CSC 411: Introduction to Machine Learning

CSC 411 Lecture 24: Review & Outlook

Mengye Ren and Matthew MacKay

University of Toronto

Review

Supervised learning: regression, classification

- Choose model, loss function, optimizer
  - MSE
  - cross entropy
  - hinge loss
  - exponential loss
  - Iterative optimization vs. closed-form solutions
- Loss function + regularization vs. Bayesian
  - MLE, MAP, posterior, posterior predictive
  - Bayes rule & conjugate prior

Review

Supervised learning: regression, classification

- Parametric vs. nonparametric
  - Parametric: LR, DT, NN, SVM, Bayesian LR
  - Non-parametric: k-NN, GP, kernel trick (kernelized version of parametric models)
  - Time and space complexity
- Generative vs. discriminative
  - Generative: Naive Bayes, Gaussian Bayes
  - Decision boundary

Review

- Unsupervised learning: learning useful representation from data without label
  - dimensionality reduction: PCA, autoencoder
  - latent variable models: k-means, GMM
  - EM algorithm, likelihood lower bound
  - matrix factorization, sparse coding
- Reinforcement learning: learning through interaction with an environment
  - MDP, value function, Bellman operator
  - planning: value iteration, policy iteration
  - learning: Q-learning, deep Q-learning

CSC421: Neural Nets

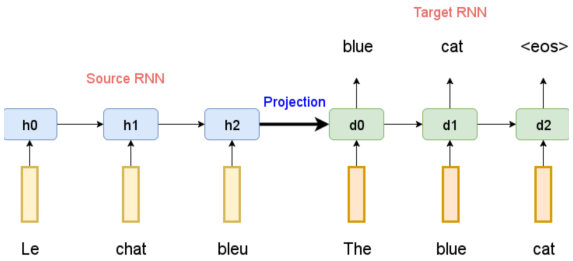
- This course covered some fundamental ideas, most of which are more than 10 years old.
- Big shift of the past decade: neural nets and deep learning
  - 2010: neural nets significantly improved speech recognition accuracy (after 20 years of stagnation)
  - 2012–2015: neural nets reduced error rates for object recognition by a factor of 6
  - 2016: a program called AlphaGo defeated the human Go champion
  - 2016: neural nets bridged half the gap between machine and human translation
  - 2015–2018: neural nets learned to produce convincing high-resolution images

CSC421: Automatic Differentiation

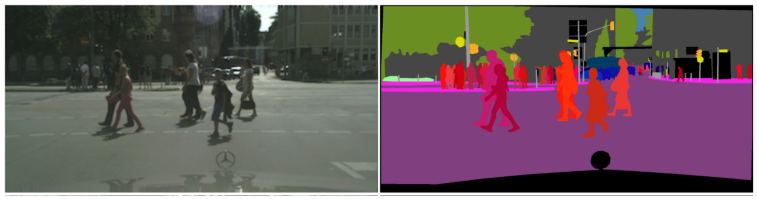
- In this course, you derived update rules by hand
- Backprop is totally mechanical. Now we have automatic differentiation tools that compute gradients for you.
- In CSC421, you learn how an autodiff package can be implemented
  - Lets you do fancy things like differentiate through the whole training procedure to compute the gradient of validation loss with respect to the hyperparameters.
- With TensorFlow, PyTorch, etc., we can build much more complex neural net architectures that we could previously.

CSC421: Beyond Scalar/Discrete Targets

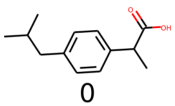
- This course focused on regression and classification, i.e. scalar-valued or discrete outputs. Sometimes we would like more structured output:
- text (e.g. image captioning, machine translation)



- dense labels of images (e.g. semantic segmentation)



- graphs (e.g. molecule design)

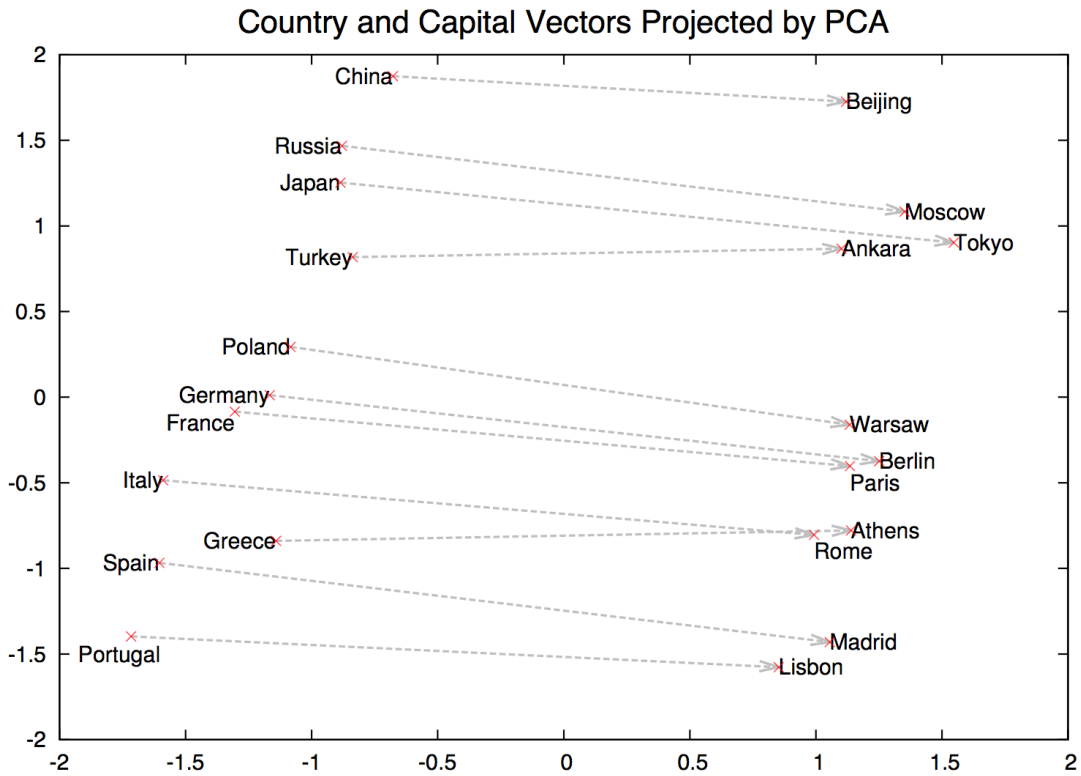


CSC421: Representation Learning

- We talked about neural nets as learning feature maps you can use for regression/classification
- More generally, want to learn a representation of the data such that mathematical operations on the representation are semantically meaningful
- Classic (decades-old) example: representing words as vectors
  - Measure semantic similarity using the dot product between word vectors (or dissimilarity using Euclidean distance)
  - Represent a web page with the average of its word vectors

CSC421: Representation Learning

- Here's a linear projection of word representations for cities and capitals into 2 dimensions (part of a representation learned using word2vec)
- The mapping $\text{city} \rightarrow \text{capital}$ corresponds roughly to a single direction in the vector space:



Mikolov et al., 2018, "Efficient estimation of word representations in vector space"

CSC421: Representation Learning

- In other words, $\text{vec}(\text{Paris}) - \text{vec}(\text{France}) \approx \text{vec}(\text{London}) - \text{vec}(\text{England})$
- This means we can analogies by doing arithmetic on word vectors:
  - e.g. "Paris is to France as London is to _____"
  - Find the word whose vector is closest to $\text{vec}(\text{France}) - \text{vec}(\text{Paris}) + \text{vec}(\text{London})$
- Example analogies:

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

Mikolov et al., 2018, "Efficient estimation of word representations in vector space"

CSC421: Representation Learning

One of the big goals is to learn *disentangled* representations, where individual dimensions tell you something meaningful

(a) Baldness (-6, 6)

(b) Face width (0, 6)

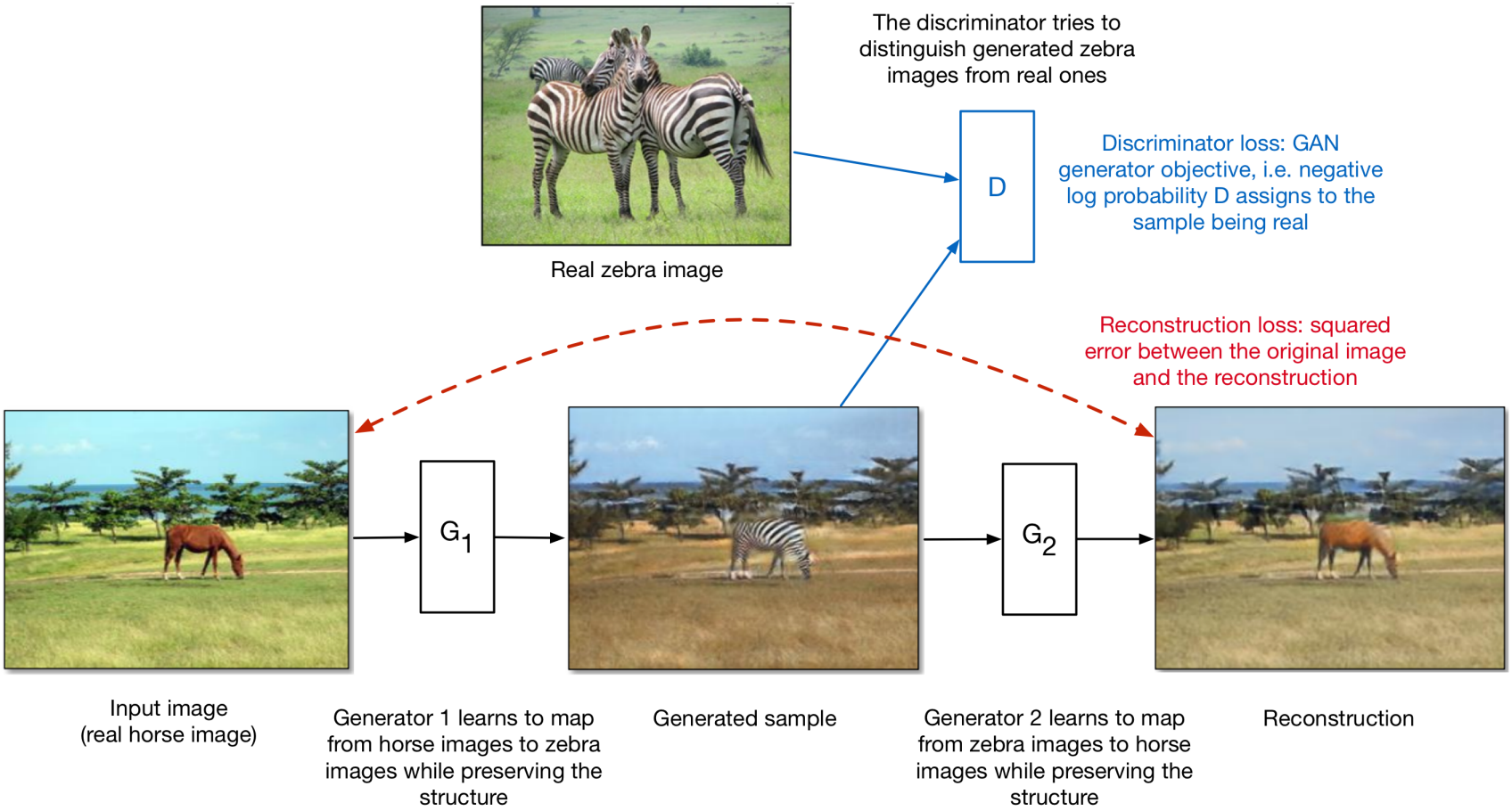
(c) Gender (-6, 6)

(d) Mustache (-6, 0)

Chen et al., 2018, "Isolating sources of disentanglement in variational autoencoders"

CSC421: Image-to-Image Translation

Due to convenient autodiff frameworks, we can combine multiple neural nets together into fancy architectures. Here's the CycleGAN.

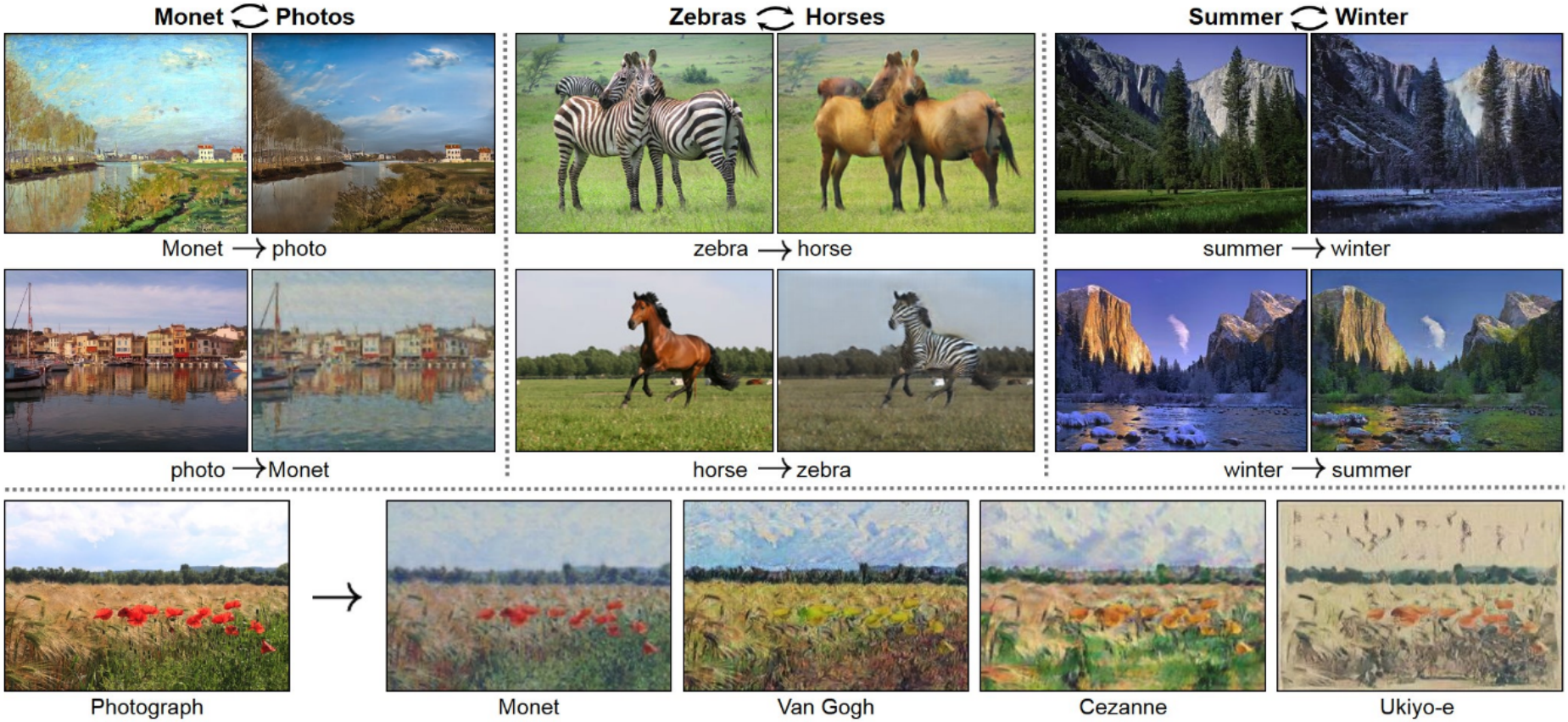


Total loss = discriminator loss + reconstruction loss

Zhu et al., 2017, "Unpaired image-to-image translation using cycle-consistent adversarial networks"

CSC421: Image-to-Image Translation

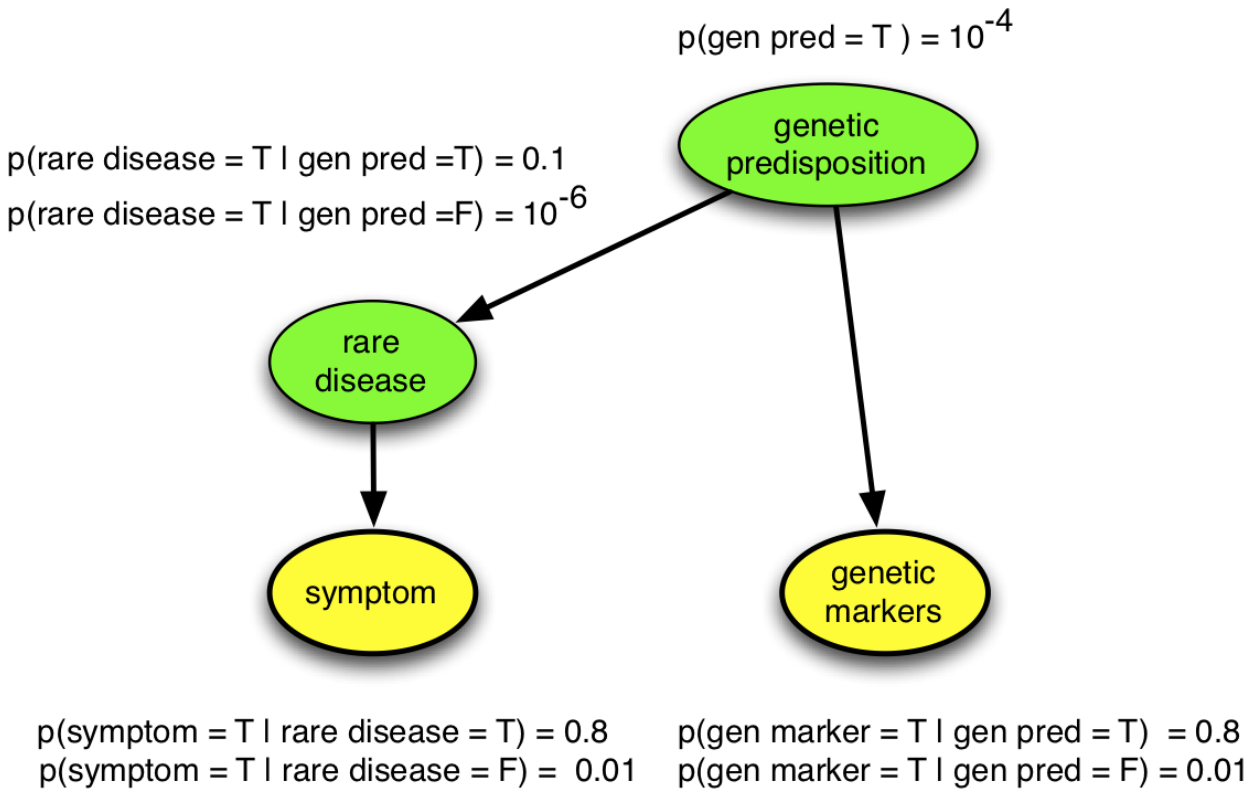
Style transfer problem: change the style of an image while preserving the content.



Data: Two unrelated collections of images, one for each style

CSC412: Probabilistic Graphical Models

- In this course, we just scratched the surface of probabilistic models.
- Probabilistic graphical models (PGMs) let you encode complex probabilistic relationships between lots of variables.



Ghahramani, 2015, "Probabilistic ML and artificial intelligence"

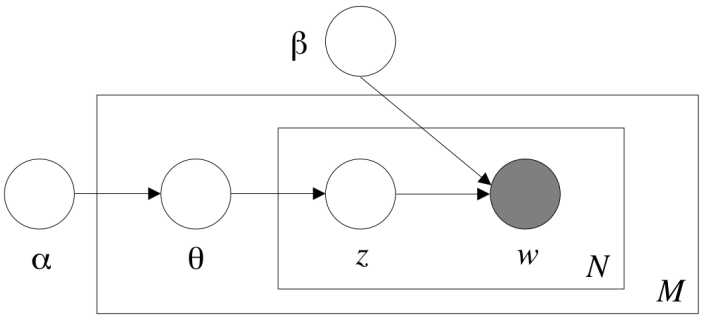
CSC412: PGM Inference

- We derived inference methods by inspection for some easy special cases (e.g. GDA, naïve Bayes)
- In CSC412, you'll learn much more general and powerful inference techniques that expand the range of models you can build
  - Exact inference using dynamic programming, for certain types of graph structures (e.g. chains)
  - Markov chain Monte Carlo
    - forms the basis of a powerful probabilistic modeling tool called Stan
  - Variational inference: try to approximate a complex, intractable, high-dimensional distribution using a tractable one
    - Try to minimze the KL divergence
    - Based on the same math from our EM lecture

CSC412: Beyond Clustering

- We've seen unsupervised learning algorithms based on two ways of organizing your data
  - low-dimensional spaces (dimensionality reduction)
  - discrete categories (clustering)
- Other ways to organize/model data
  - hierarchies
  - dynamical systems
  - sets of attributes
  - topic models (each document is a mixture of topics)
- Motifs can be combined in all sorts of different ways

Latent Dirichlet Allocation (LDA)



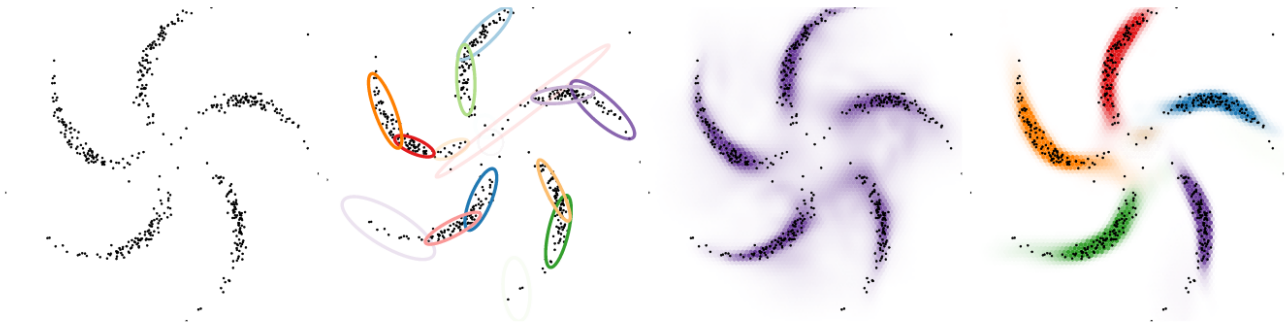
"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Blei et al., 2003, "Latent Dirichlet Allocation"

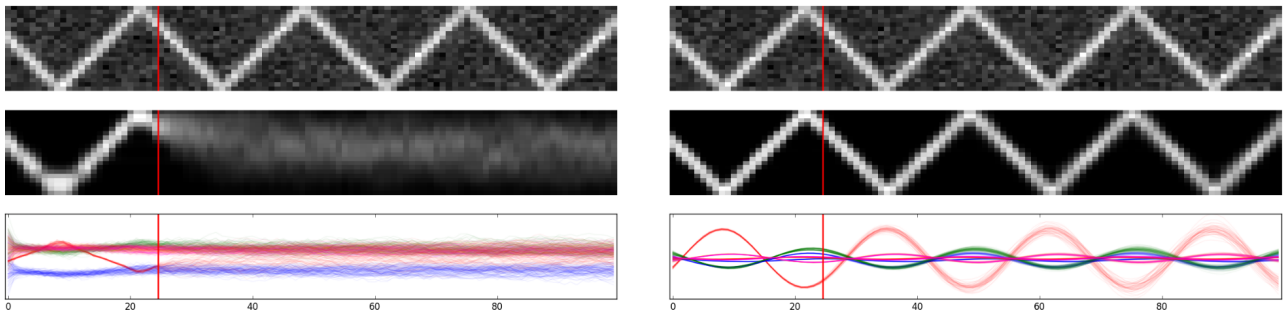
CSC412: Beyond Clustering

Interpretable latent structure:



(a) Data (b) GMM (c) Density net (VAE) (d) GMM SVAE

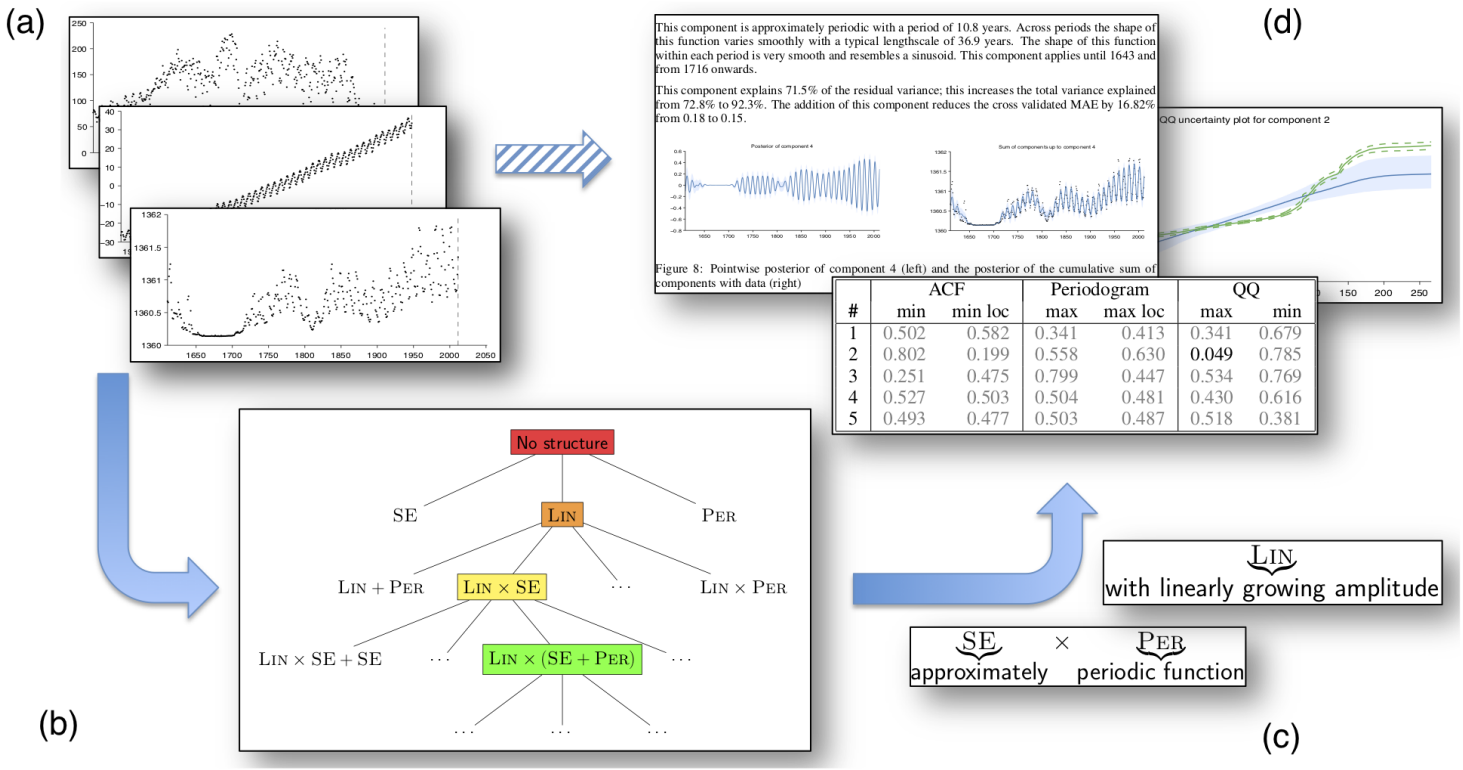
Dynamical system:



(a) Predictions after 200 training steps. (b) Predictions after 1100 training steps.

Johnson et al., 2016, "Composing graphical models with neural networks for structured representations and fast inference"

Automatic search over Gaussian process kernel structures



#	ACF min	ACF min loc	Periodogram max	Periodogram max loc	QQ max	QQ min
1	0.502	0.582	0.341	0.413	0.341	0.679
2	0.802	0.199	0.558	0.630	0.049	0.785
3	0.251	0.475	0.799	0.447	0.534	0.769
4	0.527	0.503	0.504	0.481	0.430	0.616
5	0.493	0.477	0.503	0.487	0.518	0.381

Duvenaud et al., 2013, "Structure discovery in nonparametric regression through compositional kernel search"
Image: Ghahramani, 2015, "Probabilistic ML and artificial intelligence"

Resources

Continuing with machine learning

- Courses
  - CSC 421/2516, "Neural Networks and Deep Learning"
  - CSC 412/2506, "Probabilistic Learning and Reasoning"
  - Various topics courses (varies from year to year)
- Videos from top ML conferences (NIPS/NeurIPS, ICML, ICLR, UAI)
  - Tutorials and keynote talks are aimed at people with your level of background (know the basics, but not experts in a subfield)
- Try to reproduce results from papers
  - If they've released code, you can use that as a guide if you get stuck
- Lots of excellent free resources avaiable online!