems

Societal history of representing meaning with symbols.

Cuneiform
c. 3200 BCE



Hieroglyphs
c. 3100 BCE

Chinese characters
c. 1200 BCE - today

Alphabetic writing
c. 700 BCE - today

A B C ...

Digital-era symbols
20th-21st century

@ # :) 😂 🌍

Learned meaning
different fonts, same symbol

A A A A A A A A

Machine character encodings

Computers encode characters as integer numbers.

ASCII
early character numbers

```
A → 65
B → 66
a → 97
```

Unicode code points
global character IDs

```
A → U+0041
ä → U+00E4
🌍 →
U+1F30D
```

UTF-8 bytes
stored byte sequences

```
A → 41
ä → C3 A4
🌍 → F0 9F
8C 8D
```

What's the result of "A" + "B" in C++? Try it out via cpp.sh

```cpp
std::cout << ('A' + 'B') << std::endl;
```

Word Tokenizer

Language Models tokenize text

Question: How many "r"s are in Strawberry?

Our view	`S t r a w b e r r y`
Model view	`[Str] [aw] [berry]`
Token IDs	`[3504, 1134, 19772]`

Tokenization depends on the tokenizer.

Modern LLM tokenizers use vocabularies of roughly 100k-200k unique tokens.

Try it yourself: platform.openai.com/tokenizer

Test: Strawberry · Schifffahrt · Poppelsdorf · Poppelsdorfer Allee · She sells seashells.

We also learned to read text in chunks

Hmuans do not raed ervey wrod letetr by letetr. We raed in chnuks, ptaterns, and expcetations. As lnog as the frist and lsat lettres are in the rghit palce, the mnidele can be sracmbeled and the txet is sitll surprisingly easy to raed.

We also use learned representations — not raw pixels or letters alone.

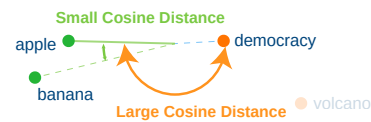
Word Embeddings

From token IDs to vectors

token	ID	vector
strawberry	8123	[0.21, -0.11, ...]
apple	15421	[-0.35, 0.42, ...]
banana	9821	[0.19, 0.05, ...]

Token IDs are arbitrary labels. Embedding vectors are learned representations that can capture meaning.

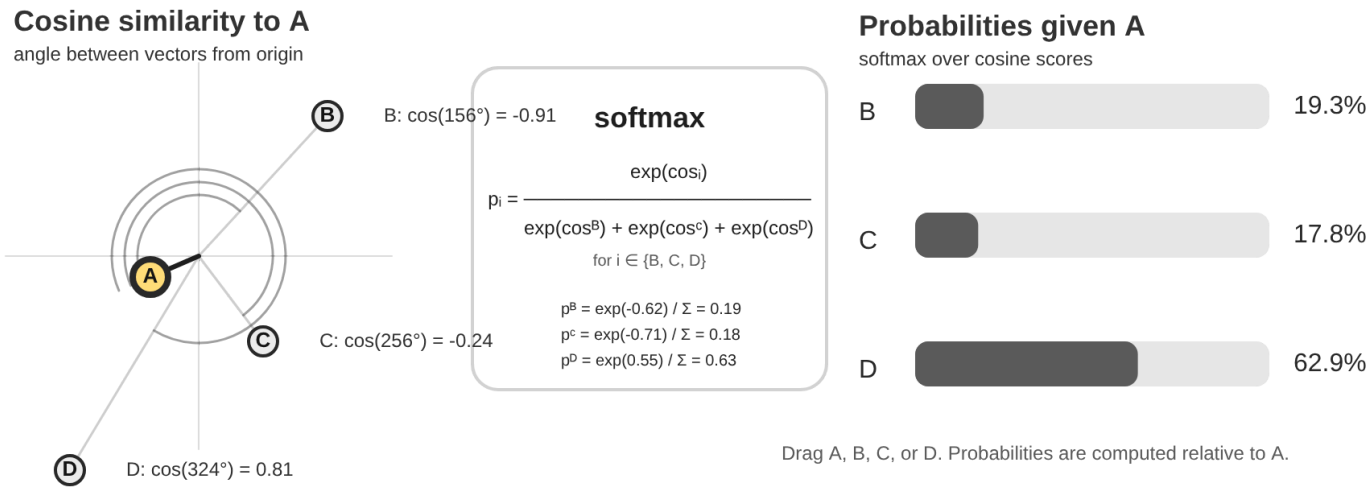
Words are points in a semantic space



Points are defined by embedding vectors

Similarity can be measured with cosine similarity.

Softmax: Cosine Similarity to Probabilities



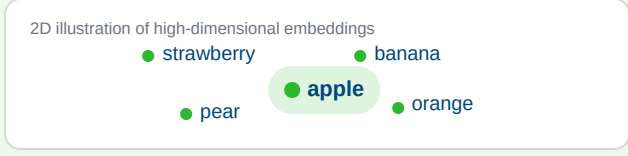
Context matters

The same word can move to different semantic neighborhoods depending on context.

Fruit context

I ate an **apple**.

`apple | "I ate an ..." → fruit-like vector`

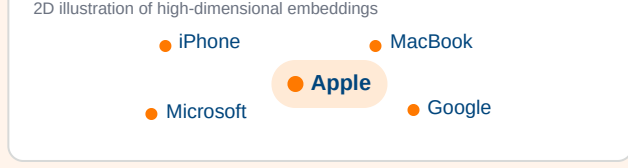


nearby words: fruit, sweet, edible

Technology context

I bought the new **Apple**.

`Apple | "I bought the new ..." → company`



nearby words: device, software, company

Language Models predict the next token.

Static embedding

`apple → one vector`

Contextual embedding

`apple in "I ate an ..." → fruit-like vector`
`Apple in "I bought the new ..." → company-like vector`

Inputs

1. word to predict: <CLS>
2. context: other text

→

Statistical language model

`p(<CLS> | context)`

→

Output probabilities

`fruit: 0.42`
`company: 0.31`
`device: 0.18`
`other: 0.09`

Try masked word prediction: huggingface.co/spaces/ysdede/fill-mask-demo

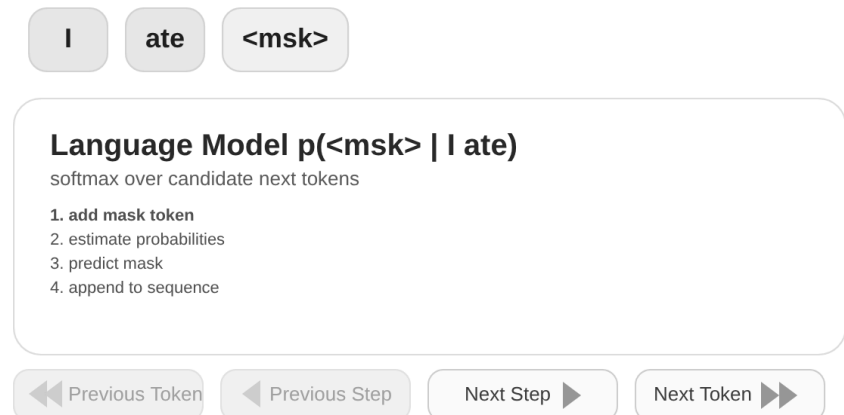
Autoregressive next token prediction

One token at a time

Language models predict the next token from a special masked position and the context that comes before it.

After a token is predicted, it is appended to the sequence. The model then repeats the same prediction step for the next position.

> **Autoregressive generation means predicting one token after another.**



Training signal

Models are trained with superised learning to minimize cross-entropy between predicted token probabilities and the actual next token.

Language Model Training

Pre-training

Language models are pre-trained to predict the next tokens as accurately as possible given an existing text corpus.

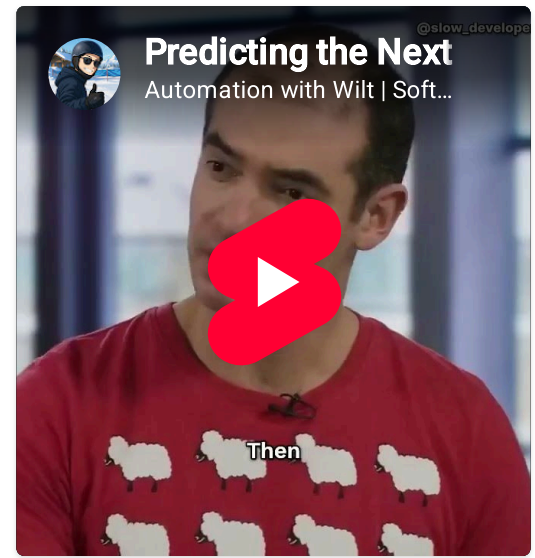
> **Predicting the next token requires world knowledge.**

Fine-tuning

They are then further fine-tuned to match developer-defined prompts.

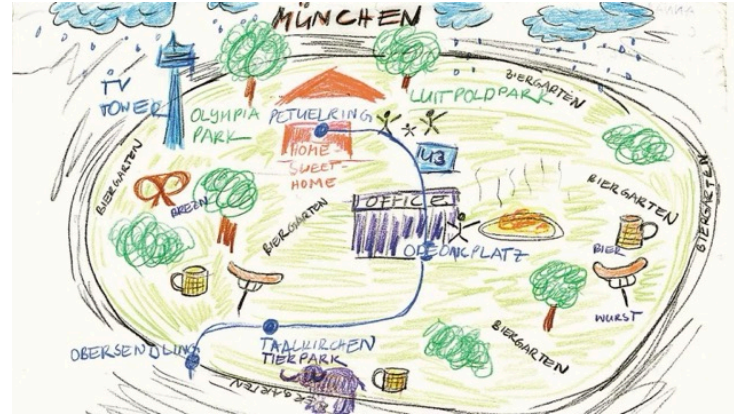
> **Fine-tuning creates AI assitants like ChatGPT, Claude, Mistral.**

More details: **Andrew Karpathy, “Deep Dive into LLMs like ChatGPT”**



NVIDIA. (2023). *Fireside Chat With Ilya Sutskever and Jensen Huang: AI Today and Vision of the Future.* YouTube short edited by @automationwithwilt .

Geospatial Representations

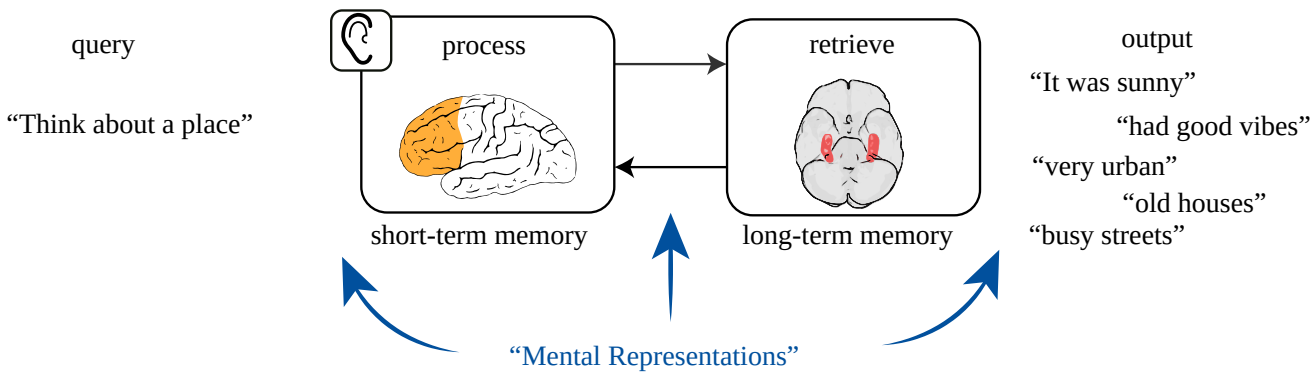


Die Stadt in meinem Kopf. (2016). *Süddeutsche Zeitung.*

Task: Think of a place

Think of a place you visited on vacation.

Not the country or city name first — think of the actual place. What do you remember?



We feel mental representations as "intuitions" and "memories" of things/places.

Preston, A. R., & Eichenbaum, H. (2013). Interplay of hippocampus and prefrontal cortex in memory. *Current Biology*, 23(17), R764-R773.

A place is more than a coordinate

Coordinate

50.7374° N, 7.0982° E

- reference
- position
- index

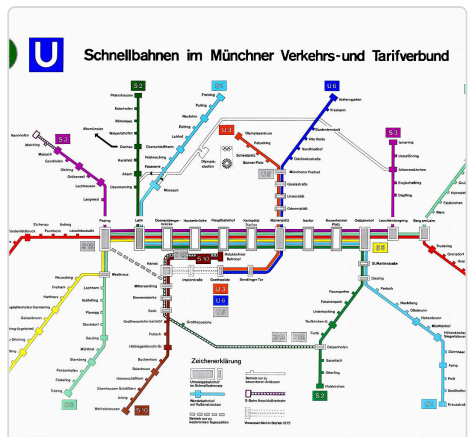
Place

- home
- routes
- landmarks
- memories
- context
- meaning

A coordinate tells us where something is. A representation tells us what that place means.

Representations of Geospatial Information

Same city, different maps



We need to select or learn a suitable representation for the problem-at-hand.

How do we experience and recognize places?

We infer location from many weak cues.

vegetation

road markings

architecture

signs and language

terrain

climate

infrastructure

We do not recognize places from one signal. We combine many imperfect signals into a coherent spatial intuition.

Spatial Representations for Earth and Society

Modern geospatial AI aims to learn representations of our planet across space, time, scale, and modality.

🌍 Environmental Systems

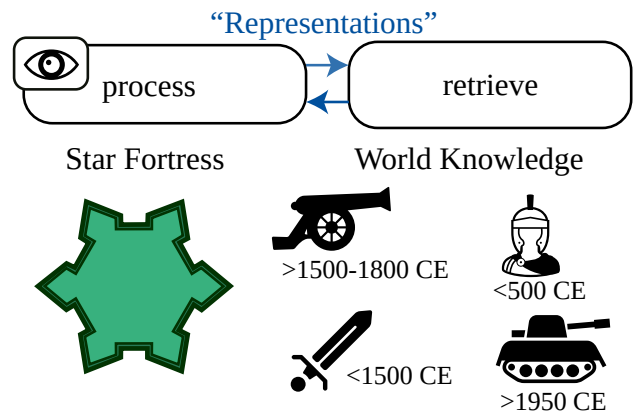
- 🌲 Deforestation and forest degradation
- 🌡️ Climate change and global warming
- 🌊 Flooding and sea level rise
- 🏜️ Desertification and drought monitoring
- 🧊 Glacier retreat and polar ice loss
- 🌾 Agriculture and crop monitoring
- 🐟 Marine ecosystems and coastal change
- 🗑️ Environmental pollution and plastic waste
- 🔥 Wildfire detection and risk prediction
- 🐘 Biodiversity and species distribution

🏙️ Societal & Human Systems

- 🏗️ Urbanization and city growth
- 🚗 Navigation and route planning
- 🚦 Traffic and mobility analysis
- 🚑 Disaster response and humanitarian aid
- 🚧 Infrastructure and construction monitoring
- ⚡ Energy systems and resource management
- 🏠 Population density and settlement mapping
- 🌍 Migration and human mobility
- 🛡️ Security and border monitoring
- 📦 Logistics and supply chain optimization

History: When was this city founded?

- ☐ 282 CE (Roman/Antique)
- ☐ 1153 CE (Midieval)
- ☑ **1593 CE (Renaissance)**
- ☐ 1975 CE (Modern)



Livability: How livable is this neighborhood?

Livability: How livable is this neighborhood?

- ☐ simple
- ☐ middle
- ☐ good
- ☑ **very good**

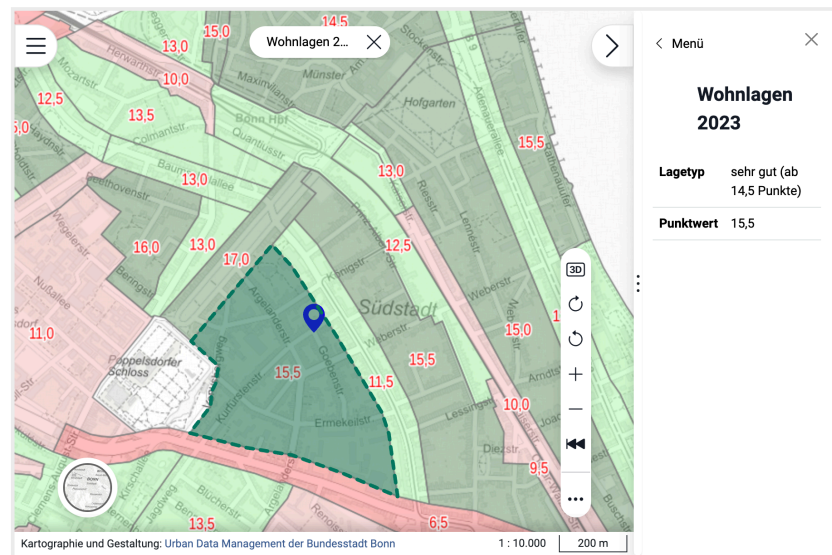
Based on which features in the image, and based on what world knowledge, did you decide?

Livability: Qualitative and Quantitative Factors

The residential location map has four levels: **simple**, **middle**, **good**, and **very good**.

It takes into account:

- Infrastructure
- Urban green space
- **Streetscape**
- Transport connection
- Appreciation
- Burden
- Centrality



Wohnlagen/Mietspiegel. Gutachterausschuss Bonn 2023.

Biodiversity: Which Agricultural Landscape supports more Diverse Bird Populations?

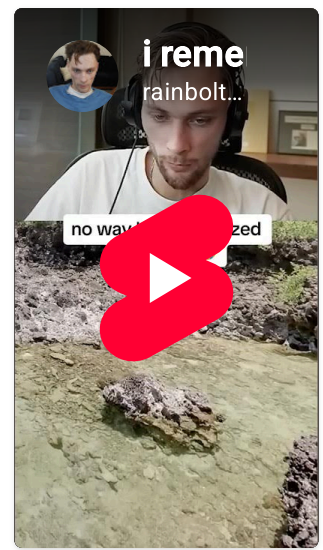
Engist, D., Finger, R., Knaus, P., Guélat, J., & Wuepper, D. (2023). Agricultural systems and biodiversity: evidence from European borders and bird populations. *Ecological Economics*, 209, 107854.

Takeaway: We learn our spatial environment

Everybody has an intuitive spatial mental representation of the world.

Some people can learn to become extremely good at mentally mapping the world.

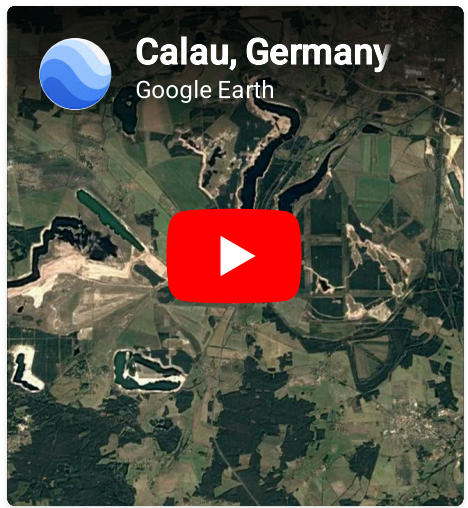
{class="max-h-full w-full object-contain rounded-xl"}



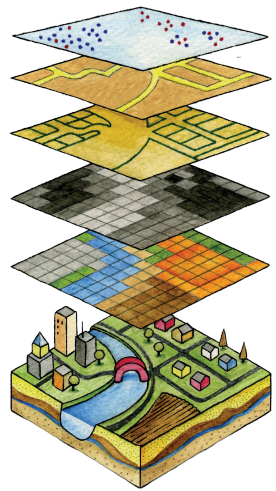
Data of our Spatial Environment

Virtual Reconstructions

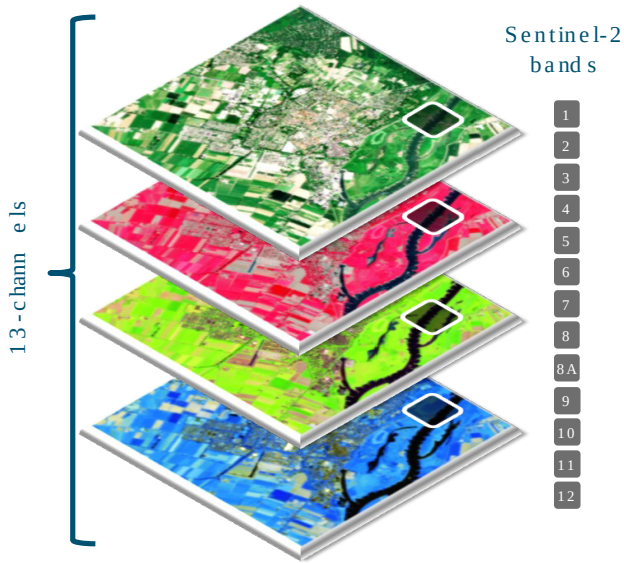
Dense Dynamic Measurements



Geospatial data layers

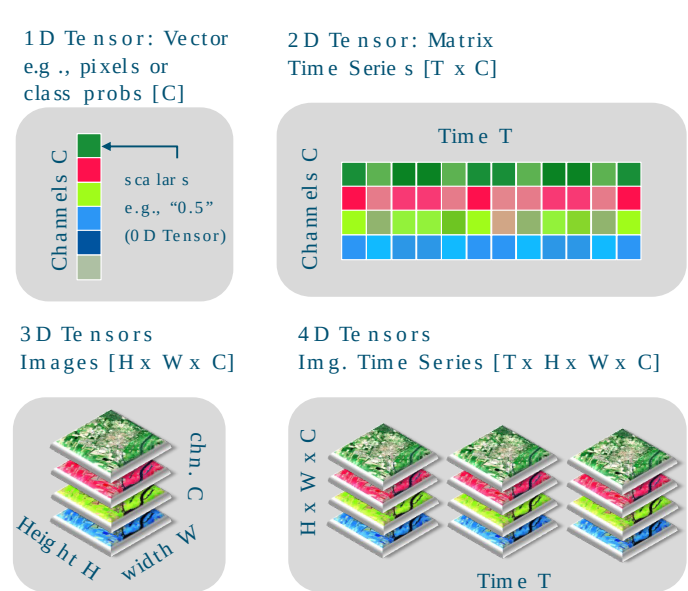


Geodata are Data Tensors



{class="h-[330px] w-full object-contain"}

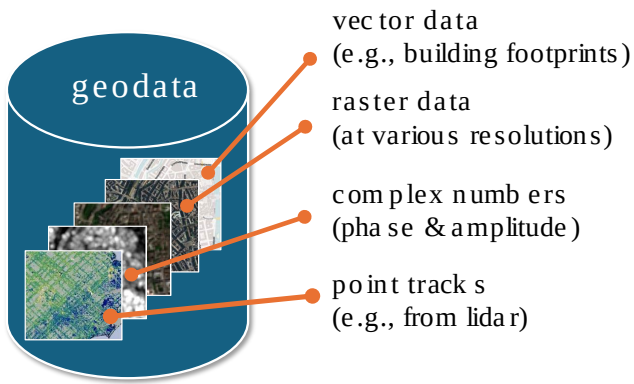
Examples of different data tensors



{class="mt-4 h-[300px] w-full object-contain"}

Heterogeneous Datasets in Geodatabases

Geodata is heterogeneous



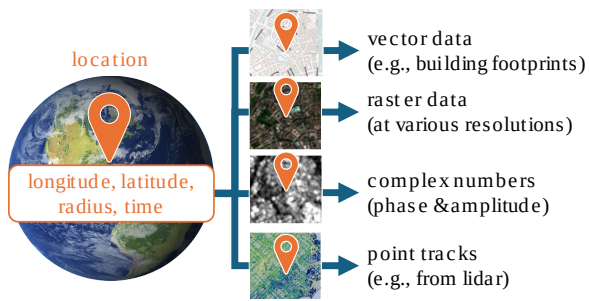
{class="mt-5 h-[330px] w-full object-contain"}

Implementation examples

- **Spatial database:** PostGIS extends PostgreSQL with spatial types, indexes, and functions.
- **Geolocated raster:** GeoTIFF stores georeferenced imagery in TIFF files.
- **Multidimensional arrays:** NetCDF stores scientific data across dimensions such as time, depth, latitude, and longitude.

Different sources use different schemas, coordinate systems, resolutions, timestamps, and file or database abstractions.

Location and Coordinates join all Geodata



Any place on Earth can be represented by coordinates

- **Longitude**
- **Latitude**
- **Elevation**
- **Time**

Coordinates join heterogeneous geodata sources together because they reference the same place and time.

Takeaways Mental Representations

Mental representations

Human memories, intuitions, expectations, and place knowledge built from experience.

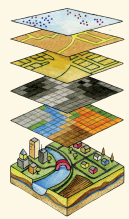
Die Stadt in meinem Kopf. (2016). *Süddeutsche Zeitung*.

Geospatial representations

Mental representations

Human memories, intuitions, expectations, and place knowledge built from experience.

Geospatial databases



Rasters, vectors, polygons, and geodata layers that encode explicit spatial structure.

Die Stadt in meinem Kopf. (2016). *Süddeutsche Zeitung*.

Closing and Takeaways

Takeaway: We Navigate the World through Representations

Biological

Perception selects and transforms reality into signals we can act on.

Engineered

Maps, coordinates, images, diagrams, and numbers make selected structure visible.

Learned

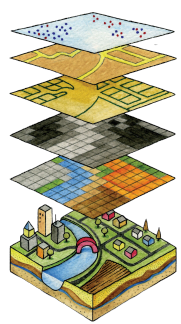
Neural networks learn representations that make patterns useful for prediction.

A representation is a useful reduction of reality: it preserves what matters for a task and hides what does not.

Choosing or Learning Representations

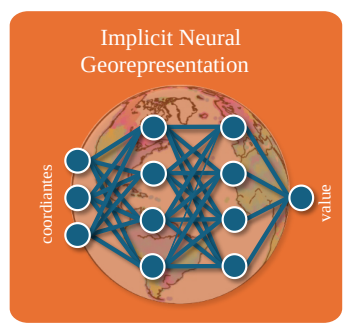
Designed representations

We choose structures that make operations easier.



Learned representations

We train models to discover useful structure.



This course focuses on learning representations.

PRACTICAL 1

EXPLORING GEOSPATIAL REPRESENTATIONS

Geospatial Representation Learning

PRESS → OR SPACE.

UNIVERSITÄT BONN