Statistics Faculty Lightning Talks

August 30, 2019

Statistical and Computational Methods for Studying Genome Regulation

Keleș Group Research
keles@stat.wisc.edu

August 30, 2019

The Big Picture

Discovering genome parts and **understand**ing how they work by **data science**

The New York Times

Years of Education Influenced by Genetic Makeup, Enormous Study Finds

This Genetic Mutation Makes People Feel Full — All the Time

No 'Gay Gene', but Study Finds Genetic Links to Sexual Behavior

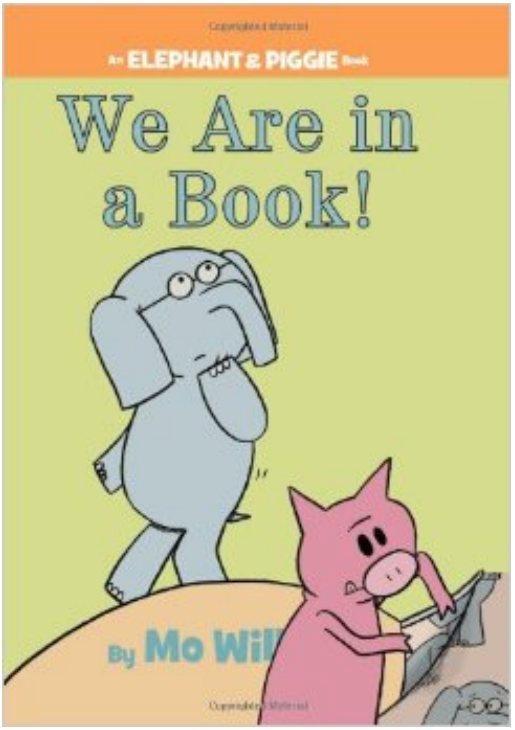
THE ‘GENO-ECONOMISTS’ SAY DNA CAN PREDICT OUR CHANCES OF SUCCESS

Genome

ACGTAAGTCCC… a sequence of 3 billion letters.

Genome

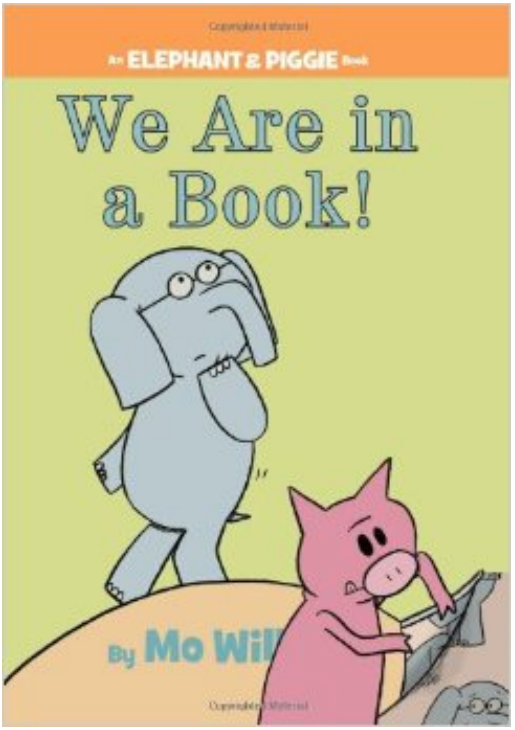
ACGTAAGTCCC… a sequence of 3 billion letters.



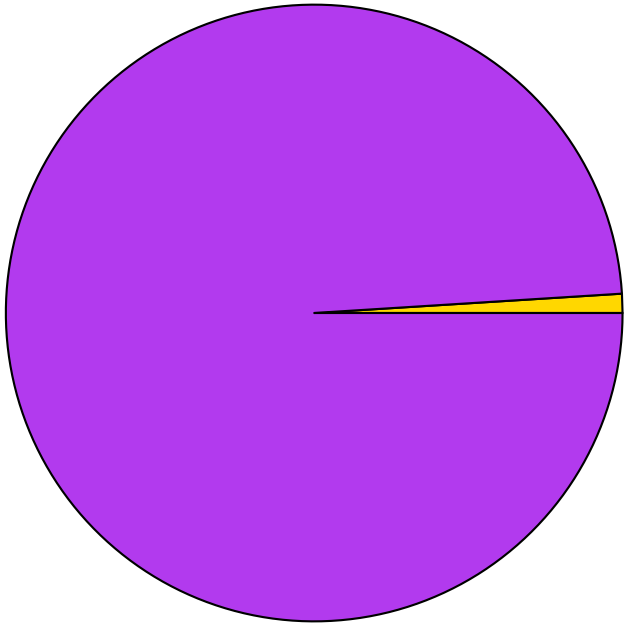
That is 10 million Piggy and Gerald books!!!

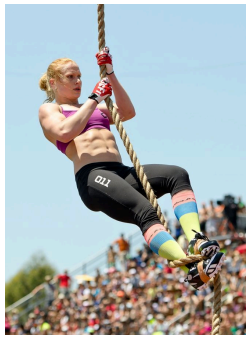
Genome

ACGTAAGTCCC… a sequence of 3 billion letters.



That is 10 million Piggy and Gerald books!!!



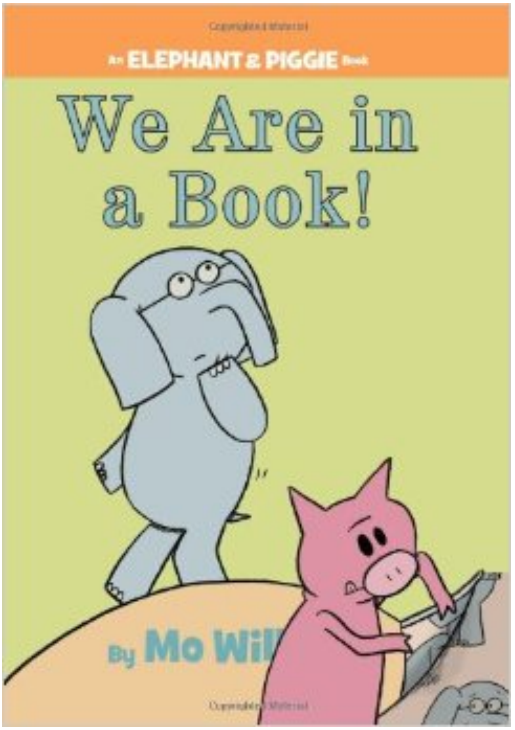


Coding for genes: < 2%

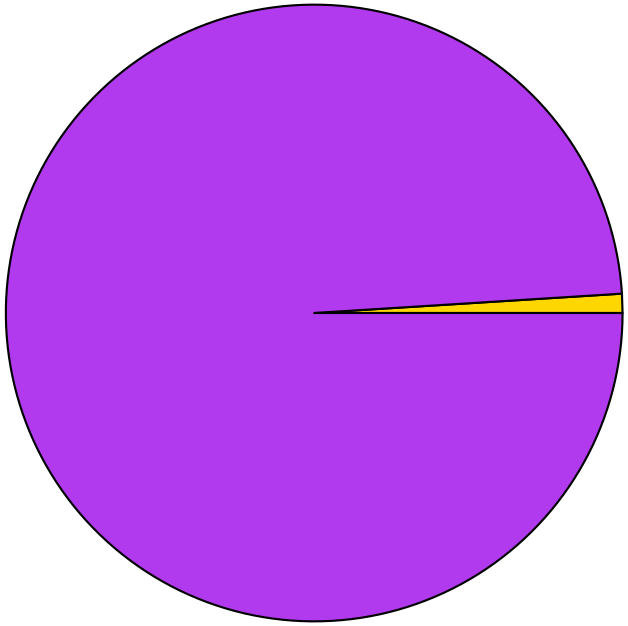
# of genes: 20,000

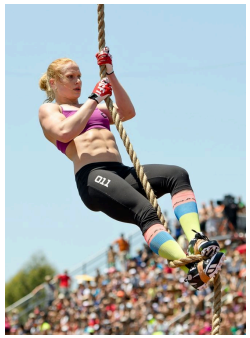
Genome

ACGTAAGTCCC… a sequence of 3 billion letters.



That is 10 million Piggy and Gerald books!!!



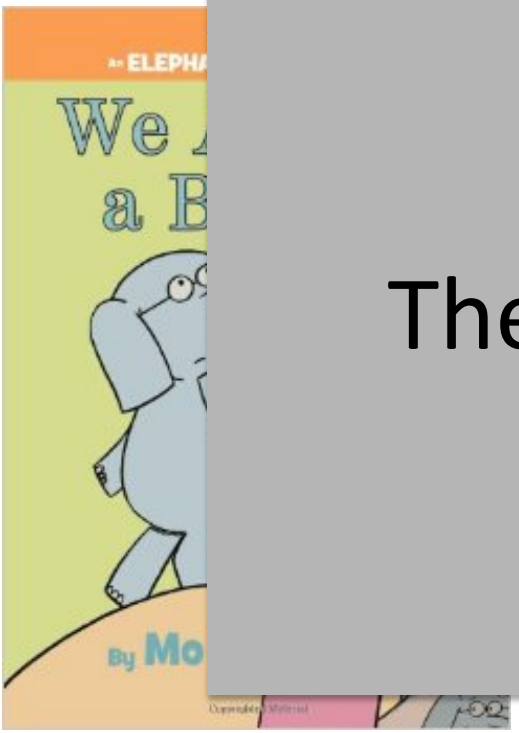


Coding for genes: < 2%

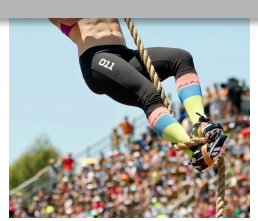
# of genes: 28,000 20,000

Genome

ACGTAAGTCCC… a sequence of 3 billion letters.



The 98% regulates the 2%.



genes: < 2%

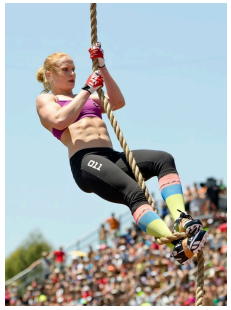
# of genes: 28,000 20,000

Where does “important” genomic variation in humans reside? Within genes?

Super stable

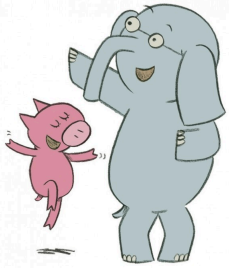
Can fall at any second

Where does “important” genomic variation in humans reside?

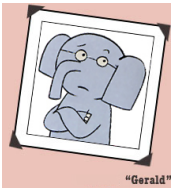


Variants

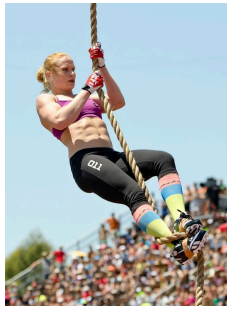
Association variants in genes:
< 5%



Association variants in non-genic regions:
≥ 95%

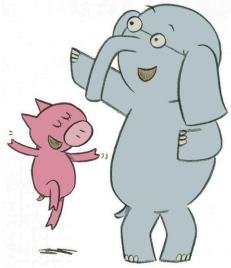


Where does “important” genomic variation in humans reside?

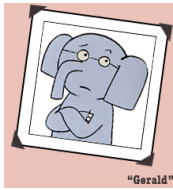


Variants

Association variants in genes:
< 5%



Association variants in non-genic regions:
≥ 95%

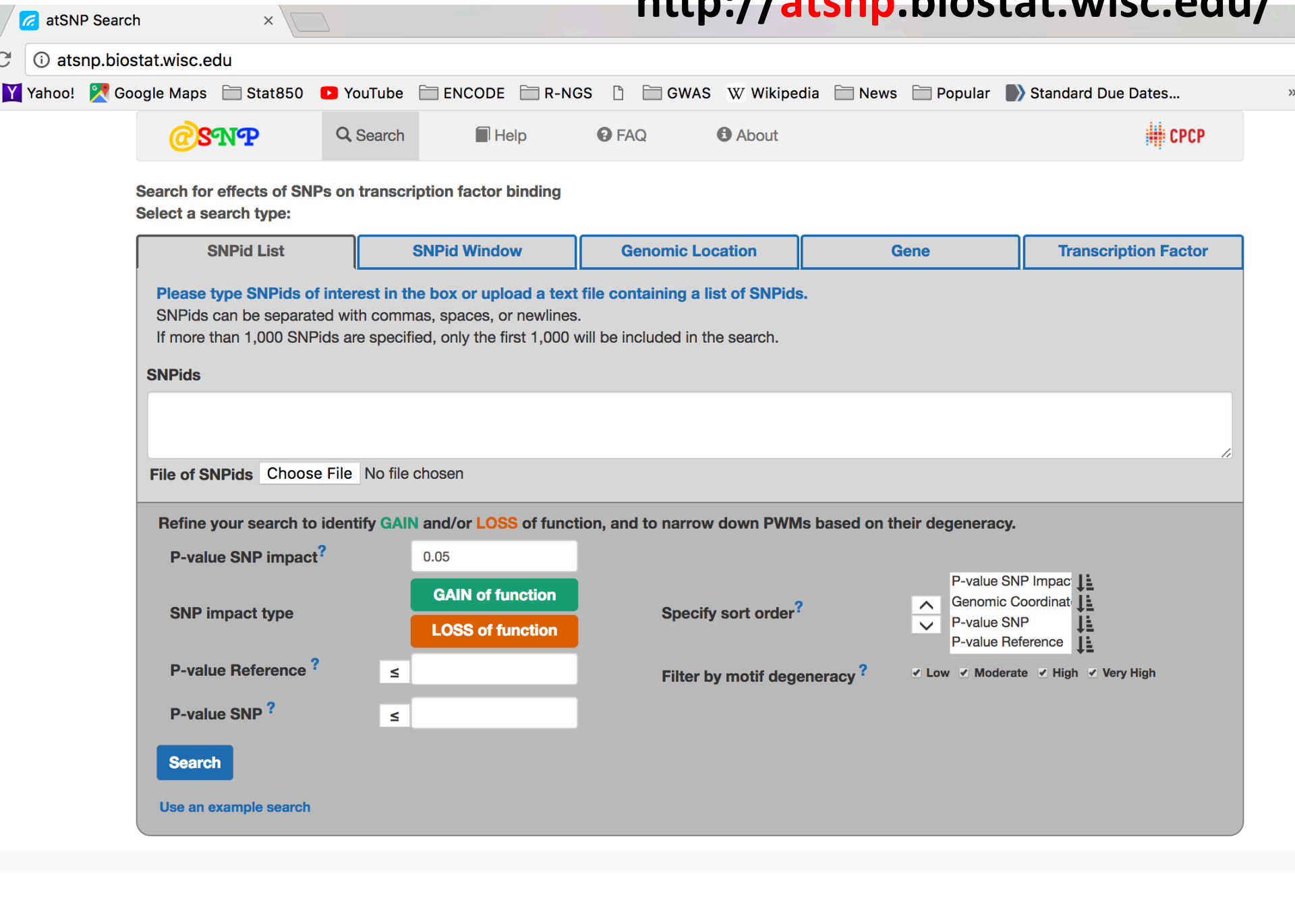


What do the variants in noncoding regions do?

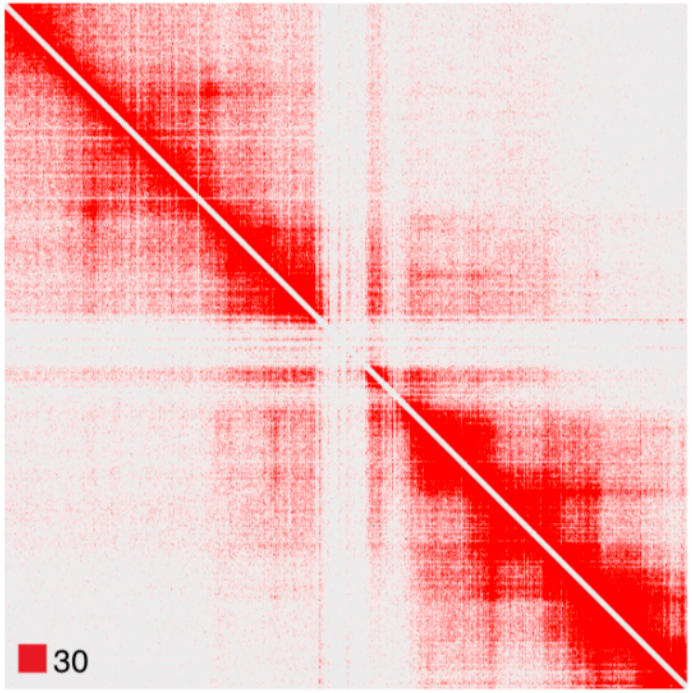
	With annotation	Without annotation
Model	$Y = \mathbf{G}\beta + \epsilon, \quad \beta = \mathbf{A}\gamma + \eta$	$Y = \mathbf{G}\beta + \epsilon,$
Objective Function	$\frac{1}{2n}\|\|Y - \mathbf{G}\beta\|\|_2^2 + \lambda_1\|\|\eta\|\|_1 + \lambda_2\|\|\gamma\|\|_1$ s.t. $\beta = \mathbf{A}\gamma + \eta$	$\frac{1}{2n}\|\|Y - \mathbf{G}\beta\|\|_2^2 + \lambda\|\|\beta\|\|_1$
Estimation Error $\|\|\beta - \hat{\beta}\|\|_2^2$	$O_p\left(\frac{(s_\gamma + s_\eta)\sigma^2 \log p}{n}\right)$	$O_p\left(\frac{s_\beta \sigma^2 \log p}{n}\right)$

Key assumption: Existence of SNPs sharing similar annotations and with similar effects: $s_\beta >> s_\gamma + s_\eta$

http://atsnp.biostat.wisc.edu/



Signals from certain regions are under-represented in the genome

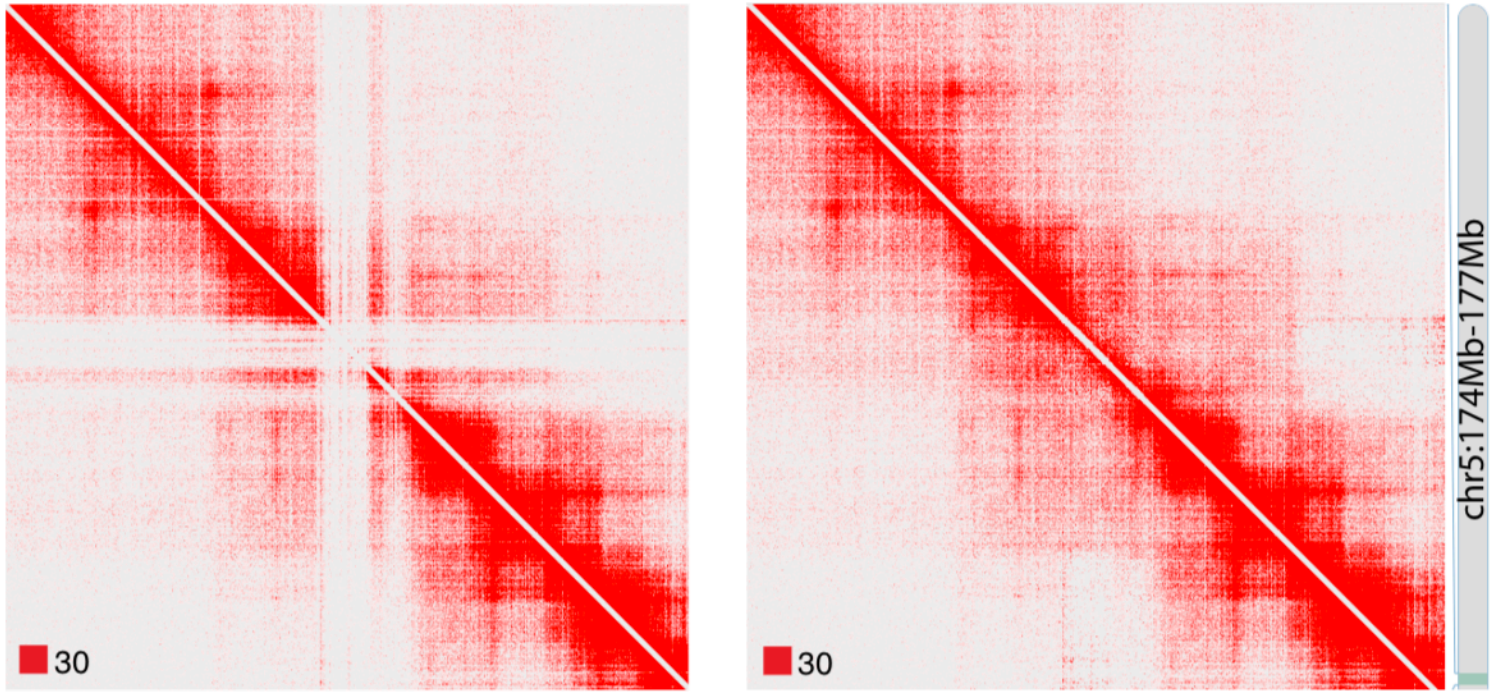


Observed data from standard processing

Biotechnology allow generation of large scale data at the genomic level for discovering and understanding genome parts.

Signals from certain regions are under-represented in the genome

After mHi-C magic with probabilistic inference



Observed data from standard processing

$$P(Z_{i,(j,k)} = 1 \mid Y_{i,(j',k')}, \forall j', k')$$

eLIFE elifesciences.org TOOLS AND RESOURCES

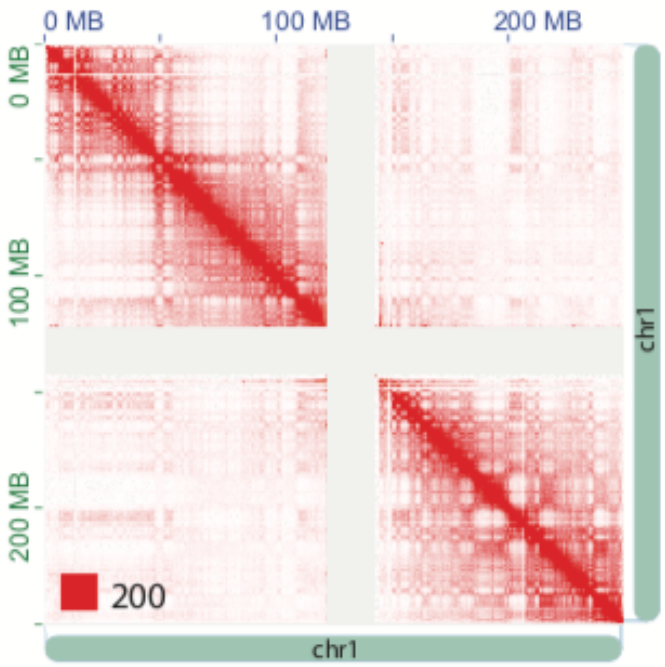
Generative modeling of multi-mapping reads with mHi-C advances analysis of Hi-C studies

Ye Zheng[1], Ferhat Ay[2,3], Sunduz Keles[1,4]*

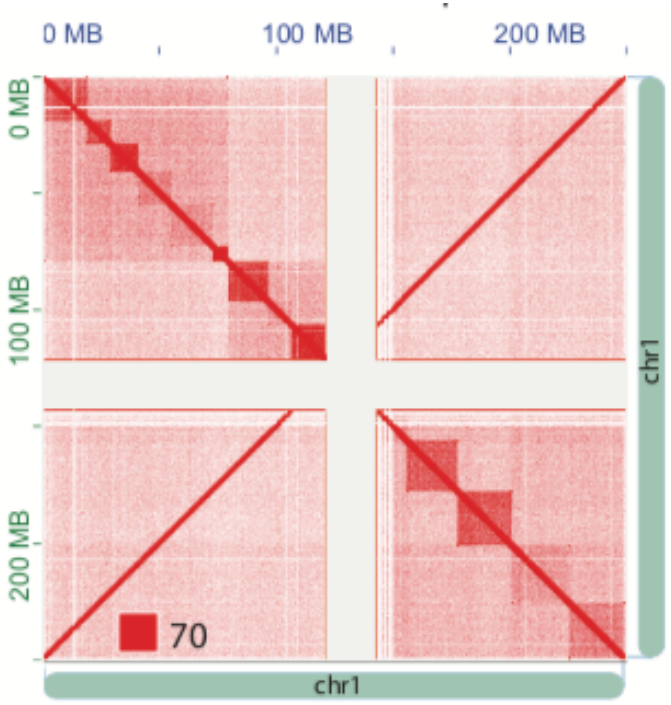
[1]Department of Statistics, University of Wisconsin-Madison, Madison, United States; [2]La Jolla Institute for Allergy and Immunology, La Jolla, United States; [3]School of Medicine, University of California, San Diego, La Jolla, United States; [4]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, United States

Realistic simulations are *key weapons* of statistical genomics

Observed biological data

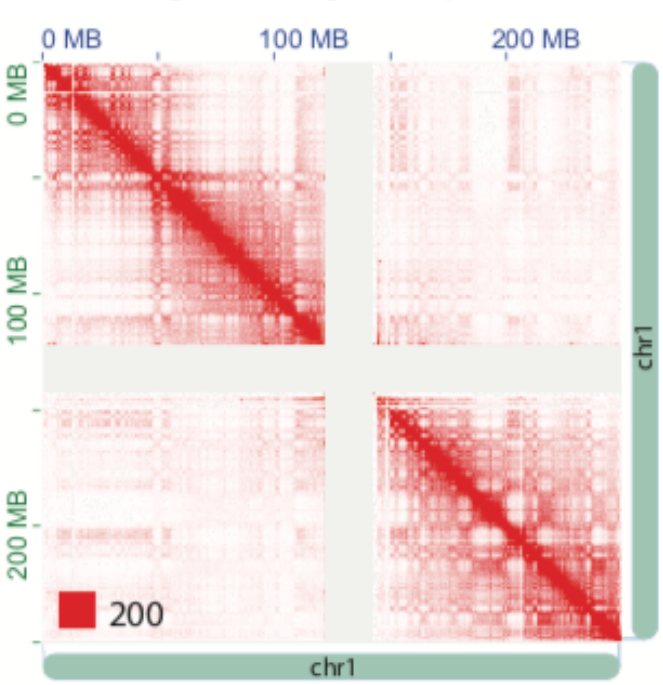


Simulated data from existing simulator

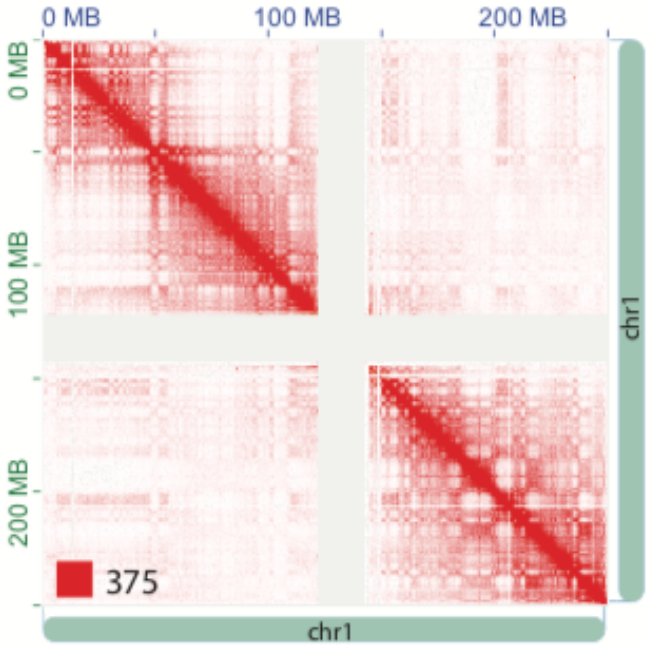


FreeHi-C: A non-parametric Hi-C simulator

Observed biological data



FreeHi-C simulated data



FreeHi-C: high fidelity Hi-C data simulation for benchmarking and data augmentation

Ye Zheng[1] & **Sündüz Keleş**[1,2,*]

[1]Department of Statistics, University of Wisconsin - Madison, Madison, WI 53706, USA

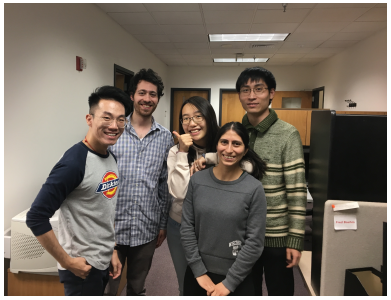
Nature Methods, 2019, In press.

Email me (keles@stat) if you are interested in learning more.

We are hiring RAs!









Courtesy of Ye Zheng

Cécile Ané

Statistics and Botany

statistical **methods** to answer questions in evolutionary biology

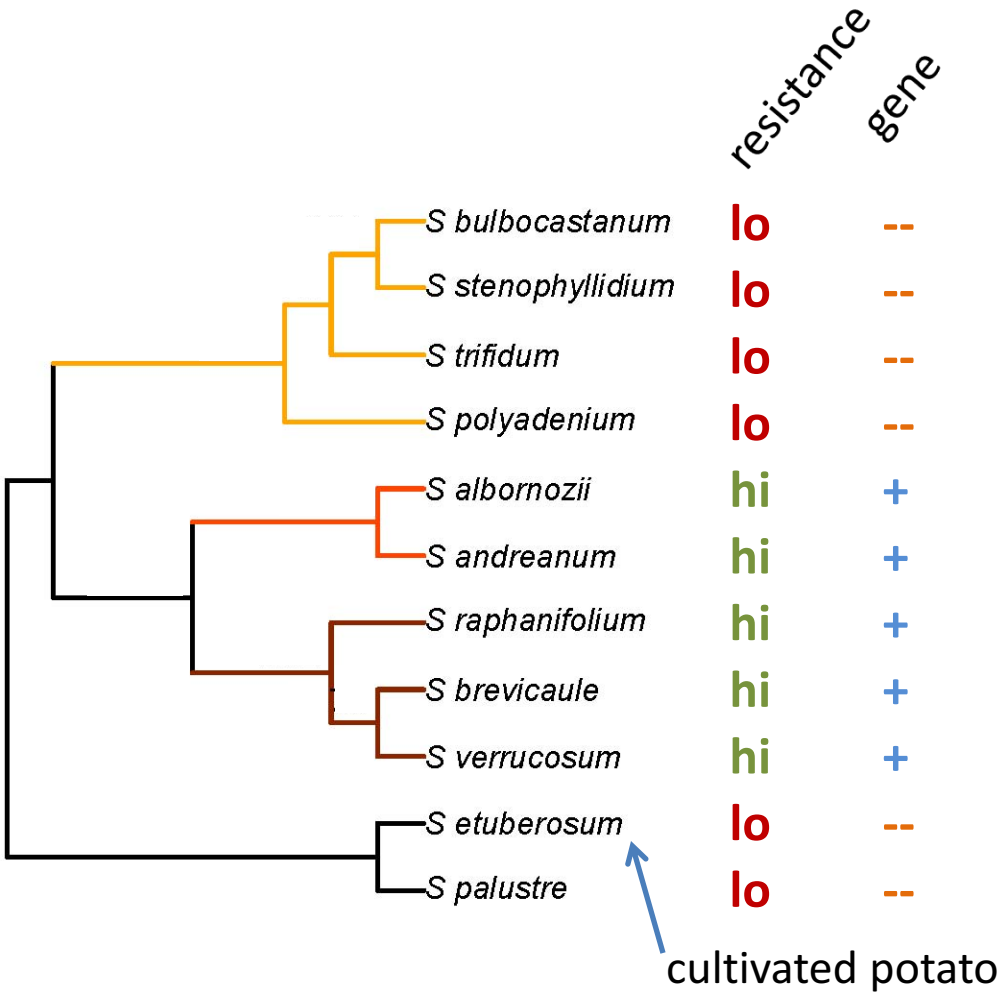
- genealogical relationships
- past hybridization / gene transfers
- shifts during evolution
- use biodiversity to find new genes

associated **software** (C++, R, Julia)
reproducibility

phylo networks

trees: data not “i.i.d”

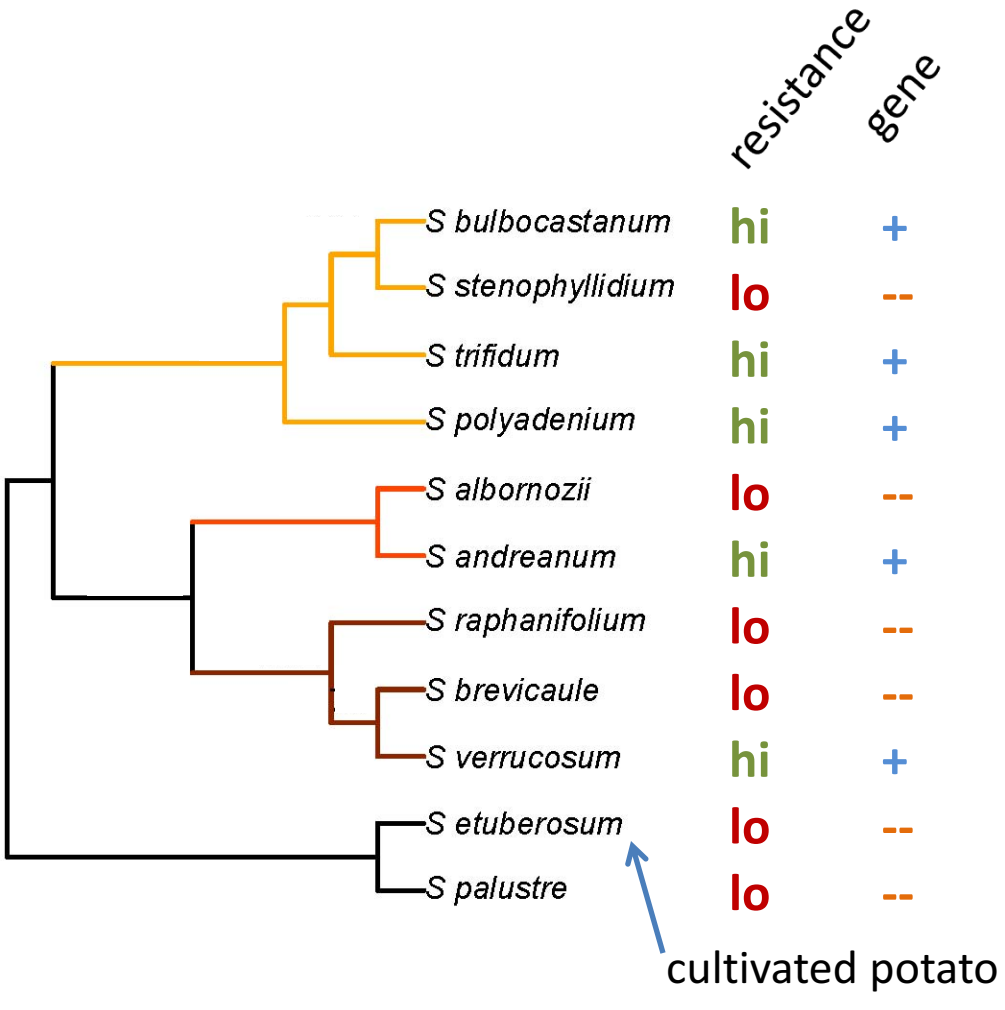
- data on species, or languages, or culture
- true correlation, or only shared ancestry?



	+	--
hi	5	0
lo	0	6

trees: data not “i.i.d”

- data on species, or languages, or culture
- true correlation, or only shared ancestry?



	+	--
hi	5	0
lo	0	6

infer trees & networks from DNA sequences

- space of trees: discrete, finite but huge

  – not standard $\mathbb{R}^n$
- 1000's genes, 100's species
- hierarchical models

- space of networks: discrete, infinite
- many open problems:

  identifiability

  model selection for network complexity

  how to scale to many species (or languages)

computing is an exciting part

- High performance: large cluster
- High throughput: 100s or 1000s of jobs run remotely

PhyloNetworks.jl

dev

Search docs

PhyloNetworks is a Julia package for the manipulation, visualization, inference of phylogenetic networks, and their use for trait evolution.

How to get help

- the package wiki has a step-by-step tutorial, done for the 2018 MBL workshop, with background on networks and explanations.

```
               _
   _       _ _(_)_     |  Documentation: https://docs.julialang.org
  (_)     | (_) (_)    |
   _ _   _| |_  __ _   |  Type "?" for help, "]?" for Pkg help.
  | | | | | | |/ _` |  |
  | | |_| | | | (_| |  |  Version 1.2.0 (2019-08-20)
 _/ |\__'_|_|_|\__'_|  |  Official https://julialang.org/ release
|__/                   |

julia> using RDatasets

julia> iris = dataset("datasets", "iris")
150×5 DataFrame
```

Row	SepalLength	SepalWidth	PetalLength	PetalWidth	Species
	Float64	Float64	Float64	Float64	Categorical…
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
⋮					
144	6.8	3.2	5.9	2.3	virginica
145	6.7	3.3	5.7	2.5	virginica

RA funding: starting fall 2020

- theory
- modelling
- computing
  software development
- collaboration

contact:

- cecile.ane@wisc.edu
- current & previous grad students

many collaborators on campus:

- stat, math
- biology (viruses, bacteria, plants, insects…)
- language sciences

Professor Michael Newton

https://arxiv.org/abs/1908.07176

(Mostly) Applied Machine Learning and Deep Learning

Applications in Biometrics, Computational Biology

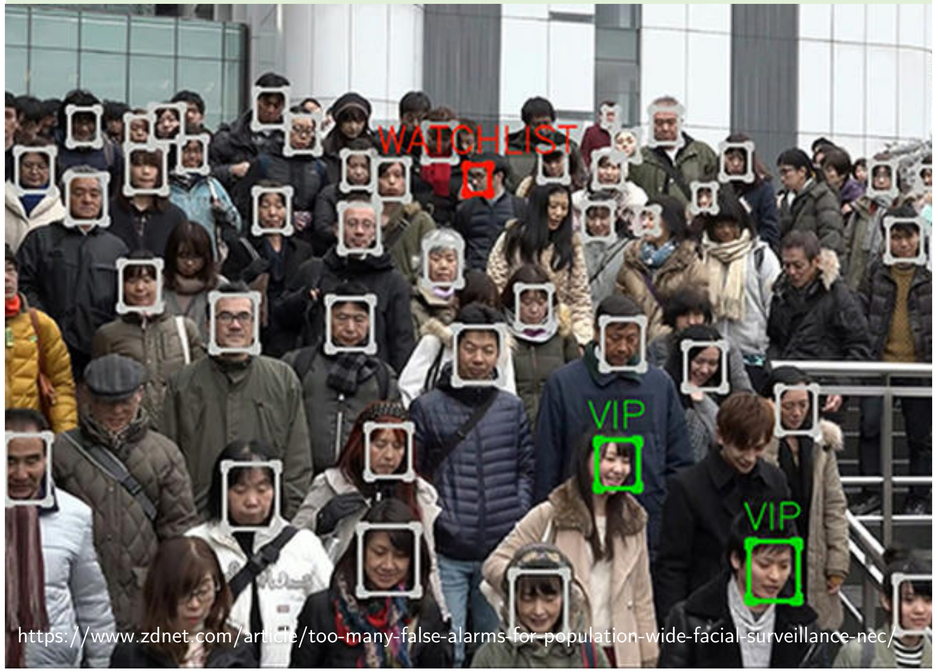
Sebastian Raschka, Ph.D.
Assistant Professor
Department of Statistics

sraschka@wisc.edu
http://stat.wisc.edu/~sraschka/

Applications of Biometric (Face) Recognition

A. Identification

Determine identity of an unknown person
1-to-*n* matching



https://www.zdnet.com/article/too-many-false-alarms-for-population-wide-facial-surveillance-nec/

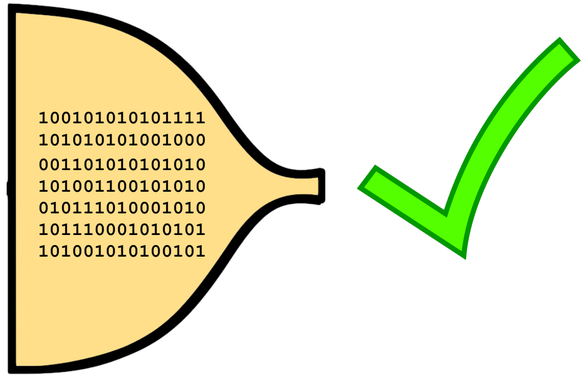
B. Verification

Verify claimed identity of a person
1-to-1 matching

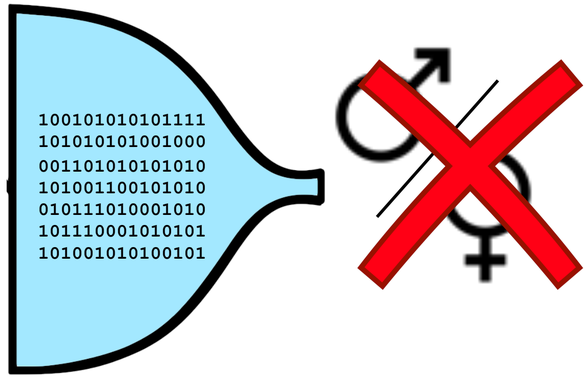
https://www.vosizneias.com/205132/2015/06/04/toronto-foreigners-entering-canada-to-undergo-biometric-screening/

Soft-biometric Attributes: Issues and Concerns

1. **Identity theft:** combining soft biometric info with publicly available data
2. **Profiling:** e.g., gender/race based profiling
3. **Ethics:** extracting data without users' consent

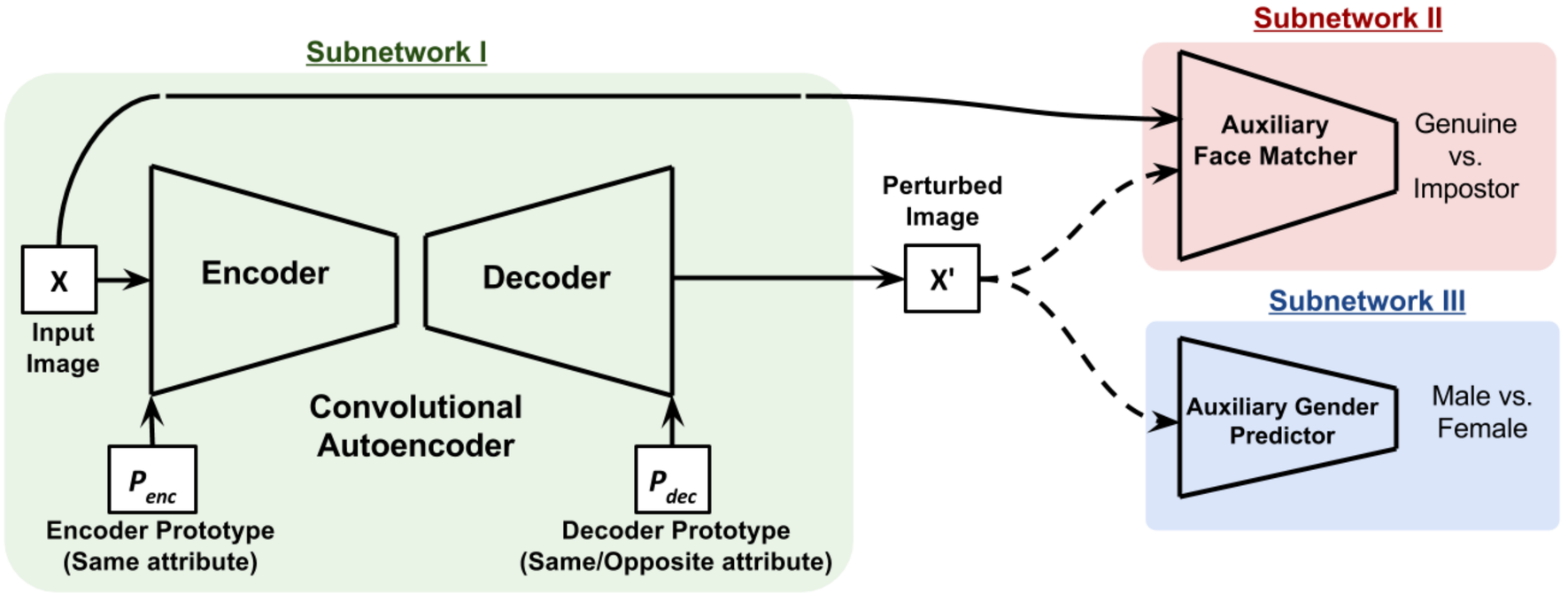


Face Matcher



Gender Classifier

Semi-adversarial Network (SAN), consisting of convolutional subnetworks for constrained optimization



Vahid Mirjalili, Sebastian Raschka, Anoop Namboodiri, and Arun Ross (2018) *Semi-adversarial networks: Convolutional autoencoders for imparting privacy to face images.* Proc. of 11th IAPR International Conference on Biometrics (ICB 2018)

Gender privacy: an ensemble of semi-adversarial networks for confounding arbitrary gender classifiers

Improvements to construct a more diverse set of SAN models for better generalizability via ensembling

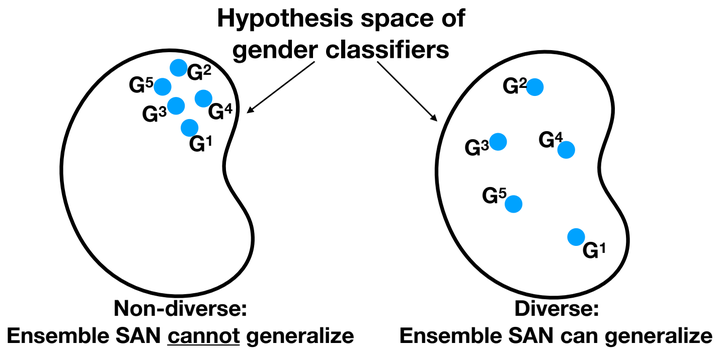


Figure 1: Diversity in an ensemble SAN can be enhanced through its auxiliary gender classifiers (see Figure 2). When the auxiliary gender classifiers lack diversity, ensemble SAN cannot generalize well to arbitrary gender classifiers.

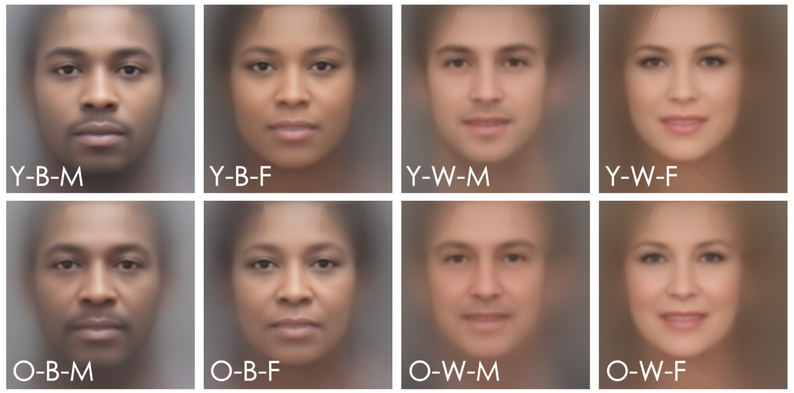
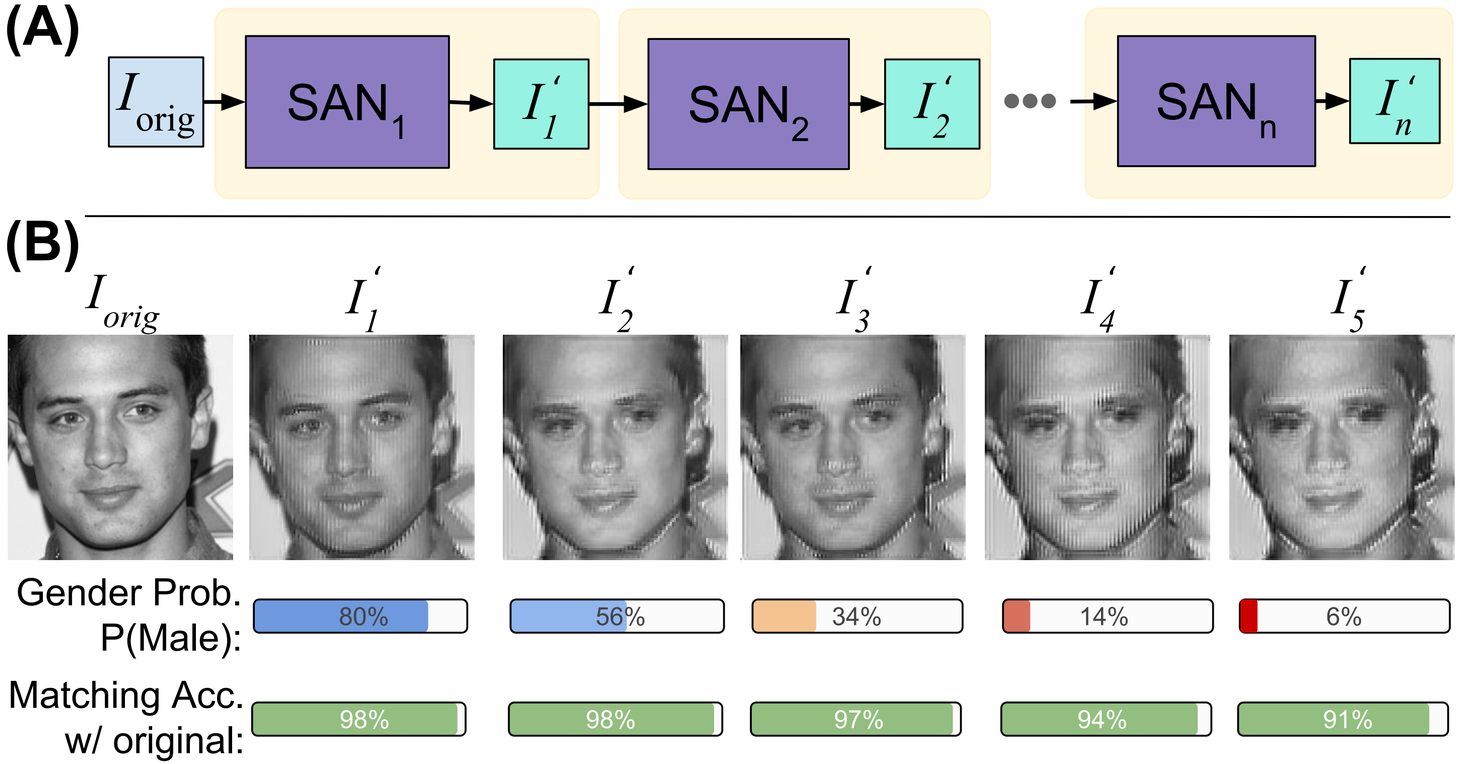


Figure 4: Face prototypes computed for each group of attribute labels. The abbreviations at the bottom of each image refer to the prototype attribute-classes, where Y=young, O=old, M=male, F=female, W=white, B=black.

Vahid Mirjalili, Sebastian Raschka, and Arun Ross (2018) *Gender Privacy: An Ensemble of Semi Adversarial Networks for Confounding Arbitrary Gender Classifiers.* 9th IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS 2018)

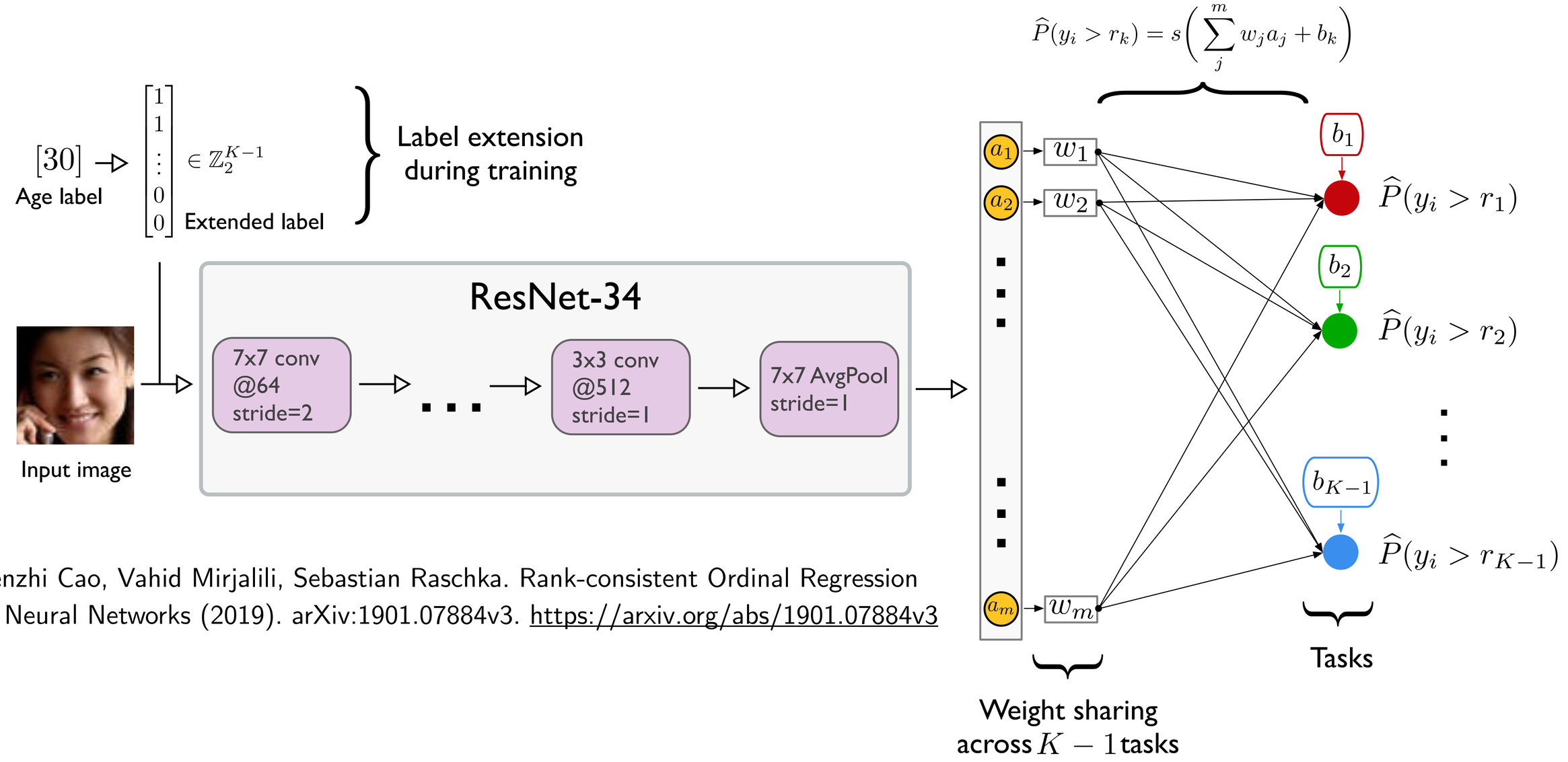
FlowSAN: Privacy-enhancing semi-adversarial networks to confound arbitrary face-based gender classifiers



Improvements to better control the perturbations and enhance the removal of soft-biometric information

Vahid Mirjalili, Sebastian Raschka, Arun Ross (2019)
FlowSAN: Privacy-enhancing Semi-Adversarial Networks to Confound Arbitrary Face-based Gender Classifiers
IEEE Access 2019, 10.1109/ACCESS.2019.2924619

Consistent Rank Logits (CORAL) for ordinal regression with neural networks



Wenzhi Cao, Vahid Mirjalili, Sebastian Raschka. Rank-consistent Ordinal Regression for Neural Networks (2019). arXiv:1901.07884v3. https://arxiv.org/abs/1901.07884v3

Niu, Z., Zhou, M., Wang, L., Gao, X., & Hua, G. (2016). Ordinal Regression with Multiple Output CNN for Age Estimation. CVPR.

Cao, W., Mirjalili V., Raschka S. (2019). Rank-consistent Ordinal Regression for Neural Networks. arXiv: 1901.07884v3. https://arxiv.org/abs/1901.07884v3

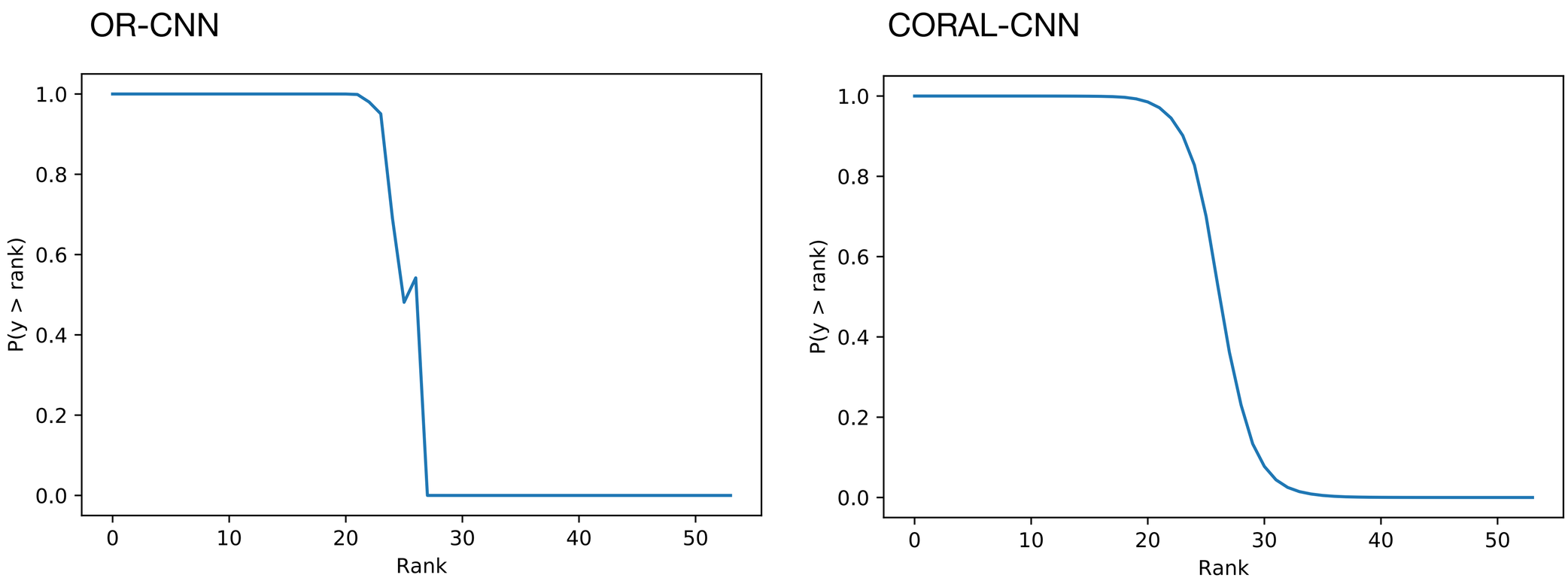
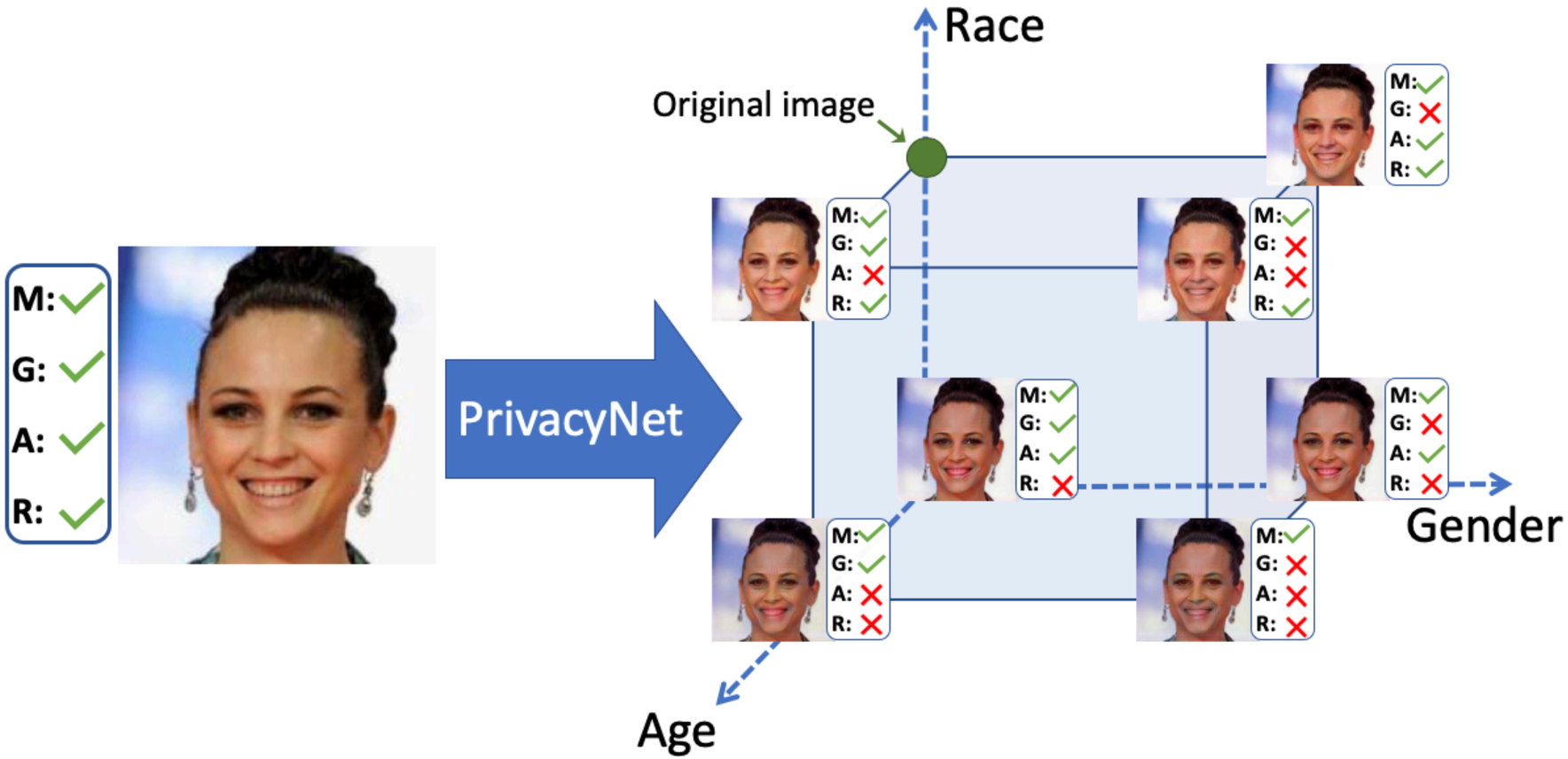


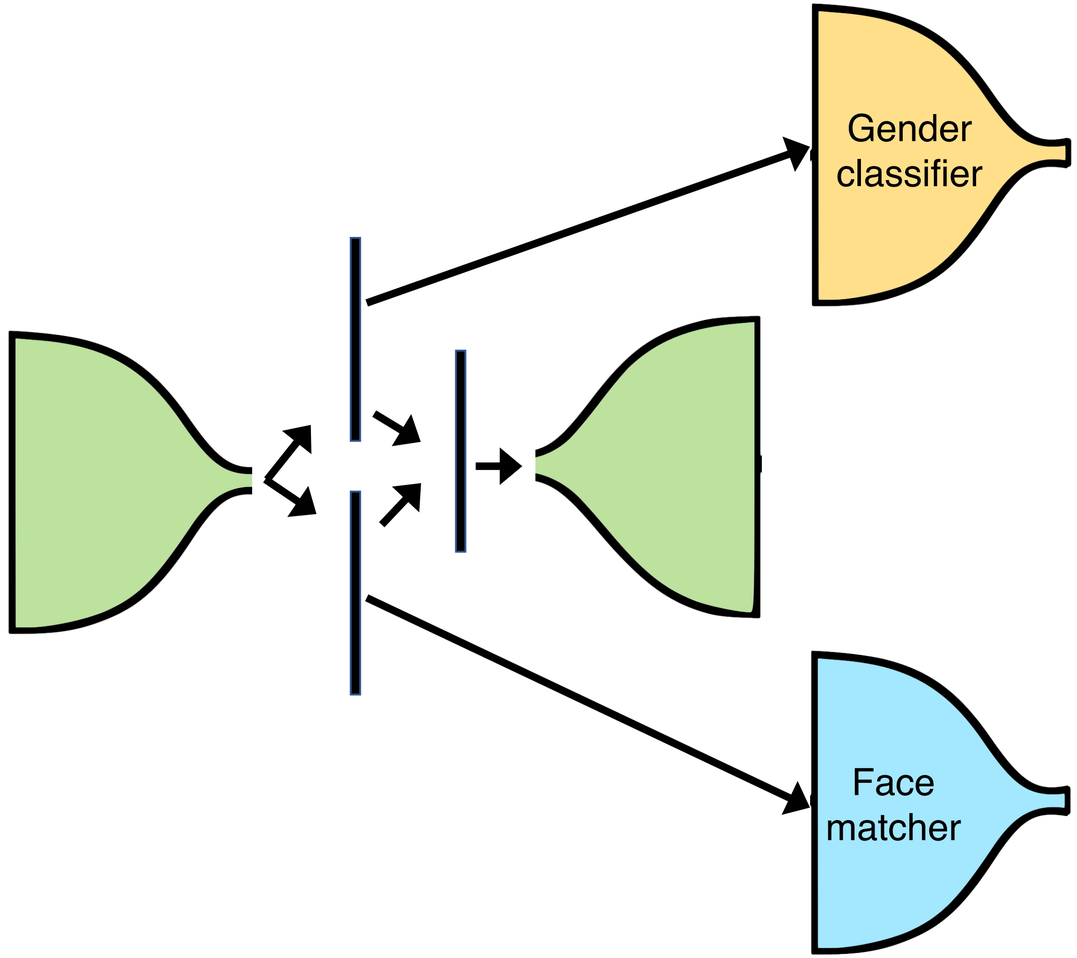
Figure S2: Plots show graphs of the predicted probabilities for each binary classifier task on one test data point in the MORPH dataset by OR-CNN (left subpanel) and CORAL-CNN (right subpanel). In this example, the ordinal regression CNN has an inconsistency at rank 26. The CORAL-CNN does not suffer from inconsistencies such that the rank prediction is a cumulative distribution function.

PrivacyNet: semi-adversarial networks for multi-attribute selective privacy



Vahid Mirjalili, Sebastian Raschka, and Arun Ross (2019) *PrivacyNet: Semi-Adversarial Networks for Multi-attribute Differential Privacy* (manuscript in prep.)

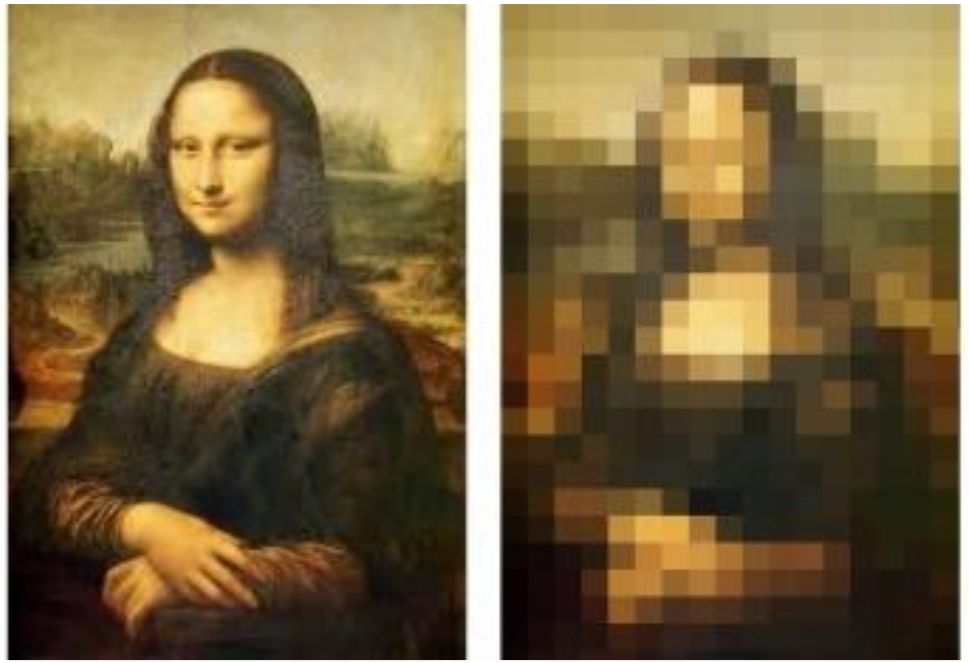
Better control over facial attributes
(e.g., better, controllable gender perturbation)
by manipulating latent space in (variational) autoencoders



One of the biggest issues/challenges in biometrics:

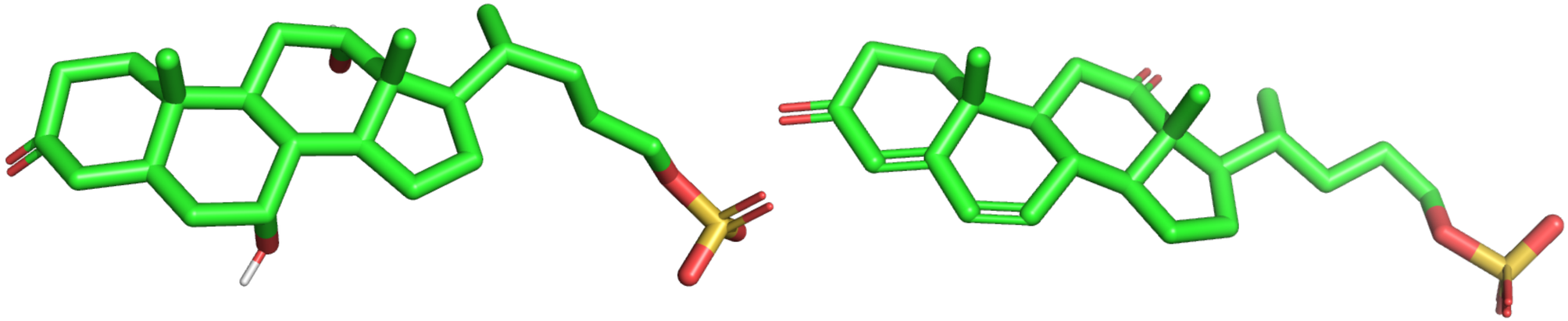
Face matching with low-resolution & noisy images, and images partially occluded faces

- **Constrained optimization approach (face matching)**
- **Face matching with graph convolutional networks**



Semi-adversarial Networks for molecule design

<u>Currently:</u> Focus on design and comparison of different graph-convolutional approaches for capturing molecular properties



<u>Future goal:</u> Ligand design (e.g., modify drug to increase solubility; synthesize drug such that binding affinity is maximized)

Research on "structure learning and statistical inference for large-scale data"

Chunming Zhang

http://www.stat.wisc.edu/~cmzhang/

Department of Statistics
University of Wisconsin-Madison

Overview

Earlier research work

- non-parametric and semi-parametric modeling, curve estimation and inference,
- with applications to **financial econometrics**, **longitudinal data analysis** and traffic forecasting in transportation.

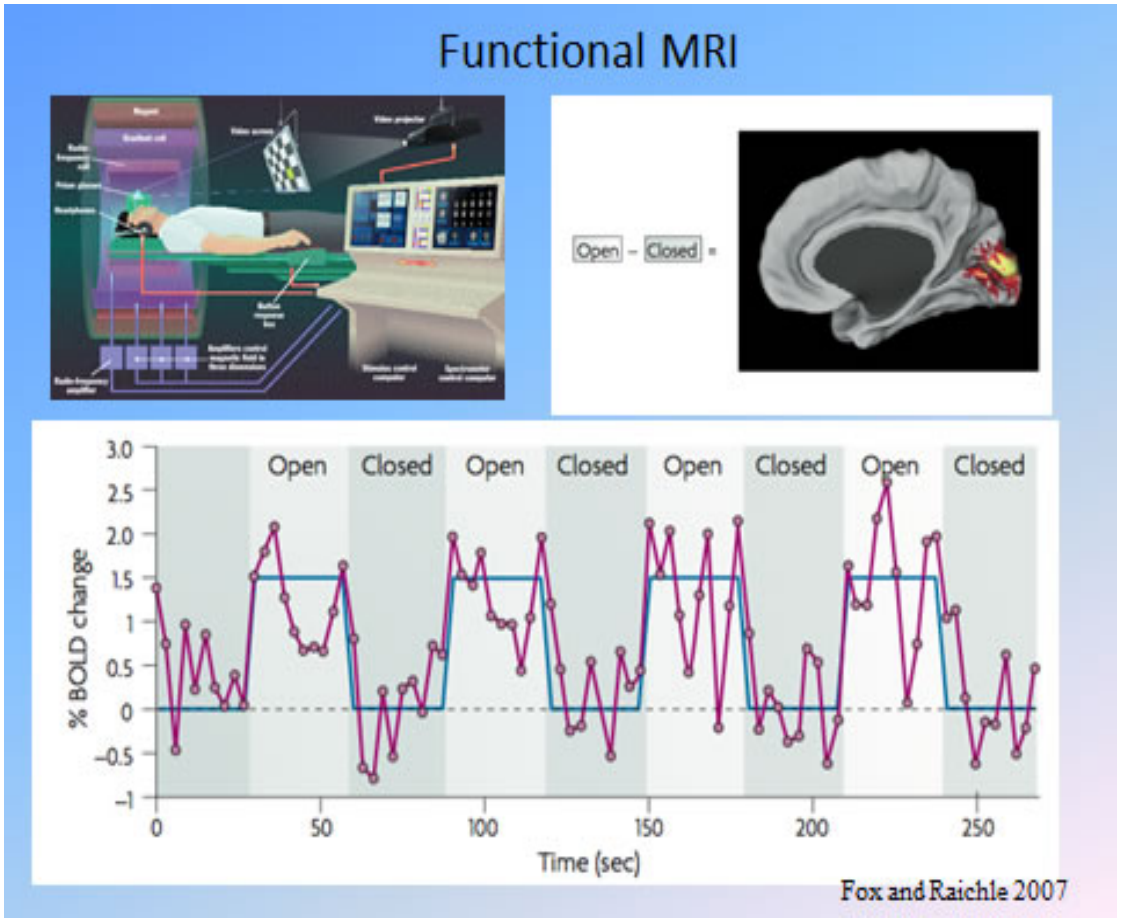
- Hall, P., Minnotte, M.C. and Zhang, C.M. (2004). "Bump hunting with non-Gaussian kernels". *Annals of Statistics*, 32, 2124–2141.
- Zhang, C.M. (2003). "Calibrating the degrees of freedom for automatic data smoothing and effective curve checking". *Journal of the American Statistical Association*, 98, 609–628.
- Fan, J. and Zhang, C.M. (2003). "A reexamination of diffusion estimators with applications to financial model validation". *Journal of the American Statistical Association*, 98, 118–134.
- Fan, J., Zhang, C.M., and Zhang, Jian (2001). "Generalized likelihood ratio statistics and Wilks phenomenon". *Annals of Statistics*, 29, 153–193.

Current research topic

- new developments in the area of large-scale **structure learning** and **statistical inference** procedures,
- with applications in neuroscience, biology, machine learning, and causal inference.

Applications to brain imaging data

functional Magnetic Resonance Imaging (fMRI) data:



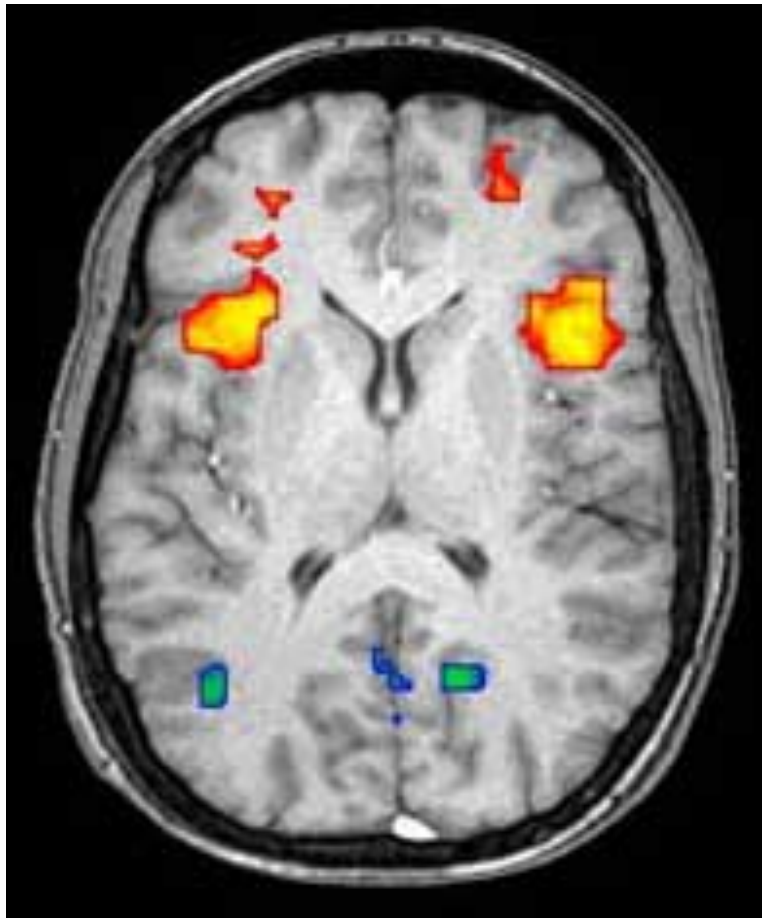


Figure 1: *Left:* **fMRI** *experiment; right: an* fMRI *scan with yellow areas showing increased activity.*

- Guo, X. and Zhang, C.M. (2015). “Estimation of the error autocorrelation matrix in semiparametric model for fMRI data”. *Statistica Sinica*, 25, 475–498.
- Zhang, C.M. and Yu, T. (2008). “Semiparametric detection of significant activation for brain fMRI”. *Annals of Statistics*, 36, 1693–1725.
- Zhang, C.M., Jiang, Y. and Yu, T. (2007). “A comparative study of one-level and two-level semiparametric estimation of hemodynamic response function for fMRI data”. *Statistics in Medicine*, 26, 3845–3861.

Diffusion Tensor Imaging (DTI) data:

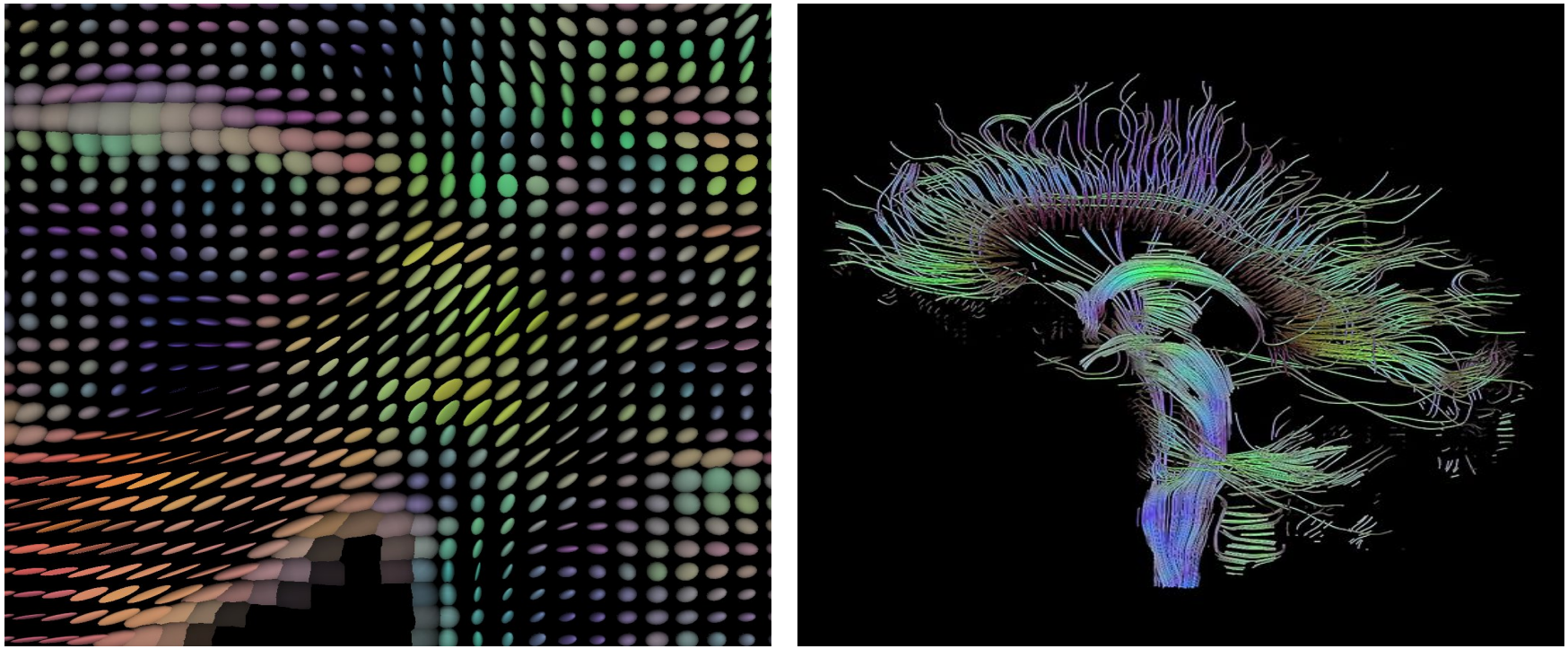


Figure 2: *Left: Visualization of* DTI *data with ellipsoids; right: Tractographic reconstruction of neural connections via* DTI.

- Yu, T., Zhang, C.M., Alexander, A.L., and Davidson, R.J. (2013). "Local tests for identifying anisotropic diffusion areas in human brain with DTI". *Annals of Applied Statistics*, 7, 201–225.

neuron spike train data:

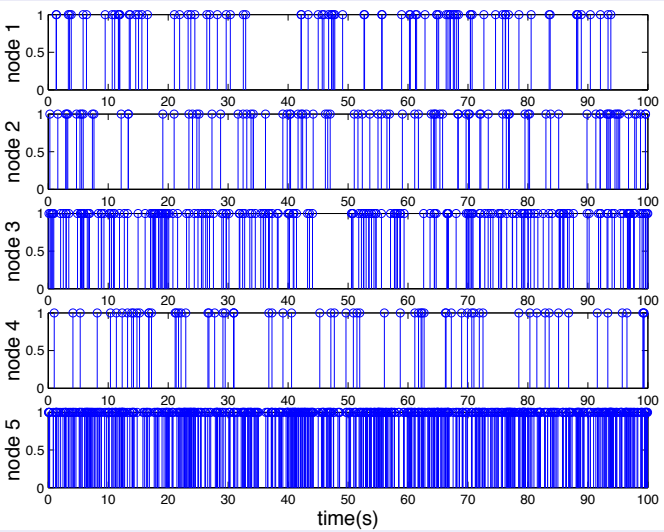


Figure 5: Samples generated from a random network with $V = 5$ and maximum number of parents is 2

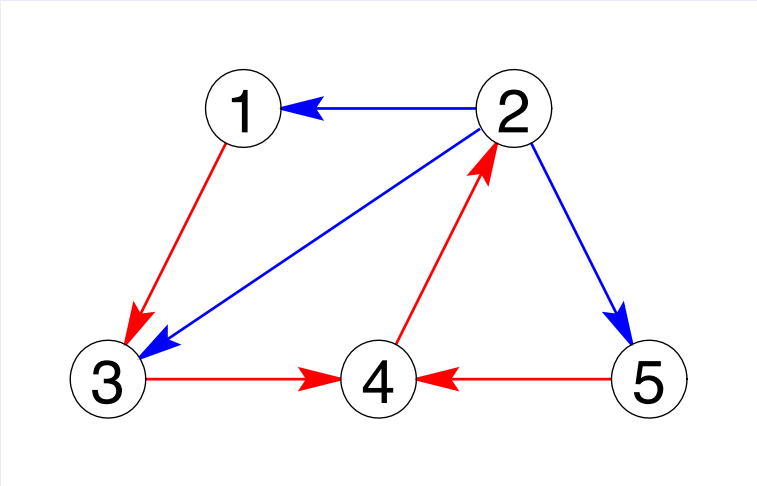


Figure 3: *Neuron spikes can trigger or inhibit neighboring neurons. Each node of the network graph in the right panel corresponds to a point process in the left panel. Arrows indicate interactions.* ***This directed graph has cycles.***

- Zhang, C.M., Chai, Y., Guo, X., Gao, M., Devilbiss, D.M. and Zhang, Z. (2016). "Statistical learning of neuronal functional connectivity". *Technometrics*, 58, 350–359.

Electro-Encephalo-Graphy (EEG) data:

Background

A commonly encountered but highly under-determined problem, "**blind source separation**" (in neuroscience, genetics, finance), aiming to separate

$$\text{"hidden components"} \quad S_1(t), \ldots, S_N(t)$$

from

$$\text{"observed signals" (i.e., "mixed signals")} \quad X_1(t), \ldots, X_p(t),$$

with very little information about either the components or the mixing mechanism.

- Guo, R.S., Zhang, C.M. and Zhang, Z.J. (2019). "Maximum independent component analysis with application to EEG data". *Statistical Science*, invited minor revision.

Multiple testing with applications to GWAS and fMRI

Motivation

- In **GWAS**, hundreds of thousands of highly correlated genetic markers (SNPs) are collected, with the purpose of identifying the **subset of SNPs** associated with a disease.
- In **fMRI**, thousands of spatially correlated brain voxels are scanned while subjects are performing certain tasks, with the purpose of detecting **activated voxels** in response to cognitive activity.

- Liu, J., Zhang, C.M., and Page, D. (2016). "Multiple testing under dependence via graphical models". *Annals of Applied Statistics*, 10, 1699–1724.
- Du, L. and Zhang, C.M. (2014). "Single-index modulated multiple testing". *Annals of Statistics*, 42, 1262–1311.
- Zhang, C.M., Fan, J. and Yu, T. (2011). "Multiple testing via FDR_L for large-scale imaging data". *Annals of Statistics*, 39, 613–642.

More technical parts

Formulation:

$$\widehat{\boldsymbol{\theta}} \quad = \quad \arg \min_{\boldsymbol{\theta}=(\theta_1,\ldots,\theta_p)^T} \left\{ \text{loss}(\boldsymbol{\theta}; D_1, \ldots, D_n) + \text{penalty}(\boldsymbol{\theta}) \right\}. \tag{1}$$

Issues:

- new **optimization** techniques for computationally intensive regularized estimators:

  how to numerically solve $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_p)^T$?

- related **statistical inference** for parameters in high-dimensional settings:

  how to do interval estimation and hypothesis testing of θ_j and $\boldsymbol{a}^T\boldsymbol{\theta}$ via $\widehat{\theta}_j$?

- procedures of **controlling false discoveries** in large-scale simultaneous inference:

  how to handle multiplicity issue in simultaneous inference of $\theta_1, \ldots, \theta_p$?

Research overview

Po-Ling Loh

ploh@stat.wisc.edu

SGSA meeting

August 30, 2019

Brief intro to robust statistics

- Robust statistics introduced in 1960s (Huber, Tukey, Hampel, et al.)

Brief intro to robust statistics

- Robust statistics introduced in 1960s (Huber, Tukey, Hampel, et al.)
- **Goals:**
  1. Develop estimators $T(\cdot)$ that are reliable under deviations from model assumptions
  2. Quantify performance with respect to deviations

Brief intro to robust statistics

- Robust statistics introduced in 1960s (Huber, Tukey, Hampel, et al.)
- **Goals:**
  1. Develop estimators $T(\cdot)$ that are reliable under deviations from model assumptions
  2. Quantify performance with respect to deviations
- Local stability captured by *influence function*

$$IF(x; T, F) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)F + \epsilon\Delta_x) - T(F)}{t}$$

Brief intro to robust statistics

- Robust statistics introduced in 1960s (Huber, Tukey, Hampel, et al.)
- **Goals:**
  1. Develop estimators $T(\cdot)$ that are reliable under deviations from model assumptions
  2. Quantify performance with respect to deviations
- Local stability captured by *influence function*

$$IF(x; T, F) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)F + \epsilon \Delta_x) - T(F)}{t}$$

- Global stability captured by *breakdown point*

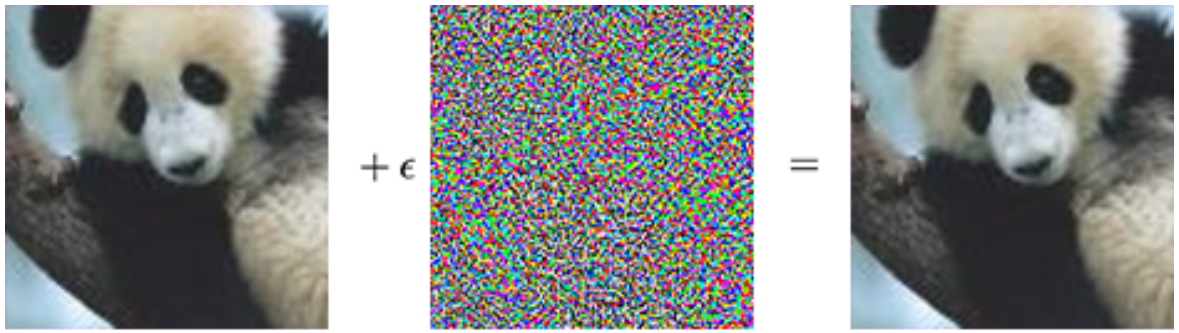
$$\epsilon^*(T; X_1, \ldots, X_n) = \min \left\{ \frac{m}{n} : \sup_{X^m} \| T(X^m) - T(X) \| = \infty \right\}$$

Adversarial contamination

- Instead of drawing i.i.d. data from an ϵ-contaminated mixture $(1-\epsilon)F + \epsilon\Delta_x$, draw i.i.d. data points $\{x_1, \ldots, x_n\}$ and **arbitrarily contaminate** an ϵ-fraction

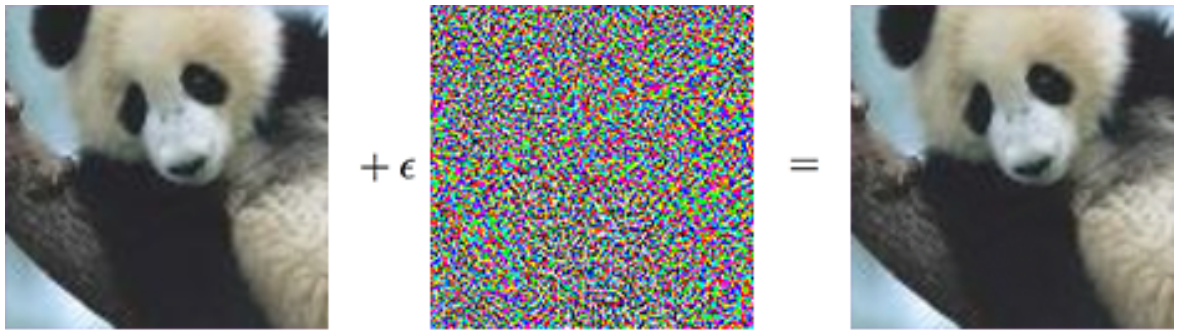
Adversarial contamination

- Instead of drawing i.i.d. data from an ϵ-contaminated mixture $(1-\epsilon)F + \epsilon\Delta_x$, draw i.i.d. data points $\{x_1, \ldots, x_n\}$ and **arbitrarily contaminate** an ϵ-fraction
- "Adversarial machine learning": Targeted attacks to neural networks



Adversarial contamination

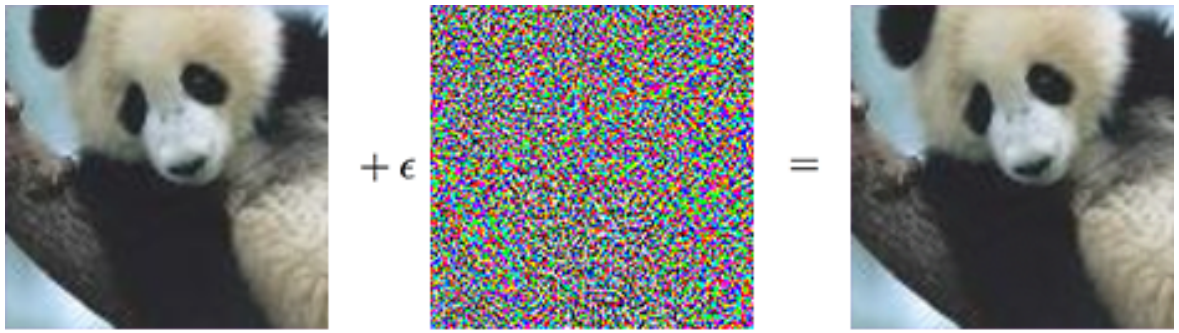
- Instead of drawing i.i.d. data from an ϵ-contaminated mixture $(1-\epsilon)F + \epsilon\Delta_x$, draw i.i.d. data points $\{x_1, \ldots, x_n\}$ and **arbitrarily contaminate** an ϵ-fraction
- "Adversarial machine learning": Targeted attacks to neural networks



- **Key question:** When can tools from classical robust statistics be leveraged to counteract adversarially contaminated data?

Adversarial contamination

- Instead of drawing i.i.d. data from an ϵ-contaminated mixture $(1-\epsilon)F + \epsilon\Delta_x$, draw i.i.d. data points $\{x_1, \ldots, x_n\}$ and **arbitrarily contaminate** an ϵ-fraction
- "Adversarial machine learning": Targeted attacks to neural networks

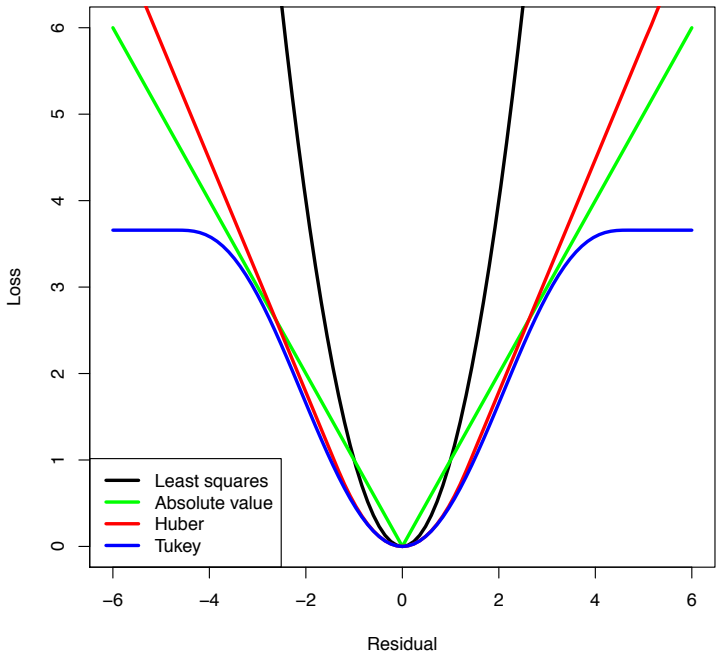


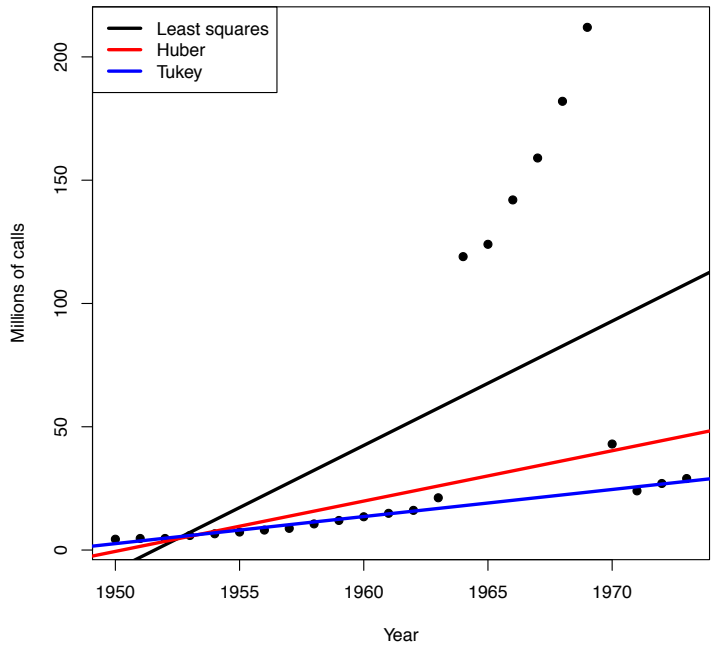
- **Key question:** When can tools from classical robust statistics be leveraged to counteract adversarially contaminated data?
- What about other types of contamination?

"Robust" M-estimation

- Generalization of OLS suitable for heavy-tailed/contaminated errors:

$$\widehat{\beta} \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(x_i^T \beta - y_i) \right\}$$



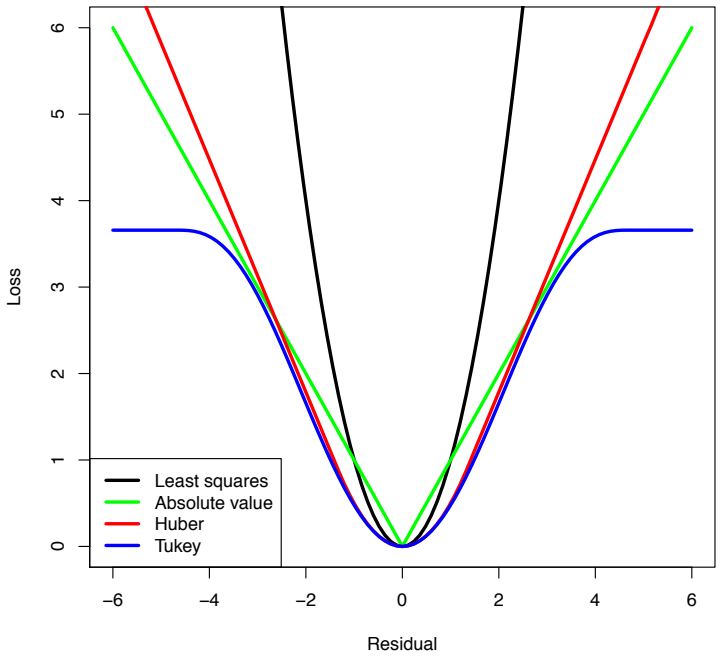


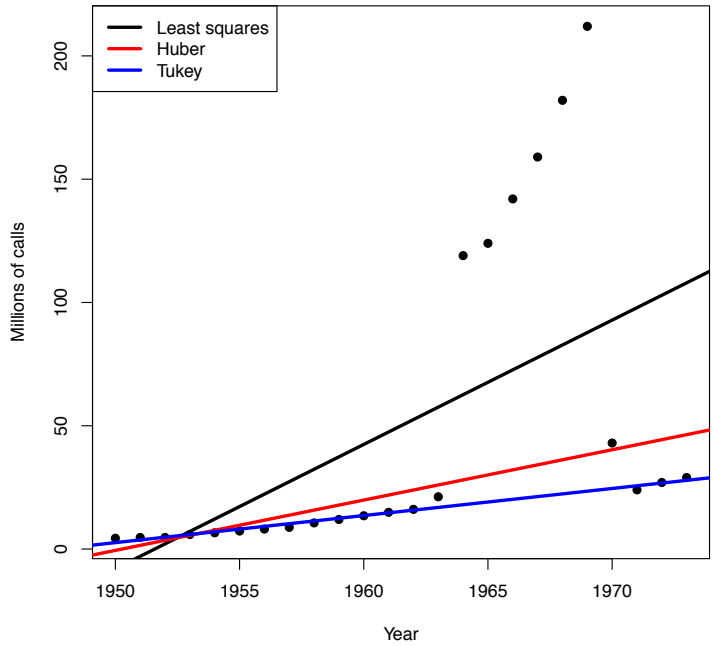
"Robust" M-estimation

- Generalization of OLS suitable for heavy-tailed/contaminated errors:

$$\widehat{\beta} \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(x_i^T \beta - y_i) \right\}$$

- Extensive theory (consistency, asymptotic normality) for p fixed, $n \to \infty$





High-dimensional M-estimators

- Natural idea: For $p > n$, use **regularized** version:

$$\widehat{\beta} \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(x_i^T \beta - y_i) + \lambda \|\beta\|_1 \right\}$$

High-dimensional M-estimators

- Natural idea: For $p > n$, use **regularized** version:

$$\widehat{\beta} \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(x_i^T \beta - y_i) + \lambda \|\beta\|_1 \right\}$$

Questions:

- Optimization for nonconvex/nonsmooth ℓ?

High-dimensional M-estimators

- Natural idea: For $p > n$, use **regularized** version:

$$\widehat{\beta} \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(x_i^T \beta - y_i) + \lambda \|\beta\|_1 \right\}$$

Questions:

- Optimization for nonconvex/nonsmooth ℓ?
- Statistical theory: Are certain losses provably better than others?

High-dimensional M-estimators

- Natural idea: For $p > n$, use **regularized** version:

$$\widehat{\beta} \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(x_i^T \beta - y_i) + \lambda \|\beta\|_1 \right\}$$

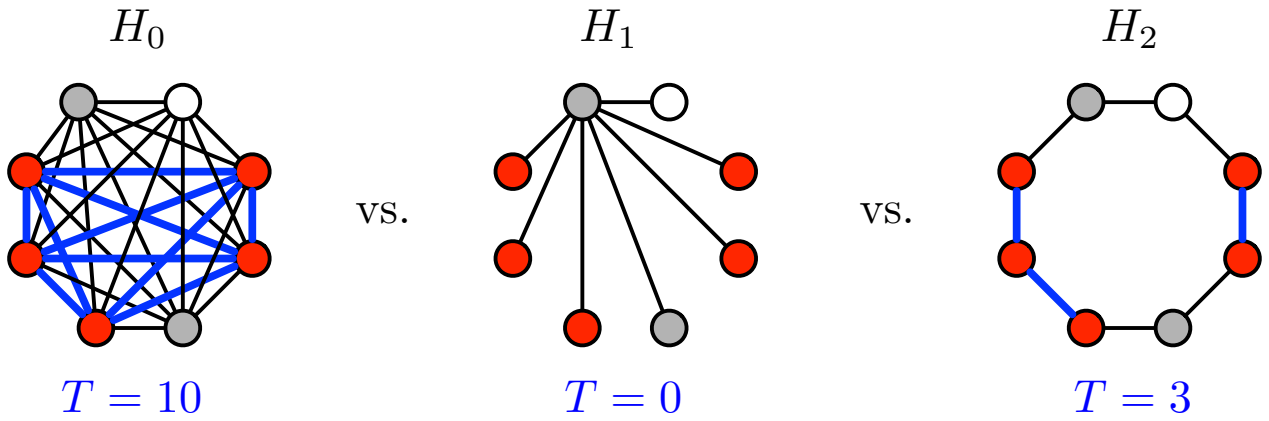
Questions:

- Optimization for nonconvex/nonsmooth ℓ?
- Statistical theory: Are certain losses provably better than others?

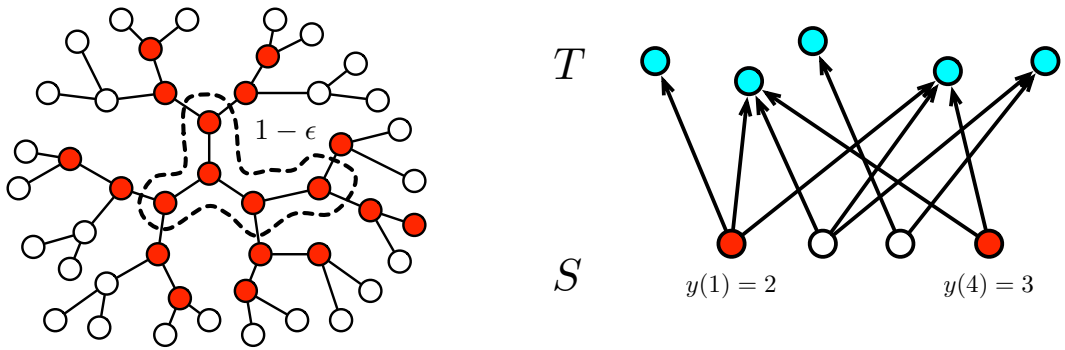
- Can we do better using non-M-estimators?

Other research interests (Network science)

- Given data from a network, how do we estimate the network?



- How do we model dynamic processes over a network?



- How do we perform efficient search over a network?

Thank you!

Vivak Patel

Fundamentally, my work uses statistical concepts to solve problems arising in linear algebra and optimization.

Numerical Linear Algebra

(1) Suppose we want to solve

$$\underbrace{A}_{\mathbb{R}^{n\times d}}\ \underbrace{x}_{\mathbb{R}^{d}} = \underbrace{b}_{\mathbb{R}^{n}}$$

for x.

(2) When n, d are very large (e.g., simulations, optimization, data assimilation), we can use randomized methods to solve these problems quickly

Numerical Linear Algebra

(3) IN OUR WORK, we leverage previously observed information from basic randomized methods to speed up convergence

* with mathematical guarantees
* without sacrificing practical considerations

Linear Regression

(1) From the basic linear system, we can move on to linear regression, in which we want to solve

$$\min_{x \in \mathbb{R}^d} \| \underbrace{A}_{\mathbb{R}^{n \times d}} x - \underbrace{b}_{\mathbb{R}^n} \|_2$$

(2) These problems are very hard when n, d are large

Linear Regression

(3) Many people use randomized methods to solve these problems quickly (e.g. SGD)

(4) But most of these methods

* lack mathematical guarantees
* ignore practical considerations

(5) IN OUR WORK, we design methods that have both

Extensions

(1) Generalized linear models

(2) Quasi likelihood models (e.g, DNN)

(3) Problems arising in engineering

* Power systems
* Finance
* ?

Fundamental Tools

(1) Measure-theoretic Probability
   * Martingales
   * Markov Chains (general state space)

(2) Linear Algebra
   * Theory
   * Numerical Analysis

(3) Optimization
   * Geometric Fonctional Analysis
   * Numerical Methods

Fundamental Tools

(4) Programming

* Julia
* Python (3)
* C | C++

Thank You!

www.vivakpatel.org

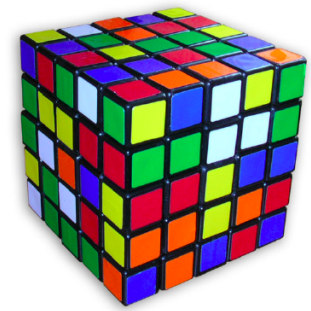
Beyond matrices: tensor decomposition and its application in sciences

Miaoyan Wang

Department Faculty lighting talk

Thanks: NSF DMS 1915978 and OVCRG Grant

August 30, 2019



Who am I?

Originally from China,
B.S. in Mathematics,
Fudan University

My research

Statistical machine learning:

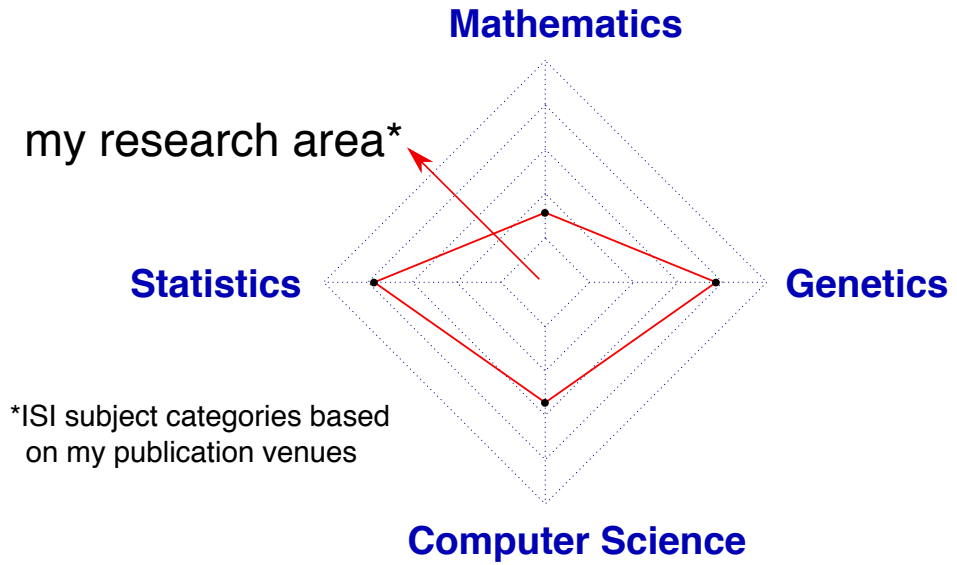
- tensor/matrix decomposition, high-dimensional statistics.

Applied Mathematics:

- numerical algebra, multilinear optimization, combinatorics.

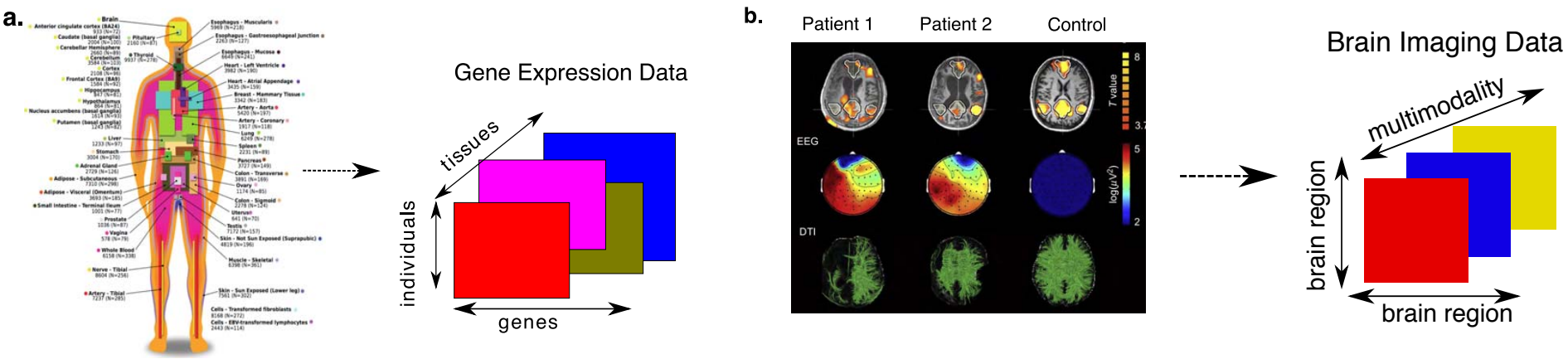
Genetics:

- genetic association studies, gene expression, neuroimaging.



Tensors in biology

- Many biomedical datasets come naturally in a multiway form.
- Multi-tissue, multi-individual gene expression measures could be organized as an order-3 tensor $\mathcal{A} = [\![a_{git}]\!] \in \mathbb{R}^{n_G \times n_I \times n_T}$.



Multi-way Clustering in Gene Expression

To identify subsets of genes that are similarly expressed within subsets of individuals and tissues, we seek local blocks in the expression tensor.

Tensors in statistical modeling

"Tensors are the new matrices" that tie together a wide range of areas:

- Longitudinal social network data $\{\boldsymbol{Y}_t : t = 1, ..., n\}$
- Spatio-temporal transcriptome data
- Joint probability table of a set of variables $\mathbb{P}(X_1, X_2, X_3)$
- Higher-order moments in single topic models
- Markov models for the phylogenetic tree $K_{1,3}$

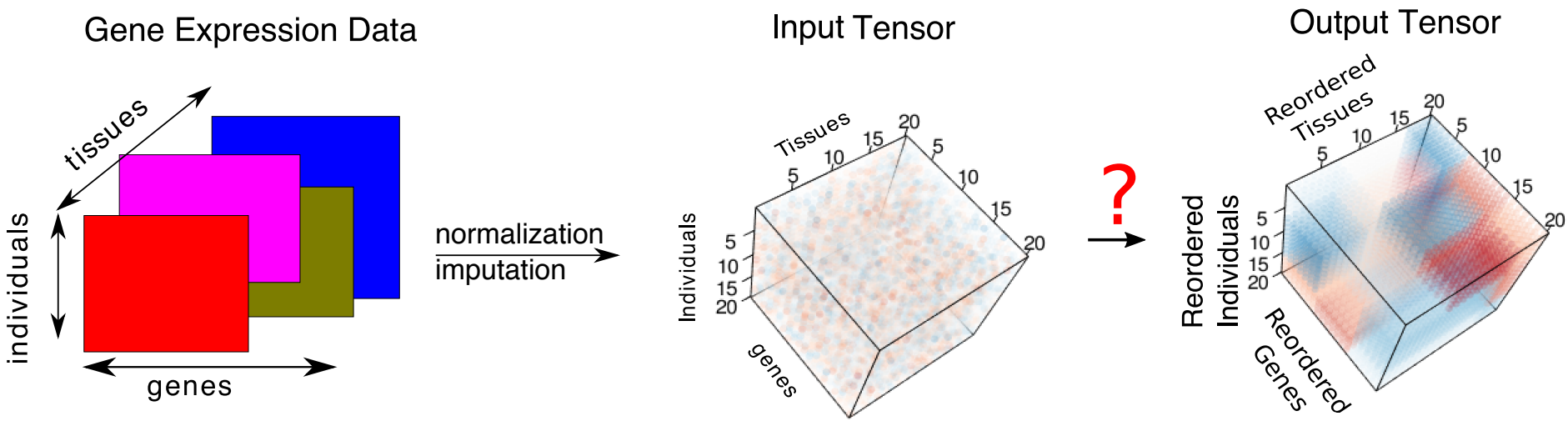
M. Yuan et al 2017, P. Hoff 2015, Montanari-Richard 2014

Anandkumar et al 2014, Mossel et al 2004, P. McCullagh 1987

Why study tensors?

Tensors provide a rich source of

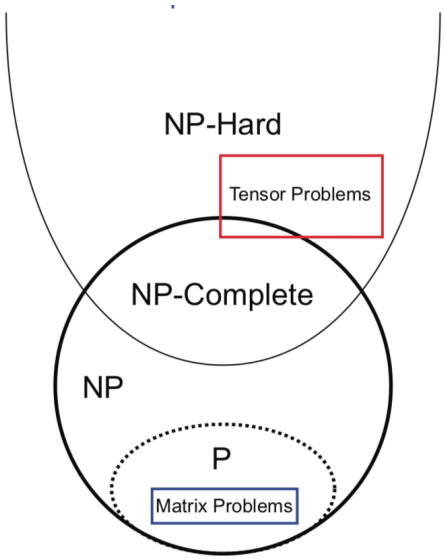
- fundamental problems in data science.
- new tools for long-standing questions.
- huge potentials for new applications.



My research

Prohibitive Computational Complexity

Most higher-order tensor problems are NP-hard [Hillar & Lim, 2013].



Fortunately, the tensors sought in statistical and machine learning applications are often specially structured:

- Low-rankness
- Sparsity
- Non-negativity
- ...

Breaking previous limits

My group is developing a framework of statistical models, efficient algorithms, and fundamental theory to analyze large-dimension large-scale tensor/matrix data.

For potential Ph.D students

Good niche if you are

- comfortable with mathematical statistics, probability theory, and/or computational statistics.
- enthusiastic about using your quantitive skills to advance our understanding in sciences.
- actively interested in learning about nearby research areas that interest you. (Which areas they are are up to you.)

As an advisor, I will

- provide you the skills, the experiences, and the connections that make you excel in your future career.
- be a demanding advisor, probably more demanding than most :)
- not be a good fit for you if you are not interested in regularly attending talks, or if you do not like getting technical in a serious way.

If the above sounds interesting to you, drop by my office and talk with me!

Optimal transport

...or 5-7 minute propaganda for Stat 992

What do all these things have in common?

- GANs (generative adversarial networks).
- Shape comparison.
- Shape interpolation.
- Fokker Planck equations.
- Aggregation equations.
- Fair district drawing.

What do all these things have in common?

- **GANs (generative adversarial networks).**
- Shape comparison.
- Shape interpolation.
- Fokker Planck equations.
- Aggregation equations.
- Fair district drawing.

*taken from https://www.analyticsvidhya.com

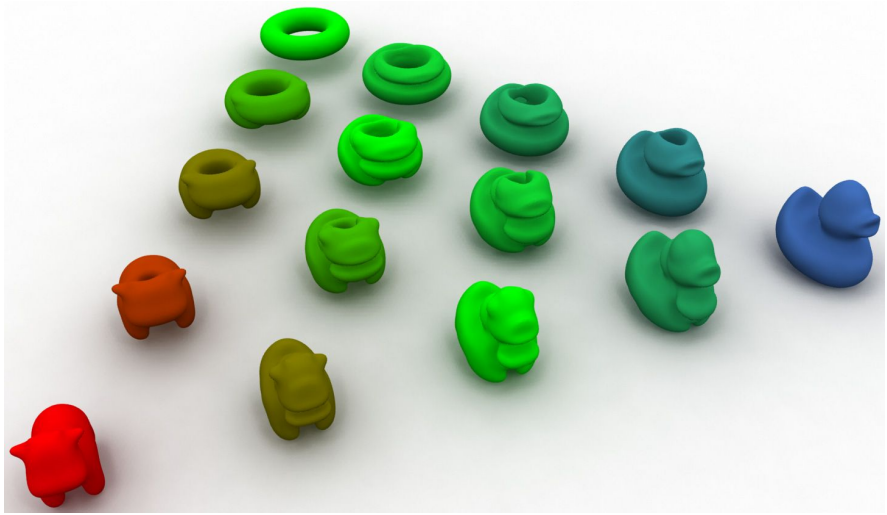
What do all these things have in common?

- GANs (generative adversarial networks).
- **Shape comparison.**
- Shape interpolation.
- Fokker Planck equations.
- Aggregation equations.
- Fair district drawing.

***from** David Gu

What do all these things have in common?

- GANs (generative adversarial networks).
- Shape comparison.
- **Shape interpolation.**
- Fokker Planck equations.
- Aggregation equations.
- Fair district drawing.



*from Su et al

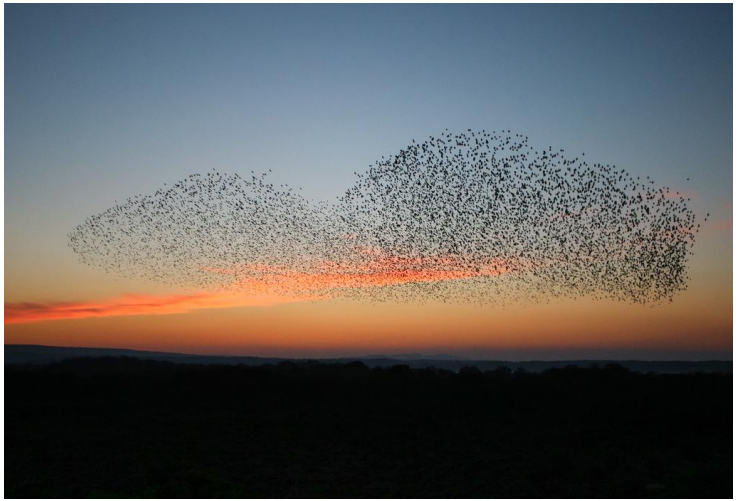
What do all these things have in common?

- GANs (generative adversarial networks).
- Shape comparison.
- Shape interpolation.
- **Fokker Planck equations.**
- Aggregation equations.
- Fair district drawing.

$$\partial_t \rho = -\Delta \rho$$

What do all these things have in common?

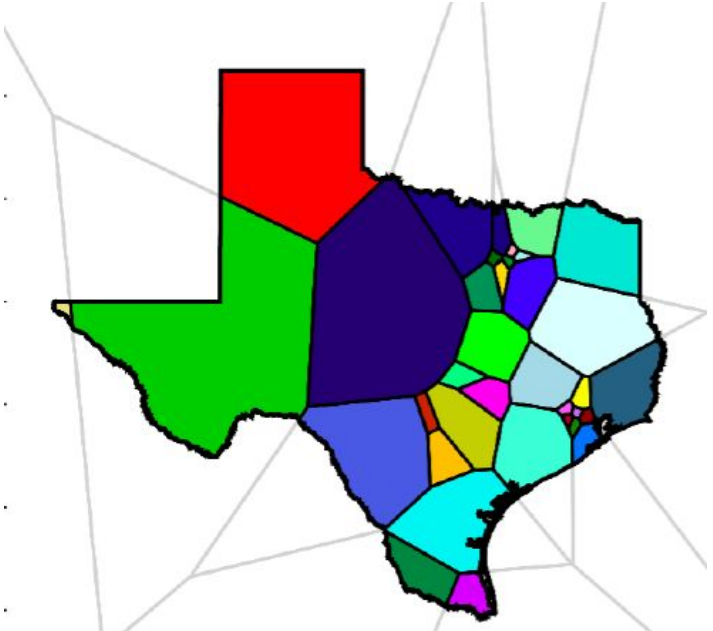
- GANs (generative adversarial networks).
- Shape comparison.
- Shape interpolation.
- Fokker Planck equations.
- **Aggregation equations.**
- Fair district drawing.



***from** Gudrun Wallentin

What do all these things have in common?

- GANs (generative adversarial networks).
- Shape comparison.
- Shape interpolation.
- Fokker Planck equations.
- Aggregation equations.
- **Fair district drawing.**

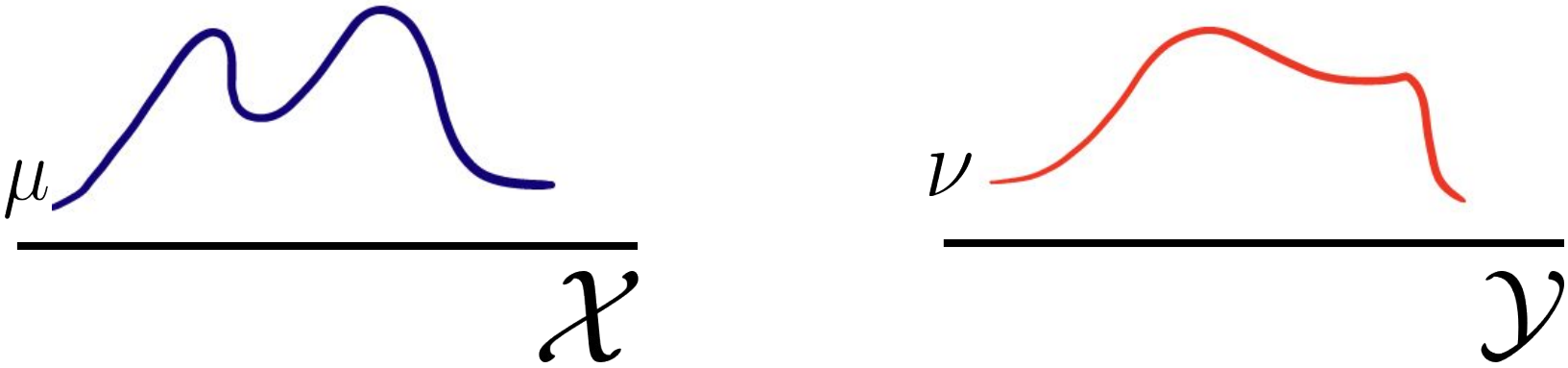


*from Cohen-Addad et al

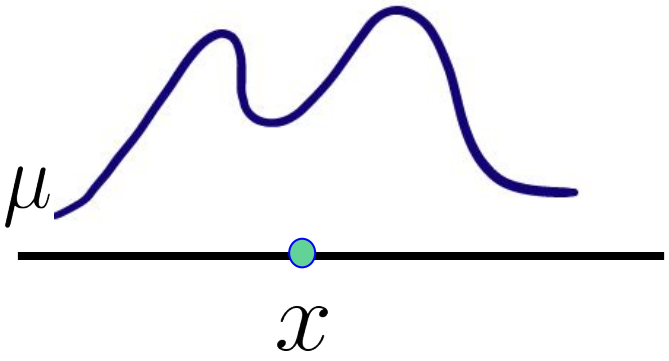
Optimal transport

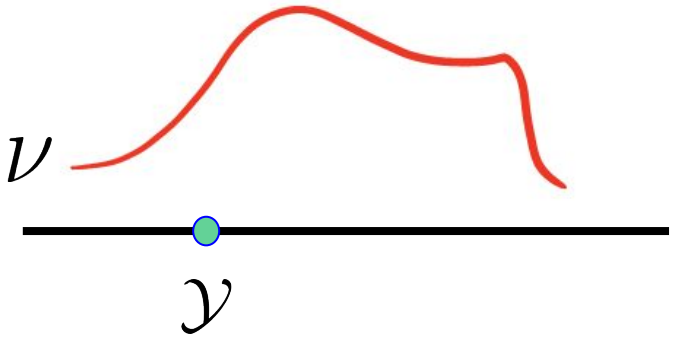
- For modelling
- Use it as an objective function
- Use it to define spaces where one can take “gradients”
- For understanding, and to help prove theorems

But what is optimal transport?



But what is optimal transport?

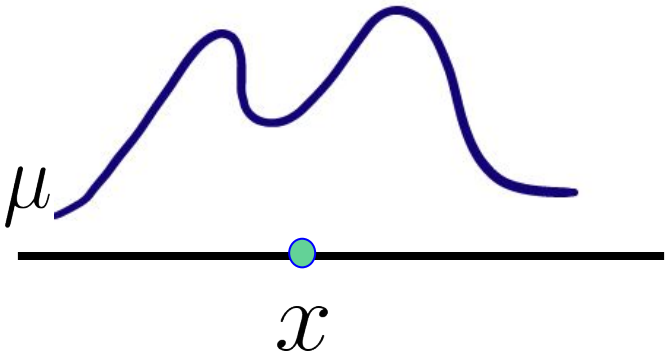


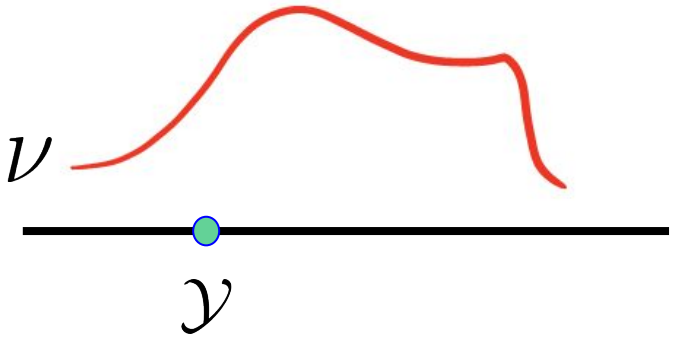


1. Transport plan $\prod$

$\Pi(x, y)$

But what is optimal transport?





1. Transport plan $\prod$ $\Pi(x, y)$
2. Transport cost $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ $c(x, y)$

But what is optimal transport?

$$\Pi \longrightarrow \int_x \int_y c(x,y) d\Pi(x,y)$$

$$\min_{\Pi} \int_{\mathcal{X}} \int_{\mathcal{Y}} c(x, y) d\Pi(x, y)$$

$$\min_{\Pi} \int_{\mathcal{X}} \int_{\mathcal{Y}} c(x,y) d\Pi(x,y)$$

Optimal transport can be considered as a subfield of:

-Linear programming

-Probability and Statistics

-Geometry

-Partial differential equations

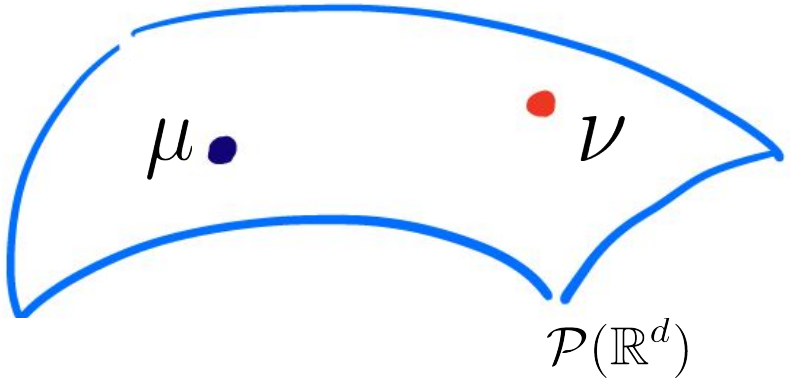
-...

Let’s consider a typical setting in applications: Euclidean case

$$\mathcal{X} = \mathbb{R}^d = \mathcal{Y} \qquad c(x, y) = |x - y|^2$$

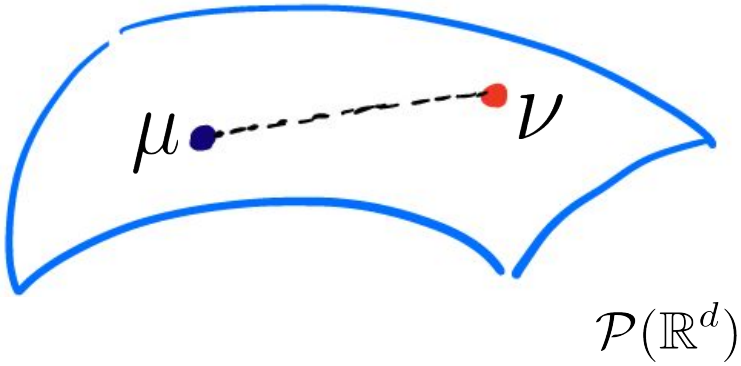
Let’s consider a typical setting in applications: Euclidean case

$$\mathcal{X} = \mathbb{R}^d = \mathcal{Y} \qquad c(x, y) = |x - y|^2$$



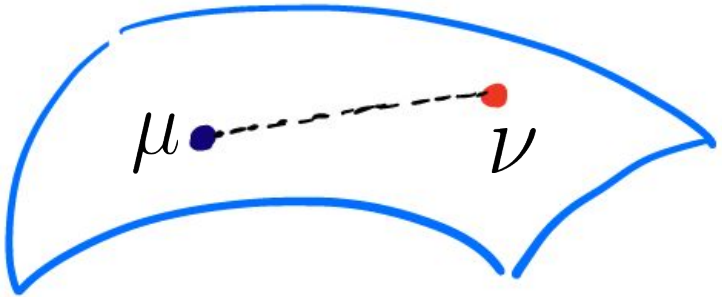
Let’s consider a typical setting in applications: Euclidean case

$$\mathcal{X} = \mathbb{R}^d = \mathcal{Y} \qquad c(x, y) = |x - y|^2$$



Let’s consider a typical setting in applications: Euclidean case

$$\mathcal{X} = \mathbb{R}^d = \mathcal{Y} \qquad c(x, y) = |x - y|^2$$

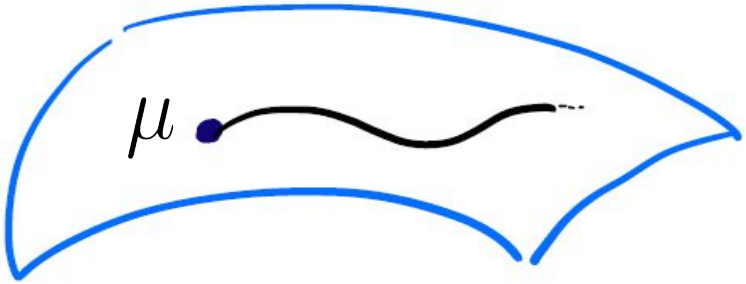


$\mathcal{P}(\mathbb{R}^d)$

$$\partial_t \rho + \operatorname{div}(\rho \vec{V}) = 0$$

Let’s consider a typical setting in applications: Euclidean case

$$\mathcal{X} = \mathbb{R}^d = \mathcal{Y} \qquad c(x, y) = |x - y|^2$$



$$\partial_t \rho + \operatorname{div}(\rho \vec{V}) = 0$$

Conclusions

-OT is highly interdisciplinary.

-Many potential applications.

-Many interesting theoretical questions.

-Interested? Enroll in Stat 992.

Research interests

- Nonlinear financial time series analysis
- Extreme value analytics for big data
- Risk analysis in finance, insurance, environmental studies, and seismic data
- Nonlinear/asymmetric causal inference
- Stochastic optimization for non-convex and non-smooth problems and simulation technique
- Hi-dimensional inference and machine learning
- Bayesian inference for time series
- Medical statistics
- Global monetary system
- Cancer research projects

Quotient correlation of coefficient

- **Ways to measure relative positions:**
  - the difference (quantitative change) $X - Y$;
  - the quotient ('qualitative' change) X/Y for positive variables X and Y.
    - $X/Y = 0$ 'means' no relation.
    - $X/Y = 1$ means they are identical.
- **In a 'Normal' world:** Pearson correlation coefficient can be written as the inner product of the Z scores,

$$r_n =< \frac{X - \bar{X}1_n}{||X - \bar{X}1_n||_2}, \frac{Y - \bar{Y}1_n}{||X - \bar{Y}1_n||_2} >, \quad r_n \xrightarrow{\mathrm{P}} \rho.$$

- Quotient correlation coefficients are based on the maxima of the quotients of the Fréchet scores, AoS 2008, EJS 2011, SS 2017.

$$\mathrm{q}_n = \frac{\max_{i\le n}\{Y_i/X_i - 1\} + \max_{i\le n}\{X_i/Y_i - 1\}}{\max_{i\le n}\{Y_i/X_i\} \times \max_{i\le n}\{X_i/Y_i\} - 1}, \ \mathrm{q}_n \xrightarrow{\mathrm{P}} \lambda(?). \qquad (1)$$

$$\lambda = \lim_{x\to\infty} P(X > x \mid Y > x).$$

A new look of a twice-told tale:

$$\text{var}(Y) = \text{var}(\text{E}(Y|X)) + \text{E}(\text{var}(Y|X)). \tag{2}$$

- $\text{Var}\big(E(Y|X)\big)$ measures the spread of the conditional mean (center) of Y given X
- $\text{Var}\big(E(Y|X)\big)/\text{Var}(Y)$ can certainly be interpreted as the explained variance of Y by X.

$$\text{SEV}(Y|X) = \frac{\text{var}(\text{E}(Y|X))}{\text{var}(Y)} = 1 - \frac{\text{E}(\text{var}(Y|X))}{\text{var}(Y)} = 1 - \frac{\text{E}[\{Y - \text{E}(Y|X)\}^2]}{\text{var}(Y)}.$$

Correlation ratios: Kendall and Stuart (1979), Doksum and Samarov (1995), Wang (2001), Zheng, Shi, and Zhang (JASA, 2012)

Classical VaR v.s. Dynamic MMVaR

- From a mark to market viewpoint, the risk of a portfolio can be that with a given confidence $1-\alpha$, the probability that the price of the portfolio drops first time below a given price level p either at the end of the first day, at the end of the second day, and so on till at the end of the 10th day.
- Such a scenario is equivalent to that the cumulative return of the portfolio drops first time below a given return level c either at the end of the first day, at the end of the second day, and so on till at the end of the 10th day. Then we can define our mark to market VaR (MMVaR, JoE 2018) by

$$MMVaR_{S_{10}^X}(\alpha) = -\sup\{c:\ P(\cup_{i=1}^{10}\{\sum_{j=1}^{l} X_j > c,\ l=1,\ldots,i-1, \sum_{j=1}^{i} X_j < c\}) \le \alpha\}, \quad (3)$$

or for any given holding period of k time unites by

$$MMVaR_{S_k^X}(\alpha) = -\sup\{c:\ P(\cup_{i=1}^{k}\{\sum_{j=1}^{l} X_j > c,\ l=1,\ldots,i-1, \sum_{j=1}^{i} X_j < c\}) \le \alpha\}. \quad (4)$$

Copula Structured Max-stable Processes, JoE 2016

$$Y_{td} = \max\left(W_{td}^{1/\beta_d}, \max\left[\mathbf{A}_{td} \cdot \mathbf{Z}_t\right]\right), \quad d = 1, \ldots, D, \quad -\infty < t < \infty, \tag{5}$$

where $\beta_d > 0,\ d = 1, \ldots, D$.
$\{\mathbf{W_t},\ -\infty < t < \infty\} = \{(W_{t1}, \cdots, W_{tD}),\ -\infty < t < \infty\}$ is a sequence of IID D-dimensional random vectors following logistic distribution (also called Gumbel-Hougaard copula with unit Fréchet marginals in the literatrue) defined as

$$G_{\log}(\mathbf{x}; \gamma) = \exp\left\{-\left(\sum_{d=1}^{D} x_d^{-\gamma}\right)^{1/\gamma}\right\}, \tag{6}$$

with $\mathbf{x} = (x_1, \ldots, x_D)$ and $\gamma \geq 1$.

A new idea of the mixture - new fashion, JRSSB 2018

- Let $\mathbf{U}_1 = (U_{11}, \ldots, U_{1d}) \sim \mathbf{C}_1$, $\mathbf{U}_2 = (U_{21}, \ldots, U_{2d}) \sim \mathbf{C}_2$, $X \sim Bernoulli(c)$ and $(\mathbf{U}_1, \mathbf{U}_2, X)$ are mutually independent,

$$\mathbf{U} = (U_1, \ldots, U_d) = \max(\mathbf{U}_1^{\frac{1}{X}}, \mathbf{U}_2^{\frac{1}{1-X}}) \sim \mathbf{C}_{mix}, \tag{7}$$

  where $\max(\cdot)$ is a pairwise max function, i.e. $U_i = \max(U_{i1}^{\frac{1}{X}}, U_{i2}^{\frac{1}{1-X}})$. Here we define $1/0{=}\infty$.

- Observation: For any random variables $X \in [0, 1]$, the pairwise max framework (1) generates a new copula

$$\mathbf{C}_{\mathbf{U}}(u_1, \ldots, u_d) = \mathbb{E}\left(\mathbf{C}_1(u_1^X, \ldots, u_d^X) \cdot \mathbf{C}_2(u_1^{1-X}, \ldots, u_d^{1-X})\right).$$

Unite mixture