The Core of AI is Statistics

The future is ours - if we want it

Mattias Villani

Department of Statistics
Stockholm University
and
Department of Computer and Information Science
Linköping University

Overview

- The role of (modern) **Statistics in AI**
- **Predicting the bankruptcy of firms**
- **Real-time robotic search**
- **Evolution of airline networks**
- **Finding bugs in computer code**

The role of (modern) Statistics in AI

- **Artificial Intelligence** and **Machine Learning** is:
  - **computationally efficient statistical inference** using
  - **flexible models** with a focus on
  - **prediction** and
  - **decision-making under uncertainty** using
  - **large-scale data**, often with
  - **real-time** requirements

- **Different focus** than traditional statistics.

- **Statistics at the center of AI**, but only if we really embrace it.

- Very important to **keep our core**:
  - Probability models
  - Rigorous statistical inference
  - Proper data analysis

Updated Statistics education

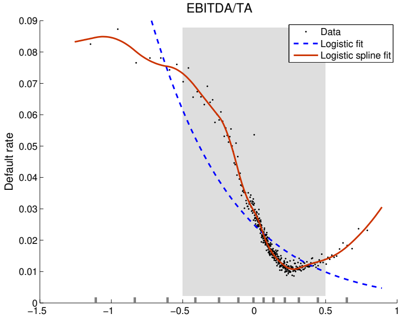
- Master program **Statistics and Machine Learning** at **Linköping University**:
  - Machine Learning, 9 hp
  - Advanced Machine Learning, 6 hp
  - Bayesian Learning, 6 hp
  - Text Mining, 6 hp
  - Big Data Analytics, 6 hp
  - Computational Statistics, 6 hp
  - R programming, 6 hp
  - Python, 3 hp
  - Deep learning, 3 hp
  - Decision theory, 6 hp
  - Statistical Methods, 6 hp
  - Probability Theory, 6 hp
- Joint with engineering master **AI and Machine Learning**.
- Plan for new master courses at **Stockholm University**:
  - Probabilistic Machine Learning, 7.5 hp
  - Bayesian Learning, 7.5 hp
  - R programming, 7.5 hp

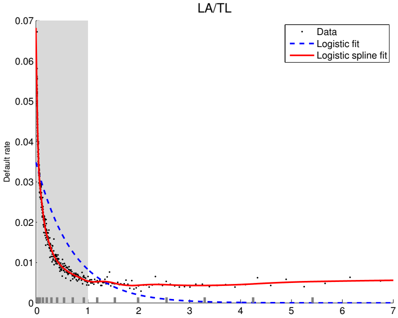
- Quarterly **data** on all Swedish cooperations 1990-2016.
  - **Large data**: 4.7 million observations
  - binary response (bankruptcy)
  - 8 covariates: financial ratios and macro variables.
- **Logistic regression**

$$\Pr(y_i = 1|\mathbf{x}_i) = \frac{1}{1 + \exp(-\mathbf{x}_i^T \beta)}.$$

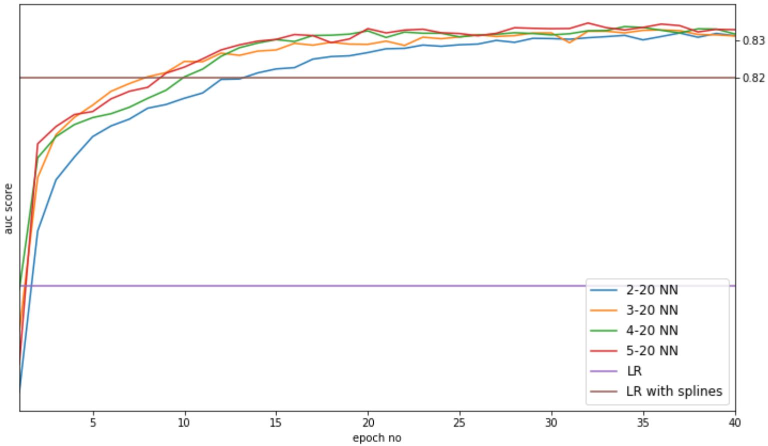
- **Linear decision boundaries** because of linear predictor $\mathbf{x}^T \beta$.
- **Non-linear logistic**: replace $\mathbf{x}^T \beta$ by nonlinear function $f(\mathbf{x})$.
  - **Splines**
  - **Deep neural networks**

Bankruptcy prediction requires nonlinear models

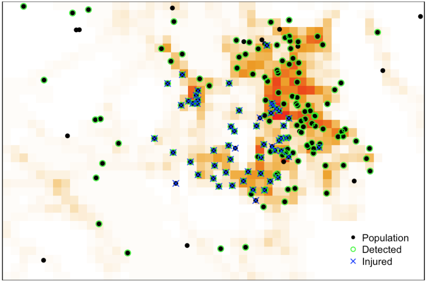




Predictive performance (AUC)



- **Scenario**: Terrorist attack in the city of Gamleby ...
- **Aim**: use flying drones to quickly find injured people.



Real-time robotic search with flying drones

- **Log Gaussian Cox Process** (LGCP) for **number of persons** in $\tilde{S} \subset S$

$$N_{y^\star}(\tilde{S})|\lambda \sim \text{Poisson}\left(\int_{\mathbf{s}\in\tilde{S}} \lambda(\mathbf{s})d\mathbf{s}\right)$$

$$\log \lambda(\mathbf{s}) = \alpha_\lambda + \underbrace{\mathbf{x}_\lambda^\top(\mathbf{s})}_{GIS}\boldsymbol{\beta}_\lambda + \underbrace{\xi_\lambda(\mathbf{s})}_{\text{Gaussian process in } 2D}$$

- The **number of detected** persons by a thinned LGCP

$$N_y(\tilde{S})|r,\lambda \sim \text{Poisson}\left(\int_{\mathbf{s}\in\tilde{S}} r(\mathbf{s})\lambda(\mathbf{s})d\mathbf{s}\right)$$

$$\log r(\mathbf{s}) = \mathbf{x}_r^\top(\mathbf{s})\boldsymbol{\beta}_r$$

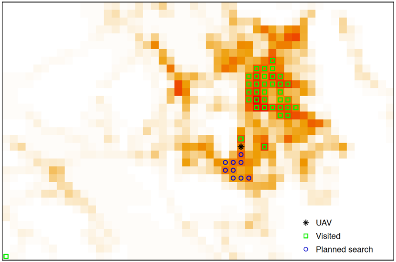
- **Probability of injury**

$$w_i|q \sim \text{Bernoulli}\left(q(\mathbf{y}_i)\right),$$

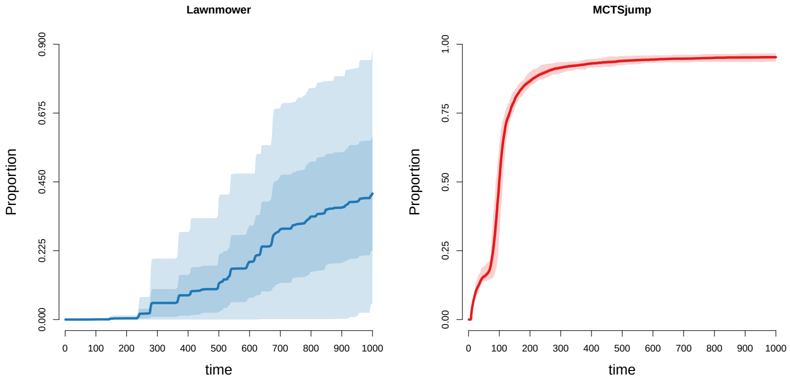
$$\operatorname{logit} q(\mathbf{s}) = \alpha_q + \mathbf{x}_q^\top(\mathbf{s})\boldsymbol{\beta}_q + \xi_q(\mathbf{s})$$

Real-time decision making under uncertainty

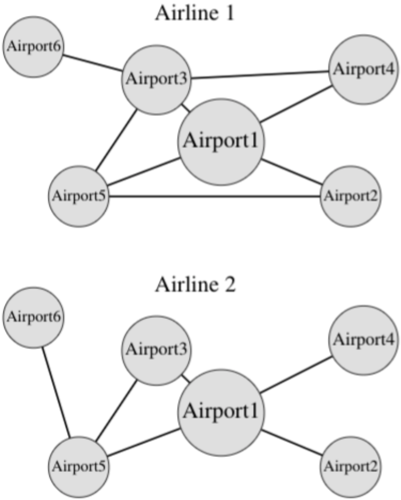
- Challenges
  1. **missing data** - point pattern is only partially observed
  2. **real-time sequential high-dimensional inference**
  3. **real-time decision making** under uncertainty
- Solutions
  1. Strong priors based on **GIS data**
  2. **Warm-started INLA** for **Bayesian inference**
  3. Tailored **Monte Carlo Tree Search** for decisions
- Video: `https://www.youtube.com/watch?v=wyDoO5hF5tE`



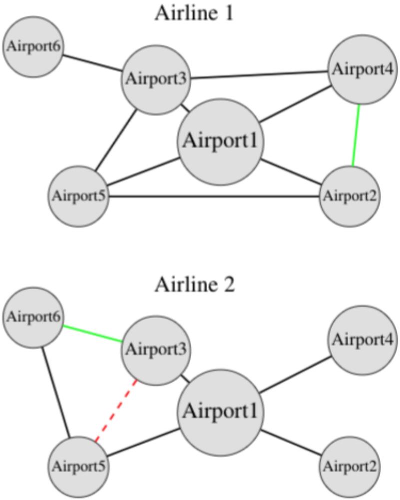
We find injured a lot faster than lawnmower

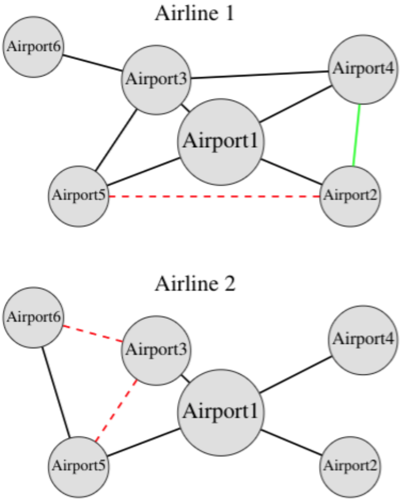


Airline network evolution – time 1



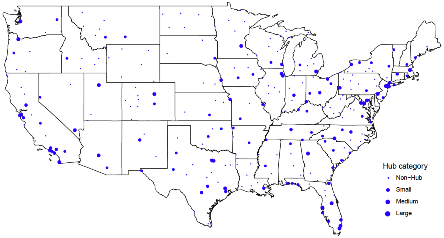
Airline network evolution - time 2

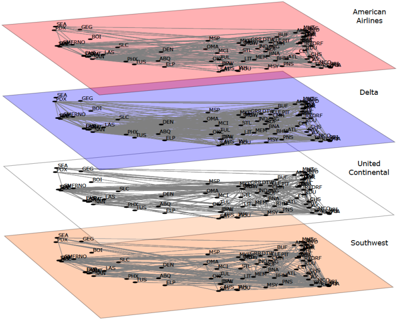




Airline 1
Airport6
Airport3
Airport4
Airport1
Airport5
Airport2
Airline 2
Airport6
Airport3
Airport4
Airport1
Airport5
Airport2

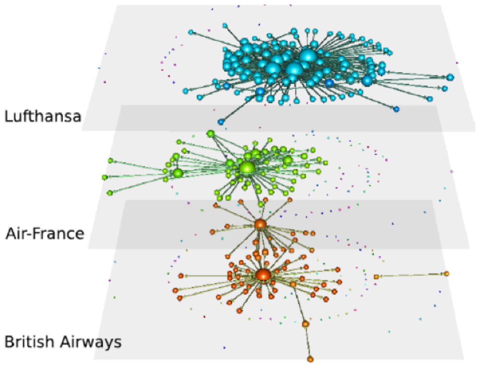
Airline network evolution - US data





Airline network evolution

- **Aim**: **Predict** the evolution of airline **networks over time**.
- **Data**: Quarterly world-wide networks for all airlines.
- **Model**: **Dynamic multi-layered networks** driven by latent processes



- **Static Bernoulli model** for adjacency matrix **Y**

$$Y_{uv}(t)|\pi \overset{iid}{\sim} \mathrm{Bern}(\pi)$$

- **Dynamic Bernoulli** with **global latent Gaussian process**

$$\begin{aligned} Y_{uv}(t)|\pi(t) &\sim \mathrm{Bern}\left(\pi(t)\right) \\ \mathrm{Logit}\left[\pi(t)\right] &= z(t), \\ z(t) &\sim \mathrm{GaussianProcess} \end{aligned}$$

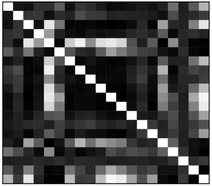
- **Dynamic Bernoulli** with **latent Gaussian processes at nodes**

$$\begin{aligned} Y_{uv}(t)|\pi_{uv}(t) &\sim \mathrm{Bern}\left[\pi_{uv}(t)\right] \\ \mathrm{Logit}\left[\pi_{uv}(t)\right] &= z(t) - \left\| x_u(t) - x_v(t) \right\|, \\ z(t) &\sim \mathrm{GaussianProcess} \\ x_u(t) &\sim \mathrm{GaussianProcess}, \ u = 1, \ldots, N. \end{aligned}$$

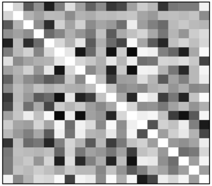
- How to '**scale to large data**'? Many airports, many airlines.

Learning a dynamic multi-layer networks

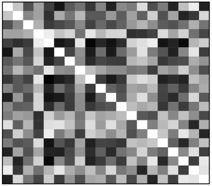
Sampled Link Probabilities at Layer 1, Time 1



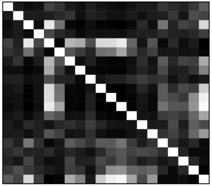
Sampled Link Probabilities at Layer 1, Time 10



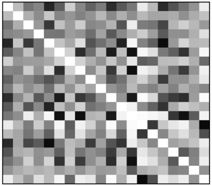
Sampled Link Probabilities at Layer 1, Time 22



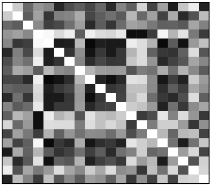
Estimated Link Probabilities at Layer 1, Time 1



Estimated Link Probabilities at Layer 1, Time 10

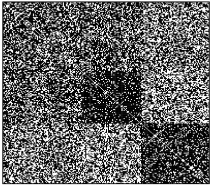


Estimated Link Probabilities at Layer 1, Time 22

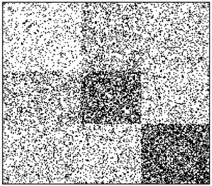


Block-structured multi-layer networks

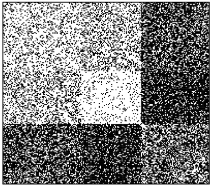
Simulated Adjacency Matrix at Layer 1, Time 1



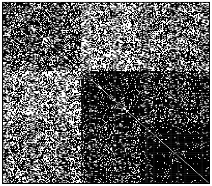
Simulated Adjacency Matrix at Layer 1, Time 9



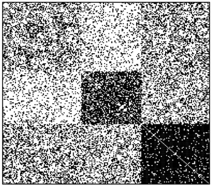
Simulated Adjacency Matrix at Layer 1, Time 18



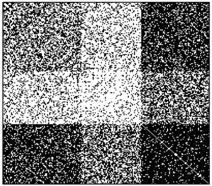
Simulated Adjacency Matrix at Layer 2, Time 1



Simulated Adjacency Matrix at Layer 2, Time 9

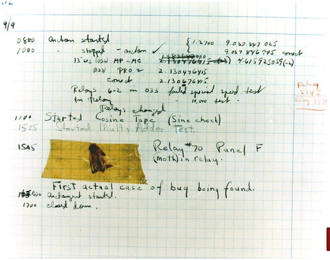


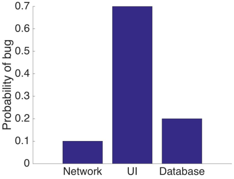
Simulated Adjacency Matrix at Layer 2, Time 18



PREDICTING BUG LOCATION FROM BUG REPORTS

```
#-----------------------------------------

def Network(networkInputs):

    # CODE
    # MORE CODE
    # TOO MUCH CODE

    return(networkOutputs)

#-----------------------------------------

def UI(UIinputs):

    # CODE
    # MORE CODE
    # TOO MUCH CODE

    return(UIOutputs)

#-----------------------------------------

def Database(DBinputs):

    # CODE
    # MORE CODE
    # TOO MUCH CODE

    return(DBOutputs)

#-----------------------------------------
```





Data

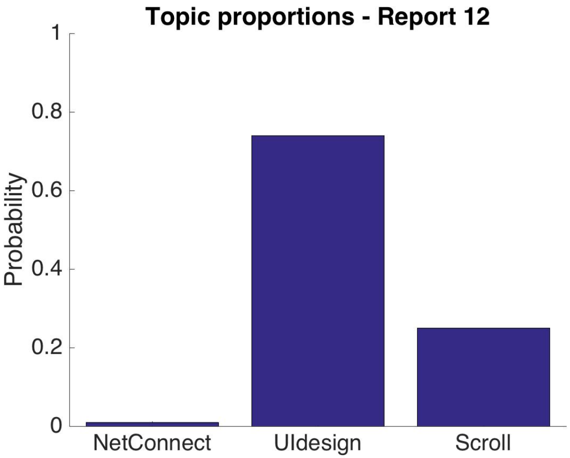
Dataset	No. Bug reports	No. classes	Vocabulary size
Mozilla	15,000	118	3505
Eclipse	15,000	49	3367
Telecom	9,778	26	5286

Summarize a bug report with topic models

- Summarize a collection of text documents into **topics**
- **Probabilistic model**
- Inputs:
  - a large collection of text documents.
  - Number of topics, K.
- Outputs:
  - a set of K **topics** that the documents talks about.
  - a numeric vector for each document with K **topic proportions**.

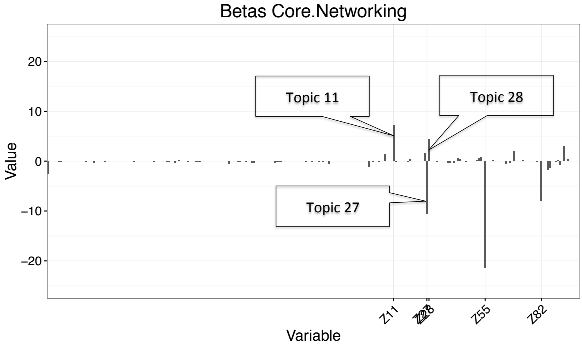
Topic	Topic label	Top 10 words in topic
11	HTTP	proxy server http network connection request connect error www host
27	Layout	div style px background color border css width height element html
28	Connection Headers	http cache accept en public localhost gmt max modified alive
55	Search	search google bar results box type find engine enter text
82	Scrolling	scroll page scrolling mouse scrollbar bar left bottom click content

Topic proportions



The effect of topics on the classes

- Topic proportions are used as covariates in **multinomial regression**.
- $\beta_{\text{topic,class}}$ is the effect of `topic` on `class`.
- **Horseshoe shrinkage prior** on $\beta_{\text{topic,class}}$ to sort out important topics for each class.



Interpretable predictions

- **DOLDA - Diagonal Orthant Latent Dirichlet Allocation**. Supervised. Topics are directly related to classes.

- System:

  - I am ***very certain*** *that the bug is in UI code*

  - ***because*** *report talks a lot about UIdesign and Scroll and very little about NetConnect.*

  - Sending the bug report to the UI-team.

- System:

  - I am ***very uncertain*** *where the bug is*

  - ***because*** *bug report contains a jumble of topics.*

  - Don't trust me. Please ask human.

Interpretable prediction without loss of accuracy

Dataset	# Classes	DOLDA	StackingLDA
Mozilla	118	45%	39%
Eclipse	49	61%	55%
Telecom	26	71%	75%