A Tour of Probabilistic and Deep Approaches for Unsupervised Learning

Rémi Emonet

2016-12-13

Université Jean Monnet – Saint Étienne
Laboratoire Hubert Curien

LABORATOIRE
HUBERT CURIEN
UMR • CNRS • 5516 • SAINT-ETIENNE

Disclaimer

Some notations are atypical.

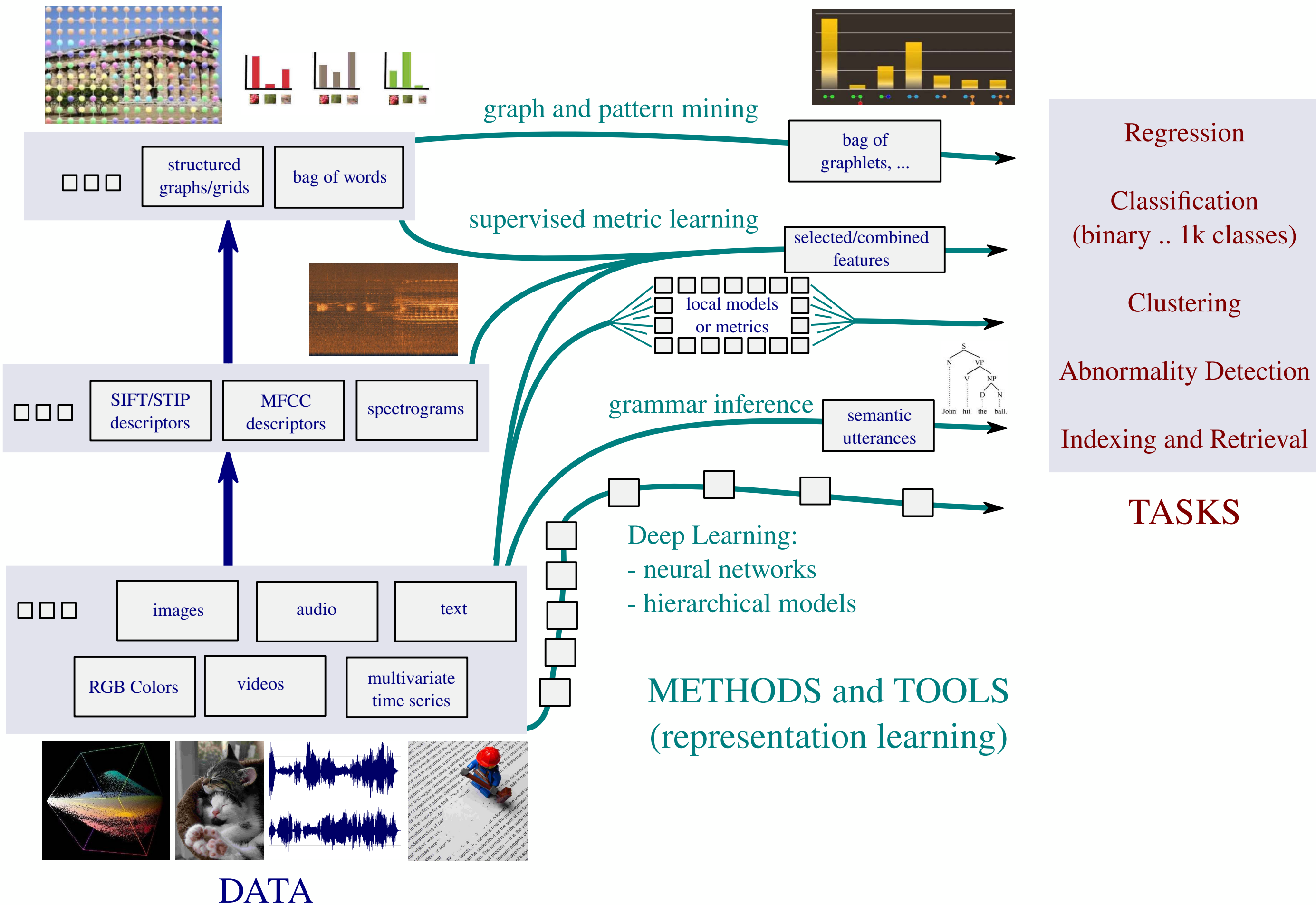
I will, almost surely, skip sections.

Don't hesitate to ask questions lives.

- Unsupervised Representation Learning
- Notations and problem formulation
- Probabilistic (graphical) models
- Auto-encoders
- Generative Adversarial Networks
    - adversarial examples and training
    - GANs
- Focus on …
    - optimization
    - space and time convolutions
    - *depth*, *breadth/width*
    - semantics
    - sequential/temporal aspects
    - recent GANs
- Wrap up

- Unsupervised Representation Learning
- Notations and problem formulation
- Probabilistic (graphical) models
- Auto-encoders
- Generative Adversarial Networks
  - adversarial examples and training
  - GANs
- Focus on …
  - optimization
  - space and time convolutions
  - *depth*, *breadth/width*
  - semantics
  - sequential/temporal aspects
  - recent GANs
- Wrap up

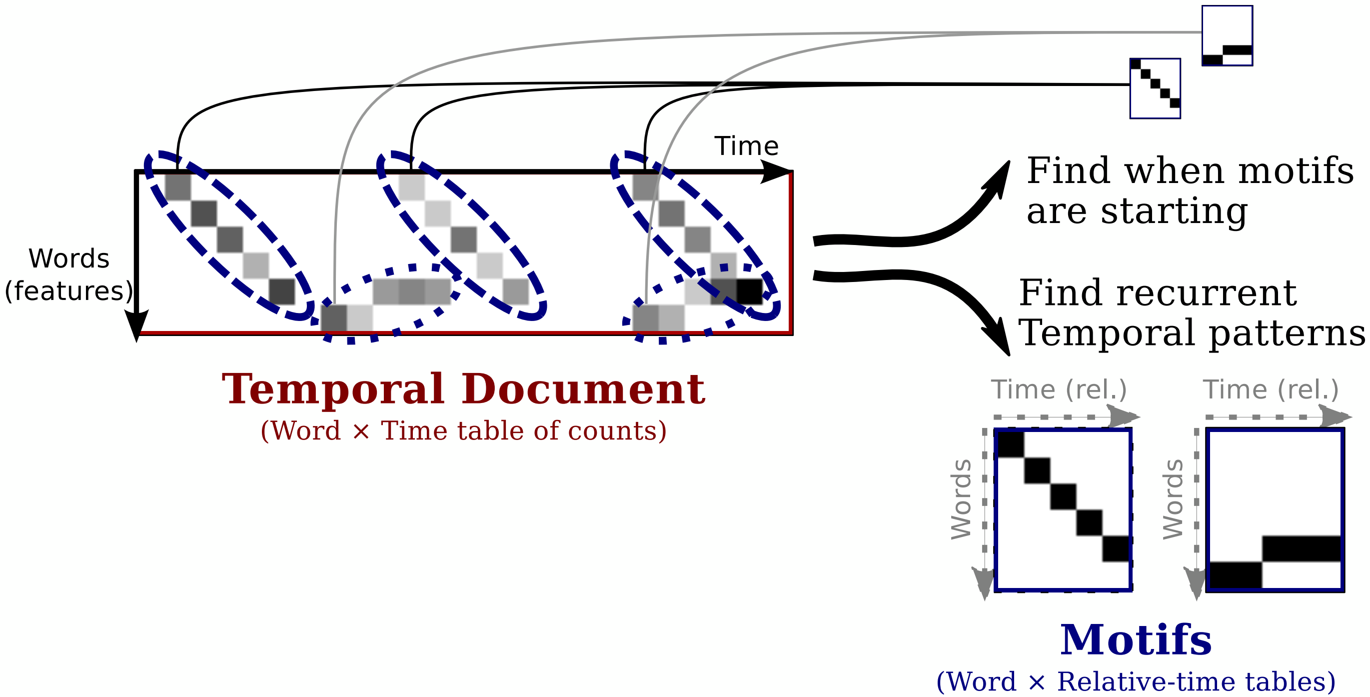
Representation Learning at Hubert Curien



Unsupervised (Representation) Learning

- No labels available
- Learning intermediate features or representations
- Task agnostic
- Related to (data) density estimation
- Related to compression

Example: motif mining in videos / temporal data



- Key points: structure? compression? density estimation?

- Unsupervised Representation Learning
- Notations and problem formulation
- Probabilistic (graphical) models
- Auto-encoders
- Generative Adversarial Networks
  - adversarial examples and training
  - GANs
- Focus on …
  - optimization
  - space and time convolutions
  - *depth*, *breadth/width*
  - semantics
  - sequential/temporal aspects
  - recent GANs
- Wrap up

Notations and problem formulation

Notations and Problem Formulation

LABORATOIRE HUBERT CURIEN
UMR • CNRS • 5516 • SAINT-ETIENNE

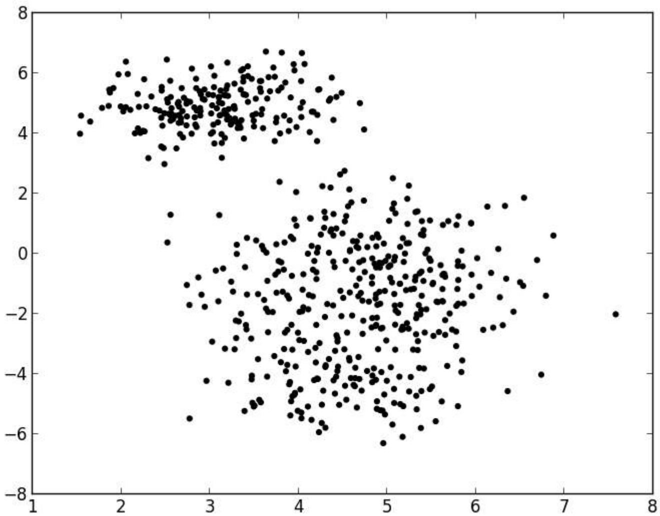
- Notations
  - x : data (observations)
  - y : value to predict (for supervised cases)
  - z : unknown, unobserved latent information
  - θ (or W) : model parameters
- Unsupervised learning
  - only x is given
  - need to find the parameters (θ, W)
  - may want to further infer the latent variables (z)

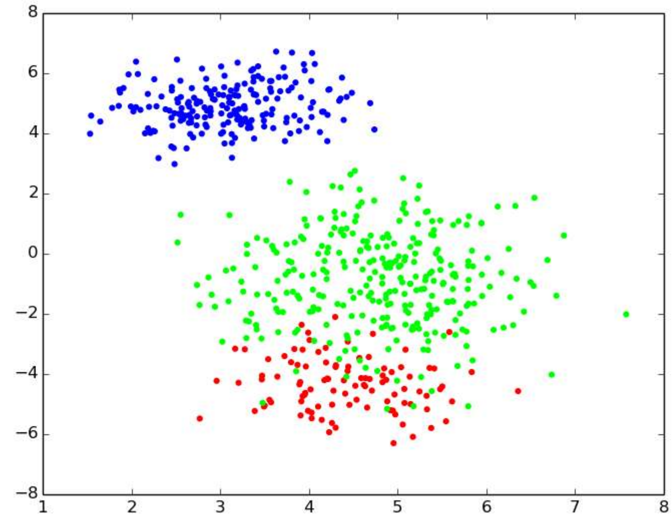
- Unsupervised Representation Learning
- Notations and problem formulation
- Probabilistic (graphical) models
- Auto-encoders
- Generative Adversarial Networks
    - adversarial examples and training
    - GANs
- Focus on …
    - optimization
    - space and time convolutions
    - *depth*, *breadth/width*
    - semantics
    - sequential/temporal aspects
    - recent GANs
- Wrap up

Probabilistic (graphical) models

Generative Model, Parameters, Latent Vars...

- Observations / Data





- Supposition, we have a mixture of 3 gaussians
- Challenge
  - gaussians have unknown *parameters*
  - which point belongs to which component is *not observable*

Probabilistic Modeling: principle

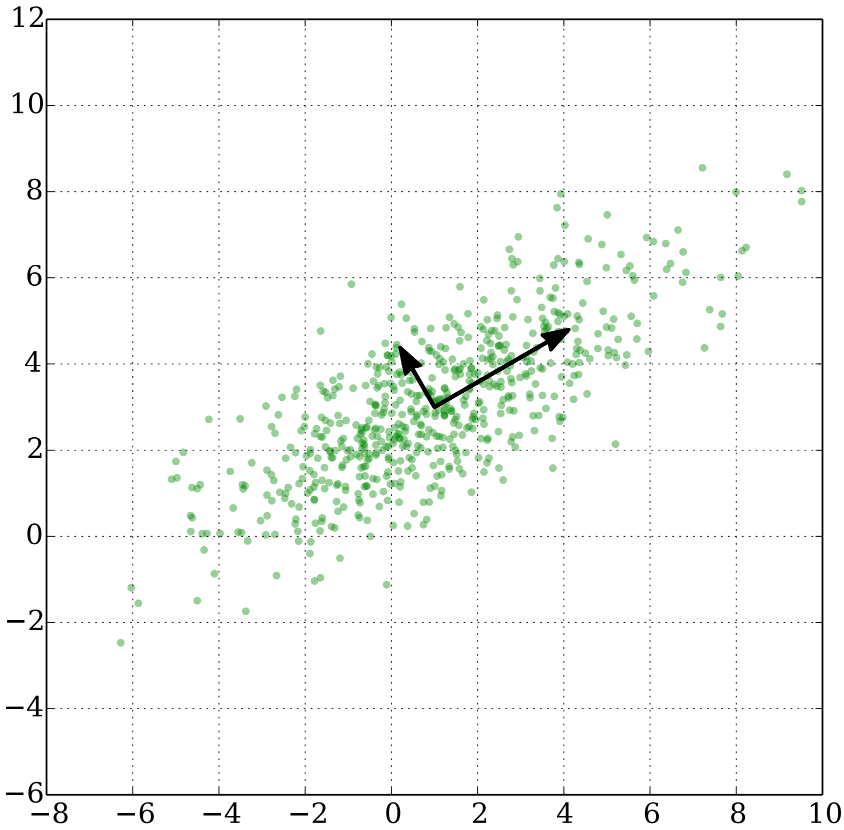
- Adopting a generative approach
  - think about how the world generated the data
  - describe it in a "generative model"
- Formalize your assumptions about the observations (data)
  - choose/design a model
  - a model formulates how *some unknown variables* that are "responsible" for the *observations* (data)
  - set some priors on the unknown variables
- Naming convention: different types of unknowns
  - parameters: unknown global parameters of the model
  - latent variables: unknown observation-specific variables
- With a mixture of Gaussians
  - parameters: mean and covariances (and weight) of all Gaussians
  - latent variables: which Gaussian each data points comes from

Probabilistic Model Learning

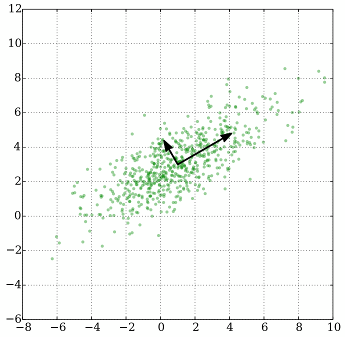
LABORATOIRE HUBERT CURIEN
UMR • CNRS • 5516 • SAINT-ETIENNE

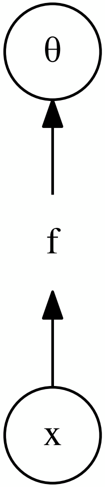
- The model is generative
  - describes how the data (x) gets generated
  - “forward model”
  - the probability of the observations: $p(x|\theta)$
- Finding the unknowns (parameters, latent var.) is challenging
  - reversing the generative process
  - finding (or maximizing) $p(\theta|x)$ or $p(\theta, z|x)$ or $p(z|x, \theta)$
  - high dimensional parameter/latent spaces
  - highly non-convex functions

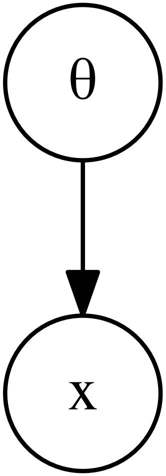
M1 – PCA: intuition



- Principle Component Analysis (eigen-*)
  - dimensionality reduction
  - capture the maximum amount of data variance
- PCA probabilistic view
  - observations come from a single low-dimensional gaussian distribution
  - … and are transformed with a linear transformation (rotation + scale),
  - … and have added noise noisy
- Over-generic graphical representation
  - θ is linear transformation
  - data points x depend on θ
  - no *explicit* latent variables ✎
- Inference problem: f
  - dedicated algorithms (covariance matrix eigenvalues, iterative methods, …)

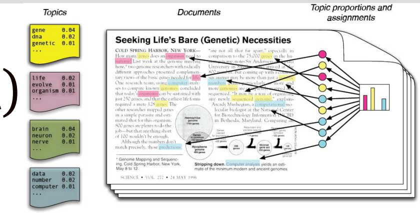






M2 – Topic Modeling: matrix factorization

LABORATOIRE HUBERT CURIEN
UMR • CNRS • 5516 • SAINT-ETIENNE



- Probabilistic Latent Semantic Analysis (PLSA)
  - matrix decomposition
  - non-negative
  - probabilistic formulation
    - $p(w|d) = \sum_z p(w|z) \times p(z|d)$
    - or $x^i = \theta^T \cdot z^i$ (for a document i)

word ↓, document →

	d1	d2	d3
	2	2	2
	1	0	0
	1	0	1
	1	1	1
	0	1	1
	0	1	0

=

topic →

	#1	#2
	?	?
	?	?
	?	?
	?	?
	?	?
	?	?

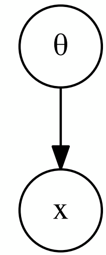
×

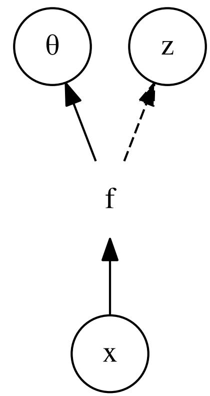
document →, topic ↑

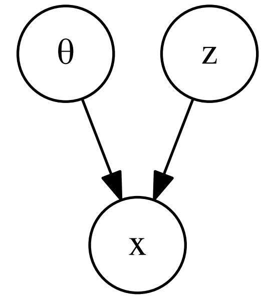
topic	d1	d2	d3
#1	?	?	?
#2	?	?	?

M2: Topic Models

- LDA, topic models …
  - Latent Dirichlet Allocation
  - mixture of discrete distributions (categorical/multinomial)
- Bayesian formulation of
  - LSA, LSI (Latent Semantic Indexing)
- Probabilistic formulation of
  - NMF (non-negative matrix factorization)
- $x^i = \theta^T \cdot z^i$ (for a document i)
- Learning/Inference, f
  - Gibbs sampling
  - EM: expectation maximization
  - variational inference





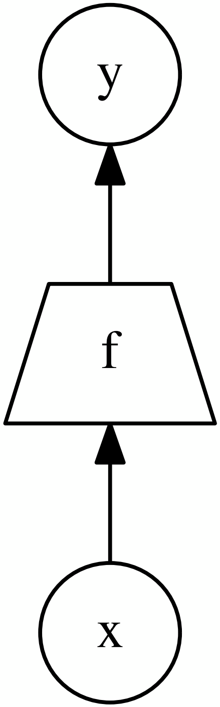


- Unsupervised Representation Learning
- Notations and problem formulation
- Probabilistic (graphical) models
- Auto-encoders
- Generative Adversarial Networks
  - adversarial examples and training
  - GANs
- Focus on …
  - optimization
  - space and time convolutions
  - *depth*, *breadth/width*
  - semantics
  - sequential/temporal aspects
  - recent GANs
- Wrap up

Auto-encoders (not yet)

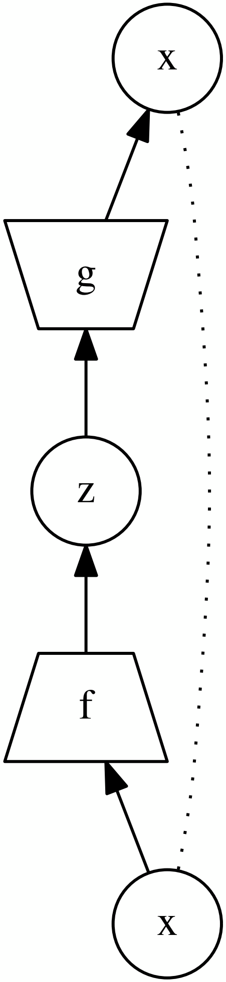
Feed-forward Neural Networks (supervised)

- Supervised learning (regression, classification, ...)
  - the x are given
  - the corresponding labels y are given
- Building blocks of "neural nets"
  - a neuron computes a weighted sum of its inputs
  - the sum is followed by an "activation" σ
  - weights are learned (W)
  - $f^o(x^i, W) = \sigma\left(\sum_d W_{o,d} \times x^i_d\right) = \sigma\left(W^T_{o,.} \cdot x^i\right)$
- Define a network architecture (class of functions)
  - number and dimension of layers
  - activation functions (sigmoid, tanh, ReLU, ...)
  - ... actually any composition of differentiable functions
- Learning with stochastic gradient descent (SGD) and variants



M3: Autoencoders

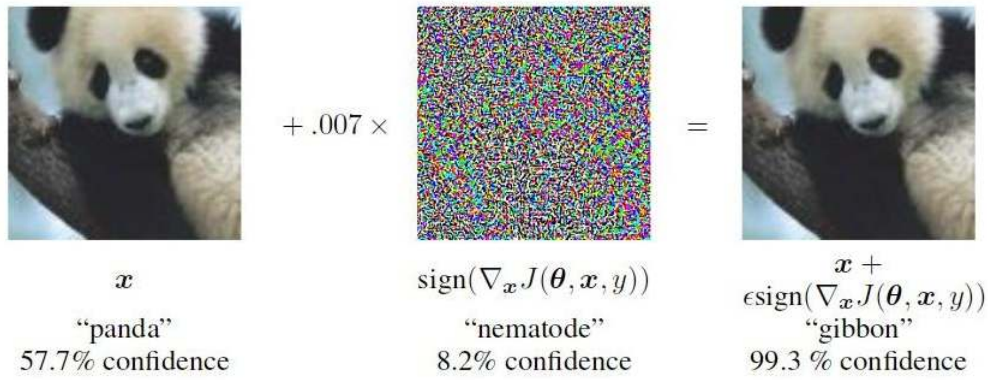
- Idea: use a feed-forward approach
  - ... for unsupervised learning (no labels)
  - to learn a compact data representation
- Principle
  - try to predict the input form the input
  - have a latent **bottleneck**: limited model capacity
  - **encoder** f: from the input x to the latent z
  - **decoder** g: from the latent z to the input x
- Learning principles of f and g
  - mean square reconstruction error: $\|g(f(x)) - x\|^2$
  - SGD (like any neural net)
  - sparsifying regularization: sparse activations (z, $f(x)$)
  - add noise to the input (denoizing autoencoders)



- Unsupervised Representation Learning
- Notations and problem formulation
- Probabilistic (graphical) models
- Auto-encoders
- Generative Adversarial Networks
  - adversarial examples and training
  - GANs
- Focus on …
  - optimization
  - space and time convolutions
  - *depth*, *breadth/width*
  - semantics
  - sequential/temporal aspects
  - recent GANs
- Wrap up

Generative Adversarial Networks

Adversarial Examples (Goodfellow, 2014)



- In high dimensional spaces
  - a huge part of the input space is never seen / irrelevant
  - models are easy to fool
  - models are wrongly calibrated (bad confidence estimation)
- Goal
  - build machine learning methods robust to adversarial examples
  - (relation to anomaly detection)
- Idea of adversarial training
  - generate adversarial examples automatically
  - train also using these examples

GAN Intuition

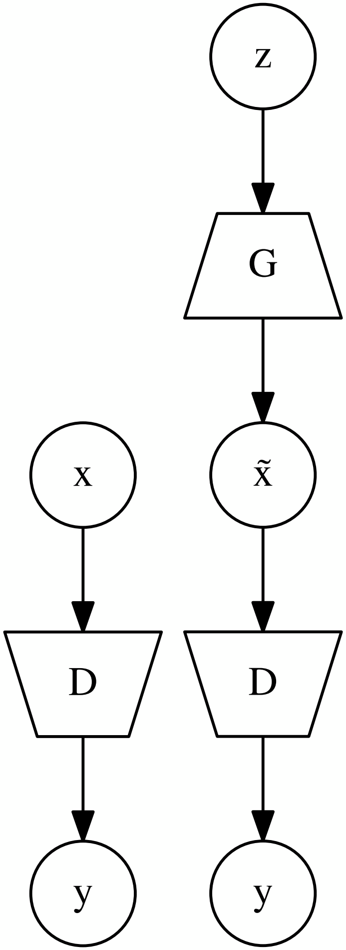
Ongoing struggle between two players:

one that makes fake samples,

one that tries to detect them.

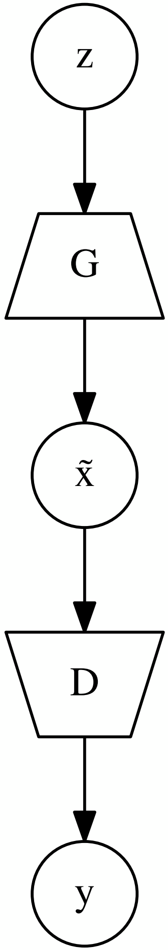
M4: Generative Adversarial Networks

- Principle: train two networks
  - G: to generate samples from noise
  - D: to discriminate between true and generated samples
  - NB: G will try to fool D
- Elements
  - x: a training sample (real)
  - z: a random point in a latent space
  - $\tilde{x}$: a generated sample (fake)
  - y: a binary “fake” (0) or “real” (1) value
- GAN is a minimax game
  - $\min_G \max_D V(D,G)$
  - $V(D,G) = \mathbb{E}_x[log(D(x))] + \mathbb{E}_z[log(1 - D(G(z)))]$



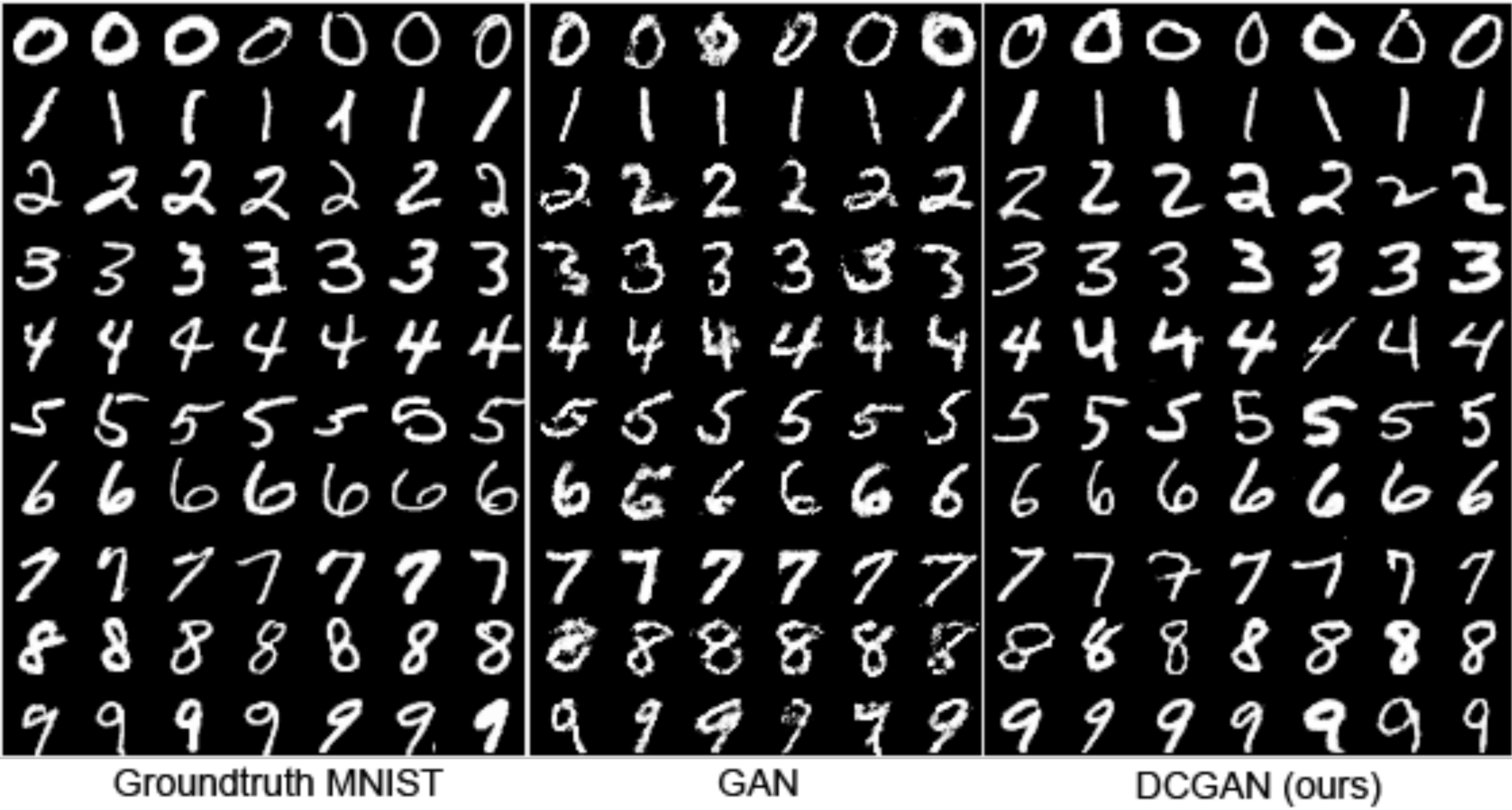
GAN Target

- GAN optimization is a minimax
  - $\min_G \max_D V(D, G)$
  - $V(D, G) = \mathbb{E}_x[log(D(x))] + \mathbb{E}_z[log(1 - D(G(z)))]$
  - find a G that minimizes the accuracy of the **best** D
- Equilibrium and best strategies
  - D ideally computes $D(x) = \dfrac{p_{data}(x)}{p_{data}(x) + p_{gen}(x)}$
  - thus G should ideally fit $p_{data}(x)$
  - … G samples for $p_{data}(x)$
- Optimization in practice
  - alternate optimization of G and D
  - warning: $\min \max$ is not $\max \min$
  - saddle point finding (hot topic)



Example of GAN-generated Digits

- DCGAN, Radford et al., 2015/2016



Example of GAN-generated Images

- Unsupervised Representation Learning
- Notations and problem formulation
- Probabilistic (graphical) models
- Auto-encoders
- Generative Adversarial Networks
  - adversarial examples and training
  - GANs
- Focus on …
  - optimization
  - space and time convolutions
  - *depth*, *breadth/width*
  - semantics
  - sequential/temporal aspects
  - recent GANs
- Wrap up

How is all This Optimized

- Probabilistic models ✎
  - Gibbs sampling
  - Expectation Maximization
  - Variational Inference
  - Black-box variational inference (e.g., Edward)
- Deep models (composition of differentiable function)
  - … using “back-propagation” (chain rule)
  - (S)GD
  - SGD with momentum
  - SGD with adaptation: RMSProp, ADAM, …
  - batch normalization trick
  - link: other tricks for learning GANs
  - are local minima any good?
  - link: which optimizer?

- Unsupervised Representation Learning
- Notations and problem formulation
- Probabilistic (graphical) models
- Auto-encoders
- Generative Adversarial Networks
  - adversarial examples and training
  - GANs
- Focus on …
  - optimization
  - space and time convolutions
  - *depth*, *breadth/width*
  - semantics
  - sequential/temporal aspects
  - recent GANs
- Wrap up

Convolution Models

- Extensions of topic models
  - replace topic with motifs (with temporal structure)
  - PLSM, HDLSM (Emonet et al., 2014)
- Convolutional Neural Networks
  - most of Christian's talk (ConvNets)
  - pixelRNN, …

- Unsupervised Representation Learning
- Notations and problem formulation
- Probabilistic (graphical) models
- Auto-encoders
- Generative Adversarial Networks
  - adversarial examples and training
  - GANs
- Focus on ...
  - optimization
  - space and time convolutions
  - *depth*, *breadth/width*
  - semantics
  - sequential/temporal aspects
  - recent GANs
- Wrap up

Depth in Unsupervised Learning

LABORATOIRE HUBERT CURIEN
UMR • CNRS • 5516 • SAINT-ETIENNE

- Neural Network depth = Hierarchical probabilistic models
- Neural Networks
  - "deep learning"
  - adding layers
  - handling depth with ReLU
  - handling depth with "ResNets", Residual Networks (Deep residual learning for image recognition, He et al. 2015)
- Hierarchical probabilistic models
  - Topic Models (LDA, Blei, Ng, Jordan, 2003)
  - Deep exponential families (Ranganath et al., 2015)
  - Deep Gaussian Processes (Damianou, Lawrence, 2013)

- Unsupervised Representation Learning
- Notations and problem formulation
- Probabilistic (graphical) models
- Auto-encoders
- Generative Adversarial Networks
    - adversarial examples and training
    - GANs
- Focus on …
    - optimization
    - space and time convolutions
    - *depth*, *breadth/width*
    - semantics
    - sequential/temporal aspects
    - recent GANs
- Wrap up

Width in Unsupervised Learning

- Width
  - Topic model: number of topics
  - Autoencoder: number of neurons in the hidden layer
  - GAN: size of z
- Non-parametric approaches, HDP, HDLSM (Emonet et al., 2014)
- Gaussian process as an infinitely wide NN layer (Damianou, Lawrence, 2013)
  - universal function approximator
- Autoencoders with group sparsity (Bascol et al., 2016)
  - allow for many hidden units
  - penalize the use of too many of them

- Unsupervised Representation Learning
- Notations and problem formulation
- Probabilistic (graphical) models
- Auto-encoders
- Generative Adversarial Networks
  - adversarial examples and training
  - GANs
- Focus on ...
  - optimization
  - space and time convolutions
  - *depth*, *breadth/width*
  - semantics
  - sequential/temporal aspects
  - recent GANs
- Wrap up

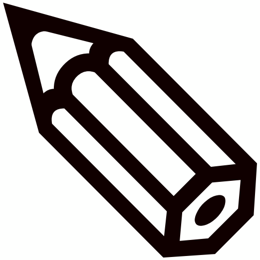
Semantics in Unsupervised Learning

- Probabilistic models
  - inference is difficult
  - consider the "explains away" principle
  - lead to better interpretability (meaningful z)
- Simpler feed-forward model
  - independent processing
  - inhibitory feedback is difficult
- Bascol et al, 2016 ...
  - group-sparsity on filters
  - local activation inhibition
  - global activation entropy maximization
  - AdaReLU: activation function that zeroes low-energy points

- Unsupervised Representation Learning
- Notations and problem formulation
- Probabilistic (graphical) models
- Auto-encoders
- Generative Adversarial Networks
  - adversarial examples and training
  - GANs
- Focus on …
  - optimization
  - space and time convolutions
  - *depth*, *breadth/width*
  - semantics
  - sequential/temporal aspects
  - recent GANs
- Wrap up

Sequential and Temporal Modeling

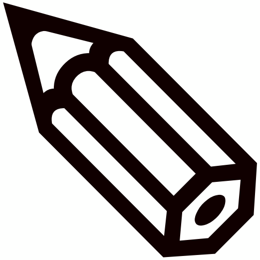
- cf. Christian Wolf's talk
- HMM, CRF =?= RNN
- LSTM =?= HSMM

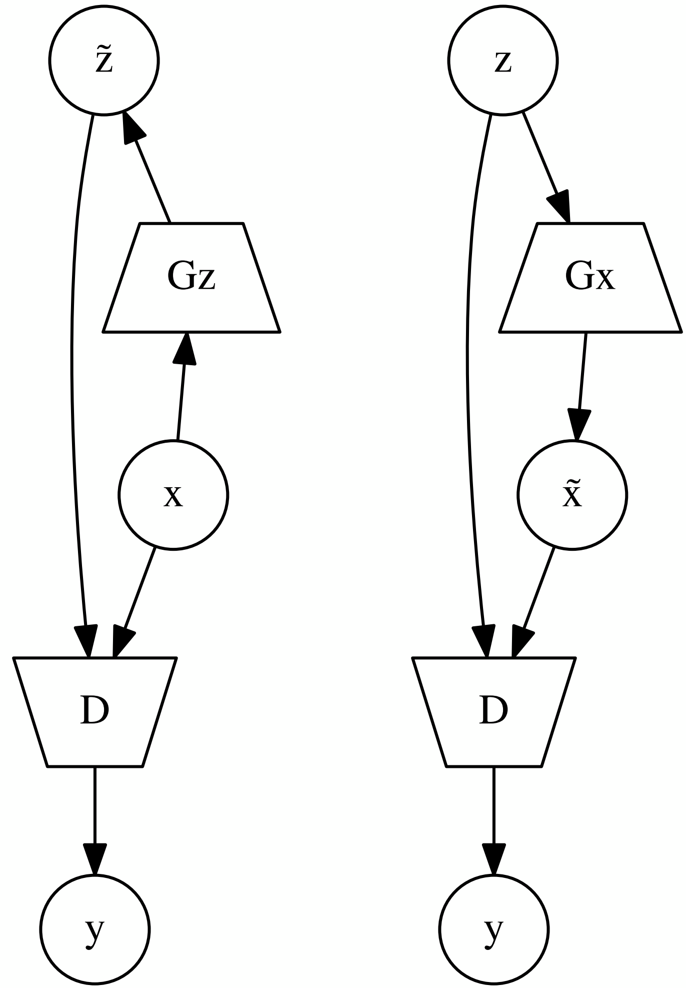


- Unsupervised Representation Learning
- Notations and problem formulation
- Probabilistic (graphical) models
- Auto-encoders
- Generative Adversarial Networks
  - adversarial examples and training
  - GANs
- Focus on …
  - optimization
  - space and time convolutions
  - *depth*, *breadth/width*
  - semantics
  - sequential/temporal aspects
  - recent GANs
- Wrap up

Recent GAN Works

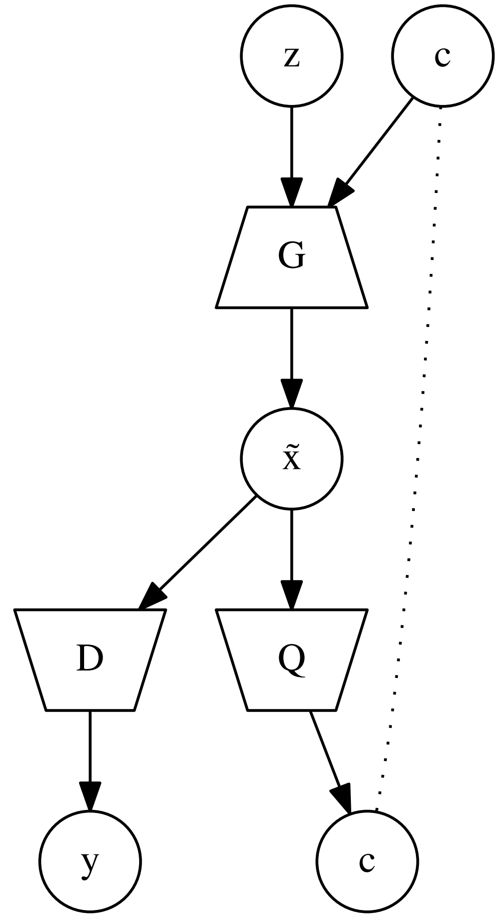
M5: BiGAN, ALI (2016)





M6: InfoGAN (2016)

- GAN noise (z)
  - is unstructured
  - can be partly ignored by G
- InfoGAN idea and principle
  - part of the noise is a code c
  - enforce high mutual information between c and $\tilde{x}$
  - in practice, predict c from $\tilde{x}$
  - use a coder Q
- Structure in the code ✎
  - Cartesian product of anything
  - (categorical, continuous, ...)



- $\min_{G,Q} \max_{D} V_{InfoGAN}(D,G,Q) = V_{InfoGAN}(D,G) - \lambda L_I(G,Q)$

InfoGAN: some results

(a) Varying c_1 on InfoGAN (Digit type)

(b) Varying c_1 on regular GAN (No clear meaning)

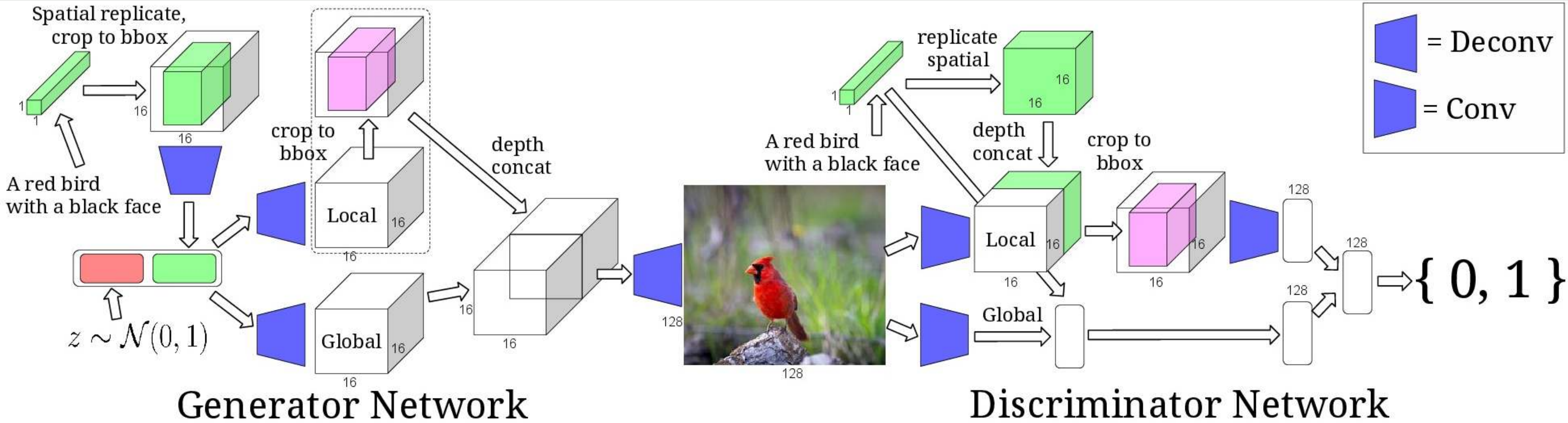
(c) Varying c_2 from -2 to 2 on InfoGAN (Rotation)

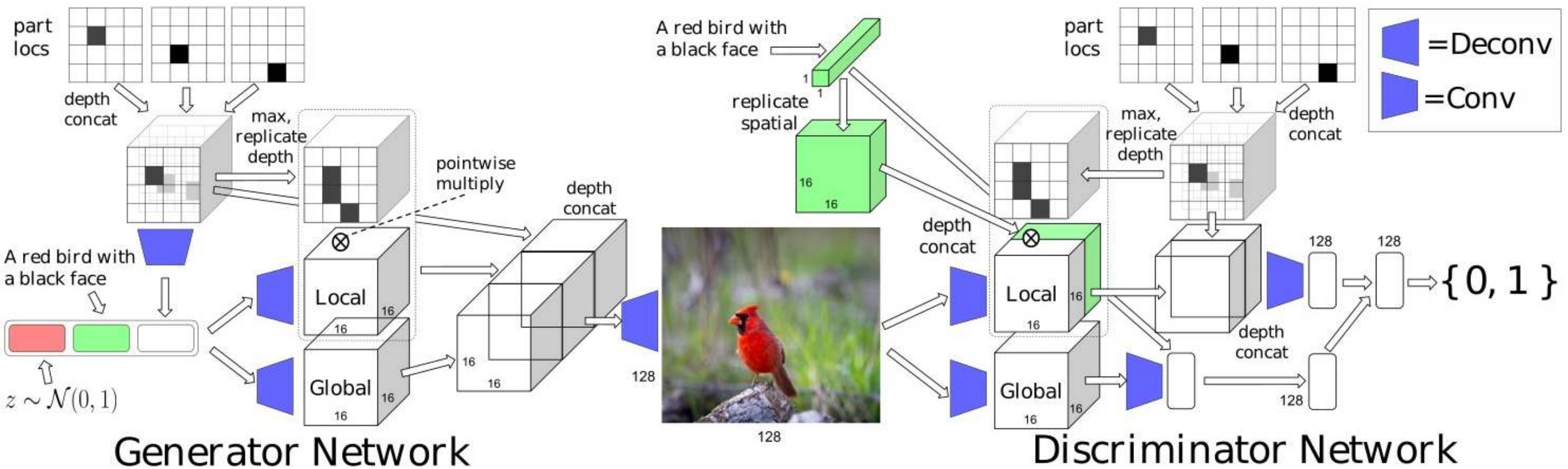
(d) Varying c_3 from -2 to 2 on InfoGAN (Width)

*GAN as a Modeling Tool

- Conditional GANs and variants (2016)
  - the GAN process is conditioned on some data
  - e.g., image generation condition on a semantic mask
  - e.g., image conditioned on a text sentence
  - e.g., audio conditioned on a text sentence
  - e.g., image conditioned on class and keypoints
  - …
- Very complex (and operational) setups

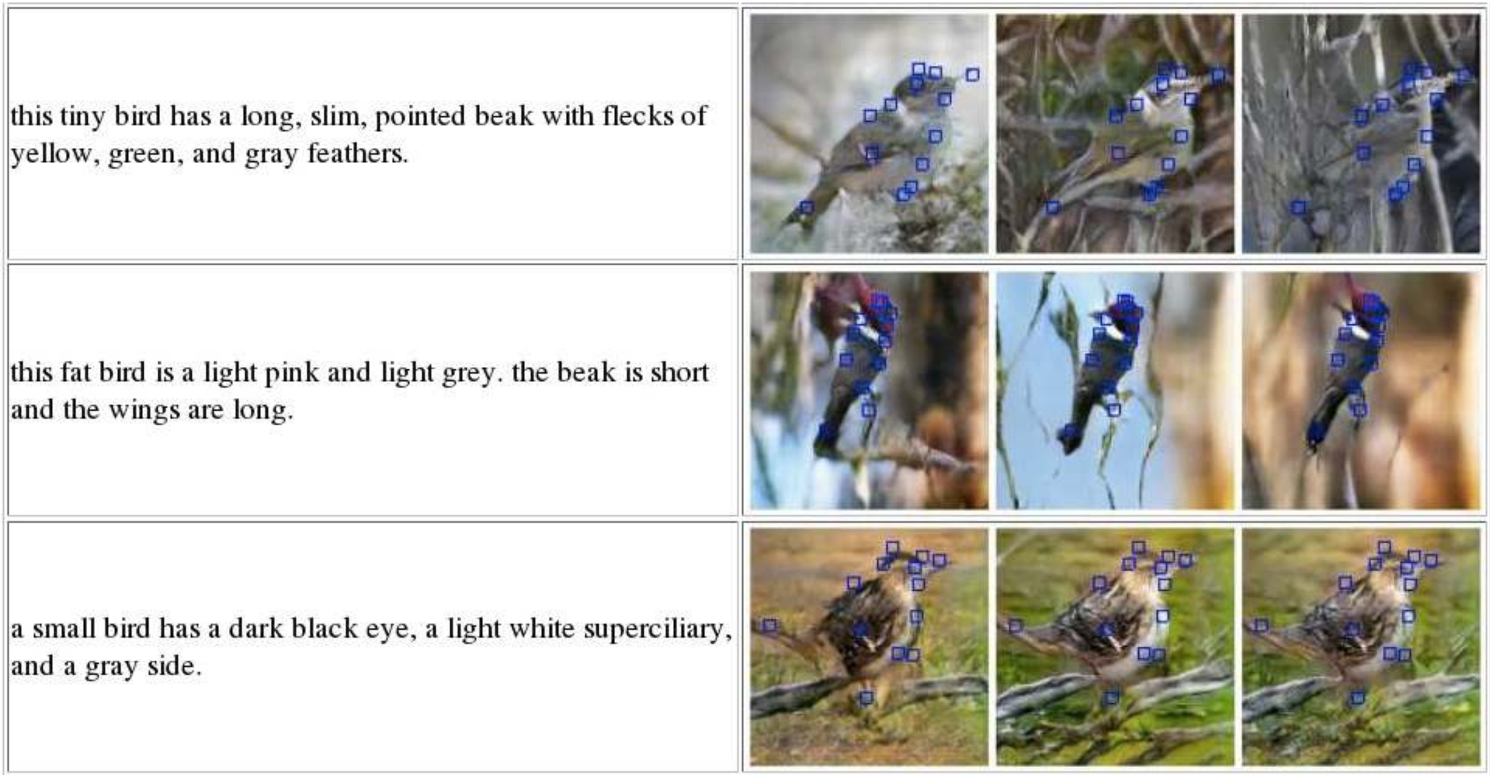
Ex: Learning What and Where to Draw





Ex: Learning What and Where to Draw

- Scott Reed et al.



- Unsupervised Representation Learning
- Notations and problem formulation
- Probabilistic (graphical) models
- Auto-encoders
- Generative Adversarial Networks
    - adversarial examples and training
    - GANs
- Focus on …
    - optimization
    - space and time convolutions
    - *depth*, *breadth/width*
    - semantics
    - sequential/temporal aspects
    - recent GANs
- Wrap up

Take-home Message

Return of the generative approaches.

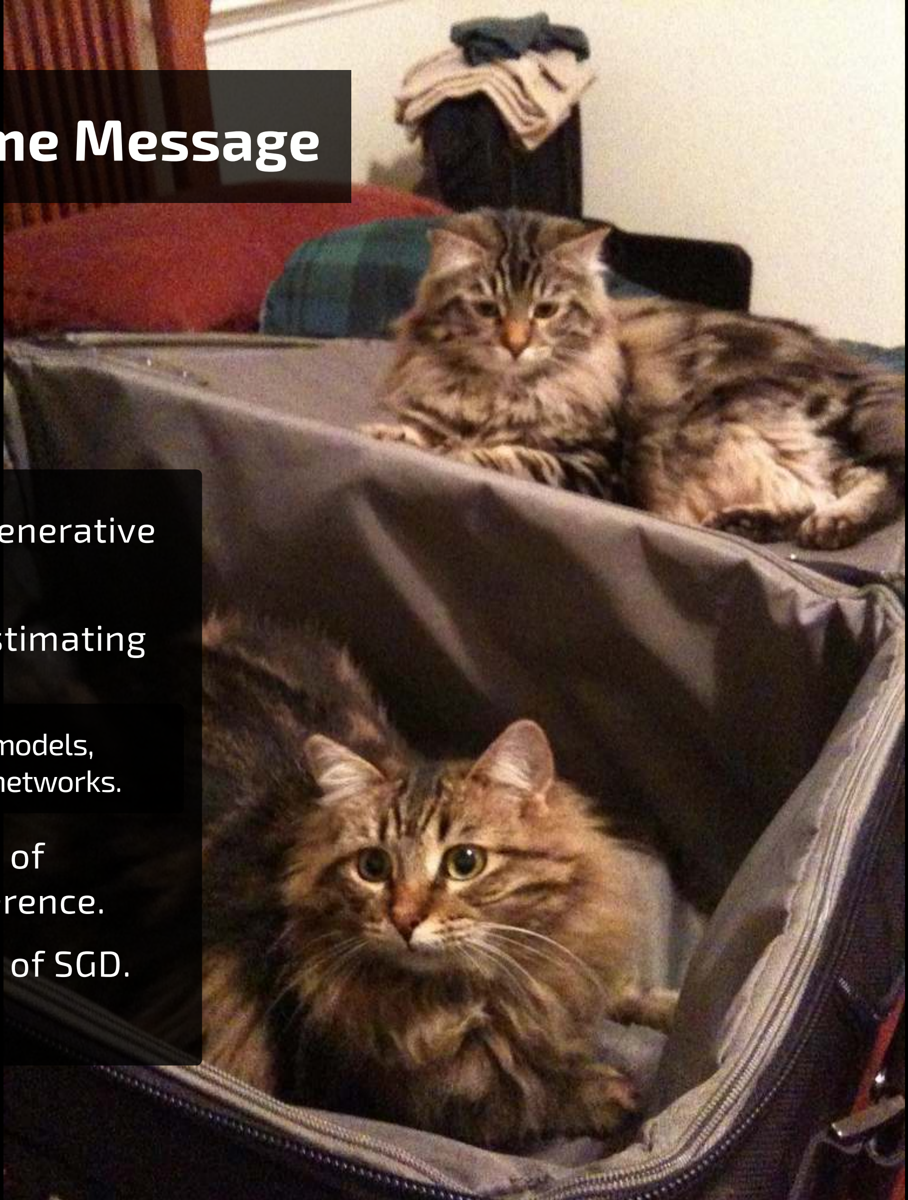
Two ways of estimating densities

- generative models,
- generative networks.

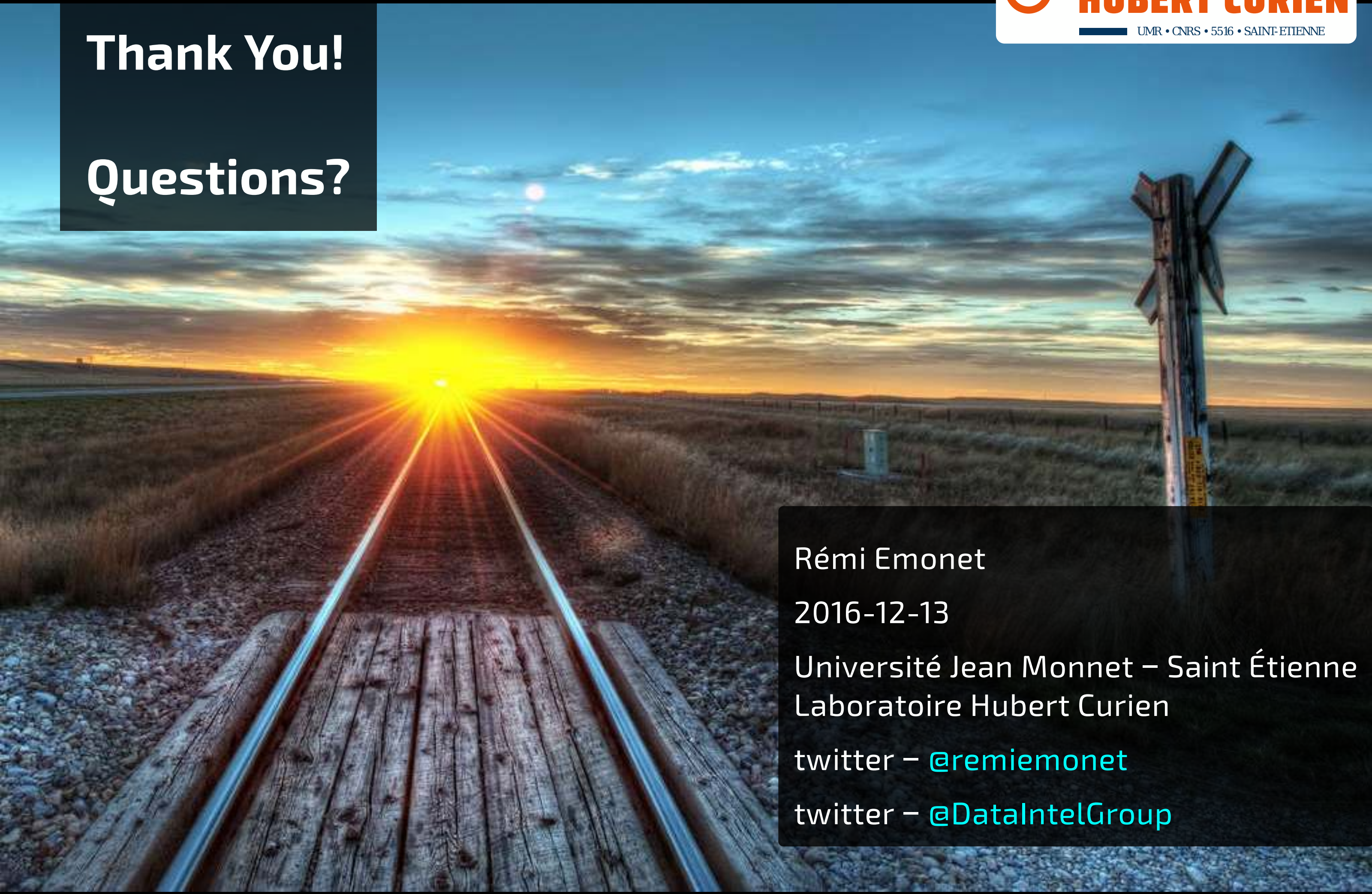
The golden age of Variational Inference.

The golden age of SGD.

Saddle points!



LABORATOIRE HUBERT CURIEN



Thank You!

Questions?

Rémi Emonet

2016-12-13

Université Jean Monnet – Saint Étienne
Laboratoire Hubert Curien

twitter – @remiemonet

twitter – @DataIntelGroup

Attributions


sayumiQ
CC by sayumiQ (Flickr)


NEXT 96 km
seefit

govan riverside


BCE ECB ЕЦБ EZB EKP EKT EKB BCE EBC 2013
5
EURO
EYPΩ
5
GorissenM
CC by GorissenM (Flickr)

someToast


ShutterRunner
CC by ShutterRunner (Flickr)