41000: Bonus Lecture

Tengyuan Liang

Econometrics and Statistics

CHICAGO BOOTH
The University of Chicago Booth School of Business

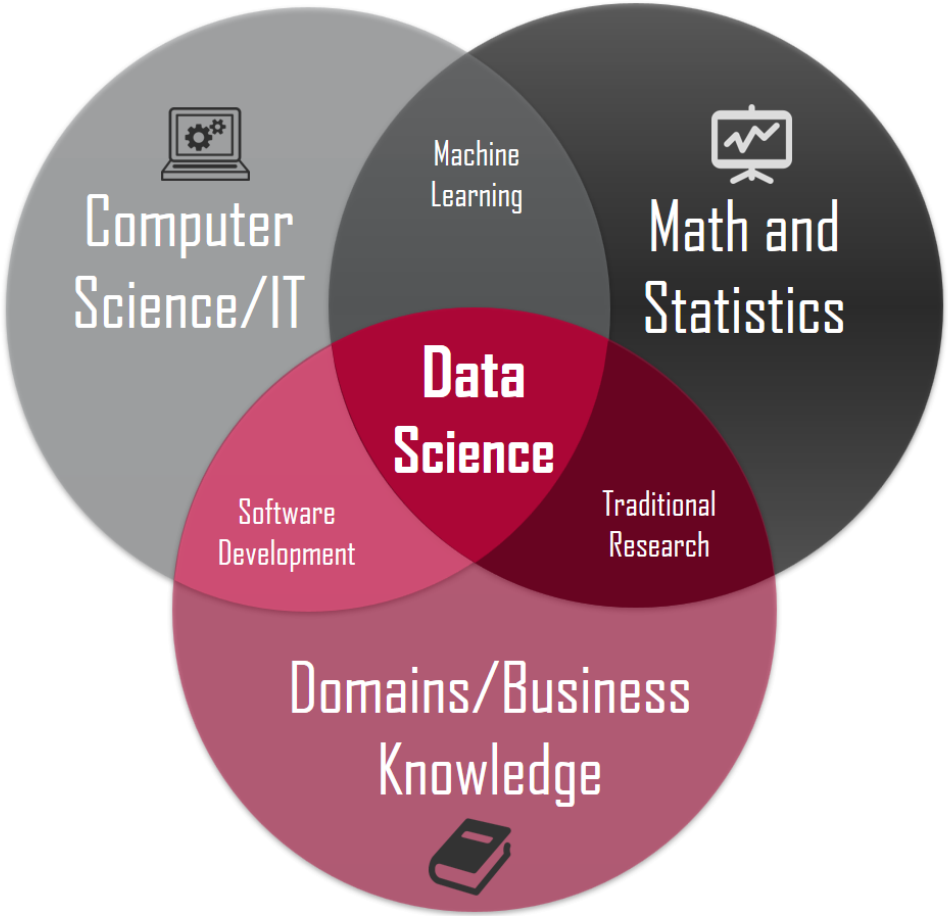
Disclaimer

This document contains images obtained by routine Google Images searches. Some of these images may perhaps be copyright. They are included here for educational noncommercial purposes and are considered to be covered by the doctrine of Fair Use.

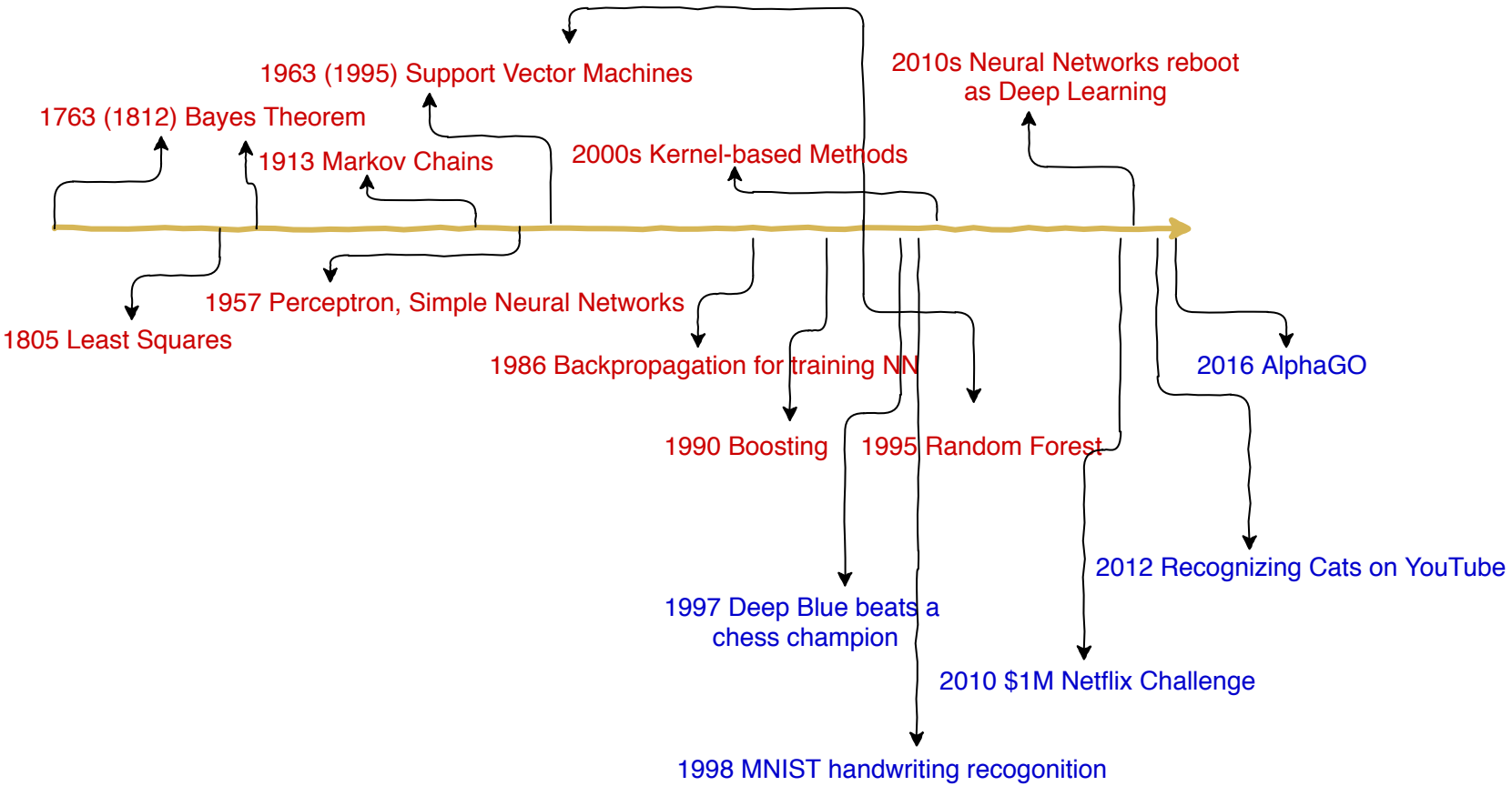
Its not feasible to give full scholarly credit to the creators of these images. We hope they can be satisfied with the positive role they are playing in the educational process.

A LONG WAY TO GO BETWEEN STAT 101 AND RESEARCH

A LONG WAY TO GO BETWEEN STAT 101 AND RESEARCH



Brief history



	from input ***x***, output:
unsupervised	summary ***z***
supervised	prediction ***y***
reinforcement	action ***a*** to maximize reward ***r***

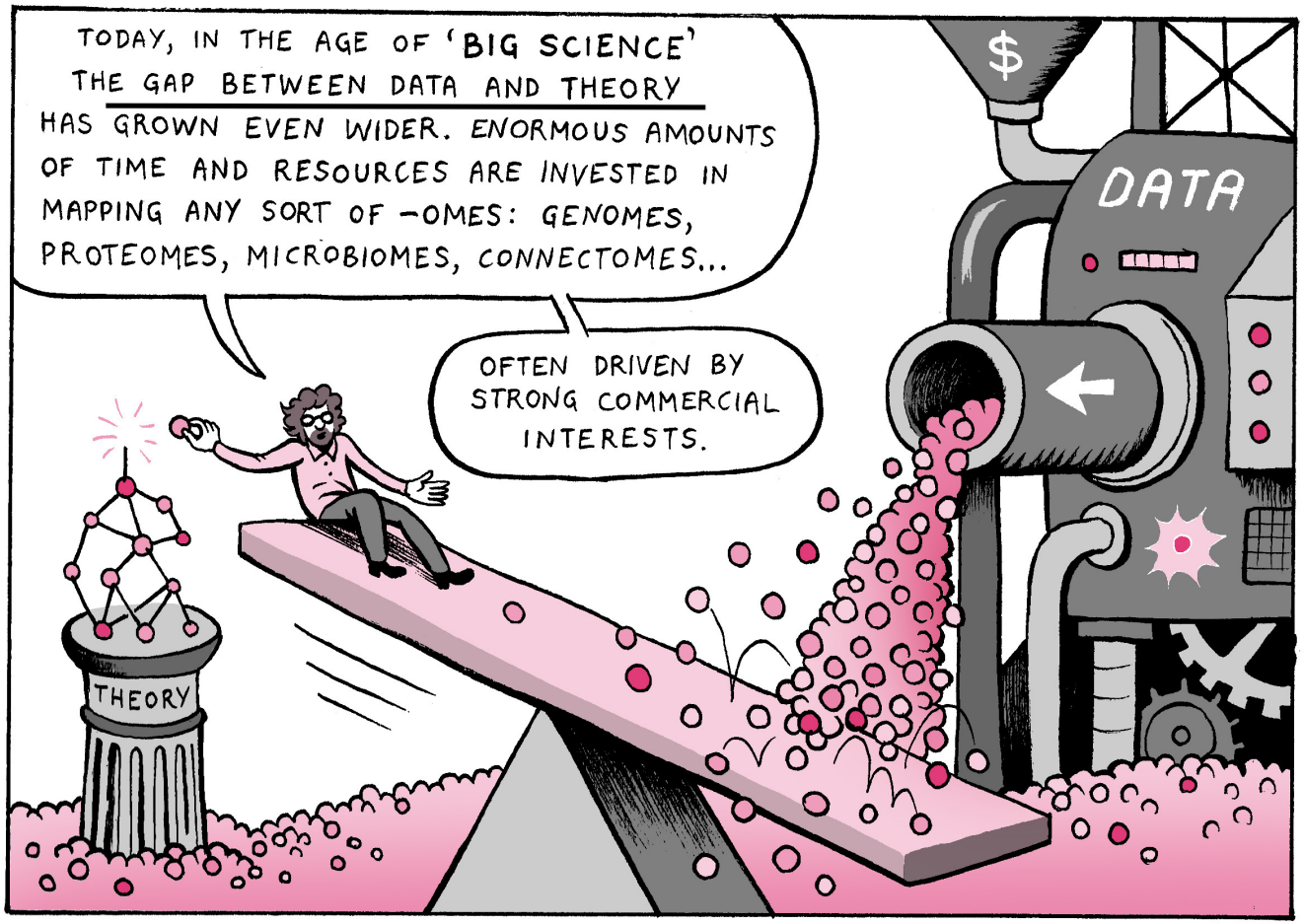
Picture from Ben Recht's blog.

- Supervised learning: regression tasks, learning a function that maps an input to an output based on example (x, y) pairs
  - classification
  - regression

- Unsupervised learning: structure/knowledge discovery, identifies commonalities in the data
  - distributions of data
  - clustering
  - latent factors
  - network structures

- Reinforcement learning: take actions in an environment so as to maximize some notion of cumulative reward

What I do?


TODAY, IN THE AGE OF 'BIG SCIENCE' THE GAP BETWEEN DATA AND THEORY HAS GROWN EVEN WIDER. ENORMOUS AMOUNTS OF TIME AND RESOURCES ARE INVESTED IN MAPPING ANY SORT OF -OMES: GENOMES, PROTEOMES, MICROBIOMES, CONNECTOMES...
OFTEN DRIVEN BY STRONG COMMERCIAL INTERESTS.
$
DATA
THEORY

Deep learning without theoretical understanding?

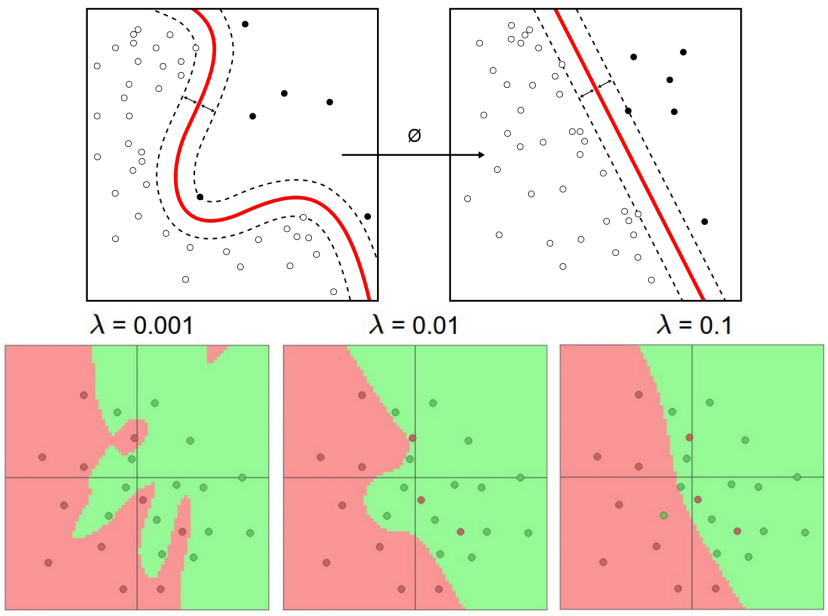
Picture from Ali Rahimi's talk.

Regression and Classification

$$y = f(x)$$

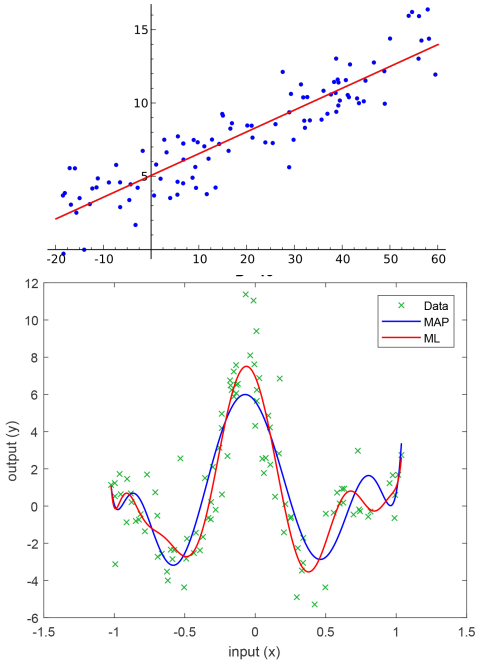
Classification

document classification, spam filtering,
image/handwriting recognition ...



Regression

stock market price, predict age of viewer
watching a given video on YouTube ...



Model complexity

Model complexity of f matters: simpler model may not be as powerful (large bias), but it is more robust (less variance).

Model selection: choose the complexity such that average error of unseen/future data – generalization error – is small.

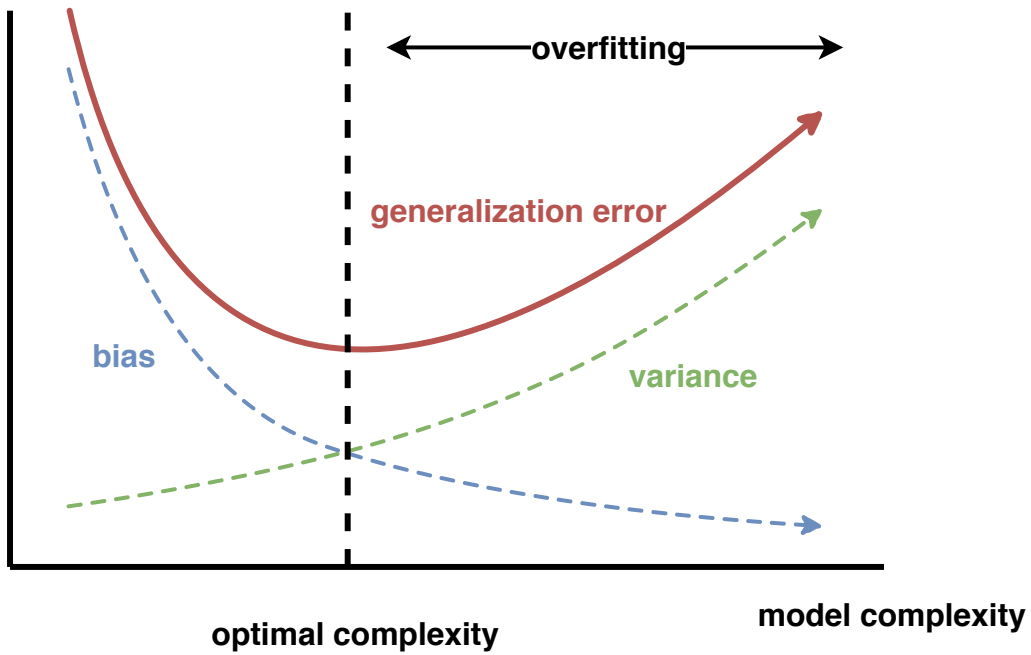
Model complexity

Model complexity of f matters: simpler model may not be as powerful (large bias), but it is more robust (less variance).

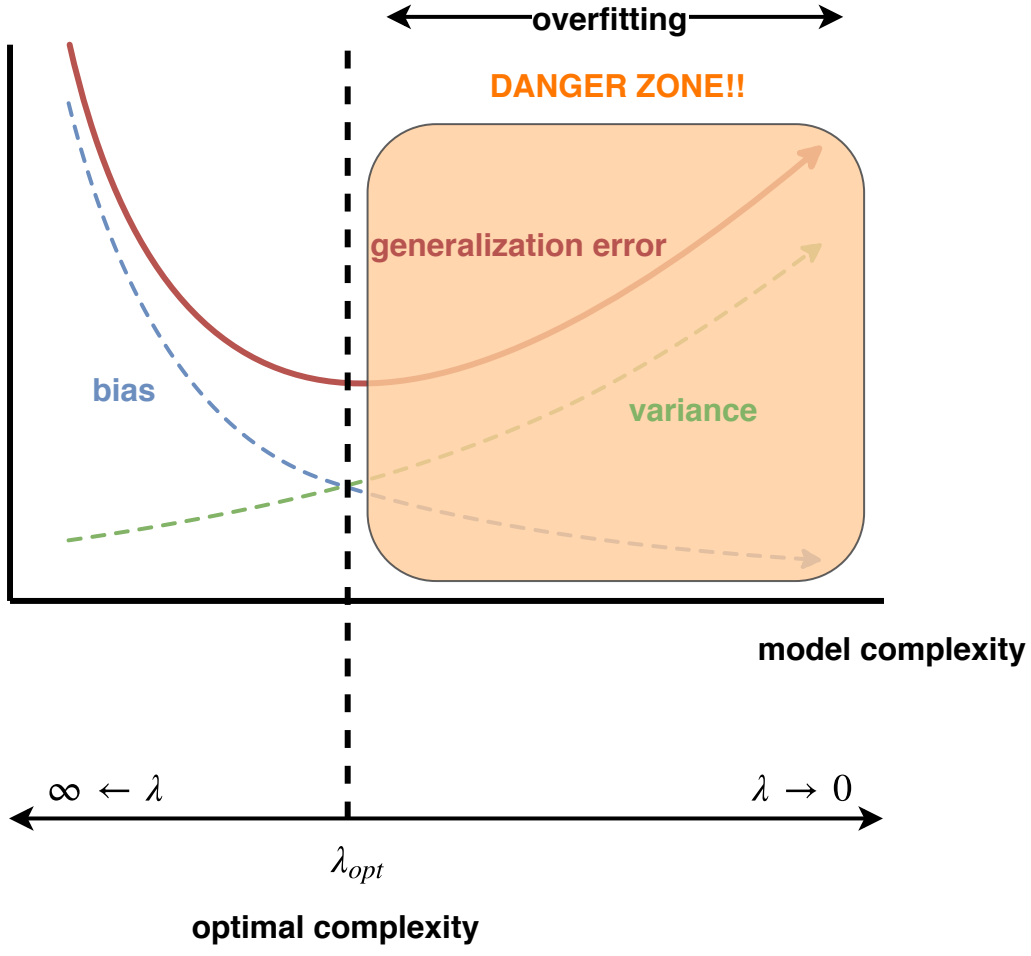
Model selection: choose the complexity such that average error of unseen/future data – generalization error – is small.

generalization error = bias + variance

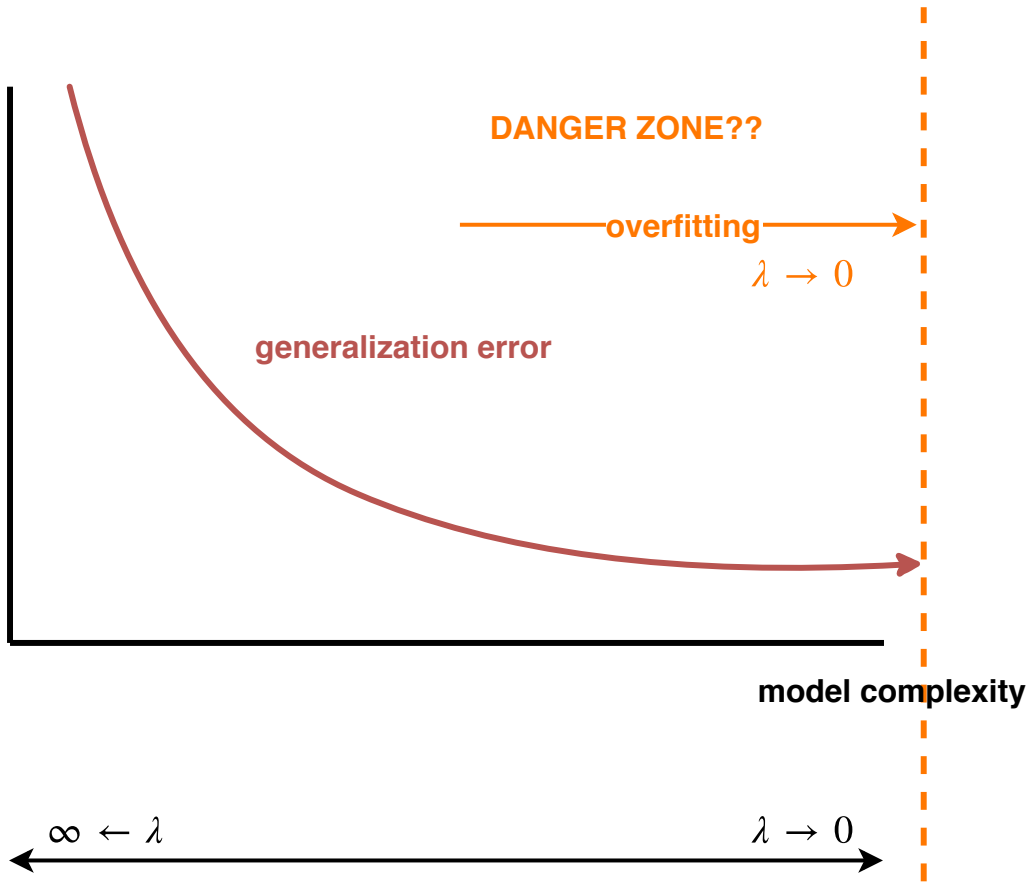
How do we teach stat/ml?



How do we teach stat/ml?



Is this really what's happening in practice?

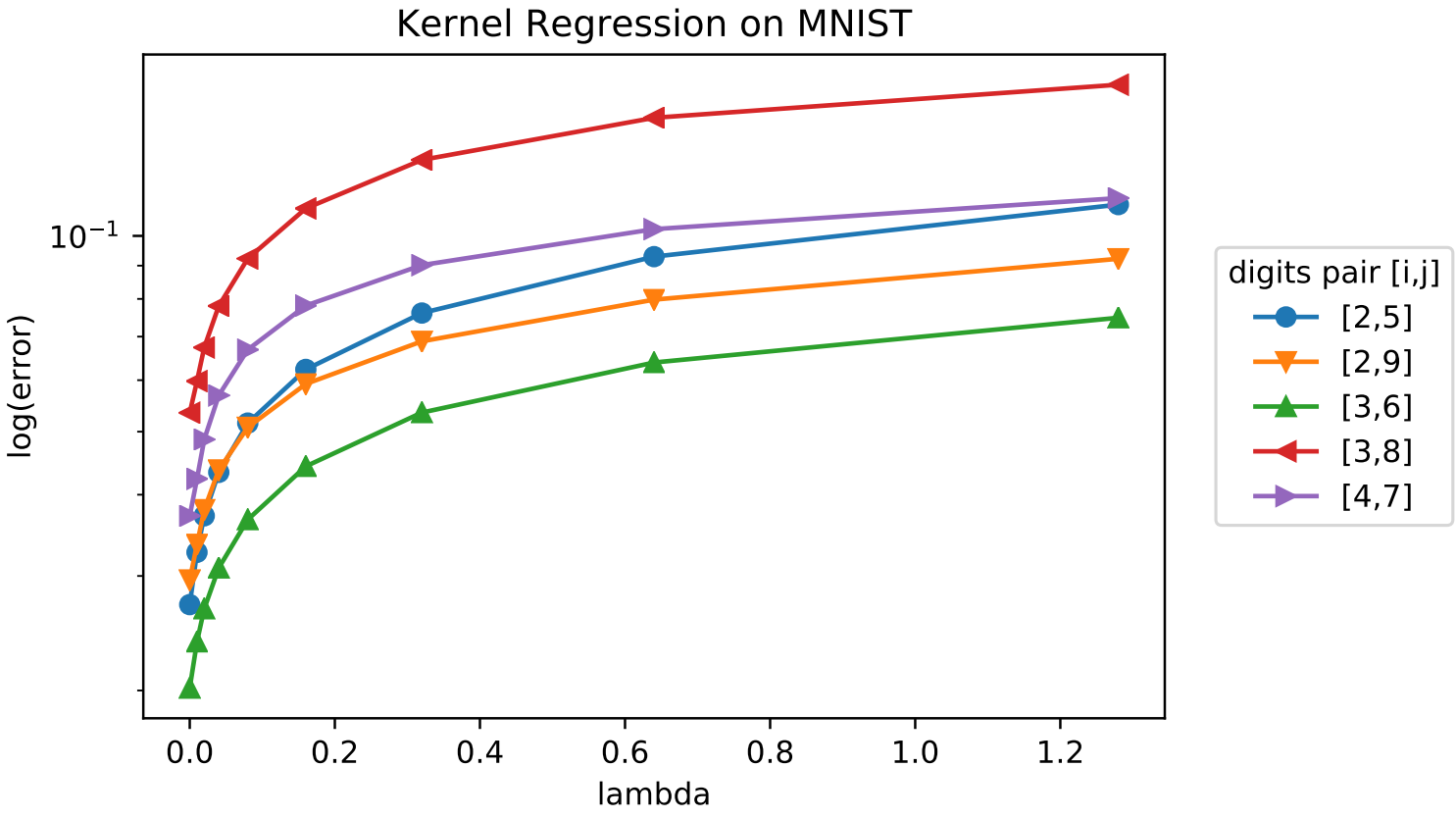


Is explicit regularization λ_{opt} really needed?

Is explicit regularization λ_{opt} really needed?

Is interpolation really bad for statistics and machine learning?

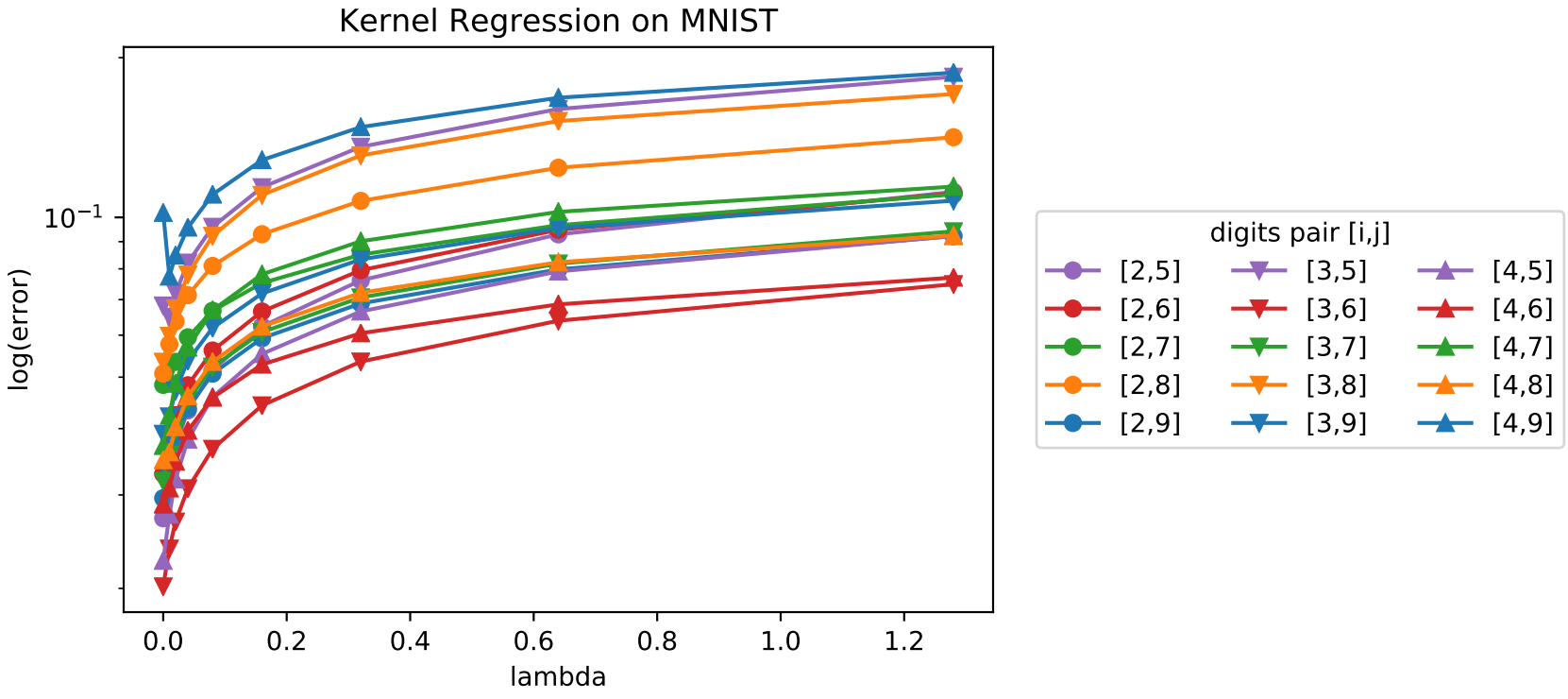
An empirical example



$\lambda = 0$: the interpolated solution, perfect fit on training data.

MNIST data from Yann LeCun

An empirical example



$\lambda = 0$: the interpolated solution, perfect fit on training data.

Isolated phenomenon? No

Understanding deep learning requires rethinking generalization

Chiyuan Zhang*
Massachusetts Institute of Technology
chiyuan@mit.edu

Samy Bengio
Google Brain
bengio@google.com

Moritz Hardt
Google Brain
mrtz@google.com

Benjamin Recht†
University of California, Berkeley
brecht@berkeley.edu

Oriol Vinyals
Google DeepMind
vinyals@google.com

Table 1: The training and test accuracy (in percentage) of various models on the CIFAR10 dataset. Performance with and without data augmentation and weight decay are compared. The results of fitting random labels are also included.

model	# params	random crop	weight decay	train accuracy	test accuracy
Inception	1,649,402	yes	yes	100.0	89.05
		yes	no	100.0	89.31
		no	yes	100.0	86.03
		no	no	100.0	85.75
(fitting random labels)		no	no	100.0	9.78
Inception w/o BatchNorm	1,649,402	no	yes	100.0	83.00
		no	no	100.0	82.00
(fitting random labels)		no	no	100.0	10.12
Alexnet	1,387,786	yes	yes	99.90	81.22
		yes	no	99.82	79.66
		no	yes	100.0	77.36
		no	no	100.0	76.07
(fitting random labels)		no	no	99.82	9.86
MLP 3x512	1,735,178	no	yes	100.0	53.35
		no	no	100.0	52.39
(fitting random labels)		no	no	100.0	10.48
MLP 1x512	1,209,866	no	yes	99.80	50.39
		no	no	100.0	50.51
(fitting random labels)		no	no	99.34	10.61

To understand deep learning we need to understand kernel learning

Mikhail Belkin, Siyuan Ma, Soumik Mandal
Department of Computer Science and Engineering
Ohio State University
{mbelkin, masi}@cse.ohio-state.edu, mandal.32@osu.edu

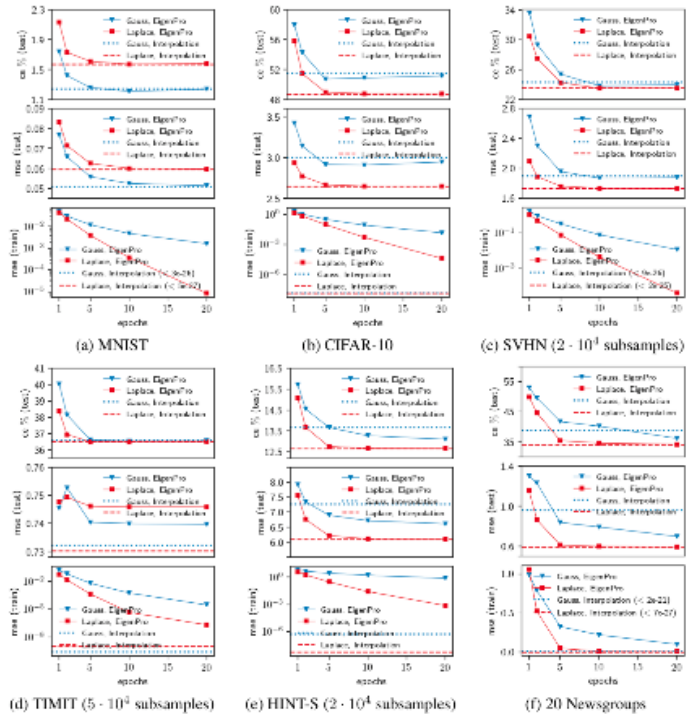


Figure 1: Comparison of approximate classifiers trained by EigenPro-SGD [MB17] and interpolated classifiers obtained from direct method for kernel least squares regression.
| All methods achieve 0.0% classification error on training set. † We use subsampled dataset to reduce the computational complexity and to avoid numerically unstable direct solution.

- Methodology: deep learning, kernel learning, nonparametric regression, AdaBoost, random forests
- Datasets: MNIST, CIFAR-10, others

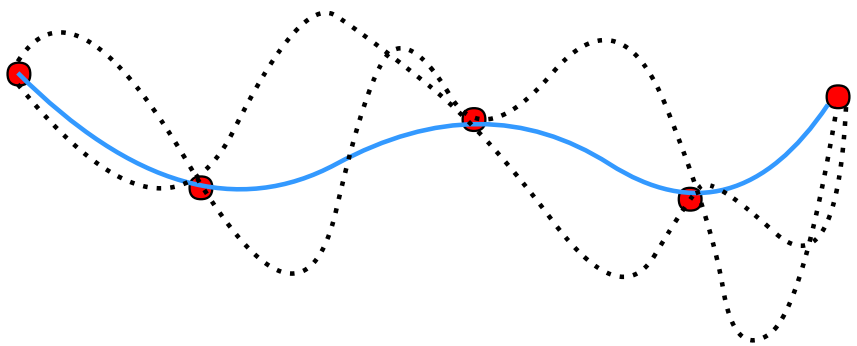
https://playground.tensorflow.org/

PUZZLES

Interpolated solutions performs very well in practice
for many (modern) methodology and datasets!

Q: what is happening? "Overfitting" is not that bad ...

There are many functions that behave exactly the same on training data, but the method prefer certain functions



Our work

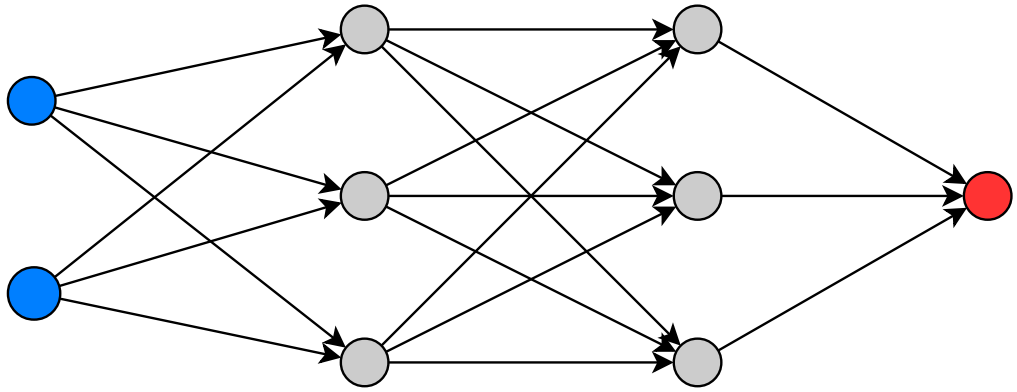
Geometric properties of the data design X, high dimensionality, and **curvature of the kernel** $\Rightarrow$ interpolated solution generalizes.

Liang, T. & Rakhlin, A. (2018). — Just Interpolate: Kernel "Ridgeless" Regression Can Generalize

Feed-Forward Neural Networks

A set of Units:

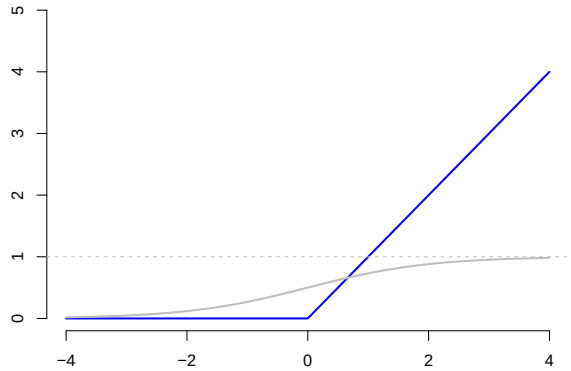
- $d = \dim(\mathbf{X})$ input units
- One output unit, Y
- U hidden units between



Units are arranged into layers:

- According to a directed, acyclic graph
- Number of layers = L = *depth*
- Unit is in layer l if it has a predecessor in $l-1$ and none for any $l' \geq l$
- Unit receives some $\tilde{\mathbf{x}}_l'\mathbf{w}_l + b_l$, returns $\sigma(\tilde{\mathbf{x}}_l'\mathbf{w}_l + b_l)$
- Dimension of $\tilde{\mathbf{x}}_l$ is *width*
- Final layer is $\widehat{f}_{\text{DNN}}(\mathbf{x}) = \tilde{\mathbf{x}}_L(\mathbf{x})'\mathbf{w}_L + b_L$

Non-linear activation function $\sigma(\cdot)$



Deep Neural Networks

Nonstandard Regression Estimator

$$\widehat{f}_{\text{DNN}} := \underset{f_\theta \in \mathcal{F}_{\text{DNN}}}{\arg\min} \sum_{i=1}^{n} \ell\left(f, y_i, \mathbf{x}_i\right), \qquad \text{e.g. } \ell\left(f, y_i, \mathbf{x}_i\right) = \frac{1}{2}\left(y - f(\mathbf{x})\right)^2$$

- What is $\mathcal{F}_{\text{DNN}}$? What is "θ"?
- Intuitively: a series estimator but with data-dependent basis
- We get something like: $\widehat{f}(\mathbf{x}) = \widehat{p}(\mathbf{x})'\widehat{\gamma}$

Computation: (Stochastic) Gradient Descent on θ

- Layer by layer, back-propagation implements chain rule
- θ = all *weights* and *biases* $\{(\mathbf{w}_l', b_l)\}_{l=1}^{L}$; W total parameters

WHAT IS THE PROBLEM?

Celebrated empirical success for deep neural networks. Very little theory on how good is $\widehat{f}(\mathbf{x})$ statistically.

New Avenue in Semiparametric Research for Causal Inference

- Machine learning as a first step
- New results on locally/doubly robust estimation
- Recent literature has focused on lasso and trees/forests

Causal Effects

- Average treatment effects (ATE)
- Profit
- Compare strategies
- Gender discrimination

Our contribution: use **deep neural networks** to **estimate and infer** causal effects with mathematical guarantees.

Farrell, M., **Liang, T.** & Misra, S. (2018). — Deep Neural Networks for Estimation and Inference: Application to Causal Effects and Other Semiparametric Estimands

Empirical Application

Direct Mail Marketing

- Data from a large consumer-goods retailer
  - Direct to customers only
- Treatment is **catalog mailing**
- Outcome is consumer spending
- Questions:
  - Does getting a catalog increase spending (ATE)?
  - Who should be mailed a catalog (profit)?

The Data

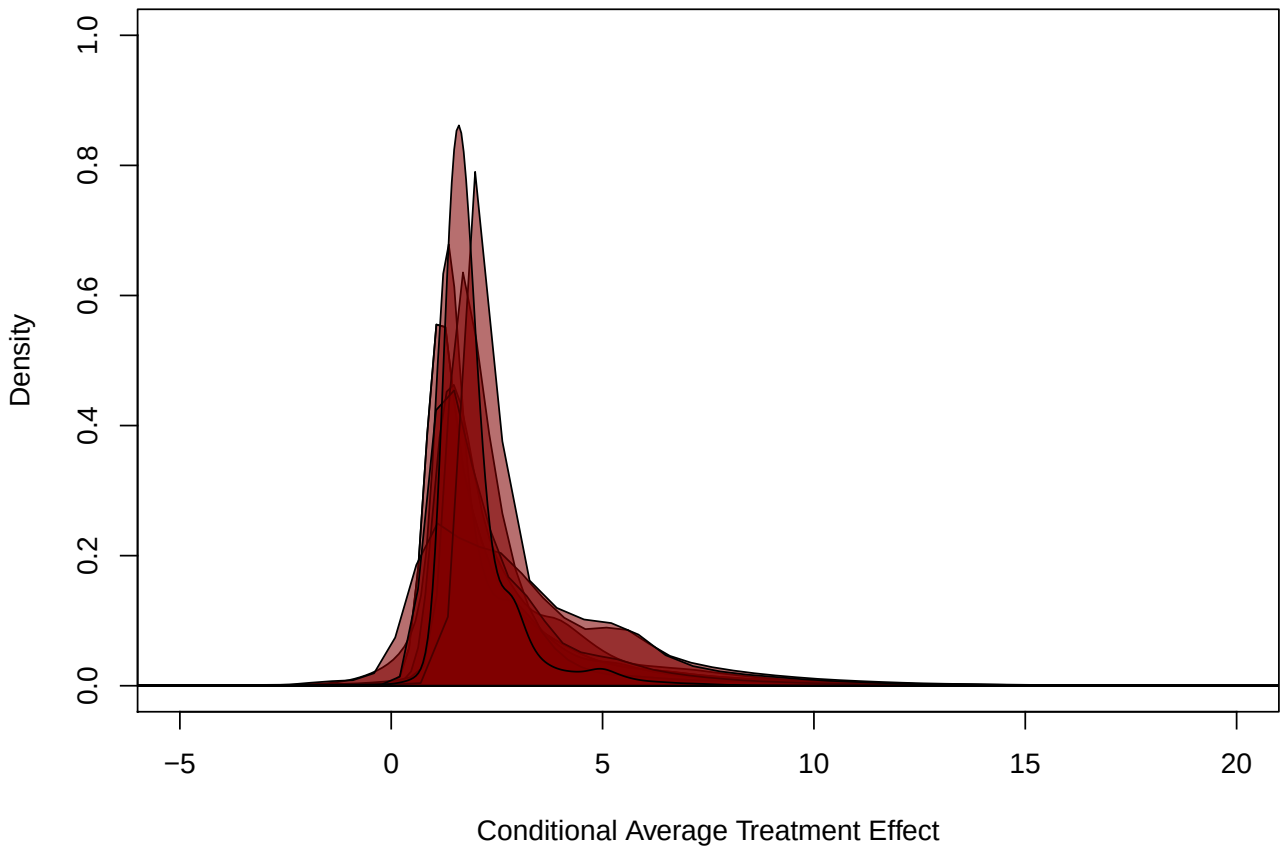
- $n = 292,657$
- Randomized: $\mathbb{P}[T = 1 \mid X] = 2/3$
- ≈ 150 covariates
- Y = sales from all channels

Deep Network Architectures

Learning Rate	Widths $[H_1, H_2, ...]$	Dropout $[H_1, H_2, ...]$	Total Parameters	Validation Loss	Training Loss
0.0003	[60]	[0.5]	8702	1405.62	1748.91
0.0003	[100]	[0.5]	14502	1406.48	1751.87
0.0001	[30, 20]	[0.5, 0]	4952	1408.22	1751.20
0.0009	[30, 10]	[0.3, 0.1]	4622	1408.56	1751.62
0.0003	[30, 30]	[0, 0]	5282	1403.57	1738.59
0.0003	[30, 30]	[0.5, 0]	5282	1408.57	1755.28
0.0003	[100, 30, 20]	[0.5, 0.5, 0]	17992	1408.62	1751.52
0.00005	[80, 30, 20]	[0.5, 0.5, 0]	14532	1413.70	1756.93

One Goodness of Fit Measure

- Getting a catalog shouldn't make you buy less
- Plot $\hat{\tau}(\mathbf{x}_i)$



Semiparametric Results

- Lines up closely with difference in means (RCT)
- Study ATE and profits from different mailing strategies
  - Loyalty: $s(\mathbf{x}_i) = 1$ if purchased in prior calendar year

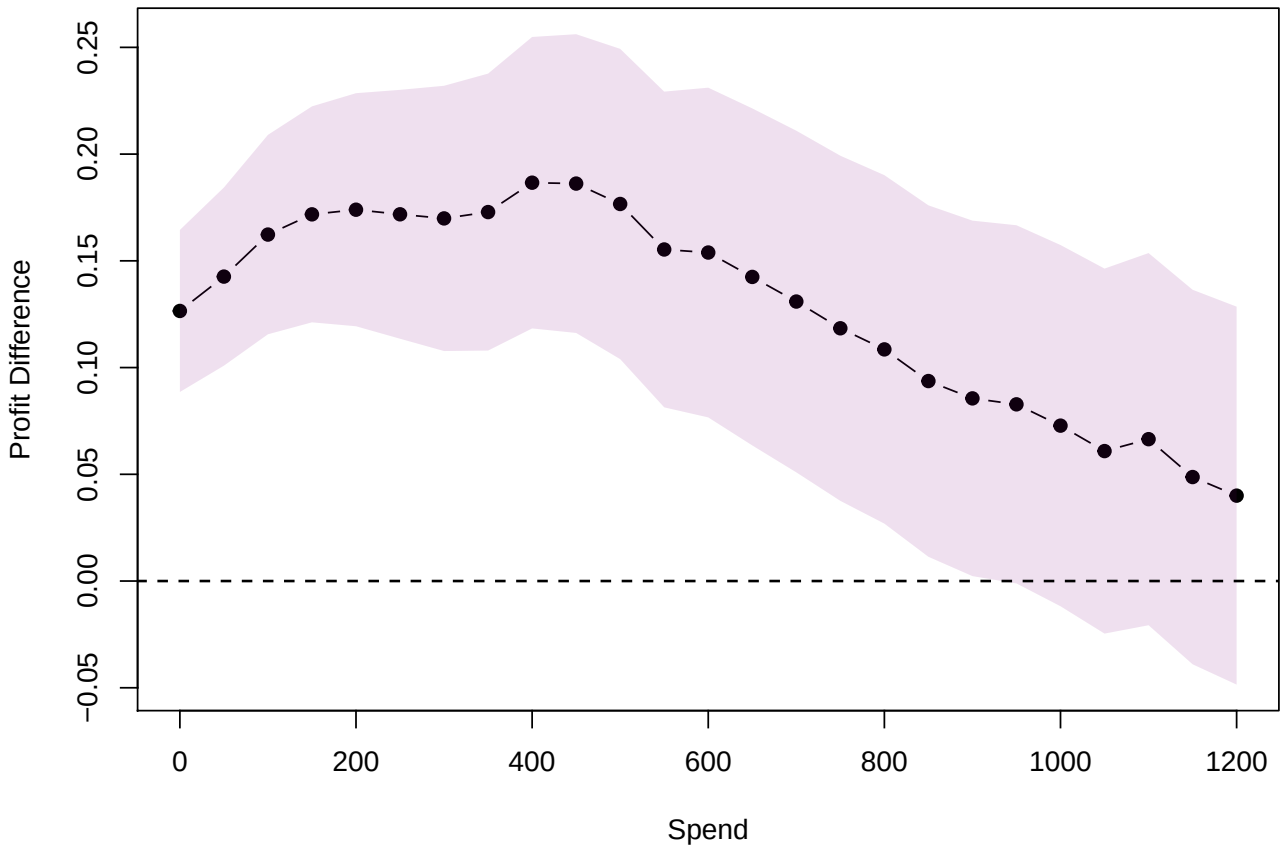
From Network 3:

Estimand:	$\hat{\pi}(s)$	**95% CI**
ATE	2.547	[2.223 , 2.872]
π (never treat)	2.027	[1.934 , 2.120]
π (always treat)	2.224	[2.152 , 2.296]
π (loyalty policy)	2.358	[2.283 , 2.434]

Targeting Strategy

Should we target bigger spenders?

- Study $\{\pi(\text{spend} > \bar{y}) - \pi(\text{always treat})\}$
- *Pointwise* 95% confidence band
- Profits increase in $\bar{y}$ until roughly \$500, then too few are targeted



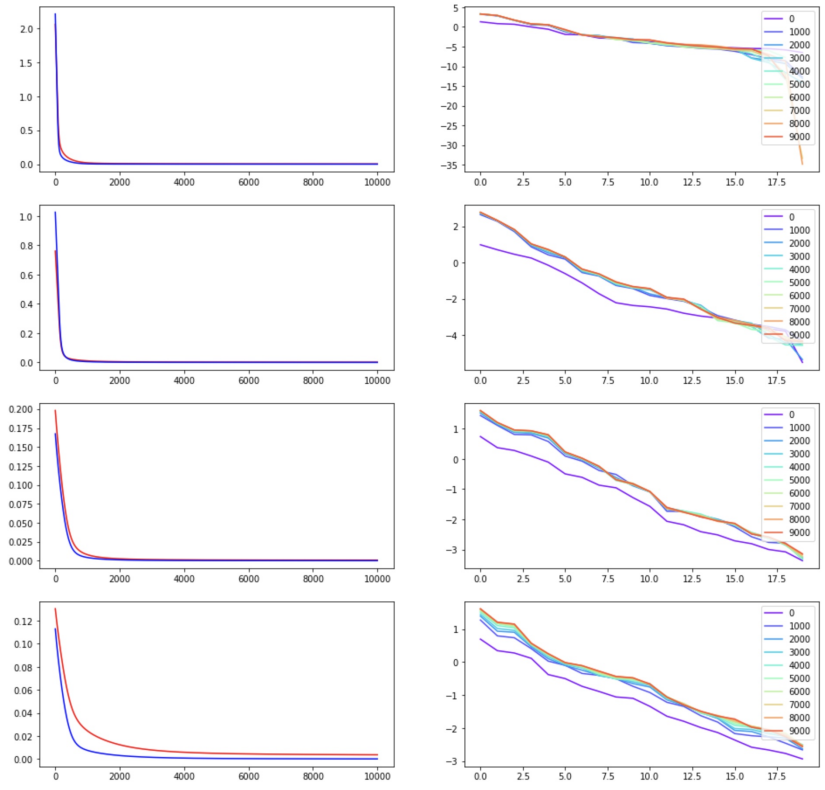
Advantages of Neural Networks?

Neural networks: data-dependent basis, an adaptive representation learned from data

Classic nonparametric statistics or statistical learning: fixed basis from functional analysis, not adaptive to data

Heuristic justification, but what does it really mean mathematically?

Our recent work tries to explain mathematically the advantage of data-dependent adaptive representation!



Dou, X. & **Liang, T.** (2018+). — Training Neural Networks as Learning Data-dependent Kernels: Interpolation, Approximation and Representation Benefits, *working paper*

My Current Interest

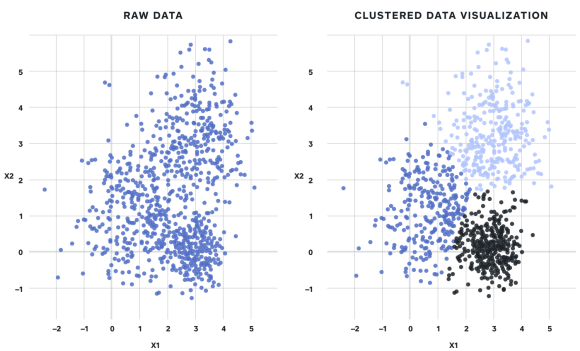
Study complex models such as deep neural networks and kernel machines, enrich the understanding both statistically and computationally.

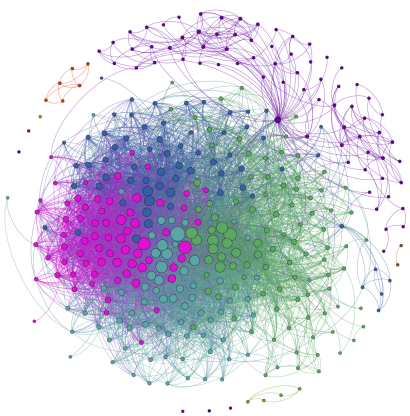
Summarizing Data Clouds (in high dim)

$$X_1, X_2, \ldots X_n \in \mathbb{R}^d, \quad d \text{ thousands of dimensions}$$

Clustering

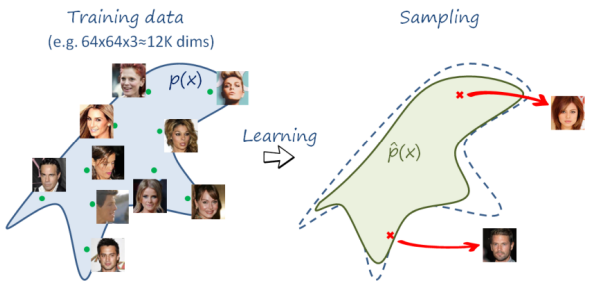
community detection, data visualization, topic modeling ...

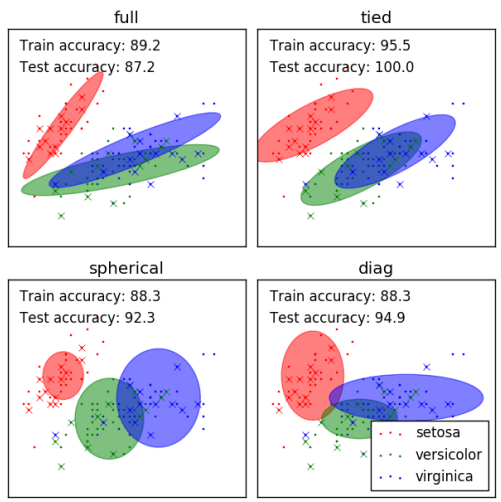




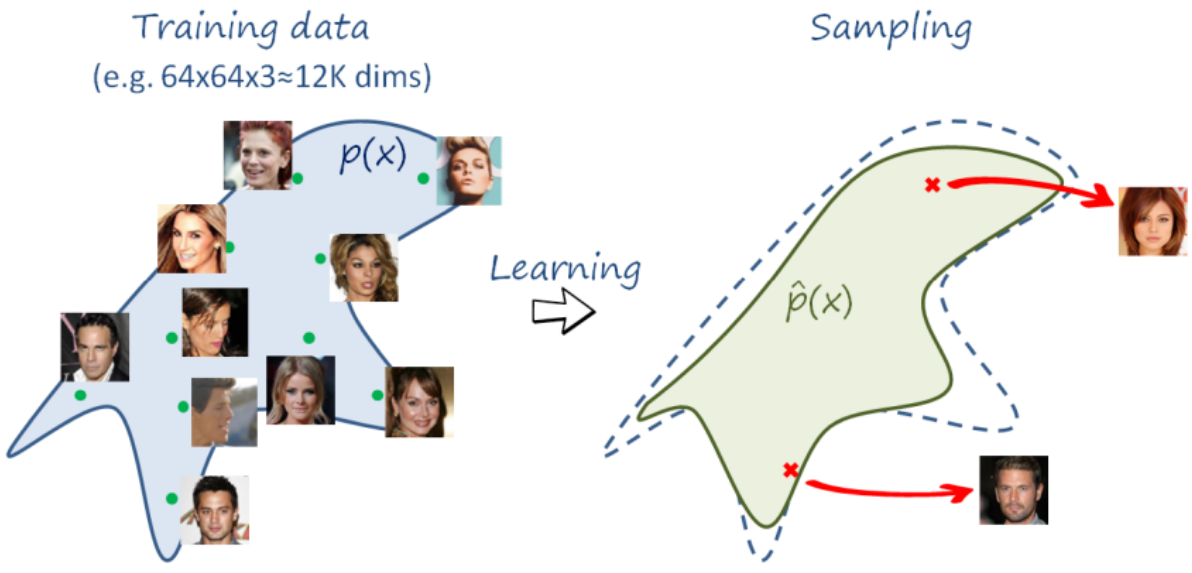
Learn distributions

generative adversarial networks (GANs), density estimation, probabilistic modeling ...

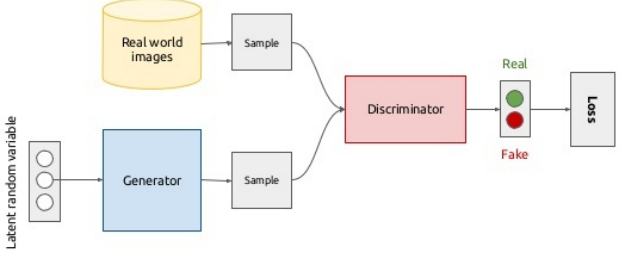


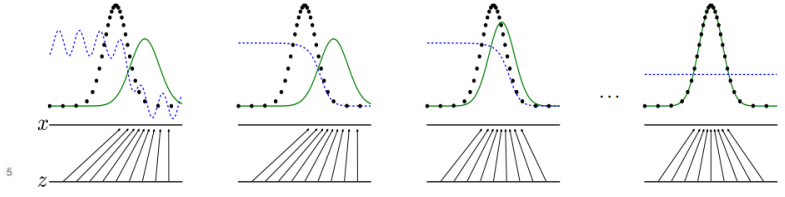


GANs

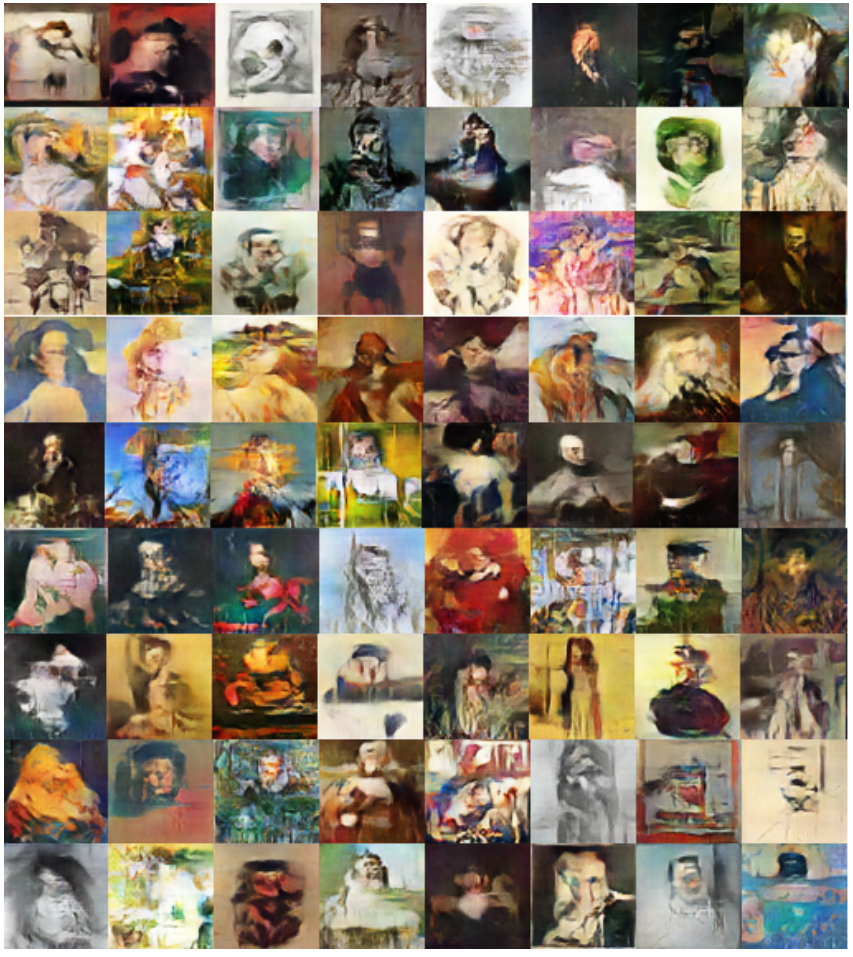


Generative adversarial networks (conceptual)





https://poloclub.github.io/ganlab/



You get some **dark** stuff when there is **no THEORY**!

Our work

Generator g_θ, Discriminator f_ω

$$U(\theta, \omega) = \mathbf{E}_{X \sim \mathcal{P}_{\text{real}}} h_1(f_\omega(X)) - \mathbf{E}_{Z \sim \mathcal{P}_{\text{input}}} h_2(f_\omega(g_\theta(Z)))$$

$$\min_\theta \max_\omega U(\theta, \omega)$$

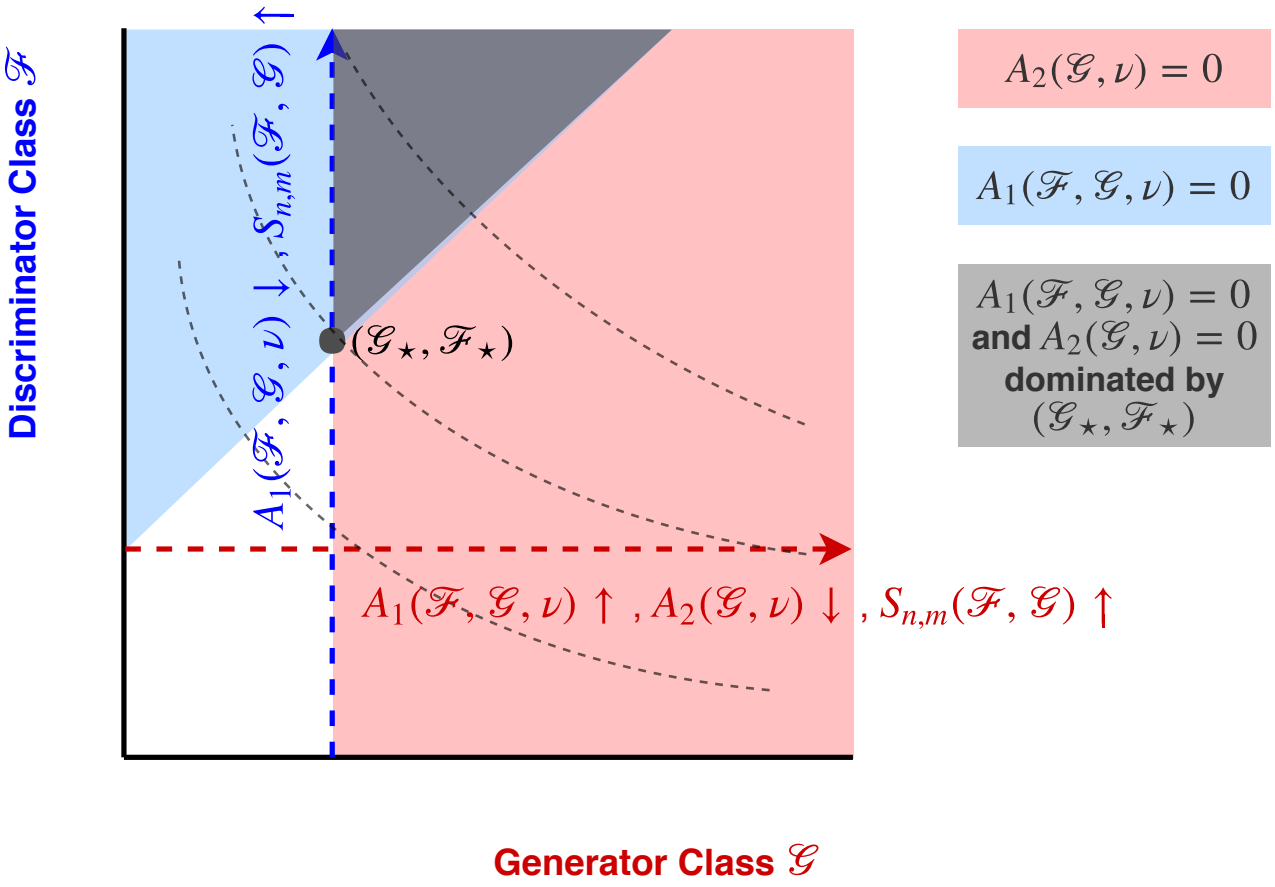
Statistically: understand statistical theory for GANs, why it can be better than classic density learning in statistics.

Computationally: what training dynamics can stablize GANs training, how things converge locally.

Liang, T. (2018). — On How Well Generative Adversarial Networks Learn Densities: Nonparametric and Parametric Results

Liang, T. & Stokes, J. (2018). — Interaction Matters: A Note on Non-asymptotic Local Convergence of Generative Adversarial Networks

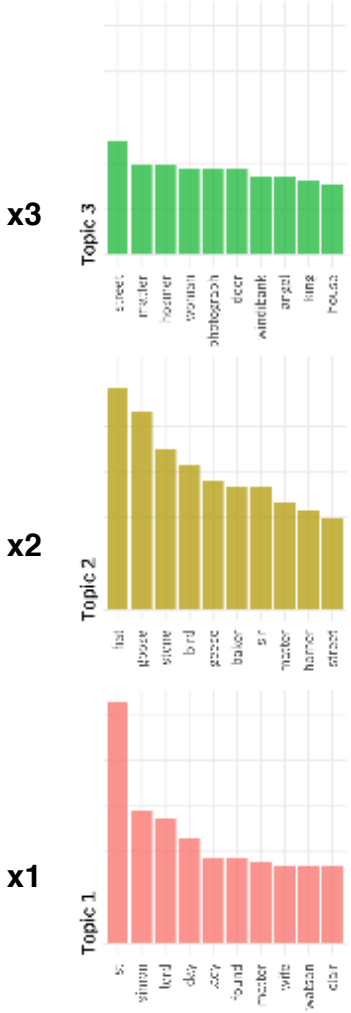
Complexity/Regularization Again



Textual Factors

Unstructured Data

Decompose



Structured Features

Textual Factors

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

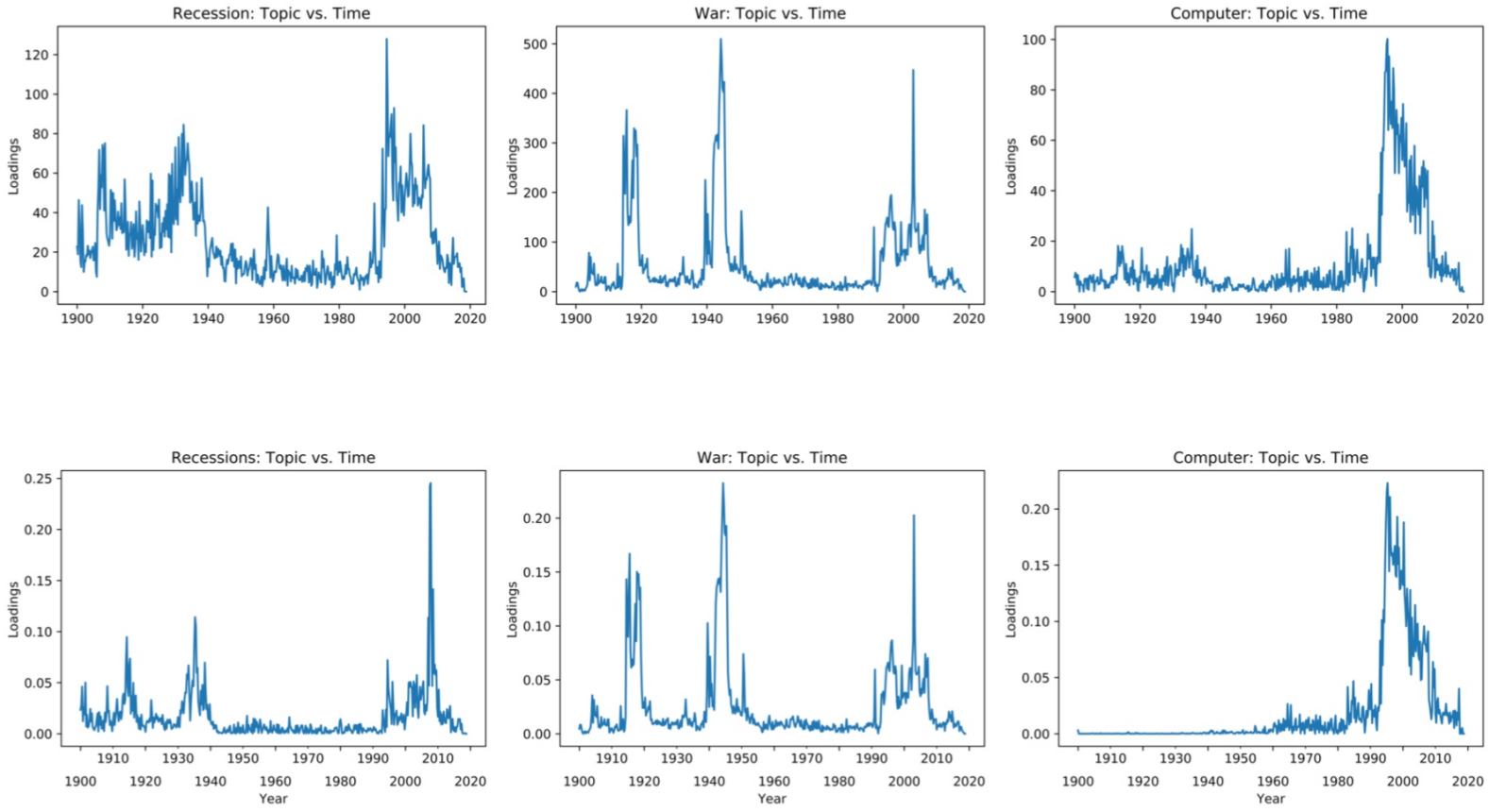
The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

We want to learn the topic/basis/factors automatically from the data.

We develop a general **algorithmic** framework for analyzing **large-scale text-based data**, which captures complex linguistic structures while ensuring computational scalability and economic interpretability.

Cong, W. L., **Liang, T.**, Zhang, X. (2017+). — Textual Factors: A Scalable, Interpretable, and Data-driven Approach to Analyzing Unstructured Information, *working paper*

Topic loadings over time



Cluster	Support
Topic ID: 62, Prob: 0.20071%	**washington**, tax, business, york, labor, letter, bulletin, wire, report, old, many, big, **president**, like, long, economic, **price**s, time, ago, federal, outlook, city, get, high, sales, white, house, back, people, even, state, just, home, world, much, **american**, man, next, **government**, job, million, still, work, companies, workers, economy, men, three, little
Topic ID: 1272, Prob: 0.17438%	stock, dividend, steel, business, **american**, oil, common, market, york, **earnings**, months, outlook, **cent**s, made, record, way, chicago, share, company, united, net, time, **president**, rate, **price**s, increase, railroad, states, june, **price**, **general**, review, shares, declared, july, report, cotton, preferred, sales, **washington**, present, large, month, regular, production, exchange, pacific, cars, quarterly, september
Topic ID: 1828: Prob: 0.11747%	steel, states, business, united, outlook, review, railroad, stock, way, market, york, country, time, **president**, great, made, **american**, **price**s, copper, increase, **earnings**, corporation, public, **government**, per, national, **general**, since, **washington**, cotton, crop, bank, report, months, state, much, commission, present, **cent**, railroads, rate, conditions, **price**, large, street, ago, letter, pacific, trade, three

Table 3: Sample plain-vanilla LDA clusters.

Cluster	Support
Topic ID: 62, Prob: 0.20071%	**washington**, tax, business, york, labor, letter, bulletin, wire, report, old, many, big, **president**, like, long, economic, **price**s, time, ago, federal, outlook, city, get, high, sales, white, house, back, people, even, state, just, home, world, much, **american**, man, next, **government**, job, million, still, work, companies, workers, economy, men, three, little
Topic ID: 1272, Prob: 0.17438%	stock, dividend, steel, business, **american**, oil, common, market, york, **earnings**, months, outlook, **cent**s, made, record, way, chicago, share, company, united, net, time, **president**, rate, **price**s, increase, railroad, states, june, **price**, **general**, review, shares, declared, july, report, cotton, preferred, sales, **washington**, present, large, month, regular, production, exchange, pacific, cars, quarterly, september
Topic ID: 1828: Prob: 0.11747%	steel, states, business, united, outlook, review, railroad, stock, way, market, york, country, time, **president**, great, made, **american**, **price**s, copper, increase, **earnings**, corporation, public, **government**, per, national, **general**, since, **washington**, cotton, crop, bank, report, months, state, much, commission, present, **cent**, railroads, rate, conditions, **price**, large, street, ago, letter, pacific, trade, three

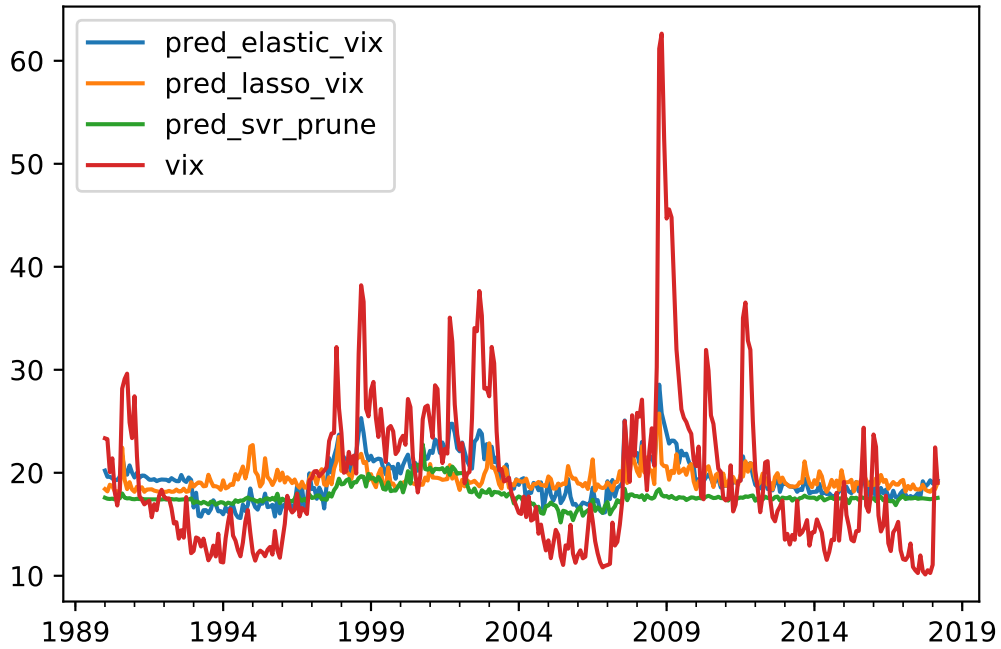
Table 3: Sample plain-vanilla LDA clusters.

Cluster	Support
Tax	quotas, visa, harvestable, import, preferential, abolished, tariffs, quota, sanction, compulsory, tariff, compulsorily, stipulating, fisheries, cess, exports, pricing, export, telcos, exporters, import, liberalization, preferential, excise, tariffs, tax, tariff, importers, deregulation, antidumping, subsidy
Oil	refiners, refiner, refineries, refinery, petrochemical, feedstock, refiners, pipelines, smelters, crudes, oil, bpd, gasoline, refiner, petrochemicals, petroleum, refining, ethanol, refineries, tankers, refinery, coker, petrochemical, ethylene, feedstock, crude
Unemployment	stimulus, foreclosures, recession, claimants, workweek, unemployed, housing, unemployment, jobless, economy, workers
Volatility	correction, uptrend, readjustment, reversal, retest, revision, divergence, retrenchment, steepening, selloff, rebalancing, bearish, pullbacks, corrective, correcting, reversion, stabilization, selldown, snapback, reassessment, volatility, pullback, bull, corrections, bottoming, downtrend
Exports	consignments, foodstuffs, exports, tins, cargo, goods, warehouses, equipments, importers, exporting, containers, tonnages, exporters, import, imports, perishable, cartons, cargoes, export, adulterated, tankers, pallets, wholesalers, demurrage, customs, transporters, consignment, consignee, exported
Investment	development, capitalization, differentiation, invest, macro, optionality, strategic, capex, macroeconomic, countercyclical, investments, investing, outperformance, diversification, equity, arbitrage, diversify, cyclicality, underperformance, diversifying, expansion, diversified, geographies, reinvest, specialization, profitability, deleveraging, consolidation, renewables, volatility, investment, liquidity, growth, maximization, sector, cyclical, synergy, reinvesting, investors, reinvestment
Stimulus	appropriation, moneys, underfunded, money, reauthorization, subsidies, budget, fundings, budgeted, allocations, budgets, budgetary, stimulus, funded, appropriations, funds, grant, nonfederal, appropriated, earmarked, infrastructure, reauthorized, assistance, unfunded, funding, financing, grants, monies, support, underfunding
Disasters	disturbances, occurance, instances, recur, disasters, incidences, occur, occurence, occurrences, causes, occurred, occurrence, phenomenon, earthquakes, anomaly, outbreaks, accidents, incidents, emergencies, observations, tragedies, ultramafic, catastrophes, polymetallic, anomalous, occurence, outbreaks, disturbances, incidences, oc-

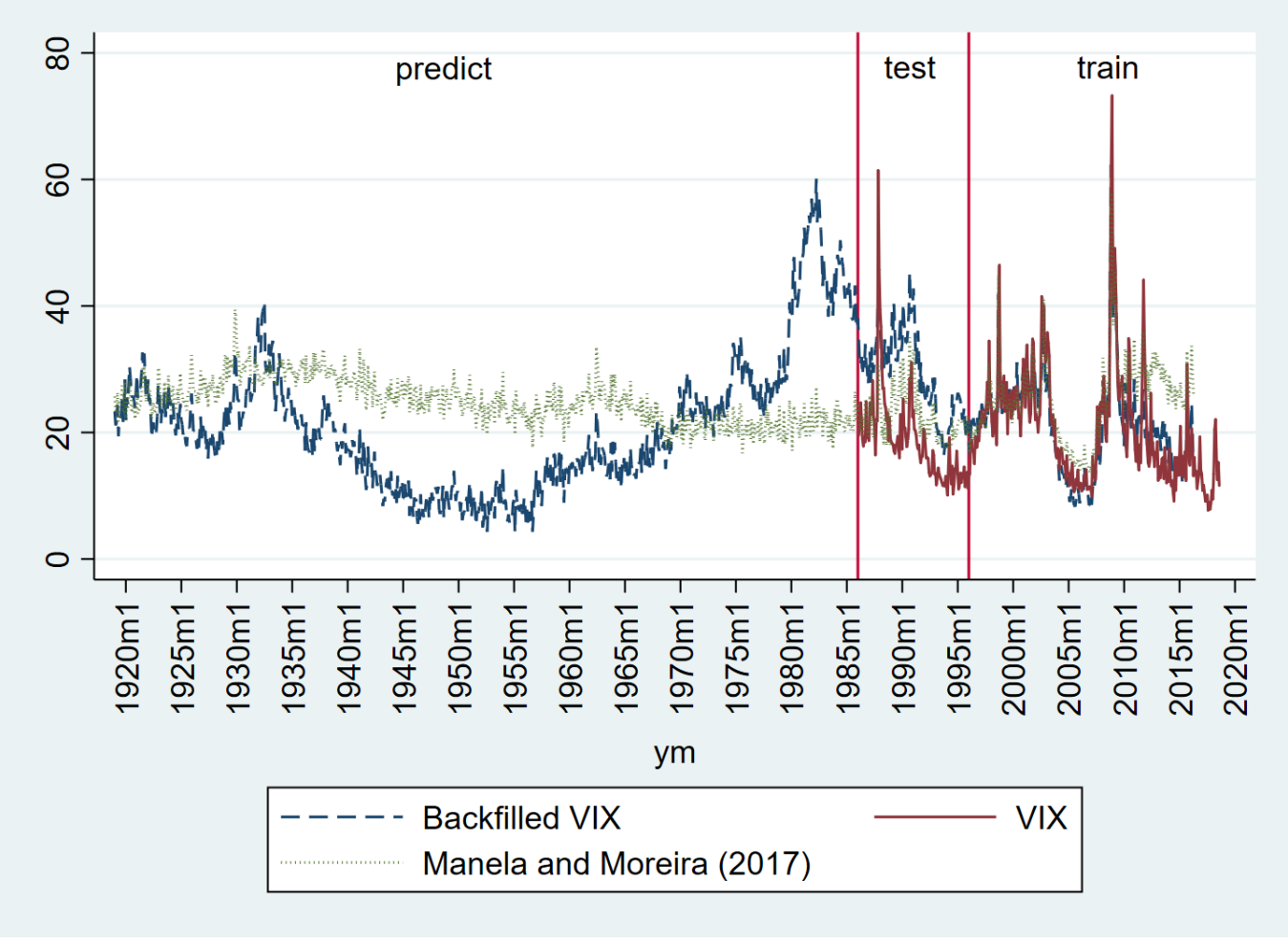
Empirical application: VIX

Applied LASSO to 2000 topics/factors, cross-validation (and model selection AIC/BIC), only **one topic** is selected.

['adversities' 'adversity' 'hardships' 'challenges' 'calamities' 'obstacles' 'difficulties' 'crises' 'disappointments' 'disasters' 'catastrophes' 'disadvantages' 'rigors' 'upheavals' 'negativity' 'turmoil' 'catastrophe' 'infirmities' 'prejudices' 'crisis' 'handicaps' 'anxieties' 'thalassemia' 'incompatibilities' 'drawbacks' 'globalization' 'disease']



Backfilling VIX



My Current Interest

Study complex regression models (such as deep neural networks and kernel machines), enrich the understanding both statistically and computationally

Study how to represent unstructured data (knowledge/structure discovery), enrich the understanding both statistically and computationally

with Economic and Business applications

Thank you all!