Data Mining and Science

Padhraic Smyth
Information and Computer Science
University of California, Irvine

July 2004

SC4DEVO-1 Astronomy Workshop, Caltech

Outline

- Introductory comments on data mining
- Data mining and science
- Hot topics in data mining
- Data mining using probabilistic models
  - Modeling/clustering non-vector data
  - Automatically extracting topics from text documents
- Concluding comments

Technological Driving Factors

- Larger, cheaper memory
  - rapid increase in disk densities
  - storage cost per byte falling rapidly
- Faster, cheaper processors
  - the CRAY of 10 years ago is now on your desk
- Success of Relational Database technology
  - everybody is a “data owner”
- Flexible modeling paradigms
  - generalized linear models, decision trees, etc
  - rise of data-driven, computationally-intensive, statistics

What is data mining?

What is data mining?

“the art of fishing over alternative models ….”

M. C. Lovell,
The Review of Economics and Statistics
February 1983

What is data mining?

"The magic phrase to put in every funding proposal written to NSF, DARPA, NASA, etc"

What is data mining?

“The magic phrase used to sell …….
- database software
- statistical analysis software
- parallel computing hardware
- consulting services”

What is data mining?

"Data-driven discovery of models and patterns from massive observational data sets"

What is data mining?

"Data-driven discovery of models and patterns from massive observational data sets"

Statistics,
Inference

What is data mining?

"Data-driven discovery of models and patterns from massive observational data sets"

Statistics, Inference

Languages and Representations

What is data mining?

"Data-driven discovery of models and patterns from massive observational data sets"

Statistics,
Inference

Languages
and
Representations

Engineering,
Data Management

What is data mining?

"Data-driven discovery of models and patterns from massive observational data sets"

Statistics, Inference

Retrospective Analysis

Languages and Representations

Engineering, Data Management

Implications

- Data mining algorithms span a range of disciplines
  - models/representations: mathematics, probability, CS
  - score functions: statistics
  - search/optimization: numerical methods, OR, AI
  - data management: data structures, databases
  - evaluation: domain knowledge

- Thus,…..
  - A data miner should have some grasp of all of these topics
  - as well as understanding the “art/engineering” of how to integrate all of these components together given a particular data analysis problem
  - (has important implications for education)

Two Types of Data

- Experimental Data
  - Hypothesis H
  - design an experiment to test H
  - collect data, infer how likely it is that H is true
  - e.g., clinical trials in medicine

- Observational or Retrospective or Secondary Data
  - massive non-experimental data sets
    - e.g., human genome, climate data, sky surveys, etc
  - assumptions of experimental design no longer valid
    - e.g., no a priori hypotheses
  - how can we use such data to do science?
    - e.g., use data to simulate experimental conditions

Data-Driven Science

- Assumptions
  - observational data is cheap, experimental data is expensive
  - observational data is massive

- Basic concepts
  - simulate experimental setup
    - random sample for data/model exploration and building
    - random sample for model evaluation
  - data-driven techniques
    - cross-validation, bootstrap, etc
  - finally (important!), conduct an actual experiment to verify final results

Themes in Current Data Mining

- Predictive Modeling
  - e.g., for multivariate data -> predict Y given <u>X</u>
  - Classification, regression, etc
  - well-proven technology, many business applications

- Data Exploration
  - clustering, pattern discovery, dependency models
  - metrics for success are not so clear
  - more suited to interactive exploration and discovery

- "Non-vector data"
  - text, images, etc
  - many innovative ideas, e.g., automated extraction of topics from text documents

- Scalable algorithms
  - General purpose data structures/algorithms for massive data
  - e.g., see Brigham Anderson's talk

Applications of Data Mining

- Business
  - spam email: naïve Bayes, logistic regression classifiers
  - finance: automated credit scoring
  - telecommunications: fraud detection
  - marketing: ranking of customers for catalog mailing
  - Internet advertising: customization of ads during Web browsing

- Sciences
  - Outside of bioinformatics, relatively few clear success stories
  - Why?
    - Scientific data is more complex: time, space
      - existing multivariate DM tools are inadequate
    - Scientific models require more than just prediction
      - interpretability + predictive power

Data Mining: Science vs. Business

- Business applications:
  - Predictive power is most important
  - Interpretability -> not so important
    - “black box” models are ok

- Scientific applications:
  - Both predictive power and interpretability are important
  - Role of data mining algorithm is often to suggest new scientific hypotheses
  - Data mining = “scientific assistant” (rather than being the end goal)

- However…..
  - Historically, data mining has emphasized the business side
    - That’s where most of the funding/profit/jobs are
  - e.g., ACM SIGKDD conference: most papers oriented towards business applications

Hot Topics in Data Mining

- Flexible predictive modeling
  - random forests, boosting, support vector machines, etc
- Engineering of scale
  - scaling up statistics to massive data sets
- Pattern finding
  - discovering associations, rules, bumps, sequential patterns
- Probabilistic modeling and learning
  - Use of hidden variable models
    - mixtures, HMMs, independent components, factors, etc
- “Non-Vector” Data
  - text, Web, multimedia (video/audio), graphs/networks, etc

Topics that are not Hot (but should be!)

- Software environments for data-driven scientific modeling
  - not just off-the-shelf tools for empirical modeling
  - instead:
    - full support for data-driven mechanistic modeling
    - high-level languages for model specification
      - scientist focuses on model structure
      - software takes care of estimation details

- Spatio-temporal data mining
  - richer spatio-temporal data representations and tools
  - e.g., object-level inference versus grid-modeling

- Integration of prior knowledge
  - flexible practical techniques for representing what we know

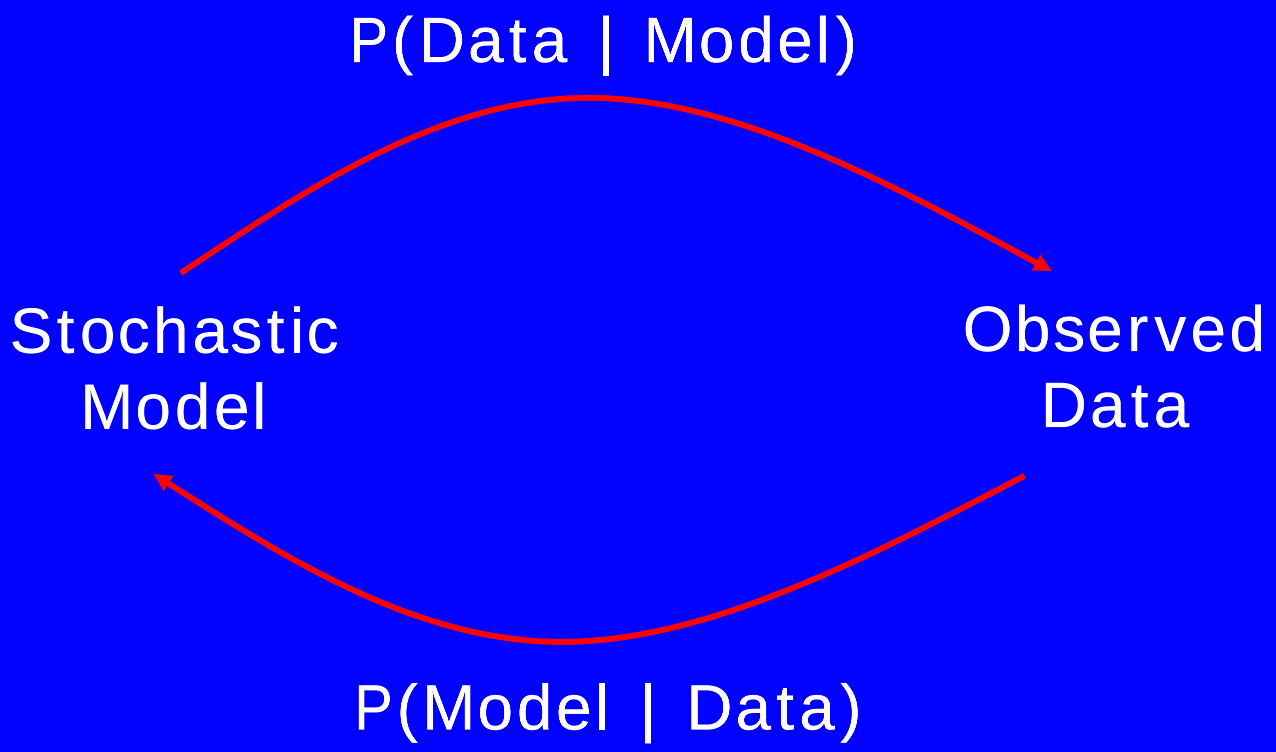
Data Mining using Probabilistic Models

P(Data | Model)

Stochastic Model

Observed Data

P(Model | Data)



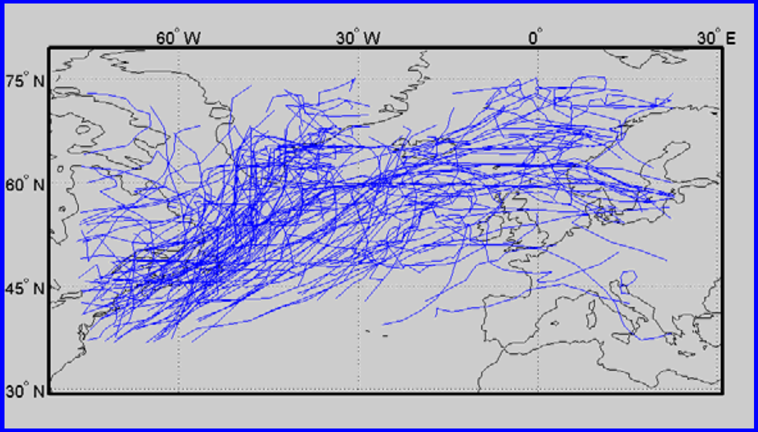
“All models are wrong, but some are useful”
G. E. P. Box

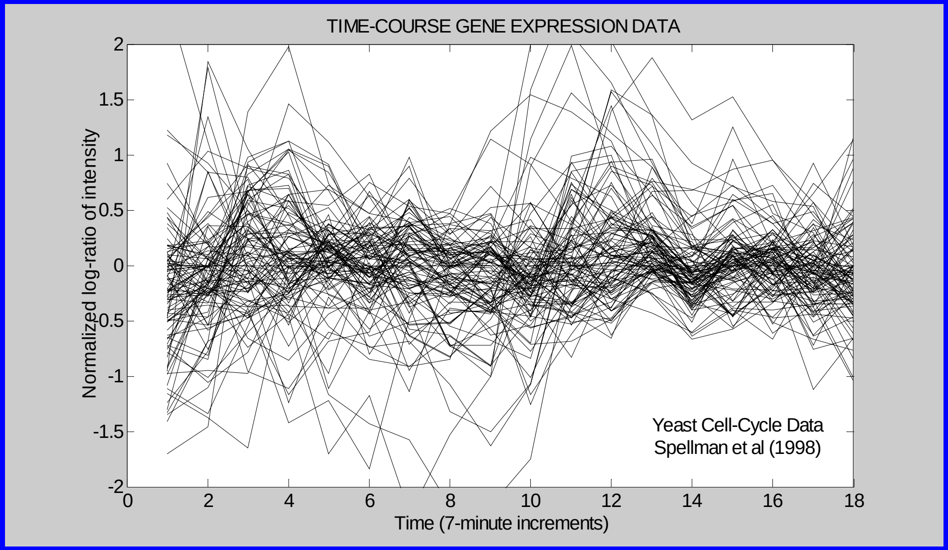
Data Mining with Probabilistic Models

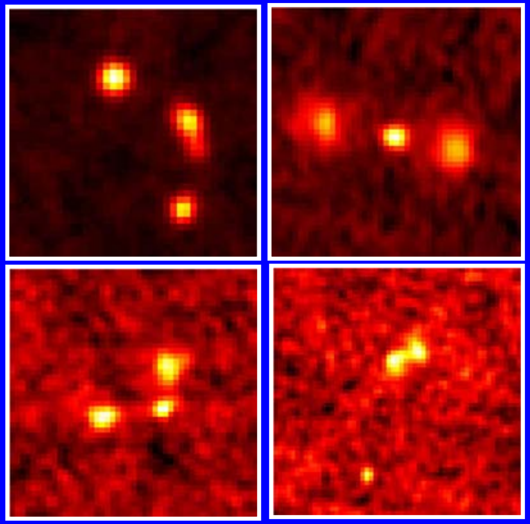
- Advantages
  - Can leverage wealth of ideas from statistical literature
    - Parameter estimation
    - Missing data
    - Hidden variables
  - Very useful for integrating multiple data sources
  - Provides a general and principled language for inference

- Potential disadvantages
  - Traditionally used on small data sets: scalable?
  - Requires explicit model assumptions

“Non-Vector Data”







How can we cluster such data?

(for general modeling of such data see Eric Mjolsness’ talk)

Clustering “non-vector” data

- Challenges with the data….
  - May be of different “lengths”, “sizes”, etc
  - Not easily representable in vector spaces
  - Distance is not naturally defined a priori

- Possible approaches
  - “convert” into a fixed-dimensional vector space
    - Apply standard vector clustering – but loses information
  - use hierarchical clustering
    - But $O(N^2)$ and requires a distance measure
  - probabilistic clustering with mixtures
    - Define a generative mixture model for the data
    - Learn distance and clustering simultaneously

More generally.....

$$p(D_i) = \sum_{k=1}^{K} p(D_i \mid c_k)\,\alpha_k$$

Generative Model

- select a component c_k for individual i
- generate data according to $p(D_i \mid c_k)$
  - $p(D_i \mid c_k)$ can be very general
  - e.g., sets of sequences, spatial patterns, etc

[Note: given $p(D_i \mid c_k)$, we can usually define an EM algorithm for learning]

Mixtures as "Data Simulators"

For i = 1 to N

$\text{class}_i \sim p(\text{class})$

$\mathbf{x}_i \sim p(\mathbf{x} \mid \text{class}_i)$

end

Mixtures with Markov Dependence

For i = 1 to N

$\text{class}_i \sim p(\text{class} \mid \text{class}_{i-1})$

$\mathbf{x}_i \sim p(\mathbf{x} \mid \text{class}_i)$

end

Current class depends on previous class (Markov dependence)

This is a hidden Markov model

Mixtures of Sequences

For i = 1 to N

    $class_i \sim p(class)$

    while non-end state

        $x_{ij} \sim p(x_j \mid x_{j-1}, class_i)$

    end

end

Produces a variable length sequence

Markov sequence model

Mixtures of Curves

For $i = 1$ to N

$class_i \sim p(class)$

$L_i \sim p(L \mid class_i)$

Length of curve

for $j = 1$ to L_i

$y_{ij} \sim f(y \mid x_j, class_i) + e_j$

end

end

Independent variable x

Class-dependent curve model

Mixtures of Spatial Objects

For i = 1 to N

$class_i \sim p(class)$

$scale_i \sim p(scale|class_i)$ ← Global scale

for j = 1 to number of landmarks

$(x,y)_{ij} \sim p(location_j \mid scale_i\ ,\ class_i)$

$features_{ij} \sim p(features_j \mid scale_i, class_i)$

end

end

Prescription for generative modeling…

1. Forward modeling (probability):
   construct a generative probabilistic model that could generate the data of interest

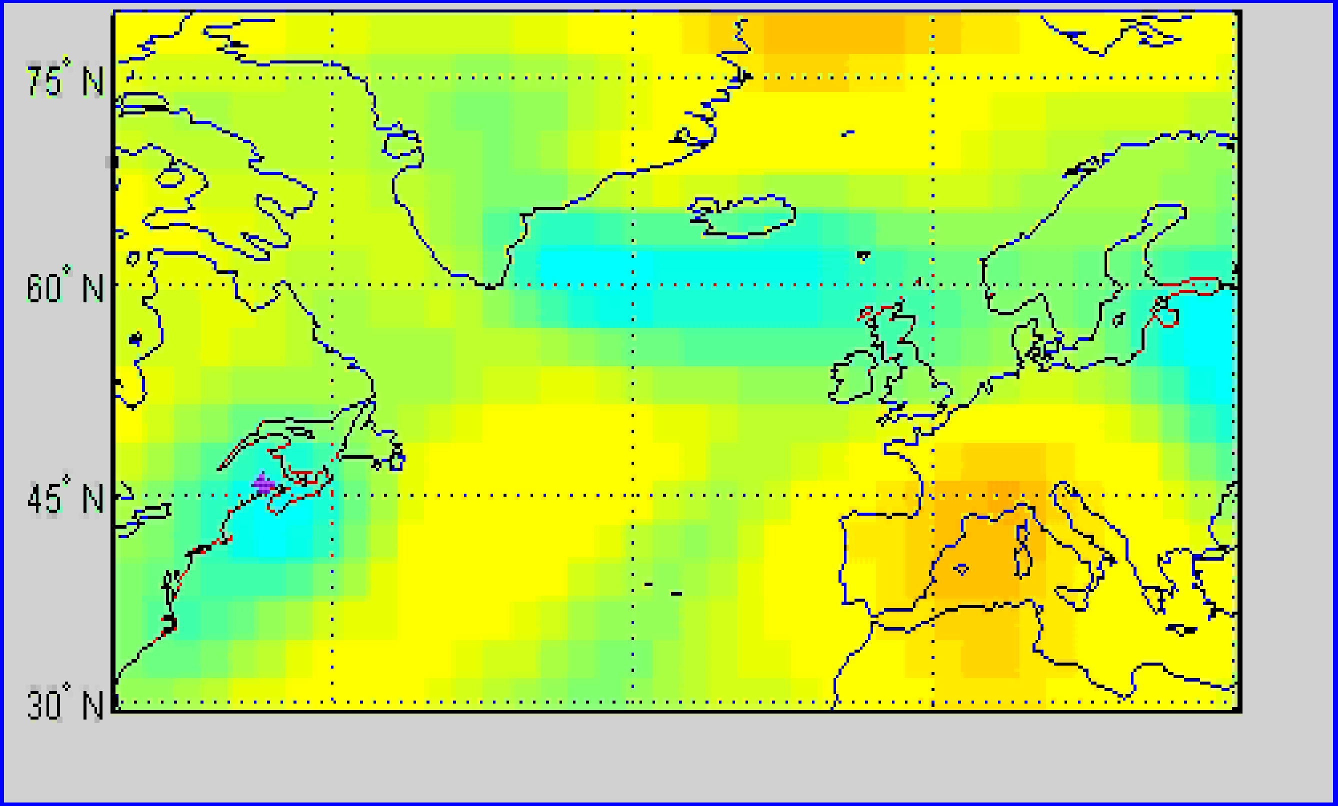
2. Inverse inference (learning, statistics):
   given observed data, now infer the parameters of our model (e.g., using EM)

Clustering of Cyclone Trajectories

[with Scott Gaffney (UCI), Andy Robertson (IRI/Columbia), Michael Ghil (UCLA)]

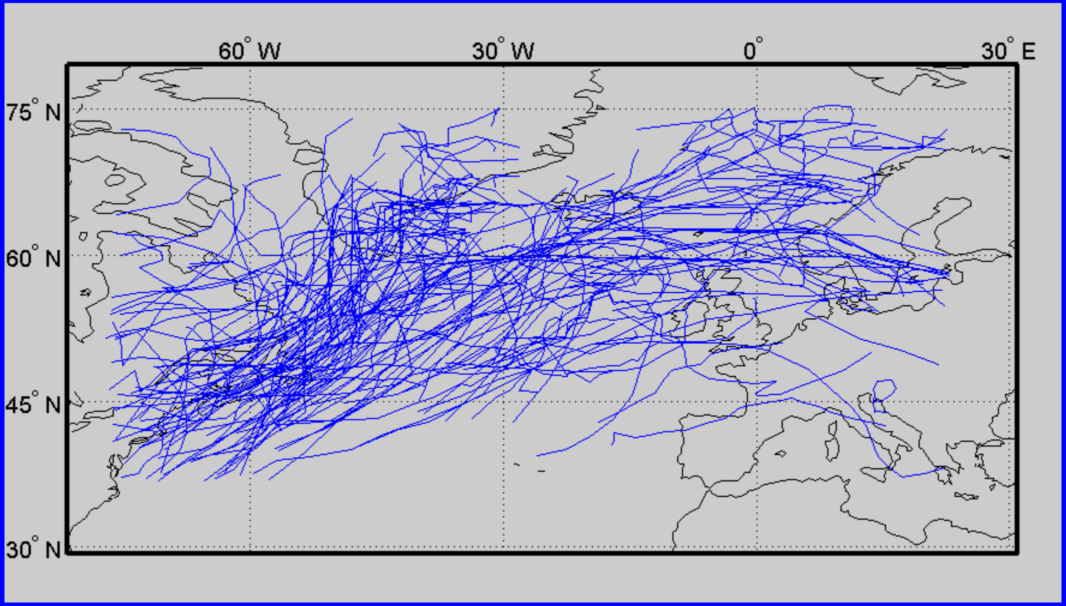
Data

- Sea-level pressure on a global grid
- Four times a day, every 6 hours, over 20 to 30 years



Blue = low pressure

Extra-Tropical Cyclones



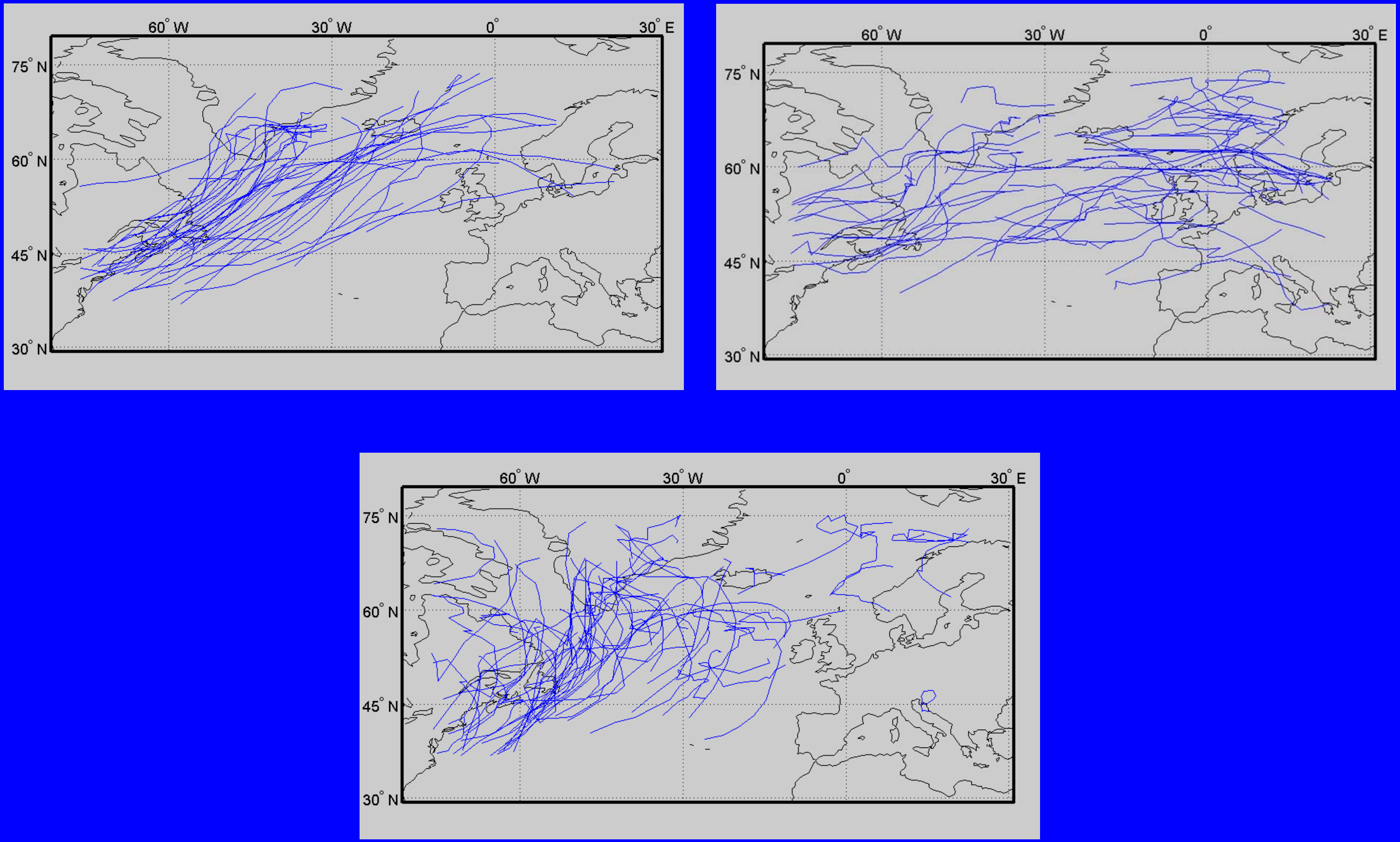
- Importance
  - Highly damaging weather over Europe
  - Important water-source in Western US
  - Influence of climate on cyclone frequency, strength, etc.
  - Impact of cyclones on local weather patterns

Clustering Methodology

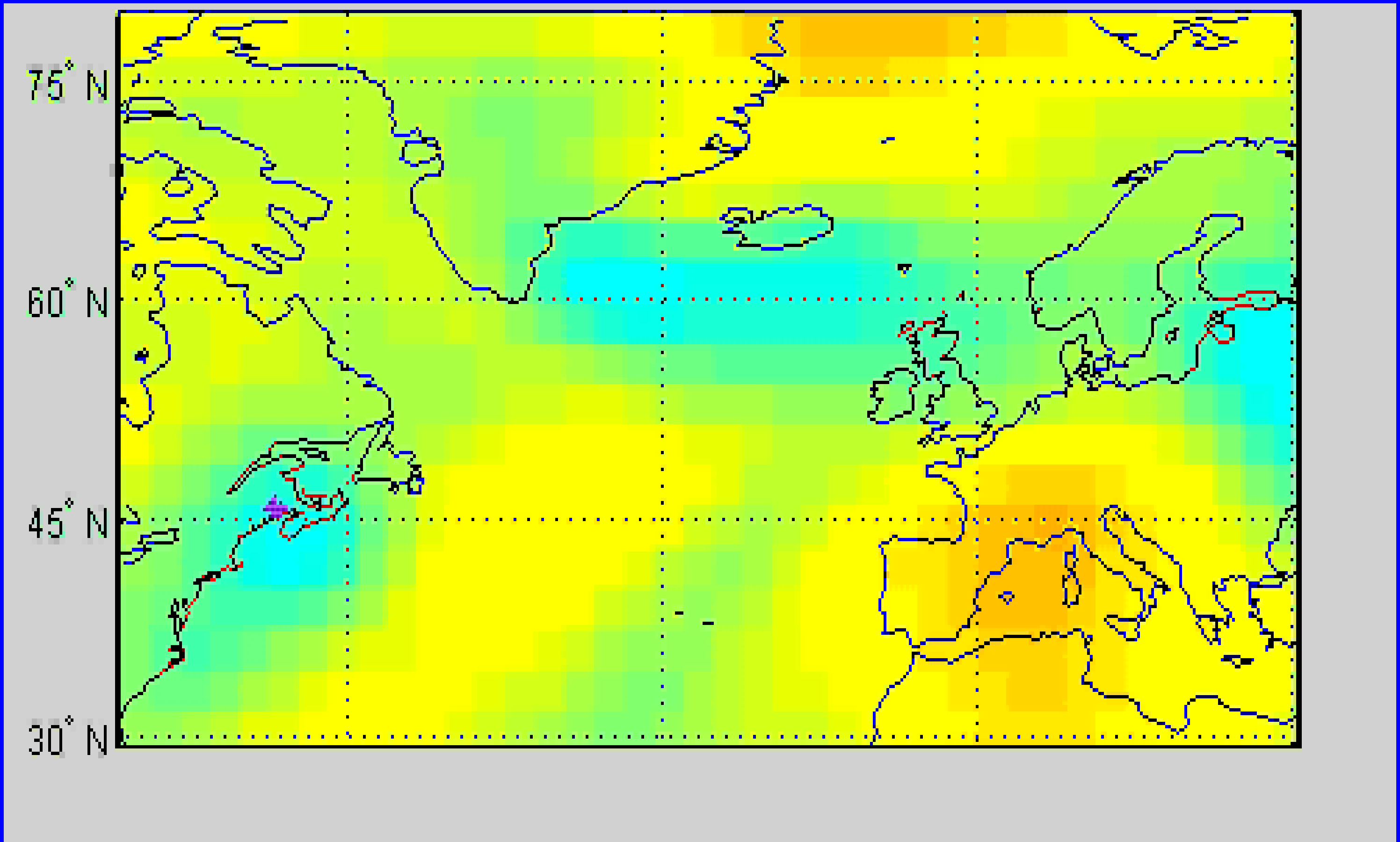
- Mixtures of polynomials
  - model as mixtures of noisy regression models
  - 2d (x,y) position as a function of time
    - $x_k(t) = a_k + b_k t + c_k t^2$
    - could also use AR or state-space models
  - use the model as a first-order approximation for clustering

- Compare to vector-based clustering...
  - allows for variable-length trajectories
  - allows coupling of other “features” (e.g., intensity)
  - provides a quantitative (e.g., predictive) model
  - can handle missing measurements

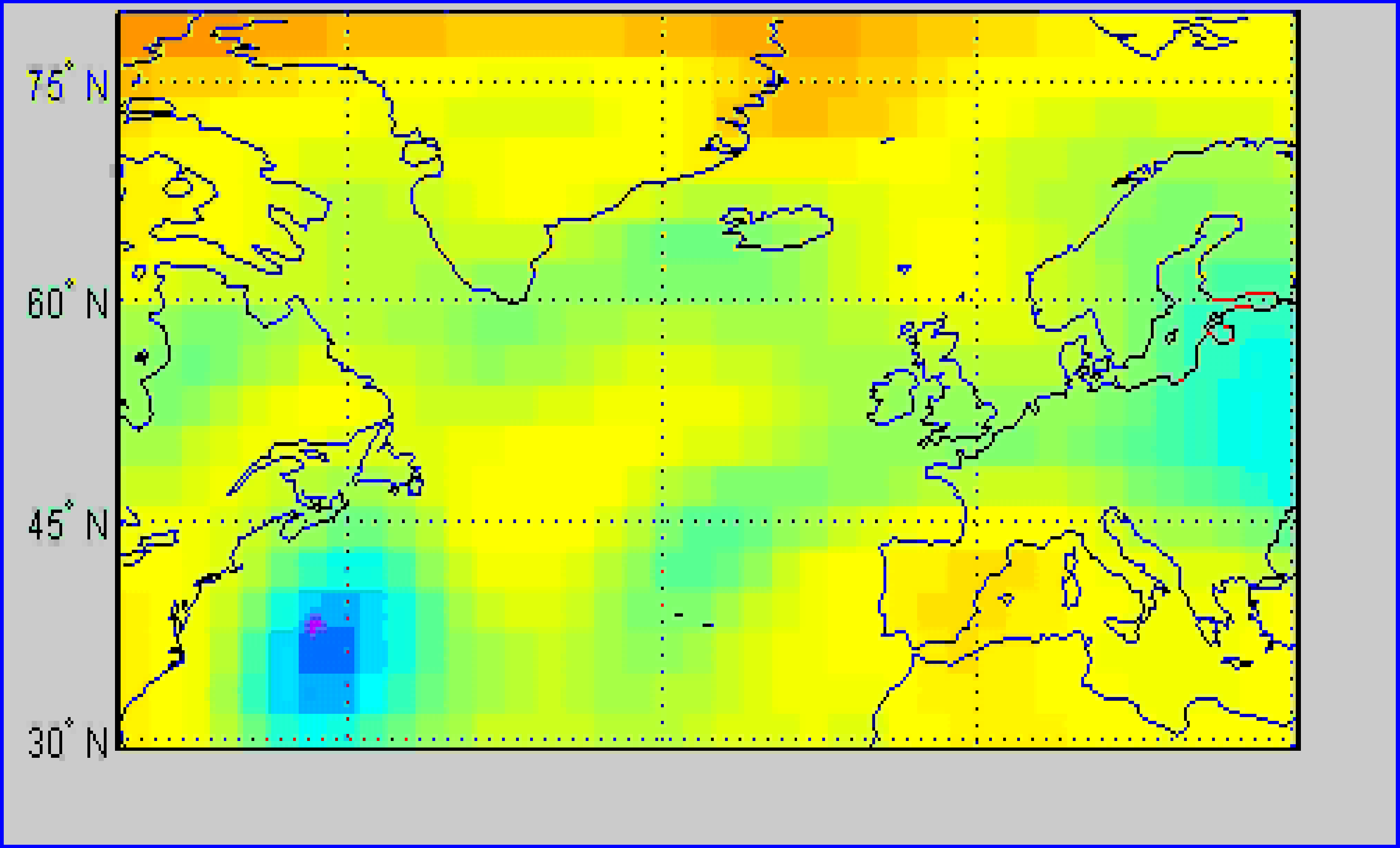
Clusters of Trajectories



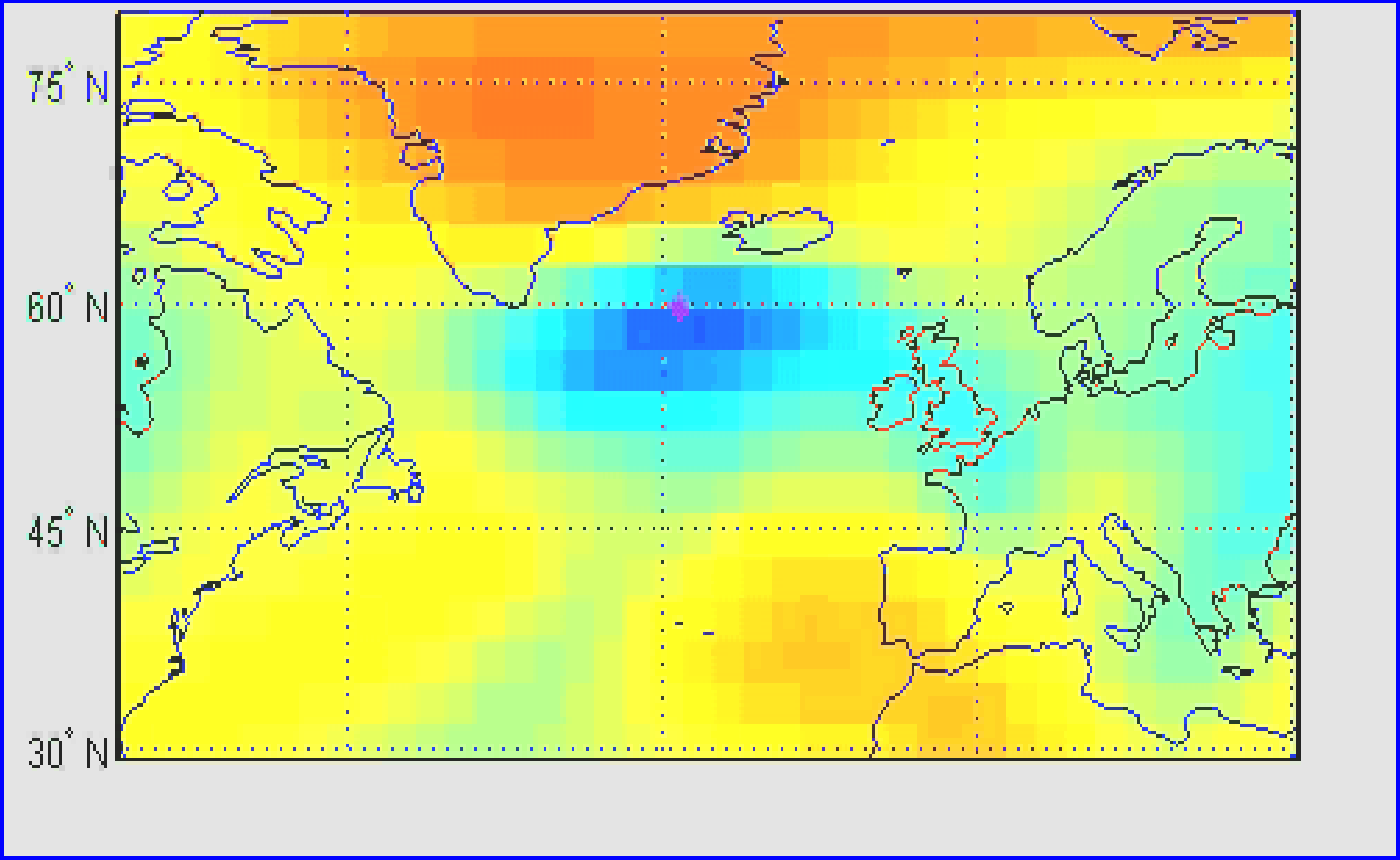
“Iceland Cluster”



“Northern Europe Cluster”



"Greenland Cluster"



Why is this useful to the scientists?

- Visualization and Exploration
  - improved understanding of cyclone dynamics
- Change Detection
  - can quantitatively compare cyclone statistics over different era’s or from different models
- Linking cyclones with climate and weather
  - correlation of clusters with NAO index
  - correlation with windspeeds in Northern Europe

Extensions

- More flexible curve models
  - mixtures of splines
  - random effects/hierarchical Bayes
  - mixtures of dynamical systems

- Other additions
  - background models for noisy trajectories
  - random shifts/offsets
  - coupling of other features: intensity, vorticity

Generalizations to Other Problems

- Mixtures of Markov chains
  - used to cluster variable-length categorical sequences
  - Clustering and visualization of Web users at msnbc.com
    - Cadez et al, 2000 and 2003
    - Algorithm is part of latest version of SQLServer

- Mixtures of spatial image patches
  - Used to cluster and align images of “double-bent galaxies”
    - Kirshner et al, 2002, 2003

- Mixtures of synthesis-decay equations
  - Used for clustering time-course gene expression data
  - Chudova, Mjolsness, Smyth, 2003

Statistical Data Mining of Text Data

[with Michal Rosen-Zvi, Mark Steyvers (UCI), Thomas Griffiths (Stanford)]

Statistical Data Mining of Text Documents

- Probabilistic models for text
  - Represent each document as a vector of word counts
  - A topic is a probability distribution on words
  - Documents are generated stochastically by mixtures of topics
    - Multiple topics can be active on a single document

- Forward model:
  - Probabilistic model with hidden/unknown topic variables

- Inverse Learning:
  - Can learn topic models in a completely unsupervised manner
    - e.g., using EM or Gibbs sampling

A topic is represented as a (multinomial) distribution over words

TOPIC 209	
WORD	**PROB.**
PROBABILISTIC	0.0778
BAYESIAN	0.0671
PROBABILITY	0.0532
CARLO	0.0309
MONTE	0.0308
DISTRIBUTION	0.0257
INFERENCE	0.0253
PROBABILITIES	0.0253
CONDITIONAL	0.0229
PRIOR	0.0219
....	...

TOPIC 289	
WORD	**PROB.**
RETRIEVAL	0.1179
TEXT	0.0853
DOCUMENTS	0.0527
INFORMATION	0.0504
DOCUMENT	0.0441
CONTENT	0.0242
INDEXING	0.0205
RELEVANCE	0.0159
COLLECTION	0.0146
RELEVANT	0.0136
...	...

$P(w \mid z)$

Data and Experiments

- Text Corpora
  - CiteSeer: 160K abstracts, 85K authors, 20 million word tokens
  - NIPS: 1.7K papers, 2K authors
  - Enron: 115K emails, 5K authors (sender)

- Removed stop words; no stemming

- Word order is irrelevant, just use word counts

- Learning the model:

  Nips: 2000 Gibbs iterations → 12 hours on PC workstation

  CiteSeer: 2000 Gibbs iterations → 1 week

  But querying the model (once learned) can be done in real-time

4 Examples of CiteSeer Topics (300 in total)

TOPIC 205	
WORD	**PROB.**
DATA	0.1563
MINING	0.0674
ATTRIBUTES	0.0462
DISCOVERY	0.0401
ASSOCIATION	0.0335
LARGE	0.0280
KNOWLEDGE	0.0260
DATABASES	0.0210
ATTRIBUTE	0.0188
DATASETS	0.0165
AUTHOR	**PROB.**
Han_J	0.0196
Rastogi_R	0.0094
Zaki_M	0.0084
Shim_K	0.0077
Ng_R	0.0060
Liu_B	0.0058
Mannila_H	0.0056
Brin_S	0.0054
Liu_H	0.0047
Holder_L	0.0044

TOPIC 209	
WORD	**PROB.**
PROBABILISTIC	0.0778
BAYESIAN	0.0671
PROBABILITY	0.0532
CARLO	0.0309
MONTE	0.0308
DISTRIBUTION	0.0257
INFERENCE	0.0253
PROBABILITIES	0.0253
CONDITIONAL	0.0229
PRIOR	0.0219
AUTHOR	**PROB.**
Friedman_N	0.0094
Heckerman_D	0.0067
Ghahramani_Z	0.0062
Koller_D	0.0062
Jordan_M	0.0059
Neal_R	0.0055
Raftery_A	0.0054
Lukasiewicz_T	0.0053
Halpern_J	0.0052
Muller_P	0.0048

TOPIC 289	
WORD	**PROB.**
RETRIEVAL	0.1179
TEXT	0.0853
DOCUMENTS	0.0527
INFORMATION	0.0504
DOCUMENT	0.0441
CONTENT	0.0242
INDEXING	0.0205
RELEVANCE	0.0159
COLLECTION	0.0146
RELEVANT	0.0136
AUTHOR	**PROB.**
Oard_D	0.0110
Croft_W	0.0056
Jones_K	0.0053
Schauble_P	0.0051
Voorhees_E	0.0050
Singhal_A	0.0048
Hawking_D	0.0048
Merkl_D	0.0042
Allan_J	0.0040
Doermann_D	0.0039

TOPIC 10	
WORD	**PROB.**
QUERY	0.1848
QUERIES	0.1367
INDEX	0.0488
DATA	0.0368
JOIN	0.0260
INDEXING	0.0180
PROCESSING	0.0113
AGGREGATE	0.0110
ACCESS	0.0102
PRESENT	0.0095
AUTHOR	**PROB.**
Suciu_D	0.0102
Naughton_J	0.0095
Levy_A	0.0071
DeWitt_D	0.0068
Wong_L	0.0067
Chakrabarti_K	0.0064
Ross_K	0.0061
Hellerstein_J	0.0059
Lenzerini_M	0.0054
Moerkotte_G	0.0053

More example topics from CiteSeer

TOPIC 10	
WORD	**PROB.**
SPEECH	0.1134
RECOGNITION	0.0349
WORD	0.0295
SPEAKER	0.0227
ACOUSTIC	0.0205
RATE	0.0134
SPOKEN	0.0132
SOUND	0.0127
TRAINING	0.0104
MUSIC	0.0102
AUTHOR	**PROB.**
Waibel_A	0.0156
Gauvain_J	0.0133
Lamel_L	0.0128
Woodland_P	0.0124
Ney_H	0.0080
Hansen_J	0.0078
Renals_S	0.0072
Noth_E	0.0071
Boves_L	0.0070
Young_S	0.0069

TOPIC 209	
WORD	**PROB.**
PROBABILISTIC	0.0778
BAYESIAN	0.0671
PROBABILITY	0.0532
CARLO	0.0309
MONTE	0.0308
DISTRIBUTION	0.0257
INFERENCE	0.0253
PROBABILITIES	0.0253
CONDITIONAL	0.0229
PRIOR	0.0219
AUTHOR	**PROB.**
Friedman_N	0.0094
Heckerman_D	0.0067
Ghahramani_Z	0.0062
Koller_D	0.0062
Jordan_M	0.0059
Neal_R	0.0055
Raftery_A	0.0054
Lukasiewicz_T	0.0053
Halpern_J	0.0052
Muller_P	0.0048

TOPIC 87	
WORD	**PROB.**
USER	0.2541
INTERFACE	0.1080
USERS	0.0788
INTERFACES	0.0433
GRAPHICAL	0.0392
INTERACTIVE	0.0354
INTERACTION	0.0261
VISUAL	0.0203
DISPLAY	0.0128
MANIPULATION	0.0099
AUTHOR	**PROB.**
Shneiderman_B	0.0060
Rauterberg_M	0.0031
Lavana_H	0.0024
Pentland_A	0.0021
Myers_B	0.0021
Minas_M	0.0021
Burnett_M	0.0021
Winiwarter_W	0.0020
Chang_S	0.0019
Korvemaker_B	0.0019

TOPIC 20	
WORD	**PROB.**
STARS	0.0164
OBSERVATIONS	0.0150
SOLAR	0.0150
MAGNETIC	0.0145
RAY	0.0144
EMISSION	0.0134
GALAXIES	0.0124
OBSERVED	0.0108
SUBJECT	0.0101
STAR	0.0087
AUTHOR	**PROB.**
Linsky_J	0.0143
Falcke_H	0.0131
Mursula_K	0.0089
Butler_R	0.0083
Bjorkman_K	0.0078
Knapp_G	0.0067
Kundu_M	0.0063
Christensen-J	0.0059
Cranmer_S	0.0055
Nagar_N	0.0050

4 Examples of NIPS Topics (100 in total)

TOPIC 19	
WORD	**PROB.**
LIKELIHOOD	0.0539
MIXTURE	0.0509
EM	0.0470
DENSITY	0.0398
GAUSSIAN	0.0349
ESTIMATION	0.0314
LOG	0.0263
MAXIMUM	0.0254
PARAMETERS	0.0209
ESTIMATE	0.0204
AUTHOR	**PROB.**
Tresp_V	0.0333
Singer_Y	0.0281
Jebara_T	0.0207
Ghahramani_Z	0.0196
Ueda_N	0.0170
Jordan_M	0.0150
Roweis_S	0.0123
Schuster_M	0.0104
Xu_L	0.0098
Saul_L	0.0094

TOPIC 24	
WORD	**PROB.**
RECOGNITION	0.0400
CHARACTER	0.0336
CHARACTERS	0.0250
TANGENT	0.0241
HANDWRITTEN	0.0169
DIGITS	0.0159
IMAGE	0.0157
DISTANCE	0.0153
DIGIT	0.0149
HAND	0.0126
AUTHOR	**PROB.**
Simard_P	0.0694
Martin_G	0.0394
LeCun_Y	0.0359
Denker_J	0.0278
Henderson_D	0.0256
Revow_M	0.0229
Platt_J	0.0226
Keeler_J	0.0192
Rashid_M	0.0182
Sackinger_E	0.0132

TOPIC 29	
WORD	**PROB.**
REINFORCEMENT	0.0411
POLICY	0.0371
ACTION	0.0332
OPTIMAL	0.0208
ACTIONS	0.0208
FUNCTION	0.0178
REWARD	0.0165
SUTTON	0.0164
AGENT	0.0136
DECISION	0.0118
AUTHOR	**PROB.**
Singh_S	0.1412
Barto_A	0.0471
Sutton_R	0.0430
Dayan_P	0.0324
Parr_R	0.0314
Dietterich_T	0.0231
Tsitsiklis_J	0.0194
Randlov_J	0.0167
Bradtke_S	0.0161
Schwartz_A	0.0142

TOPIC 87	
WORD	**PROB.**
KERNEL	0.0683
SUPPORT	0.0377
VECTOR	0.0257
KERNELS	0.0217
SET	0.0205
SVM	0.0204
SPACE	0.0188
MACHINES	0.0168
REGRESSION	0.0155
MARGIN	0.0151
AUTHOR	**PROB.**
Smola_A	0.1033
Scholkopf_B	0.0730
Burges_C	0.0489
Vapnik_V	0.0431
Chapelle_O	0.0210
Cristianini_N	0.0185
Ratsch_G	0.0172
Laskov_P	0.0169
Tipping_M	0.0153
Sollich_P	0.0141

ENRON Email: two example topics (T=100)

TOPIC 10	
WORD	**PROB.**
BUSH	0.0227
LAY	0.0193
MR	0.0183
WHITE	0.0153
ENRON	0.0150
HOUSE	0.0148
PRESIDENT	0.0131
ADMINISTRATION	0.0115
COMPANY	0.0090
ENERGY	0.0085
SENDER	**PROB.**
NELSON, KIMBERLY (ETS)	0.3608
PALMER, SARAH	0.0997
DENNE, KAREN	0.0541
HOTTE, STEVE	0.0340
DUPREE, DIANNA	0.0282
ARMSTRONG, JULIE	0.0222
LOKEY, TEB	0.0194
SULLIVAN, LORA	0.0073
VILLARREAL, LILLIAN	0.0040
BAGOT, NANCY	0.0026

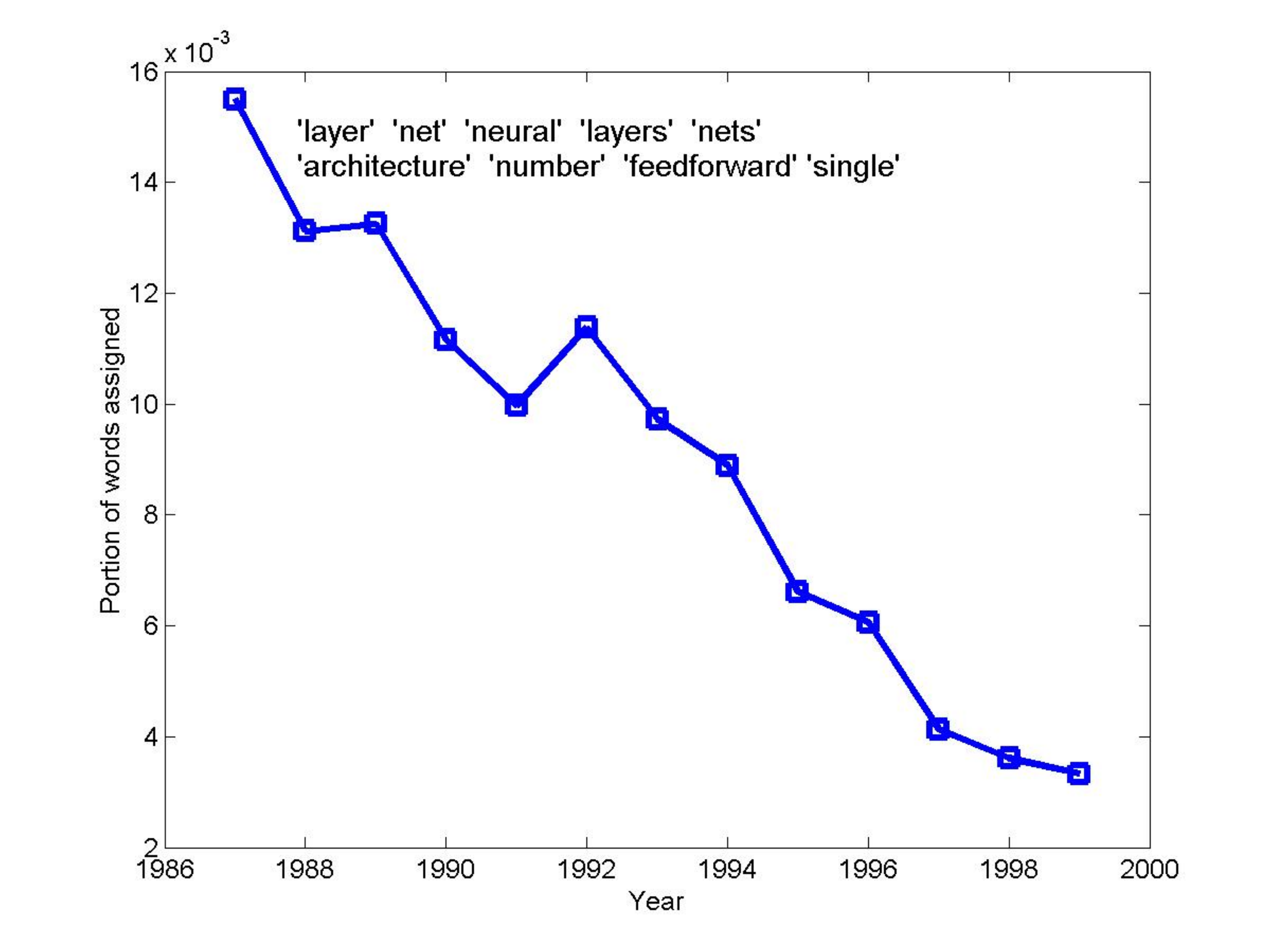
TOPIC 32	
WORD	**PROB.**
ANDERSEN	0.0241
FIRM	0.0134
ACCOUNTING	0.0119
SEC	0.0065
SETTLEMENT	0.0062
AUDIT	0.0054
CORPORATE	0.0053
FINANCIAL	0.0052
JUSTICE	0.0052
INFORMATION	0.0050
SENDER	**PROB.**
HILTABRAND, LESLIE	0.1359
WELLS, TORI L.	0.0865
DUPREE, DIANNA	0.0825
ARMSTRONG, JULIE	0.0316
DENNE, KAREN	0.0208
SULLIVAN, LORA	0.0072
N..M.SZAFRANSKI@US.ANDERSEN.COM	0.0026
WILSON, DANNY	0.0016
HU, SYLVIA	0.0013
MATHEWS, LEENA	0.0012

ENRON Email: two topics not about Enron

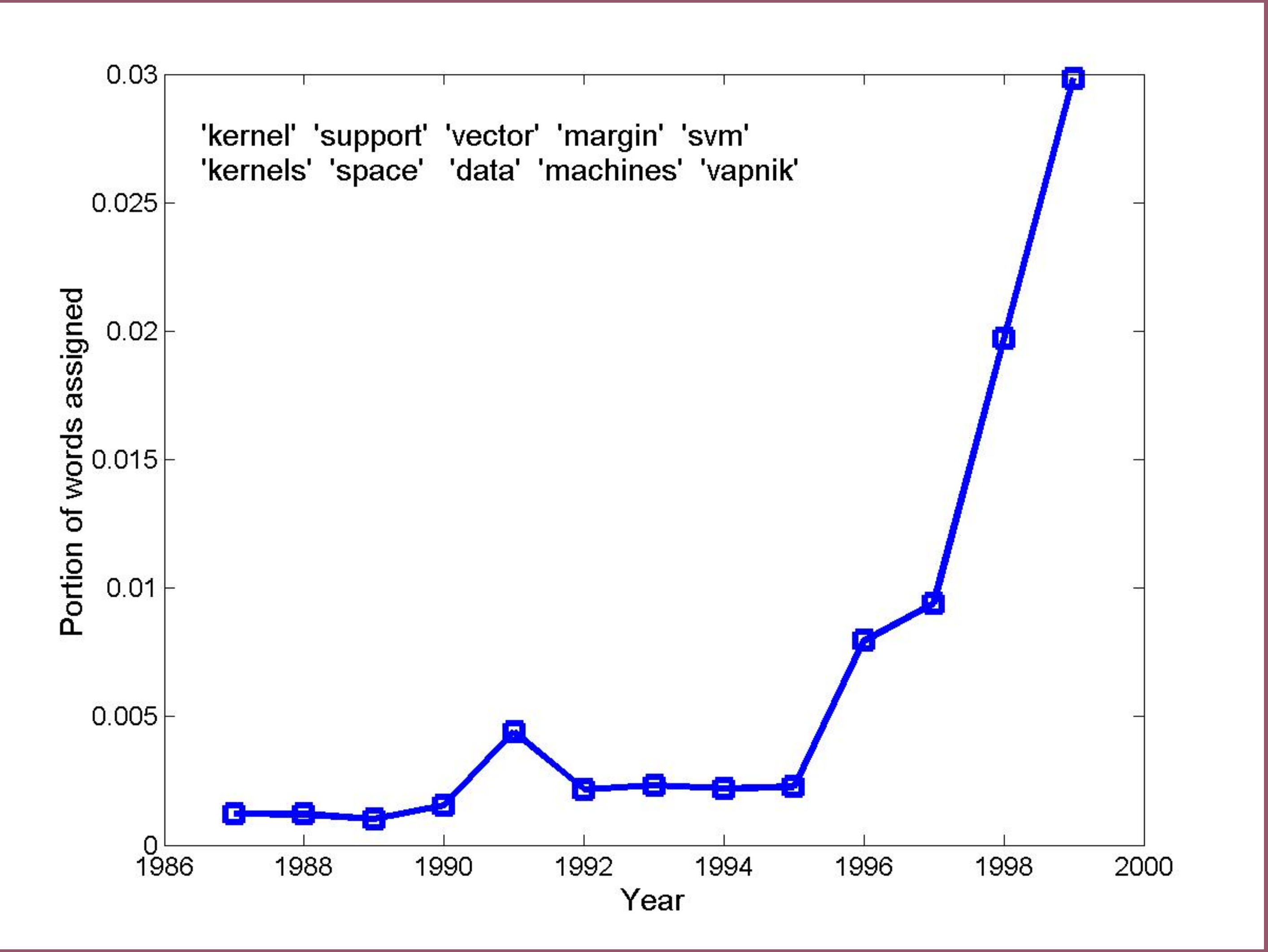
TOPIC 38	
WORD	**PROB.**
TRAVEL	0.0161
ROUNDTRIP	0.0124
SAVE	0.0118
DEALS	0.0097
HOTEL	0.0095
BOOK	0.0094
SALE	0.0089
FARES	0.0083
TRIP	0.0072
CITIES	0.0070
SENDER	**PROB.**
TRAVELOCITY MEMBER SERVICES	0.0763
BESTFARES.COM HOT DEALS	0.0502
<DEALS@BESTFARES.COM>	0.0315
LISTS.COOLVACATIONS.COM	0.0151
CHEAP TICKETS	0.0111
EXPEDIA FARE TRACKER	0.0106
TRAVELOCITY.COM	0.0096
HOTDEALS@MAIL.HOTELRESNETWORK.COM	0.0088
LUCKY@ICELANDAIR.IS	0.0066
LASTMINUTE.COM	0.0051

TOPIC 25	
WORD	**PROB.**
NEWS	0.0245
MAIL	0.0182
NYTIMES	0.0149
YORK	0.0128
PAGE	0.0095
TIMES	0.0090
HEADLINES	0.0079
BUSH	0.0077
DELIVERY	0.0070
HTML	0.0068
SENDER	**PROB.**
THE NEW YORK TIMES DIRECT	0.3438
<NYTDIRECT@NYTIMES.COM>	0.0104
THE ECONOMIST	0.0029
@TIMES - INSIDE NYTIMES.COM	0.0015
JHILLIN@ENRON.COM	0.0011
AMAZON.COM DELIVERS BESTSELLERS	0.0009
NYTIMES.COM	0.0009
HYATT, JERRY	0.0008
NEWSLETTER_TEXT	0.0008
CHRIS LONG	0.0007

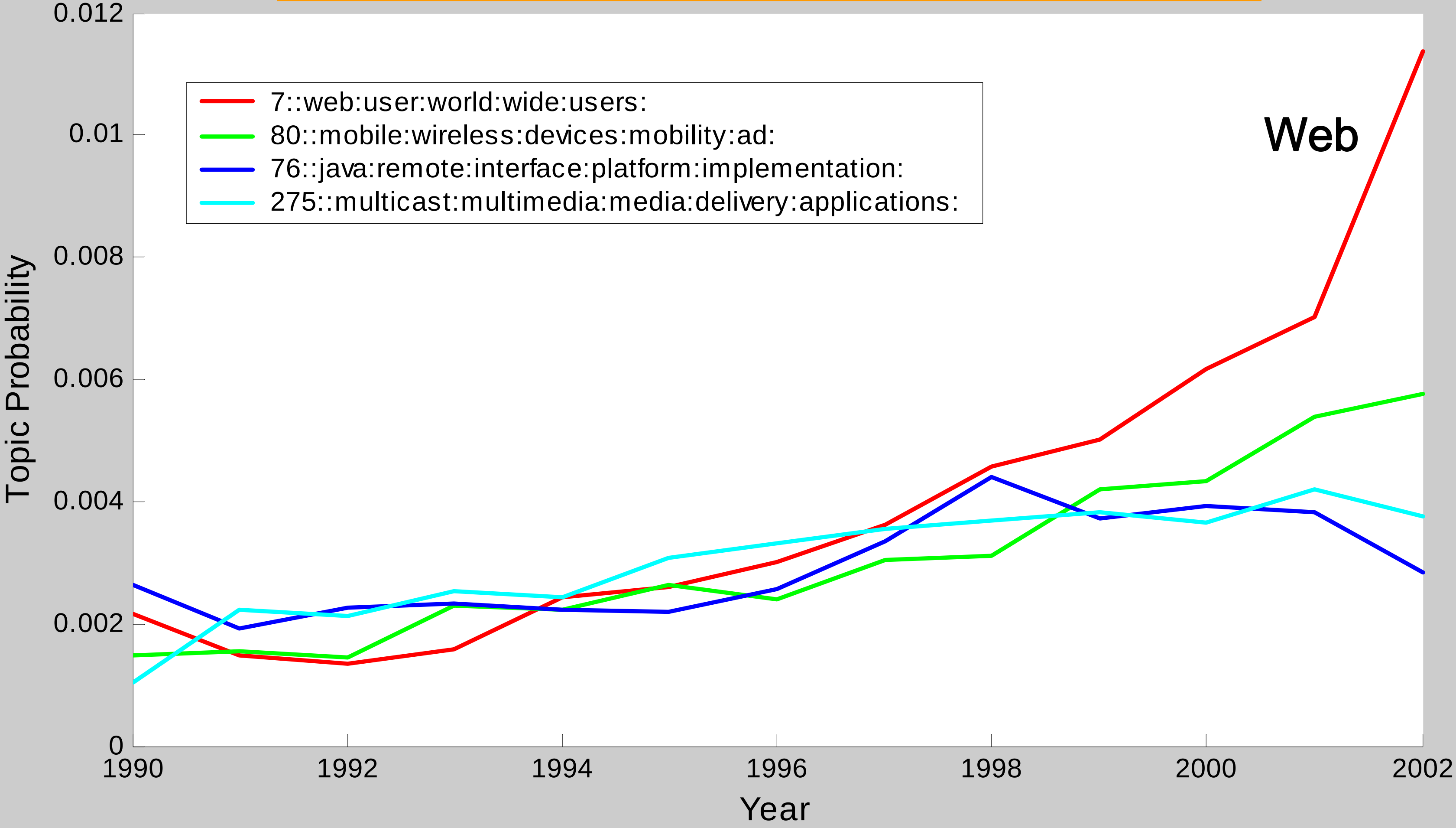
NIPS cold topic…



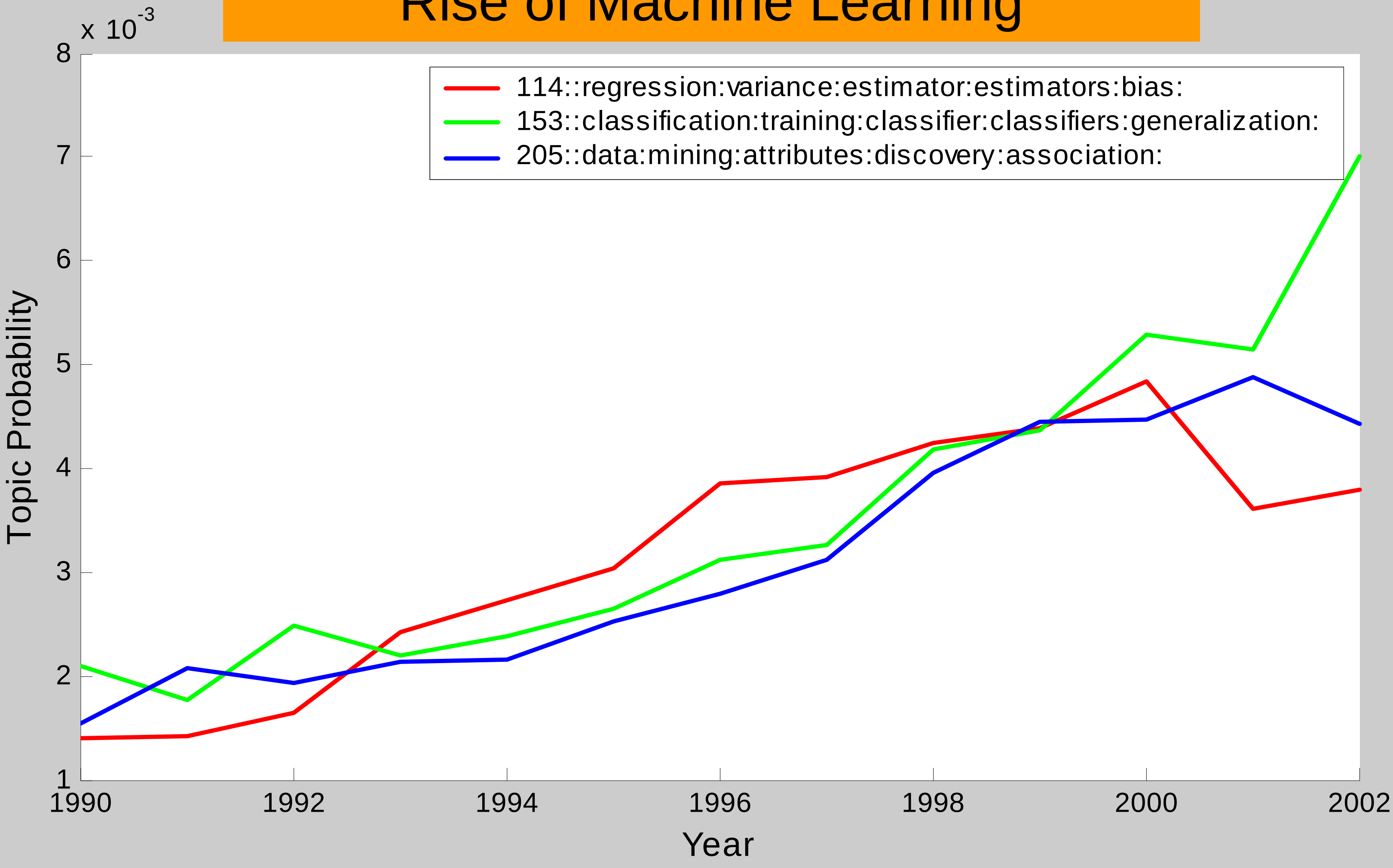
Very hot topic…SVM/Kernel Methods



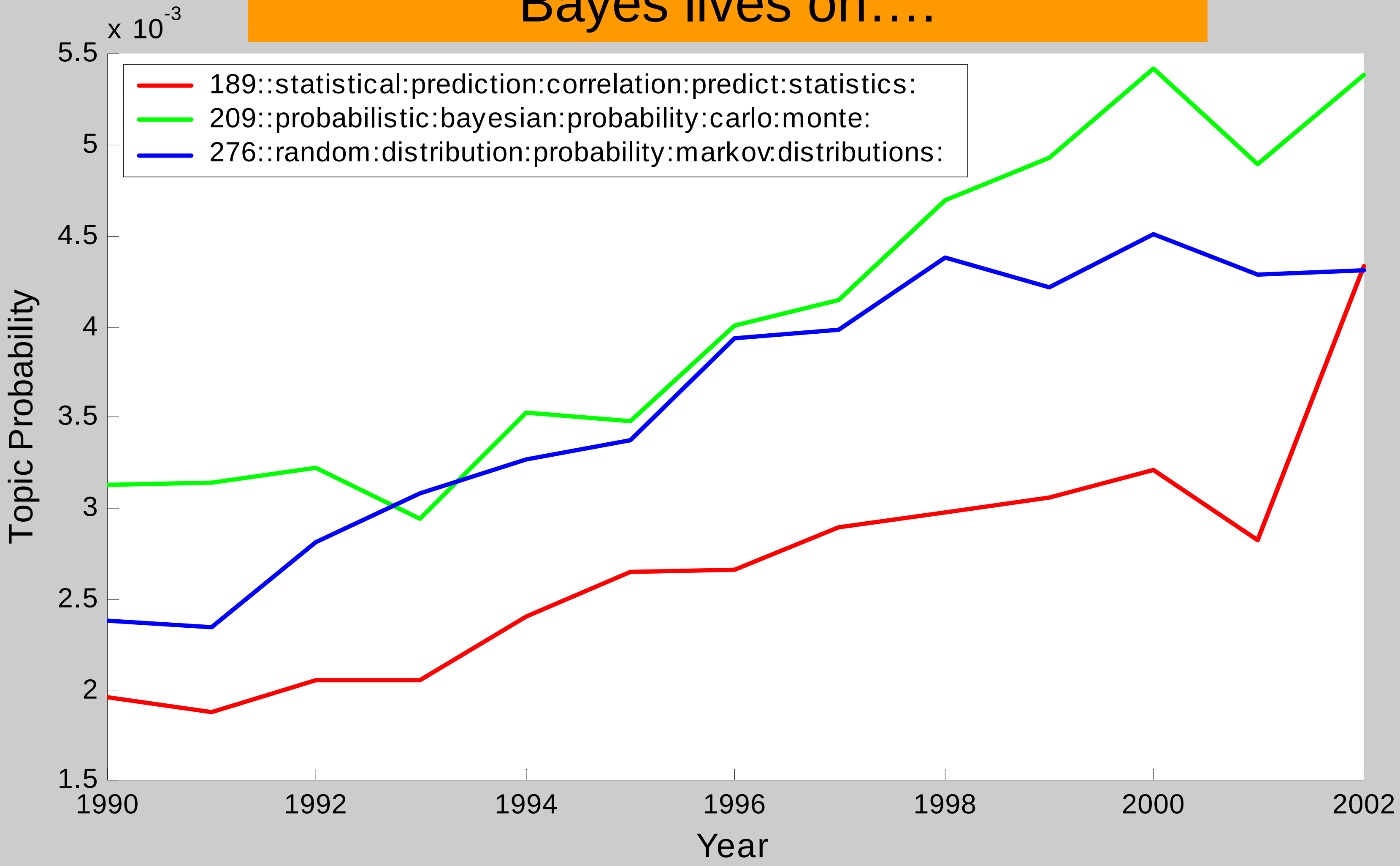
Rise in Web, Mobile, JAVA



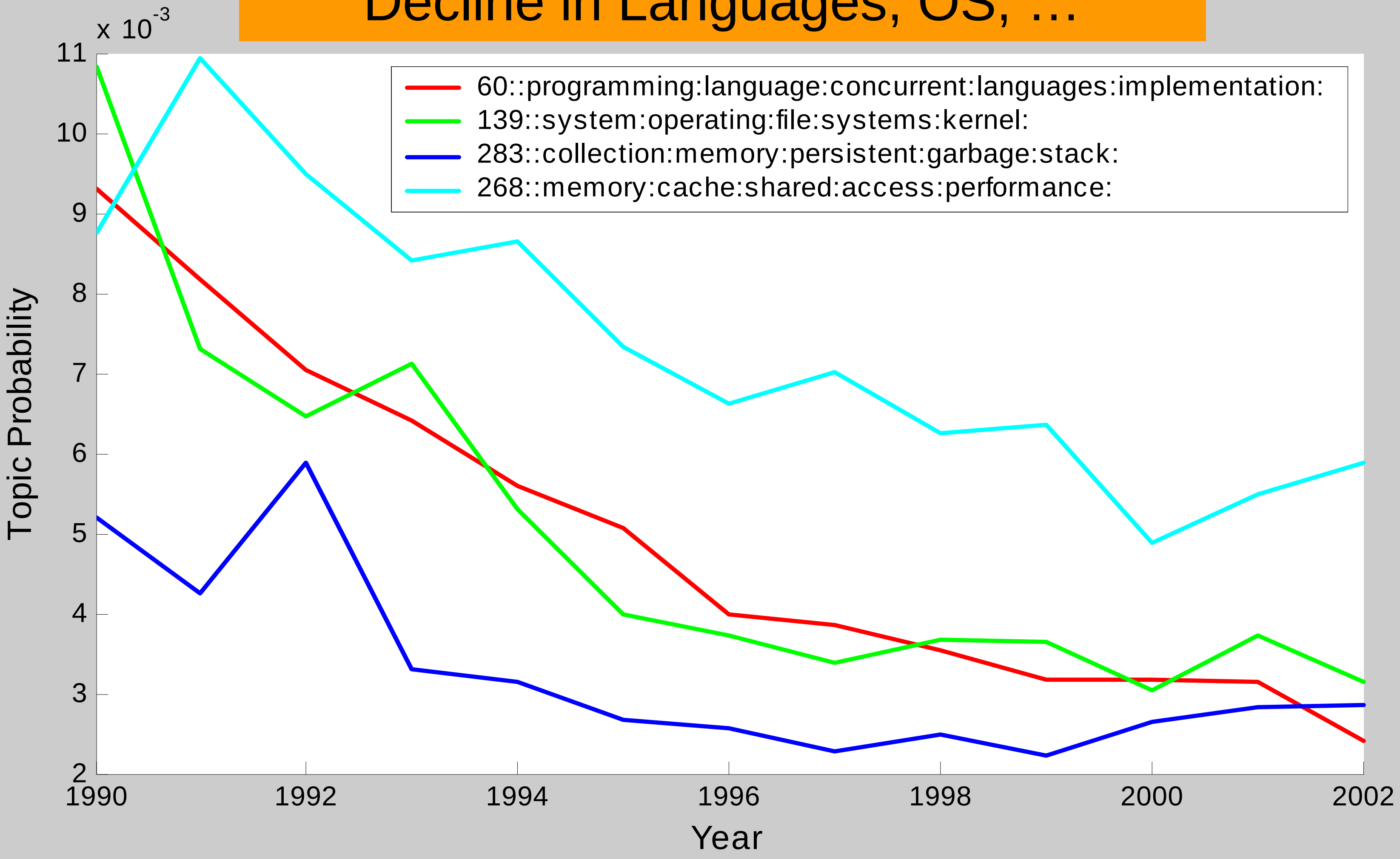
Rise of Machine Learning



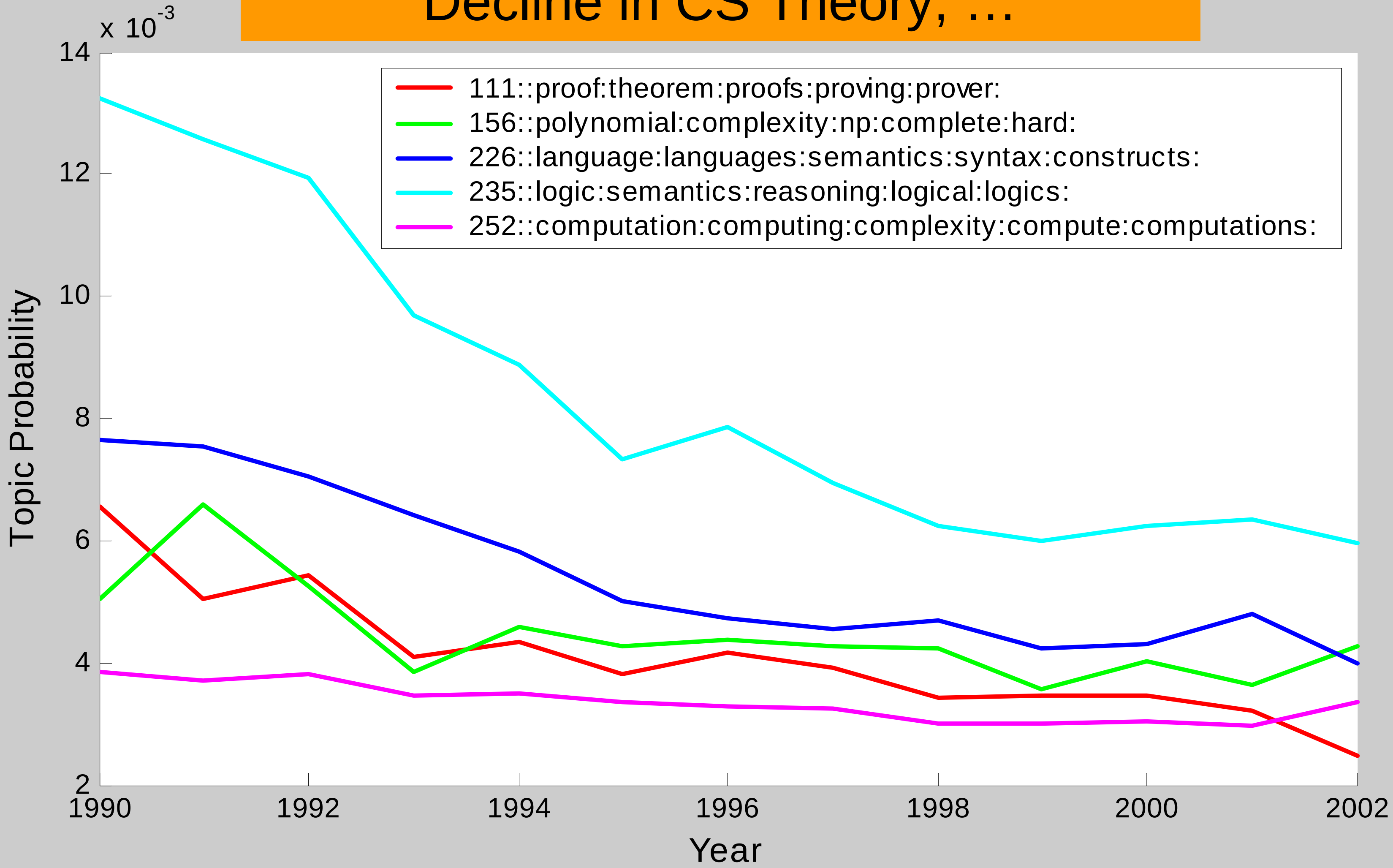
Bayes lives on….



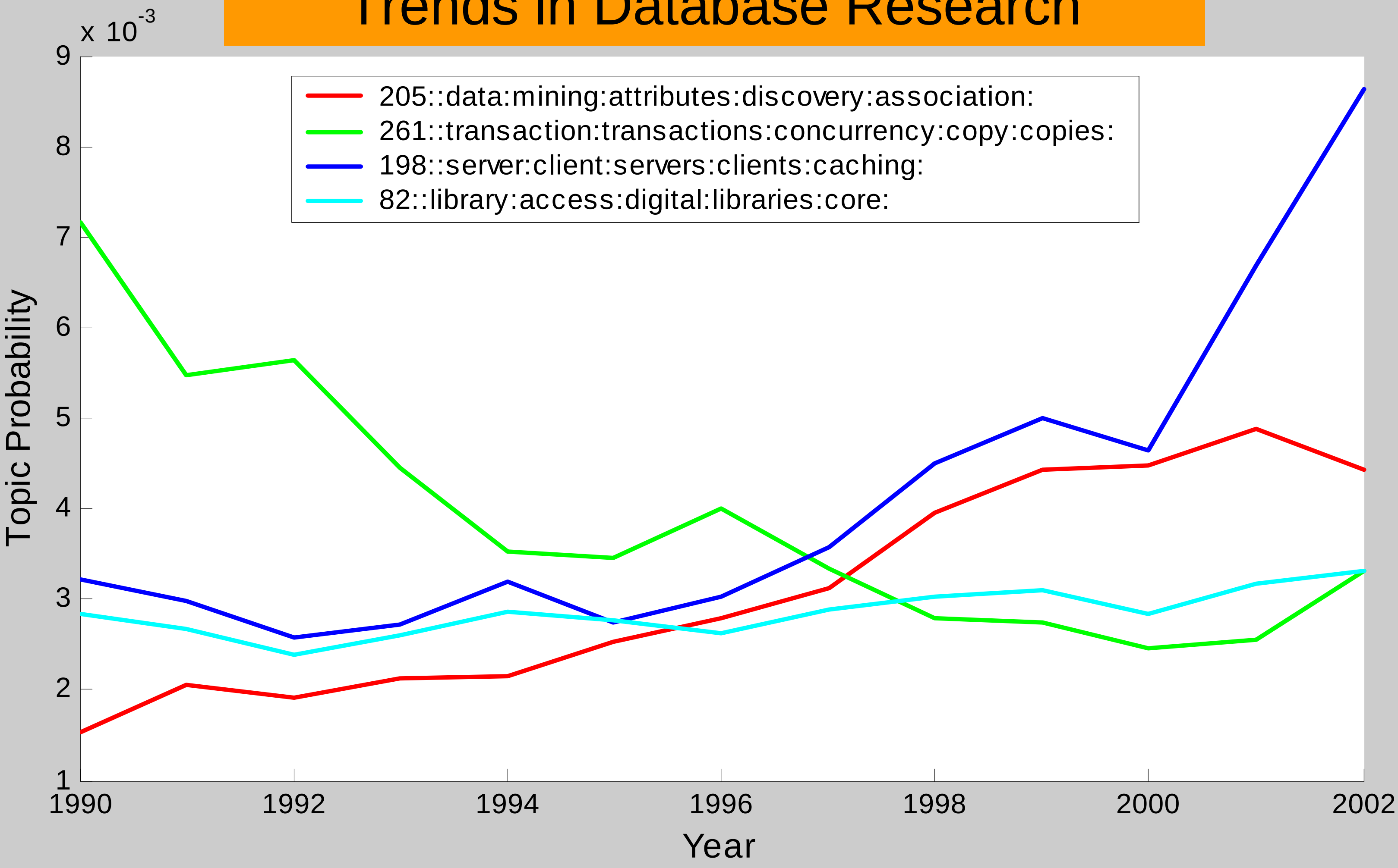
Decline in Languages, OS, …



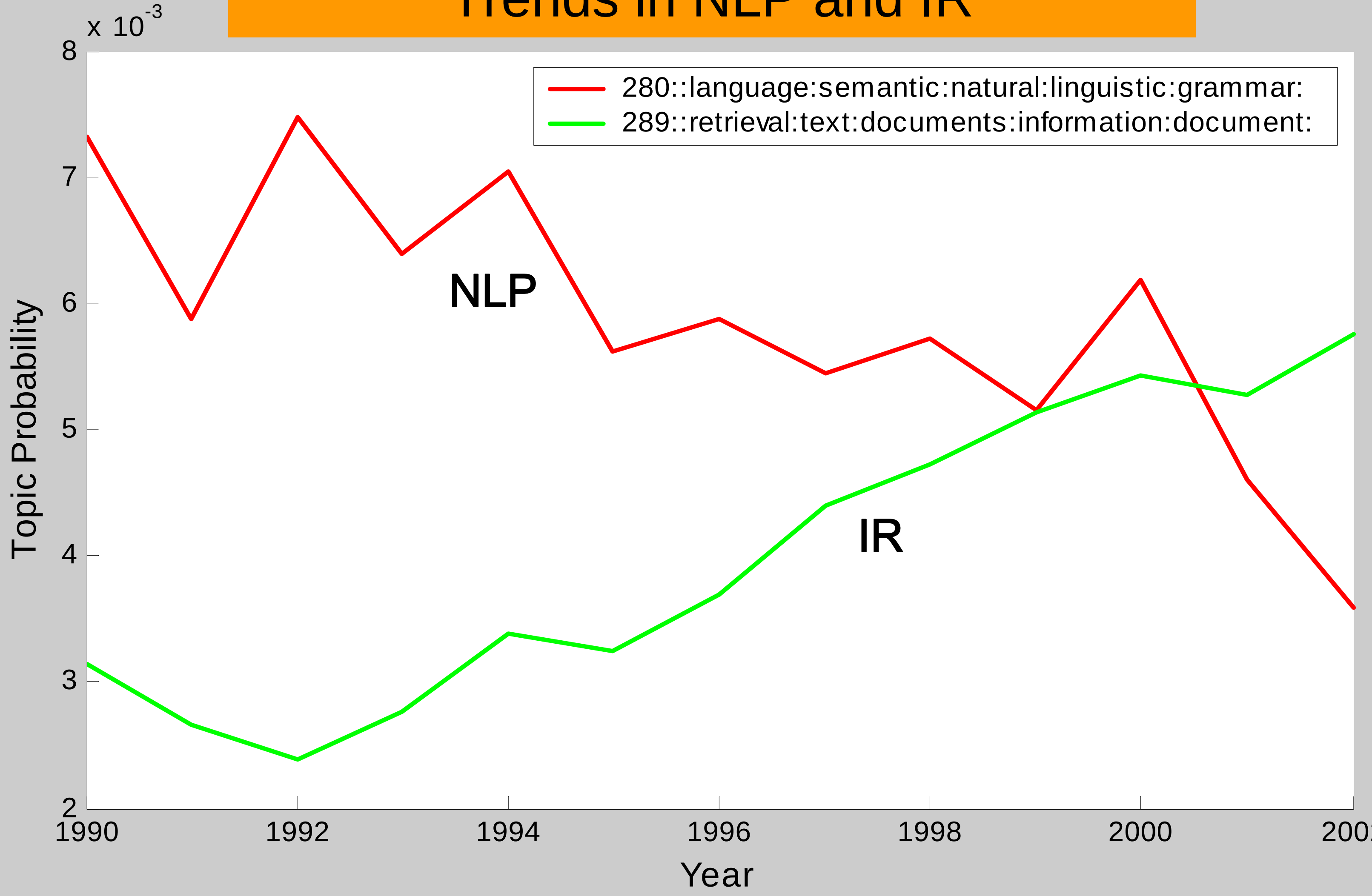
Decline in CS Theory, …



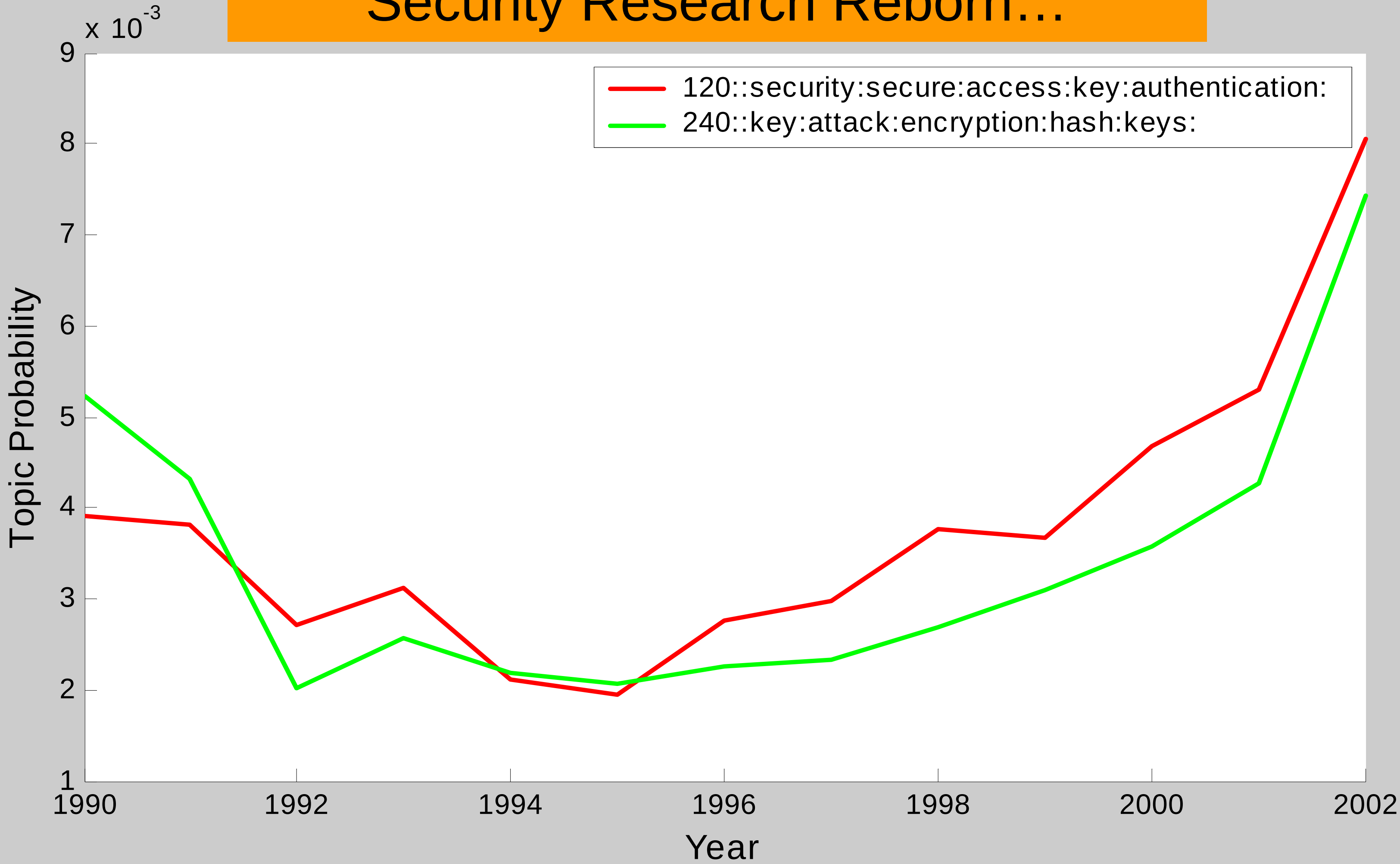
Trends in Database Research



Trends in NLP and IR



Security Research Reborn…



Decline in use of Greek Letters ☹

Extensions/Applications

- Software tool for summarizing text document collections
  - General query-answering capabilities
    - Who writes on what topics?
    - What is the “topic map”?
  - Applications
    - Federal funding databases
    - Enron email archives
    - …..

- Prototype reviewer recommender system
  - Provide system with abstract of a paper
  - Returns list of potential reviewers, based on topic models

General Comments on Data Mining Software

- Spectrum of environments:
  - From restrictive-simple-efficient to general-complex-inefficient
- Examples
  - Data mining directly using SQL
  - High-level modeling standards
    - CRISP-DM
    - PMML (industry consortium)
    - SEMMA (SAS)
  - Public-domain packages
    - WEKA
  - General purpose data analysis environments
    - R, BUGS, MATLAB, etc
- Problems
  - Difficult to know in advance what a data analyst may wish to do
  - Packages are good at algorithms, but poor at process support

Final Comments

- Successful data mining requires integration/understanding of
  - statistics
  - computer science
  - the application discipline

- Current practice of data mining:
  - algorithmic-orientation
  - often focused on business applications
  - little support for iterative scientific process
  - considerable “hype” factor

- Research Directions
  - Interface of statistics and computer science
  - Scaling up statistical ideas to massive, non-traditional data

References

- Papers:
  - `www.ics.uci.edu/~smyth`
  - e.g., "Data mining: data analysis on a grand scale?", P. Smyth, (2000), *Statistical Methods in Medical Research.*
  - Specific papers on curve clustering, author-topic models, etc
- Texts
  - *Principles of Data Mining*
    - D. J Hand, H. Mannila, P. Smyth, MIT Press, 2001
  - *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*
    - Hastie, Tibshirani, and Friedman, Springer-Verlag, 2001
- Web sites
  - www.kdnuggets.com