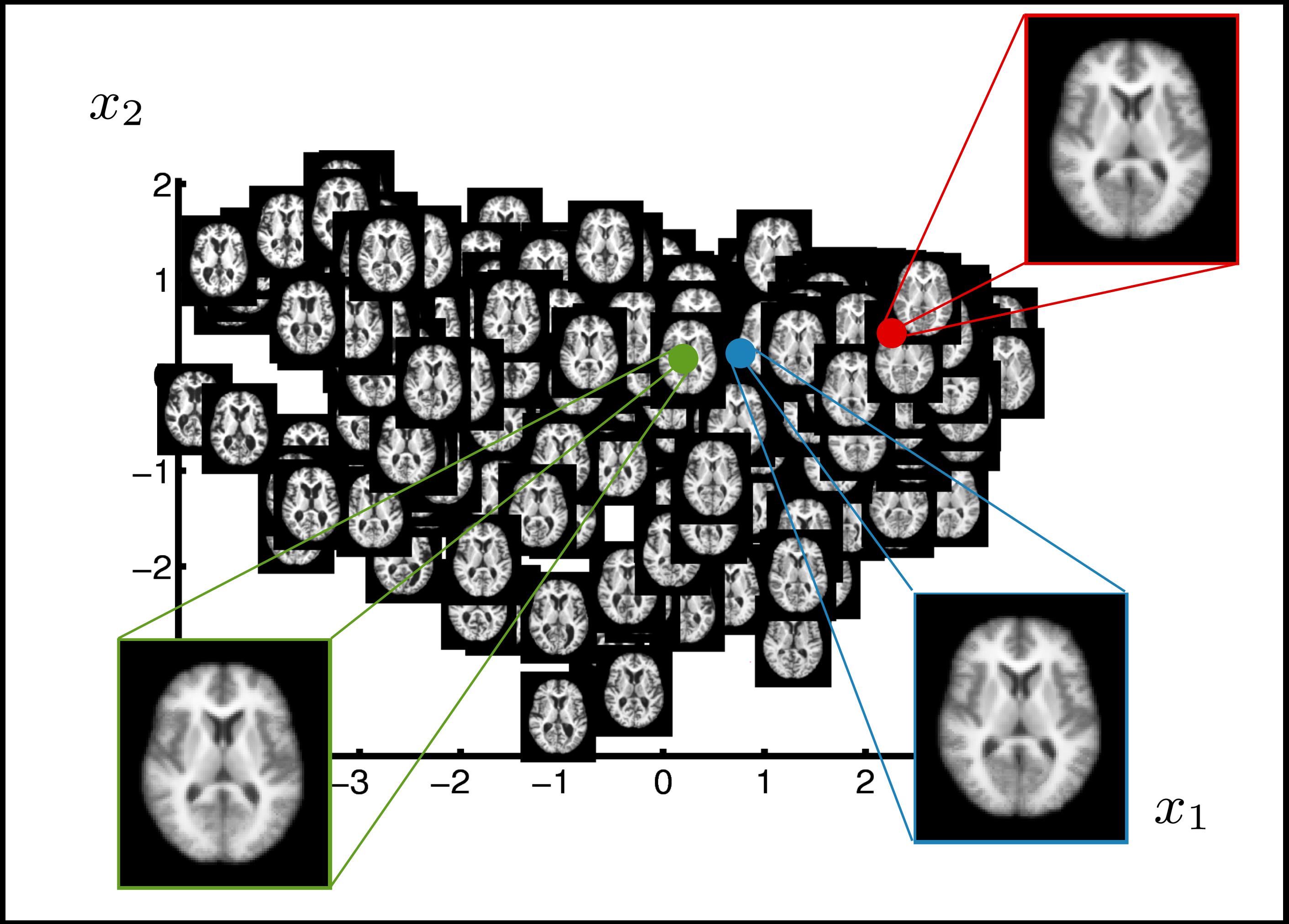
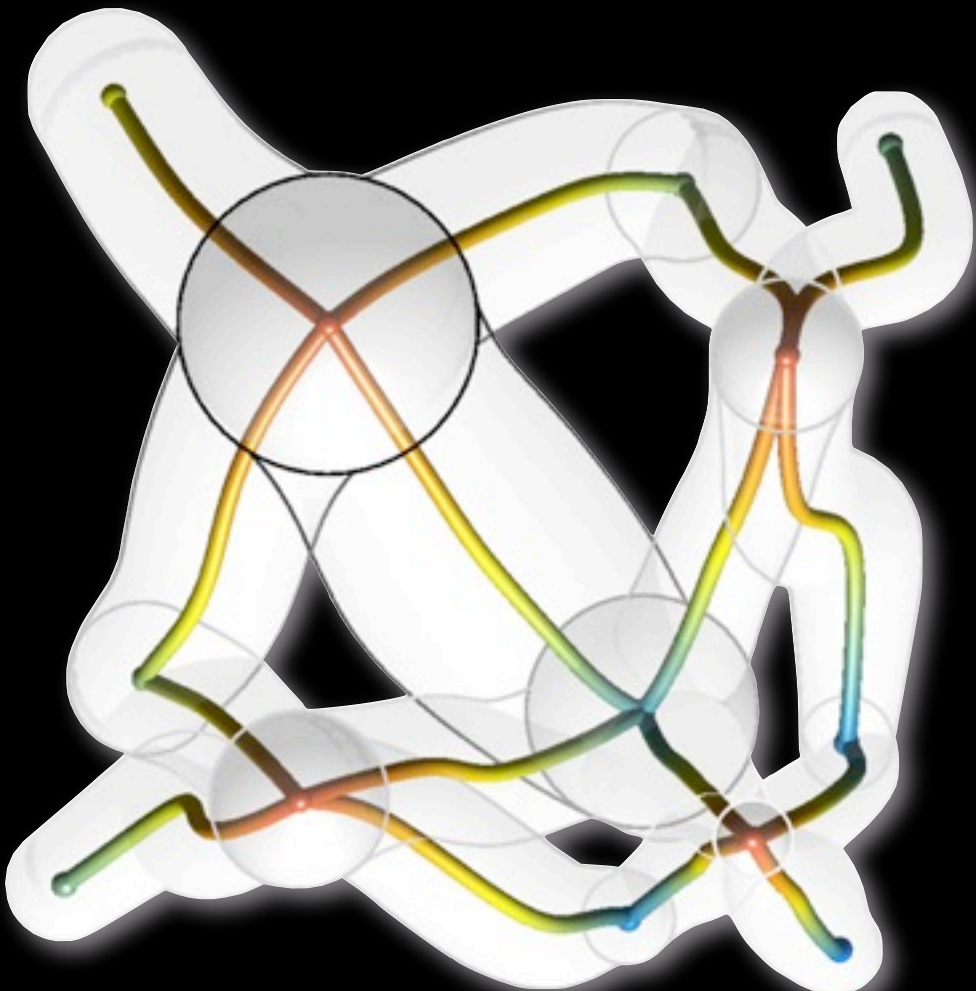
Dimension Reduction

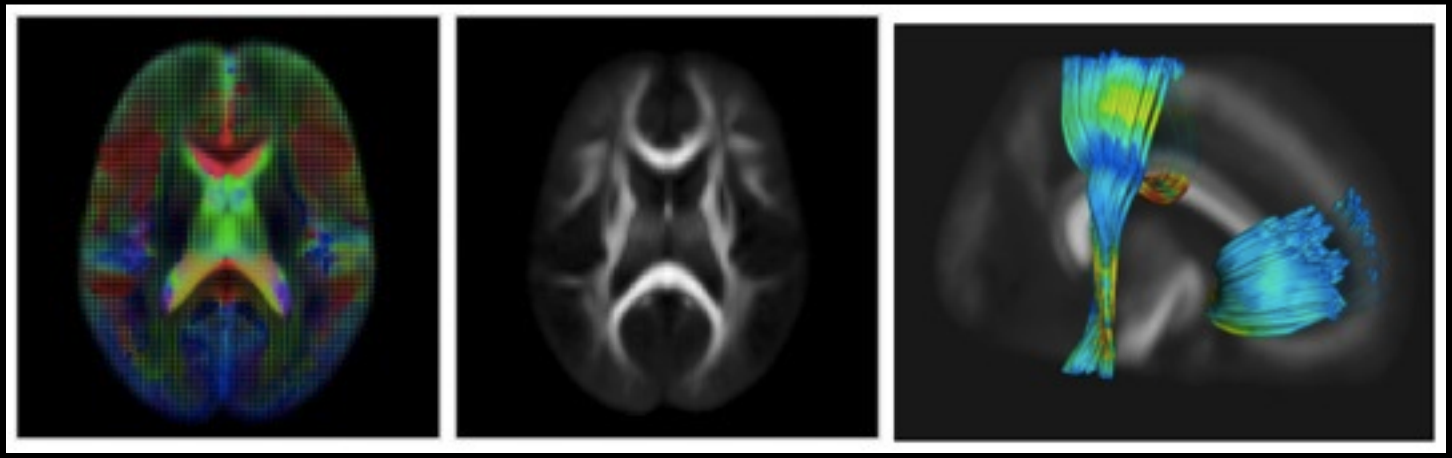




Samuel Gerber, University of Utah

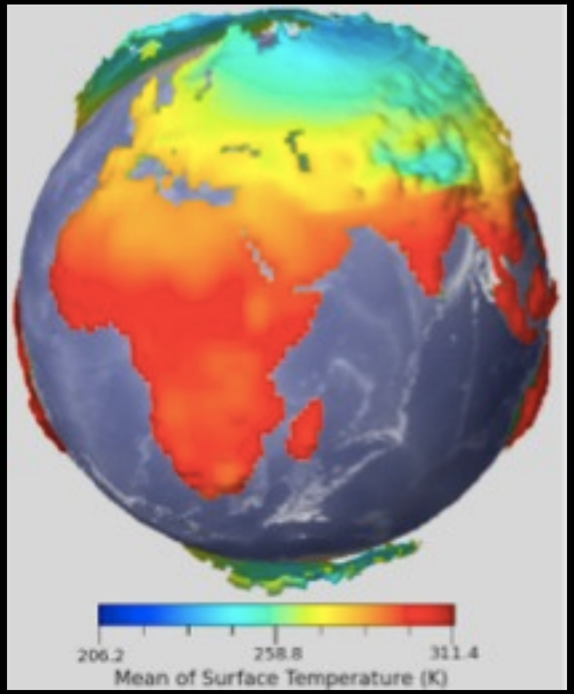
SCI INSTITUTE

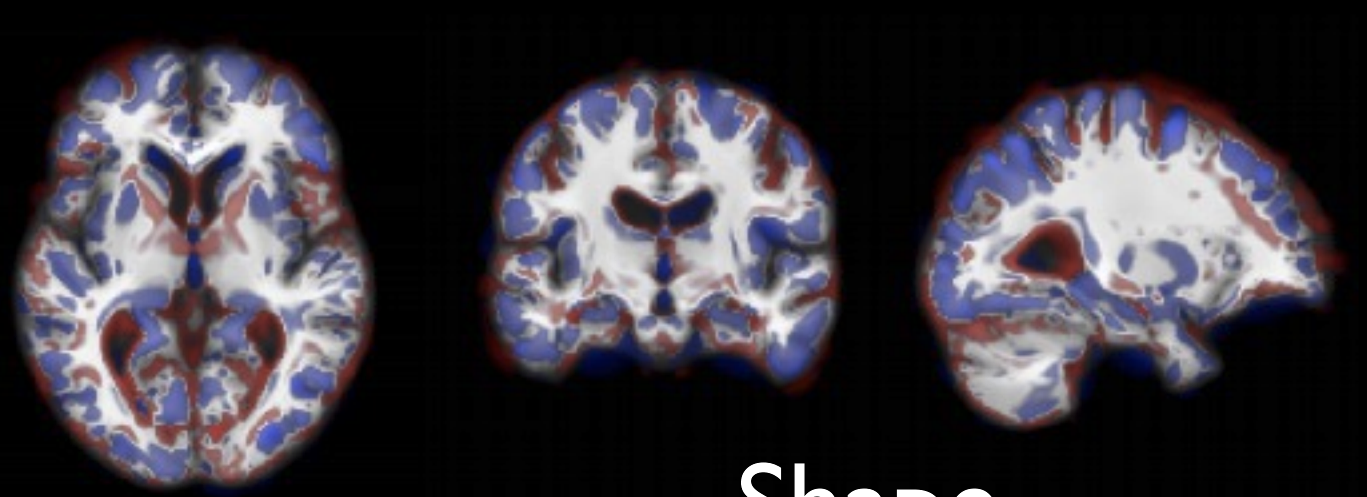
High-Dimensional Data



Diffusion Tensor Images

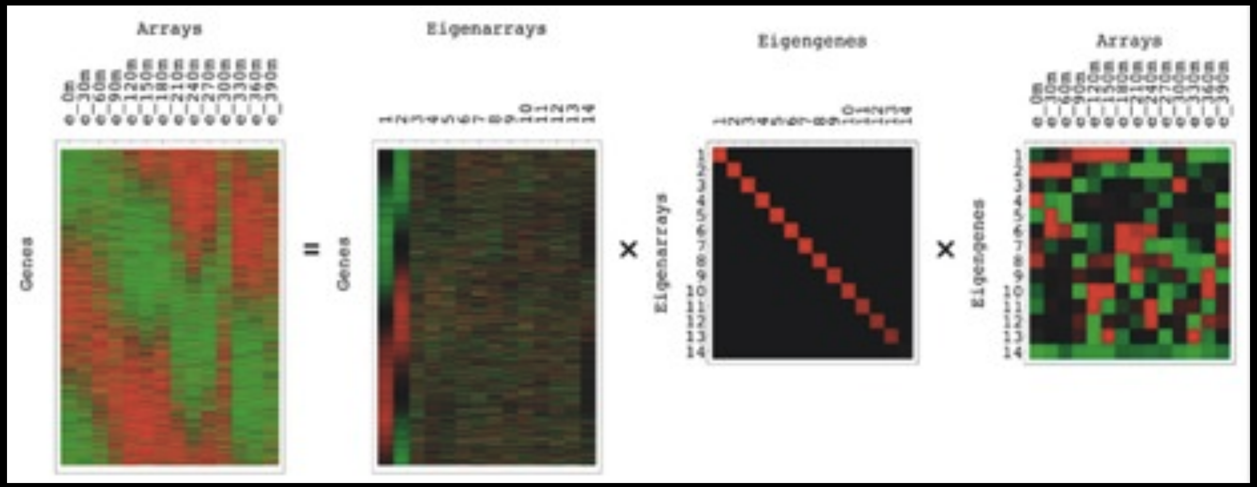
Simulations





Shape

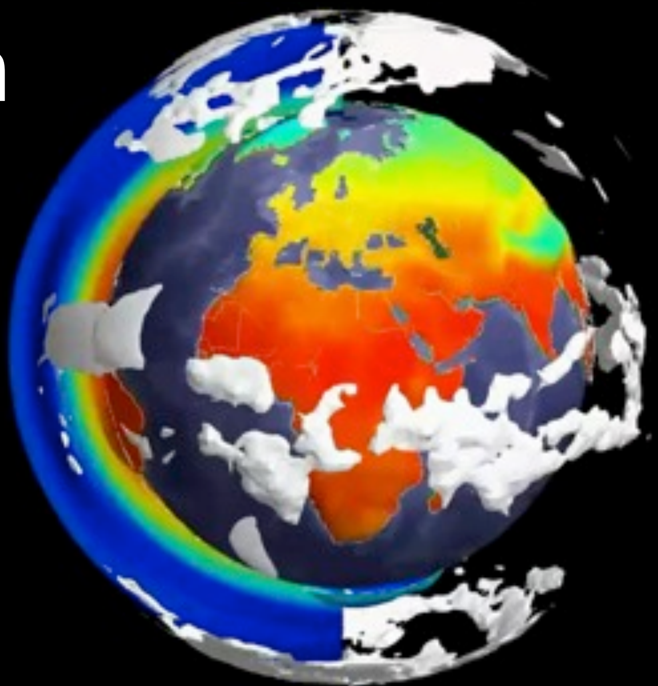
Gene Expression



High-Dimensional Data

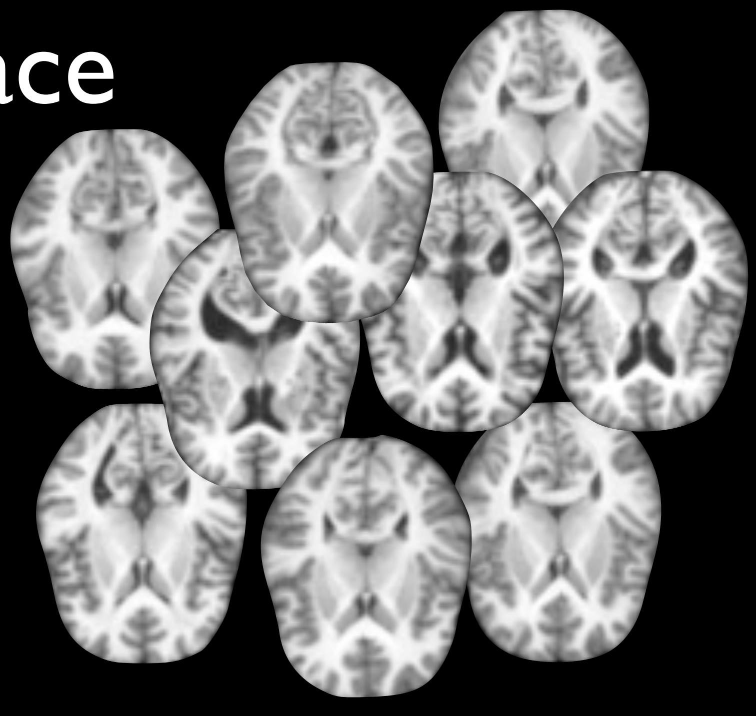
- Supervised: Characterize the function

$$y_i = f(\mathbf{x}_i) \qquad \mathbf{x} \in \mathbf{R}^d$$



- Unsupervised: Characterize the space

$$\mathbf{x}_i \qquad \mathbf{x} \in \mathbf{R}^d$$

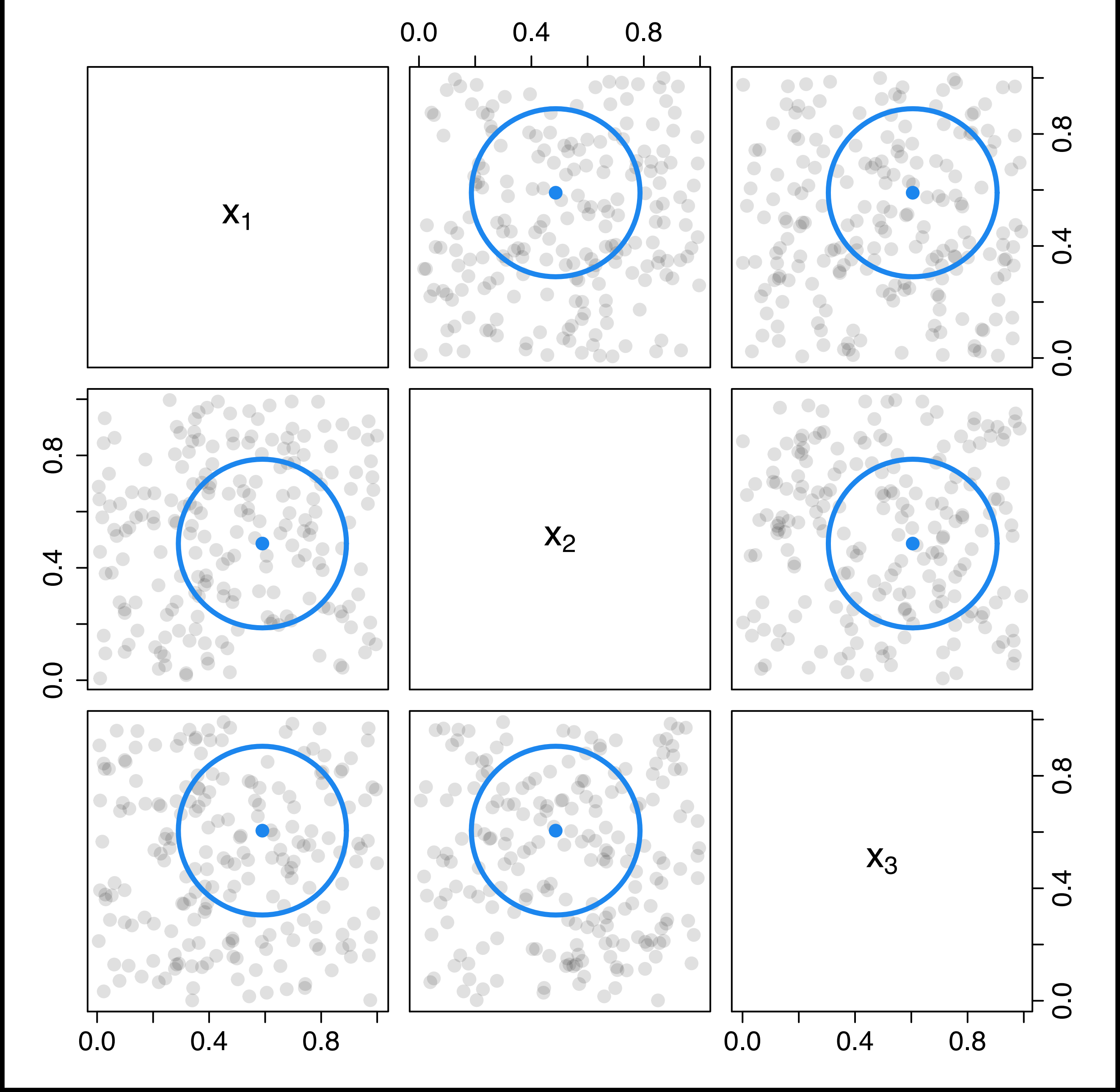


Exploratory Analysis

- Feedback about the structure of the data
- Supports building and validating models
- Formation of new hypotheses about process/system that generated the data

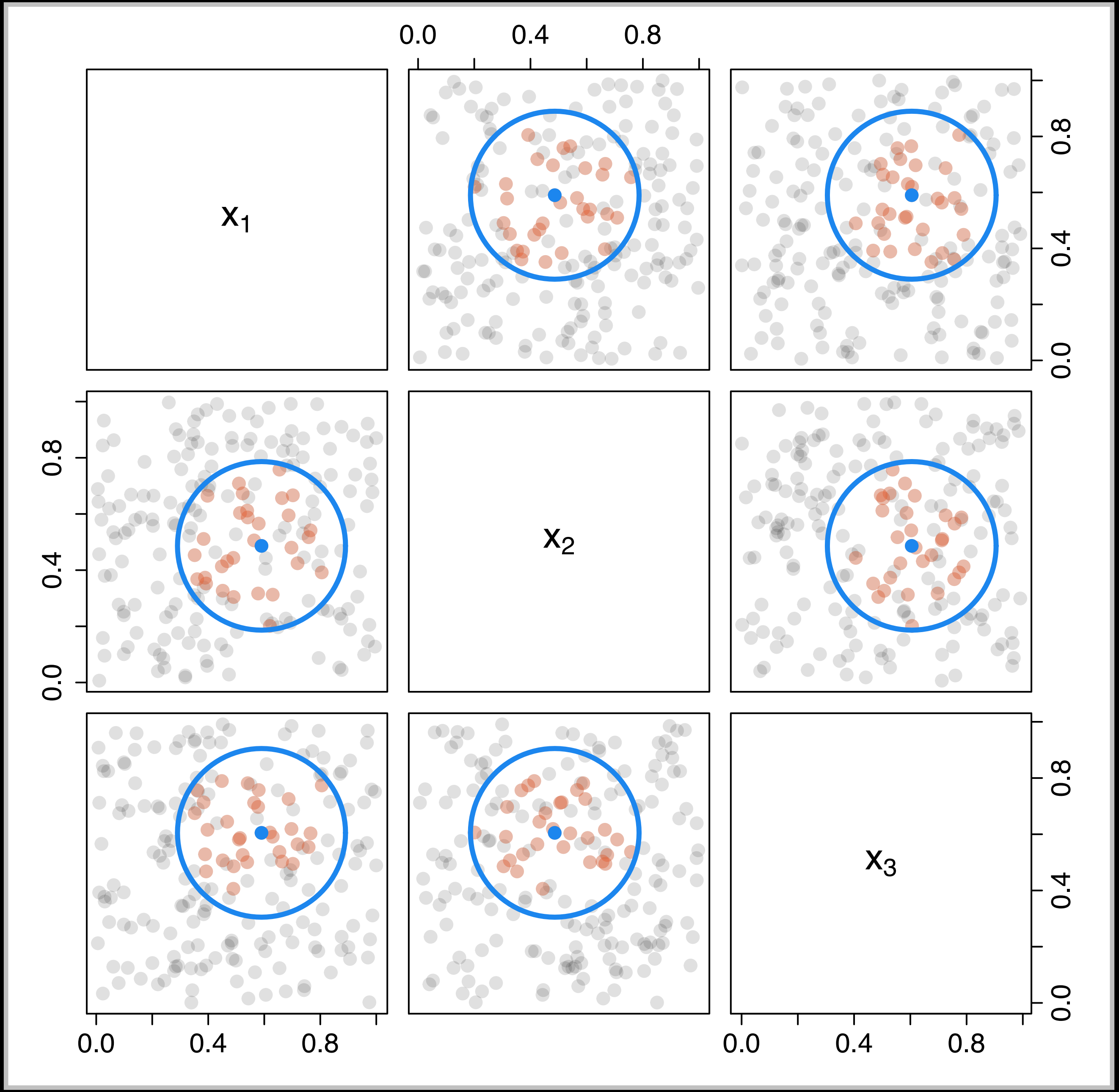
Exploratory Analysis in High-D

- Plotting of raw data or projections



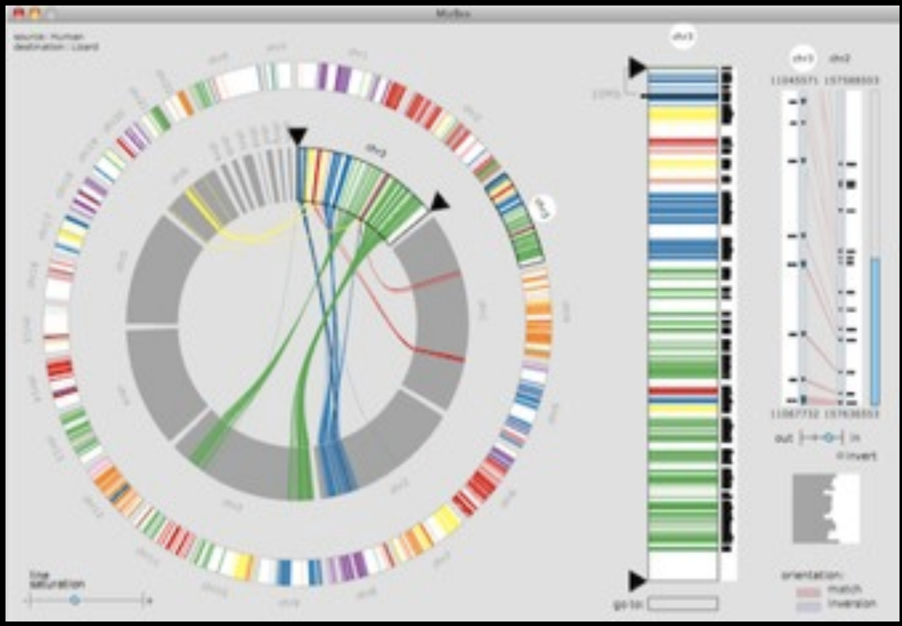
Exploratory Analysis in High-D

- Plotting of raw data or projections

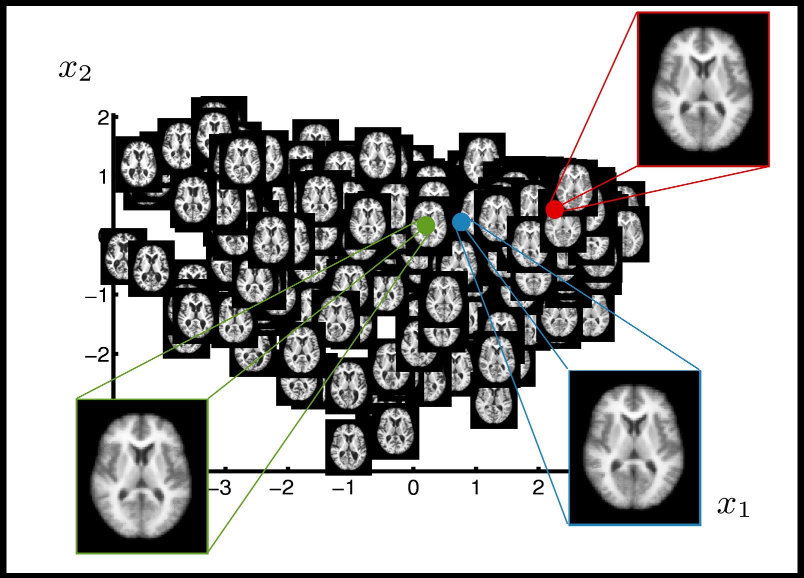


Exploratory Analysis in High-D

- High-dimensional data requires methods that provide a holistic view
- Two (not exclusive) approaches:
  - Clever visualizations of raw data

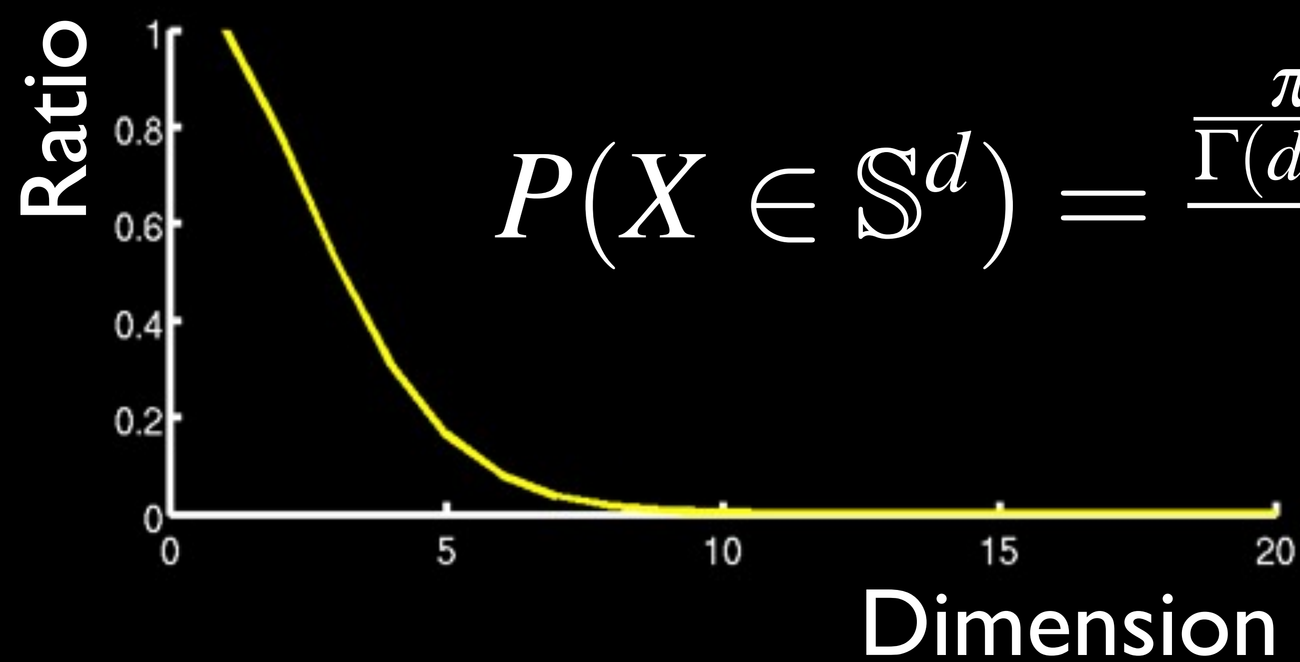


  - *Summary* representations

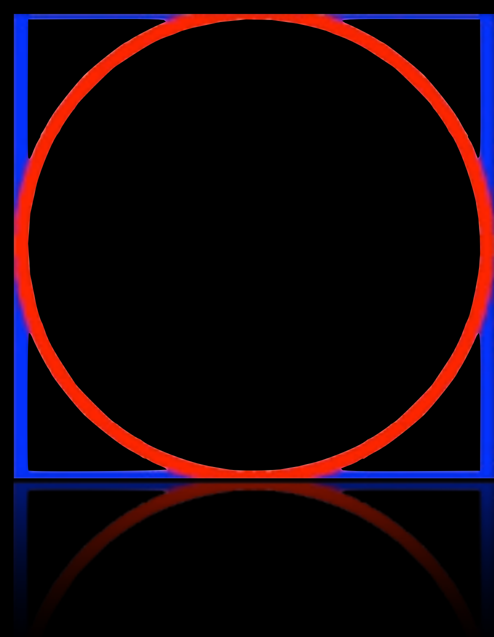


Curse of Dimensionality

- Volume increases exponentially with dimensionality
- Example: Ratio of (hyper) sphere to cube volume



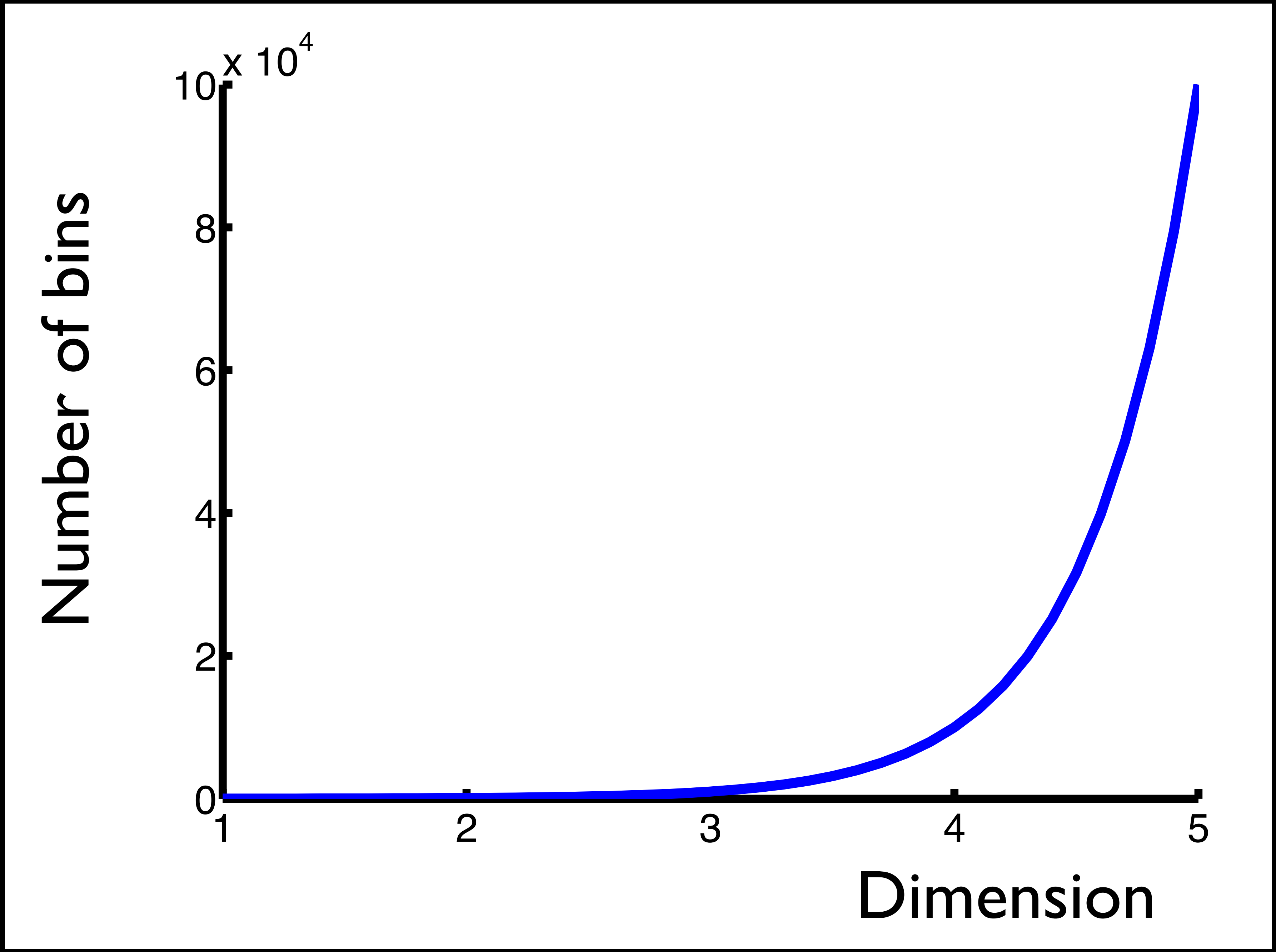
$$P(X \in \mathbb{S}^d) = \frac{\frac{\pi^{d/2}}{\Gamma(d/2+1)}}{2^d}$$



- $\lim_{d \to \infty} P(X \in \mathbb{S}^d) = 0$

Curse of Dimensionality

- Example: Density estimation, histogram with 10 bins per dimension



Curse of Dimensionality

- Bellman 1961: Sample size grows exponentially with dimensionality for approximation of functions.
- Barron 1993: Sample size independent of dimensionality with additional *regularity* conditions
- Complexity limited by sample size
  - Restriction to *simple* functions (supervised)
  - Reduce dimensionality of domain (un-/supervised)

Bellman, R. **Adaptive control processes: a guided tour.** 1961

Barron, A. R. **Universal approximation bounds for superpositions of a sigmoidal function.** 1993

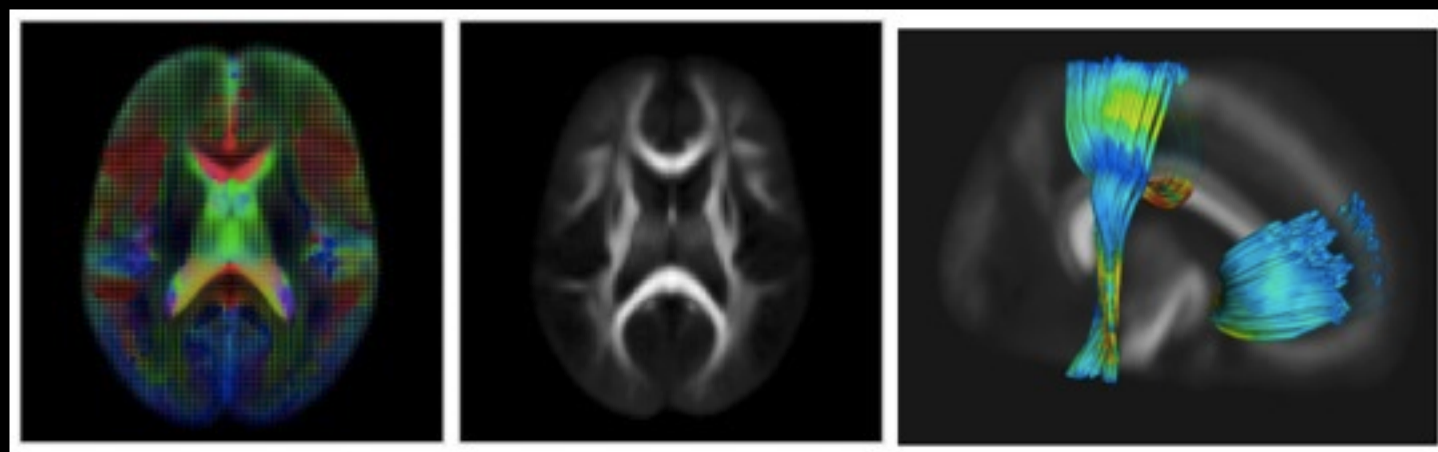
Curse of Dimensionality

- Bellman 1961: Sample size grows exponentially with dimensionality for approximation of functions.
- Barron 1993: Sample size independent of dimensionality with additional *regularity* conditions
- Complexity limited by sample size
  - Restriction to *simple* functions (supervised)
  - Reduce dimensionality of domain (un-/supervised)

Bellman, R. **Adaptive control processes: a guided tour.** 1961

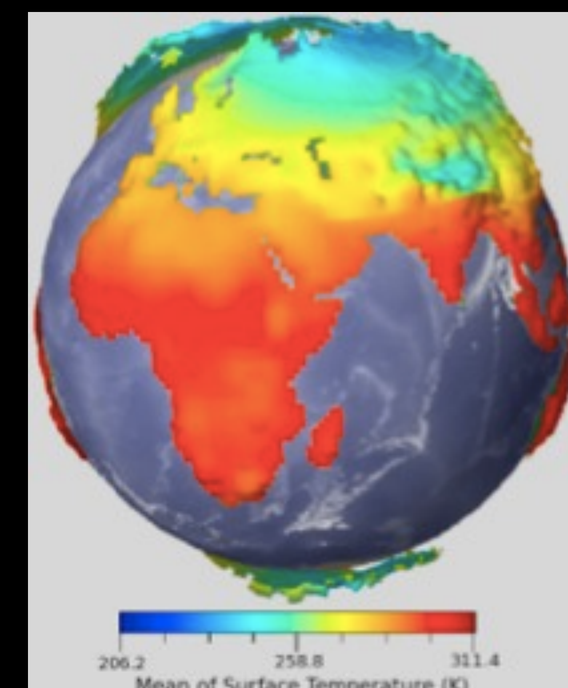
Barron, A. R. **Universal approximation bounds for superpositions of a sigmoidal function.** 1993

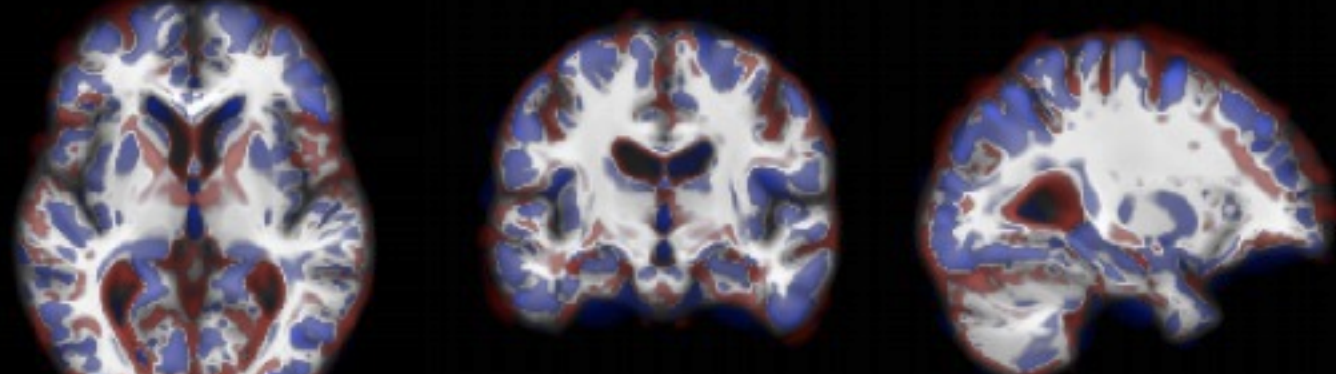
High-Dimensional Data



DTI - Constrained PCA

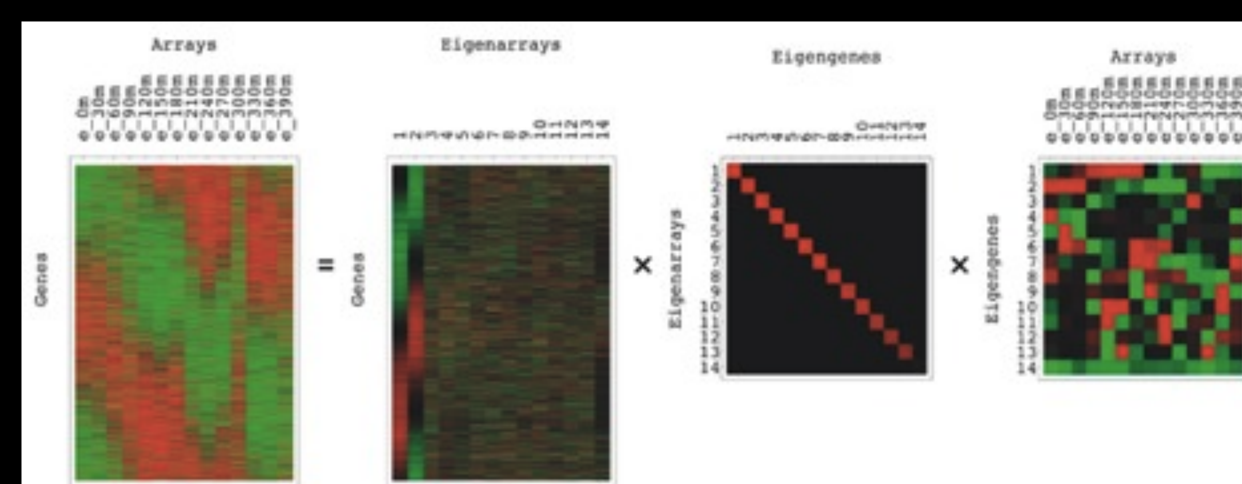
Simulations





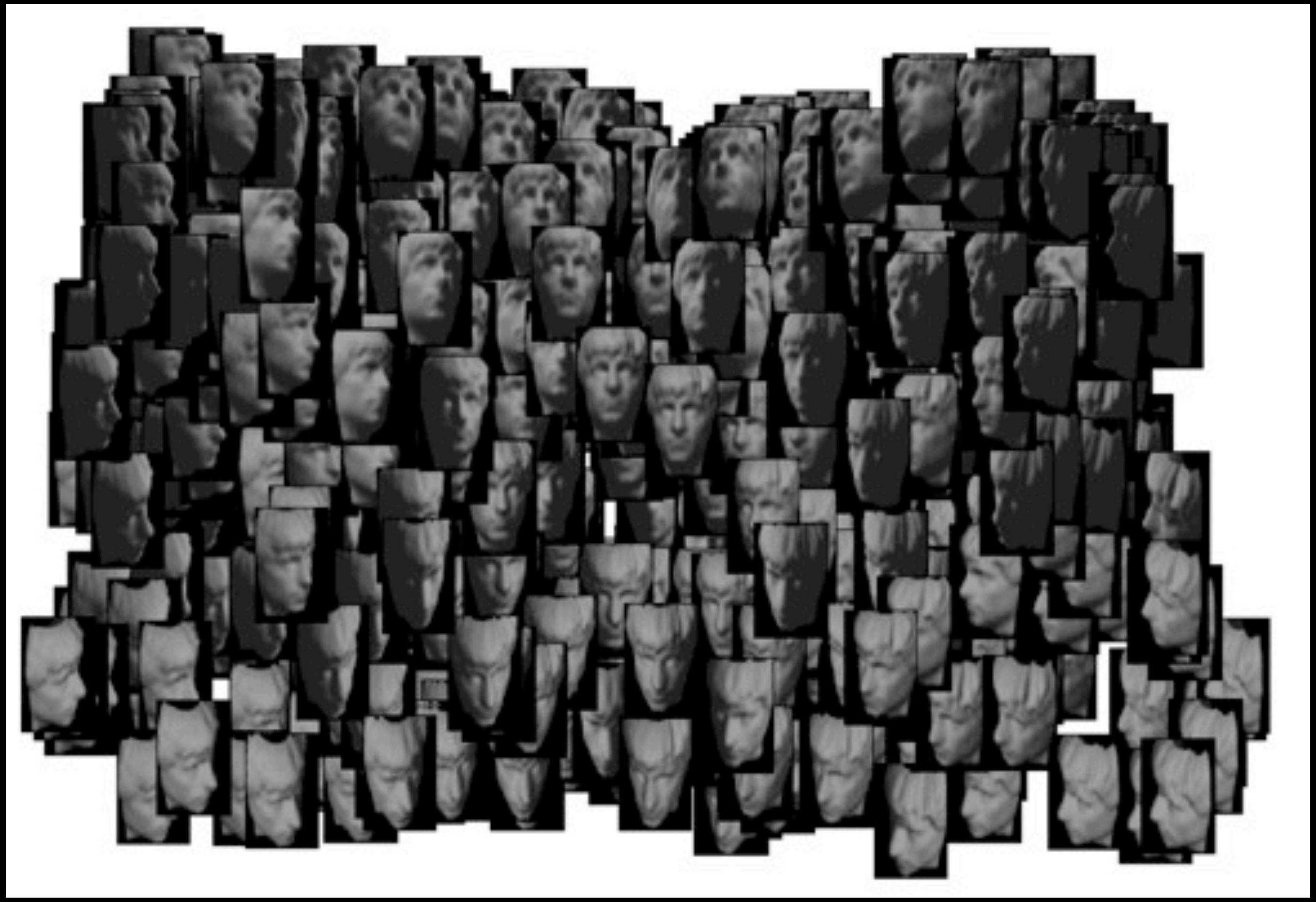
Shape - Partial least squares
(Linear subspace)

Genes - PCA



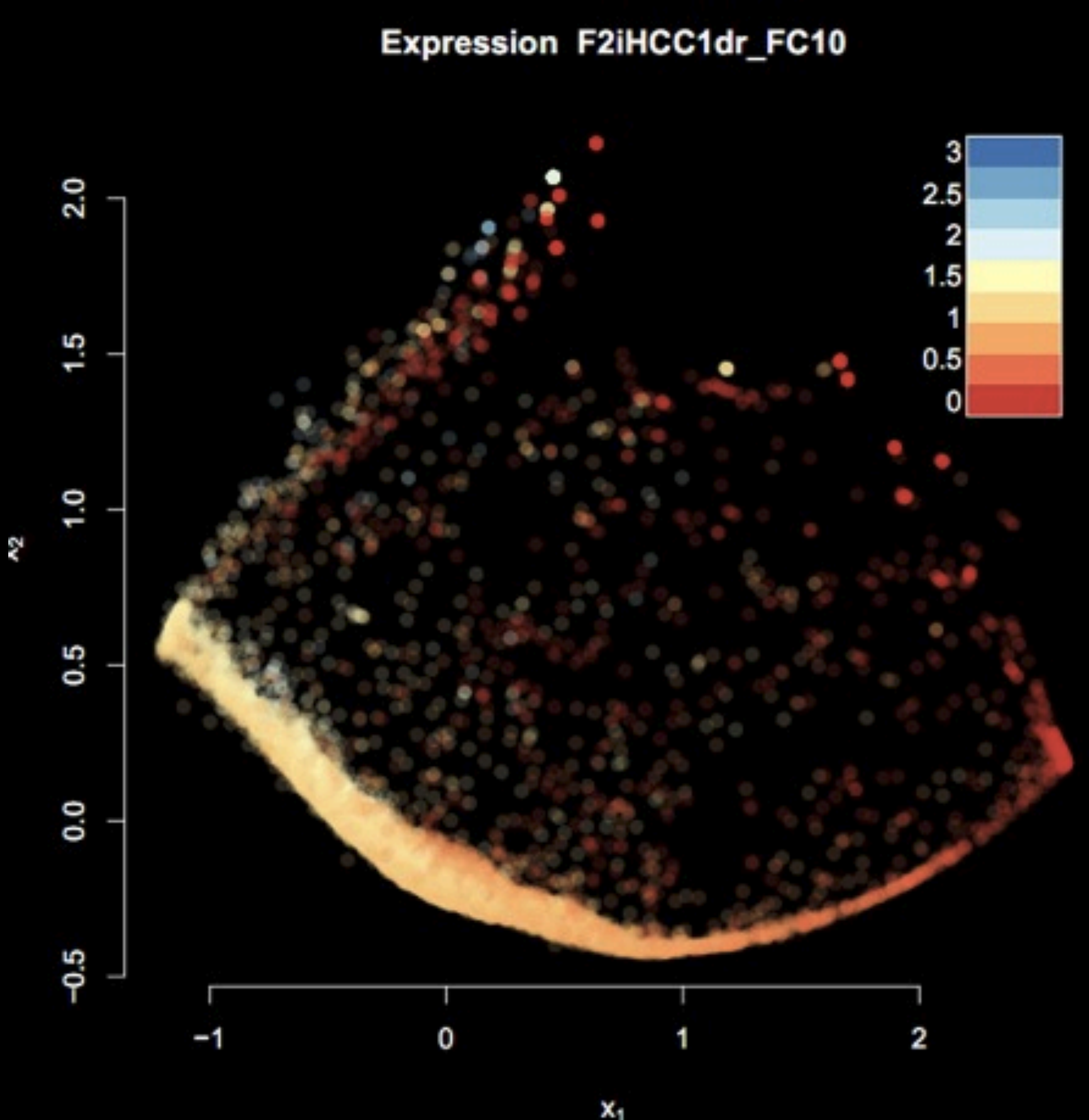
Dimension Reduction

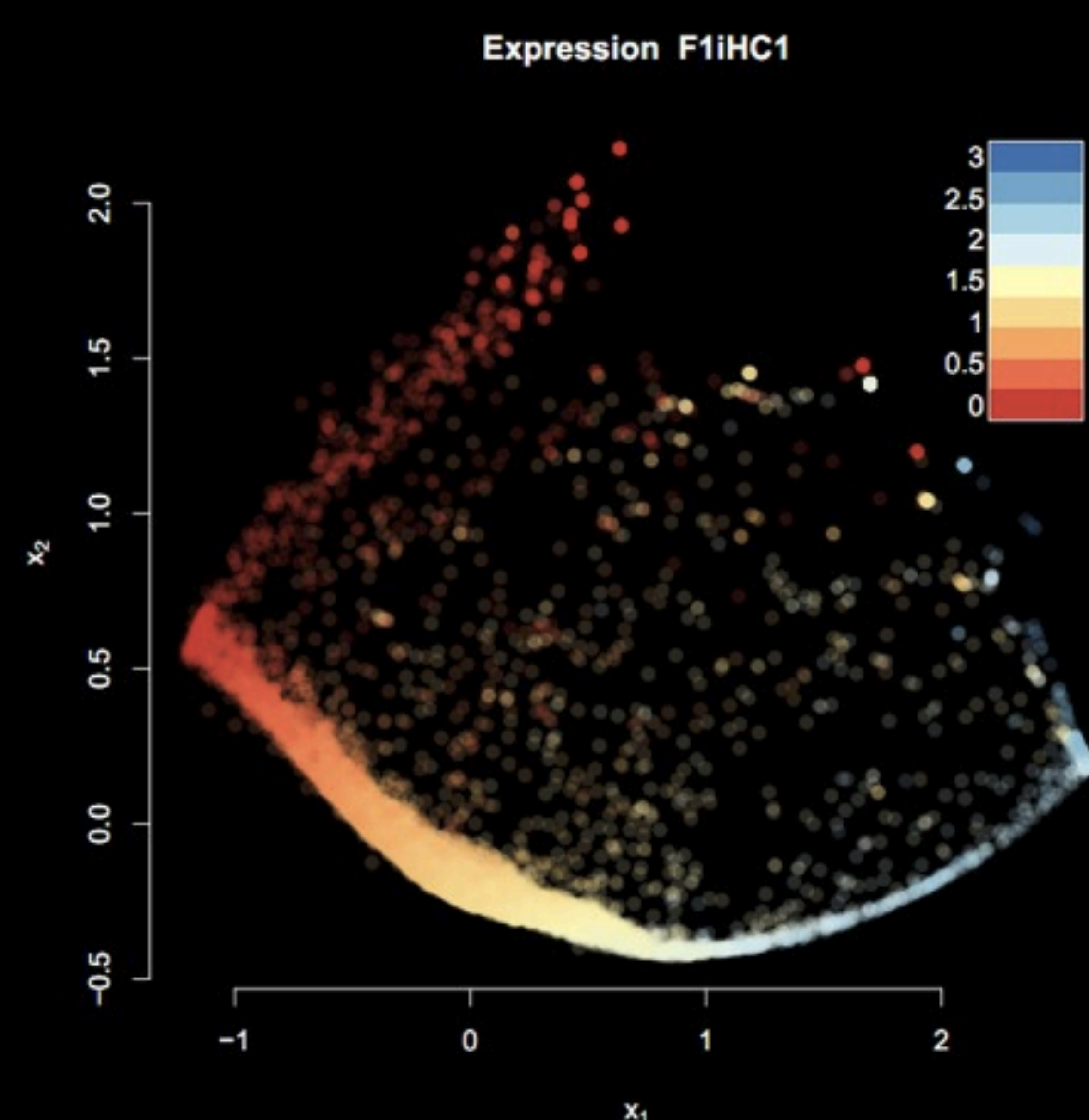
- Dimensionality of observation $\neq$ Variability
- Find low-dimensional representation of variability



Use Cases for Vis

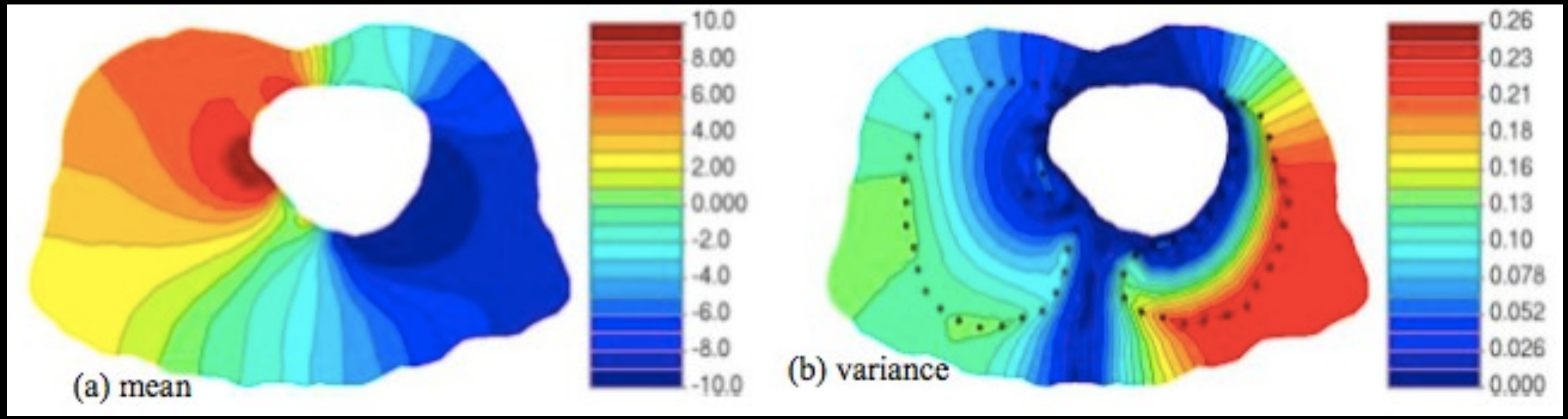
- Scaffolding for an interactive exploratory visualization

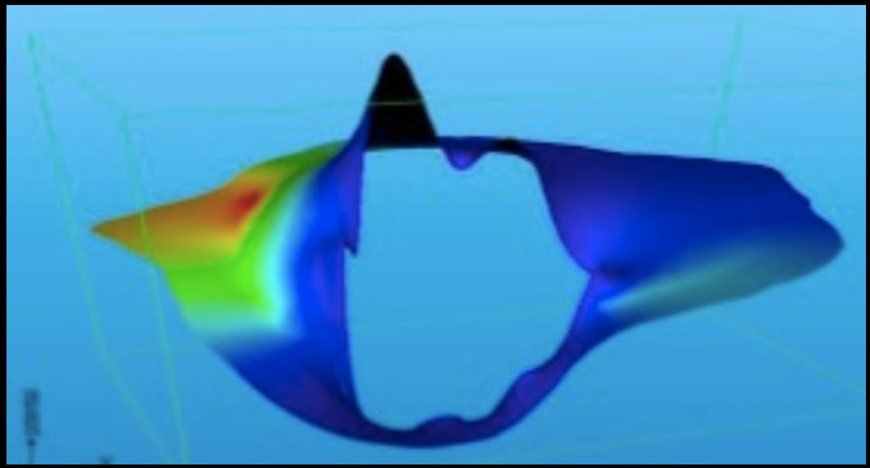




Use Cases for Vis

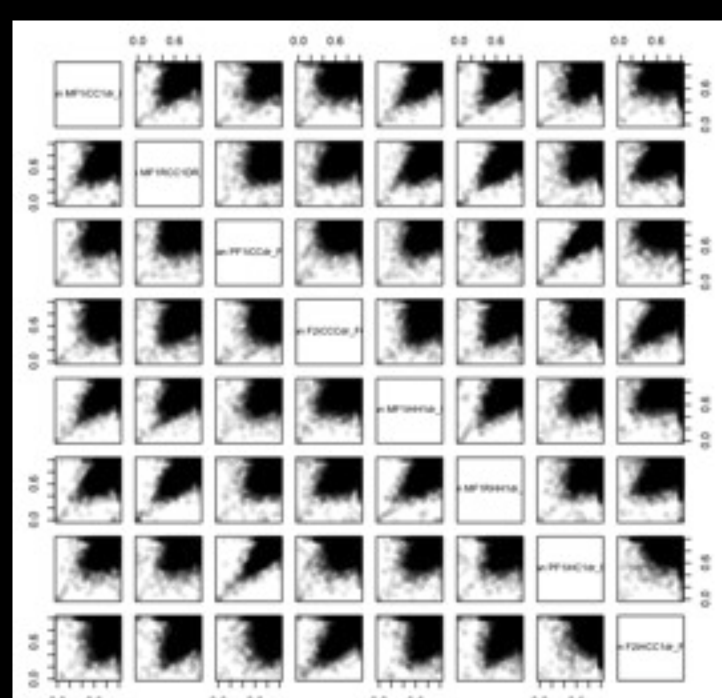
- Hot topic: Uncertainty visualization
  - Distribution over possible data realizations / visualizations
- Typical framework. Encode a *most likely* solution and augment with *variability*
  - Major challenge: Perceptual efficient encoding of appropriate summary statistics of the data set

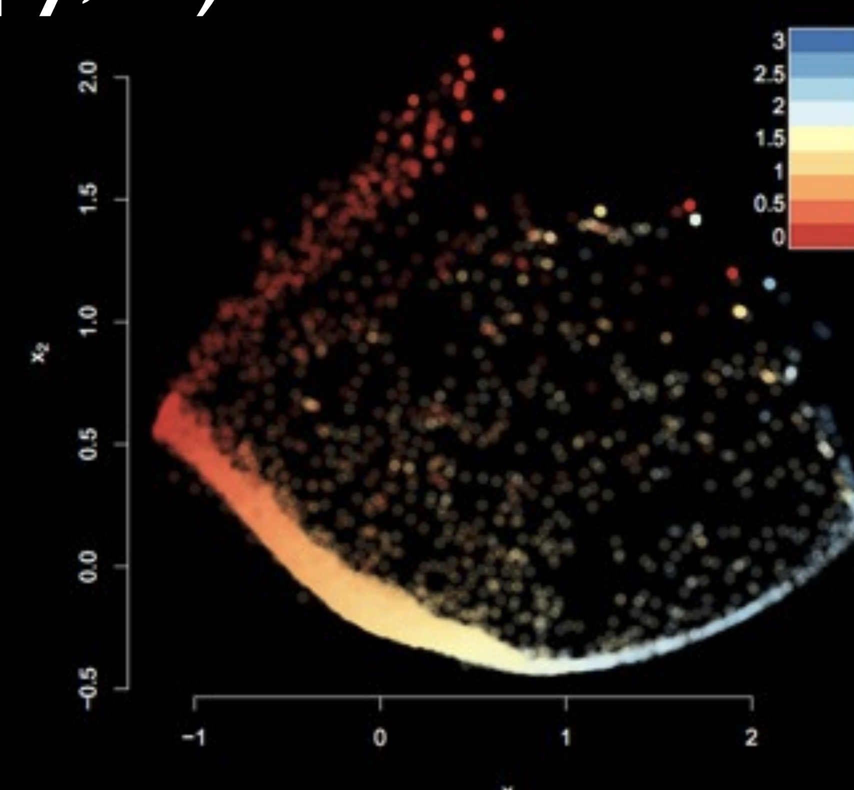




Dimension Reduction

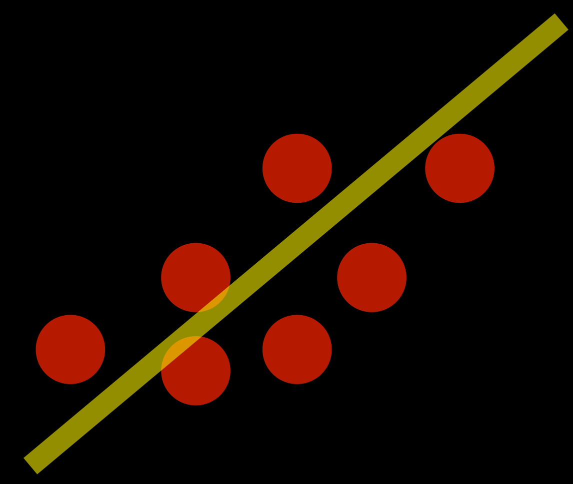
- Reduce m-dimensional data set Y to n-dimensional representation X.
- Preserve *information* in Y in reduced space X
  - Geometry (distances, angles, ...)
  - Statistics (density, entropy, ...)



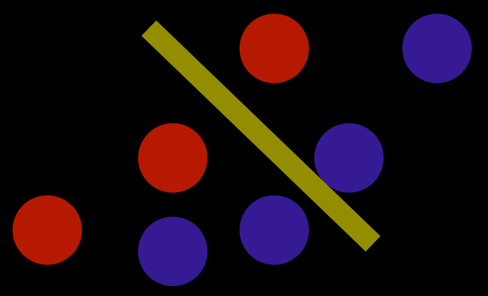


Some Linear Methods

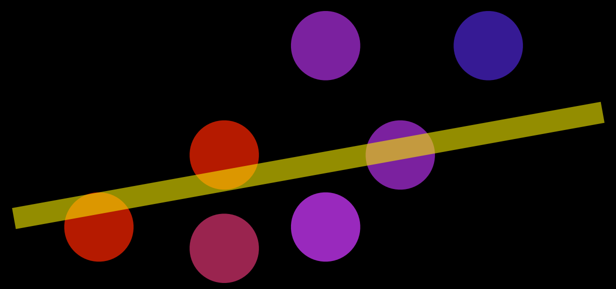
- Principal Component Analysis:
  - Preserve variance / Minimize orthogonal distance to data points



- Linear Discriminant Analysis:
  - Discrimination between groups

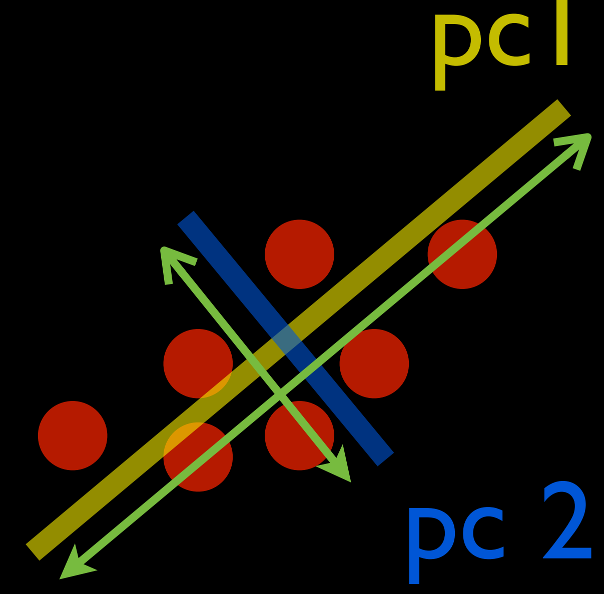


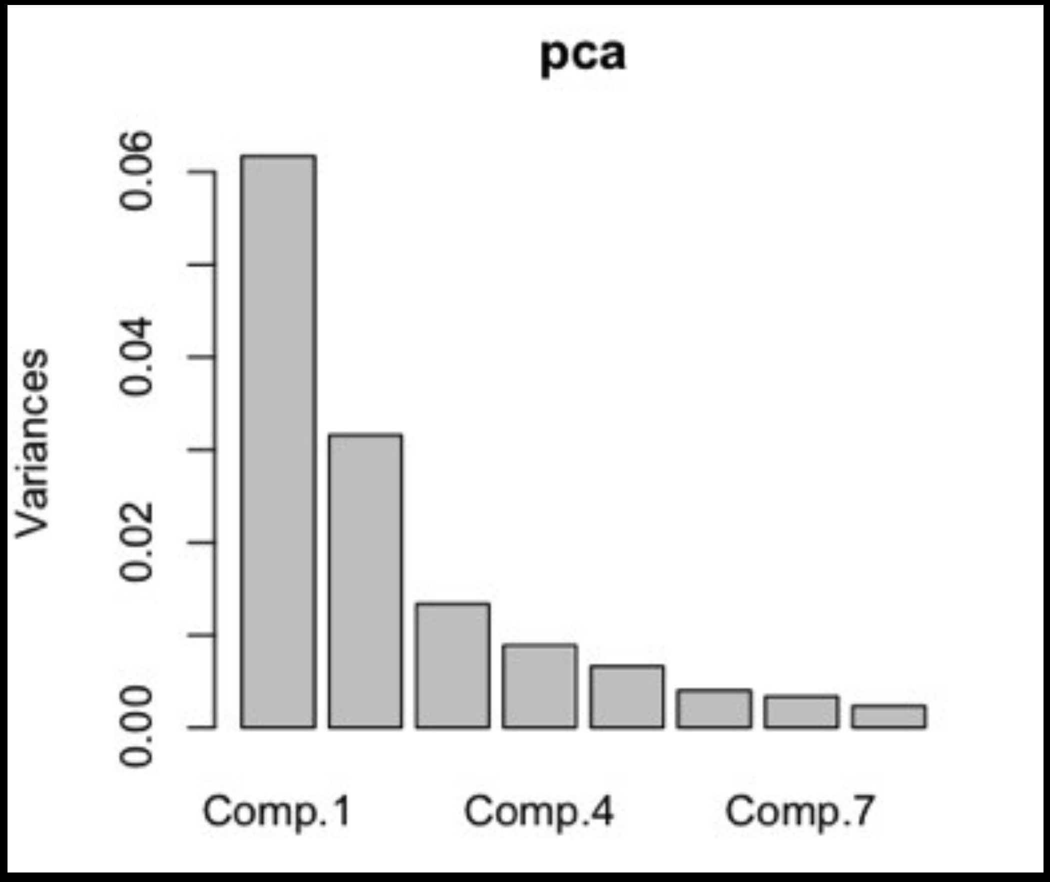
- Sliced inverse regression:
  - Linear prediction



Principal Components

- Closed from solution
  - Eigen-decomposition of covariance matrix
  - Set of orthogonal vectors (principal components)
- Eigenvalues are the amount of variance in each component

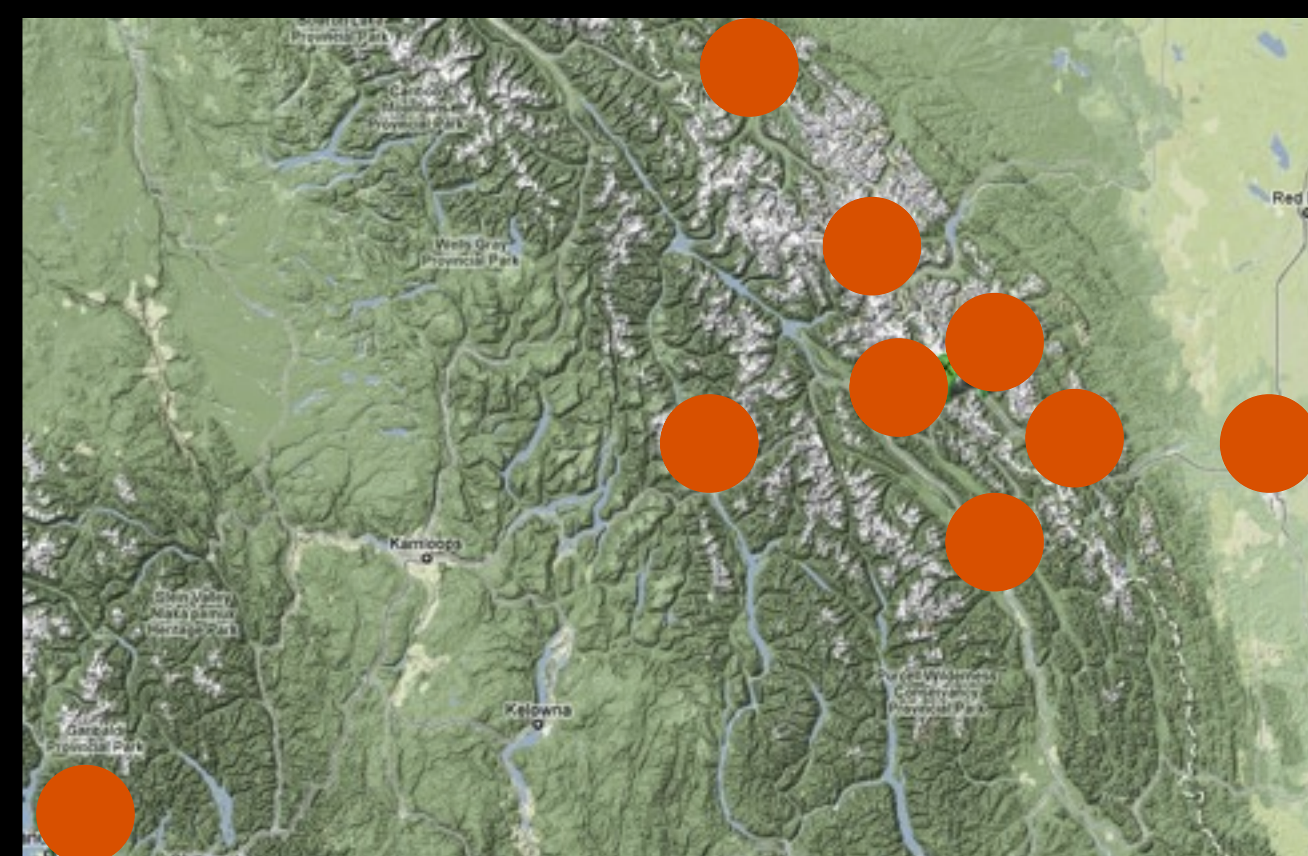




Multidimensional Scaling

- Input: N pairwise distances
- Finds configuration of N points in low dimensional space such that pairwise distance are preserved

	Banff	Calgary	Columbia Icefield	Edmonton	Field, B.C.	Jasper	Lake Louise	Radium Hot Springs	Golden	Revelstoke
Calgary	128									
Columbia Icefield	188	316								
Edmonton	423	295	461							
Field, B.C.	85	213	157	508						
Jasper	291	419	100	361	260					
Lake Louise	58	186	130	481	27	233				
Radium Hot Springs	132	260	261	555	157	361	130			
Golden	134	262	207	557	49	307	76	105		
Revelstoke	282	410	355	705	197	455	224	253	148	
Vancouver	856	984	928	1279	771	798	794	818	713	565

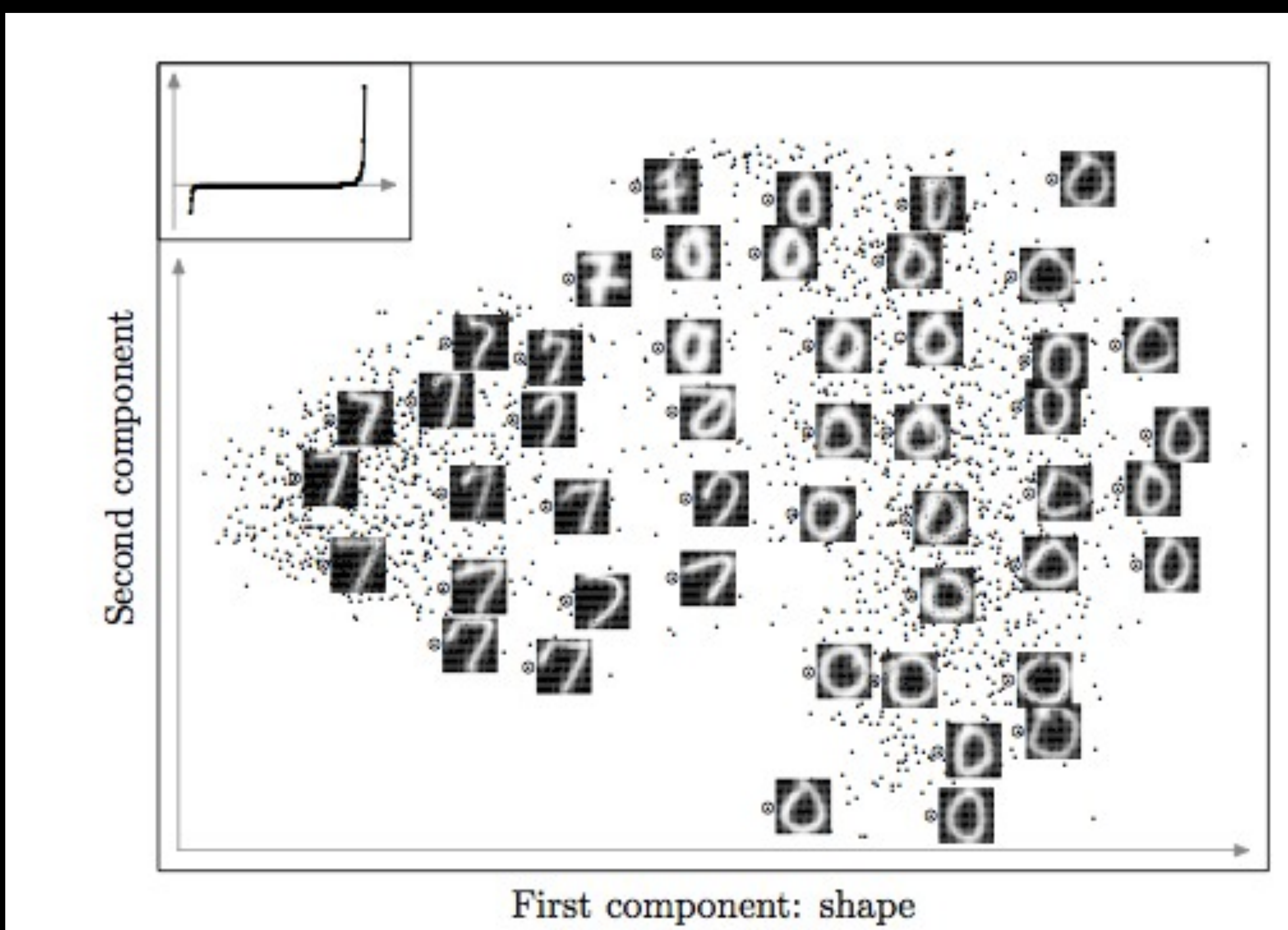


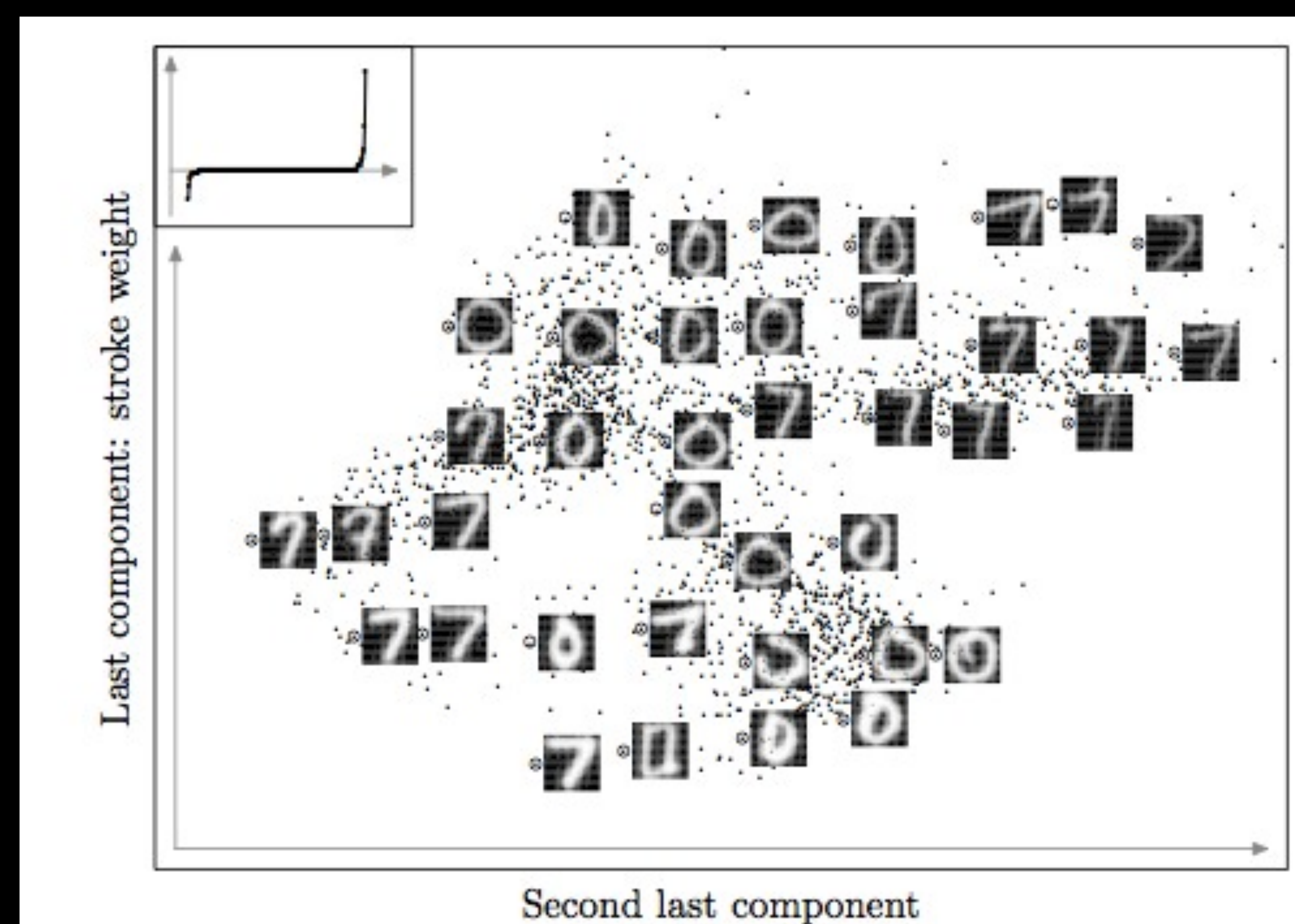
Multidimensional Scaling

- What if there is no good set of points?
  - Distortion
- Many algorithms that emphasize differently
  - Small versus large distances
- Classical MDS - Closed form solution
  - Equivalent to PCA if distances from points in Euclidean space

Classical MDS

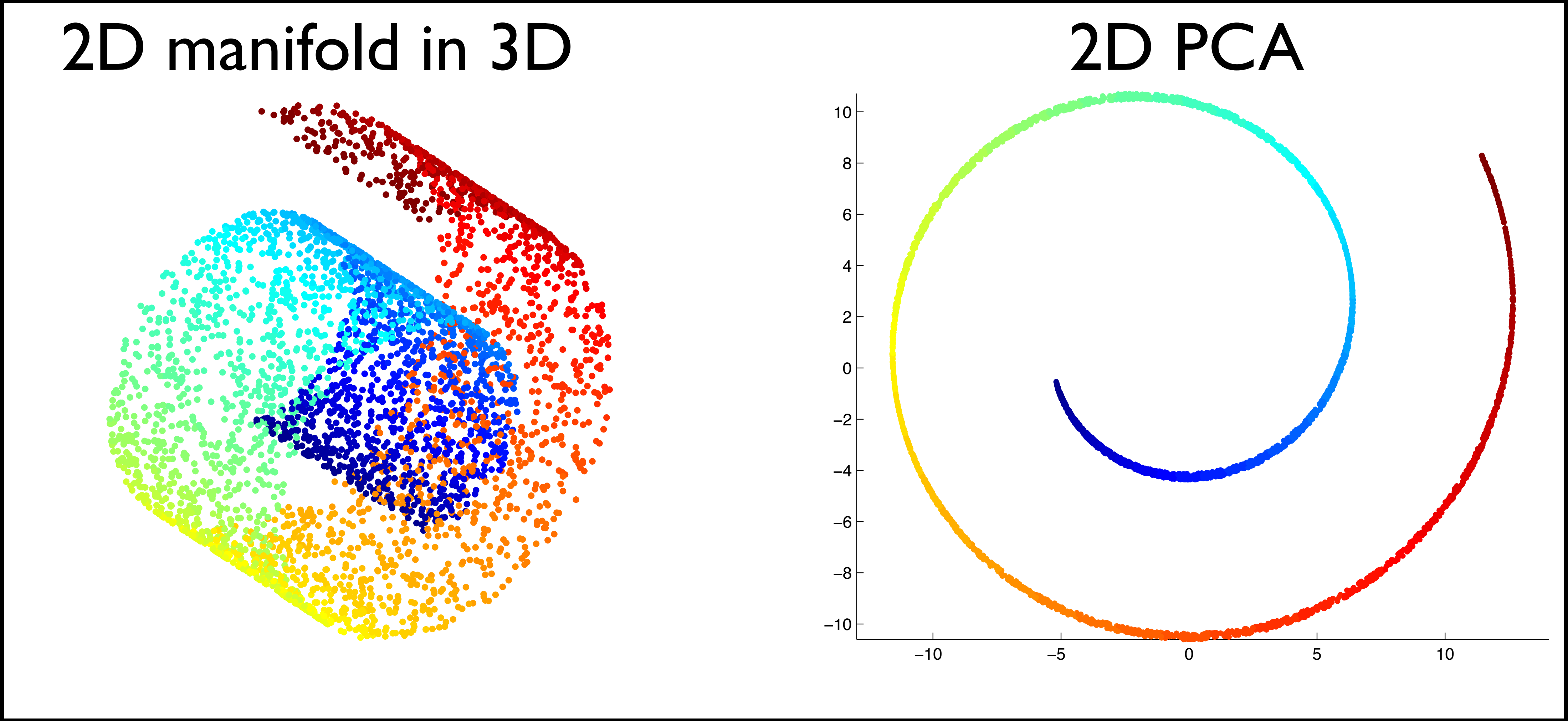
- Possible negative eigenvalues
  - Not embeddable in Euclidean space
  - Can contain important information





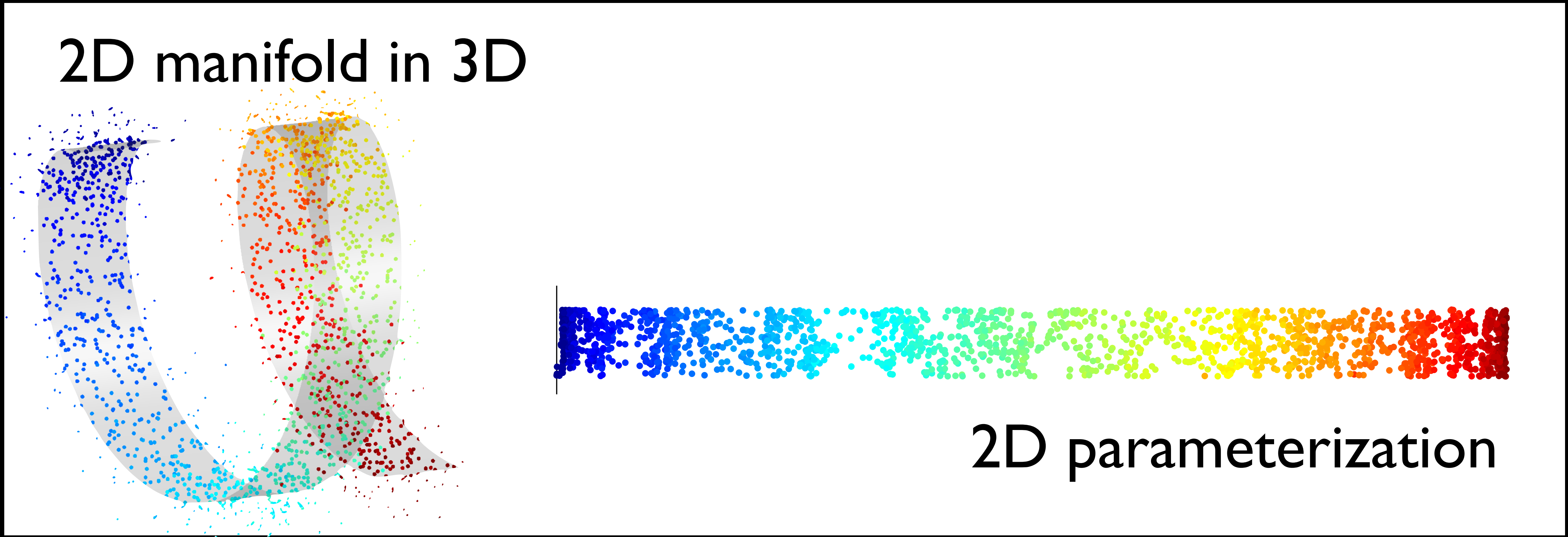
Nonlinear Structure

- PCA "fails" for data sets with non-linear structure

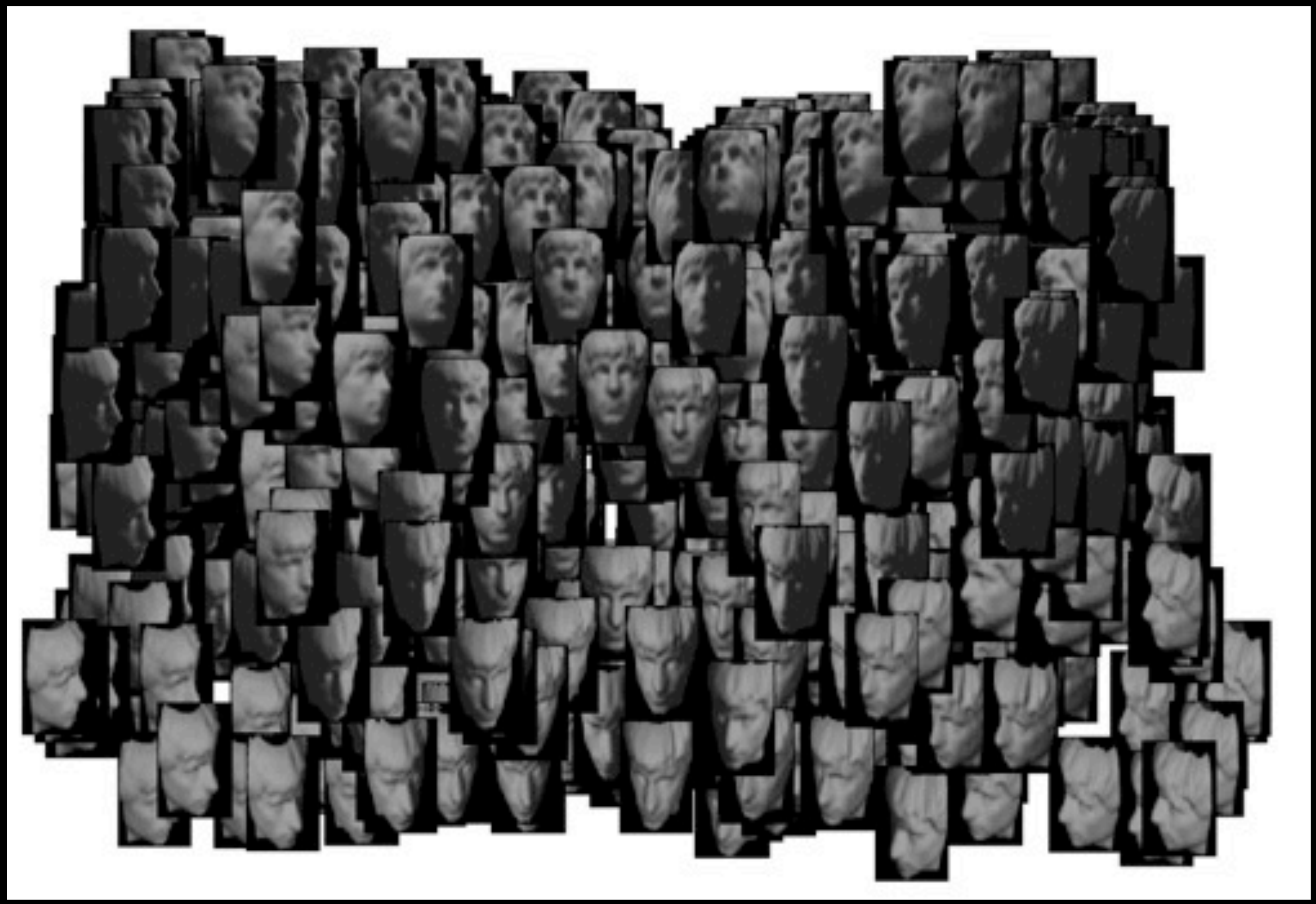


Manifold Learning

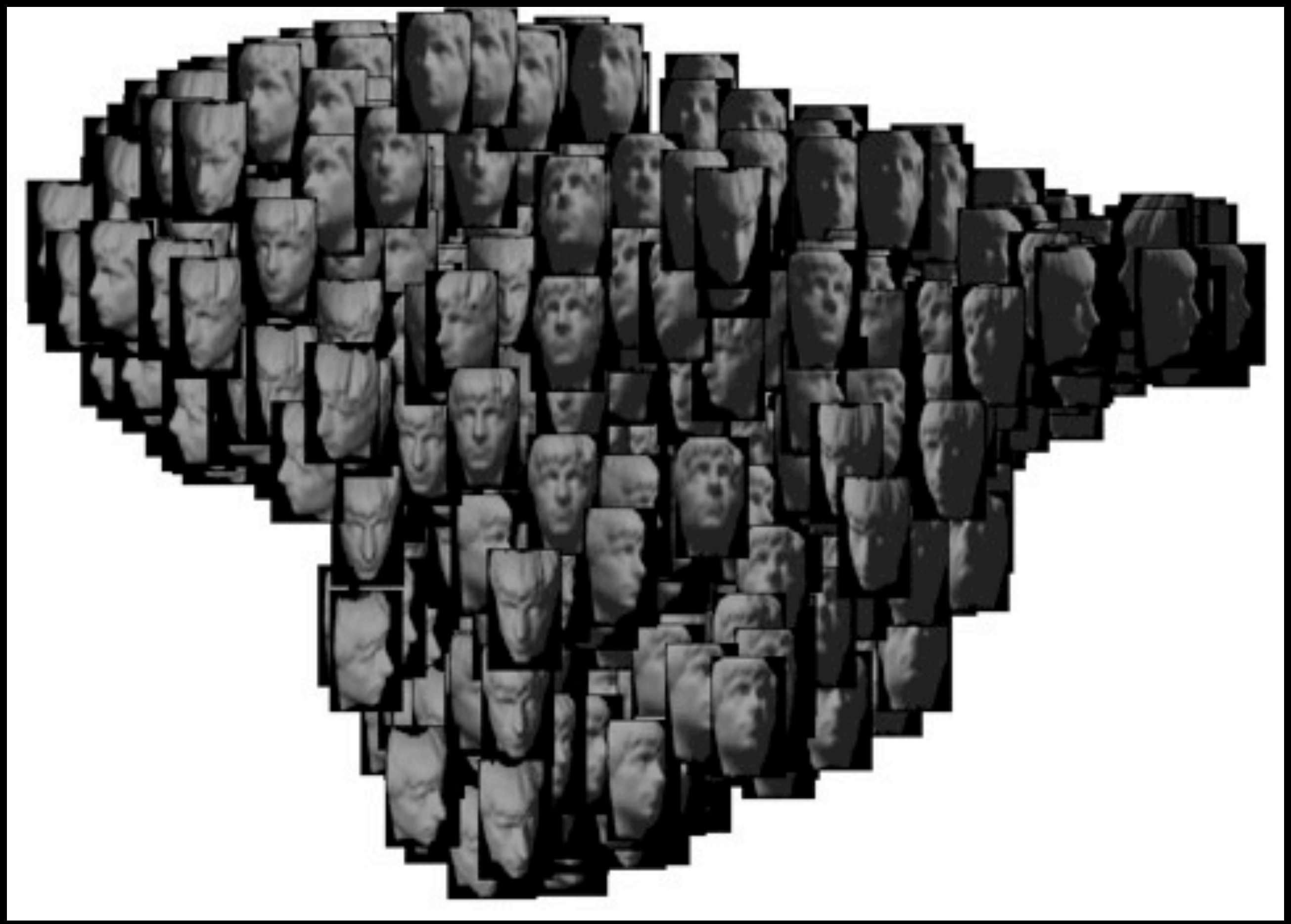
- <u>Assumption</u>: Y sampled from a low dimensional manifold embedded in a high dimensional space



Manifold Learning

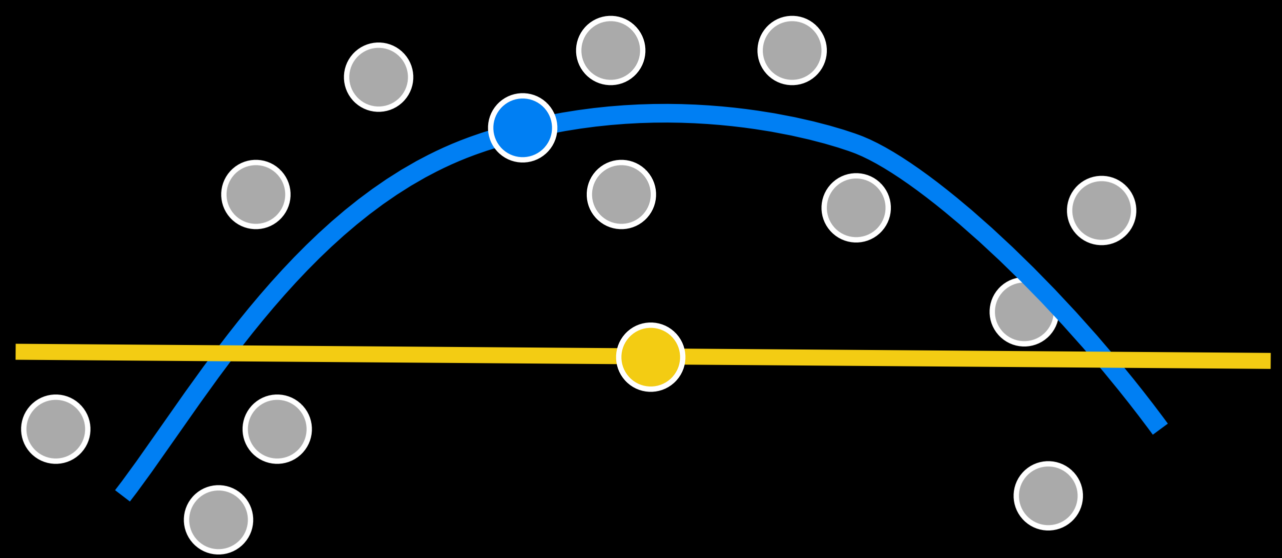


Manifold - Non-linear



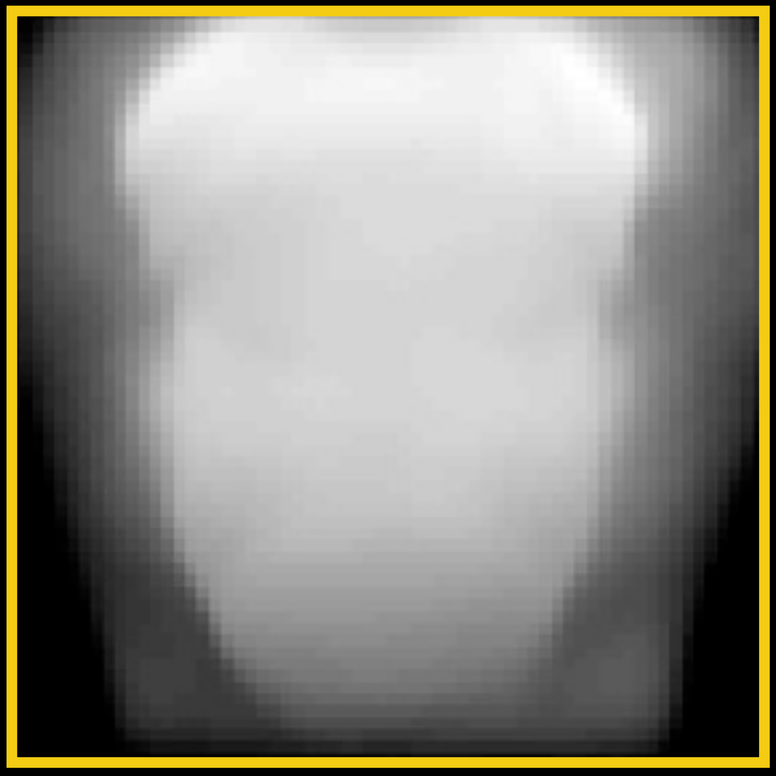
PCA - Linear

Manifold Learning



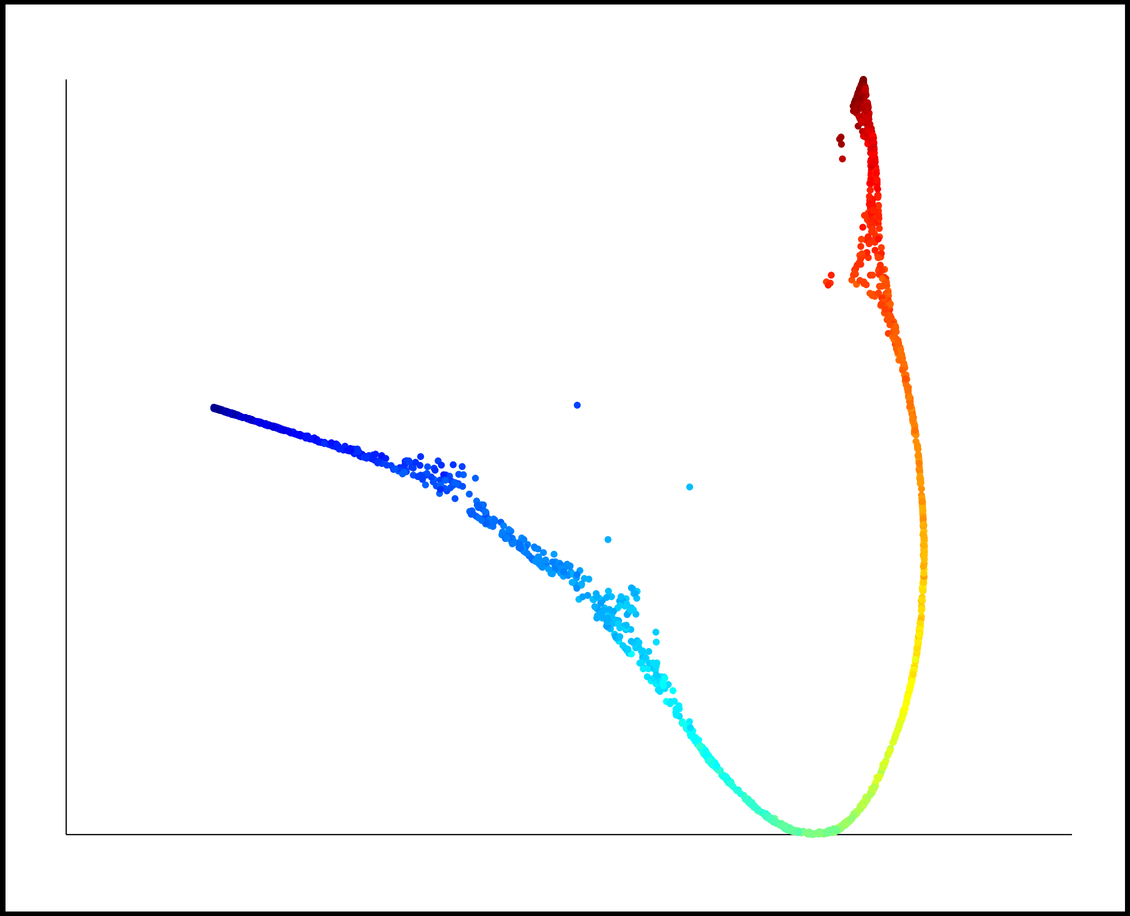
Manifold Mean

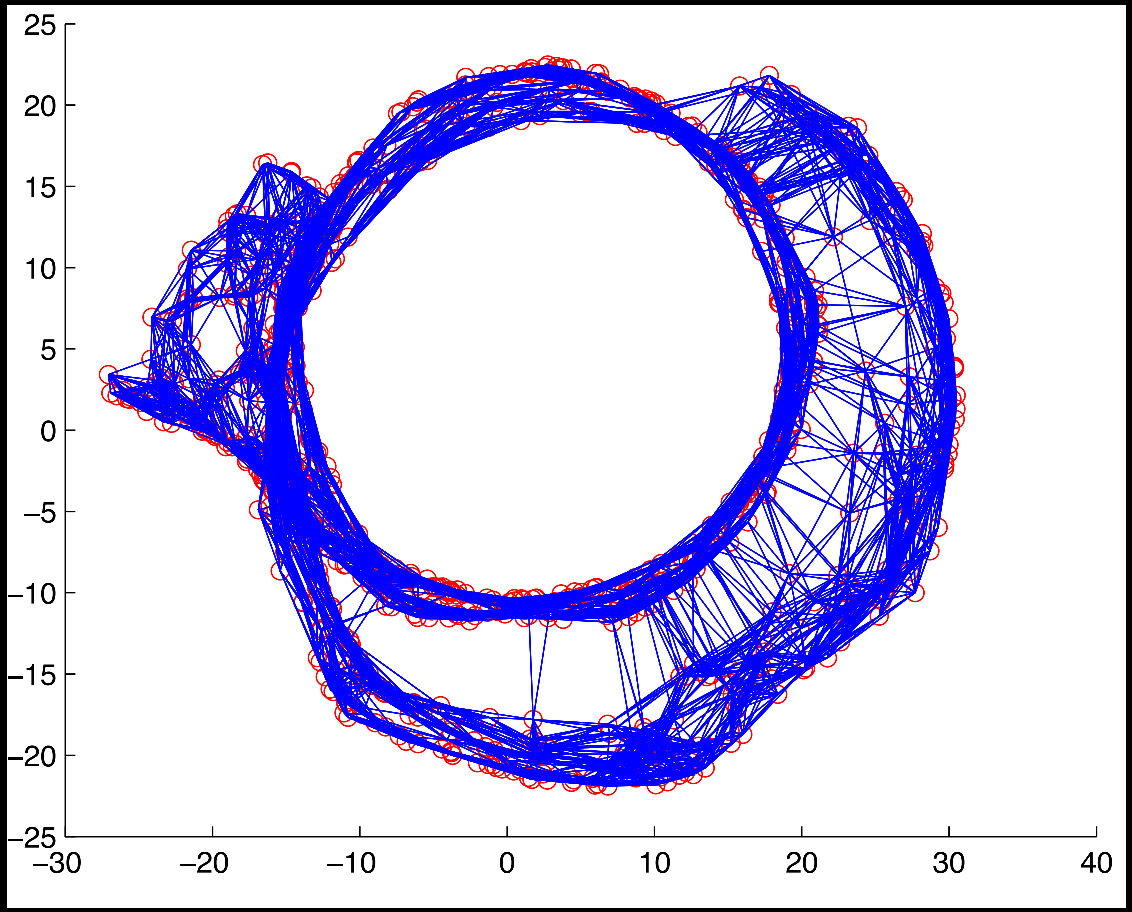
Linear Mean



Popular Algorithms

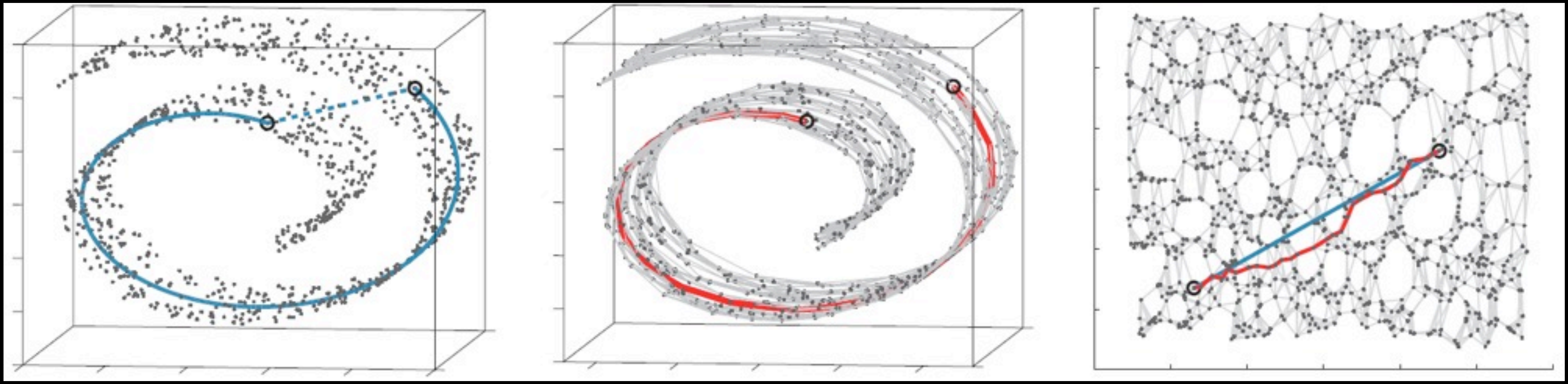
Techniques and Comments





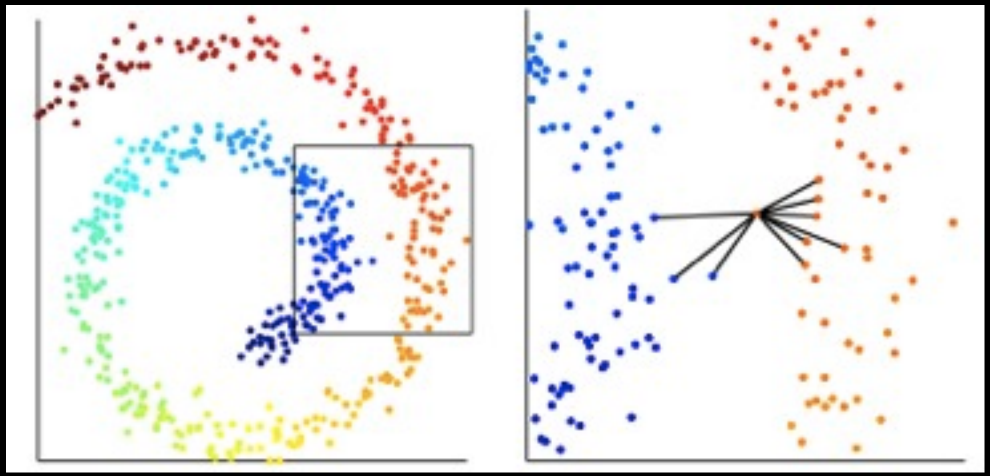
Isomap

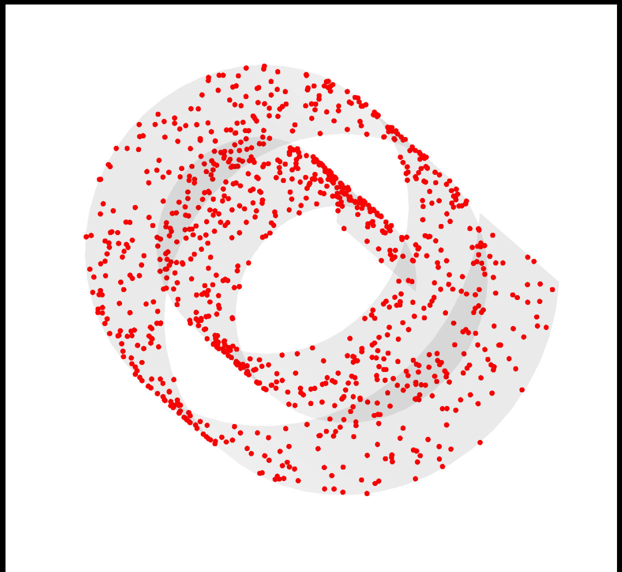
- Approximate geodesic distances $\hat{\delta}$ by shortest path in nearest neighbor graph
- Preserve approximate geodesics
  - $\min_x = \sum_{i,j} [\hat{\delta}(y_i, y_j) - d(x_i, x_j)]^2$
  - Multidimensional scaling

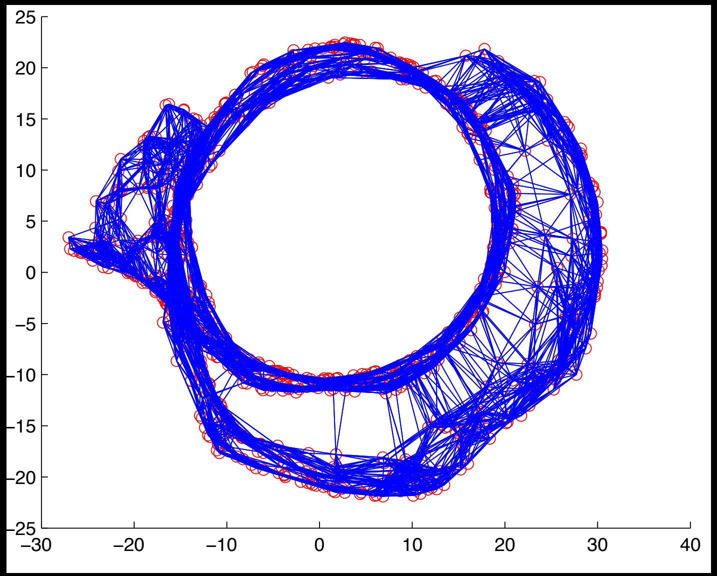


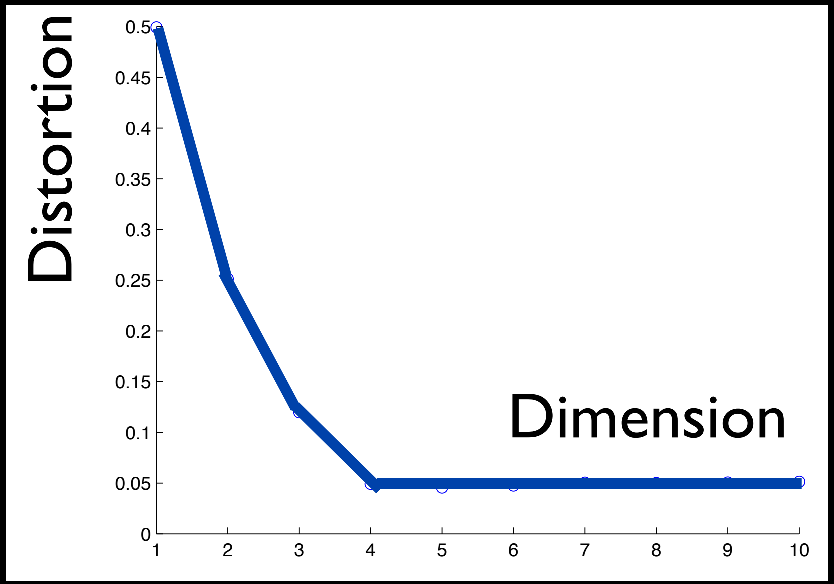
Properties

- Only relies on accurate local distances
- Shortcuts in graph - very bad approximation
- Quality measure based on graph embedding



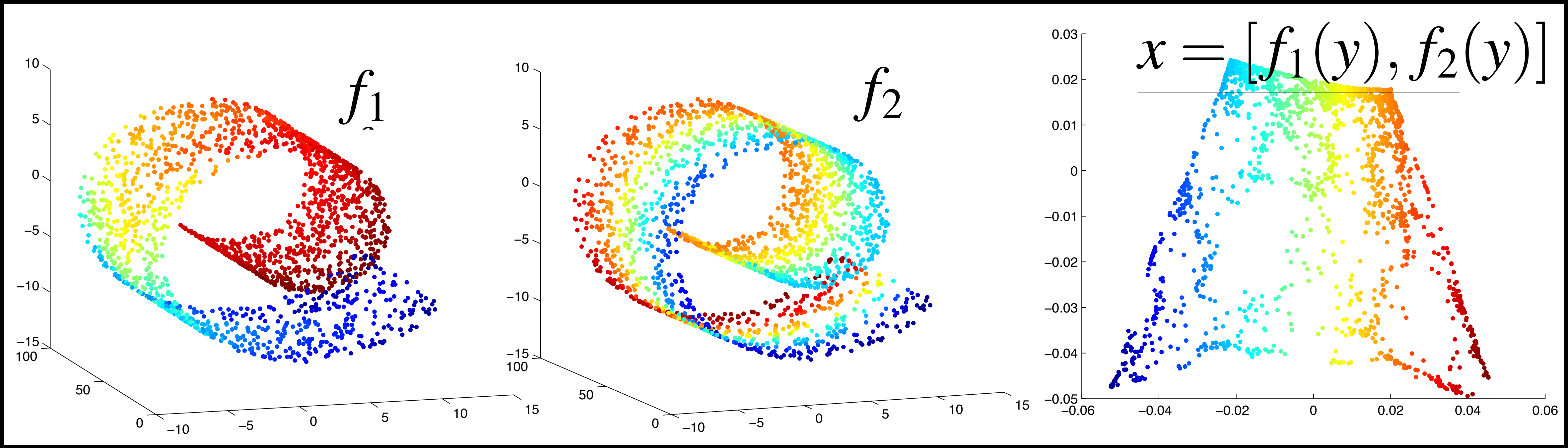






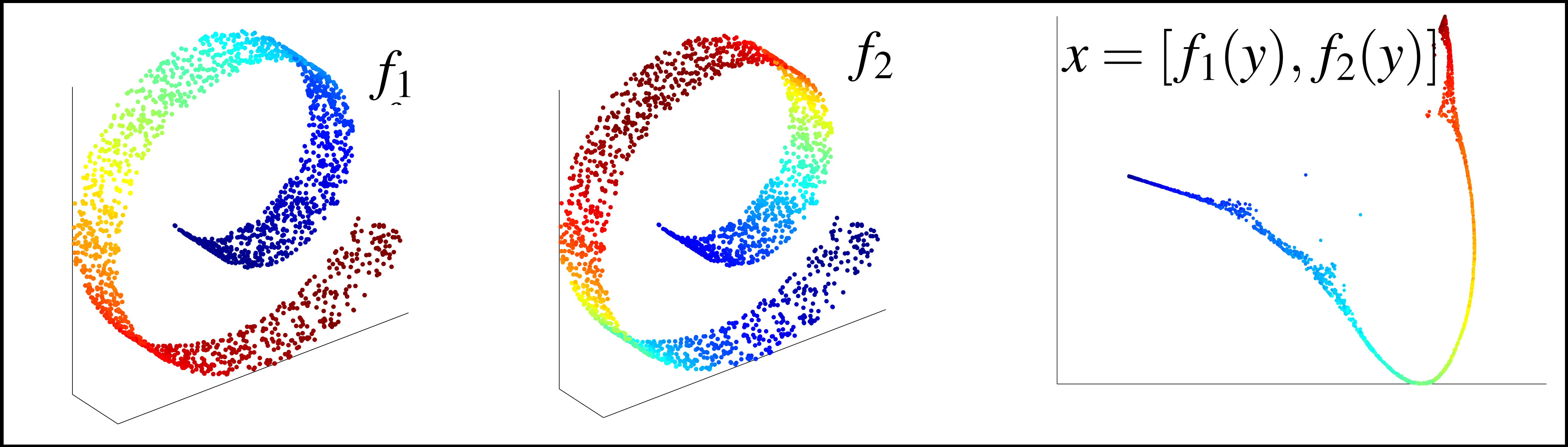
Laplacian Eigenmaps

- Given a manifold $\mathcal{M}$ find functions $f : \mathcal{M} \mapsto \mathbb{R}$ such that $\int_{\mathcal{M}} \|\nabla f(y)\|^2 dy$ is minimized
- The low dimensional embedding is $x = [f_1(y), \ldots, f_n(y)] \in \mathbb{R}^n$
- Small gradient implies that close by points will be mapped close together



Properties

- Again only local distances important
- No quality measure of the embedding

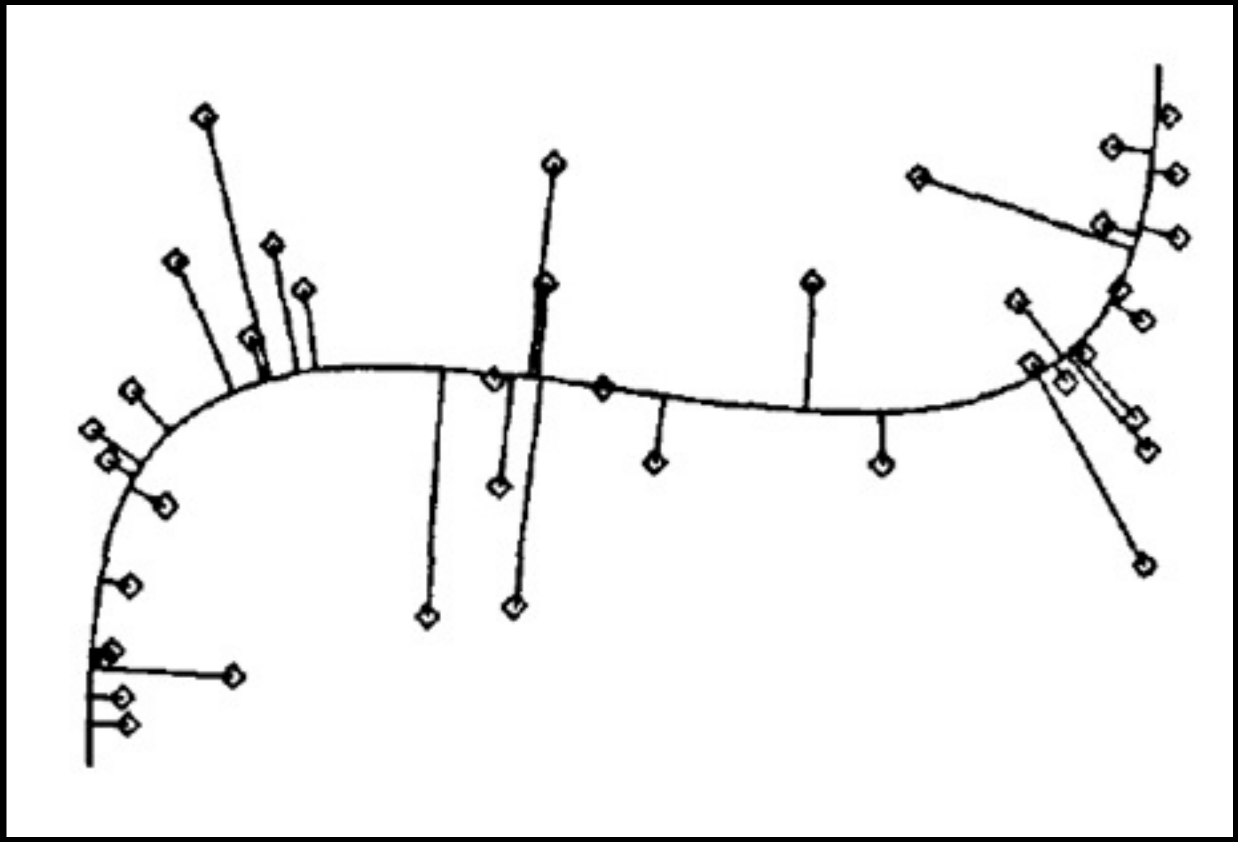


Spectral Methods

- Isomap, Laplacian eigenmaps, LLE, Kernel PCA, and a whole set of related variations
- Based on eigendecomposition of a similarity matrix
  - Closed form solution, relatively fast, simple

Principal Curves

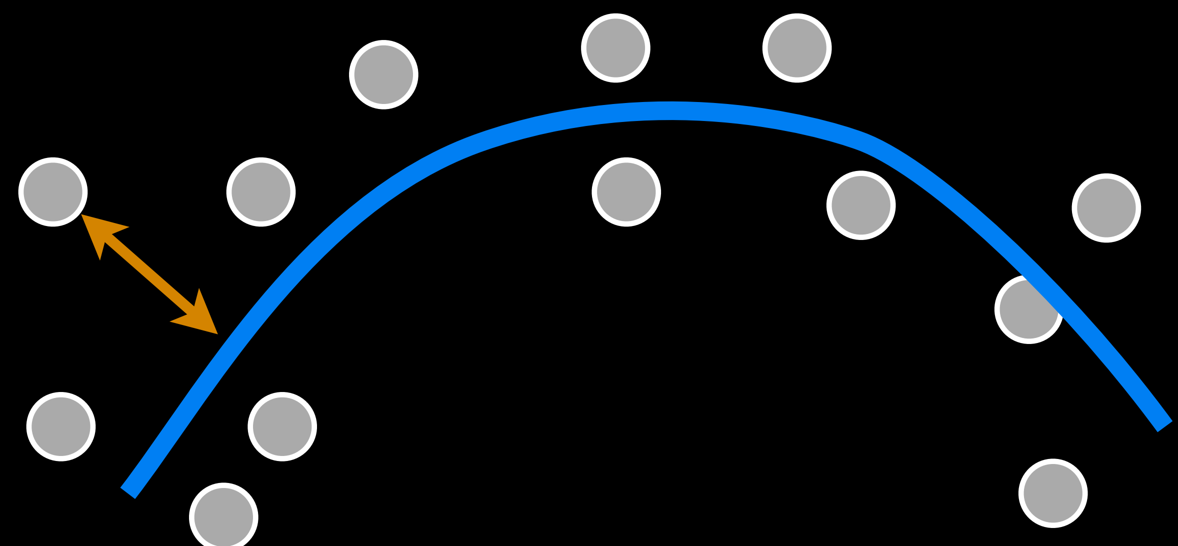
Curve that passes through the *middle* of a density



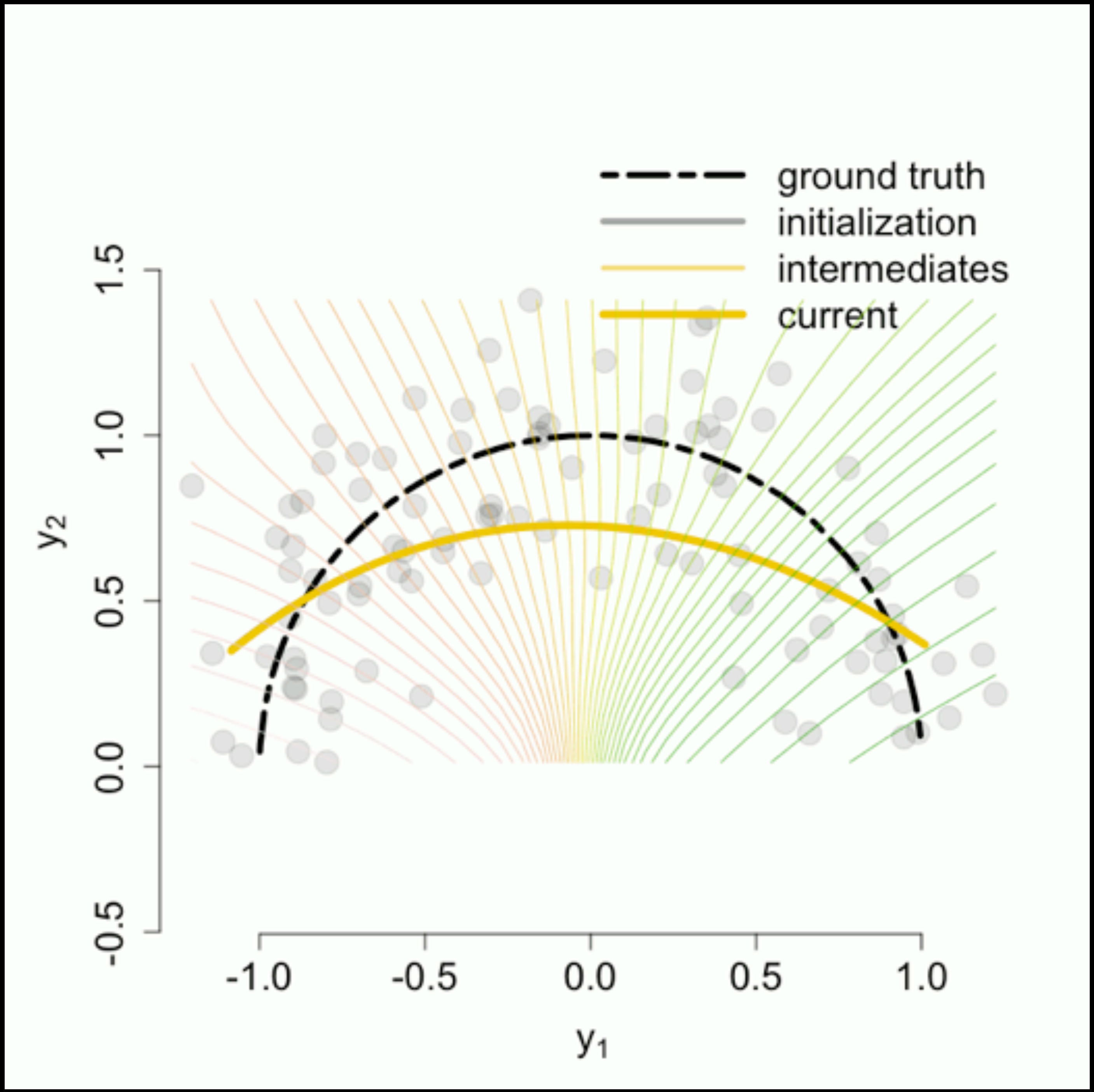
T. Hastie, W. Stuetzle, **Principal curves**
Journal of the American Statistical Association 1989

Principal Curves

- Quantitative measure of manifold fit
- Principled approach to reconstruct and project unseen data points
- Lots of methods to do fitting by minimizing reconstruction error (neural nets, regression, local PCA, ...)



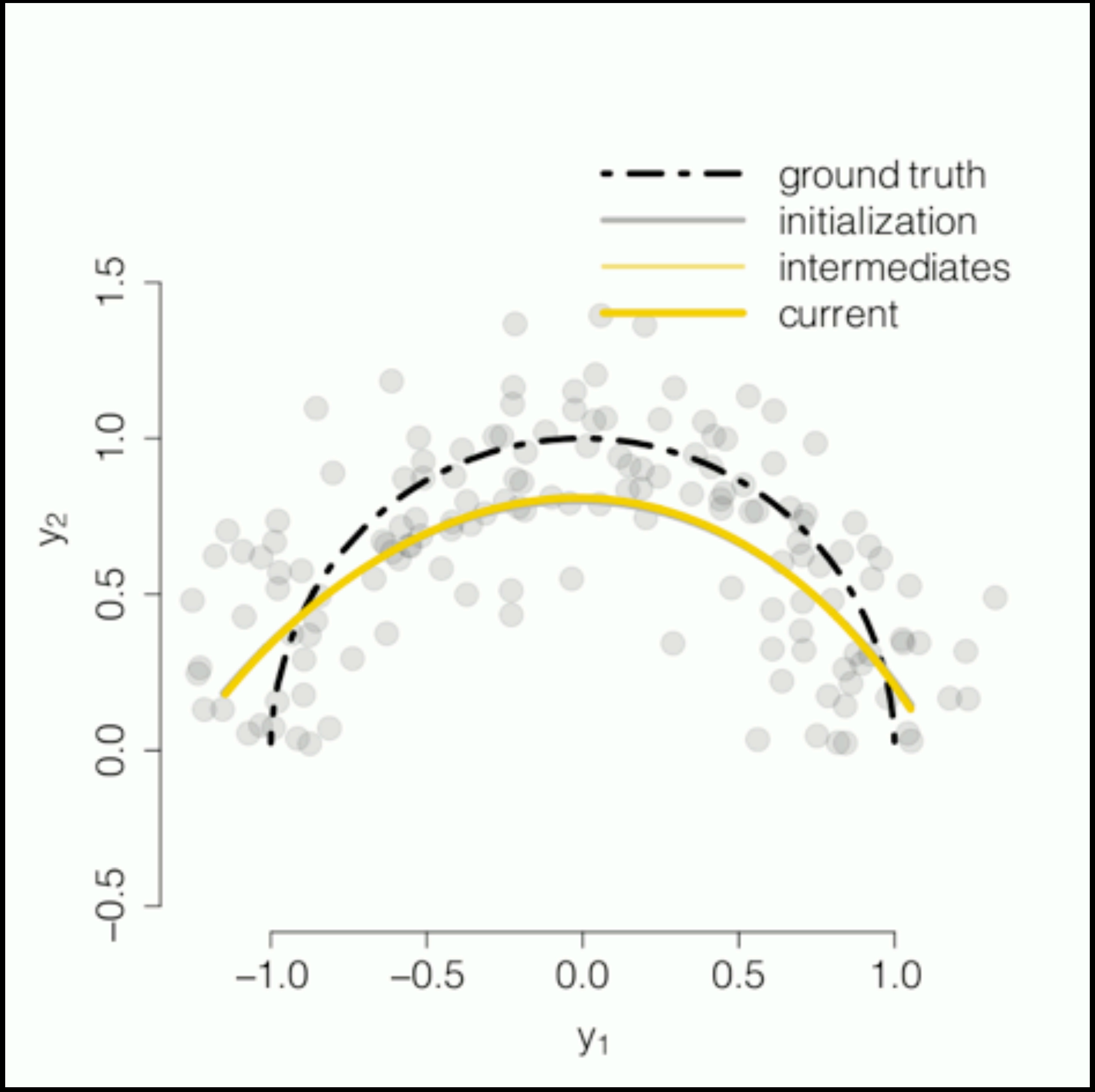
Fitting



Samuel Gerber, Tolga Tasdizen, Ross Whitaker **"Dimensionality Reduction and Principal Surfaces via Kernel Map Manifolds"**, (ICCV 2009)

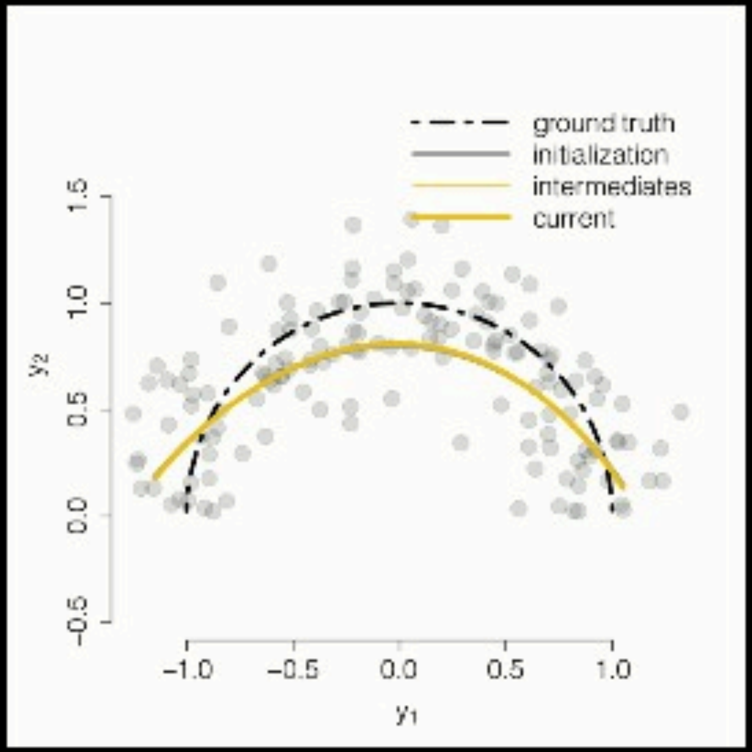
Model Complexity

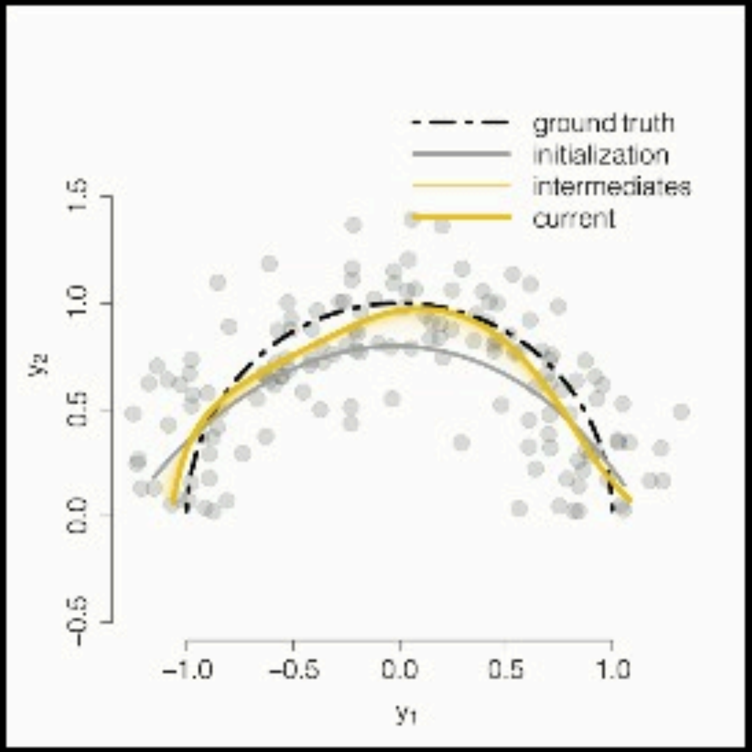
- Minimizing reconstruction error favors curves that pass close to the data points

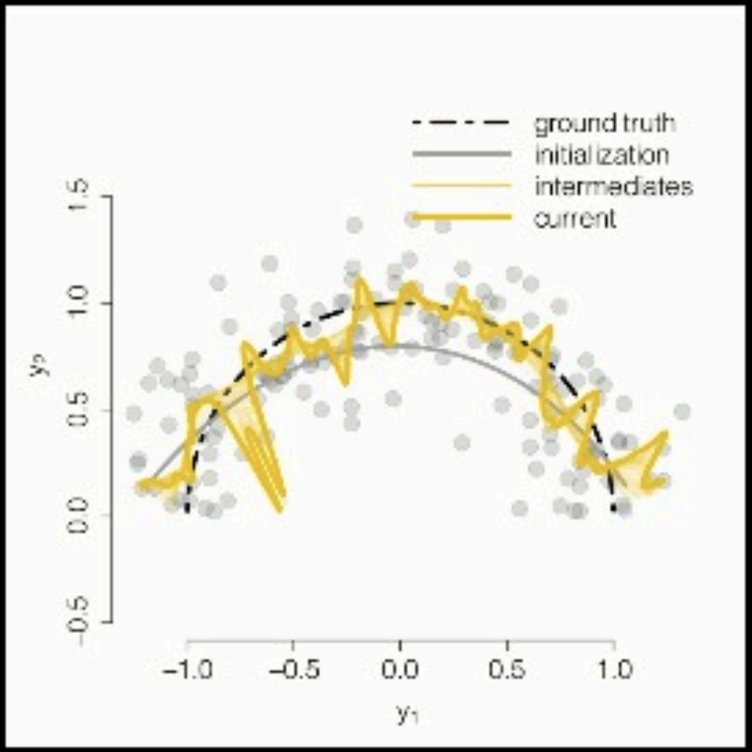


Regularization

- Defeats purpose of exploratory analysis
- Violates principal curve properties

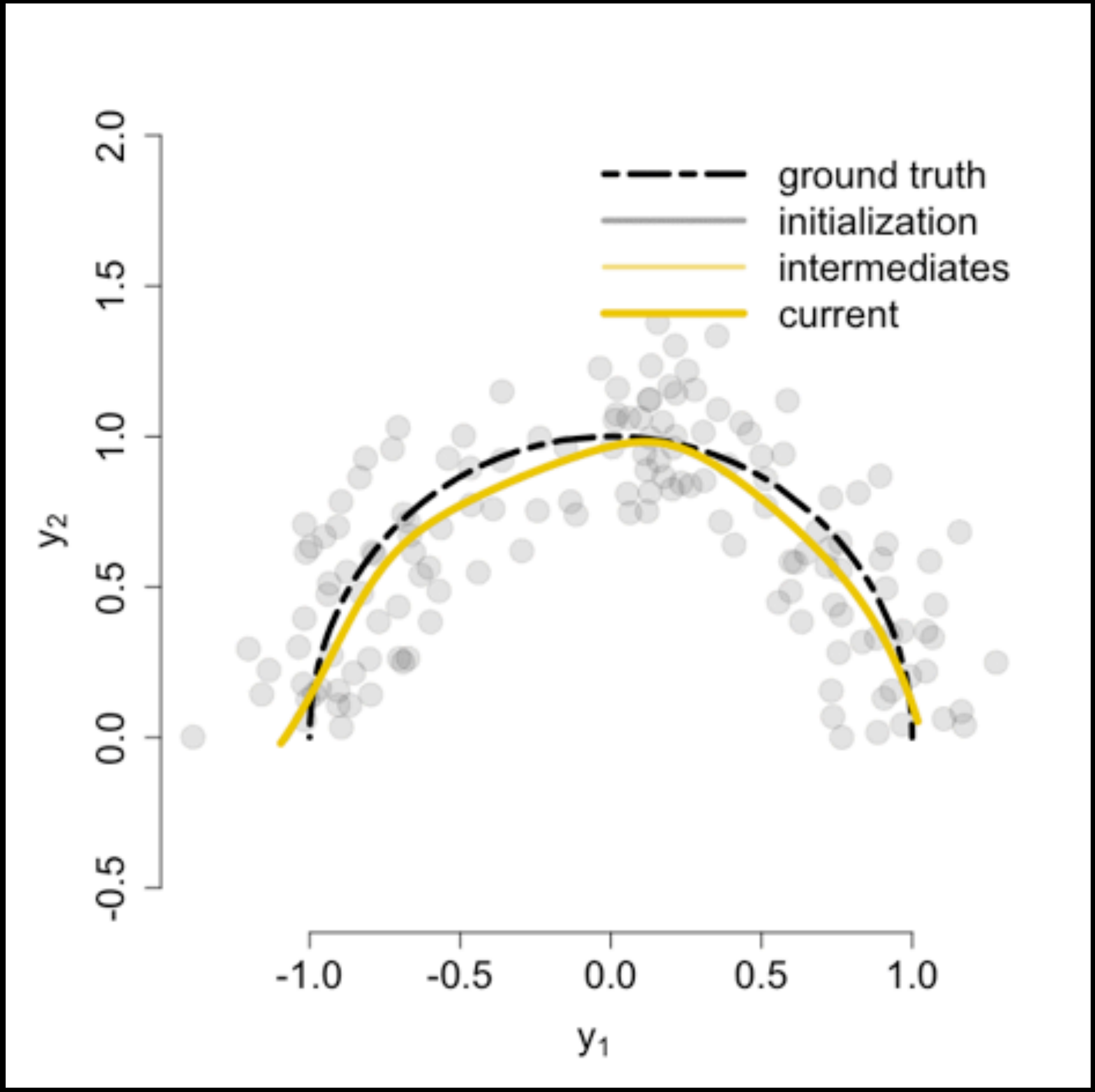


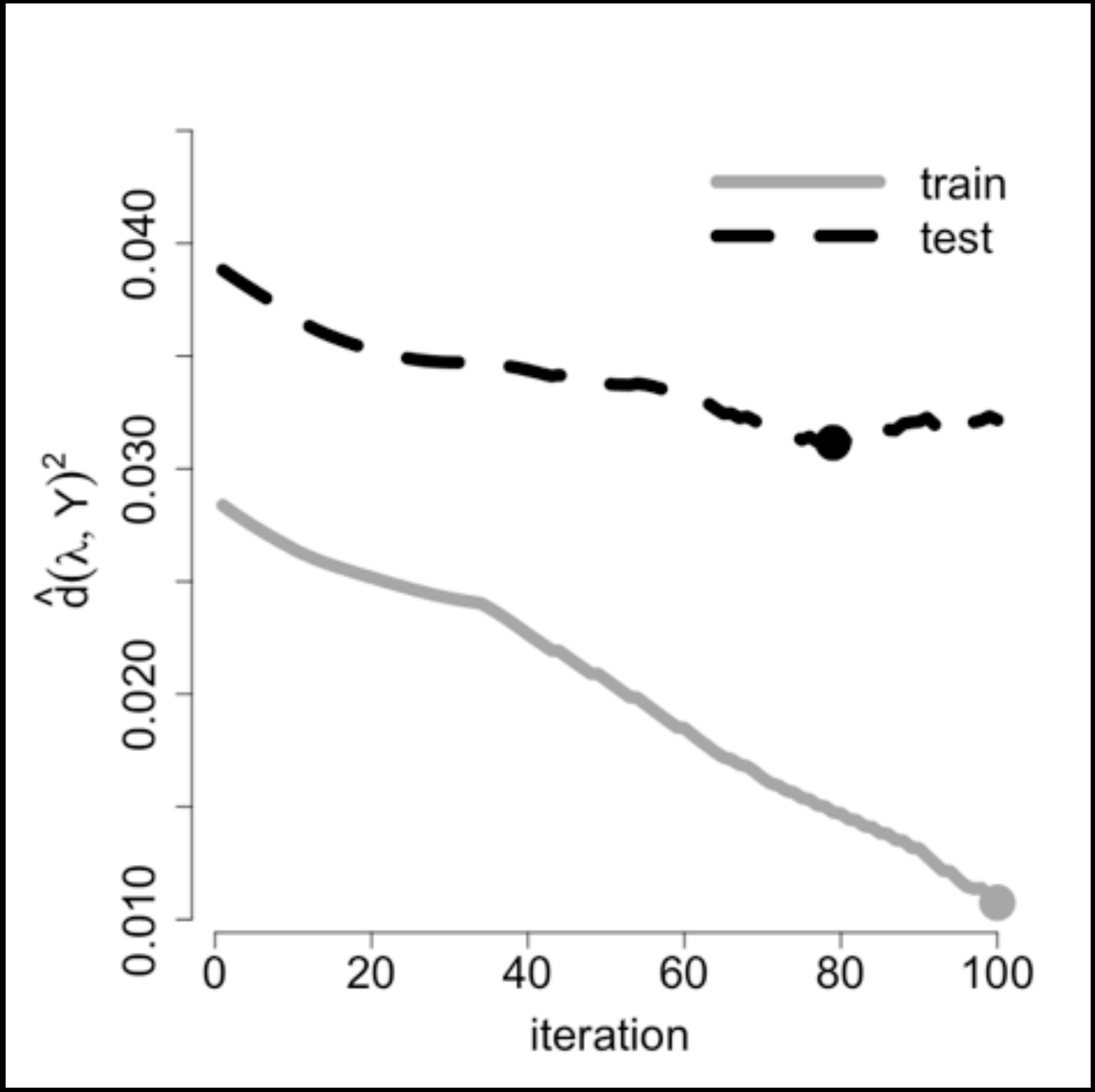




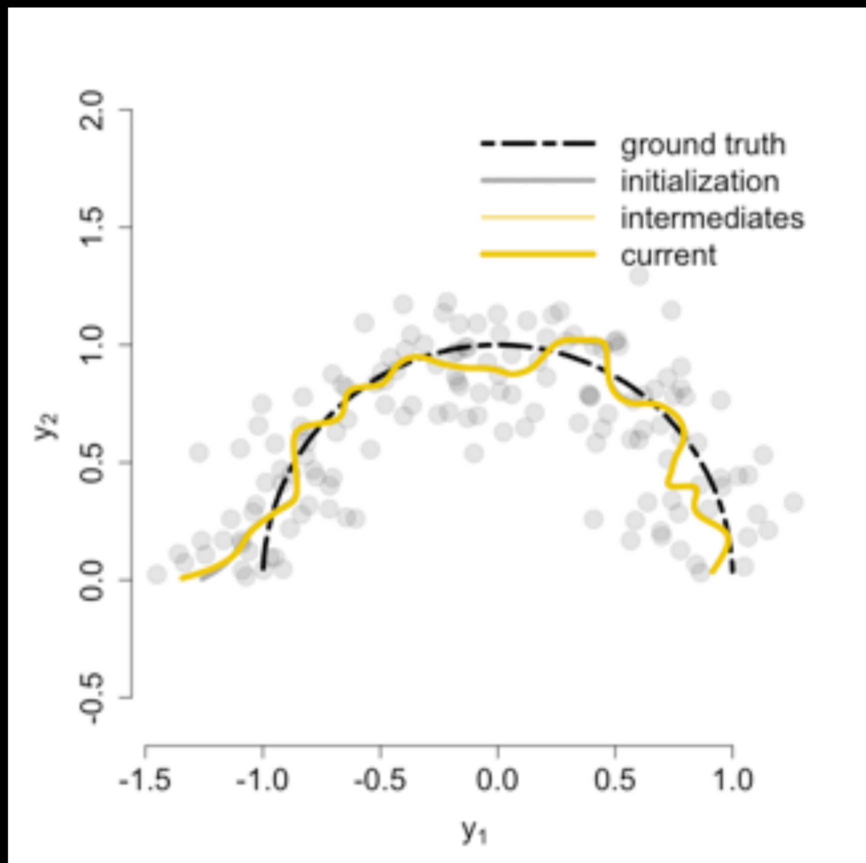
Cross-Validation

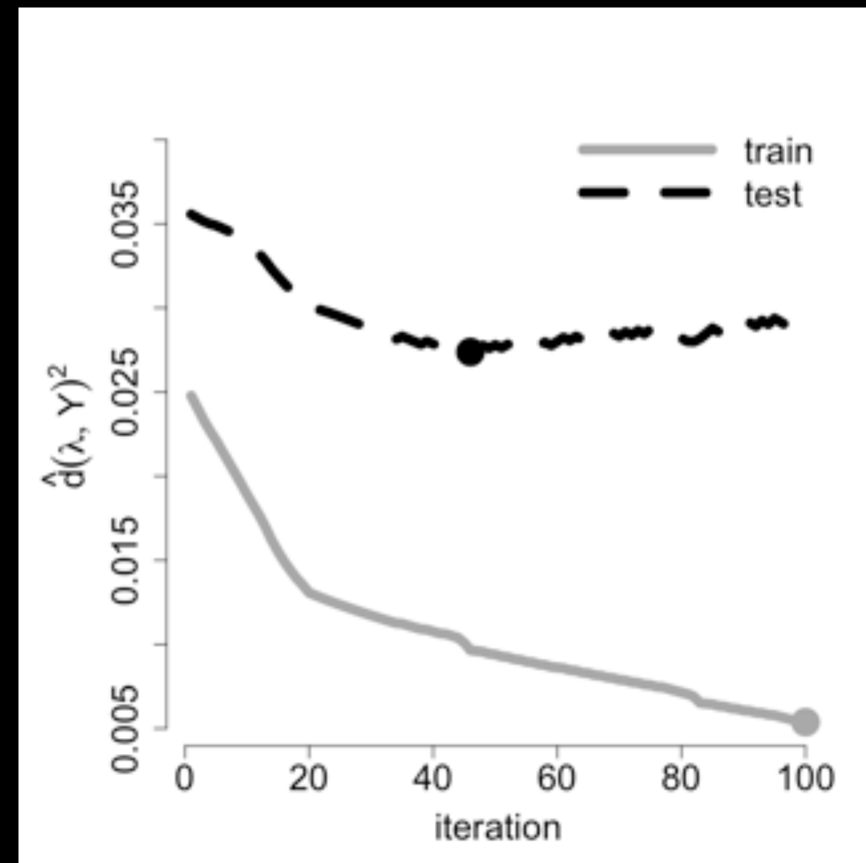
- No overfitting in traditional sense





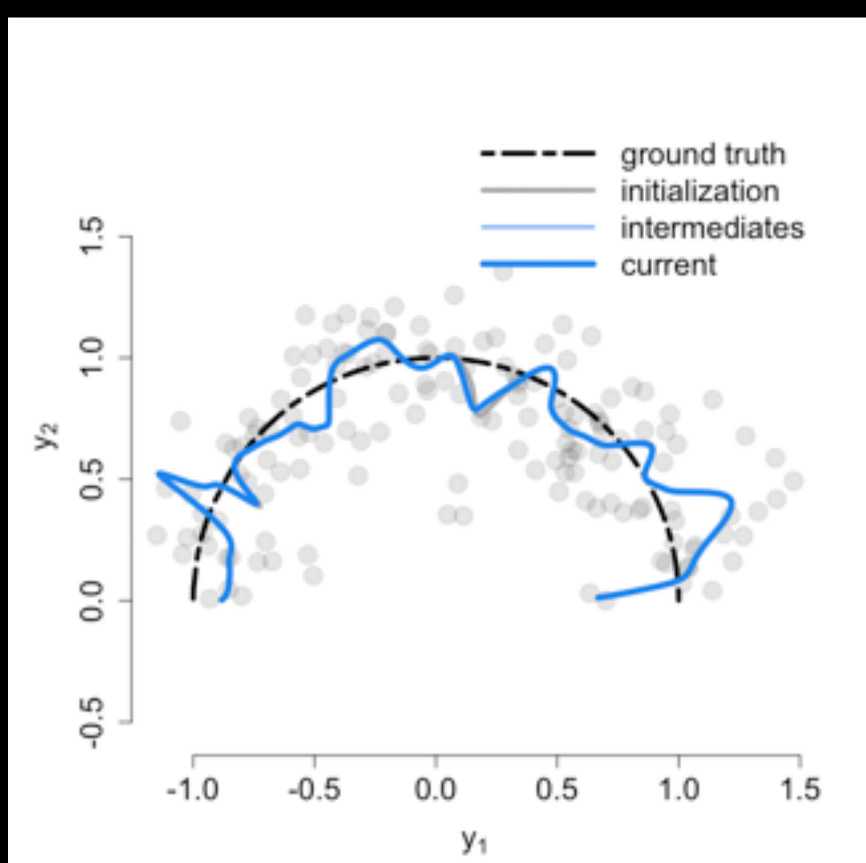
Results

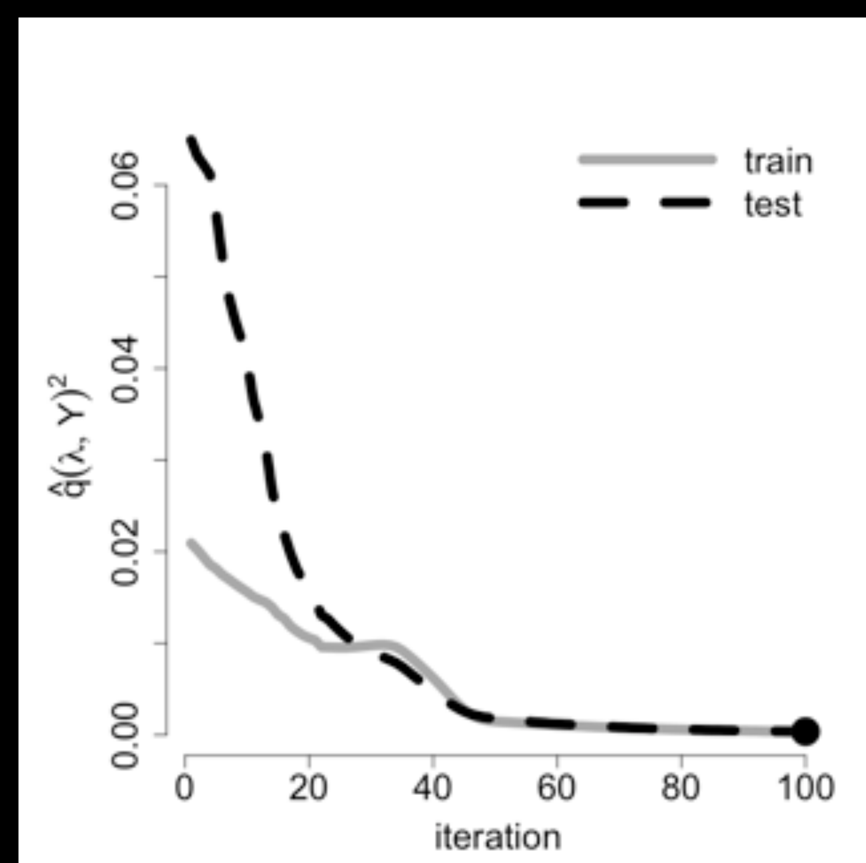




- Initialization
  - Ground truth
  - Small bandwidth

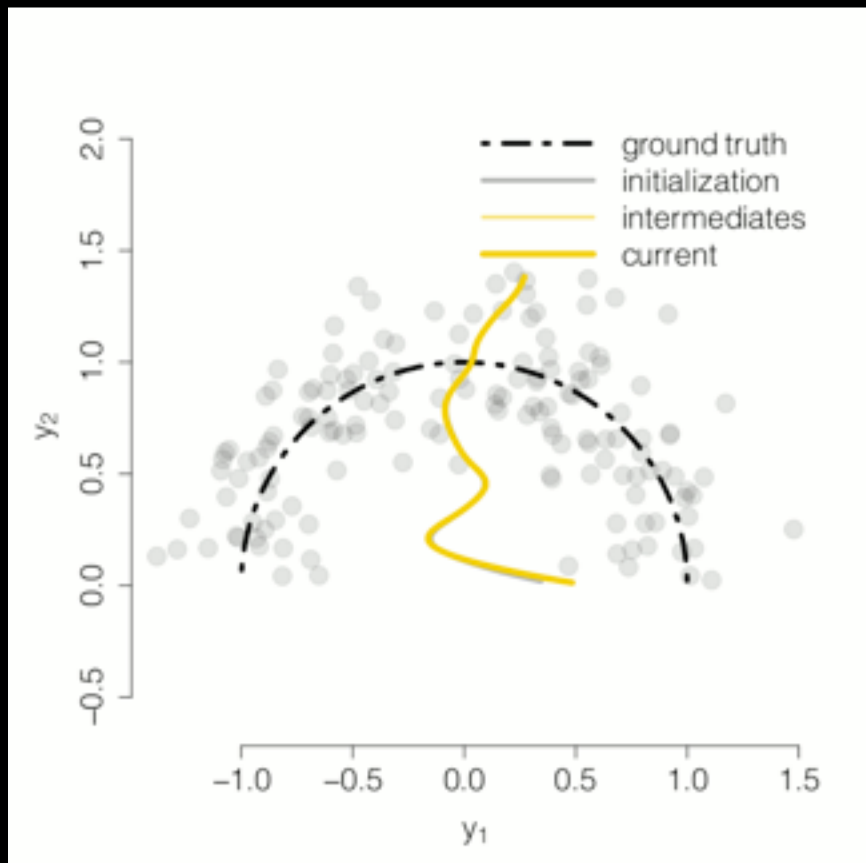
$\hat{d}(\lambda, Y)^2$

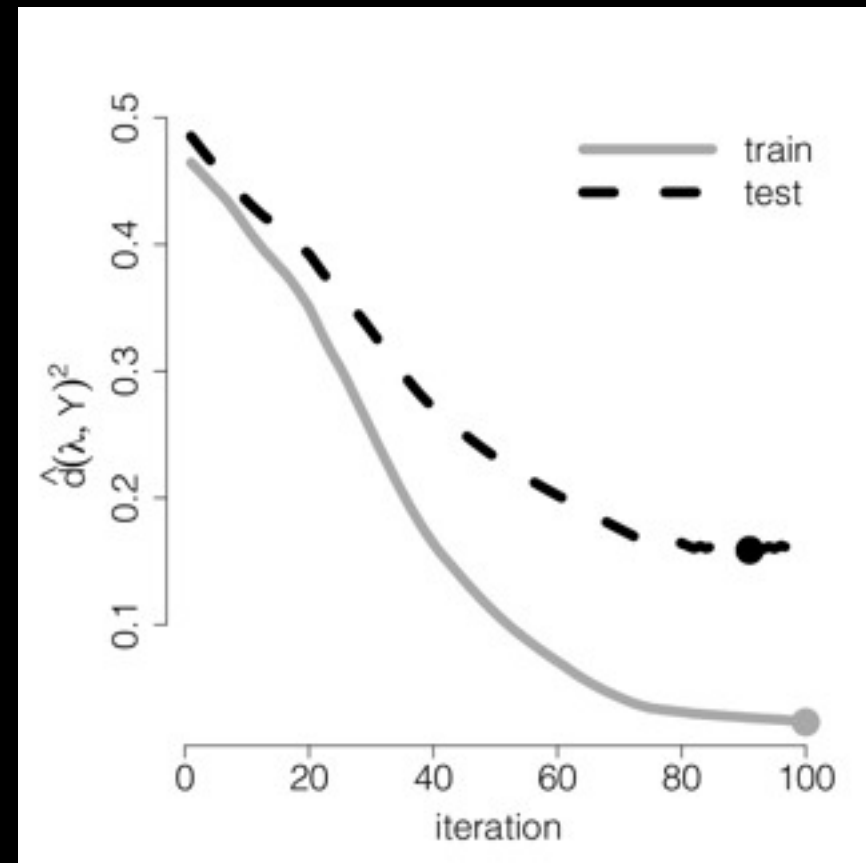




$\hat{q}(\lambda, Y)^2$

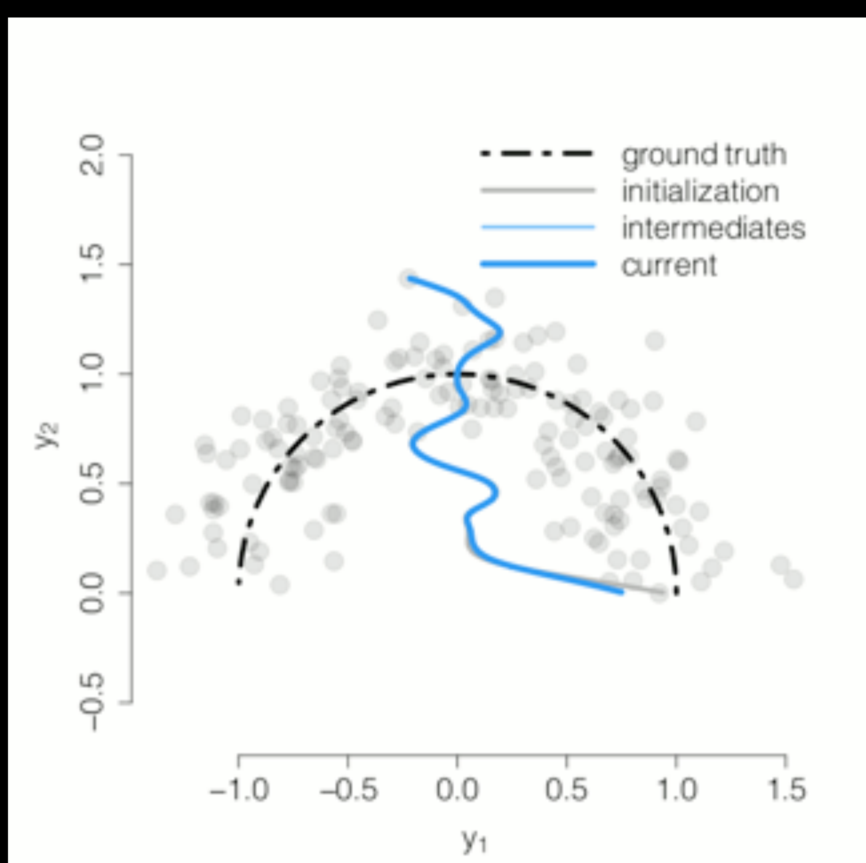
Results

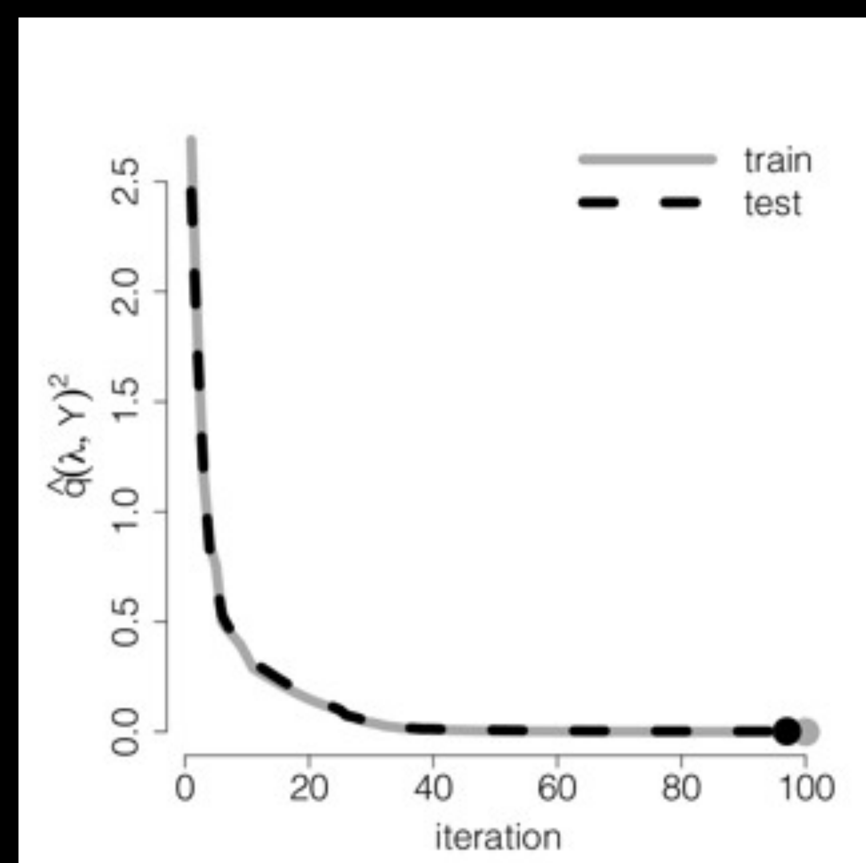




- Initialization
  - Principal component

$\hat{d}(\lambda, Y)^2$

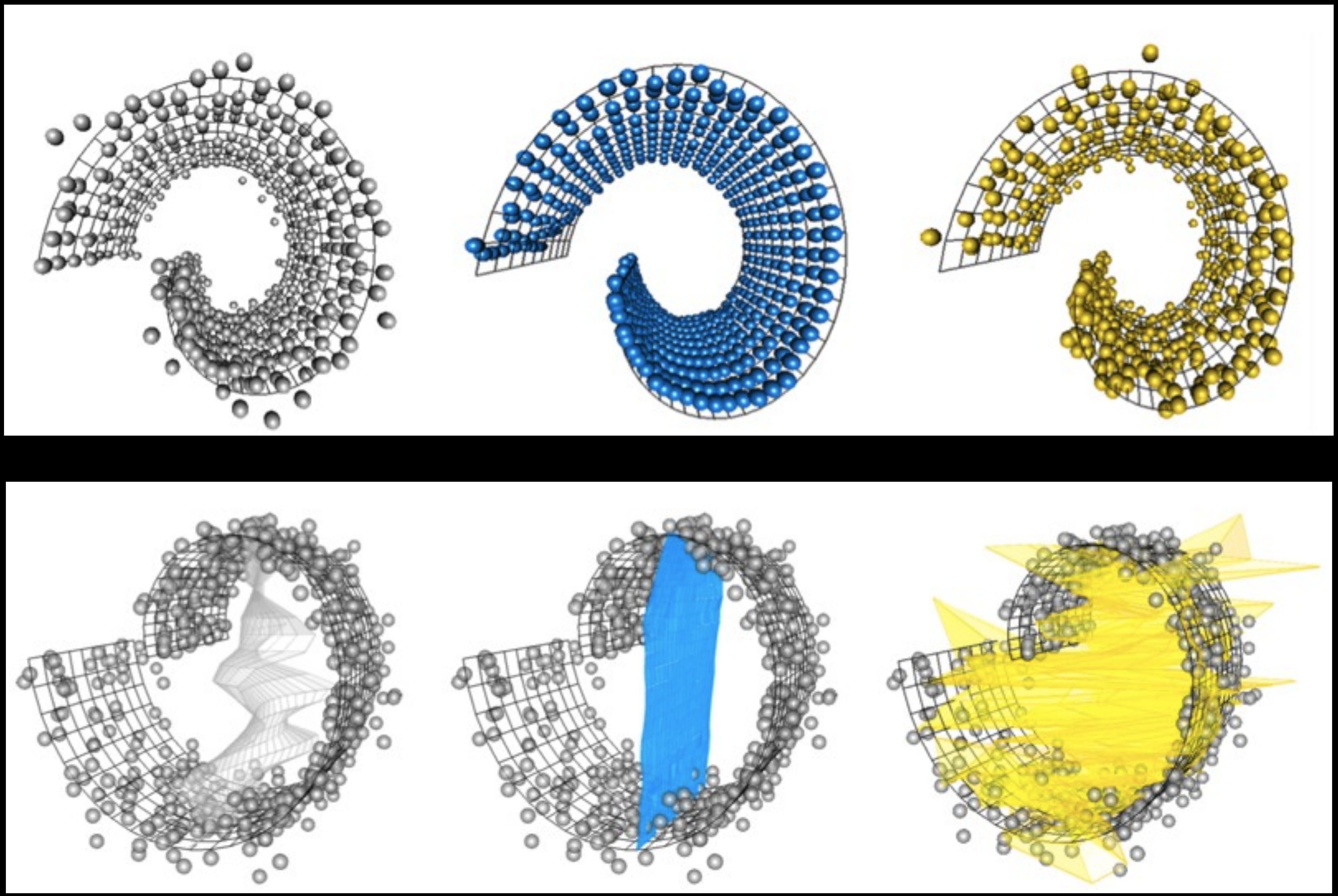




$\hat{q}(\lambda, Y)^2$

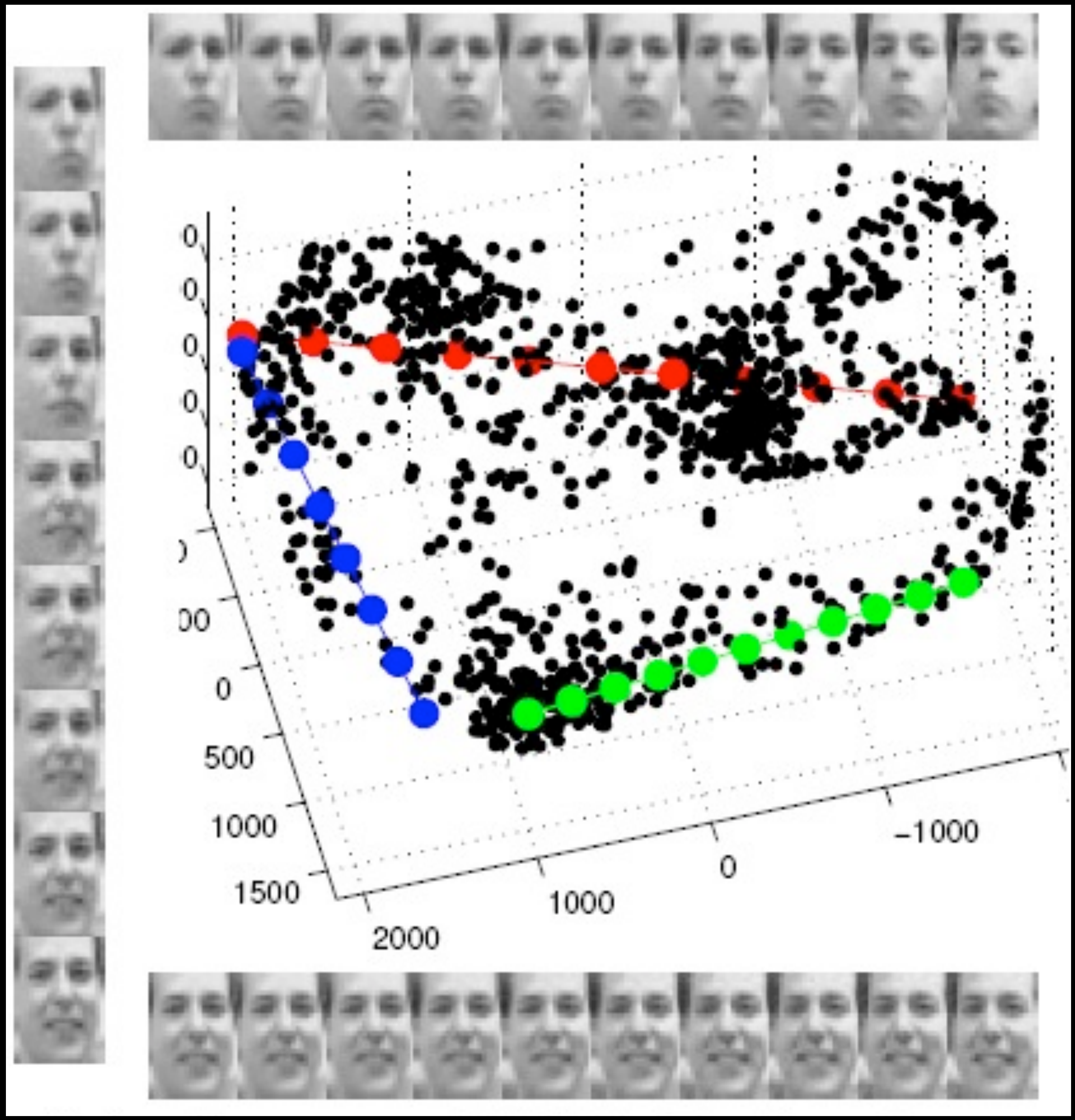
Results

$\hat{q}(\lambda, Y)^2$ $\hat{d}(\lambda, Y)^2$



Some examples

- 1965 images of different facial expression (20x28)



Some Examples

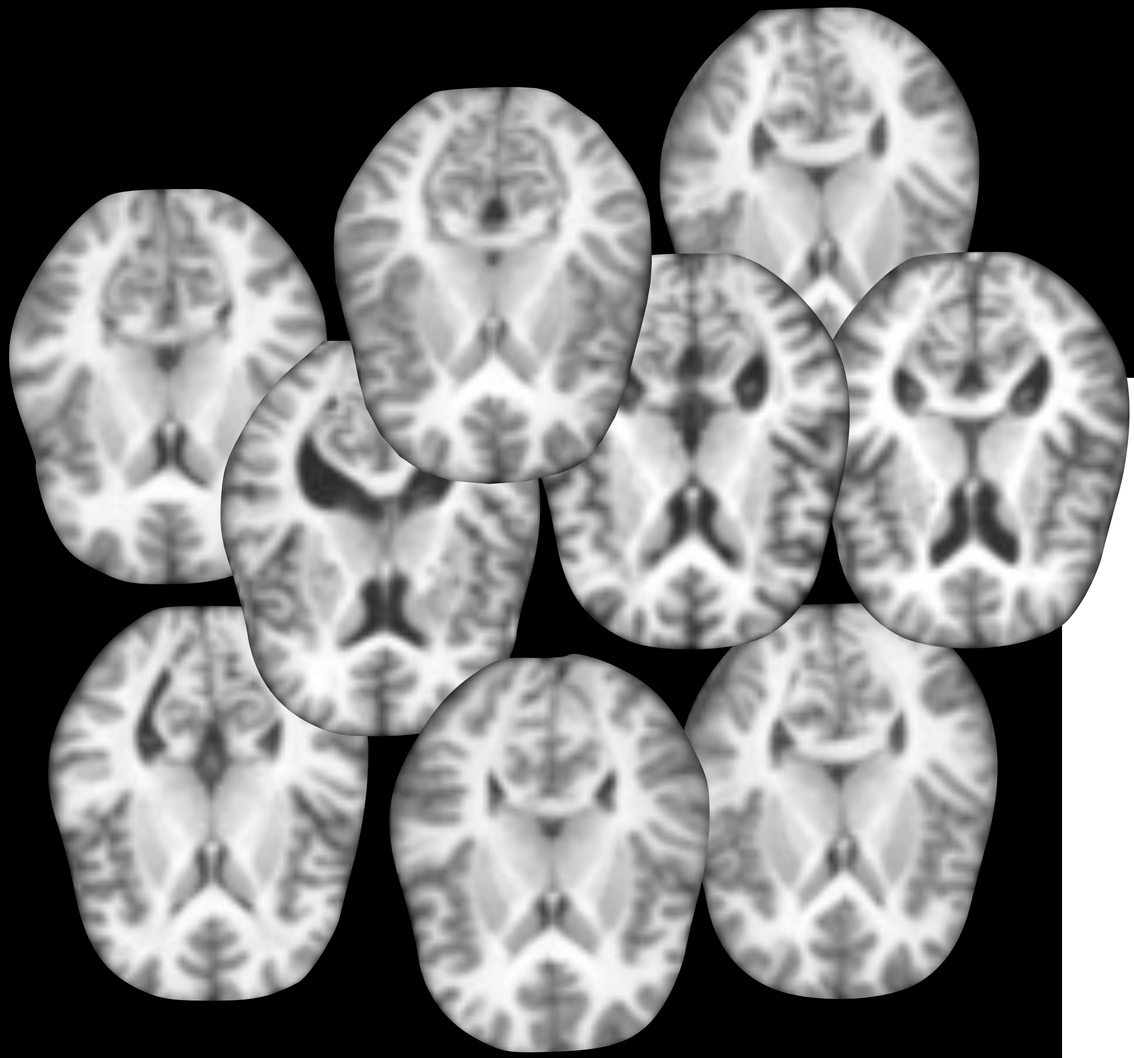
Original

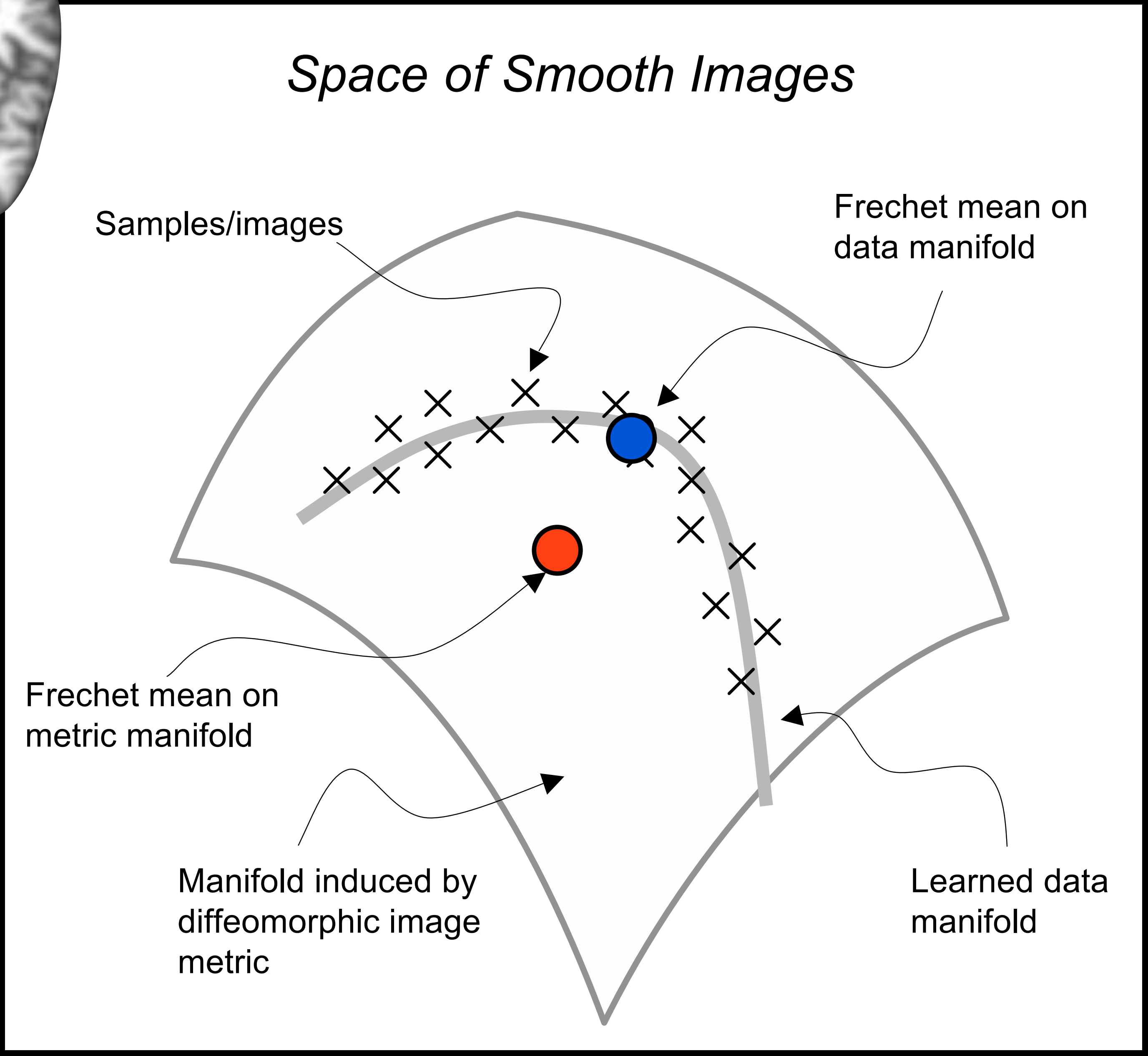
Noisy

$\hat{d}(\lambda, Y)^2$

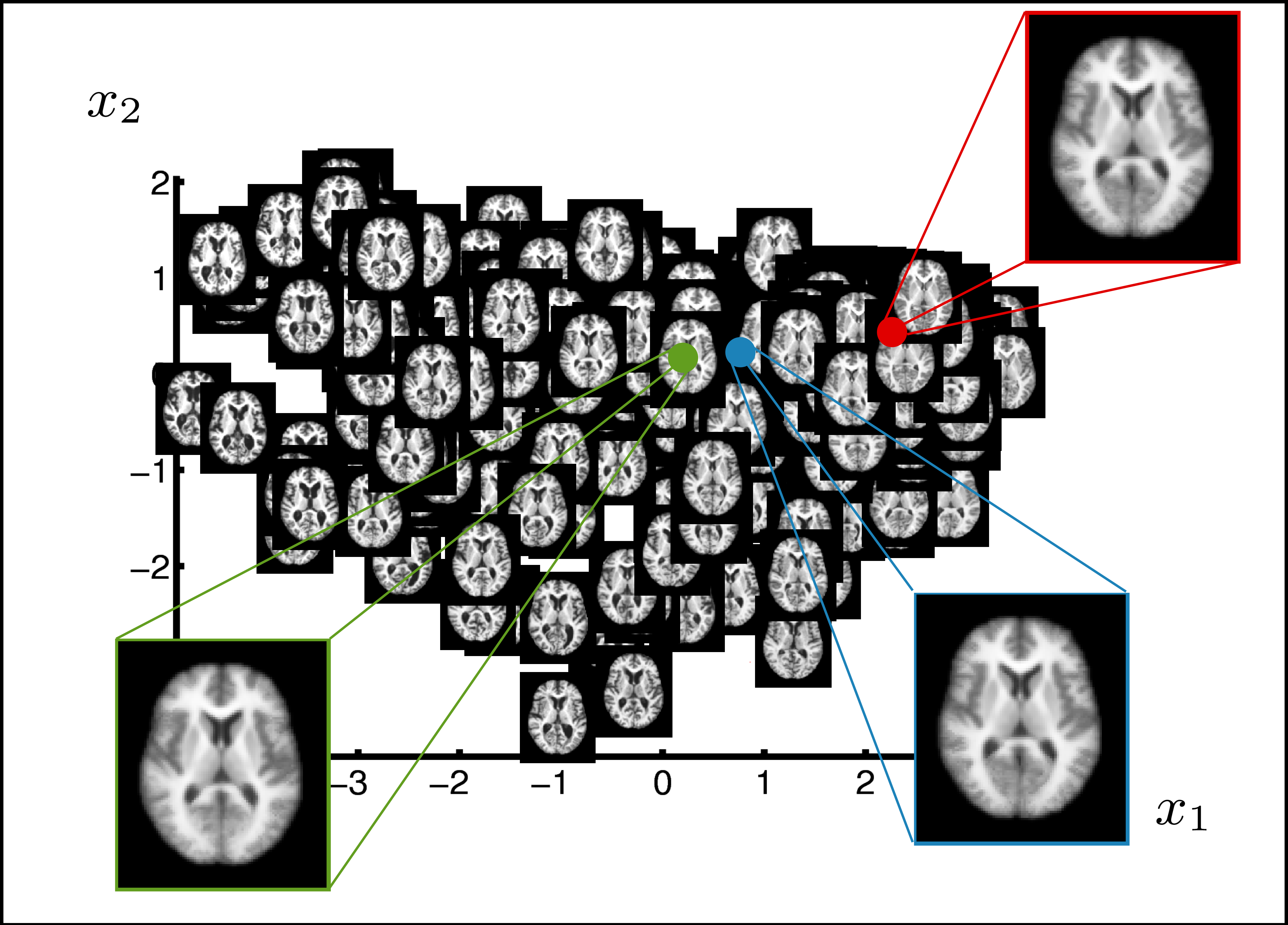
$\hat{q}(\lambda, Y)^2$

Manifold in Brain Space



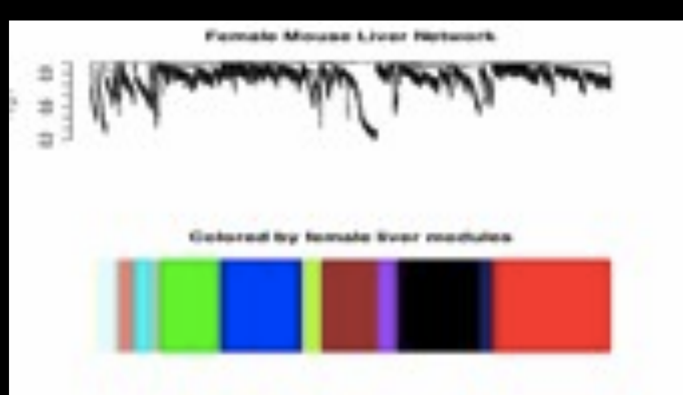


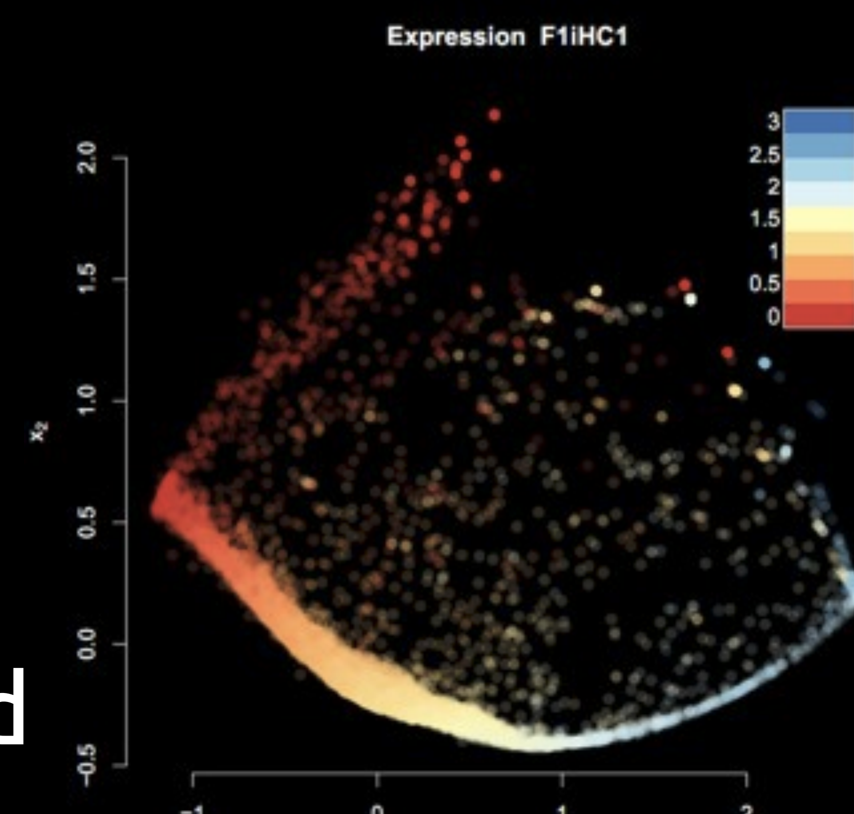
Space of Brain Images



Remarks

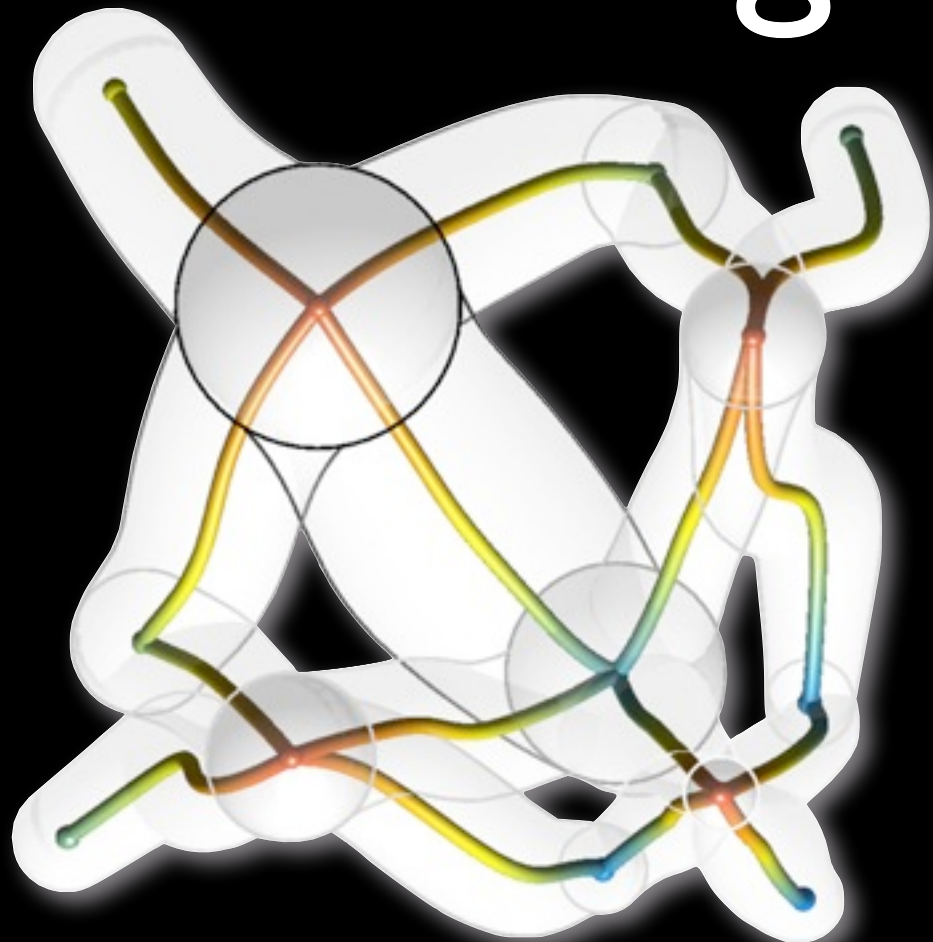
- Be aware of the specifics of the method
  - Quantify quality of summary / approximation
- How to measure distances (metric) often more important than dimension reduction method.
- Summary representation provides a cognitive model.





Gene clusters vs. manifold

Modeling High-Dimensional Scalar Functions



Visual exploration of high-dimensional scalar functions
S. Gerber, P-T. Bremer, V. Pascucci, and R. Whitaker, IEEE TVCG, 2010.

Morse-Smale regression
S. Gerber, O. Ruebel, P-T. Bremer, V. Pascucci, and R. Whitaker, Journal of Computational and Graphical Statistics, 2012 (to appear)

Data analysis with the Morse-Smale complex: The msr package for R
S. Gerber and K. Potter, Journal of Statistical Software, 2012 (to appear)

Problem Setting

- Given a sample from a scalar function

$$y_i = f(\mathbf{x}_i) \qquad \mathbf{x} \in \mathbf{R}^d$$

- Goals:
  - Visualize / explore the structure of f
  - Build an interpretable regression (approximation) model

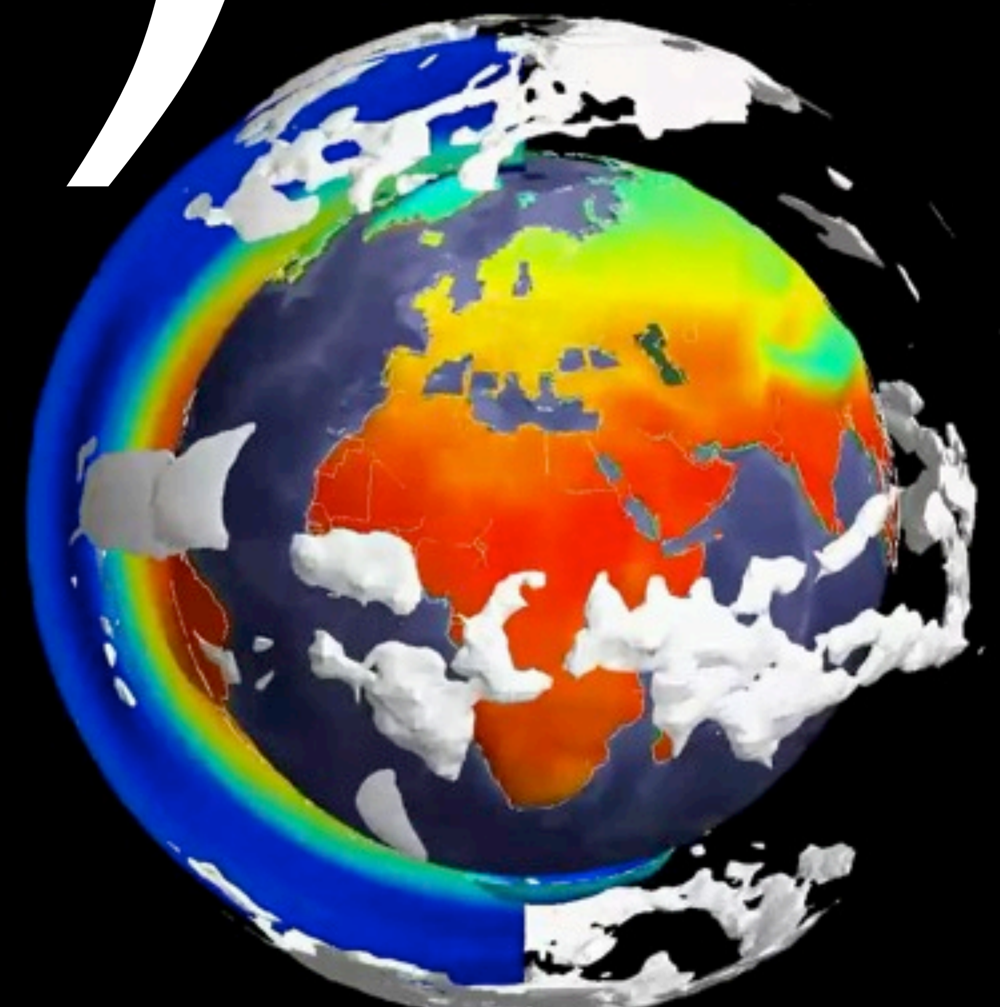
Crime Rates

f (
Population
Percentage urban
Police per population
Person per family
Income
Employment
Rent cost
Number of homeless
Household size
Poverty levels
......
) = crime rate

http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime

Climate Simulations

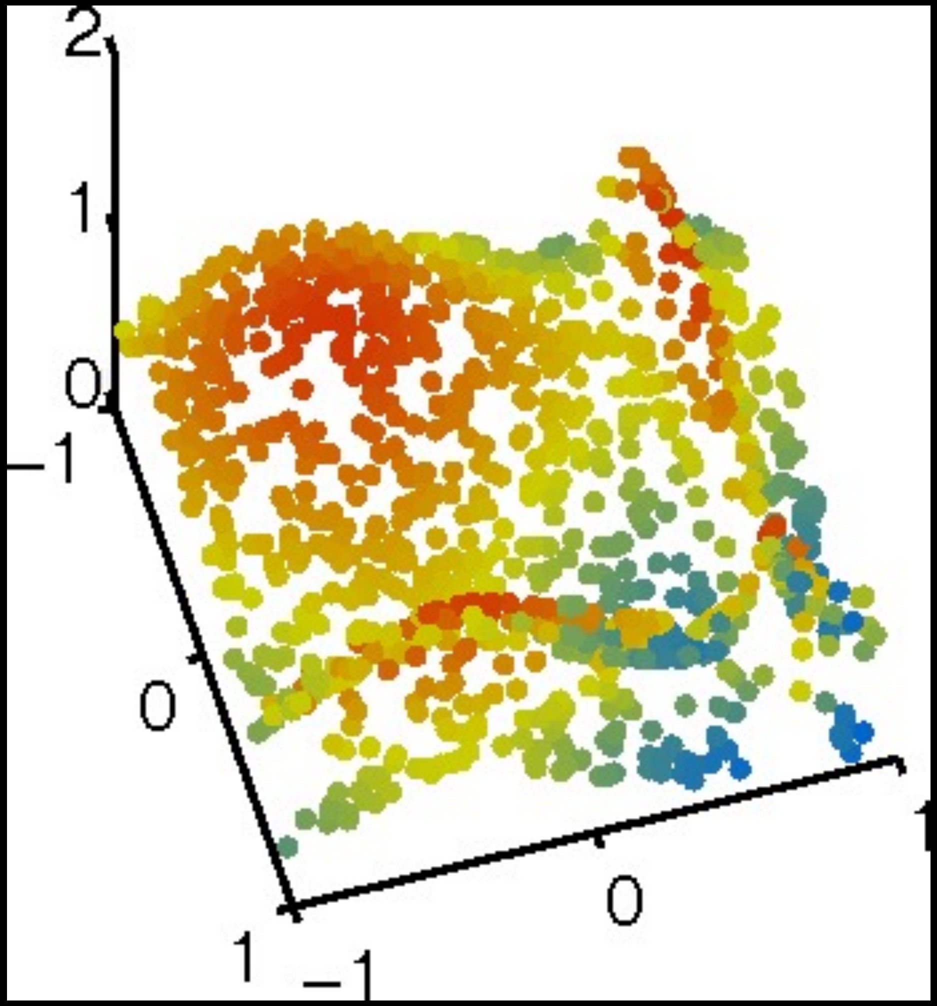
$f($
rhminh - high stable cloud formation
rhminl - low stable cloud formation
rligice - liquid drop size sea ice
rligland -liquid drop size land
rligocean-liquid drop size ocean
isf - ice stokes factor
capnc - particle density land/ocean
capnsi - particle density sea ice
capnw - particle density warm land
conke - evaporation
icritic - cold ice conversion
icritw - warm ice conversion
r3lcrit - liquid conversion
ricr - richardson number
c0_hk - shallow convection efficiency
cmftau - characteristic time scale
alfa - cloud downdraft flux
c0_zm - deep convection efficiency
dmpdz - fractional mass
ke - air entertainment
tau - consumption time scale
$) =$ flux



Visus image courtesy of K. Potter

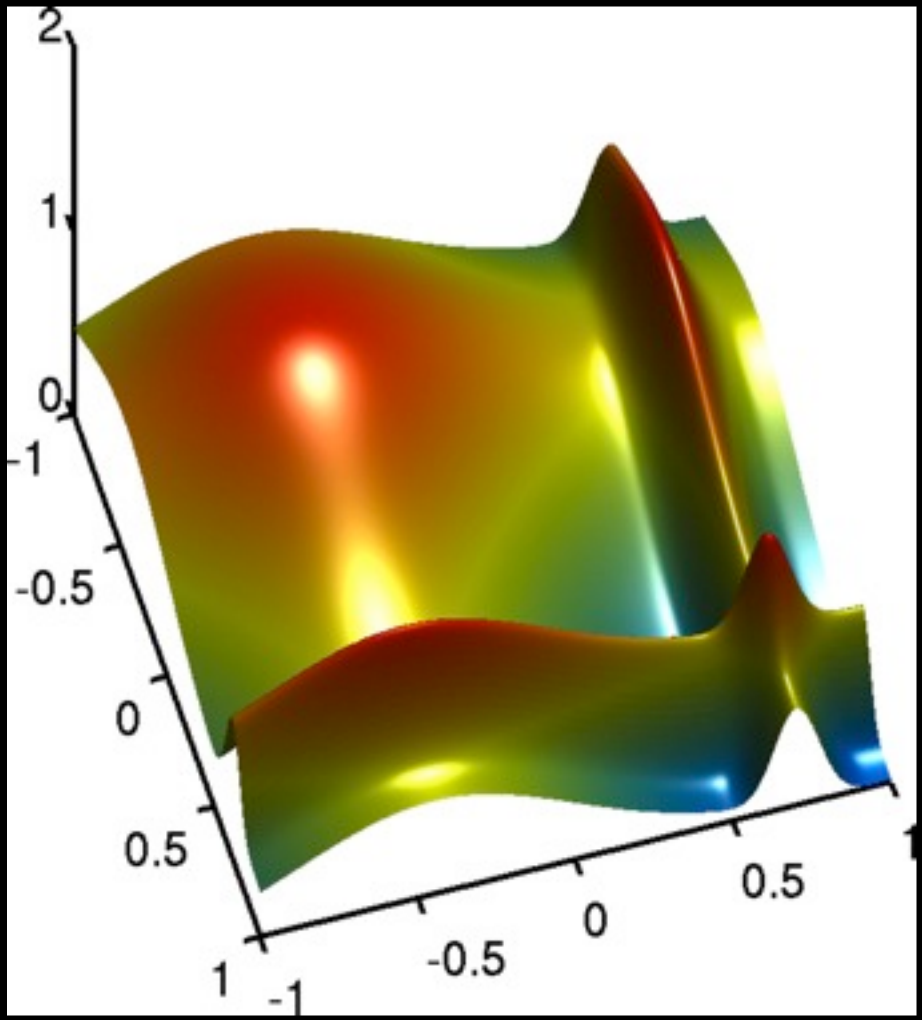
Visualization / Exploration

- Simple in 2D: Regression of y_i on $\mathbf{x}_i$



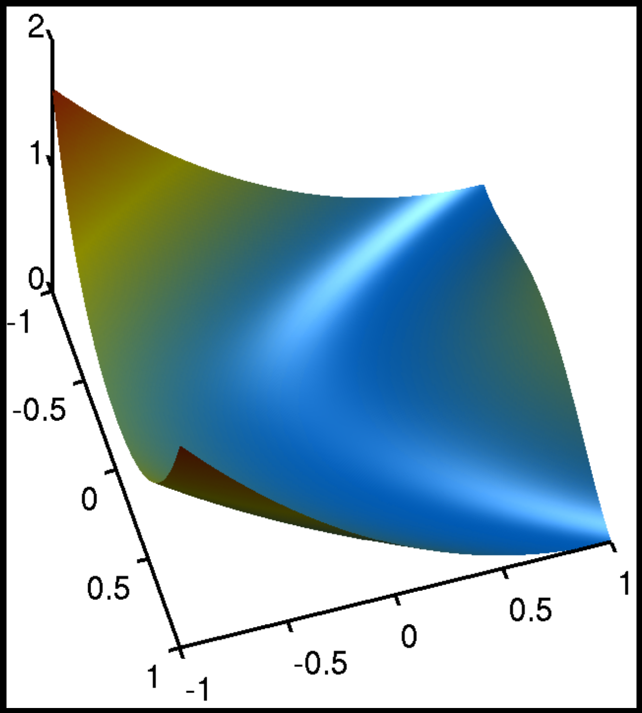
What to Visualize?

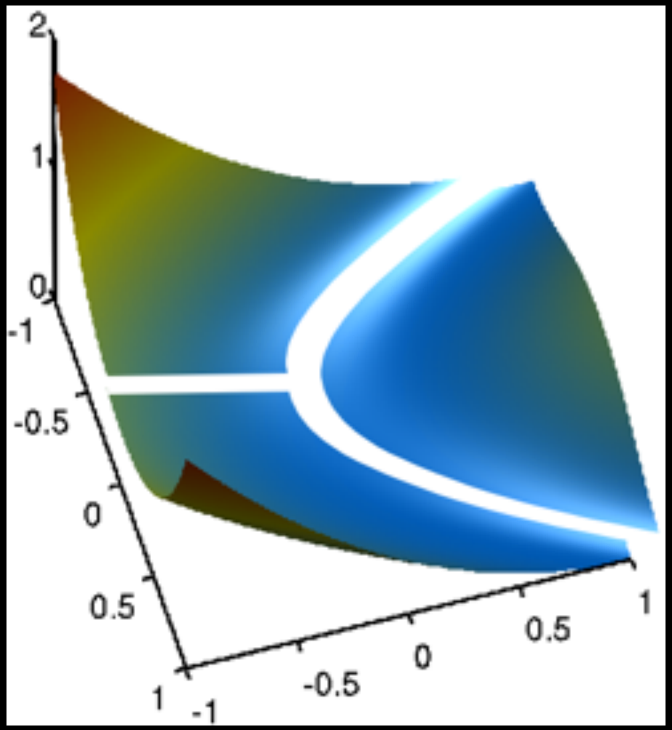
- **Goal:**
  - Summary representation of geometry of height field
- **Strategy:**
  - Extract important features
    - Extremal Points
    - Regions of similar behavior
  - Summary representation & Dimension reduction

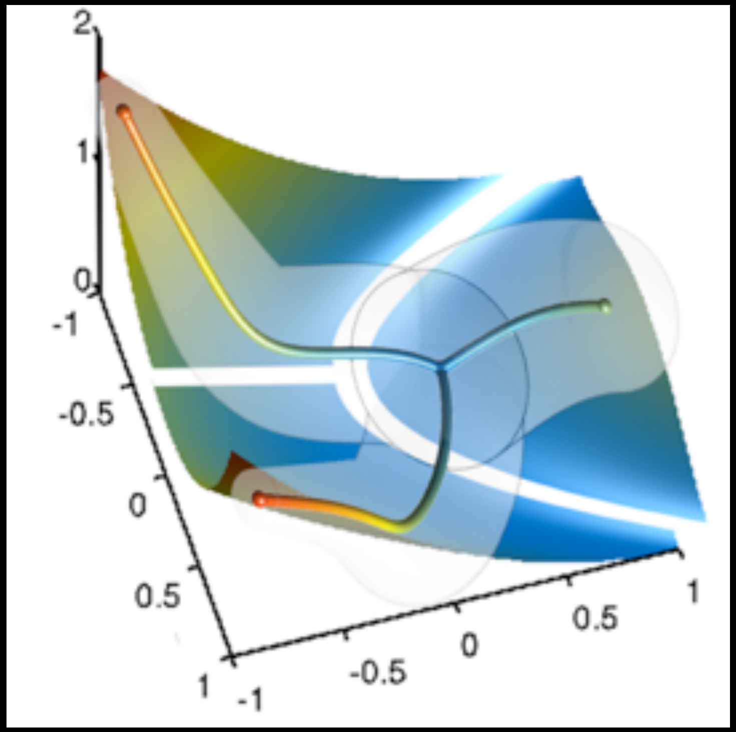


Methodology

- Segment *surface* (Morse-Smale complex)
- Summary for each partition of the segmentation
- Embed in 2D/3D for visualization

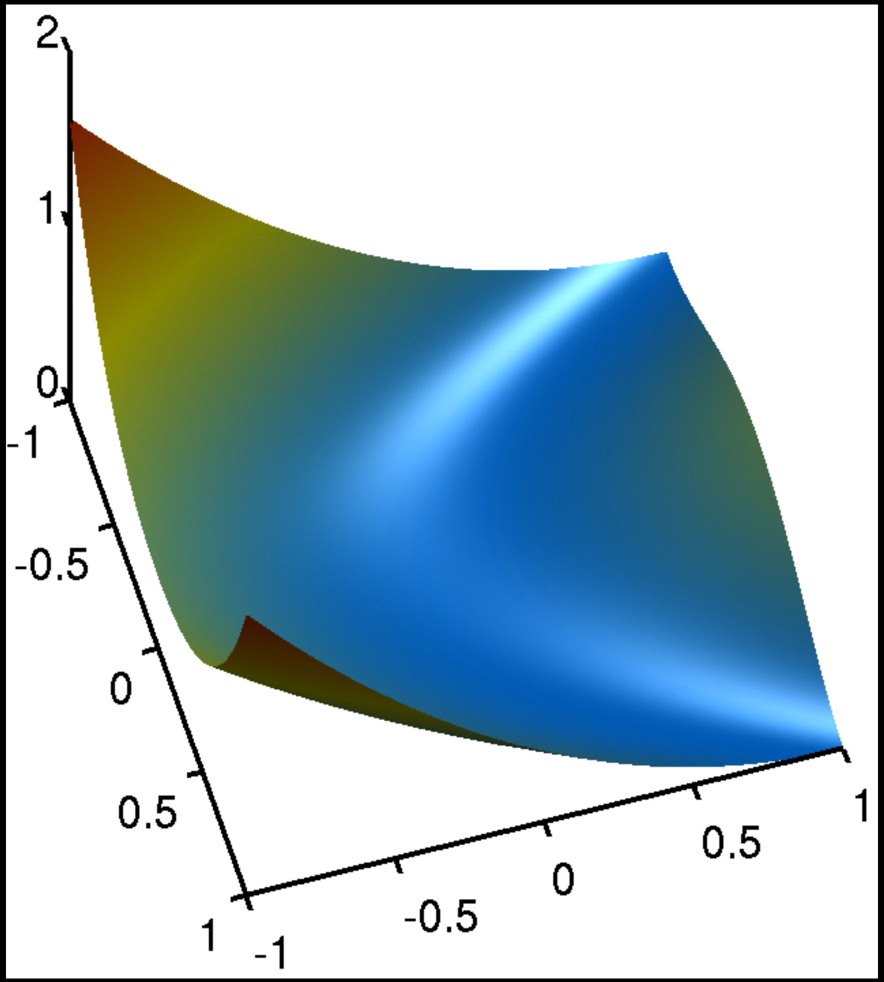


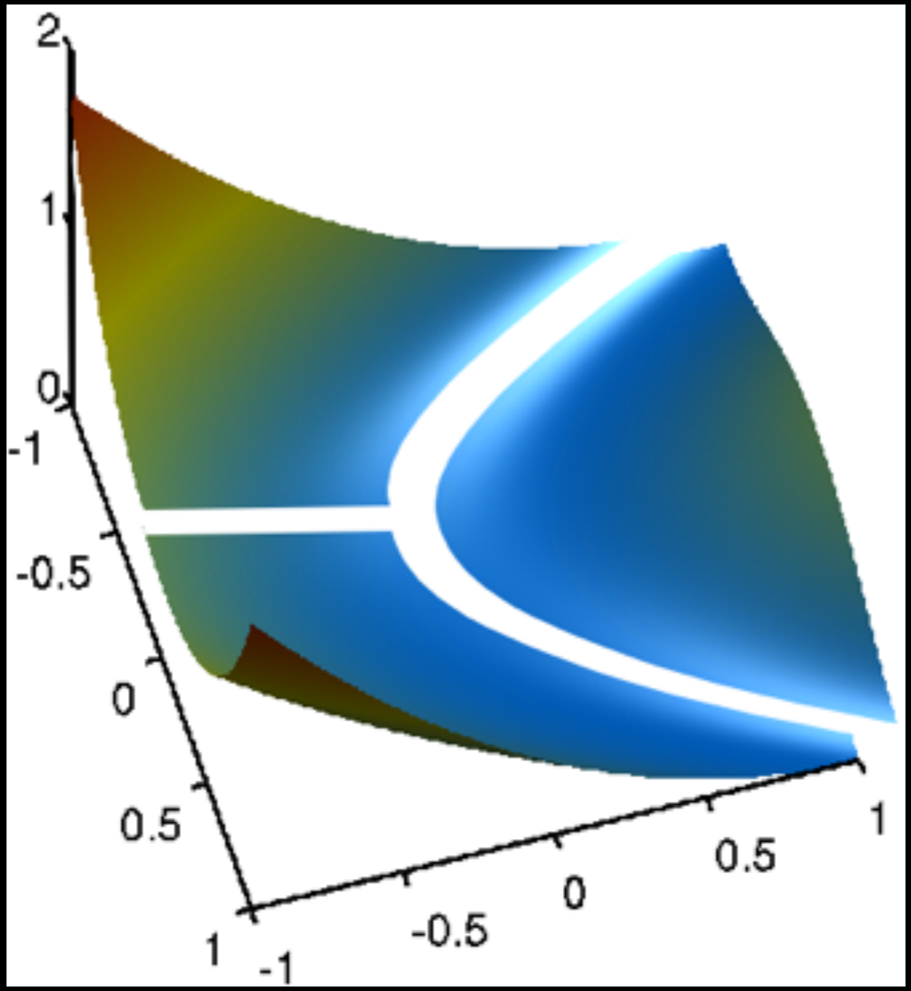




Morse-Smale Complex

- Segmentation of the domain
- Regions with similar gradient flow with a single source (minimum) and sink (maximum)





Segmentation Summaries

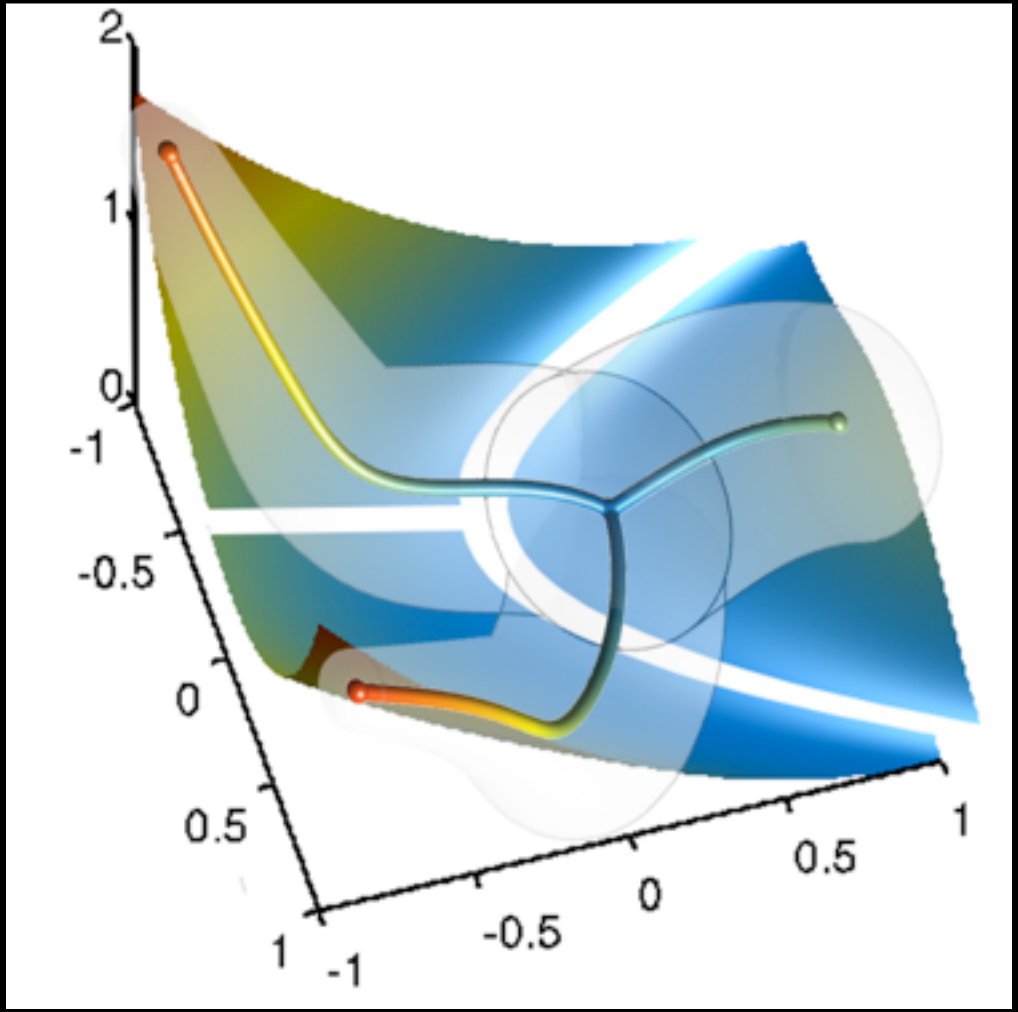
- 1D domain summary with regression curve for each partition

$$r_i(y) = E[X \in C_i | Y = y] \quad \in \mathbf{R}^d$$

  - Estimate with kernel regression

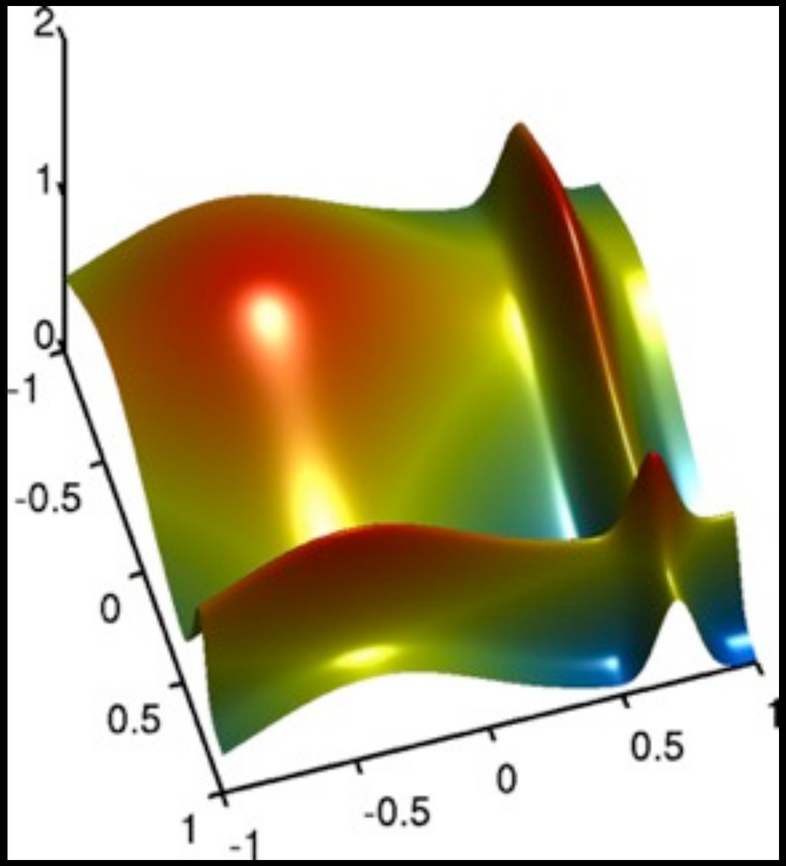
- **Not** estimating the surface

$$y_i = f(\mathbf{x}_i)$$



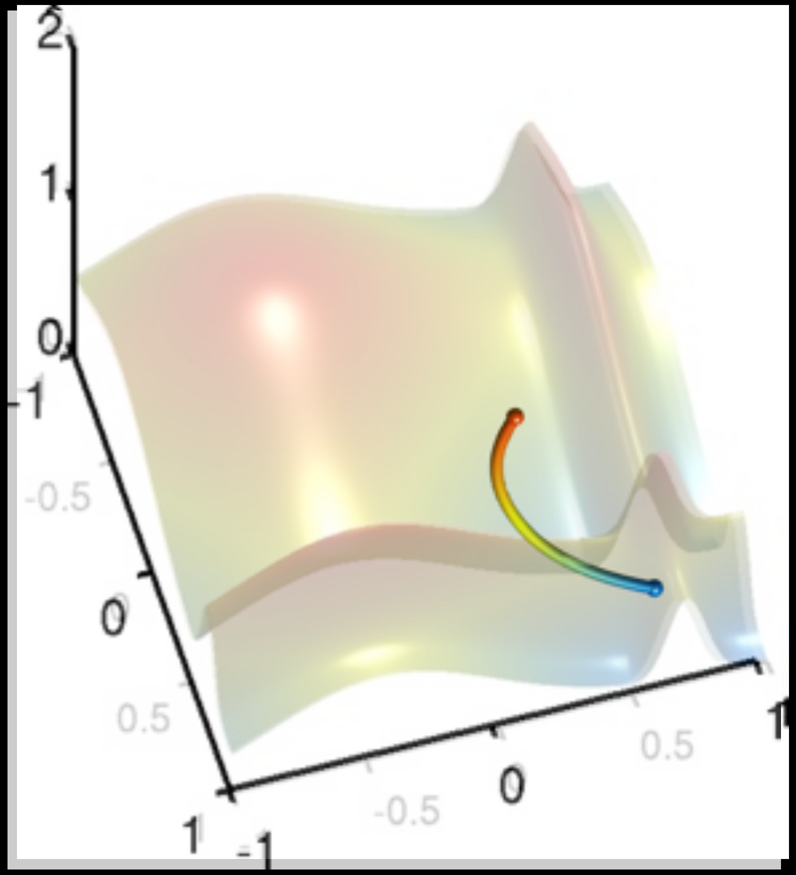
Segmentation Summaries

- Morse-Smale decomposition preserves extrema



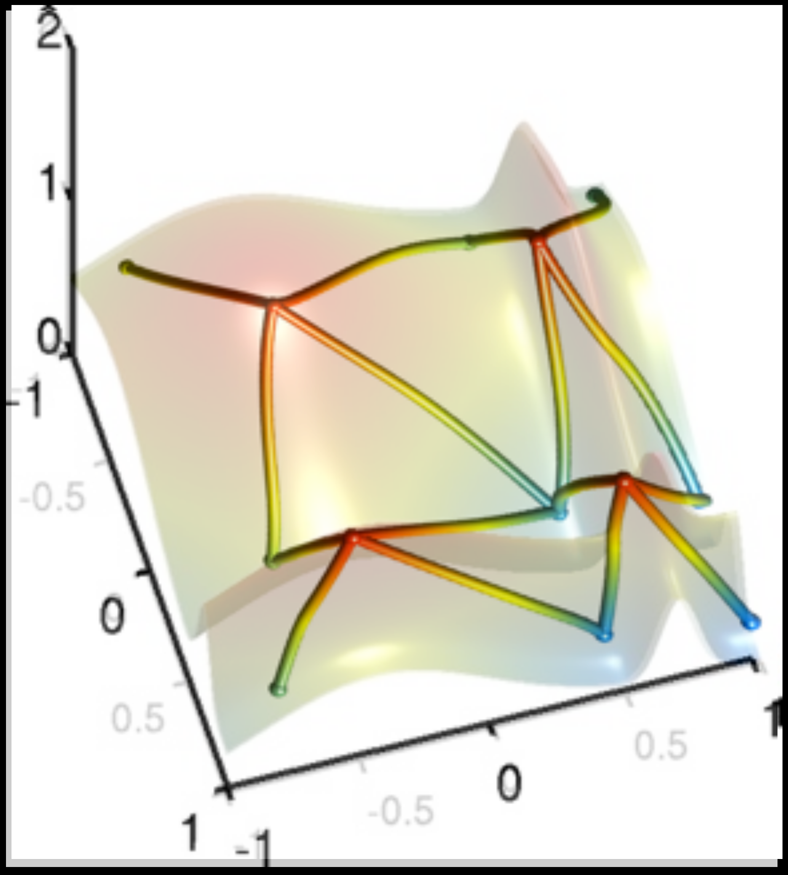
Complete Domain

$$g(y) = E[X|Y=y]$$



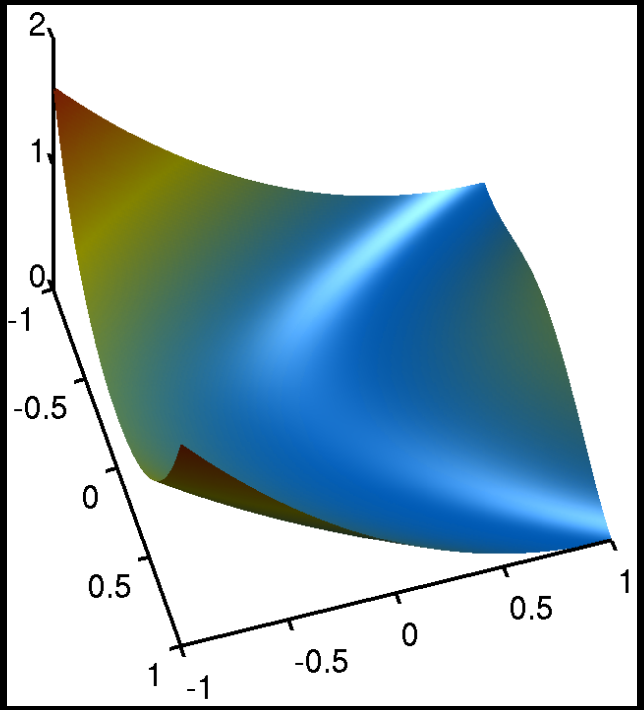
Per partition

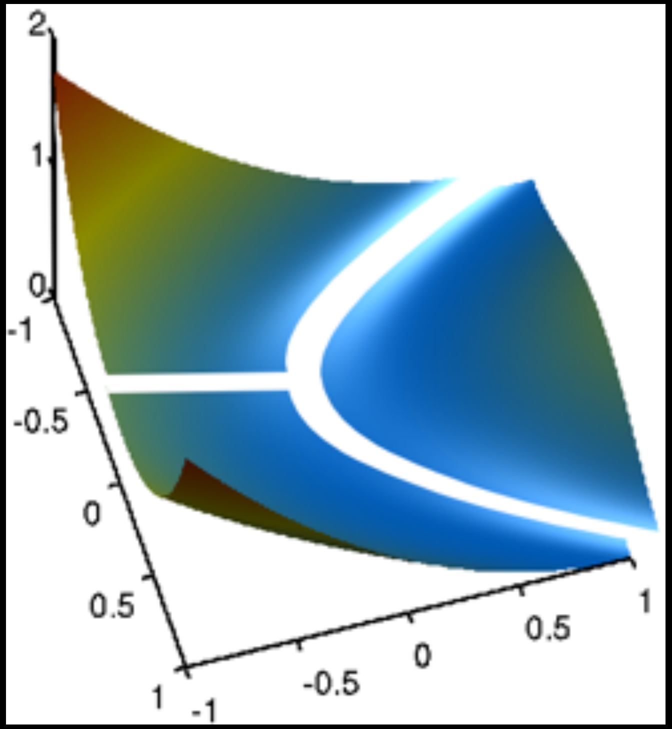
$$r_i(y) = E[X \in C_i | Y = y]$$

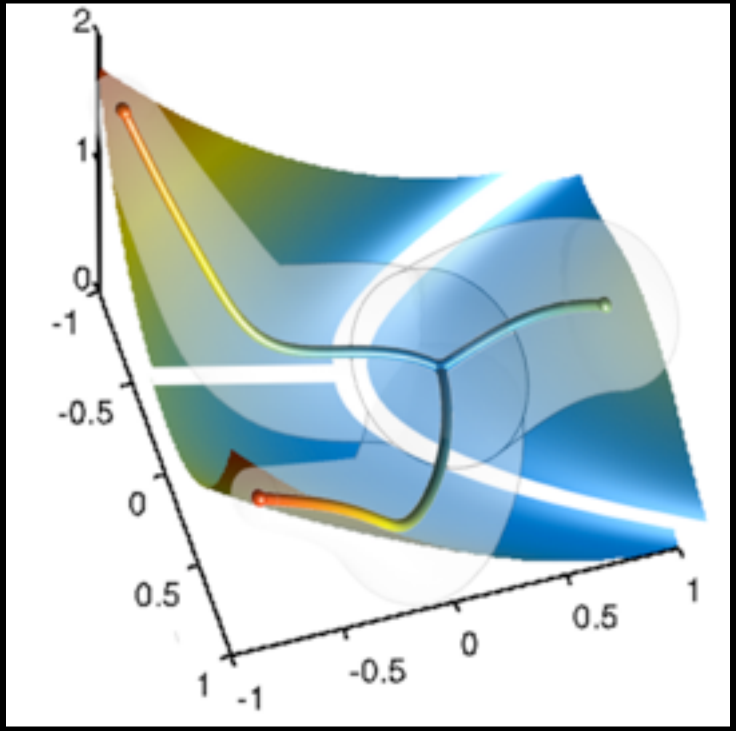


So Far ...

- Segmentation based on Morse-Smale Complex
  - Monotonic regions
- Kernel regression 1D summary for each partition
- Regression curves still in $\mathbf{R}^d$

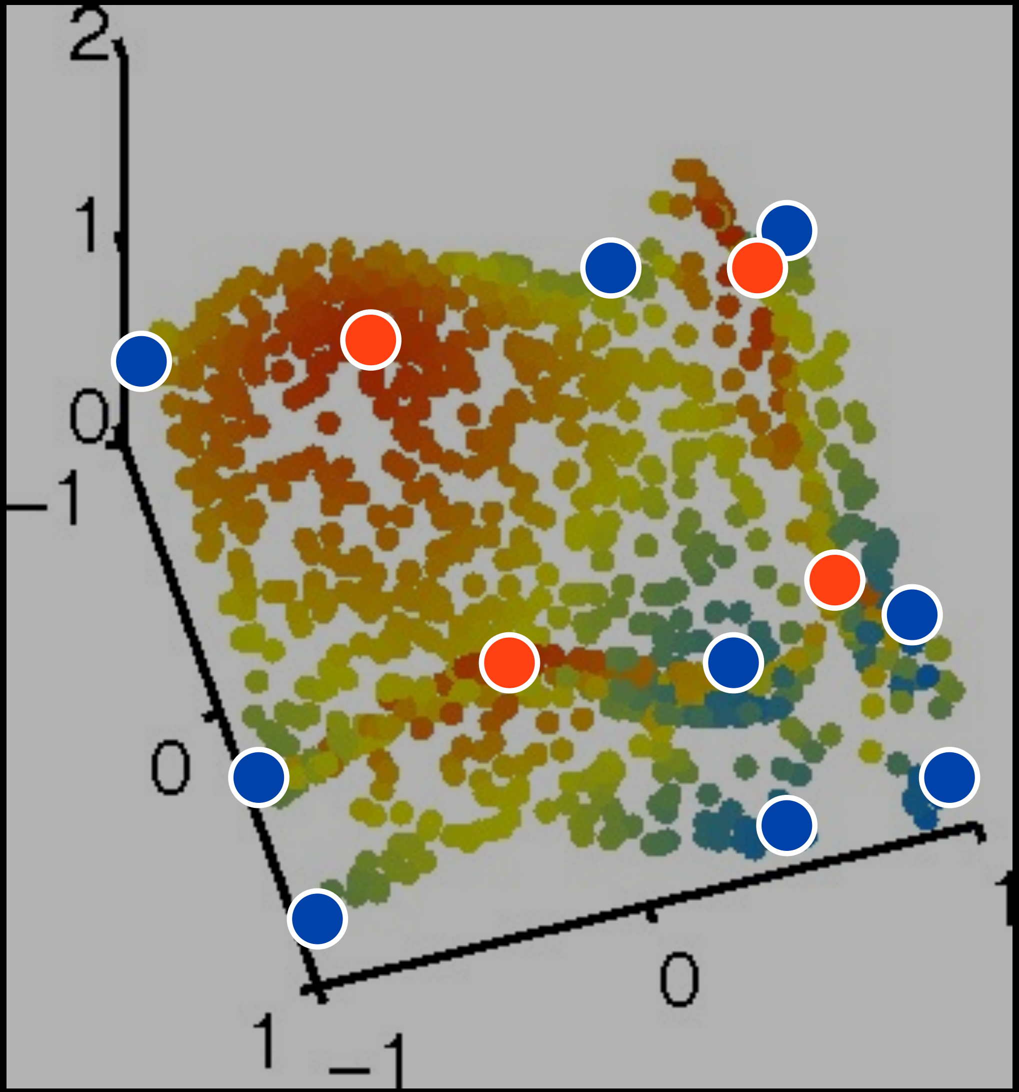






Embedding in 2D

- **1. Embed extremal points**
  - Preserve geometric location as best possible
  - No underlying structure - 2D principal component analysis

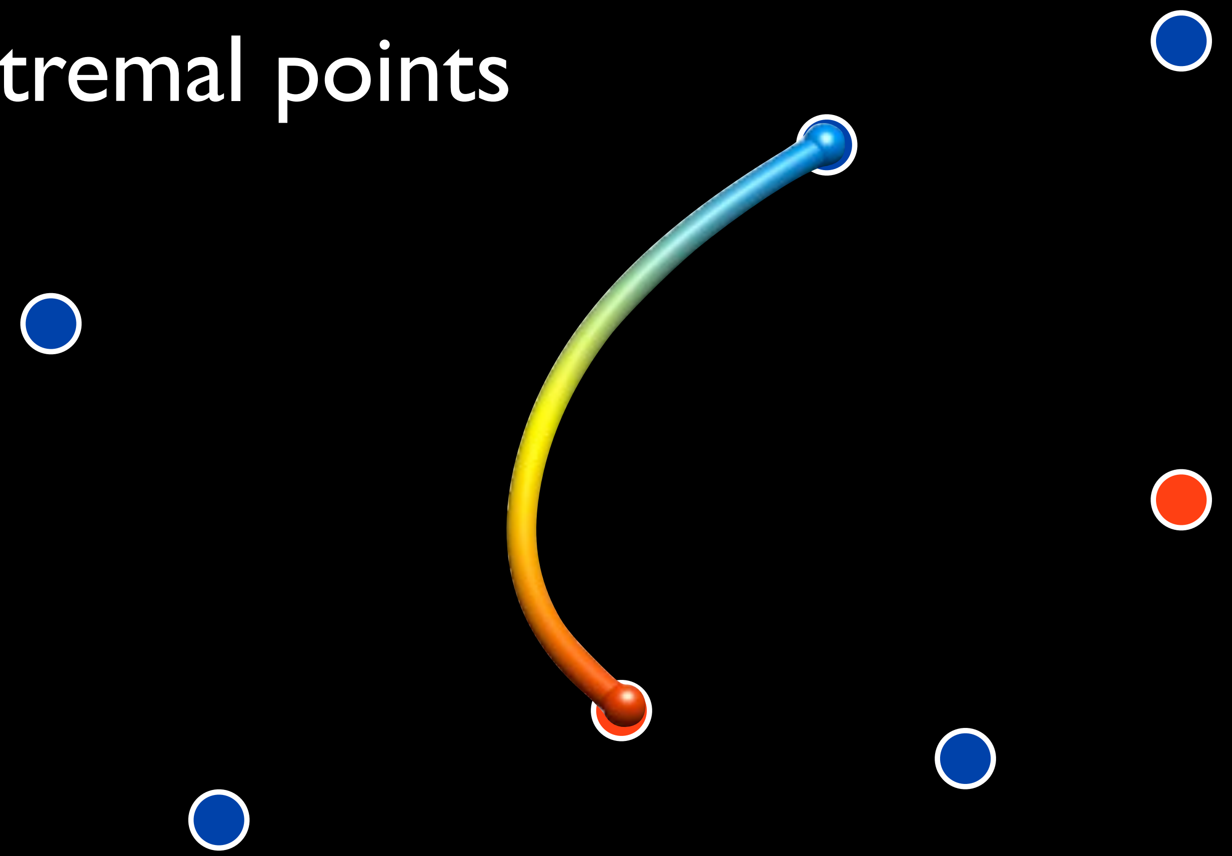


Embedding in 2D

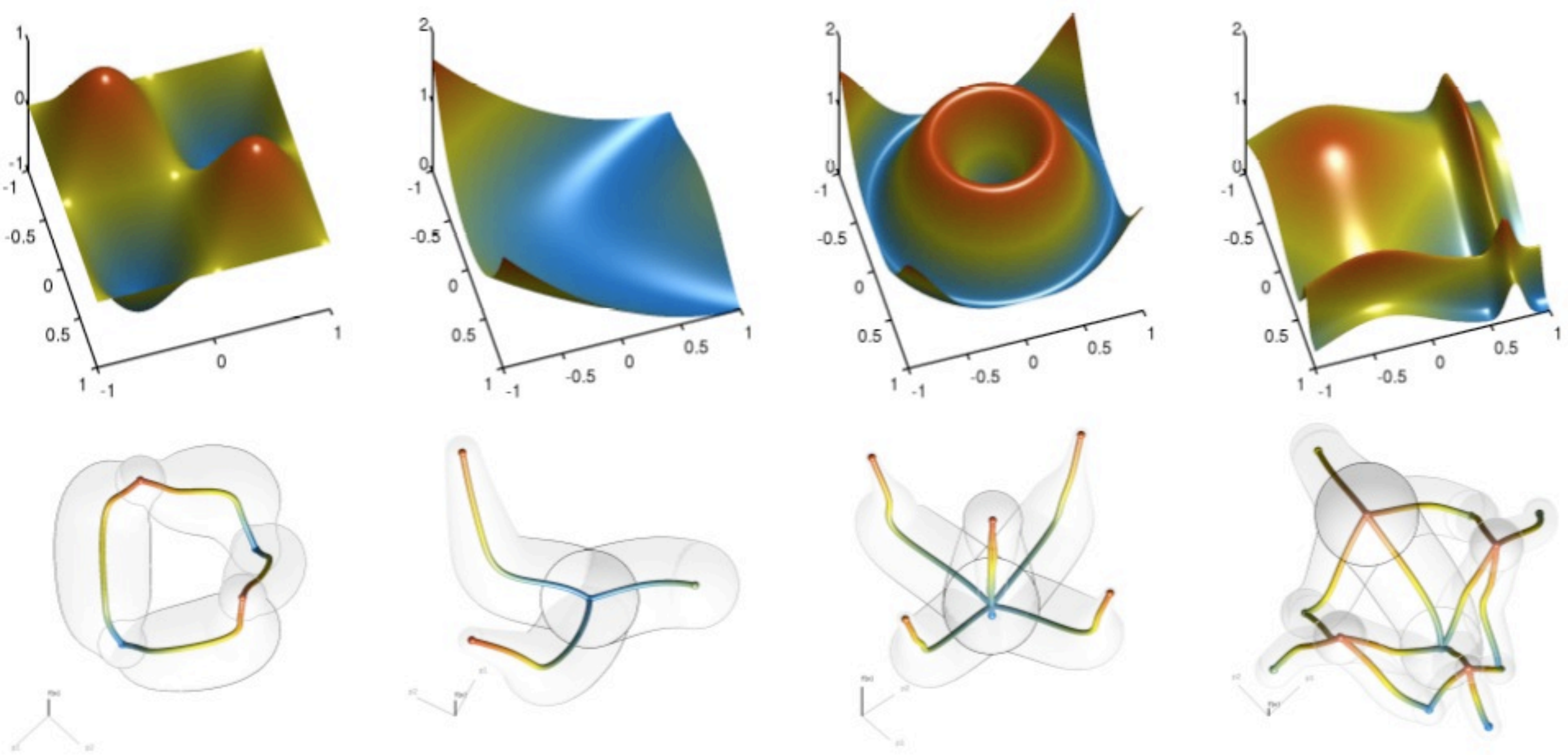
- **2. Embed regression curves separately**
  - Sample regression curves - piecewise linear approximation

$$r_i(y) = E[X \in C_i | Y = y]$$

  - Preserve as much of the geometry as possible - 2D principal component analysis for each curve
  - Transform to match to extremal points

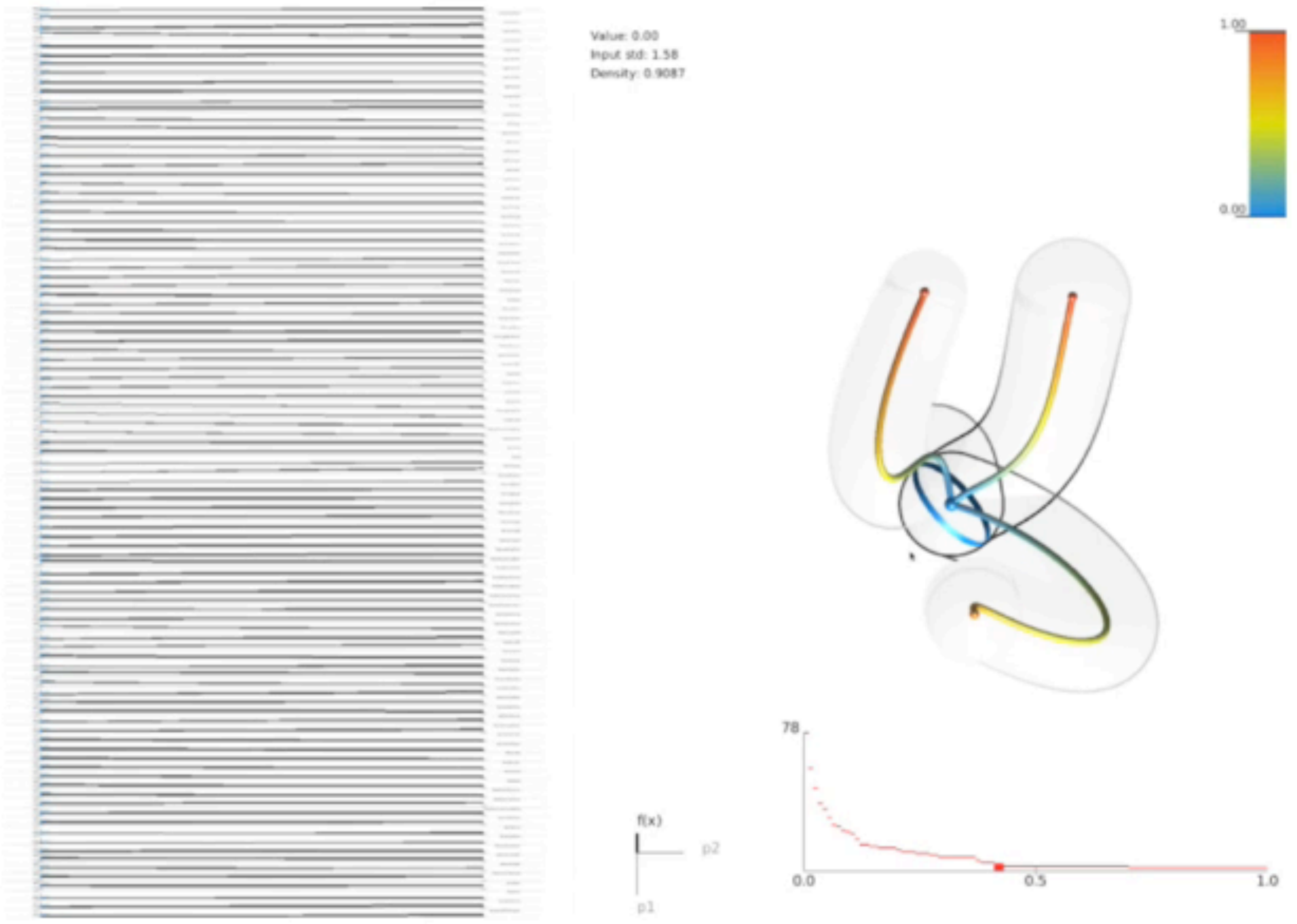


Some 2D Examples



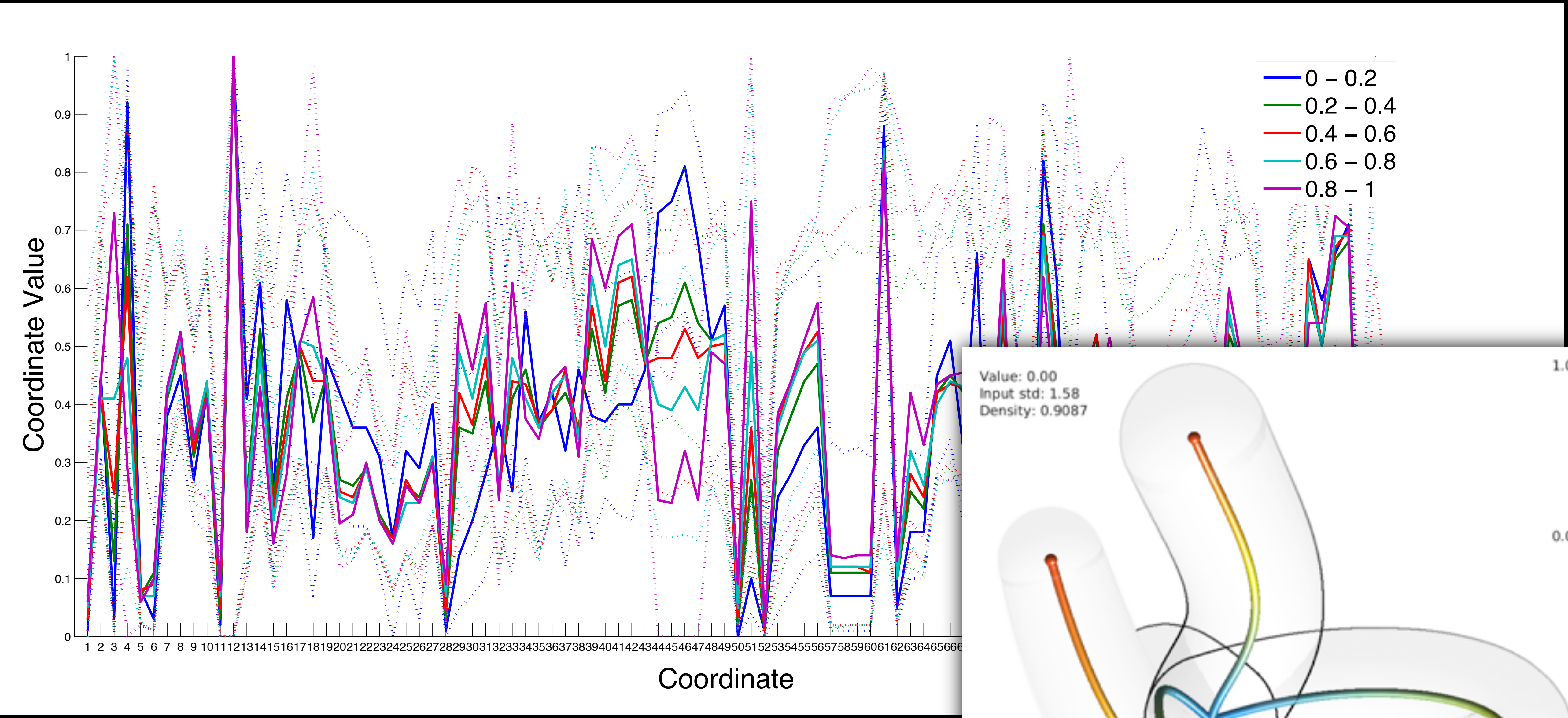
An intuitive description of the overall structure of the function

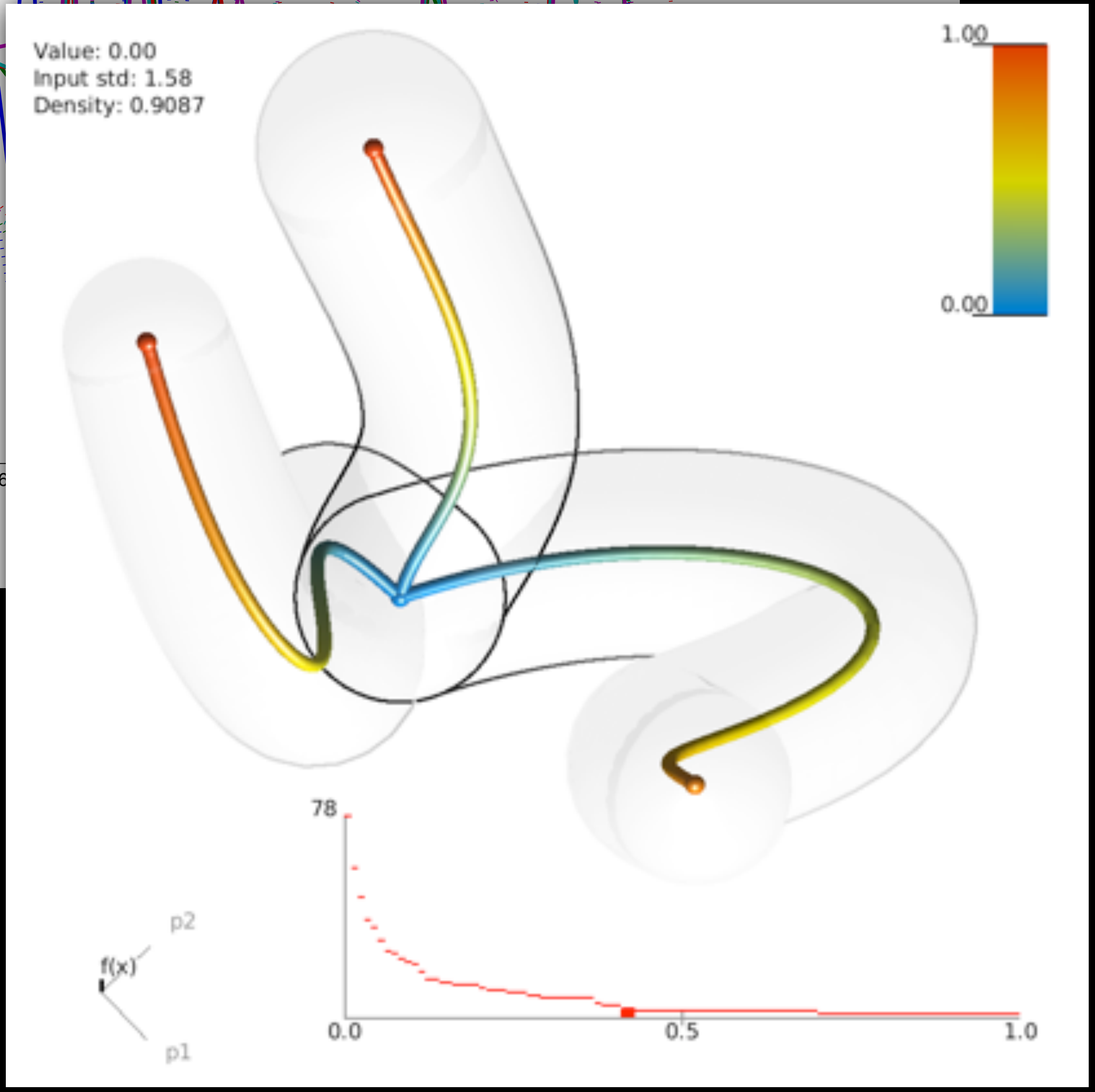
Crime data



Multiple peaks indicate different factors leading to high crime rates.

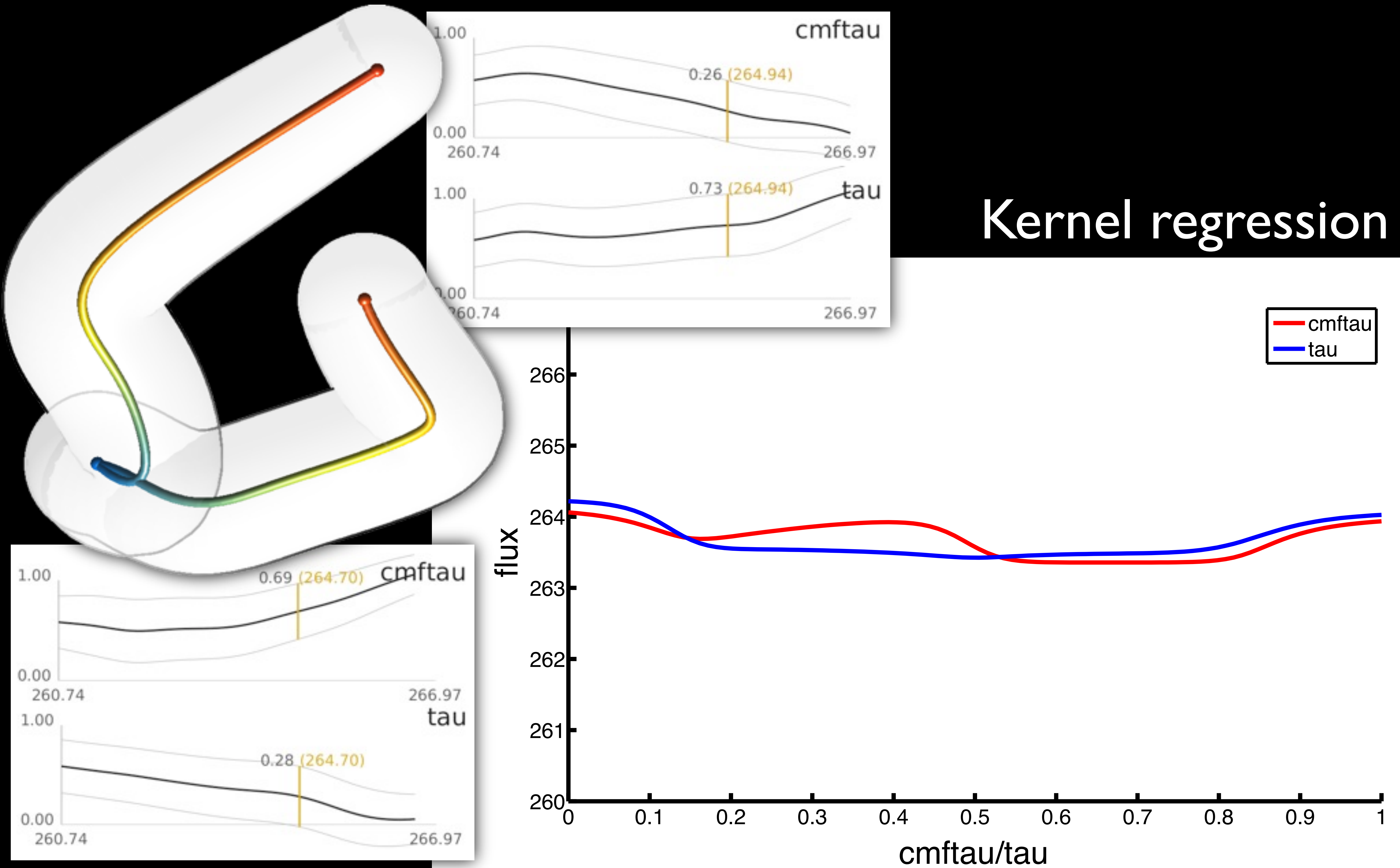
Crime data





Parallel coordinates

Climate Simulations



administrivia

- **for final projects that have been approved, email me:**
  - working title
  - group member names
  - two or three sentence description

PROJECTS

OVERVIEW

Your final project will be in one of two flavors: **programming,** where you will implement a visualization system of your own; or **analysis,** where you will analyze a data set of your choosing using a variety of existing visualization systems. In both cases you will be responsible for acquiring data, creating a data and task abstraction of the problem your visualization system(s) will address, surveying existing visualization methods, and analyzing the results -- depending on your type of project you will emphasize these aspects to different degrees.

You may do the projects individually or in teams of two. The total amount of work done must be commensurate with the size of the group. Note that research novelty is not a requirement for a course project.

PROGRAMMING PROJECT: For a programming project, you will implement a visualization system you design and develop yourself. Common varieties of programming projects are:

- *problem-driven design studies* based around a specific dataset and questions of interest
- *technique-driven explorations* of how to carry out specific visual encoding or interaction ideas
- *implementations of previous published algorithms*

You may use existing components as the base for your system, as well as any programming language or toolkit of choice. A (not necessarily complete) list of visualization languages and toolkits can be found on the resources page.

ANALYSIS PROJECT: For an analysis project, you will analyze a dataset/problem of your choosing using a variety of existing visualization tools, as well as analyze the strengths and weaknesses of those tools and discuss in detail whether they are effective for the data and task that you have chosen. These projects are *problem-driven design studies* in nature. No serious programming is required, so this option is suitable for non-CS students. You may need to write some scripts to change data formats, however. A (not necessarily complete) list of visualization tools can be found on the resources page. This style of project will require a much more extensive survey of previous work than a programming project.

IMPORTANT DATES

-meetings, **Feb 14 - 23**
-proposals due, **March 9 at noon**
-project update presentation, **March 27, 29, and April 3**
-final presentations, **May 1 from 1-3pm**
-process books due, **May 3 at noon**

PROPOSAL

Written proposal length: several pages
Written proposal format: PDF

Prior to submitting your proposal you will meet with me in person to discuss your project at least once before submitting a proposal. **It may take more than one meeting for me to sign off that you're ready to move on to the proposal writeup stage.** You'll need to meet with me between **Feb 14 and Feb 23** at the latest, and earlier would be better.

I advise that you start by thinking about a domain/dataset/visualization method that you are interested in exploring more deeply. The key is to find some domain and task that both interests you and presents an opportunity for infovis. That is, there is some task where a human needs to understand the structure of a large dataset. You're welcome to link the infovis project to another class or research project. Keep in mind that you're submitting a proposal, not a specification -- it's natural that your plans will change somewhat as you refine your ideas. But your proposal should be based on an idea that we've discussed and I've approved. When you come talk to me about your proposal, I'll give you some pointers to background reading in the area of your interest.

LAST TIME

THE SMARTPHONE CHALLENGE

part 6

- get back into large groups
- share sketches
- turn in abstraction and individual sketches

L13: Tabular Data

REQUIRED READING

Hierarchical Parallel Coordinates for Exploration of Large Datasets

Ying-Huey Fua, Matthew O. Ward and Elke A. Rundensteiner
Computer Science Department
Worcester Polytechnic Institute
Worcester, MA 01609
{yingfua,matt,rundenst}@cs.wpi.edu *

Abstract

Our ability to accumulate large, complex (multivariate) data sets has far exceeded our ability to effectively process them in search of patterns, anomalies, and other interesting features. Conventional multivariate visualization techniques generally do not scale well with respect to the size of the data set. The focus of this paper is on the interactive visualization of large multivariate data sets based on a number of novel extensions to the parallel coordinates display technique. We develop a multiresolutional view of the data via hierarchical clustering, and use a variation on parallel coordinates to convey aggregation information for the resulting clusters. Users can then navigate the resulting structure until the desired focus region and level of detail is reached, using our suite of navigational and filtering tools. We describe the design and implementation of our hierarchical parallel coordinates system which is based on extending the XmdvTool system. Lastly, we show examples of the tools and techniques applied to large (hundreds of thousands of records) multivariate data sets.

Keywords: Large-scale multivariate data visualization, hierarchical data exploration, parallel coordinates.

1 Introduction

- Dimensional embedding techniques, such as dimensional stacking [16] and worlds within worlds [6].
- Dimensional subsetting, such as scatterplots [5].
- Dimensional reduction techniques, such as multidimensional scaling [20, 15, 29], principal component analysis [12] and self-organizing maps [14].

Most of these techniques do not scale well with respect to the size of the data set. As a generalization, we postulate that any method that displays a single entity per data point invariably results in overlapped elements and a convoluted display that is not suited for the visualization of large data sets. The quantification of the term "large" varies and is subject to revision in sync with the state of computing power. For our present application, we define a large data set to contain 10^6 to 10^9 data elements or more.

Our research focus extends beyond just data display, incorporating the process of data exploration, with the goal of interactively uncovering patterns or anomalies not immediately obvious or comprehensible. Our goal is thus to support an active process of discovery as opposed to passive display. We believe that it is only through data exploration that meaningful ideas, relations, and subsequent inferences may be extracted from the data. The major hurdles we need to overcome are the problems of display density/clutter (too

Metric-Based Network Exploration and Multiscale Scatterplot

Yves Chiricota*
Université du Québec à Chicoutimi, Canada

Fabien Jourdan, Guy Melançon†
LIRMM UMR CNRS 5506, Montpellier, France

Abstract

We describe an exploratory technique based on the direct interaction with a 2D modified scatterplot computed from two different metrics calculated over the elements of a network. The scatterplot is transformed into an image by applying standard image processing techniques resulting into blurring effects. Segmentation of the image allows to easily select *patches* on the image as a way to extract sub-networks. We were inspired by the work of Wattenberg and Fisher [21] showing that the blurring process builds into a multiscale perceptual scheme, making this type of interaction intuitive to the user. We explain how the exploration of the network can be guided by the visual analysis of the blurred scatterplot and by its possible interpretations.

CR Categories: I.3.6 [Computer Graphics]: Methodology and Techniques—Interaction Techniques I.3.3 [Computer Graphics]: Picture / Image Generation—Viewing algorithms I.4.3 [Image Processing]: Enhancement—Smoothing

Keywords: Graph navigation, exploration, scatterplot, multiscale perceptual organization, clustering, filtering, blurring

1 Introduction

Part of the research activity in Information Visualization is devoted to exploratory techniques [4, 12]. Indeed, when designing a tool it is important to distinguish whether the user is facing familiar data and is actually using it for a specific task (annotating it or consulting it, for instance) or if she/he is exploring the data trying to find patterns

is to specify a threshold by moving the cursor down (or up) and filter out nodes or edges with a value above (or not exceeding) the threshold. This hiding method gains effectiveness when coupled with a colour map as the elements that are filtered out have a lighter hue and/or lesser intensity, are thiner, etc.

The use of multiple range sliders can help the exploration of a dataset by filtering elements based on a combination of criterion. Williamson and Schneiderman [24] have successfully applied this technique when exploring a real estate database, enabling a user to specify a price range and number of bedrooms, for instance. Barry Becker's MineSet [2] is a tool supporting the exploration of multidimensional databases, helping the user to navigate the data through the selection of range values on several dimensions.

It is unclear whether range selectors are as effective when dealing with less intuitive metrics. What if the values correspond to a *theoretical measure* computed over all nodes of the network, such as for example the so-called clustering index used to define small world networks [22, 23] or the pagerank index of web pages [17] ? What if the values are unevenly distributed over the range they cover ? How should a user manipulate the range selectors to correctly monitor the threshold (filter) ? These observations become even more relevant when dealing with two-dimensional metrics. Situations that are hardly predictable may appear where one slider requires finer tuning depending on the values that were selected using the other. Section 2 provides examples and a more detailed discussion on these issues that were one of the starting point of our work.

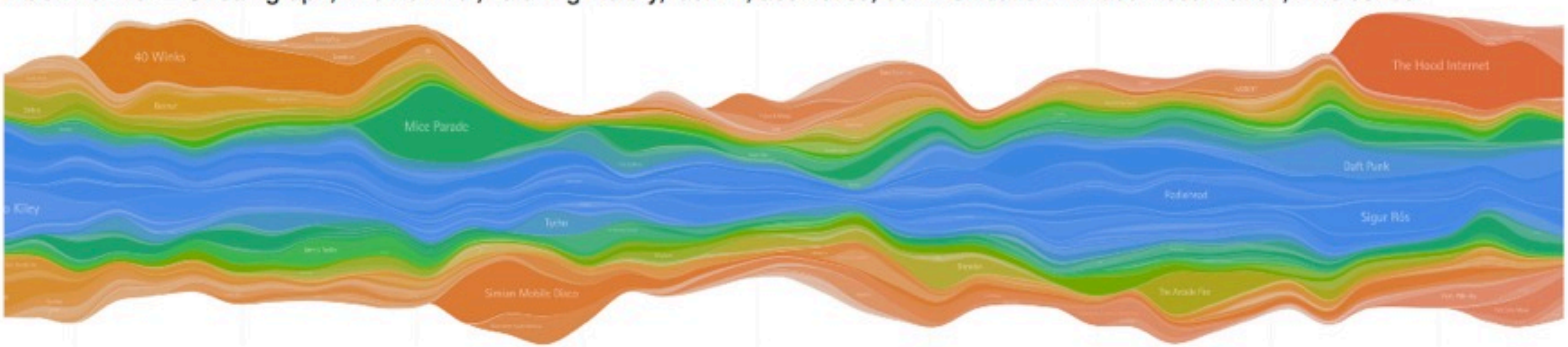
The technique we put forward in this paper gives the user direct access to the 2D set of values through a *modified* scatterplot view. More precisely, the view the user acts on is obtained from the actual

Stacked Graphs – Geometry & Aesthetics

Lee Byron & Martin Wattenberg

Abstract — In February 2008, the New York Times published an unusual chart of box office revenues for 7500 movies over 21 years. The chart was based on a similar visualization, developed by the first author, that displayed trends in music listening. This paper describes the design decisions and algorithms behind these graphics, and discusses the reaction on the Web. We suggest that this type of complex layered graph is effective for displaying large data sets to a mass audience. We provide a mathematical analysis of how this layered graph relates to traditional stacked graphs and to techniques such as ThemeRiver, showing how each method is optimizing a different "energy function". Finally, we discuss techniques for coloring and ordering the layers of such graphs. Throughout the paper, we emphasize the interplay between considerations of aesthetics and legibility.

Index Terms — Streamgraph, ThemeRiver, listening history, last.fm, aesthetics, communication-minded visualization, time series.



1 Introduction

In February 2008, The New York Times stirred up a debate. The famous newspaper is no stranger to controversy, but this time the issue was not political bias or anonymous sources—it was an unusual graph of movie ticket sales. On information design blogs, opinions of the chart ranged from "fantastic" to "unsavory." Meanwhile, on other online forums and blogs, hundreds of people posted insights and questions spurred by the visualization.

The story of the design process and algorithms behind this engag-

graphic and accompanying online interactive visualization of the box office revenue for 7500 movies over a 21-year period.

In this paper we first provide a case study of the New York Times and last.fm visualizations. We pay special attention to the response on the web and the role of aesthetics in the appeal of visualizations. Second, we perform a detailed analysis of the algorithms that define these graphs. A key theme is the role of aesthetics in visualization design, and the process and trade-offs necessary to create engaging