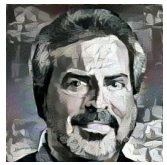
Deep Learning for Text

Take two LSTMs and call me in the morning

HOLBERTON
school()

Class 3 - Q2 - 2016



Gregory Renard
@redo
https://www.linkedin.com/in/gregoryrenard

Louis Monier
@louis_monier
https://www.linkedin.com/in/louismonier

Natural Language Processing (NLP)

Parsing words

Spell checking

Finding synonyms

Part of Speech (POS) Tagging

Classification
- Encoding and language detection
- Sentiment analysis
- Spam detection
- Matching ad to content

Extracting entities (people, places…)

Full-text Search

Summarization

Automated Translation

Question Answering

Virtual Assistant
- Siri, Watson, Alexa, … QA
- Her, HAL, Sarah, … Empathy & Emotion

How to Represent Words?

String: “p l a t y p u s”

Opaque ID:

One-hot vector: [0 0 … **1** … 0 0]

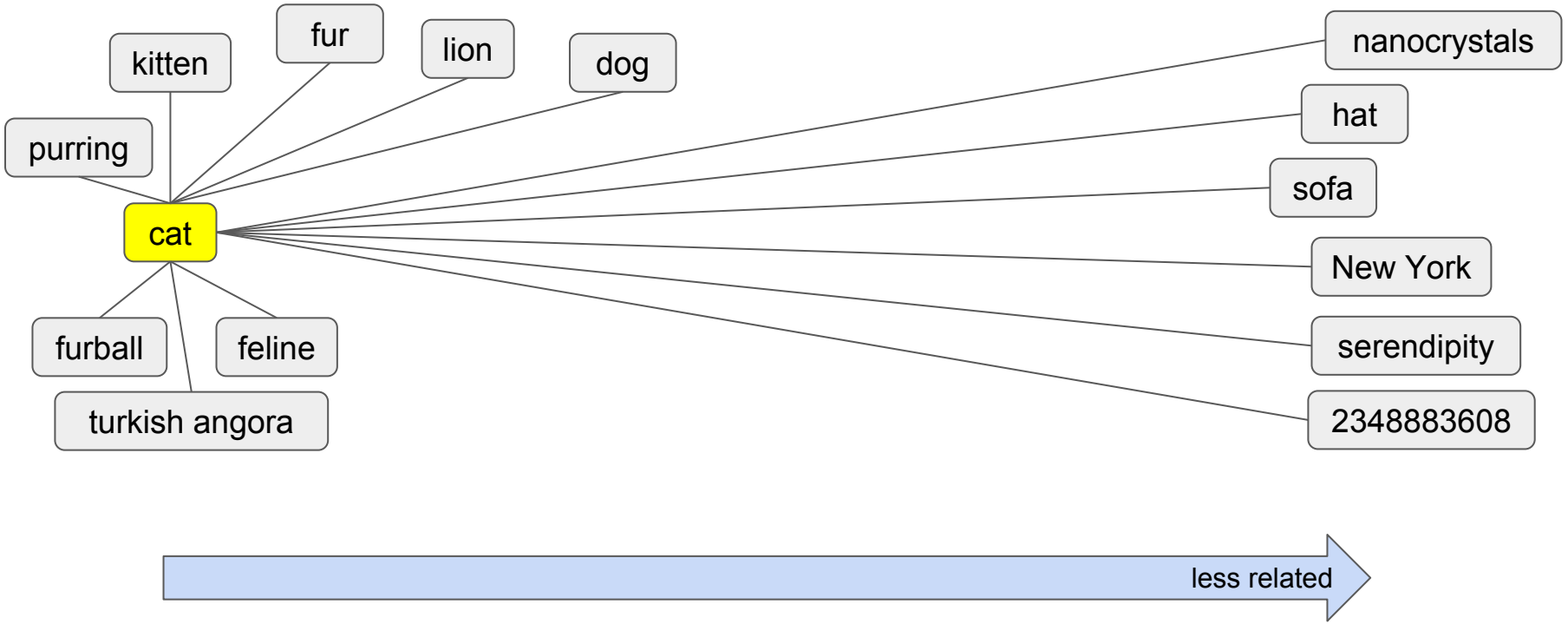
Distributed Representation

[
mammal
striped tail
hand-like front paws
washes its food
facial mask
non-retractile claws
omnivorous
…
]

Semantic distance

Distributed Representations
Word Embeddings
Word Vectors

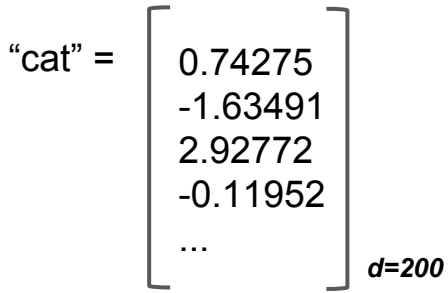
Semantic Distance between Words

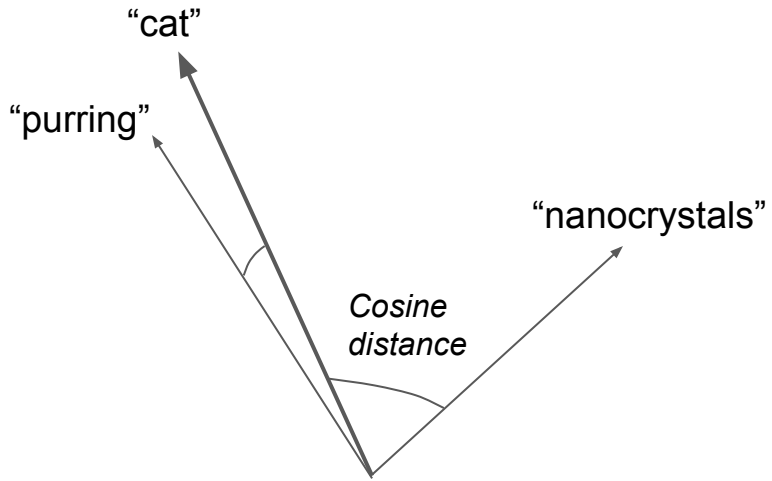


Not to scale :)

Word2vec - Tomas Mikolov - 2013

Lots of text, pick a dimension (~50 to 4000) => set of vectors





Terms similar to ***Champagne***

french champagne, cognac, champagne's, champagnes, veuve clicquot, cremant, louis roederer, rosé, taittinger, fine champagne, champagne wine, sparkling wines, dom pérignon, dom perignon, pol roger, vintage champagne, bubblies, pommery, rose wine, pink wine, blancs, french wine, cliquot, beaujolais nouveau, sancerre, sparkling, burgundy, chateau, chablis, cognacs, pink champagne, domaine, moët, methode champenoise, burgundy wines, apéritif, armagnac, chandon, champenoise, beaujolais, heidsieck, marnier, wine, bourgogne, aperitif, chateau margaux, demi-sec, moelleux, champagne cocktail, crémant, half-bottle, cuvée, brut, ruinart, champagne flute, st emilion, white wine, loire valley, wine cocktail, veuve, drinking champagne, french wines, blanc, chardonnay wine, champagne glass, cuvées, mauzac, roederer estate, laurent-perrier, puligny, negociant, prosecco, rose wines, gloria ferrer, red wine, musigny, coteaux, corton-charlemagne, fine wine, dessert wine, bordeaux, champagne glasses, cheval blanc, champagne flutes, cuvees, champange, four wines, montlouis, rémy martin, primeur, fine wines, lirac, d'yquem, burgundy wine, red bordeaux, brandy, cuvee, white burgundy, chardonnay, chambolle-musigny, cheverny, great vintages, yquem, special wines, wonderful wines, burgundies, half bottles, grand marnier, grand cru, primeurs, sauterne, minervois, pouilly-fuissé, sauternes, chambertin, white bordeaux, vougeot, epernay, vin gris, chalonnaise, quaffer, loire, sweet white wine, d'aunis, côtes, gevrey-chambertin, limoux, english wine, chateaux, château haut-brion, blanche, pinot meunier, six glasses, mâconnais, épernay, bourbon, sparkler, volnay, white wines, chassagne-montrachet, burgundys, vin jaune, claret, beaune, grande champagne, white grapes, bordeaux wine, dessert wines, crème de cassis, pinot noir grapes, chardonnay grapes, armand de brignac, select wines, calvados, country wine, muscadet, leflaive, reisling, cointreau, own wine, caveau, clos de vougeot, inexpensive wines, vosne-romanée..., expensive wines, red burgundy, barsac, delicious wine, wine flight, puligny-montrachet, rousanne, châteauneuf-du-pape, liqueur, schramsberg, touraine, montrachet, arbois, lanson, vintage wine, chateauneuf, blanquette, non-vintage, orange wine, three wines, wine.the, banyuls, merlot wine, vendange, red table wine, sweet wines, santenay, languedoc, moscato d'asti …

Terms similar to ***Brad Pitt***

angelina jolie, george clooney, cameron diaz, julia roberts, leonardo dicaprio, matt damon, tom cruise, nicole kidman, reese witherspoon, charlize theron, jennifer aniston, halle berry, kate winslet, jessica biel, ben affleck, bruce willis, scarlett johansson, uma thurman, matthew mcconaughey, jake gyllenhaal, sandra bullock, oscar winner, gwyneth paltrow, sean penn, demi moore, naomi watts, colin farrell, mickey rourke, orlando bloom, bradley cooper, natalie portman, jennifer garner, tom hanks, dicaprio, jessica chastain, robert de niro, julianne moore, leo dicaprio, channing tatum, kirsten dunst, jessica alba, emily blunt, salma hayek, ryan gosling, mark wahlberg, renee zellweger, drew barrymore, renée zellweger, gerard butler, hilary swank, ryan phillippe, john malkovich, nicolas cage, kate hudson, sharon stone, sienna miller, *new movie*, kim basinger, robert downey jr, keira knightley, ryan reynolds, johnny depp, jennifer connelly, edward norton, emma stone, don cheadle, marisa tomei, jason statham, eva mendes, kate beckinsale, *oscar-winner*, katie holmes, kelly preston, denzel washington, zac efron, clive owen, *oscar-winning*, forest whitaker, penelope cruz, ashton kutcher, sigourney weaver, rachel weisz, billy bob thornton, catherine zeta-jones, benicio del toro, keanu reeves, *new film*, ewan mcgregor, jeremy renner, hugh grant, liam neeson, scarlett johannson, jude law, russell crowe, jodie foster, harrison ford, meryl streep, justin theroux, john travolta, christian bale, emile hirsch, adrien brody, jonah hill, nick nolte, dennis quaid, liv tyler, kate bosworth, *hollywood star*, amber heard, javier bardem, robert deniro, evan rachel wood, helen mirren, milla jovovich, blake lively, james franco, vince vaughn, joaquin phoenix, diane kruger, *upcoming movie*, robert pattinson, michael douglas, courteney cox, richard gere, daniel craig, sylvester stallone, *latest movie*, rachel mcadams, josh brolin, jennifer lawrence, brangelina, *oscar winners*, hugh jackman, zoe saldana, *oscar nominee*, dakota fanning, josh hartnett, annette bening, mila kunis, emma watson, david fincher, megan fox, quentin tarantino, ben stiller, a-lister, kristen stewart, charlie sheen, christoph waltz, christopher walken, michelle pfeiffer, phillip seymour hoffman, thandie newton, amanda seyfried, ethan hawke, liam hemsworth, morgan freeman, robert downey, owen wilson, olivia wilde, costars, paula patton, casey affleck, kevin costner, clooney, clooneys, andrew garfield …

Terms similar to ***greenish***

bluish, pinkish, yellowish, reddish, brownish, purplish, grayish, yellow-green, orange-yellow, yellow-brown, yellowish green, reddish brown, orange-red, pale green, whitish, reddish-brown, greenish yellow, mottled, pale yellow, greenish-brown, greenish-yellow, yellow-orange, orangish, red-brown, bluish-green, dark brown, greyish, yellowish-green, bluish-black, reddish-orange, orange-brown, yellowish-orange, yellowish-white, brownish red, pale orange, bright yellow, deep yellow, blue-green, paler, brownish-red, bluish-grey, blueish, green-brown, pinkish-brown, golden yellow, blotches, yellowish-brown, brownish-yellow, golden-yellow, pale, grayish-white, coppery, creamy yellow, greyish-white, pale gray, purple-brown, olive-green, pale brown, blackish, brownish yellow, tinge, dark purple, light yellow, red-orange, dark red, rusty brown, brownish black, purplish-red, mottling, bluish-gray, yellowish brown, greyish-green, dull red, dark green, creamy white, purple-black, yellow brown, pinkish red, greenish-blue, reddish purple, bright red, reddish-purple, grayish-green, greenish-white, pale cream, creamy-white, brownish-gray, white spots, silvery, dark grey, dark orange, purplish-black, grayish-blue, purple-blue, greenish-black, yellow spots, bluish-white, purple-red, pure white, light brown, various shades, grey-brown, pale grey, orange-pink, brownish-black, brick-red, purplish-brown, olive-brown, brown colour, speckling, pale blue, brownish gray, deep orange, grayish-brown, blue-black, darker spots, brown-red, yellow patches, gray-black, coloration, reddish color, bluish-purple, green patches, pale red, chestnut-brown, brown streaks, yellow green, lemon yellow, pinkish-red, flecks, dark reddish brown, black spots, grey-black, lemon-yellow, pinkish-white, deep red, brownish-grey, dull black, purple spots, darker green, red spots, blue-grey, splotches, grey-green, pink-purple, greenish-gray, violet-blue, silvery grey, chocolate-brown, yellowish color, cream-coloured, orange brown, small white spots, light orange, brown-grey, violaceous, dark-brown, streaked, green veins, olive brown, olive green, brown markings, gray-green, pale pink, dark blotches, light green, grey-white, dark markings, brilliant red, light violet, blackish-brown, greyish-brown, color ranges, brown-black, orange red, yellow colour, yellow color, red brown, orange markings, small black spots, veined, brick red …

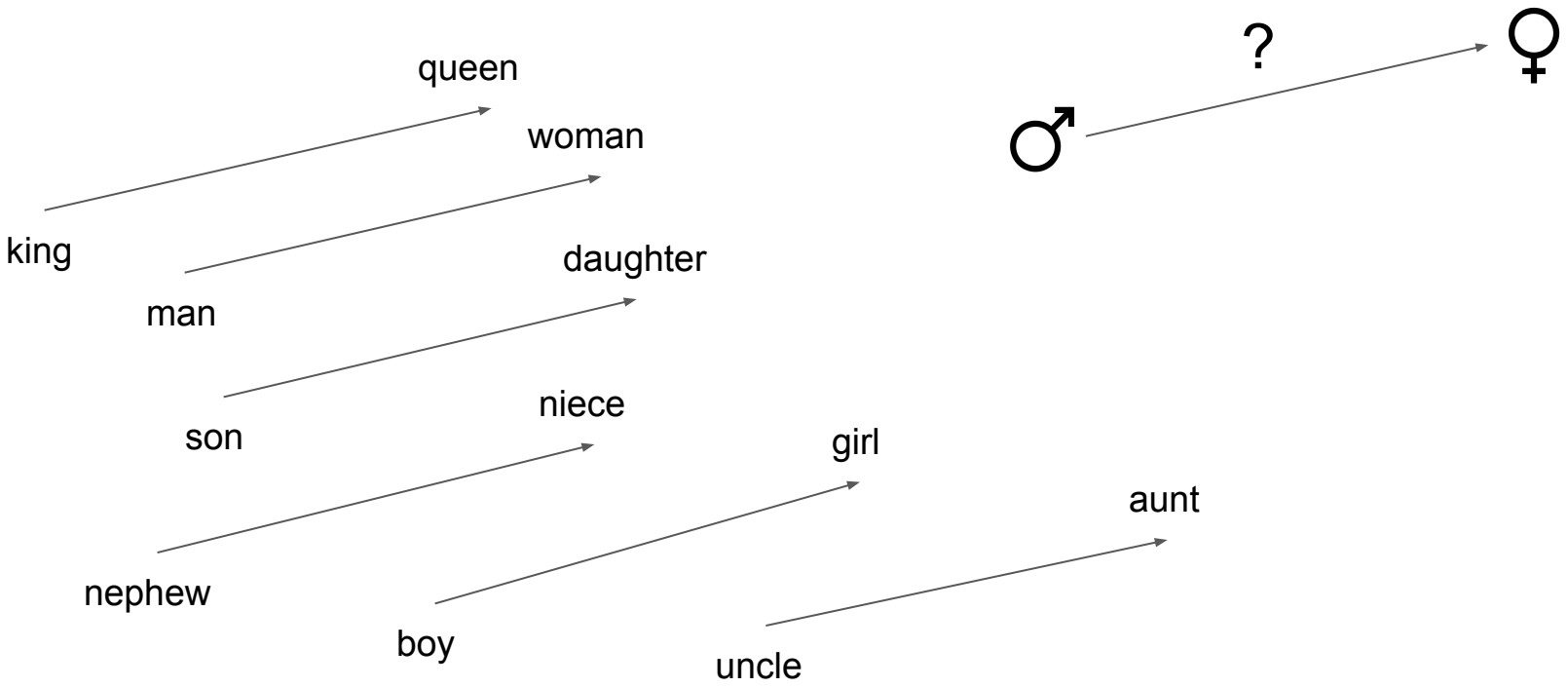
Terms similar to ***worse***

even worse, far worse, very bad, horrible, terrible, awful, horrendous, bigger problem, suffer, things worse, horribly, unfortunate, better, worst, bad, complain, real problem, after all, unfortunately, no good, too, lousy, atrocious, even less, even so, very poor, far more serious, miserable, intolerable, terribly, serious problem, trouble, worrying, bothering, blame, no better, worsened, bother, worse off, dreadful, hardly, horrid, big problem, real concern, fortunately, main problem, sooner, major problem, hopeless, excuse, serious problems, way worse, complaining, horrendously, abysmal, better off, worried, inevitable, wrong, marginally, even, rid, frankly, anymore, bothered, bothers, worry, uglier, sadly, even more, worsen, severe, serious, unacceptable, badly, nasty, different story, worse problems, main reason, worst thing, far less, go away, hurt, obviously, seriously, serious trouble, hurting, gotten, anyone else, worse.it, anyway, happen, worst cases, *say nothing*, appalling, main concern, somehow, obvious reason, troubling, simple fact, unbearable, problematic, huge problem, worst one, exacerbated, afraid, tired, blaming, painfully, suffers, much, ironically, do anything, embarrassing, worse things, inevitably, same problems, bad problems, anything, real reason, everyone else, atrociously, unpleasant, thing, worse again, apparent reason, needlessly, ignore, seemed, horrifically, worth noting, biggest problem, real issue, even more serious, dreadfully, worsening, useless, even though, probably more, some people, pitiful, worrisome, far more, because, deplorable, point out, but, stupid, admittedly, pudgenet, worst part, less so, little improvement, grossly, make things, unnecessarily, too bad, crap, bad thing, laughable, problem, might, trying, exaggerating, pretty much, lot, doing anything, ridiculous, little reason, misguided, exact opposite, worse not better, even when, weren't, inconsequential, simple reason, expect, avoided, something wrong, counter-productive, dismal, appallingly, far more likely, ugly, almost everyone, shame, wonder why, less, polfbroekstraat, worse here, plagued, worse though, honestly, bad situation, nobody, pathetic, certainly, plain wrong, almost nothing …

Terms similar to ***confabulation***

confabulations (0.78), false memories (0.69), autobiographical memory (0.69), psychogenic amnesia (0.69), false memory (0.67), memory retrieval (0.66), delusions (0.65), amnesia (0.65), capgras (0.65), hindsight bias (0.65), temporal lobe epilepsy (0.64), human memory (0.64), hallucination (0.64), reality testing (0.64), hallucinations (0.63), elizabeth loftus (0.63), semantic dementia (0.63), conversion disorder (0.63), perseveration (0.62), reduplicative (0.62), cognitive processing (0.62), cognitive errors (0.62), episodic memory (0.62), veridical (0.61), memory processes (0.61), false memory syndrome (0.61), dissociative disorder (0.61), thought disorder (0.61), ideomotor (0.61), psychological phenomena (0.61), dissociative disorders (0.61), perceptual experiences (0.61), cognitive bias (0.60), inattentional blindness (0.60), faulty memory (0.60), suggestibility (0.60), post-hypnotic (0.60), misattribution (0.60), semantic memory (0.60), visual hallucinations (0.60), neurolinguistic (0.60), auditory hallucinations (0.60), retrograde amnesia (0.60), anterograde amnesia (0.60), thought content (0.60), suspiciousness (0.60), cognitive processes (0.59), frontal lobe (0.59), palinopsia (0.59), malingering (0.59), memory traces (0.59), poor memory (0.59), psychogenic (0.59), explicit memory (0.59), hallucinatory (0.59), pathological lying (0.59), anosognosia (0.59), psychosis (0.59), autistic behavior (0.59), implicit memory (0.59), non-psychotic (0.59), aboulia (0.58), thought processes (0.58), hypnopompic (0.58), confusional (0.58), cognitive disorder (0.58), brain dysfunction (0.58), conscious states (0.58), cognitive biases (0.58), confirmation bias (0.58), selective amnesia (0.58), hyper-vigilance (0.58), recollection (0.58), tulving (0.58), multiple personality (0.58), recovered memory (0.58), psychological explanations (0.58), temporal lobe (0.58), mutism (0.58), long term memory (0.58), factitious disorder (0.58), simultanagnosia (0.58), somatoform (0.58), multiple-personality (0.58), recognition memory (0.58), how memory works (0.58), visuo-spatial (0.58), declarative memory (0.58), neuroimaging studies (0.58), dream content (0.58), episodic memories (0.58), conscious perception (0.58), delusional disorder (0.58), fundamental attribution error (0.57), memory errors (0.57), fmri studies (0.57), preconscious (0.57), inner speech (0.57), conscious awareness (0.57), cognitive deficit (0.57), angular gyrus (0.57), delusion (0.57),

Additive property



Works across many other “meanings”

Paris : France :: London : England

Cat : Kitten :: Dog : Puppy

Driver : Driving :: Priest : Praying

Windows : Microsoft :: Android : Google

Copper : Cu :: Gold : Au

Japan : Sushi : Germany : Bratwurst

Works across grammar

King : Kings :: Queen : Queens

Good : Bad :: Better : Worse

Try : Trying :: Eat : Eating

Solve : Solved :: See : Seen

Soft : Softly :: Rough : Roughly

Possible : Impossible :: Ethical : Unethical

Shows stereotypes, inherited from the corpus

Man : Doctor :: Woman : Nurse Practitioner

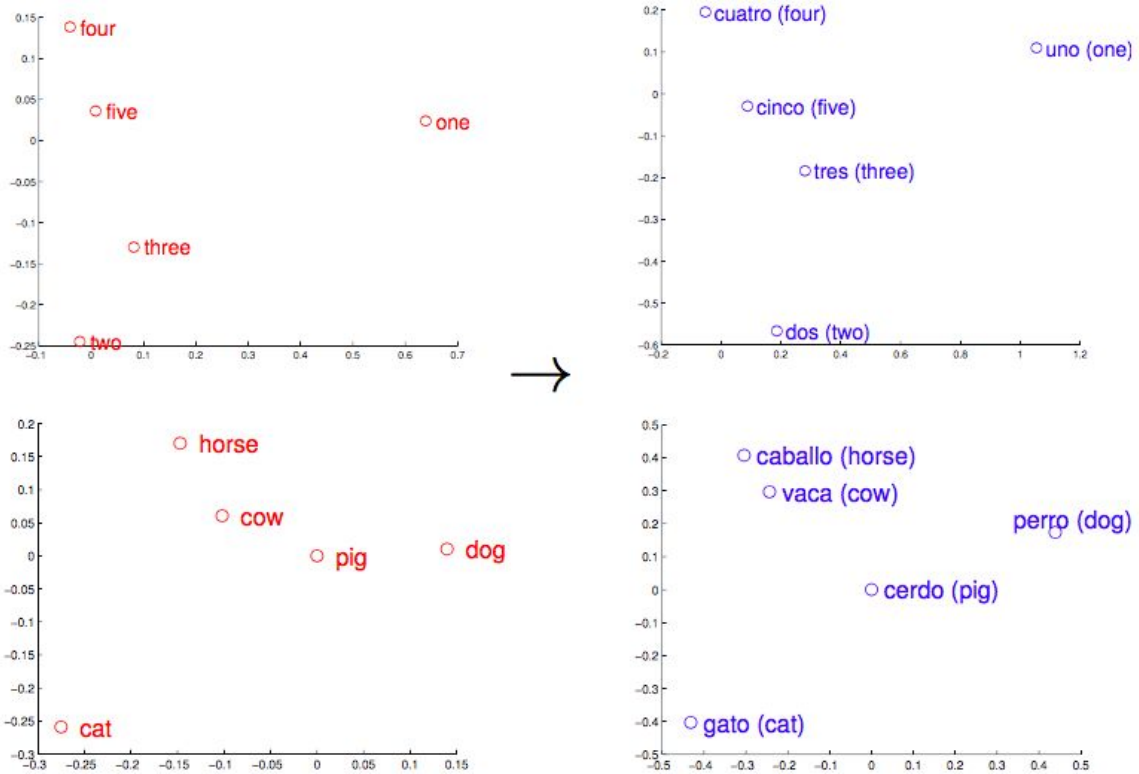
He : Doctor :: She : Nurse Practitioner

Man : Disgusting :: Woman : Disgusted

She : Eclairs :: He : Sausage

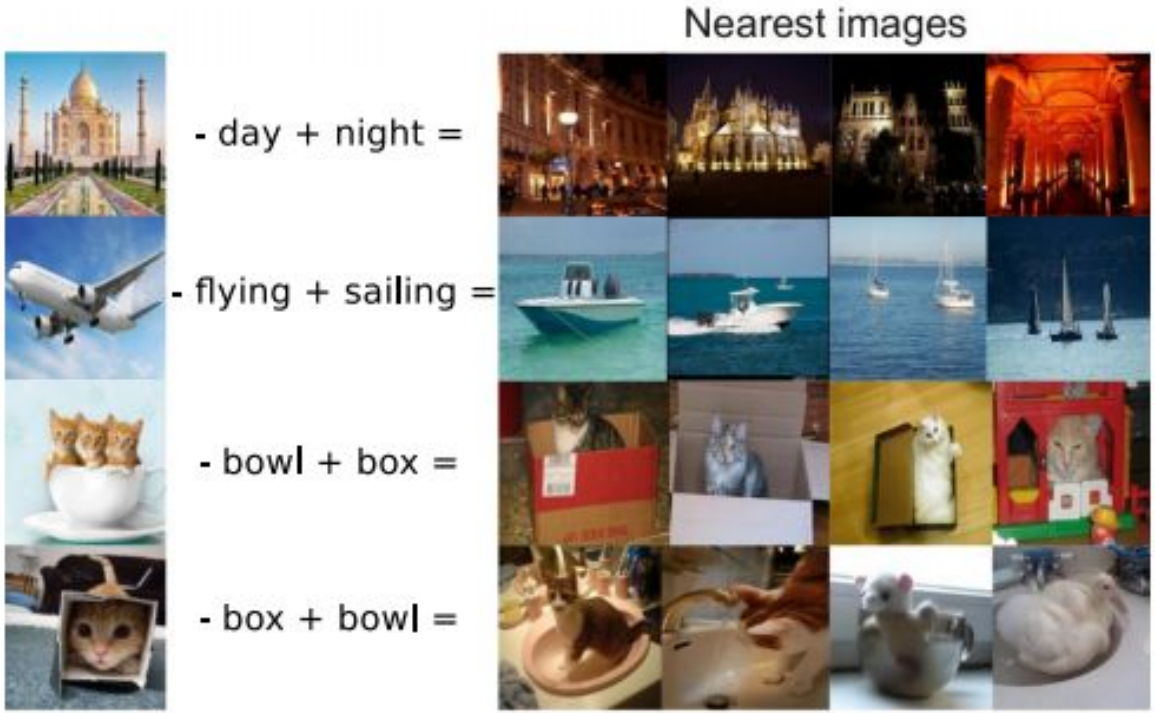
She : Feminist :: He : Misogynist

Works across Languages (?!?)



Exploiting Similarities among Languages for Machine Translation, Mikolov & al, 2013

Works with Images (you are kidding, right?)



Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models, 2014

Word Vectors are a Very Powerful Representation

Animal	**Vegetal**	**Mineral**
Dog	Grass	Rock
Cat	Carrot	Coal
Bird	Redwood	Sand
Platypus	Sycamore	Tar
Fox	Tomato	Diamond
Bear	Celeri	Amethyst
Wolf	Zucchini	
Ladybug	Peach tree	
Shark	Apple tree	
Fox		
Coyote		

Training Set

Raccoon

Without semantic distance

Raccoon

Similarity	
Bobcat:	**67%**
Squirrel:	**67%**
Coyote:	**67%**
Skunk:	**63%**
Possum:	**58%**
Porcupine:	**58%**
Fisher cat:	**59%**
Cat:	**62%**
Grass:	22%
Sand:	9%

Building a set of embedded vectors

Option 1:

- Download existing vectors from Google, GloVe, PolyGlot

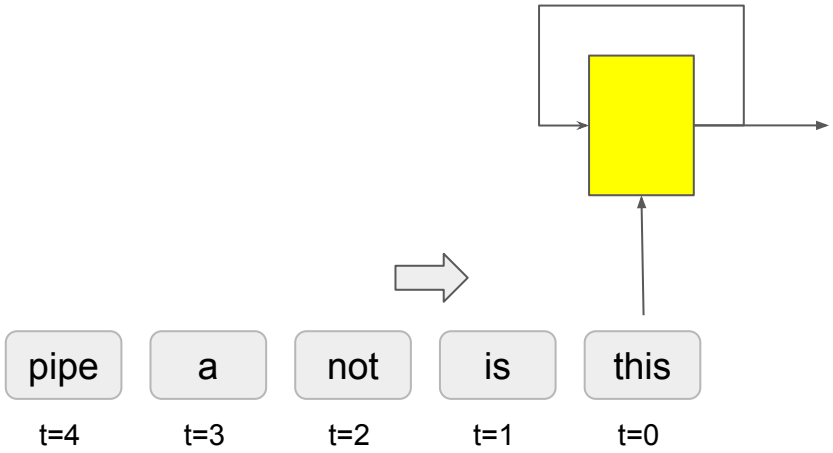
Option 2:

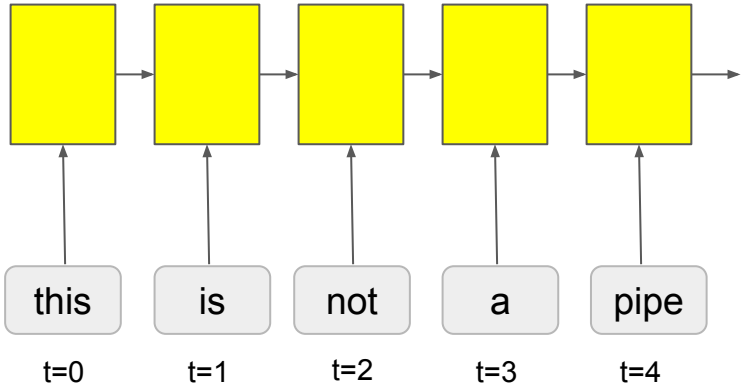
- Download a big corpus (Wikipedia, Web pages from Common Crawl, news…)
- Clean diverse text is best.
- Pick a vocabulary (words and phrases), or just most frequent words.
- Pick a vector size: 50-1000, 200 is a sweet spot.
- Run **word2vec** (in gensym), or similar (GloVe…).

Recurrent Neural Networks

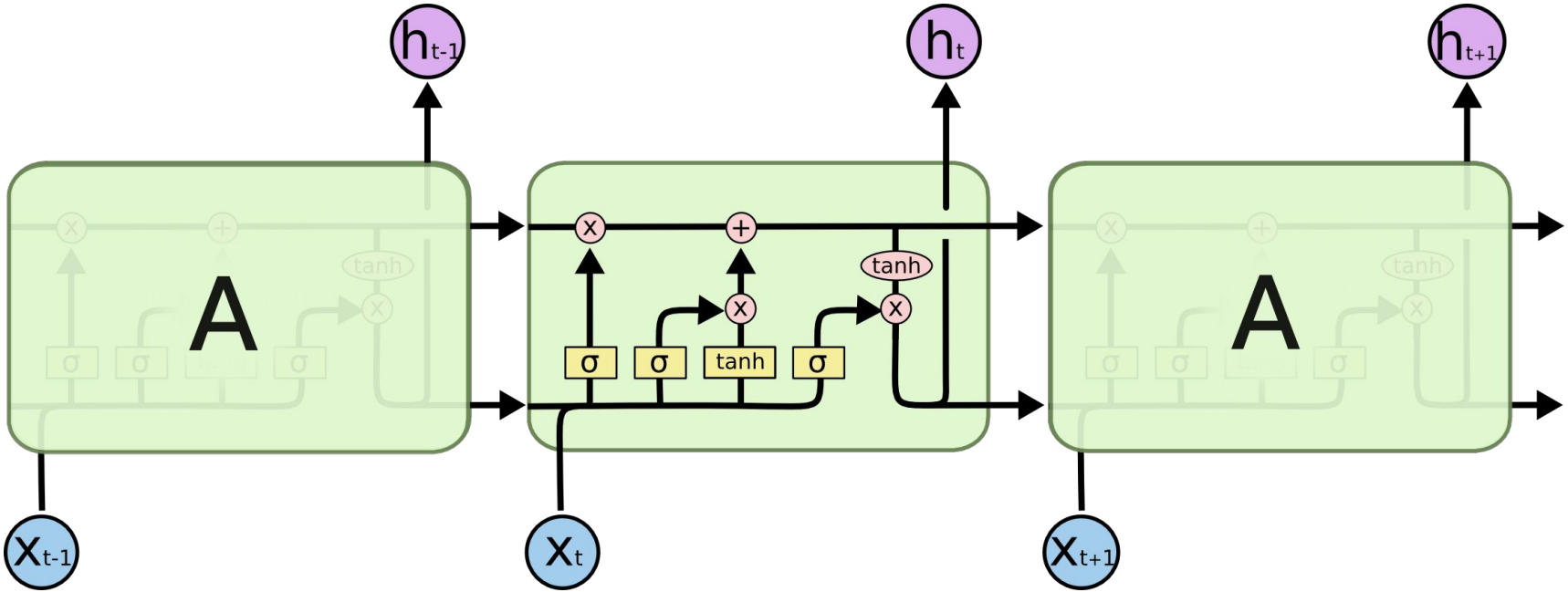
Recurrent Neural Networks (RNN)

Great for sequences!





Long Short-Term Memory - LSTM



Credit: Christopher Olah

4 parameters, 2 recurring connections per cell

Simple example: char-nn by Andrej Karpathy

Feed text one character at a time.
Learn to predict the next character.
Then **hallucinate** new text from noise.

```
PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.
```

```c
/*
 * Increment the size file of the new incorrect
 * of the size generatively.
 */
static int indicate_policy(void)
{
  int error;
  if (fd == MARN_EPT) {
    /*
     * The kernel blank will coeld it to userspa
     */
    if (ss->segment < mem_total)
      unblock_graph_and_set_blocked();
    else
      ret = 1;
    goto bail;
  }
  segaddr = in_SB(in.addr);
  selector = seg / 16;
  setup_works = true;
  for (i = 0; i < blocks; i++) {
    seq = buf[i++];
    bpf = bd->bd.next + i * search;
    if (fd) {
      current = blocked;
    }
  }
```

What is being learned?

Cell that turns on inside quotes:

```
"You mean to imply that I have nothing to eat out of.... On the
contrary, I can supply you with everything even if you want to give
dinner parties," warmly replied Chichagov, who tried by every word he
spoke to prove his own rectitude and therefore imagined Kutuzov to be
animated by the same desire.

Kutuzov, shrugging his shoulders, replied with his subtle penetrating
smile: "I meant merely to say what I said."
```

Cell that robustly activates inside if statements:

```
static int __dequeue_signal(struct sigpending *pending, sigset_t *mask,
        siginfo_t *info)
{
  int sig = next_signal(pending, mask);
  if (sig) {
    if (current->notifier) {
      if (sigismember(current->notifier_mask, sig)) {
        if (!(current->notifier)(current->notifier_data)) {
          clear_thread_flag(TIF_SIGPENDING);
          return 0;
        }
      }
    }
    collect_signal(sig, pending, info);
  }
  return sig;
}
```

Credit: Andrej Karpathy

Cell that is sensitive to the depth of an expression:

```
#ifdef CONFIG_AUDITSYSCALL
static inline int audit_match_class_bits(int class, u32 *mask)
{
  int i;
  if (classes[class]) {
    for (i = 0; i < AUDIT_BITMASK_SIZE; i++)
      if (mask[i] & classes[class][i])
        return 0;
  }
  return 1;
}
```

A large portion of cells are not easily interpretable. Here is a typical example:

```
/* Unpack a filter field's string representation from user-space
 * buffer. */
char *audit_unpack_string(void **bufp, size_t *remain, size_t len)
{
  char *str;
  if (!*bufp || (len == 0) || (len > *remain))
    return ERR_PTR(-EINVAL);
  /* Of the currently implemented string fields, PATH_MAX
   * defines the longest valid length.
   */
```

Sequence to Sequence Learning

A freakishly good idea

The Problem: Map seq(n) to seq(m), for any n, m

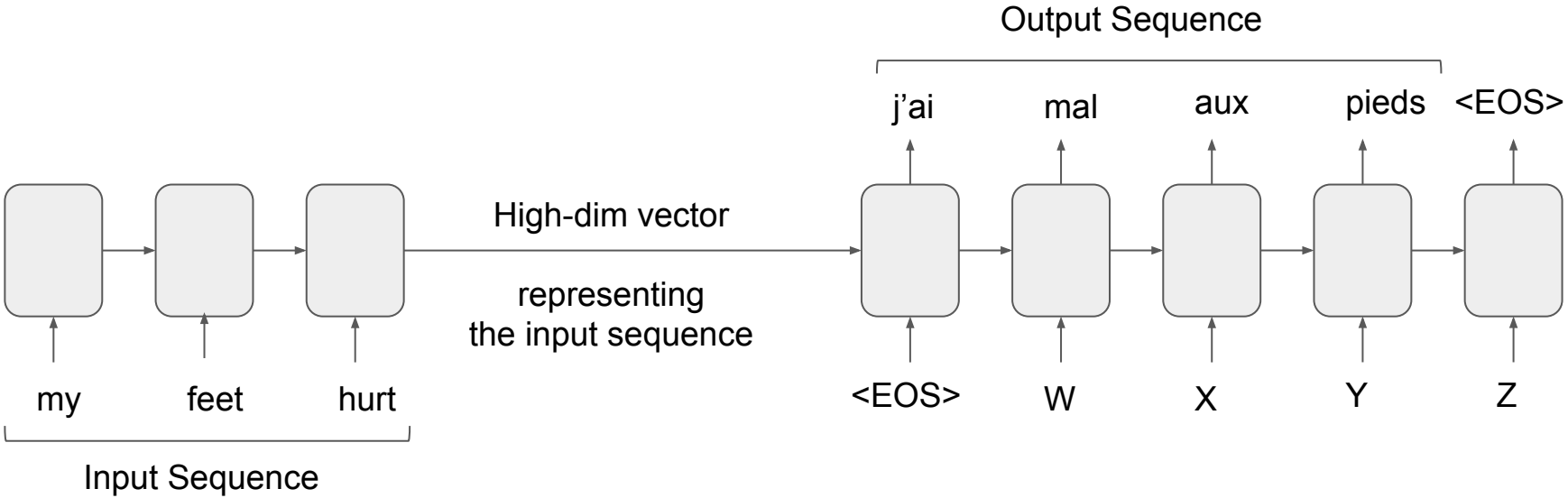
Very general problem: machine translation, question answering, speech.

(-) DNNs work on fixed-sized inputs and outputs.

(+) RNNs can accept and output sequences, but have issues.

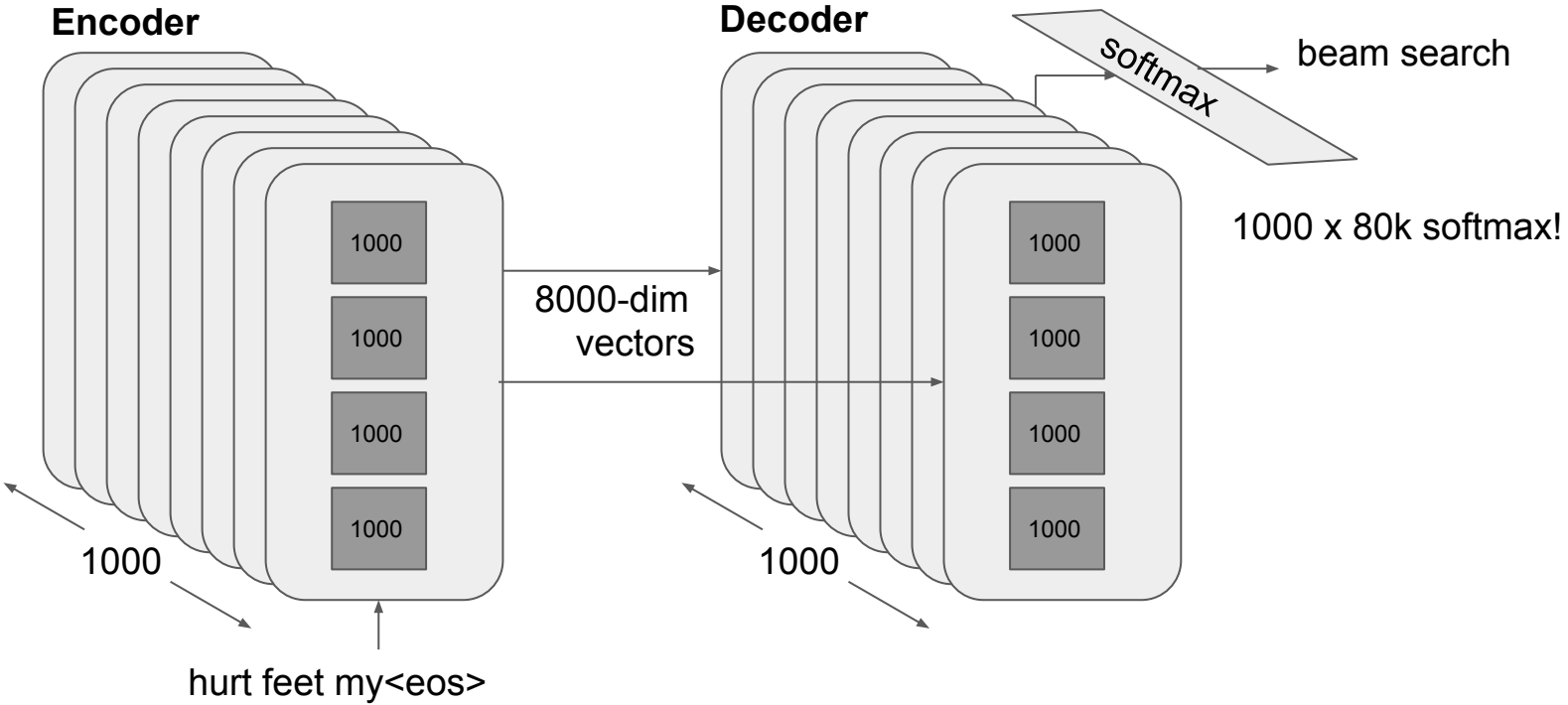
(+) LSTMs can learn large sets of long sequences, with no issues.

The Big Idea (unrolled in time)



What Google built

12M sentence pairs
vocab: 160K English words, 80K French
2 x 4-deep LSTMs, 384M parameters (?)
10 days to train on 8 GPUs



OK to mix and match

LSTM on images

- OCR
- Hand-writing recognition
- ConvNet + LSTM for image caption

ConvNet on text

- Combo ConvNet / LSTM for classification

Big Wins for NLP using Deep Learning

Parsing

POS tagging

Classification (e.g. sentiment, spam)

Translation