



S. C. Big Data Health Sciences Conference

Object Oriented Data Analysis

J. S. Marron

Dept. of Statistics and Operations Research,
University of North Carolina

February 26, 2020



How does Statistics relate to Data Science?

From My Elementary Courses:

- ❑ Gaining Insight from Numbers

Similar to “Data Science” Definitions

- ❑ The Science of Managing Uncertainty

Where Probability Modeling Is Vital

This is Why Statistics is Fundamental



More Currently Popular Terminology

Big Data

- Isn't It Just Statistics?
- Yes, But More Needed Too
- Maybe Bigger Challenge:

Optimization:
Machine Learning

Complex Data



An Idea Worth Spreading?

UNC, Stat & OR

Well Understood Concept:

Great science now done by teams
with complementary skill sets

- Biology
- Chemistry
- Engineering
- Quantitative Work

⋮

Common Current Idea:
1 Team member



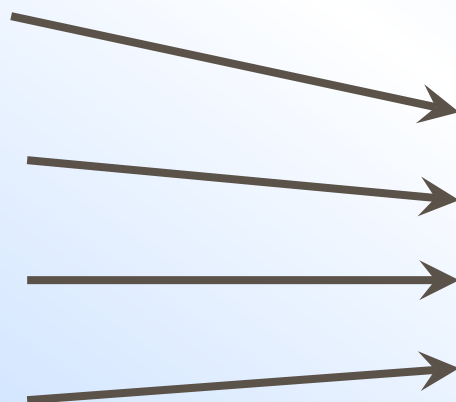
An Idea Worth Spreading?

UNC, Stat & OR

Extension of this:

Great *Quantitative Work* needs teams
with complementary skill sets

- Statistics
- Imaging
- Optimization
- Data Base
- ...



Big
Data



An Idea Worth Spreading?

Proposed New Approach:

Team Data Science

- ❖ Teams with Complimentary Skillsets
- ❖ Education of Team Members:
 - ❖ Bring Valued (Deep) Skill
 - ❖ Know Enough to Communicate
 - ❖ Give Opportunities to Practice



Object Oriented Data Analysis

What is the “atom” of a statistical analysis?

- 1st Course: Numbers
- Multivariate Analysis Course : Vectors
- Functional Data Analysis: Curves
- More generally: **Data Objects**



Object Oriented Data Analysis

Original Thought:

OODA = Mathematical Framework

(containing wide variety
of interesting cases)



Object Oriented Data Analysis

Original Thought:

OODA = Mathematical Framework

Current View:

OODA = Focal Point

{For discussions (interdisciplinary)
about tackling serious analyses}



Object Oriented Data Analysis

Original Thought:

OODA = Mathematical Framework

Current View:

OODA = Focal Point

What should be the Data Objects?



Principal Component Analysis

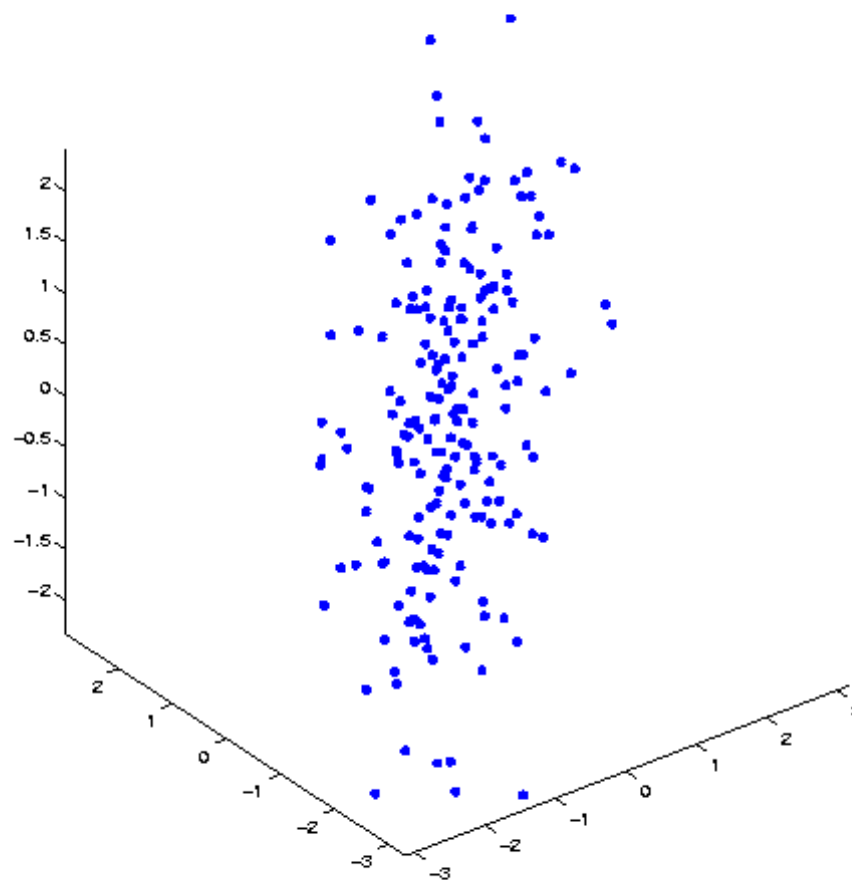
More Than *Dimensionality Reduction*:

- Visualization
 - Relationships Between Objects (Scores)



Demo PCA in $\mathbb{R}^3$

Raw Toy Data Set in $\mathbb{R}^3$

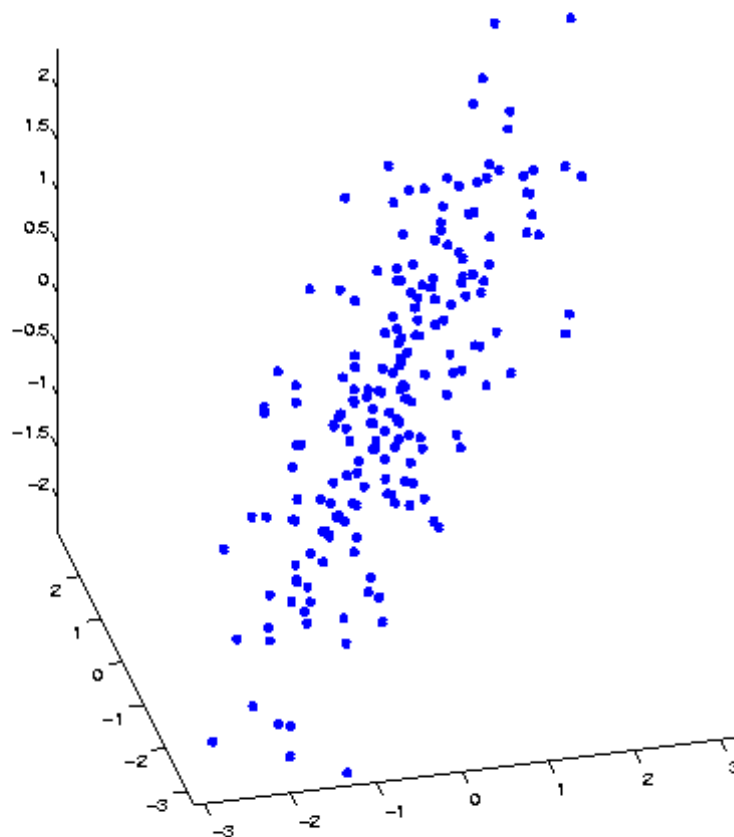




UNC, Stat & OR

Demo PCA in $\mathbb{R}^3$

Raw Toy Data Set in $\mathbb{R}^3$

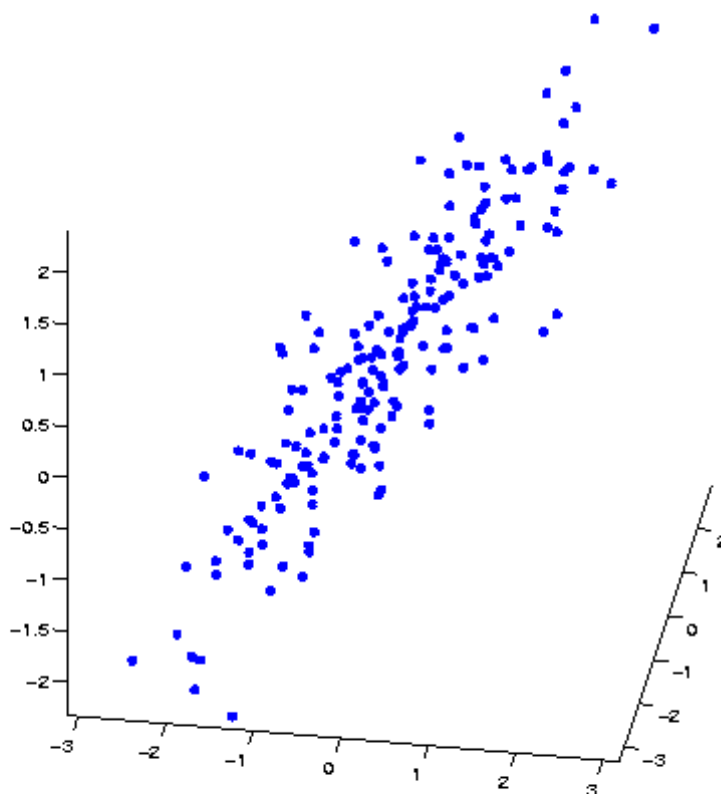




UNC, Stat & OR

Demo PCA in $\mathbb{R}^3$

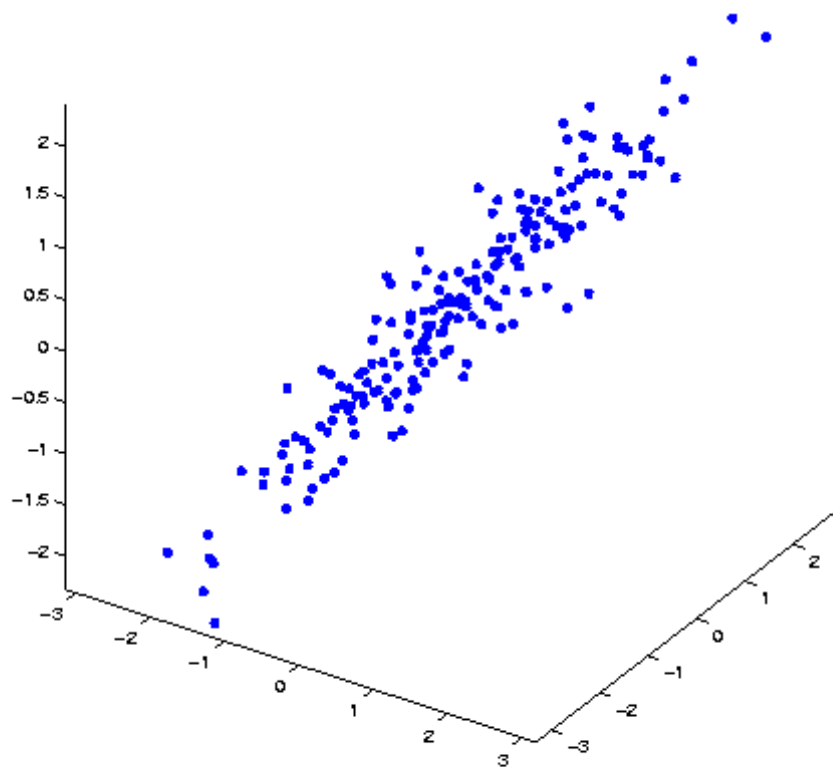
Raw Toy Data Set in $\mathbb{R}^3$





Demo PCA in $\mathbb{R}^3$

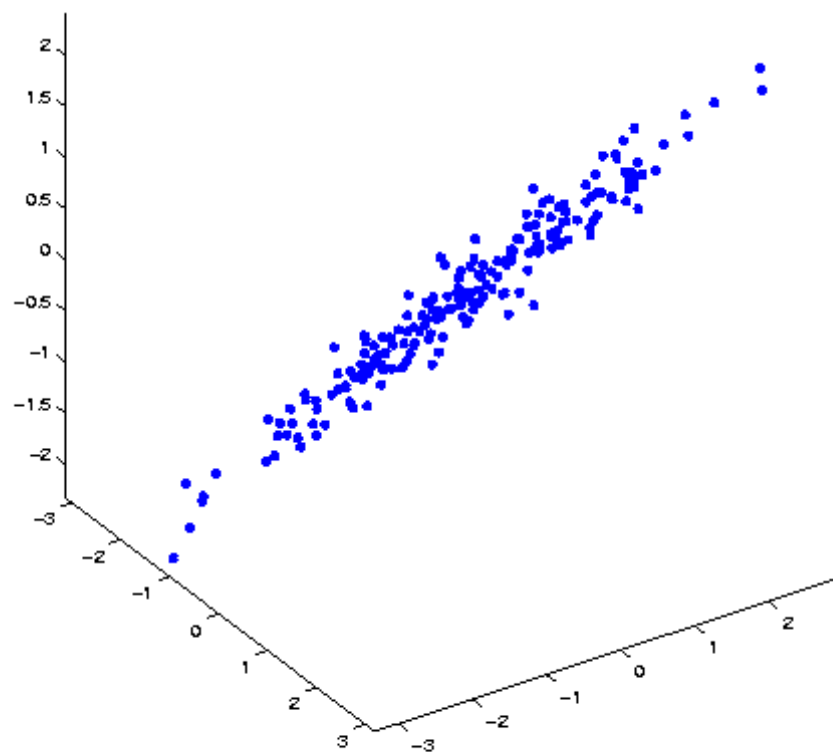
Raw Toy Data Set in $\mathbb{R}^3$





Demo PCA in $\mathbb{R}^3$

Raw Toy Data Set in $\mathbb{R}^3$

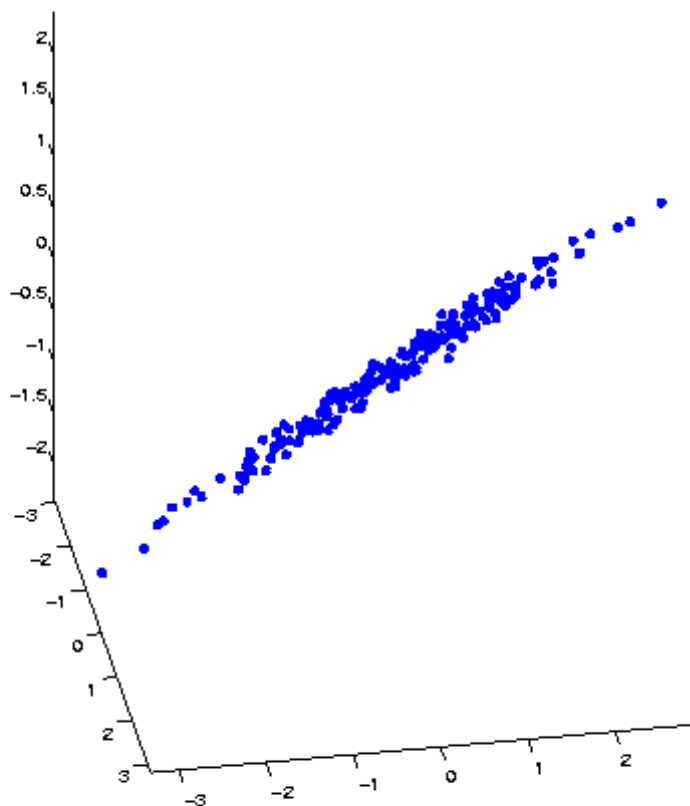




Demo PCA in $\mathbb{R}^3$

UNC, Stat & OR

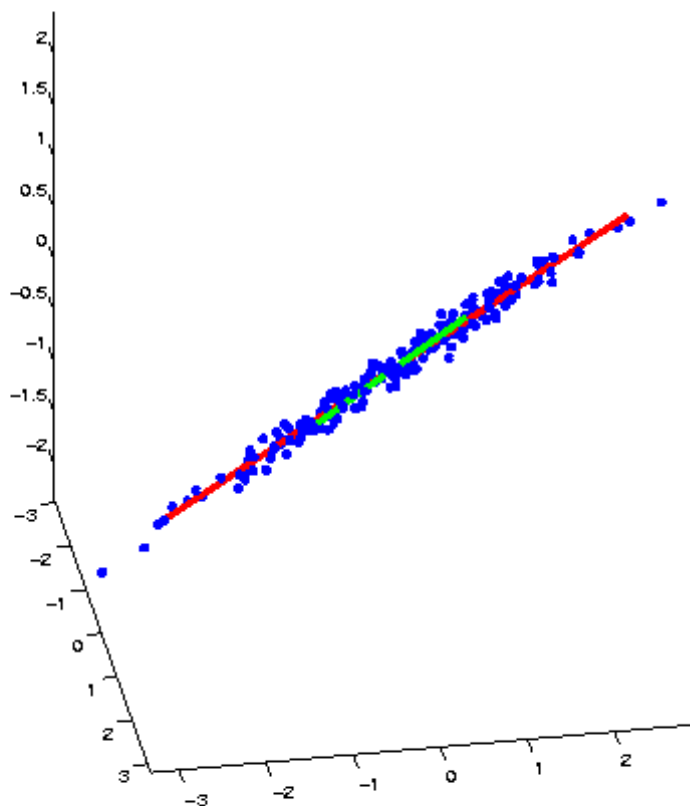
Raw Toy Data Set in $\mathbb{R}^3$





Demo PCA in $\mathbb{R}^3$

PCA Eigenvectors Give Useful Coordinates

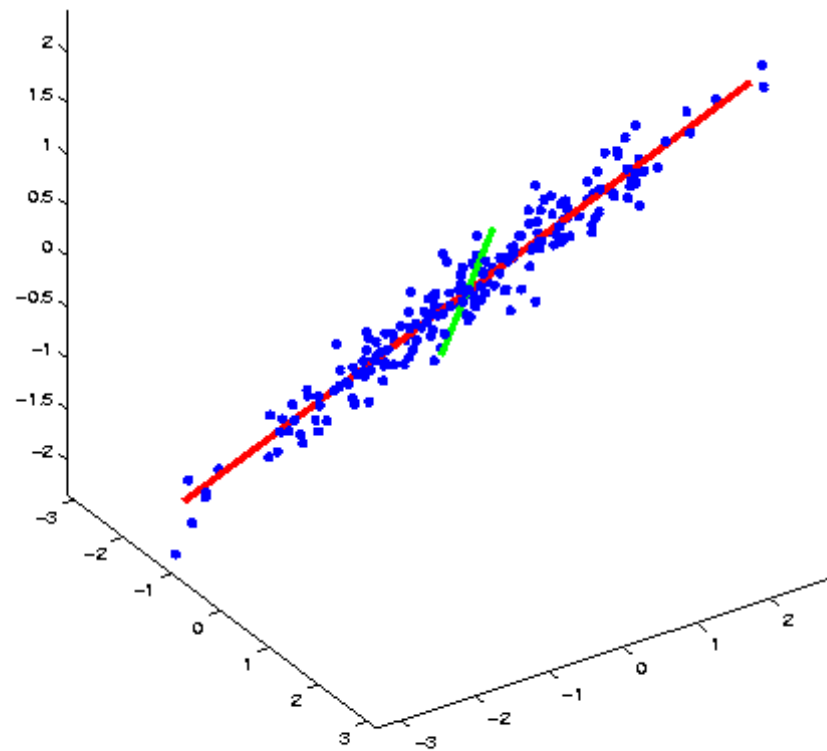




Demo PCA in $\mathbb{R}^3$

UNC, Stat & OR

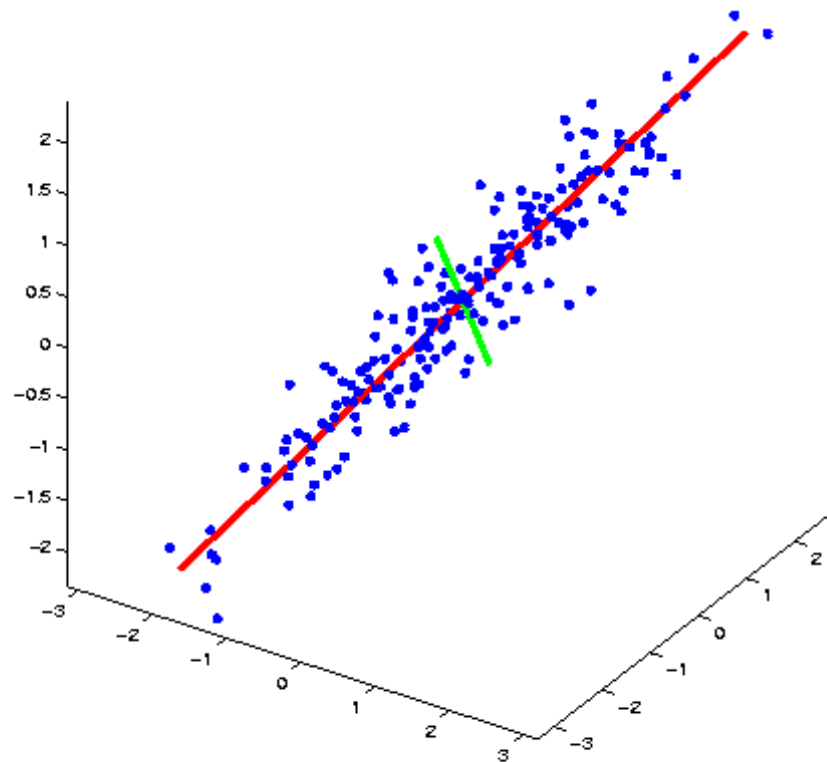
PCA Eigenvectors Give Useful Coordinates





Demo PCA in $\mathbb{R}^3$

PCA Eigenvectors Give Useful Coordinates

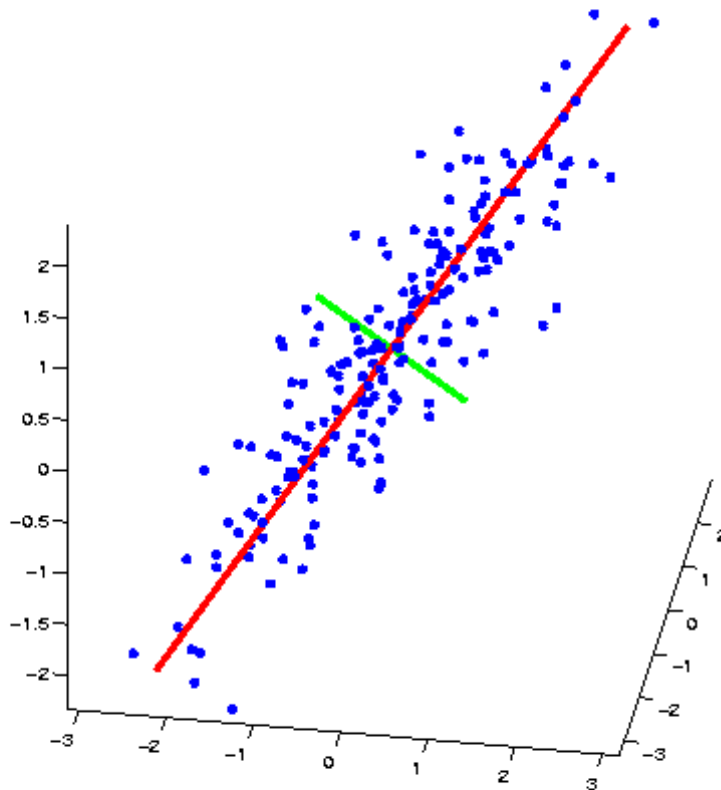




Demo PCA in $\mathbb{R}^3$

UNC, Stat & OR

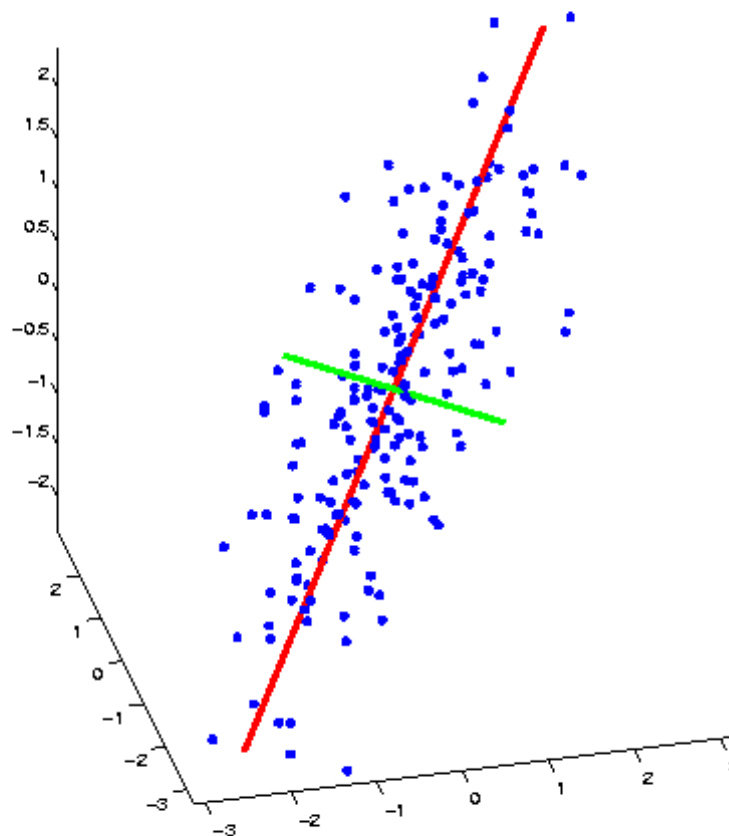
PCA Eigenvectors Give Useful Coordinates





Demo PCA in $\mathbb{R}^3$

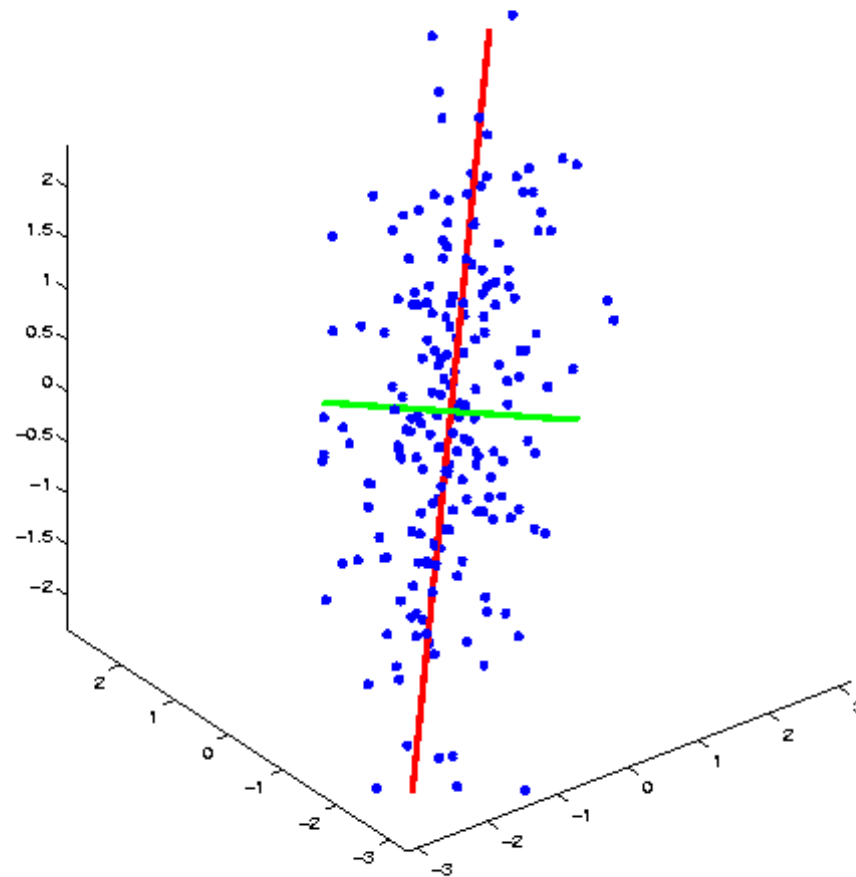
PCA Eigenvectors Give Useful Coordinates





Demo PCA in $\mathbb{R}^3$

PCA Eigenvectors Give Useful Coordinates





Functional Data Analysis

Interesting Real Data Example

- Genetics (Cancer Research)
- RNAseq (Next Generation Sequencing)
- Deep look at “gene components”

Microarrays: Single number (per gene)

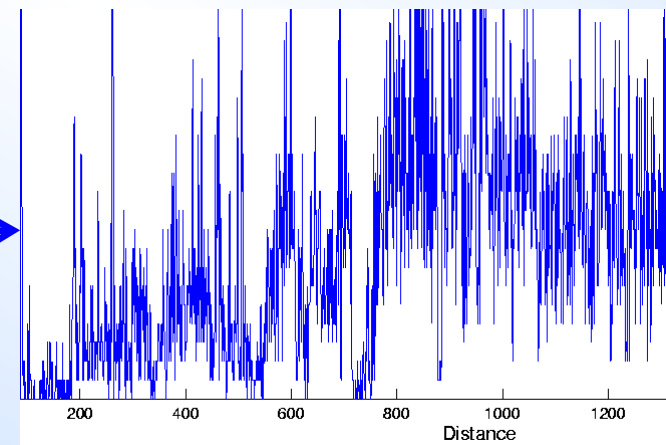
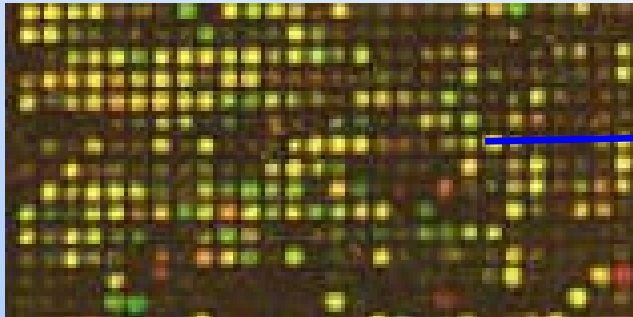
RNAseq: Thousands of measurements



Functional Data Analysis

Interesting Real Data Example

- Genetics (Cancer Research)
- RNAseq (Next Generation Sequencing)
- Deep look at “gene components”





Functional Data Analysis

Interesting Real Data Example

- Genetics (Cancer Research)
- RNAseq (Next Generation Sequencing)
- Deep look at “gene components”
- Gene studied here: CDKN2A
- Goal: *Study Alternate Splicing*
- Sample Size, $n = 180$
- Dimension, $d = \sim 1700$

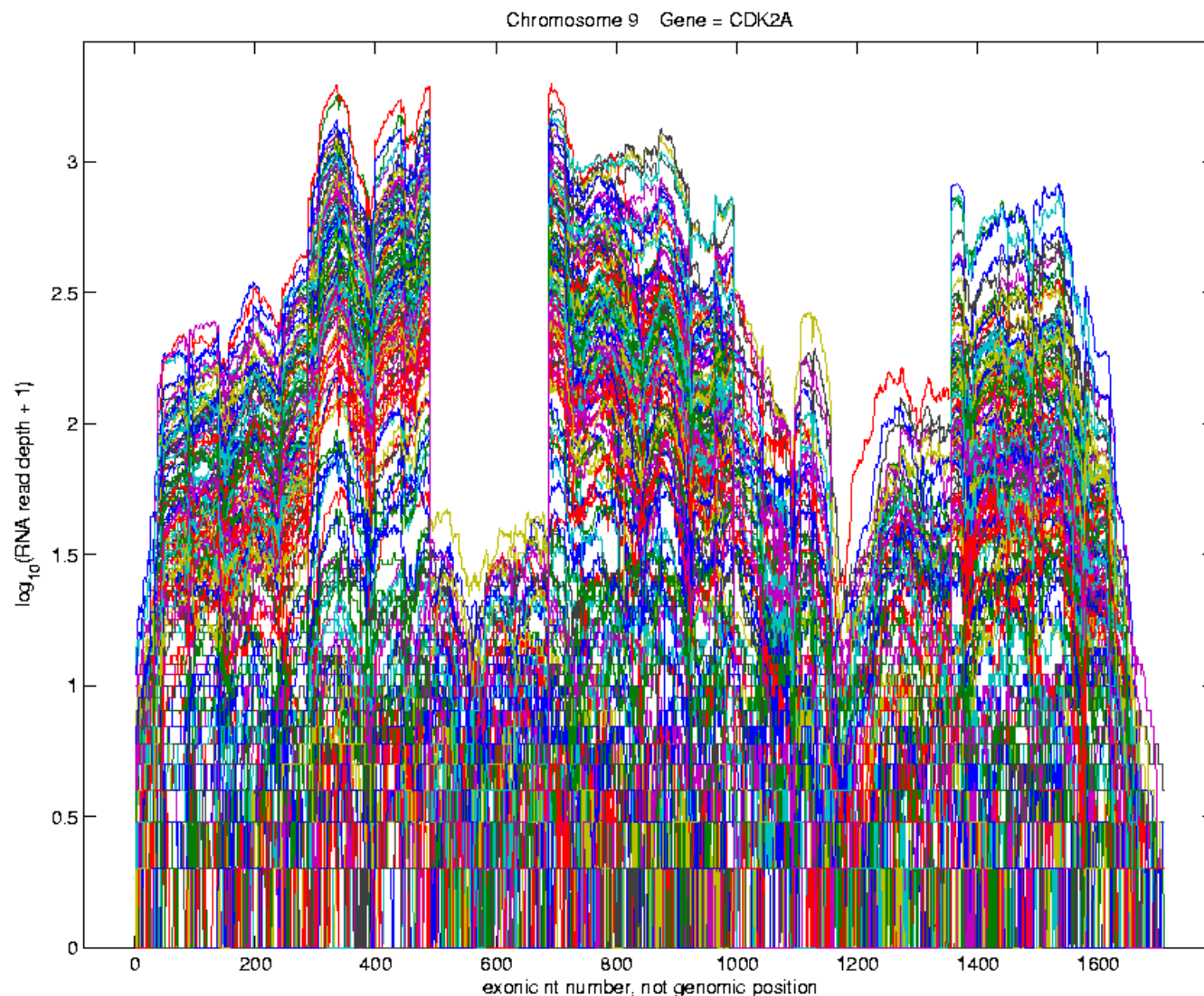


Functional Data Analysis

UNC, Stat & OR

Simple
1st
View:
Curve
Overlay
(log
scale)

Thanks to
Matt Wilkerson

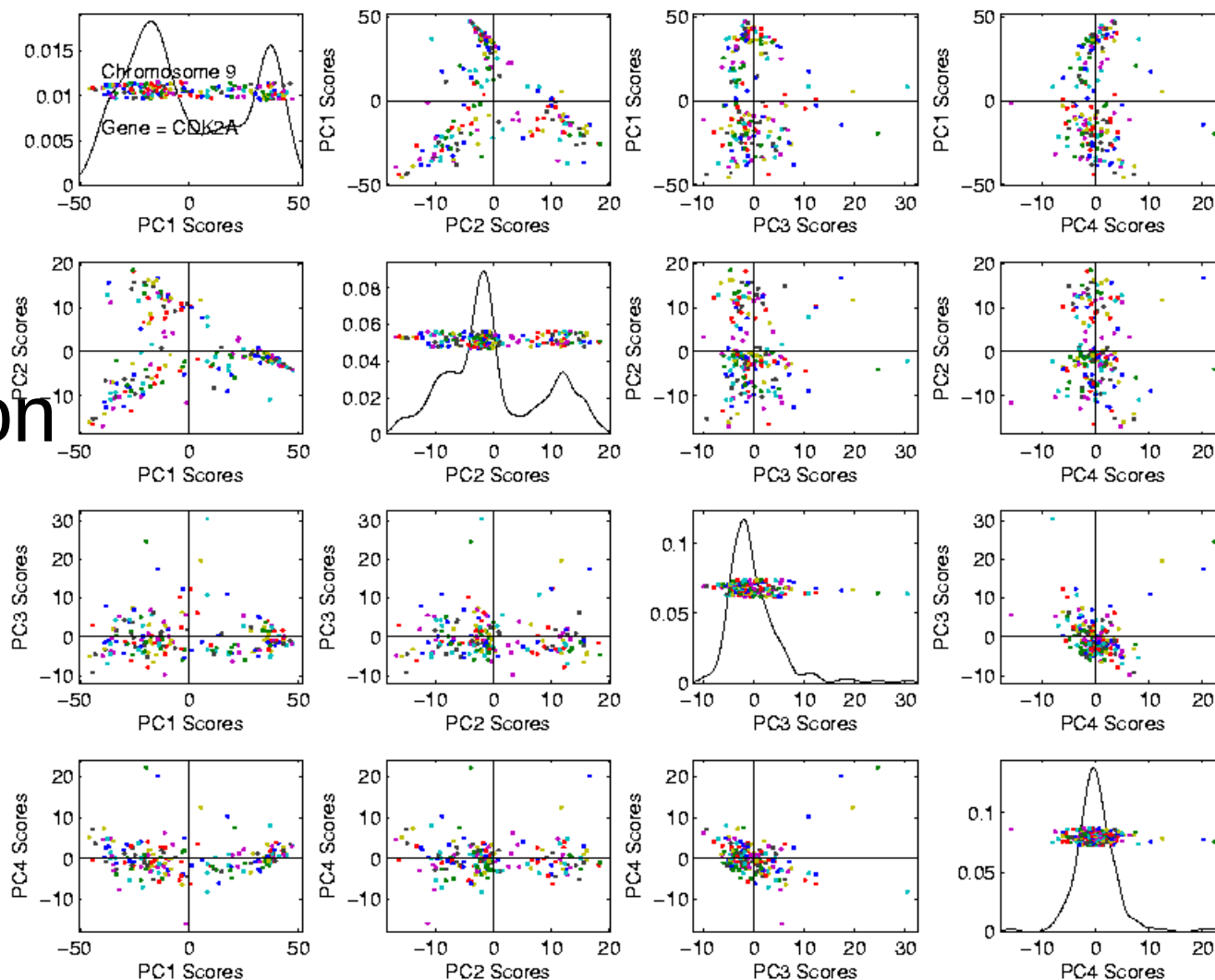




Functional Data Analysis

UNC, Stat & OR

Often
Useful
Population
View:
PCA
Scores

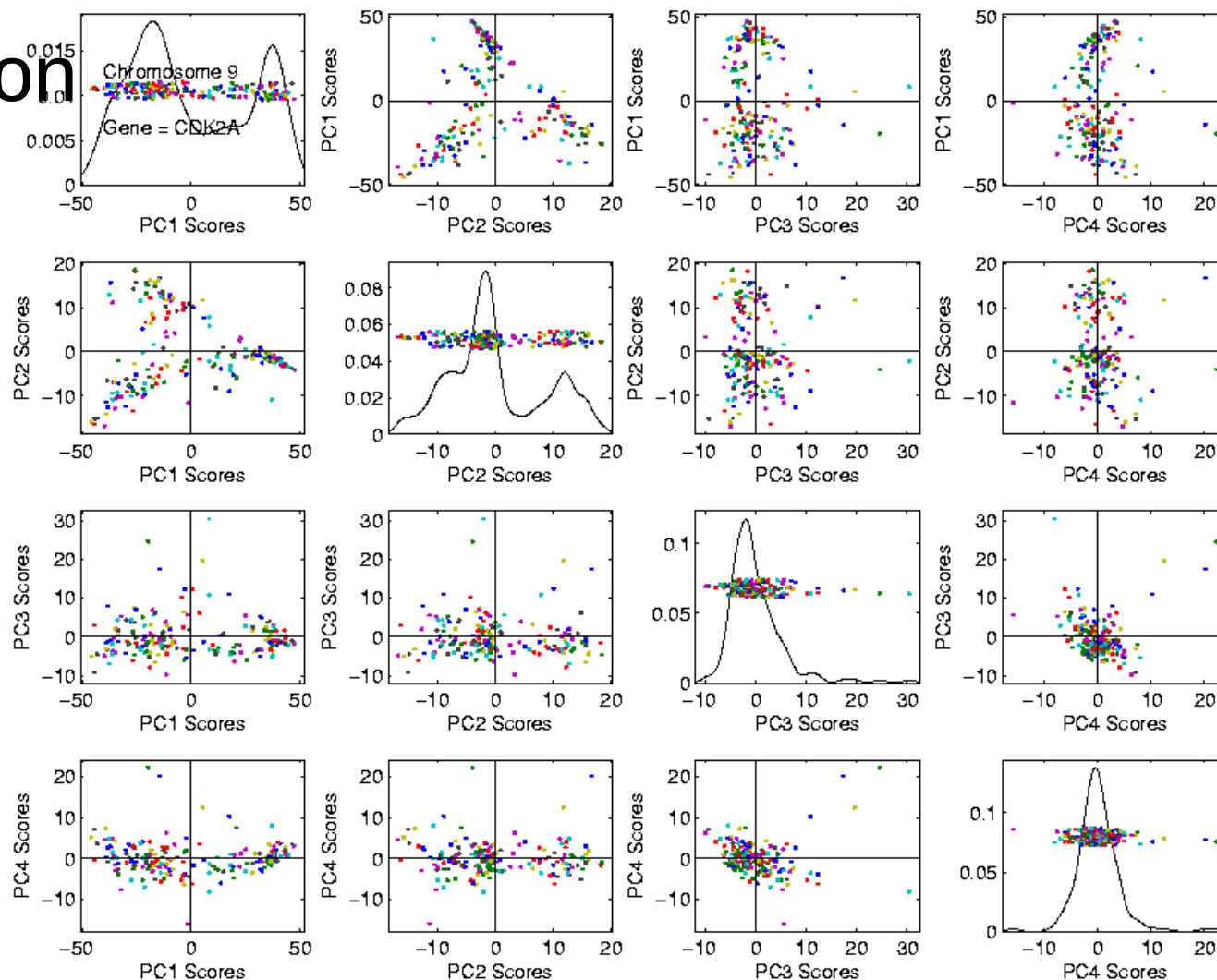




Functional Data Analysis

UNC, Stat & OR

Suggestion
Of
Clusters
???



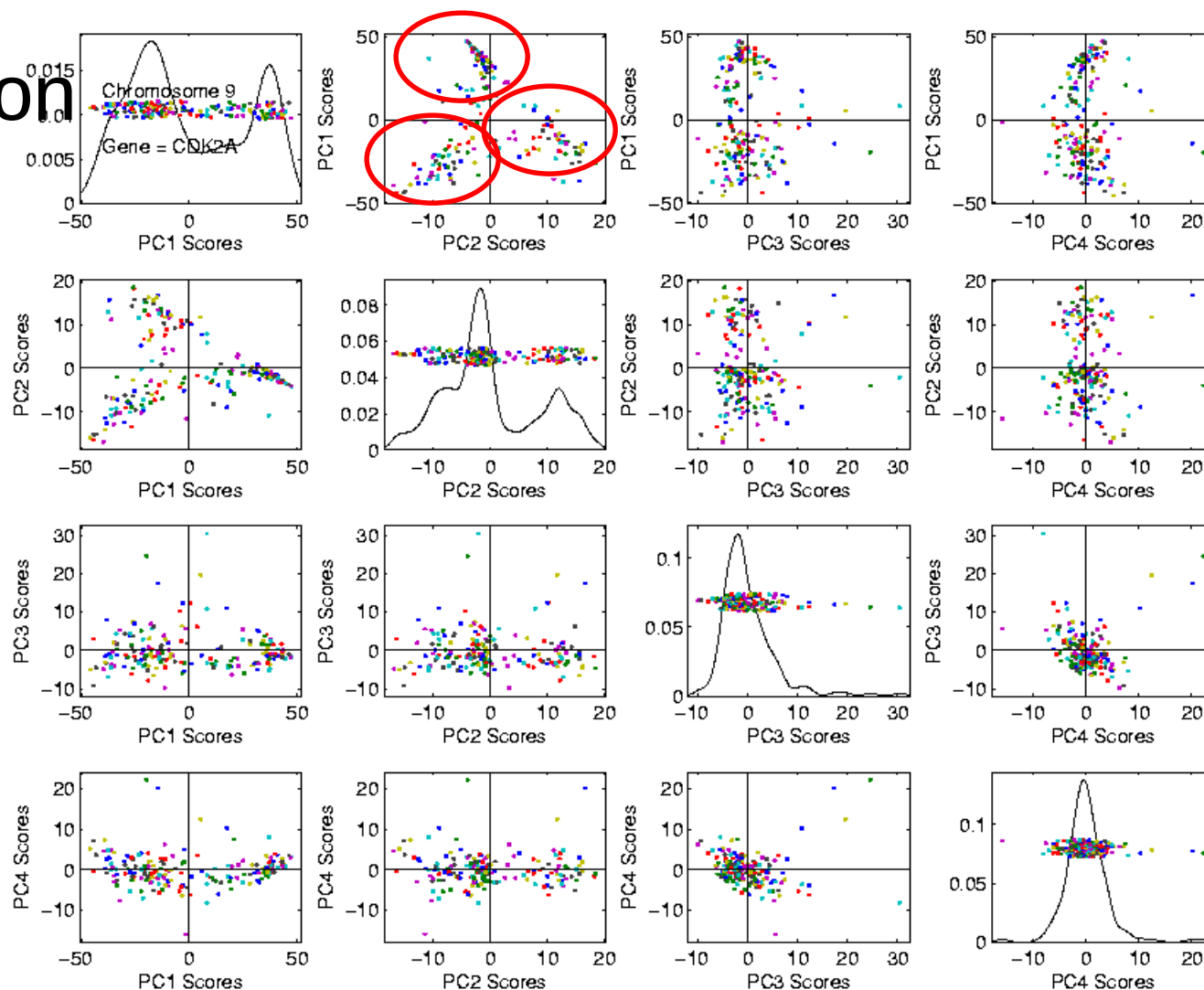


Functional Data Analysis

UNC, Stat & OR

Suggestion
Of
Clusters

Which
Are
These?

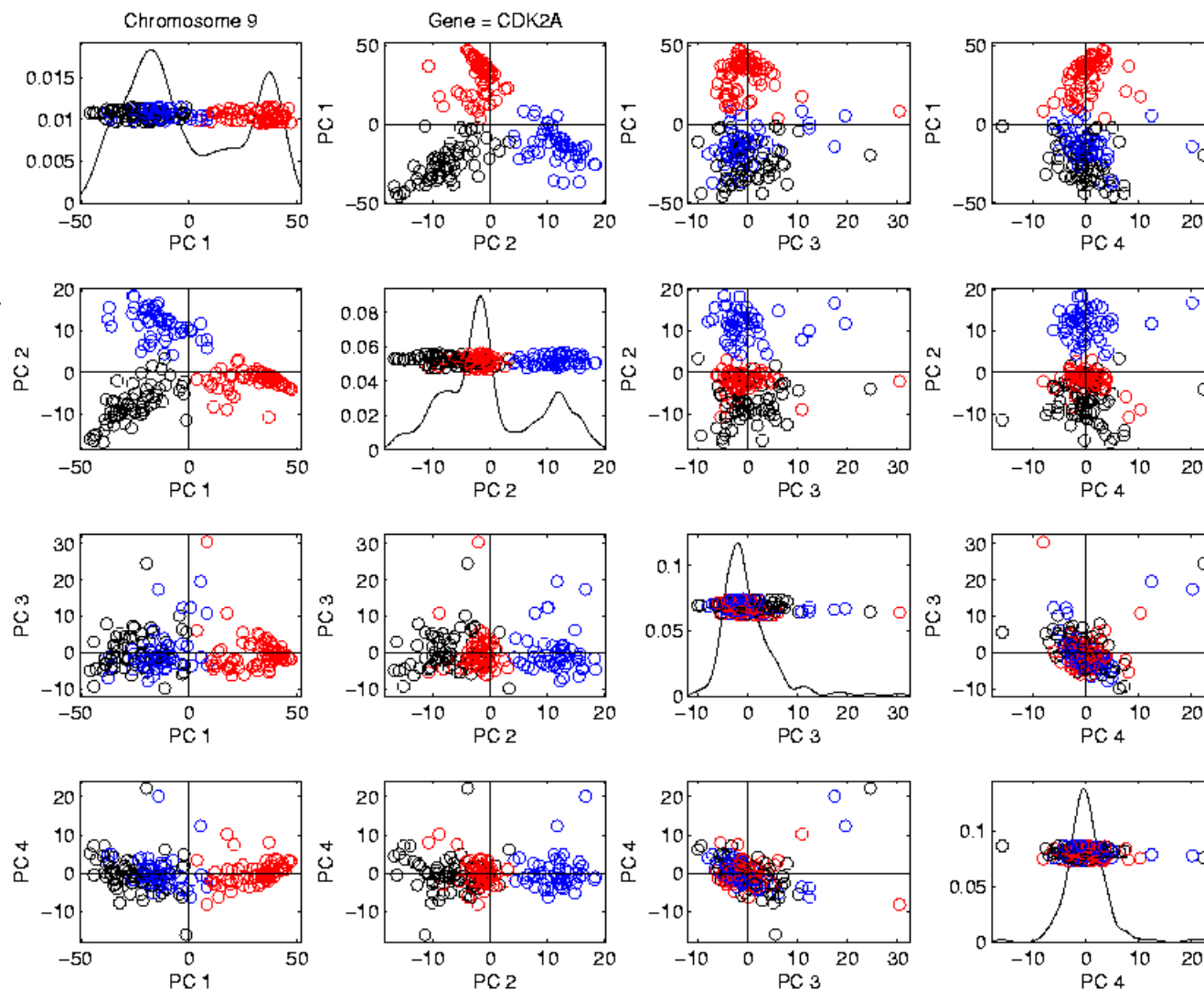




Functional Data Analysis

UNC, Stat & OR

Manually
Brush
Clusters



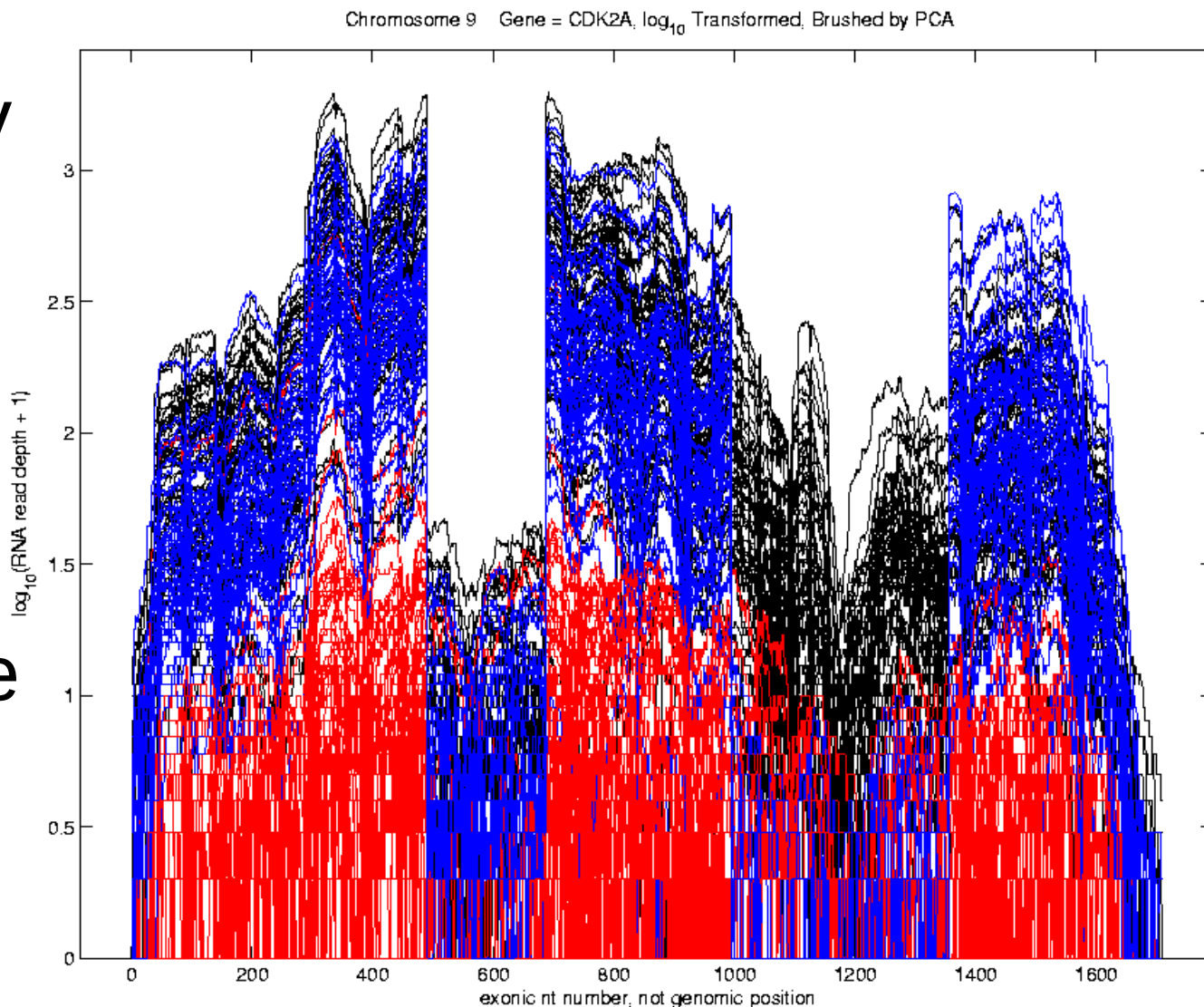


Functional Data Analysis

UNC, Stat & OR

Manually
Brush
Clusters

Clear
Alternate
Splicing





Functional Data Analysis

Consequences of this Visualization:

- ✓ Lead to Full Genome Screening Method
SigFuge
- ✓ Important Component: SigClust
(Which Clusters are *Really There*?)
- ✓ Found New Splices
(Now Been Biologically Verified)



Object Oriented Data Analysis

What is the “atom” of a statistical analysis?

- 1st Course: Numbers
- Multivariate Analysis Course : Vectors
- Functional Data Analysis: Curves
- More generally: **Data Objects**



Personal Motivating Contexts

Interdisciplinary Areas:

- Cancer Genetics
- Medical Image Analysis
- Evolutionary Biology
- Drug Discovery



Some Special Cases of OODA

Data Object Types:

- Curves (Functional Data Analysis)
- Spectra (Non-Negative!)
- Images
- Shapes
- Trees
- Movies (Functional MRI)

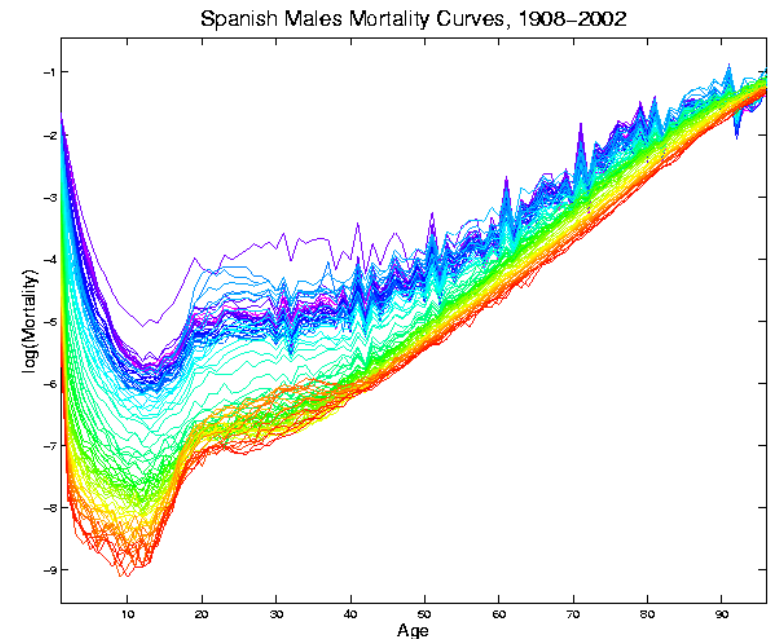
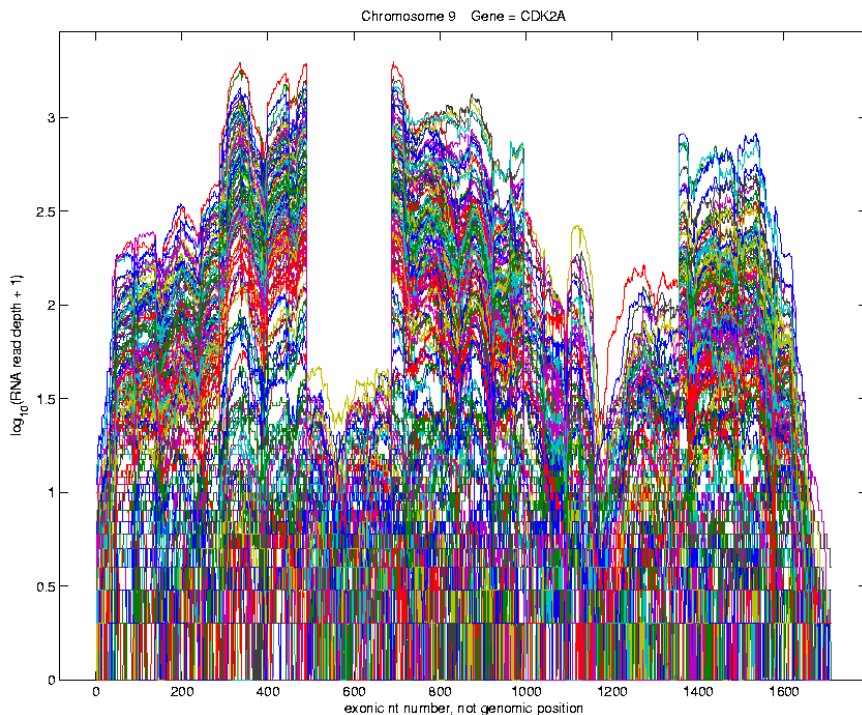
⋮



Curves as Data Objects (FDA)

UNC, Stat & OR

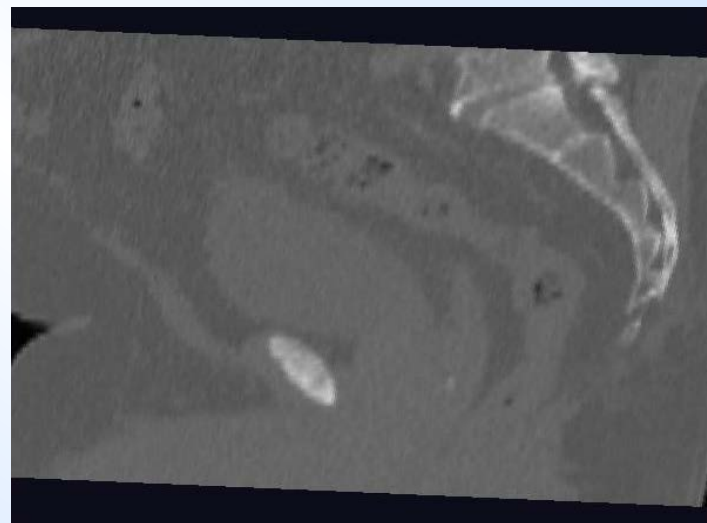
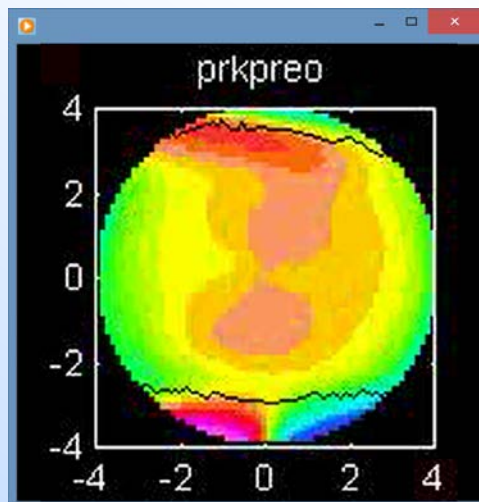
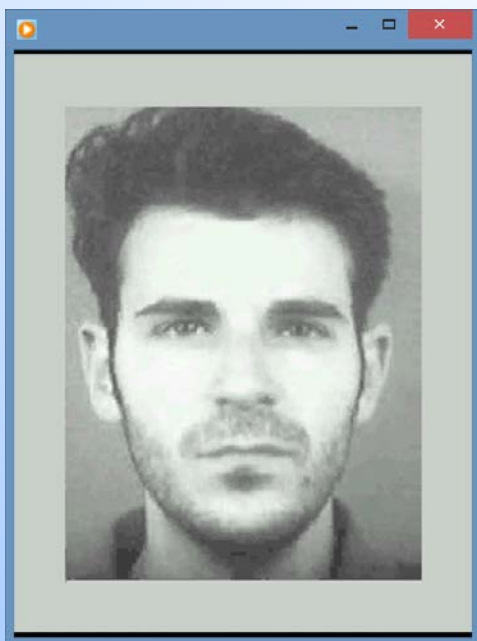
Generally Euclidean: Use standard methods





Images as Data Objects

Challenge: High Dimension, Low Sample Size





Shapes as Data Objects

Challenge: Data Lie in (Curved) Manifold

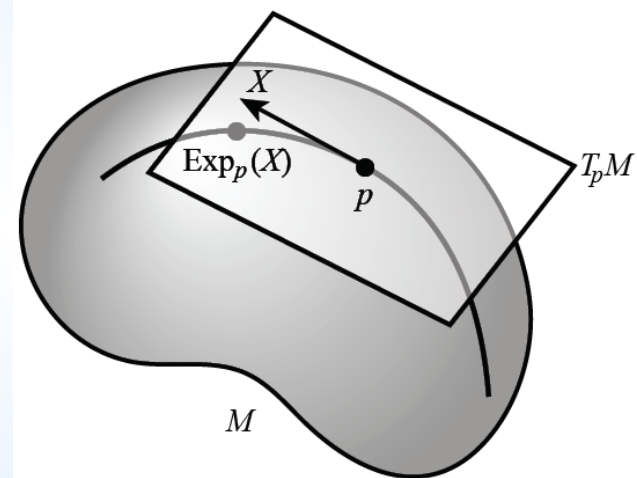
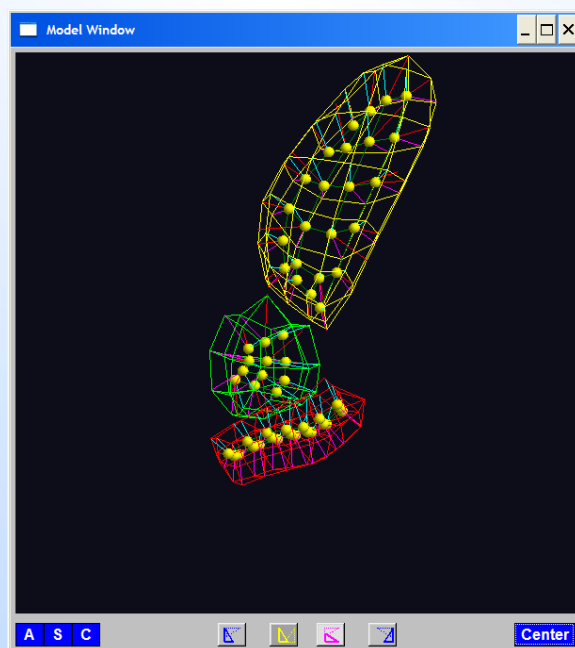
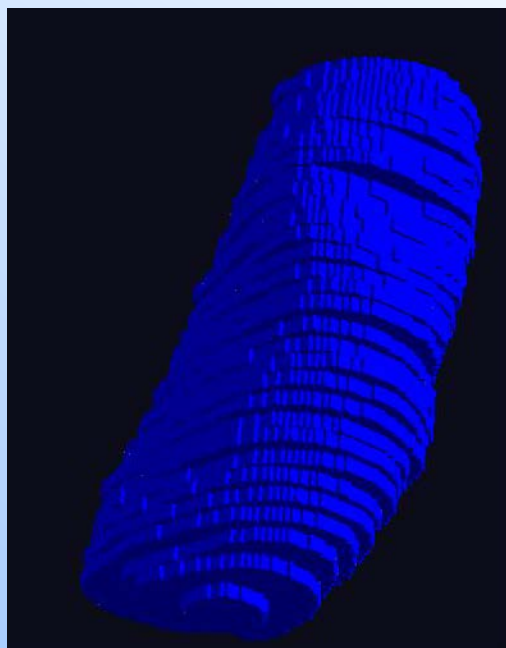


Figure 2.2: The Riemannian exponential map.



Shapes as Data Objects

Challenge: Data Lie in (Curved) Manifold

{Tackle With Differential Geometry}

Important General Development:

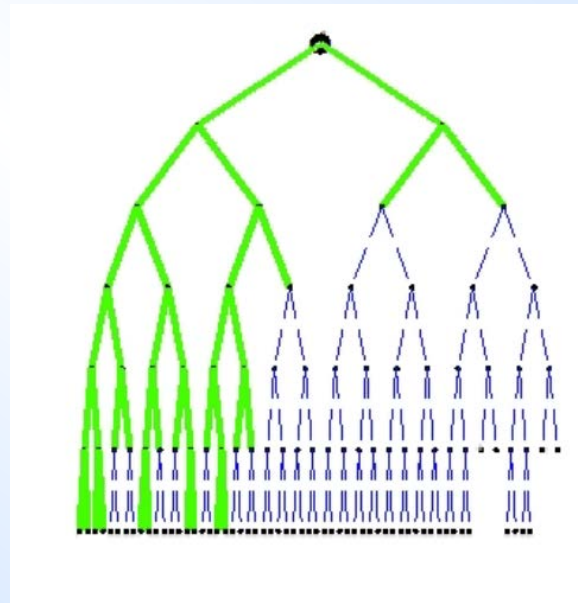
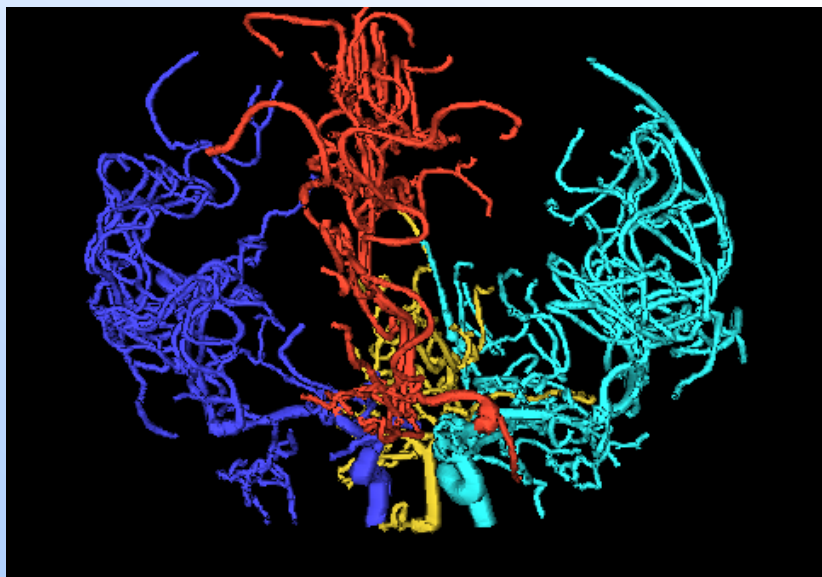
Backwards PCA



Trees as Data Objects

Challenge: More Complicated Data Space

Manifold Stratified Space





Trees as Data Objects

Challenge: More Complicated Data Space

Manifold Stratified Space

Surprisingly (?!?) Useful Approach:

Topological Data Analysis

Persistent Homology



Advertisement

Short Course on OODA & TDA

International Biometrics Conference Seoul July 2020



Moo K. Chung, U. Wisc.



Yuan Wang, U. S. C.



Carolina Breast Cancer Study

UNC, Stat & OR

The Carolina Breast Cancer Study

Phase III: The Jeanne Hopkins Lucas Study

Carolina Breast Cancer Study
University of North Carolina-Chapel Hill Lineberger Comprehensive Cancer Center
Funded by
*University Cancer Research Fund
National Cancer Institute
Susan G. Komen*

NATIONAL CANCER INSTITUTE

susan g. komen.

CAROLINA BREAST CANCER STUDY

UNC CANCER CARE

[Home](#) [Breast Cancer Resources](#) [Meet the Staff](#) [For Participants](#) [For Researchers](#)

Thanks to: **Iain Carmichael** (Deep Learning, AJIVE)
Melissa Troester (Head, CBCS)
Joseph Geradts, Benjamin Calhoun (Pathology)
Katie Hoadley, Chuck Perou (Genomics)





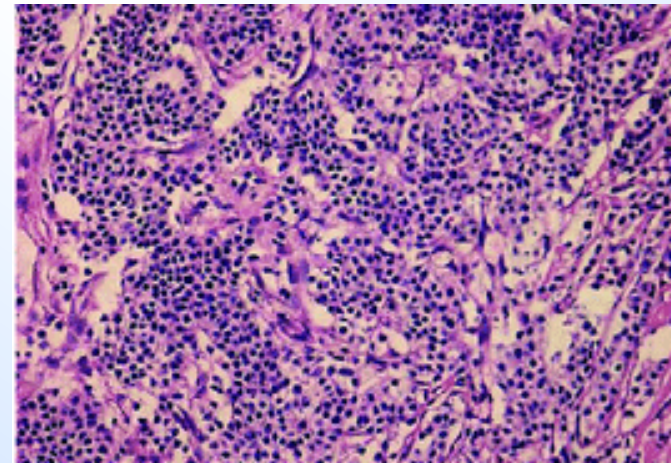
Carolina Breast Cancer Study

UNC, Stat & OR

Clinical Diagnosis of Cancer:
Pathologist Views Tissue Under Microscope,
Tissue Stained with Hematoxylin & Eosin (H&E)



Thanks to BBC.CO.UK
and Reseachgate.net





Carolina Breast Cancer Study

UNC, Stat & OR

Analysis Goal:

Clinical Diagnostic Standard

Understand How Images & Genomics

- Work Together
- Work Separately

Deep Insight &
New Treatments

Approach: JIVE

As of Jan. 2020: ~70
Drugs Approved for
Breast Cancer Therapy



Joint & Individual Variation Explained

(Angle Based)



JIVE Collaborators

UNC, Stat & OR



Eric Lock

Qing Feng



Andrew Nobel



Jan Hannig





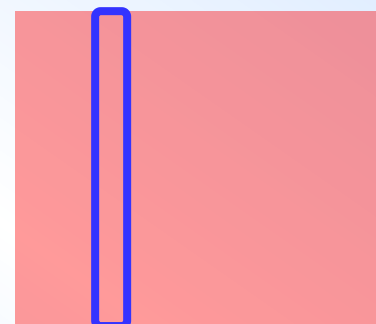
JIVE Data Structure

JIVE Organizational Model:
Multiple Matrices

(Data Types, i.e. “Blocks”)

With common

Columns as Data Objects

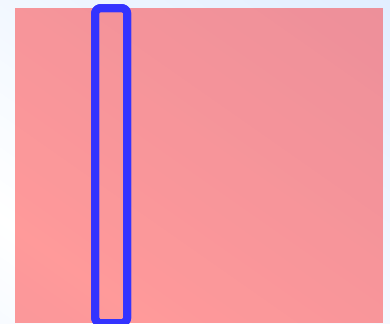
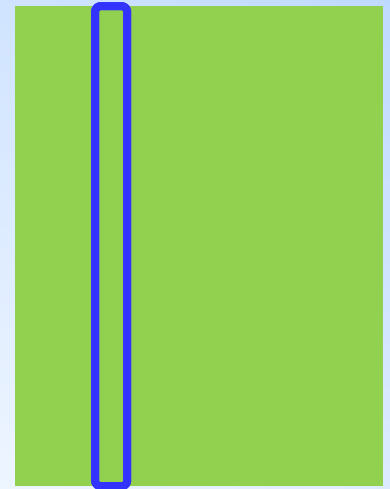




JIVE Analytic Goals

Explore & Quantify Variation

In spirit of PCA
(Principal Component Analysis)





Carolina Breast Cancer Study

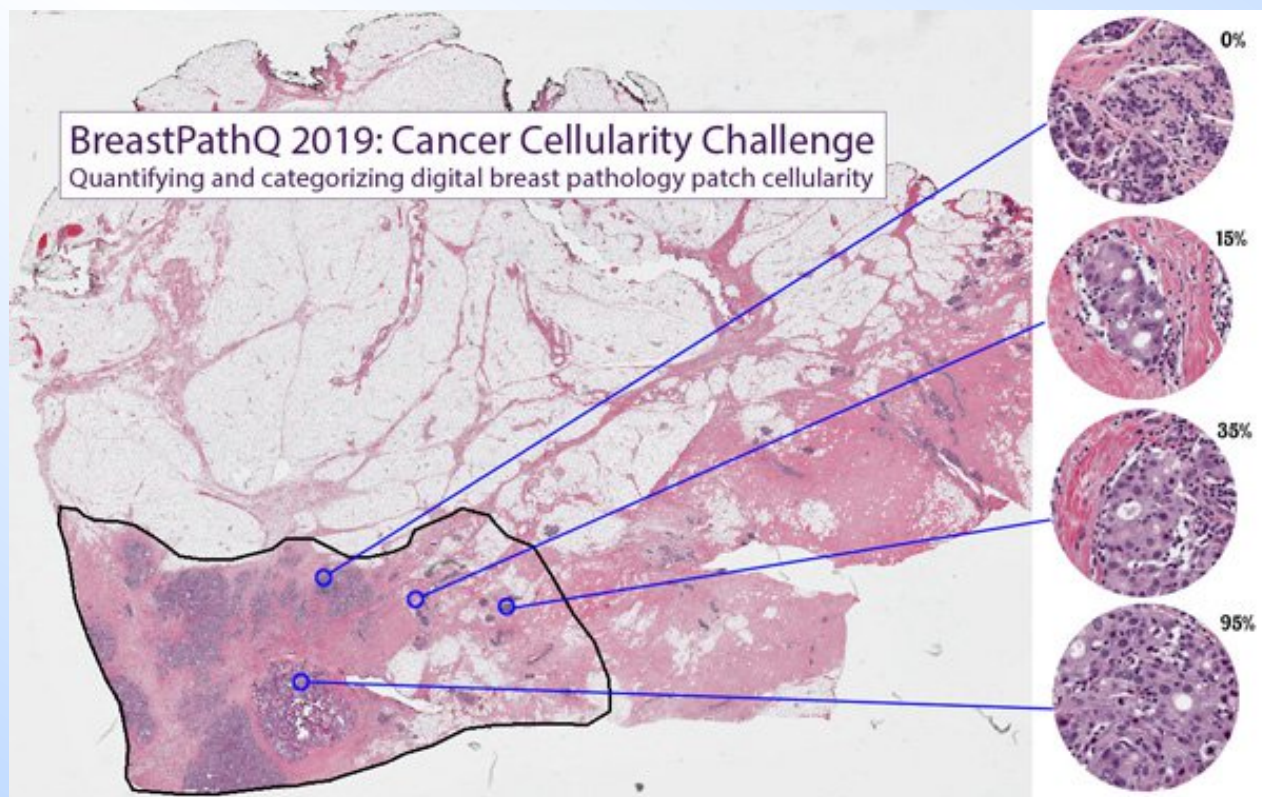
UNC, Stat & OR

Tissue Micro Array Data:

Extract Small

(1mm diam.)

“Cores”



Thanks to SPIE.ORG

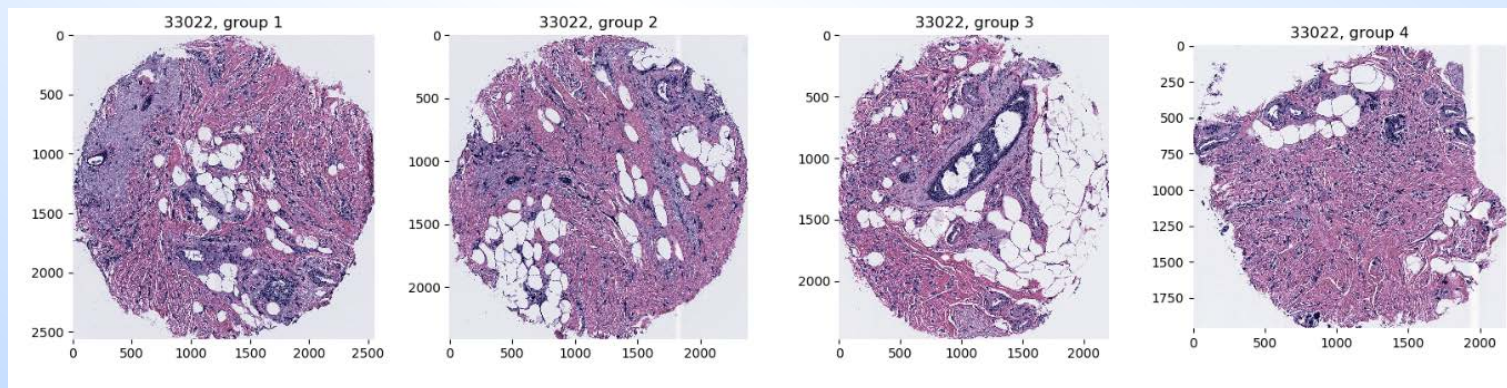


Carolina Breast Cancer Study

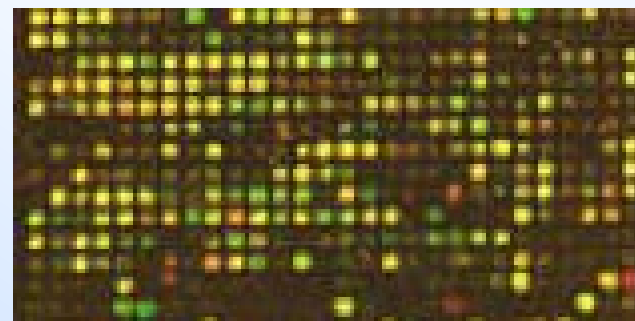
UNC, Stat & OR

Experimental Design: $n = 1191$ people

For Each: 1 – 4 TMA Cores



PAM 50 Gene Expression:





Carolina Breast Cancer Study

UNC, Stat & OR

PAM 50 Gene Expression:

Early Technology, More Recent
RNAseq → 10,000s Genes

- ☐ Set of 50 Genes
- ☐ Measured mRNA Expression Level
- ☐ Good at Separating SubTypes
 - ☐ Basal
 - ☐ Her2
 - ☐ Luminal A
 - ☐ Luminal B

Perou Discovery:
No Benefit From
Chemo-Therapy





Carolina Breast Cancer Study

UNC, Stat & OR

Deep Learning Image Representation

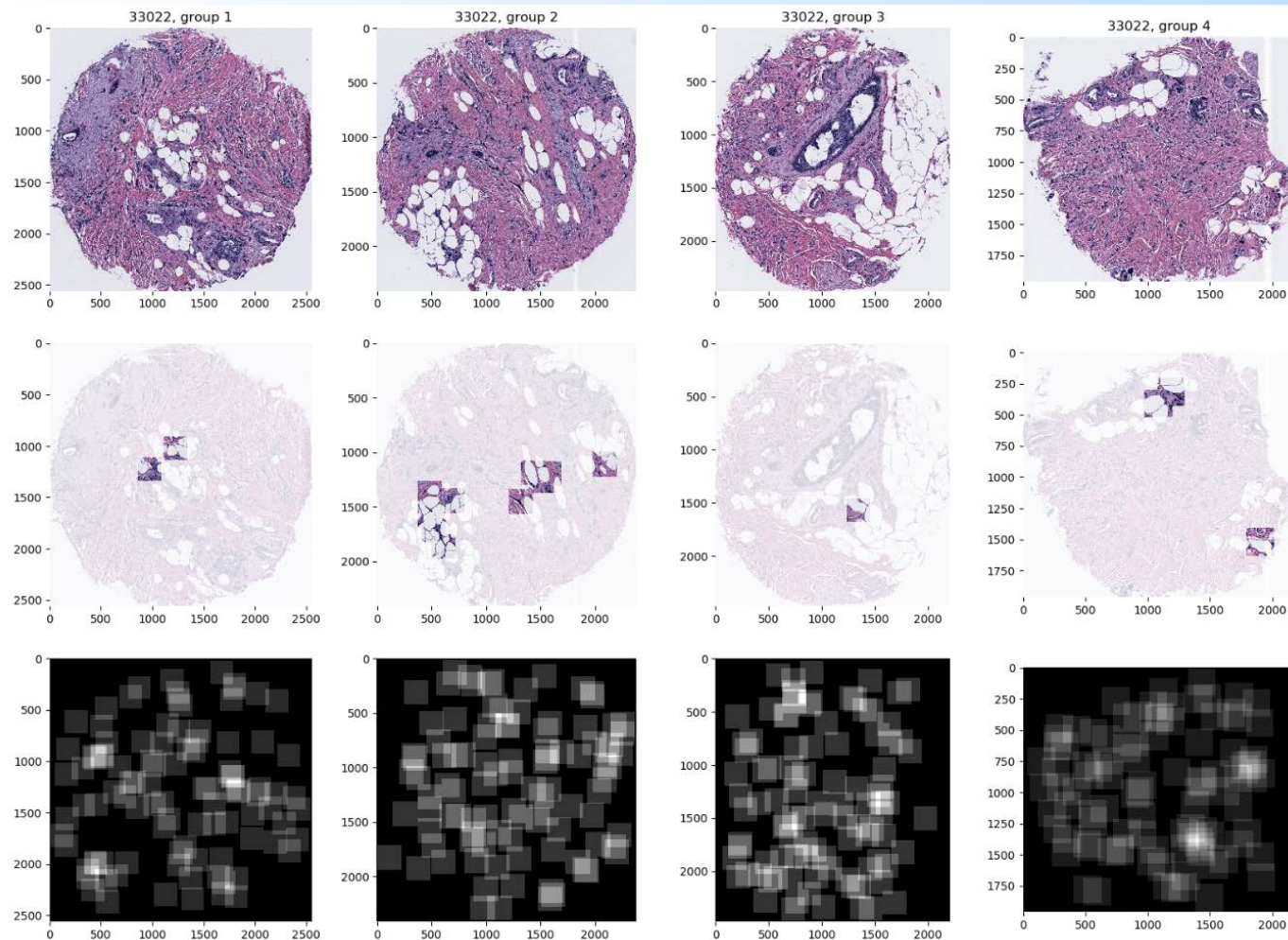
Each Core:

Randomly

Select 100

224×224

Patches





Carolina Breast Cancer Study

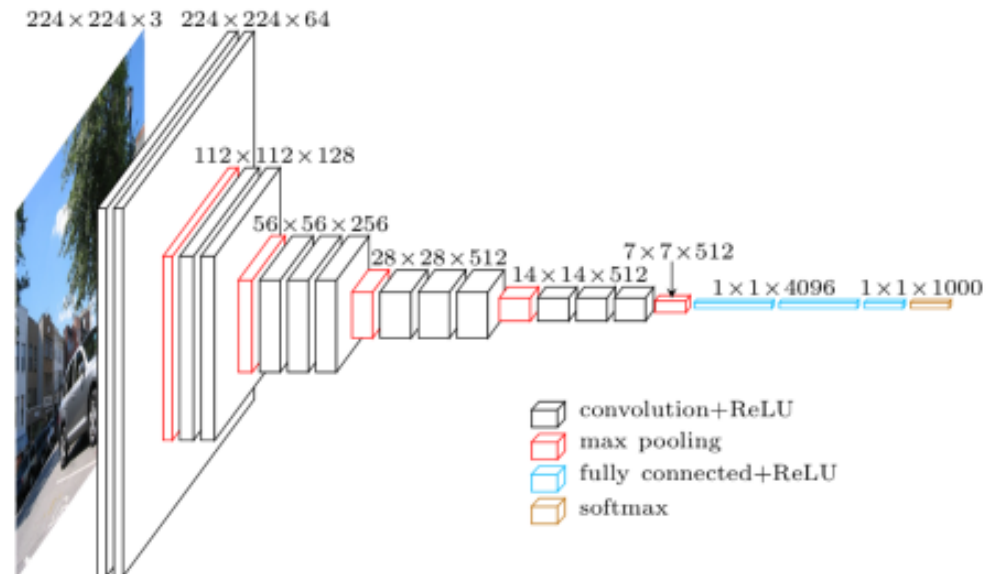
UNC, Stat & OR

Deep Learning Image Representation

Reduce Each Patch to 512 Features

Using Transfer Learning From VGG16:

(Trained on Many
Natural Images)



Thanks to pyimagesearch.com



Carolina Breast Cancer Study

UNC, Stat & OR

Deep Learning Image Representation

For Each Core:

Aggregate Patches by Averaging

(Damps Out “Location” Information)

Then Average Cores For Each Person



Carolina Breast C

UNC, Stat & OR

JIVE: Common Normal

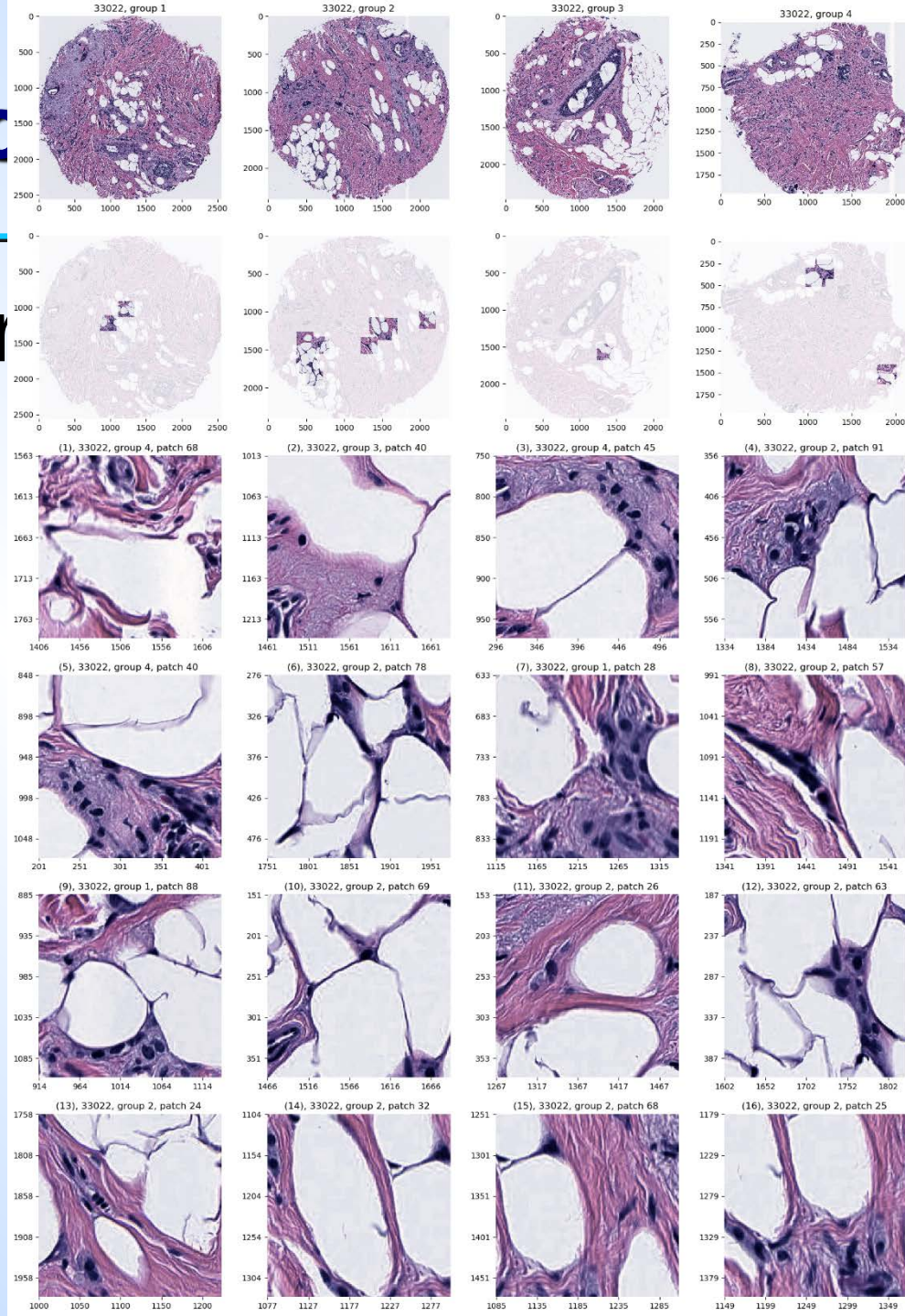
Look at Extremes

Negative End

Top Person

Top 16 Patches

Fat Cells & Stroma





Carolina Breast C

UNC, Stat & OR

JIVE: Common Normal

Look at Extremes

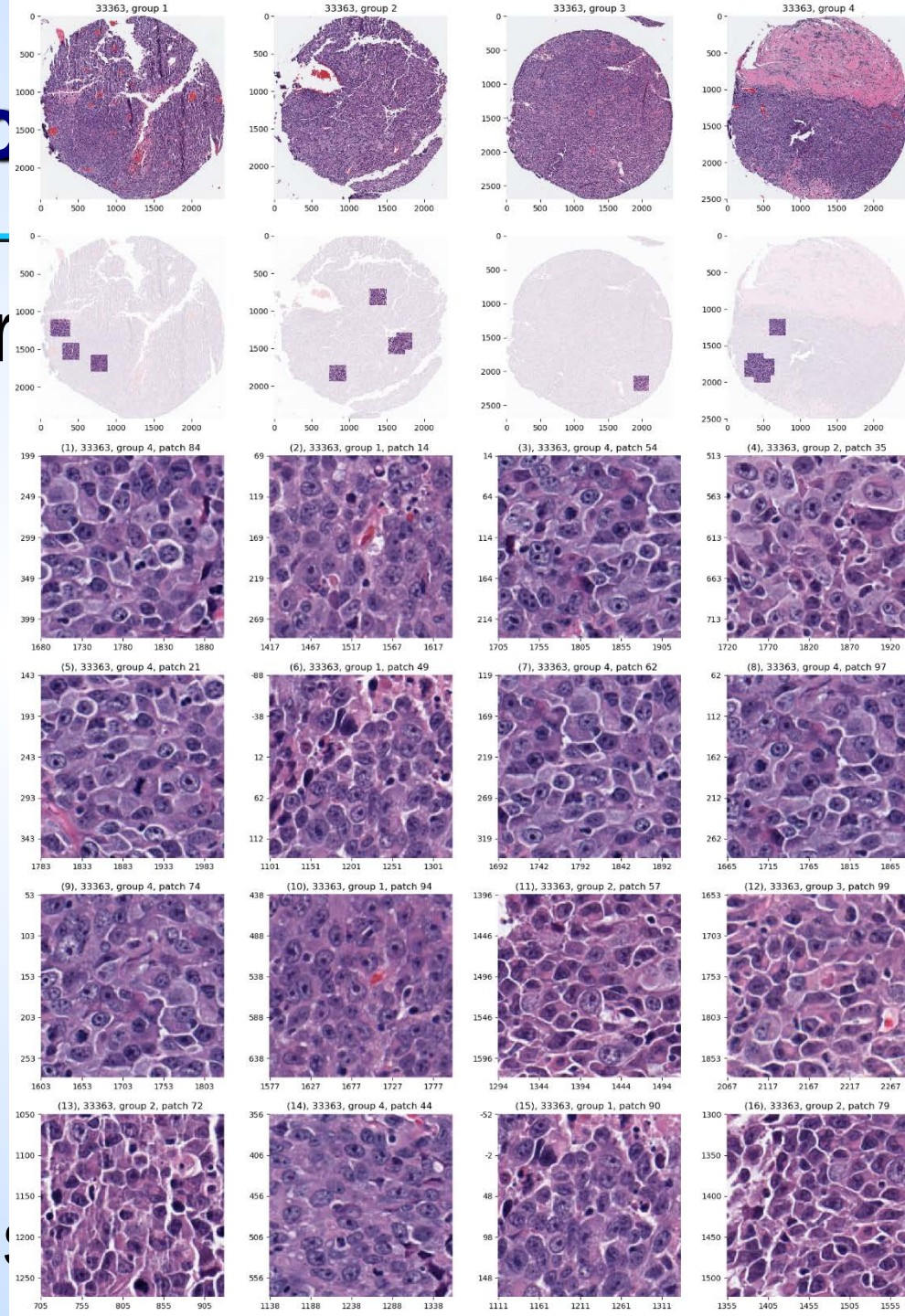
Positive End

Top Person

Top 16 Patches

Highly “Cellular”

Markedly Atypical Cells





Carolina Breast Cancer Study

UNC, Stat & OR

JIVE: Gene Expression

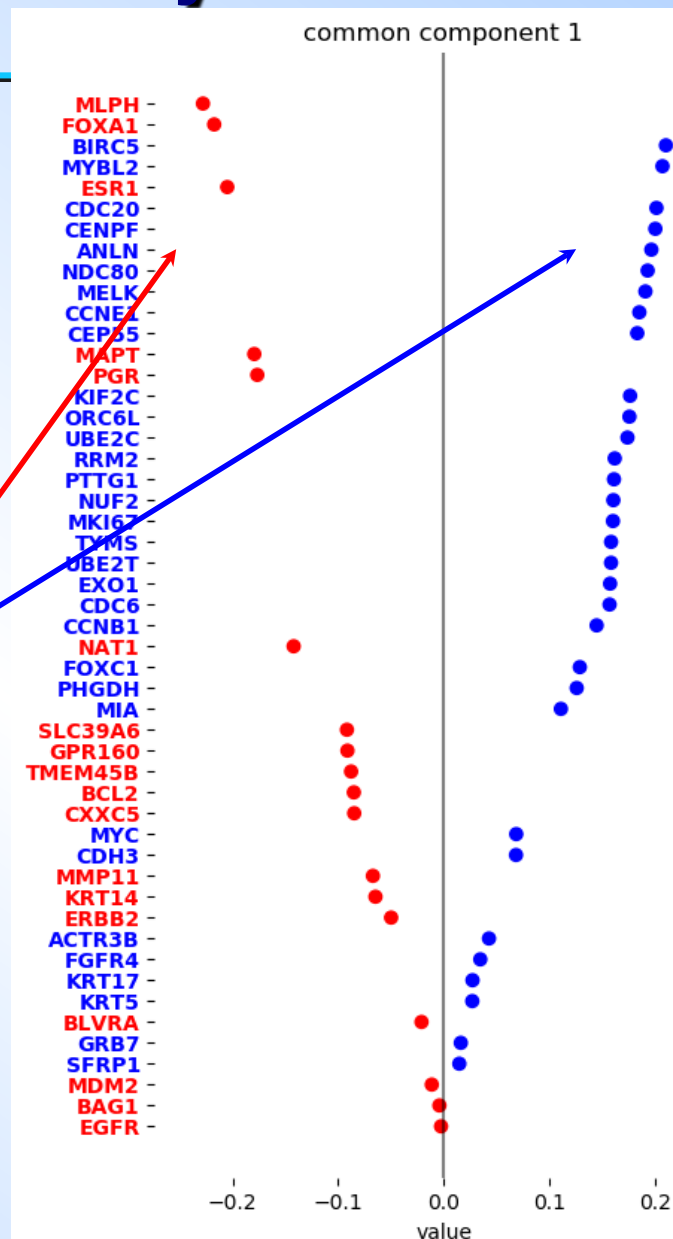
Common Normalized

Loadings 1

“High Proliferation” Genes

“Low Prolif.” – Lum A Genes

Very Consistent w/ Image





Carolina Breast C

UNC, Stat & OR

JIVE: Common No

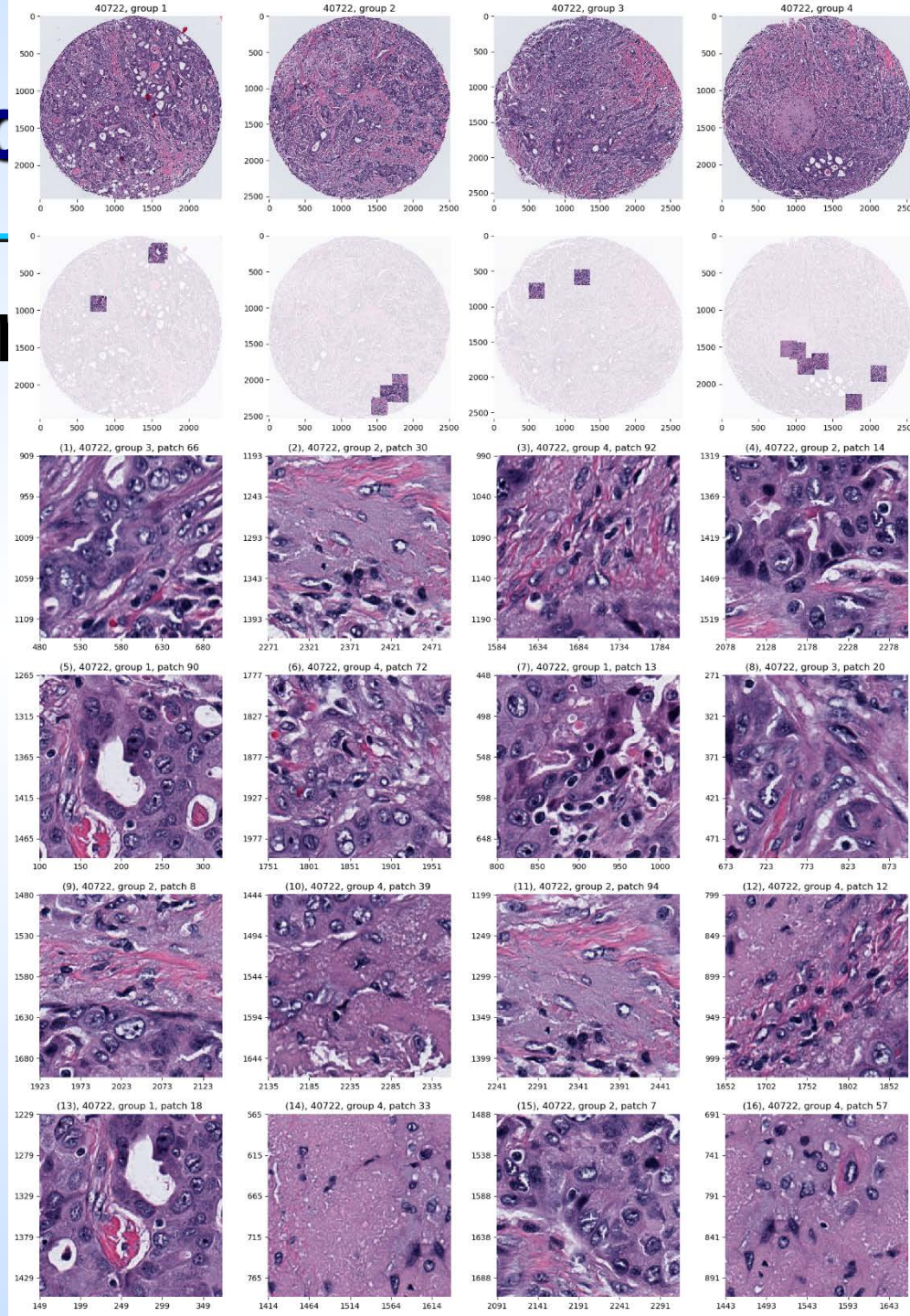
Look at Extremes

Negative, Top

Top 16 Patches

High Nuclear Grade

Atypical Cells





Carolina Breast C

UNC, Stat & OR

JIVE: Common Normal

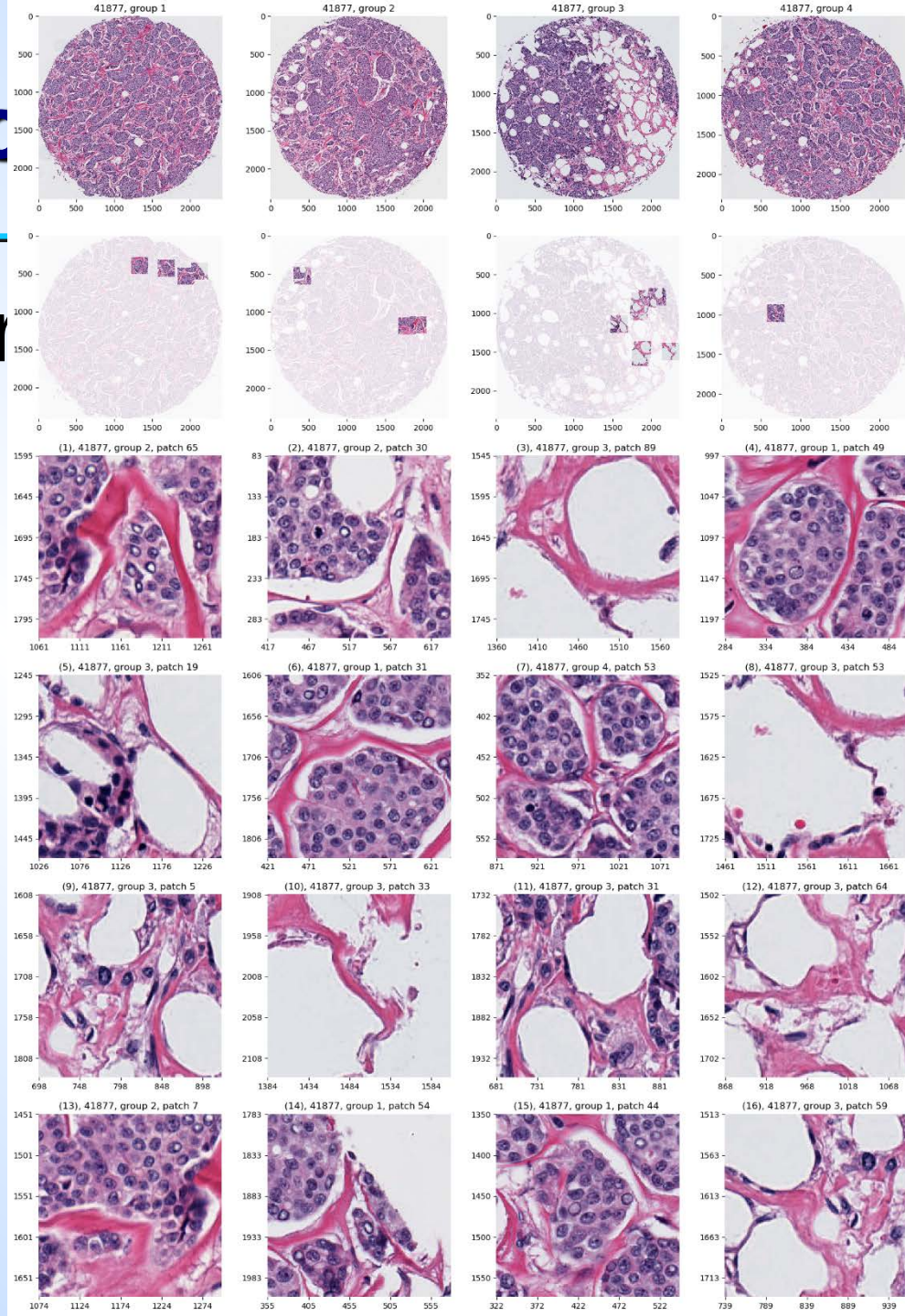
Look at Extremes

Positive, Top

Top 16 Patches

Lower Nuclear Grade

Stroma, Sclerosis





Carolina Breast Cancer

UNC, Stat & OR

JIVE: Gene Expression

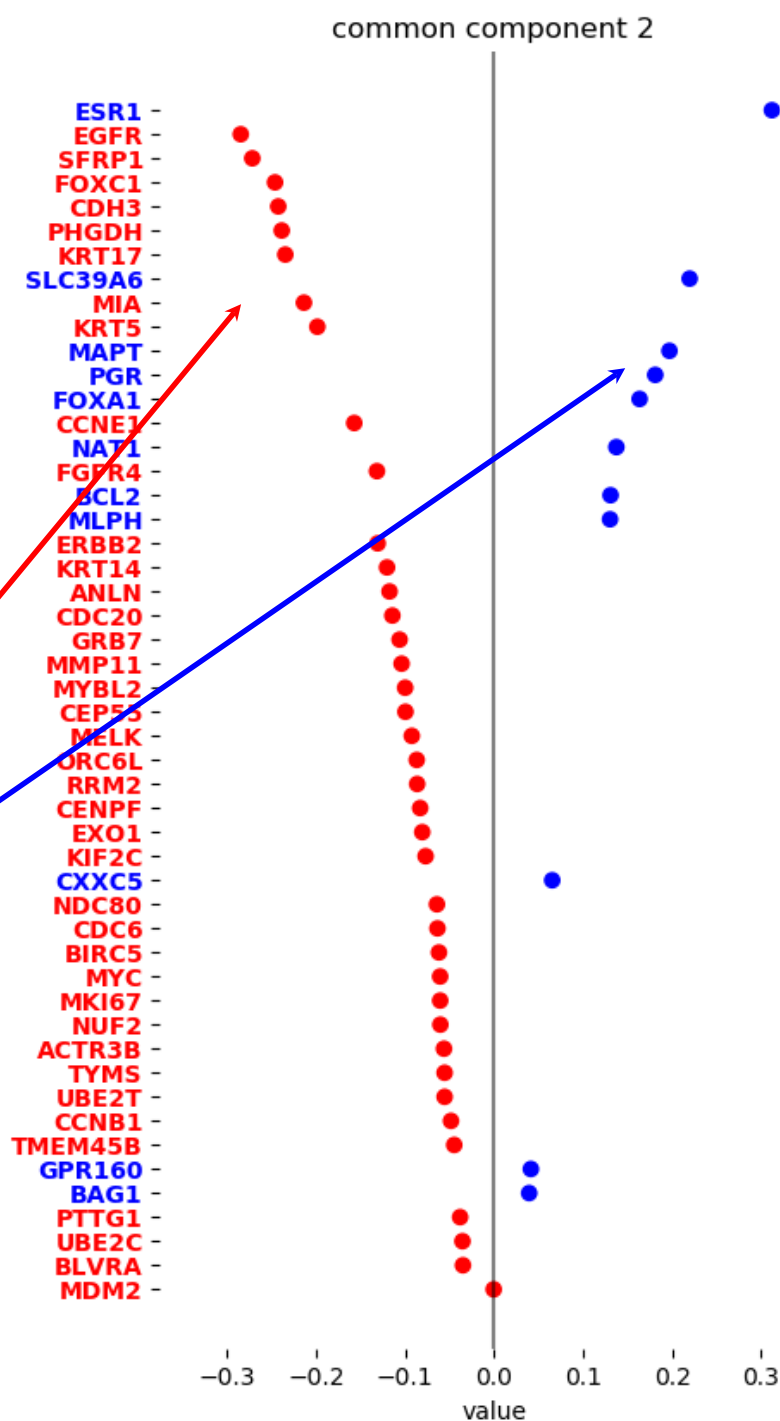
Common Normalized

Loadings 2

Luminal Subtype Genes

Basal Subtype Genes

Very Consistent w/ Image





Carolina Breast Cancer

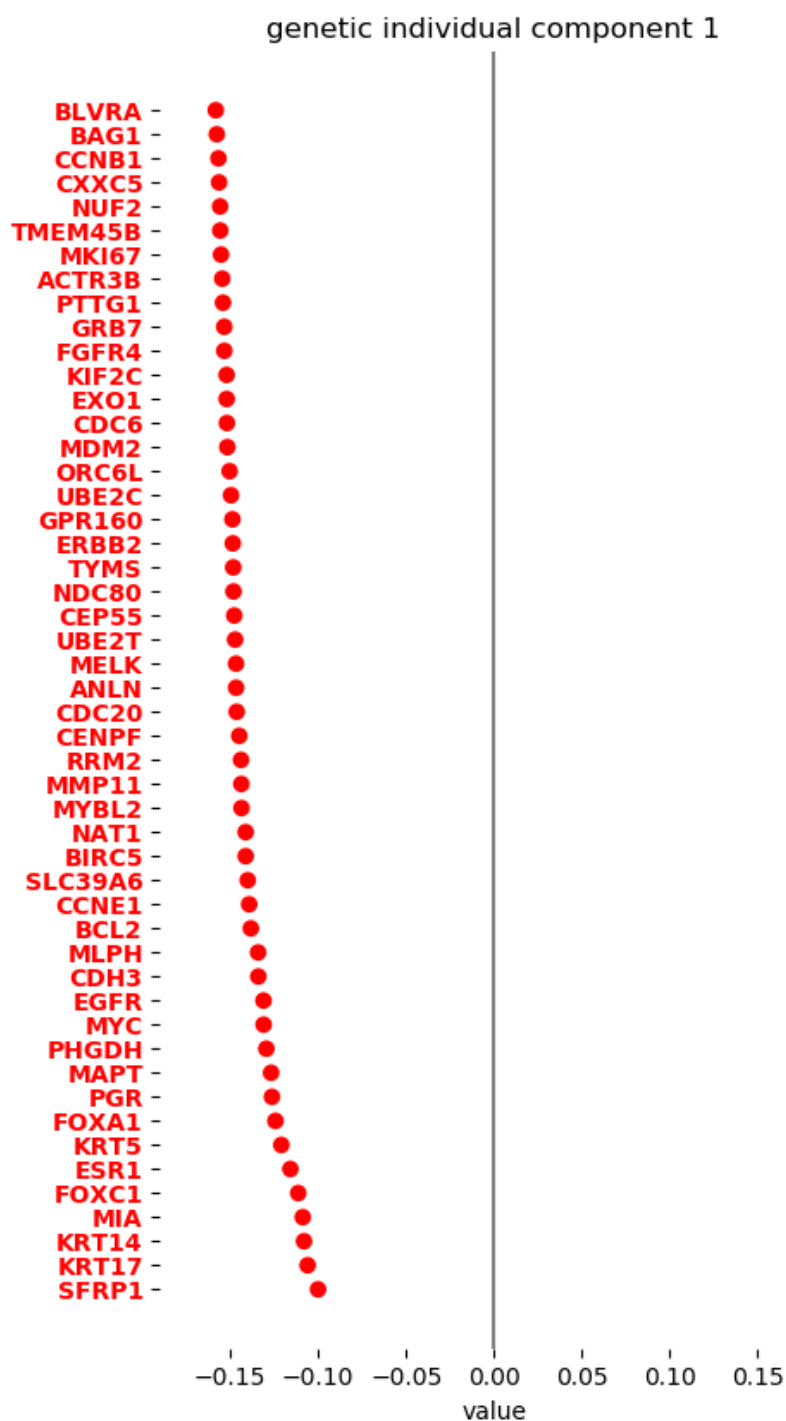
UNC, Stat & OR

JIVE, Genes, Individual

Overall Up & Down

Together

Not Subtype Related!





Carolina Breast Cancer

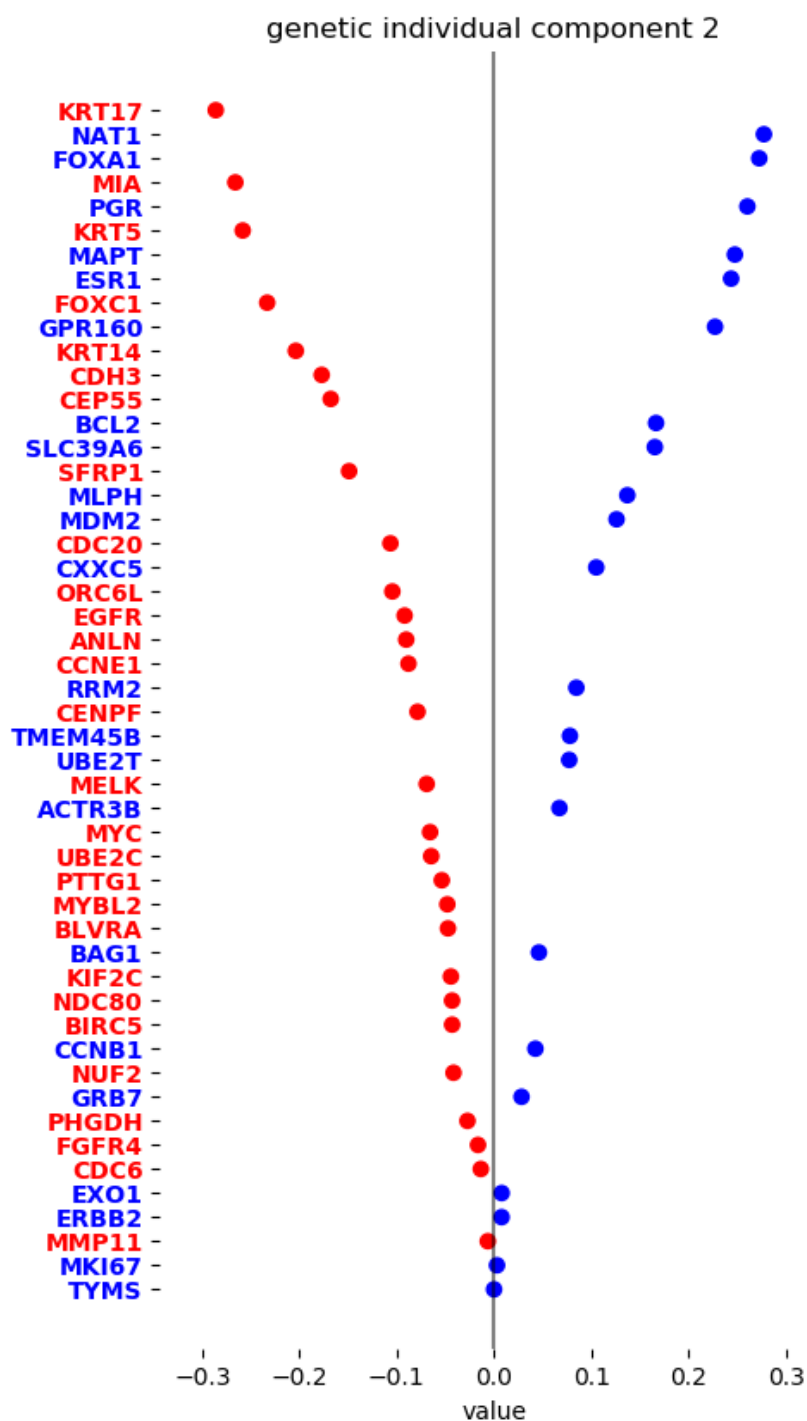
UNC, Stat & OR

JIVE, Genes, Individual

Luminal vs. Basal

“Contrast”?

“Unbalanced Types”?





Carolina Breast C

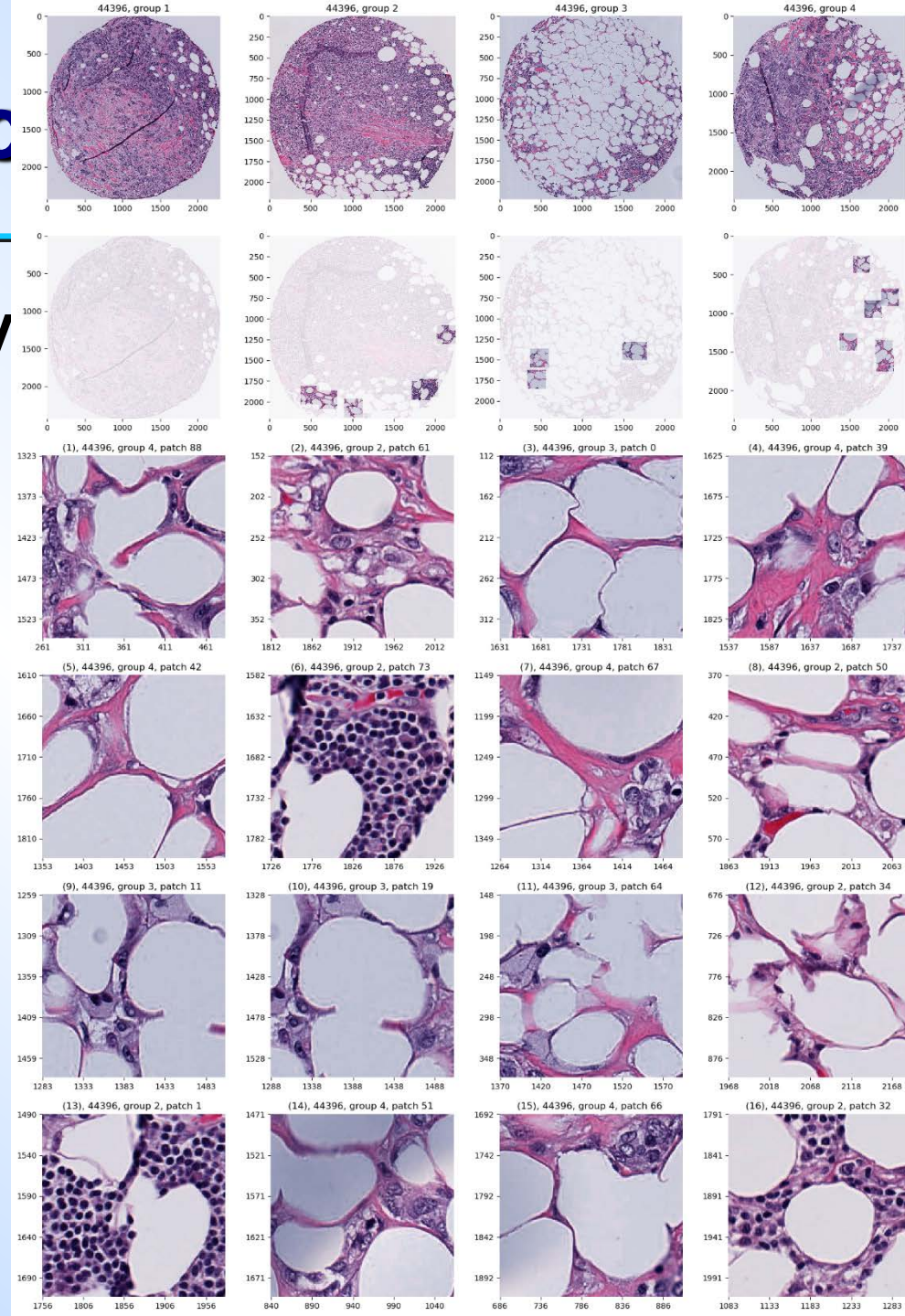
UNC, Stat & OR

JIVE, Images, Individ

Negative

Mostly Fat Cells

Few Nuclei





Carolina Breast C

UNC, Stat & OR

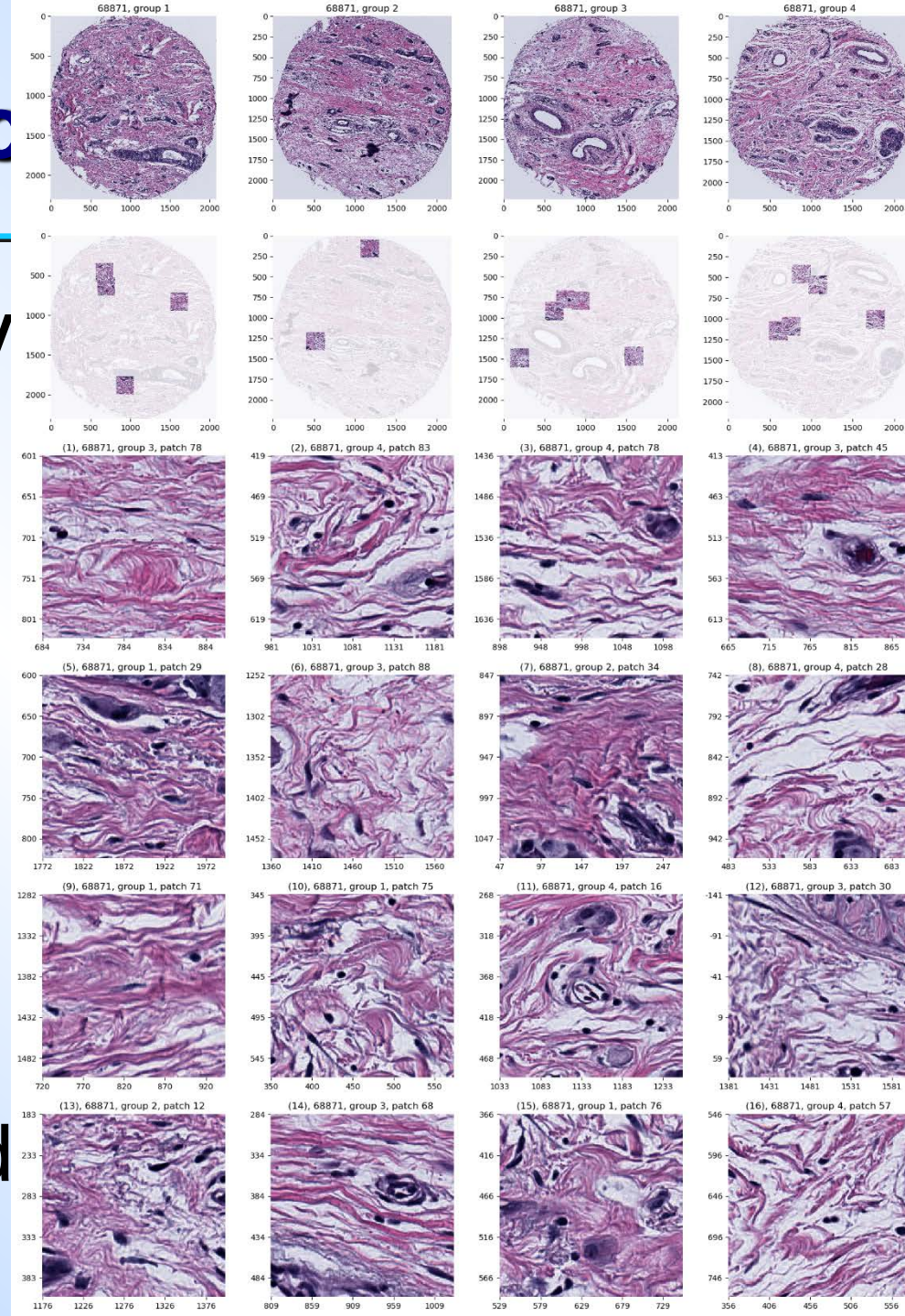
JIVE, Images, Individ

Positive

Reactive Stroma,

Few Nuclei

Little Gene Connected





Carolina Breast C

UNC, Stat & OR

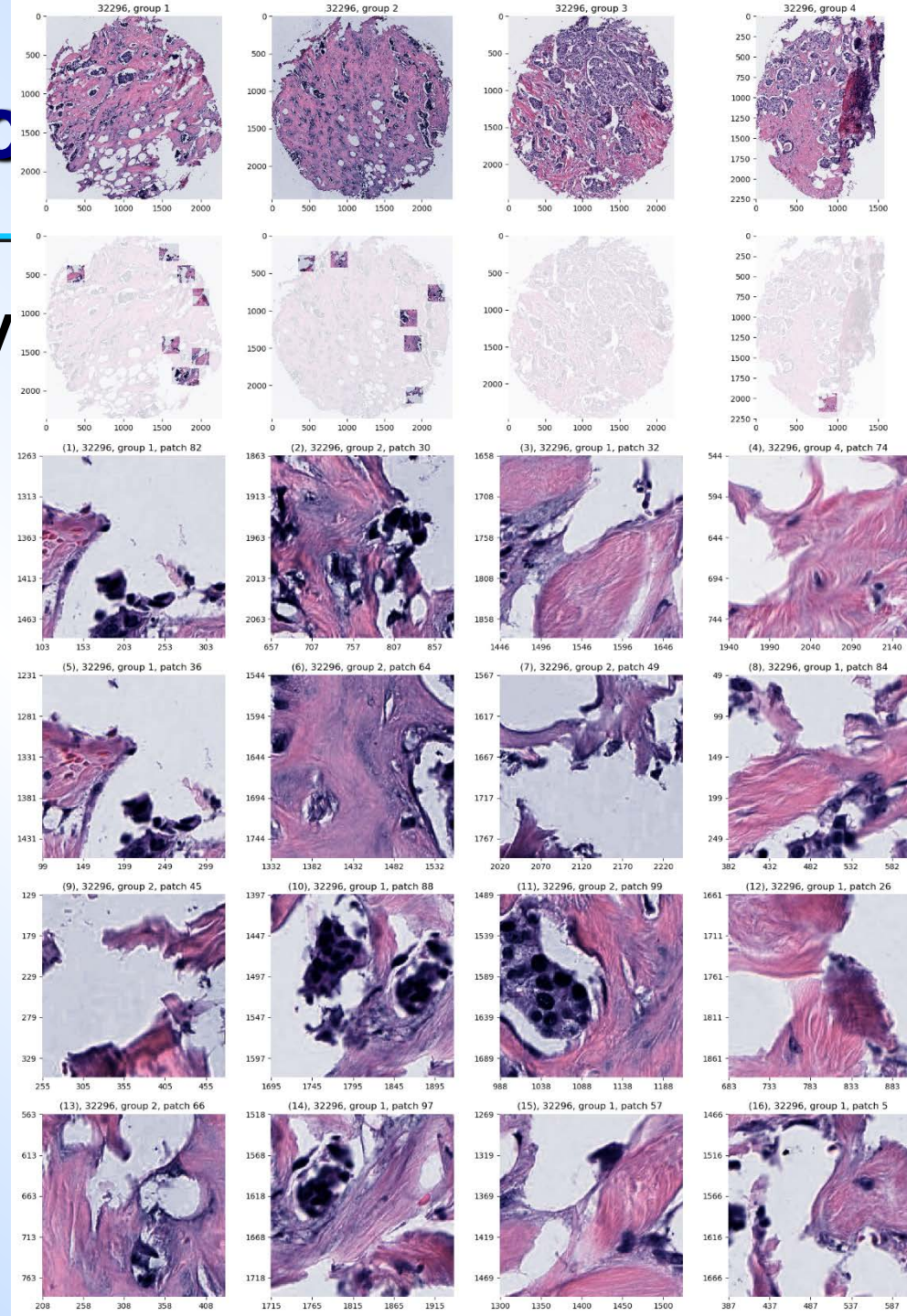
JIVE, Images, Individ

Negative

Mucinous & Micro-

Papillary Carcinoma

Not PAM50 Related





Carolina Breast C

UNC, Stat & OR

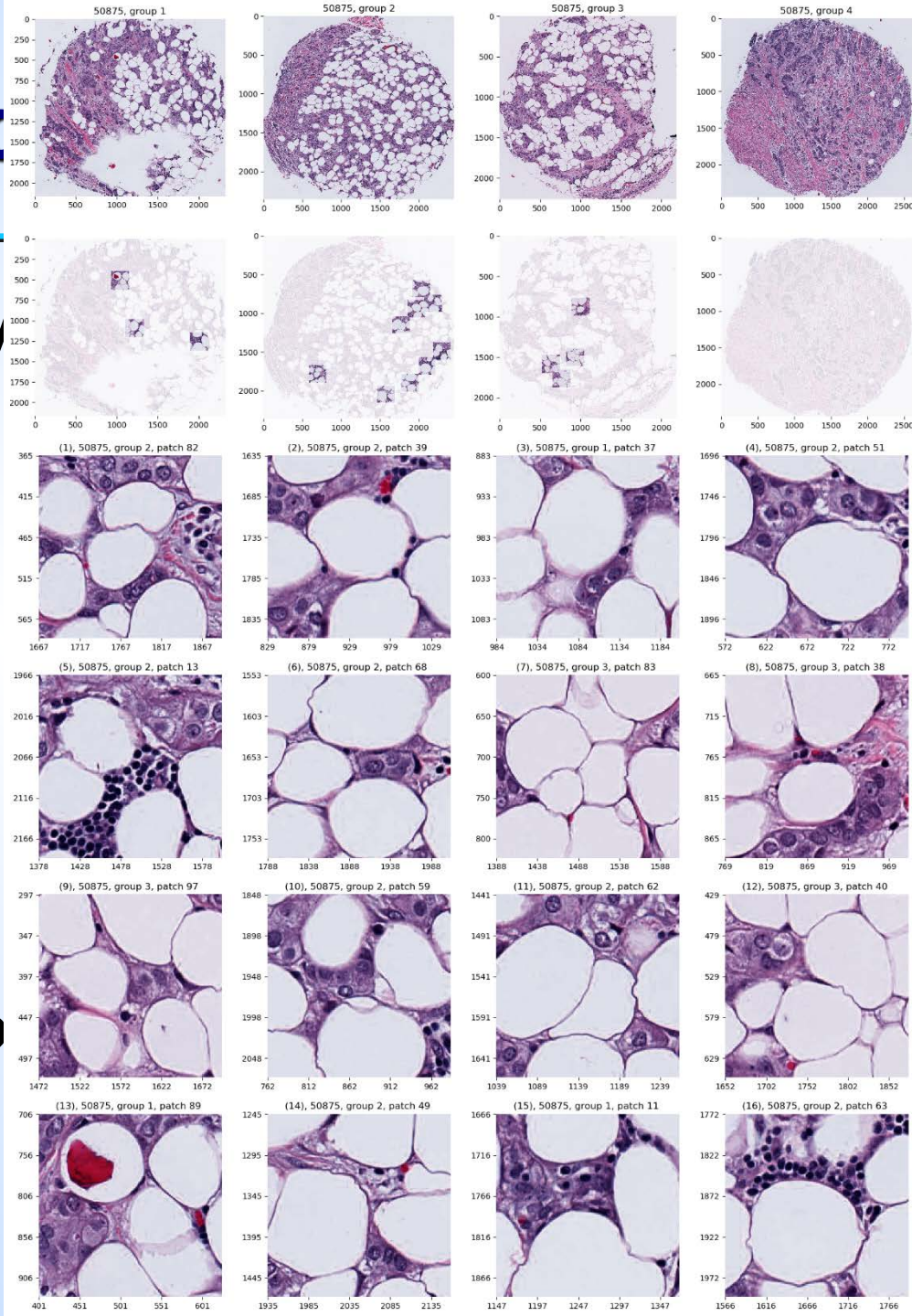
JIVE, Images, Individ

Positive

Fat Cells

Common Endpoint?

“Center”?





Next Generation Data Integration

Data Integration Via Subspace Analysis (DIVAS)



DIVAS / JIVE Collaborators

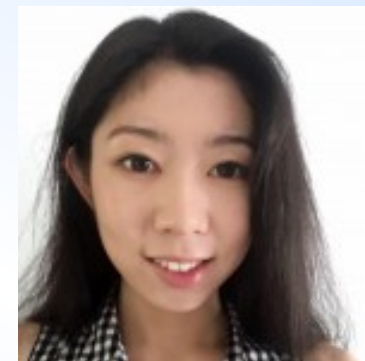
Jan Hannig



Meilei Jiang



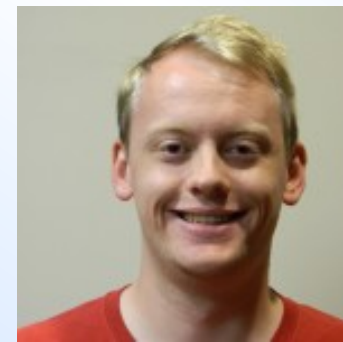
Xi Yang



Iain Carmichael



Jack Prothero





DIVAS Improves JIVE

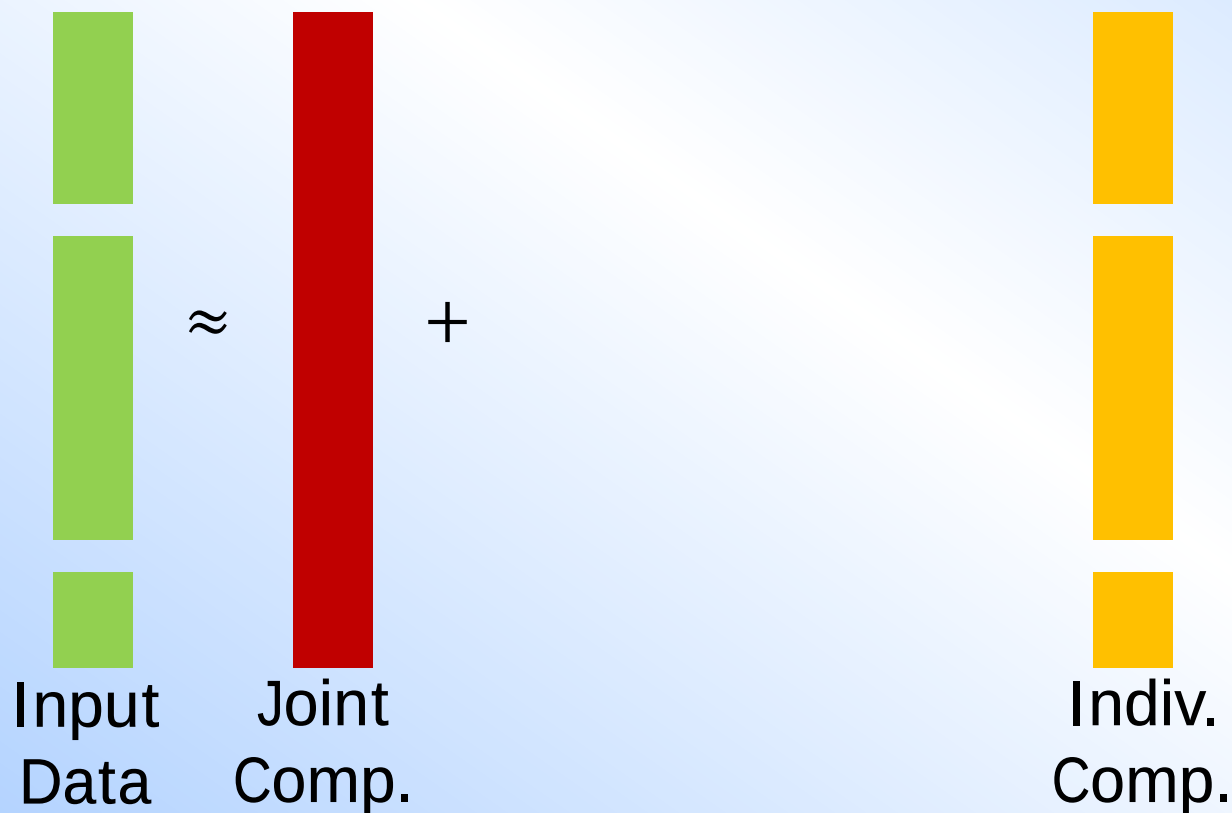
1. Partially Shared Blocks



DIVAS Improves JIVE

1. Partially Shared Blocks

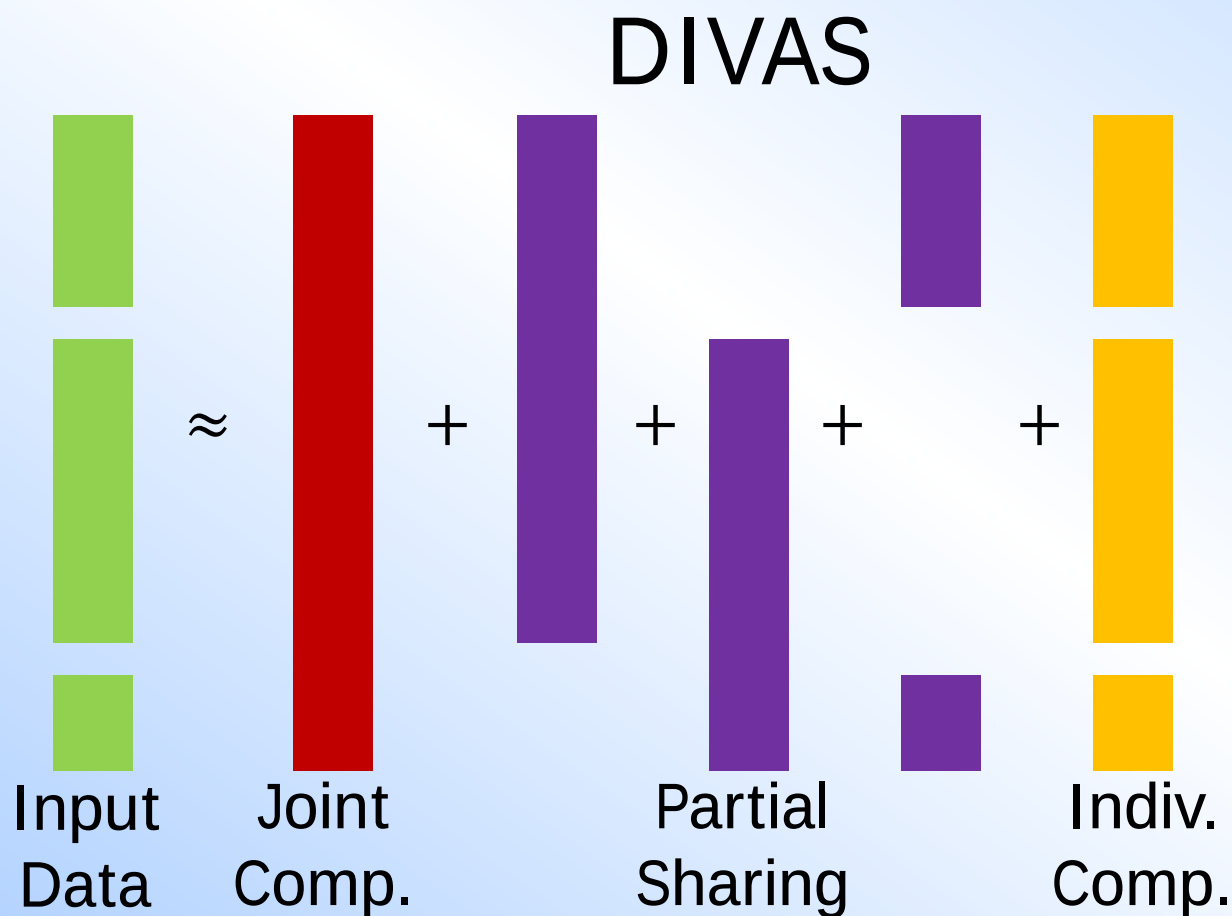
JIVE





DIVAS Improves JIVE

1. Partially Shared Blocks





DIVAS Motivation

UNC, Stat & OR

The Cancer
Genome
Atlas

Multiple Blocks

People Common
Across Blocks

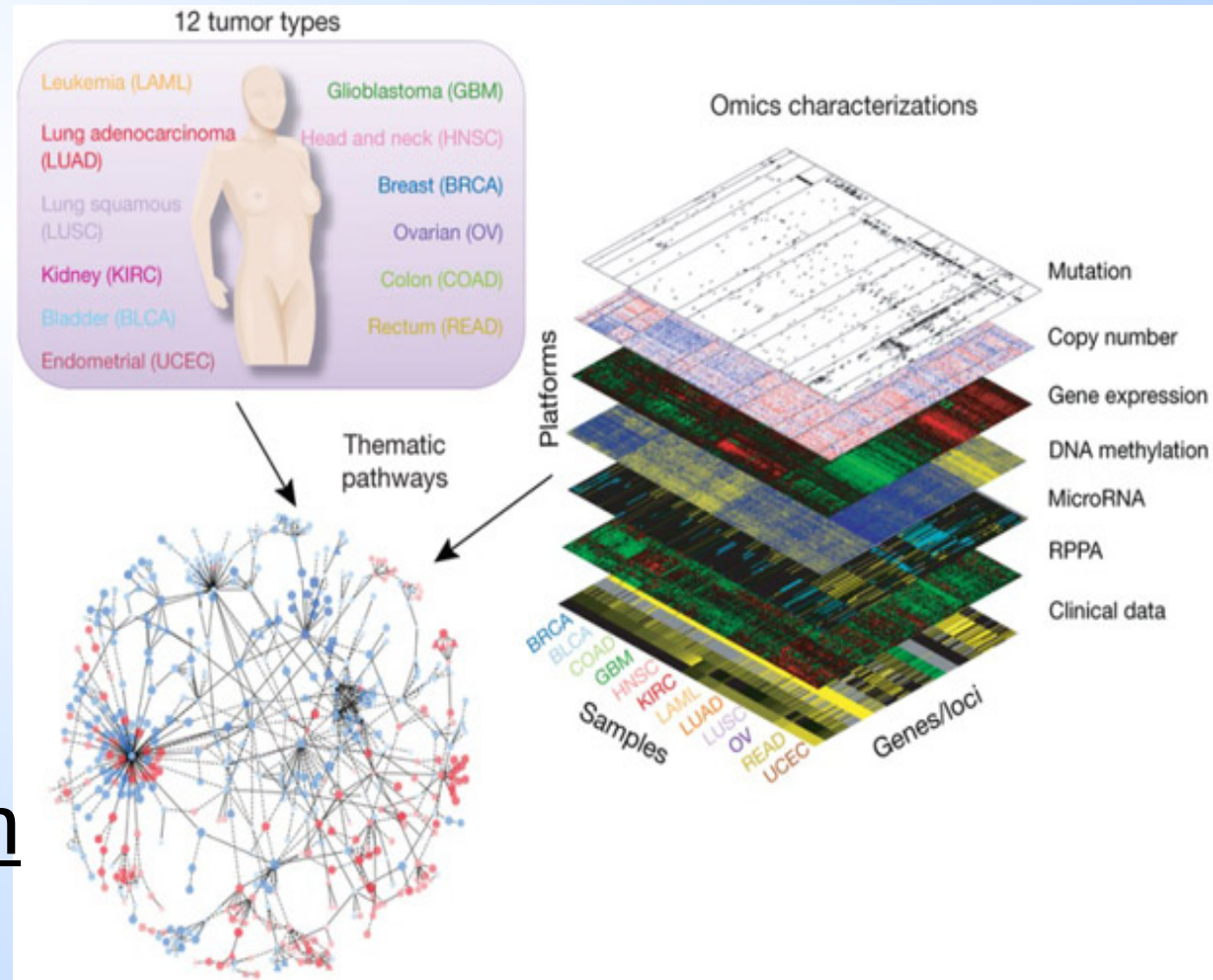


Figure : The Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013)

The Cancer Genome Atlas Pan-Cancer analysis project.



DIVAS on TCGA Data

Breast Cancer

- ☐ Gene Expression (GE) [16615 x 616]
- ☐ Copy Number (CN) [24174 x 616]
- ☐ Protein Exp. (RPPA) [187 x 616]
- ☐ Mutation Status (MU) [18256 x 616]



DIVAS on TCGA Data

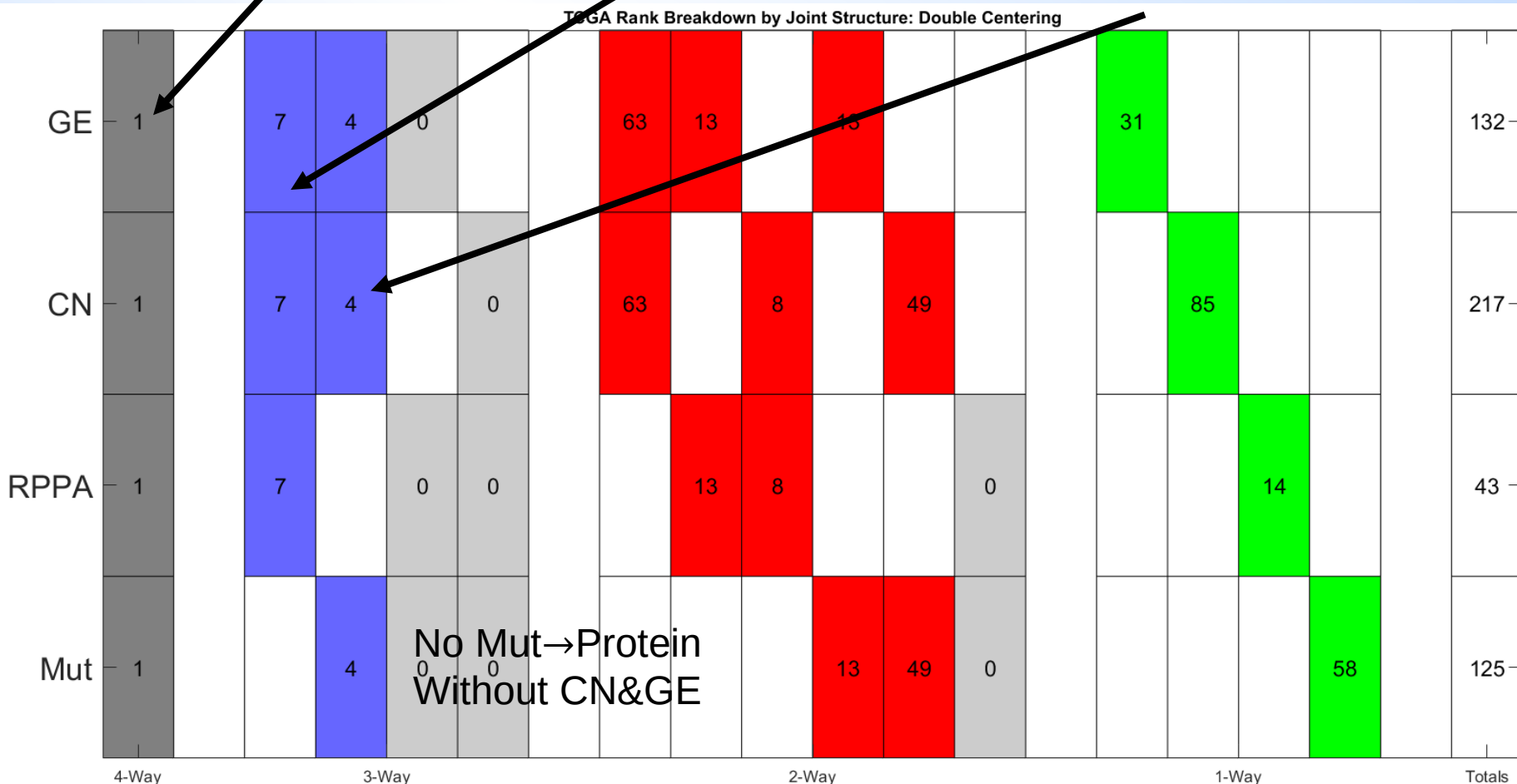
UNC, Stat & OR

Single 4-way mode, as in AJIVE Analysis

Statistical Significance Summary

CN→GE→Protein Common

Mut→CN→GE Sensible





DIVAS on TCGA Data

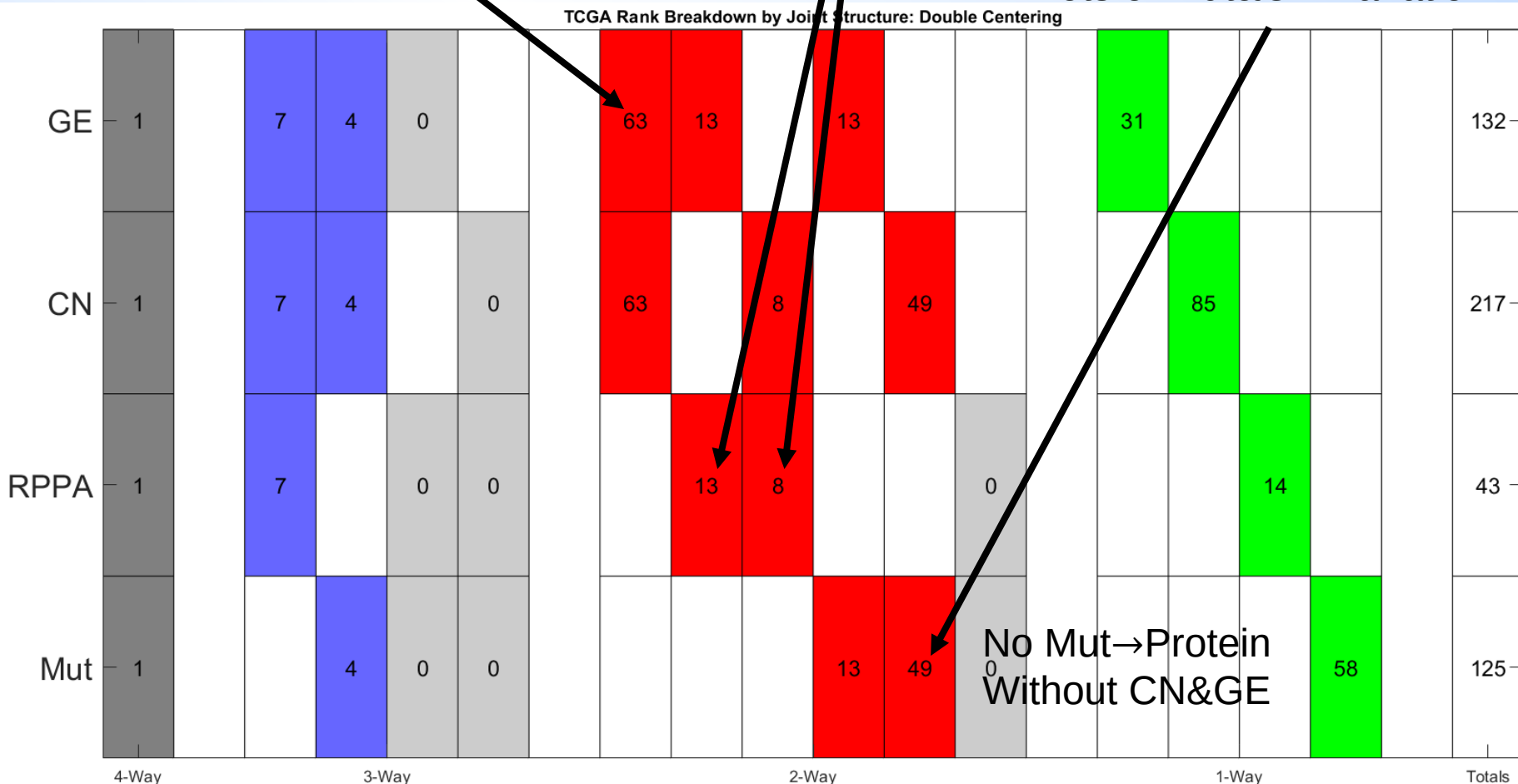
UNC, Stat & OR

Major CN→GE Variation

GE or CN→ Proteins

Statistical Significance Summary

Lots of Mut&CN Variation





Carry Away Concept

OODA is more than a “framework”

It Provides a Focal Point

Highlights Pivotal Choices:

What should be the Data Objects?

How should they be Represented?