Education in Statistics

Time for change? Yes, definitely.

Mattias Villani[1]

Department of Statistics
Stockholm University

Department of Computer and Information Science
Linköping University

[1] Slides: http://mattiasvillani.com/news

My recent article in Qvintensen 2020:1

Statistikämnet är hotat – tack och lov

Data finns numera överallt och är en enorm drivkraft för modern industri. I denna blomstringstid för dataanalys finns paradoxalt nog en oro för statistikämnets död. Vi utmanas av närliggande ämnesområden som maskininlärning. Det är en välbehövlig katalysator till förnyelse av vårt ämne.

Jag har under åren 2011–2019 lett avdelningen för statistik och maskininlärning vid Linköpings universitet med ambitionen att integrera statistik med signalbehandling och datavetenskap, speciellt maskininlärning. I det arbetet har jag tillsammans med kollegor utvecklat och undervisat på ett stort antal kurser i maskininlärning för statistiker och ingenjörer, och forskat i gränslandet mellan statistik, maskininlärning och artificiell intelligens. Jag upplever därför inget hot från angränsande ämnen, utan ser tvärtom stora möjligheter till samarbeten och framför allt en vitalisering av statistikämnet.

Jag ska här föreslå sex områden där jag upplever att vårt ämne bör förändras om vi ska fortsätta att spela en central roll i det nya informationssamhället. Jag är väl medveten om att andra universitet, speciellt de som inte har en närhet till en teknisk högskola, kommer behöva välja en delvis annan väg än den vi valde i Linköping. Men jag vill hävda att mina sex förslag är applicerbara på alla utbildningar i statistik och matematisk statistik; de handlar om modernisering av undervisning i statistisk dataanalys och inferens. Som utgångspunkt för mitt inlägg ska jag först beskriva vad jag själv lägger i begrepp som artificiell intelligens, maskininlärning och data science.

Vad är artificiell intelligens och maskininlärning?

Artificiell intelligens (AI) handlar om att ge maskiner eller mjukvara någon form av människoliknande kognitiv förmåga. Mer konkret handlar AI om att göra det möjligt för en maskin att lära känna sin miljö och fatta beslut som maximerar sannolikheten att uppnå uppsatta mål. Moderna AI-system använder data för att lära upp dessa system med hjälp av maskininlärning. Under senare år har det skett stora framsteg inom den enklare formen av artificiell intelligens där maskiner utför mycket specialiserade uppgifter som att känna igen bilder eller förstå talat språk. Begreppen "känna igen" och "förstå" ska inte övertolkas här, dagens AI kan inte sägas förstå språk i någon djupare semantisk mening, men har lärt sig mönster i språket genom att tränas på enorma mängder text och ljud. Genombrott i s k generell AI där maskiner, likt människor, kan generalisera kunskaper till helt nya domäner har visat sig oerhört svårare. En del AI-forskare föredrar därför att tala om utökad intelligens (eng. augmented intelligence) där maskiner (t ex mobiltelefoner) utökar en människas intelligens.

Maskininlärning (ML) är vetenskapen om metoder och algoritmer som gör det möjligt för en maskin eller mjukvara att använda data för att lära sig om världen och fatta optimala beslut under osäkerhet. Det klassiska exemplet är en robot som lär sig om omgivningen från ett antal sensorer och kontinuerligt fattar beslut för att uppnå ett mål. Ämnet överlappar med reglerteknik och signalbehandling. Både namnet maskininlärning och detta klassiska exempel målar dock upp en alltför snäv bild av ämnesområdet. Den stora majoriteten av tillämpningar av maskininlärning involverar inte fysisk hårdvara, utan handlar snarare om olika former av mjukvara med automatiserade prediktioner och beslut. Det kan t ex handla om s.k. appar som rekommenderar filmer åt filmkonsumenten, ger läkaren diagnosförslag baserat på patientmätningar, hjälper jordbrukaren med råd om gödsling från sensorer på åkern, eller utför automatiserad försäljning av aktier. ML har även börjat användas inom samhällsvetenskap och (digital) humaniora.

Under senare år har även termen *Data Science* (DS) populariserats och det finns en stor arbetsmarknad för s k data scientists. DS verkar betyda olika saker för olika personer. Det finns en stor spridning i examina på personer som titulerar sig som data scientists, allt från datavetare utan några större statistiska ämneskunskaper till personer med master- eller t o m doktorsexamen i statistik/matematisk statistik. Det finns också alternativa titlar, som data analyst och data engineer, på olika delar av spektrumet från enkel datainsamling till avancerad dataanalys, som tolkas lite olika beroende vem man frågar. Kvalificerade statistiker som använder titeln data scientists eller liknande har i allmänhet större kunskaper i datahantering, programmering och algoritmer än den genomsnittlige statistikern.

I den vidaste definitionen handlar maskininlärning i princip om alla former av datadriven prediktion och beslutsfattande under osäkerhet. Och någonstans här börjar termen bli kontroversiell för oss statistiker. Det har ju vi statistiker hållit på med i minst 100 år! Jag är den förste att hålla med om att mycket inom ML är traditionella statistiska metoder under nya namn. Jag är också rätt less på den överdrivna tilltron till maskininlärning som ett slags trollstav, som kan omvandla brusiga skräpdata till precisa prediktioner. Men jag välkomnar också hur nya ämnen som ML utmanar statistikämnet med nya perspektiv, och tvingar oss åtminstone till självreflektion, men förhoppningsvis också till förändring. För även om statistik och ML arbetar med samma palett av inferens, prediktion och beslut, så är betoningen på olika delområden ofta radikalt olika. ML har mycket att lära sig från oss statistiker vad gäller modellering och inferensteori. Men vi statistiker kan också lära oss en hel del från ML om effektiva beräkningar, storskalig datahantering och om nya högintressanta tillämpningar och probabilistiska modeller för komplexa problem.

My background

- **PhD in Statistics** from Stockholm University, 2000.
- Researcher and adviser Sveriges Riksbank, 2004-11.
  Developed models and methods for **monetary policy**.
- Interested in **Computational Statistics**.
- Flexible models and **Machine Learning** since 2006.
- Built up **Division of Statistics and Machine Learning** within **Computer Science** dept at LiU, 2011-19.
- Developed a large number of **machine learning courses** for statistics master, engineering programs and industry.[2]
- **Application areas** in research: economics, neuroscience, robotics, software engineering, text analysis, transportation.[3]

[2] Slides: http://mattiasvillani.com/teaching

[3] Slides: http://mattiasvillani.com/research

1. Prediction and decision

- **Statistics**:
  - focus on **estimating** and **interpreting parameters**
  - very little focus on prediction
  - even less on decisions under uncertainty

- **Machine learning**:
  - focus on **prediction**
  - automated **decisions** are central

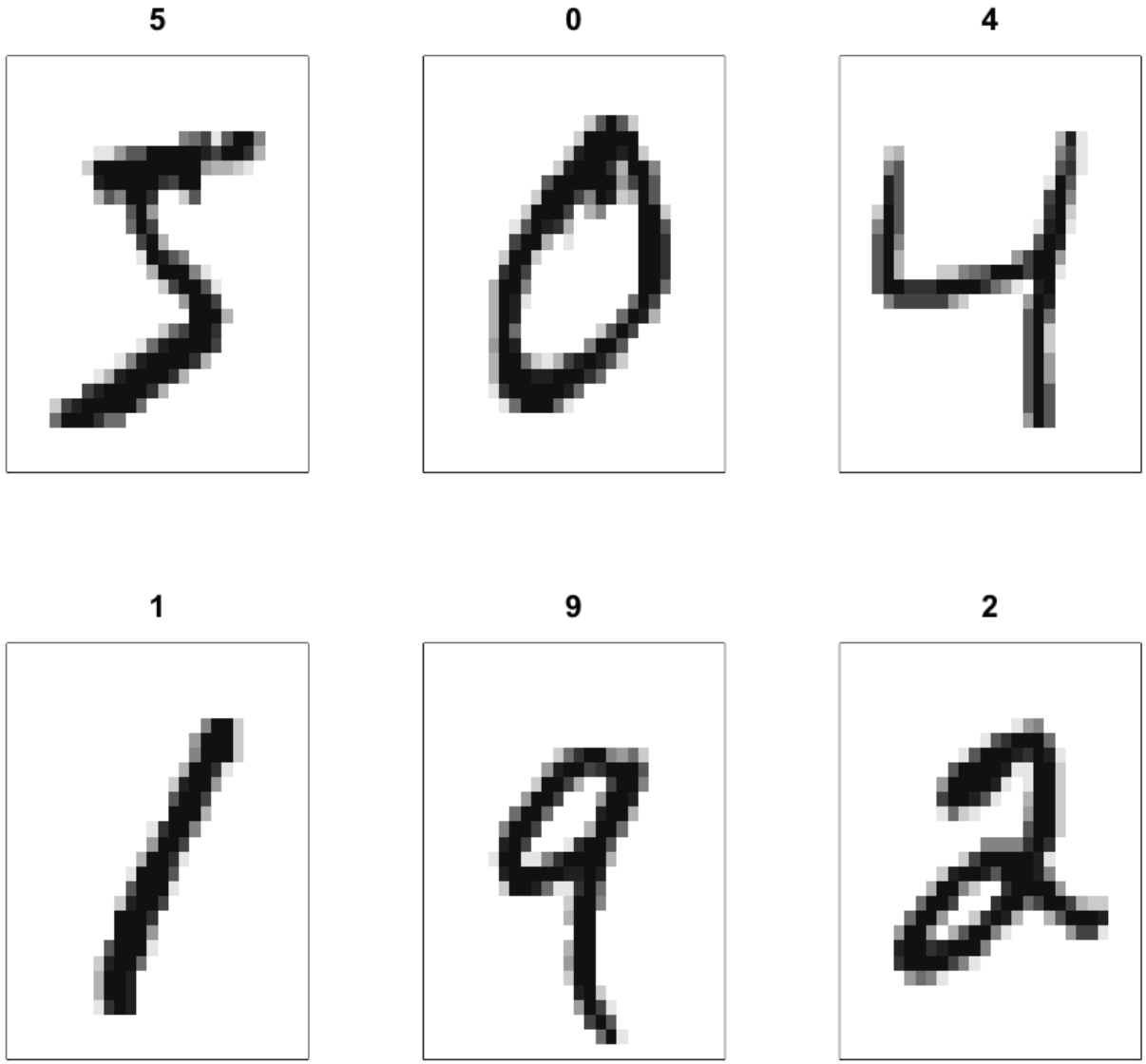
Leo Breiman's (2001) "Statistical Modeling: The Two Cultures." An energetic and passionate argument for the "algorithmic culture" ... Breiman turned out to be more prescient than me: ***pure prediction algorithms have seized the statistical limelight in the twenty-first century*** *...*

Brad Efron, "Prediction, Estimation, and Attribution", Discussion paper JASA, June 2020.

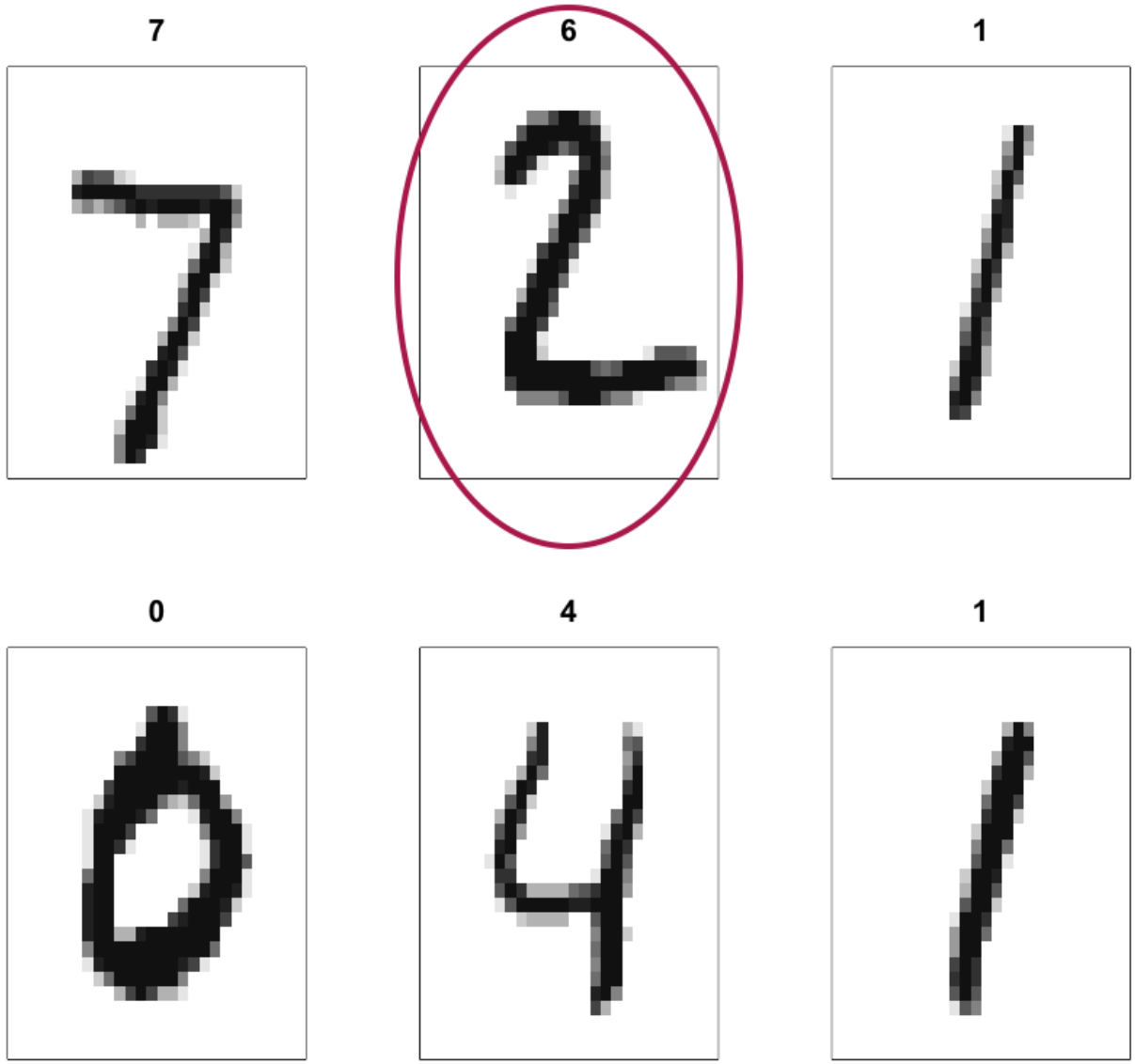
Let's keep the discipline, but shift our focus

- Not a fan: 'pure prediction algorithms', e.g. random forests.
- Flexible probabilistic models for prediction and decisions.
- Attribution and Causality are absolutely crucial
  ... but that rarely means $H_0 : \theta \neq 0$.
- Which factors are important for predictions and decisions?
- Statistics should insist on using probability models and well-founded inference methods...
- ... but we need to shift focus to stay relevant
- ... especially when data and computing is everywhere.

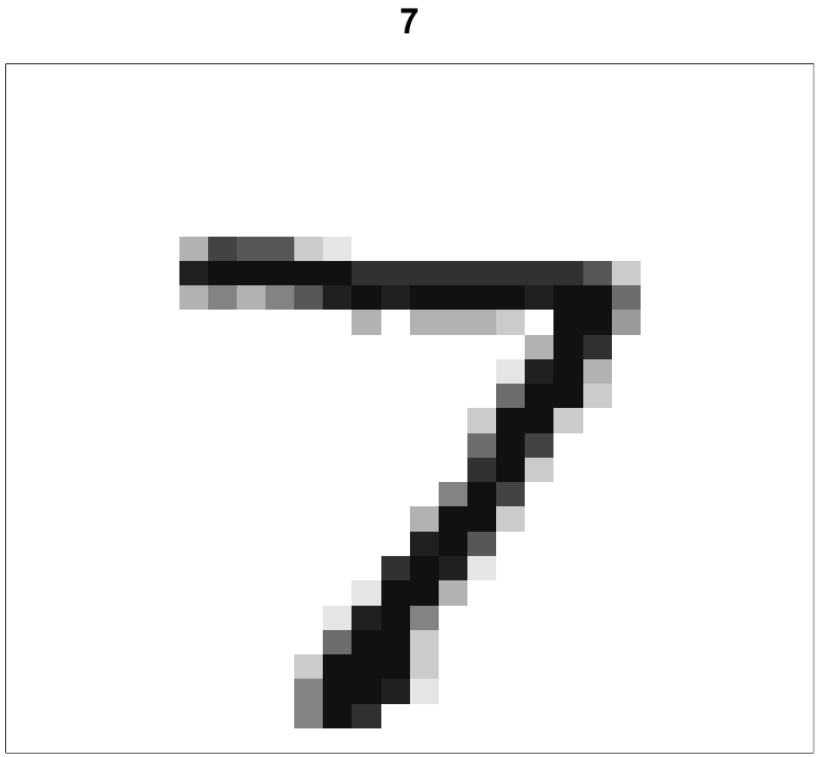
Prediction and decisions are great for teaching



Predicting new handwritten digits



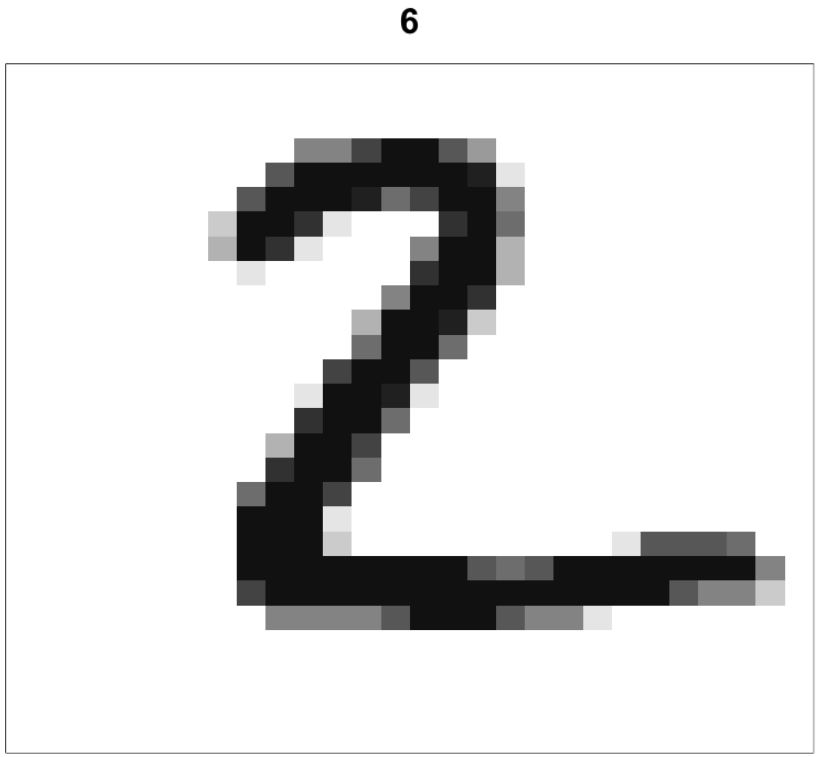
Probabilities and uncertainty



- The system is **very certain** about this prediction

	0	1	2	3	4	5	6	7	8	9
Probability				1%				99%		

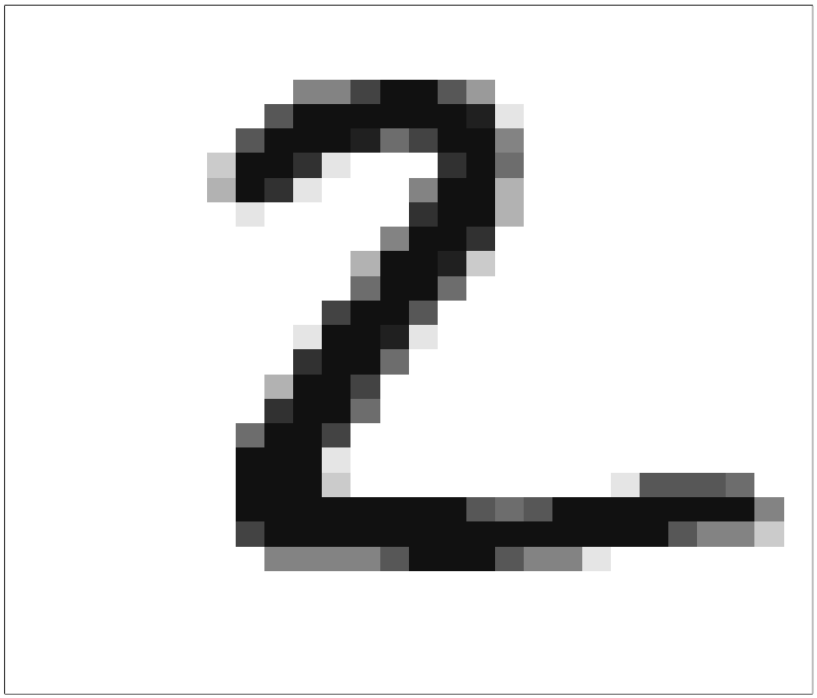
Probabilities and uncertainty



- The system is **very uncertain** about this prediction

	0	1	2	3	4	5	6	7	8	9
Probability	10%		32%			7%	46%		2%	

More training data improves performance



- The system is now getting it right

	0	1	2	3	4	5	6	7	8	9
Probability	6%		64%	4%		5%	20%		1%	

1. Prediction and decision

- Above example **motivates the study of**:
  - Probability and probability models
  - Inference
  - Probabilistic prediction
  - Decision making
- Complement/replace with example from students' domain.
- Changes in statistics education:
  - **Much larger focus on prediction**. Early, for motivation.
  - Compare models' **predictive performance**. Cross-validation.
  - **Regression early** and a lot of it.
  - **Less tests**. A lot less.
  - More focus on **decision making** as the **final aim**.

2. Flexible models and regularization

- **Flexible models**
  - **Richly parametrized mean**/moments.

    Example: $\mathbb{E}(y|\mathrm{x})$ modelled by spline basis/neural net/tree.

    Example: $\mathbb{V}(y|\mathrm{x})$ is a flexible function of x.
  - **Flexible functional form**

    Example: $p(y|\mathrm{x})$ is a mixture density

    Example: $p(\mathrm{y}|\mathrm{x})$ is given by a multivariate copula.
- Overfitting is clearly an issue.
- But linear models may underfit.
- **Regularization**: flexible models, but shrink parameters.

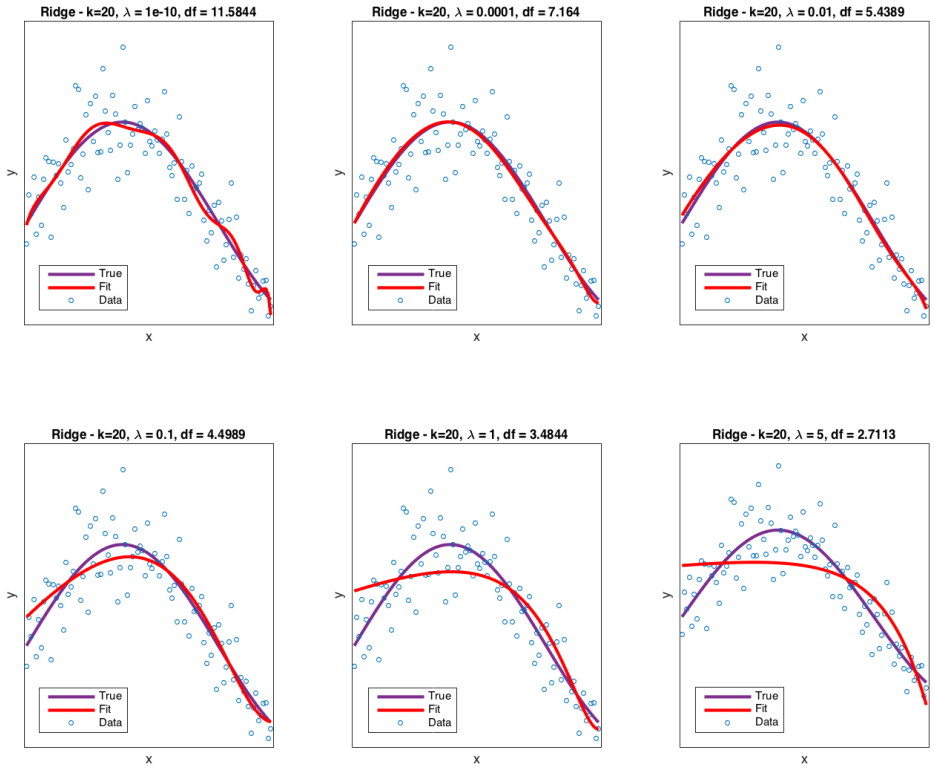
2. Flexible models and regularization

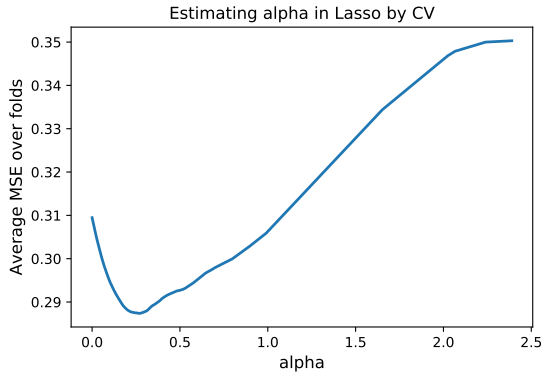
- Model: $y = f_\theta(x) + \varepsilon, \quad \varepsilon \overset{\text{iid}}{\sim} N(0, \sigma^2)$.
- Maximum likelihood with **L2-regularization** minimizes

$$\sum_{i=1}^{n} (y_i - f_\theta(x_i))^2 + \lambda \|\theta\|^2$$

- Linjär modell: $f_\theta(x_i) = x^\top \theta$, then $\hat{\theta}_{\text{RR}} = (X^\top X + \lambda I_n)^{-1} X^\top y$.
- **L1-regularization**: **Lasso**. Shrinkage and variable selection.
- Large λ: **Effective flexibility** $\ll$ nominal number of θ's.
- Linear fit: $\hat{y} = \text{H}y$. **Degrees of freedom** of a model: $\text{tr}(\text{H})$.
- Problems with $n < p$ is solved by regularization.

L2-regularization on polynomial of order 20





2. Flexible models and regularization

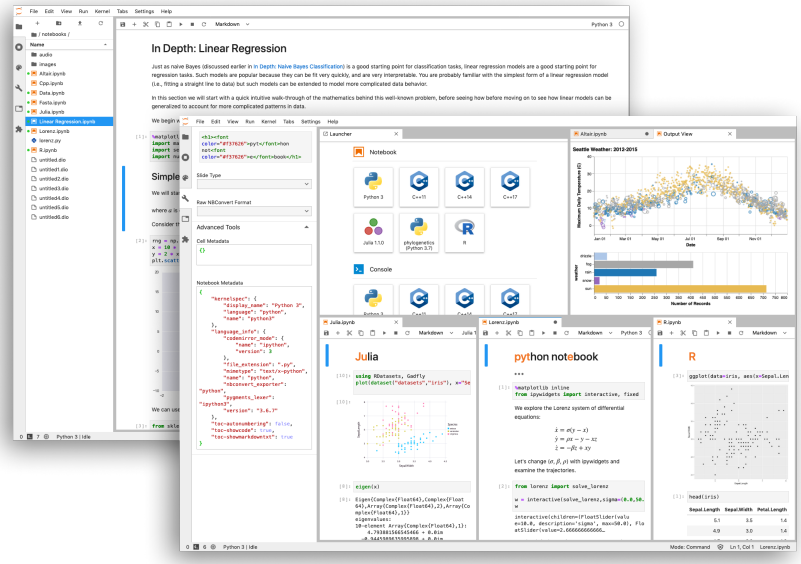
- Let's teach about:
  - **Regularization**. Effective flexibility $\neq$ Nominal flexibility
  - **Predictive performance**. **Cross-validation**. **Overfitting**.
  - **Interpreting nonlinear models**.
  - **Interpretable predictions** and **decisions**.
  - **Modern variable selection**. Lasso and beyond.

3. Computations matter - programming

- My current advice:
  - **learn R** well
  - learn **basics of Python** (and/or Julia!)
  - learn **interoperability** (how to call Python code from R etc)
- **Hallmarks of Science**:
  - transparent
  - open to critique
  - reproducible
  - sharing knowledge
- So universities should teach with **open source software**.

3. Competing with data scientist - tools matter

- **JuPyteR notebooks**: interactive documents with
  - **text** (markdown)
  - **math** (LaTeX)
  - **code** (>40 languages)
- Great for teaching, exercises and computer exams.



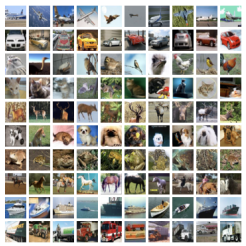
- Package development, Version control (Git), Testing, Profiling.
- Change we must: (R) **programming early**. More **tools**.

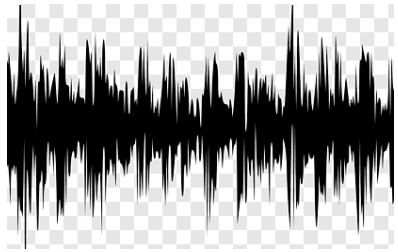
4. Computational efficiency and large-scale data

- Statistician's **strength**: **statistical efficiency**.
- Statistician's **weakness**: **computational efficiency**.
- **Computational complexity** of a method is often crucial.
- **Large-scale (big) data**
  - Memory access and storage
  - Streaming data. Real-time constraints.
  - **Scalable statistical methods**
- Ch-ch-changes:
  - Teach awareness that **computations matters**.
  - **Analysis of algorithms** from computer science.
  - **Performant code, parallel computing** and all that.

5. New data types, noise and anomalies

- Modern data are **not tables**:
  - Images
  - Text
  - Computer generated log files
  - Sound bites

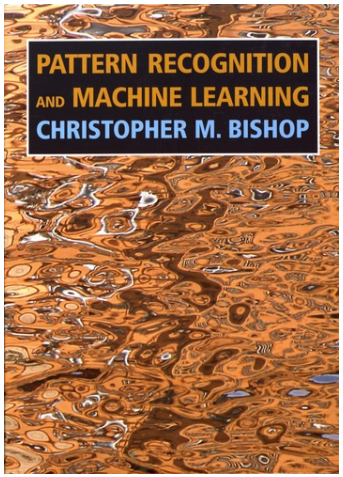


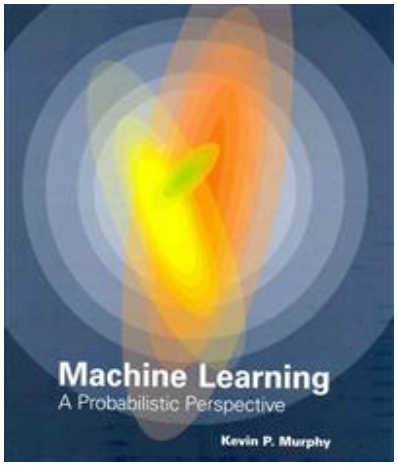


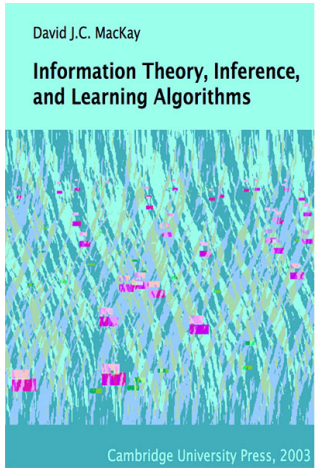
- Let's get messy:
  - More **examples** and exercises with **non-standard data** types
  - **Case studies** with messy, noisy, and slightly inappropriate data
  - **Data access tools** and **pre-processing** pipelines

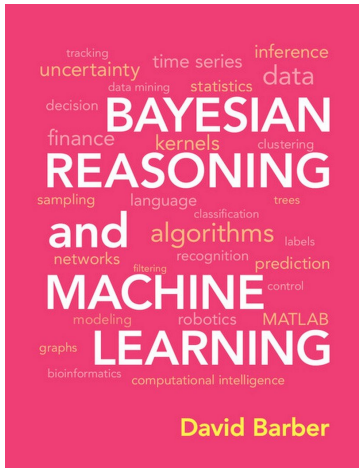
6. and yes, Bayes is popular in Machine Learning

- Many popular course books in ML take a Bayesian approach.









- Why is Bayes so popular in ML?
  - More **engineering** than pure science. **Subjective** is OK.
  - Using more information (**priors**) is better.
  - **Prediction** and **decision-making** are natural for Bayes.
  - Bayes has **regularization** built in (smoothness priors).
  - **Algorithmic** implementations (MCMC etc).
- **Deep learning** - Bayes to **quantify uncertainty**.

6. Bayes is popular in applied statistics

- Current issue of **JASA Applications & Case Studies**:

6. Bayes is popular in applied statistics

- Current issue of **Annals of Applied Statistics**:

ISSN 1932-6157 (print)
ISSN 1941-7330 (online)

THE ANNALS
of
APPLIED
STATISTICS

An Official Journal of the
Institute of Mathematical Statistics

Articles

SHOPPER: A probabilistic model of consumer choice with substitutes and complements . . . . . Francisco J. R. Ruiz, Susan Athey and David M. Blei 1
BART with targeted smoothing: An analysis of patient-specif c stillbirth risk . . . . . Jennifer E. Starling, Jared S. Murray, Carlos M. Carvalho, [illegible] 29
Integrative survival analysis with uncertain event times in application to a suicide risk study . . . . . Wenjie Wang, Robert Aseltine, Kun Chen and Jun Yan 51
Eff cient real-time monitoring of an emerging inf uenza pandemic: How feasible? . . . . . Paul J. Birrell, Lorenz Wernisch, Brian D. M. Tom, Leonhard Held, Gareth O. Roberts, Richard G. Pebody and Daniela De Angelis 74
Modeling microbial abundances and dysbiosis with beta-binomial regression . . . . . Bryan D. Martin, Daniela Witten and Amy D. Willis 94
A statistical analysis of noisy crowdsourced weather data . . . . . Arnab Chakraborty, Soumendra Nath Lahiri and Alyson Wilson 116
Surface temperature monitoring in liver procurement via functional variance change-point analysis . . . . . Zhenguo Gao, Pang Du, Ran Jin and John L. Robertson 143
Assessing wage status transition and stagnation using quantile transition regression . . . . . Chih-Yuan Hsu, Yi-Hau Chen, Ruoh-Rong Yu and Tsung-Wei Hung 160
TFisher: A powerful truncation and weighting procedure for combining *p*-values . . . . . Hong Zhang, Tiejun Tong, John Landers and Zheyang Wu 178
Modifying the Chi-square and the CMH test for population genetic inference: Adapting to overdispersion . . . . . Kerstin Spitzer, Marta Pelizzola and Andreas Futschik 202
A hierarchical Bayesian model for predicting ecological interactions using scaled evolutionary relationships . . . . . Mohamad Elmasri, Maxwell J. Farrell, T. Jonathan Davies and David A. Stephens 221
Bayesian factor models for probabilistic cause of death assessment with verbal autopsies . . . . . Tsuyoshi Kunihama, Zehang Richard Li, Samuel J. Clark and Tyler H. McCormick 241
Estimating the health effects of environmental mixtures using Bayesian semiparametric regression and sparsity inducing priors . . . . . Joseph Antonelli, Maitreyi Mazumdar, David Bellinger, David Christiani, Robert Wright and Brent Coull 257
Feature selection for generalized varying coeff cient mixed-effect models with application to obesity GWAS . . . . . Wanghuan Chu, Runze Li, Jingyuan Liu and Matthew Reimherr 276
Optimal asset allocation with multivariate Bayesian dynamic linear models . . . . . Jared D. Fisher, Davide Pettenuzzo and Carlos M. Carvalho 299

Continued on back cover

Vol. 14, No. 1—March 2020

ISSN 1932-6157 (print)
ISSN 1941-7330 (online)

THE ANNALS
of
APPLIED
STATISTICS

An Official Journal of the
Institute of Mathematical Statistics

Articles—*Continued from front cover*

Modeling wildf re ignition origins in southern California using linear network point processes . . . . . Medha Uppala and Mark S. Handcock 339
Regression for copula-linked compound distributions with applications in modeling aggregate insurance claims . . . . . Peng Shi and Zifeng Zhao 357
Estimating and forecasting the smoking-attributable mortality fraction for both genders jointly in over 60 countries . . . . . Yicheng Li and Adrian E. Raftery 381
Measuring human activity spaces from GPS data with density ranking and summary curves . . . . . Yen-Chi Chen and Adrian Dobra 409
A comparison of principal component methods between multiple phenotype regression and multiple SNP regression in genetic association studies . . . . . Zhonghua Liu, Ian Barnett and Xihong Lin 433
Estimating causal effects in studies of human brain function: New models, methods and estimands . . . . . Michael E. Sobel and Martin A. Lindquist 452
A hierarchical dependent Dirichlet process prior for modelling bird migration patterns in the UK . . . . . Alex Diana, Eleni Matechou, Jim Griffin [illegible] 473
Bayesian mixed effects models for zero-inf ated compositions in microbiome data analysis . . . . . Boyu Ren, Sergio Bacallado, Stefano Favaro, Tommi Vatanen, Curtis Huttenhower and Lorenzo Trippa 494

Correction

Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects . . . . . Trang Quynh Nguyen and Elizabeth A. Stuart 518

Vol. 14, No. 1—March 2020

Summary

- **Prediction** and **decision making**
- **Regularization**
- **Programming** and computational tools matter
- **Computational complexity** is a statistical issue
- **Modern data** are not nice tables.
- (Bayes)

Jo, jag menar det!