

# Statistik möter datavetenskap - erfarenheter från Linköpings universitet

**Mattias Villani**

**Avdelningen för statistik och maskininläring  
Institutionen för datavetenskap  
Linköpings universitet**

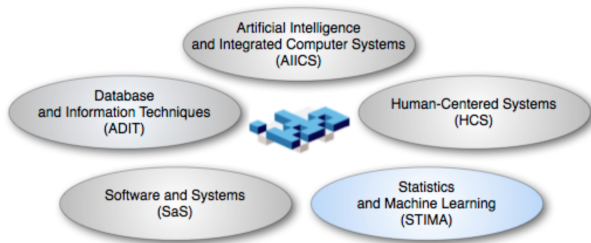
# Översikt

- ▶ Organisation
- ▶ Undervisning
- ▶ Forskning

Slides: <https://github.com/mattiasvillani/Talks/raw/master/MatStat.pdf>

# Statistikämnet tillhör datavetenskaplig institution

- **Avdelningen för statistik och maskininläring** är sedan 2008 en av fem avdelningar vid institutionen för datavetenskap.



- *Frontiers in Massive Data Analysis (US National Research Council):*

*“**Computer scientists** involved in building big-data systems must develop a deeper awareness of inferential issues, while **statisticians** must concern themselves with scalability, algorithmic issues, and real-time decision-making.”*

# Masterprogrammet Statistics and Machine Learning

- ▶ 2-årigt internationellt **masterprogram i statistik** med start 2008.
- ▶ **Studenter** från statistik, datavetenskap, ingenjörsvetenskaper, tillämpad matematik.
- ▶ Försiktig början med större inslag av data mining- och databaskurser från andra avdelningar.
- ▶ Under senare år mer **probabilistiska modeller** och **likelihoodbaserade metoder**.
- ▶ **Prediktivt fokus.**
- ▶ **Programmering. Datorlaborationer** och **datortentor.**
- ▶ Kurser samläses med **masterprofil inom AI och Maskininlärning på ingenjörsprogram.**

# Kurs: Sannolikhetslära och statistik, 6 hp

- ▶ **Grundläggande kurs** i Sannolikhetslära och statistik för **ingenjörer**.
- ▶ Sannolikhetslära + Inferens + **Prediktion** + **Beslut**
- ▶ Både **frekventistisk** och **Bayesiansk inferens**
- ▶ Tre **regjäla datorlaborationer**. Simulering för att förstå teorin.
- ▶ **Exempel från maskininlärning/AI** från dag 1 för motivationen.
- ▶ Kurssida:  
<https://www.ida.liu.se/~TDAB01/info/courseinfo.sv.shtml>.

# Machine learning, 9 hp

- ▶ Avancerad nivå.
- ▶ Bred översiktskurs.
- ▶ Moment:
  - ▶ Basic Concepts in Machine Learning
  - ▶ Regression, Regularization and Model Selection
  - ▶ Classification Methods
  - ▶ Dimensionality Reduction and Uncertainty Estimation
  - ▶ Kernel Methods and Support Vector Machines.
  - ▶ Neural Networks and Deep Learning
  - ▶ Model Inference and Variable Selection
  - ▶ Ensemble Methods and Mixture Models
  - ▶ Online Learning
  - ▶ Splines and Additive Models
  - ▶ High-Dimensional Problems
- ▶ Kurssida:  
<https://www.ida.liu.se/~732A95/info/courseinfo.en.shtml>.

# Bayesian learning, 6 hp

- ▶ Avancerad nivå.
- ▶ **Fyra moment:**
  - ▶ The Bayesics
  - ▶ Bayesian Regression and Classification
  - ▶ Bayesian Computations: MCMC and Variational Bayes
  - ▶ Model Inference and Variable Selection
- ▶ **Bayesiansk inferens** passar ML:
  - ▶ **Prediktion** och **beslut** på ett naturligt sätt
  - ▶ **Simuleringsvänligt** (MCMC etc)
  - ▶ **Regularisering** av flexibla icke-linjära modeller via mjukhetspriors
- ▶ Kombinerad **dator- och papperstenta**.
  - ▶ Tenta i datorsal, med speciellt mjukvarusystem för datortenta.
  - ▶ 3/4 löses med dator, 1/4 löses med papper och penna.
  - ▶ Studentens labbrapporter finns tillgängliga under tentan.
- ▶ Kurssida: <https://www.ida.liu.se/~732A91/info/courseinfo.en.shtml>.

# Advanced Machine learning, 6 hp

- ▶ Avancerad nivå.
- ▶ Djupdykning i mindre antal probabilistiska modeller. Bayes.
- ▶ Moment:
  - ▶ Hidden Markov Models
  - ▶ State Space Models
  - ▶ Graphical Models and Bayesian Networks
  - ▶ Gaussian Process Regression and Classification
- ▶ Kurssida: <https://www.ida.liu.se/~732A96/info/courseinfo.en.shtml>.



# Gaussiska processer för ML

- ▶ Bok: **Gaussian Processes for Machine Learning** (Rasmussen-Williams).
- ▶ Innehåll:
  - ▶ Resultat om **multivariat normal** (täthet, marginella och betingade fördelningar, linj.transf).
  - ▶ **Definition Gaussisk process** (GP) som sannolikhetsfördelning över funktioner.
  - ▶ **Regression** med GPs
  - ▶ **Klassifikation** med GPs
  - ▶ **Optimering** med GPs.
  - ▶ Numeriskt stabil implementation av GPs.

# Text Mining, 6 hp

- ▶ Avancerad nivå.
- ▶ Samarbete mellan STIMA, datorlingvistik och databasgruppen.
- ▶ Hela pipeline:
  - ▶ **web-scraping**
  - ▶ **linguistisk pre-processing**
  - ▶ **probabilistisk modellering.**
- ▶ Moment:
  - ▶ Information Retrieval
  - ▶ Natural Language Processing
  - ▶ Statistical Analysis of Textual Data
- ▶ Kurssida: <https://www.ida.liu.se/~732A92/courseinfo.en.shtml>.

# Forskning

## ► Grundtema:

- statistisk analys baserat på sannolikhetsmodeller
- med fokus på prediktion och beslutsfattande
- genom effektiva skalbara beräkningar för
- stora komplexa datamängder.

## ► Ex på STIMA-publikationer under senaste två åren:

- *Journal of Computational and Graphical Statistics* (3 st)
- *Journal of Machine Learning Research*
- *IEEE Transactions on Pattern Analysis and Machine Intelligence*
- *Journal of the American Statistical Association*
- *Annals of Applied Statistics*
- *Proceedings of National Academy of Sciences (PNAS)* (2 st).
- *NeuroImage* (3 st)
- *Human Brain Mapping*

# Övrig forskningsverksamhet inom ML

- ▶ STIMA leder **IDA Machine Learning Research Group** vid institutionen för datavetenskap (IDA).  
<https://liu.se/machinelearning/>
- ▶ **IDA Machine Learning Seminars**. STIMA-ledd internationell månatlig seminarieriserie i maskininlärning med världsledande forskare. <https://liu.se/machinelearning/seminars>
- ▶ **LiU Seminars in Statistics and Mathematical Statistics**.  
Gemensam seminarieriserie tillsammans med MatStat.
- ▶ **AI, Autonomous Systems and Software Program (WASP)**.
  - ▶ **WASP-doktorand**. *Bayesian Learning for Spatio-Temporal Models in Transportation*.
  - ▶ **WASP-doktorand**. *Methods for Scalable and Safe Robot Learning*.
  - ▶ **WASP industridoktorand** tillsammans med Ericsson Research. *Machine Learning for 5G System – Control and Automation*.
- ▶ Medlem av styrelsen för **Nationellt Superdatorcentrum**.

# Vad krävs för att närma sig ML? Vill vi?

- ▶ Statistiker måste vara **genuint intresserade** av skalbara beräkningar och algoritmer (och inte bara av AI-miljarder).
- ▶ **Helhetssyn** på
  - ▶ sannolikhetsmodeller
  - ▶ inferens
  - ▶ **prediktion** och **beslut**
  - ▶ **skalbarhet** och **beräkningseffektivt**.
- ▶ **Forskningsverksamhet** inom computational statistics och maskininlärning. ML-tillämpningar i undervisning.
- ▶ **Behålla vår statistiska integritet**, men stå ut med att hoppa över vissa eleganta härledningar. Inte alla egenskaper måste utforskas.
- ▶ Mer fokus på **prediktiv inferens**, mycket mindre på hypotestest och deras asymptotiska egenskaper.
- ▶ **Ett bayesiansk perspektiv** underlättar.