Stochastic modeling and algorithms for structured data and distributed systems

Long Nguyen

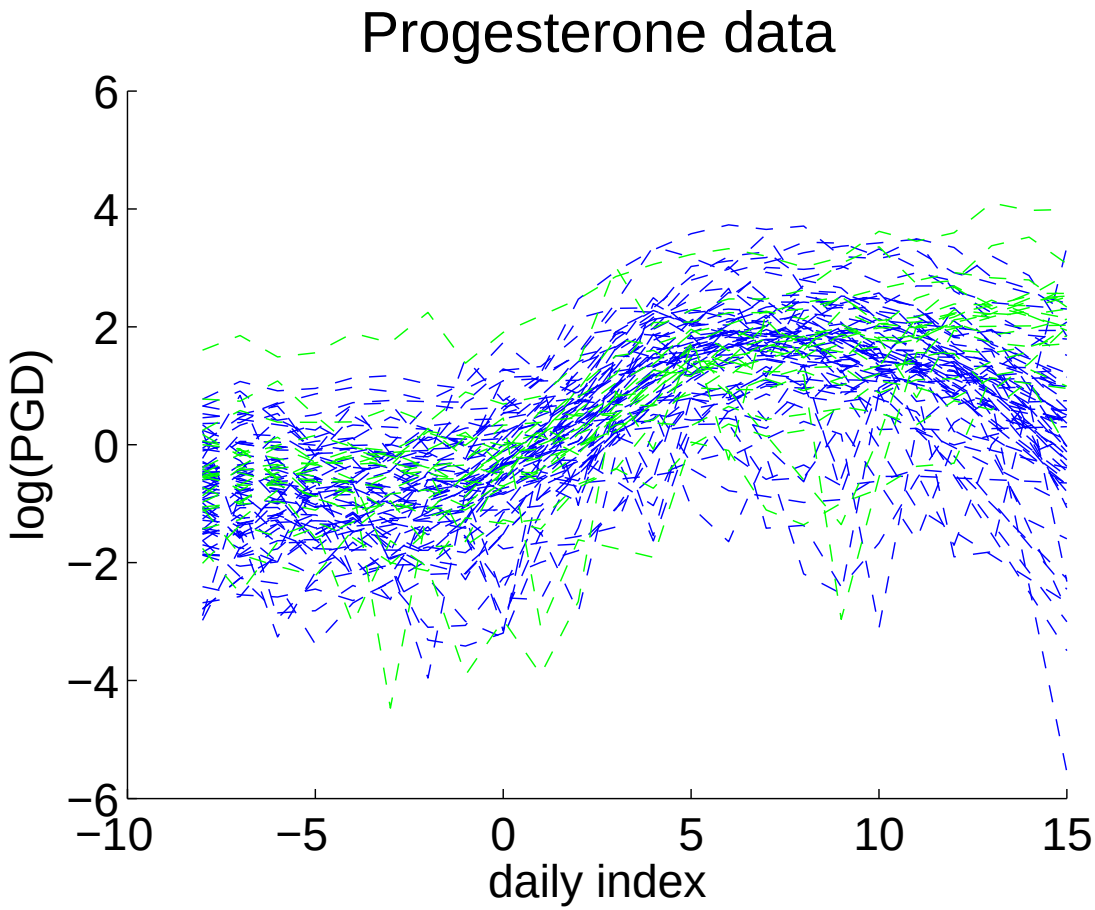
Department of Statistics
Department of Electrical Engineering and Computer Science
University of Michigan

Structured data

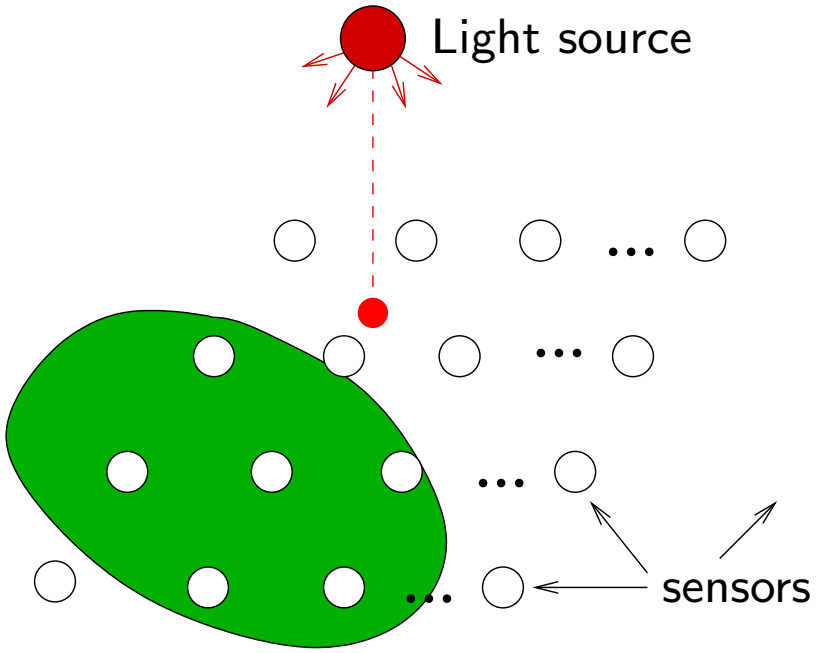
Data that are rich in contextual information:

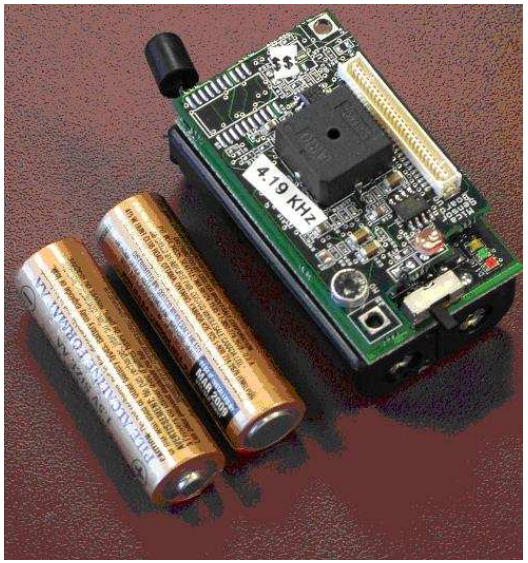
- time/sequence
- space
- network-driven
- etc (other domain knowledge)

Example: Time series signals/curves



Example: Multi-mode sensor networks



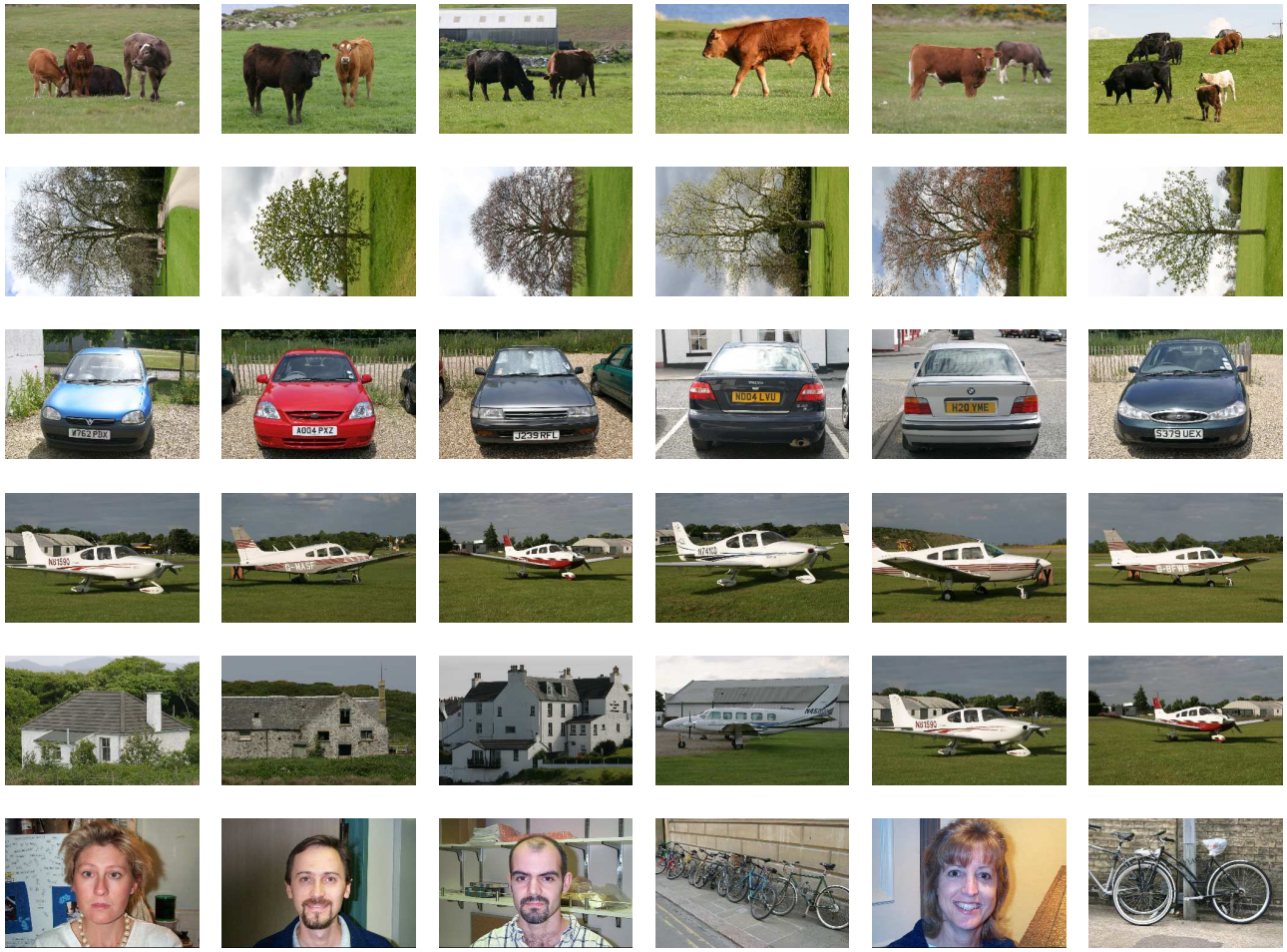


- applications: anomaly detection, environmental monitoring

Example: Sensors distributed over large geographical area

- traffic monitoring and forecast

Example: Natural images



- image segmentation, clustering, ranking

Other data examples we have/are working on

- Ecology: forest populations and species compositions in Eastern US
  - effects of climate change on evolution of species over time and a large geographical area
  - fine-grained aspects of species competition

- Neuroscience: fMRI data of human subjects
  - activity/connectivity analysis
  - neurobiological pathways underlying various risk behaviors

- Information retrieval: social network data

Drawing inference from structured data

- the key step for a statistician (machine learner/data miner) is to systematically translate such known structures into statistically/mathematically rich and yet computationally tractable models

  - borrow "statistical strengh" from one subpopulation/system/task to learn about other subpopulations/systems/tasks

  - aggregage statistical strengh across subpopulations to obtain useful, often "global", patterns

- statistical models provide the right language to describe data, but clever algorithms and data structures are the needed vehicles to help us extract useful patterns

Example: "Bag-of-word" model in IR

- the structure being exploited here is that the "words" are not independent; moreover, they are *exchangeable*
- de Finetti's theorem: If the sequence of random variables $X_1, \ldots, X_n, \ldots$ is infinitely exchangeble, the joint distribution for $X_1, \ldots, X_n$ can be expressed by a mixture model:

$$p(X_1, \ldots, X_n) = \int \prod_{i=1}^{n} p(X_i|\theta)\pi(\theta)d\theta$$

  for some prior distribution π over θ
  - θ plays the role of "latent" topics (e.g., probalistic Latent Semantic Indexing model, Latent Dirichlet Allocation model)
- mixture modeling strategy extends generally to the very rich hierarchical modeling methodology

Beyond exchangeability: injecting spatial/graphical dependence to hierarchical models

- exchangeability assumption is useful for uncovering aggregated and global aspects of data
  - clustering based on latent topics

- but *not* suitable for prediction, extrapolation of local aspects of data
  - segmentation, part-of-speech tagging

- exchangeability assumption is too restrictive in temporal-spatial data, data with non-stationary or asymmetric structures

- other modeling tools are available: Markov random fields (a.k.a. probablistic graphical models), multivariate analysis techniques

Beyond finite dimensionality: Nonparametric Bayesian methods

- in the mixture representation,

$$p(X_1, \ldots, X_n) = \int \prod_{i=1}^{n} p(X_i|\theta)\pi(\theta)d\theta.$$

  the latent (topic) variable θ can be taken to be unbounded (infinite dimensional): As there are more data items, more relevant topics emerge!

- the topics can be organized by random and hierarchical structures

- learning over these random and potentially unbounded topic hierarchies is very natural using tools from stochastic processes (e.g., Dirichlet processes, Levy processes)

Some current works

- Dirichlet labeling process mixture model was developed to account for spatial/sequential dependency (Nguyen & Gelfand, 2009)
  - applied to clustering curves and images, image segmentation

- Graphical Dirichlet process mixture model was developed to learn graphically dependent clustering distributions (Nguyen, 2010)
  - connectivity analysis in social networks, and in human brains

- A great deal of attention is paid to balancing between statistical richness of model and computational tractability
  - better sampling algorithms
  - variational inference motivated from convex optimization

Decision-making in data-driven distributed systems

- communicational and computational bottleneck
- real-time constraints in decision-making
- marrying statistical and computational modeling with constraints driven by distributed systems is an exciting challenge in our research agenda