Tutorial: Introduction to Machine Learning

Arindam Banerjee
banerjee@cs.umn.edu

Dept of Computer Science & Engineering
University of Minnesota, Twin Cities

2011 NASA Conference on Intelligent Data Understanding

October 19, 2011

Success Stories

Parent of a baby

Today's Recommendations For You

Here's a daily sample of items recommended for you. Click here to **see all recommendations**.

Page 1 of 44

Pampers Sensitive 3X Wipes 192 Count (Pack of 4)
(233) $25.94
Fix this recommendation

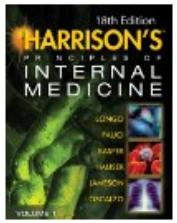
Munchkin 6 Pack Soft-Tip Infant Spoon, Colors M...
(222) $4.49
Fix this recommendation

Playtex Diaper Genie II Refill (Pack of 3)
(109) $18.00
Fix this recommendation

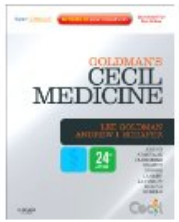
Fisher-Price Brilliant Basics Baby's First Blocks
(212) $8.79
Fix this recommendation

New For You®

Page 1 of 12



Harrison's Principles of Internal M... (Hardcover) by Dan Longo
(2) $142.97
Fix this recommendation



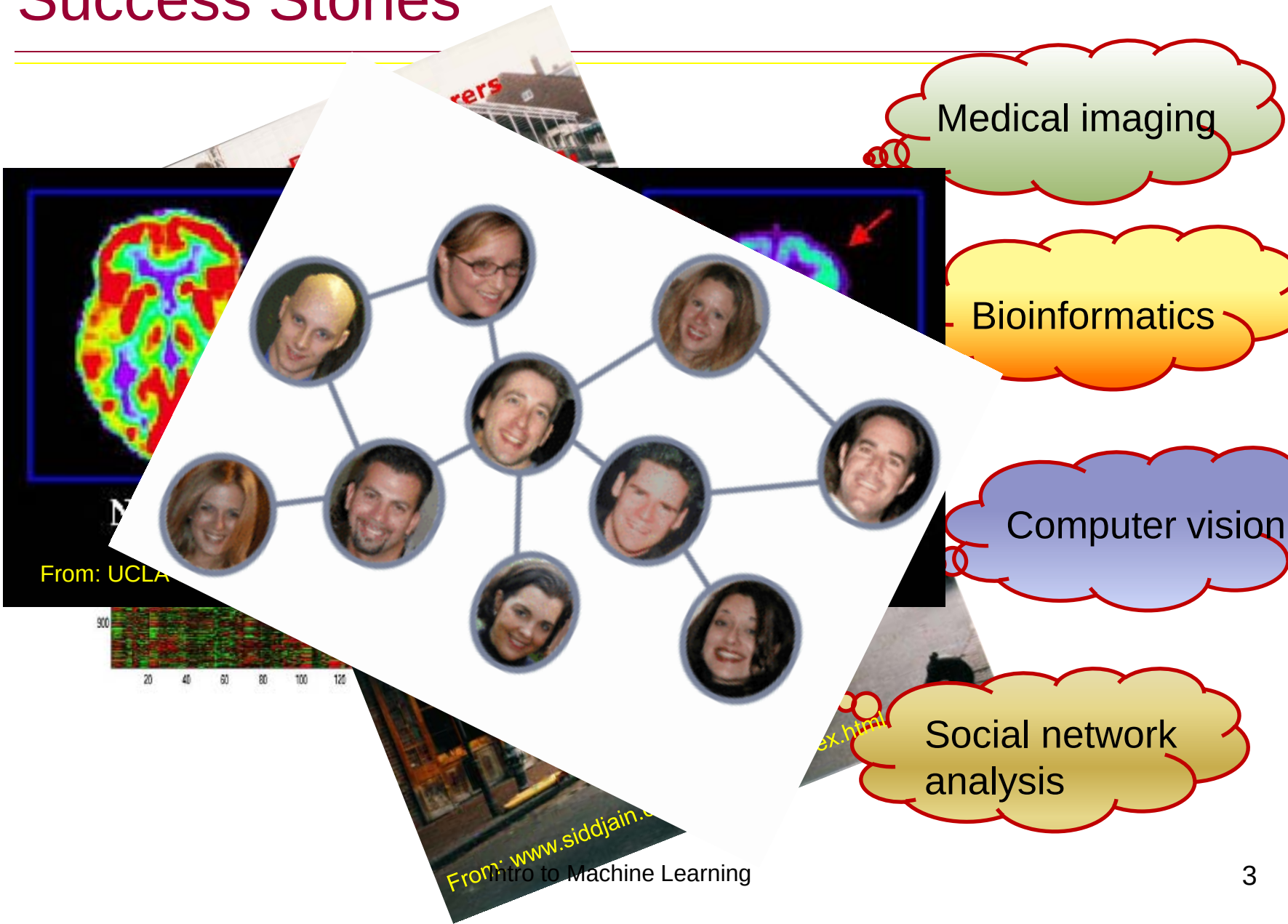
Goldman's Cecil Medicine: Expe... (Hardcover) by Lee Goldman MD
$157.39
Fix this recommendation

Tap into Your Friends BETA

Connect to Facebook to get Amazon recommendations for you and discover your friends' Favorites and Likes

Medical Doctor

Success Stories



Medical imaging

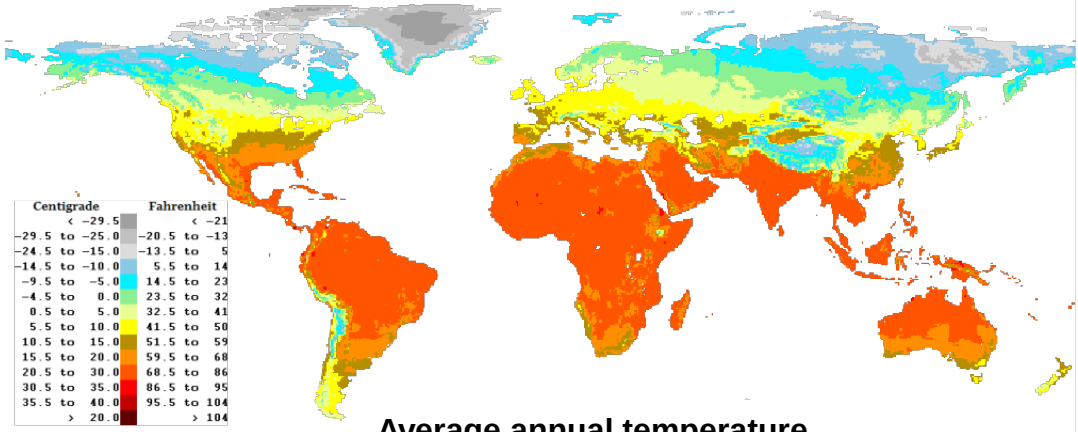
Bioinformatics

Computer vision

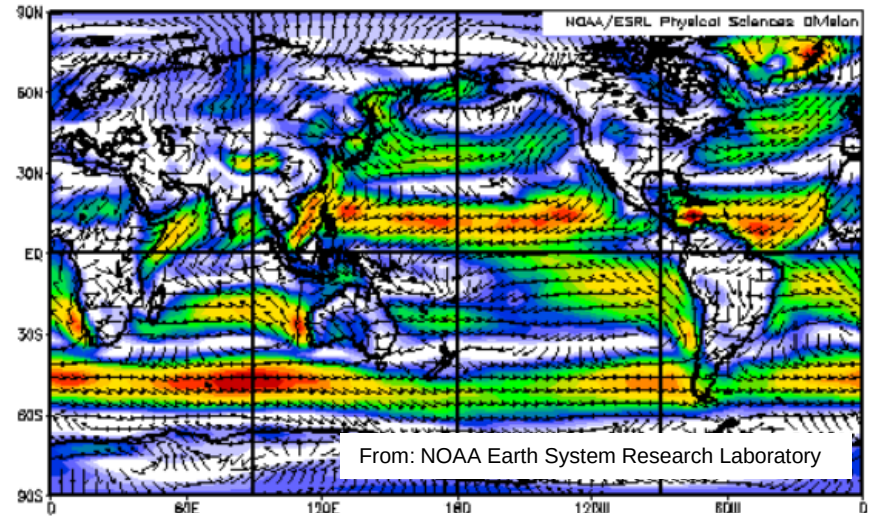
Social network analysis

Types of Data

- Data Types:
  - Discrete, Ordinal, Continuous
  - Univariate, Multi-variate
  - Structured, Temporal, Spatial, Spatiotemporal
- Data Dependencies
  - Independent and Identically Distributed (IID)
  - Graphical Models
  - Linear vs Non-linear dependencies
- Other Considerations
  - Observable vs Latent variables



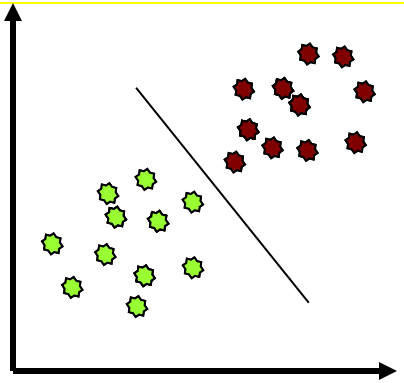
Average annual temperature



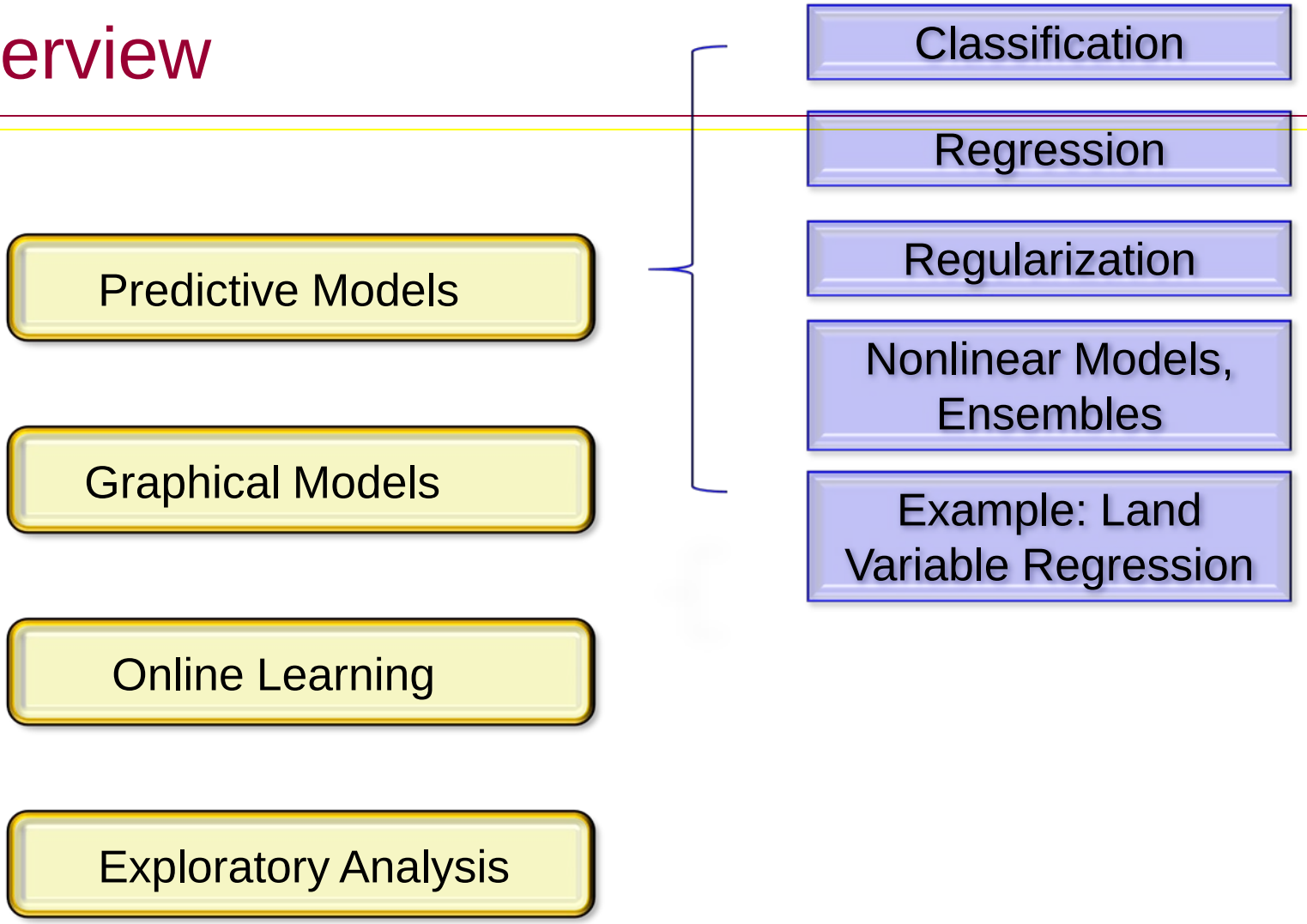
Average wind direction

Types of Learning

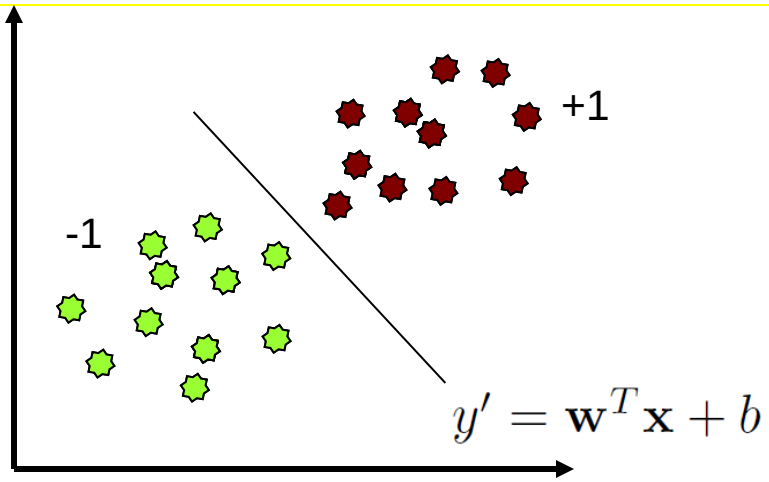
- Supervised:
  - Given $\{\mathbf{x}_i, y_i\}$, learn $f : X \rightarrow Y$
  - Examples : Classification, Regression



Overview

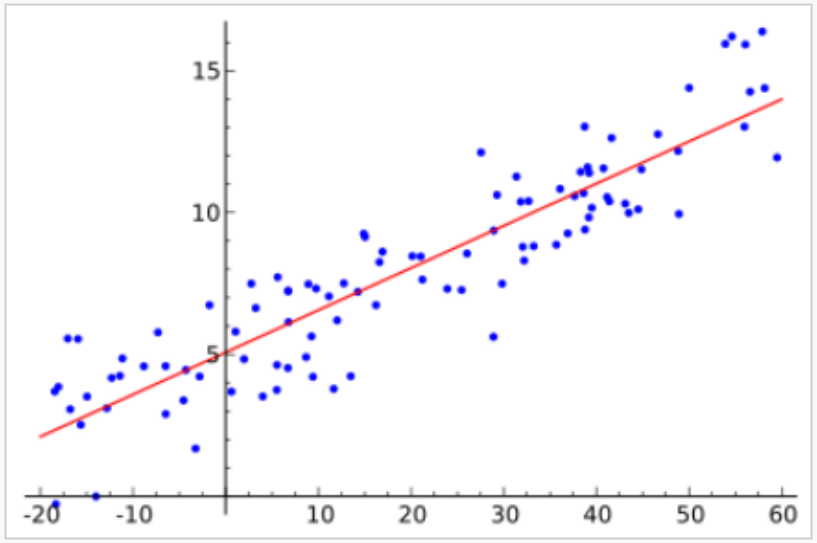

Predictive Models
Graphical Models
Online Learning
Exploratory Analysis
Classification
Regression
Regularization
Nonlinear Models, Ensembles
Example: Land Variable Regression

Classification: Linear Models



- Linear Classification Rule
  - If $\mathbf{w}^T\mathbf{x}_i + b \geq 0$, then $y_i = +1$
  - If $\mathbf{w}^T\mathbf{x}_i + b < 0$, then $y_i = -1$
- Family of methods:
  - Support Vector Machines, Perceptrons, Decision Stumps
- Two Aspects: Optimization, Statistical guarantees

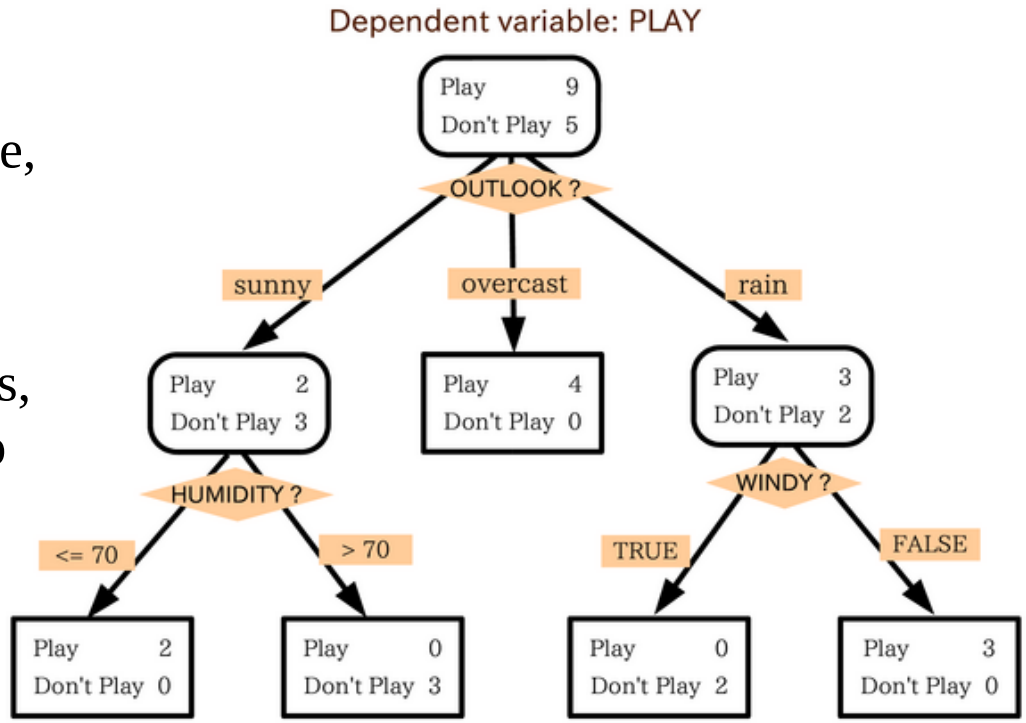
Regression: Linear Models



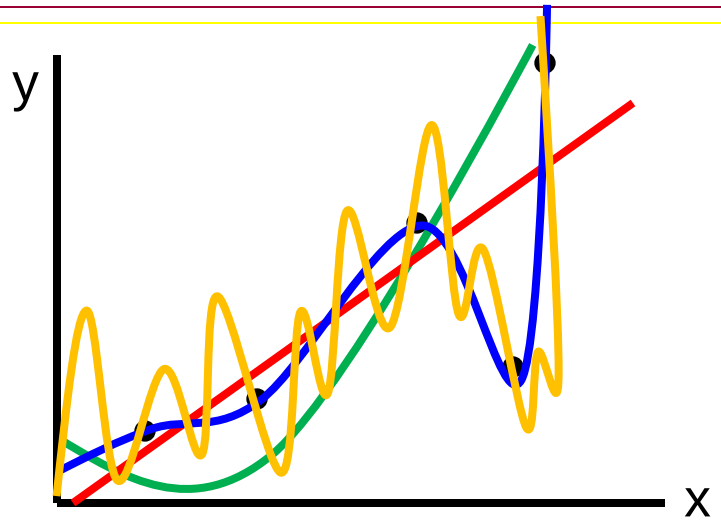
- Linear Regression:
  - Model: $y_i = \mathbf{w}^T \mathbf{x}_i + w_0 + \epsilon_i$
- Family of methods/models:
  - Least squares, Ridge regression, Sparse regression
  - Generalized linear models, Exponential family noise
- Two Aspects: Optimization, Statistical Guarantees

Hierarchical Linear Models

- Hierarchical Models
  - Each node is a linear model
  - Node outcome (decision, value, etc.) is fed into next node
- Family of models/methods:
  - Classification/Regression trees, Multi-layer Perceptrons, Deep Belief Networks
- Two Aspects:
  - Learning/Optimization
  - Statistical Guarantees



Regularization, Generalization



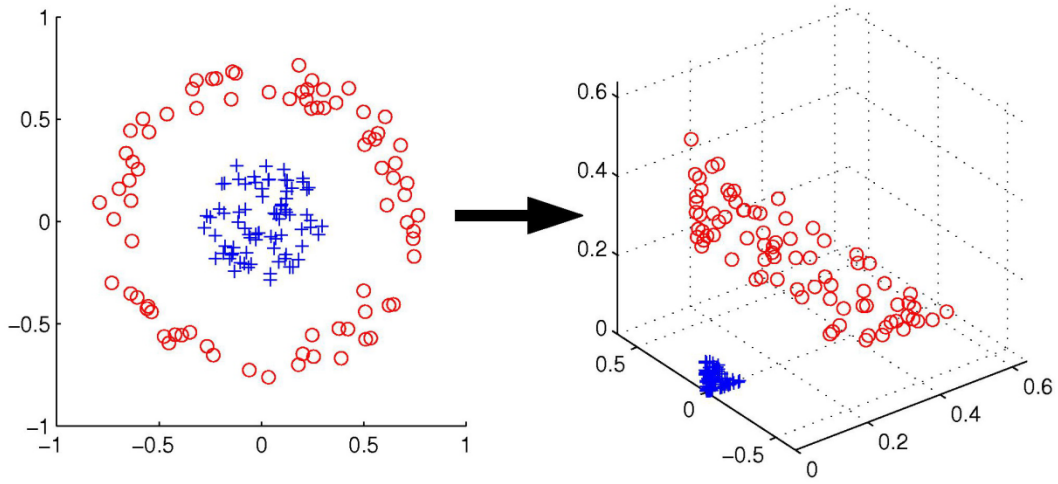
- Controlling complexity of predictor is key

$$L(\mathbf{w}, \{\mathbf{x}_i, y_i\}) + \lambda R(\mathbf{w})$$

Loss term Regularization term

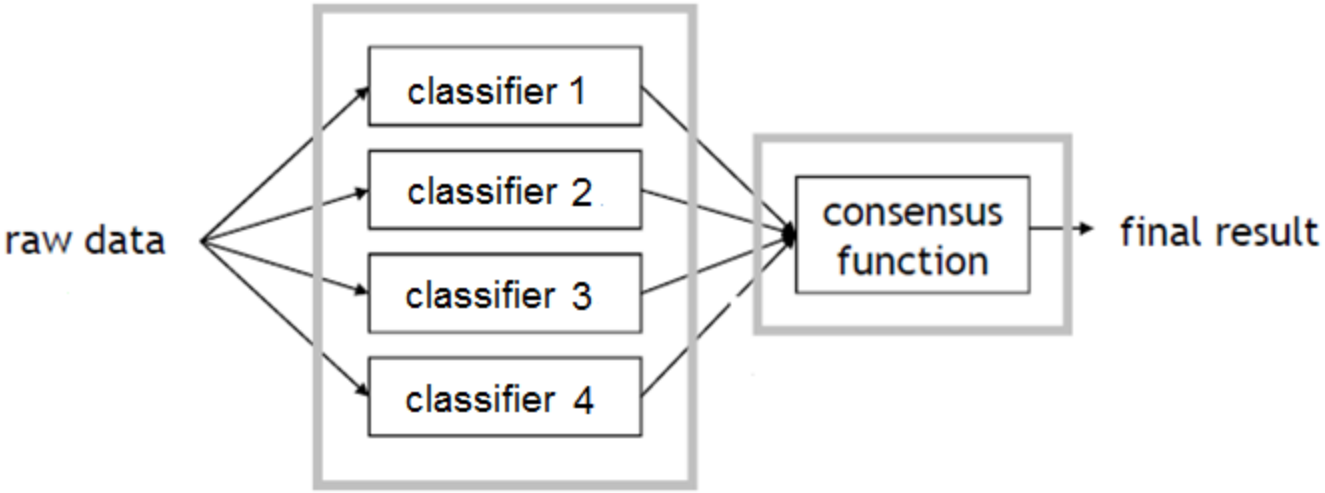
  - Low loss with low complexity predictor guarantees good generalization
- Examples:
  - L_2 (SVM, Ridge Regression), L_1 (Lasso)
  - Bayesian models: Prior penalizes complex models, $R(\mathbf{w}) = -\log p(\mathbf{w})$

Non-linear Methods



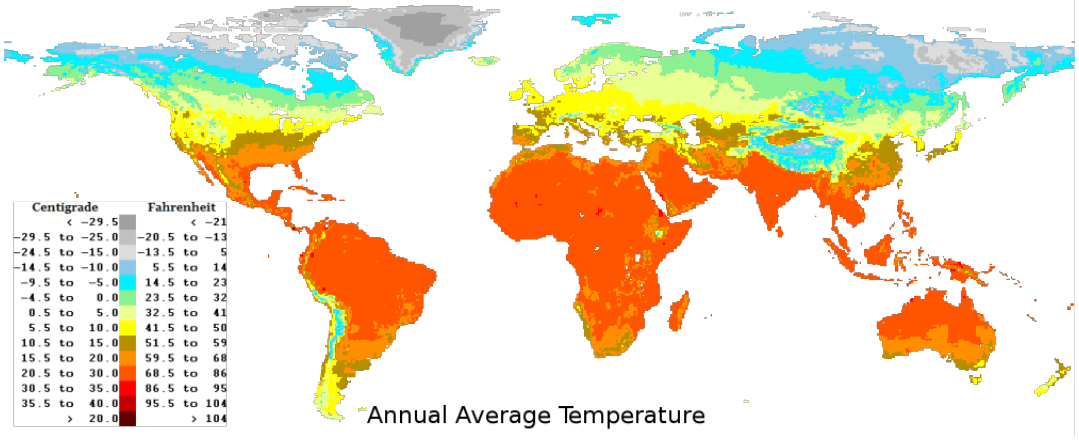
- The Kernel Trick
  - Map data to high-dimensional space (“RKHS”)
  - Linear model in RKHS ≡ Nonlinear model in data space
  - Avoid explicit modeling in high-dimensional spaces
- Models and Methods
  - Kernelized Classification, Regression, Dimensionality Reduction
  - Multiple Kernel Learning, Combination of Kernels

Ensembles

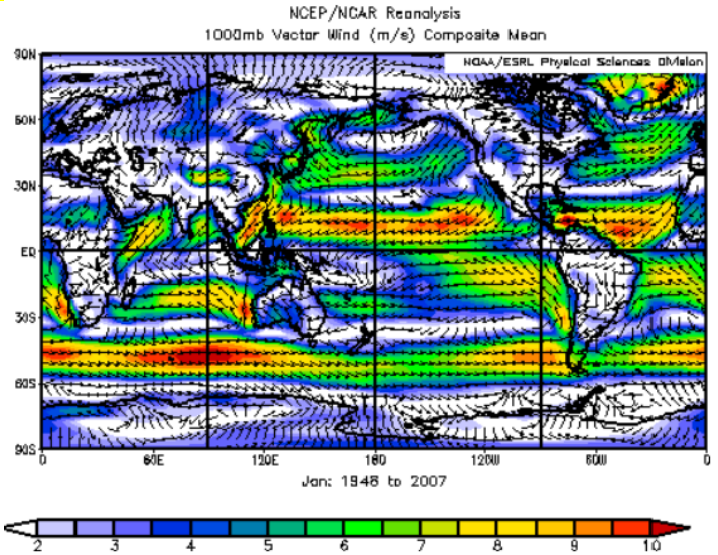


- Main idea:
  - Combination reduces loss, does not increase complexity
  - Bayesian model averaging
- Examples:
  - Bagging, Boosting, Random forests
  - Base models on samples/reweighted versions of points, features

High Dimensional Modeling



Annual Average Temperature

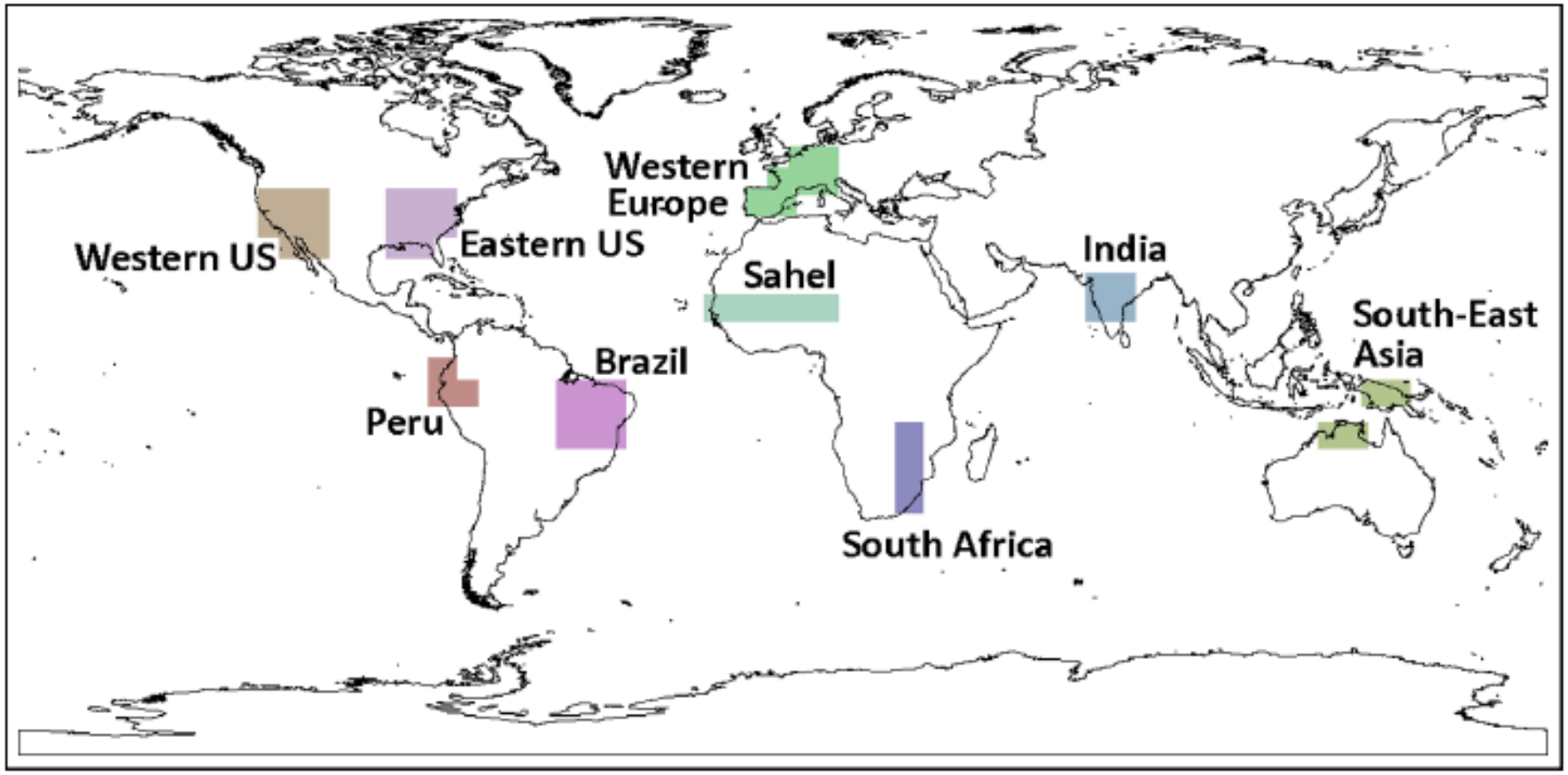


- The “Small n, High p” regime
  - Underspecified, difficult to fit models with guarantees
- “Low Dimensional” models
  - Small number of features are relevant: Feature Selection
  - Low intrinsic dimensionality: Manifold embedding
  - Models/Methods: Sparse regression, Non-parametric regression

Land Variable Regression

- Land-Sea variable interactions
  - How do sea variables affect land variables?
  - Are proximal locations important?
  - Are there long range spatial dependencies (tele-connections)?
- NCEP/NCAR Reanalysis 1: Monthly means for 1948-2010
  - Covariates: Temperature, Sea Level Pressure, Precipitation, Relative Humidity, Horizontal Wind Speed Vertical Wind Speed
  - Response: Temperature, Precipitation at 9 locations
- High-dimensional Regression
  - Dimension $p = Lm$, L locations, m variables/location
  - Two considerations:
    - Not all locations are relevant
    - Even if a location is relevant, not all variables are relevant

Land Regions for Prediction: Temp, Precip



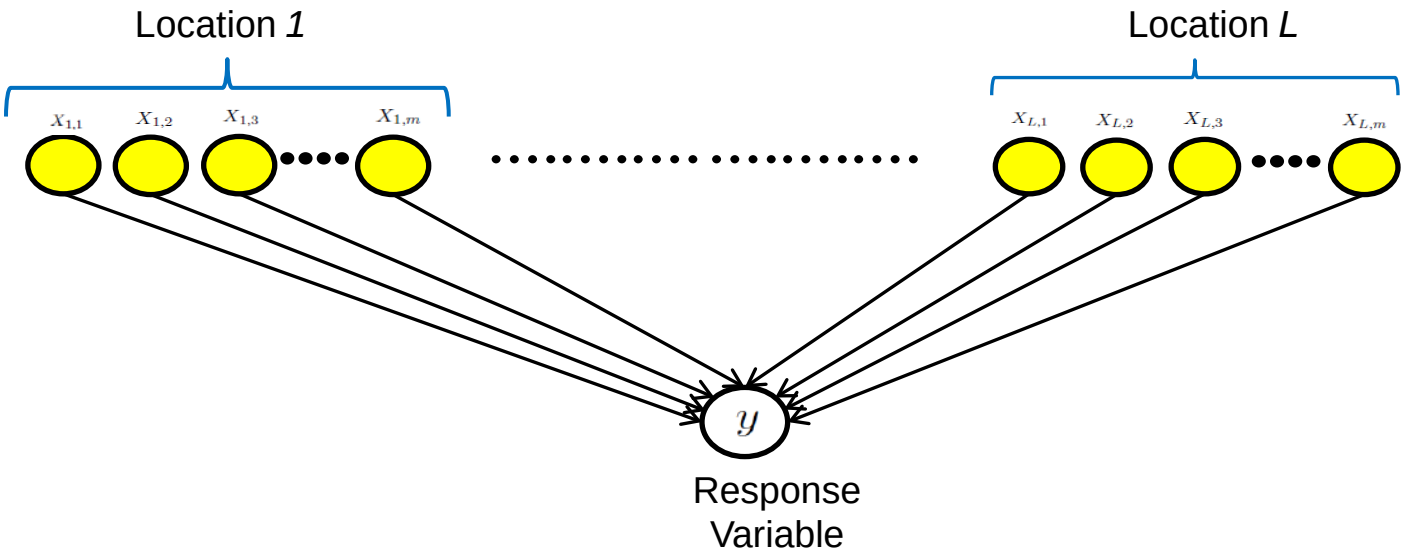
Land regions chosen for predictions

Ordinary Least Squares

- Naive method: Use *ordinary least squares* (OLS) :

$$\hat{\theta} := \arg min_{\theta \in \mathbb{R}^{mL}} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 \right\}$$

- For $n < mL$, OLS estimate $\hat{\theta}$ non-unique
- In general, y depends on all mL covariates: complex model



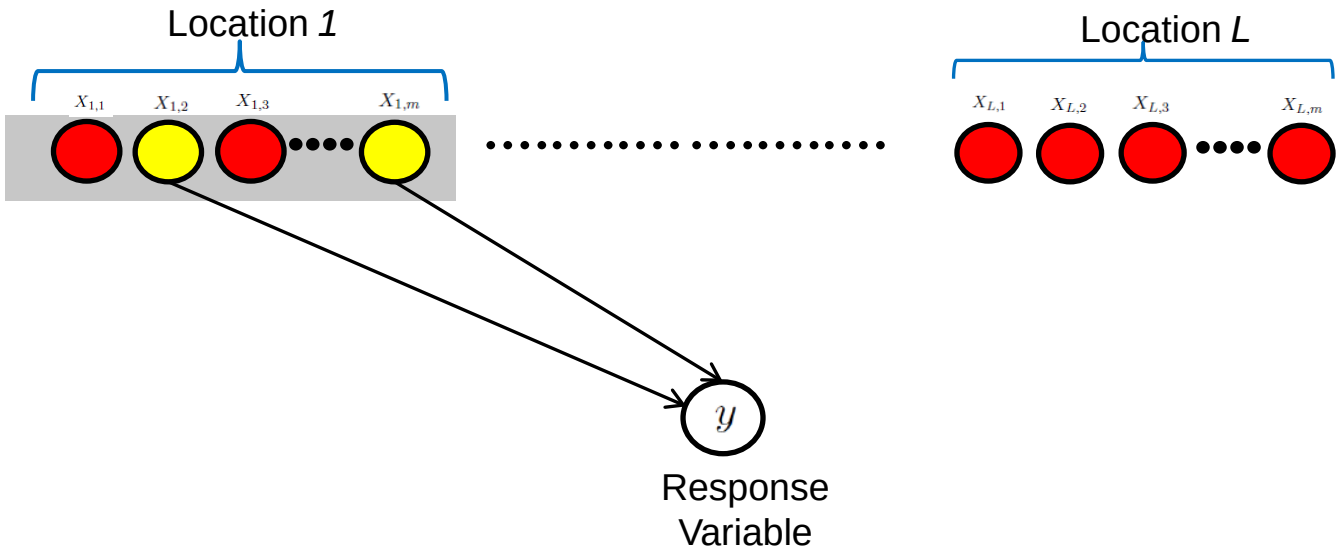
Sparse Group Lasso (SGL)

- High-dimensional Regression with “Sparse” “Group Sparsity”

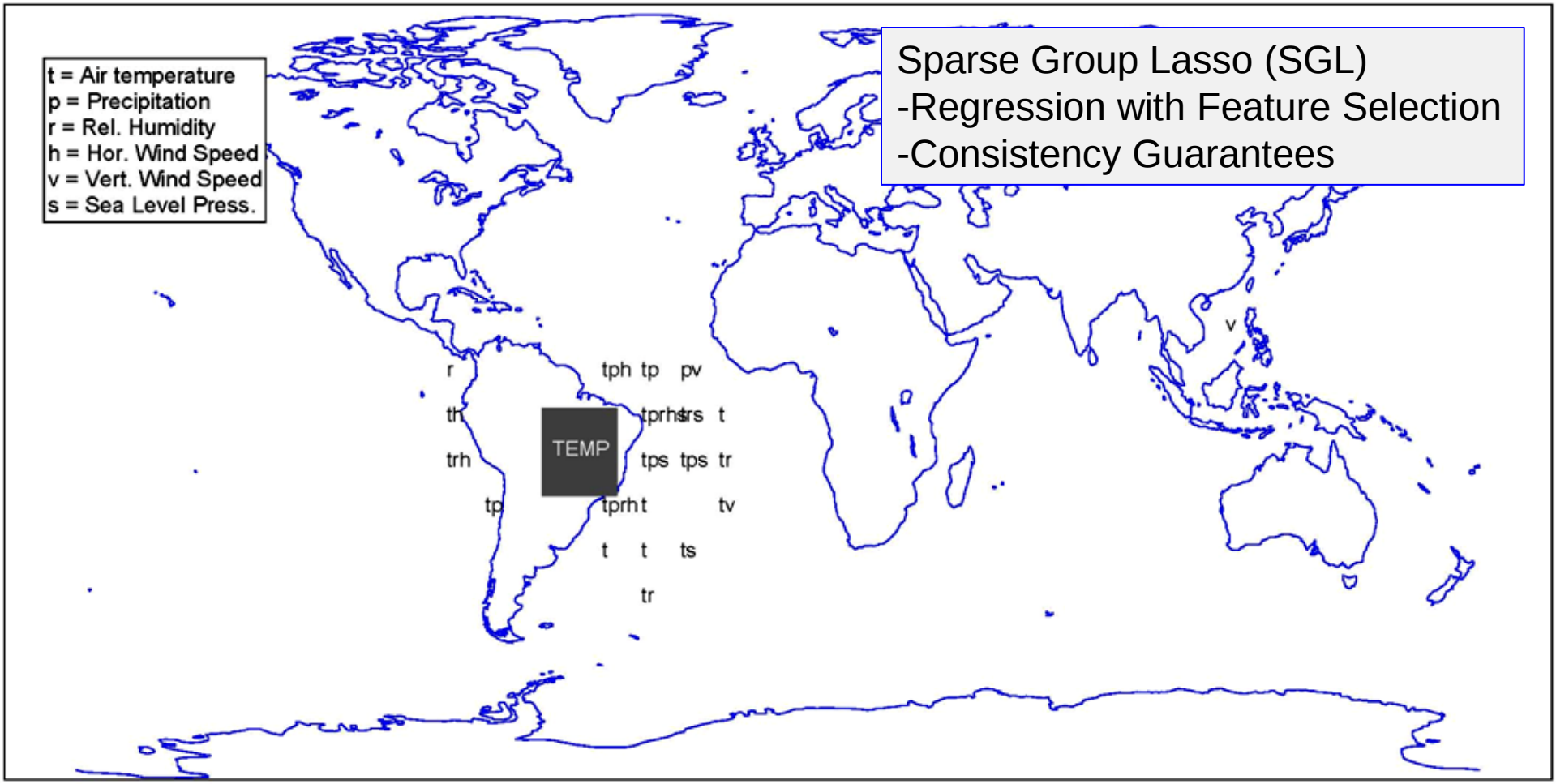
$$\hat{\theta}_{SGL} = \arg min_{\theta \in \mathbb{R}^{mL}} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda_1 \|\theta\|_1 + \lambda_2 \|\theta\|_{1,\mathcal{G}} \right\}$$

$$\|\theta\|_1 = \sum_{i=1}^{mL} |\theta_i| \qquad \|\theta\|_{1,\mathcal{G}} = \sum_{k=1}^{L} \|\theta_{G_k}\|_2$$

$$\mathcal{G} = \{G_1, \ldots, G_L\}\text{: groups of } m \text{ variables at } L \text{ locations}$$



Example: Land Variable Regression

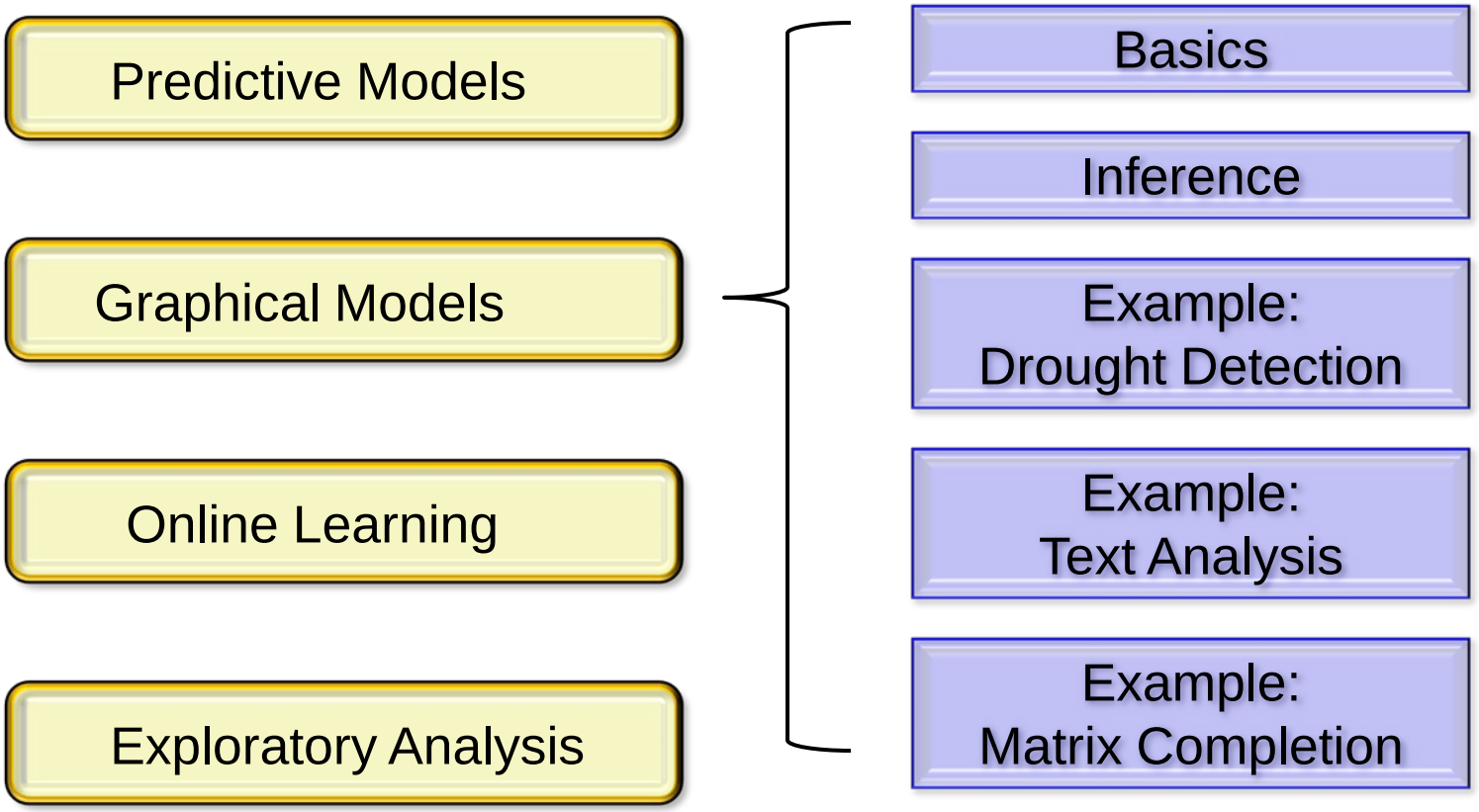


Relevant Variables for Temperature Prediction in Brazil

Error (RMSE): SGL, NC, OLS

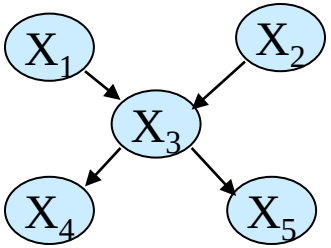
Variable	Region	Sparse Group Lasso	Network Clusters	Ordinary Least Squares
Air Temperature	Brazil	**0.198**	0.534	0.348
	Peru	**0.247**	0.468	0.387
	West USA	**0.270**	0.767	0.402
	East USA	**0.304**	0.815	0.348
	W Europe	**0.379**	0.936	0.493
	Sahel	**0.320**	0.685	0.413
	S Africa	**0.136**	0.726	0.267
	India	**0.205**	0.649	0.300
	SE Asia	**0.298**	0.541	0.383
Precipitation	Brazil	**0.261**	0.509	0.413
	Peru	**0.312**	0.864	0.523
	West USA	**0.451**	0.605	0.549
	East USA	**0.365**	0.686	0.413
	W Europe	**0.358**	0.45	0.551
	Sahel	**0.427**	0.533	0.523
	S Africa	**0.235**	0.697	0.378
	India	**0.146**	0.672	0.264
	SE Asia	**0.159**	0.665	0.312

Overview

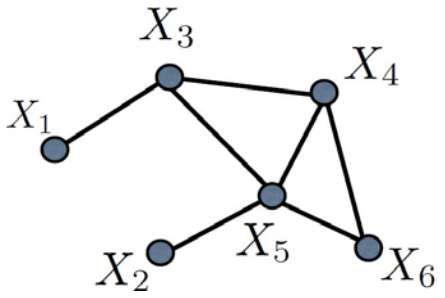

Predictive Models
Graphical Models
Online Learning
Exploratory Analysis
Basics
Inference
Example:
Drought Detection
Example:
Text Analysis
Example:
Matrix Completion

Graphical Models: What and Why

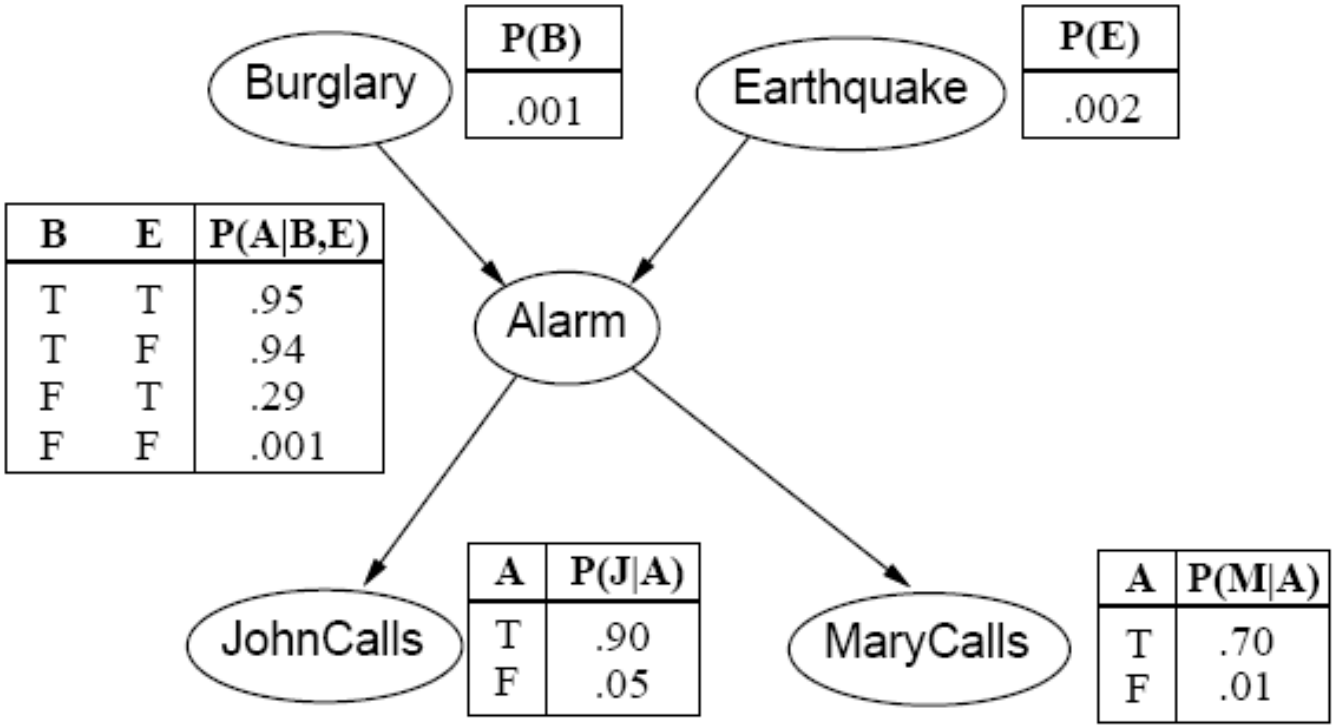
- Graphical models
  - Dependencies between (random) variables, avoid I.I.D. assumptions
  - Closer to reality, learning/inference is much more difficult
- Basic nomenclature
  - Node = Random Variable, Edge = Statistical Dependency
- Directed Graphs
  - A *directed* graph between random variables
  - Example: Bayesian networks, Hidden Markov Models
  - Joint distribution is a product of P(child|parents)



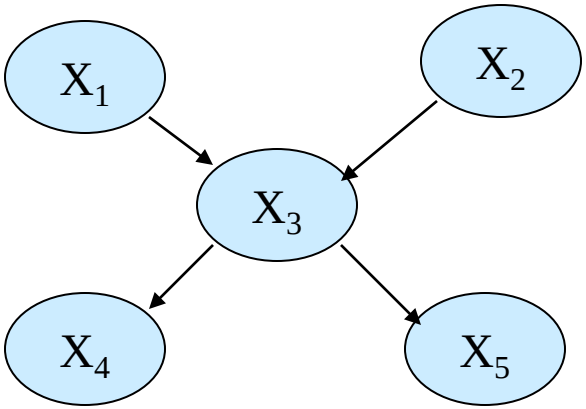
- Undirected Graphs
  - An *undirected* graph between random variables
  - Example: Markov/Conditional random fields
  - Joint distribution in terms of potential functions



Example: Burglary Network



Bayesian Networks

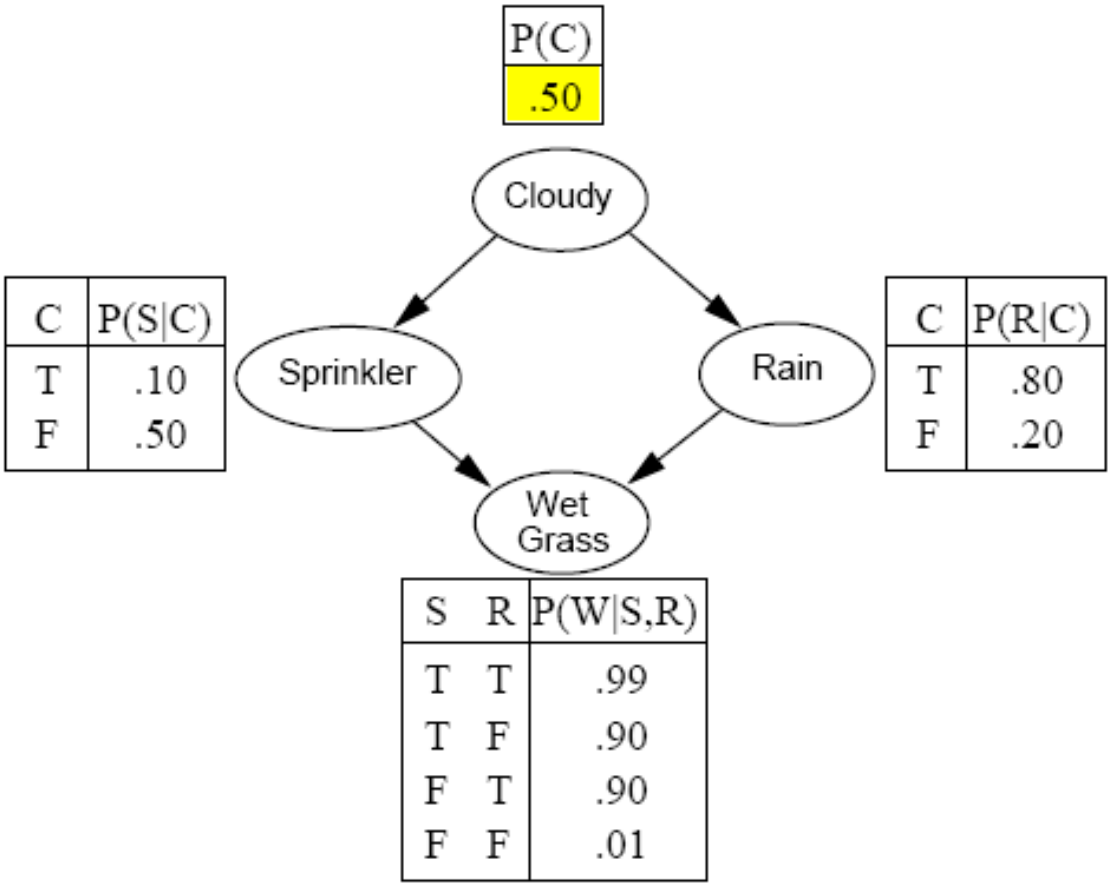


- Joint distribution in terms of *P(X|Parents(X))*

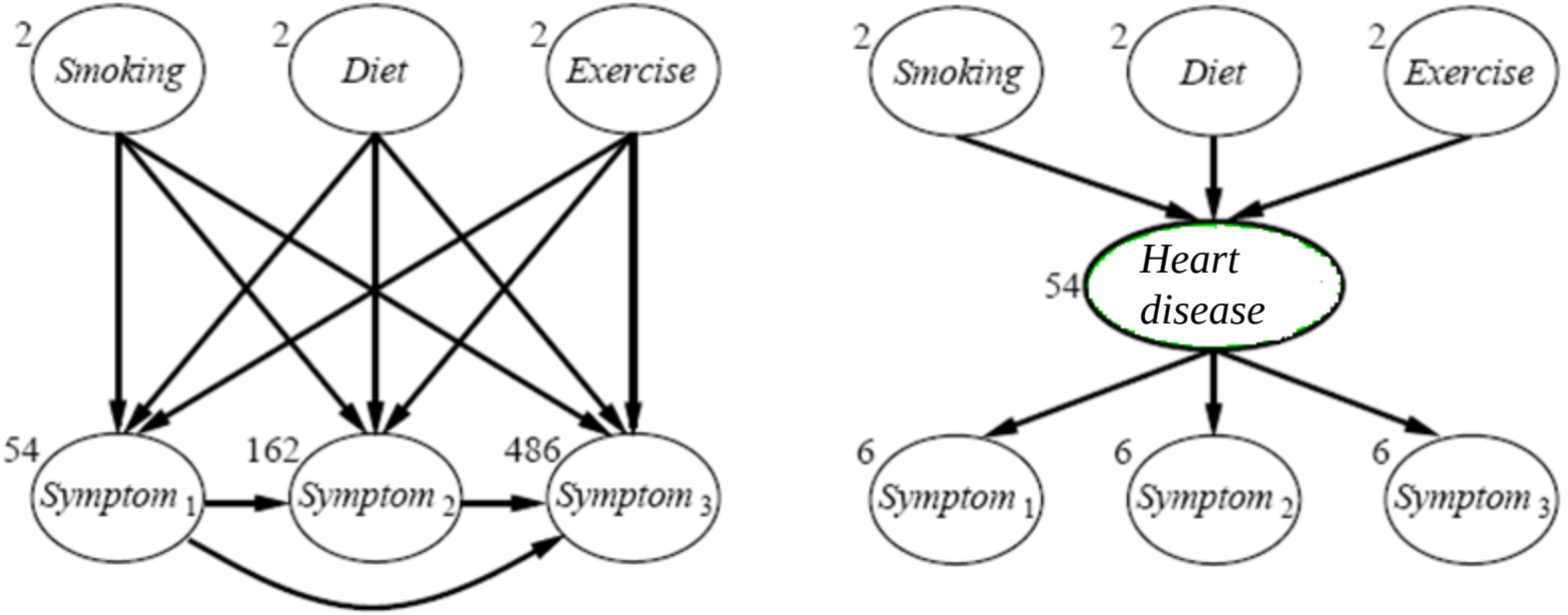
$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | X_1, \ldots, X_{i-1})$$

$$= \prod_{i=1}^{n} P(X_i | Parents(X_i))$$

Example II: Rain Network

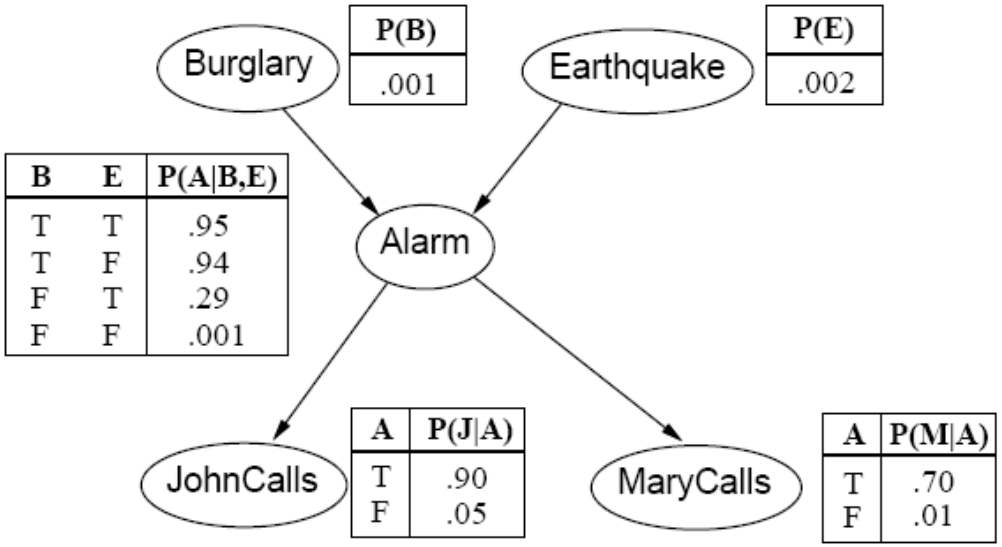


Latent Variable Models



- Bayesian network with hidden variables
  - Semantically more accurate, less parameters

Inference



- Some variables in the Bayes net are observed
  - the evidence/data, e.g., John has not called, Mary has called
- Inference
  - How to compute value/probability of other variables
  - Example: What is the probability of Burglary, i.e., $P(b|\neg j,m)$

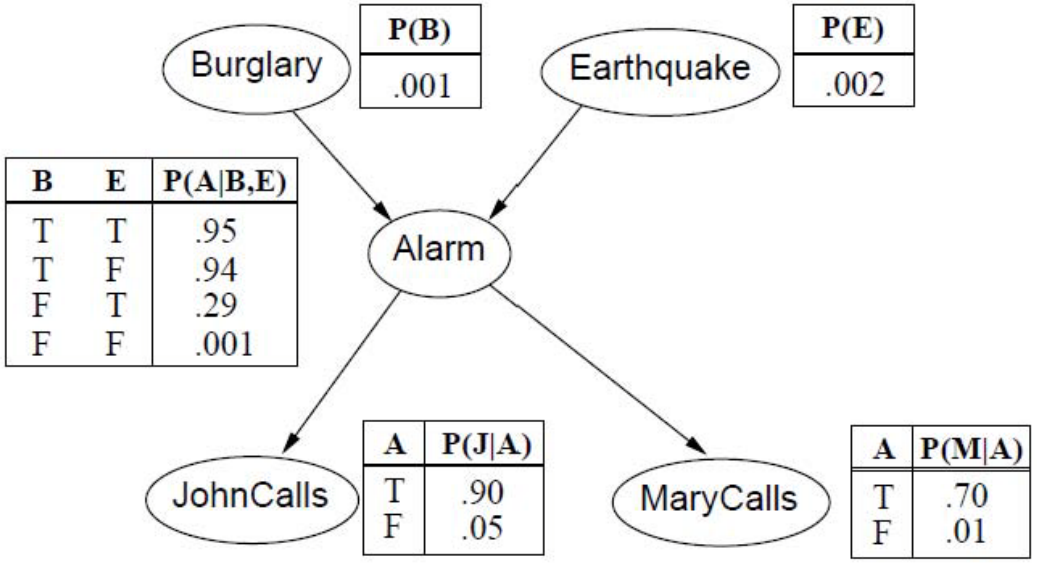
Inference Algorithms

- Graphs without loops: Tree-structured Graphs
  - Efficient exact inference algorithms are possible
  - Sum-product algorithm, and its special cases
    - Belief propagation in Bayes nets
    - Forward-Backward algorithm in Hidden Markov Models (HMMs)
- Graphs with loops
  - Junction tree algorithms
    - Convert into a graph without loops
    - May lead to exponentially large graph
  - Sum-product/message passing algorithm, ‘disregarding loops’
    - Active research topic, correct convergence ‘not guaranteed’
    - Works well in practice
  - Approximate inference

Approximate Inference

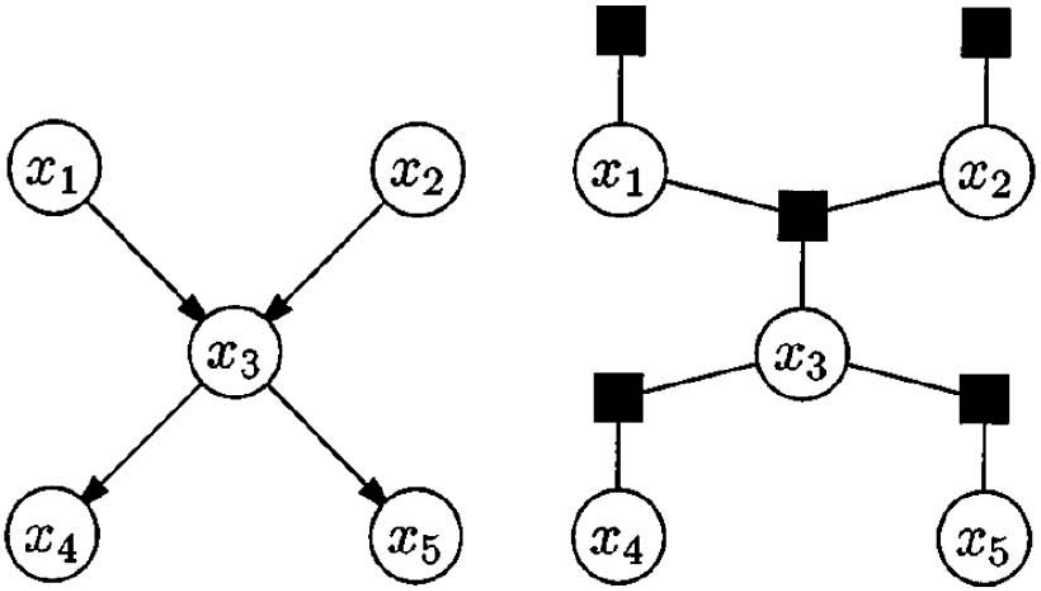
- Variational Inference
  - Deterministic approximation
  - Approximate complex true distribution over latent variables
  - Replace with family of simple/tractable distributions
    - Use the best approximation in the family
  - Examples: Mean-field, Bethe, Kikuchi, Expectation Propagation
- Stochastic Inference
  - Simple sampling approaches
  - Markov Chain Monte Carlo methods (MCMC)
    - Powerful family of methods
  - Gibbs sampling
    - Useful special case of MCMC methods

The Inference Problem



How can we compute $P(b|j, m)$?

Bayes Nets to Factor Graphs

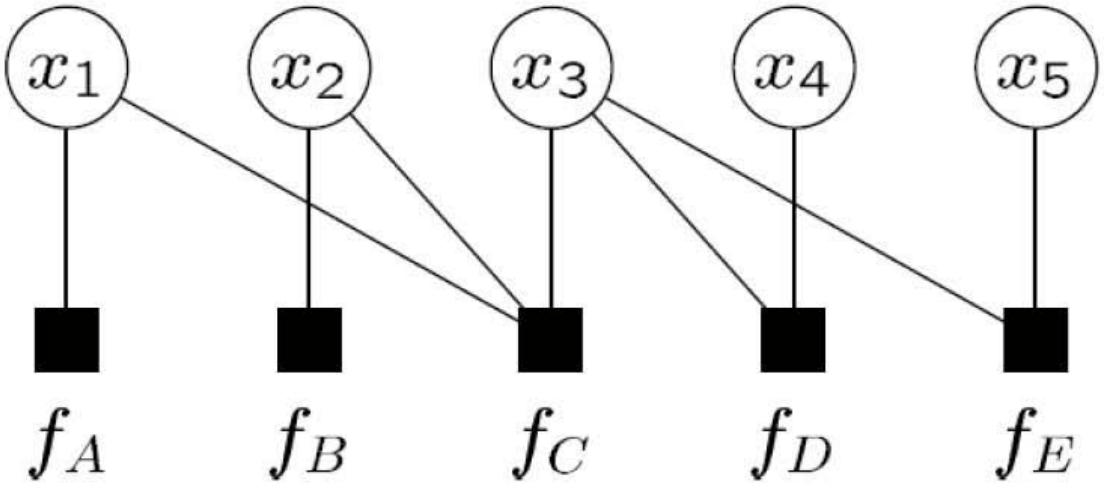


$$f_A(x_1) = p(x_1) \qquad f_B(x_2) = p(x_2) \qquad f_C(x_1, x_2, x_3) = p(x_3|x_1, x_2)$$

$$f_D(x_3, x_4) = p(x_4|x_3) \qquad f_E(x_3, x_5) = p(x_5|x_3)$$

Factor Graphs: Product of Local Functions

$$g(x_1, x_2, x_3, x_4, x_5) = f_A(x_1) f_B(x_2) f_C(x_1, x_2, x_3) f_D(x_3, x_4) f_E(x_3, x_5)$$



Marginalize Product of Functions (MPF)

- Marginalize (or maximize) product of functions

$$g(x_1, x_2, x_3, x_4, x_5) = f_A(x_1) f_B(x_2) f_C(x_1, x_2, x_3) f_D(x_3, x_4) f_E(x_3, x_5)$$

- For $g_1(x_1)$, we have

$$g_1(x_1) = f_A(x_1) \sum_{\sim x_1} \left(\cdots \left(\sum_{\sim x_3} f_E(x_3, x_5) \right) \right)$$

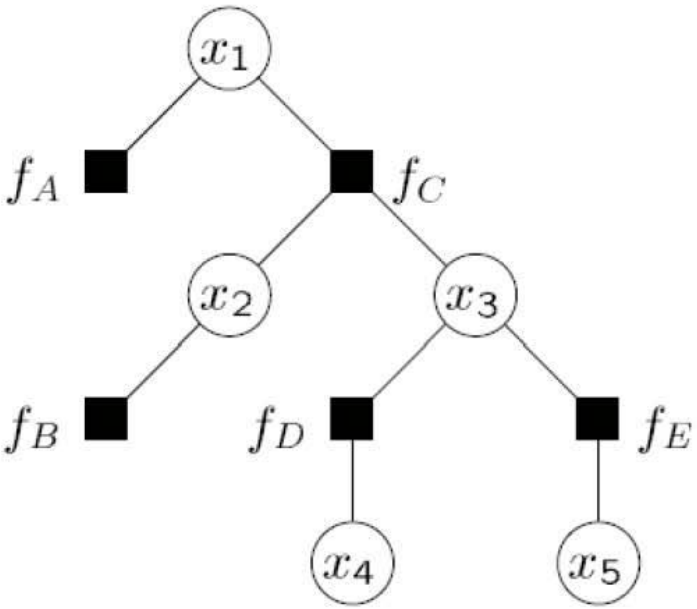
Distributive Law: ab + ac = a(b+c)

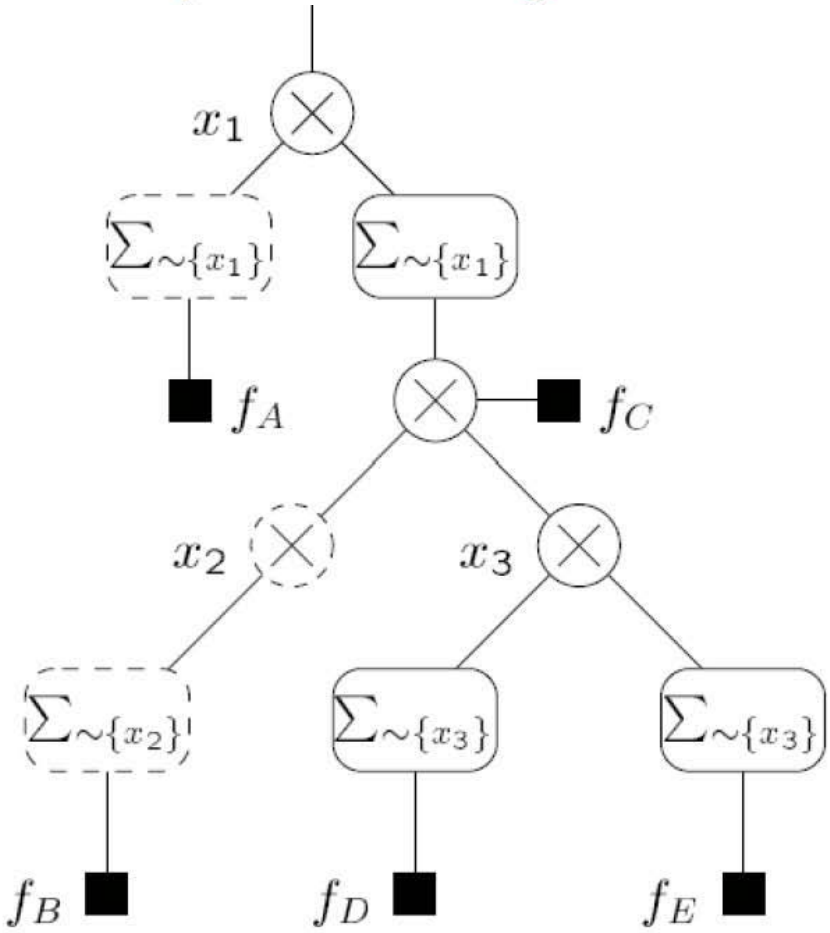
- For $g_3(x_3)$, we have

$$g_3(x_3) = \left(\sum_{\sim x_3} f_A(x_1) f_B(x_2) f_C(x_1, x_2, x_3) \right) \left(\sum_{\sim x_3} f_D(x_3, x_4) \right) \left(\sum_{\sim x_3} f_E(x_3, x_5) \right)$$

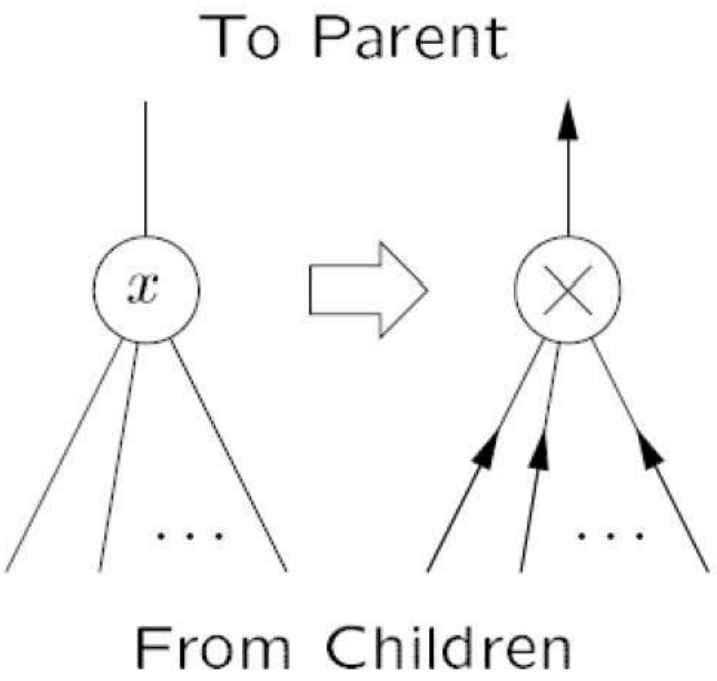
Example: Computing $g_1(x_1)$

$$g_1(x_1) = f_A(x_1) \sum_{\sim x_1} \left(f_B(x_2) f_C(x_1, x_2, x_3) \left(\sum_{\sim x_3} f_D(x_3, x_4) \right) \left(\sum_{\sim x_3} f_E(x_3, x_5) \right) \right)$$





Message Passing



Compute product of descendants

The Sum-Product Algorithm

- To compute $g_i(x_i)$, form a tree rooted at x_i
- Starting from the leaves, apply the following two rules
  - Product Rule:

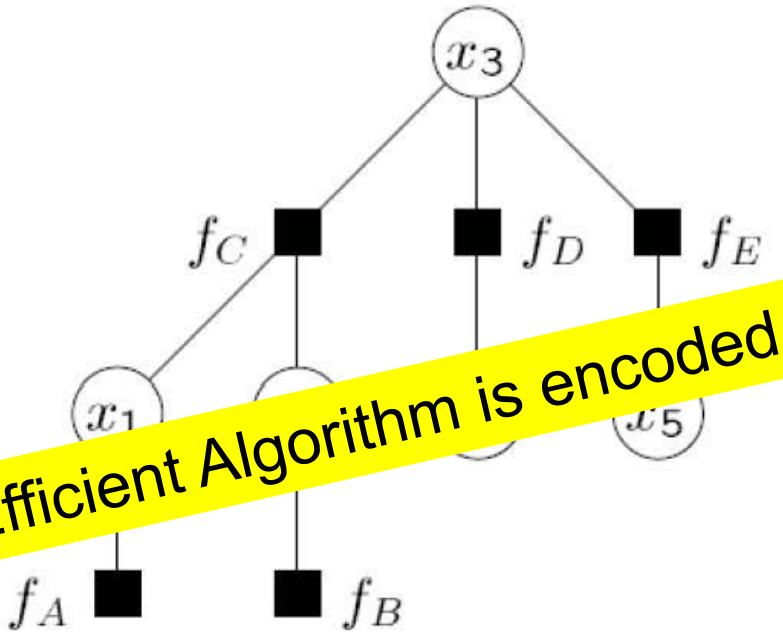
    At a variable node, take the product of descendants
  - Sum-product Rule:

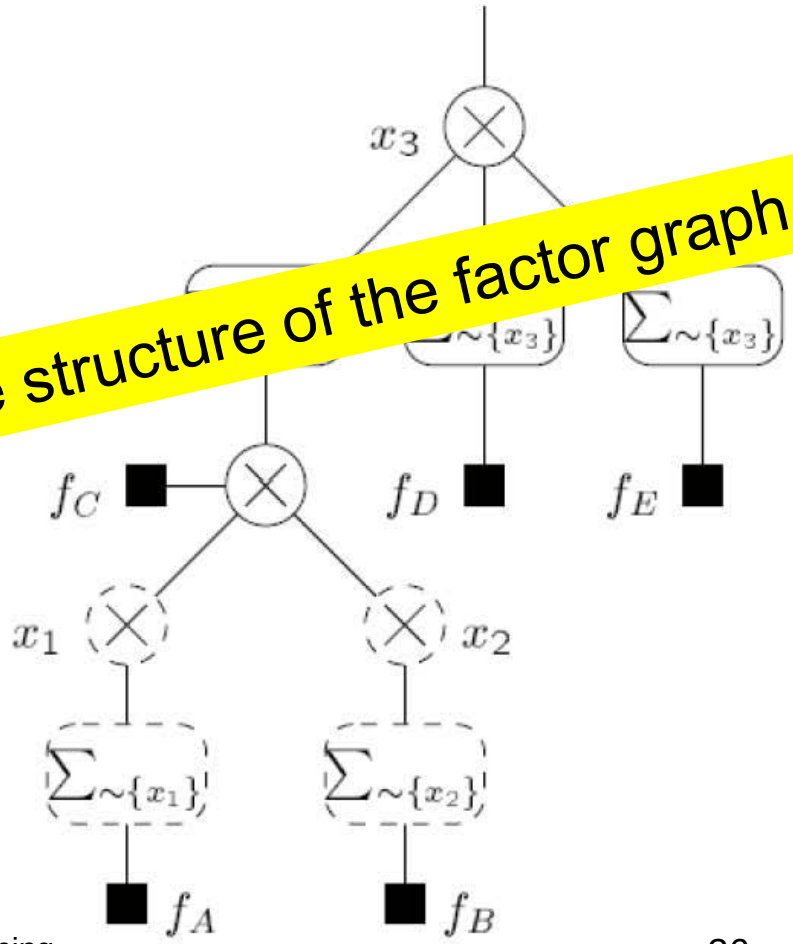
    At a factor node, take the product of f with descendants;

    then perform not-sum over the parent node
- To compute all marginals
  - Can be done one at a time; repeated computations, not efficient
  - Simultaneous message passing following the sum-product algorithm
  - Examples: Belief Propagation, Forward-Backward algorithm, etc.

Example: Computing $g_3(x_3)$

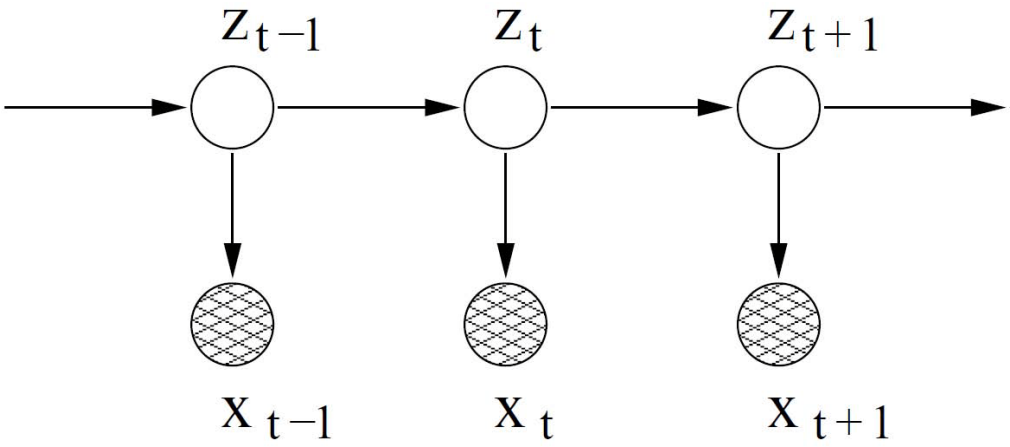
$$g_3(x_3) = \left(\sum_{\sim x_3} f_A(x_1) f_B(x_2) f_C(x_1, x_2, x_3)\right)\left(\sum_{\sim x_3} f_D(x_3, x_4)\right)\left(\sum_{\sim x_3} f_E(x_3, x_5)\right)$$





Efficient Algorithm is encoded in the structure of the factor graph

Hidden Markov Models (HMMs)

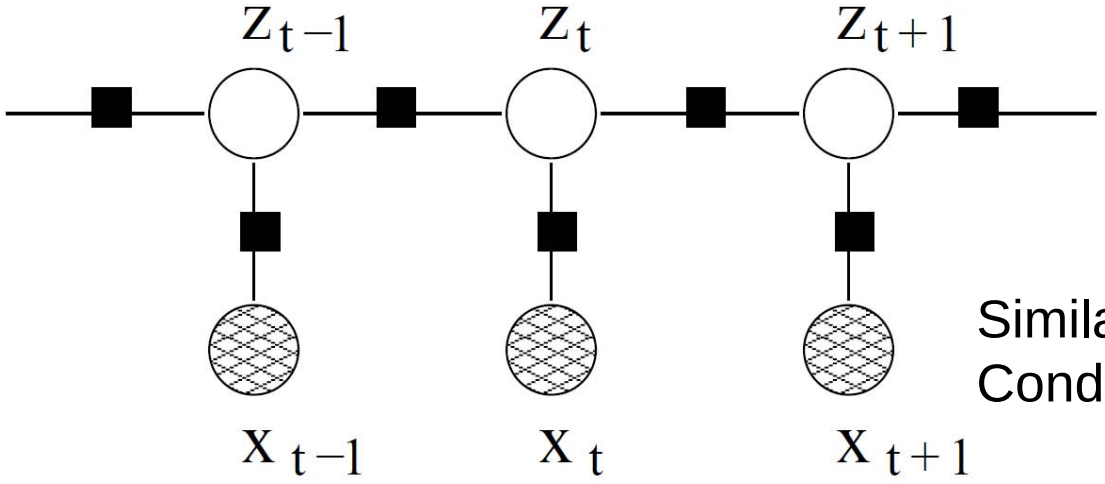


Latent variables:

$z_0, z_1, \ldots, z_{t-1}, z_t, z_{t+1}, \ldots, z_T$

Observed variables:

$x_1, \ldots, x_{t-1}, x_t, x_{t+1}, \ldots, x_T$



Inference Problems:

1. Compute $p(x_{1:T})$
2. Compute $p(z_t|x_{1:T})$
3. Find $\max_{z_{1:T}} p(z_{1:T}|x_{1:T})$

Similar problems for chain-structured Conditional Random Fields (CRFs)

Distributive Law on Semi-Rings

- Idea can be applied to any commutative semi-ring
- Semi-ring 101
  - Two operations (+,×): Associative, Commutative, Identity
  - Distributive law: a×b + a×c = a×(b+c)

	K	"$(+,0)$"	"$(\cdot,1)$"	short name
1.	A	$(+,0)$	$(\cdot,1)$	
2.	$A[x]$	$(+,0)$	$(\cdot,1)$	
3.	$A[x,y,\ldots]$	$(+,0)$	$(\cdot,1)$	
4.	$[0,\infty)$	$(+,0)$	$(\cdot,1)$	sum-product
5.	$(0,\infty]$	$(\min,\infty)$	$(\cdot,1)$	min-product
6.	$[0,\infty)$	$(\max,0)$	$(\cdot,1)$	max-product
7.	$(-\infty,\infty]$	$(\min,\infty)$	$(+,0)$	min-sum
8.	$[-\infty,\infty)$	$(\max,-\infty)$	$(+,0)$	max-sum
9.	$\{0,1\}$	$(\mathtt{OR},0)$	$(\mathtt{AND},1)$	Boolean
10.	2^S	$(\cup,\emptyset)$	$(\cap,S)$	
11.	Λ	$(\vee,0)$	$(\wedge,1)$	
12.	Λ	$(\wedge,1)$	$(\vee,0)$.	

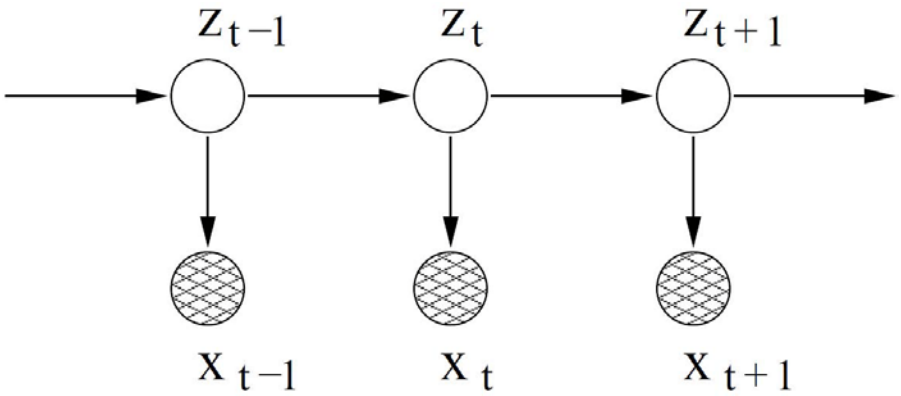
- Belief Propagation in Bayes nets
- MAP inference in HMMs
- Max-product algorithm
- Alternative to Viterbi Decoding
- Kalman Filtering
- Error Correcting Codes
- Turbo Codes
- …

Message Passing in General Graphs

- Tree structured graphs
  - Message passing is guaranteed to give correct solutions
  - Examples: HMMs, Kalman Filters
- General Graphs
  - Active research topic
    - Progress has been made over the past 10 years
  - Loopy Message passing
    - May not converge
    - May converge to a ‘local minima’ of ‘Bethe variational free energy’
  - New approaches to convergent and correct message passing
- Applications
  - True Skill: Ranking System for Xbox Live
  - Turbo Codes: 3G, 4G phones, satellite comm, Wimax, Mars orbiter

Temporal Models

- State Space Models



State Space, State Transitions

Hidden Markov Models
Kalman Filters
Linear Dynamical Systems

- Time Series Analysis

Time Domain Models

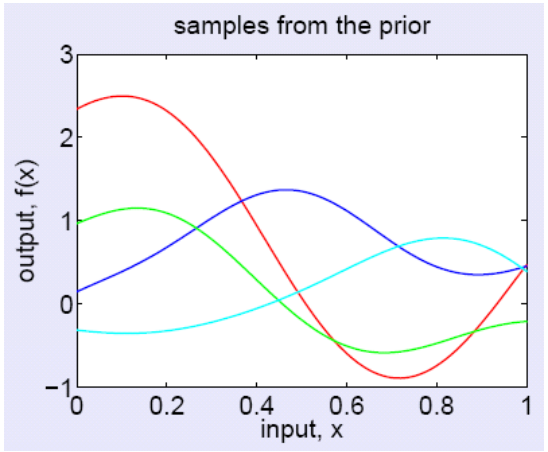
AR/MA/ARMA models
ACF, and Partial ACF
ARIMA, and Seasonal ARIMA
Estimation, Forecasting

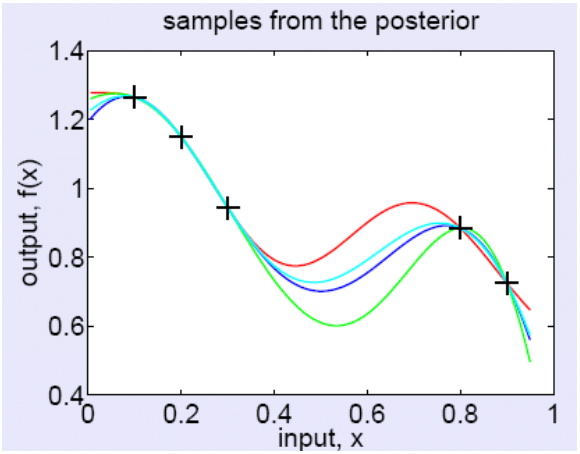
Spectral Analysis

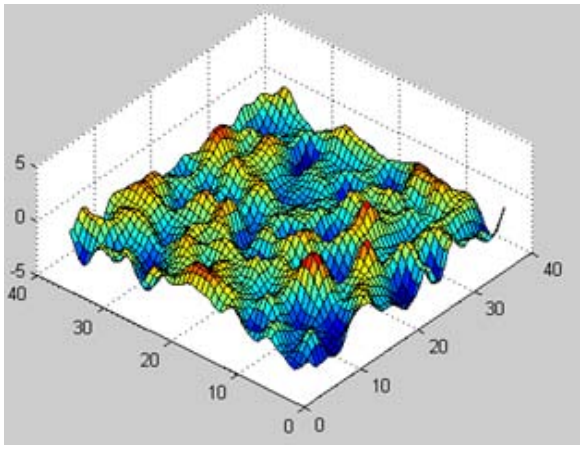
Spectral Density
Fourier Transforms
Wavelet Transforms
Filtering

Spatial Models

Gaussian Process







From: H. Wallach Introduction to Gaussian process regression, 2005

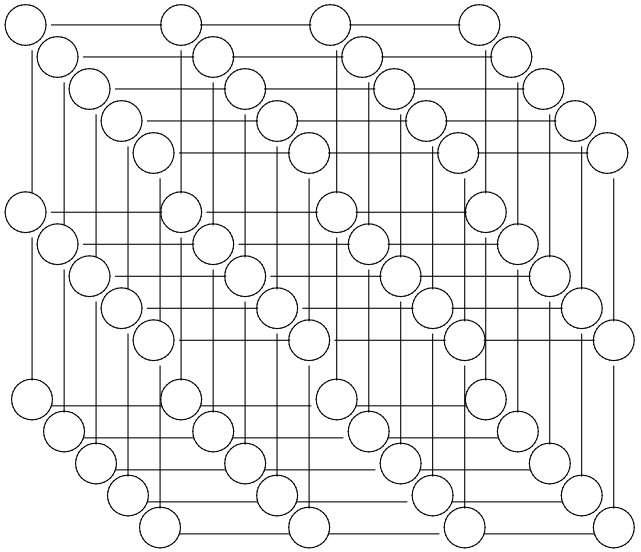
- Models for Continuous Data:
  - Gaussian Processes: Prior over functions GP(0,K)
  - Kriging, Co-Kriging, Linear Models of Coregionalization (LMCs)
- Models for Discrete Data:
  - Markov Random Fields (MRFs)
- Structure Learning:
  - Finding statistical dependencies

Drought Detection

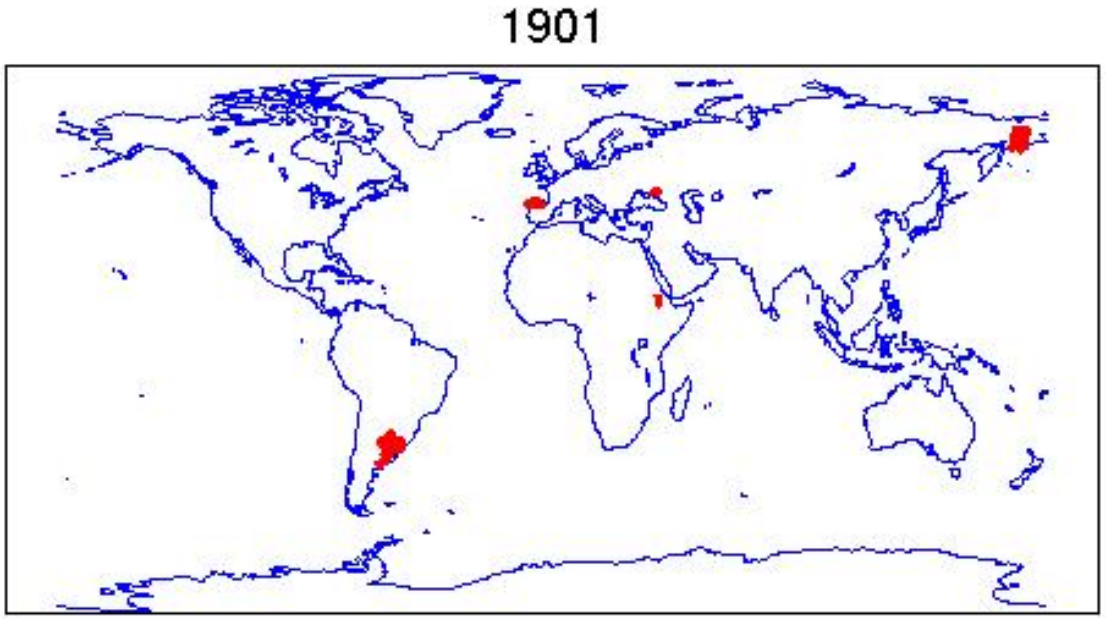
- Significant droughts
  - Persistent over space and time
  - Catastrophic consequences
- Examples:
  - Late 1960s Sahel drought
  - 1930s North American Dust Bowl
- Dataset: Climate Research Unit
  - http://data.giss.nasa.gov/precip_cru/
  - Monthly precipitation over land
  - Duration: 1901 to 2006
  - Resolution: 0.5 latitude × 0.5 longitude

Detection with Markov Random Field (MRF)

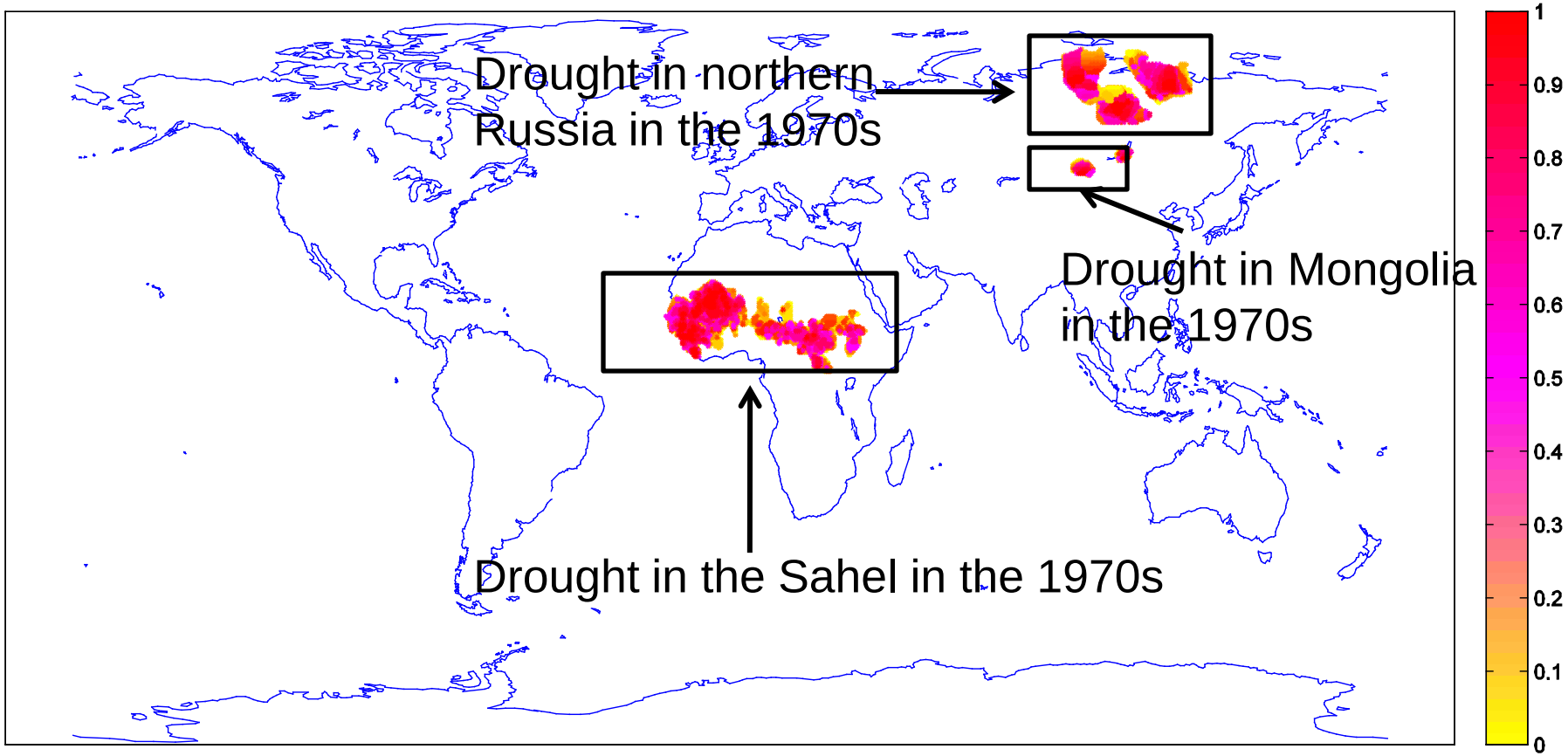
- Model dependencies using a 4-nearest neighbor grid
  - Replicate grid over time
  - Each node can be 0 (normal) or 1 (dry)
- Toy example:
  - m = 3, n = 4, N = 12, T = 5
  - Total # States: $2^{60} = 1.1529 \times 10^{18}$
- CRU data:
  - m = 720, n = 360, N = 67,420, T = 106
  - Total # States: $2^{7,146,520} > 10^{2,382,200}$
- MAP Inference:
  - Find the most likely "state" of the system
  - Integer programming with LARGE number of states
  - Postprocess MAP state to identify significant droughts



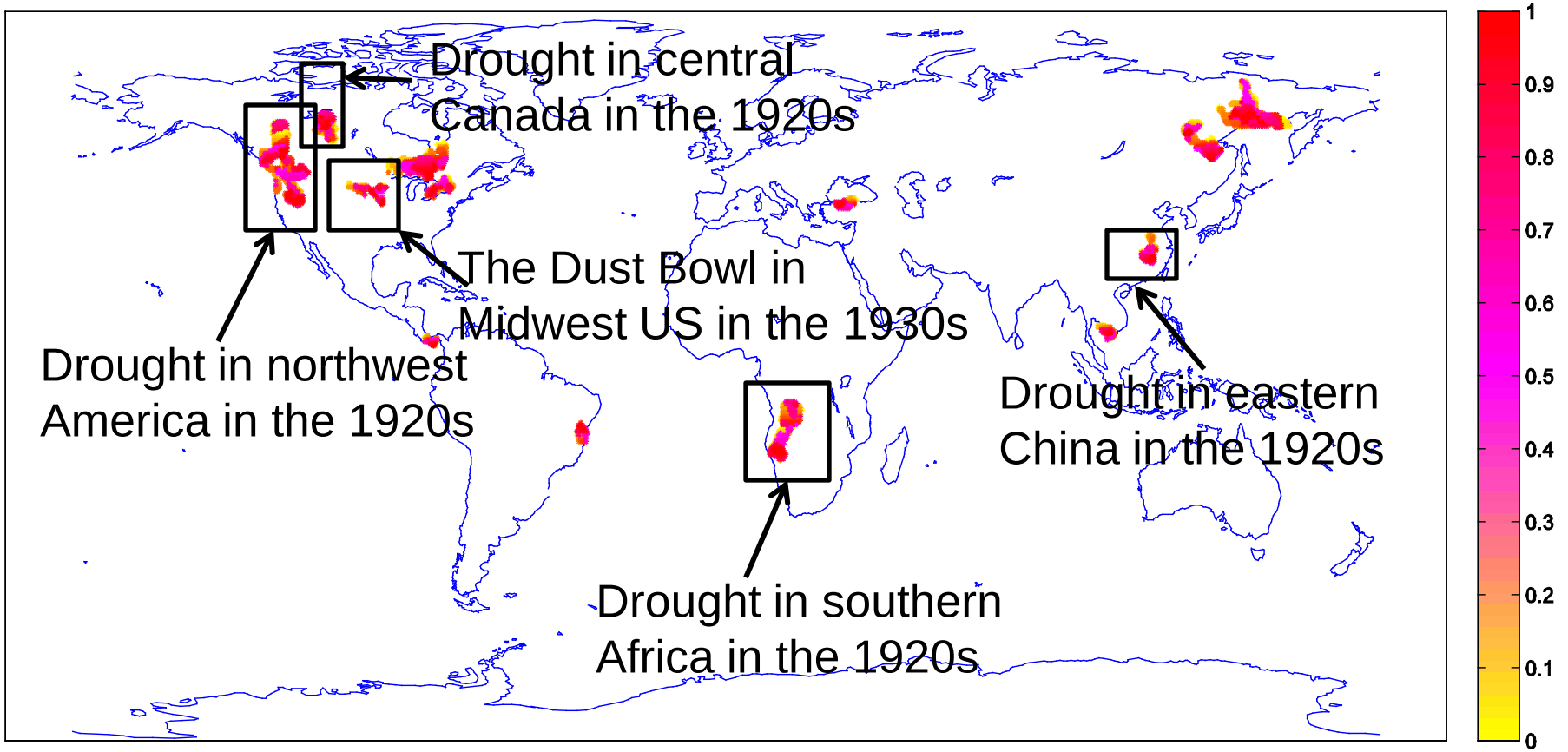
Drought Regions from MRFs



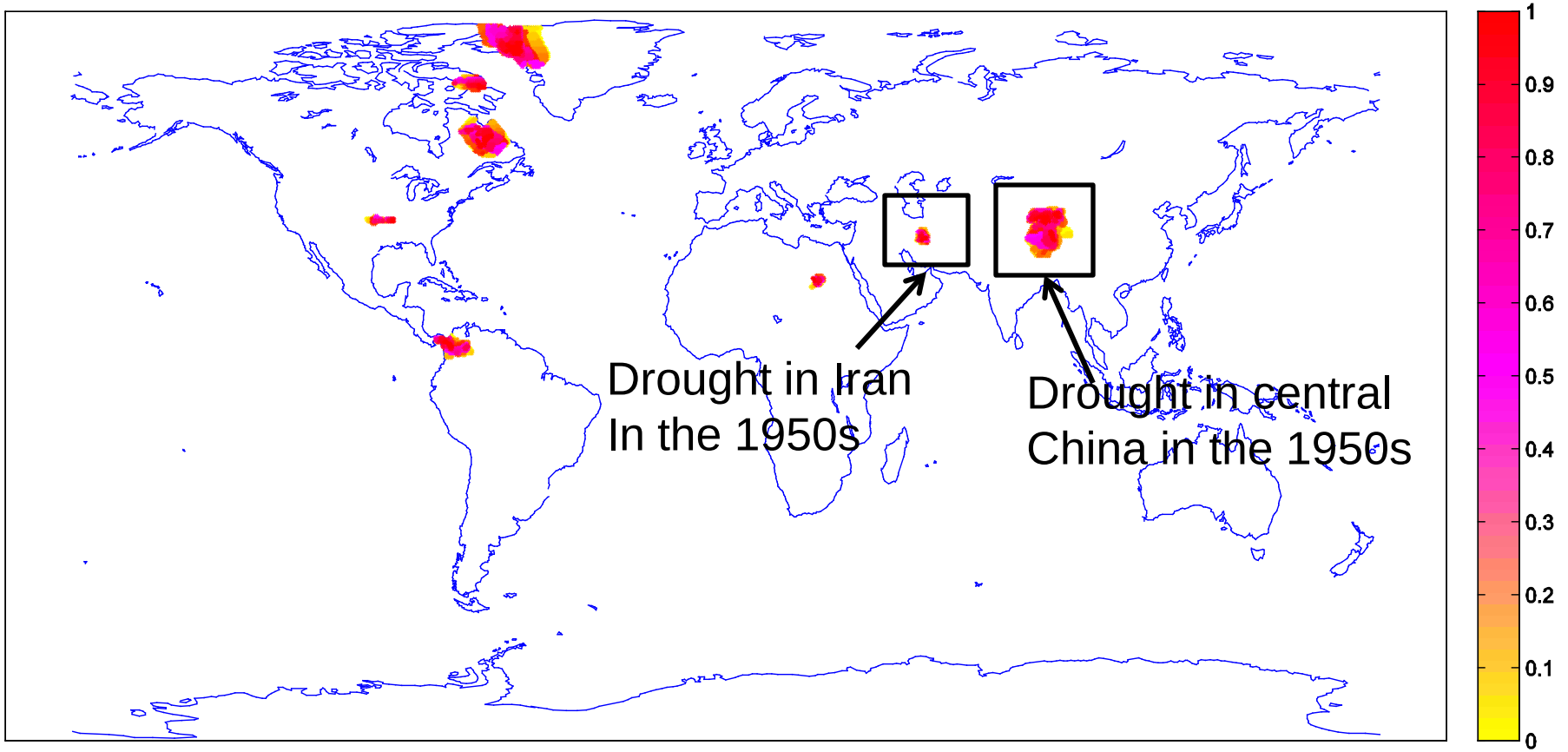
Results: Drought Detection



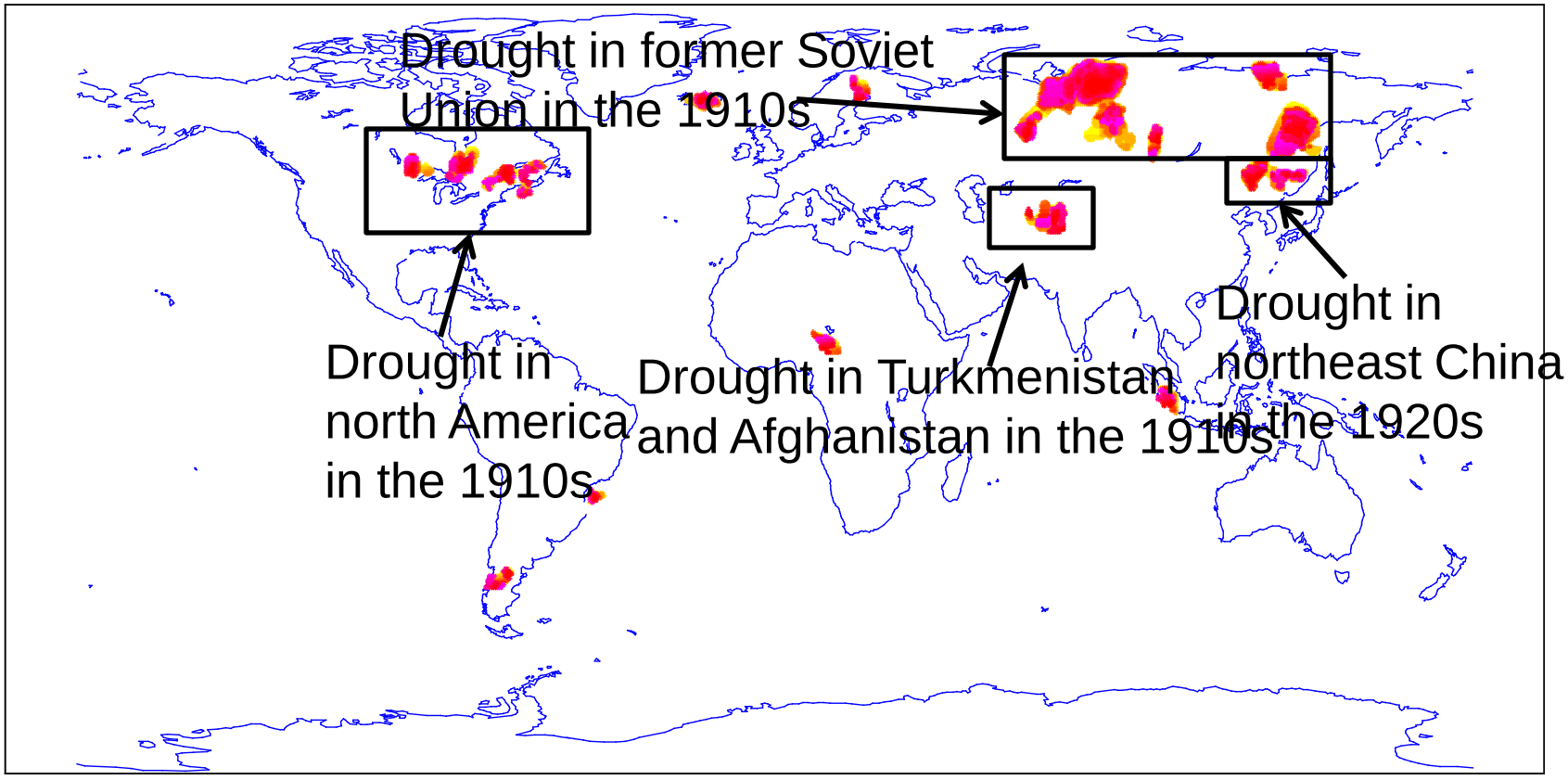
Results: Drought Detection



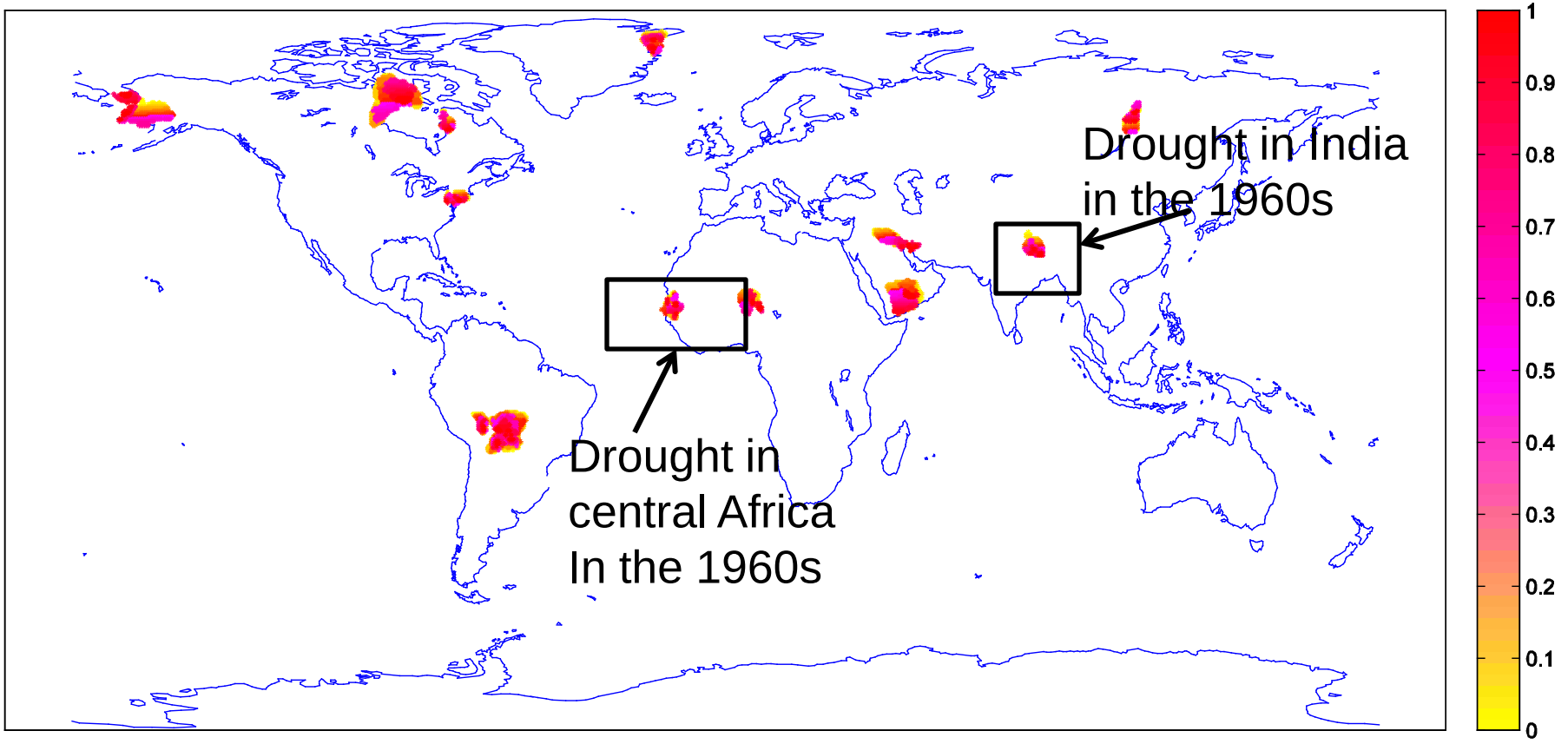
Results: Drought Detection



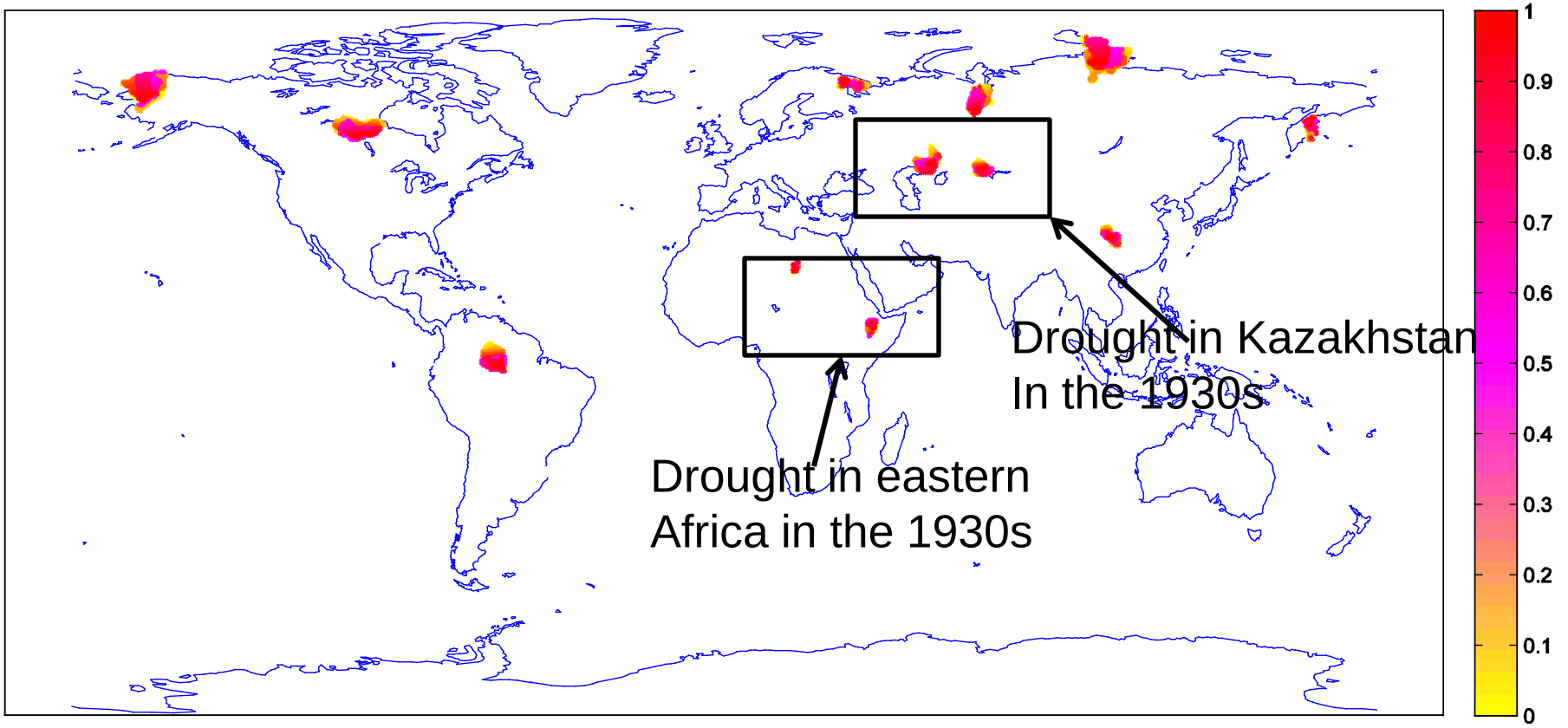
Results: Drought Detection



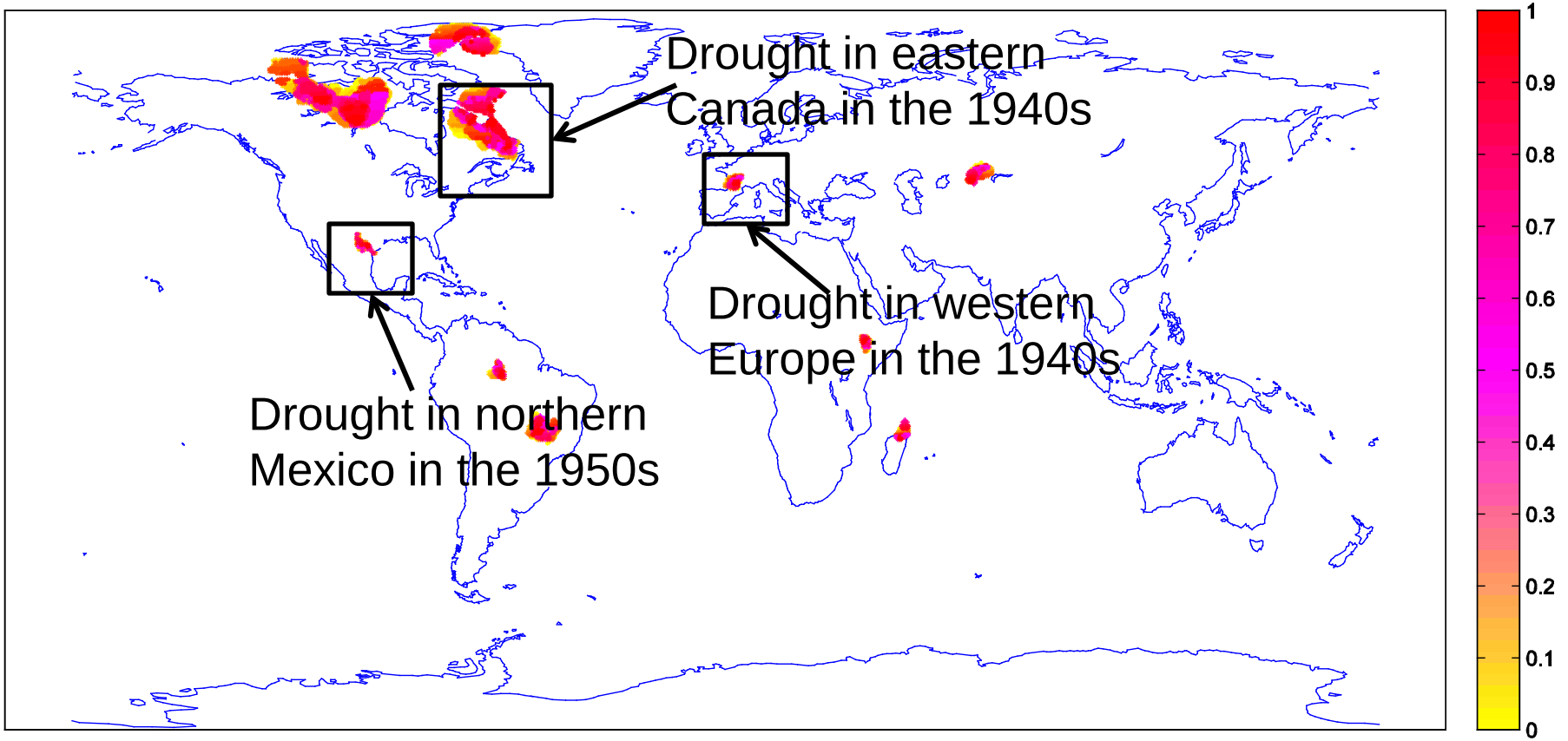
Results: Drought Detection



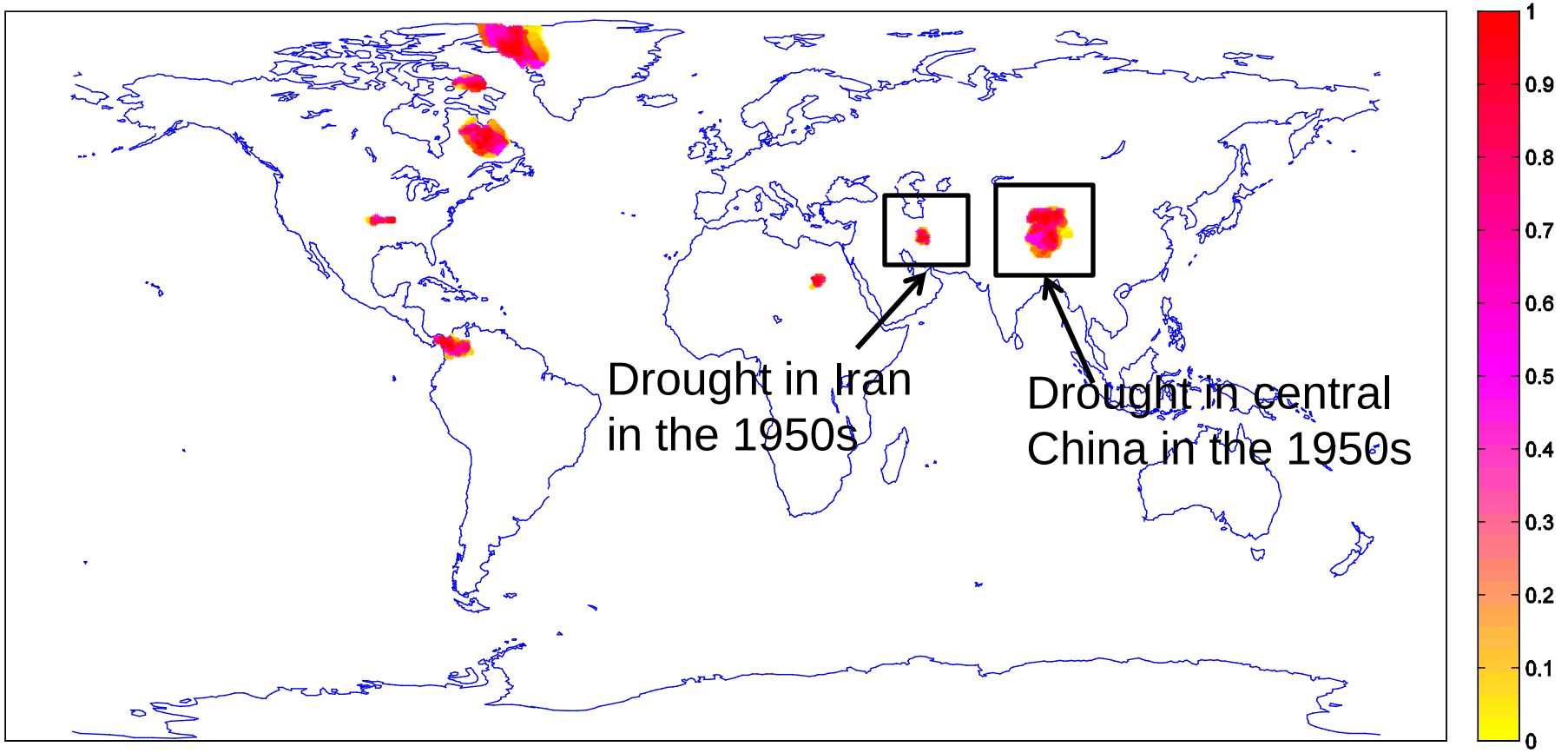
Results: Drought Detection



Results: Drought Detection



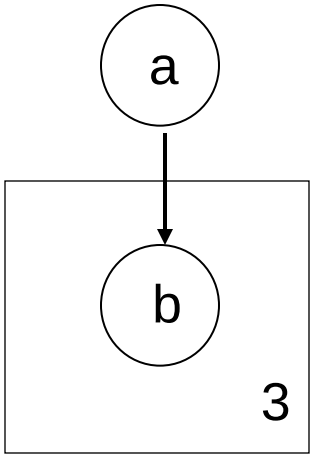
Results: Drought Detection

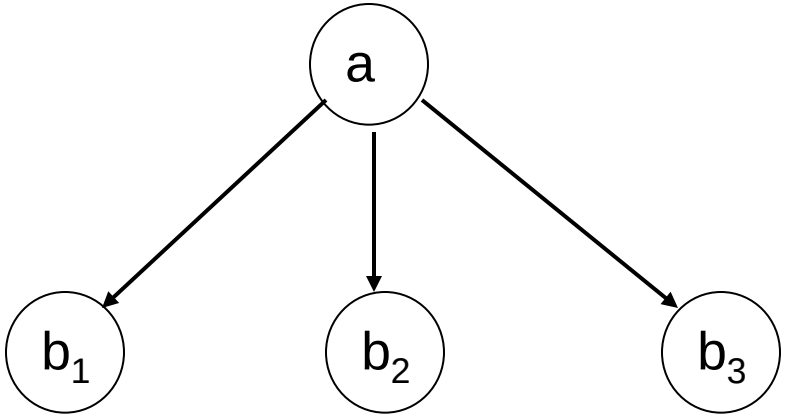


Graphical Models (Contd.)

- Basics, Inference
- Markov Random Fields
  - Example: Draught Detection
- Mixed Membership Models
  - Example: Text Analysis and Topic Modeling
- Probabilistic Matrix Factorization
  - Example: Matrix Completion and Recommendation Systems

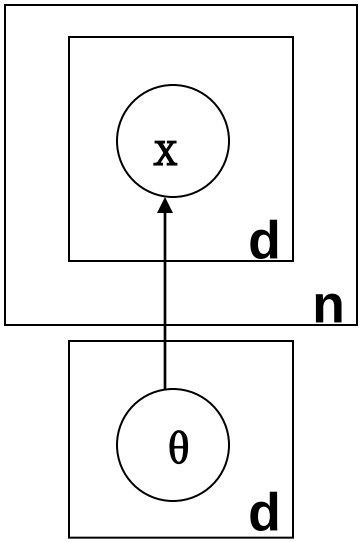
Background: Plate Diagrams

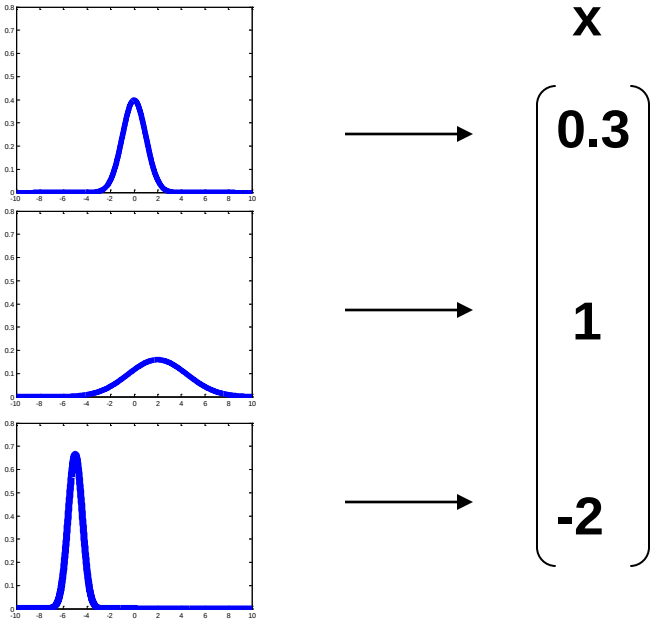




Compact representation of large Bayesian networks

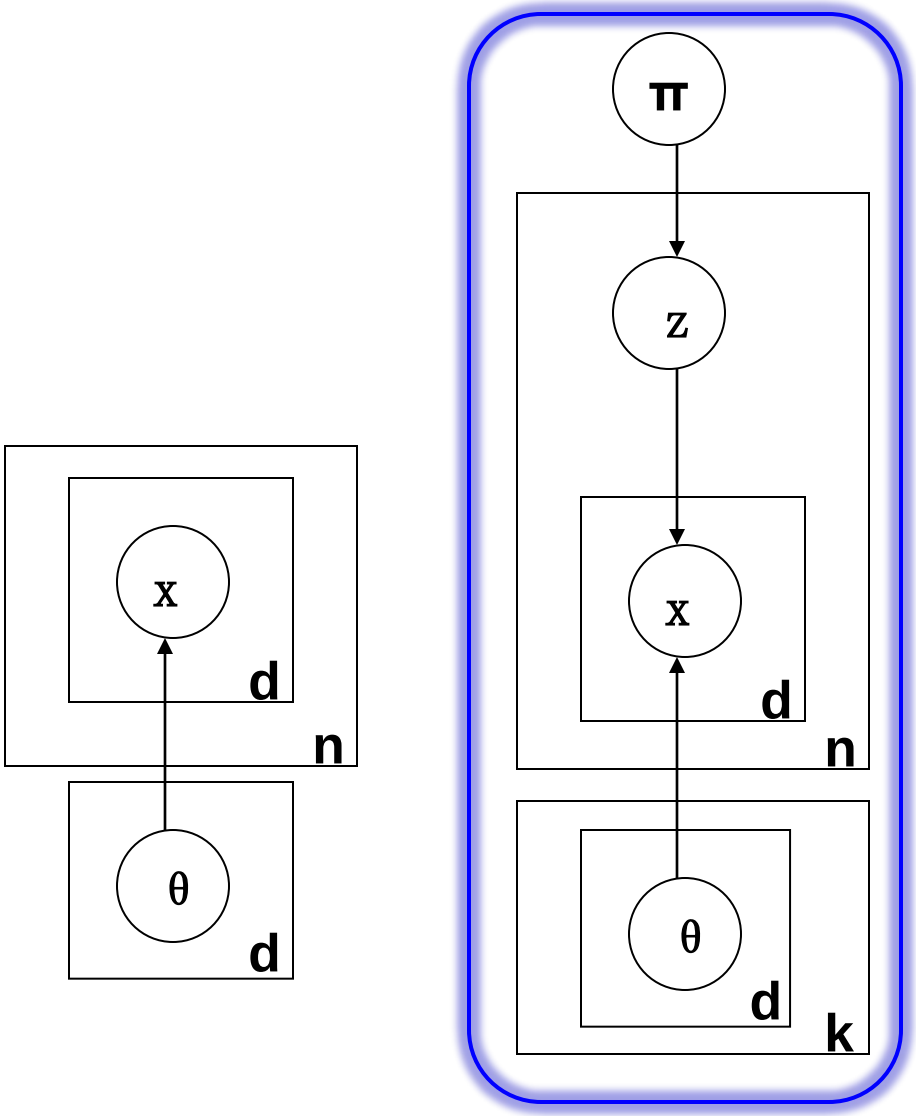
Model 1: Independent Features



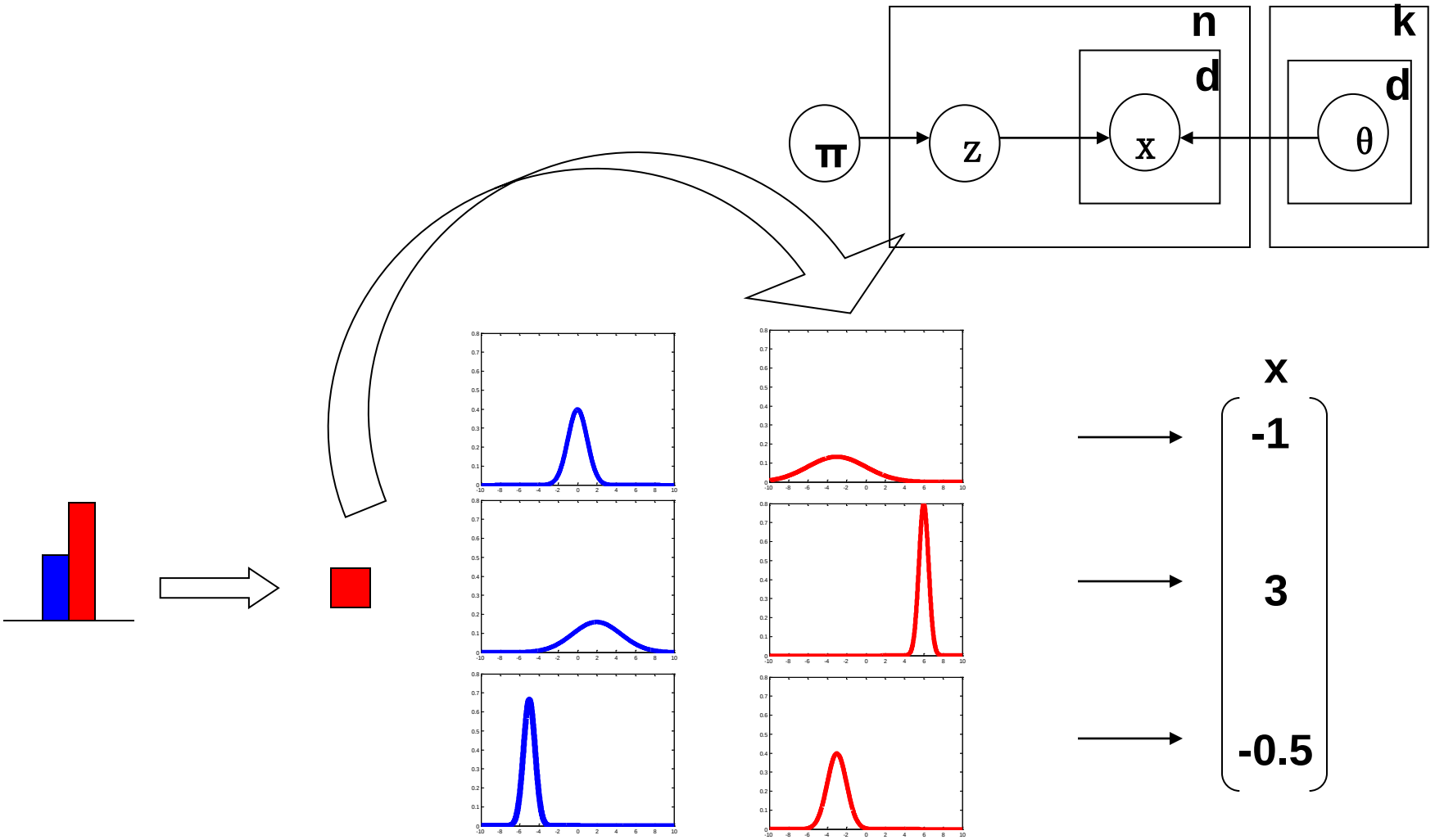


d=3, n=1

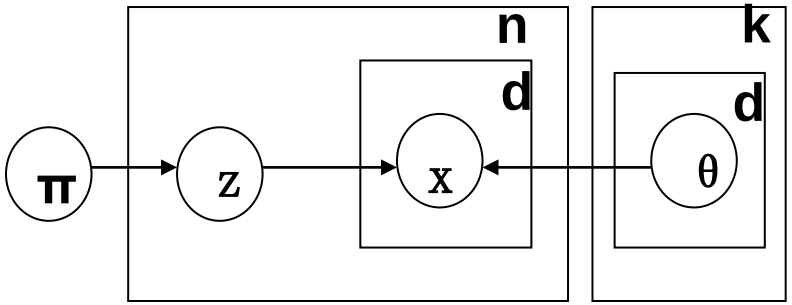
Model 2: Naïve Bayes (Mixture Models)

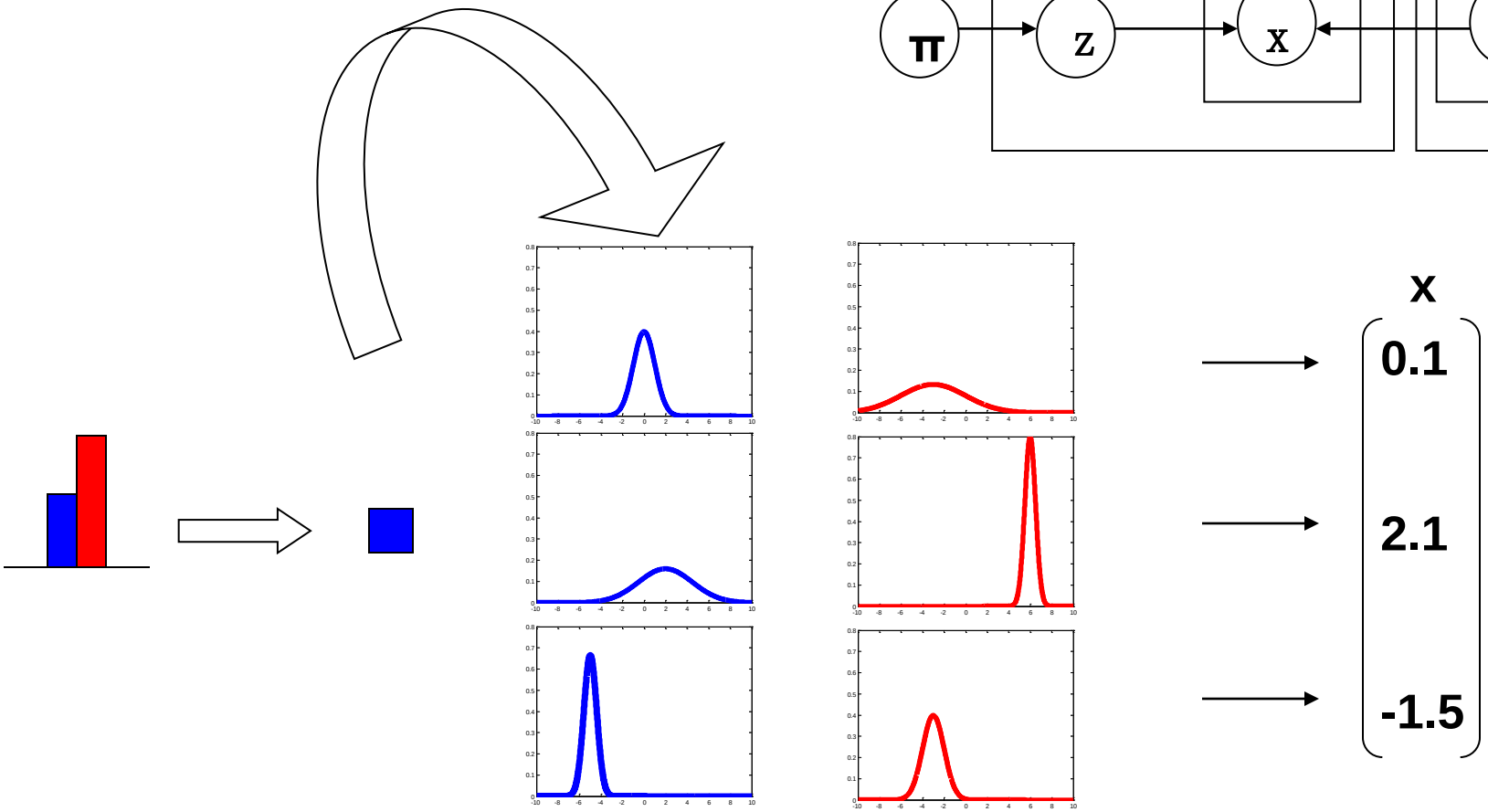


Naïve Bayes Model

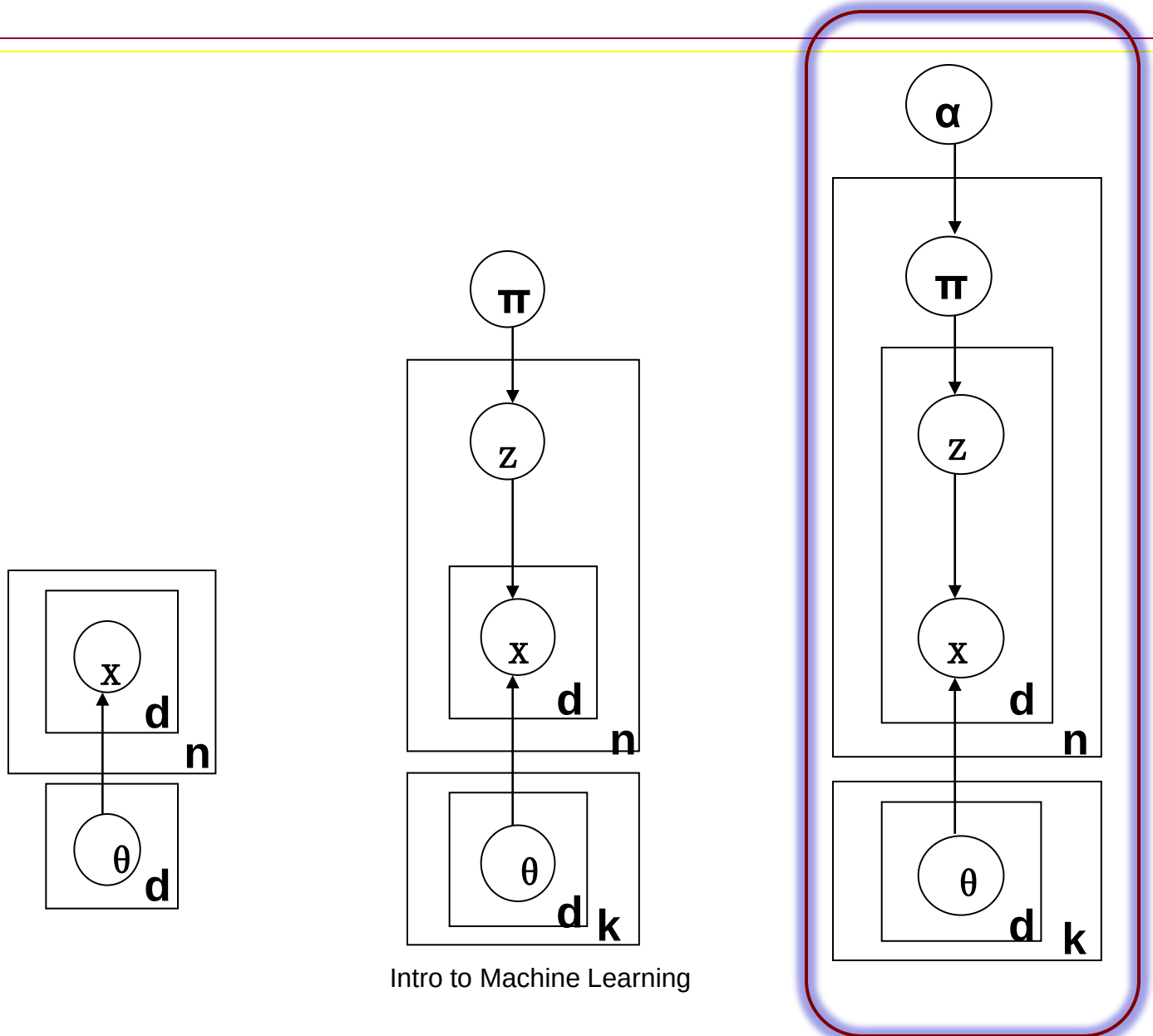


Naïve Bayes Model

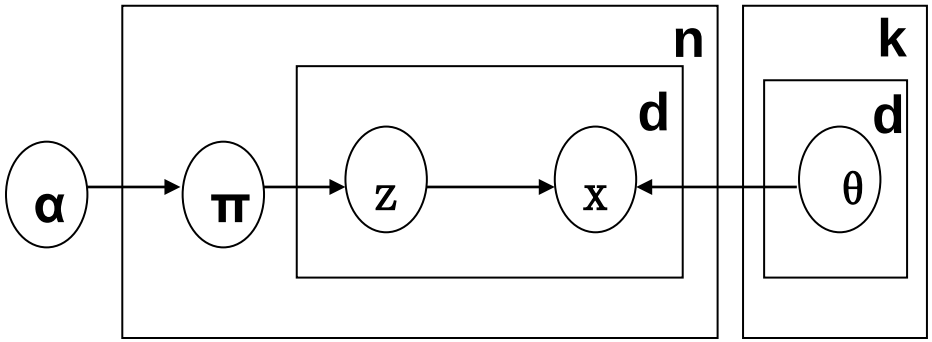


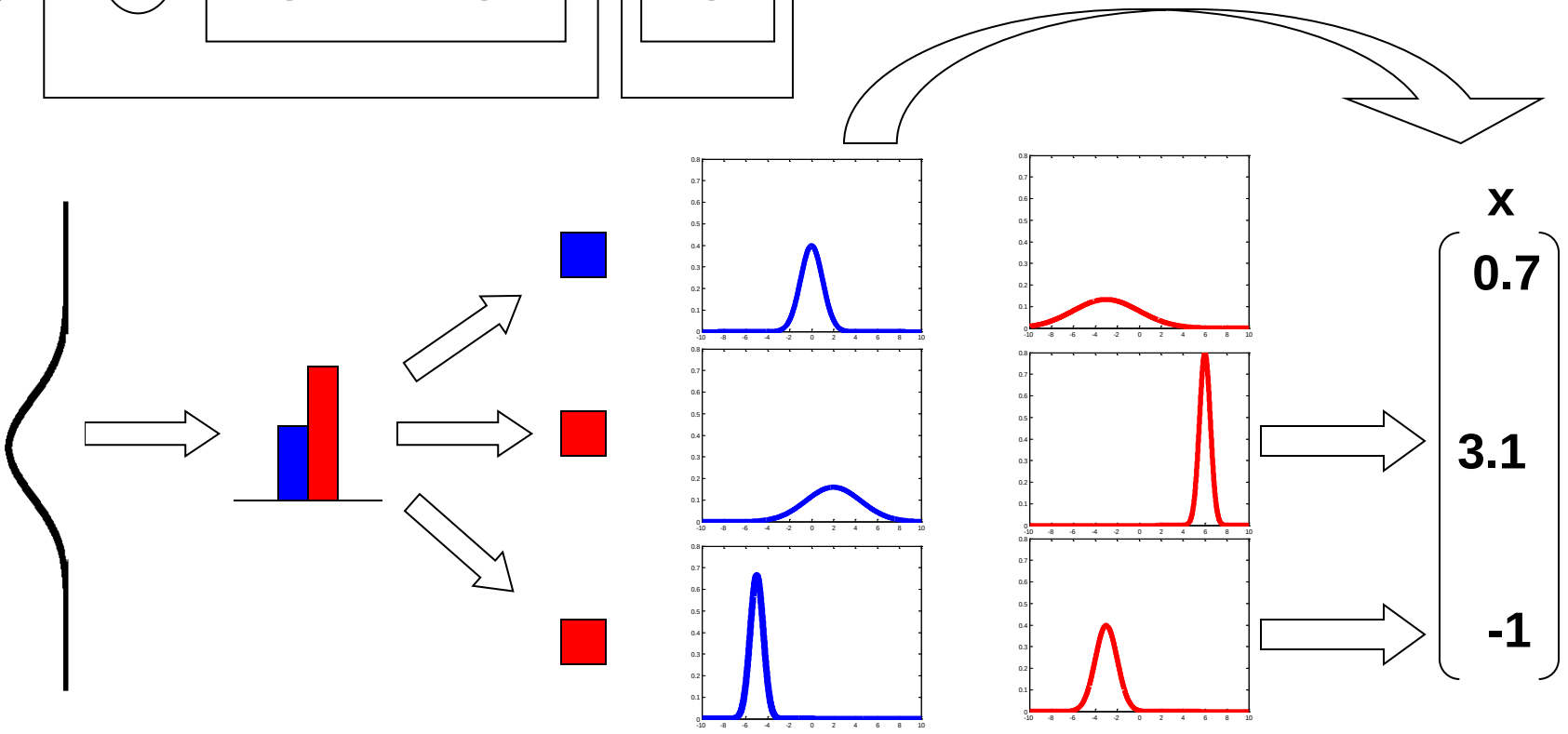


Model 3: Mixed Membership Model

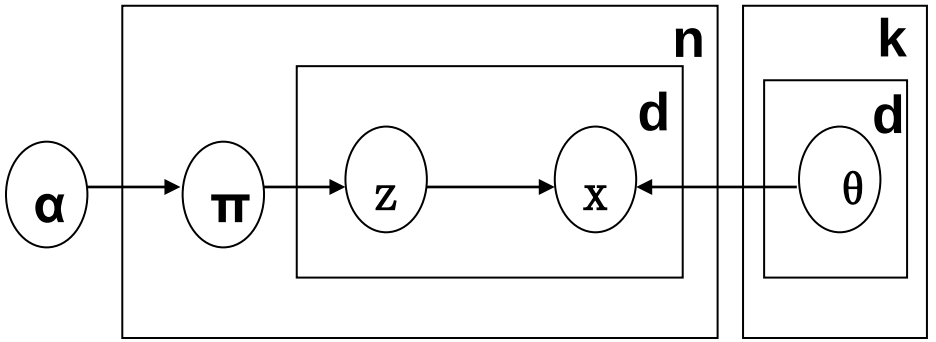


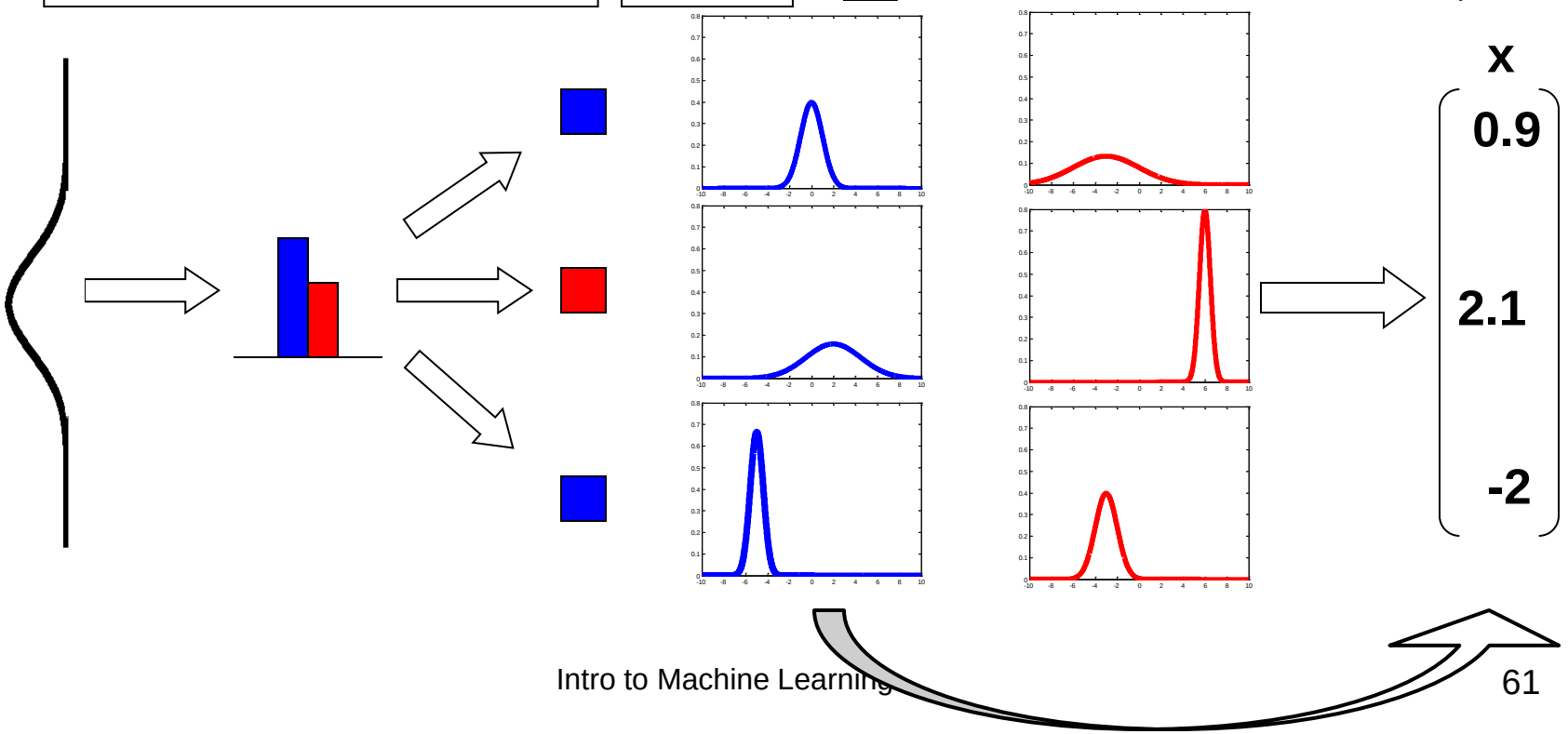
Mixed Membership Models



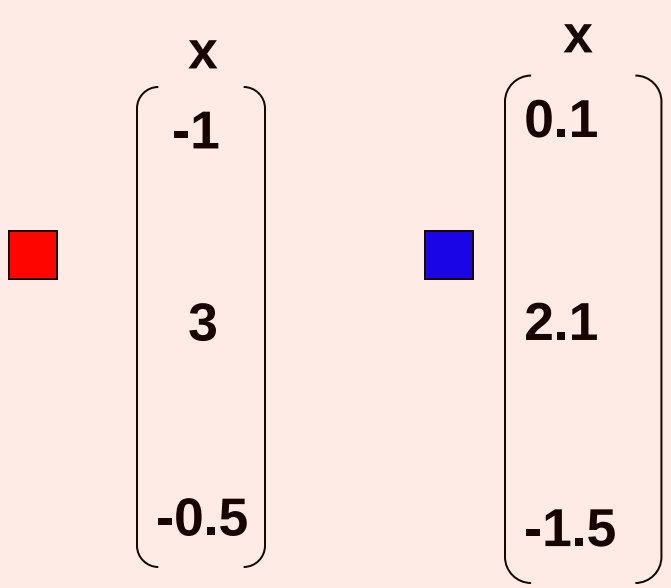


Mixed Membership Models

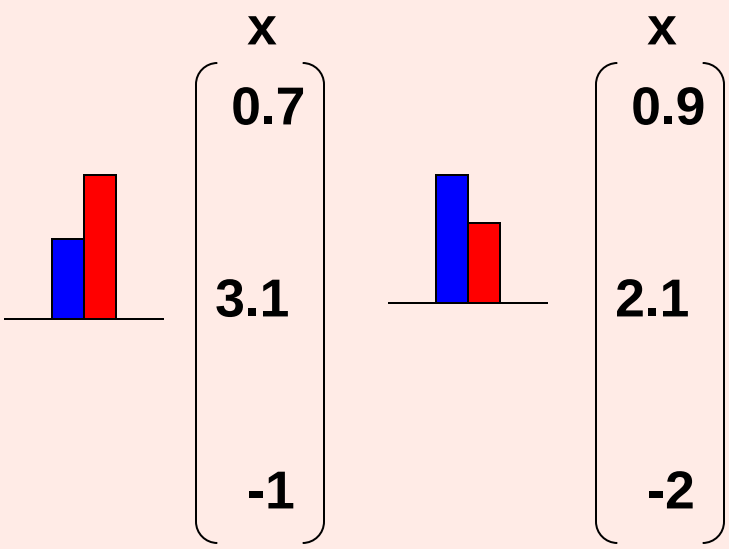




Mixture Model vs Mixed Membership Model

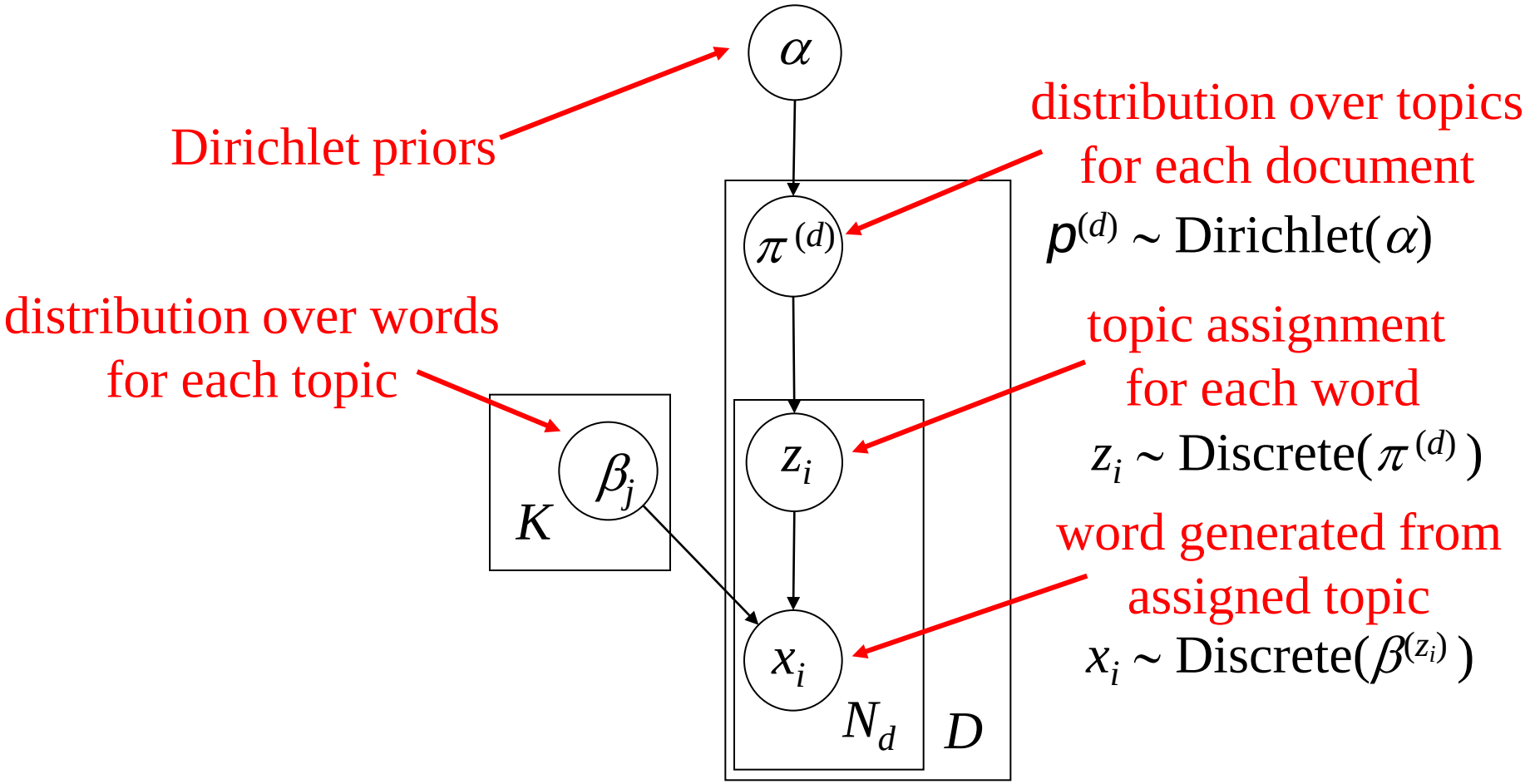


Single component membership

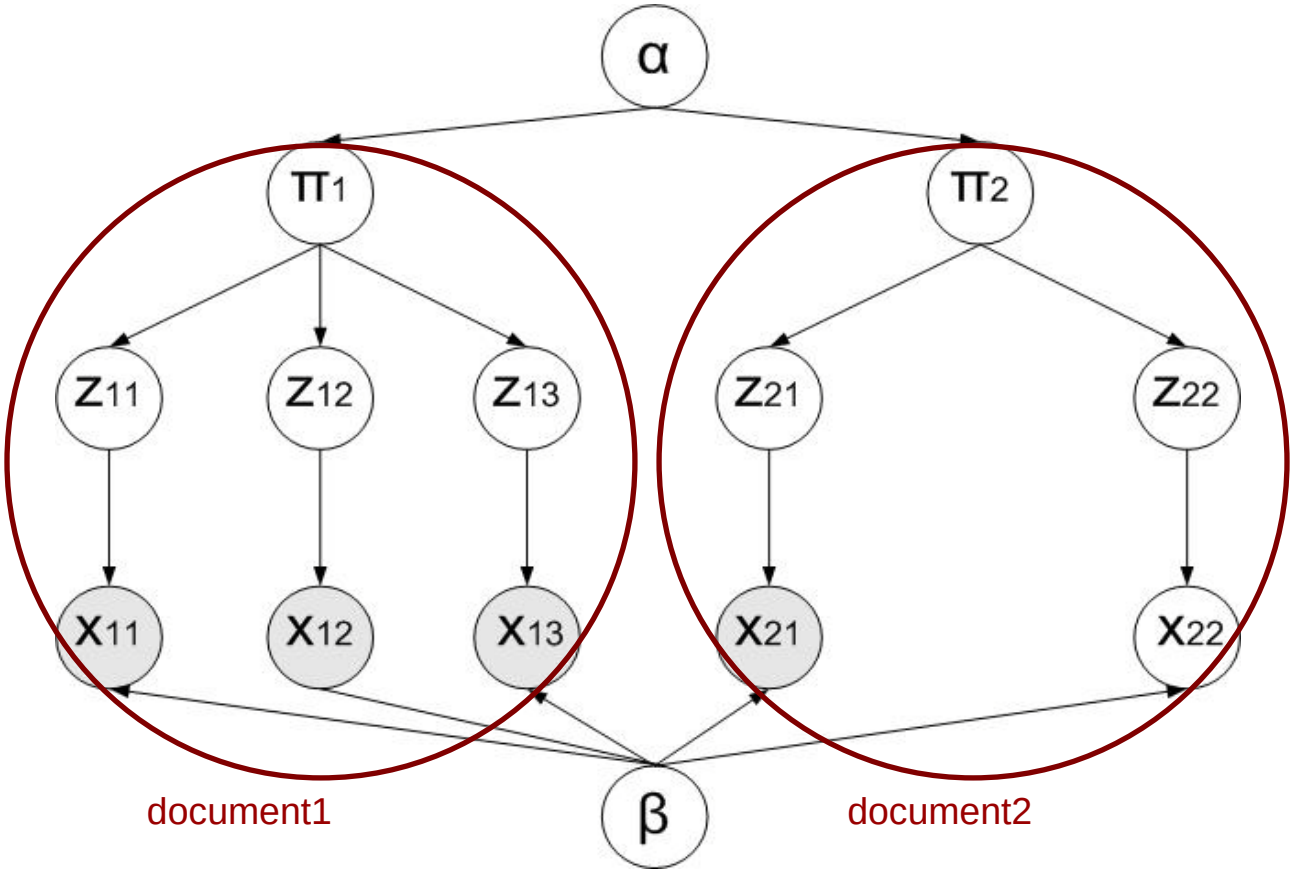


Multi-component mixed membership

Latent Dirichlet Allocation (LDA)



LDA Generative Model



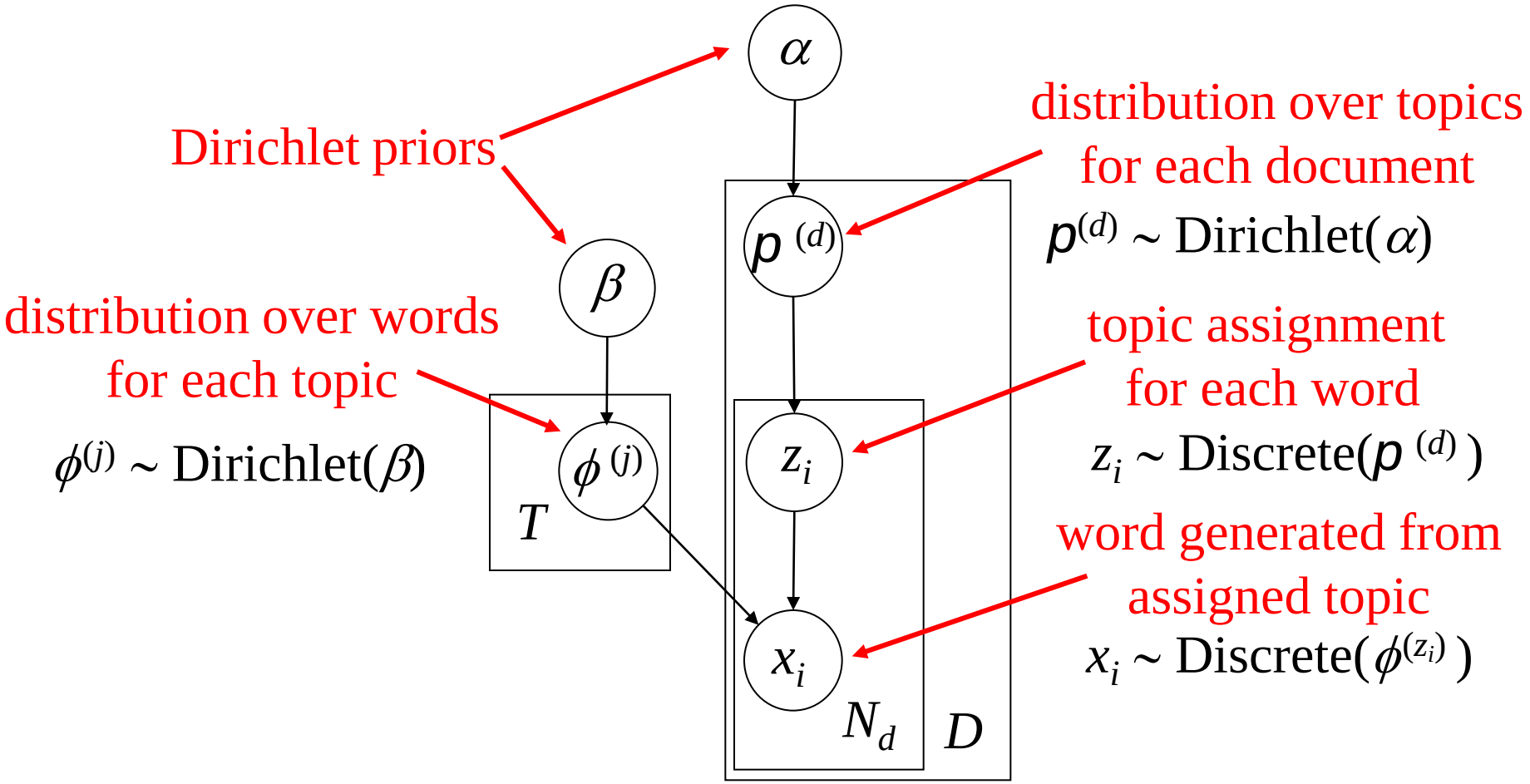
Learning: Inference and Estimation

- Learning
  - Estimate model parameters (α, β) to maximize log-likelihood
  - Infer 'mixed-memberships' of documents

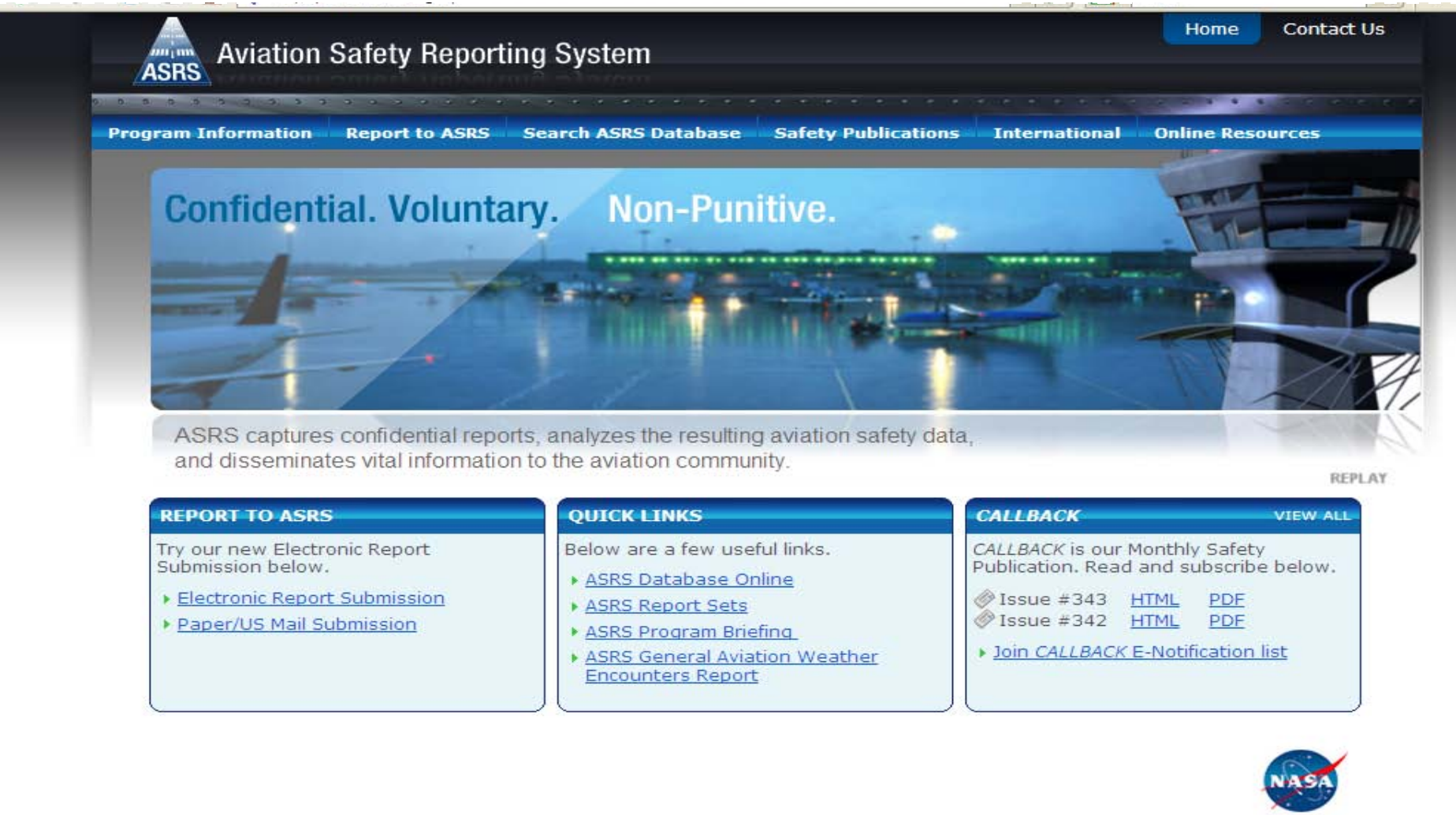
Variational Inference

- Introduce a variational distribution $q(\pi, z|\gamma, \phi)$ to approximate $p(\pi, z|\mathbf{x}, \alpha, \beta)$

Smoothed Latent Dirichlet Allocation



Aviation Safety Reports (NASA)



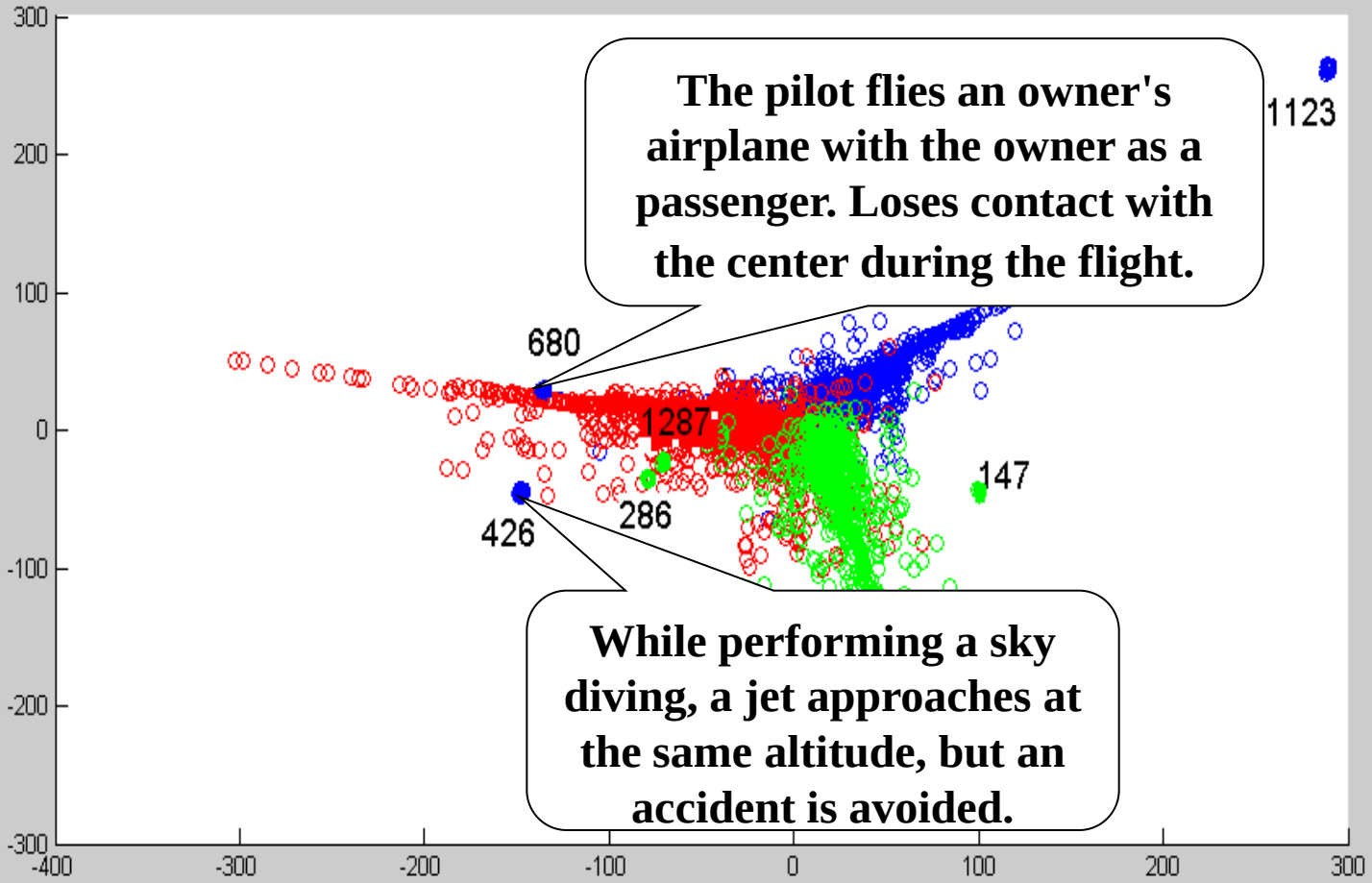
Results: NASA Reports I

Arrival Departure	Passenger	Maintenance
runway	passenger	maintenance
approach	attendant	engine
departure	flight	mel
altitude	seat	zzz
turn	medical	air craft
tower	captain	installed
air traffic control	attendants	check
heading	lavatory	inspection
taxi way	told	fuel
flight	police	Work

Results: NASA Reports II

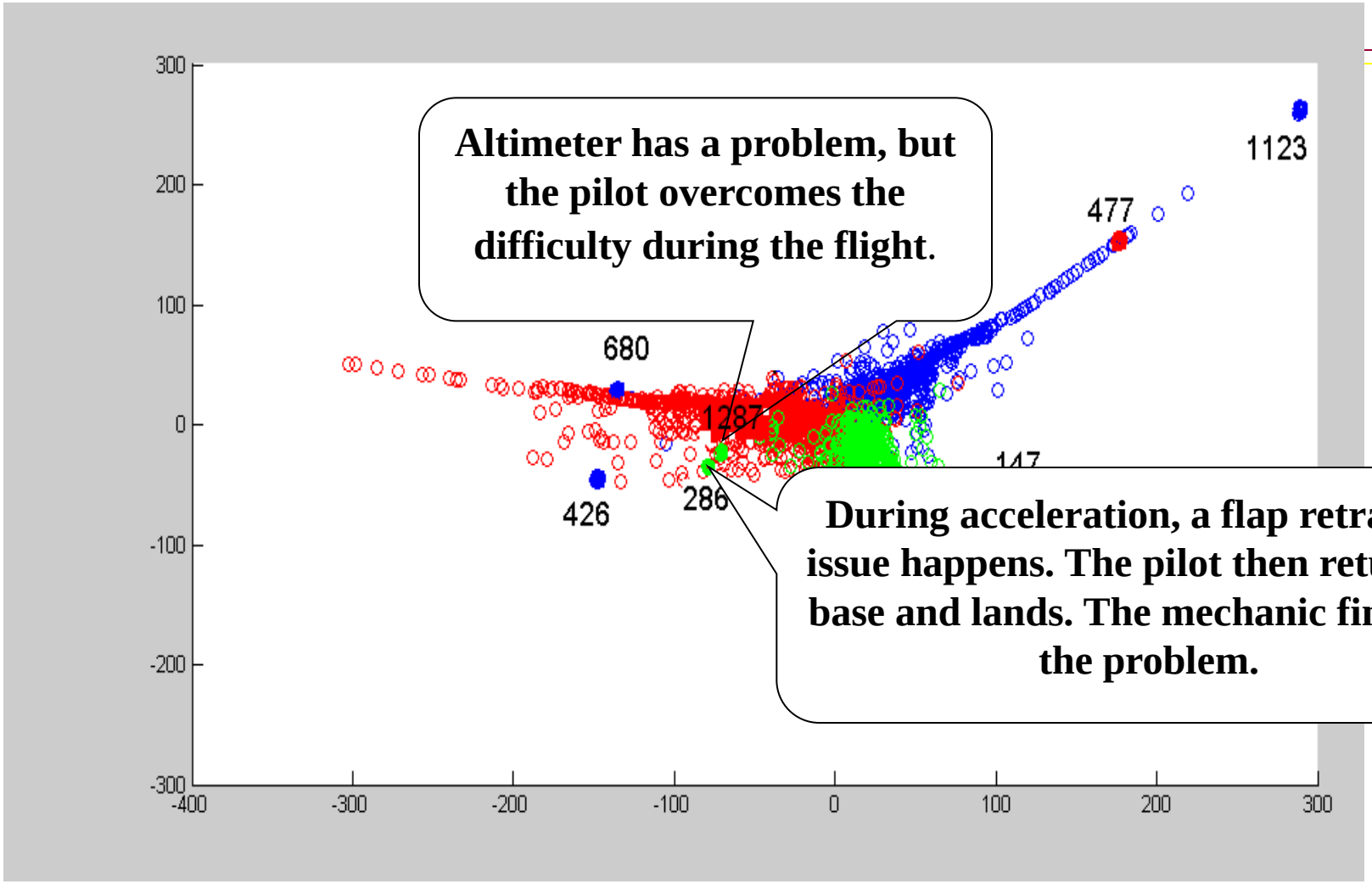
Medical Emergency	Wheel Maintenance	Weather Condition	Departure
medical	tire	knots	departure
passenger	wheel	turbulence	sid
doctor	assembly	aircraft	dme
attendant	nut	degrees	altitude
oxygen	spacer	ice	climbing
emergency	main	winds	mean sea level
paramedics	axle	wind	heading
flight	bolt	speed	procedure
nurse	missing	air speed	turn
aed	tires	conditions	degree

Two-Dimensional Visualization for Reports



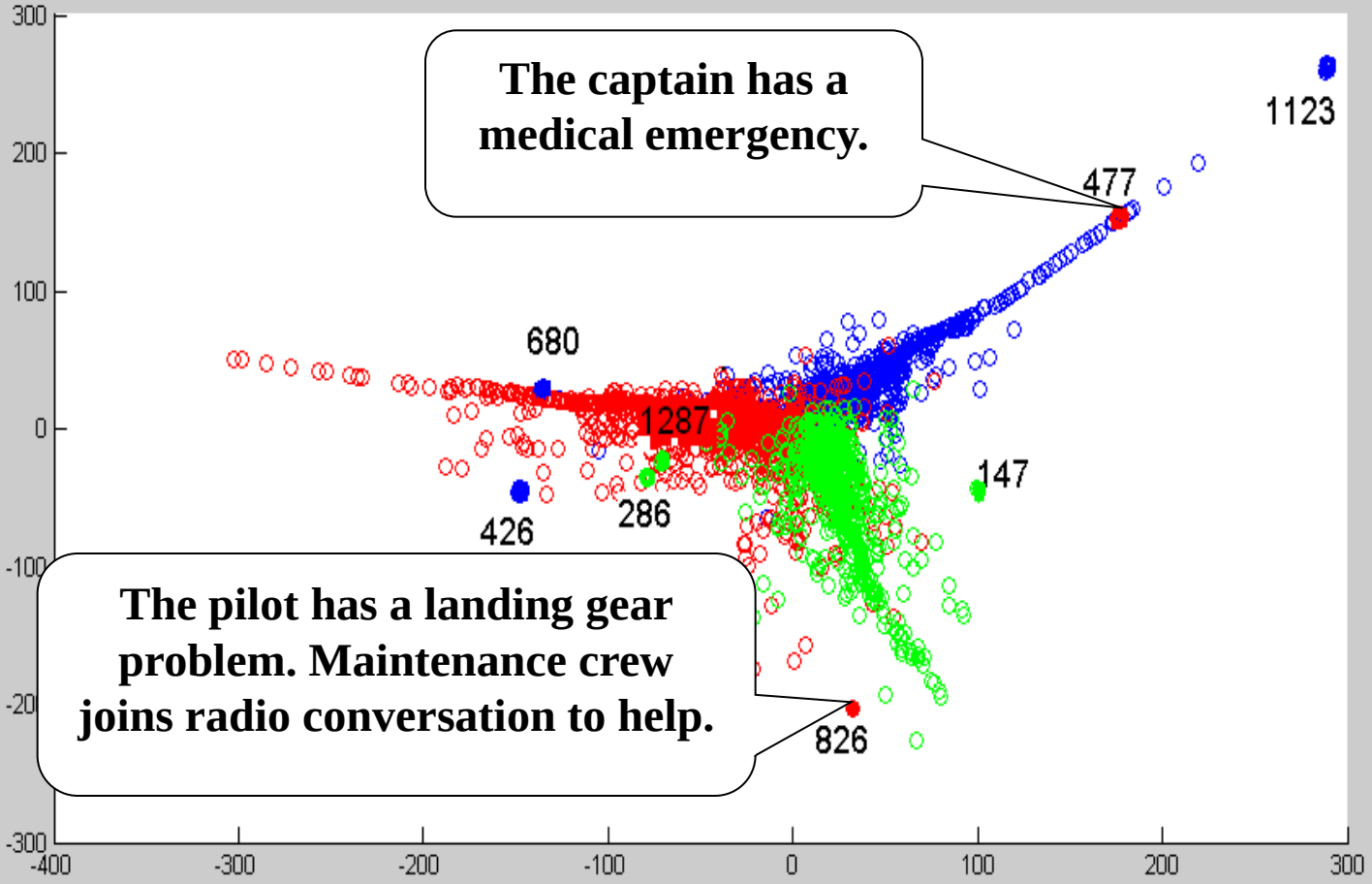
Red: Flight Crew Blue: Passenger Green: Maintenance

Two-Dimensional Visualization for Reports



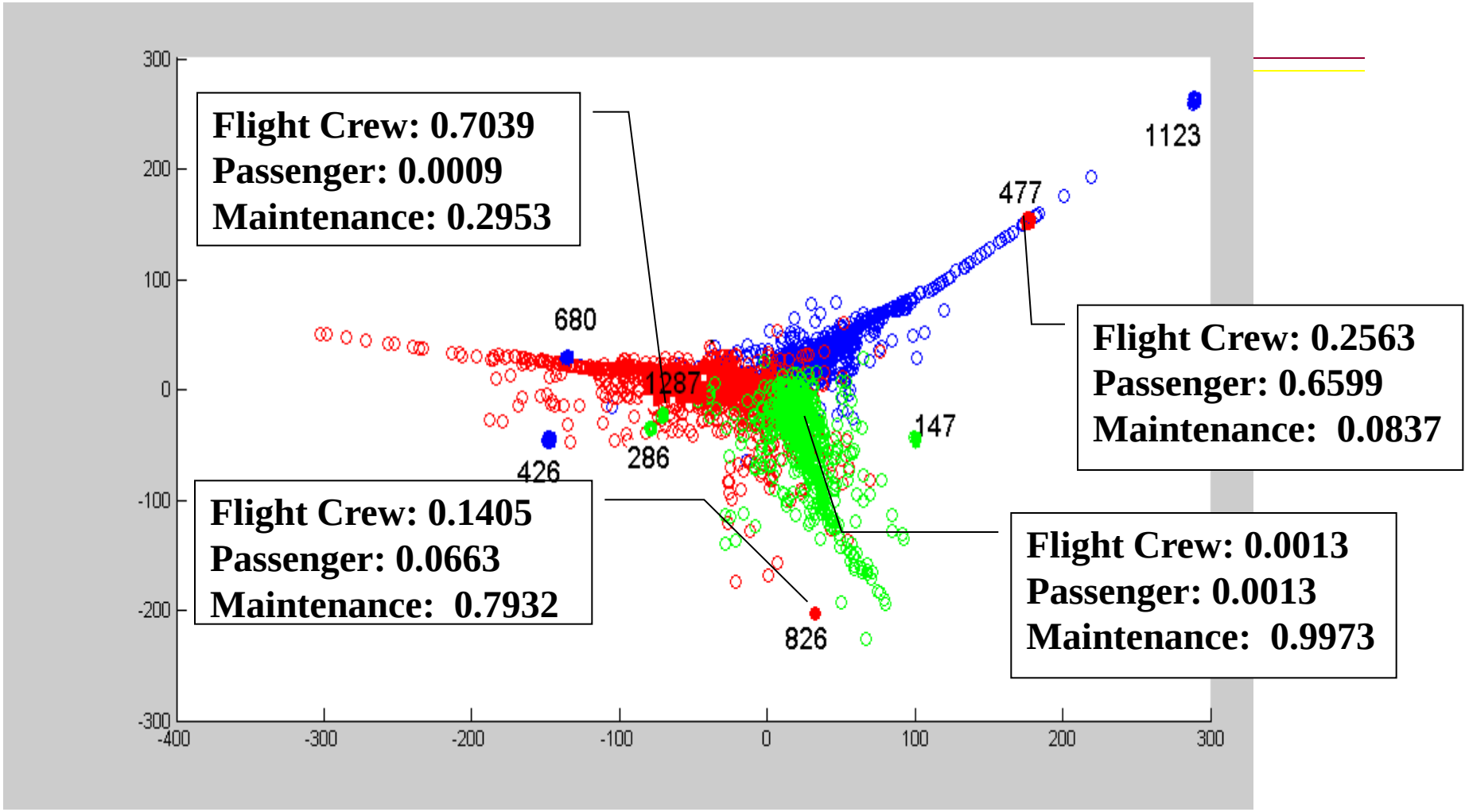
Red: Flight Crew Blue: Passenger Green: Maintenance

Two-Dimensional Visualization for Reports



Red: Flight crew Blue: Passenger Green: Maintenance

Mixed Membership of Reports



Red: Flight Crew Blue: Passenger Green: Maintenance

Generalizations

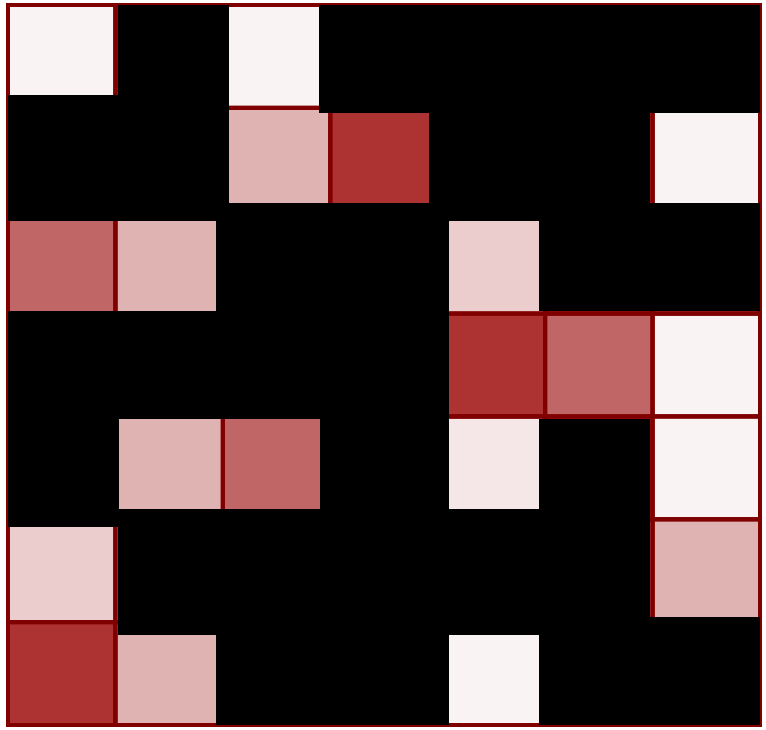
- Generalized Topic Models
  - Correlated Topic Models
  - Dynamic Topic Models, Topics over Time
  - Dynamic Topics with birth/death
- Mixed membership models over non-text data, applications
  - Mixed membership naïve-Bayes
  - Discriminative models for classification
  - Cluster Ensembles
- Nonparametric Priors
  - Dirichlet Process priors: Infer number of topics
  - Hierarchical Dirichlet processes: Infer hierarchical structures
  - Other priors: Gaussian Process, Beta Process, Pitman-Yor Process, etc.

Recommendation Systems

Title: Gone with the wind
Release year: 1940
Cast: Vivien Leigh, Clark Gable
Genre: War, Romance
Awards: 8 Oscars
Keywords: Love, Civil war
…

Movies

Users



Age: 28
Gender: Male
Job: Sales man
Interest: Travel
…

Movie ratings matrix

Advertisements on the Web

Category: Sports shoes
Brand: Nike
Ratings: 4.2/5
…

Category: Baby
URL: babyearth.com
Content: Webpage text
Hyperlinks:

Webpages

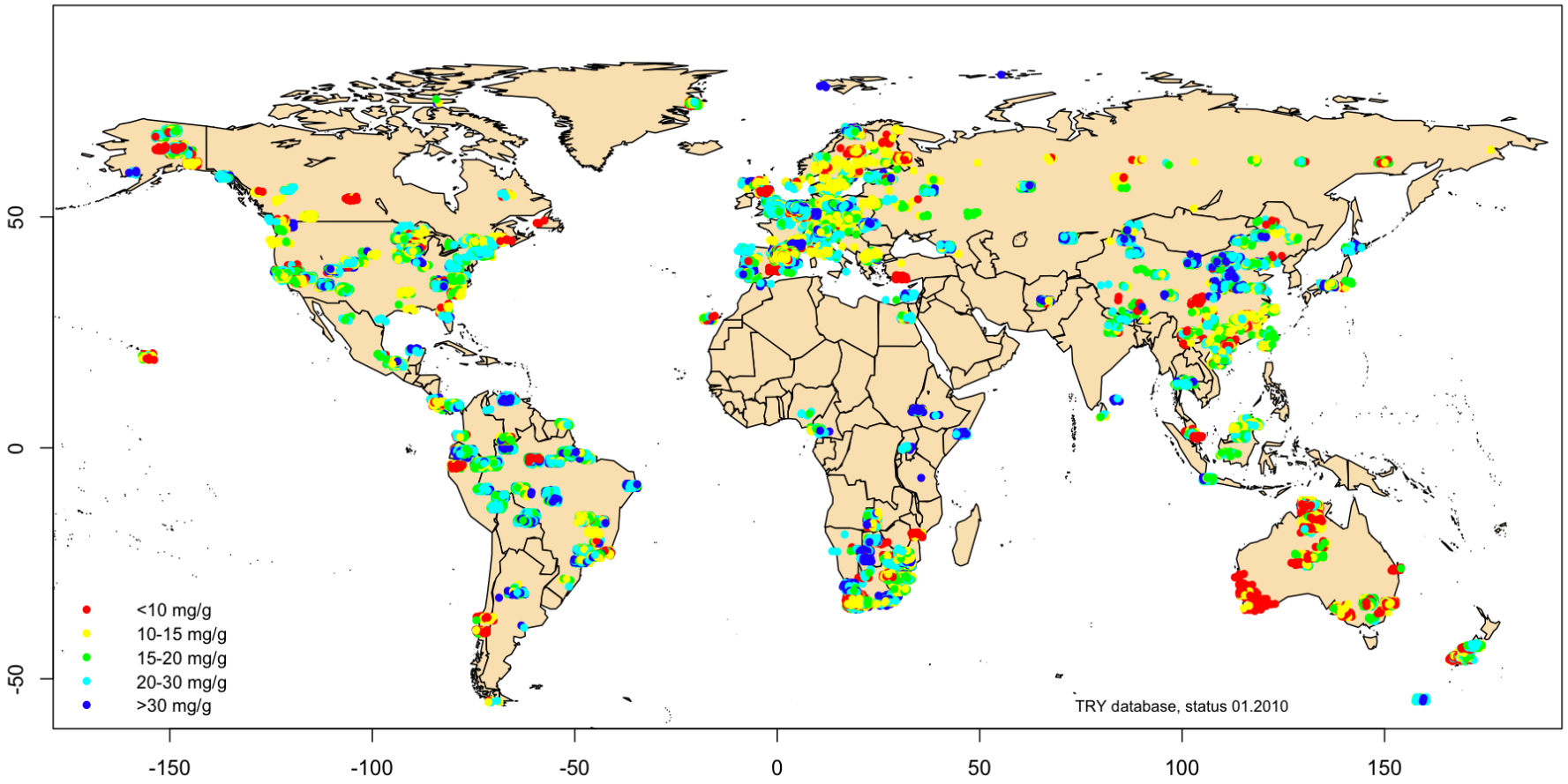
Products

…

1%	2%			0.01%		…
0.1%		2%	3%			…
2%	2%			0.5%		…
		0.2%	0.3%	1.5%	2%	…
				2.5%	1%	…
		1.5%	1%		0.04%	…

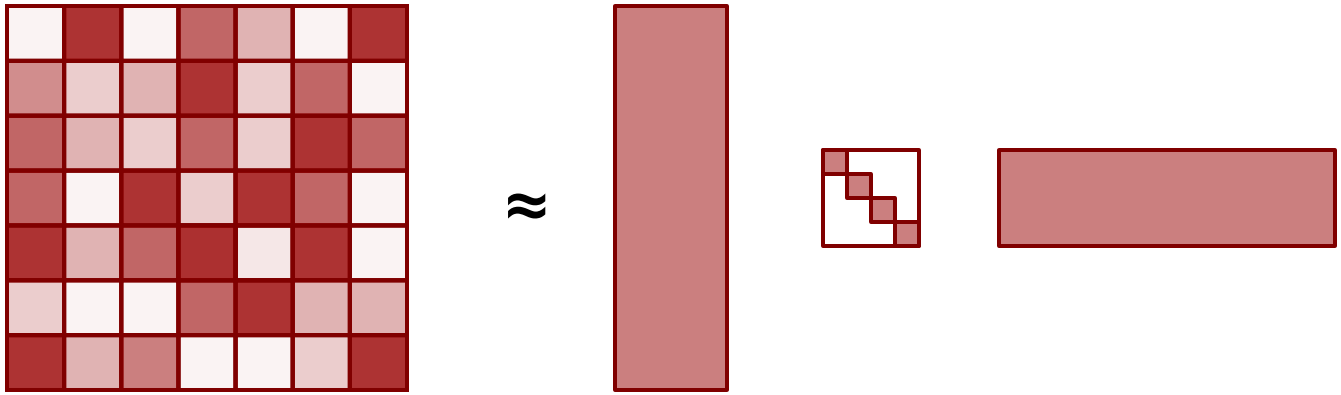
Click-Through-Rate matrix

Forest Ecology

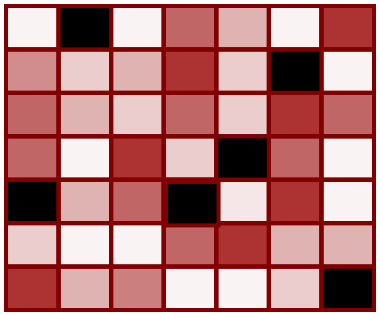


Matrix Factorization

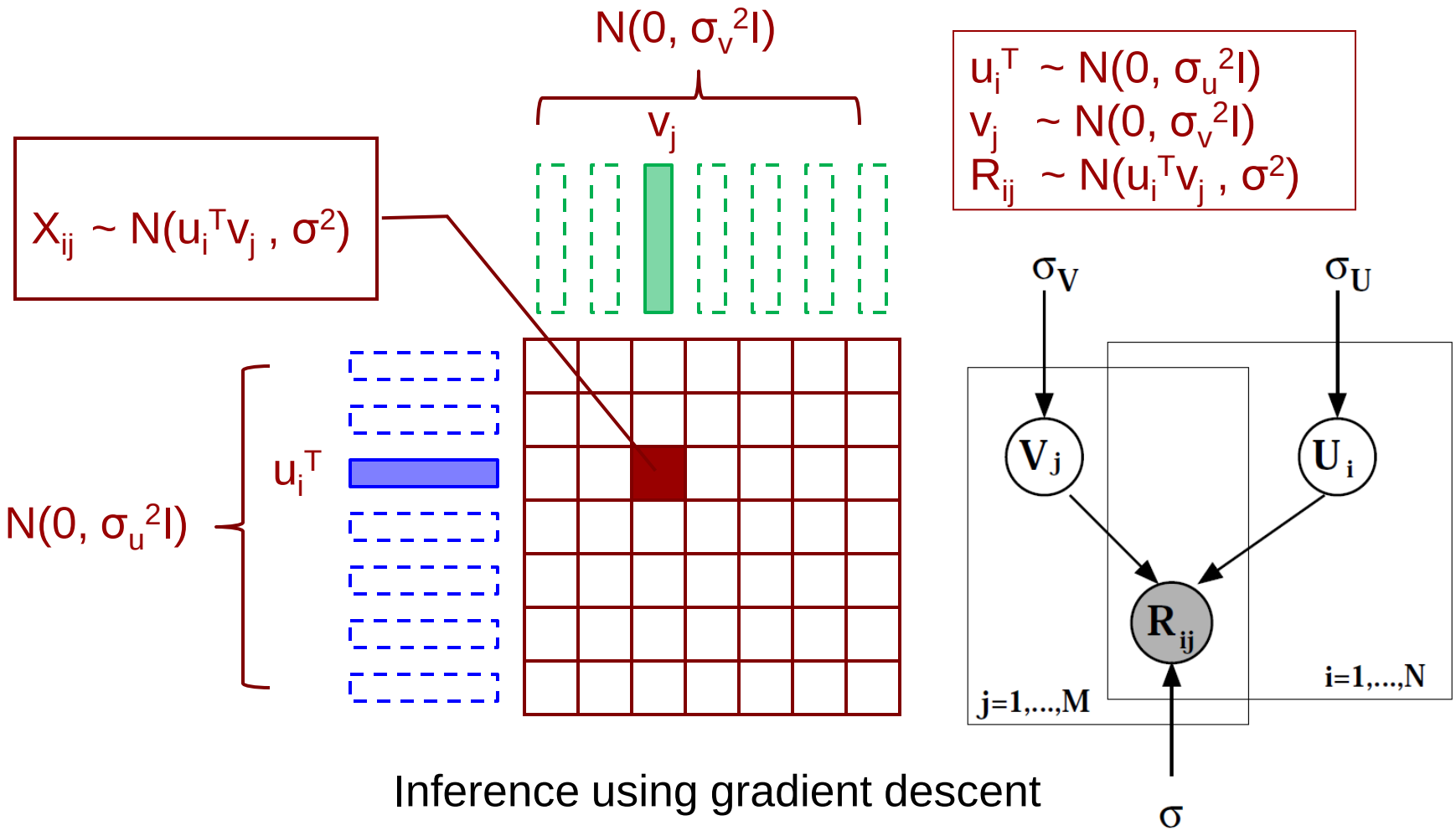
- Singular value decomposition



- Problems
  - Large matrices, with millions of row/columns
    - SVD can be rather slow
  - Sparse matrices, most entries are missing
    - Traditional approaches cannot handle missing entries



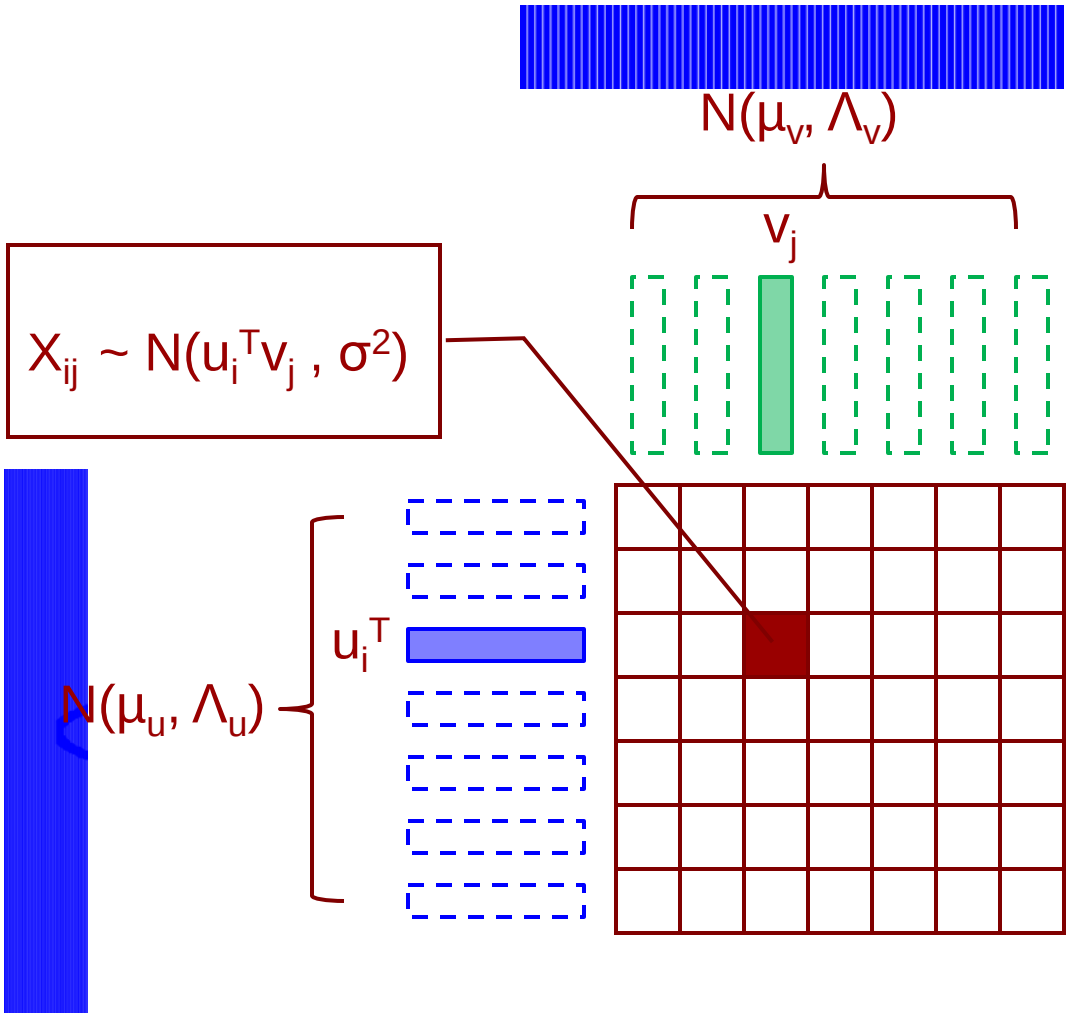
Probabilistic Matrix Factorization (PMF)

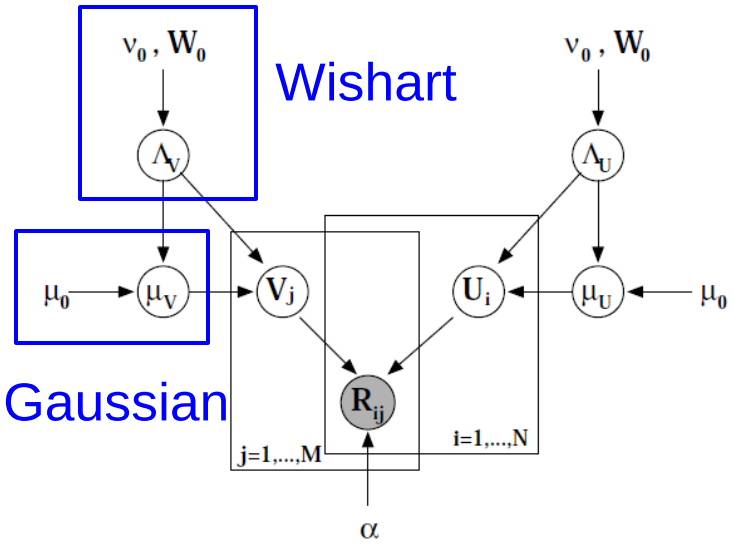


Inference using gradient descent

Bayesian Probabilistic Matrix Factorization

$\mu_u \sim N(\mu_0, \Lambda_u), \Lambda_u \sim W(v_0, W_0)$
$\mu_v \sim N(\mu_0, \Lambda_v), \Lambda_v \sim W(v_0, W_0)$
$u_i \sim N(\mu_u, \Lambda_u)$
$v_j \sim N(\mu_v, \Lambda_v)$
$R_{ij} \sim N(u_i^T v_j, \sigma^2)$

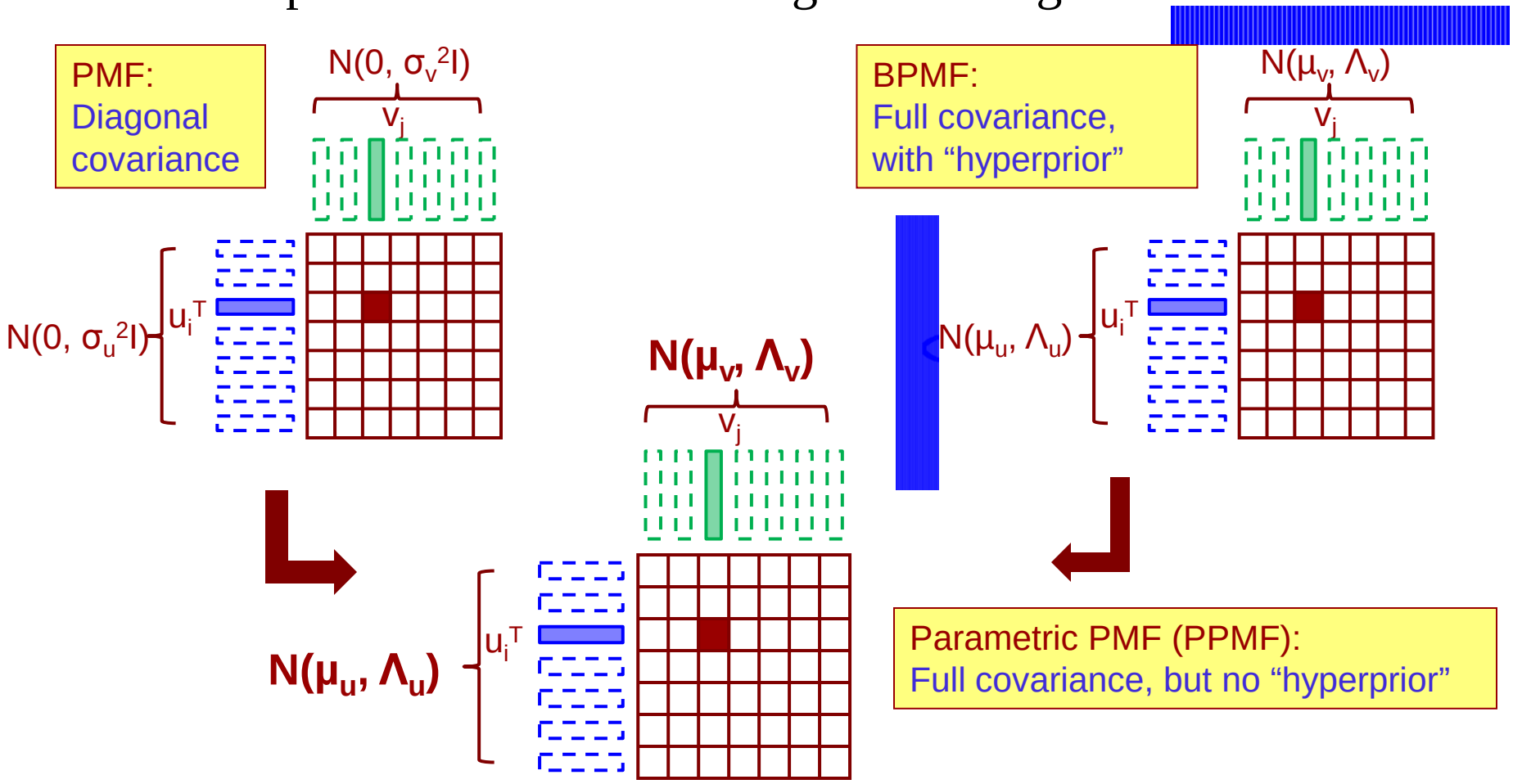




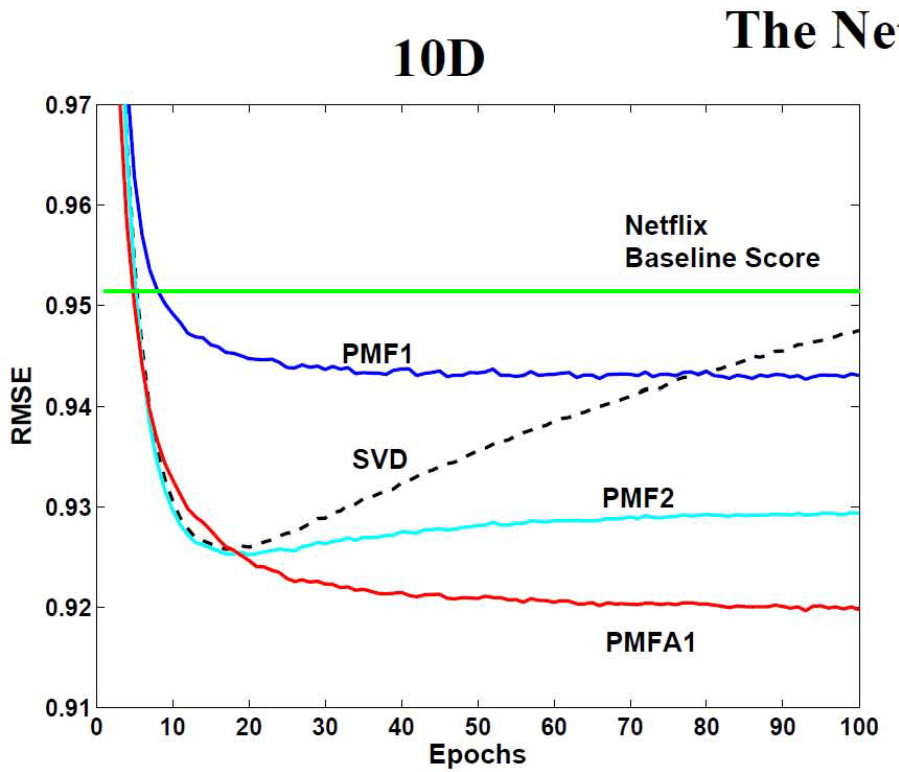
Inference using MCMC

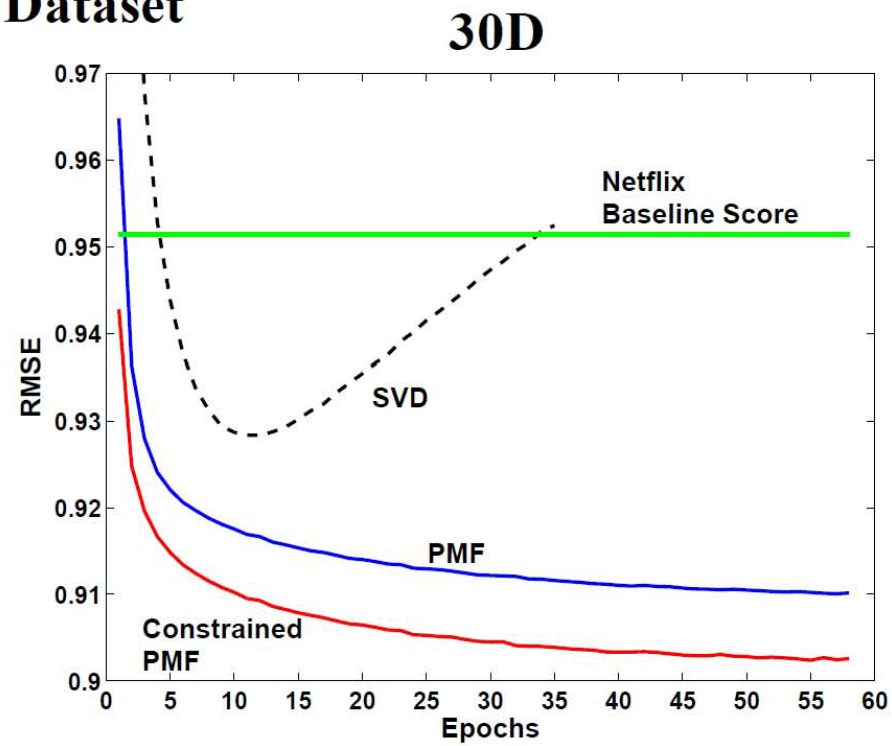
Parametric PMF (PPMF)

- Are the priors suitable? Can we get faster algorithms?

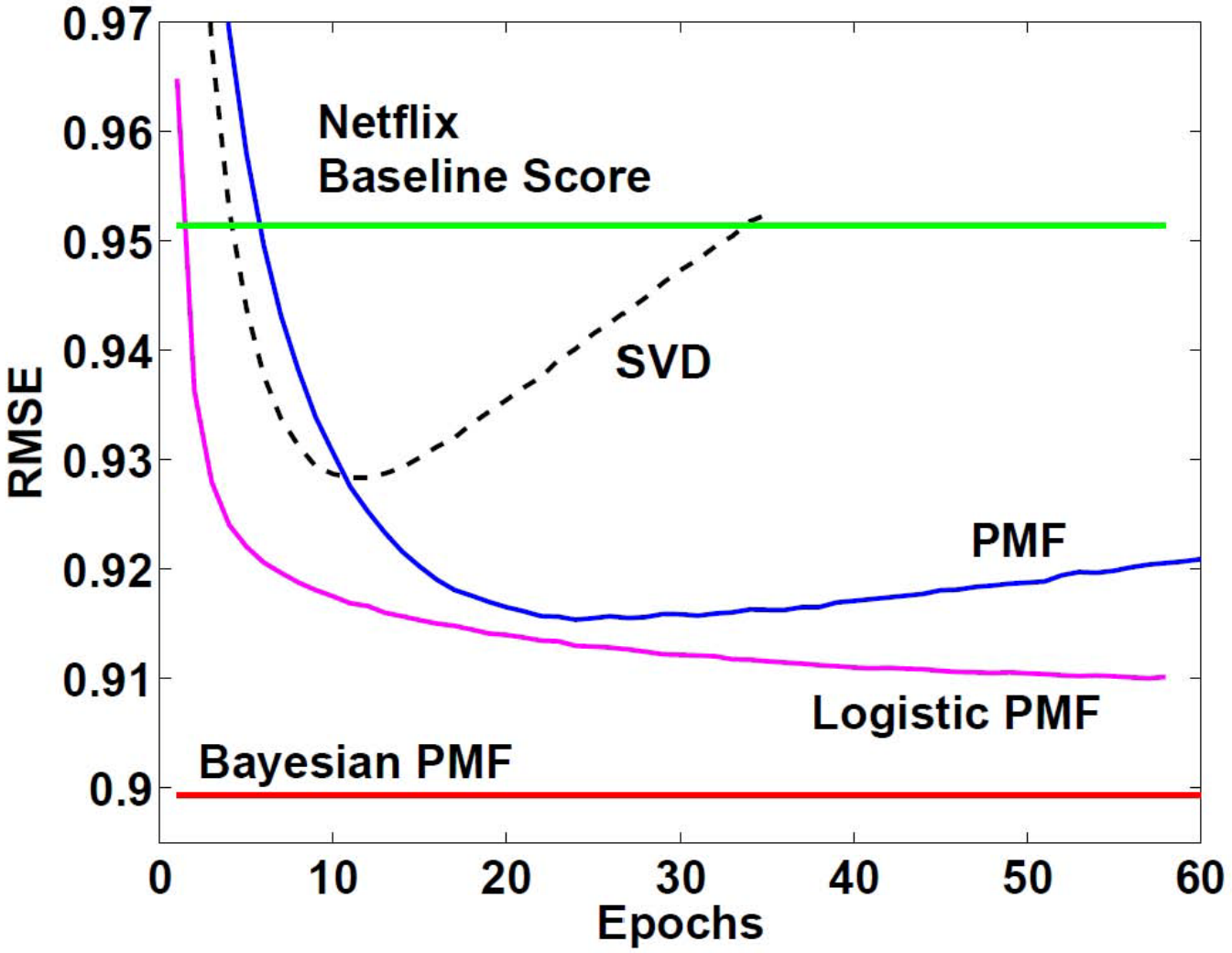


Results: PMF on Netflix

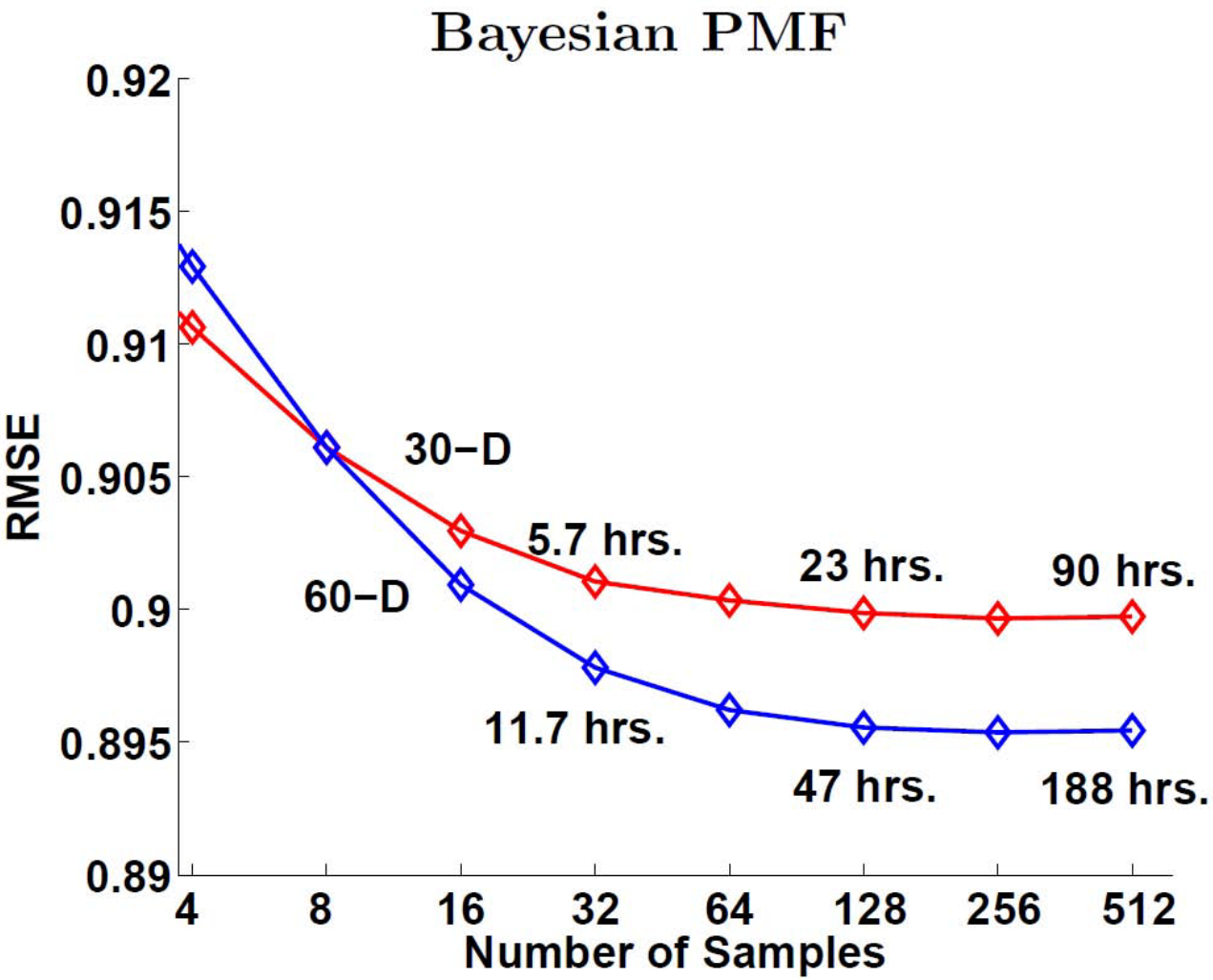




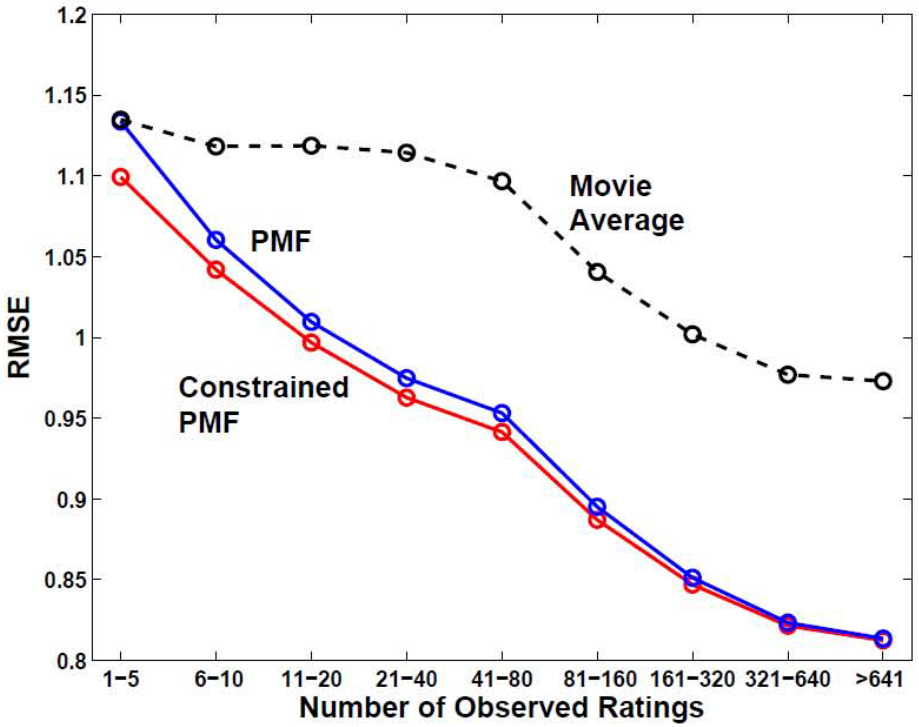
Results: Bayesian PMF on Netflix

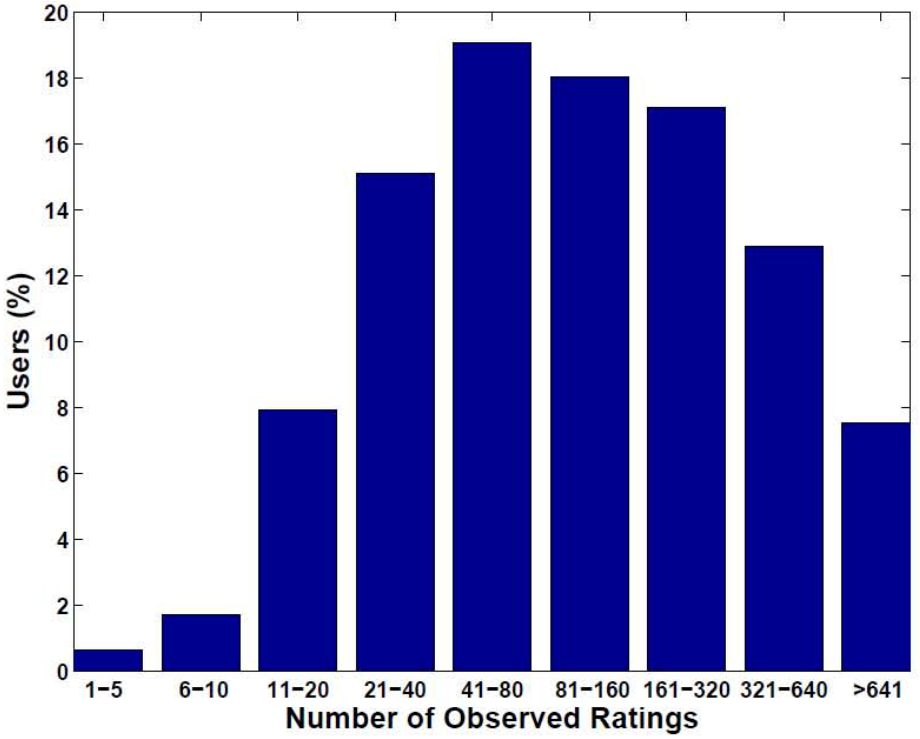


Results: Bayesian PMF on Netflix

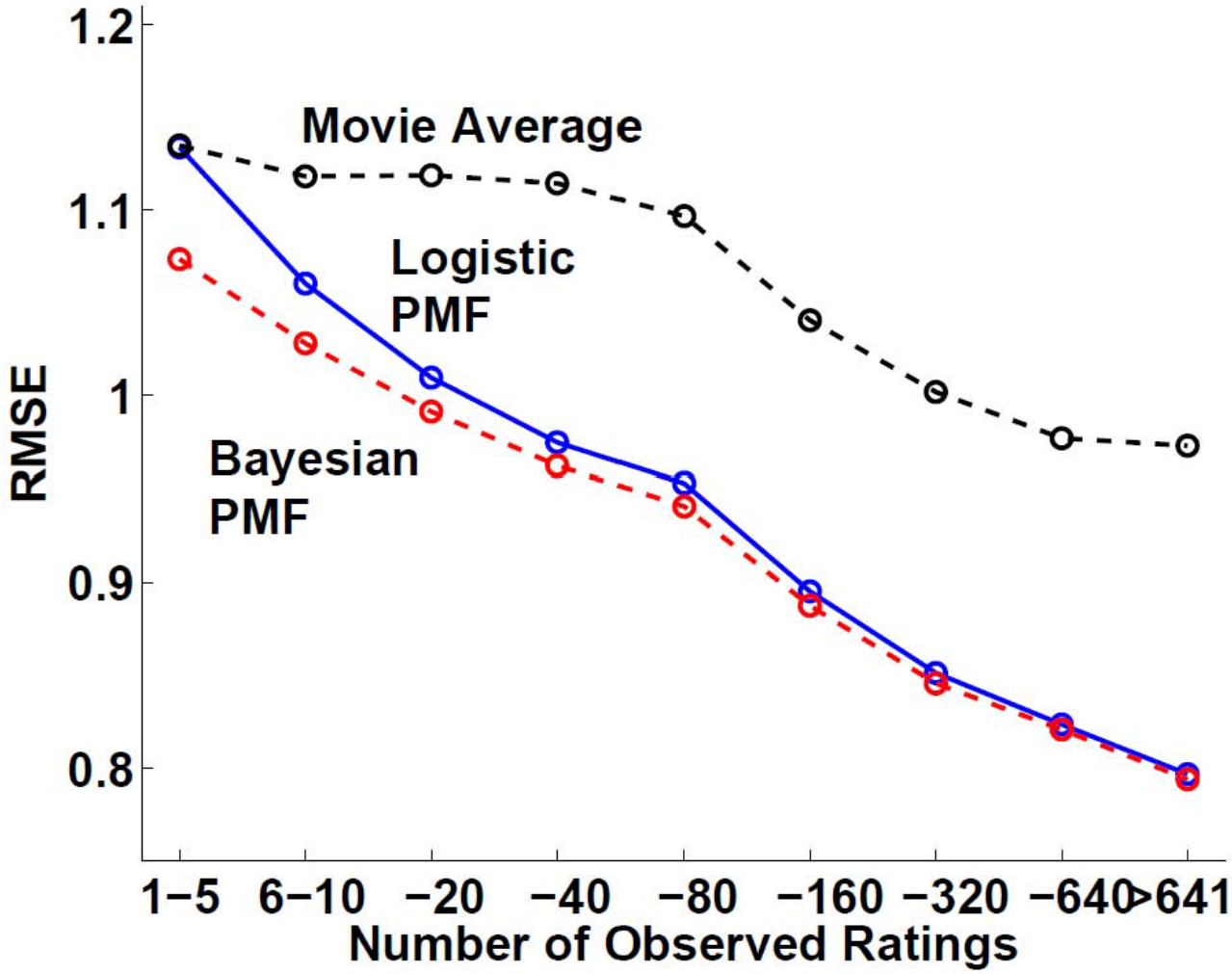


Results: PMF on Netflix





Results: Bayesian PMF on Netflix

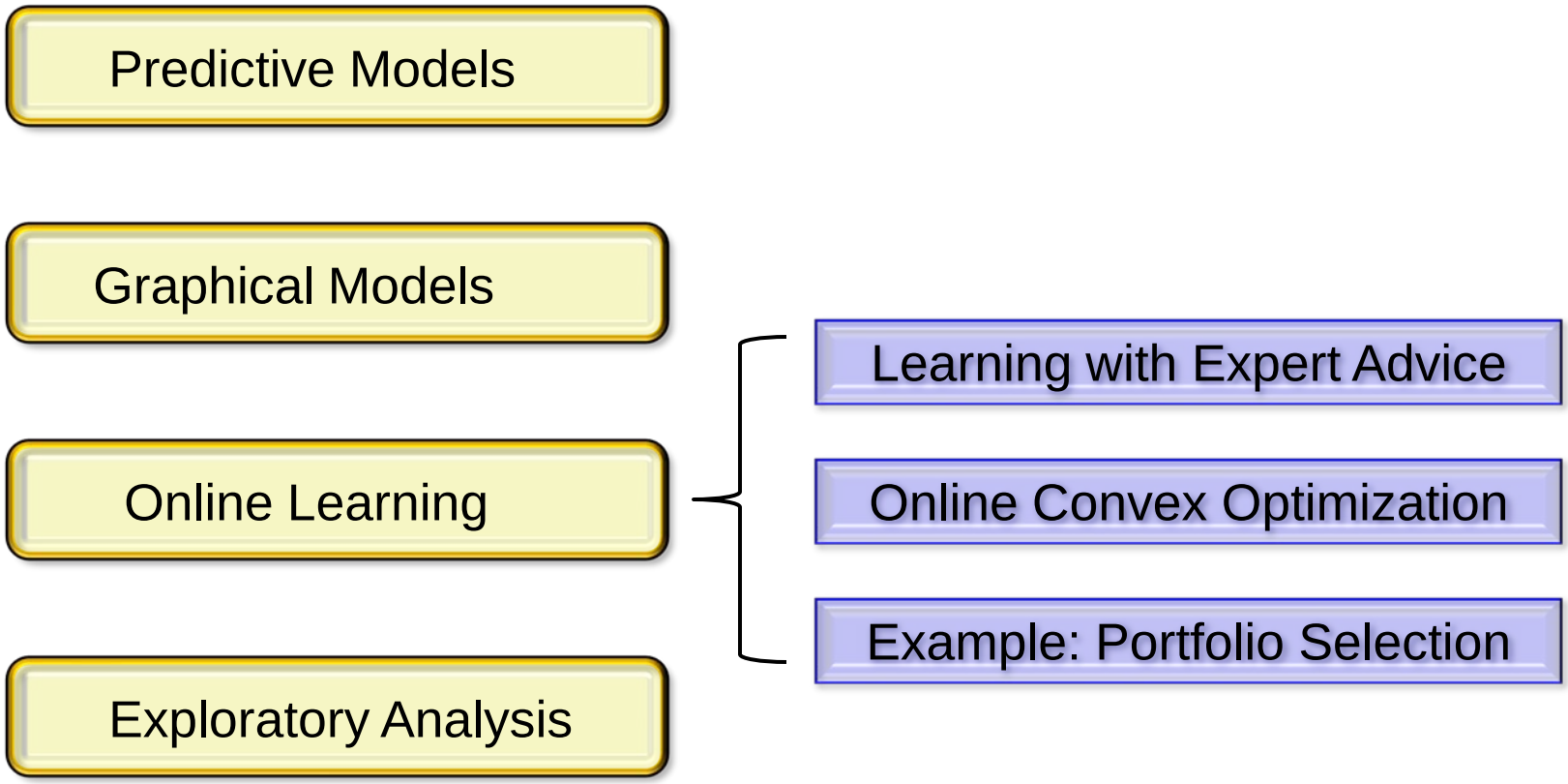


PPMF vs. PMF, BPMF

k	5	10	15	20	25	30
PMF	**0.07996 ±0.00044**	0.08040 ±0.00036	0.08238 ±0.00070	0.08343 ±0.00067	0.08362 ±0.00047	0.08431 ±0.00059
BPMF	0.08517 ±0.00055	0.08685 ±0.00058	0.08970 ±0.00062	0.09382 ±0.00063	0.09879 ±0.00067	0.10525 ±0.00076
PPMF	0.08003 ±0.00057	**0.07837 ±0.00051**	**0.07849 ±0.00050**	**0.07840 ±0.00055**	**0.07855 ±0.00054**	**0.07832 ±0.00059**

PPMF mostly achieves higher accuracy

Overview


Predictive Models
Graphical Models
Online Learning
Exploratory Analysis
Learning with Expert Advice
Online Convex Optimization
Example: Portfolio Selection

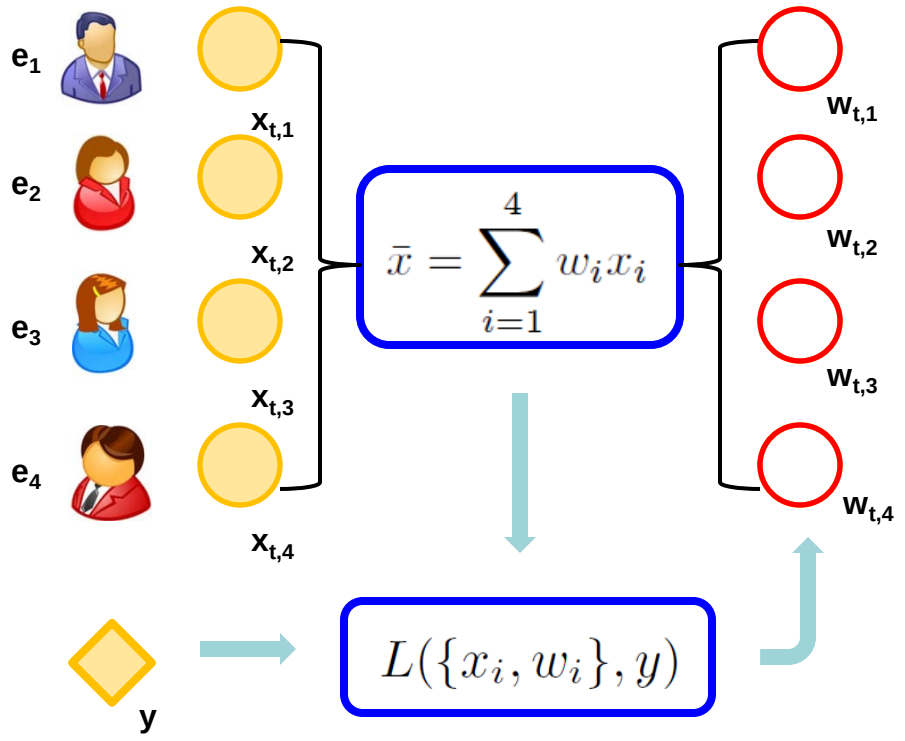
Learning with Expert Advice

- Prediction by adaptive combination of expert outputs
  - At round t:

    Expert prediction $\mathbf{x}^t$,

    Weight on Experts: $\mathbf{w}^t$
  - Combined prediction $\bar{x} = \mathbf{w}^t \cdot \mathbf{x}^t$
  - Nature reveals truth: y^t
  - Loss of expert i: $L\left(x_i^t, w_i^t, y^t\right)$
  - Update "confidence" on experts:

$$w_i^{t+1} = \frac{w_i^t \exp(-\eta L(x_i^t, w_i^t, y^t))}{Z^{t+1}}$$



Time Machine (TM) vs Online Learning (OL)

Goal: Maximize prediction accuracy

TM: Choose best w* in hindsight

Base Models

x_1 x_2 x_3 x_4 x_5 . . . . . x_T

x_1 x_2 x_3 x_4 . . . . . x_T

w_1 w_2 w_3 w_4 w_T

OL: Choose w_t before seeing x_t

Guarantee: OL will be competitive with TM

Online Concave Optimization (OCO)

- Batch Concave Optimization:
  - Given a concave function ϕ and a feasible set P

$$\max_{\mathbf{z} \in P} \phi(\mathbf{z})$$

OCO: Good News and Bad News

- The regret guarantee for OCO

$$\sum_{t=1}^{T} \phi_t(\mathbf{z}_t) \geq \max_{\mathbf{z} \in P} \sum_{t=1}^{T} \phi_t(\mathbf{z}) - o(T)$$

Portfolio Selection: Notation and Definitions

- Assume some unit of time: 'day'
- Consider n stocks over T days

The Portfolio Selection Problem

- The total multiplicative gain in wealth:

$$S_T(\mathbf{p}_{1:T}, \mathbf{x}_{1:T}) = \prod_{t=1}^{T} \mathbf{p}_t^T \mathbf{x}_t$$

- The log-gain

$$LS_T(\mathbf{p}_{1:T}, \mathbf{x}_{1:T}) = \sum_{t=1}^{T} \log(\mathbf{p}_t^T \mathbf{x}_t)$$

Constant Rebalanced Portfolio (CRP)

- The total multiplicative log-gain in wealth:

$$LS_T(\mathbf{p}_{1:T}, \mathbf{x}_{1:T}) = \sum_{t=1}^{T} \log(\mathbf{p}_t^T \mathbf{x}_t)$$

- CRP: Use a fixed portfolio $\mathbf{p}$ over time

Regret: Competing with the Best CRP

Regret:
Difference of log wealth between Best CRP and Online Strategy

On-Line

Best CRP

$$\sum_{t=1}^{T} \log(w_t \cdot x_t) \geq \sum_{t=1}^{T} \log(w^* \cdot x_t) - o(T)$$

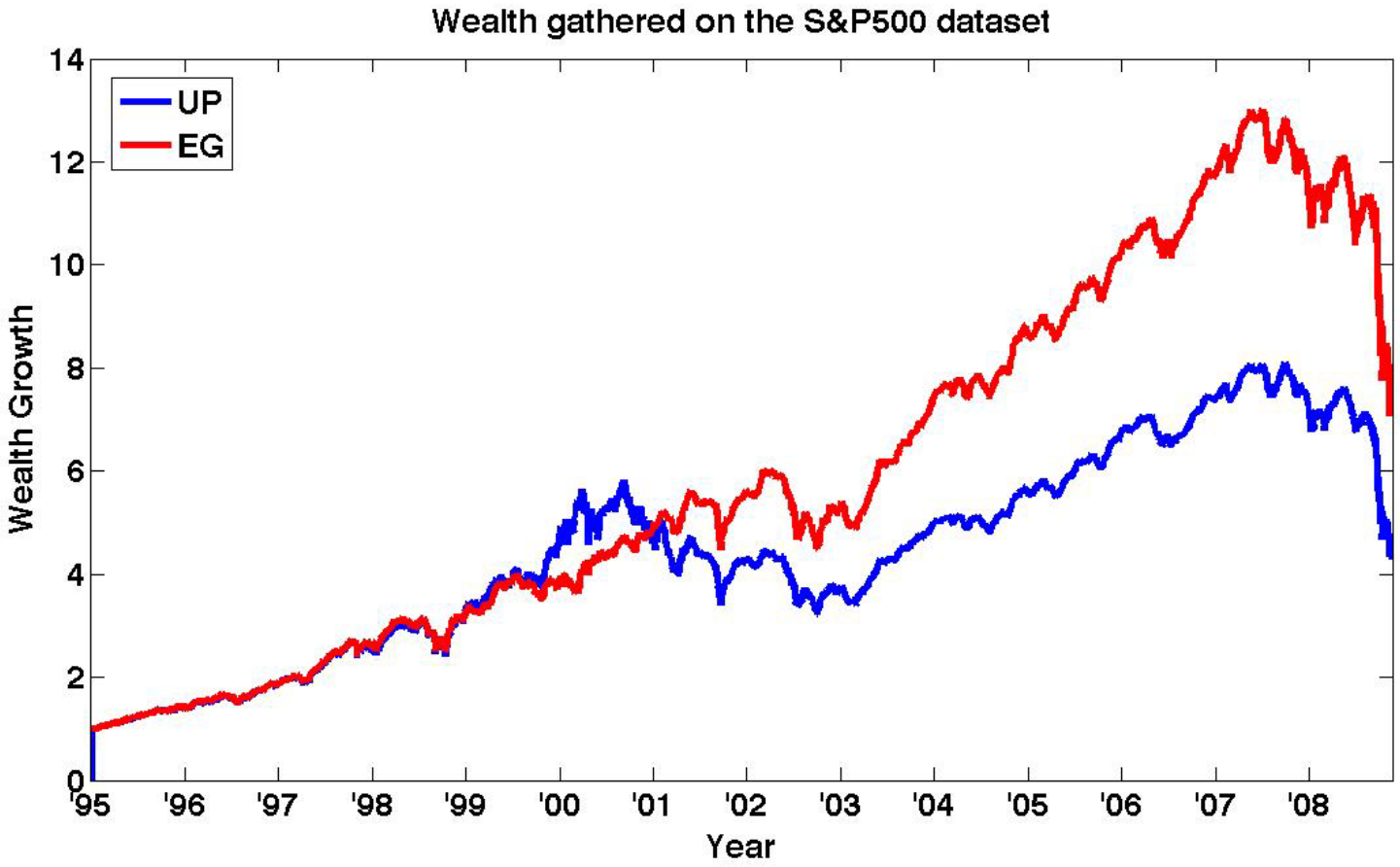
$$\frac{1}{T}\sum_{t=1}^{T} \log(w_t \cdot x_t) \geq \frac{1}{T}\sum_{t=1}^{T} \log(w^* \cdot x_t) - \varepsilon$$

Datasets

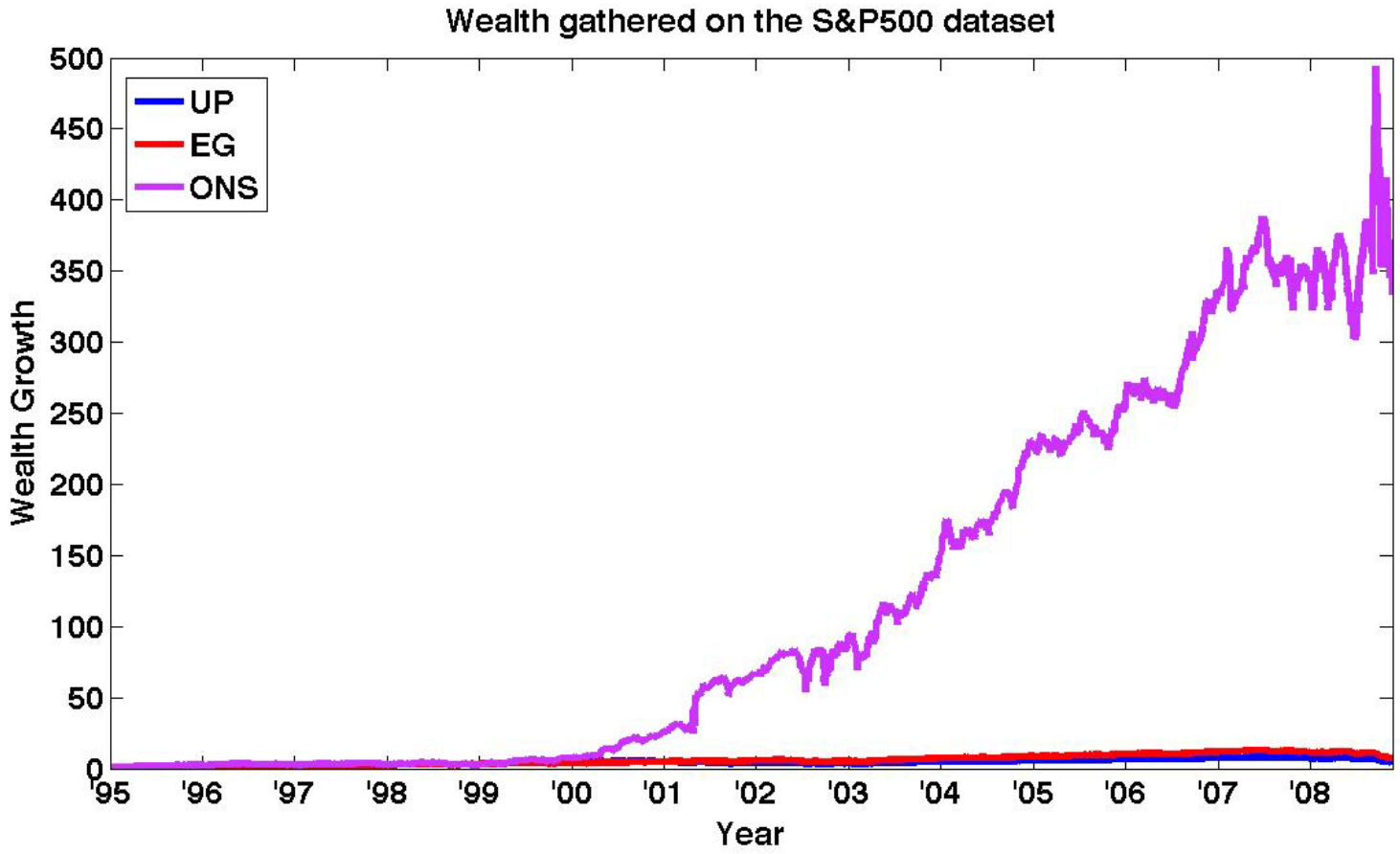
- S&P 500
  - S&P 500 Index: 500 large cap actively traded stocks
  - All stocks that were in S&P 500 from Jan, 1995 to Nov, 2008
  - Total 385 stocks over 14 years at daily resolution
  - Two financial meltdowns: Dot com (Mar 2000), Housing (Oct 2007)

- NYSE
  - Widely used dataset in most earlier papers
  - 36 large cap stocks over 22 years at daily resolution
  - From July, 1962 to Dec, 1984
  - One major bear market: 1973-1974

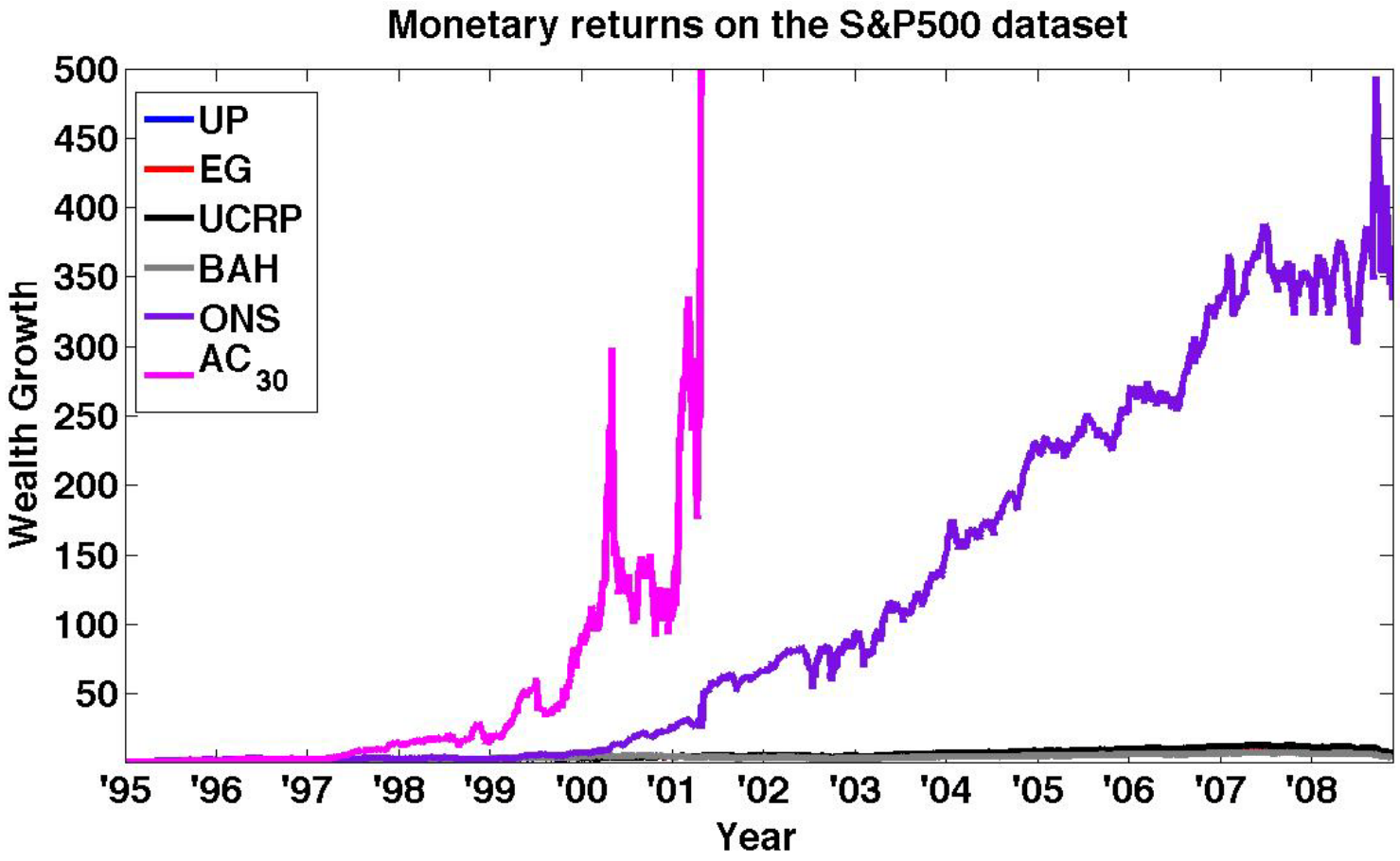
Results: UP, EG



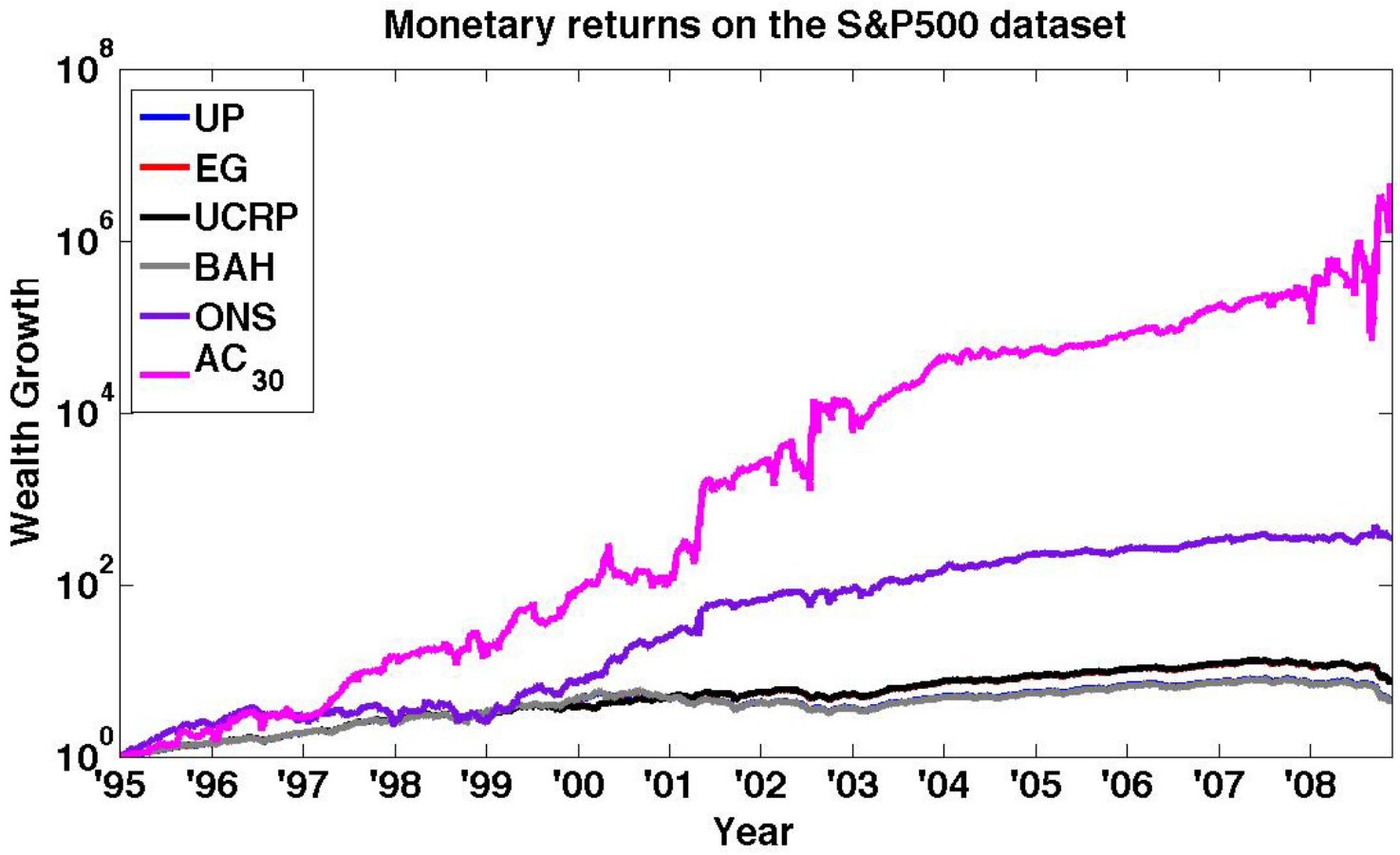
Results: UP, EG, ONS



Results: UP, EG, ONS, AC_{30}



Results: UP, EG, ONS, AC_{30}



Motivation for Meta Algorithms

Universal Algorithms

Competitive with best CRP

Can be outdone by simple heuristics

Heuristic based methods:

Strong Empirical Performance

No performance guarantee

Best of both worlds

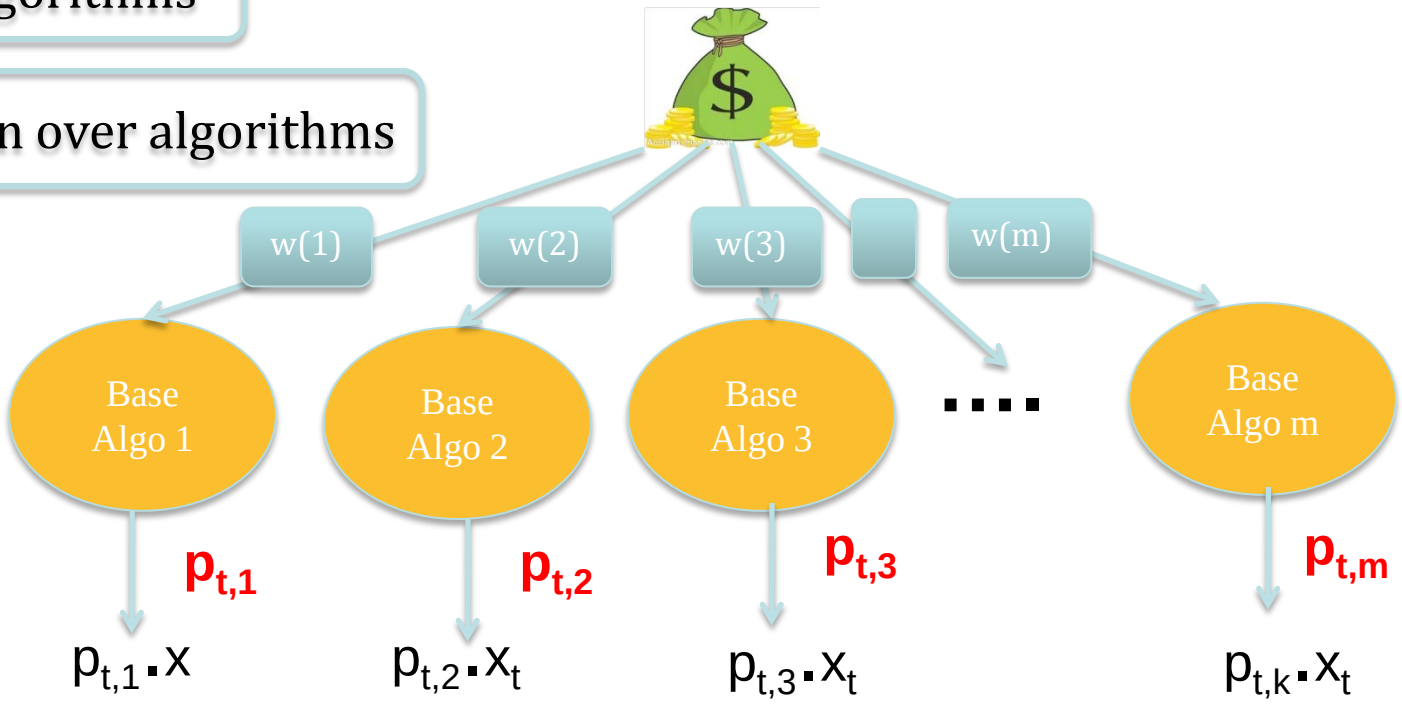
Competitive with best CRP

Strong empirical performance

Meta Algorithm

m base algorithms

w_t : distribution over algorithms



Day t

$\mathbf{x_t}$

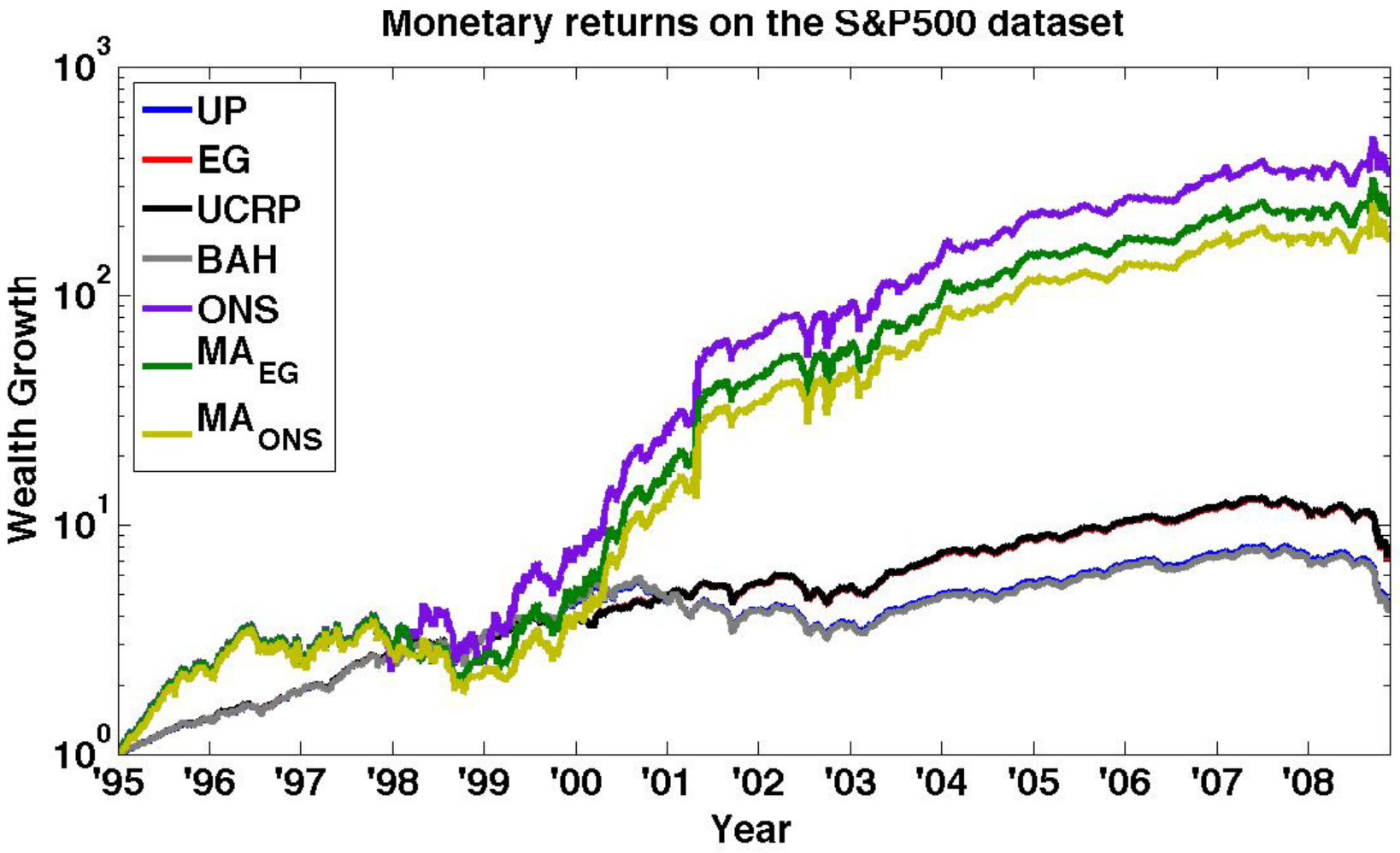
Wealth Gain

$r_{t,j} = p_{t,j} \cdot x_t$

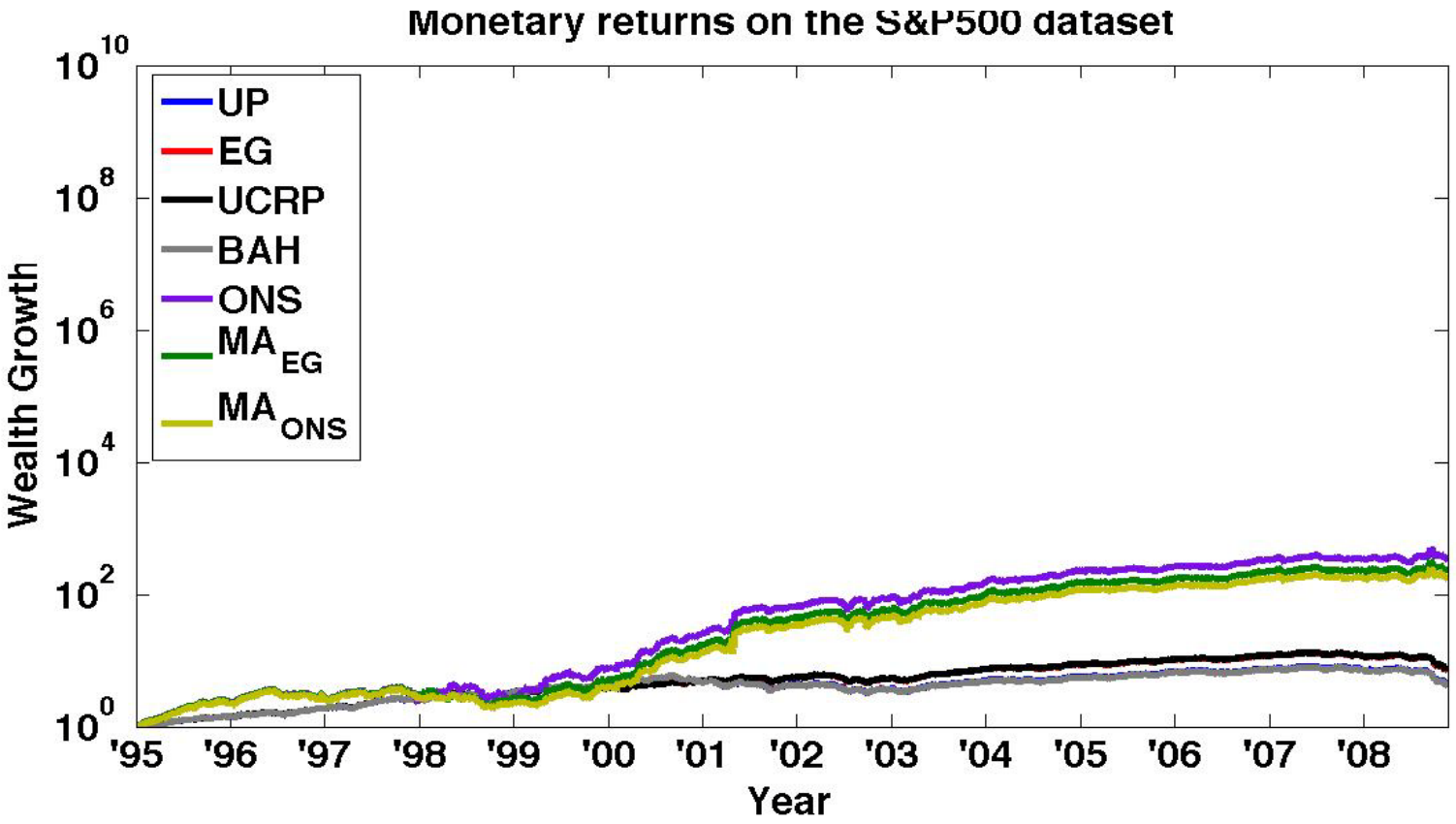
Weight update

- w_t updated based on r_t
- Good algorithms get more money

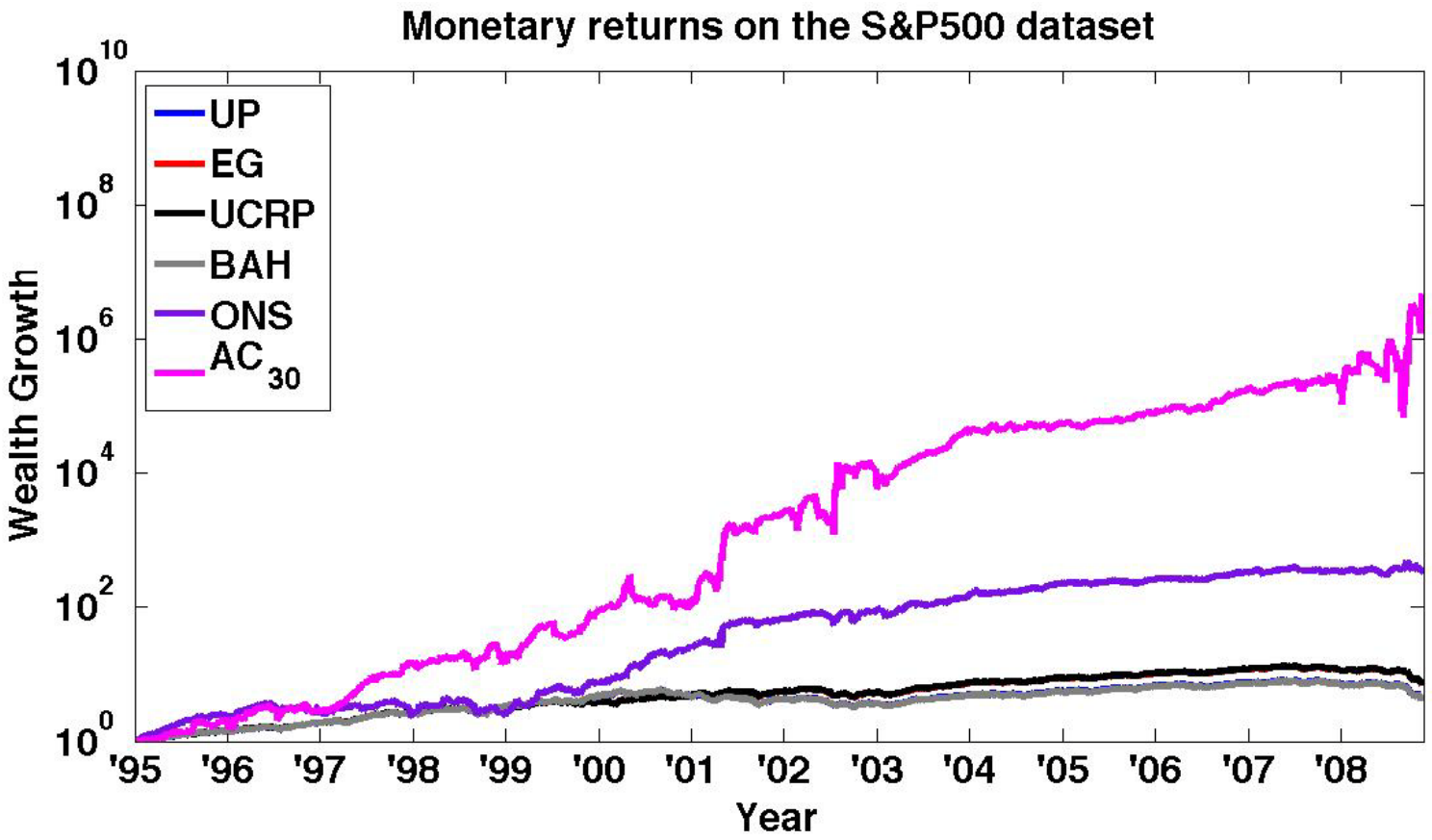
Results: MA_{EG}, MA_{ONS}



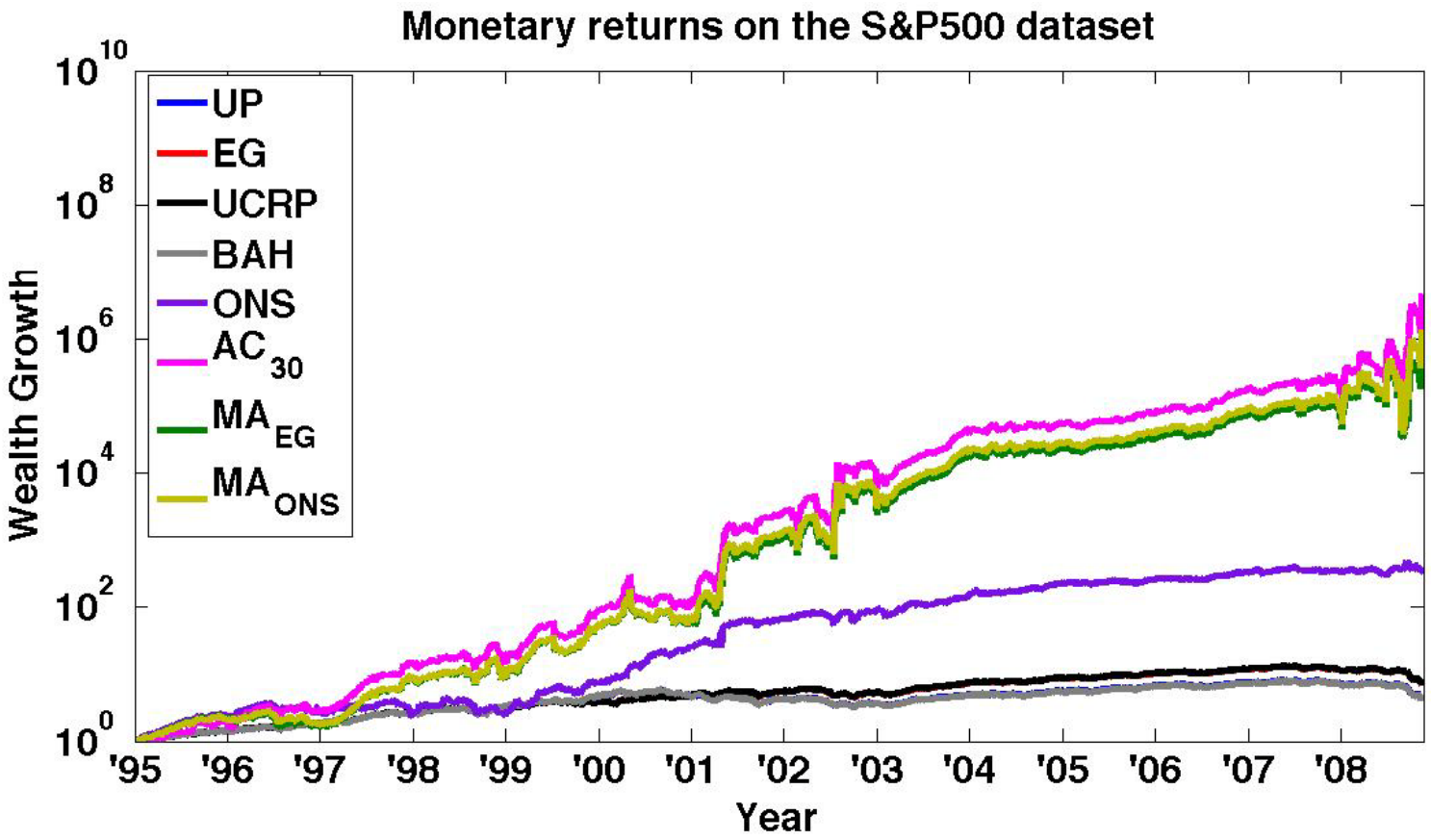
Results: MA_{EG}, MA_{ONS}



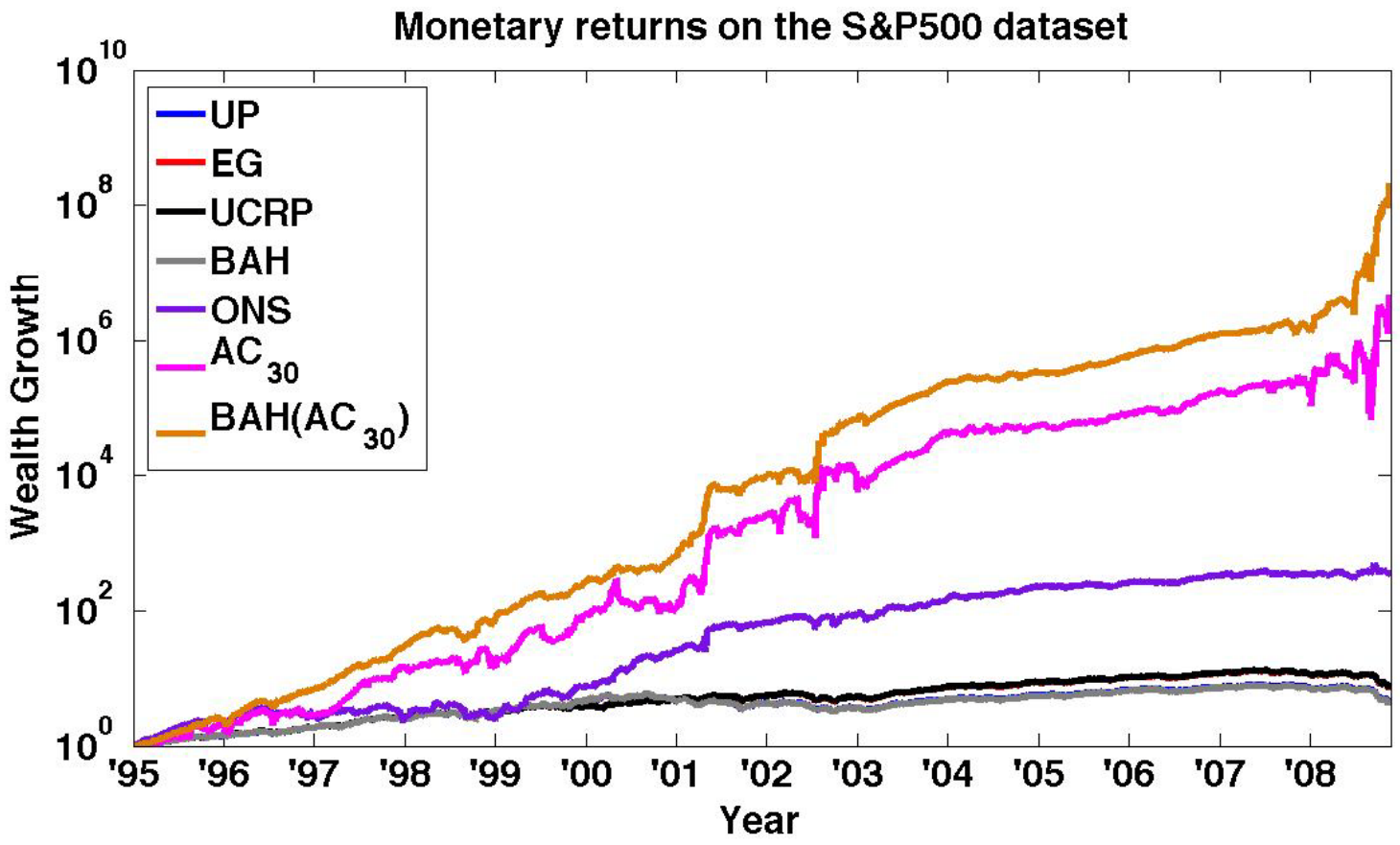
Results: UP, EG, ONS, AC_{30}



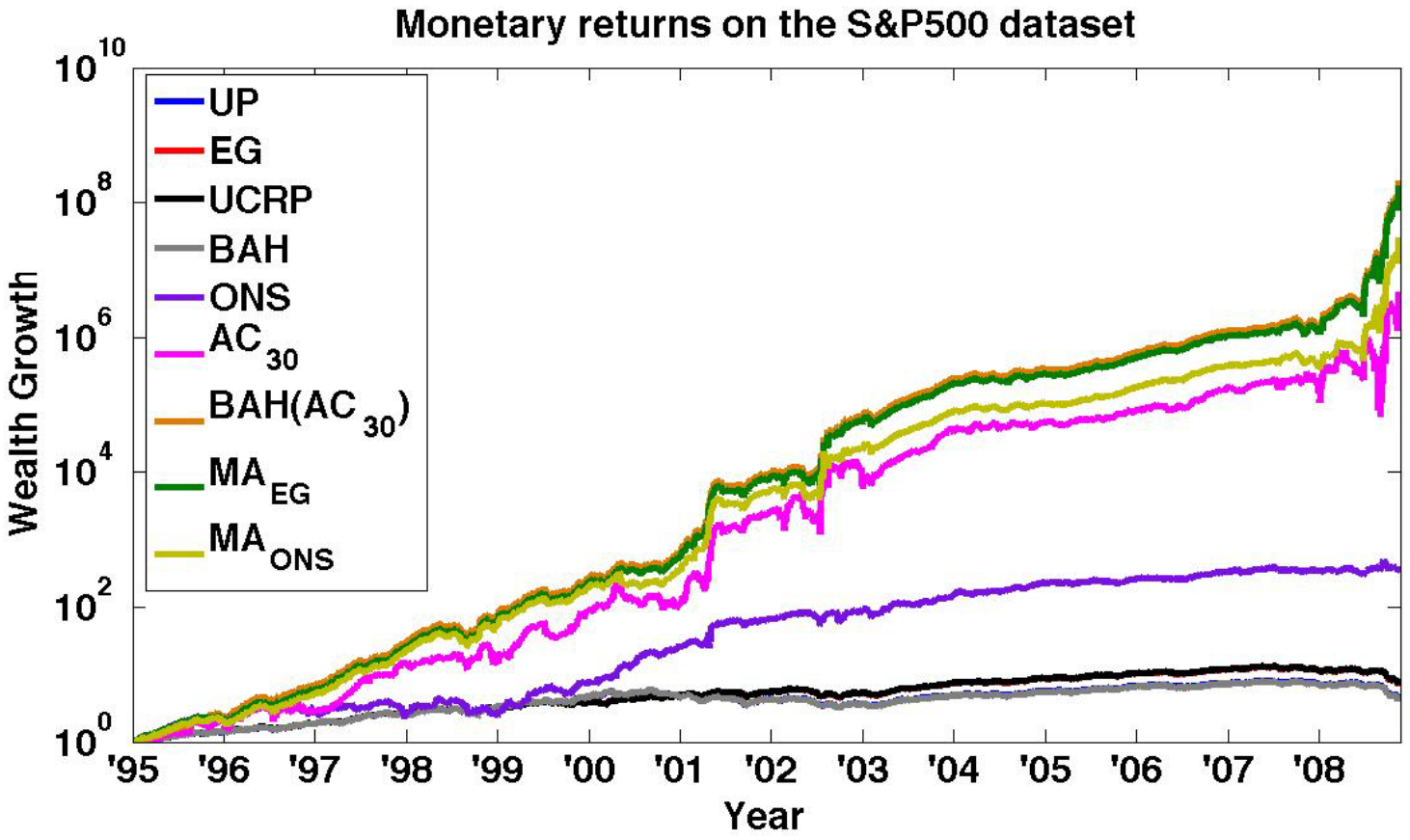
Results: MA_{EG}, MA_{ONS}



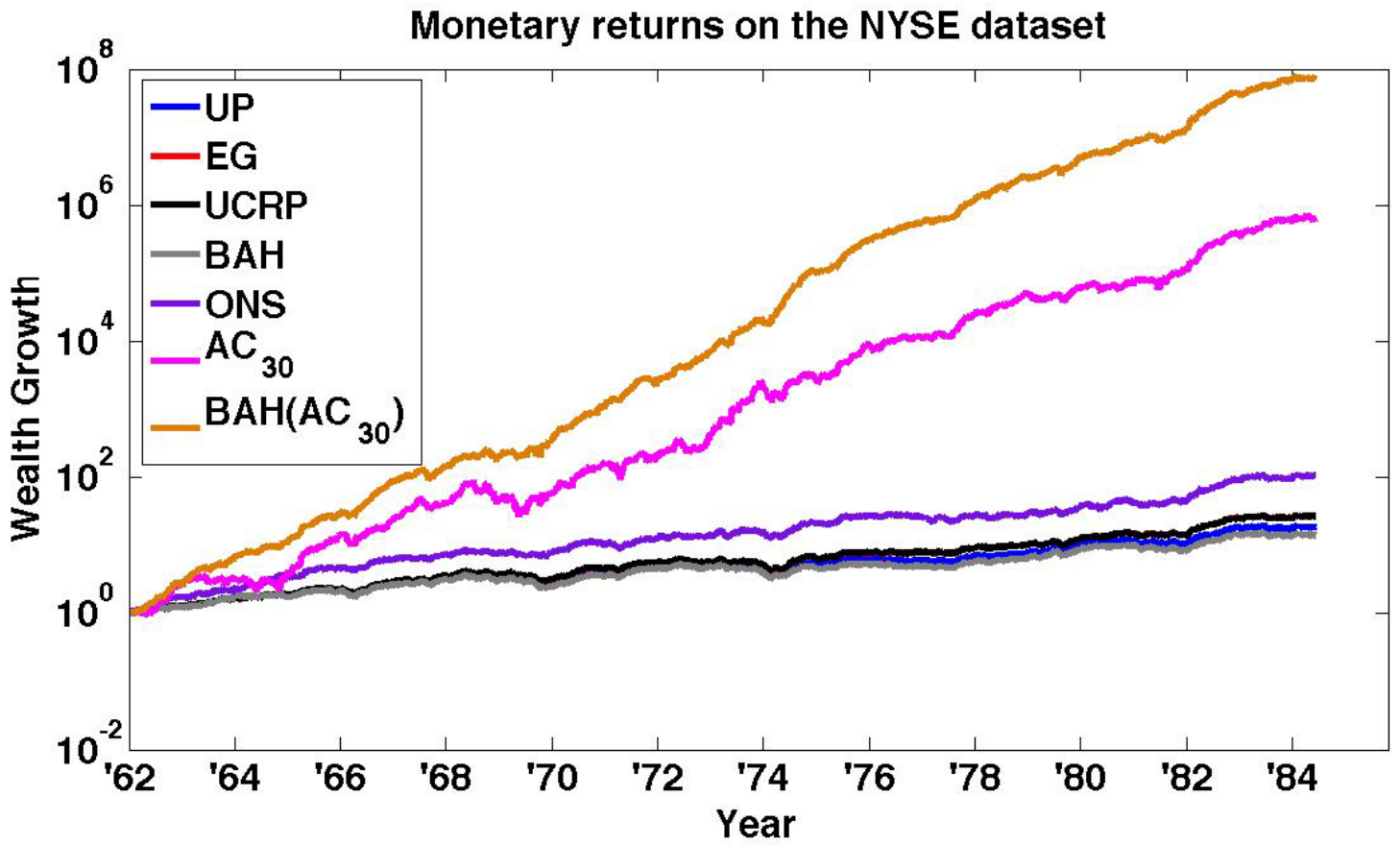
Results: New Base Algorithm BAH(AC_{30})



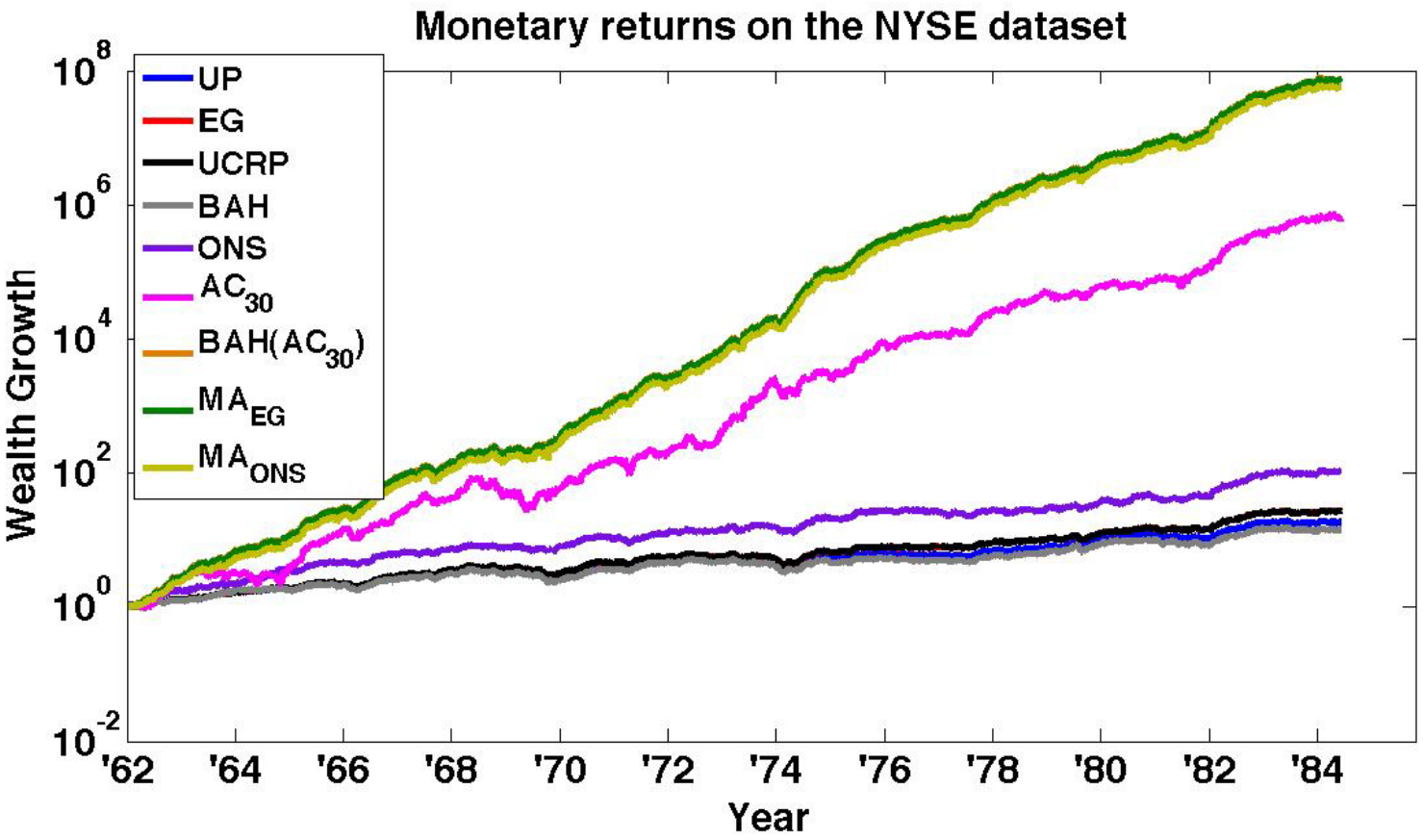
Results: MA_{EG}, MA_{ONS}



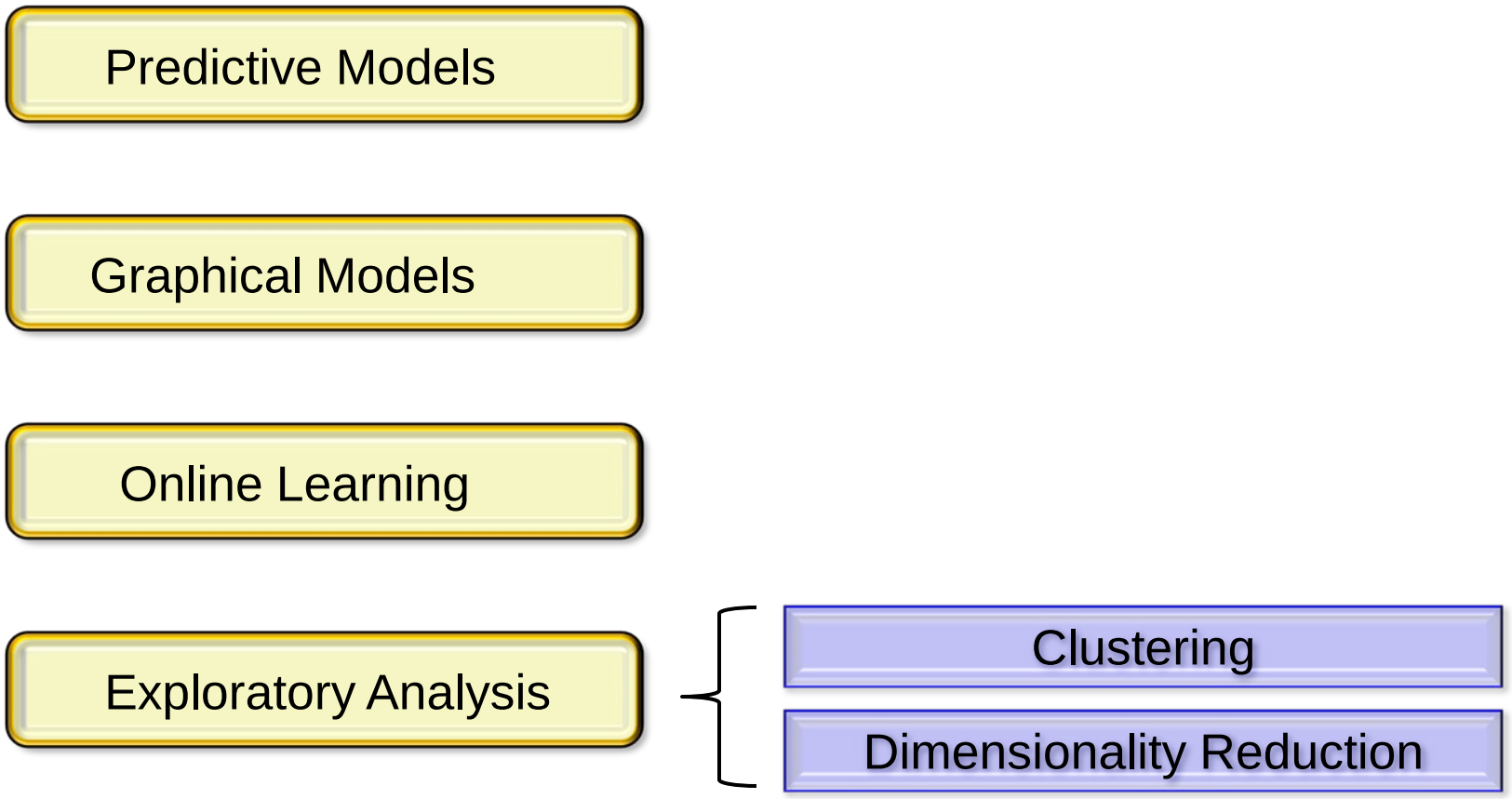
Results: NYSE Base Algorithms



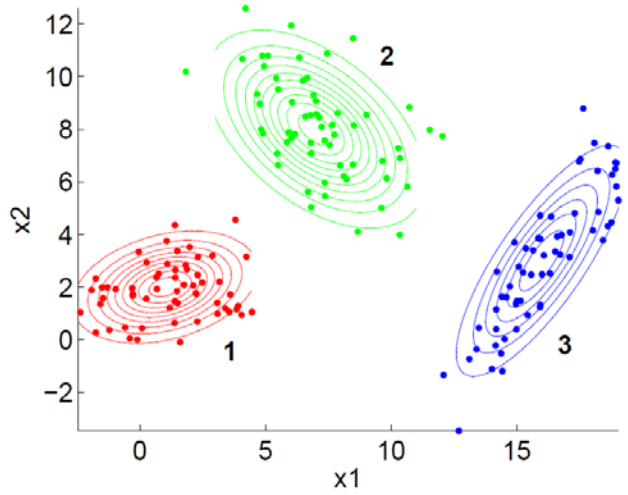
Results: MA_{EG}, MA_{ONS}

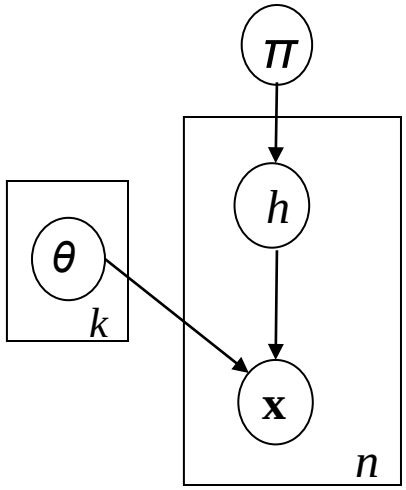


Overview


Predictive Models
Graphical Models
Online Learning
Exploratory Analysis
Clustering
Dimensionality Reduction

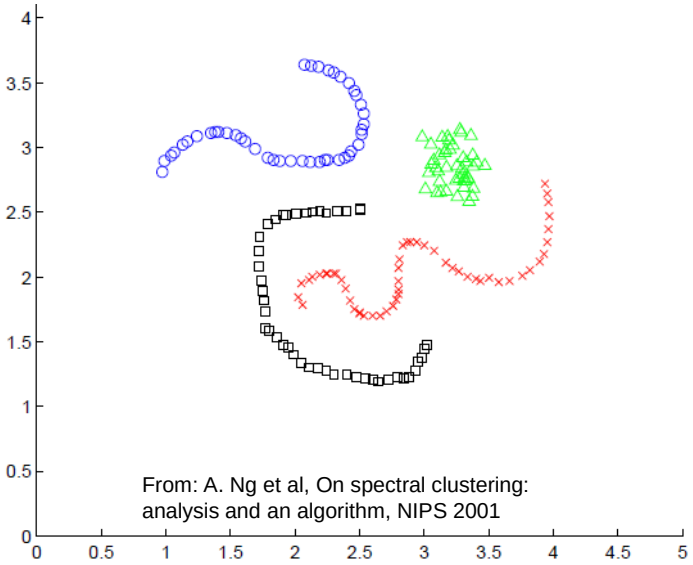
Model-based Clustering





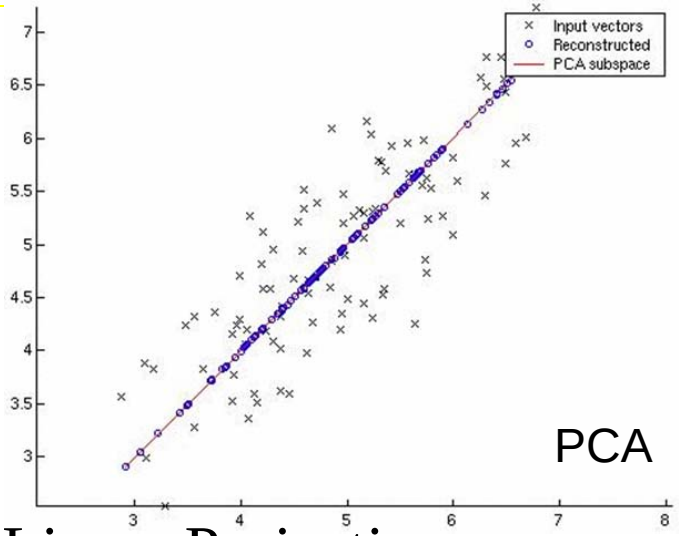
- Mixture Modeling
  - $p(X|\pi, \Theta) = \sum_{h=1}^{k} \pi_h p(\mathbf{x}_i|\theta_h)$
  - Expectation Maximization (EM), or Spectral/Random projections
- Kmeans and Related Approaches
  - $\sum_{h=1}^{k} \sum_{\mathbf{x}_i \in C_h} d(\mathbf{x}_i, \boldsymbol{\mu}_h)$
  - Special (limit) case of mixture modeling
  - Kmeans algorithm: Iteration Relocation

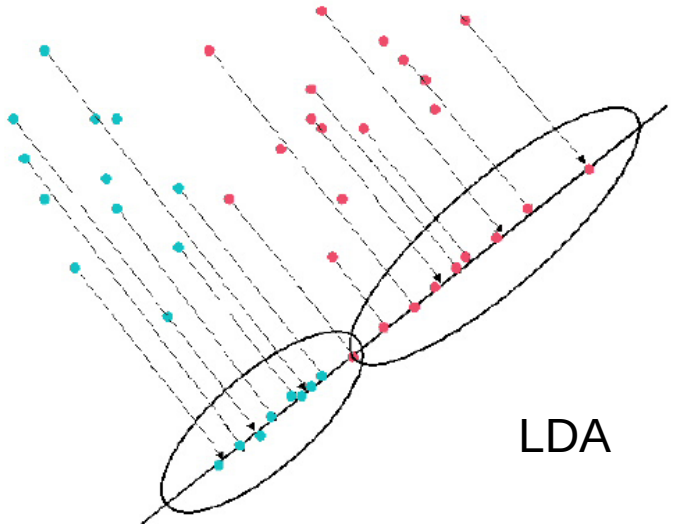
Spectral Clustering



- Weighted Graph $G = (V, E)$, with weights W
- Spectral Clustering
  - Graph Laplacian: $L = I - D^{-1/2} W D^{-1/2}$
  - k-segments $\equiv$ k-eigenvalues are (close to) 0
  - Methods: Spectral Approaches. Kernel K-means

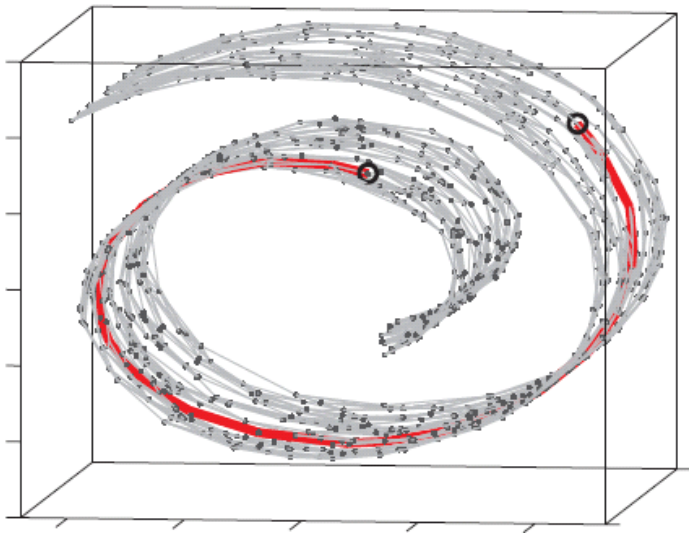
Linear Dimensionality Reduction

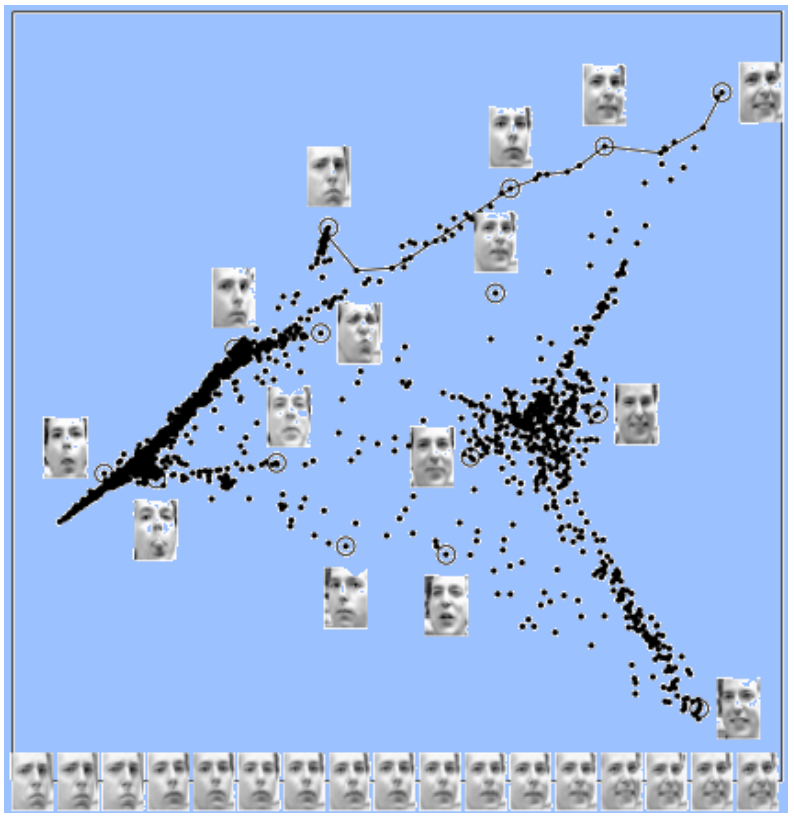




- Linear Projections
  - Projection: $\mathbf{z}_i = W\mathbf{x}_i$, such that $\{\mathbf{z}_i\}$ optimizes some criterion
  - Solve suitable (generalized) eigenvalue problem
- Principal Component Analysis (PCA)
  - Variance of $\{\mathbf{z}_i\}$ is maximized, best hyperplane approximation
- Fisher's Linear Discriminant Analysis (LDA)
  - Separation of classes is maximized

Nonlinear Dimensionality Reduction



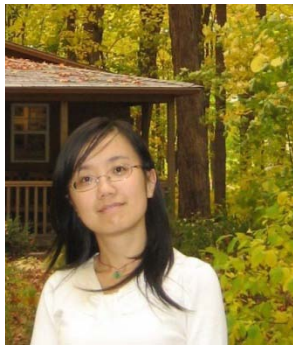


- Manifold Embedding
  - Preserve local geometric structure
- Geometric Approaches:
  - Isomap, Locally Linear Embedding, Laplacian Eigenmaps
- Probabilistic Approaches:
  - Probabilistic PCA, Gaussian Process Latent Variable Models

Summary and Conclusions

- Success of Machine Learning
  - Internet applications, Bioinformatics, Social Network Analysis,…
- Four Themes:
  - Predictive + Graphical Models, Online Learning, Exploratory Analysis
  - Numerous other active and relevant themes. e.g., Sparsity
- Key challenges: Methods
  - High-dimensional problems, nonlinear, nonstationary dependencies
  - Spatial and Temporal effects, Gradual vs Abrupt changes
  - Modeling typical behavior vs extremes/anomalies
- Key challenges: Science/Domain Questions
  - Feynman's 3-step 'algorithm' for problem solving
  - Questions+Expertise+Data from Domain Scientists/Experts

Acknowledgements



Hanhuai Shan

Puja Das

Qiang Fu

Soumyadeep Chatterjee

Amrudin Agovic

Thank you!