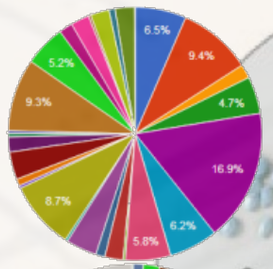
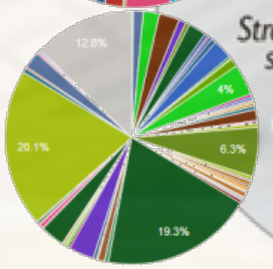
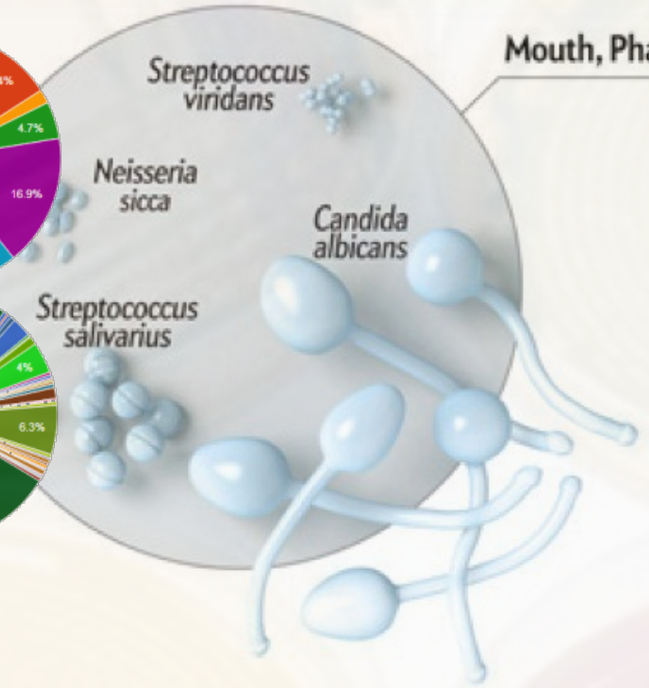
Feature Subset Selection for Inferring Relative Importance of Taxonomy

Gregory Ditzler and Gail Rosen
Drexel University
Dept. of ECE
Philadelphia, PA 19104
gregory.ditzler@gmail.com, gailr@ece.drexel.edu

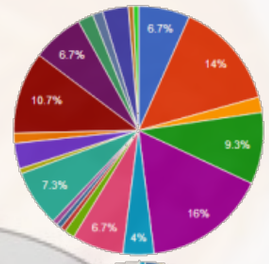
Healthy vs. Unhealthy

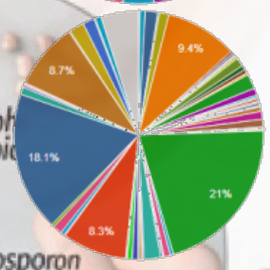




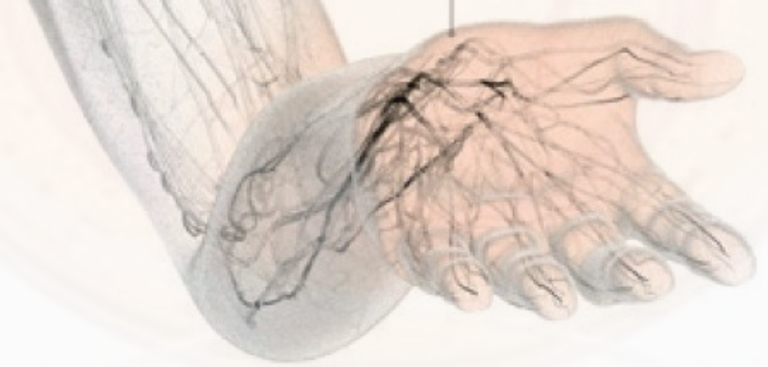


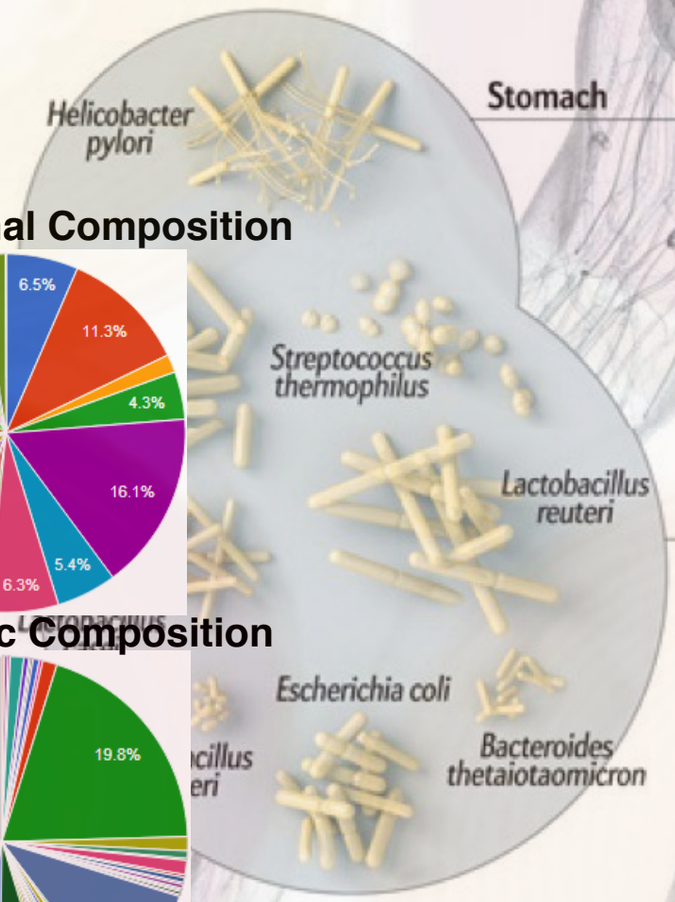
Mouth, Pharynx, Respiratory System





Skin



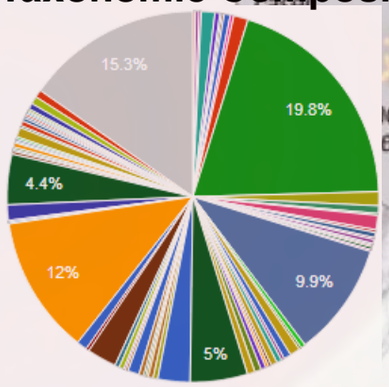


Stomach

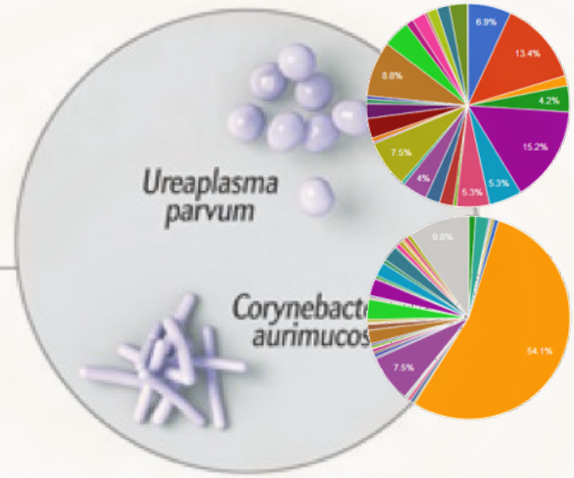
Intestines

Functional Composition

Taxonomic Composition



Urogenital tract



Microbes are Everywhere

(even the news)

CNN Health

Antibiotics may mess with your baby's metabolism

By **Jacque Wilson** and **Val Willmingham**, CNN
updated 2:36 PM EDT, Fri August 15, 2014

Junk food does more than make you fat, and other studies you missed

By **Jen Christensen**, CNN
updated 5:42 PM EDT, Fri August 29, 2014

shots - health news

Could Detectives Use Microbes To Solve Murders?

npr

How A Change In Gut Microbes Can Affect Weight

Staying Healthy May Mean Learning To Love Our Microbiomes

Diverse Gut Microbes, A Trim Waistline And Health Go Together

shots - health news

Can Probiotics Help Soothe Colicky Babies?

January 20, 2014 • As many as 15 percent of babies have colic, which can cause bouts of inconsolable crying

The New York Times

Some of My Best Friends Are Germs

Our Microbiome May Be Looking Out for Itself

By CARL ZIMMER

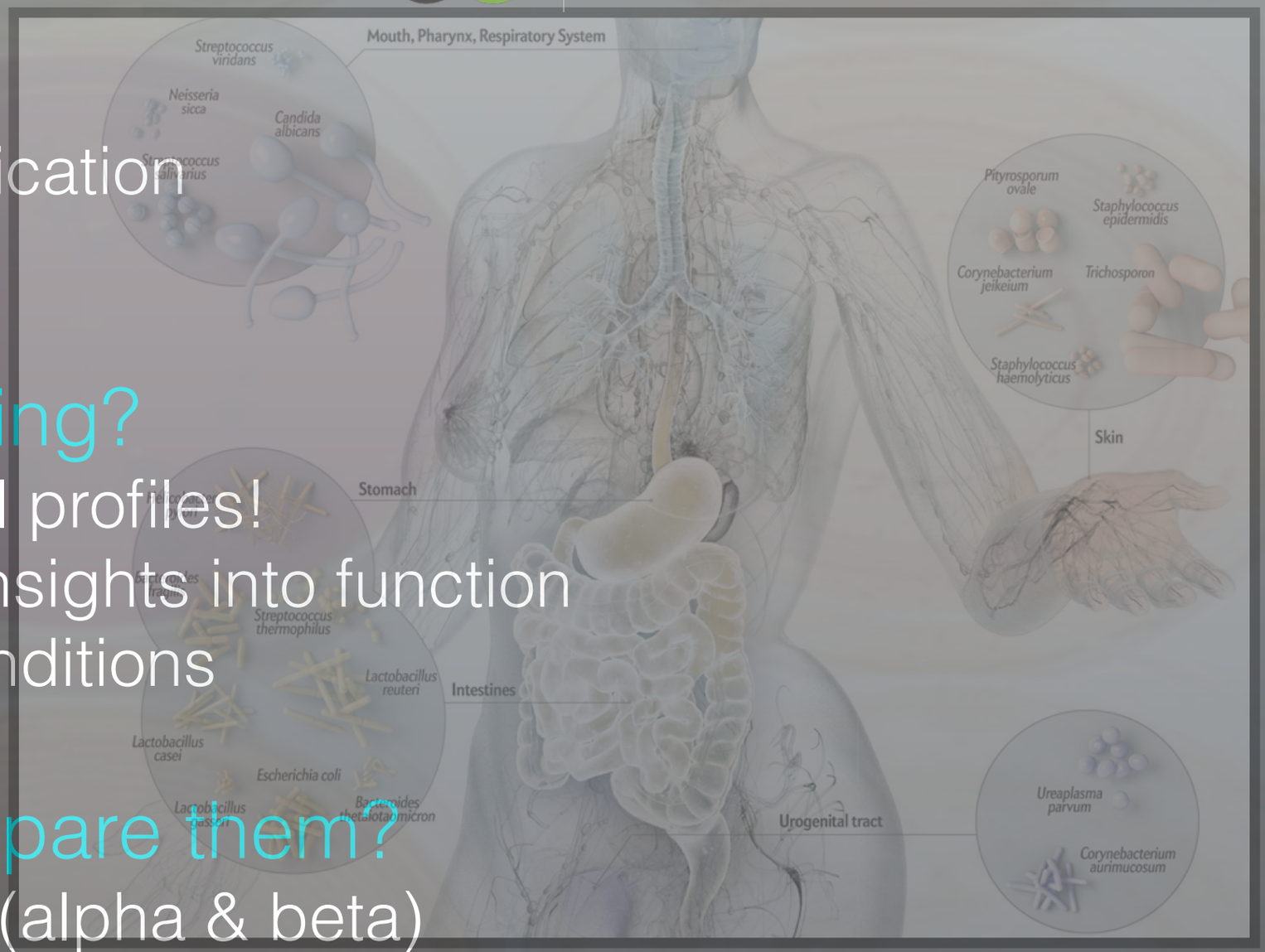
In Good Health? Thank Your 100 Trillion Bacteria

By GINA KOLATA
Published: June 13, 2012

- Everyday microbes
  - Oxygen/Carbon cycles
- Extreme microbes
  - Psychrophile (Antarctic)
  - Halophile (Dead Sea)
- Human Health
  - Lean/Obese
  - Inflammatory Bowel Disease

Who, Why, and How?

- Who is there?
  - Taxonomic classification
  - Novelty detection
- What are they doing?
  - Genes! Functional profiles!
  - Transcripts offer insights into function under specific conditions
- How can we compare them?
  - Diversity analysis (alpha & beta)
  - Machine learning & data mining

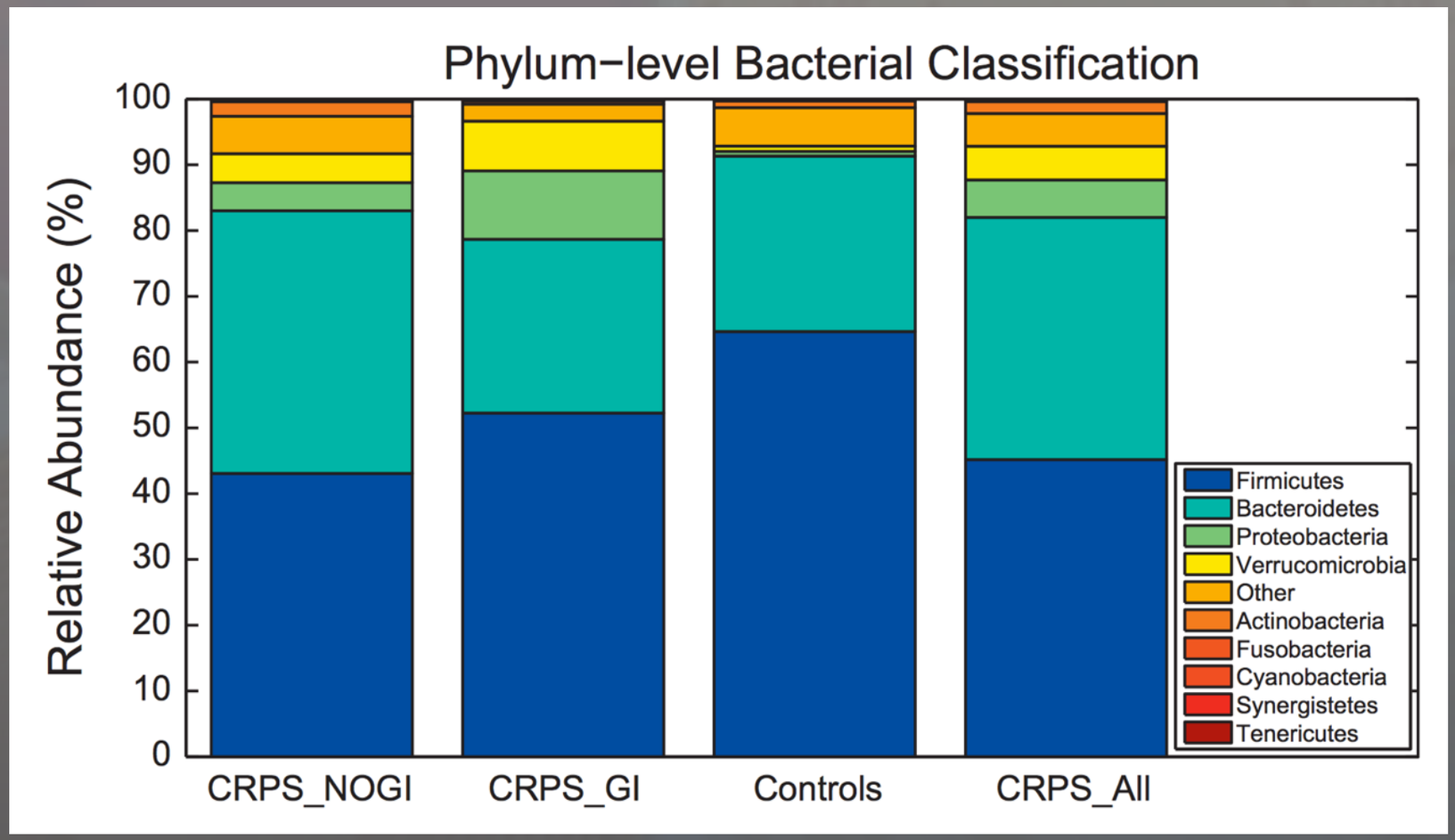


$$\hat{\mathbf{x}} = \mathbf{E}^{\mathsf{T}}(\mathbf{x} - \mathbf{m})$$

Complex Regional Pain Syndrome

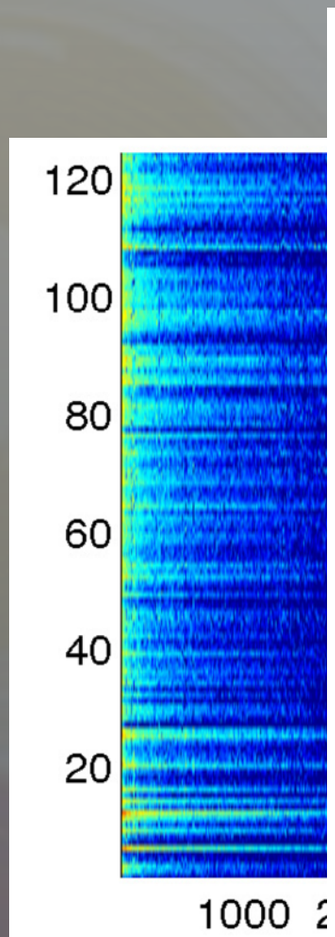
- Which bacteria can best represent the differences between patients with CRPS?
  - *Are the most abundant the most informative?*

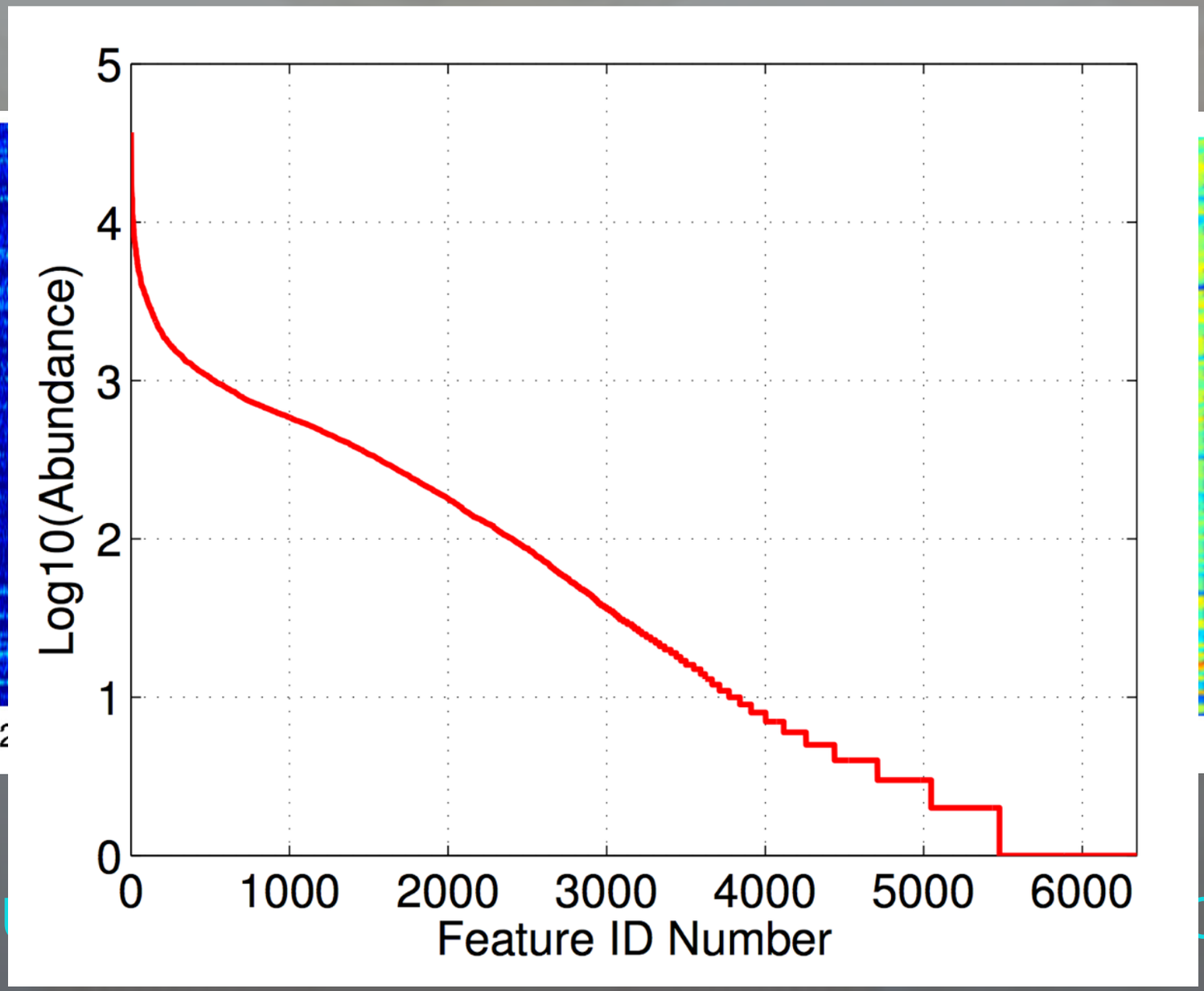
Relationships are complex and a framework to handle uncertainty is needed.

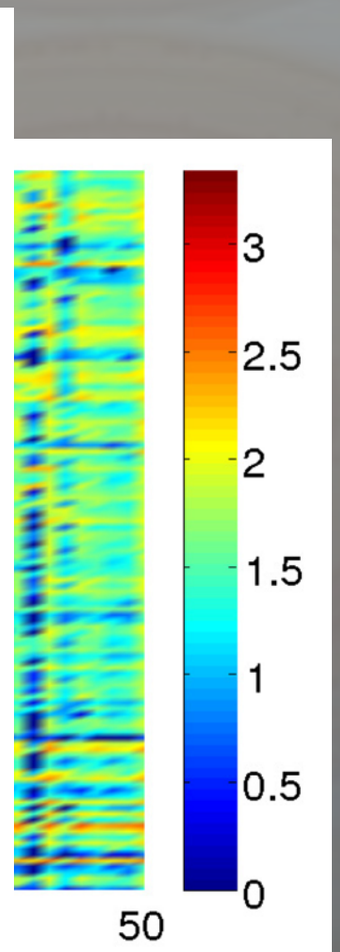


Reichenberger et al., "Establishing a relationship between bacteria in the human gut and Complex Regional Pain Syndrome," in *Brain, Behavior, and Immunity*, vol. 29, 2013, pp. 62—29.

A Study of IBD Patients



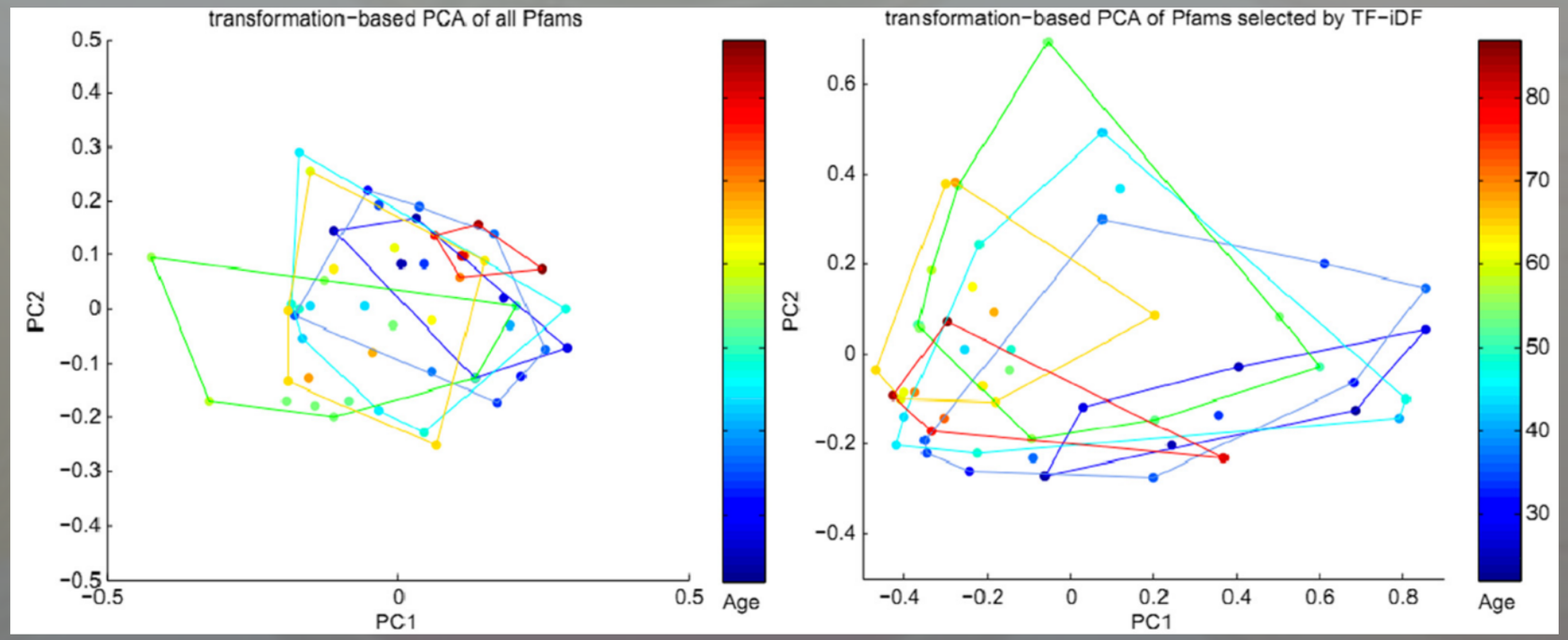




Feat ction

G. Ditzler, Y. Lan, J.-L. Bouchot, and G. Rosen, “Feature selection for metagenomic data analysis,” *Encyclopedia of Metagenomics*, 2014

Functional Importance & Age in the Microbiome



- Findings
  - Down regulation of B-12 biosynthesis family with aging
  - Down regulation of a broad range of reductases with aging (including protection from oxidative stress)

Y. Lan, A. Kriete, and G. Rosen. "Selecting age-related functional characteristics in the human gut microbiome," *BMC Microbiome*, Jan. 2013.

Why Feature Selection?

- Rapid inference about variable importance
  - Which OTUs/PFAMS/(etc) best differentiate multiple populations?
  - How can we mathematically define variable “importance”?
- Scalable and versatile for genomic data
  - Are there more variables than data (i.e., underdetermined system)?
  - Scaling & normalizations of abundance matrices
- Extensions for BigData
  - Recent thrusts in scaling feature selection for massive data (typically larger than the HMP, EMP and AG can provide)
  - Millions/Billions of features & observations from heterogenous data sources

Wrapper Approaches

- Wrapper feature selection approaches attempt to find a subset of features that minimize the loss of a classifier
  - Choose a subset of features, build/evaluate a classifier, and measure a loss
  - Adapt feature subset & repeat
- Typically classifiers have a small loss; however, they are prone to *overfitting* and *computationally burdensome*!
- Classifier dependent!
- Not of interest for our purposes!

Embedded Approaches

- Jointly optimize the parameters of a classifier and feature selector at the same time
  - Note the subtle difference between the embedded approach
  - Embedded approaches are typically of lower complexity than wrappers
- Examples: Lasso, Elastic-nets, …
  - Commonly performed with l_1 minimization problems
- Both embedded and wrappers tie themselves to a classifier
  - Added *complexity* for microbial ecologists?
  - Added *complexity* for general problems of simple knowledge discovery

Filter Approaches

- Filter methods decouple the feature subset optimization from the classifier optimization
  - Assign feature sets a measure of importance or value using a function that is not classifier loss
  - Examples: mutual information, correlation, any other set function that is not error
- Filters are known to be quite fast compare to wrappers and embedded methods
  - Filters cannot guarantee minimum loss (though neither can wrappers and embedded methods)
  - Not ideal for data where the feature set size dwarfs the feature subset size

Some Take Away Notes

- What are the assumptions!
  - Every algorithm makes assumptions, but what they are and how much they can be tolerated is up to the user
  - Kind of like: “Show me the constants!” in computational learning theory
- How big is my data?
  - Not all feature subset selection algorithms scale the same to the number of observations, or features
- Is classification the end goal?
  - Classifiers == Added Complexity
  - Even classifiers make assumptions!
- Your solution will be custom to your problem

What does this mean for microbial ecologists?

- The obvious: A mathematical framework to detect the relative importance of taxa, Pfams, etc.
- The subtle: Discovering and detecting the key factors (mathematically speaking) that differentiate multiple populations
  - There is always the possibly of an known unknown affecting the outcome of subset selection

LASSO & Elastic Nets

- Least Absolute Shrinkage and Selection Operator (Lasso)
  - Assumes a linear relationship between the input and output
  - Works for small sample size & large feature set

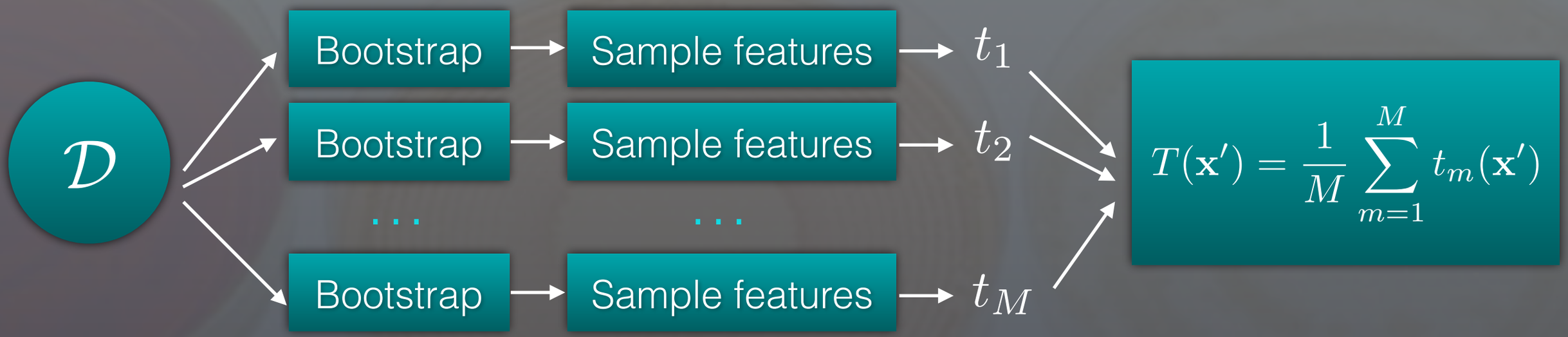
$$\theta^* = \arg\min_{\theta \in \Theta} \|\mathbf{y} - \mathbf{X}^\mathsf{T}\theta\|_2^2 + \lambda_1 \|\theta\|_1$$

- Elastic Nets
  - Gets around Lasso not working when the sample size is larger than the feature set size

$$\theta^* = \arg\min_{\theta \in \Theta} \|\mathbf{y} - \mathbf{X}^\mathsf{T}\theta\|_2^2 + \lambda_1 \|\theta\|_1 + \lambda_2 \|\theta\|_2^2$$

Random Forests

- *Simple and straightforward* bagging-like approach that generates an ensemble of decision tree for prediction
  - Capable of estimating variable importance, or the decrease in accuracy if the variable is omitted
    - permute a feature and compute the OOB error
  - Effective for large datasets and robust to overfitting
- Widely used as the tool for supervised classification with tools such as QIIME



Information Theory

A Mathematical Theory of Communication

By C. E. SHANNON

- *Information theory* provides us a convenient mathematical framework for capturing uncertainty and information in random variables.
- Mutual information provides a key quantity of measuring variable importance

$$I(X;Y) = \int_{y\in\mathcal{Y}} \int_{x\in\mathcal{X}} p_{X,Y}(x,y) \log \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} \mathrm{d}x\mathrm{d}y$$

- Designing a general objective function (Brown, 2012)

$$\mathcal{J}(X_k) = I(X_k;Y) - \alpha \sum_{j\in\mathcal{F}} I(X_k;X_j) + \beta \sum_{j\in\mathcal{F}} I(X_k;X_j|Y)$$

relevancy redundancy

Greedy Algorithms

Input: Collection of features $\mathcal{X} := \{X_i : i \in [K]\}$, scoring function $\mathcal{J}$, and phenotype variables Y.
Initialize: $\mathcal{F} = \varnothing$
while $|\mathcal{F}| < k$ **do**

- Compute next best feature

$$X^* = \arg\max_{X' \in \mathcal{X}} \mathcal{J}(X', Y, \mathcal{F}) \tag{2}$$

MIM, mRMR, JMI, …

- $\mathcal{F} \leftarrow \mathcal{F} \cup X^*$
- $\mathcal{X} \leftarrow \mathcal{X} \backslash X^*$

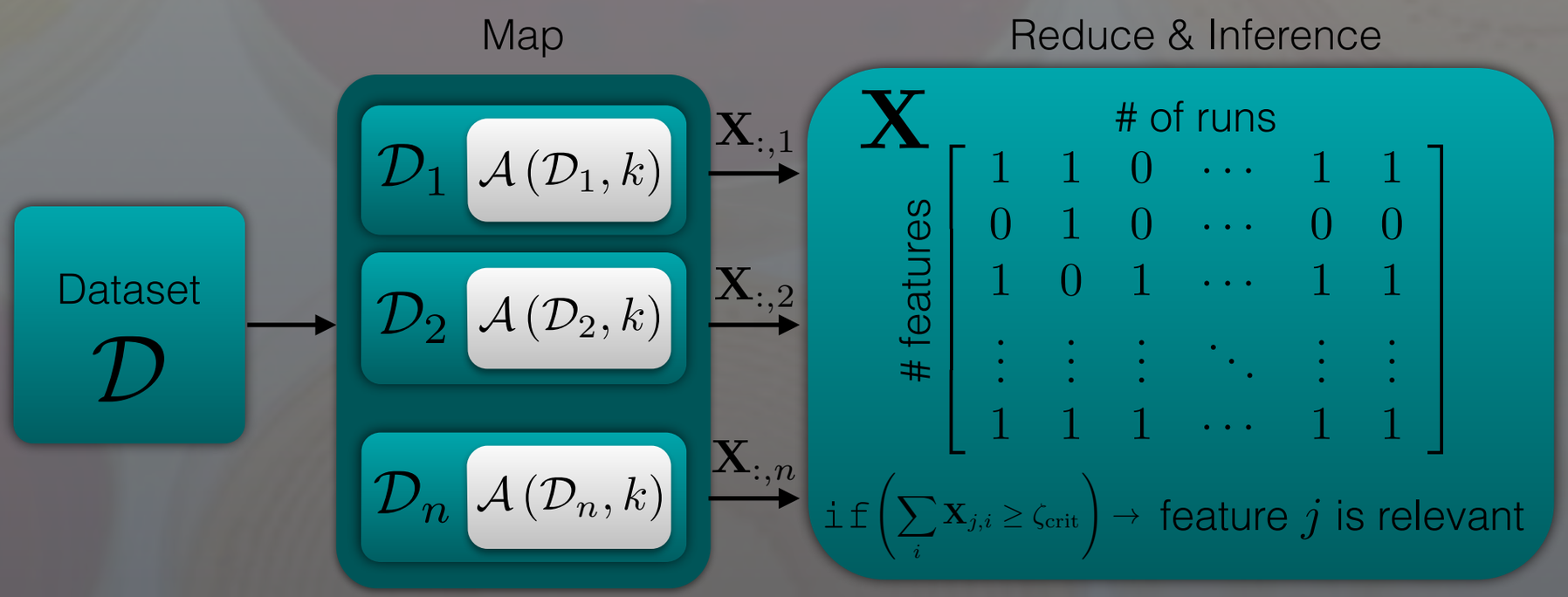
end while

Neyman-Pearson Feature Selection

- In most* situations we do not know in advance how many variables will be important
  - Ex., How many variables from a medical test are indicative of a response?
  - What if your software implementation only provides decisions of importance?
- Datasets with a large set of observations can be computationally burdensome to process all of the data at once
- Neyman-Pearson feature selection was designed to detect variable importance for a base-subset selection algorithm (i.e., MIM, or mRMR)

The NPFS Approach

- The *Neyman-Feature Feature Selection* (NPFS) approach detects feature importance from a filter's feature ranking… given no more an initial guess at how many features are important
- NPFS has some nice theoretical guarantees and has been shown to be quite effective in practice.
- We have implemented NPFS for biological data formats



G. Ditzler, R. Polikar, and G. Rosen, "A bootstrap based Neyman-Pearson test for identifying variable importance," *IEEE Transactions on Neural Networks and Learning Systems*, 2014, In Press.

About the Data

- American Gut Project
  - We isolate 469 samples from 231 females, and 238 males. Approximately 26k OTUs
  - OTUs are detected using Greengenes

- Caporaso et al. Illumina Time-Series
  - A total of 467 samples are collected from one male and one female. Approximately 17k OTUs

- Observational Study
  - How to the gut microbes of male and females differ?
  - We can use existing studies to verify any inferences made from our information-theoretic perspective

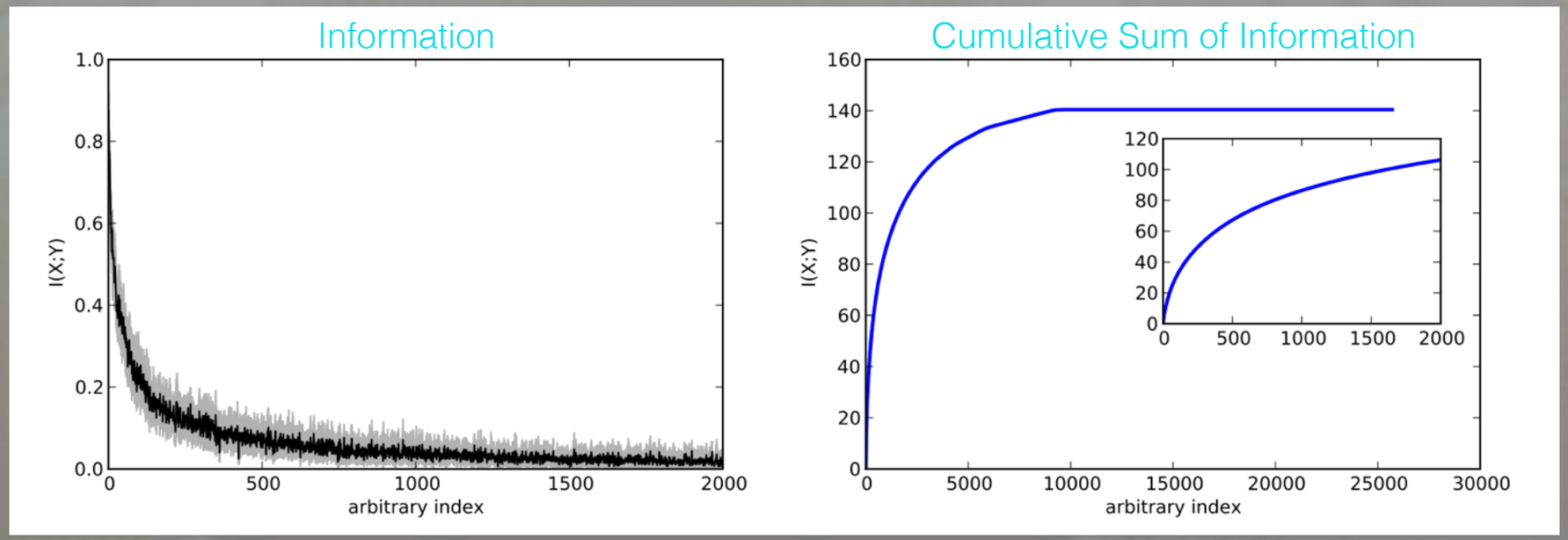
http://www.earthmicrobiome.org/
https://github.com/biocore/American-Gut

Methods

- Fizzy: *Information-Theoretic Subset Selection for Biological Data Formats*
  - Mutual Information Maximization
- NPFS: *Neyman-Pearson Feature Selection*
  - Automatically detects feature importance given an objective function. We use mutual information maximization
- Lasso: *Least Squares with l_1 regularization*
- Elastic-nets: *Least Squares with l_1 and l_2 regularization (not of much relevance, or shown)*
- Random Forests: *Ensemble of decision trees*

https://github.com/EESI/Fizzy
http://scikit-learn.org/stable/ http://qiime.org/

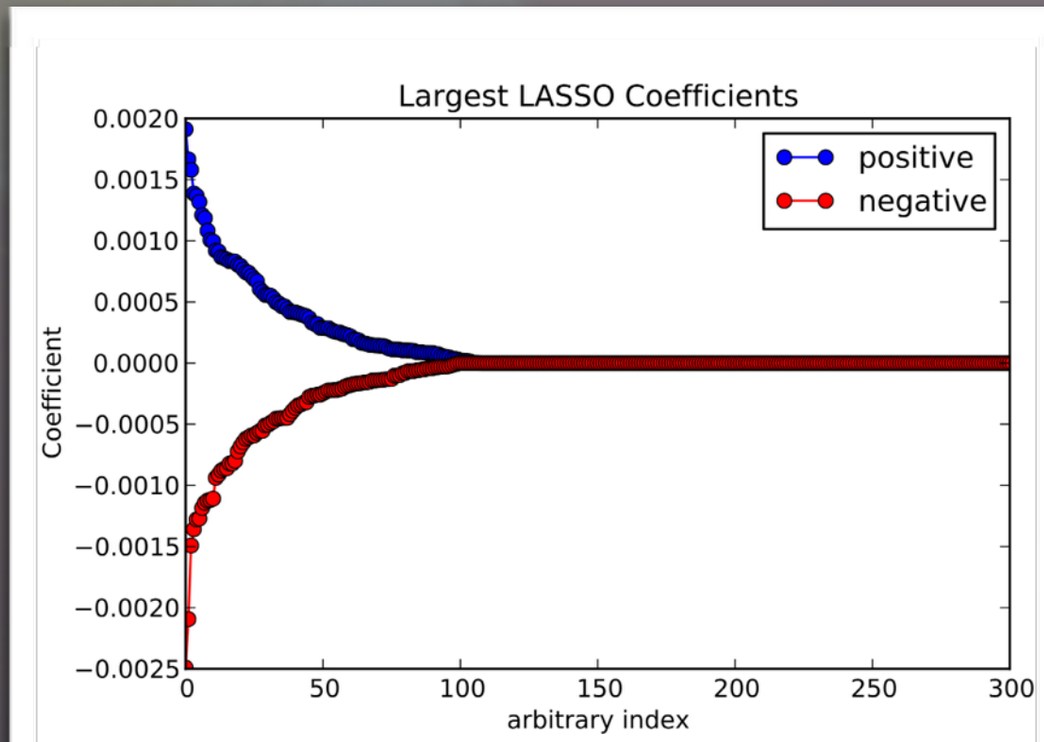
Information in Gender

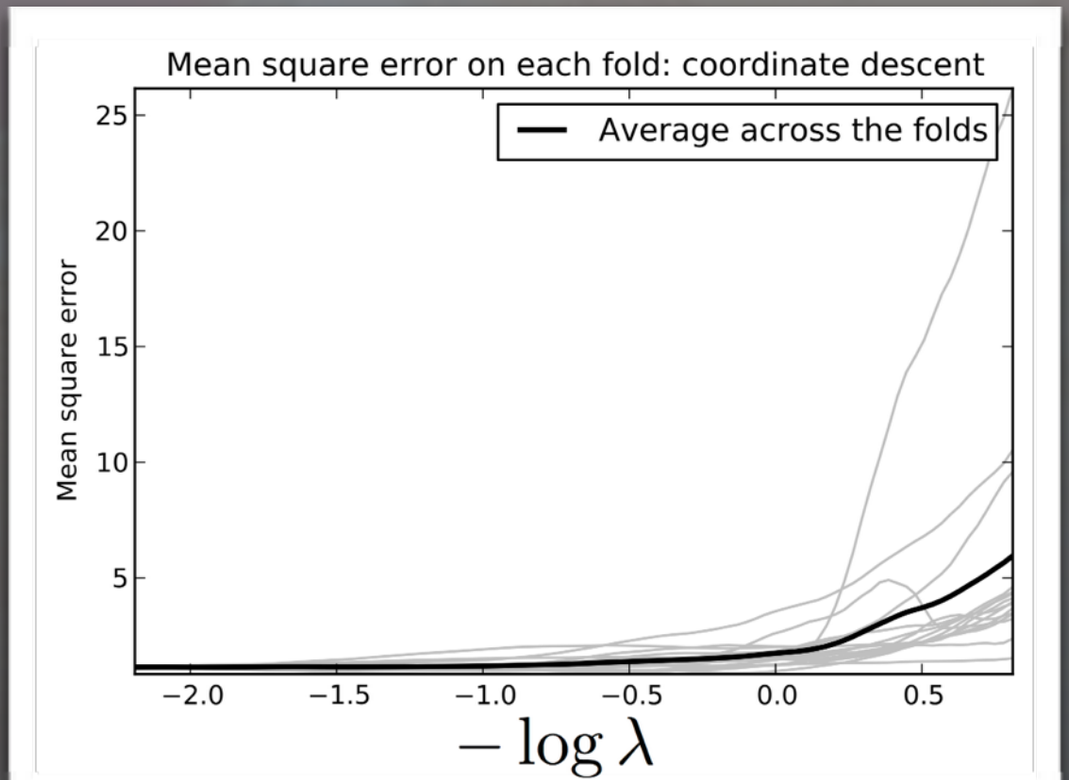


- MI is computed over bootstrap samples from the population
- Most of the *information* about Sex and the gut microbes are summarized by ~250 OTUs
- Bulk of the features are meaningless for explaining these differences

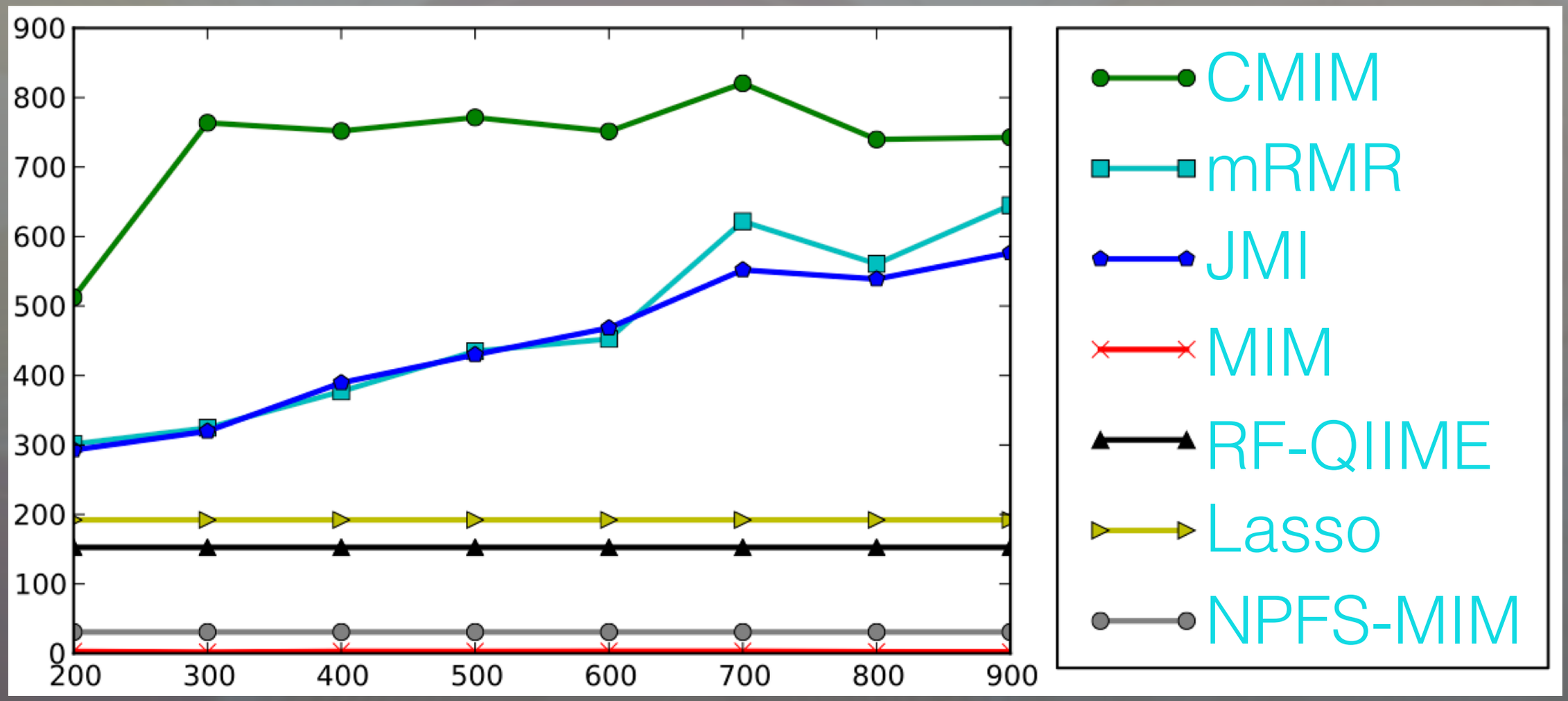
Lasso Feature Weights

- The weights from Lasso *confirm* what was discovered with mutual information
  - Relatively few OTUs appear to be responsible for the differences in gender





Timing


Runtime (sec)
900
800
700
600
500
400
300
200
100
0
200
300
400
500
600
700
800
900
# of Feature Selected
CMIM
mRMR
JMI
MIM
RF-QIIME
Lasso
NPFS-MIM

IBD & Obesity with PFAMS

	IBD features	**Obese features**
feature 1	ABC transporter (PF00005)	ABC transporter (PF00005)
feature 2	Phage integrase family (PF00589)	MatE (PF01554)
feature 3	Glycosyl transferase family 2 (PF00535)	TonB dependent receptor (PF00593)
feature 4	Acetyltransferase (GNAT) family (PF00583)	Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase (PF02518)
feature 5	Helix-turn-helix (PF01381)	Response regulator receiver domain (PF00072)

- ABC transporter is known to mediate fatty acid transport that is associated with obesity and insulin resistant states
- ATPases that catalyze dephosphorylation reactions to release energy
- Glycosyl transferase is hypothesized to result in recruitment of bacteria to the gut mucosa and increased inflammation
- More results can be found in Ditzler et al. (2014)

G. Ditzler, Y. Lan, J.-L. Bouchot, and G. Rosen, “Feature selection for metagenomic data analysis,” *Encyclopedia of Metagenomics*, 2014

Conclusions

- At least in terms of gender, there are *not many OTUs that carry a significant amount of information*
  - Current results with NPFS and MIM *go along with our intuition* about the microbiome
  - Filter methods provide results very quickly compared to some of the embedded approaches
- OTU importance results with filters are further reinforced using Lasso
  - Lasso is capable of capturing some of the inter-OTU dependencies that MIM cannot
- Subset selection offers microbial ecologists an alternative to beta diversity

Future Work

- How much information is contained in 16S and metagenomic abundance matrices?
  - From a mathematical perspective?
    - > best/worst case bounds?
  - Empirical?
- Bandits & the bag of little bootstraps for subset selection on a massive scale!
- Viewing computational metagenomics as a stream (i.e., online learning)

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. CAREER #0845827, NSF #1120622, and DOE #DE-SC0004335.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Department of Energy.

Collaborators

Gail Rosen

(Drexel)

Robi Polikar

(Rowan)

Steve Essinger

(Pandora)

Erin Reichenberger

(Drexel)

Yemin Lan

(Drexel)

Calvin Morrison

(Temple)

Steve Pastor

(Drexel)

Steve Woloszynek

(Drexel)

https://github.com/EESI

Thank You!