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

- 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

PREDICTING THE BANKRUPTCY OF FIRMS

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 $\mathbf{x}^T \boldsymbol{\beta}$.
- **Non-linear logistic**: replace $\mathbf{x}^T \boldsymbol{\beta}$ by nonlinear function $f(\mathbf{x})$.
 - Splines
 - Deep neural networks

BANKRUPTCY PREDICTION REQUIRES NONLINEAR MODELS



PREDICTIVE PERFORMANCE (AUC)



REAL-TIME ROBOTIC SEARCH WITH FLYING DRONES

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



REAL-TIME ROBOTIC SEARCH WITH FLYING DRONES



STRUCTURAL SPATIAL POINT PROCESS

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

$$N_{y^{\star}}(\tilde{S})|\lambda \sim \text{Poisson}\left(\int_{\mathbf{s}\in\tilde{S}}\lambda(\mathbf{s})d\mathbf{s}\right)$$
$$\log \lambda(\mathbf{s}) = \alpha_{\lambda} + \underbrace{\mathbf{x}_{\lambda}^{\top}(\mathbf{s})}_{GIS}\beta_{\lambda} + \underbrace{\xi_{\lambda}(\mathbf{s})}_{\text{Gaussian process in 2}}\beta_{\lambda}$$

The number of detected persons by a thinned LGCP

$$N_{y}(\tilde{S})|r, \lambda \sim \text{Poisson}\left(\int_{\mathbf{s}\in\tilde{S}}r(\mathbf{s})\lambda(\mathbf{s})d\mathbf{s}\right)$$
$$\log r(\mathbf{s}) = \mathbf{x}_{r}^{\top}(\mathbf{s})\beta_{r}$$

Probability of injury

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

 $\text{logit} q(\mathbf{s}) = \alpha_q + \mathbf{x}_q^{\top}(\mathbf{s})\beta_q + \xi_q(\mathbf{s})$

- 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 NETWORK EVOLUTION - TIME 3





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



DYNAMIC NETWORKS DRIVEN BY LATENT VARIABLES

- Static Bernoulli model for adjacency matrix \mathbf{Y} $Y_{uv}(t)|\pi \stackrel{iid}{\sim} \operatorname{Bern}(\pi)$
- Dynamic Bernoulli with global latent Gaussian process

 $egin{aligned} & Y_{uv}(t) | \pi(t) \sim \mathrm{Bern}\left(\pi(t)
ight) \ \mathrm{Logit}\left[\pi(t)
ight] = & \mathbf{Z}(t), \ & \mathbf{Z}(t) \sim \mathrm{GaussianProcess} \end{aligned}$

Dynamic Bernoulli with latent Gaussian processes at nodes

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

■ 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



PREDICTING BUG LOCATION FROM BUG REPORTS

| | | | def I | letwork(netw | orkInputs): | | | | |
|--|--|--|---|---|--|---|-----------------------|--|--|
| | | | | CODE | | | | | |
| | | | | MORE CODE | ODE | | | | |
| | | | | | | | | | |
| | | | | eturn(netwo | rkoutputs) | | | | |
| | | | # | | | | | | |
| | | | def l | I(UIinputs) | : | | | | |
| | | | | CODE | | | | | |
| | | | - 1 | TOO MUCH C | ODE | | | | |
| | | | | eturn(UIOut | puts) | | | | |
| | | | ø | | | | | | |
| | | | def I | atabase(DRi | nnuts). | | | | |
| | | | | CODE | | | | | |
| | | | | MORE CODE | 0.05 | | | | |
| | | | | TOU MUCH C | ODE | | | | |
| | | | | eturn(DBOut | puts) | | | | |
| | | | # | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| 14 | | | | | | | | | |
| ·* //= | | | | | | | | | |
| 1/9 | 1.61 | | | | | | | | |
| ~ 1/9 0 500 Autom | startil strut - auton | {1-2700 | 9.01 | 7 847 025 | | | | | |
| 1/9 0 5800 Andram 1000 Is | stanted shoped - aiction inclose MP-Me | { | 9.03 9.03 (66) 9 | 7 847 025 876 785 000 1615925059(-1) | .t | | | | |
| | stanted staged - audam ice lass AR - AR Bay PRA Concert | {/-2.700 2-13.09.70 2-13.09.70 3-13.09.70 3-13.09.70 | 9 01 9 15 (**) 5 | 2 547 025 1 876 985 - 994 1615925059(-1) | u . | 0. | 7 | | |
| 1/9 0500 Anton 1000 Is | stants stopped - autom inc loss AP and and AP and courd sons 6-2 in c | { / 2.700 2 130475223 / 5 2 1304762 / 5 3 / 7506 76475 3 / 7506 76475 4 / 7506 76475 4 / 2007 4 / 2009 2 | 9 0 9 9 0 0 9 0 9 | 2 547 625 1876 785 cmu 415925059(-1) Heal | ek Bilang Bilang | 0. | | | |
| 2/9 0500 Andan 1000 12 Fe | started stopped - autom inclose AP - AC any PPO to any PPO to any or in c any or in c | {/22,000 2. (13,09,247) 2. (13,09,247) 3. (23,09,247) 3. (23,09,24 | 9.09 9.09 1.09 1.09 1.09 1.09 1.09 1.09 | 7 1473 025 1476 945 - 1446 1475 925089(12) Head | ريد المراجع المراجع المراجع المراجع المراجع المراجع | 0. | 7 | | |
| 1/9 0 500 Ondan 1000 Is 1575 Starta 1575 Starta | stantil shaped - action in low HP-MC over J court fairs or in c fairs cho Casine Tap Casine Tap | (1-2,00) 1-136/756/5- 2-1369756/5- 3-1369756/5- 035 full spirit | 9 0 9 9 0 0 9 0 9 | 7 607 025 1946 995 - enus 1975 725059(12) Head | 14 Fellowit 3174 2 3174 2 | 0. 0.1 0.2 0.3 | 7 | | |
| 1/9 0580 andon 1/000 is 1/1/00 1/100 1 | stantil stoppid - action could - action day pro- could - could song - could days - child Cosine - Tap ed Multy Ad | (12700) + (1369750) 5- 2. 13697607 2. 13697607 2. 13697607 3. 136 | 9 0 9 8 0 9 9 0 0 9 0 9 | 9 597 025 1 596 095 000 1 595 52505(2) 1 405 | st Silver Silver | 0 ol pni ol pni | 5 | | |
| 1/9 0580 Quidan 1/000 III 1/1/1 1575 Start-1 1575 Start-1 1545 | staatul songal - auden varia (aa aa aa songal - aaa songal - aa songal - aa songal - aa songal - aa so | (13.000) 1.1309367 2.1309367 3.1309367 5035 full squart full Conception (Sinc check Relay #70 (Bott) in relay | ron en tot tot Pana | 9 m3 e25 946 f95 env 405925e9(4) 404 404 404 | At Asily 5 Sily 5 Sily 5 | :.0 1.0 0 0 | 7 | | |
| 1/4 0500 Andan 1/000 II 1/00 1/00 1/00 1/00 1/00 1/00 1/00 1/ | startil stoppi - autom coud APP-ME coud lass or in c the coud Coine stop different Coine stop Coine stop Coine stop | (1300) 11111111111 2. Isaysory 3. Isaysory out 1. Isaysory 1. Is | ros en tot tot Pana 1 | 2 W2 OLS W6 OS and W5 F25059(W) Had | 4 Sive 107 | .0 .0 .0 .0 .0 .0 | 7 | | |
| 1/9 0500 Ondom 1/000 IIS IS 1575 Start- 1575 Start- 1575 Start- 1575 Start- | stantul stoppi - autor stoppi - autor Box pro-ne Source - actual Cosine Tap athult Ad | Arthurson Arthurson 2 1.1249200 2 1.1249200 2 1.1249200 2 1.124920 2 1.1249200 2 1.124920 2 1.1249200 2 1.1249200 2 1.1249200 2 1.1249200 2 1.1249200 | Pana 1 Pana 1 Pana 1 Pana | 9 by ors We the and Wistrasol(w) Hade I F Al. | 4 Stat 2010 | obability of bug 2. 0. 0 2. 0. 0 | 7 5 5 4 | | |
| 1/9 0 580 audan 1/000 13 72 1575 Start 1575 Start 1575 Start 1585 1000 2010 | stantyl singui - autor ov nav HP-ne ov Low HP-ne ov Low HP-ne i - over i - | (13000 11309001 2130000 21300000 21300000 2130000000 213000000000000000000000000000000000000 | 9.05 9.05 1.000 1 1.0000 1.00000 1.00000 1.0000 1.0000000 1.00000000 | 9 WY ORS 1940 PRS - com 1945 92509(w) Head Head I F | 4 F)-r4 X1/2 K | Probability of bug :0 :0 :0 :0 | 7 5 1 3 2 | | |
| 199 0 500 Aulan 1999 Is 1999 Is 1999 Starts 1995 Starts 1995 Prest 1996 Starts 1996 Starts | stantil singus - autom ou low mene owy mene source - and ou our - and ou our - and outom outom outom outom actual case - shull. | (1200 11507007 2.15097007 2.15097007 2.5097007 Cost Auto apart (Since Auto (Since Auto (| 200 301 400 1 400 1 1 1 1 1 1 1 1 1 1 1 1 1 1 | 7 157 025 1976 7985 anno 1915 Fizzo 19(-1) -test F | ۲ ۲ ۲ ۲ ۲ ۲ ۲ ۲ ۲ ۲ ۲ ۲ ۲ ۲ | .0 Probability of bug | 7 5 4 3 2 | | |
| 1/9 0 Gro andon 1/000 1/1 1/1 1/1 1/1 1/1 1/1 1/1 1/1 1 | statu stratu stratu stratu stratu stratu actual casa statu st statu st st st st st st st st st st | (1220) 2 (1220) 2 (1220) 2 (1220) 2 (200) 2 | 900 900 1000 1000 1000 1000 1000 | 9 byg oris 1996 Fore Good 44 Stass Stass 44 Stass Stass 44 Stass 45 Stass 46 Stass 4 | 4 | .0 Probability of bug .0 .0 .0 .0 .0 | 7 5 4 3 2 | | |

| 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 connec- |
| | | tion request connect error www |
| | | host |
| 27 | Layout | div style px background color bor- |
| | | der css width height element html |
| 28 | Connection | http cache accept en public local- |
| | Headers | host gmt max modified alive |
| 55 | Search | search google bar results box type |
| | | find engine enter text |
| 82 | Scrolling | scroll page scrolling mouse scroll- |
| | | bar 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 β_{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% |