Strojový wikipedista? Strojovým překladem to začíná.

Ondřej Bojar
bojar@ufal.mff.cuni.cz
Ústav formální a aplikované lingvistiky
Matematicko-fyzikální fakulta
Univerzita Karlova v Praze

Osnova

- Počítačová lingvistika a její aplikace.
- Proč je strojový překlad (MT) těžký.
- Překlad do roku 2016.
- Stav strojového překladu v roce 2017.
  = Neuronový ~~neuronální~~ strojový překlad.
- Jak a co se hluboké neuronové sítě se učí.
- Budoucí role wikipedistů.
- (Budoucí role počítačových lingvistů.)

Počítačová lingvistika a její aplikace

Kontrola překlepů. Kontrola pravopisu.
Vyhledávání dokumentů (na webu). Sumarizace textů.
Syntéza a rozpoznávání řeči. Dialogové systémy.
Strojový překlad (mluvené řeči).

Obtížnost strojového překladu

I	saw	two	green	striped	cats	.
já	pila	dva	zelený	pruhovaný	kočky	.
	pily	dvě	zelená	pruhovaná	koček	
	. . .	dvou	zelené	pruhované	kočkám	
	viděl	dvěma	zelení	pruhovaní	kočkách	
	viděla	dvěmi	zeleného	pruhovaného	kočkami	
	. . .		zelených	pruhovaných		
	uviděl		zelenému	pruhovanému		
	uviděla		zeleným	pruhovaným		
	. . .		zelenou	pruhovanou		
viděl jsem			zelenými	pruhovanými		
viděla jsem			. . .	. . .		

Obtížnost strojového překladu

I	saw	two	green	striped	cats	.
já	**pila**	dva	**zelený**	pruhovaný	kočky	.
	pily	**dvě**	zelená	pruhovaná	koček	
	. . .	dvou	zelené	pruhované	kočkám	
	viděl	dvěma	zelení	pruhovaní	kočkách	
	viděla	dvěmi	zeleného	pruhovaného	kočkami	
	. . .		zelených	pruhovaných		
	uviděl		zelenému	pruhovanému		
	uviděla		zeleným	pruhovaným		
	. . .		zelenou	pruhovanou		
viděl jsem			zelenými	pruhovanými		
viděla jsem			. . .	. . .		

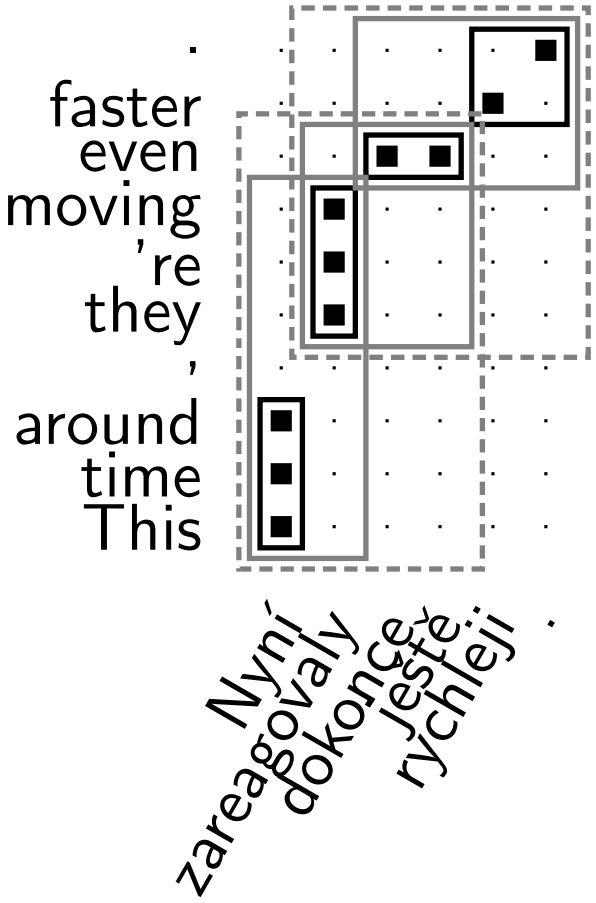
Obtížnost strojového překladu

I	saw	two	green	striped	cats	.
já	pila	dva	zelený	pruhovaný	**kočky**	.
	pily	**dvě**	zelená	pruhovaná	koček	
	. . .	dvou	**zelené**	**pruhované**	kočkám	
	viděl	dvěma	zelení	pruhovaní	kočkách	
	viděla	dvěmi	zeleného	pruhovaného	kočkami	
	. . .		zelených	pruhovaných		
	uviděl		zelenému	pruhovanému		
	uviděla		zeleným	pruhovaným		
	. . .		zelenou	pruhovanou		
viděl jsem			zelenými	pruhovanými		
viděla jsem			. . .	. . .		

Obtížnost překladu

I	saw	two	green	striped	cats	.
já	pila	dva	zelený	pruhovaný	**kočky**	.
	pily	**dvě**	zelená	pruhovaná	koček	
	. . .	**dvou**	**zelené**	**pruhované**	kočkám	
	viděl	dvěma	zelení	pruhovaní	**kočkách**	
	viděla	dvěmi	zeleného	pruhovaného	kočkami	
	. . .		**zelených**	**pruhovaných**		
zrak mi utkvěl na			zelenému	pruhovanému		
			zeleným	pruhovaným		
	. . .		zelenou	pruhovanou		
viděl jsem			zelenými	pruhovanými		
viděla jsem			. . .	. . .		

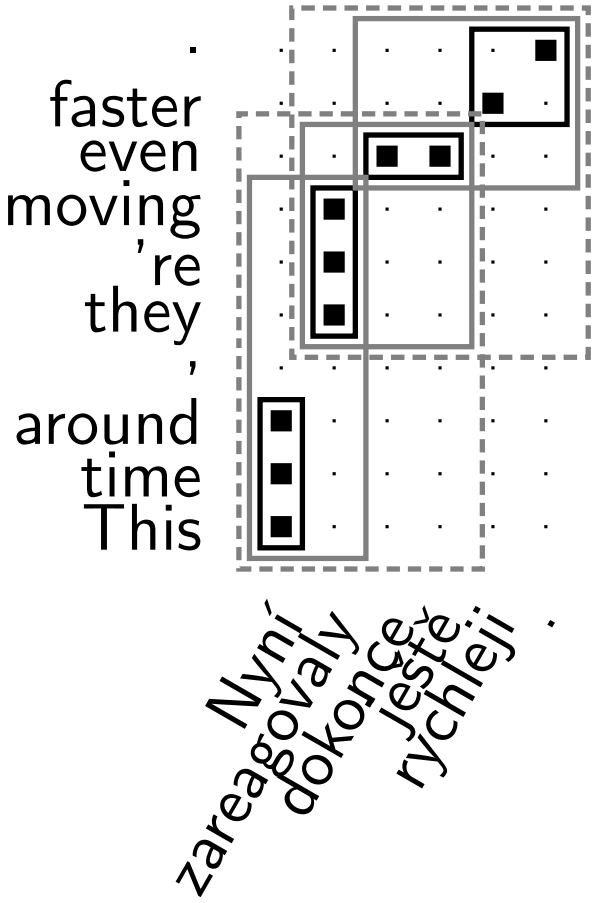
Frázový překlad



Trénovací data:

- paralelní korpus (česká věta = anglická věta)
- zarovnání slov (české slovo ∼ anglické slovo)

Frázový překlad

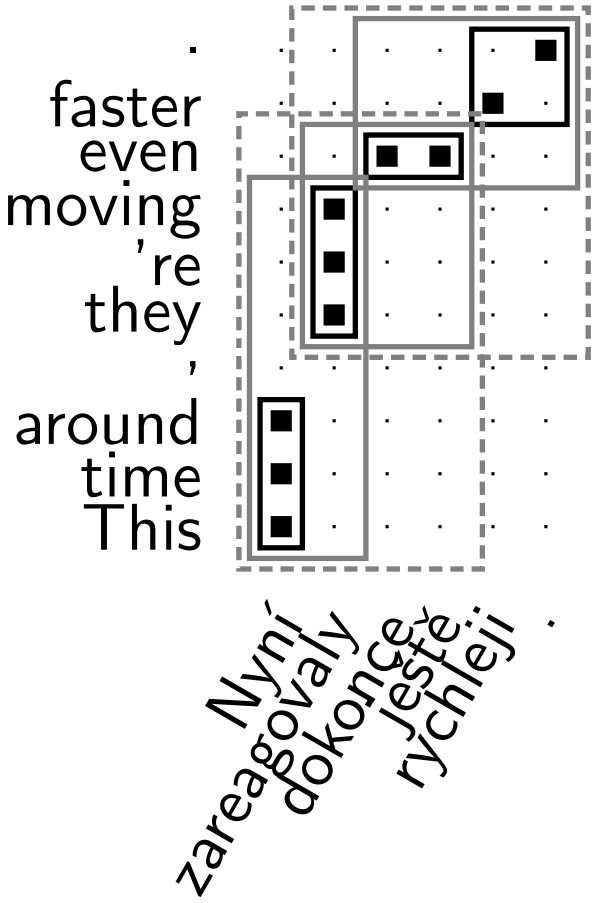


This time around	=	Nyní
they 're moving	=	zareagovaly
even	=	dokonce ještě
...	=	...
This time around, they 're moving	=	Nyní zareagovaly
even faster	=	dokonce ještě rychleji
...	=	...

Trénovací data:

- paralelní korpus (česká věta = anglická věta)
- zarovnání slov (české slovo ∼ anglické slovo)

Frázový překlad



This time around	=	Nyní
they 're moving	=	zareagovaly
even	=	dokonce ještě
...	=	...
This time around, they 're moving	=	Nyní zareagovaly
even faster	=	dokonce ještě rychleji
...	=	...

Trénovací data:

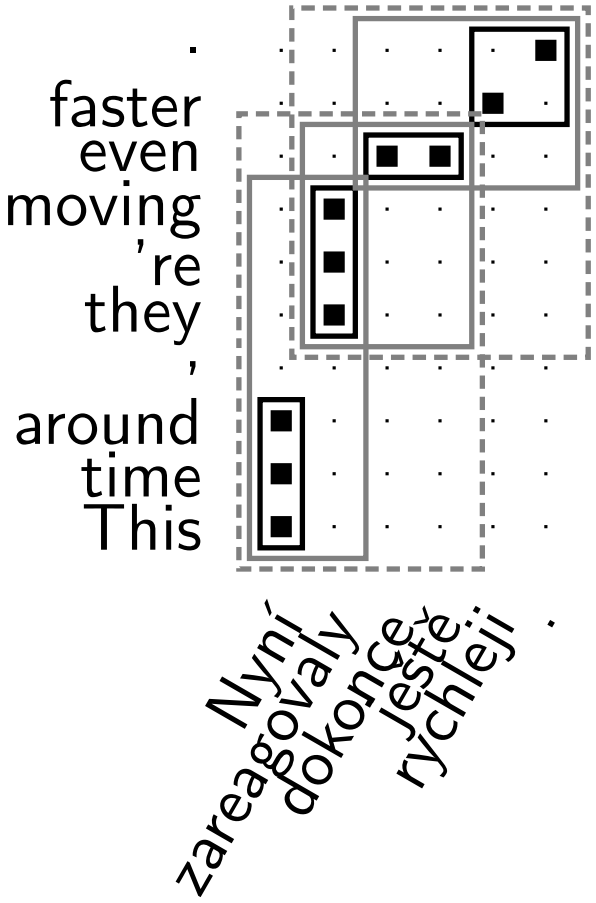
- paralelní korpus (česká věta = anglická věta)
- zarovnání slov (české slovo ∼ anglické slovo)

Při samotném překladu hledáme:

- takové dělení vstupní věty na úseky (“fráze”)
- a takové překlady frází

aby byl výstup co nejpravděpodobnější.

Frázový překlad



This time around	=	Nyní
they 're moving	=	zareagovaly
even	=	dokonce ještě
...	=	...
This time around, they 're moving	=	Nyní zareagovaly
even faster	=	dokonce ještě rychleji
...	=	...

Trénovací data: **. . . 50 mil. párů vět**

- paralelní korpus (česká věta = anglická věta)
- zarovnání slov (české slovo ∼ anglické slovo) **. . .**

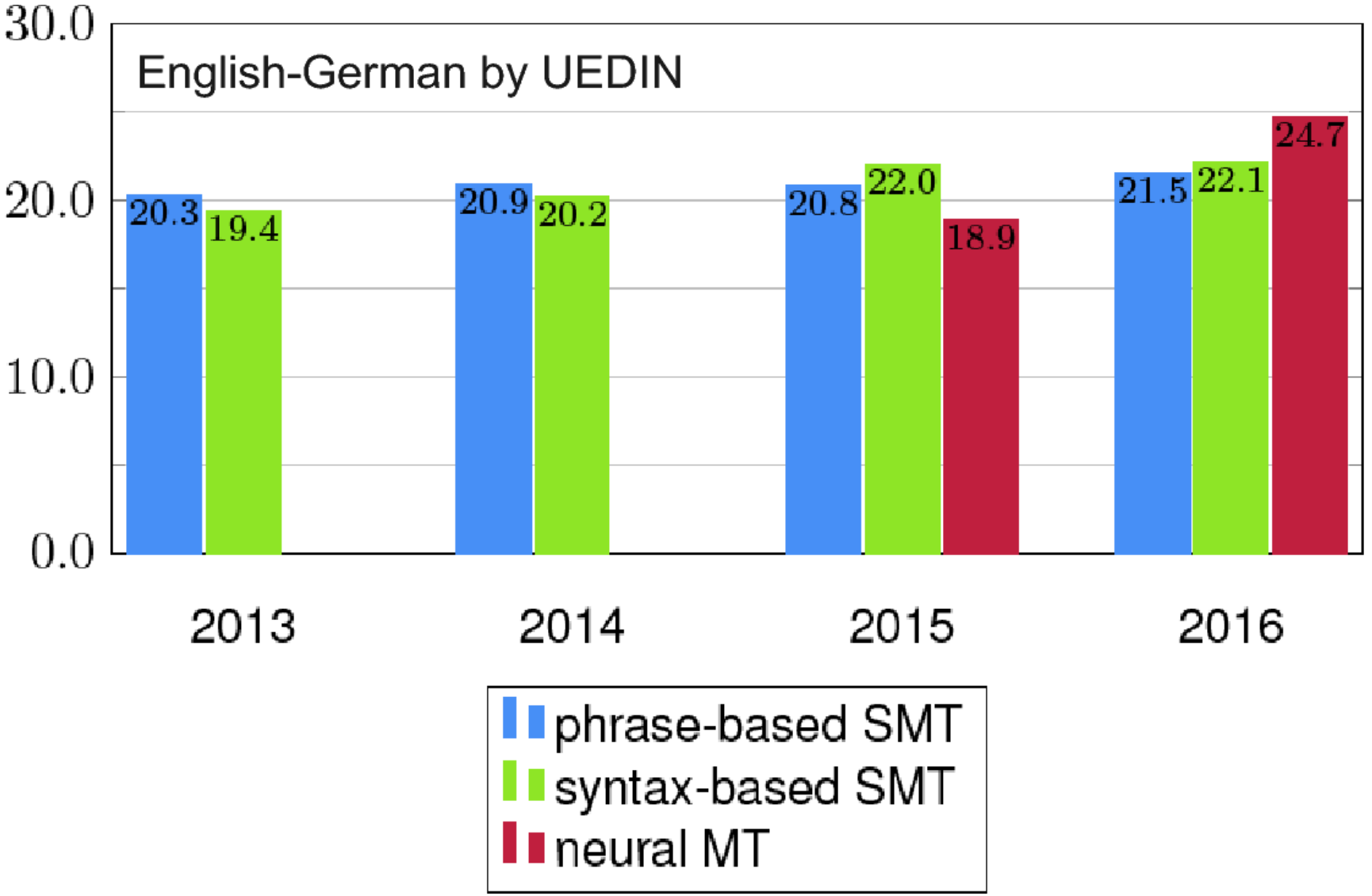
  >700 mil. slov v každé řeči

Při samotném překladu hledáme:

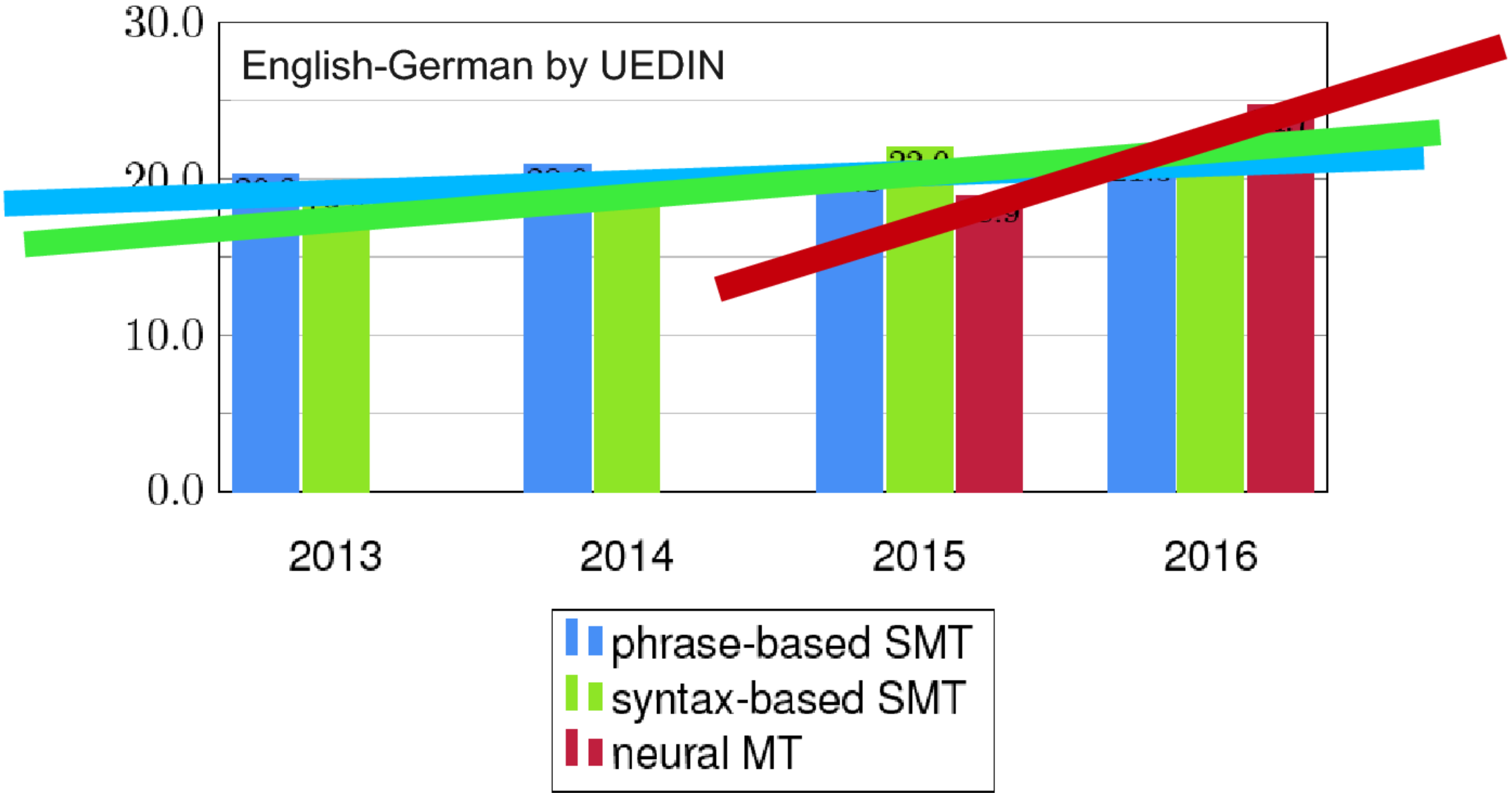
- takové dělení vstupní věty na úseky (“fráze”)
- a takové překlady frází

aby byl výstup co nejpravděpodobnější.

Neuronový strojový překlad (NMT)



Neuronový strojový překlad (NMT)



Je NMT o tolik lepší?

Výstupy nejlepšího systému v letošní soutěži: `http://matrix.statmt.org/`

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.

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.

Je NMT o tolik lepší?

Výstupy nejlepšího systému v letošní soutěži: `http://matrix.statmt.org/`

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.

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.

Naštěstí ;-) stále dělá vážné chyby

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 representativ
MT 63letý **mladík** se nyní stal zvláštním zástupcem...

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

...vlastně katastrofické chyby (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

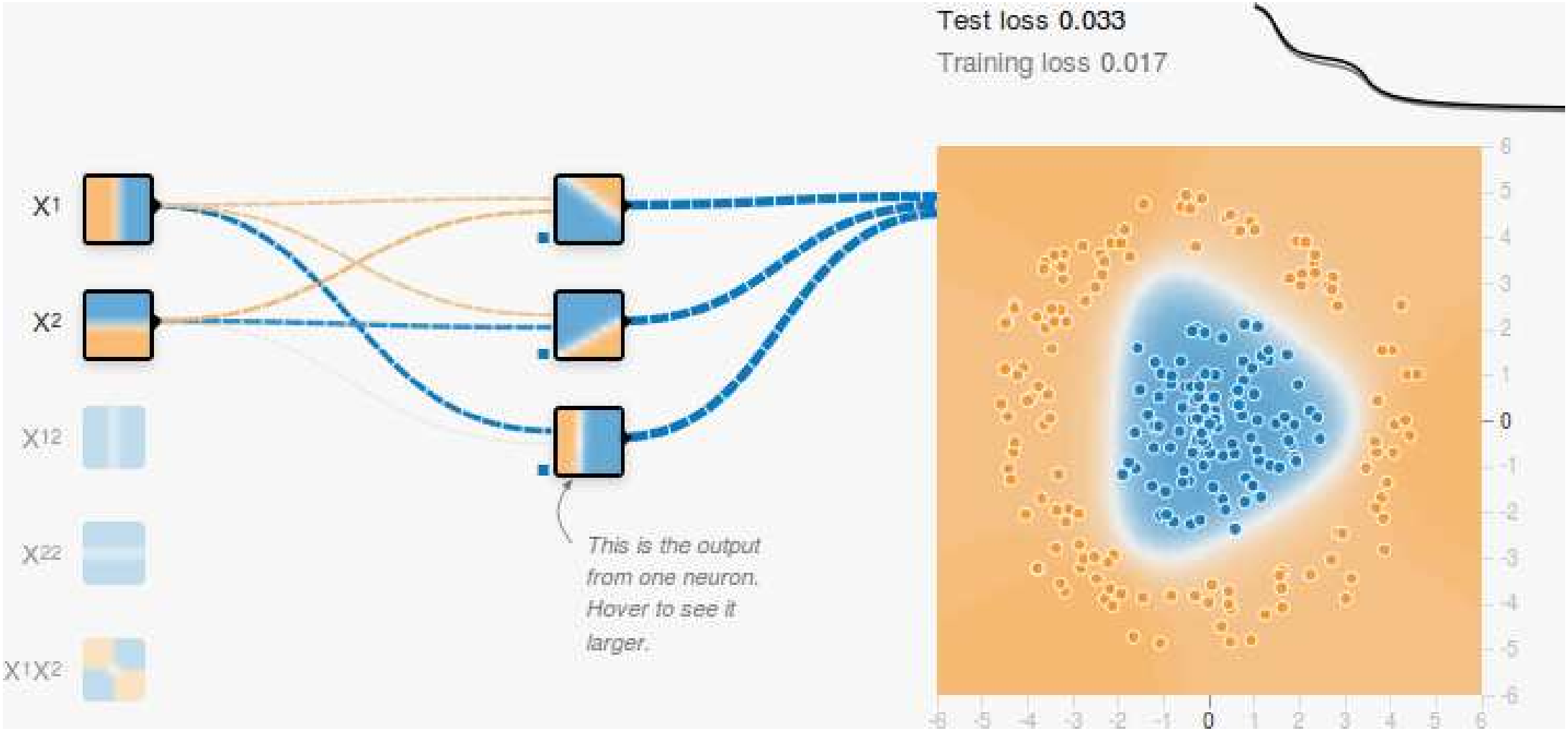
MT **Kriminalisté Minsku** hvězdu Thomase Gibsona **vyhostili** po **zásahu** producenta

SRC ...add to that its long-standing grudge...

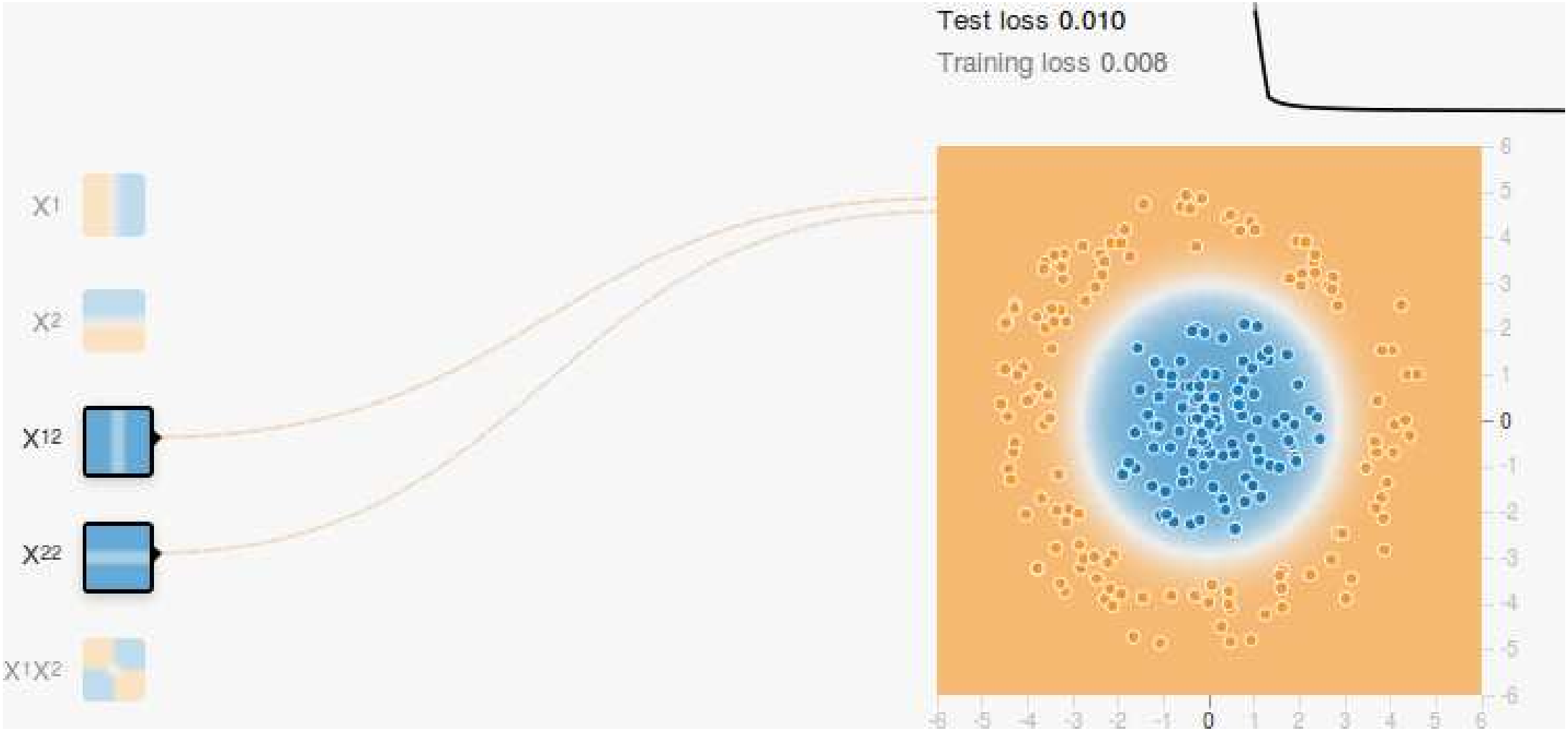
REF ...přidejte k tomu svou dlouholetou nenávist...

MT ...přidejte k tomu svou dlouholetou **záštitu**...
(grudge → zášť → záštita)

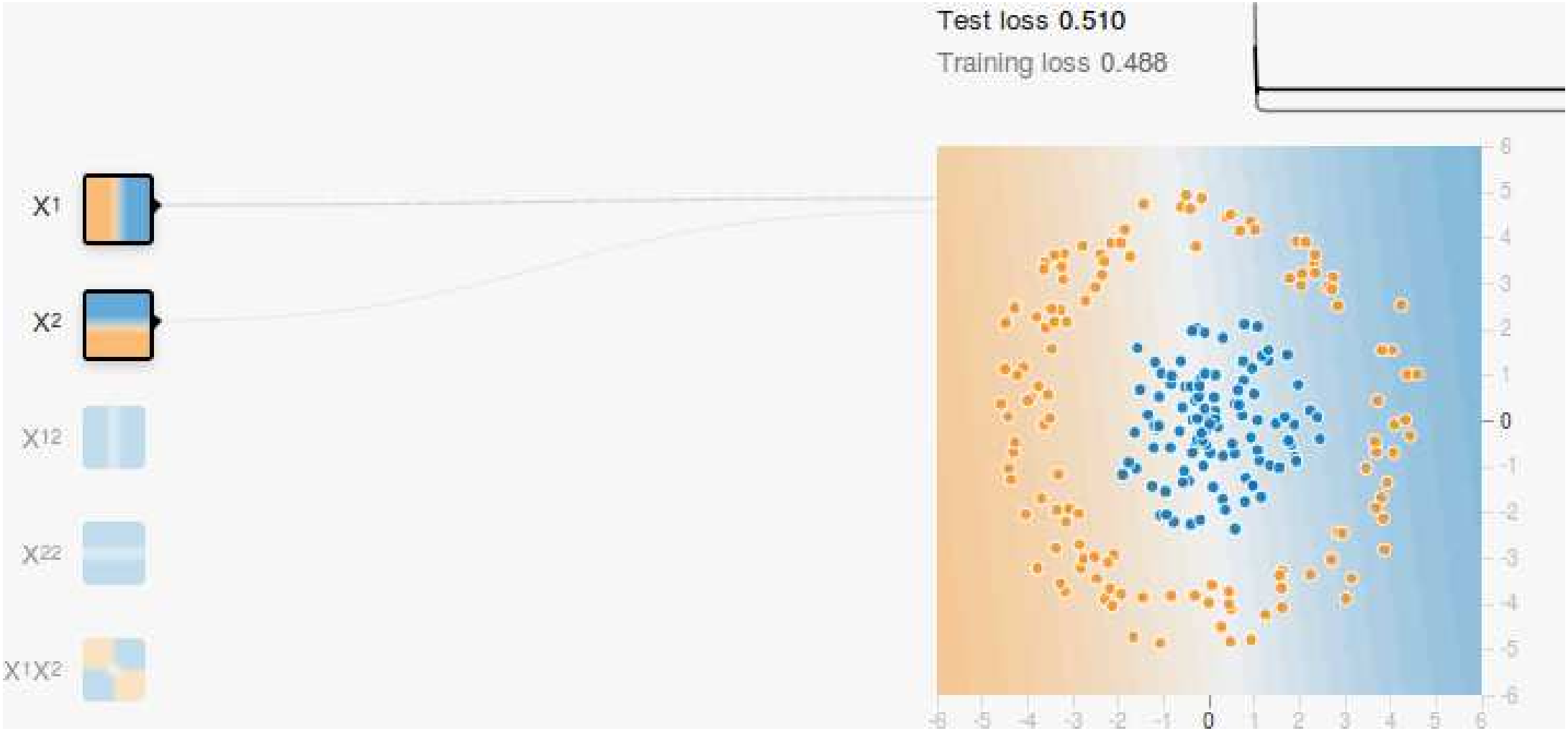
playground.tensorflow.org



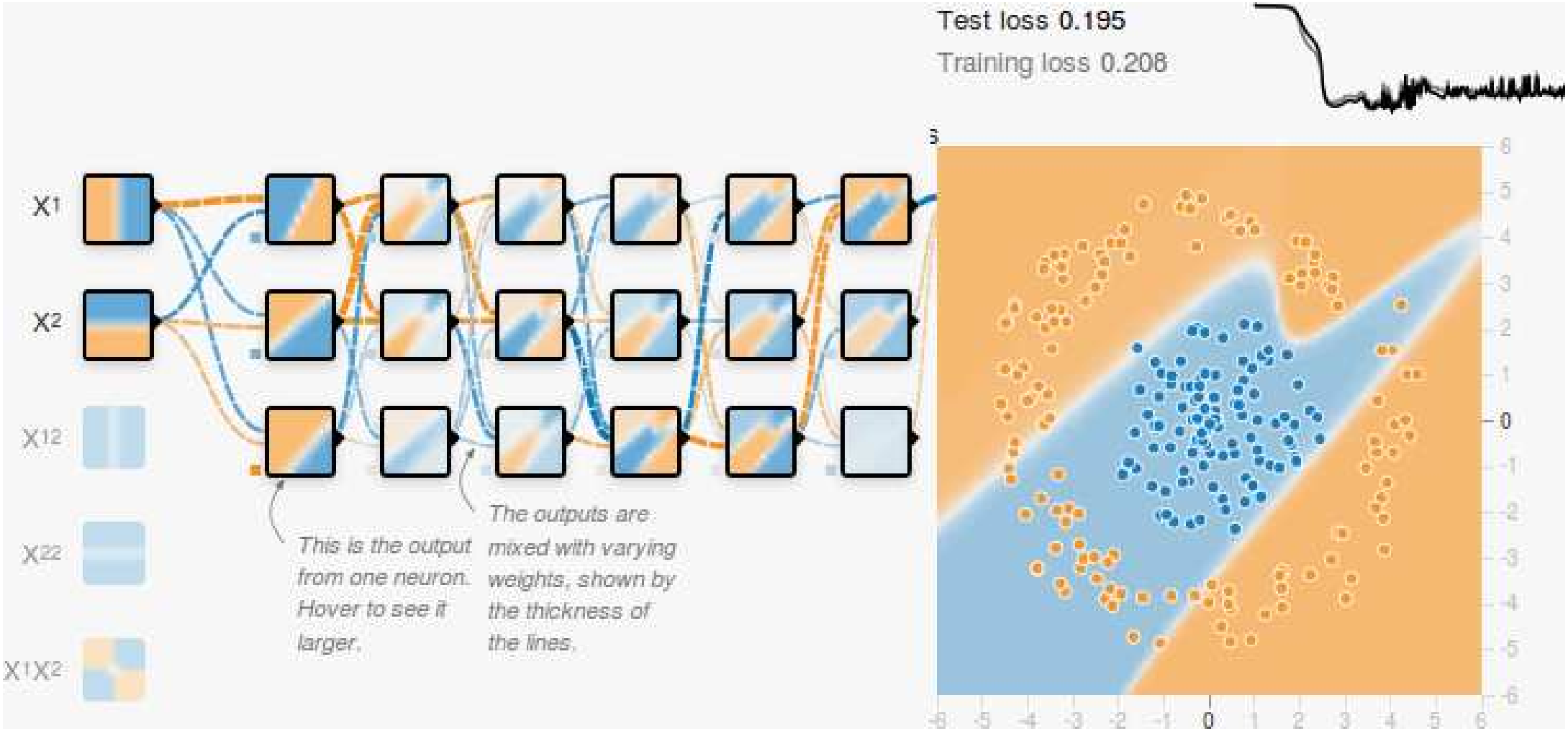
Ideální vstupy



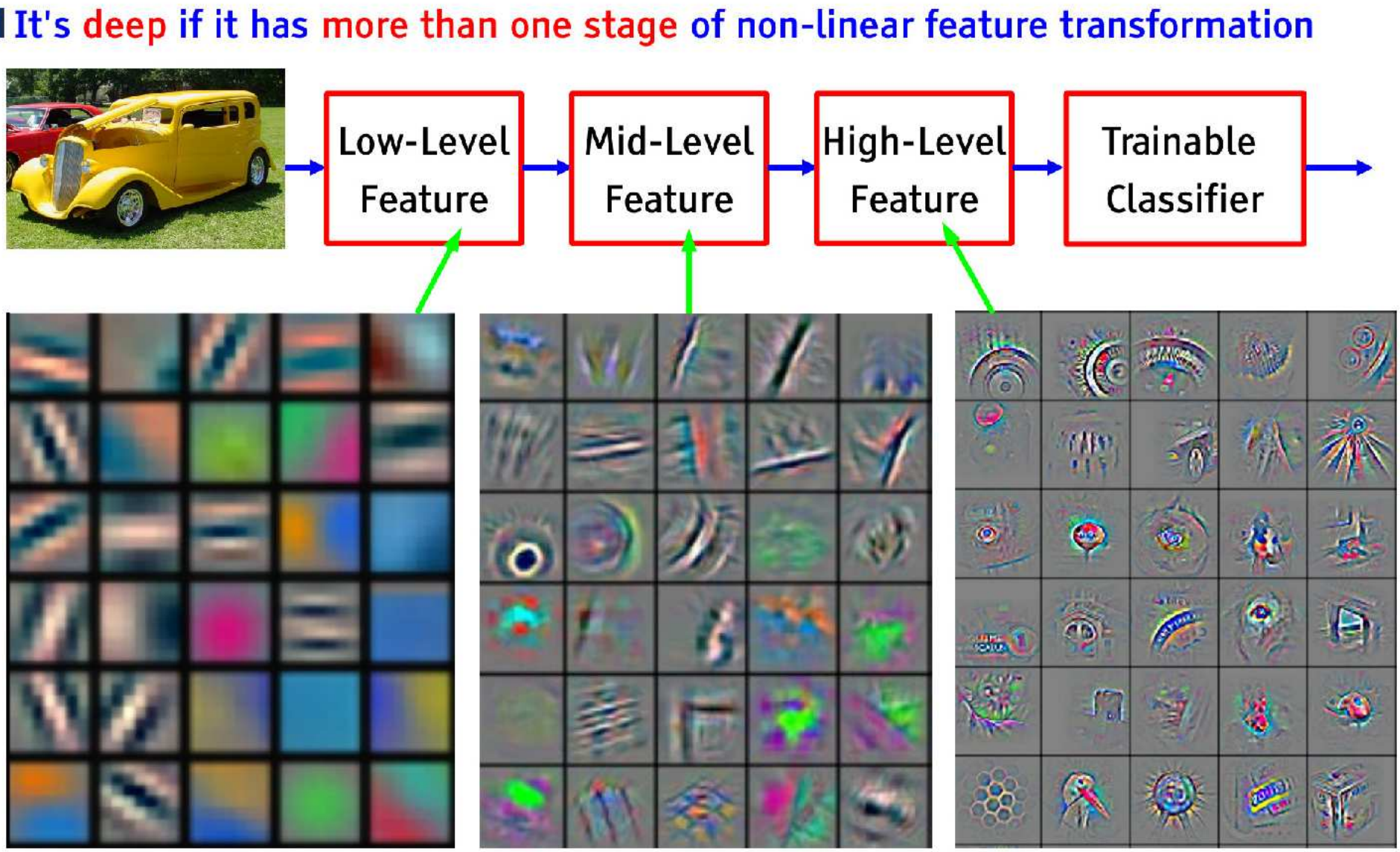
Nevhodné vstupy a malá hloubka



Příliš složitá síť se nenatrénuje



Hluboké NN pro klasifikaci obrázků



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

Hluboké sítě se učí reprezentaci

- Na základě trénovacích dat
  (ukázkové vstupy a očekávané výstupy)
- se neuronová síť (neural network, NN) sama naučí
- čeho si ve vstupech všímat.

"**Reprezentace**" je nový souřadný systém.

- Místo 3 rozměrů (x, y, barva) dostaneme
- 2000 rozměrů: (slonovitost, počet čápů, modrost, . . .)
- nalezených tak, aby nejvíc pomáhaly uhodnout výstup.

Jedna vrstva $\tanh(Wx + b)$, 2D→2D

Zkosení:

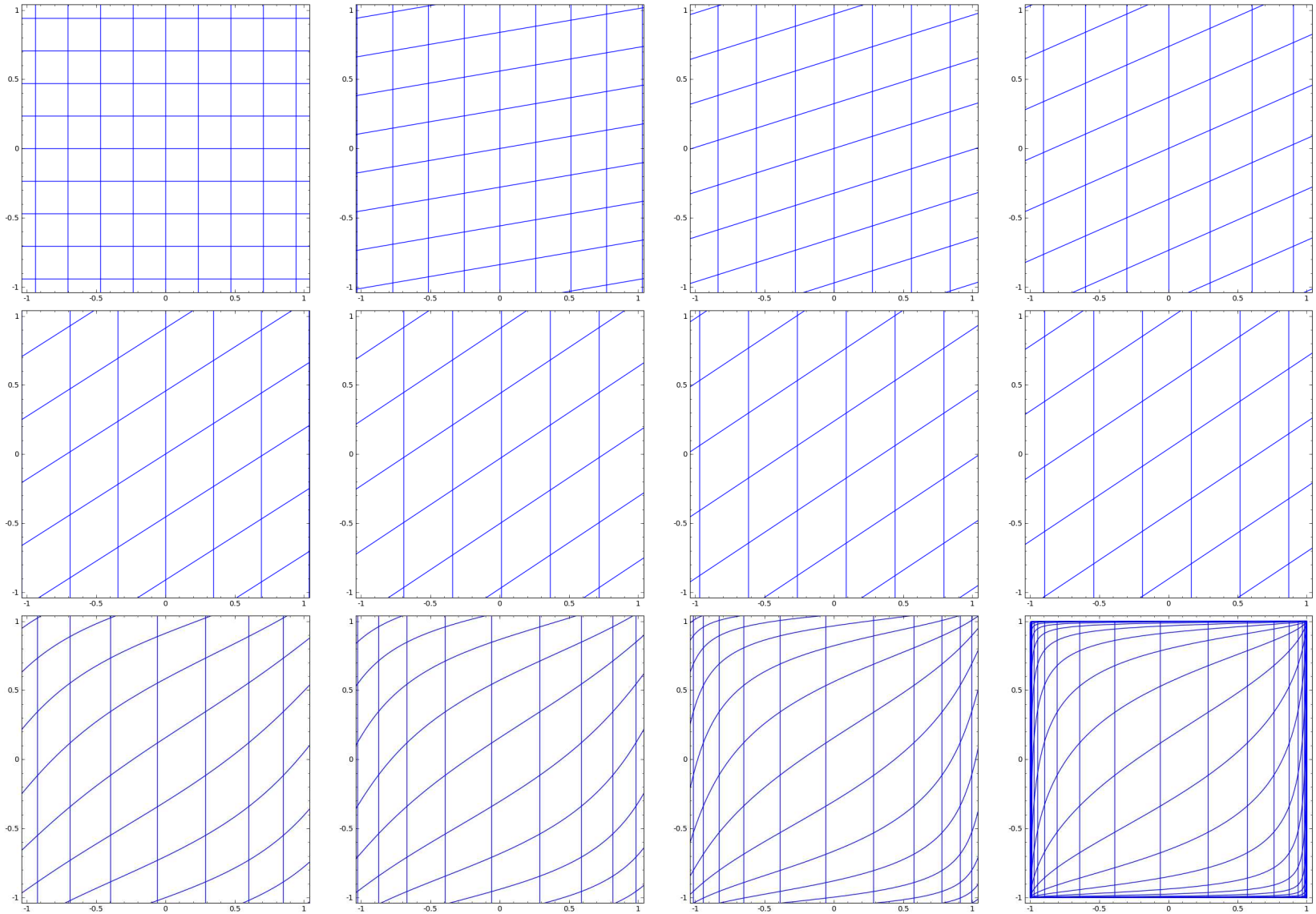
W

Posun:

b

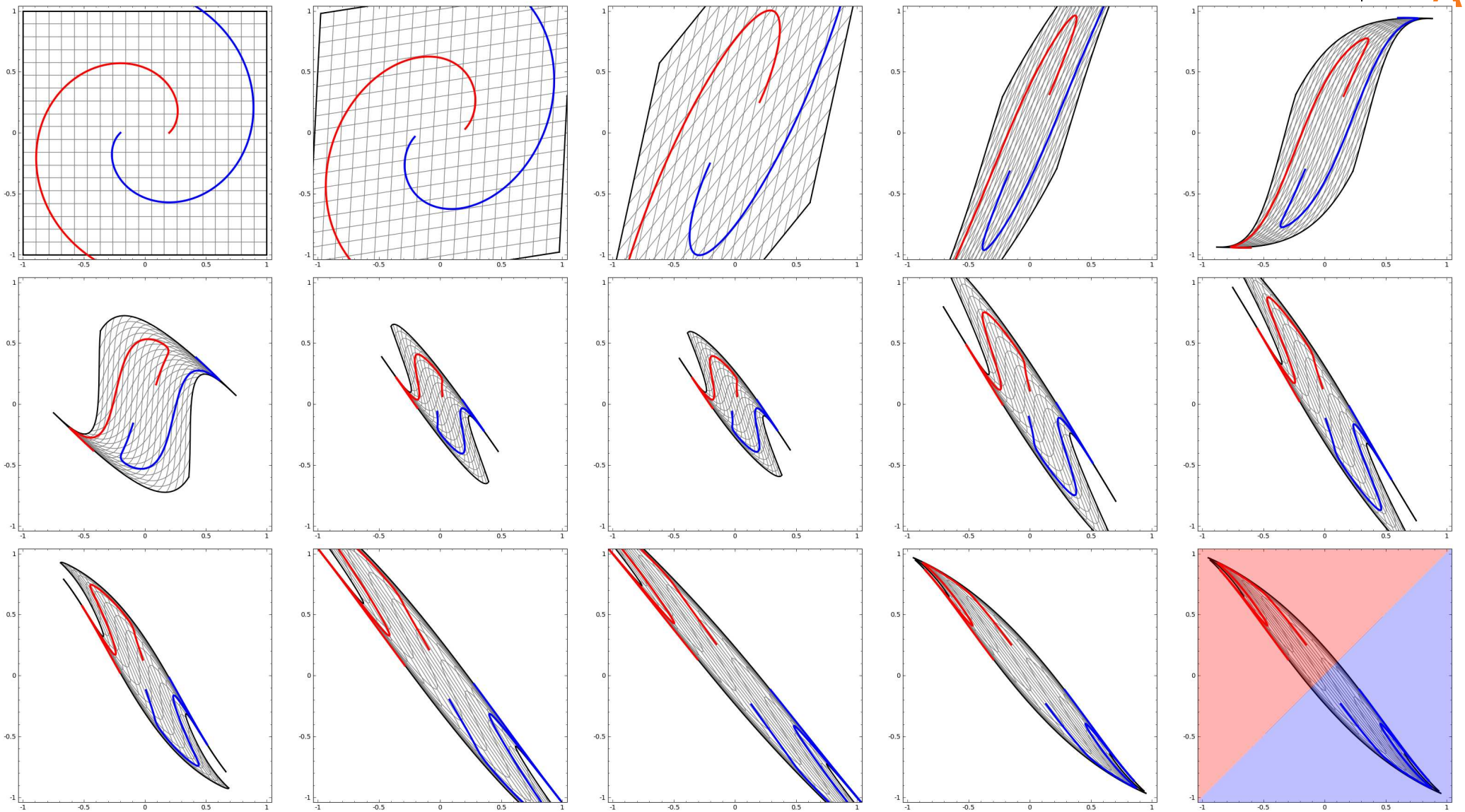
Nelinearita:

$\tanh$



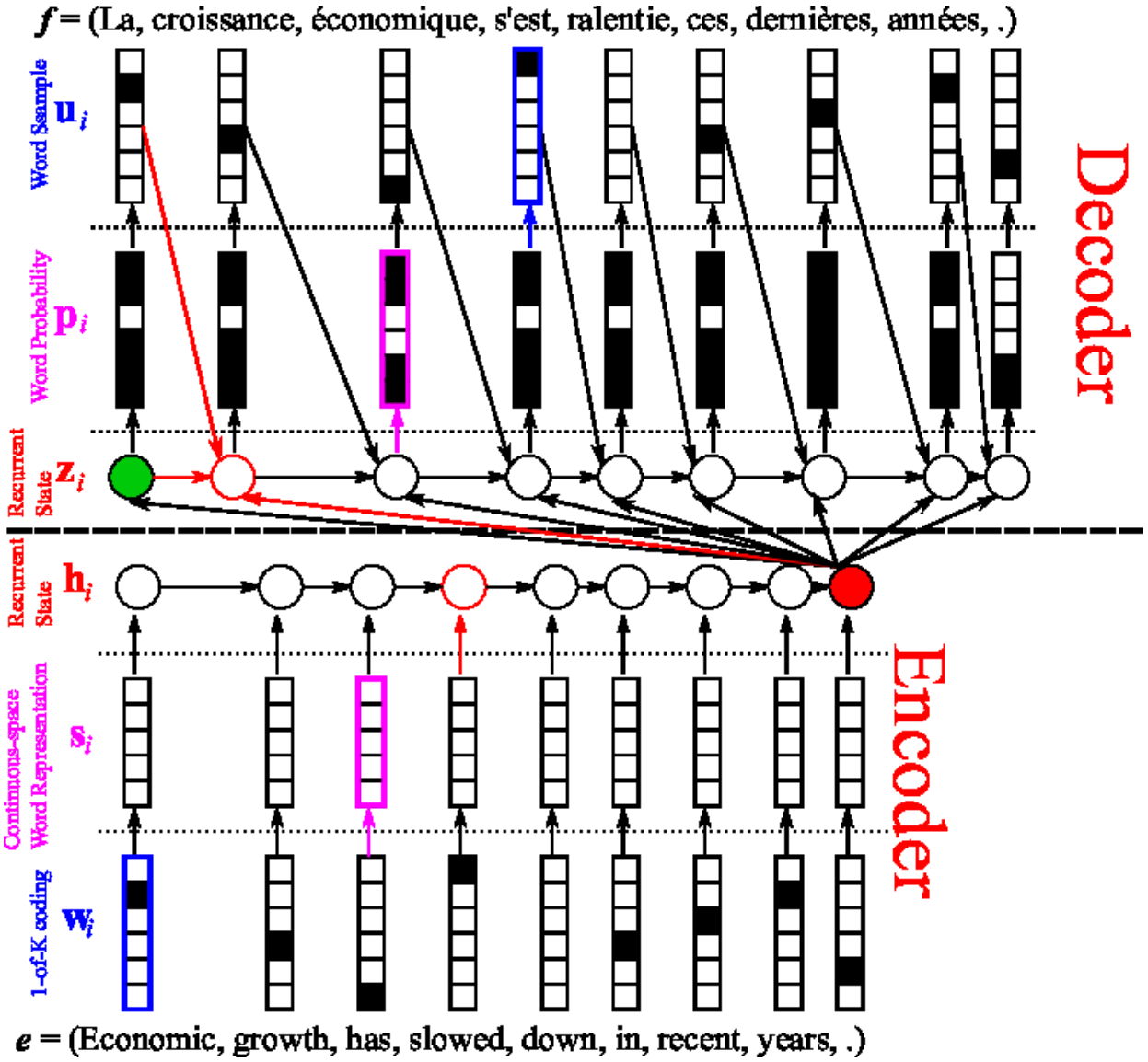
Zdroj animace: `http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/`

Čtyřvrstvá NN oddělí ramena spirály



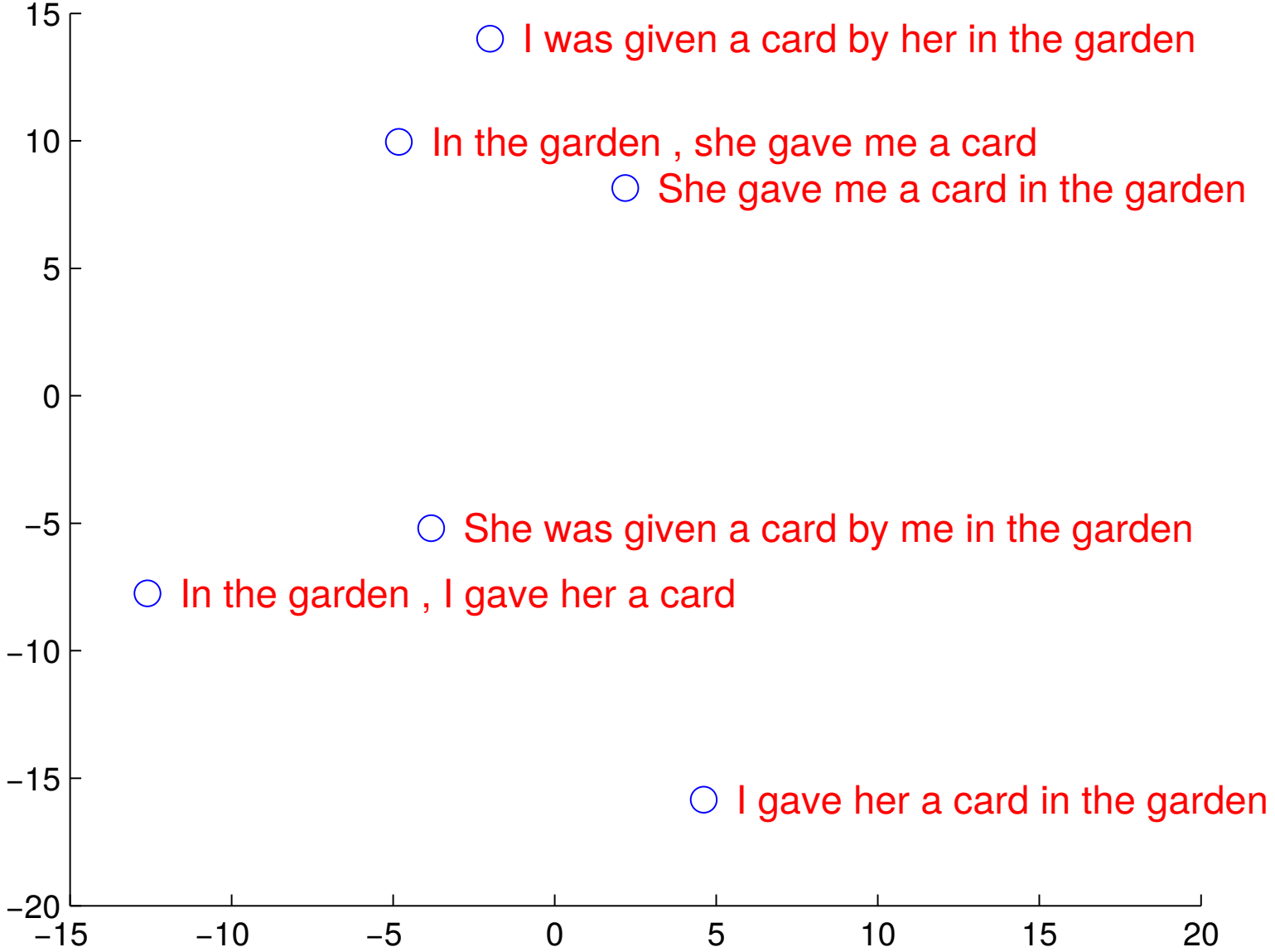
Zdroj animace: `http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/`

NMT: Enkodér-Dekodér



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

Vektorová reprezentace vět



8000-rozměrný prostor reprezentací vět promítnut do 2D (Sutskever et al., 2014).

Automatická sumarizace

Vstup:

legendární slovenská punkrocková kapela extip se letos vrátila na pódia poté, co vyšla v reedici její debutová deska pekný, škaredý deň, kterou přehraje 1. prosince na sedmičce na strahově. soubor nezanikl, i když bratislavskou punkovou scénu v devadesátých letech rozložily drogy. své zkušenosti s tím má kytarista sveto korbel, který odpovídal na otázky novinek.

Lidský výstup:

slovenská punková legenda extip se vrátila

Automatická sumarizace

Vstup:

legendární slovenská punkrocková kapela extip se letos vrátila na pódia poté, co vyšla v reedici její debutová deska pekný, škaredý deň, kterou přehraje 1. prosince na sedmičce na strahově. soubor nezanikl, i když bratislavskou punkovou scénu v devadesátých letech rozložily drogy. své zkušenosti s tím má kytarista sveto korbel, který odpovídal na otázky novinek.

Lidský výstup:

slovenská punková legenda extip se vrátila

“Sumarizace překladem:”

slovenská kapela extip se vrací do prahy

Automatická sumarizace chápe?

Vstup:

legendární slovenská punkrocková kapela extip se letos vrátila na pódia poté, co vyšla v reedici její debutová deska pekný, škaredý deň, kterou přehraje 1. prosince na sedmičce na **strahově**. soubor nezanikl, i když bratislavskou punkovou scénu v devadesátých letech rozložily drogy. své zkušenosti s tím má kytarista sveto korbel, který odpovídal na otázky novinek.

Lidský výstup:

slovenská punková legenda extip se vrátila

“Sumarizace překladem:”

slovenská kapela extip se vrací do **prahy**

Automatická sumarizace nechápe.

na strahově	slovenská kapela extip se vrací do **prahy**
v o2 aréně	slovenská kapela extip se vrací do **prahy**
na hradecku	slovenská kapela extip se vrací do **čech**
u vajgaru	slovenská kapela extip se vrací do **prahy**

Automatická sumarizace nechápe.

na strahově	slovenská kapela extip se vrací do **prahy**
v o2 aréně	slovenská kapela extip se vrací do **prahy**
na hradecku	slovenská kapela extip se vrací do **čech**
u vajgaru	slovenská kapela extip se vrací do **prahy**
ve stromovce	slovenská kapela extip se vrací na scénu. tentokrát kvůli drogám v reedici. s. s. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. i. m. . . . . m. . . . m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. . m. m. m. m. m. m. m. m. m. m.

Selhávají ale i lidé

Selhávají ale i lidé

Co čekat od hlubokého učení?

- Mimořádný skok v plynulosti výstupu.
- Velký skok i v adekvátnosti překladu.
- Často nelze odlišit lidský a strojový překlad.
- Přesto pravděpodobně dosud žádné strojové porozumění.

- Pro řadu úloh jsou dostupná velká trénovací data.
- Hluboké učení stále dostupnější.

$\Rightarrow$ Lze očekávat výrazný nástup dalších aplikací.

  – Strojové slohovky, strojové souhrny, strojové odpovědi, strojové referáty, strojové recenze, . . .

Budoucí role wikipedistů?

- Strojové zpracování přirozené řeči výrazně usnadní práci.
- Především bude nutno zajistit věcnou kontrolu.

Dvě možné cesty:

- Zpracovávat víc stránek.
- Zpracovávat stránky pečlivěji.

I v pečlivosti mohou počítače pomoct:

- Módní heslo: Fact checking.
  - Kontrola faktů proti databázím (∼`wikidata.org`).
  - Kontrola tvrzení ve větách (∼`demagog.cz`).

Shrnutí

- Strojový překlad je těžký.
- Hluboké neuronové sítě jej přesto často zvládají skvěle.

. . . I překladatelé ostatně dělají chyby.

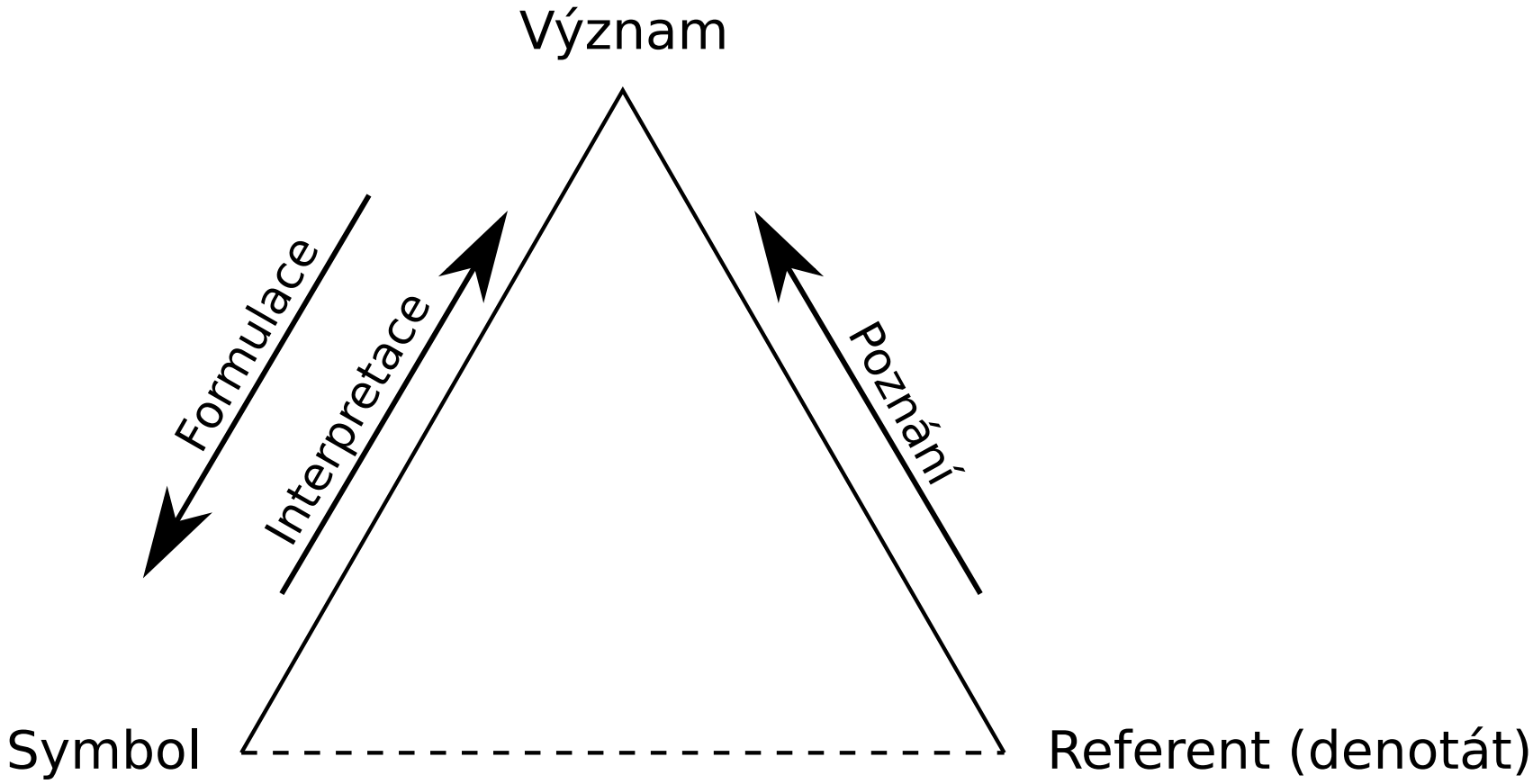
- Strojové porozumění ještě úplně nemáme.

. . . ale učení se reprezentacím má k tomu dobré předpoklady.

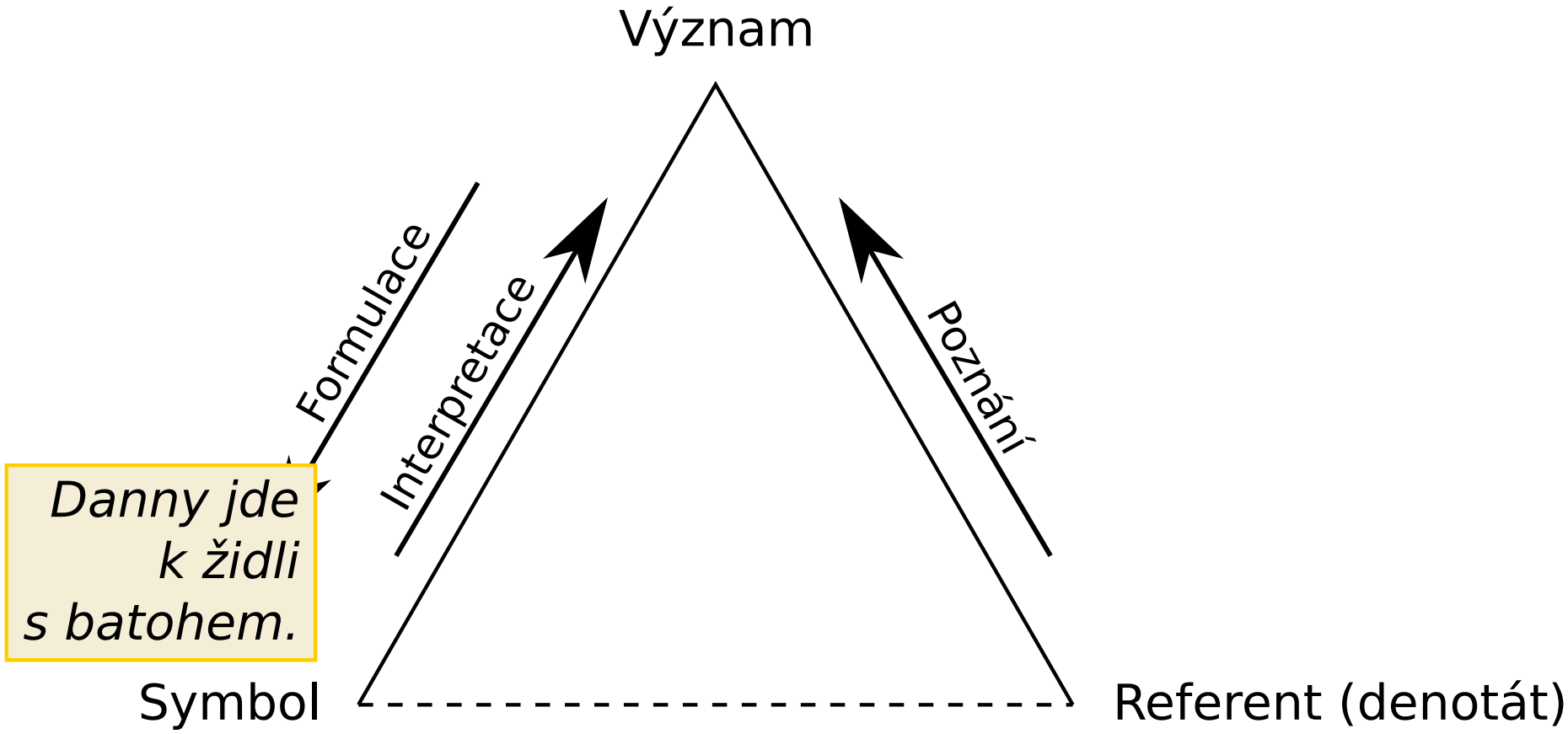
- Nabízí se produkovat více textů.

. . . ale lepší by bylo soustředit se na lepší texty.

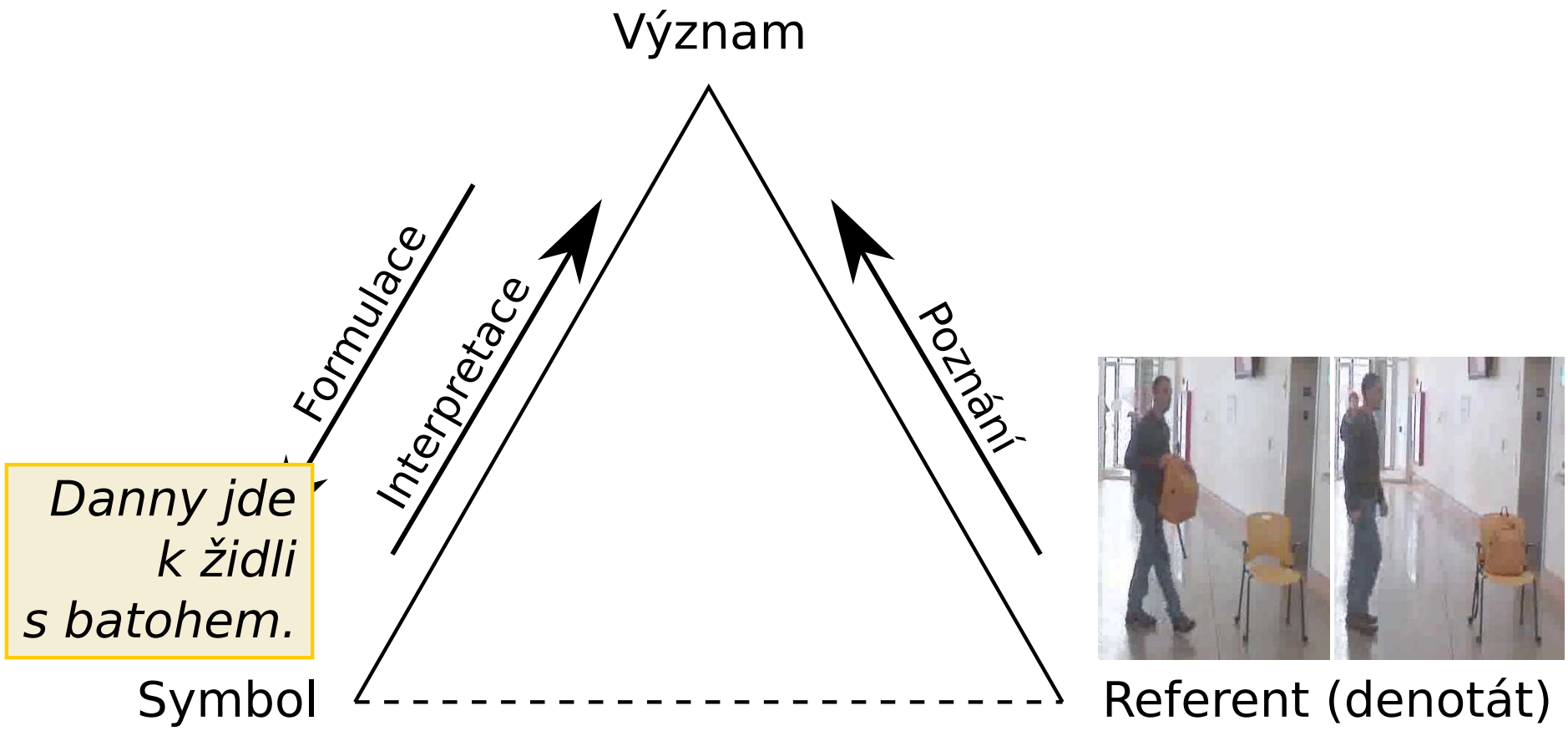
Sémantický trojúhelník



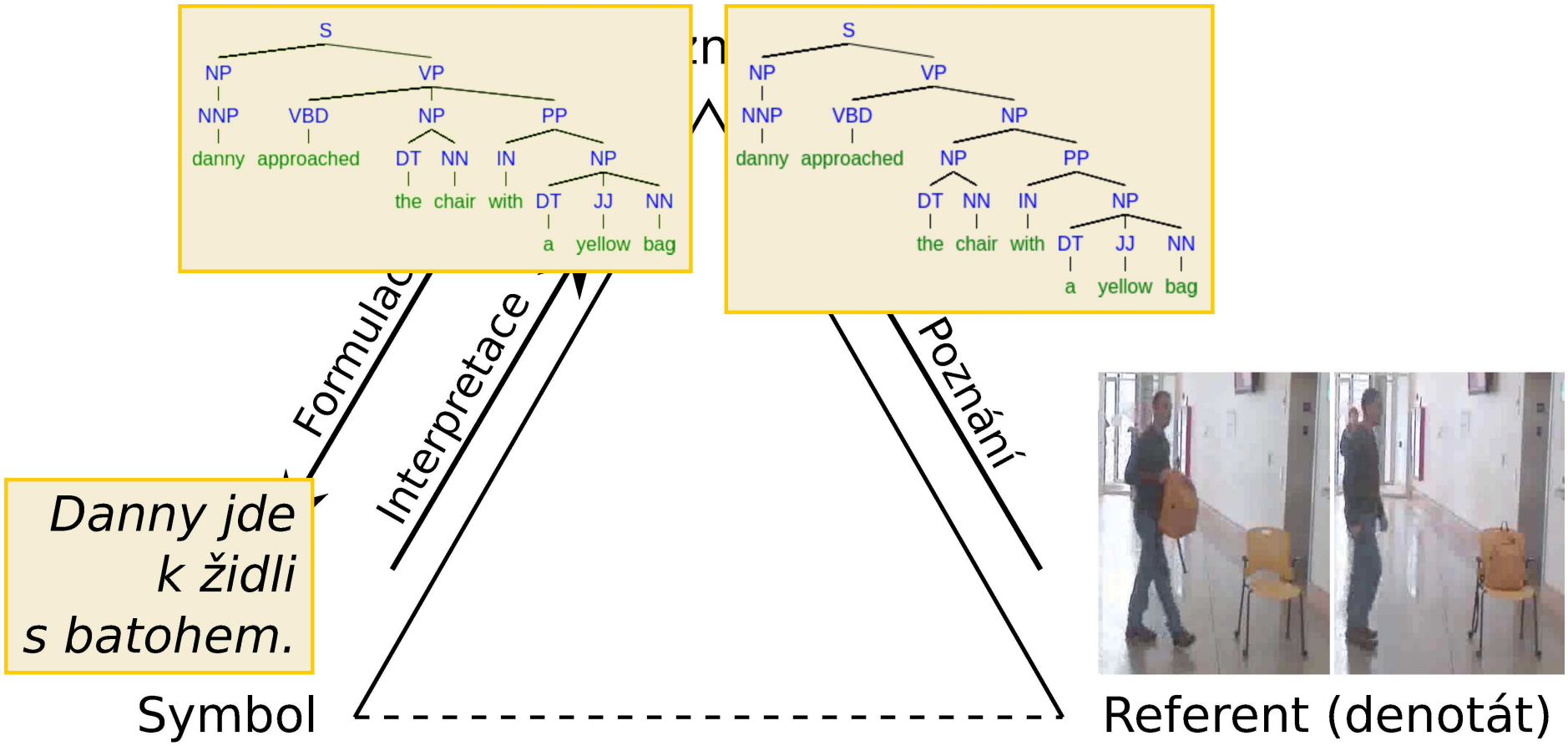
Sémantický trojúhelník



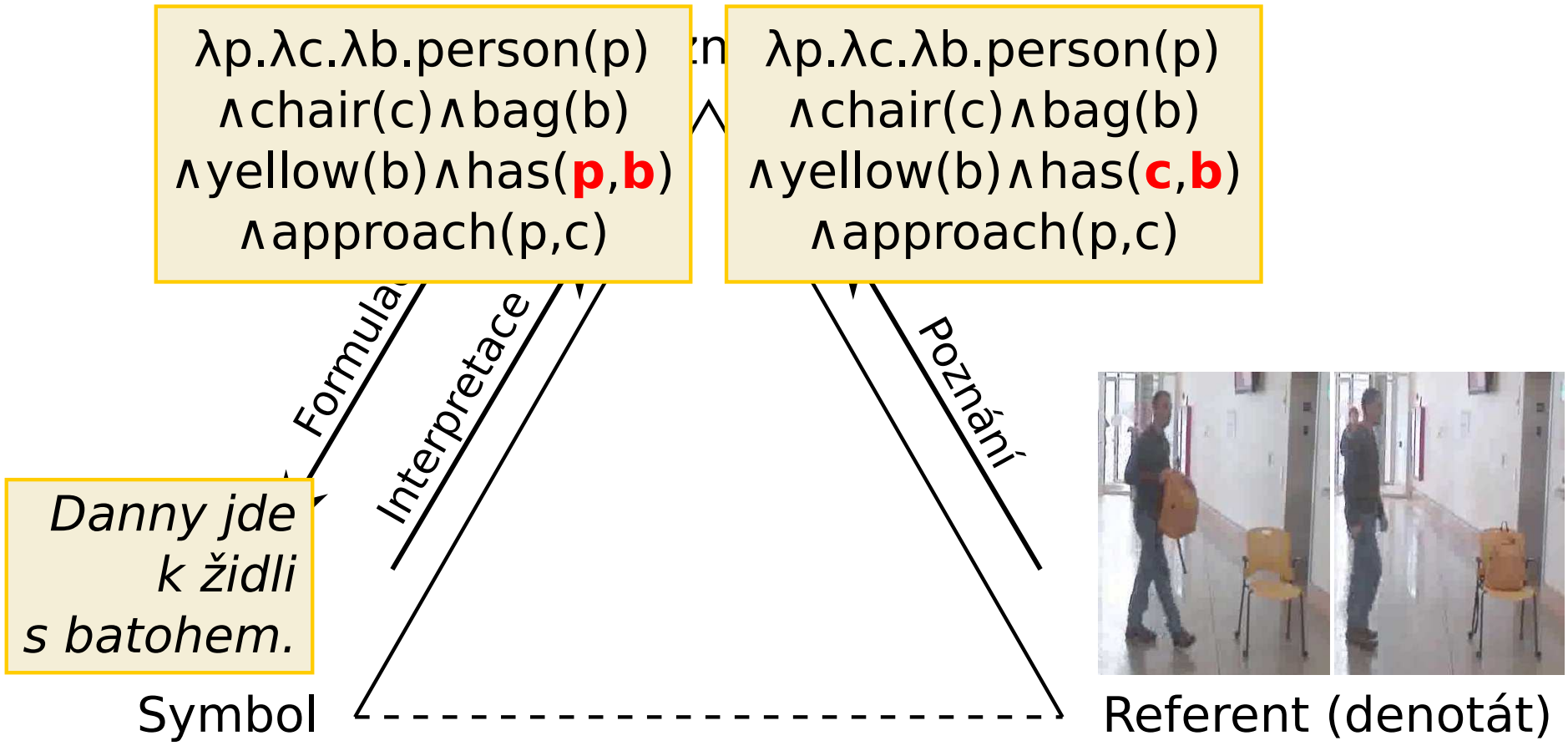
Sémantický trojúhelník



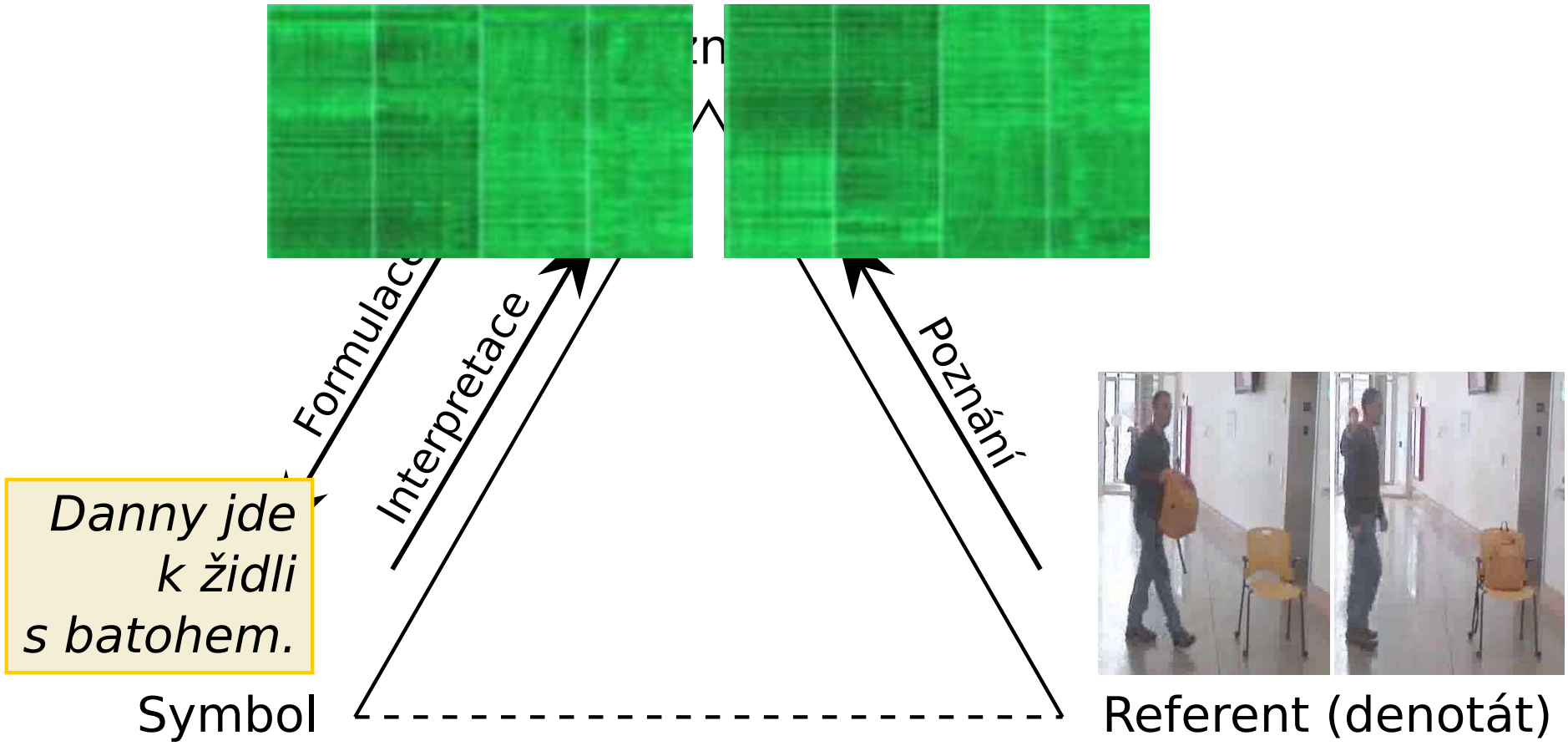
Sémantický trojúhelník



Sémantický trojúhelník



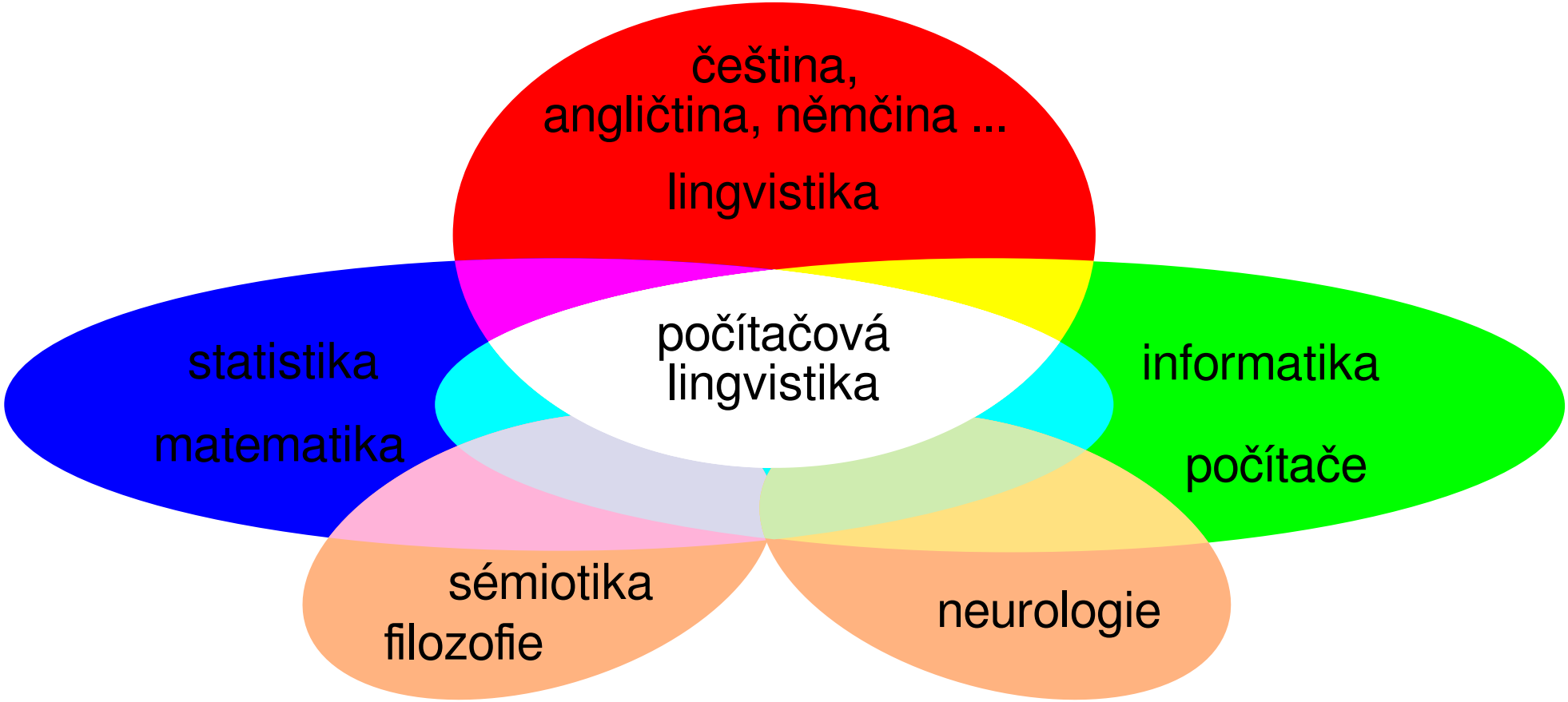
Sémantický trojúhelník



Počítačová lingvistika

...spojí filozofii...

...spojí filozofii s neurologií.



Shrnutí

- Strojový překlad je těžký.
- Hluboké neuronové sítě jej přesto často zvládají skvěle.

. . . I překladatelé ostatně dělají chyby.

- Strojové porozumění ještě úplně nemáme.

. . . ale učení se reprezentacím má k tomu dobré předpoklady.

- Nabízí se produkovat více textů.

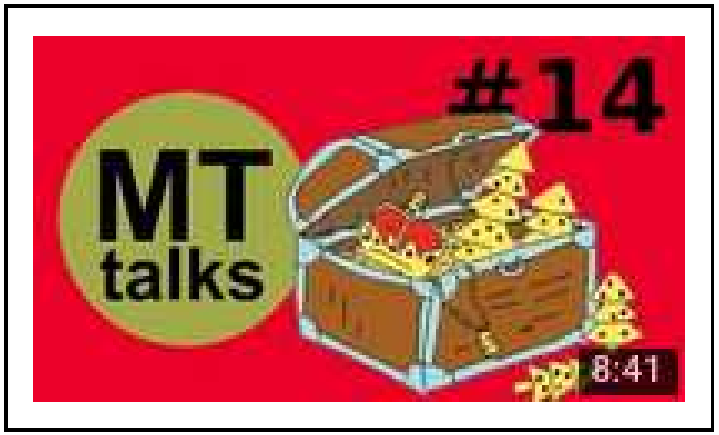
. . . ale lepší by bylo soustředit se na lepší texty.

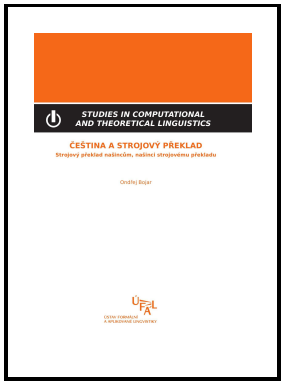
Chcete-li vědět víc

`http://ufal.mff.cuni.cz/`

O (předneuronovém) překladu:

Videa `http://mttalks.ufal.cz/`





Knížka *Čeština a strojový překlad: Strojový překlad našincům, našinci strojovému překladu.*
Knihkupectví Karolinum (Celetná 18; eshop)

References

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. pages 3104–3112.