How Many Dimensions for your Latent Model? A Cross-Domain Perspective

Victor Charpenay, Rodolphe Le Riche

IMT, CNRS – LIMOS

Data&IA @ IMT, 26 sept. 2023

Introduction I

Victor Charpenay

IMT, *Informatique et systèmes intelligents*
Knowledge representation and reasoning

MINES Saint-Étienne
Une école de l'IMT

Rodolphe Le Riche

CNRS, *Génie mathématique et industriel*
Optimization and metamodeling

cnrs

Introduction II

They both carry out research in the Laboratory of Informatics, Modelling and Systems Optimization (LIMOS)

This presentation comes from an **interdisciplinary** discussion about "latent representations" that took place in Saint-Étienne.

Notice: interdisciplinarity (computer sc./math./engineering) not about an application this time

Abstract

Latent representations are ubiquitous in data analytics and AI tasks. They are used as intermediary hidden models to go from a set of observations to decisions.
Victor Charpenay and Rodolphe Le Riche confront the perspectives of their domains about these intermediary vector representations. They identify two antagonist purposes: while the latent variables of statistical models are used to ease computation, the hidden layers of neural networks are meant to capture non-trivial regularities in the observed data. The difference has consequences on the dimension of the latent feature space: looking for regularities implies finding an optimal contraction of the input data to a smaller latent space, in contrast to the infinite-dimensional vectors used in kernel based approaches.

General problem

$$\textbf{observations} \xrightarrow{?} \textbf{decision}$$

$$\text{industrial control variables} \xrightarrow{?} \text{anomaly detection}$$

$$\text{knowledge graph} \xrightarrow{?} \text{auto-completion}$$

$$\text{wind farm topology} \xrightarrow{?} \text{productivity}$$

$$\text{airfoil shape} \xrightarrow{?} \text{lift/drag prediction}$$

$$\text{mechanical measures} \xrightarrow{?} \text{material characteristics (strength } \ldots \text{)}$$

$$\textbf{observations} \xrightarrow{\phi} \text{latent representation} \xrightarrow{\psi} \textbf{decision}$$

Importance of latent representations

(one solution)

Many learning and optimization tasks rely on latent representations. . .

(to address many problems)

. . . despite the great variety in the observed data and the nature of the decision.

Question

Is there such a diversity in the role of latent representations themselves?

Heterogeneous terminology

$$\text{observations} \xrightarrow{\phi} \textbf{latent representation} \xrightarrow{\psi} \text{decision}$$

Latent: unobserved (hidden), transitive, implicit.

Other names:

- "hidden features"
  as in "feature space" of Kriging
- "latent variables"
  as in "random variables" of Expectation-Maximization
- "embeddings" into a vector space
  as in "word embeddings" of Language Models

Notation

$$x \xrightarrow{\phi(x\ ;\ \theta_\phi)} h \xrightarrow{\psi(h\ ;\ \theta_\psi)} y$$

- single observation: input vector x
- hidden vector h
- decision: output vector y
- ϕ and ψ parameterized with θ_ϕ, θ_ψ

Example: defect detection I

A typical application of machine learning in industrial systems is **defect detection**.

Hidden Markov Models (HMMs) may be used for this task [Aggarwal, 2016, ch. 10].

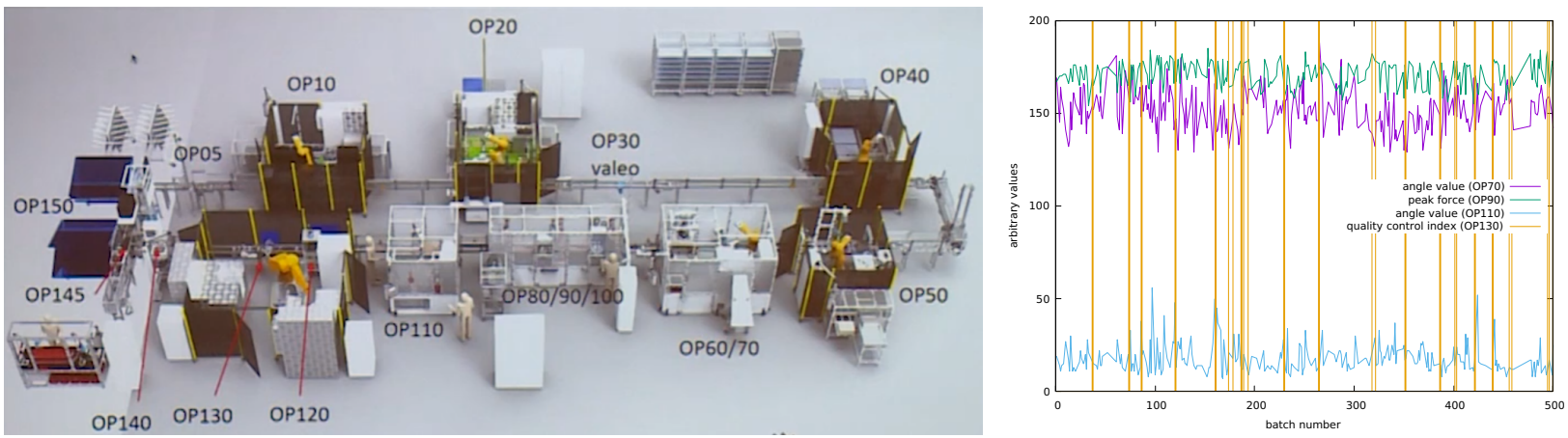


Figure: Production line (for starter motors)

Example: defect detection II

- **observations**: 14 mechanical variables characterizing work cells (angle of a robotic arm, force applied by its effector, etc.)
- **decision**: quality indicator of output product (0/1)
- ϕ: transition probability estimation
- ψ: emission probability estimation

Example: Knowledge Graph completion I

Word **embeddings** are common in Natural Language Processing.

Similar latent models can be used on structured data such as **Knowledge Graphs** to perform **inductive reasoning** [Hogan et al., 2021, ch. 5].

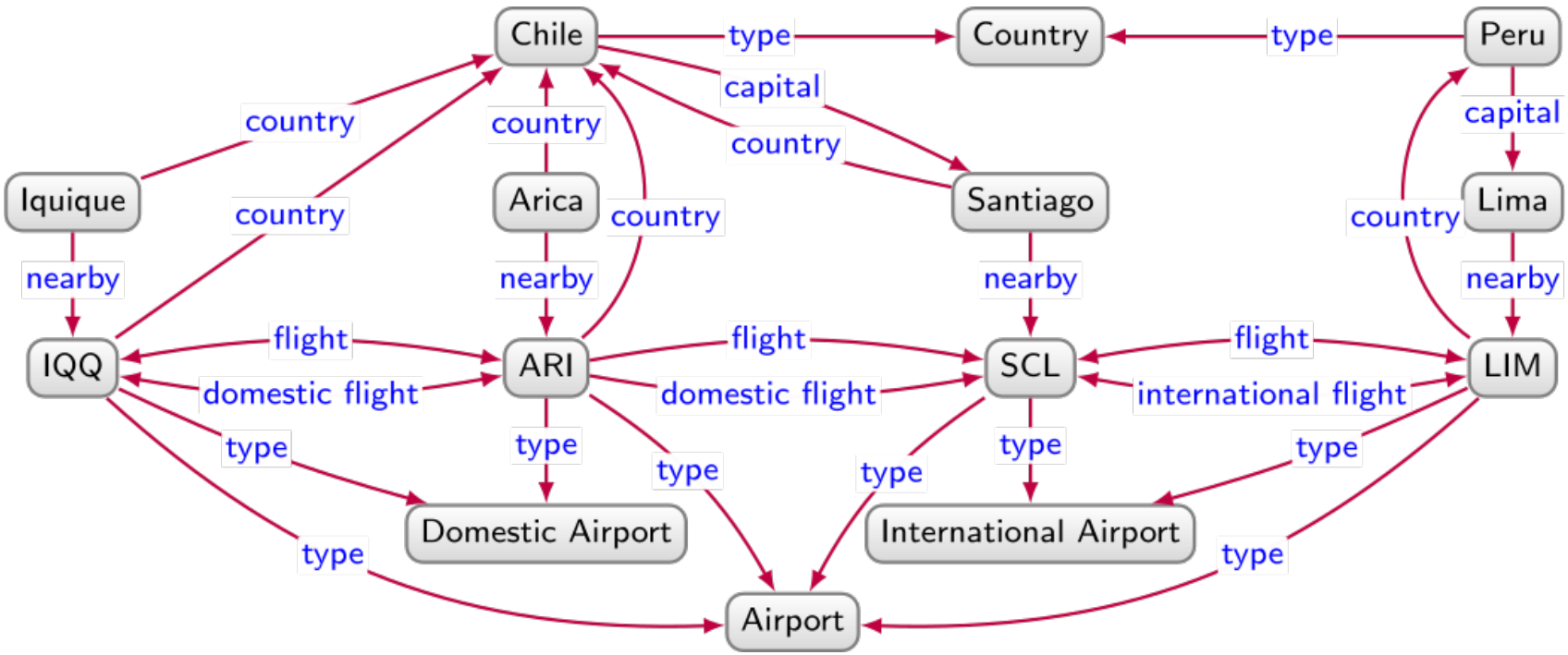


Figure: Airports Knowledge Graph

Example: Knowledge Graph completion II

- **observations**: a set of $e_1 \xrightarrow{r} e_2$ edges,
- **decision**: predict the plausibility of any edge.
- ϕ: embedding array lookup
- ψ: geometric transformation, e.g. $-\|\phi(e_1) + \phi(r) - \phi(e_2)\|$ or $\phi(e_1)\phi(r)^D\phi(e_2)^T$

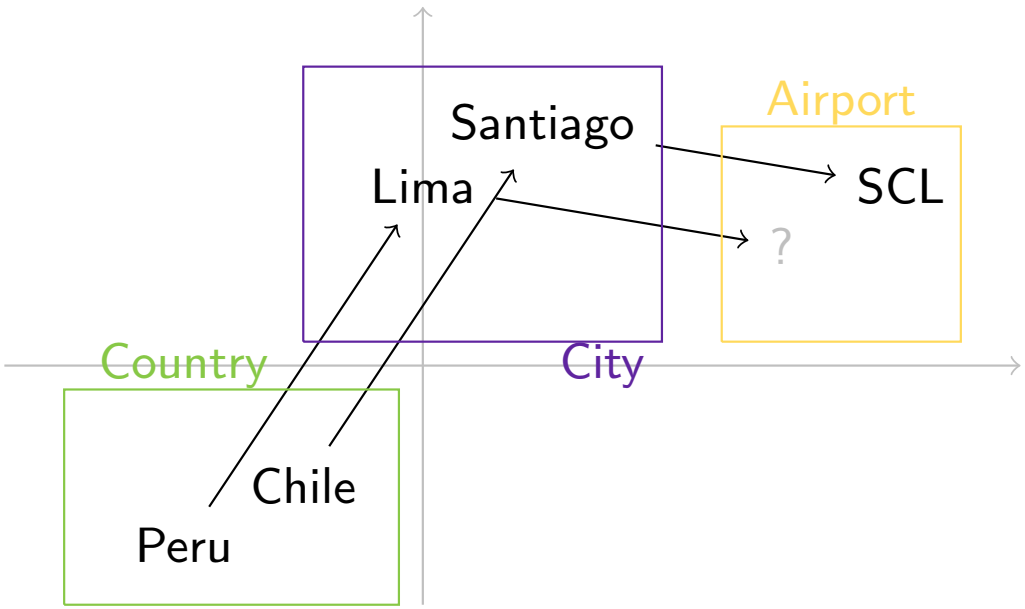
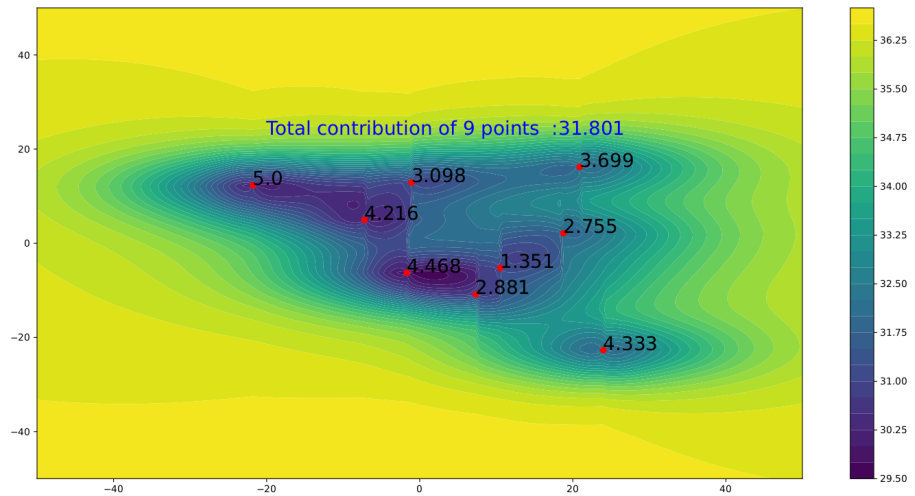


Figure: Geometric interpretation of KG embeddings

Example: wind farm power production

An example of **Kriging**, a kernel-based method.

- **observation**, x: a new set of wind turbine positions.
- ϕ : almost always **implicit** thanks to the kernel trick, $\langle \phi(x), \phi(x') \rangle_{\mathcal{H}} = k(x, x'; \theta_\phi)$ and always rely on $k()$
- ψ: Bayesian linear regression
- **decision**, y: predict the average power production from new wind turbine positions.

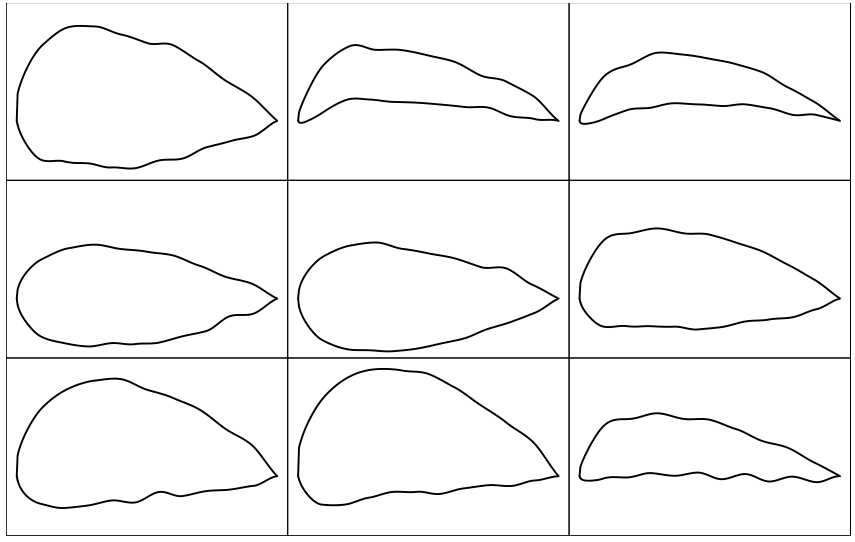


Complete study in [Sow et al., 2023]

Example : eigenshape decomposition I

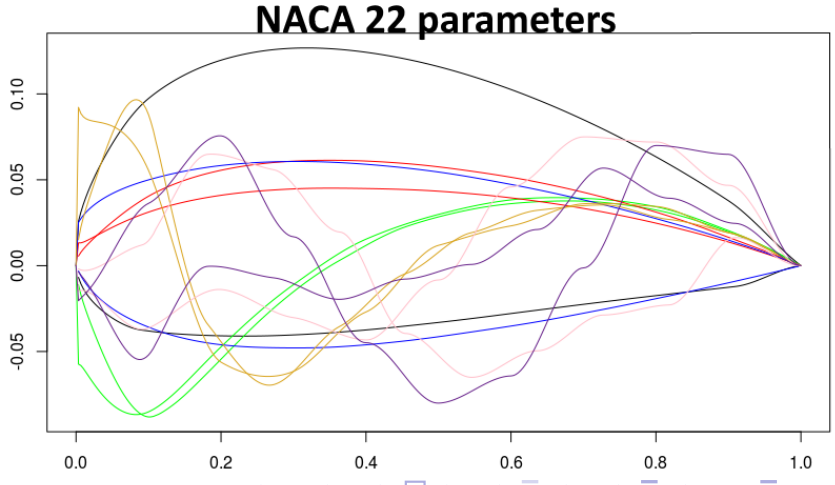
An example of Principal Components Analysis (**PCA**) use.

From a database of possible shapes,



. . .

extract a basis of most important shapes by principal component analysis, $\{V^1, \ldots, V^{\dim(h)}\}$



Example : eigenshape decomposition II

Shapes are now described with their eigencomponents h,

$$\text{shape} \approx \sum_{i=1}^{\dim(h)} h_i V^i \quad , \quad h_i = x^\top V^i$$

Then work in latent h-space, cf. [Gaudrie et al., 2020].

- **observation**: a shape x
- $h = \phi(x; \theta_\phi)$: projections of x on the basis, $\theta_\phi = \{V^1, \ldots, V^{\dim(h)}\}$
- ψ: regression in h-space
- **decision**, y: prediction of lift and drag (then optimization)

Example : latent variables in materials science I

Example of Expectation-Maximization, **EM**.

Latent random variables to describe sample variability (e.g., [Laboulfie et al., 2021]).

$$h^1 \sim p_H(h \mid \theta_\phi) \rightsquigarrow \text{sample } X^1 \qquad h^n \sim p_H(h \mid \theta_\phi) \rightsquigarrow \text{sample } X^n$$

$$h^i \neq h^j$$

In learning, average out latent variables to calculate likelihood:

$$L(\theta_\phi; X) = p(X \mid \theta_\phi) = \prod_{i=1}^{n} \int p(X^i \mid h, \theta_\phi) p_H(h \mid \theta_\phi) dh$$

Example : latent variables in materials science II

- **observations**, x: stress-strain measures on specimen
- ϕ: the probability distribution of the material parameters $h \sim p_H(. \mid \theta_\phi)$
- ψ: material probabilistic model, $p(x \mid h)$
- **decision**: likelihood of the measures

Why latent representations

- A transformation to help learning?
- Or a causality (explanation) hidden in the observations ? I.e., try to understand regularities of the data.

Comparison criteria

Various comparison criteria for latent representations:

- dimension of latent vectors
- complexity of ϕ (number of parameters)
- complexity of ψ (number of parameters)

Do not mistake the above mapping complexity for the computational complexity. Some mappings (e.g., EM) have no parameter but imply heavy calculations.

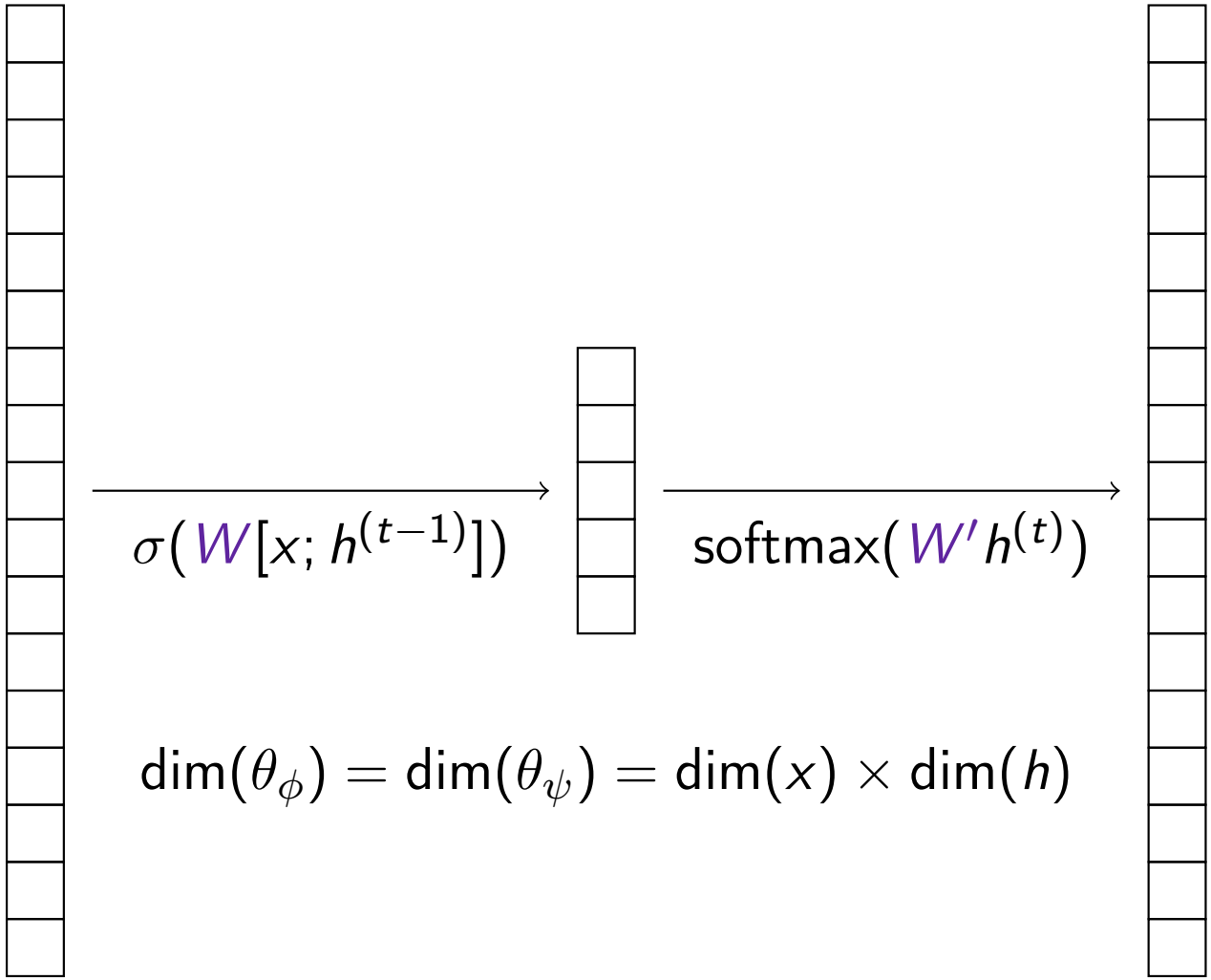
Latent representation families

In the following, we use these criteria to characterize various families of latent representations **after learning**, in the prediction phase.

Families of latent representations:

- Knowledge Graph embeddings (TransE, RESCAL)
- Word embeddings (word2vec, GloVe)
- Transformers (BERT)
- Principal Component Analysis (PCA)
- Kernel-based algorithms (Kriging, SVM)
- Expectation-maximization algorithms

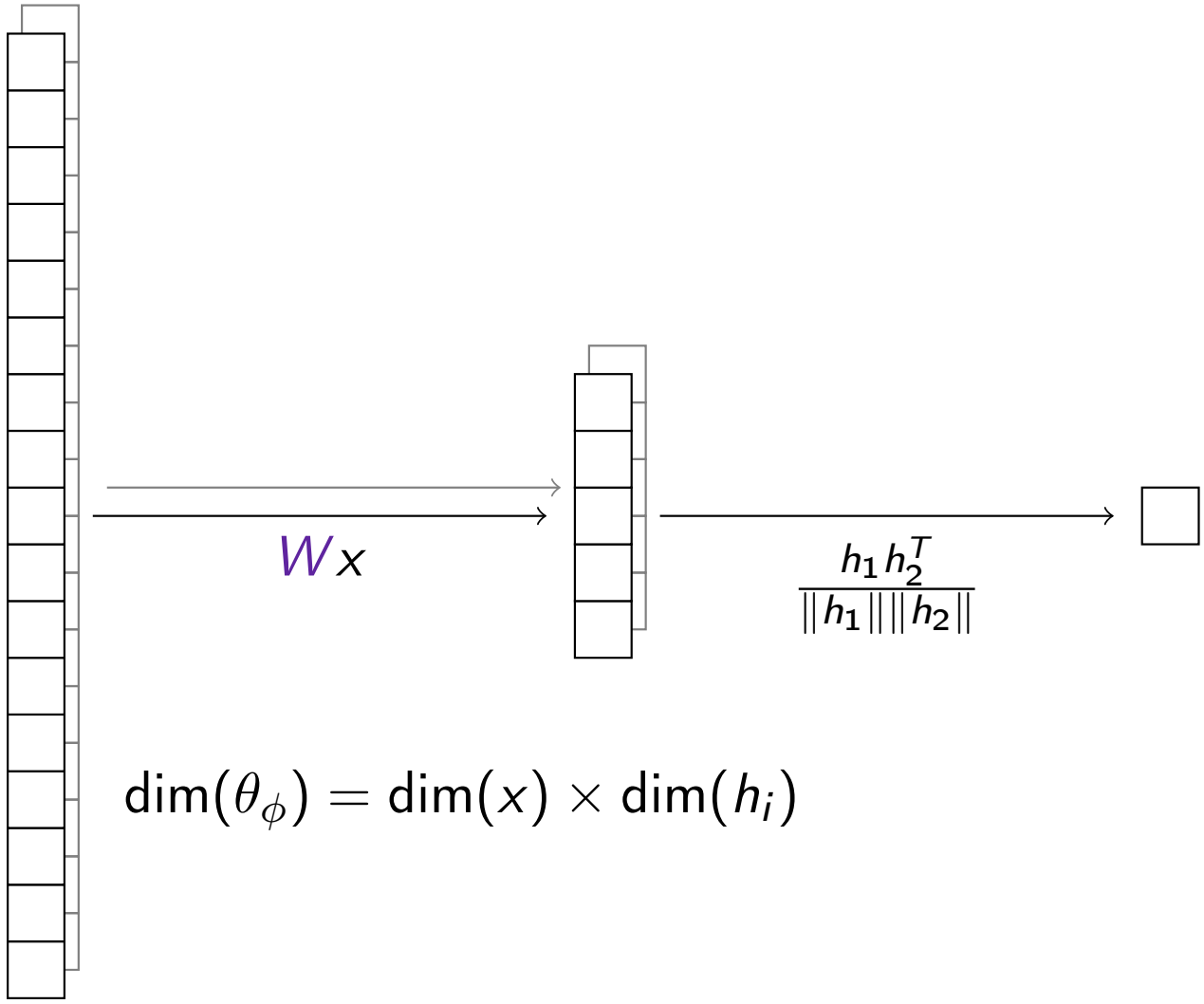
word2vec



$$\dim(\theta_\phi) = \dim(\theta_\psi) = \dim(x) \times \dim(h)$$

$\dim(x) \approx 30k$ $\dim(h) = 90$ $\dim(y) = \dim(x)$

GloVe



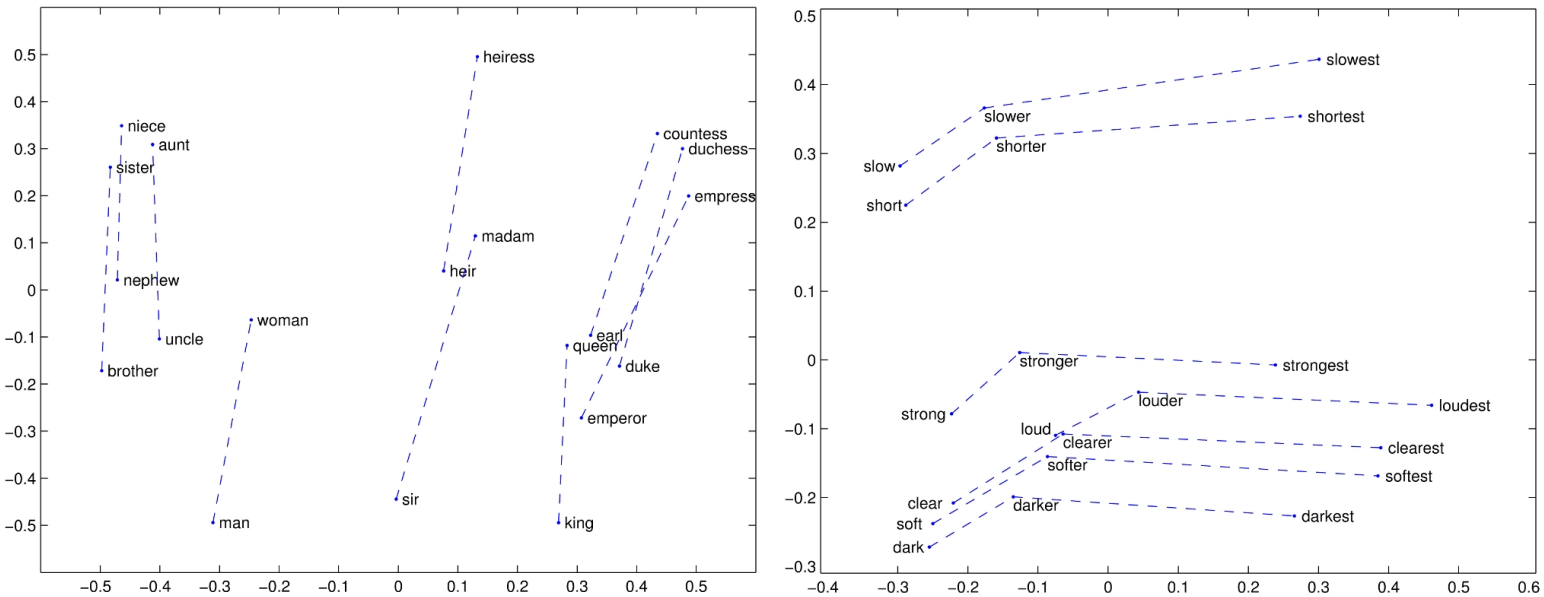
$$\dim(\theta_\phi) = \dim(x) \times \dim(h_i)$$

$\dim(x) \approx 400k$ $\dim(h_i) = 300$ $\dim(y) = 1$

GloVe (discussion) I

GloVe was designed to capture similarity between words.

Semantic relations can **explain**, a posteriori, certain regularities in the latent space.



(a) gender equivalence

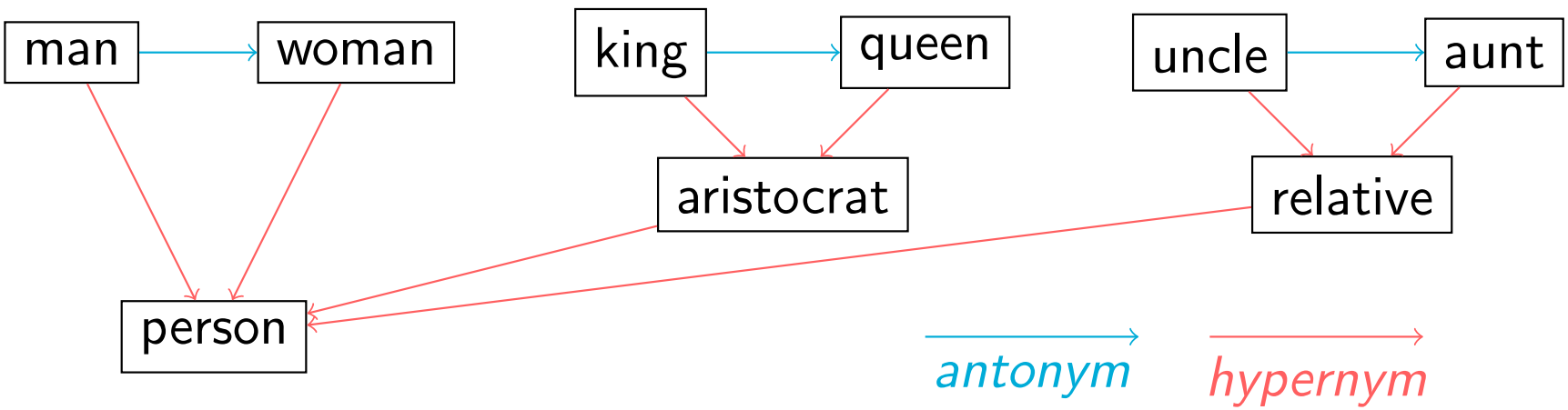
(b) comparative-superlative

Figure: Linear substructures of GloVe embeddings [Pennington et al., 2014]

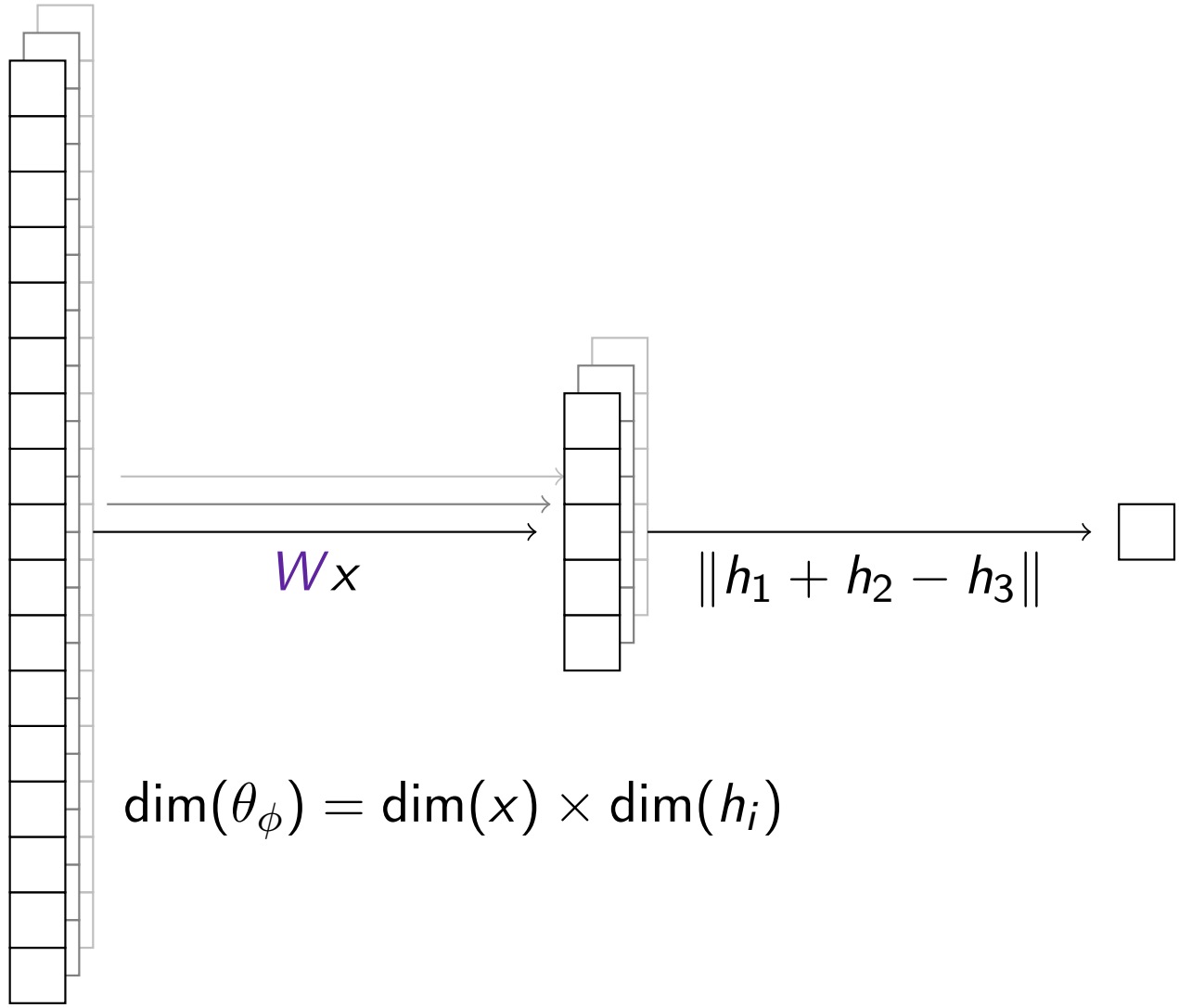
GloVe (discussion) II

But can they explain all latent space regularities?

The inverse approach consists in **embedding lexical databases** such as WordNet.



TransE



$$\dim(\theta_\phi) = \dim(x) \times \dim(h_i)$$

$\dim(x) \approx 15k$ $\dim(h_i) = 50$ $\dim(y) = 1$

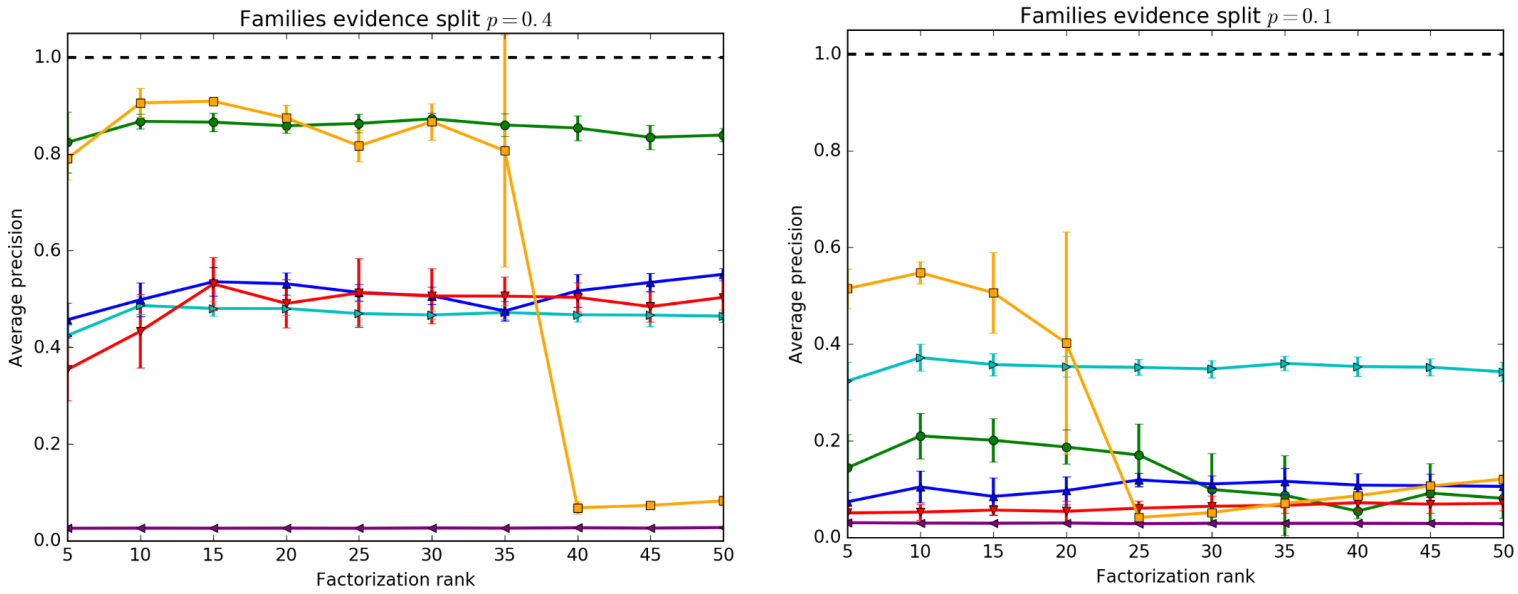
TransE (discussion) I

WordNet embeddings have **decent** true/false classification **performances**. But dim(h) varies from 50 to 500 across experiments.

Does the English language require **500 semantic relations** to discriminate all pairs of words? Probably not.

TransE (discussion) II

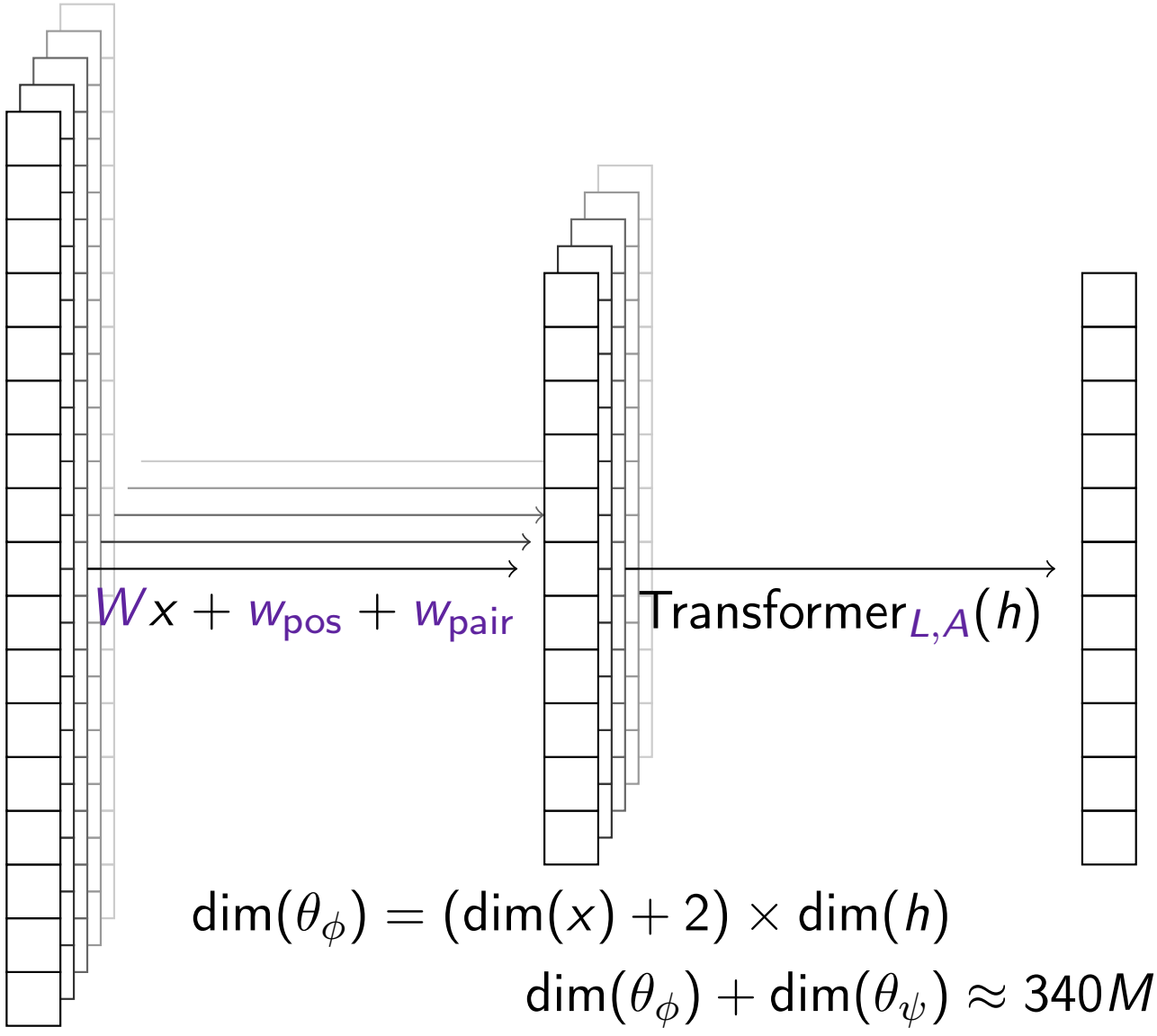
Knowledge Graph embeddings have an optimal (often unknown) dimension.



(a) 40% of data observed (b) 10% of data observed

Figure: Average true/false classification performances on a synthetic dataset (yellow: RESCAL, cyan: TransE) [Trouillon et al., 2019]; see also [Charpenay, 2023]

BERT



$$\dim(\theta_\phi) = (\dim(x) + 2) \times \dim(h)$$

$$\dim(\theta_\phi) + \dim(\theta_\psi) \approx 340M$$

$\dim(x) \approx 30k$ $\dim(h) = 1024$ $\dim(y) = \dim(h)$

BERT (discussion) I

Which part of BERT does capture latent features?

- The embedding layer?
- Any of the hidden encoder layers?
- The entire encoder?

All layers, partly.

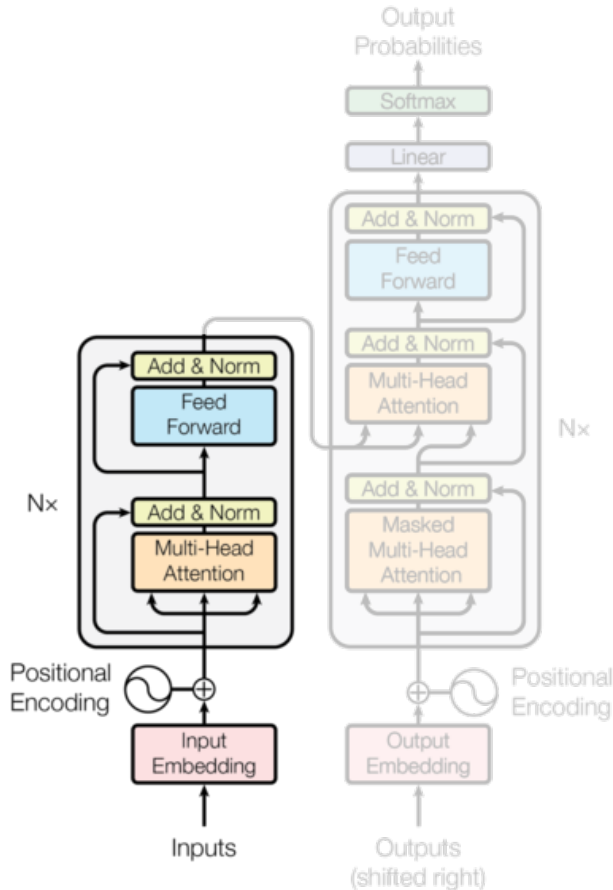


Figure: Transformer encoder [Vaswani et al., 2017]

BERT (discussion) II

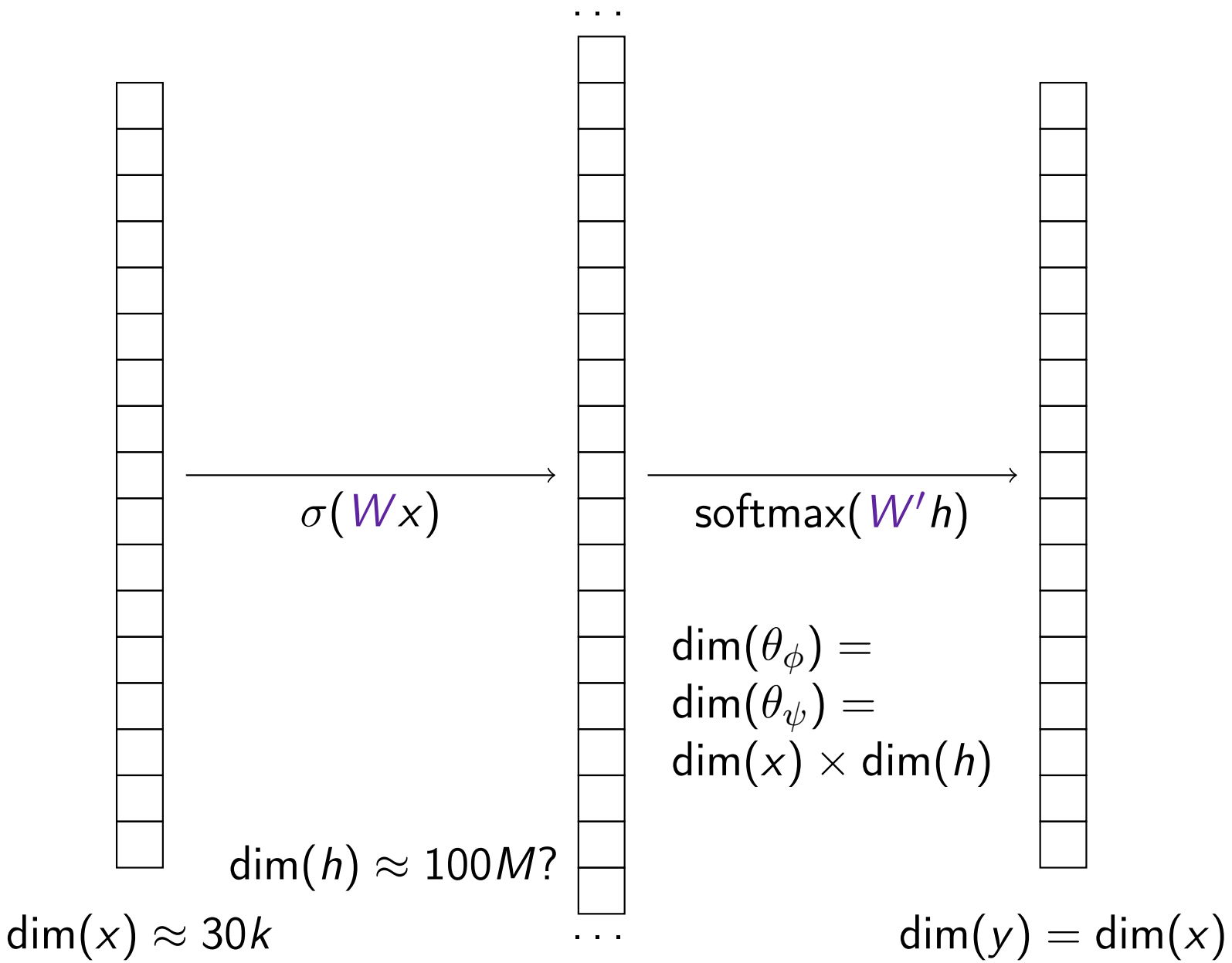
	F1 score
Fine-tuning	
BERT	96.4
Feature-based learning	
Embeddings	91.0
Last hidden	94.9
Weighted sum all 12 hidden	95.5
Second-to-last hidden	95.6
Weighted sum last four hidden	95.9
Concat last four hidden	96.1

Table: Scores on a Named Entity Recognition task with fine-tuning and feature-based learning from pre-trained BERT

BERT (discussion) III

BERT is equivalent to a (theoretical) neural network with a single hidden layer.

BERT-equivalent single-layer network



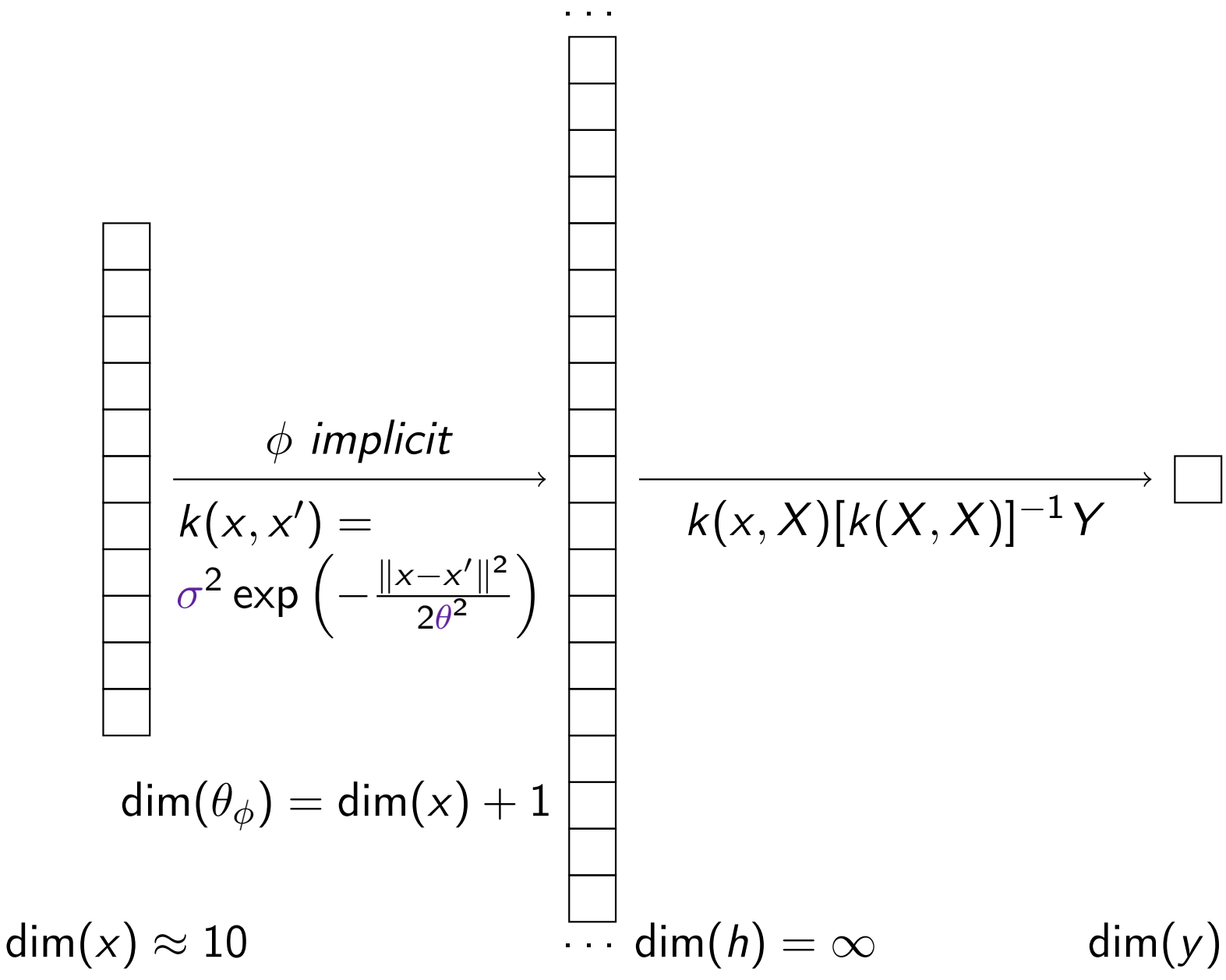
BERT-equivalent single-layer network (discussion)

There is some interdependence between $\dim(h)$ and $\text{card}(\theta_\phi) + \text{card}(\theta_\psi)$.

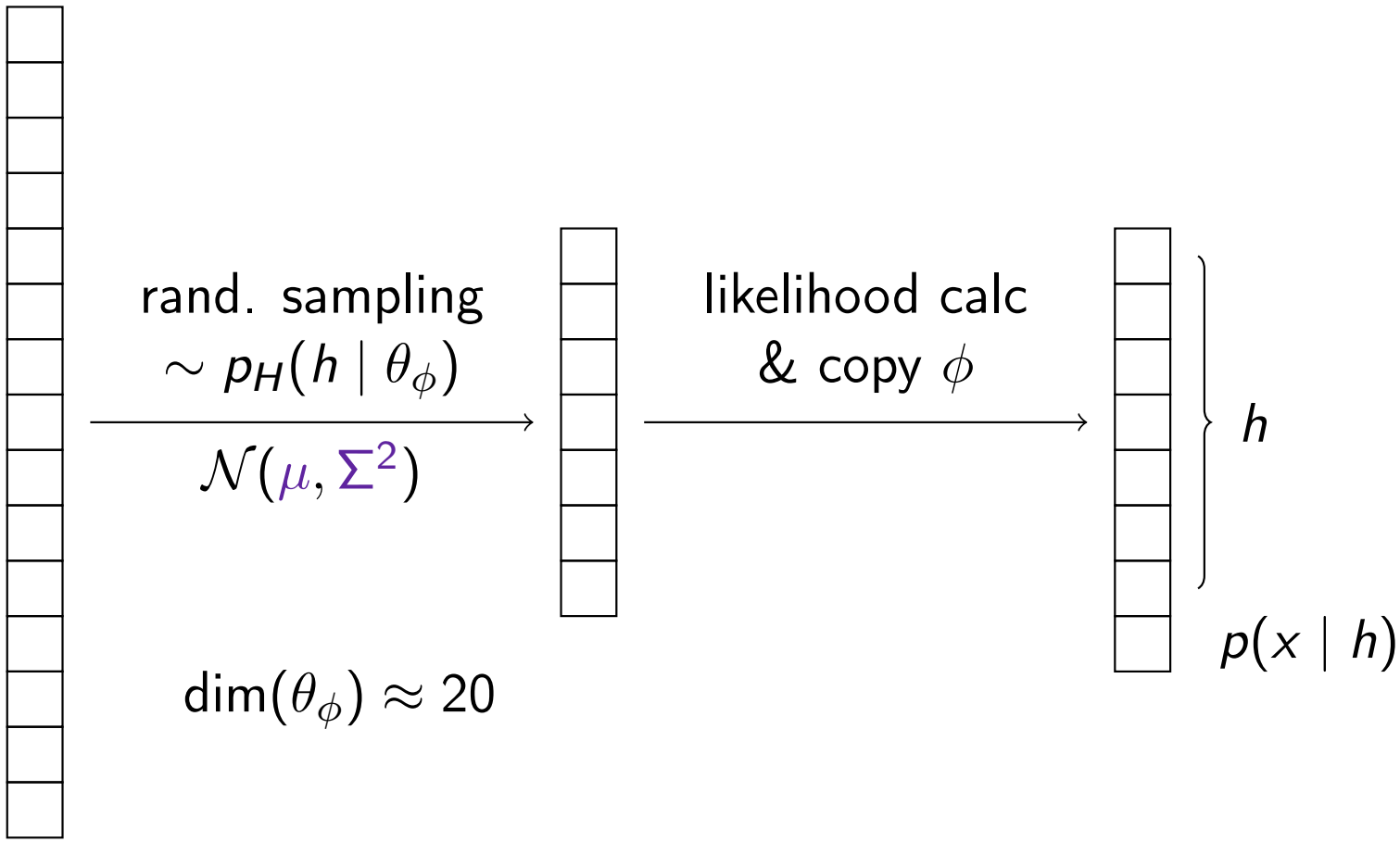
Why are Transformers used in practice instead of single-layer networks? For efficient computation.

In contrast, kernel-based methods tend to **increase** the latent dimension, to ease the calculation of ψ.

Kriging & SVM

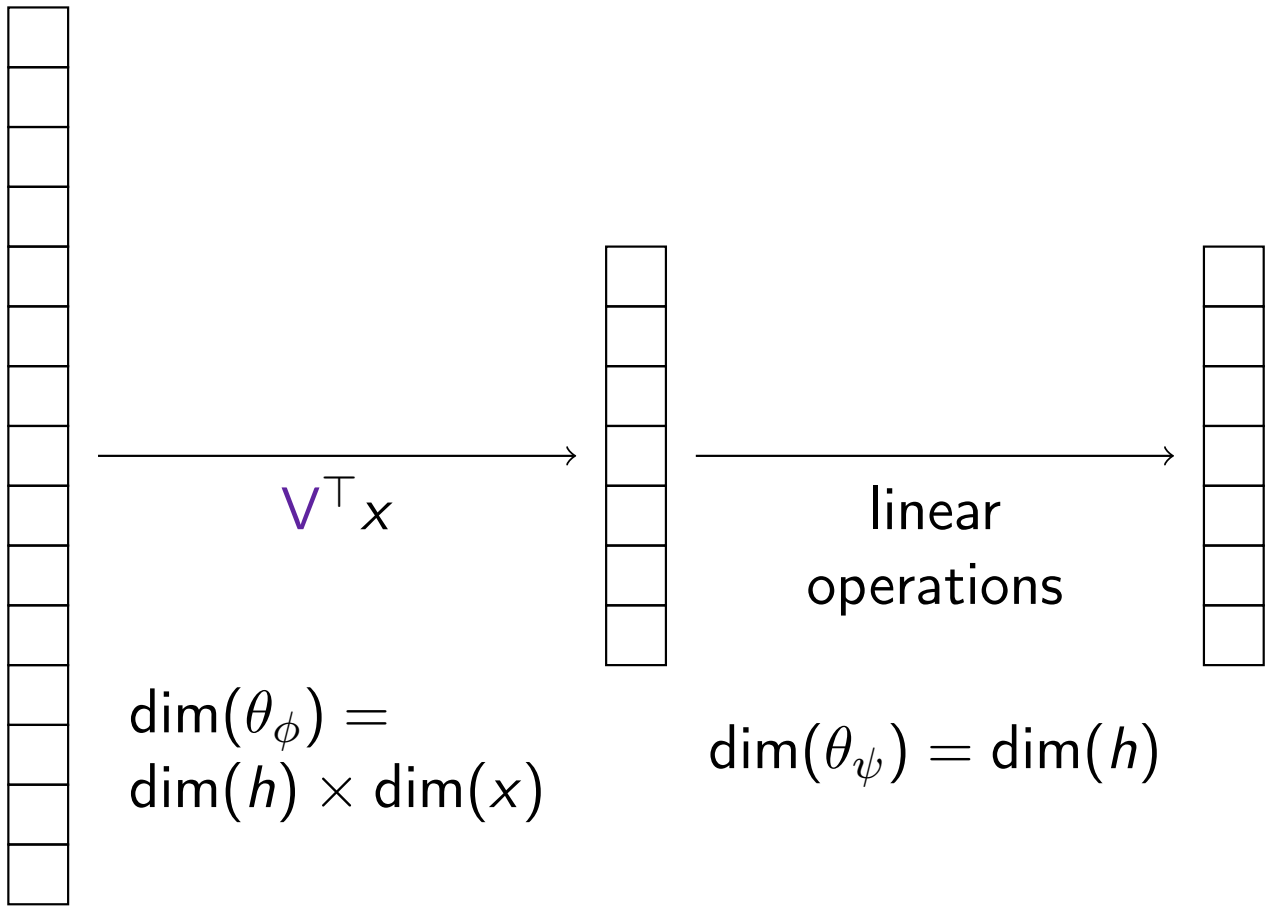

ϕ *implicit*
$k(x, x') =$
$\sigma^2 \exp\left(-\frac{\|x-x'\|^2}{2\theta^2}\right)$
$k(x, X)[k(X, X)]^{-1} Y$
$\dim(\theta_\phi) = \dim(x) + 1$
$\dim(x) \approx 10$
$\dim(h) = \infty$
$\dim(y) = 1$

Expectation-Maximization



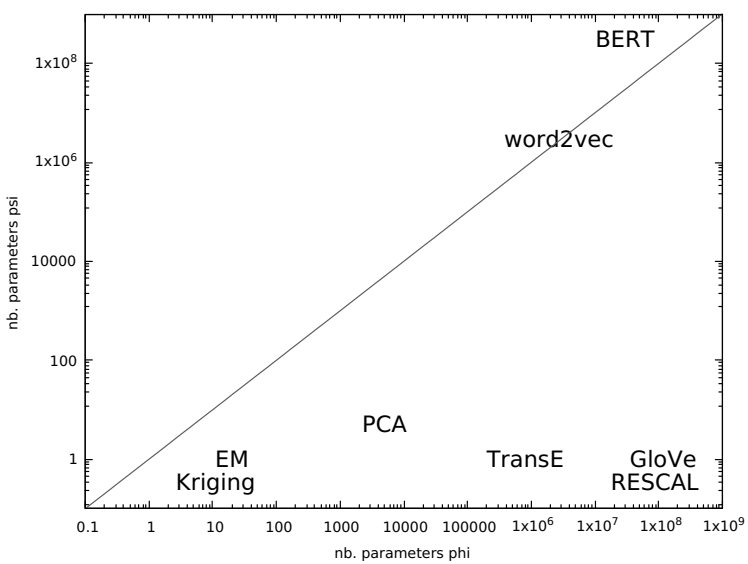
$\dim(x) \approx 500$ $\dim(h) \approx 5$ $\dim(y) = 1 + \dim(h)$

Principal Component Analysis

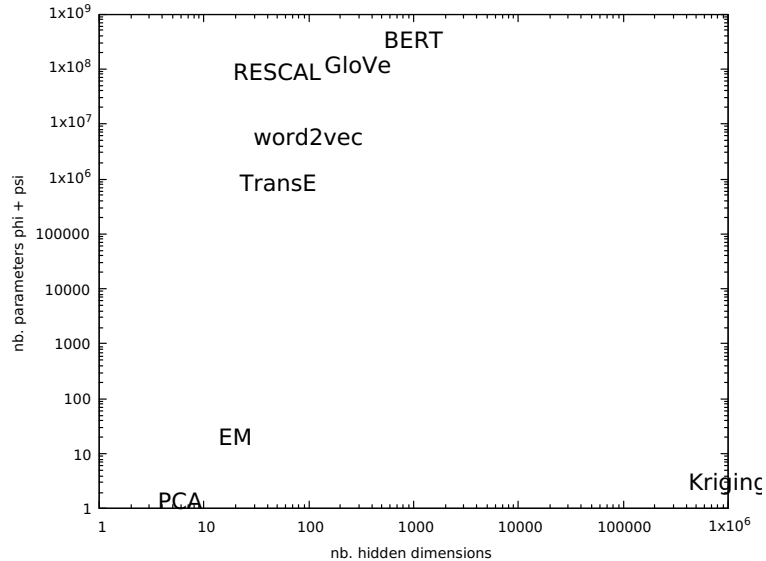


$\dim(x) \approx 1000$ $\dim(h) \approx 5$ $\dim(y) = 1$ to $\dim(h)$

Comparison I



(a) $\dim(\theta_\phi)$ vs. $\dim(\theta_\psi)$



(b) $\dim(h)$ vs. $\dim(\theta_\phi) + \dim(\theta_\psi)$

Figure: Comparison of latent representations on size and complexity criteria

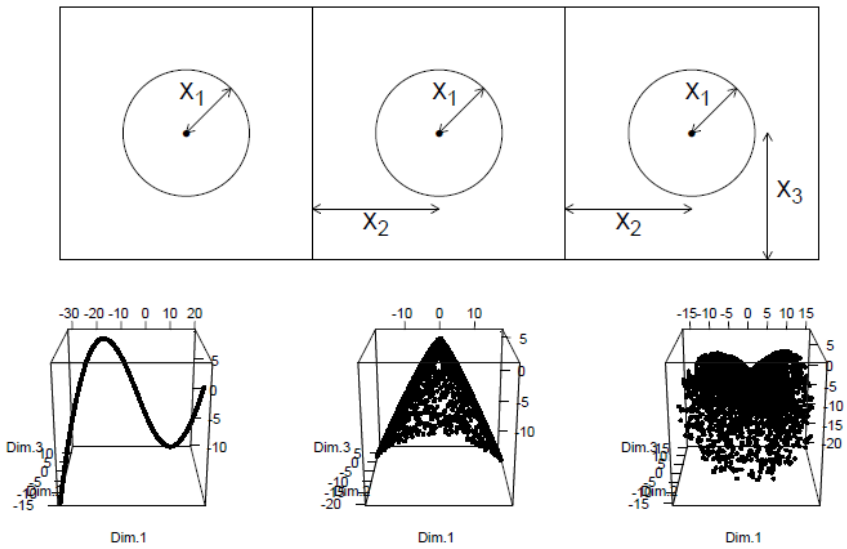
Comparison II

- Current algorithms rely mainly on latent variables as $\dim(\theta_\psi) \leq \dim(\theta_\phi)$
- The explainability of the methods depends on both the dimension of the latent space and the total number of parameters

Latent spaces as manifold, intrinsic dimension

- The latent space is a manifold.
- The smallest number of dimensions among useful latent spaces is the **intrinsic dimension** [Camastra and Staiano, 2016] of the problem

Ex: Parameterized shape families (top row) and associated (h_1, h_2, h_3)



(illustration from extended version of [Gaudrie et al., 2020] on arXiv)

Latent spaces are not unique

- If learning is repeated, the same map from x to y typically has different (ϕ, ψ) pairs.
- Mathematically, ϕ is not unique because it depends on ψ
- Example :

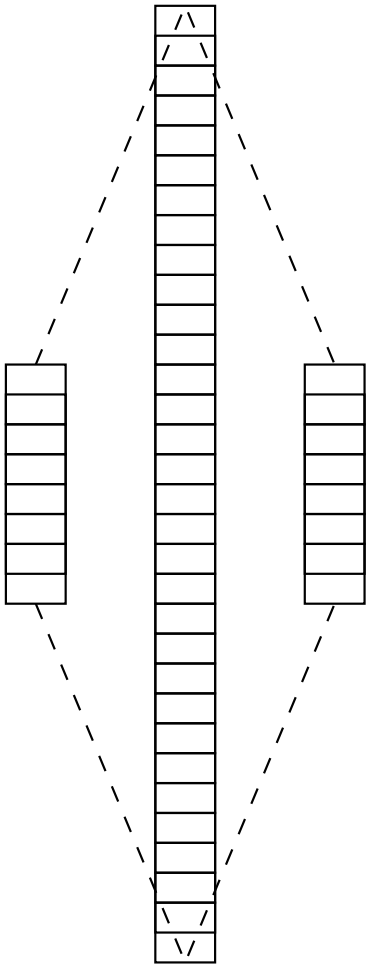
  $$\psi(\tfrac{1}{3} \times 3 \times \phi(x)) = \psi'(\phi'(x)) = \psi(\phi(x))$$
  $$\text{where } \psi'(\square) = \psi(\tfrac{1}{3}\square) \quad , \quad \phi'(\square) = 3 \times \phi(\square)$$

- More generally, for any bijection g from the h-space to itself, when $\phi'(\square) = g(\phi(\square))$ and $\psi'(\square) = \psi(g^{-1}(\square))$, $\psi(\phi(\square)) = \psi'(\phi'(\square))$.
- Account for this when comparing latent spaces.

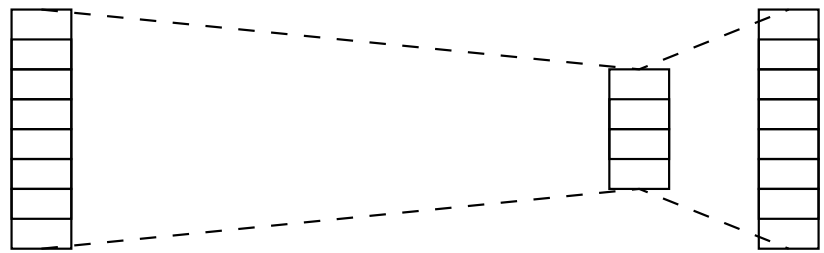
Conclusions I

- Latent variables are ubiquitous in data science, making them a topic for interdisciplinary research.
- 2 goals were identified :
  - find regularities in data, explain data $\Rightarrow$ **reduce** latent dimension to tend towards the (low) intrinsic dimension of the problem.
  - ease computation $\Rightarrow$ **increase** latent dimension to allow linear classification or regression.

Conclusions II



(a) high dim(h), low dim(θ)



(b) low dim(h), high dim(θ)

Figure: Schematic view on the dimension(s) of latent representations

Conclusions III

- The **increase in complexity** of $\psi \circ \phi$ allowed by progress in algorithms (regularization) and hardware is compensated for by the **need for explainability** that calls for low intrinsic dimensions.
- The link between the **data set size** and the **latent dimension** is an open question of practical importance.

Bibliography I


Aggarwal, C. C. (2016).
Outlier Analysis.
Springer Cham.


Camastra, F. and Staiano, A. (2016).
Intrinsic dimension estimation: Advances and open problems.
Information Sciences, 328:26–41.


Charpenay, V. (2023).
On the dimensionality of knowledge graph embeddings.
submitted to IJCKG.


Gaudrie, D., Le Riche, R., Picheny, V., Enaux, B., and Herbert, V. (2020).
Modeling and optimization with gaussian processes in reduced eigenbases.
Structural and Multidisciplinary Optimization, 61:2343–2361.


Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutiérrez, C., Kirrane, S., Labra Gayo, J. E., Navigli, R., Neumaier, S., Ngonga Ngomo, A.-C., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J. F., Staab, S., and Zimmermann, A. (2021).
Knowledge Graphs.
Number 22 in Synthesis Lectures on Data, Semantics, and Knowledge. Springer.

Bibliography II


Laboulfie, C., Balesdent, M., Brevault, L., Da Veiga, S., Irisarri, F.-X., Le Riche, R., and Maire, J.-F. (2021).
Calibration of material model using mixed-effects models.
In *4th International Conference on Uncertainty Quantification in Computational Sciences and Engineering (UNCECOMP 2021)*.


Pennington, J., Socher, R., and Manning, C. D. (2014).
GloVe: Global vectors for word representation.
available on nlp.stanford.edu.


Sow, B., Le Riche, R., Pelamatti, J., Zannane, S., and Keller, M. (2023).
Learning functions defined over sets of vectors with kernel methods.
In *5 th ECCOMAS Thematic Conference on Uncertainty Quantification in Computational Sciences and Engineering (UNCECOMP 2023)*.


Trouillon, T., Gaussier, E., Dance, C. R., and Bouchard, G. (2019).
On inductive abilities of latent factor models for relational learning.
JAIR.

Bibliography III

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017).
Attention is all you need.
In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.