Inferring Missing Data with Auto-Associators

Mark J. Embrechts (Mark.Embrechts@gmail.com)

HMGU Munich

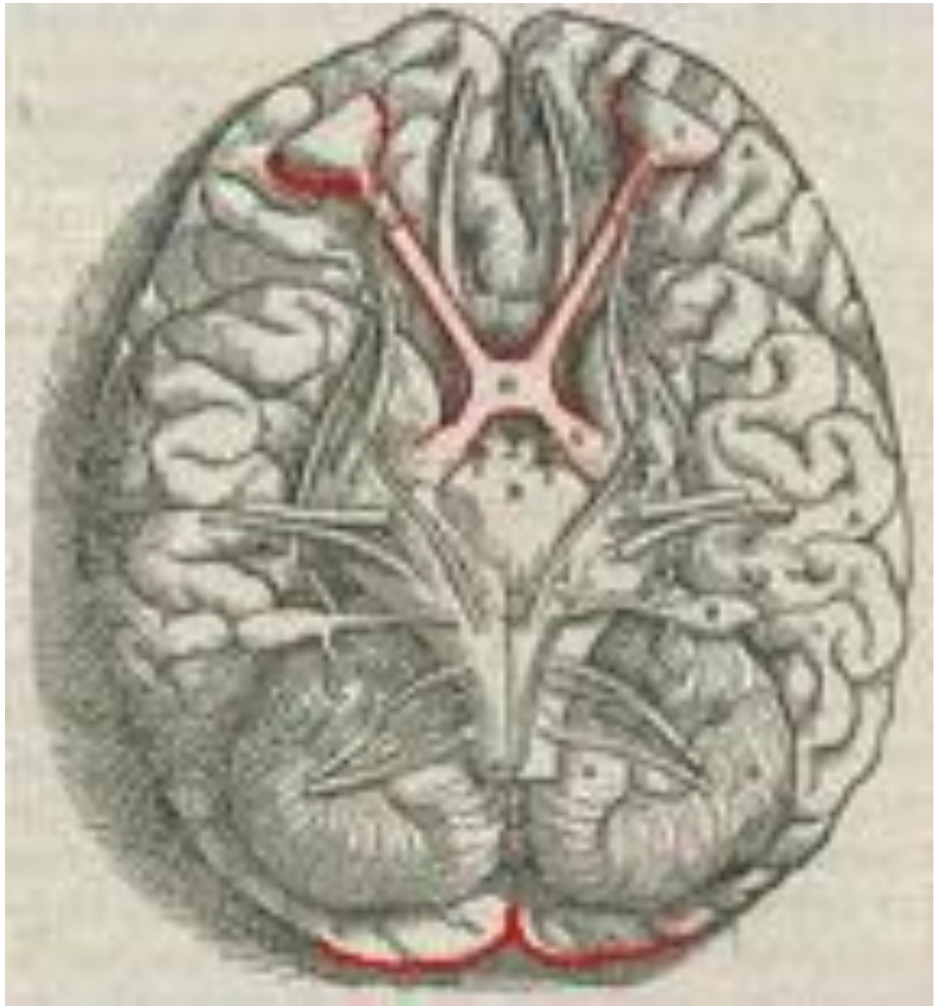
AIDD School
Artificial Intelligence for Drug Discovery
Irish College, Leuven, October 25, 2022

HELMHOLTZ MUNICH

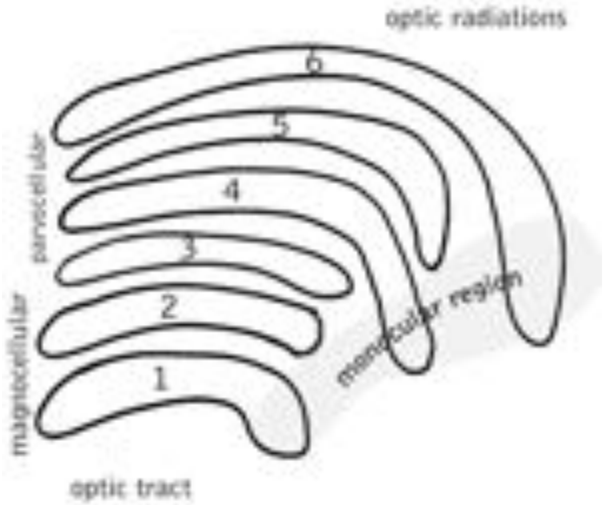
Andreas Vesalius

Born	31 December 1514 Brussels, Habsburg Netherlands
Died	15 October 1564 (aged 49) Zakynthos, Republic of Venice
Fields	Anatomy
Doctoral advisor	Johannes Winter von Andernach Gemma Frisius
Doctoral students	Matteo Realdo Colombo
Known for	*De humani corporis fabrica* or "the fabric of the human body"
Influences	Jacques Dubois Jean Fernel

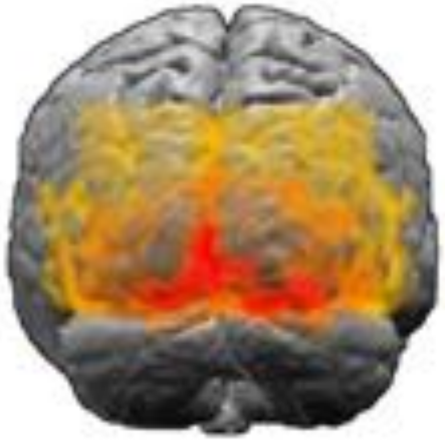
Vesalius, 1543



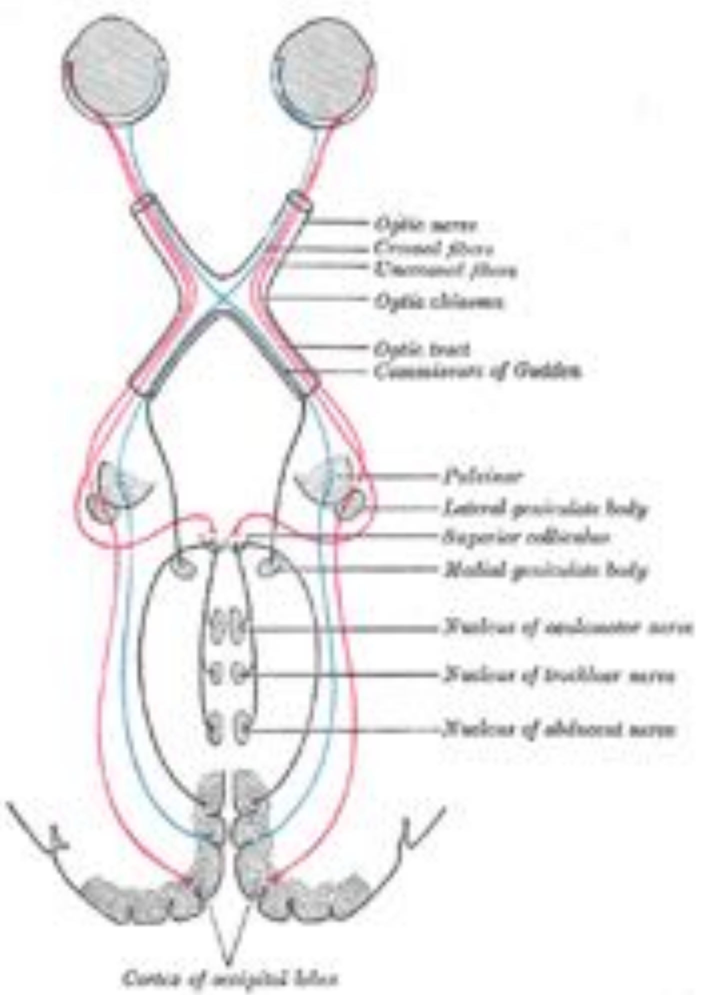
Vesalius, Anatomy, 1543



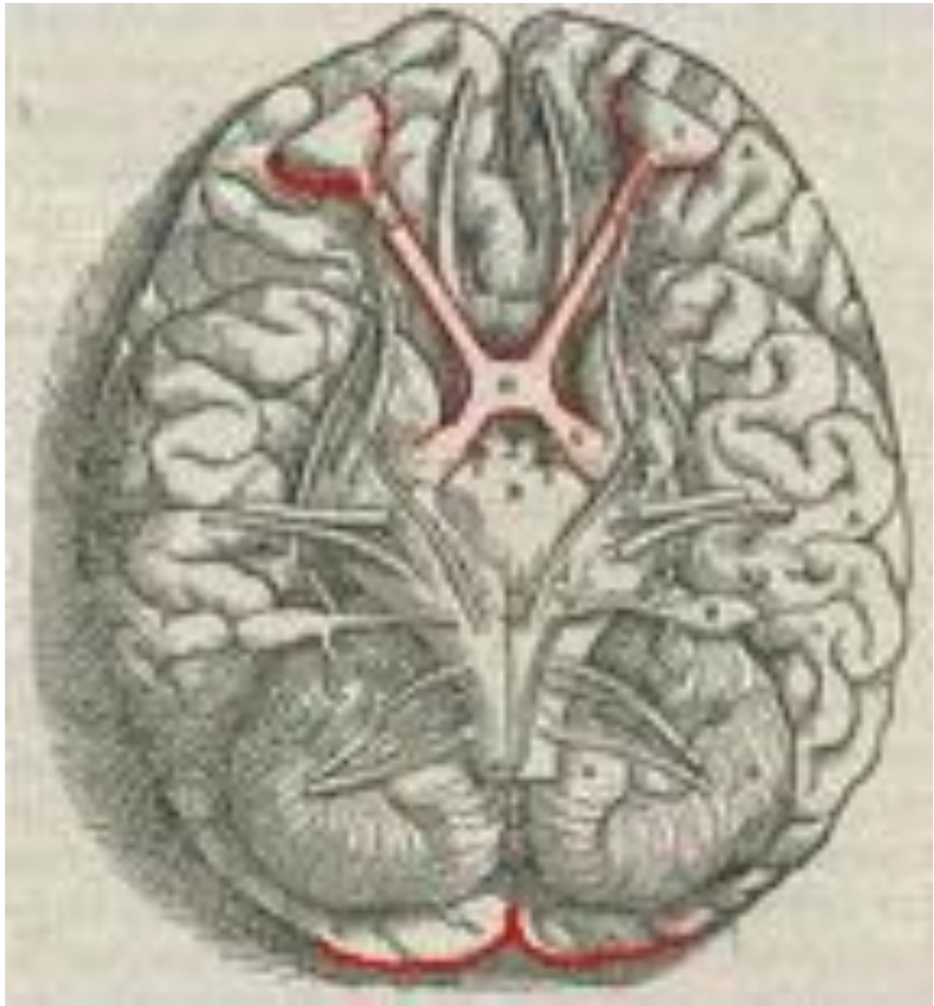
Lateral geniculate nucleus (in thalamus)



Visual cortex



© Gray: http://en.wikipedia.org/wiki/Visual_system



Vesalius, Anatomy, 1543

Overview

- Dealing with missing data
- Auto-associators
- Some results

5	3			7				
6			1	9	5			
	9	8					6	
8				6				3
4			8		3			1
7				2				6
	6					2	8	
			4	1	9			5
				8			7	9

5	3	-999		7				
6	-999	-999	1	9	5			
-999	9	8					6	
8				6				3
4			8		3			1
7				2				6
	6					2	8	
			4	1	9			5
				8			7	9

Missing Data

- What are missing data?
- Simple ways to deal with missing data
  - Eliminate descriptor (columns)
  - Eliminate records (rows
  - Replace by mean/median values
- More sophisticated ways to deal with missing data
  - Auto-associators, ...
- Remarks
  - There might be information in missing data
  - Categorical data need a more sophisticated approach

HELMHOLTZ MUNICH

Some tricky issues related to missing data

- Missing data can by systemic:
  e.g., in a survey a person might not want to mention alcohol use.
- It might be helpful to add a column indicating whether a descriptor is missing
- Data curation: suspect data might be flagged as missing (-999)
- Example of systemic issues with missing data: 9 class Italian Olive Oil Data

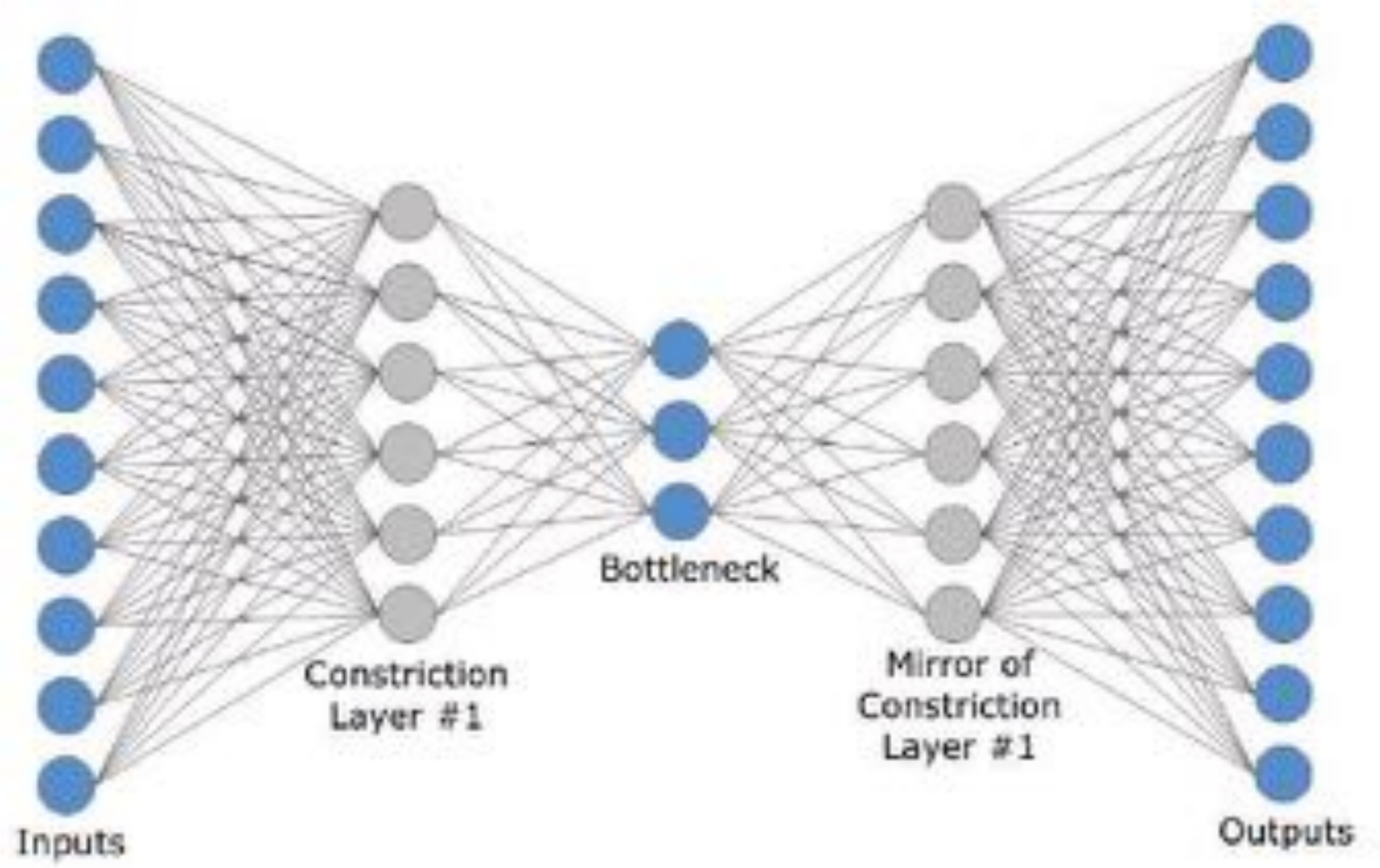
32.9	14.3	45.7	67.3	35.9	58.3	65	80.4	4	318
31.7	15.1	54.3	66	37.7	59.7	59.2	80.4	4	319
46	21.9	90.6	58.2	27.4	58.3	71.8	64.3	4	320
51.4	26.4	54.3	54.3	35.8	58.3	68.9	71.4	4	321
57.5	30.2	42.6	52.4	35.6	56.9	66	58.9	4	322
58	27.5	66.8	55.5	25	56.9	67	51.8	4	323
46.1	39.6	31.4	46.1	65	56.9	93.2	**0**	5	324
38.4	45.3	26	51	65.4	45.8	85.4	**0**	5	325
43.8	30.6	26	51.2	62.3	41.7	89.3	**0**	5	326
45.2	30.9	30.9	46.4	69.1	45.8	89.3	**0**	5	327

class

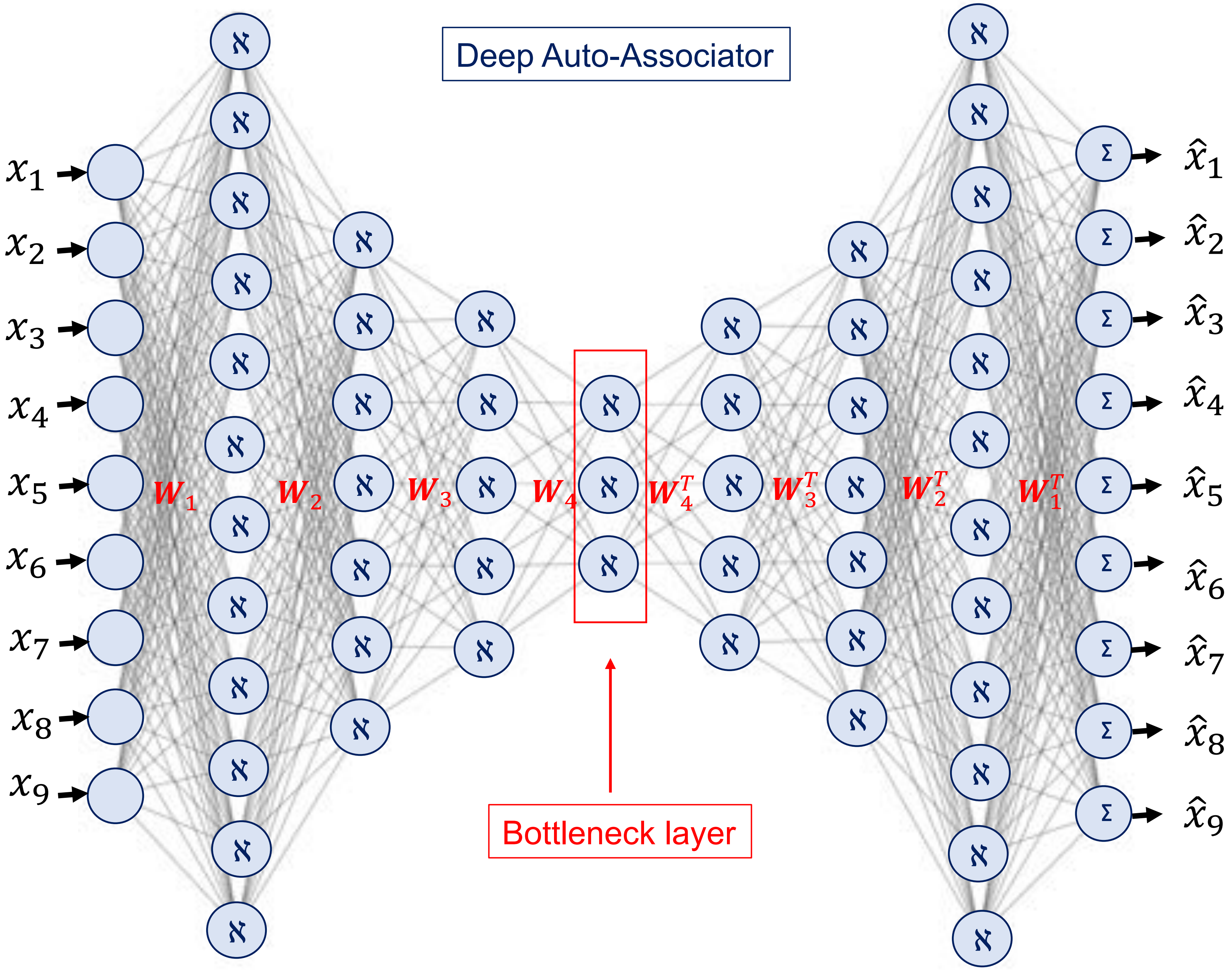
- Descriptors for certain classes are missing, but set to zero in the original data set
- We will replace the zero settings by -999, indicating they are really missing

HELMHOLTZ MUNICH

Auto-Encoders or Auto-Associative Networks

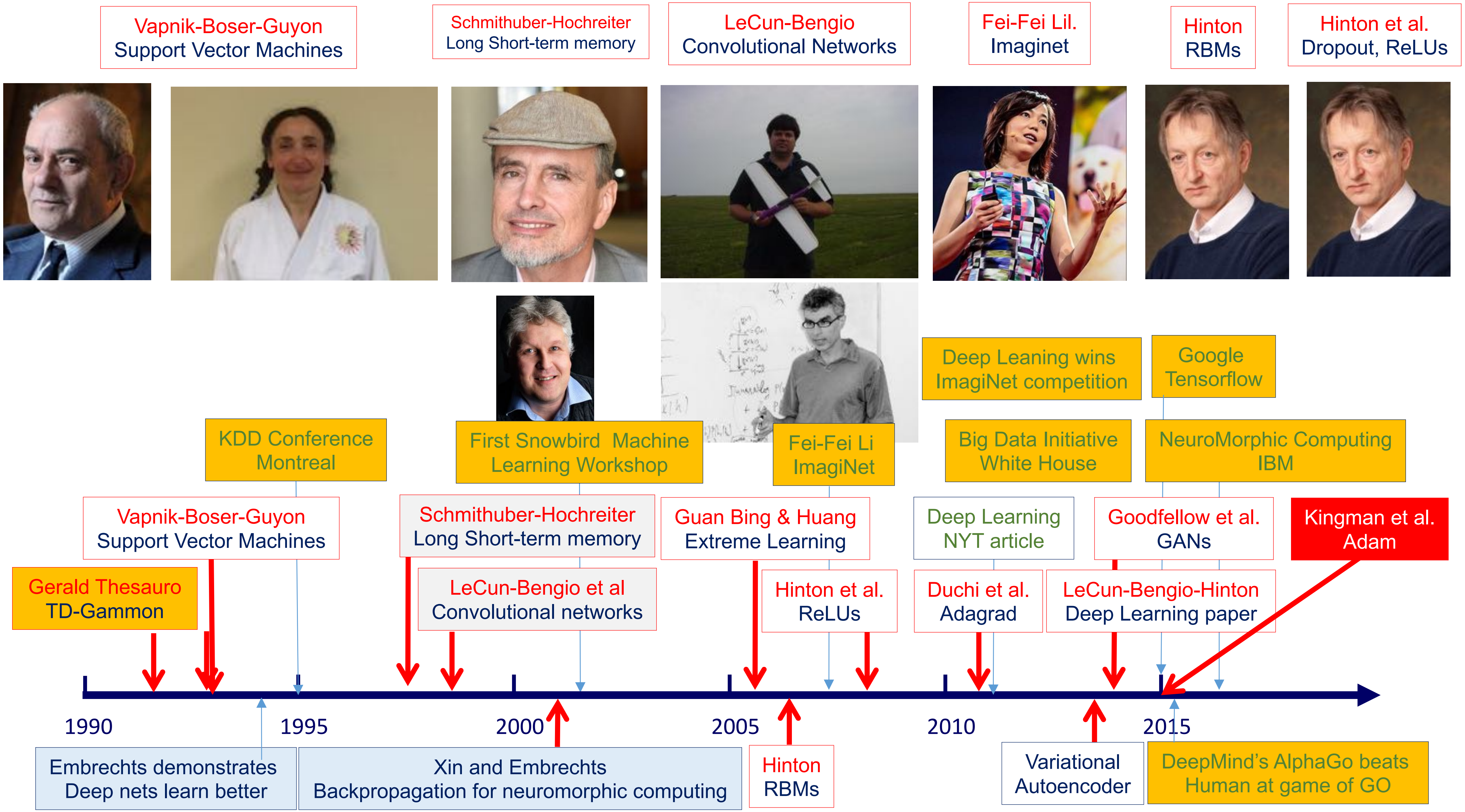


HELMHOLTZ
MUNICH



HELMHOLTZ MUNICH

More Recent History of Neural Networks

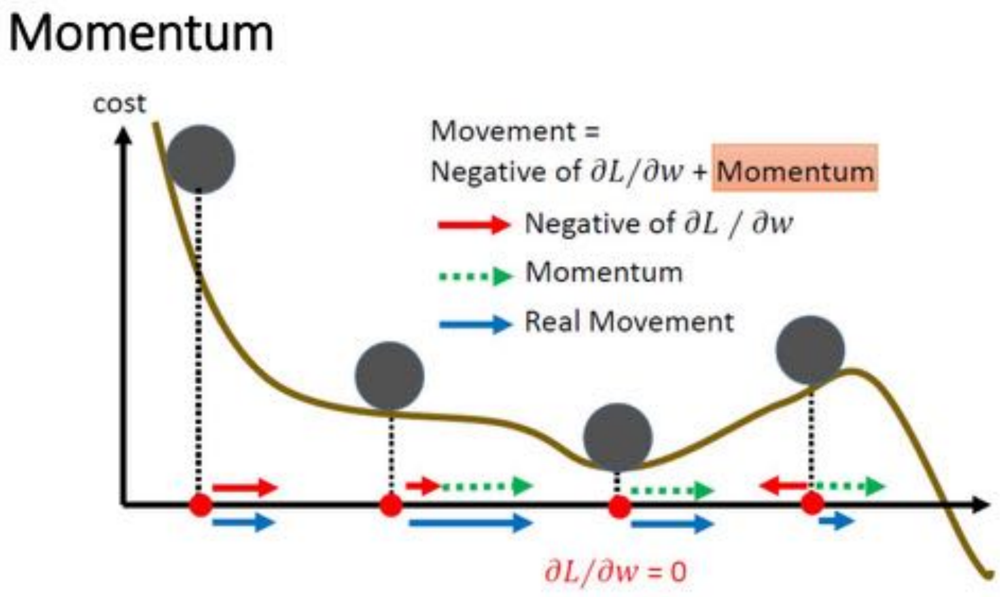


2015 Adam: Adaptive moment estimation

<u>PARAMETER SETTINGS</u>:
η -- stepsize (i.e., learning parameter η 0.0001)
β_1 -- exponential decay rate gradient (0.9)
β_2 -- exponential decay rate 2nd moment (0.999)

<u>INITIALIZATION</u>:

$$\boldsymbol{m}_0 \leftarrow 0 \ (gradient\ tensor)$$
$$\boldsymbol{v}_0 \leftarrow 0 \ (2nd\ moment\ tensor)$$
$$\boldsymbol{w}_0 \leftarrow 0 \ (weight\ tensor)$$



UPDATING RULE:

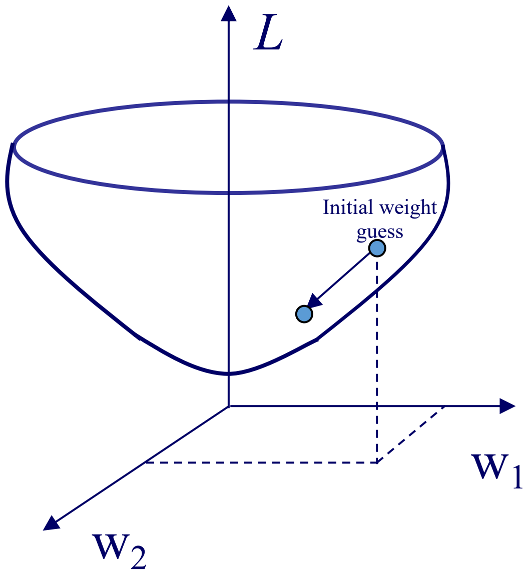
$$\boldsymbol{g}_t \leftarrow \nabla L_t(\boldsymbol{w}_t)$$
$$\boldsymbol{m}_t \leftarrow \boldsymbol{\beta}_1 \boldsymbol{m}_{t-1} + (1 - \boldsymbol{\beta}_1)\boldsymbol{g}_t$$
$$\boldsymbol{v}_t \leftarrow \boldsymbol{\beta}_2 \boldsymbol{v}_{t-1} + (1 - \boldsymbol{\beta}_2)\boldsymbol{g}_t^2$$
$$\widehat{\boldsymbol{m}}_t \leftarrow \boldsymbol{m}_t / (1 - \boldsymbol{\beta}_1^t)$$
$$\widehat{\boldsymbol{v}}_t \leftarrow \boldsymbol{v}_t / (1 - \boldsymbol{\beta}_2^t)$$
$$\boldsymbol{w}_t \leftarrow \boldsymbol{w}_{t-1} - \boldsymbol{\eta}\, \widehat{\boldsymbol{m}}_t / \sqrt{\widehat{\boldsymbol{v}}_t + \boldsymbol{\varepsilon}}$$

$$\Delta_t = \boldsymbol{\alpha} \cdot \widehat{\boldsymbol{m}}_t / \sqrt{\widehat{\boldsymbol{v}}_t} + \boldsymbol{\epsilon}$$

$|\Delta_t| \lessapprox \alpha.$

Trust-region

$$\widehat{\boldsymbol{m}}_t / \sqrt{\widehat{\boldsymbol{v}}_t} \approx \boldsymbol{E}[\boldsymbol{g}_t] / \sqrt{\boldsymbol{E}[\boldsymbol{g}_t^2]} \leq \mathbf{1}$$



Diederick P. Kingman, and Jimmy Lei Ba [2015] Adam: A method for stochastic optimization. Proceedings of the 3rd International Conference for Learning Representations ICLR15, San Diego, 2015.

Tricks to make learning faster or more effective: Adam (Adaptive moment estimation)

$$w_t = w_{t-1} - \eta \, {}^{\hat{m}_t}/_{\sqrt{\hat{v}_t}+\varepsilon}$$

$$\hat{m}_t = {}^{m_t}/_{(1-\beta_1^t)}$$

$$\boldsymbol{w_t \leftarrow w_{t-1} - \frac{\nabla E}{H}} \qquad \boldsymbol{w_t \leftarrow w_{t-1} - \eta \, {}^{\hat{m}_t}/_{\sqrt{\hat{v}_t}+\varepsilon}}$$

$$\hat{v}_t = {}^{v_t}/_{(1-\beta_2^t)}$$

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$$
$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$$

β_1 is the exponential decay rate gradient (typically, 0.9) and β_2 is the exponential decay rate for the 2nd momentum (typically 0.999)

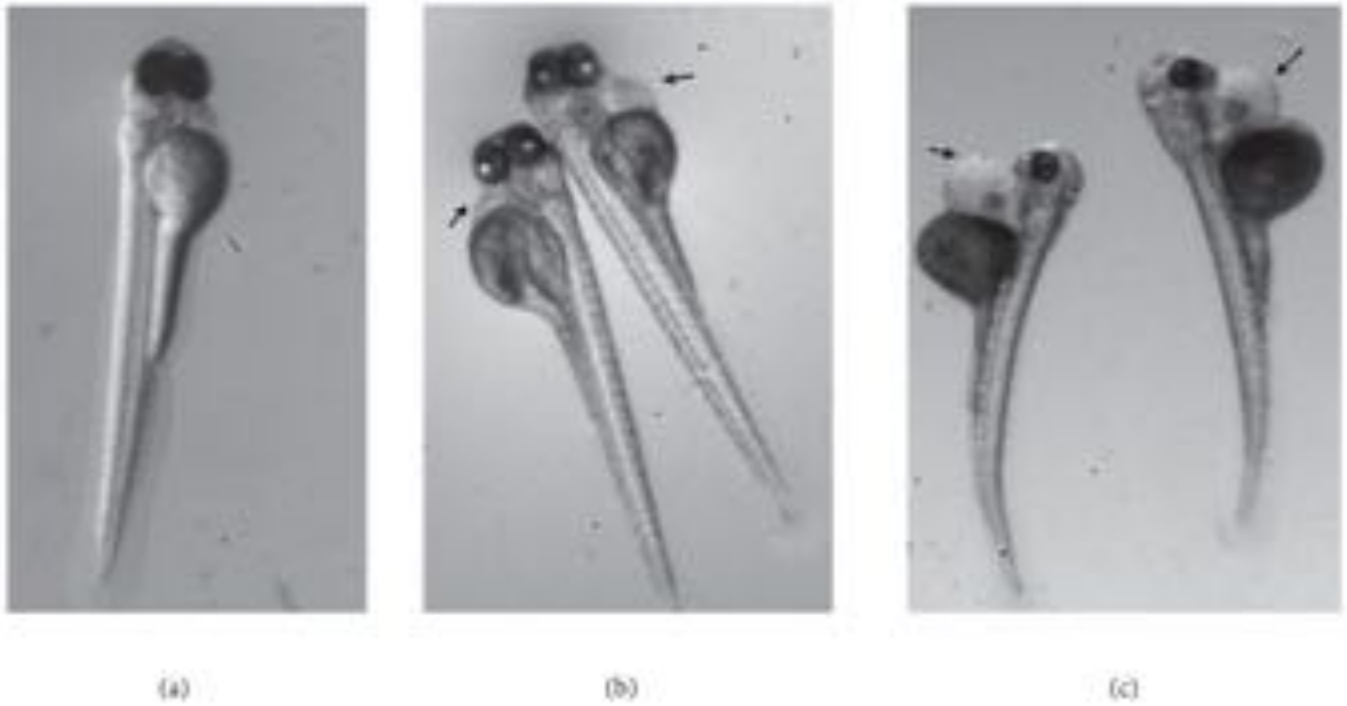
Some Practical Details

- I typically use a Mx800x400x200x100x50x10x50x100x200x800xM structure (12 hidden layers)
- Train by policy:
  - Use Adam
  - mini-batches of 30 data and 30 passes through the data
- For missing data, outputs with missing data are not backpropagated
- More details
  - tanh activation function

Case Study #1:Toxicity challenge data as an example of real-world QSAR data

- 963 training data and 120 test data
- 2223 descriptors
- Consider 6 sigma outliers as missing data

REM GET DATA
mje tox --TOX
REM EXTRACT MOE DESCRIPTORS
REM execute tox (255 1)
tox



Typical phenotype of a zebrafish embryo incubated from 24-hpf to 96 hpf in (a) embryo medium as a Negative control, in (b) 10 mm diethyl-aminobenzaldehyde (DEAB), and in (c) 100 mm DEAB. Note the Deformed embryos in DEAB: short size, scoliosis, yolk, and heart edema (black arrows).

Younes, Fatima Mraiche, Sahar I. Da'as, and Husevin C. Yalkin [2018] Using zebrafish for investigating the Molecular Mechanisms of Drug-induced cardiotoxicity. Biomedical Research International (BMRI), Vol. 2018, Article ID 1642684.

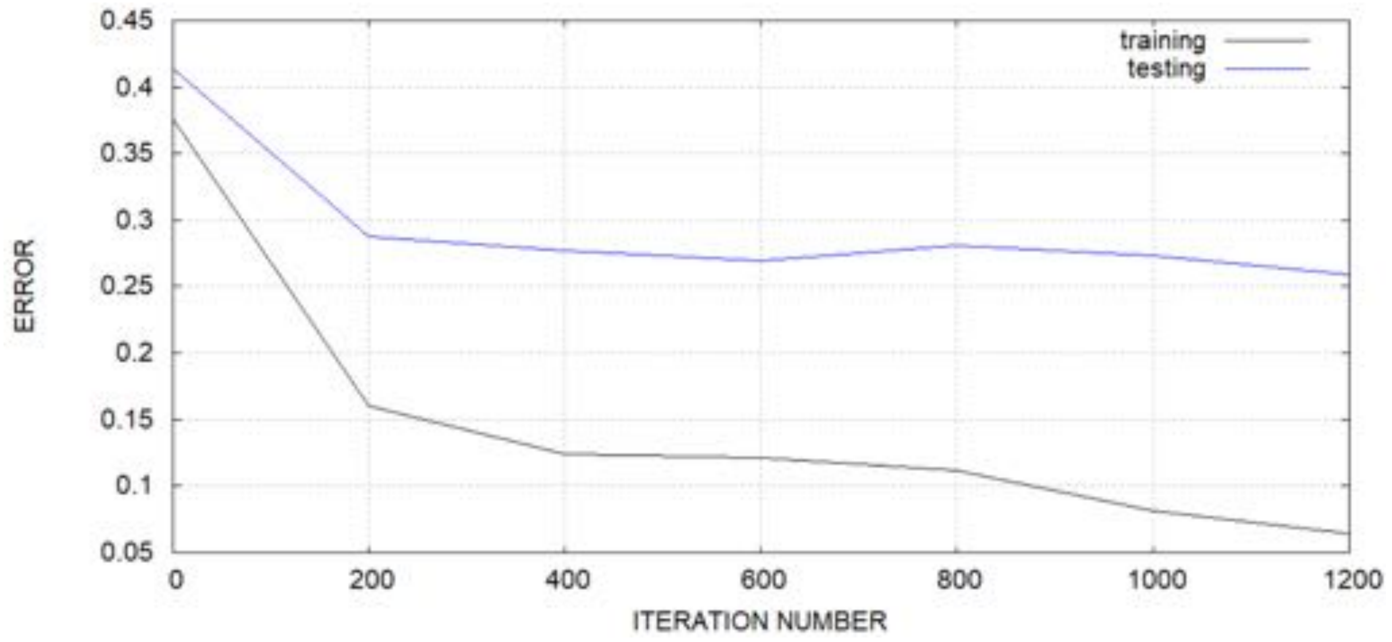
2008 Toxicity challenge data as an example of real-world QSAR data

- ICANN 2008 QSAR Toxicity Data with 2223 descriptors:
  255 MOE descriptors (255 1)
  1664 DRAGON descriptors (1664 256)
  221 SIMULATIOM PLUS descriptors (221 1920)
  60 ELECTRONIC STATE descriptors (60 2141)
  23 QUANTUM CHEMISTRY descriptors (23 2201)
- There are 644+339+110 = 1093 training and 120 test molecules

[1] Hao Zhu, Alexander Tropsha, Denis Fourches, Alexandre Varnek, Ester Papa, Paola Gramatica, Tomas Oberg, Phuong Dao, Artem Cherkasov, and Igor V. Tetko [2008] Combinatorial QSAR Modeling of Chemical Toxicants Tested against Tetrahymena pyriformis. Journal of Chemical Information and Modeling, Vol. 48 pp. 766-784.

[2] Igor V. Tetko, Iurii Sushko, Anil Kumar Pandey, Hao Zhu, Alexander Tropsha, Ester Papa, Tomas Oberg, Robrto Todeschini, Denis Fourches, and Alexandre Varnek [2008] Critical Assessment of QSAR Models of Environmental Toxicity against Tetrahymena pyriformis: Focusing on Applicability Domain of Overfitting by Variable Selection. Journal of Chemical Information and Modeling, Vol. 48, pp. 1733 – 1746.

2008 Toxicity challenge data as an example of real-world QSAR data

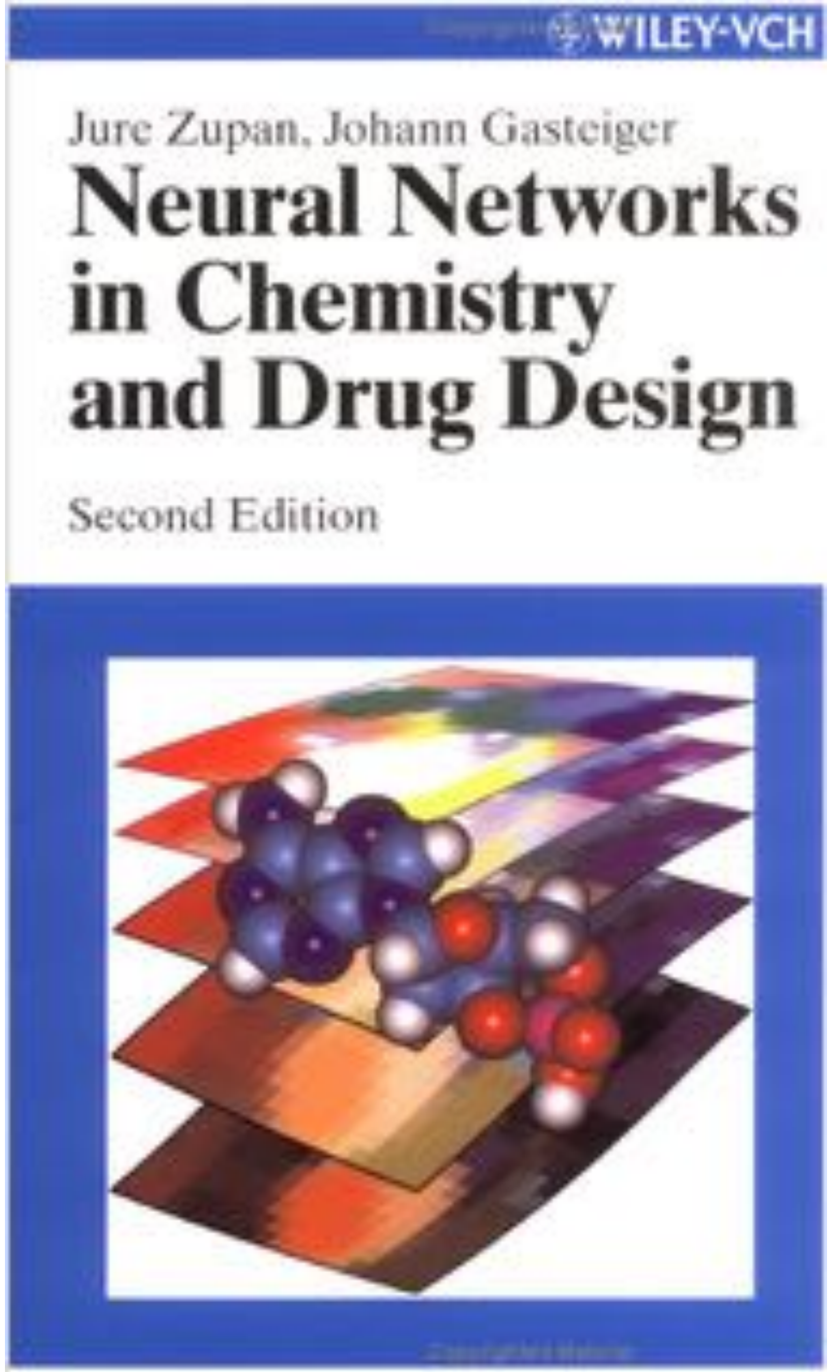


Rank	Username		Diagram	RMSE on blind test set	RMSE on known test set	Status
1	Dr Bunsen Honeydew	Final Winner		0.741	0.353	
2	embrechts	First Pass Winner		0.742	0.367	
3	oobrezanova	Final Winner		0.756	0.41	
4	Lajkonik	First Pass Winner		0.76	0.292	
5	MJE	First Pass Winner		0.765	0.395	
6	mb	First Pass Winner		0.778	0.288	

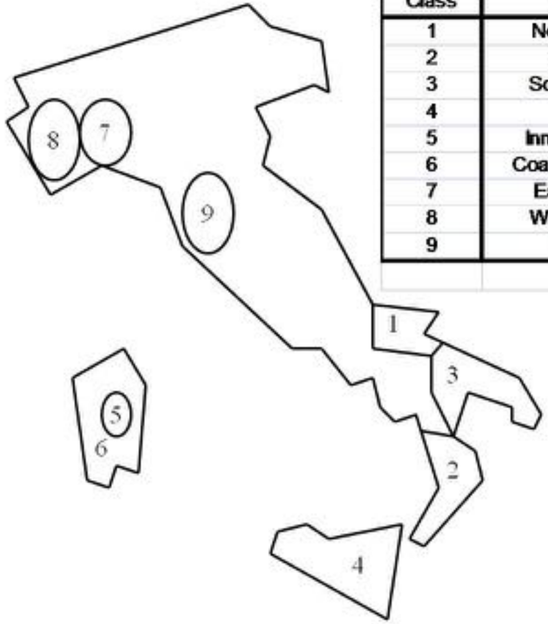
q2	Q2	MSE	MAE	#inputs	net	Note
0.4885	0.5489	0.835	0.598	2223	800-400-200-110-50	
0.4583	0.5103	0.805	0.579	629	same	95% cousin removed
0.4653	0.5300	0.820	0.610	629	same	cousin + six sigma missing
0.5776	0.6613	0.916	0.675	2223	same	six sigma missing
0.4951	0.5452	0.832	0.593	111	same	+ cousins removed
0.4421	0.4663	0.769	0.573	2223	same	six sigma corrected
0.5076	0.5681	0.849	0.607	1429		+ 97% cousin removed

- Would have ranked 6th in competition
- All training was done by policy

HELMHOLTZ MUNICH


WILEY-VCH
Jure Zupan, Johann Gasteiger
Neural Networks in Chemistry and Drug Design
Second Edition

Case study #2: Italian Olive Oil Data



Class	Region	# samples
1	North Apulia	25
2	Calabria	56
3	South Apulia	206
4	Sicily	36
5	Inner Sardinia	65
6	Coastal Sardinia	33
7	East Liguria	50
8	West Liguria	50
9	Umbria	51
		572

- 9-class Italian olive oil data for 572 Italian olive oils
  - 8 fatty acid indicators by 9 regions
  - 9-class data are not balanced by region
  - Gasteiger used 250 training data (we do same)

[1] M. Forina and C. Armanino [1981] Eigenvector projection and simplified nonlinear mapping of fatty acid content of Italian olive oils. Ann. Chem, Vol. 72, pp. 125-127.
[2] Jure Zapan and Johann Gasteiger [1999] Neural networks in chemistry and drug design (2nd edition). Wiley - VCH

Issue related to Italian Olive Oil Data

- Missing data are systemic: some classes have descriptors missing over entire class ➔ gives away answer
- Data curation: flag missing data by -999

32.9	14.3	45.7	67.3	35.9	58.3	65	80.4	4	318
31.7	15.1	54.3	66	37.7	59.7	59.2	80.4	4	319
46	21.9	90.6	58.2	27.4	58.3	71.8	64.3	4	320
51.4	26.4	54.3	54.3	35.8	58.3	68.9	71.4	4	321
57.5	30.2	42.6	52.4	35.6	56.9	66	58.9	4	322
58	27.5	66.8	55.5	25	56.9	67	51.8	4	323
46.1	39.6	31.4	46.1	65	56.9	93.2	**0**	5	324
38.4	45.3	26	51	65.4	45.8	85.4	**0**	5	325
43.8	30.6	26	51.2	62.3	41.7	89.3	**0**	5	326
45.2	30.9	30.9	46.4	69.1	45.8	89.3	**0**	5	327

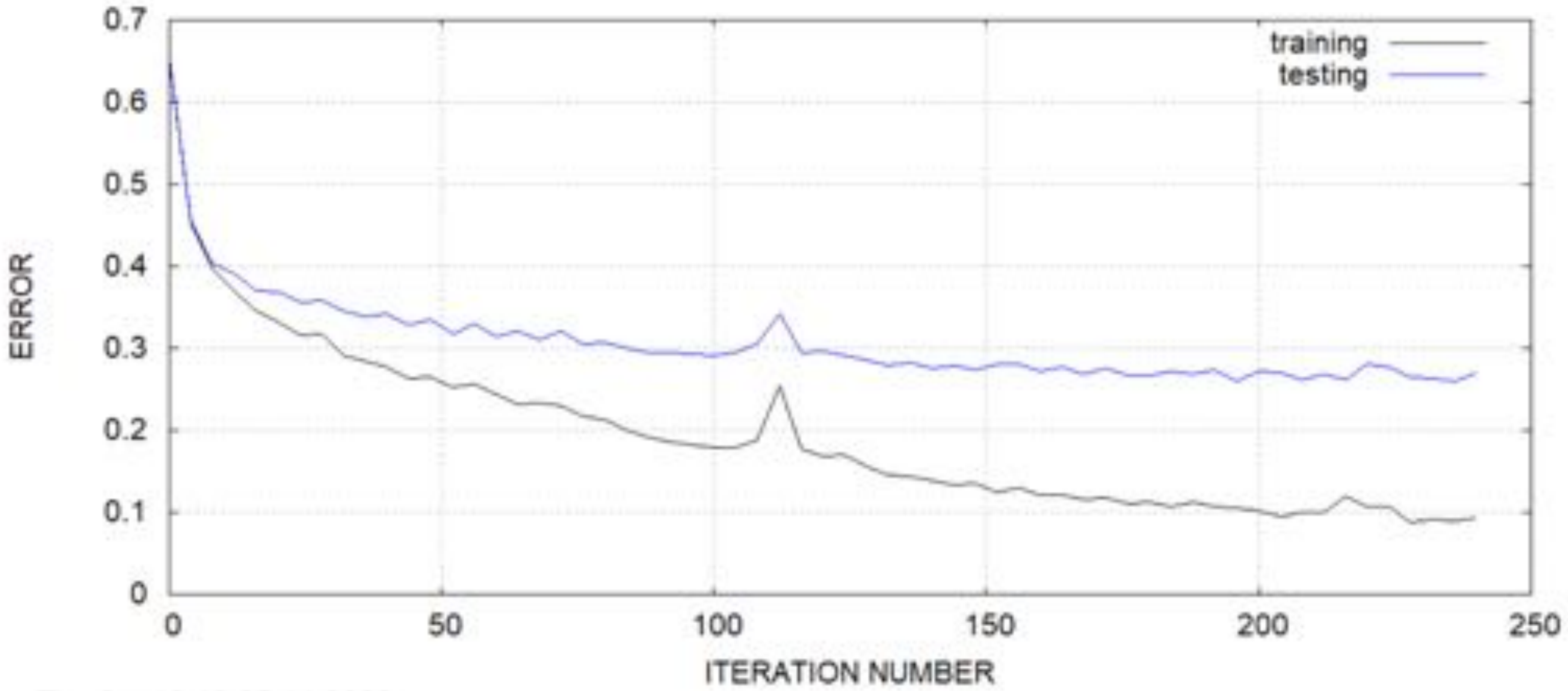
class

- Descriptors for certain classes are missing, but set to zero in the original data set

HELMHOLTZ MUNICH

Case study #2: Italian Olive Oil Data

q2	Q2	%COR	BER(%)	F1	Comment
0.0269	0.0274	94.099	92.59	0.92	Original Data (8x400x200x110x50x23x9 net) – systemic issue
0.1192	0.1230	88.820	85.73	0.85	Missing data replaced by average (8x400x200x110x50x23x9 net)
0.1372	0.1440	89.441	84.82	0.84	Missing data inferred with auto-associator
0.1018	0.1050	90.062	87.15	0.86	same but bottleneck is now 7 rather than 8 neurons
0.0975	0.1015	90.373	86.52	0.86	same but 6 neurons in bottleneck



Conclusions

- Missing data are often systemic
- Missing data are usually represented by -999 (as a tag)
- Faulty data (e.g., six-sigma outliers in MOE descriptors can be flagged as missing and then inferred
- Auto-associators are an effective trick to infer missing data
- This approach can also be used on the NETFLIX challenge