Bridging the gap between data-centric disciplines

An undergraduate education perspective

Mattias Villani

Department of Statistics
Stockholm University

Department of Computer and Information Science
Linköping University

Never tweet late in the night

Mattias Villani @matvil · Oct 21

I will regret this slide in the morning.

🔥🔥 Machine Learning vs Traditional Statistics 🔥🔥

	Stats	ML
Parameter inference	😍	😴
Prediction	😐	😍
Decision making	😳	😍
Interpreting models	😍	🙄
Interpreting decisions	😳	😘
Flexible models and regularization	😨	😍
Rigorous theory	😍	😜
Causality	😐	😐
Programming	😥	😎

40 342 1.4K

Caricatures to initiate discussions about education

- Mostly everyone's response: 😍 or 😂 or at least 😁.
- Very few 😠 , and then: **I** am an X, and **I** certainly know Y!
- Good **caricatures** are recognizable and have some truth.
- The tweet originated from thoughts about **basic courses in data-centric subjects**.
- How can we **bridge the gap** between subjects?

Who is this emoji clown? 🧔

- **BSc, MSc, PhD in Statistics** (2000), Stockholm University.
- Researcher and Adviser at Central Bank of Sweden (2003-11).
- Since then: neuroimaging, text, robotics, transportation.
- **Professor of Statistics** at LiU (2011-) and SU (2018-).
- Developed many statistics **and** machine learning courses: Machine learning, Bayesian learning, Advanced ML, Text mining, ML for industry, Statistics for engineers etc.
- Built up and **lead a statistics and machine learning division** in a **computer science department**.

"I develop computationally efficient Bayesian methods for inference, prediction and decision making with flexible probabilistic models."

🔥🔥 Machine Learning vs Traditional Statistics 🔥🔥

	Stats	ML
Parameter inference	😍	😴
Prediction	😐	😍
Decision making	😳	😍
Interpreting models	😍	🙄
Interpreting decisions	😳	😘
Checking model assumptions	😎	🤭
Flexible models and regularization	😨	😍
Rigorous theory	😍	😜
Causality	😐	😐
Programming	😥	😎
Scalability, big data	😱	😎
Real-time/Online	🤭	🤠
Data collection and experimental design	😎	🤭

Parameter inference Stats: 😍 ML: 😴

- Naming of regression coefficients (Neil Lawrence):
  - **Stats**: fancy greek letters like β. **Parameters are key players**.
  - **ML**: weights w. Parameters are **just weights in prediction**.
- Interpreting parameters $\Longrightarrow$ focus on linear models.
- **Feature construction is hard** for modern problems.
- **Nonlinear models** and methods for interpreting $\frac{\partial \mathbb{E}(y|\mathrm{x})}{\partial x_j}$.
- The anticlimax of stats: testing for uninteresting $H_0 : \theta = 0$.
- Let us at least focus on effect sizes.

Prediction

Stats: 😐 ML: 😍

- ML:
  - **Prediction** is **the** aim.
  - Models are evaluated by **predictive performance**.
  - **Training-Validation-Test** split of the data is standard.
  - **Regularized parameter-rich models** best for prediction.
- Stats:
  - Prediction is **essentially only in time series** courses.
  - Not much used for model selection (Box-Jenkins still rules).
  - Prediction is a tiny part of regression courses.
  - Almost no discussion about:
    - **generalization performance**
    - **bias-variance trade-off** for predictions
    - **cross-validation** and similar methods.

Decision making Stats: 😳 ML: 😍

- Stats (blushing emoji):
  - Decision making under uncertainty is well developed in Stats
  - ... but have forgotten about it, definitely in education.
  - Needed: **Decision making as end goal**.
  - **Bayesian inference** is key here.
- ML:
  - Automatic **decision making is in focus**.
  - Loves decisions, but **often stops short at predictions**.
  - **Uncertainty quantification** is crucial for decisions. Typically ignored in deep learning.
  - **Bayesian inference** is key here.

Interpreting models Stats: 😍 ML: 🙄

- **Stats**: loves **scientifically grounded interpretable models**.
- ML: any **Black Box** with accurate predictions is fine.
- ML debate:
  - very accurate black box

    vs
  - less accurate interpretable model.
- **SciML**:
  - scientifically grounded model (e.g. from physics or economics)

    $+$
  - added flexibility by neural networks.

Interpreting decisions Stats: 😳 ML: 😘

- Stats: interpreting decisions?
- ML: explain **why a decision** was made. **Explainable AI**.
- Hard to explain decisions from black box models.
- ML flirt: **exploring decisions locally interpretable models**.
- Focusing on **decisions gives discipline in modeling**:
  - Will the addition of this model feature affect decisions?
  - "Effect sizes" with respect to decisions.
  - AI. Real-time decisions limit the class of possible models.

Checking model assumptions Stats: 😎 ML: 🤭

- Stats:
  - **heavily trained in checking model assumptions**
  - **understanding of data quality** and its influence on modeling.
  - too much driven by getting correct distributions for tests.
- ML's focus on evaluating models by predictive performance
  - often leads to **neglect in model checking**
  - forgets to explore the data
  - **gives models that are complicated** and hard to interpret.

Flexibility & Regularization   Stats: 😨 ML: 😍

- Stats:
  - **flexibility** is often measured by the **number of parameters**.
  - regularization (e.g. Lasso) to **penalize complexity** is not part of our DNA.
  - Many statisticians develop regularization methods, but ...
    ... it is only presented in '**Statistical Learning**' courses.
  - **overfitting** with flexible models is an **exaggerated fear/concern**.
- ML:
  - Flexible models is the **standard**
  - **Regularization from day one**
  - Connection to **predictive performance** helps here

Rigorous theory Stats: 😍 ML: 😜

- Stats:
  - long history of **strong theory deeply rooted in probability**.
  - empirical success is not as highly valued, **math is king**.
  - models are often simplified to fit the maths.
    Box: '**mathematistry**'.
  - The statistical theory has led to better understanding in ML.
- ML:
  - empirical performance is everything valued, theory is less of a requirement.
  - but has its own body of theory, often proving certain guarantees on the procedure.
  - more **rigorous on algorithmic performance**.
  - ML **needs a probabilistic perspective** to make decisions correctly.

Causality

- Causality is starting to become a hot topic in both fields, at least in research.
- Scientific discoveries must be causal.
- Also robots need to learn cause and effect to learn tasks well. **Learning by interacting with environment**.
- Different approaches in Stats and ML, but perhaps merging?

Programming   Stats: 😥   ML: 😎

- Stats: **programming skills** have not been valued and taught.
- **Software development** virtually non-existent in Stats education.
- Yet, statistical practice is now very relient on these skills.
- Data scientists know useful software engineering methods and **tools** which impresses at job interviews.
- **R and its package culture** has advanced programming skills among statisticians. But it is still a weak point.
- ML is an engineering field where professional programming and software development is valued and taught.

Scalability, big data   Stats: 😱   ML: 😎

- Stats is all about **statistical efficiency** ...
  ... but very little about **computational efficiency**.
- **Big datasets** are still scary.
- **Non-tabular data** (image, text, sound) are rare in stats education.
- Needed in **stats education**:
  - basic notions about **computational complexity** of algorithms
  - **numerical maths**, e.g. computer arithmetic, lin alg, sparsity
  - handling **large datasets**.
  - handling **messy datasets**.
- **ML**:
  - often extensive training in **algorithms** and their **complexity**.
  - databases and big data frameworks
  - often at least one course in numerical methods
  - need to learn that **data quality also matters**.

Real-time/Online Stats: 🤭 ML: 🤠

- Stats:
  - rarely even thinks about **real-time requirements**.
  - **online learning** = Kalman filter ... at best.
  - **reinforcement learning** in ML is modeled by **Markov decision processes**. That is statistics, but never taught.
- ML:
  - physical systems (e.g. robots) make decisions in real-time.
  - a lot of research on online real-time learning.
  - not always really real-time for real machines in real world.
  - perpetually young field since (embedded) hardware develops rapidly.

Data & experimental design   Stats: 😎   ML: 🫢

- Stats:
  - **design of experiments** is an old speciality.
  - **survey sampling** studies **data collection**.
  - **data quality** is well understood.
  - **GIGO** well understood.
- ML:
  - massive data in industry, but often collected for other reasons.
  - collected **data** often **unable to answer relevant questions**.
  - poor data quality, and **poor understanding of GIGO**.
  - Statistics/ML is not a magic wand.
- Statisticians in experimental design and survey sampling need to get involved in ML problems.
- **active learning** = sequential experimental designs.

Unifying data-centric disciplines

- **Suggestions for Statistics education:**
  - More focus on prediction and decisions.
  - (Much) less focus on hypothesis testing.
  - Regression early and a lot.
  - Nonlinear models and regularization early.
  - Downplay unbiasedness. Bias-Variance trade-off.
  - Computational thinking, scalability and real-time problems.
  - Make programming second nature. Tools matter.
  - Bayesian (likelihood) inference is unifying across disciplines.
- **Suggestions for Machine learning/Data Science education:**
  - Data quality and experimental design.
  - Model assessment beyond predictive performance.
  - Probability distributions for data instead of cost minimization.
  - Uncertainty quantification for decision making.