Deep Turnaround in Machine Translation

Ondřej Bojar
bojar@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University, Prague

October 30, 2017

Outline

- WMT17 Challenge.
- SMT: MTUs and Lack of Context.
- Neural Networks as Universal Approximators.
- Processing Words and Sentences.
- Encoder-Decoder with Attention.
- Learning Meaning Representations.

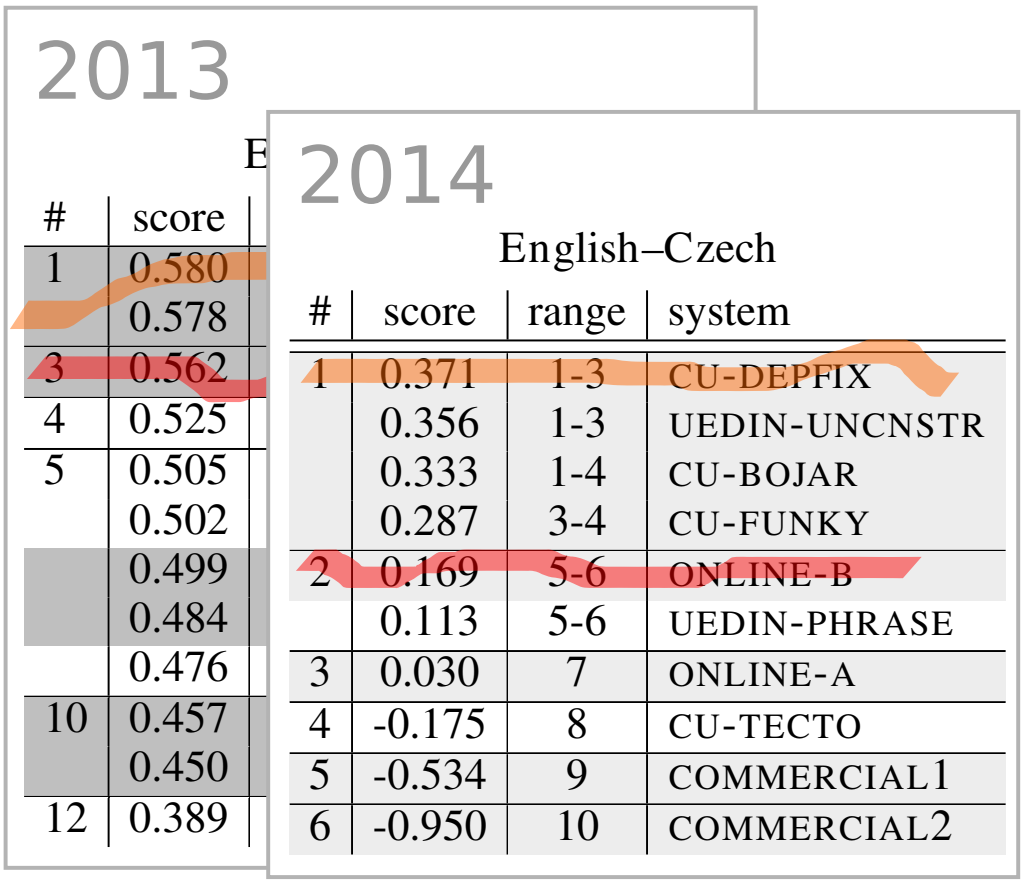
Recent WMT History

2013

English-Czech

#	score	range	system
1	0.580	1-2	CU-BOJAR
	0.578	1-2	CU-DEPFIX
3	0.562	3	ONLINE-B
4	0.525	4	UEDIN
5	0.505	5-7	CU-ZEMAN
	0.502	5-7	MES
	0.499	5-8	ONLINE-A
	0.484	7-9	CU-PHRASEFIX
	0.476	8-9	CU-TECTOMT
10	0.457	10-11	COMMERCIAL-1
	0.450	10-11	COMMERCIAL-2
12	0.389	12	SHEF-WPROA

Recent WMT History



2013

E

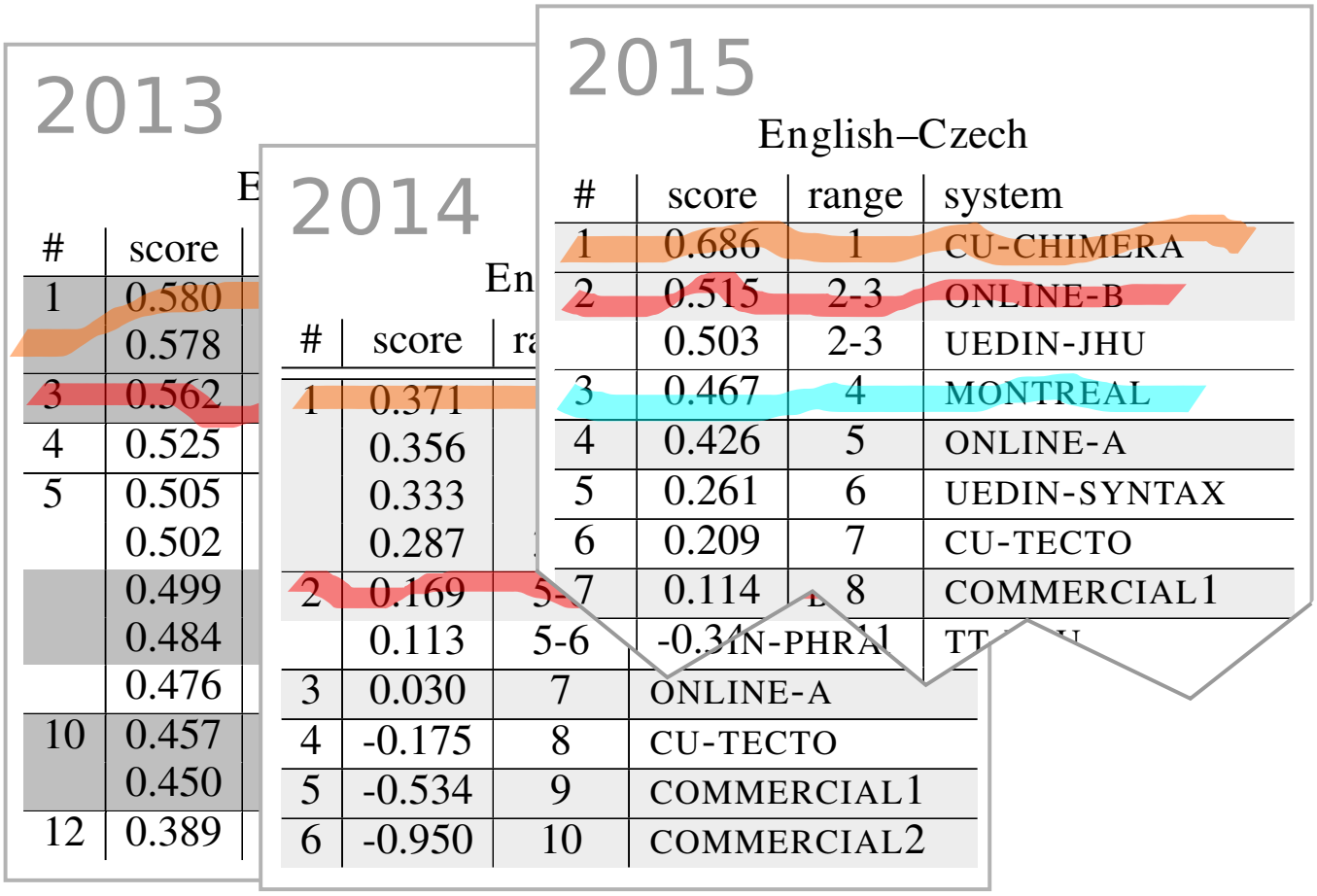
#	score
1	0.580
	0.578
3	0.562
4	0.525
5	0.505
	0.502
	0.499
	0.484
	0.476
10	0.457
	0.450
12	0.389

2014

English–Czech

#	score	range	system
1	0.371	1-3	CU-DEPFIX
	0.356	1-3	UEDIN-UNCNSTR
	0.333	1-4	CU-BOJAR
	0.287	3-4	CU-FUNKY
2	0.169	5-6	ONLINE-B
	0.113	5-6	UEDIN-PHRASE
3	0.030	7	ONLINE-A
4	-0.175	8	CU-TECTO
5	-0.534	9	COMMERCIAL1
6	-0.950	10	COMMERCIAL2

Recent WMT History



2013

#	score
1	0.580
	0.578
3	0.562
4	0.525
5	0.505
	0.502
	0.499
	0.484
	0.476
10	0.457
	0.450
12	0.389

2014

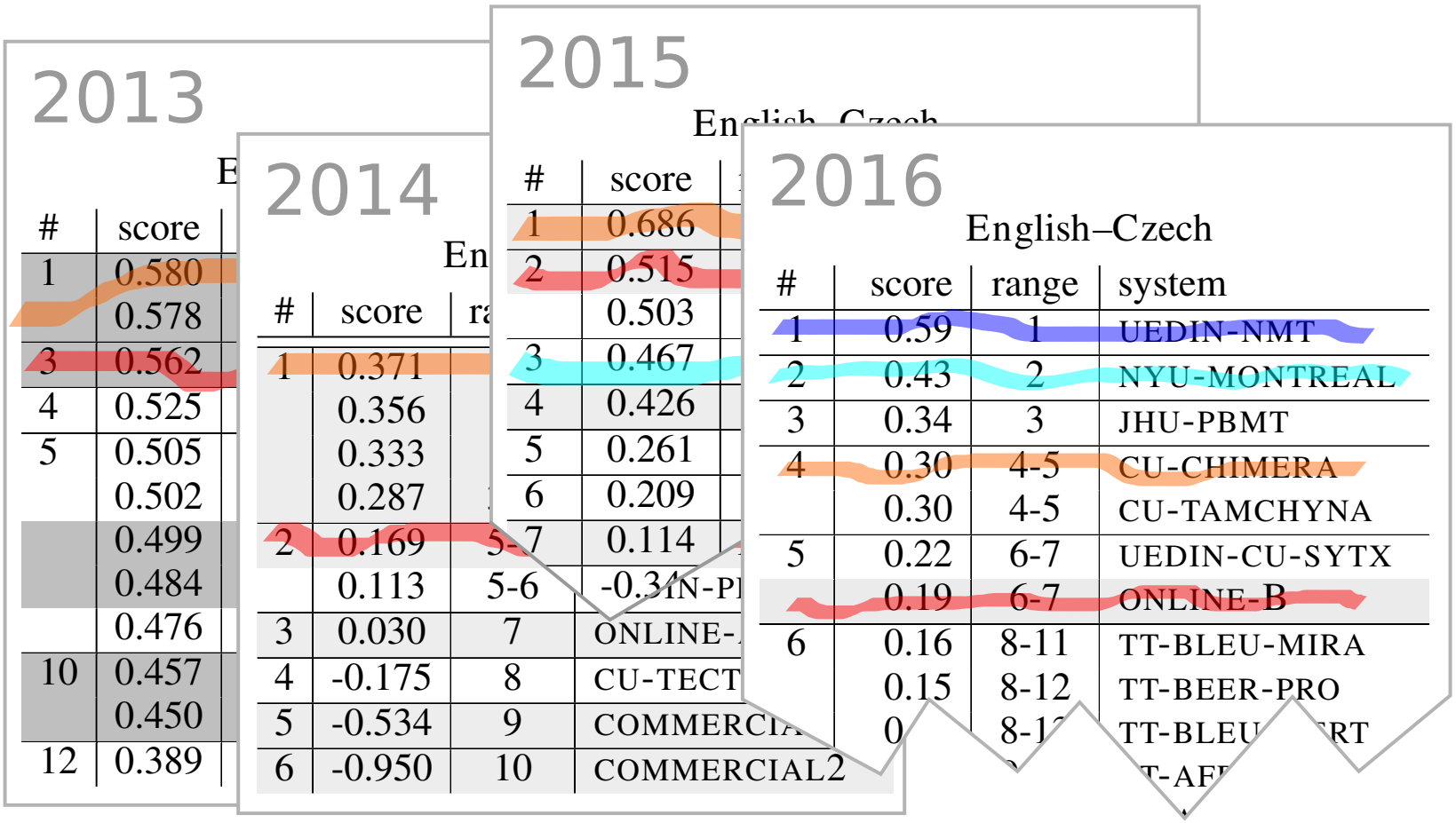
#	score	ra
1	0.371	
	0.356	
	0.333	
	0.287	
2	0.169	5-7
	0.113	5-6
3	0.030	7
4	-0.175	8
5	-0.534	9
6	-0.950	10

2015

English–Czech

#	score	range	system
1	0.686	1	CU-CHIMERA
2	0.515	2-3	ONLINE-B
	0.503	2-3	UEDIN-JHU
3	0.467	4	MONTREAL
4	0.426	5	ONLINE-A
5	0.261	6	UEDIN-SYNTAX
6	0.209	7	CU-TECTO
	0.114	8	COMMERCIAL1
	-0.34	11	N-PHRA

Recent WMT History



Our Collective Efforts for WMT17

- Neural Monkey (Helcl and Libovický, 2017).
- NMT Training Task (Bojar et al., 2017).

- BPE, Learning rate and other meta-parameters.
- Batch sizing (smaller/larger/variable).
- Additional training objective:
  - Targetting GIZA++ alignments.
  - Scoring the *set* of produced words, disregarding position.
- Minibatch bucketing.
- Curriculum learning. (Kocmi and Bojar, 2017)
- Pre-trained embeddings.
- Domain adaptation: Subsample for Testset / Each Doc.
- Neural sys combination: Concatenative/Multi-encoder.

... If Gains, then Mediocre ...

- Neural Monkey (Helcl and Libovický, 2017).
- NMT Training Task (Bojar et al., 2017).

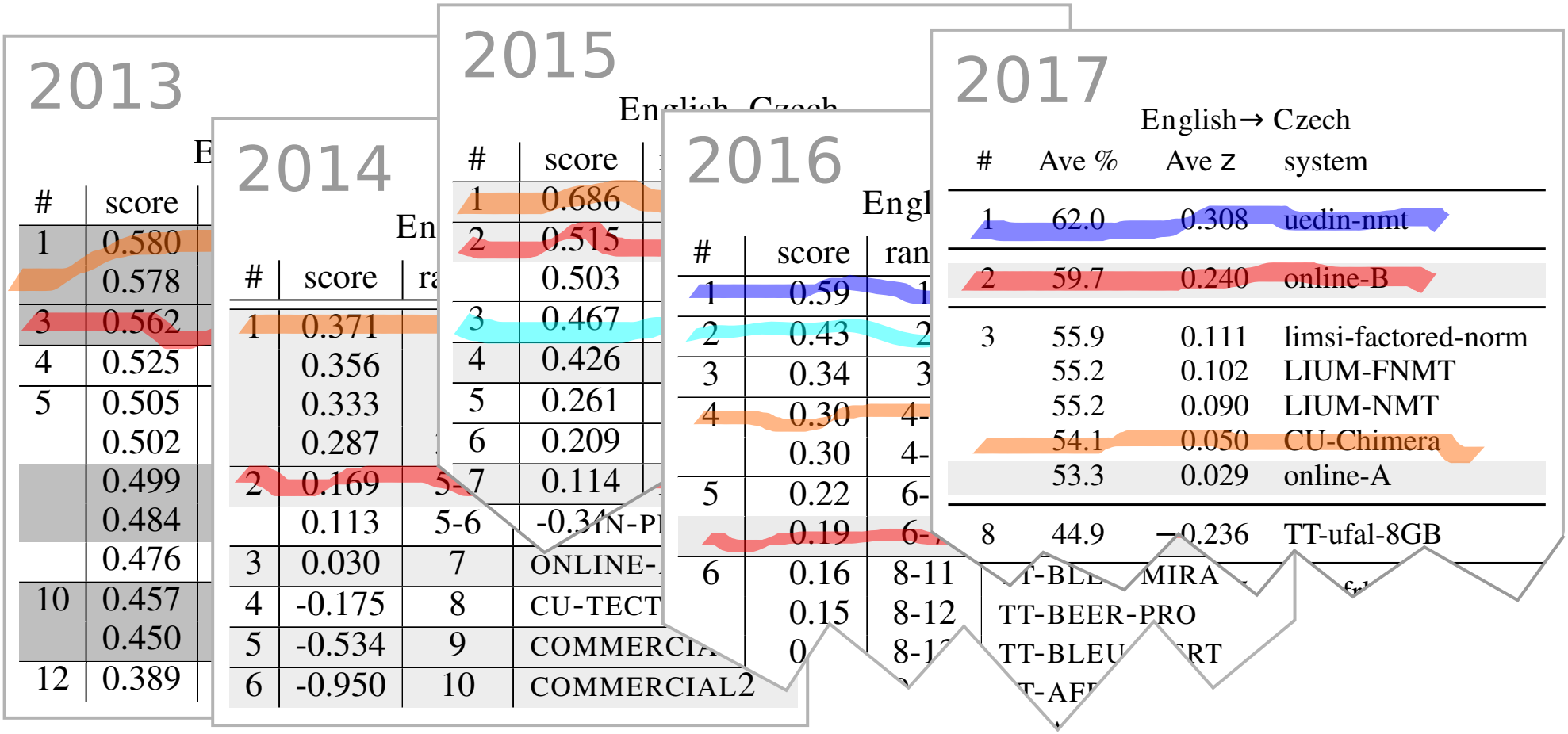
- ~~BPE, Learning rate and other meta-parameters.~~
- ~~Batch sizing (smaller/larger/variable).~~
- Additional training objective:
  - ~~Targetting GIZA++ alignments.~~
  - ~~Scoring the *set* of produced words, disregarding position.~~
- ~~Minibatch bucketing.~~
- ~~Curriculum learning.~~ (Kocmi and Bojar, 2017)
- ~~Pre-trained embeddings.~~
- Domain adaptation: Subsample for Testset / ~~Each Doc.~~
- ~~Neural sys combination: Concatenative/Multi-encoder.~~

Our WMT System

... so we decided to stick with phrase-based MT backbone:

- Train on back-translated mononews only.
- Beam search in Neural Monkey.
- **Chimera-style combination:**
  - Moses system with several phrase tables:
    - Standard corpus-based one (synthetic mononews only!).
    - Output of TectoMT for the test set.
    - **Output of Nematus 2016 and Neural Monkey 2017.**
  - Followed by Depfix.

... And the Result:



We Were Hoping to Be the Second!

	Manual		Automatic Scores				
#	Ave %	Ave z	BLEU	TER	CharacTER	BEER	System
1	**62.0**	**0.308**	**22.8**	**0.667**	**0.588**	**0.540**	uedin-nmt
2	59.7	0.240	20.1	0.703	0.612	0.519	online-B
3	55.9	0.111	20.2	0.696	0.607	0.524	limsi-factored
	55.2	0.102	20.0	0.699	-	-	LIUM-FNMT
	55.2	0.090	20.2	0.701	0.605	0.522	LIUM-NMT
	54.1	0.050	20.5	0.696	0.624	0.523	**CU-Chimera**
	53.3	0.029	16.6	0.743	0.637	0.503	online-A
8	41.9	-0.327	16.2	0.757	0.697	0.485	PJATK

Details on CU-Chimera in Sudarikov et al. (2017).
Automatic scores by `http://matrix.statmt.org/`.

Is UEDIN NMT That Much Better?

SRC	28-Year-Old Chef Found Dead at San Francisco Mall
	28letý šéfkuchař Found Dead v San Francisco Mall
	Osmadvacetiletý šéfkuchař nalezen mrtev v obchodě v San Francisku

Is UEDIN NMT That Much Better?

SRC	28-Year-Old Chef Found Dead at San Francisco Mall
MT	28letý šéfkuchař Found Dead v San Francisco Mall
REF	Osmadvacetiletý šéfkuchař nalezen mrtev v obchodě v San Francisku

Is UEDIN NMT That Much Better?

SRC	28-Year-Old Chef Found Dead at San Francisco Mall
MT	28letý šéfkuchař Found Dead v San Francisco Mall
REF	Osmadvacetiletý šéfkuchař nalezen mrtev v obchodě v San Francisku
SRC	A 28-year-old chef who had recently moved to San Francisco was found dead in the stairwell of a local mall this week.
	Osmadvacetiletý kuchař, který se nedávno přestěhoval do San Francisca, byl tento týden nalezen mrtvý na schodišti místního obchodního centra.
	Osmadvacetiletý šéfkuchař, který se nedávno přistěhoval do San Franciska, byl tento týden schodech místního obchodu.

Is UEDIN NMT That Much Better?

SRC	28-Year-Old Chef Found Dead at San Francisco Mall
MT	28letý šéfkuchař Found Dead v San Francisco Mall
REF	Osmadvacetiletý šéfkuchař nalezen mrtev v obchodě v San Francisku
SRC	A 28-year-old chef who had recently moved to San Francisco was found dead in the stairwell of a local mall this week.
MT	Osmadvacetiletý kuchař, který se nedávno přestěhoval do San Francisca, byl tento týden nalezen mrtvý na schodišti místního obchodního centra.
REF	Osmadvacetiletý šéfkuchař, který se nedávno přistěhoval do San Franciska, byl tento týden ∅ schodech místního obchodu.

Is UEDIN NMT That Much Better? (2/4)

SRC A spokesperson for Sons & Daughters said they were "shocked and devastated" by his death.

Mluvčí společnosti Sons & Daughters **uvedla**, že jsou jeho smrtí "šokováni a zdrceni".

Mluvčí restaurace Sons & Daughters **řekl**, že jsou jeho smrtí „šokováni a zničeni“.

Is UEDIN NMT That Much Better? (2/4)

SRC	A spokesperson for Sons & Daughters said they were "shocked and devastated" by his death.
MT	Mluvčí společnosti Sons & Daughters **uvedla**, že jsou jeho smrtí "šokováni a zdrceni".
REF	Mluvčí restaurace Sons & Daughters **řekl**, že jsou jeho smrtí „šokováni a zničeni“.

Is UEDIN NMT That Much Better? (2/4)

SRC	A spokesperson for Sons & Daughters said they were "shocked and devastated" by his death.
MT	Mluvčí společnosti Sons & Daughters **uvedla**, že jsou jeho smrtí "šokováni a zdrceni".
REF	Mluvčí restaurace Sons & Daughters **řekl**, že jsou jeho smrtí „šokováni a zničeni“.
SRC	"He found an apartment, he was dating a girl," Louis Galicia told KGO.
	„Našel si byt, chodil s dívkou,“ řekl Louis Galici**a** pro KGO.
	"Našel si byt, chodil s holkou," řekl Louis Galici**e** KGO.

Is UEDIN NMT That Much Better? (2/4)

SRC	A spokesperson for Sons & Daughters said they were "shocked and devastated" by his death.
MT	Mluvčí společnosti Sons & Daughters **uvedla**, že jsou jeho smrtí "šokováni a zdrceni".
REF	Mluvčí restaurace Sons & Daughters **řekl**, že jsou jeho smrtí „šokováni a zničeni".
SRC	"He found an apartment, he was dating a girl," Louis Galicia told KGO.
REF	„Našel si byt, chodil s dívkou," řekl Louis Galici**a** pro KGO.
MT	"Našel si byt, chodil s holkou," řekl Louis Galici**e** KGO.

Is UEDIN NMT That Much Better? (3/4)

SRC The police arrested two men, who on Tuesday attacked a thirty-five-year-old man with a knife and a machete.

Policie **obvinila** dva útočníky, kteří v úterý v centru Olomouce napadli nožem a mačetou pětatřicetiletého muže.

Policie **zatkla** dva muže, kteří v úterý napadli pětatřicetiletého muže nožem a mačetou.

Is UEDIN NMT That Much Better? (3/4)

SRC	The police arrested two men, who on Tuesday attacked a thirty-five-year-old man with a knife and a machete.
REF	Policie **obvinila** dva útočníky, kteří v úterý v centru Olomouce napadli nožem a mačetou pětatřicetiletého muže.
MT	Policie **zatkla** dva muže, kteří v úterý napadli pětatřicetiletého muže nožem a mačetou.

Is UEDIN NMT That Much Better? (3/4)

SRC The police arrested two men, who on Tuesday attacked a thirty-five-year-old man with a knife and a machete.

REF Policie **obvinila** dva útočníky, kteří v úterý v centru Olomouce napadli nožem a mačetou pětatřicetiletého muže.

MT Policie **zatkla** dva muže, kteří v úterý napadli pětatřicetiletého muže nožem a mačetou.

SRC There were creative differences on the set and a disagreement.

Došlo ke vzniku kreativních rozdílů na scéně a k neshodám.

Na place byly tvůrčí rozdíly a neshody.

Is UEDIN NMT That Much Better? (3/4)

SRC	The police arrested two men, who on Tuesday attacked a thirty-five-year-old man with a knife and a machete.
REF	Policie **obvinila** dva útočníky, kteří v úterý v centru Olomouce napadli nožem a mačetou pětatřicetiletého muže.
MT	Policie **zatkla** dva muže, kteří v úterý napadli pětatřicetiletého muže nožem a mačetou.
SRC	There were creative differences on the set and a disagreement.
REF	Došlo ke vzniku kreativních rozdílů na scéně a k neshodám.
MT	Na place byly tvůrčí rozdíly a neshody.

Is UEDIN NMT That Much Better? (4/4)

SRC	Economy Secretary Keith Brown visited the site today and was among the first to walk from the land on to the bridge.
	Ekonomický tajemník Keith Brown stavbu dnes navštívil a byl mezi prvními, kteří přišli z pevniny na most.
	Ministr hospodářství Keith Brown dnes místo navštívil a byl mezi prvními, kteří vyšli ze země na most.

Is UEDIN NMT That Much Better? (4/4)

SRC	Economy Secretary Keith Brown visited the site today and was among the first to walk from the land on to the bridge.
REF	Ekonomický tajemník Keith Brown stavbu dnes navštívil a byl mezi prvními, kteří přišli z pevniny na most.
MT	Ministr hospodářství Keith Brown dnes místo navštívil a byl mezi prvními, kteří vyšli ze země na most.

UEDIN at WMT17

- Our small annotation of up to 185 sentences.
- Blind mix: reference or MT.

Real MT was assumed to be:

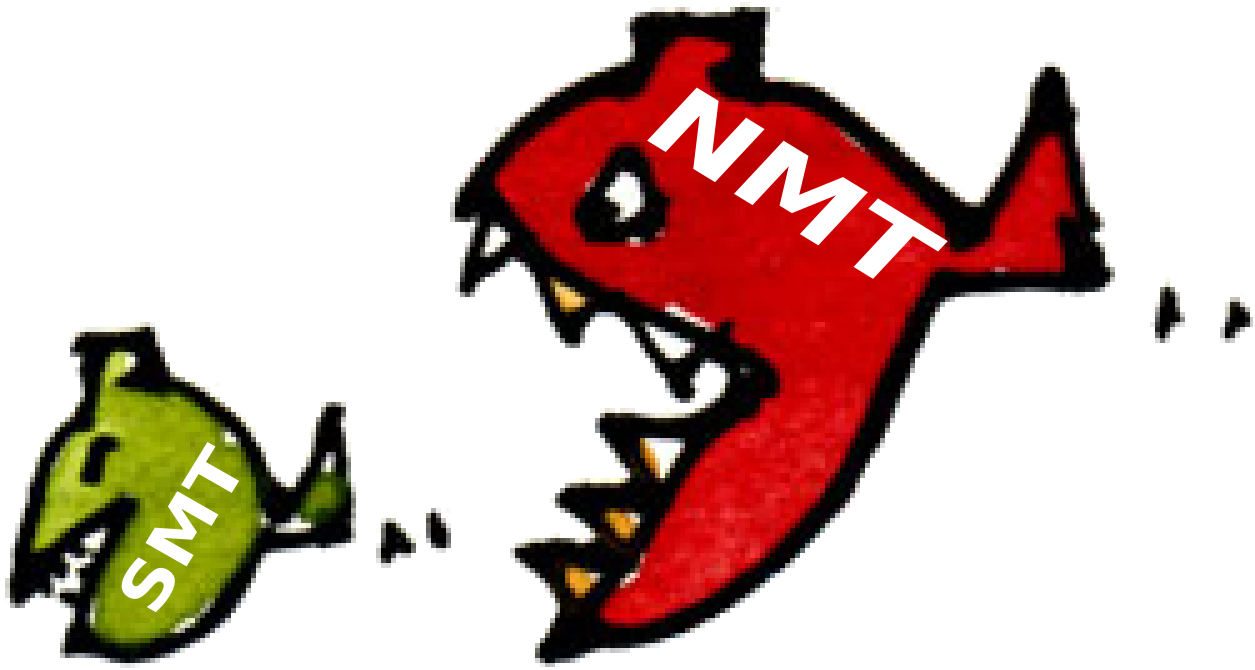
	OB	DM	DV
MT	142 (76.8 %)	86 (77.5 %)	72 (87.8 %)
didn't know	34 (18.4 %)	9 (8.1 %)	6 (7.3 %)
human	9 (4.9 %)	16 (14.4 %)	4 (4.9 %)
Total	185 (100.0 %)	111 (100.0 %)	82 (100.0 %)

$\Rightarrow$ 10–20% of outputs indistinguishable from humans.

	OB	DM	DV
almost flawless	17 (9.19 %)	2 (1.80 %)	0 (0 %)
flawless	82 (44.32 %)	37 (33.33 %)	27 (32.93 %)

$\Rightarrow$ 30–50% of outputs flawless or almost flawless.

A Clear Result



Fish by Frits Ahlefeldt

Catastropic Errors Happen

Also WMT17 UEDIN outputs (but not easy to spot):

SRC ... said Frank initially stayed in hostels...
MT ... řekl, že Frank původně zůstal v **Budějovicích**...

SRC Most of the Clintons' income...
MT Většinu příjmů **Kliniky**...

SRC The 63-year-old has now been made a special represen
MT 63letý **mladík** se nyní stal zvláštním zástupcem...

SRC He listened to the moving stories of the women.
MT Naslouchal **pohyblivým** příběhům žen.

Catastropic Errors Happen (2/2)

SRC Criminal Minds star Thomas Gibson sacked after hitting producer
REF Thomas Gibson, hvězda seriálu Myšlenky zločince, byl propuštěn po té, co uhodil režiséra
MT **Kriminalisté Minsku** hvězdu Thomase Gibsona **vyhostili** po **zásahu** producenta

SRC ...add to that its long-standing grudge...
REF ...přidejte k tomu svou dlouholetou nenávist...
MT ...přidejte k tomu svou dlouholetou **záštitu**...
(grudge → zášť → záštita)

German→Czech SMT vs. NMT

- A smaller dataset, very first (but comparable) results.
- NMT performs better on average, but occasionally:

SRC	Das Spektakel ähnelt dem Eurovision Song Contest.
REF	Je to jako pěvecká soutěž Eurovision.
SMT	Podívan**ou** připomíná hudební soutěž Eurovize.
NMT	Divadlo se podobá Eurovizi **Conview**.
SRC	Erderwärmung oder Zusammenstoß mit Killerasteroid.
REF	Globální oteplení nebo kolize se zabijáckým asteroidem.
SMT	Globální oteplování, nebo srážka s **Killerasteroid**.
NMT	Globální oteplování, nebo střet **s zabijákem**.
SRC	Zu viele verletzte Gefühle.
REF	Příliš mnoho nepřátelských pocitů.
SMT	Příliš mnoho zraněných pocit**y**.
NMT	Příliš mnoho zraněných.

Reminder: Statistical MT

Given a source (foreign) language sentence $f_1^J = f_1 \dots f_j \dots f_J$,
Produce a target language (English) sentence
$e_1^I = e_1 \dots e_j \dots e_I$.
Among all possible e_1^I, choose the most likely one:

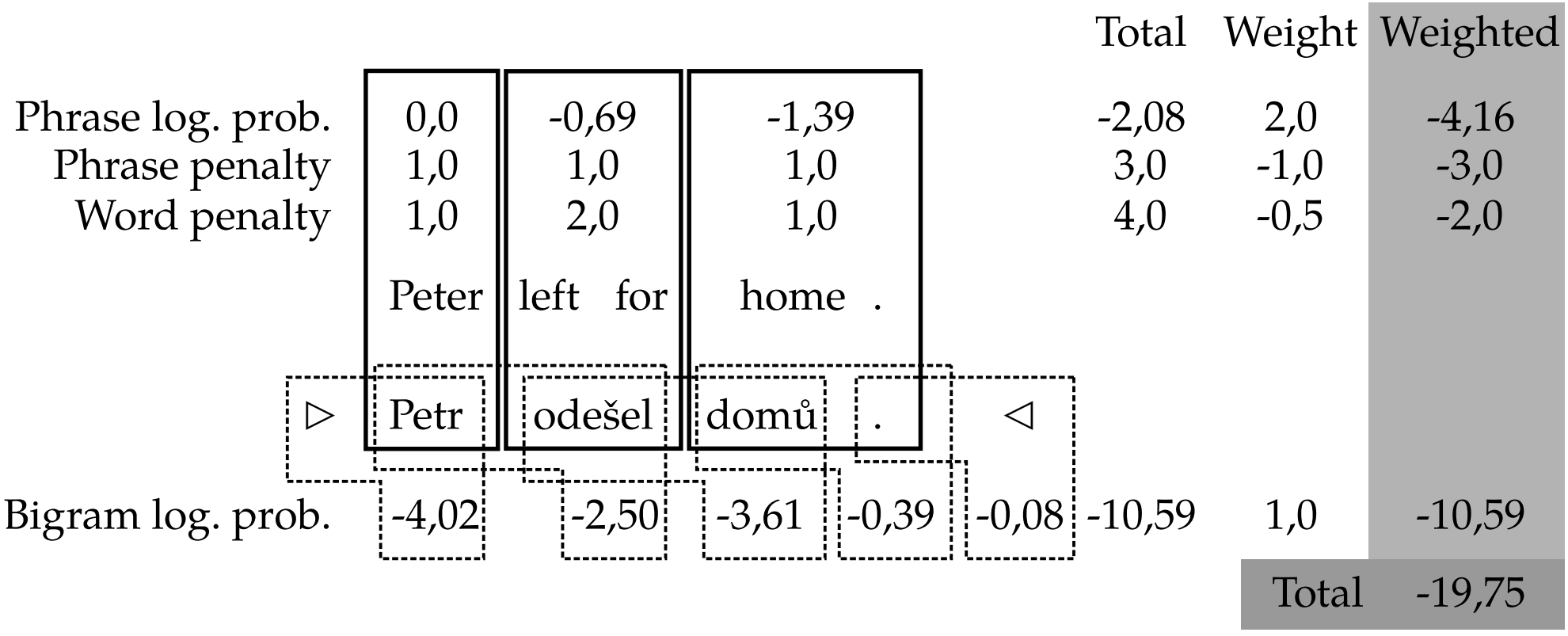
$$\hat{e}_1^{\hat{I}} = \underset{I, e_1^I}{\operatorname{argmax}}\, p(e_1^I | f_1^J) \tag{1}$$

Bayes' law divided the model into components:

$$\hat{e}_1^{\hat{I}} = \underset{I, e_1^I}{\operatorname{argmax}}\, p(f_1^J | e_1^I) p(e_1^I) \tag{2}$$

$p(f_1^J \mid e_1^I)$	Translation model (TM, "reversed", $e_1^I \rightarrow f_1^J$)
$p(e_1^I)$	Language model (LM)

LM and TM Use Smaller Units



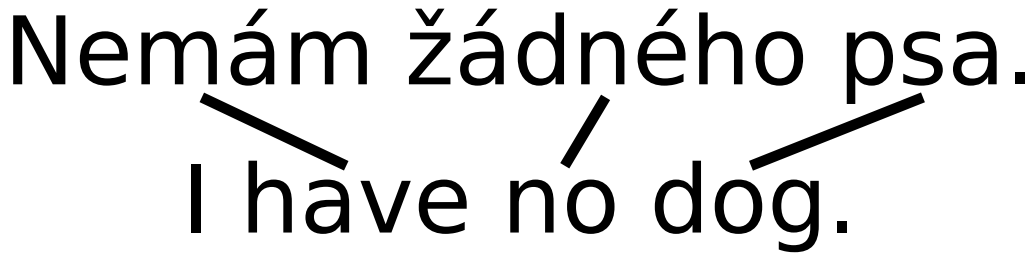
- Output composed of selected TM units (phrase pairs, solid)
- Scored with LM *n*-grams (dashed) and other features.
- Scores weighted (log-linear, not pure Bayes actually).

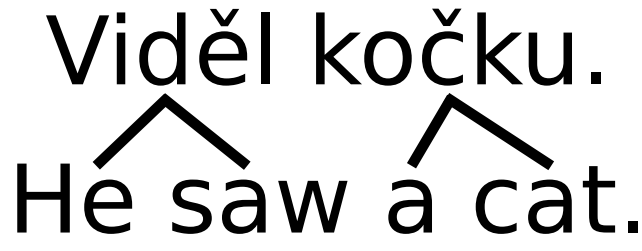
1: Align Training Sentences

Nemám žádného psa. Viděl kočku.

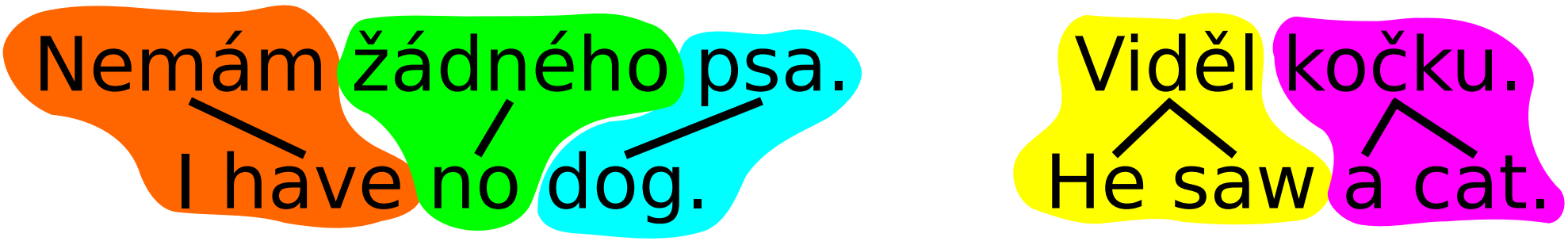
I have no dog. He saw a cat.

2: Align Words

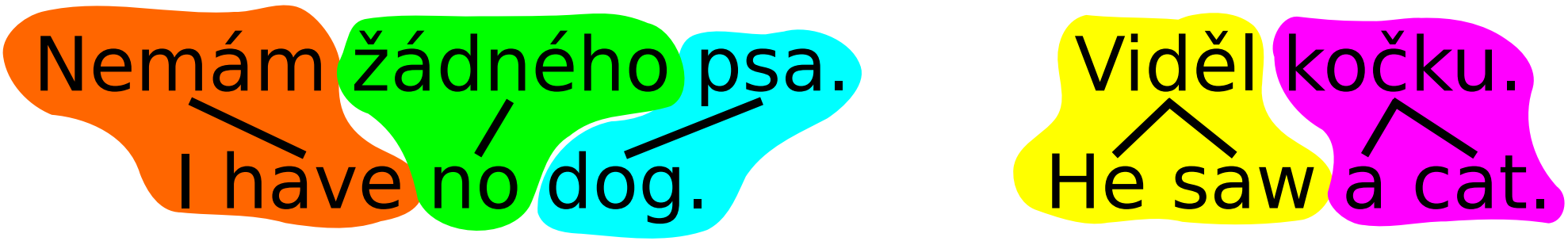




3: Extract Phrase Pairs (MTUs)

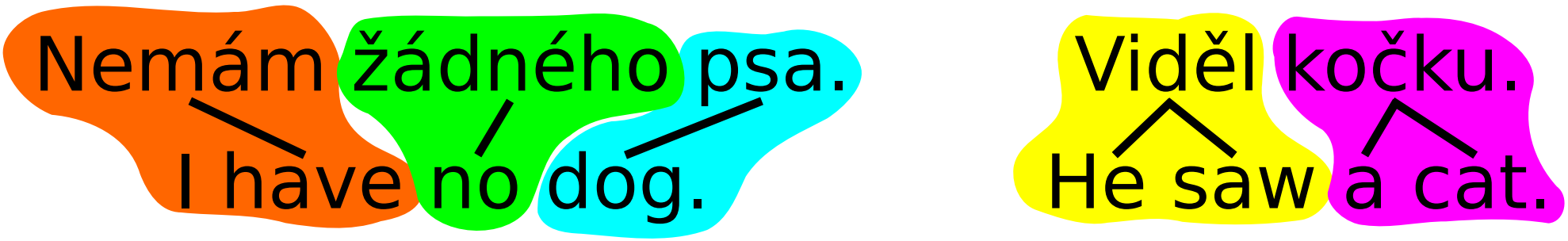


4: New Input



New input: Nemám kočku.

4: New Input



... I don't have cat.

New input: Nemám kočku.

5: Pick Probable Phrase Pairs (TM)

6: So That *n*-Grams Probable (LM)

Meaning Got Reversed!

What Went Wrong?

$$\hat{\hat{e}}_1^{\hat{I}} = \underset{I,e_1^I}{\operatorname{argmax}}\, p(f_1^J|e_1^I)p(e_1^I) = \underset{I,e_1^I}{\operatorname{argmax}} \prod_{(\hat{f},\hat{e})\in\text{phrase pairs of } f_1^J,e_1^I} p(\hat{f}|\hat{e})p(e_1^I) \tag{3}$$

- Too strong phrase-independence assumption.
  - Phrases do depend on each other.
    Here "nemám" and "žádného" jointly express one negation.
  - Word alignments ignored that dependence.
    But adding it would increase data sparseness.
- Language model is a separate unit.
  - $p(e_1^I)$ models the target sentence independently of f_1^J.

Redefining $p(e_1^I|f_1^J)$

What if we modelled $p(e_1^I|f_1^J)$ directly, word by word:

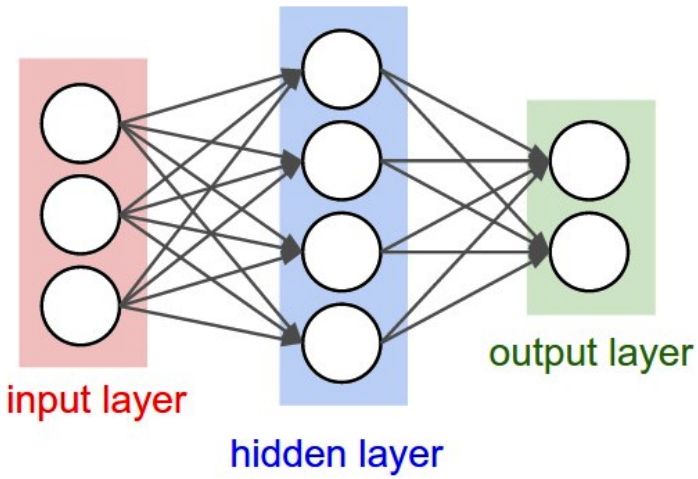
$$\begin{aligned} p(e_1^I|f_1^J) &= p(e_1, e_2, \ldots e_I|f_1^J) \\ &= p(e_1|f_1^J) \cdot p(e_2|e_1, f_1^J) \cdot p(e_3|e_2, e_1, f_1^J) \ldots \\ &= \prod_{i=1}^{I} p(e_i|e_1, \ldots e_{i-1}, f_1^J) \end{aligned} \tag{4}$$

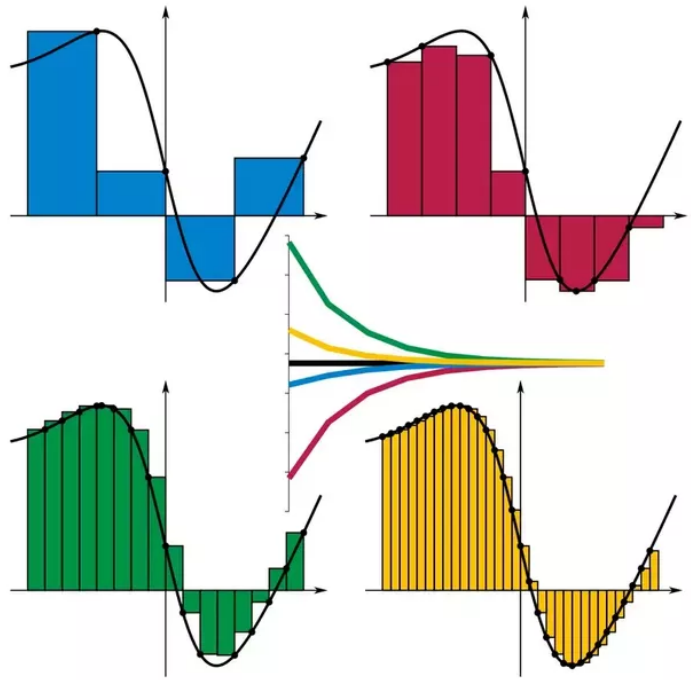
...this is "just a cleverer language model:" $p(e_1^I) = \prod_{i=1}^{I} p(e_i|e_1, \ldots e_{i-1})$

Main Benefit: All dependencies available.

But what technical device can learn this?

NNs: Universal Approximators

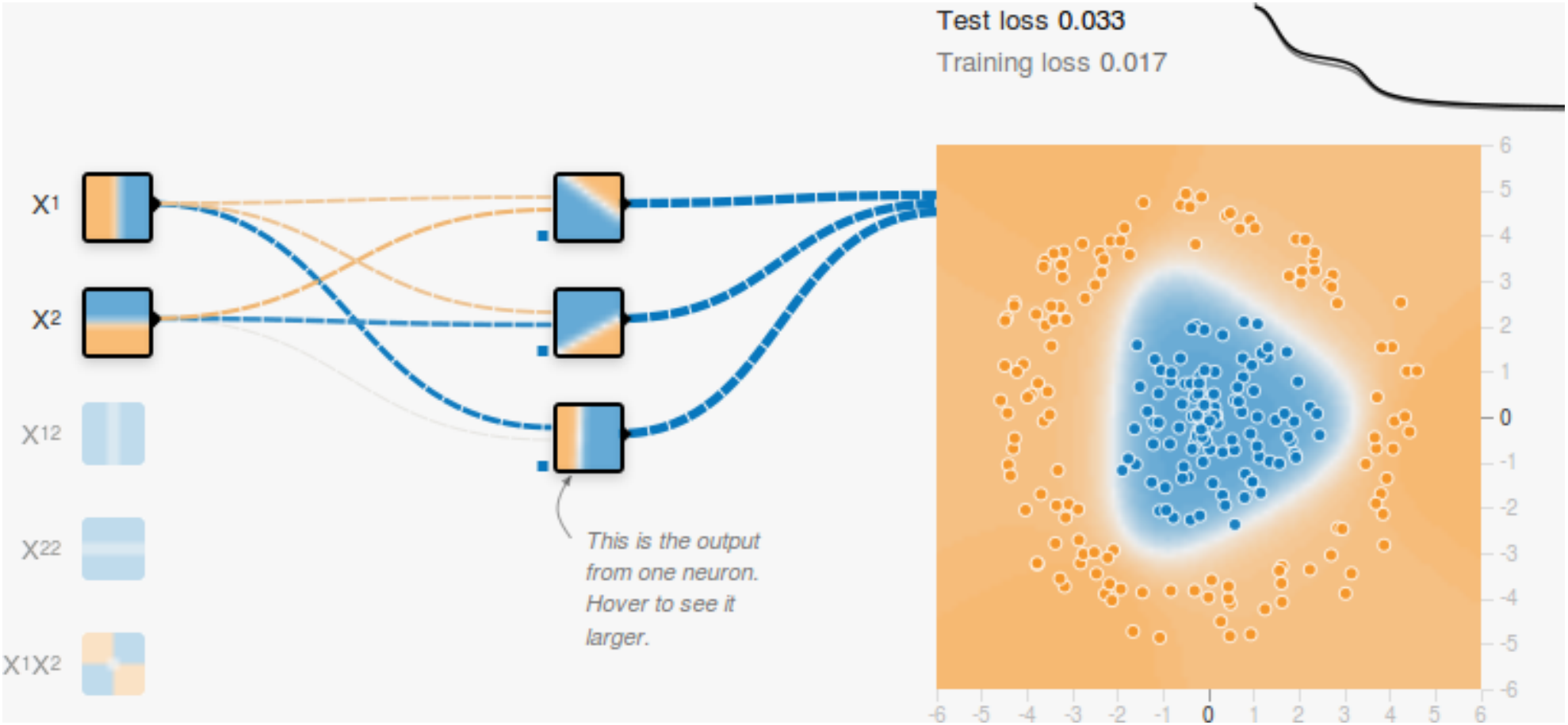




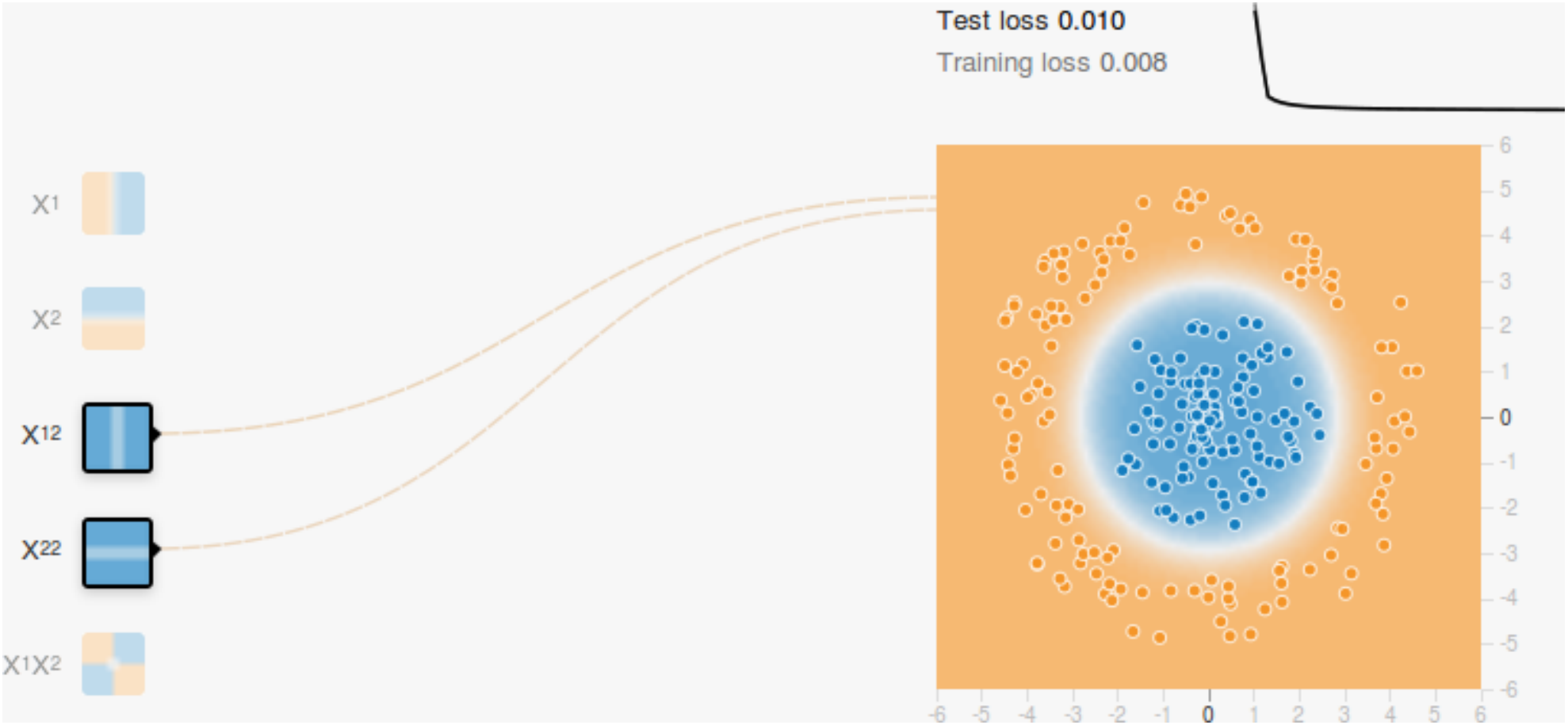
- A neural network with a single hidden layer (possibly huge) can approximate any continuous function to any precision.
- (Nothing claimed about learnability.)

`https://www.quora.com/How-can-a-deep-neural-network-with-ReLU-activations-in-its-hidden-layers-approximate-any-function`

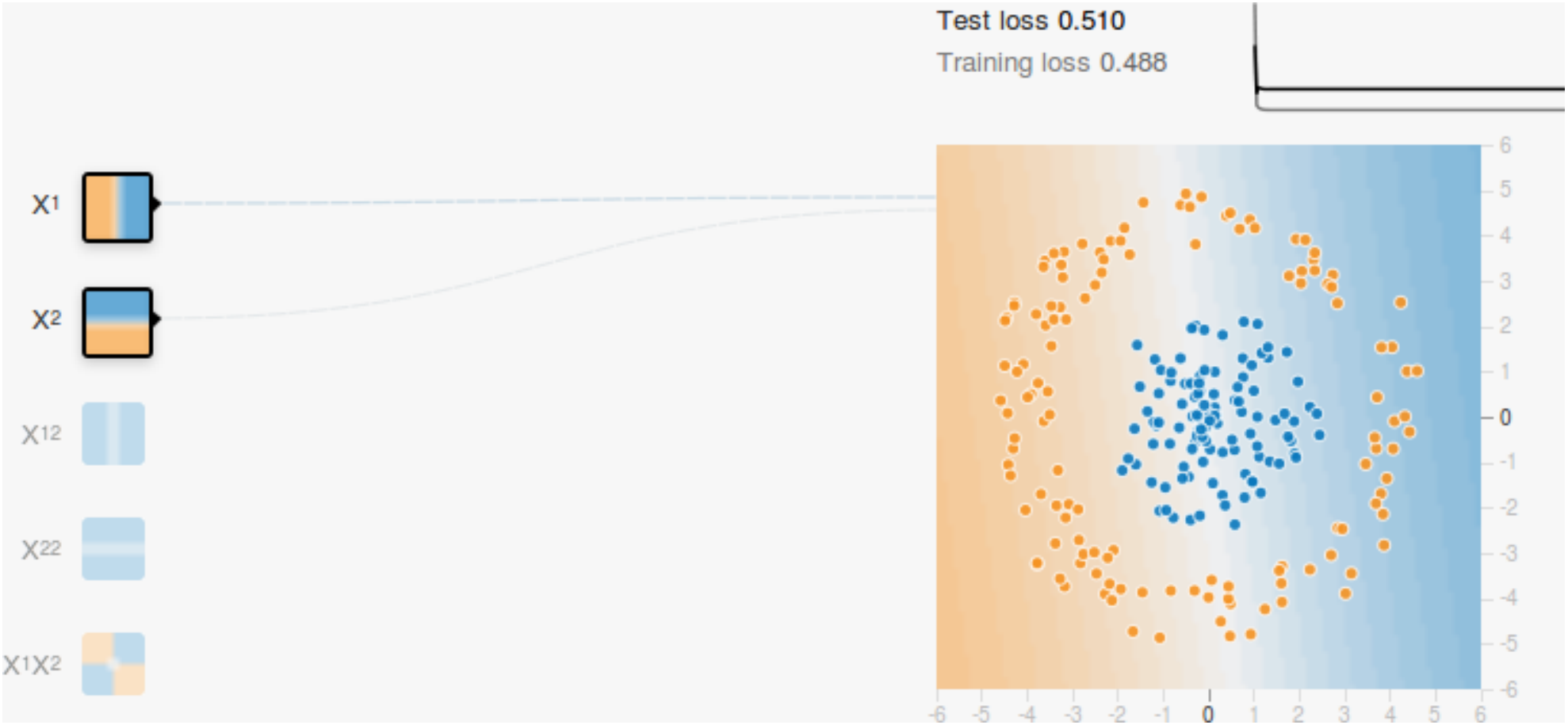
playground.tensorflow.org



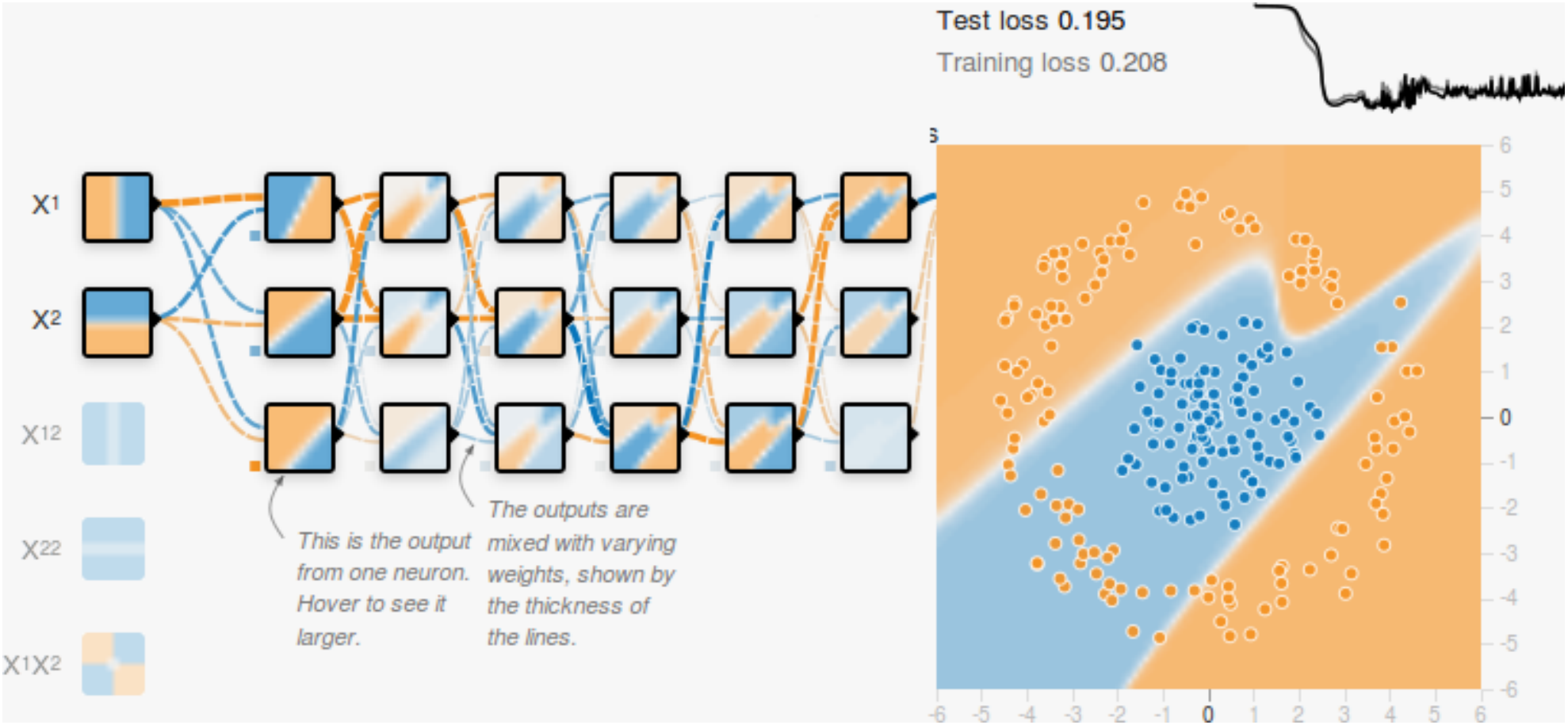
Perfect Features



Bad Features & Low Depth

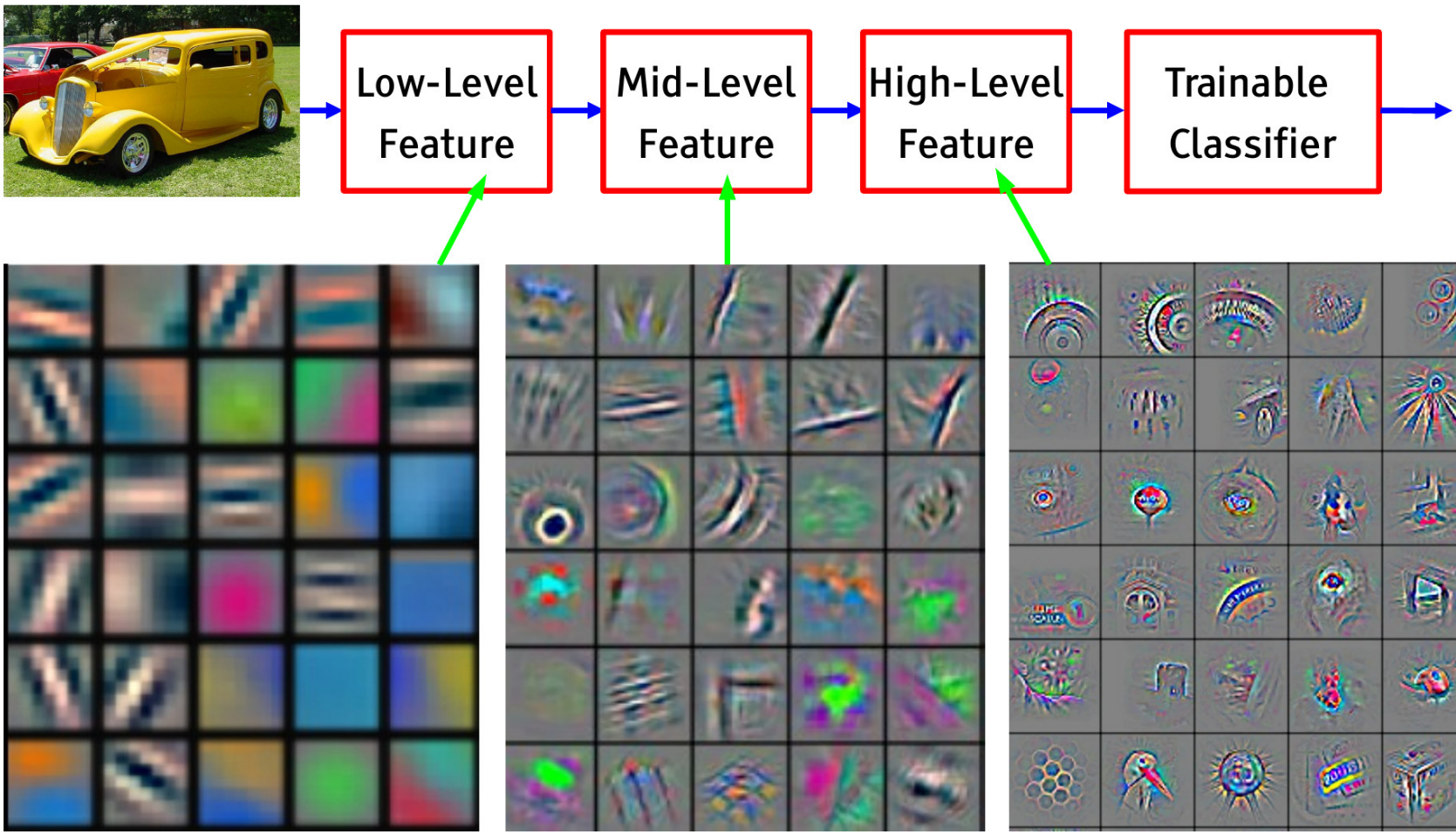


Too Complex NN Fails to Learn



Deep NNs for Image Classification

- It's deep if it has more than one stage of non-linear feature transformation



Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

Representation Learning

Representation Learning

- We saw DL finding useful features.
- We can think of these features as new coordinates.

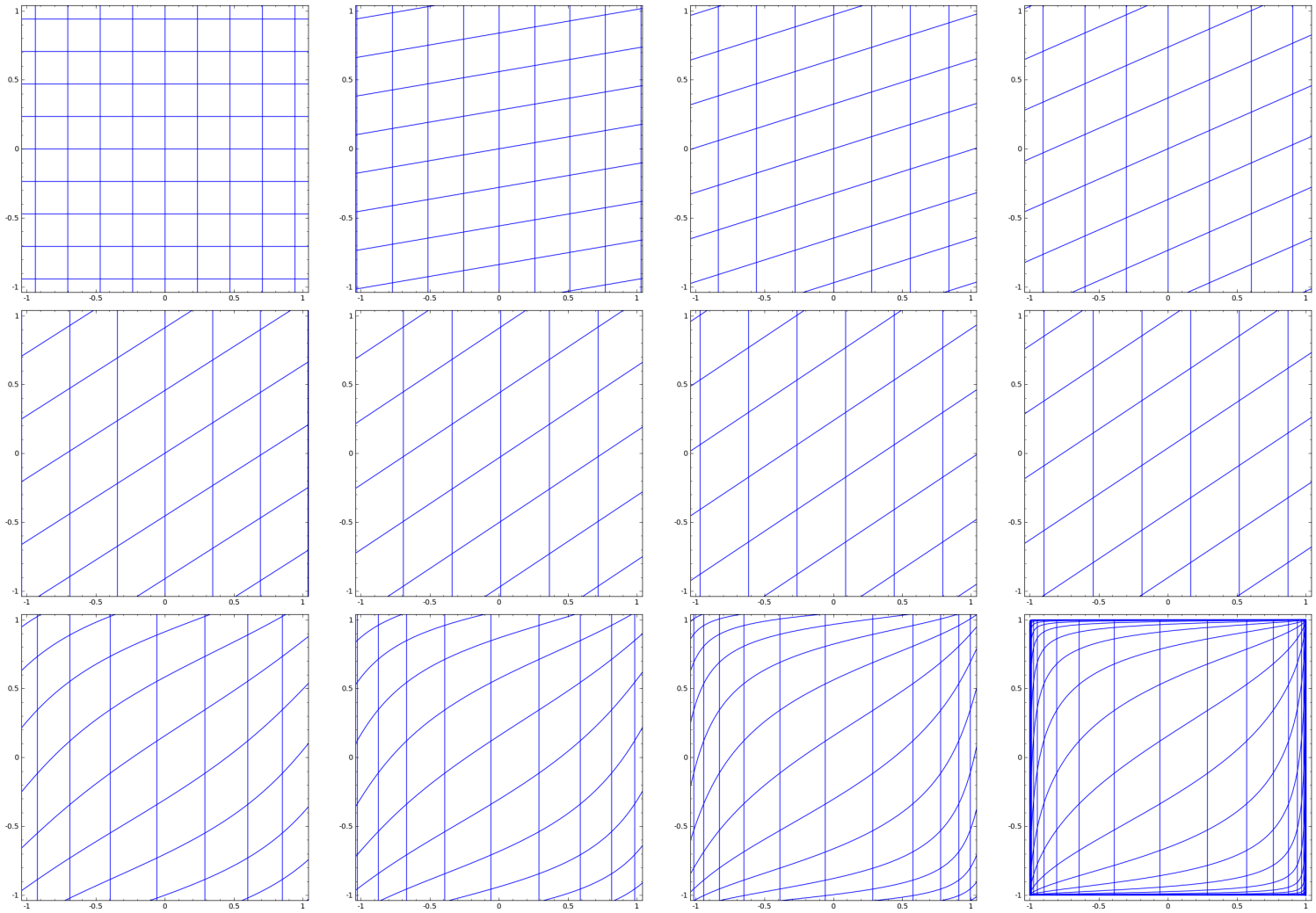
$\Rightarrow$ NNs are learning how to represent the input to make it linearly separable.

One Layer $\tanh(Wx + b)$, 2D→2D

Skew:
W

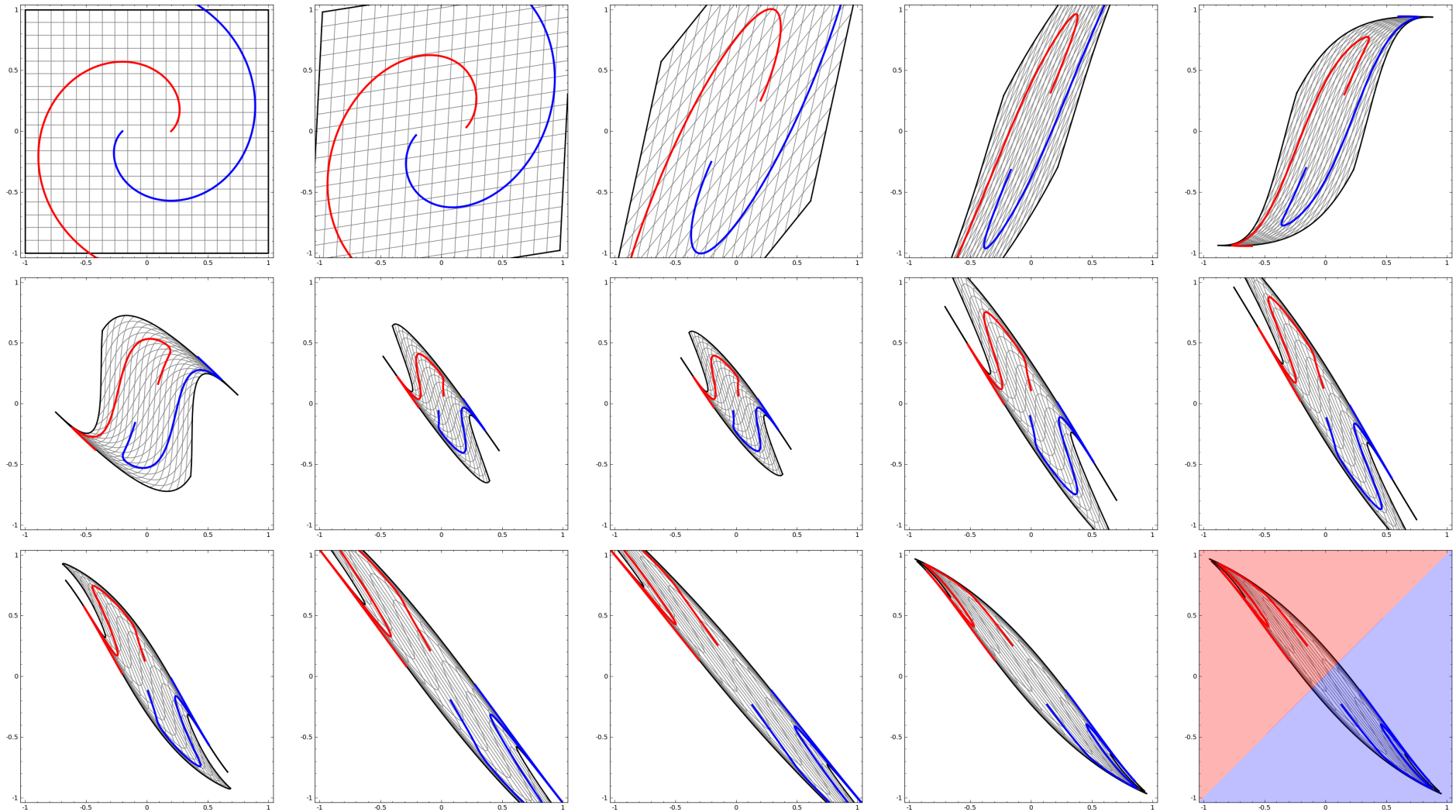
Transpose:
b

Non-lin.:
tanh



Animation by `http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/`

Four Layers, Disentagling Spirals



Animation by `http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/`

Processing Text with NNs

- Map each word to a vector of 0s and 1s ("1-hot repr."):

$$\text{cat} \mapsto (0, 0, \ldots, 0, 1, 0, \ldots, 0)$$

- Sentence is then a matrix:

		the	cat	is	on	the	mat
↑	a	0	0	0	0	0	0
	about	0	0	0	0	0	0
	...	...	...	...	...	...	...
	cat	0	**1**	0	0	0	0
	...	...	...	...	...	...	...
	is	0	0	**1**	0	0	0
	...	...	...	...	...	...	...
	on	0	0	0	**1**	0	0
	...	...	...	...	...	...	...
	the	**1**	0	0	0	**1**	0
	...	...	...	...	...	...	...
↓	zebra	0	0	0	0	0	0

Processing Text with NNs

- Map each word to a vector of 0s and 1s ("1-hot repr."):

$$\text{cat} \mapsto (0, 0, \ldots, 0, 1, 0, \ldots, 0)$$

- Sentence is then a matrix:

		the	cat	is	on	the	mat
↑	a	0	0	0	0	0	0
	about	0	0	0	0	0	0
	...	...	...	...	...	...	...
	cat	0	**1**	0	0	0	0
Vocabulary size:	...	...	...	...	...	...	...
1.3M English	is	0	0	**1**	0	0	0
2.2M Czech	...	...	...	...	...	...	...
	on	0	0	0	**1**	0	0
	...	...	...	...	...	...	...
	the	**1**	0	0	0	**1**	0
	...	...	...	...	...	...	...
↓	zebra	0	0	0	0	0	0

Processing Text with NNs

- Map each word to a vector of 0s and 1s ("1-hot repr."):

$$\text{cat} \mapsto (0, 0, \ldots, 0, 1, 0, \ldots, 0)$$

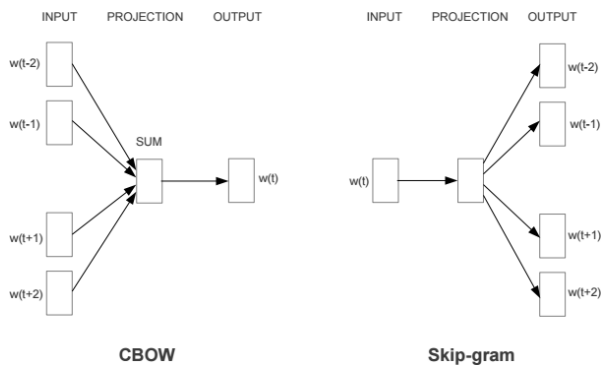
- Sentence is then a matrix:

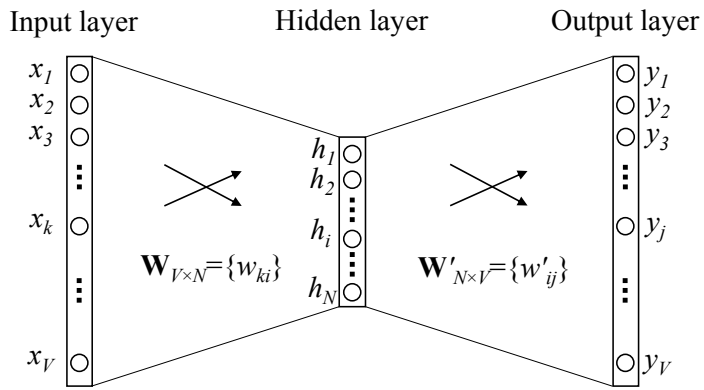
		the	cat	is	on	the	mat
↑	a	0	0	0	0	0	0
	about	0	0	0	0	0	0
	...	...	...	...	...	...	...
	cat	0	**1**	0	0	0	0
Vocabulary size:	...	...	...	...	...	...	...
1.3M English	is	0	0	**1**	0	0	0
2.2M Czech	...	...	...	...	...	...	...
	on	0	0	0	**1**	0	0
	...	...	...	...	...	...	...
	the	**1**	0	0	0	**1**	0
	...	...	...	...	...	...	...
↓	zebra	0	0	0	0	0	0

Main drawback: No relations, all words equally close/far.

Word Embeddings

- Map each word to a dense vector.
- In practice 300–2000 dimensions are used, not 1–2M.
  - The dimensions have no clear interpretation.
- Embeddings are trained for each particular task.
  - NNs: The matrix that maps 1-hot input to the first layer.
- The famous word2vec (Mikolov et al., 2013):
  - CBOW: Predict the word from its four neighbours.
  - Skip-gram: Predict likely neighbours given the word.



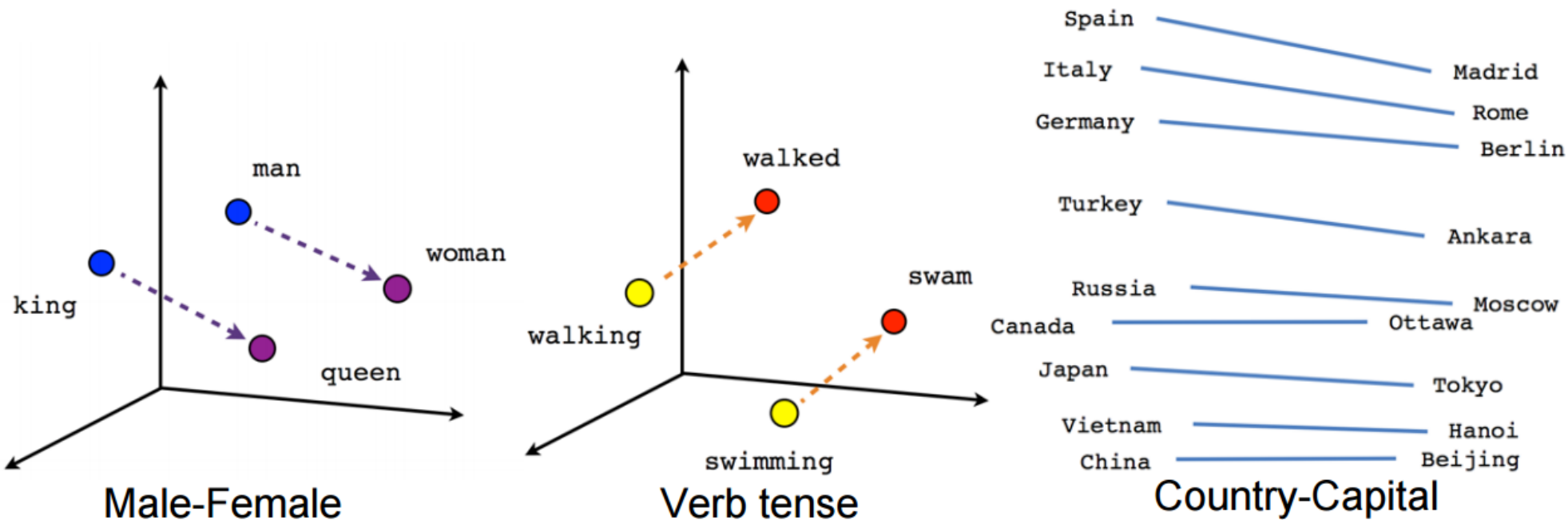


Right: CBOW with just a single-word context
(`http://www-personal.umich.edu/~ronxin/pdf/w2vexp.pdf`)

Continuous Space of Words

Word2vec embeddings show interesting properties:

$$v(\text{king}) - v(\text{man}) + v(\text{woman}) \approx v(\text{queen}) \tag{5}$$



Illustrations from `https://www.tensorflow.org/tutorials/word2vec`

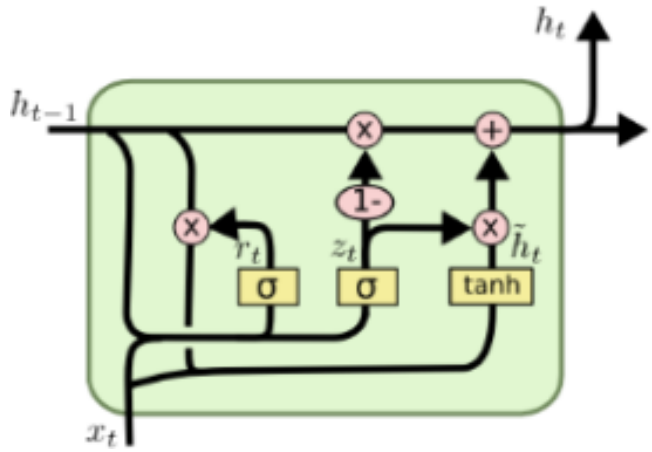
Variable-Length Inputs

Variable-length input can be handled by recurrent NNs:

- Reading one input symbol at a time.
  - The same (trained) transformation used every time.
- Unroll in time (up to a fixed length limit).

Tricks needed to train (to avoid "vanishing gradients"):

- Dropout.
- LSTM (Long Short-Term Memory Cells, Hochreiter and Schmidhuber (1997)).
- GRU (Gated Recurrent Unit Cells, Chung et al. (2014)).



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

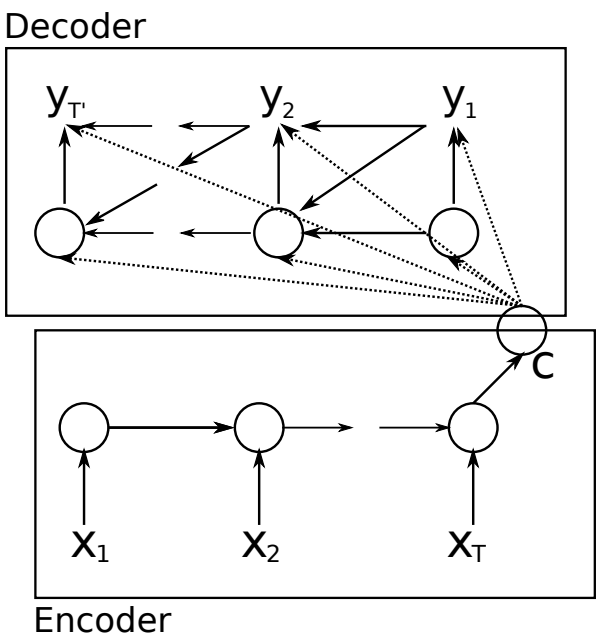
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

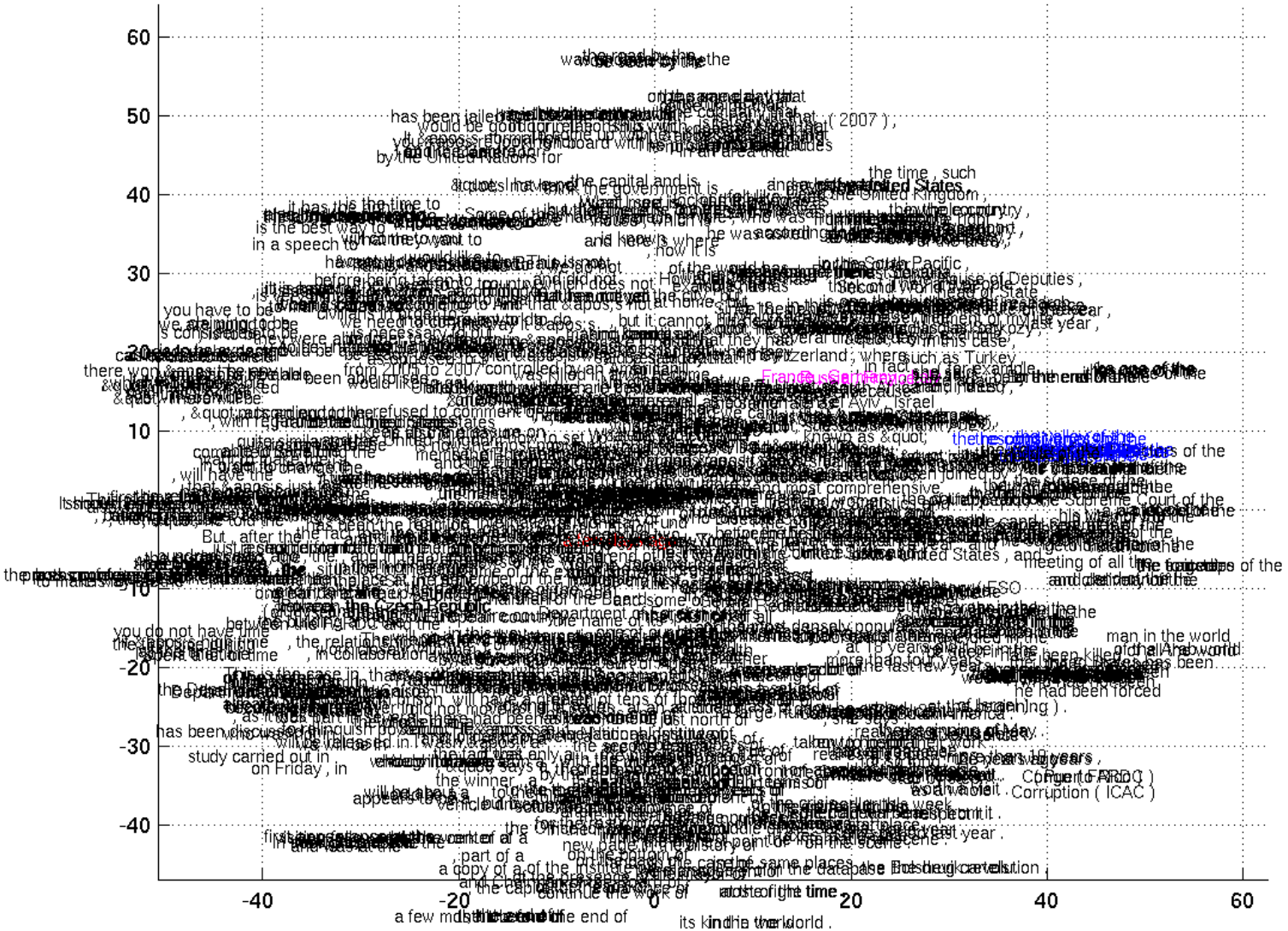
NNs as Translation Model in SMT

Cho et al. (2014) propose:

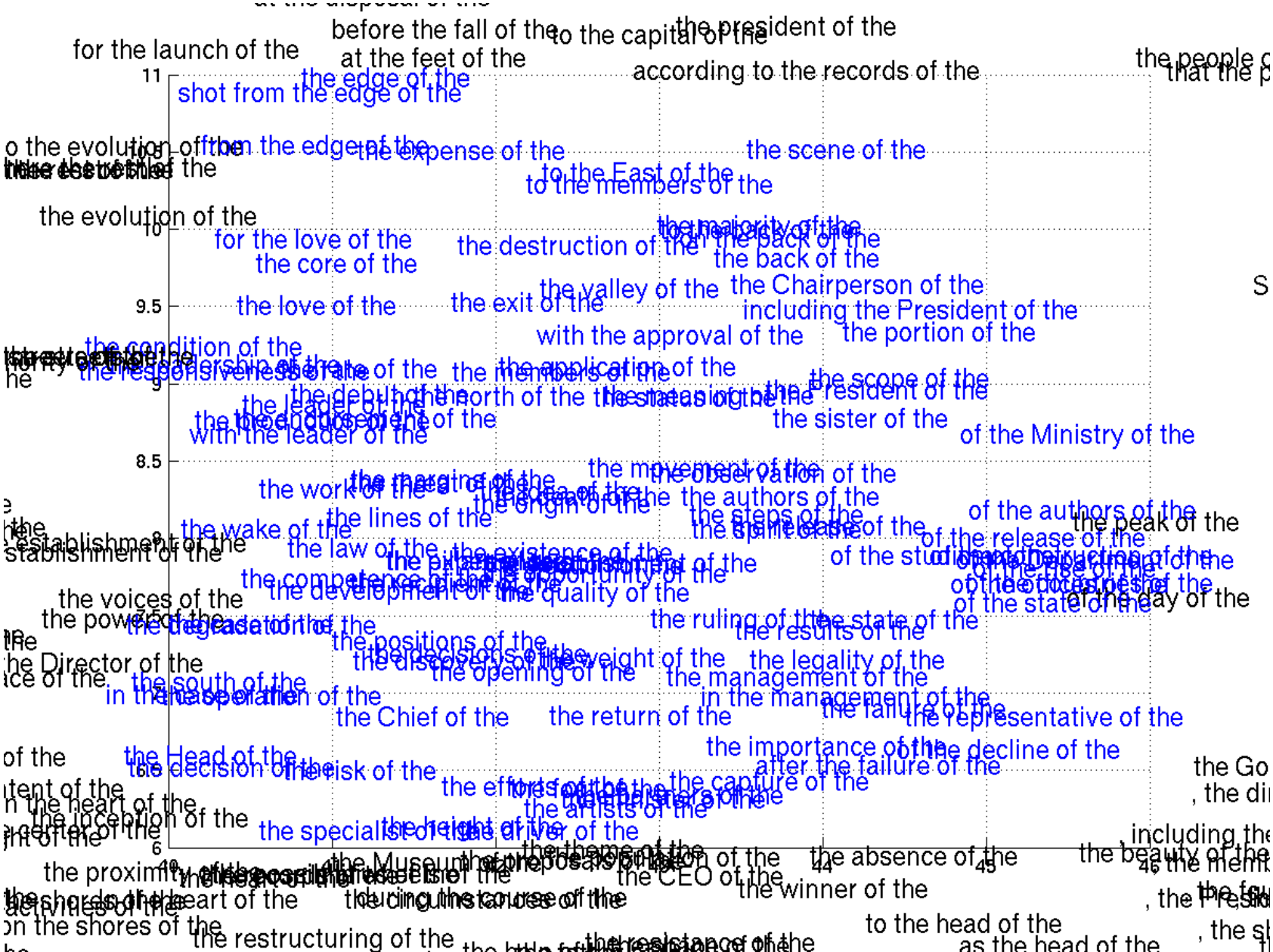
- encoder-decoder architecture and
- GRU unit (name given later by Chung et al. (2014))
- to score variable-length phrase pairs in PBMT.



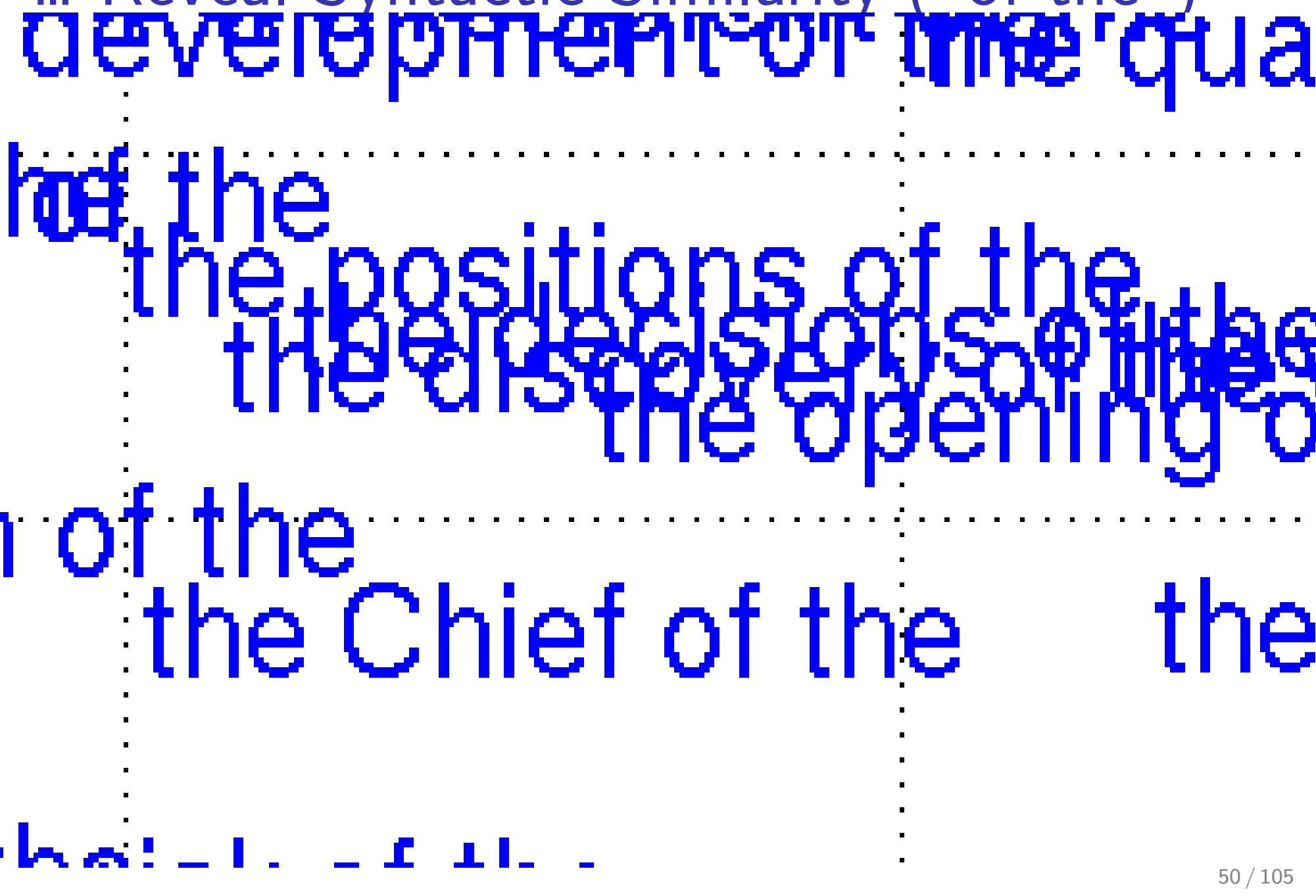
⇒ Embeddings of Phrases



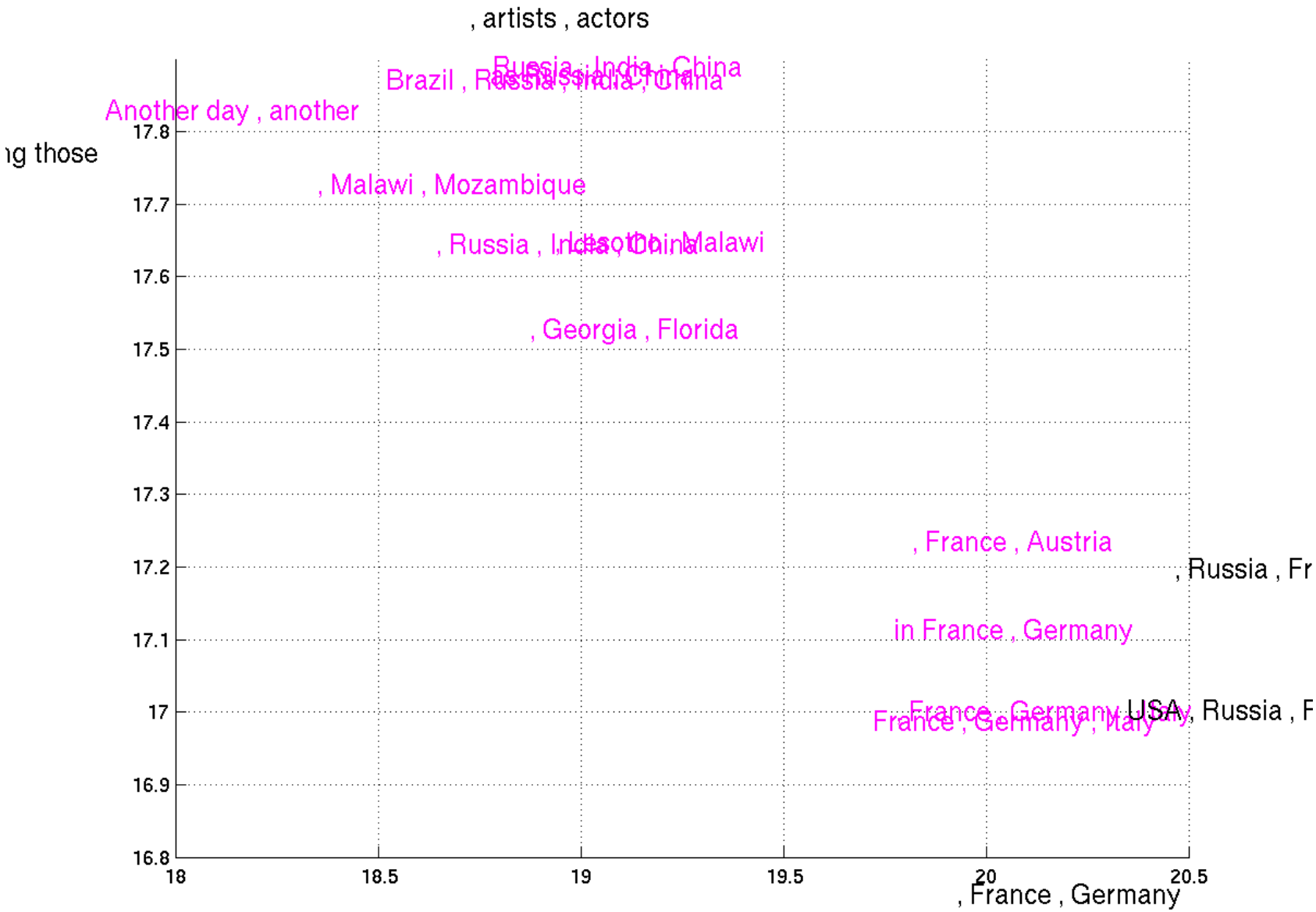
... Reveal Syntactic Similarity ("of the")



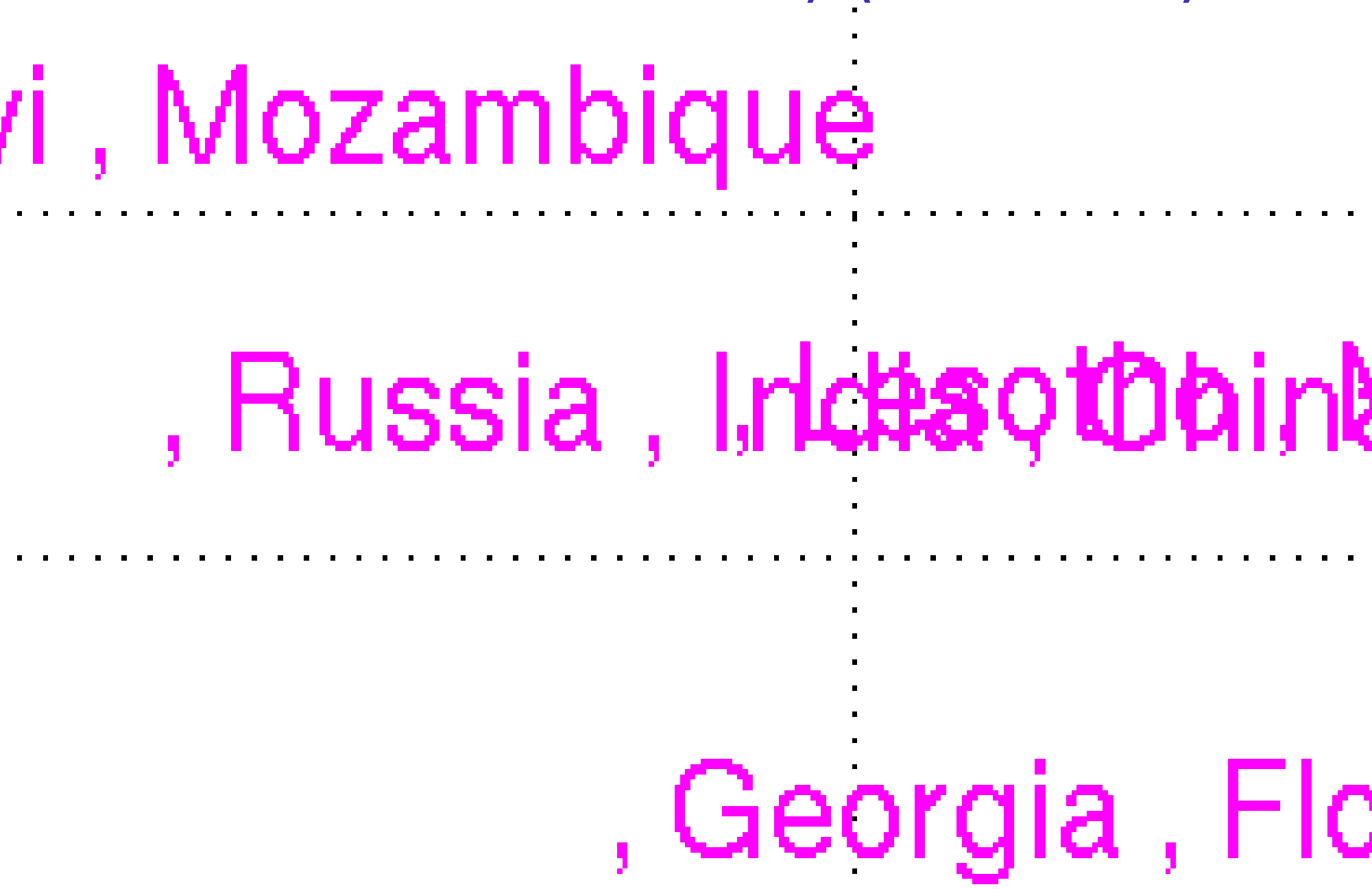
... Reveal Syntactic Similarity ("of the")



... and Semantic Similarity (Countries)



... and Semantic Similarity (Countries)

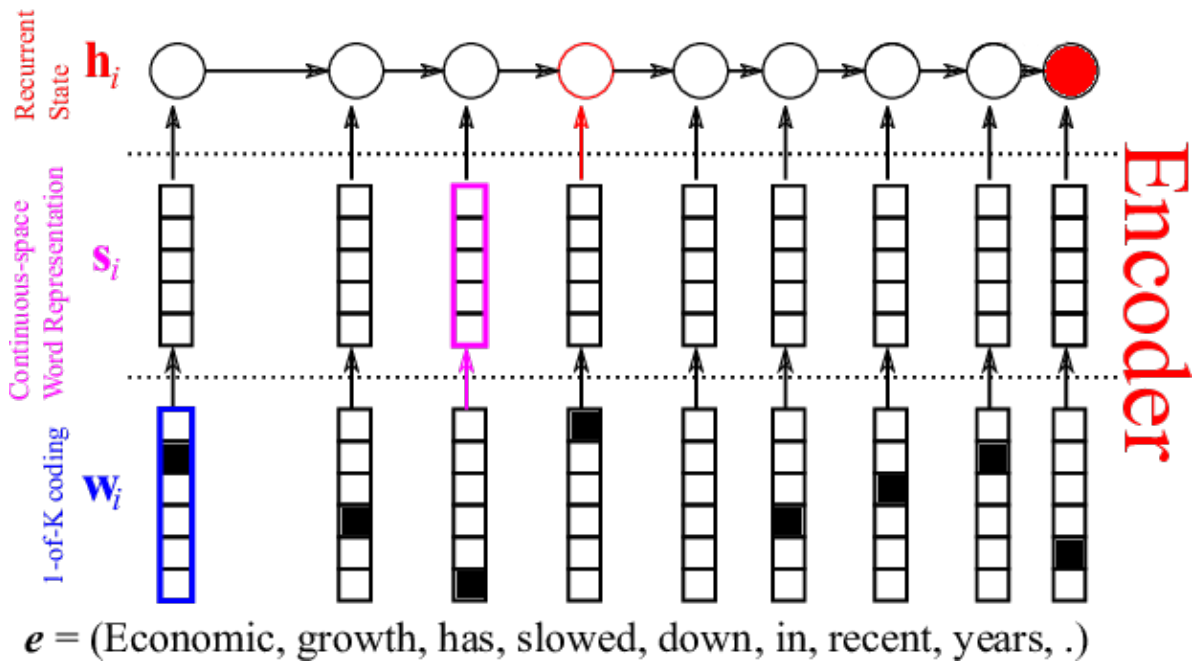


NMT: Sequence to Sequence

Sutskever et al. (2014) use:

- LSTM RNN encoder-decoder
- to consume
  and *produce* variable-length sentences.

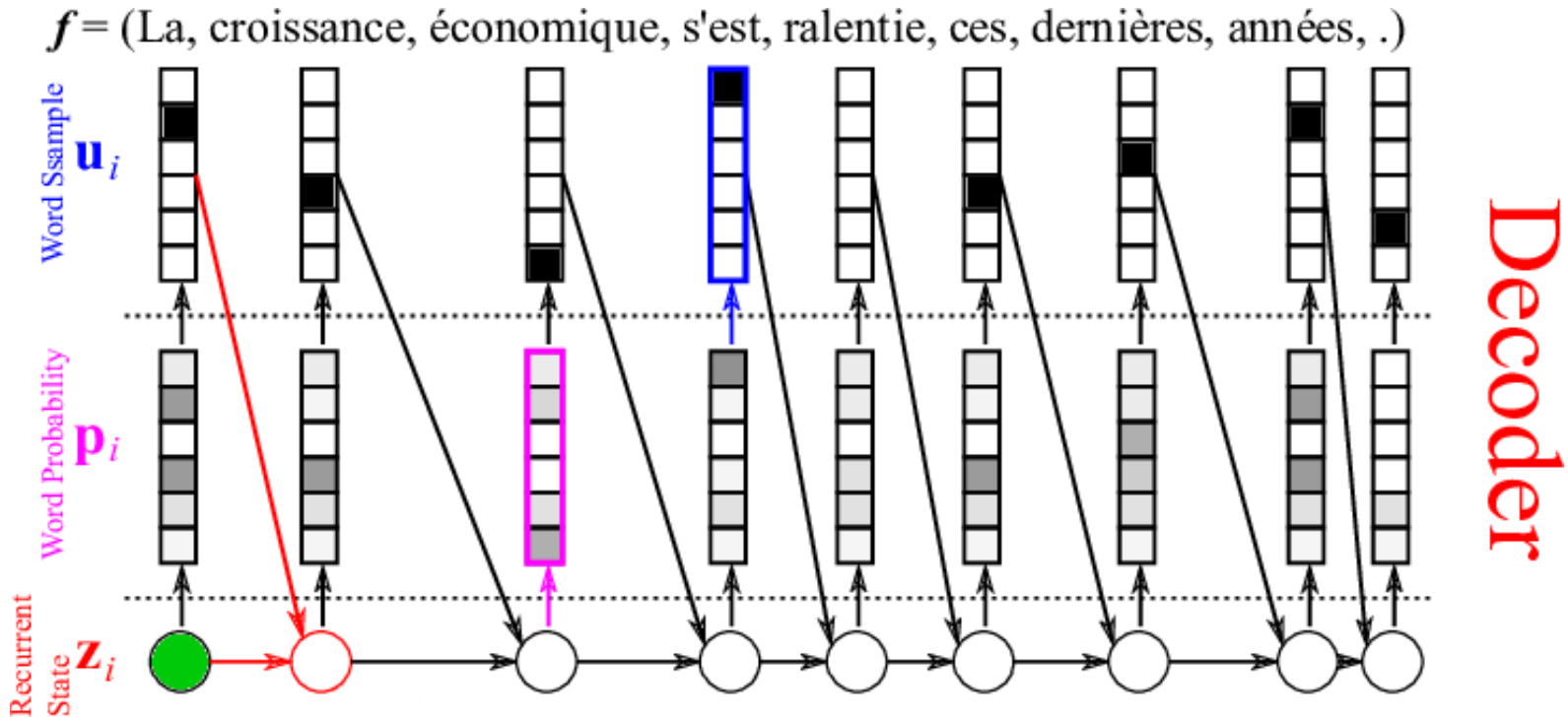
First the Encoder:



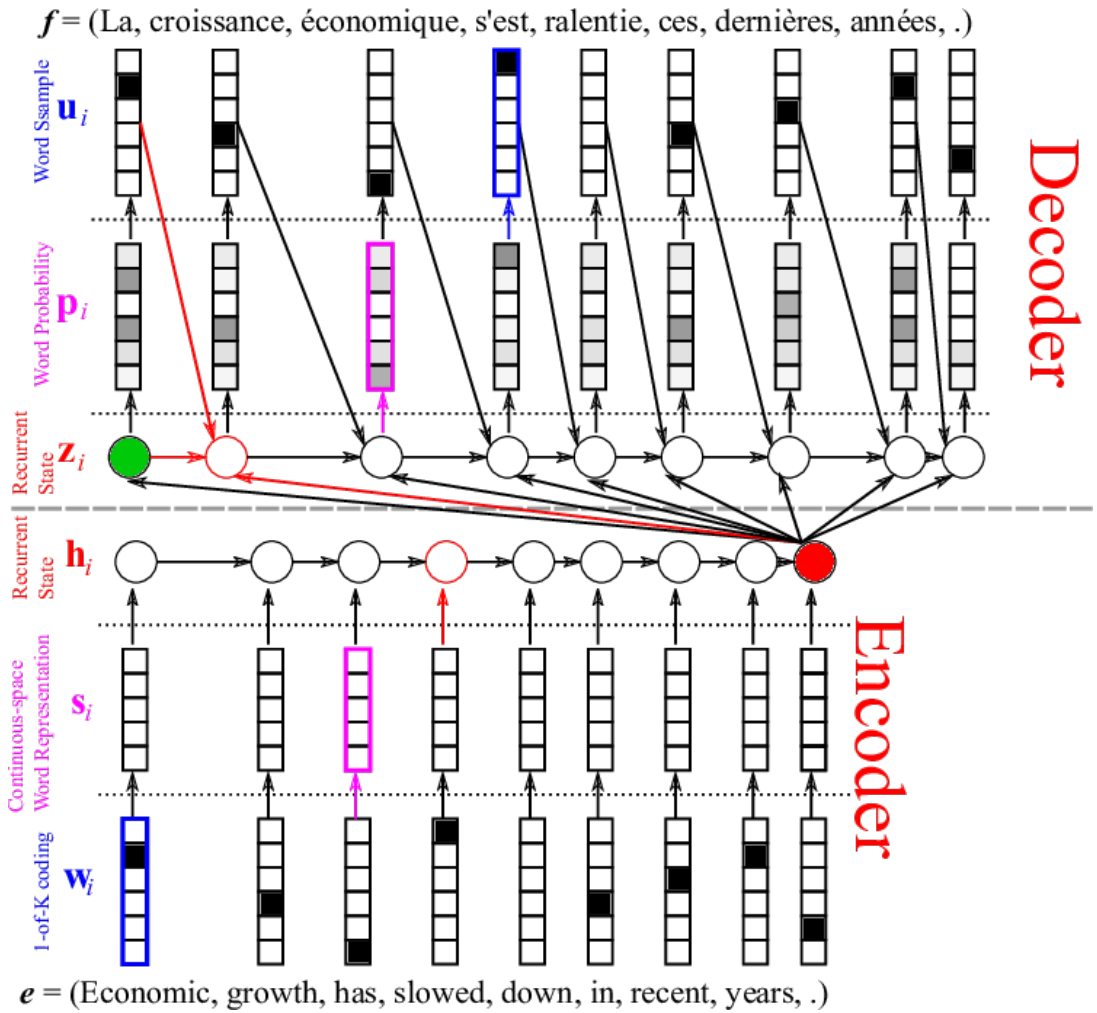
Then the Decoder

Remember: $p(e_1^I|f_1^J) = p(e_1|f_1^J) \cdot p(e_2|e_1, f_1^J) \cdot p(e_3|e_2, e_1, f_1^J) \ldots$

- Again RNN, producing one word at a time.
- The produced word fed back into the network.
  - (Word embeddings in the target language used here.)

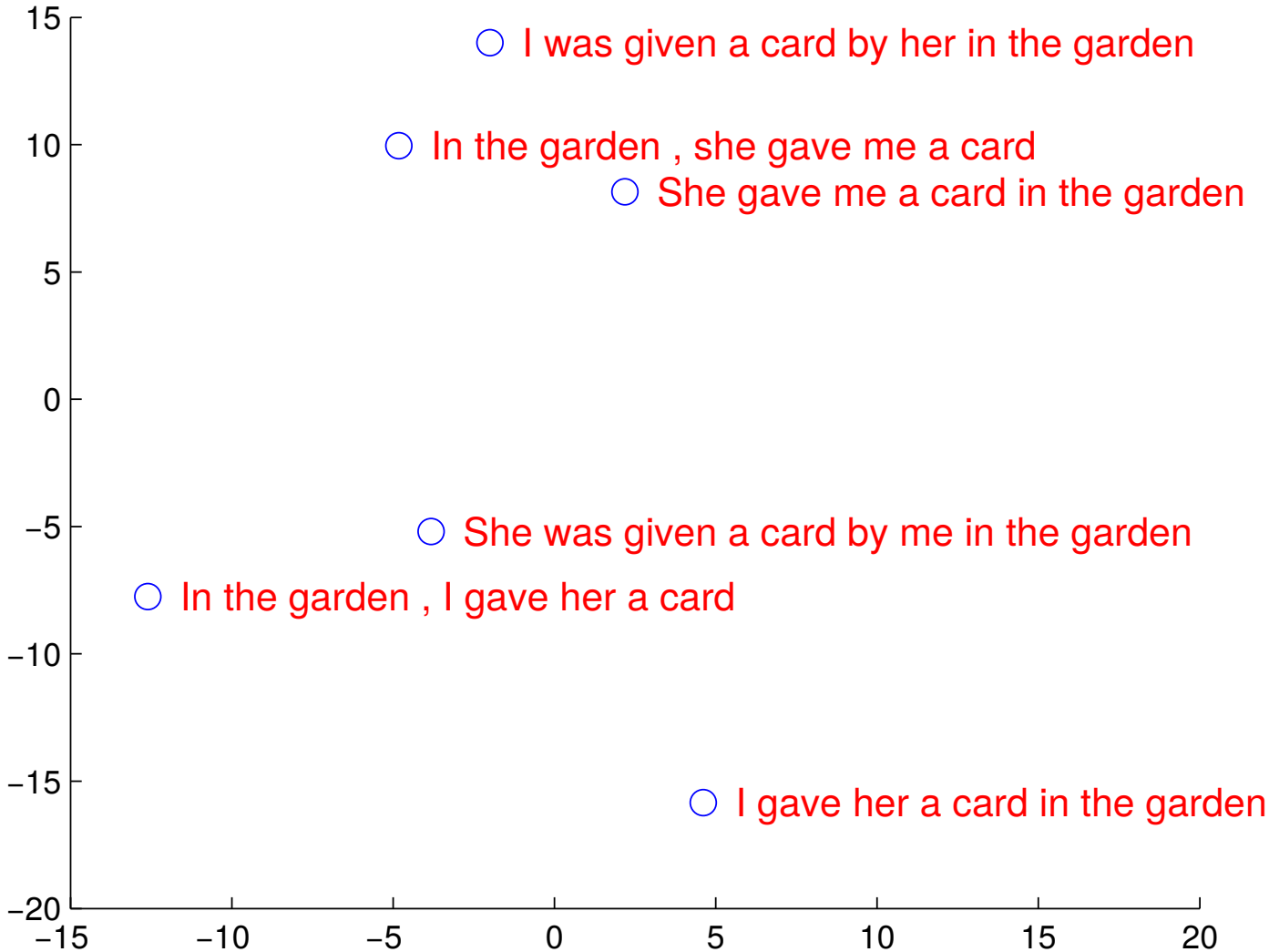


Encoder-Decoder Architecture



https://devblogs.nvidia.com/parallelforall/introduction-neural-machine-translation-gpus-part-2/

Continuous Space of Sentences

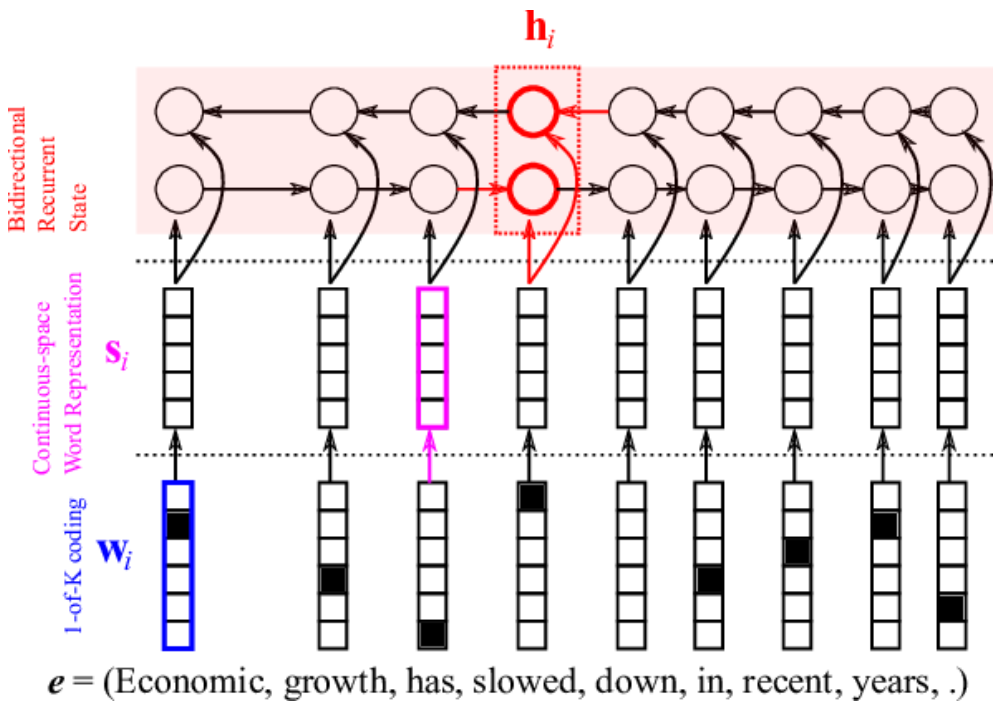


2-D PCA projection of 8000-D space representing sentences (Sutskever et al., 2014).

Attention (1/2)

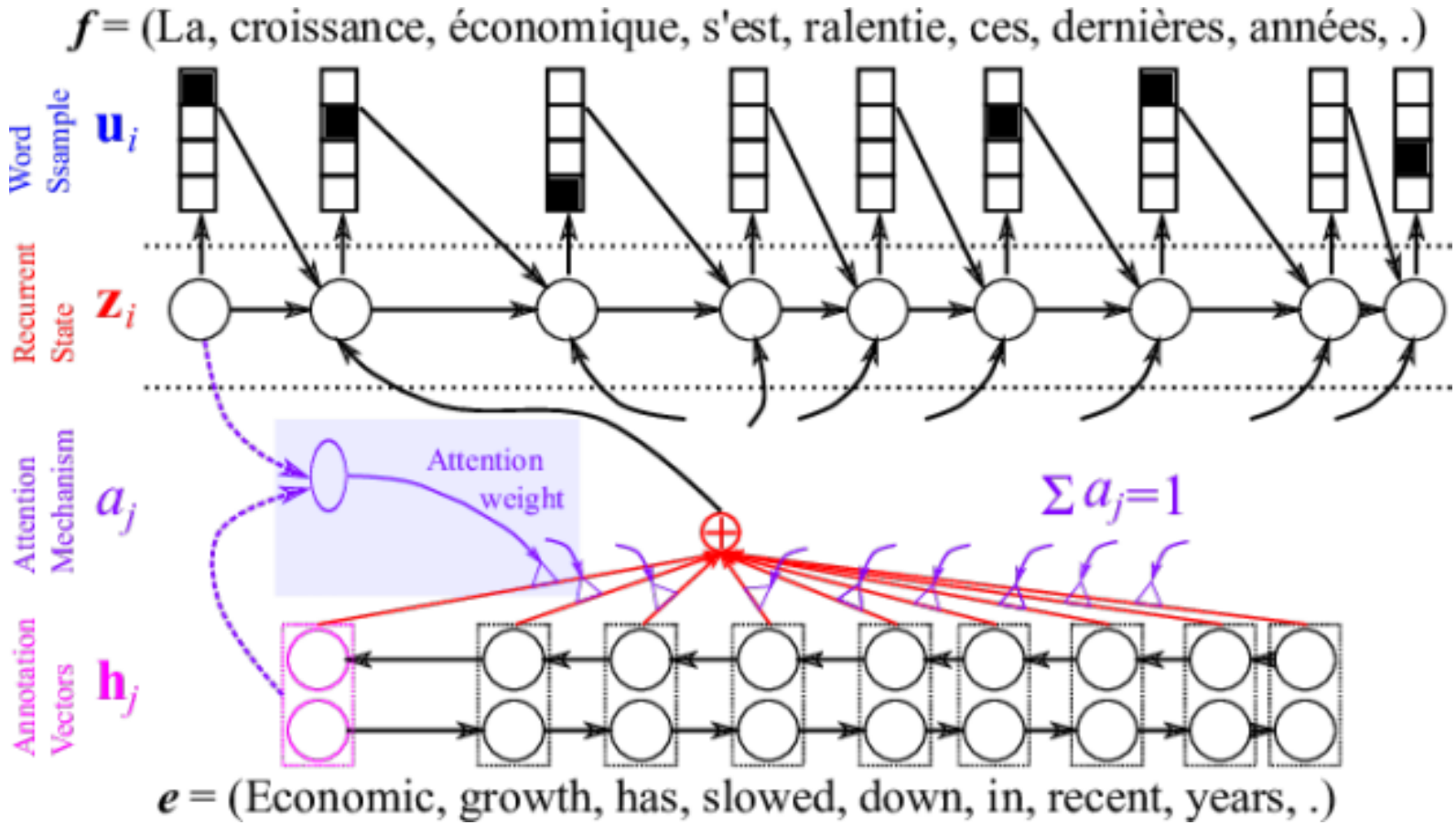
- Arbitrary-length sentences fit badly into a fixed vector.
- Reading input *backward* works better.
  - … because early words will be more salient.

$\Rightarrow$ Use Bi-directional RNN and "attend" to all states h_i.



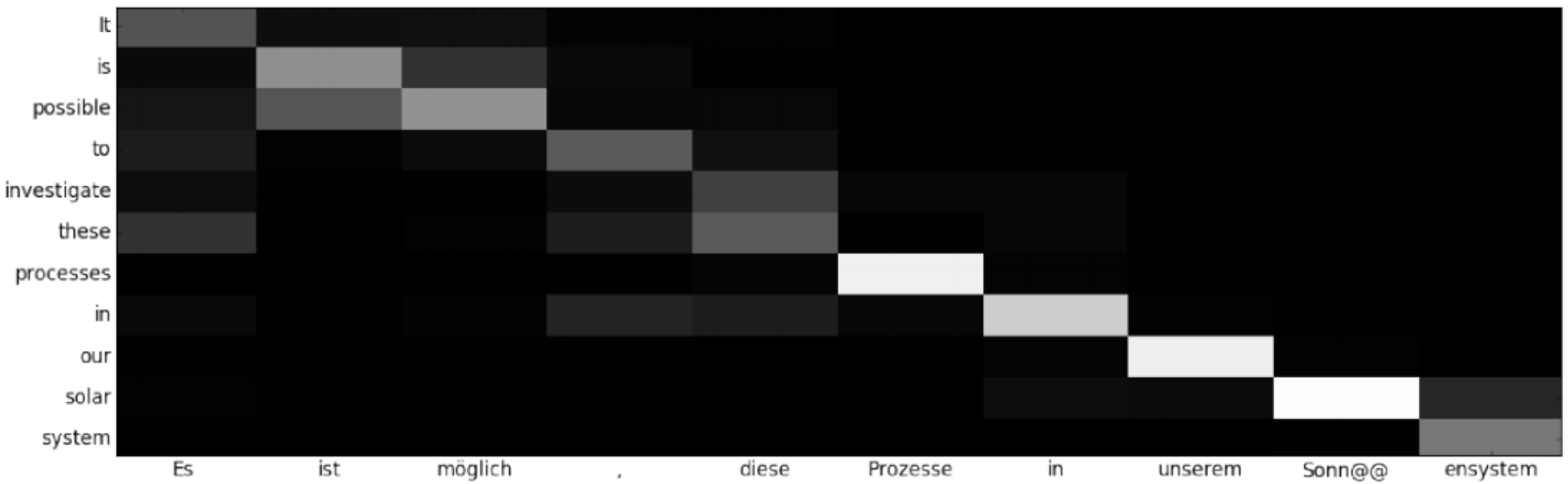
Attention (2/2)

- Add a sub-network predicing importance of source states at each step.



Attention ≈ Alignment

- We can collect the attention across time.
- Each column corresponds to one decoder time step.
- Source tokens correspond to rows.



Ultimate Goal of SMT vs. NMT

Goal of "classical" SMT:

Find **minimum translation units** $\sim$ graph partitions:

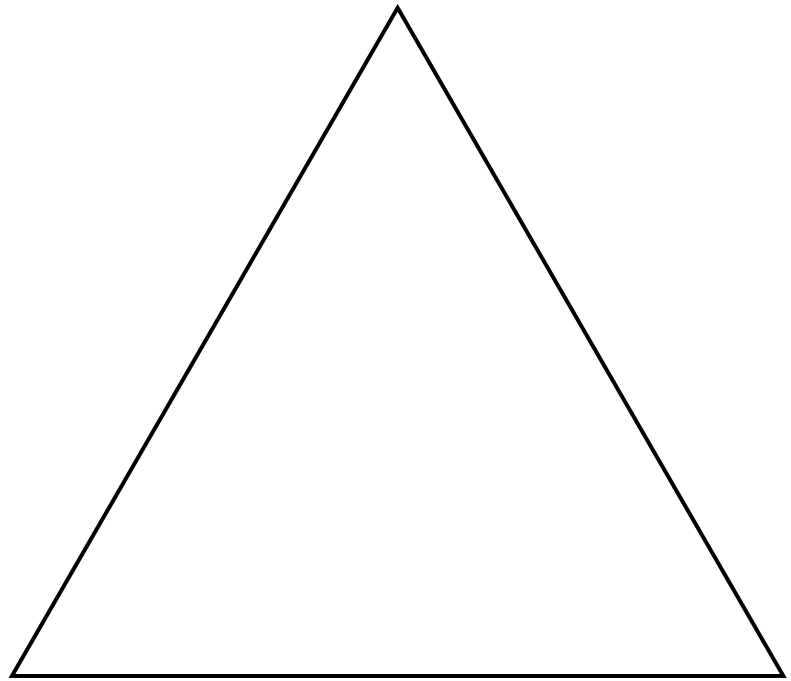
- such that they are frequent across many sentence pairs.
- without imposing (too hard) constraints on reordering.
- in an unsupervised fashion.

Goal of neural MT:

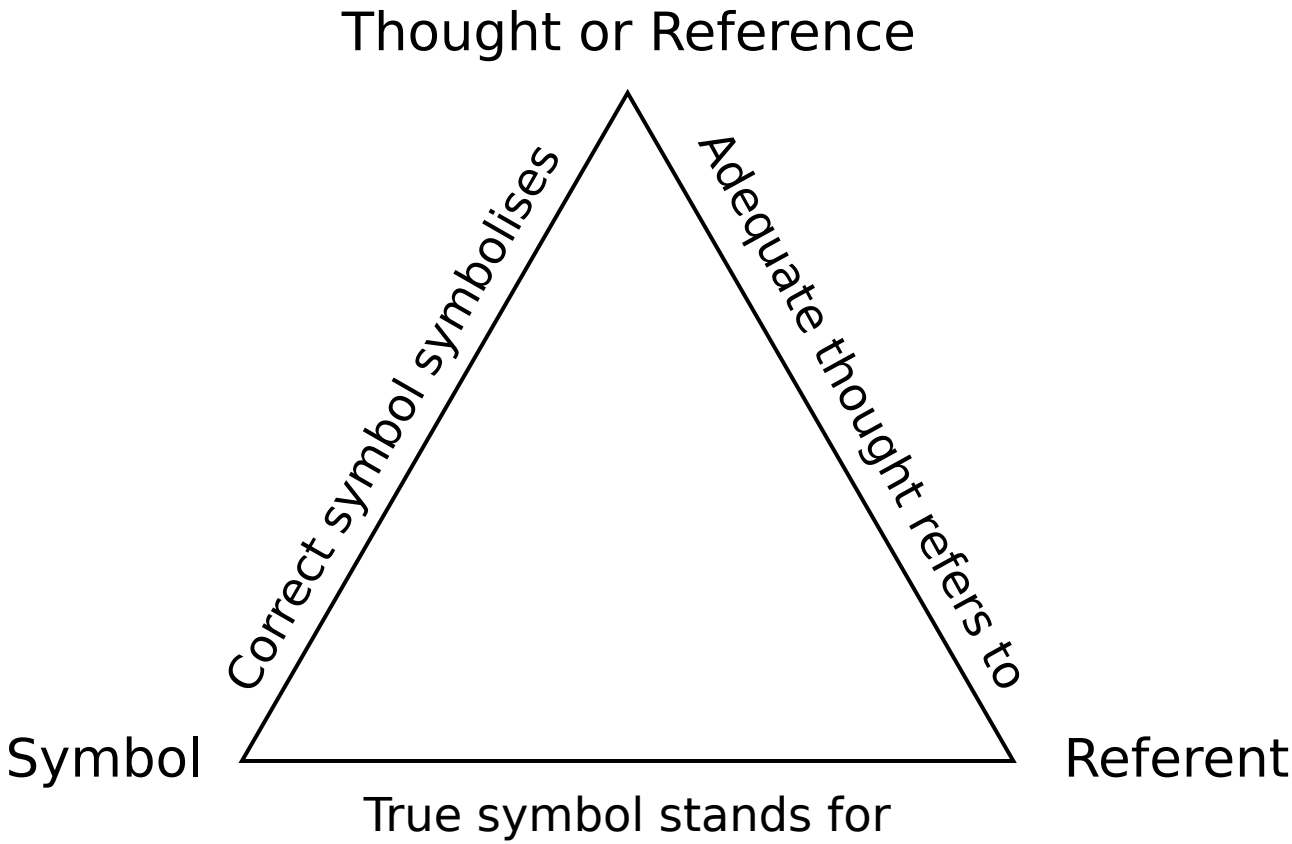
Avoid minimum translation units.

Come up with a network architecture that:

- Reads input in as original form as possible.
- Produces output in as final form as possible.
- Can be optimized end-to-end *in practice*.



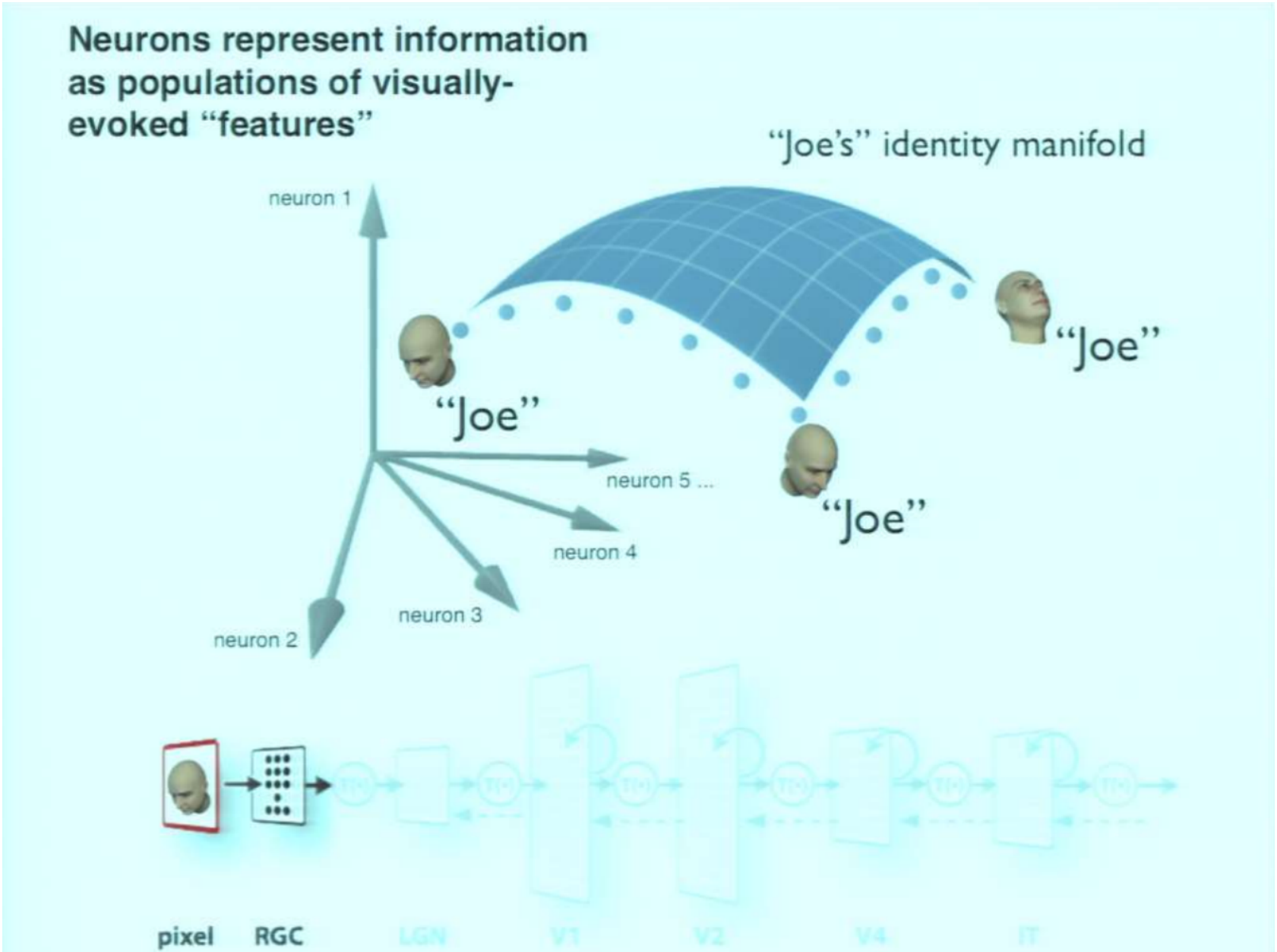
Semiotic Triangle by Ogden and Richards



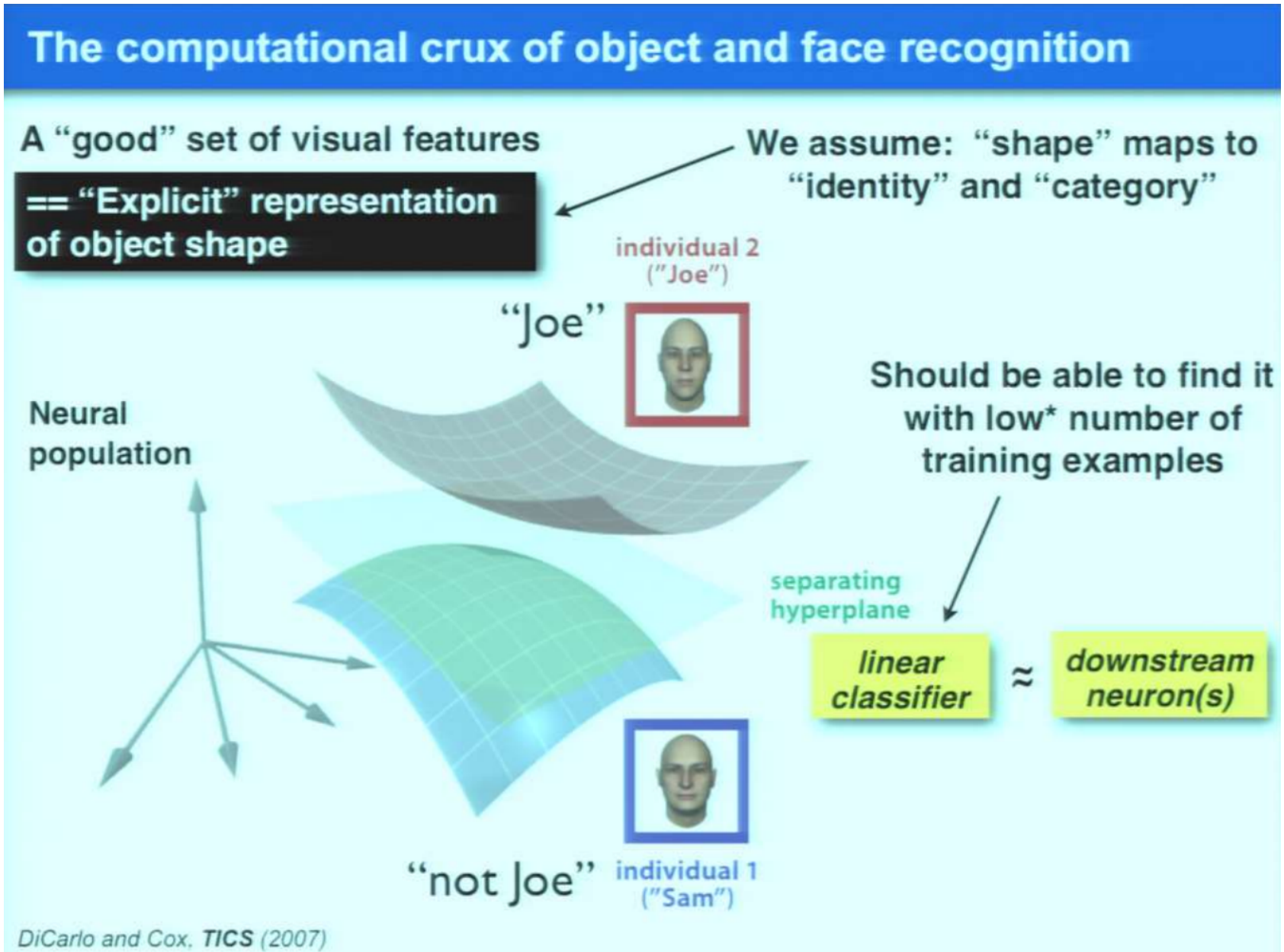
Manifold Hypothesis

- Related objects are close to each other (in a sensible representation).
  ... They live on a manifold.
- NNs are learning space transformations
  to disentagle manifolds.

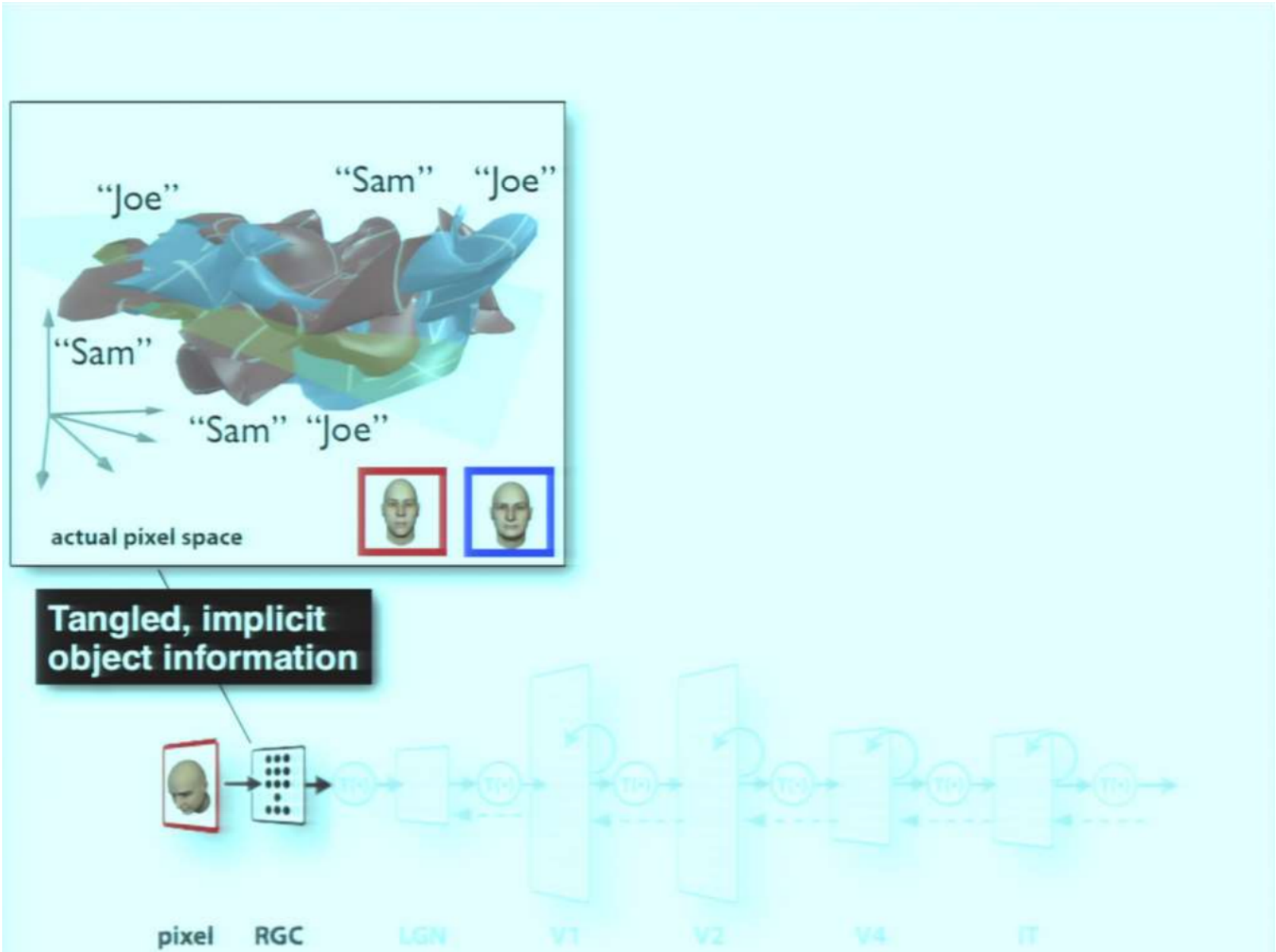
Example from Face Recognition



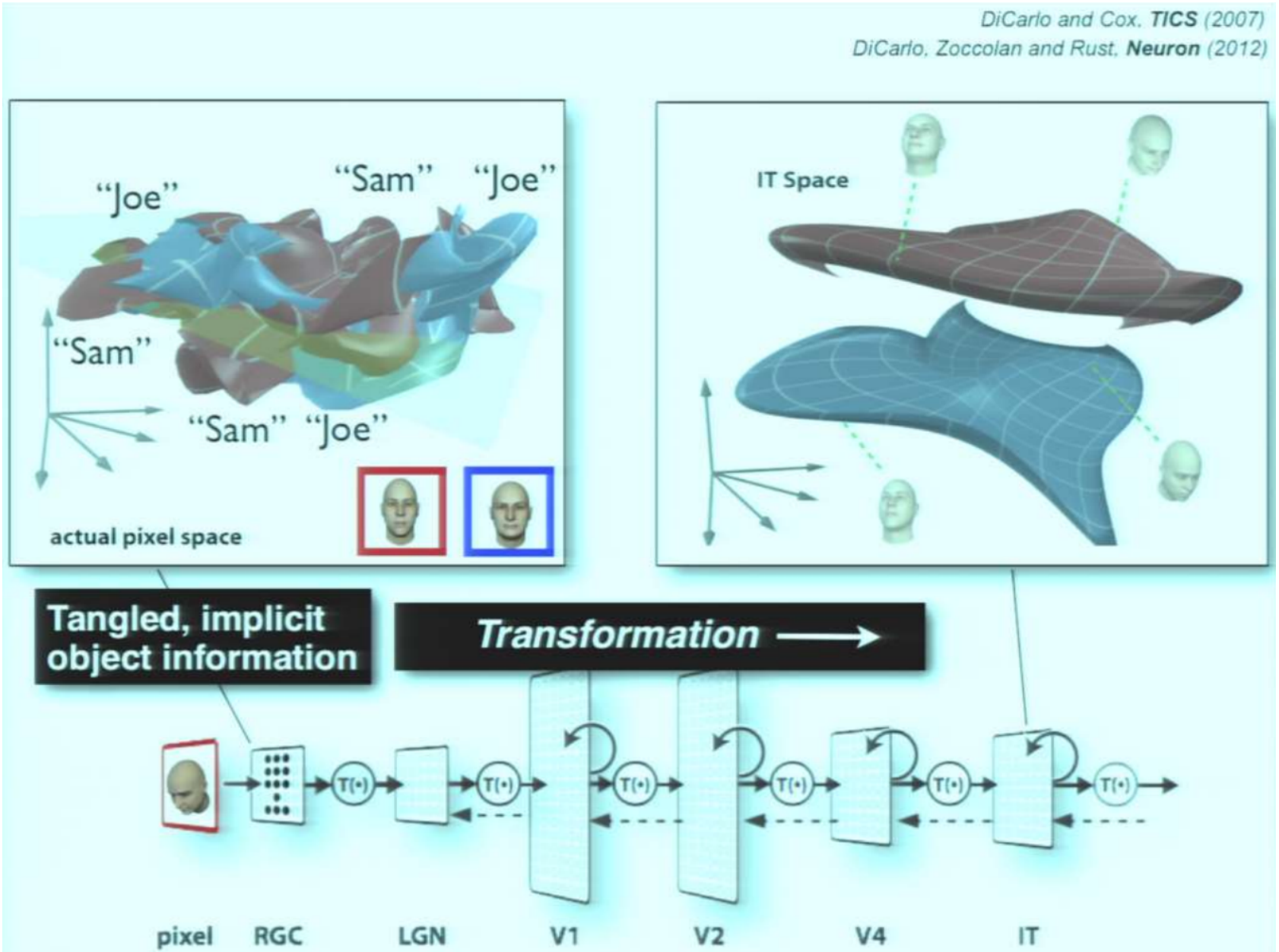
Example from Face Recognition



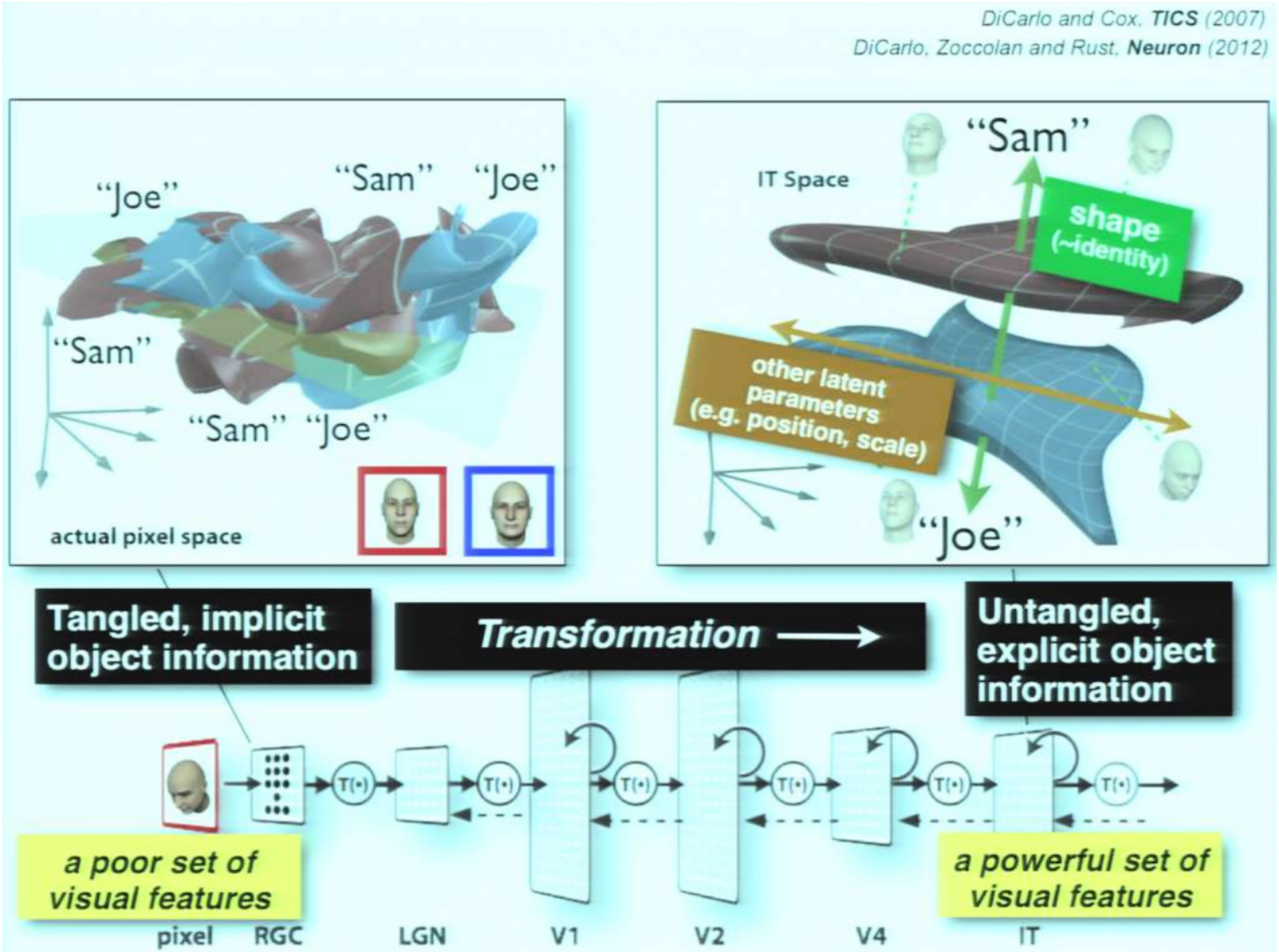
Example from Face Recognition



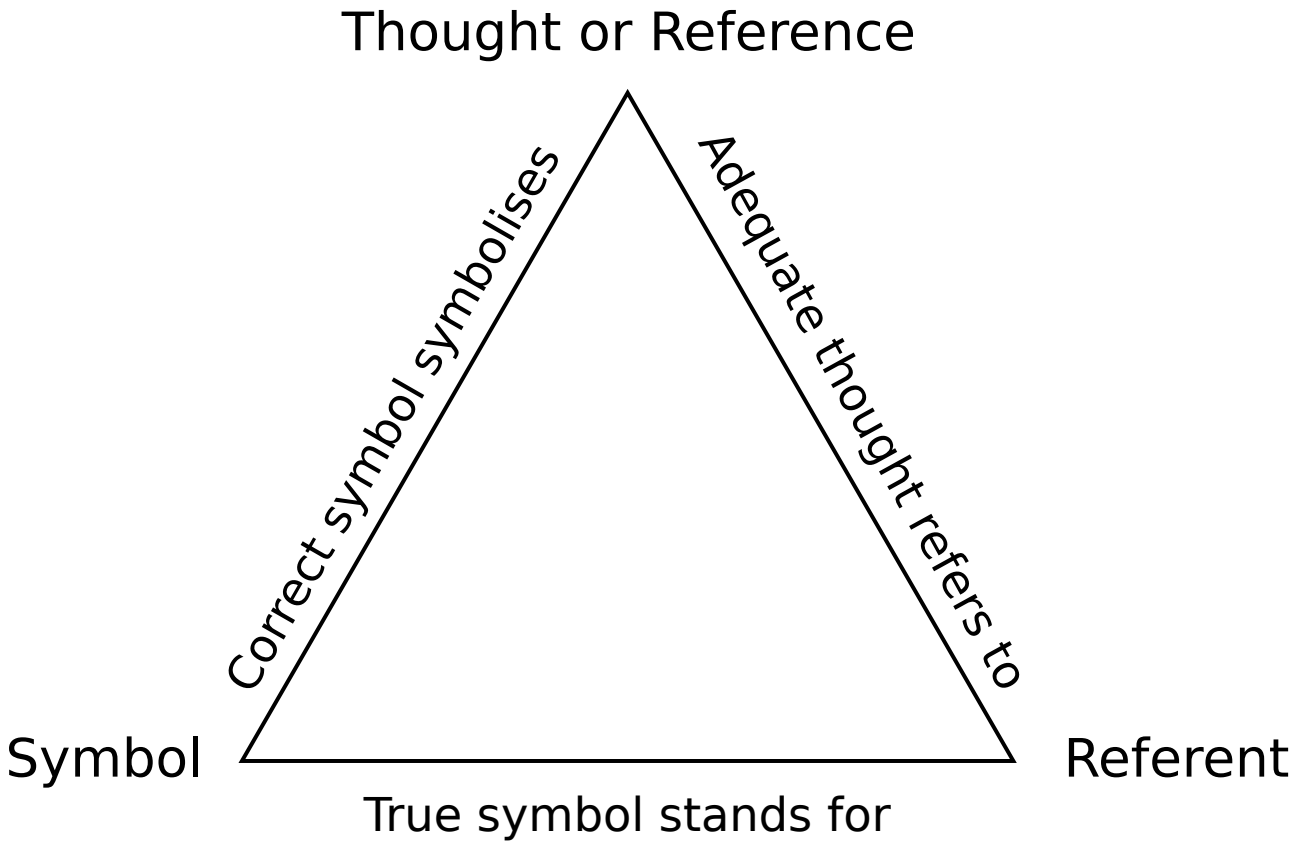
Example from Face Recognition



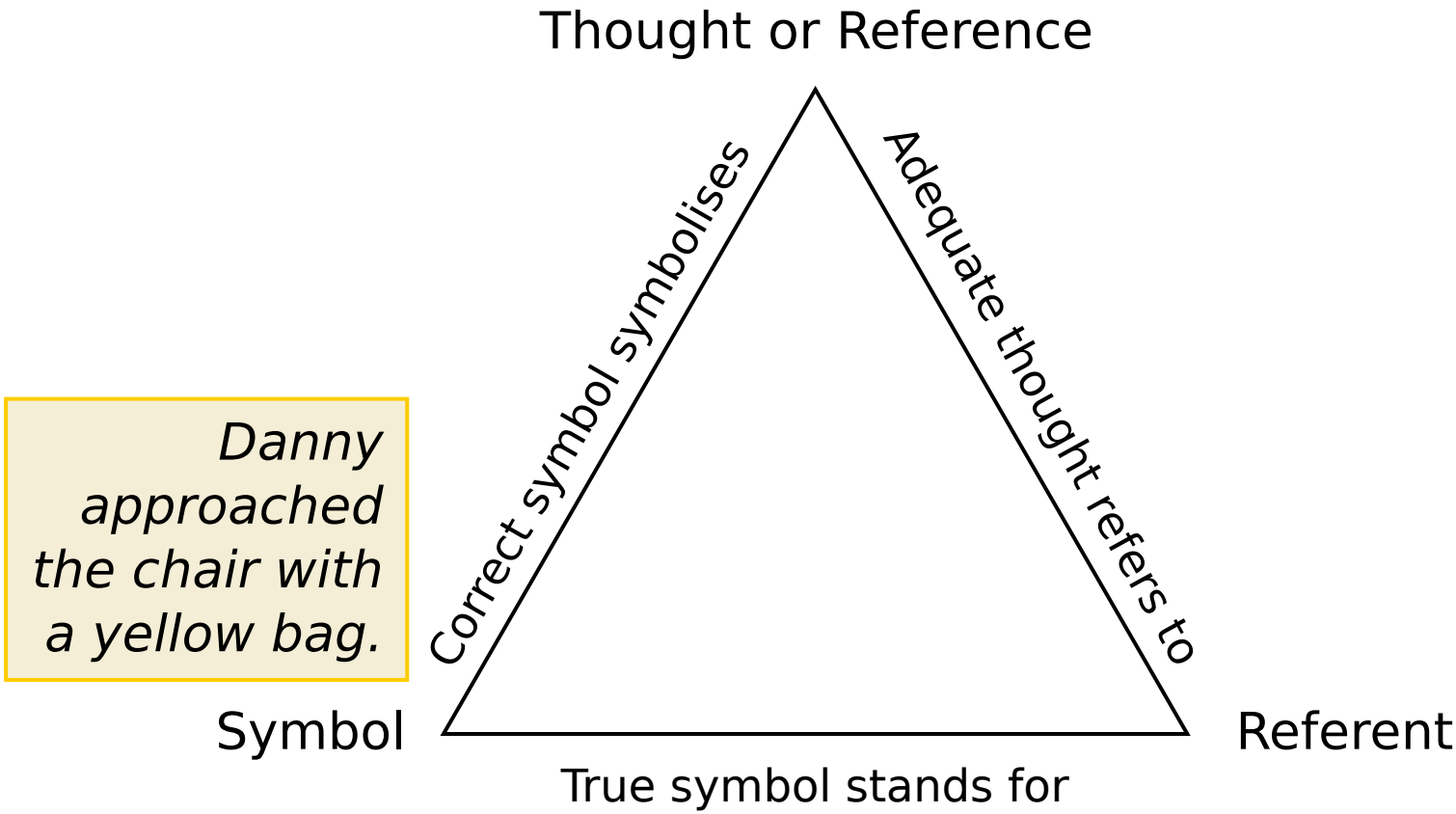
Example from Face Recognition



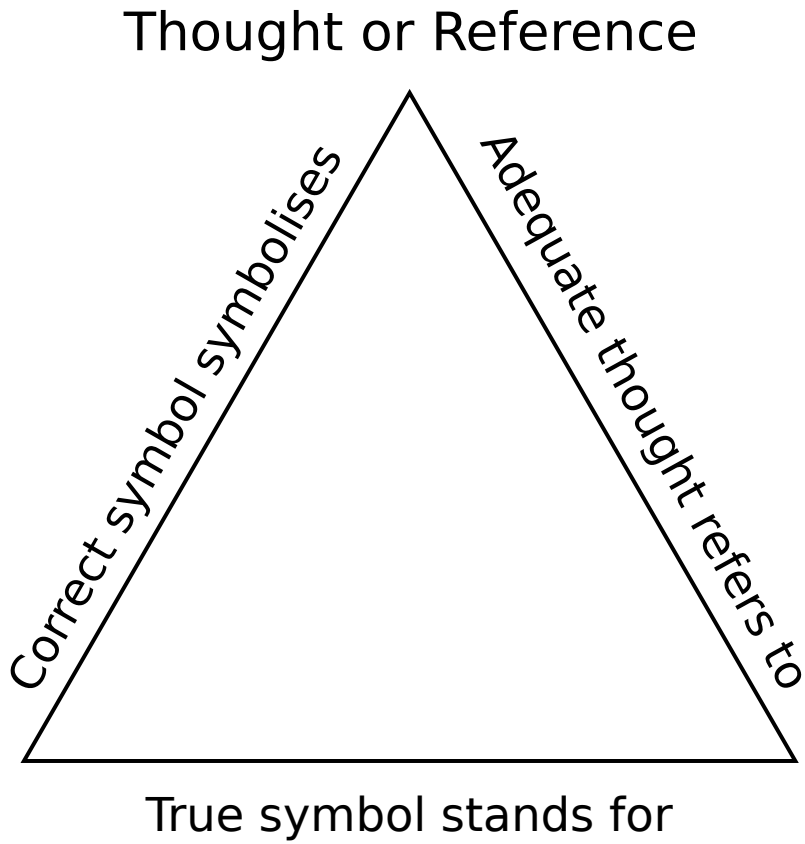
Semiotic Triangle by Ogden and Richards



Semiotic Triangle by Ogden and Richards

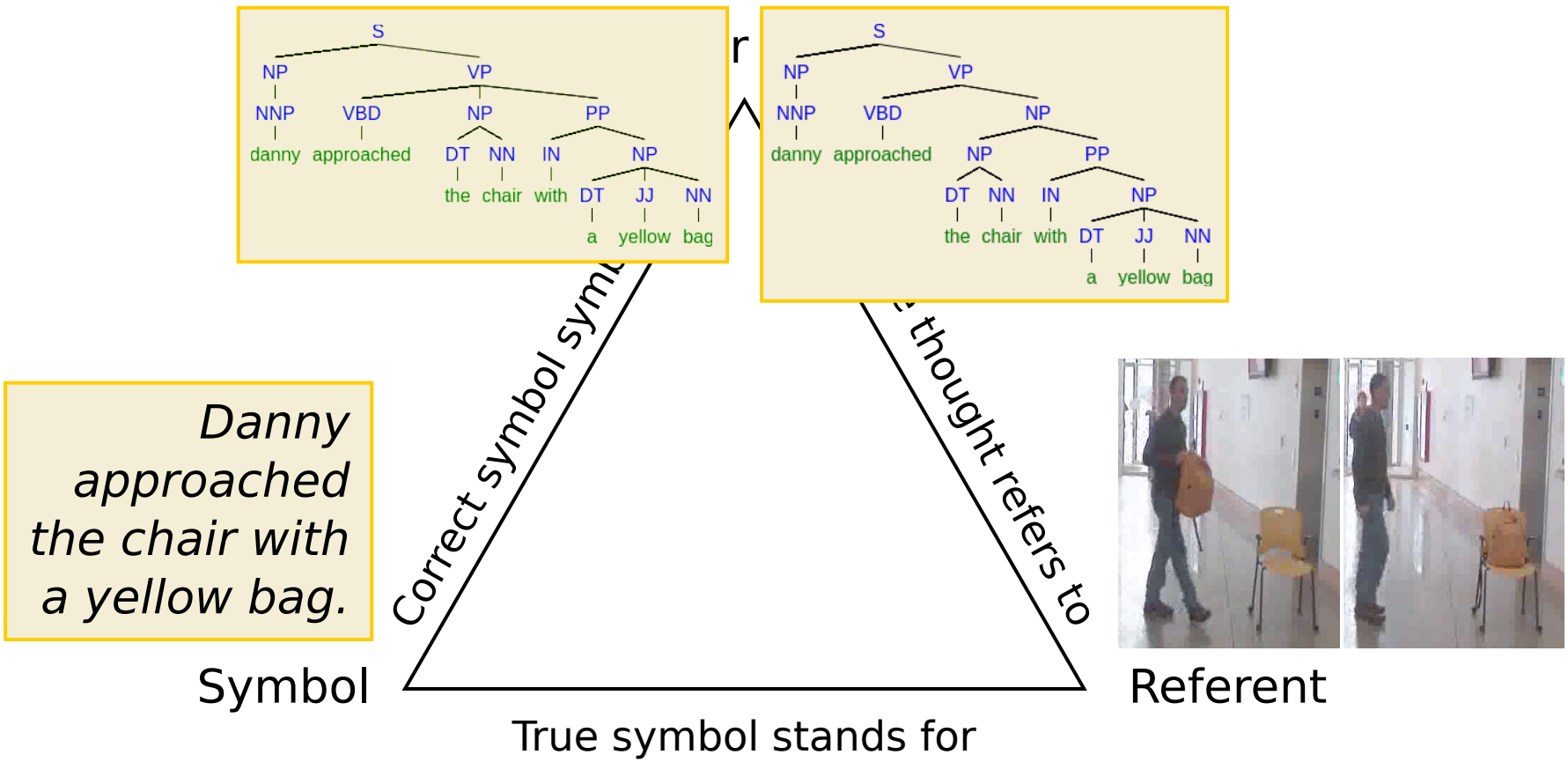


Semiotic Triangle by Ogden and Richards

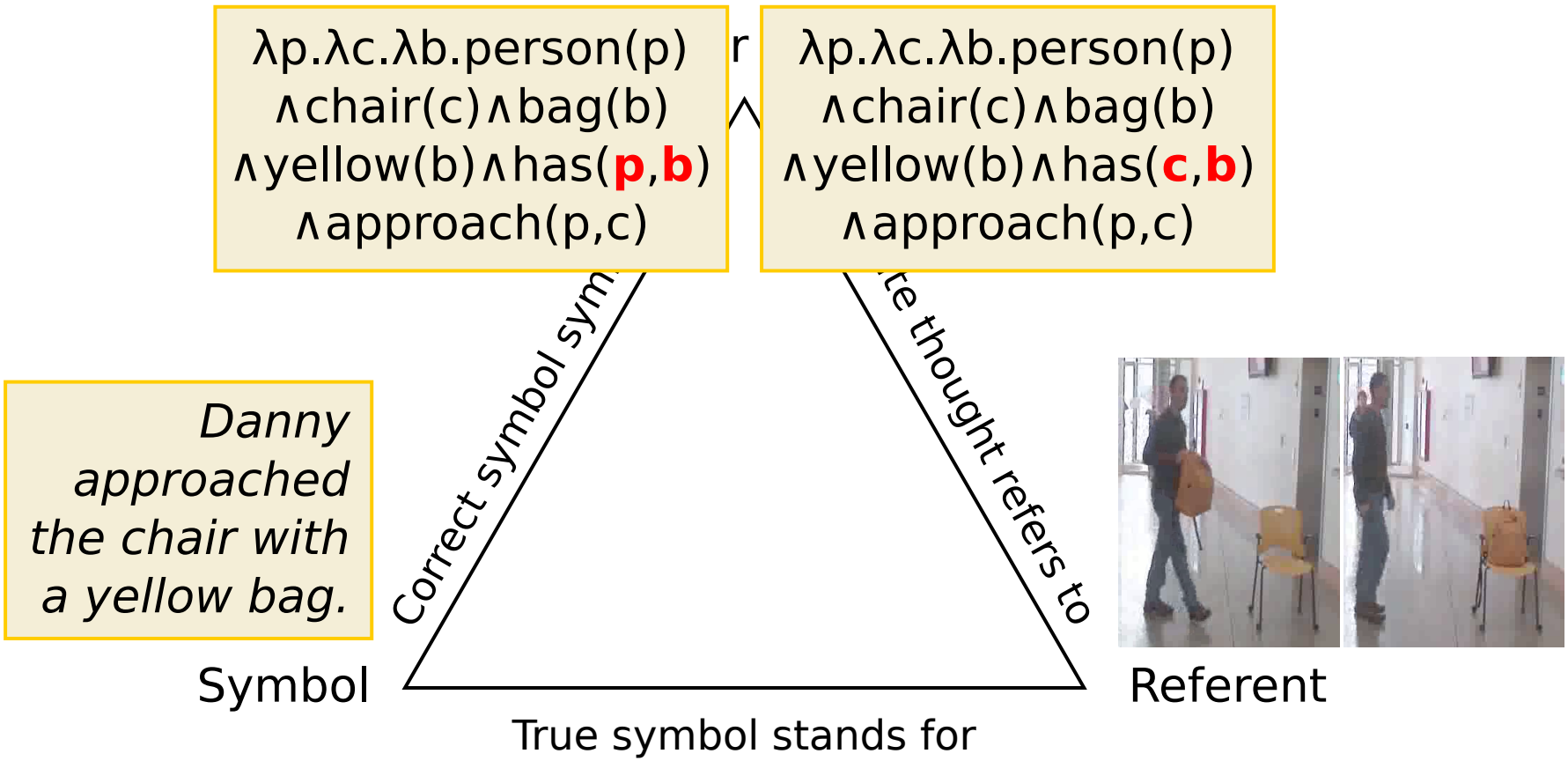


Danny approached the chair with a yellow bag.

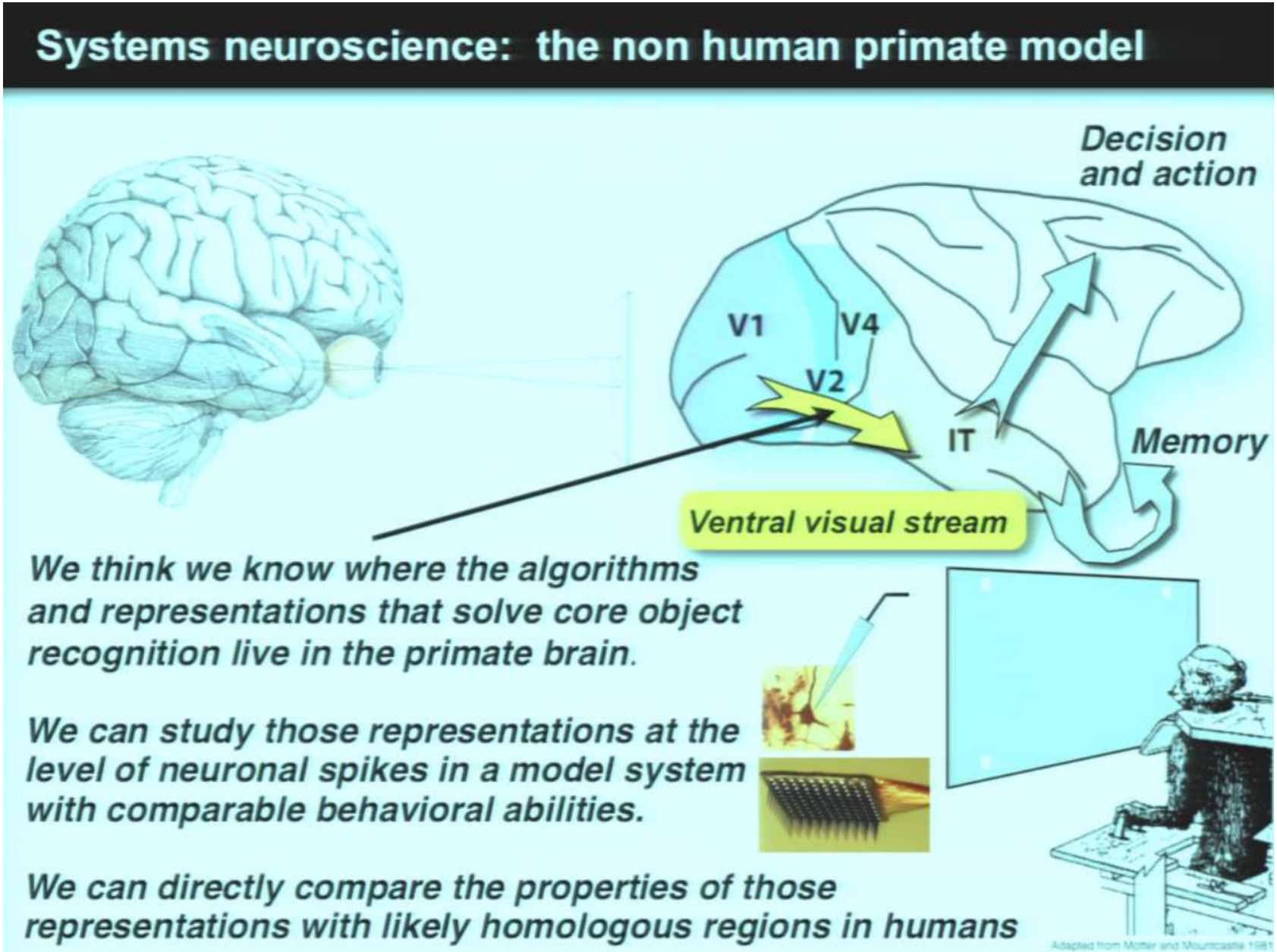
Semiotic Triangle by Ogden and Richards



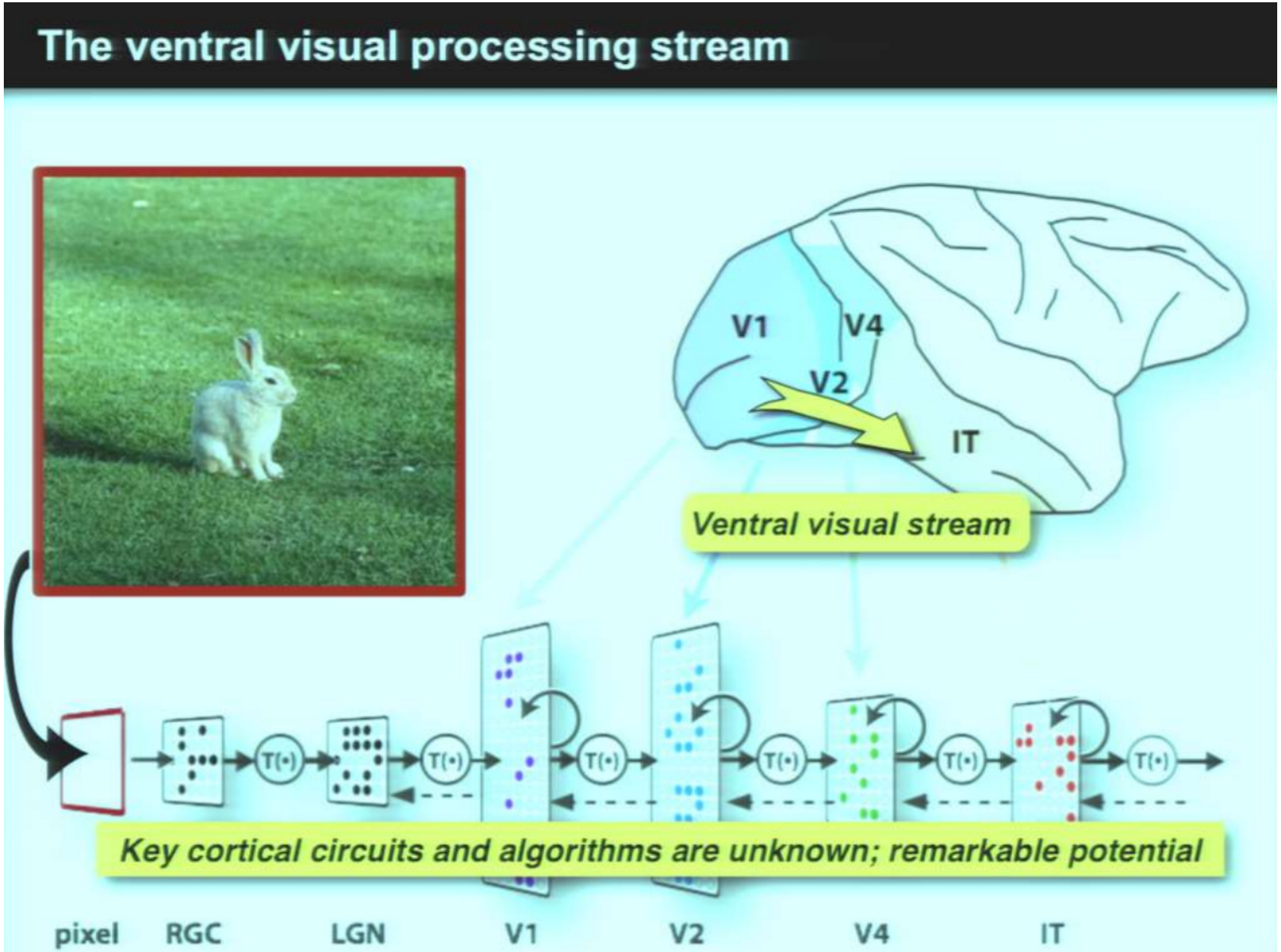
Semiotic Triangle by Ogden and Richards



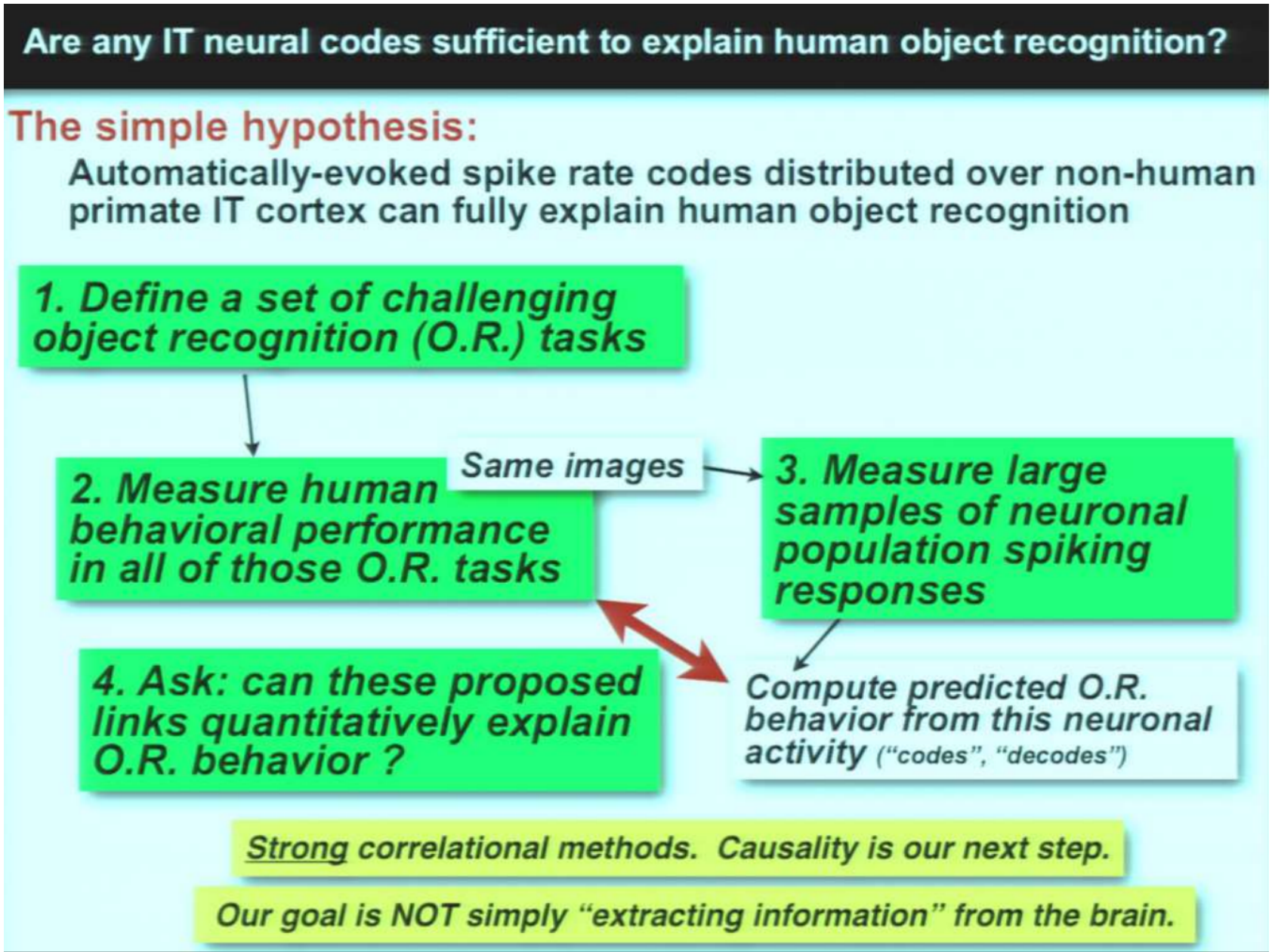
DiCarlo NIPS 2013 Tutorial on Vision



DiCarlo NIPS 2013 Tutorial on Vision



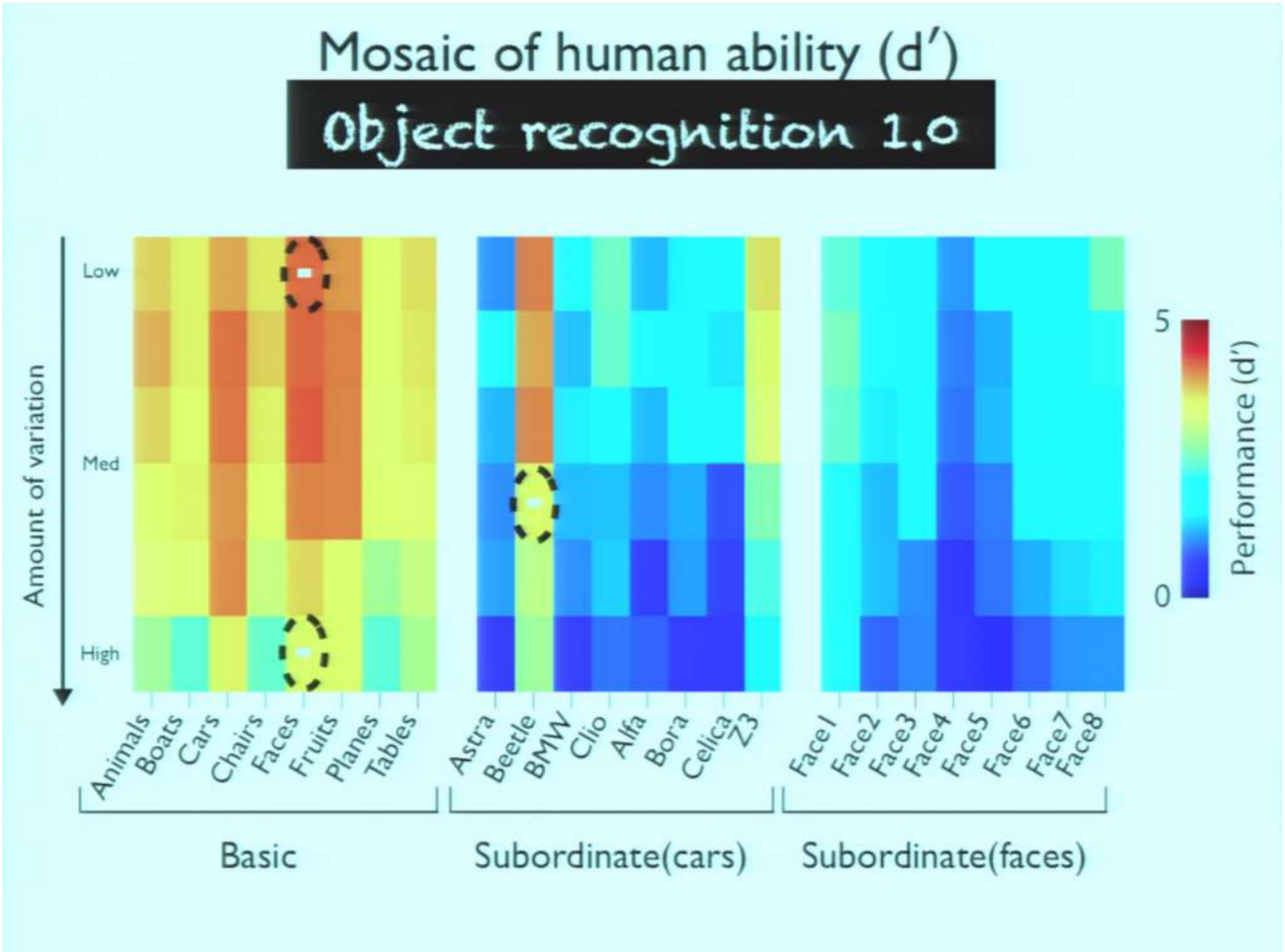
DiCarlo NIPS 2013 Tutorial on Vision



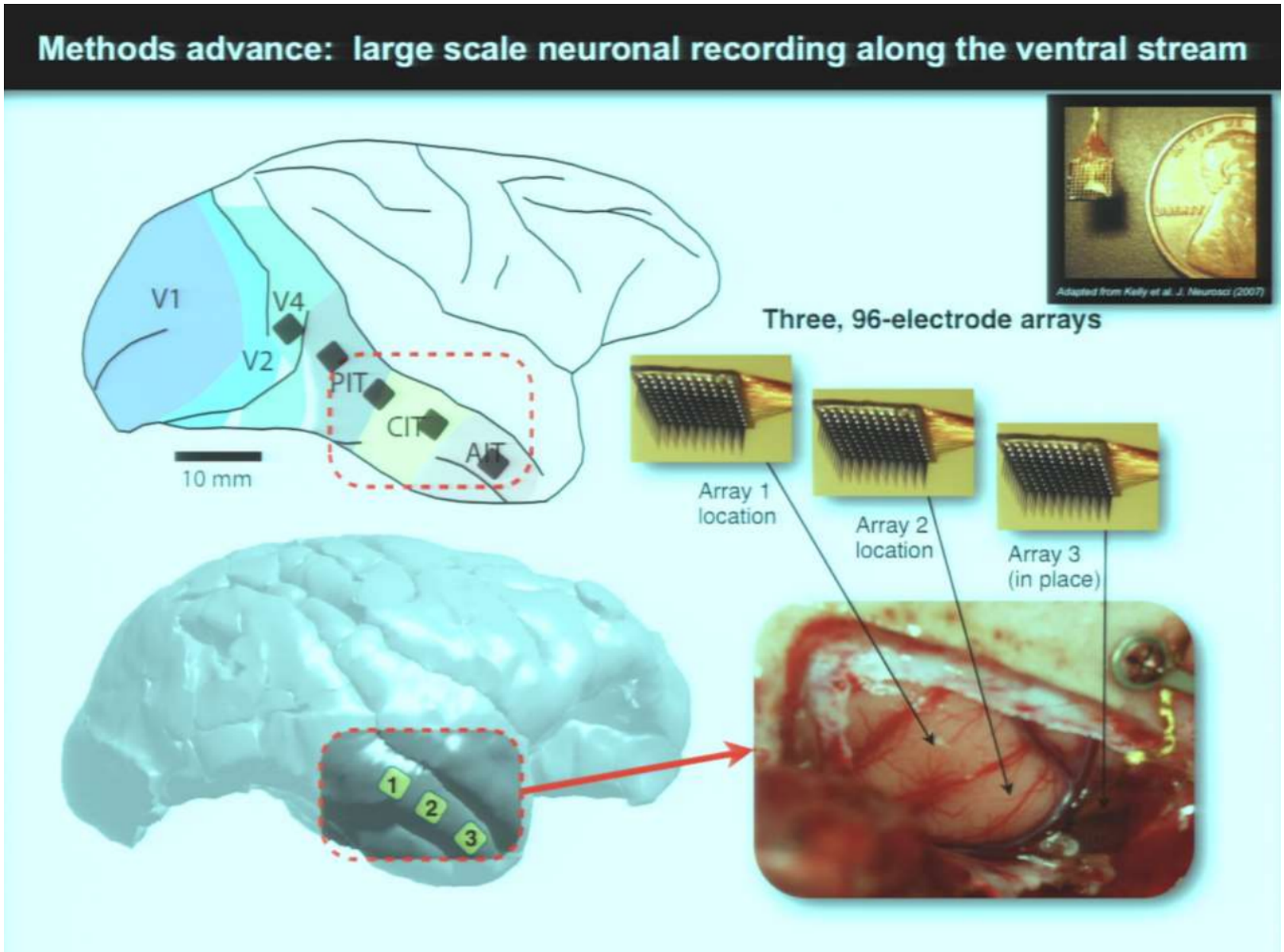
DiCarlo NIPS 2013 Tutorial on Vision

- 64 objects, can generate as many images as we like
- full parametric control
- "natural" statistics
- uncorrelated, new background every image
- not fully "natural" by design -- challenging for computer vision, doable by humans

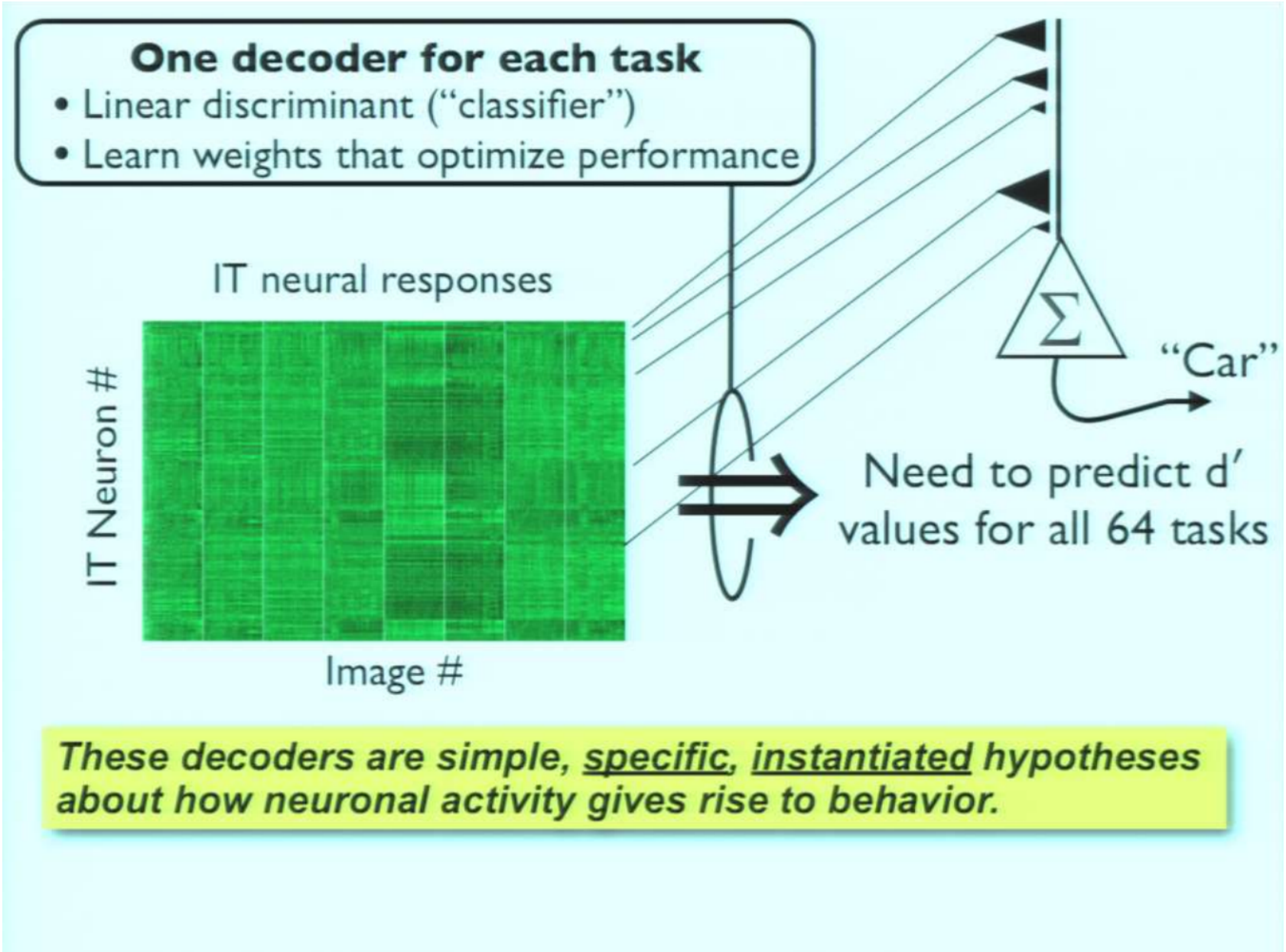
DiCarlo NIPS 2013 Tutorial on Vision



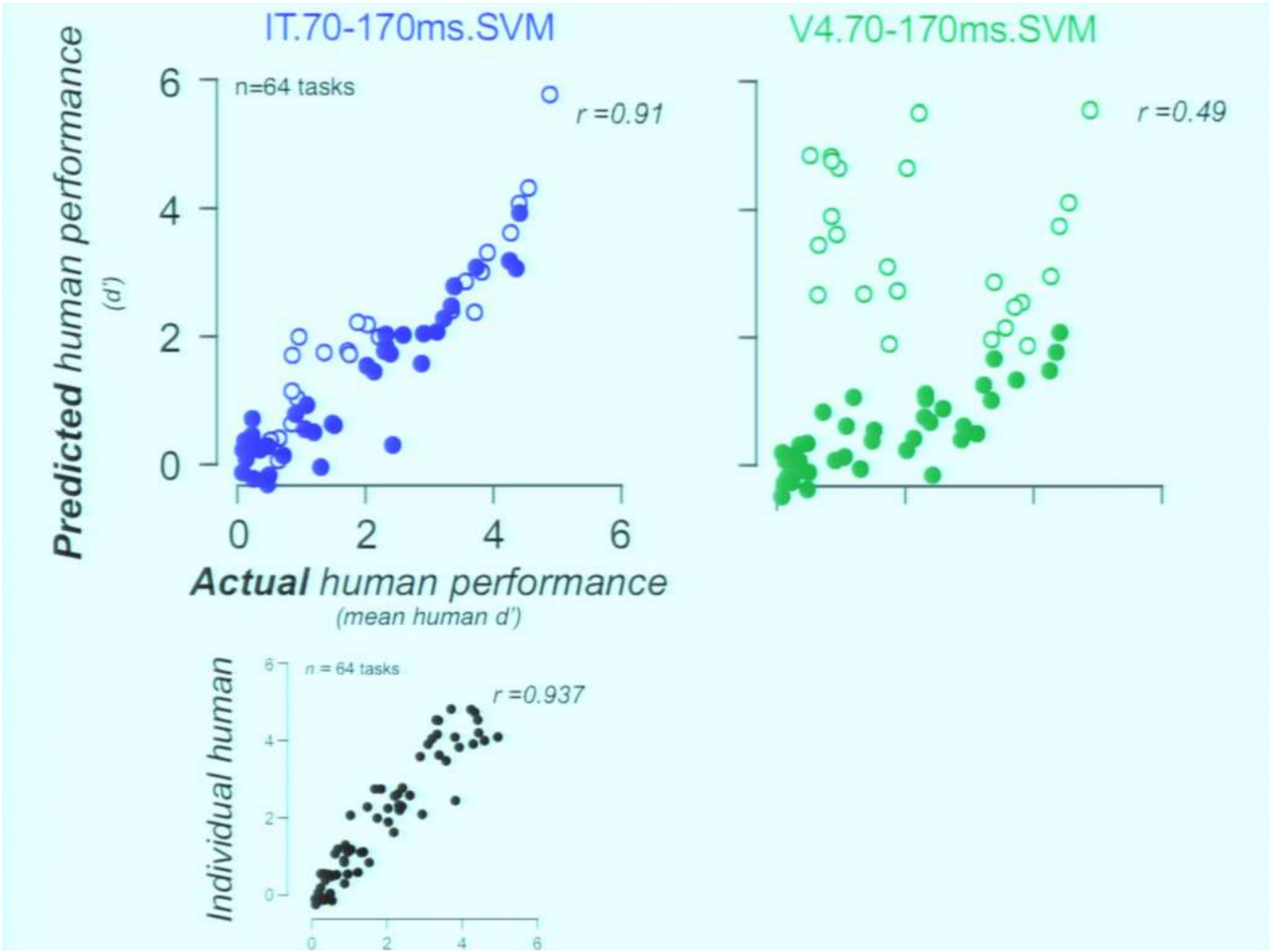
DiCarlo NIPS 2013 Tutorial on Vision



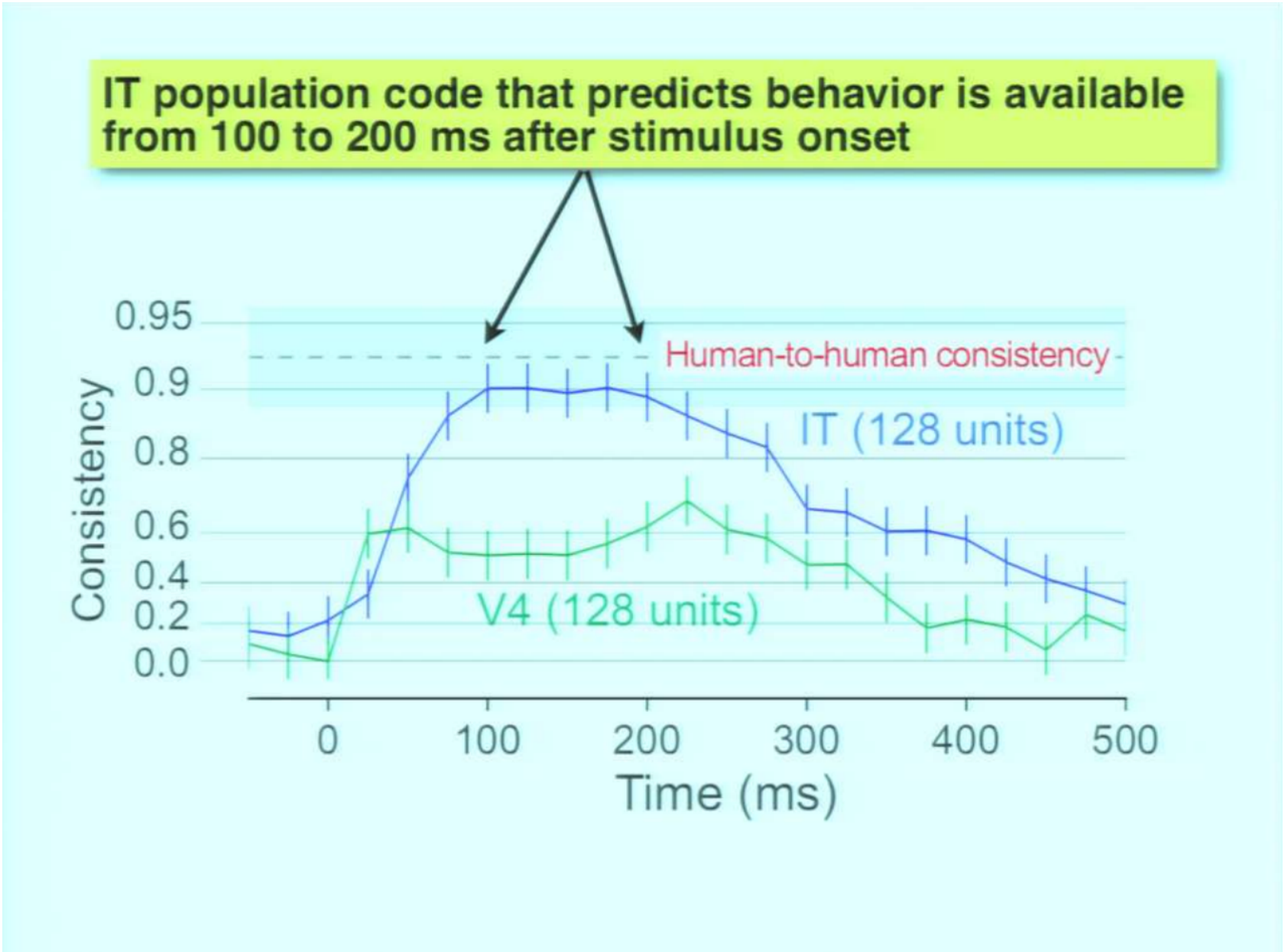
DiCarlo NIPS 2013 Tutorial on Vision



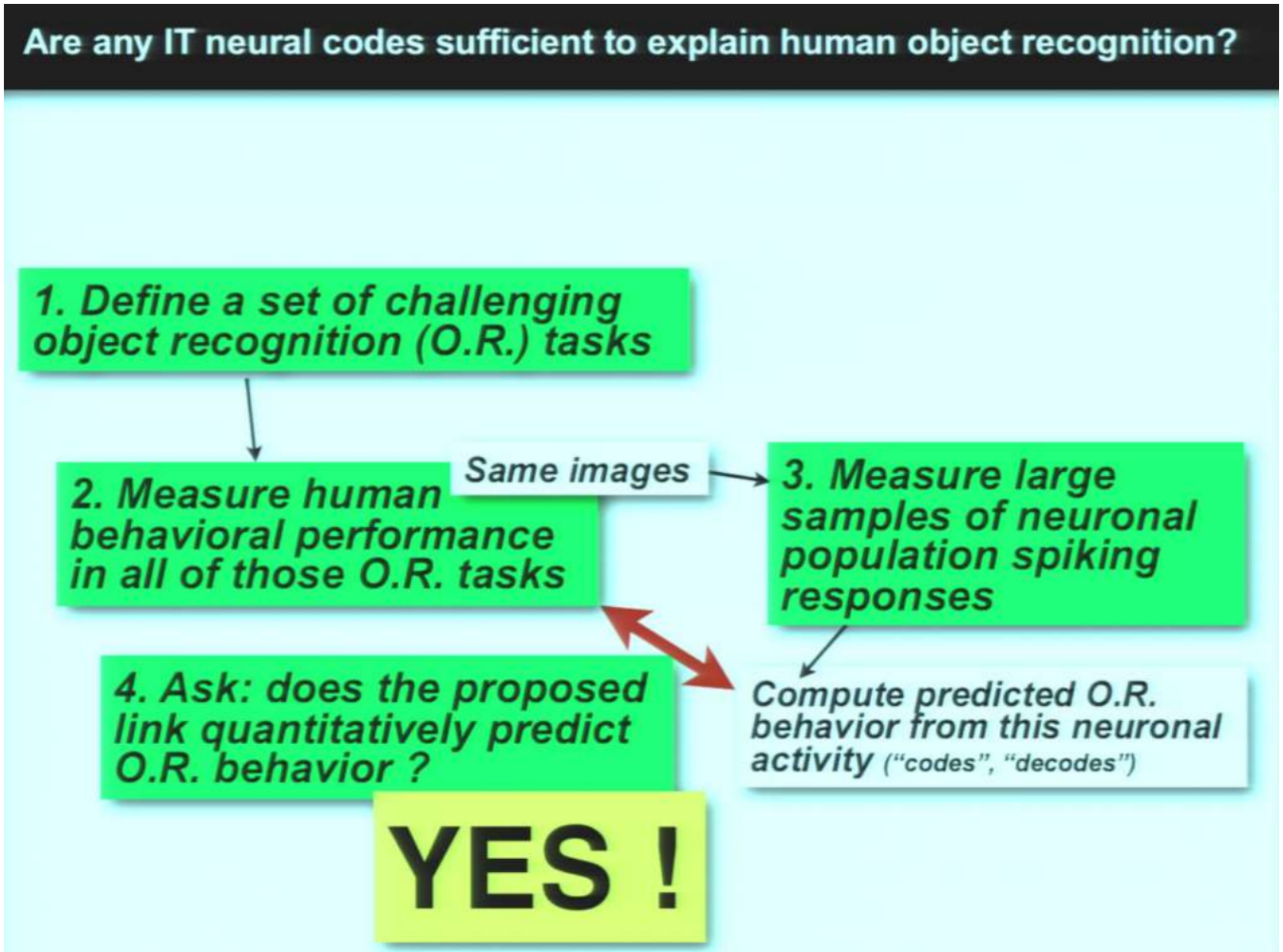
DiCarlo NIPS 2013 Tutorial on Vision



DiCarlo NIPS 2013 Tutorial on Vision



DiCarlo NIPS 2013 Tutorial on Vision



From Vision to Language

We can explain human object recognition by:

- Recording apes' neuronal activity
  and attaching a single-layer NN to interpret it
- Measuring human performance

... on the same object recognition tasks.

- and relating them.

Idea:

- Record NMT behaviour (all parameters accessible)
- and human behaviour, possibly recording:
  - Objective: reading studies, eye-tracking, ...
  - Subjective: introspection.

... on the same language processing tasks.

- and relate them.

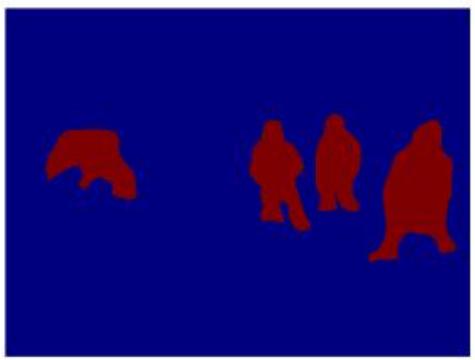
Aspects of Meaning

- Meaning is a coarsening:
  - Pictures: Semantic segmentation ("reverse raytracing")
  - Programs: The output they give (caveat: undecidable).
  - CL: Reference to real world? Speaker's intention?
- Meaning can be shifted, modified.
- Meanings can be compared.
- Meaning is generally compositional.
- *Linguistic meaning* captures the structure of expressions:
  - Morphology, syntax, ...
- Pragmatics: Named entities, numbers, anaphora...
- Expressions are ambiguous.
- Meanings are vague.
- Continuousness.
- Statefulness.

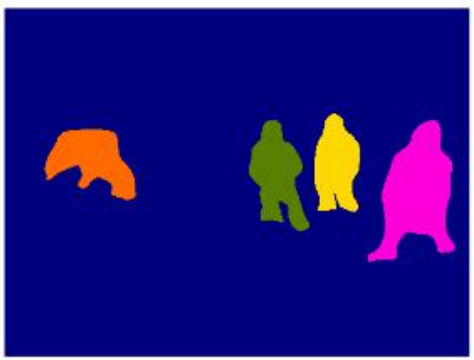
Meaning as a Coarsening

Semantic Segmentation of Pictures

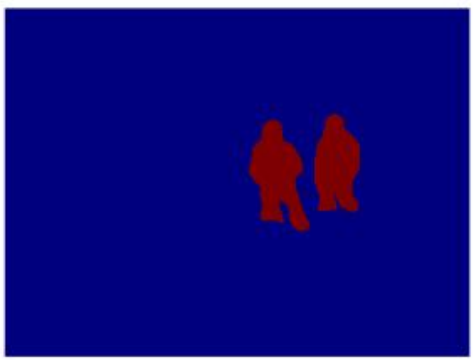
(a) input image



(b) object class segmentation of class ***people***



(c) object instance segmentation of class ***people***



(d) segmentation from expression *"people in blue coat"*

Illustration from `http://www.cs.toronto.edu/~tingwuwang/semantic_segmentation.pdf`.

Compositionality of Meaning

Manning (2015):

> *understanding novel and complex sentences crucially depends on being able to construct their meaning compositionally from smaller parts—words and multiword expressions—of which they are constituted.*

Meaning Statefulness

Stateful Meaning Representation:

- "The state of mind after having read this and produced this output so far."
- Corresponds to models with attention.
- Btw needed to interpret humour (Gluscevskij, 2017).

Stateless Meaning Representation:

- Points correspond to expressions.
  - Ambiguity representation unclear.
- Points correspond to meanings.
  - As in models without attention.

Is Sentence Meaning Continuous?

We know that one English sentence can have 70k Czech translations (Bojar et al., 2013):

And even though he is a political veteran, the Councilor Karel Brezina responded similarly.

A ačkoli ho lze považovat za politického veterána, radní Březina reagoval obdobně.
A i přestože je politický matador, radní Karel Březina odpověděl podobně.
A přestože je to politický veterán, velmi obdobná byla i reakce radního K. Březiny.
A radní K. Březina odpověděl obdobně, jakkoli je politický veterán.
Byť ho lze označit za politického veterána, Karel Březina reagoval podobně.
Byť ho můžeme prohlásit za politického veterána, byla i odpověď K. Březiny velmi po
K. Březina, i když ho lze prohlásit za politického veterána, odpověděl velmi obdobně.
Odpověď Karla Březiny byla podobná, navzdory tomu, že je politickým veteránem.
Radní Březina odpověděl velmi obdobně, navzdory tomu, že ho lze prohlásit za politi
Reakce K. Březiny, třebaže je politický veterán, byla velmi obdobná.
Velmi obdobná byla i odpověď Karla Březiny, ačkoli ho lze prohlásit za politického ve

Is Sentence Meaning Continuous?

Similarly for English (Dreyer and Marcu, 2012):

> *Premiere of Iraq Nuri al-Maliki was given an excuse by President Bush, who expressed his confidence in him, and he stated that the circumstances are complicated.*
> *President Bush said that he trusts in Nouri Maliki, head of government of Iraq, and he stated that he finds an excuse for him "because the situation is tricky".*
> *Head of cabinet of Iraq Nuri al-Maliki was given an excuse by President Bush, who expressed his trust in him, and he indicated that the circumstances are difficult.*
> *Iraq's head of cabinet Nuri al-Maliki was given a reason by President Bush, who expressed his trust in him, and he indicated that the case is tricky.*
> *President Bush said that he has faith in Iraqi head of cabinet Nouri al-Maliki, and he stated that he finds an excuse for him "for the case is complicated".*

Q: Are all these paraphrases close in sent embedding spaces?
Q: How entagled are manifolds of *different* sents?
... work in progress with Holger Schwenk.

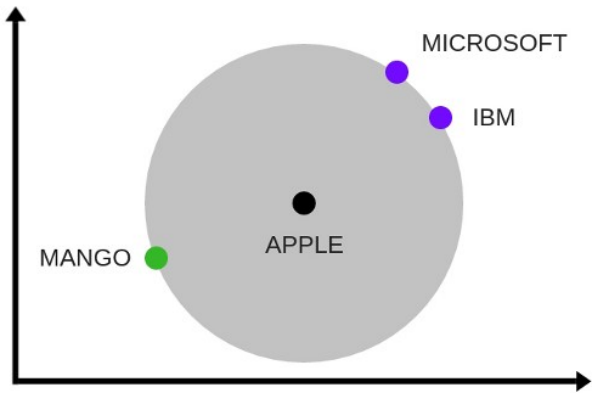
Aspects of Meaning: Symbolic vs. DL

Aspect as covered by	Symbolic Theories	DL Reprs
Meaning is a Coarsening	$\surd$	$\times$
Operations (shifting, modification)	$\sim$	$\sim$
Compositionality	$\surd$	$\sim$
Relatability	? (WIP)	$\surd$
Meaning is vague	$\times$	$\surd$
Expressions are ambiguous	$\surd$	$\times$ (WIP)
Continuous	? (WIP)	$\surd$
Statefulness	$\sim$	$\surd$

Modelling Ambiguity

Sentence-level embeddings always produced by an encoder.

- Encoder = A deterministic mapping from expression to meaning.
- Unclear how ambiguous expressions are represented.



Idea: Focus on decoder:

- Decoder maps meanings to expressions.

Work in progress with Ondřej Cífka.

Dedecoding

"Dedecoding": Given an expr., which all meanings lead to it?

Reversing the decoder:

- Non-singular matrices fully reversible.
- Injective activation functions fully reversible.
- But softmax kills reversibility.

Proposed solution:

- *Train* source sentence representation
  - using standard backpropagation
  - and fixed decoder
  - to produce a given fixed input sentence.
- Run training many times to sample the space.

First Results of Dedecoding (1/2)

- Neural Monkey S2S without attention.
- Trained for en2de on the multimodal task data (29k sents).
- Sampling for one sentence takes about 2 minutes.
- Outputs so far rather bad.
  - One sentence perfect but no variance:

Input	Ein Mann schläft in einem grünen Raum auf einem Sofa .
Dede1	Ein Mann schläft in einem grünen Raum auf einem Sofa .
Dede2	Ein Mann schläft in einem grünen Raum auf einem Sofa .
Dede3	Ein Mann schläft in einem grünen Raum auf einem Sofa .
Dede3	Ein Mann schläft in einem grünen Raum auf einem Sofa .

First Results of Dedecoding (1/2)

- Some sentences related:

Input	Eine Gruppe von Männern lädt Baumwolle auf einen Lastwagen
Dede1	Eine Gruppe von Männern lädt Vons-Einkaufswagen auf einen Lastwager
Dede2	Eine Gruppe von Männern lädt Vons-Einkaufswagen auf einen Lastwager
Dede3	Eine Gruppe von Männern lädt besucht einen Lastwagen .
Dede4	Eine Gruppe von Männern lädt besucht einen Lastwagen

- Most sentences totally off:

Input	Ein süßes Baby lächelt einem anderen Kind zu .
Dede1	Ein Mann im lila Hemd , der einen grünen Hut trägt .
Dede2	Eine Mutter mit einer großen weißen Kapuze .
Dede3	Ein junger Mann hält eine türkische Flagge .
Dede4	Eine Frau trägt einen schwarzen Ganzkörperanzug und posiert im Freien

Examining Continuous Space of Sents.

Stages of Space Mapping:

1. Propose directions of exploration.
2. Generate seed pairs of sentences for each of the directions.
3. Collect specimens along the proposed directions:
   - interpolation, a “sentence in between”,
   - extrapolation, “a sentence further in the hinted direction”.
   - Allow people to say “impossible”.
4. Validate the relations.
5. Create the partially ordered set.
6. Search for a manifold covering the ordered set.

Work in progress with Chris Callison-Burch.

Directions of Exploration (1/2)

- Politeness
- Tense
- Verity: How much the speaker believes the message.
- Modality: Willingness/Ability of the speaker to do it.
- "Counting" / Generic Numerals, Scalar adjectives
  - I saw a handful of people there. / a big crowd / a massive crowd.
  - freezing / cold / chilly
- "Negation", but not only reversing the main predicate
- Complexity / simplicity, Length.

Directions of Exploration (2/2)

- Specificity / Generality, Vagueness.
  - Geese fly / Geese migrate / Geese migrate south / The Canadian geese flew over the pond at friendly Farms in their southward migration.
  - Hammer the hook into the wall. / Put the hook on the wall. / Do the thingy in there.
- Contextual boundness.
  - Give it to him. / Give the parcel to the man at the counter. / Give your parcel to the operator at the post office.
- High/low style/English/class.
  - Hey y'all it's a nice day ain't it?
  - Greetings! Lovely weather we are having.

Thanks to Sarka Zikanova for some of the ideas.

First Results of Getting Pairs

Can you please give me a minute?	Could you leave me alone?
Close the door.	Close the damn door man
Can you help me find something?	I need you to help me get something.
May I talk to Mary?	Is Mary here?
I'm sorry-I don't believe we have met.	Who the hell are you?
Can you move so I can see the screen?	You aren't made of glass, you know.
Will you kindly exit?	I do not want you here!
Would you please get the mail?	Get the mail!
Can I help you?	What do you want?
Can you please help me with this?	Get over here and help me!
Can you make me breakfast?	Why are you not making me breakfast right
I tried to call were you busy?	You never answer your phone.

First Results of Midpointing (1/3)

Can you help me find something?

Would you help me look?
Find this for me.
Help me find something.
Please help me find something.
Will you help me?
Your assistance in finding something is required.

I need you to help me get something.

First Results of Midpointing (2/3)

Can you please give me a minute?
I'd like a minute alone.
Please wait.
Give me a minute.
One moment.
I need more time.
Come back later
Hey give me a minute.
One minute.
I need a minute to myself.
Could you leave me alone?

First Results of Midpointing (3/3)

Can you move so I can see the screen?

Blocking the view, friend.
Move your blocking the screen
Could you move a little bit, you're blocking the screen.
Can you please move?
I can't see, can you move a little?
Hey can you move.
Please move.
Can you move a bit?

You aren't made of glass, you know.

Some Techniques of NN Inspection

- MicroNNs, e.g. Shi et al. (2016) learning length.
- Lobotomy.
- Exploring representation space.
  - t-SNE and PCA for sentence pairs
  - Translation by search = similarity in meaning reflected in space
  - Attaching an NN to see if it can infer:
    - POS or morphology from NMT
    - Subject-Verb agreement (Linzen et al. TACL/EACL 2017)
- Linguistic exploration:
  - Various test suites (Burlot 2017, Burchhardt MQM).
  - Stanford Natural Language Inference (SNLI)
    `https://nlp.stanford.edu/projects/snli/`
  - Paraphrases (see above).
- Comparing representations (Nili et al., 2014)

Summary

Fish by Frits Ahlefeldt

Summary

- Neural MT reaches and can surpass humans.
  - Catastrophic errors still possible.
- As a side-effect, continuous representations are learned.
- Insight in vision thanks to relating computer and human vision.

- Computational linguistics has plenty of data.
- Other data can be relatively easily obtained.

$\Rightarrow$ Let's train NLU systems and dissect them.

References

Ondřej Bojar, Matouš Macháček, Aleš Tamchyna, and Daniel Zeman. 2013. Scratching the Surface of Possible Translations. In *Proc. of TSD 2013*, Lecture Notes in Artificial Intelligence, Berlin / Heidelberg. Západočeská univerzita v Plzni, Springer Verlag.

Ondřej Bojar, Jindřich Helcl, Tom Kocmi, Jindřich Libovický, and Tomáš Musil. 2017. Results of the WMT17 Neural MT Training Task. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark, September. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October. Association for Computational Linguistics.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555.

Markus Dreyer and Daniel Marcu. 2012. HyTER: Meaning-Equivalent Semantics for Translation Evaluation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 162–171, Montréal, Canada, June. Association for Computational Linguistics.

Dmitrij Gluscevskij. 2017. Methodological issues and prospects of semiotics of humour. *Sign Systems Studies*, 45(1/2):137–151.

Jindřich Helcl and Jindřich Libovický. 2017. Neural Monkey: An open-source tool for sequence learning. *The Prague Bulletin of Mathematical Linguistics*, 107:5–17.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November.

Tom Kocmi and Ondřej Bojar. 2017. Curriculum Learning and Minibatch Bucketing in Neural Machine Translation. In *Proceedings of Recent Advances in NLP (RANLP 2017)*.