Machine learning for sequences
Text and time series

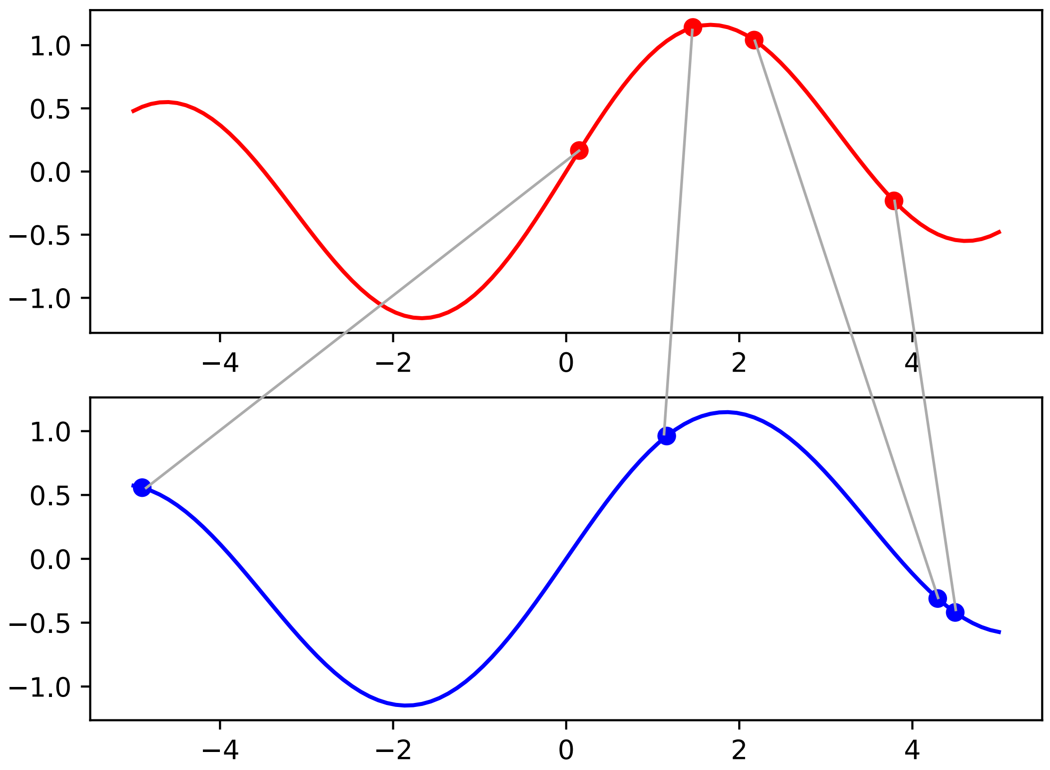
Romain Tavenard (Université de Rennes)
M2 Data Science

Machine learning for structured data (continued)

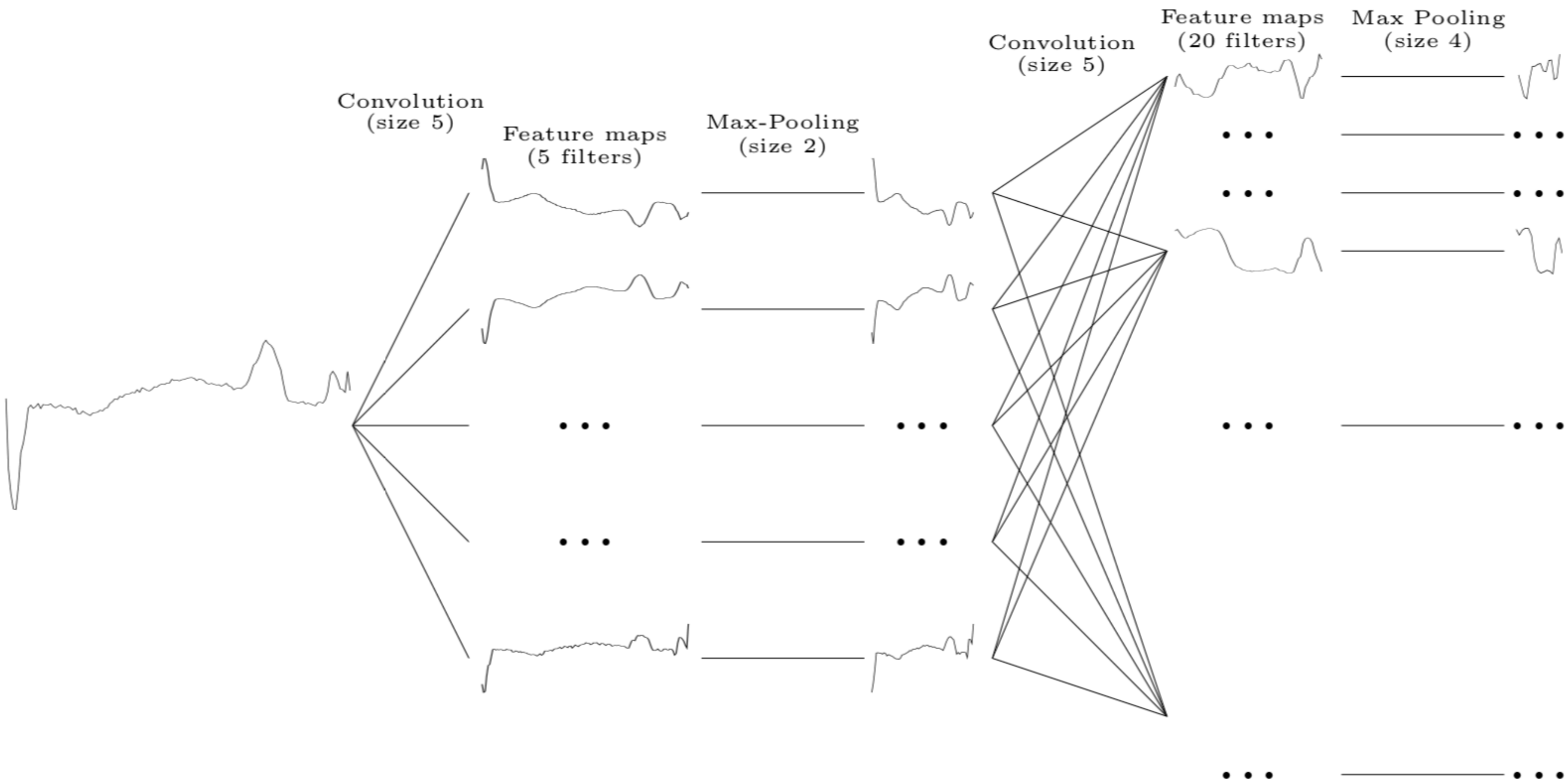
- Two options
  1. Cast the data to tabular
     - Representation based on global features (*eg.* bag of words for text or images)
  2. Use structural information in the model
     - images: 2d convolutions
     - sequences: recurrent models, 1d convolutions, temporal kernels

Standard issues with sequences

- Variable number of observations per sequence
  - `the cat eats the mouse`
  - `at the moment, the cat is eating the mouse`
- Segmentation (starting/end points)
- Irregular sampling (time series)



Solution #1: NN with 1d-convolutions

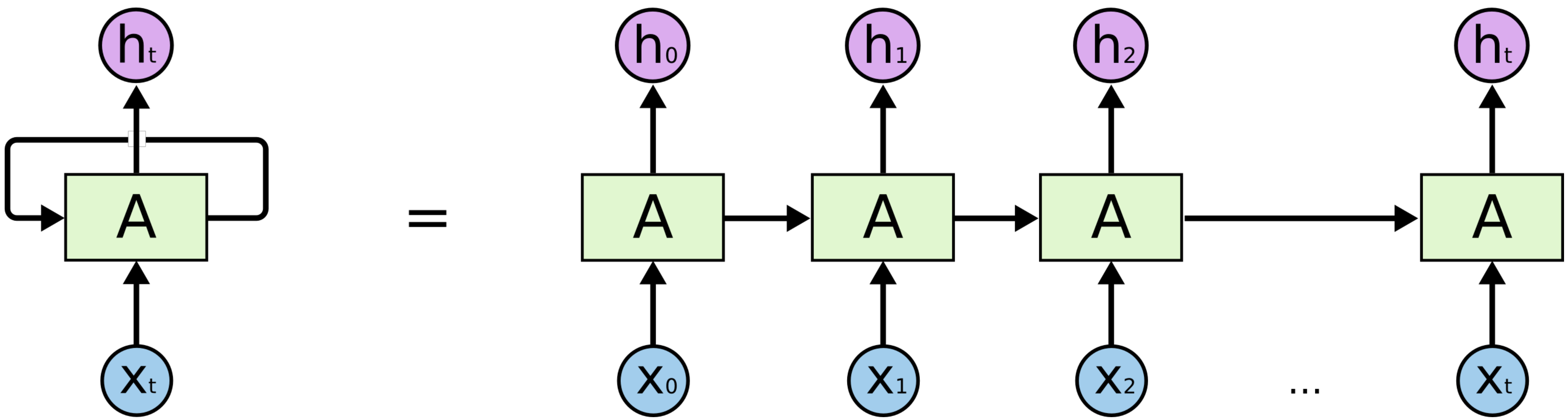


Source: [Le Guennec *et al.*, 2014]

Solution #1: NN with 1d-convolutions

- Variable number of observations per sequence
  - Pad the sequence with 0
- Segmentation (starting/end points)
  - Data augmentation
  - Global Max-Pooling
- Irregular sampling (time series)
  - Not robust to that!

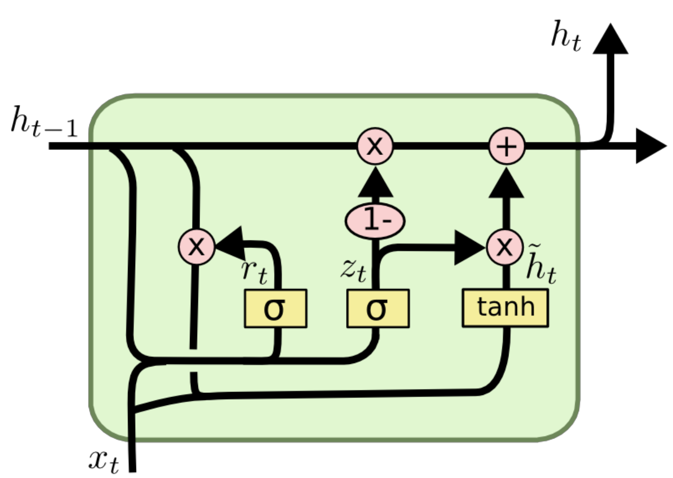
Solution #2: NN with recurrent units



Source: Christopher Olah's blog

Solution #2: NN with recurrent units

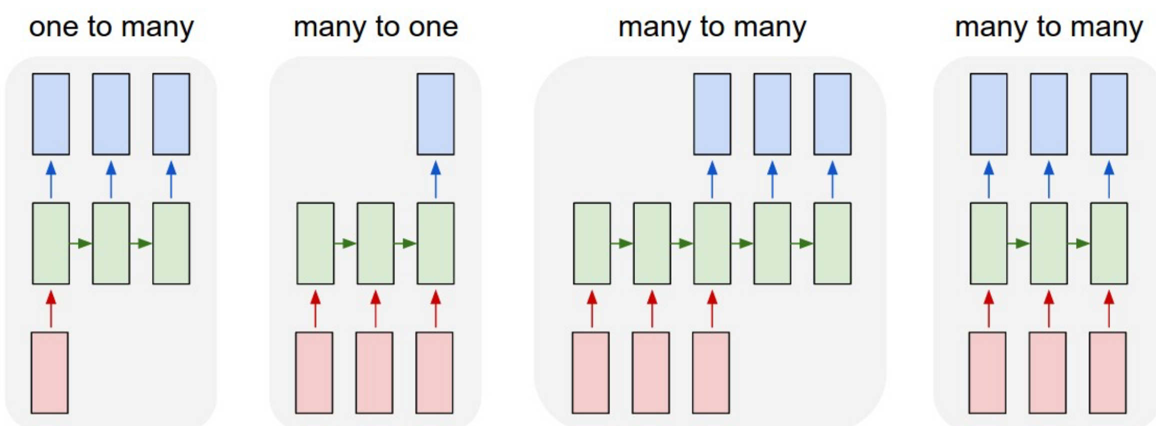
- Variants that work well in practice
  - Long Short Term Memory (LSTM)
  - Gated Recurrent Unit (GRU)
- Principle
  - At each time step, keep only part of the information



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

Illustration: GRU cell, source: Christopher Olah's blog

Solution #2: NN with recurrent units




```
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.

Second Lord:
They would be ruled after this chamber, and
my fair nues begun out of the fact, to be conveyed,
Whose noble souls I'll have the heart of the wars.
```

Sample text generated by a RNN trained on Shakespeare words

For $\bigoplus_{n=1,\dots,m}$ where $\mathcal{L}_{m_\bullet} = 0$, hence we can find a closed subset $\mathcal{H}$ in $\mathcal{H}$ and any sets $\mathcal{F}$ on X, U is a closed immersion of S, then $U \to T$ is a separated algebraic space.

Proof. Proof of (1). It also start we get

$$S = \mathrm{Spec}(R) = U \times_X U \times_X U$$

and the comparicoly in the fibre product covering we have to prove the lemma generated by $\coprod Z \times_U U \to V$. Consider the maps M along the set of points Sch_{fppf} and $U \to U$ is the fibre category of S in U in Section, ?? and the fact that any U affine, see Morphisms, Lemma ??. Hence we obtain a scheme S and any open subset $W \subset U$ in $Sh(G)$ such that $\mathrm{Spec}(R') \to S$ is smooth or an

$$U = \bigcup U_i \times_{S_i} U_i$$

which has a nonzero morphism we may assume that f_i is of finite presentation over S. We claim that $\mathcal{O}_{X,x}$ is a scheme where $x, x', s'' \in S'$ such that $\mathcal{O}_{X,x'} \to \mathcal{O}'_{X',x'}$ is separated. By Algebra, Lemma ?? we can define a map of complexes $\mathrm{GL}_{S'}(x'/S'')$ and we win. $\square$

To prove study we see that $\mathcal{F}|_U$ is a covering of $\mathcal{X}'$, and $\mathcal{T}_i$ is an object of $\mathcal{F}_{X/S}$ for $i > 0$ and $\mathcal{F}_p$ exists and let $\mathcal{F}_i$ be a presheaf of $\mathcal{O}_X$-modules on $\mathcal{C}$ as a $\mathcal{F}$-module. In particular $\mathcal{F} = U/\mathcal{F}$ we have to show that

$$\widetilde{M}^\bullet = \mathcal{I}^\bullet \otimes_{\mathrm{Spec}(k)} \mathcal{O}_{S,s} - i_X^{-1}\mathcal{F})$$

is a unique morphism of algebraic stacks. Note that

$$\mathrm{Arrows} = (Sch/S)^{opp}_{fppf}, (Sch/S)_{fppf}$$

and

$$V = \Gamma(S, \mathcal{O}) \longmapsto (U, \mathrm{Spec}(A))$$

is an open subset of X. Thus U is affine. This is a continuous map of X is the inverse, the groupoid scheme S.

Proof. See discussion of sheaves of sets. $\square$

The result for prove any open covering follows from the less of Example ??. It may replace S by $X_{spaces,étale}$ which gives an open subspace of X and T equal to S_{Zar}, see Descent, Lemma ??. Namely, by Lemma ?? we see that R is geometrically regular over S.

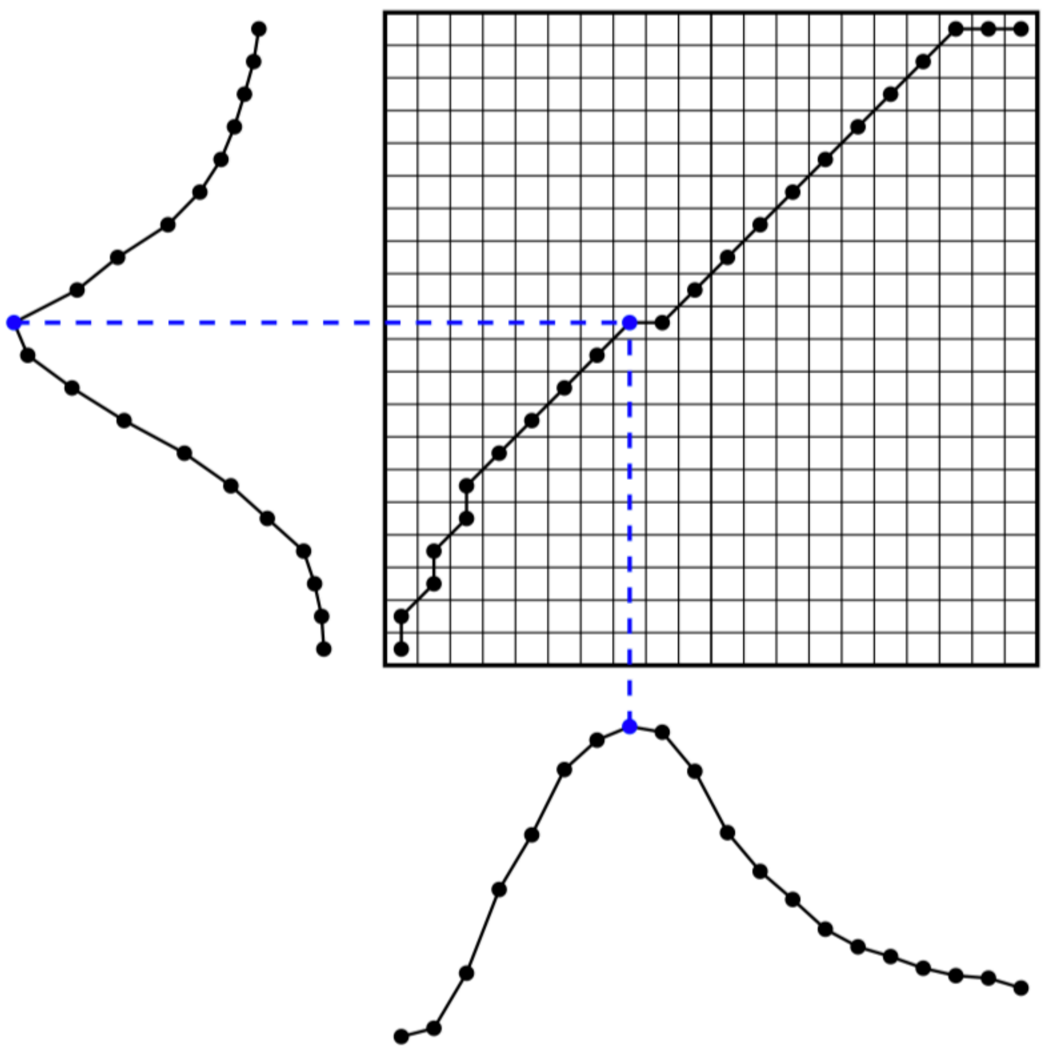
Sample LaTeX generated by a RNN trained on a book of algebraic geometry

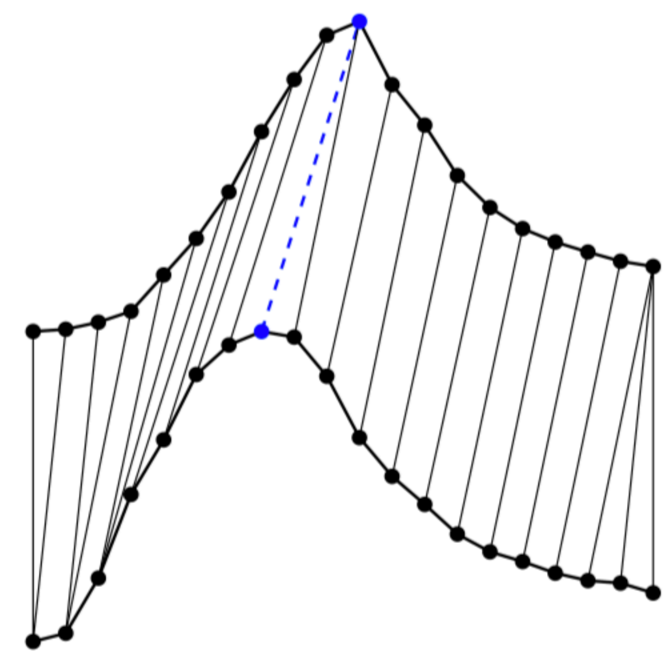
Source: Andrej Karpathy's blog, http://karpathy.github.io/2015/05/21/rnn-effectiveness/

Solution #2: NN with recurrent units

- Variable number of observations per sequence
  - OK
- Segmentation (starting/end points)
  - Not so robust to that!
- Irregular sampling (time series)
  - OK

Solution #3: knn/SVM with a temporal kernel





Source: [Dupas *et al.*, 2015]

Solution #3: knn/SVM with a temporal kernel

- Variants that work well in practice
  - Dynamic Time Warping (DTW) with knn
  - Global Alignment Kernel (GAK) with SVM

- Example Python implementation: `tslearn` [Tavenard, 2017]
  - doc for DTW+knn
  - doc for SVM+GAK

Solution #3: knn/SVM with a temporal kernel

- Variable number of observations per sequence
  - Should be OK (depending on implementation)
- Segmentation (starting/end points)
  - OK, up to some point
- Irregular sampling (time series)
  - OK

Dealing with text in practice

- Raw text data is challenging to handle
  - typos
  - what is a term?
  - lots of variants for a term
    - verb conjugation
    - plural form
    - *etc.*
  - synonyms

Dealing with text in practice

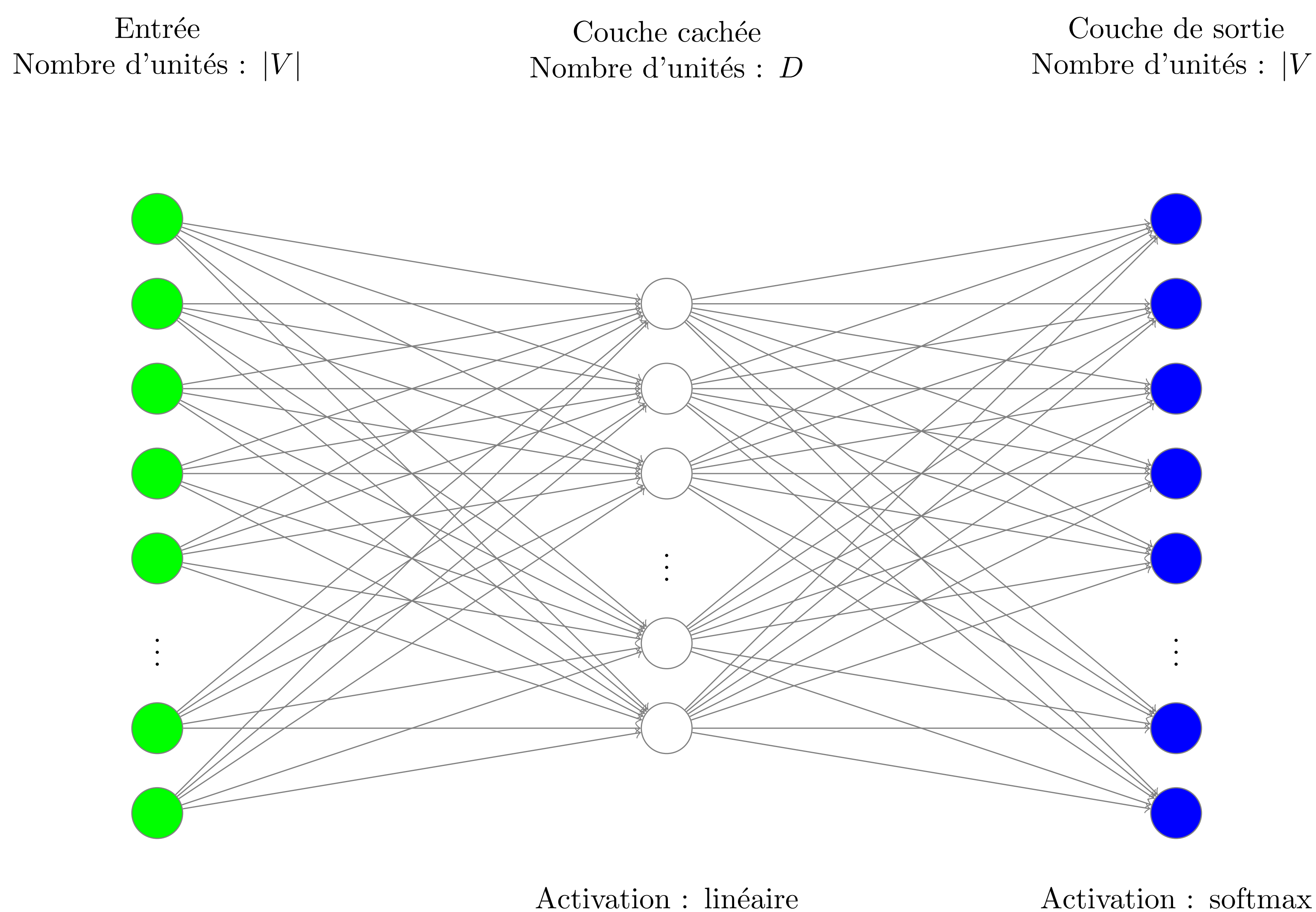
- Raw text data is challenging to handle
  - typos preprocessing
  - what is a term? tokenization
  - lots of variants for a term stemming
    - verb conjugation
    - plural form
    - *etc.*
  - synonyms word embeddings

Word embeddings

- Basic idea
  - 1 term = 1 point in multidimensional space
  - Goal: define a space such that similar terms are close

- Reference embedding
  - word2vec

word2vec



Entrée
Nombre d'unités : $|V|$
Couche cachée
Nombre d'unités : D
Couche de sortie
Nombre d'unités : $|V|$
Activation : linéaire
Activation : softmax

word2vec: Continuous Bag of Words (CBOW)

Le **chien mange un os dans sa gamelle**.

- À l'**entrée du réseau**
  - Une représentation sac de mots (vecteur de 0 et de 1) du voisinage du mot cible
- Tâche de classification : prédire le **mot central**

- Pour générer un exemple d'apprentissage
  1. On tire un mot d'un texte au hasard
  2. On fournit son voisinage de taille fixe

- Pourquoi le nom CBOW ?
  - La couche cachée est une version continue, condensée, du BoW fourni en entrée

word2vec: skip-gram

Le **chien** mange un **os** dans sa gamelle.

- À l'**entrée du réseau**
  - Un mot
- Tâche de classification : prédire le **mot du voisinage**

- Pour générer un exemple d'apprentissage
  1. On tire un mot d'un texte au hasard
  2. On tire un mot de son voisinage au hasard

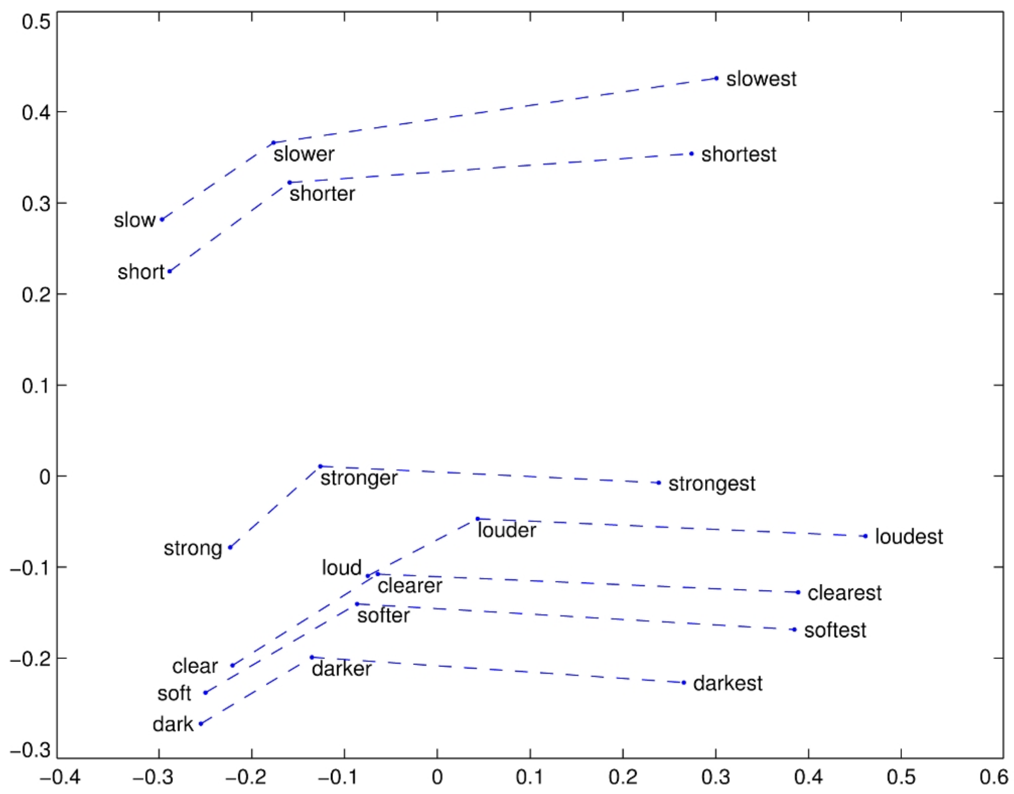
- Pourquoi le nom skip-gram ?
  - On cherche à associer des paires de mots (similaire au bi-gram)
  - On s'autorise des sauts

More about word2vec

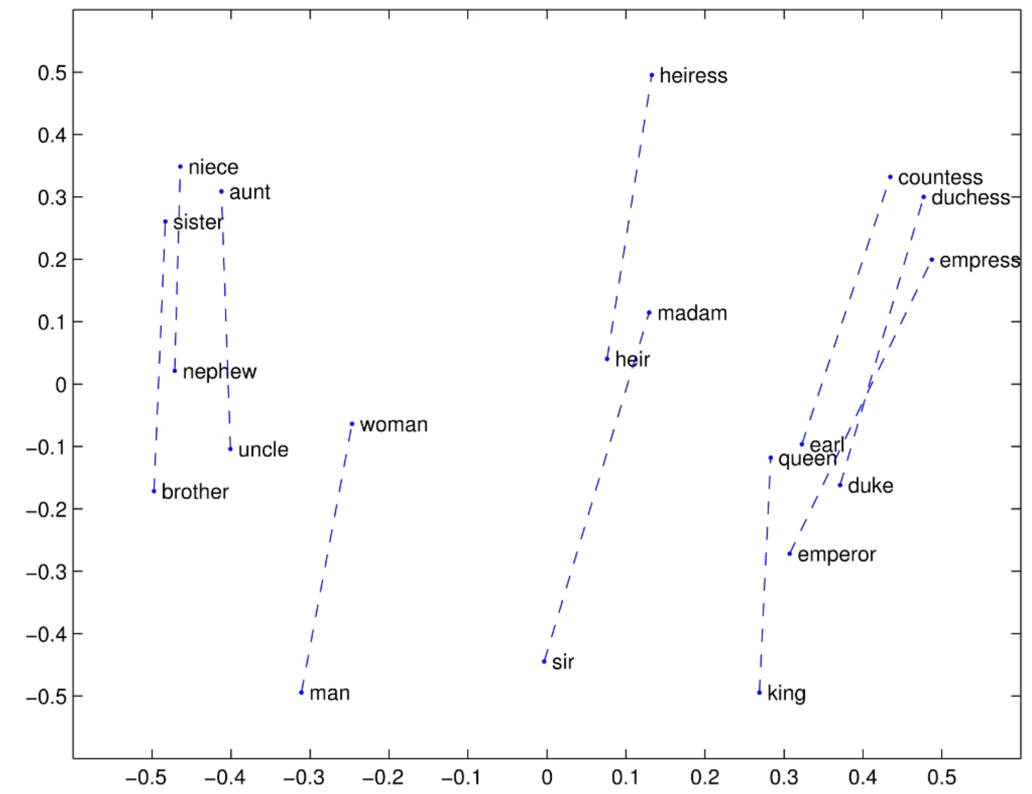
- According to authors
  - Skip-gram
    - works well with small amount of the training data
    - represents well even rare words or phrases
  - CBOW
    - several times faster to train than the skip-gram
    - slightly better accuracy for the frequent words

- To use them
  - Download pre-trained embedding in the correct language
    - https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit for English (3M terms, 300d, 1.5GB)
    - http://fauconnier.github.io/#data for French
  - Use it as a first layer in a NN
    - *cf.* https://blog.keras.io/using-pre-trained-word-embeddings-in-a-keras-model.html

Embedding visualisation



Source: Stanford NLP



- 2d-3d projections (PCA)
  - https://projector.tensorflow.org