High Dimensionality I

Manifold Methods

Talk Overview

- Define Concepts and Problems
- Paper I: *Charting A Manifold* by Matthew Brand
- Paper 2: *Maximum Likelihood Estimation of Intrinsic Dimension* by Elizaveta Levina and Peter J. Bickel
- Discussion

Common Scientific Problem

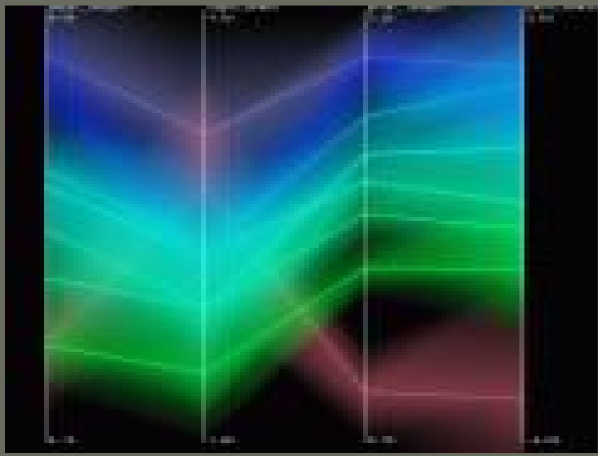
- Make N observations
- Make a series of M measurements per observation

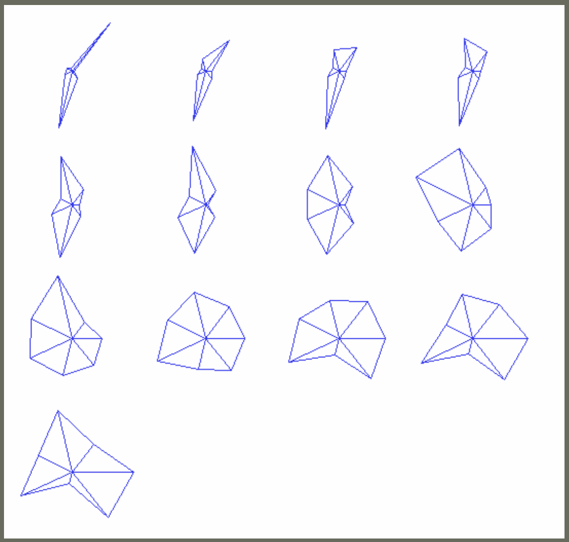
Common Scientific Problem

- Make N observations
- Make a series of M measurements per observation
- NOW WHAT?

Visualization

- Directly Visualize Dimensions
  - Parallel Coordinates
  - Glyphs
  - Star Coordinates
  - Etc.





Problem: Hidden Factors

True Dimensionality < Measured Dimensionality

Example

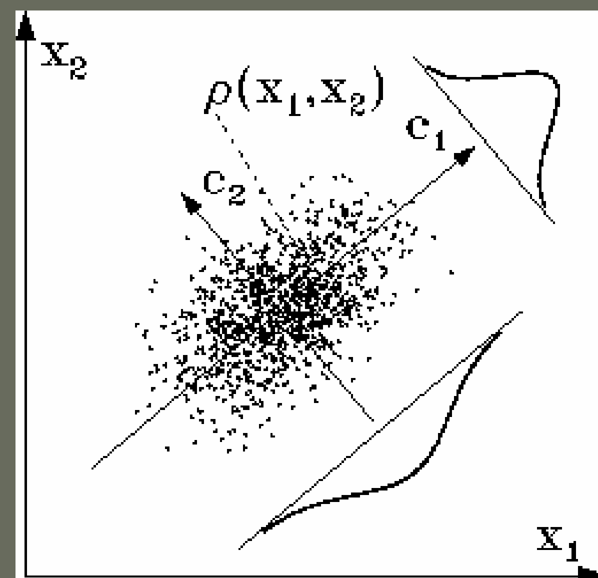
- Rotating head

- Large Number of Measured Dimensions

- Low Number of “Intrinsic” Dimensions

Solution: Dimensionality Reduction

- Find the true dimensionality
- PCA – Find Largest Axes of Variability And Construct a Plane



- MDS – Embed points based on Distances

Problem

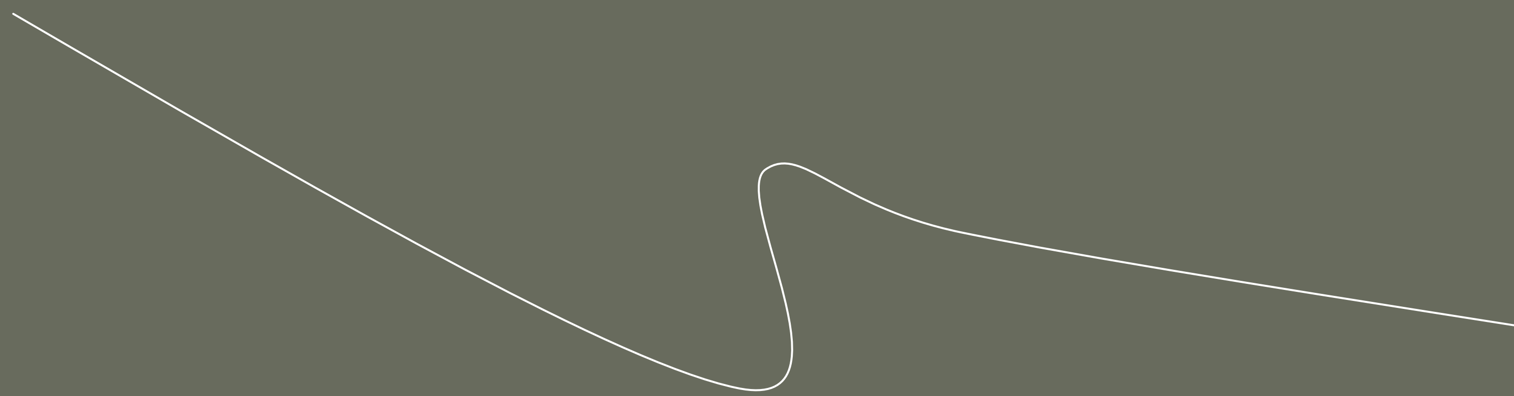
MANIFOLDS

What is a Manifold?

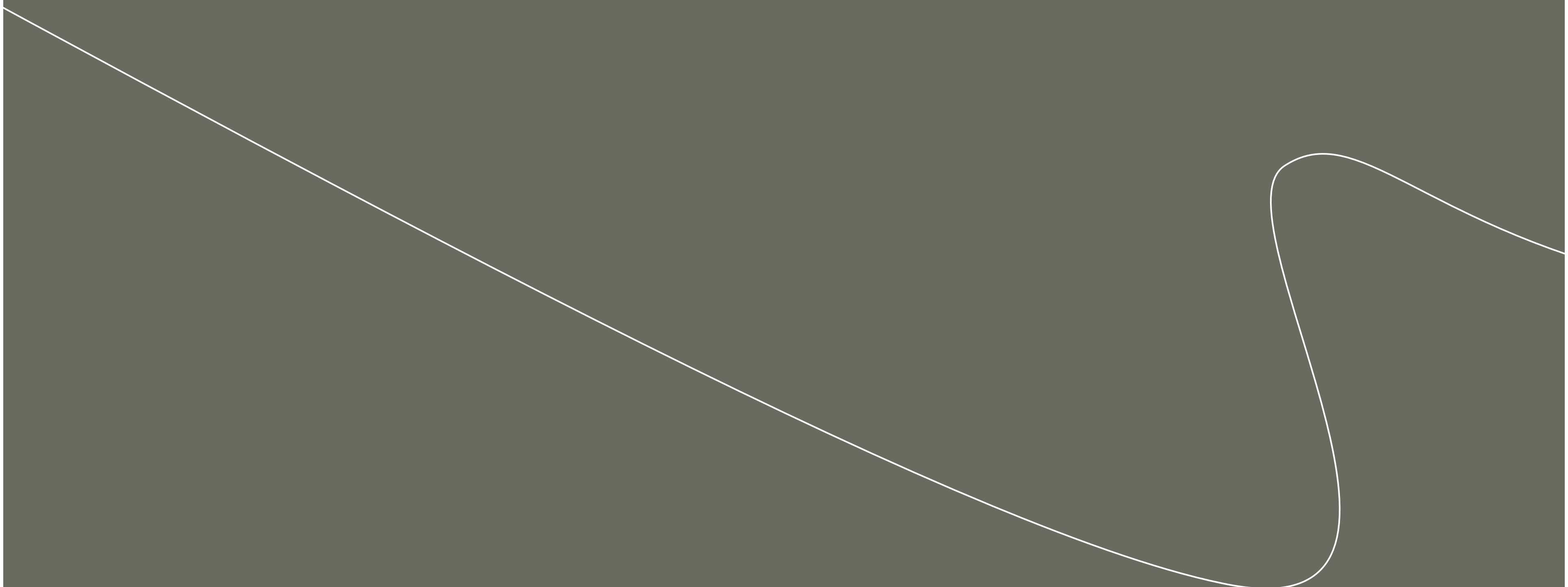
- A topological space that looks locally like the Euclidean space R^n

What is a Manifold?

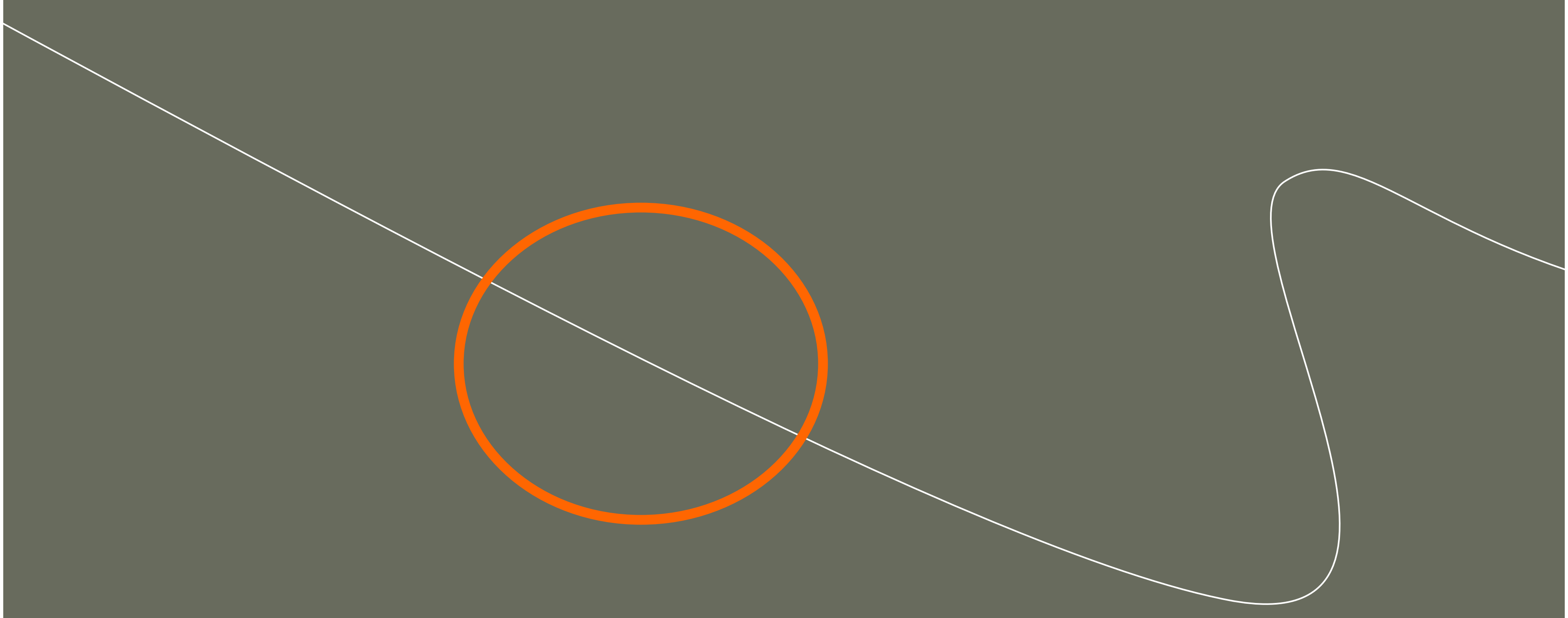
What is a Manifold?



What is a Manifold?

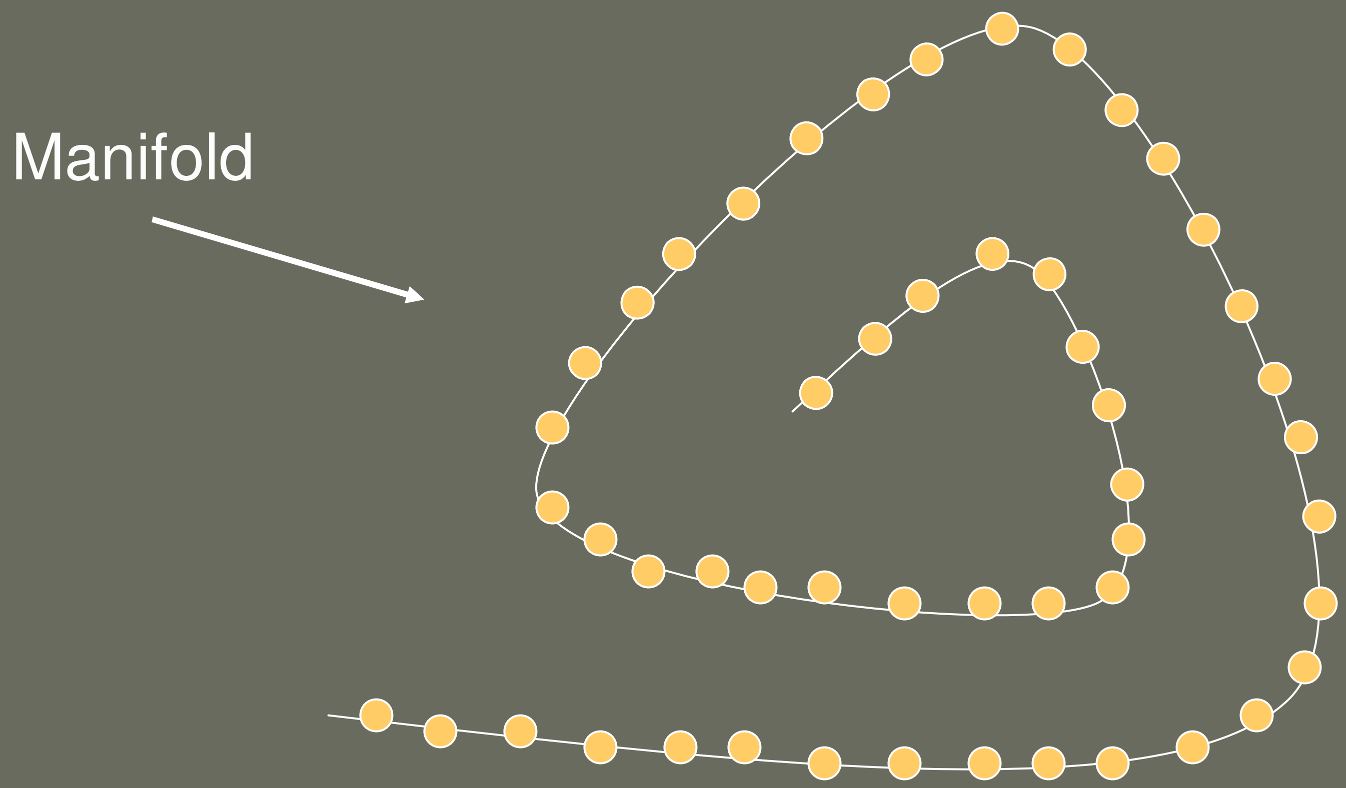


What is a Manifold?



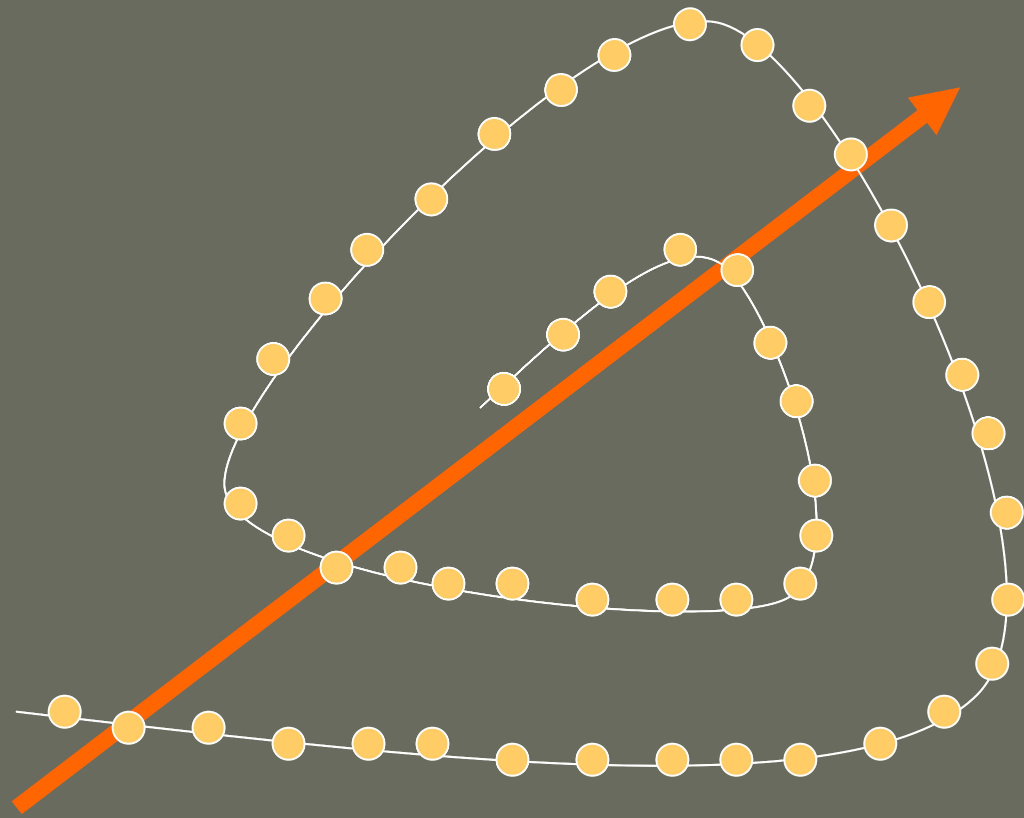
So What’s the Problem?

So What's the Problem?



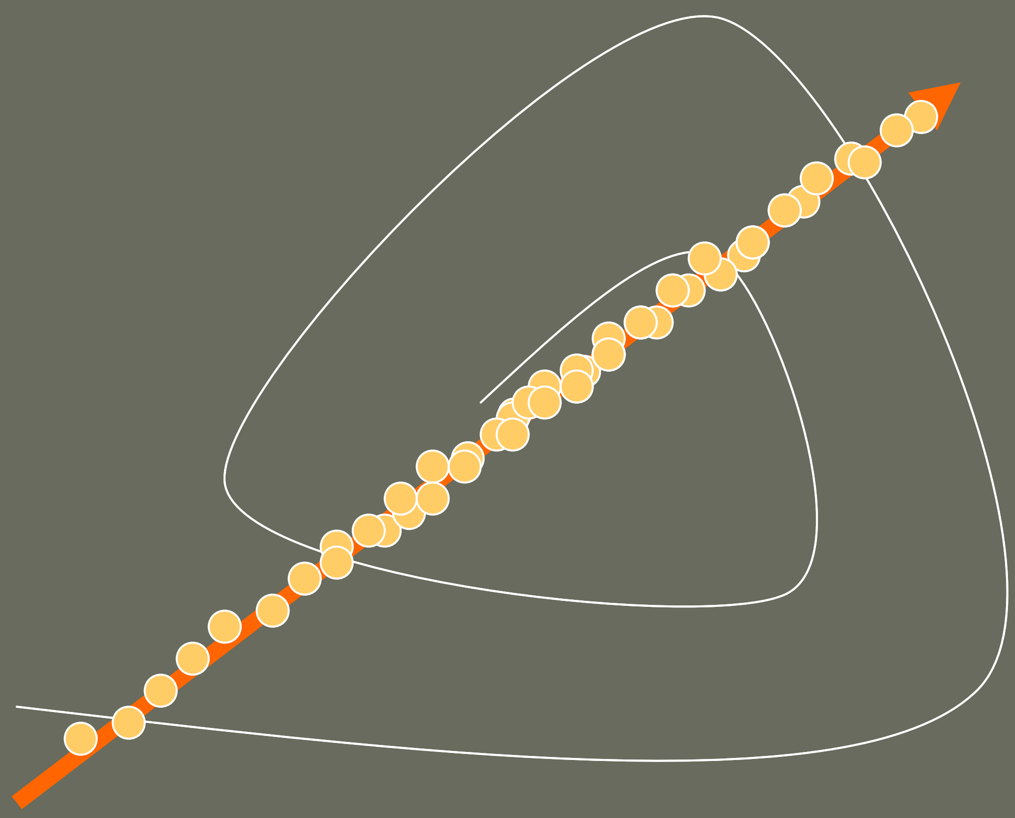
So What's the Problem?

PCA



So What's the Problem?

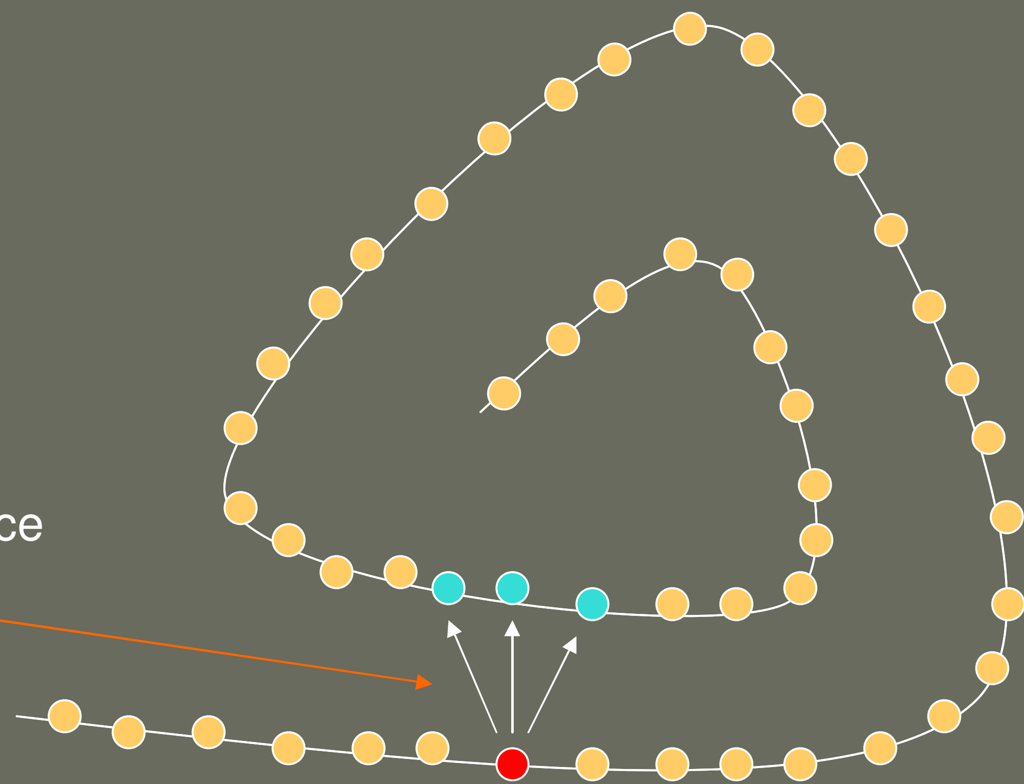
PCA



So What's the Problem?

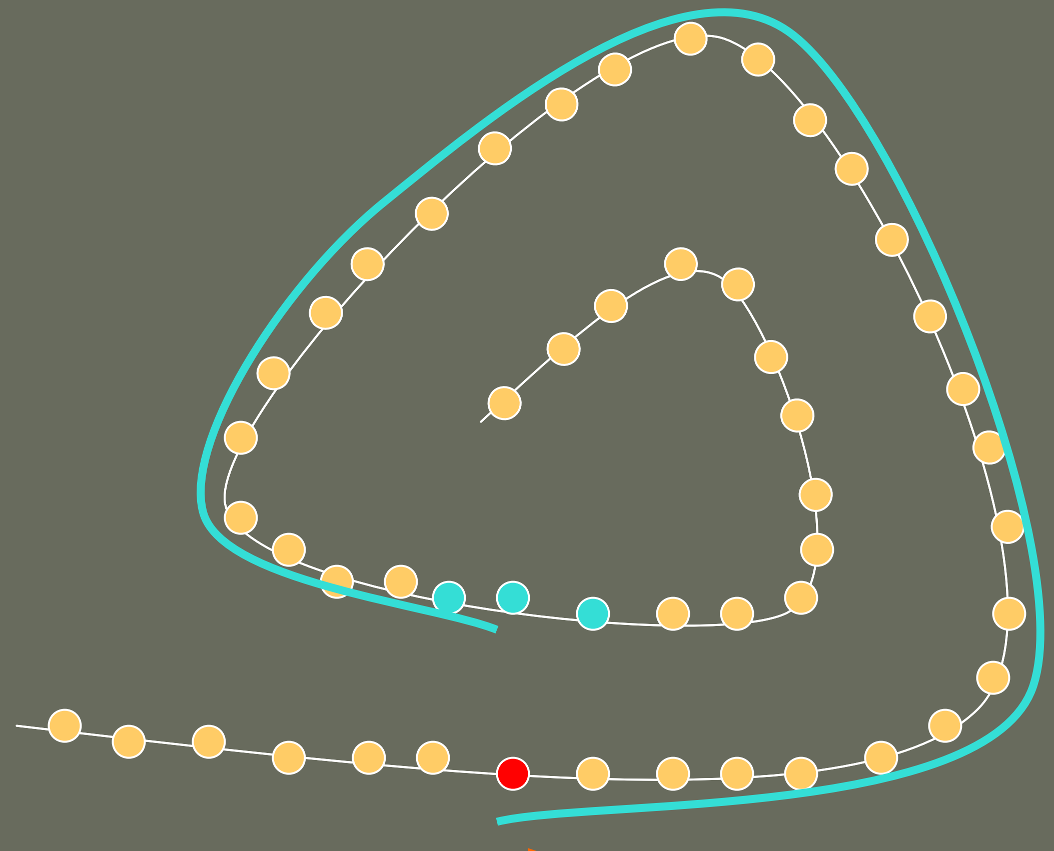
MDS

Euclidean Distance



So What's the Problem?

MDS



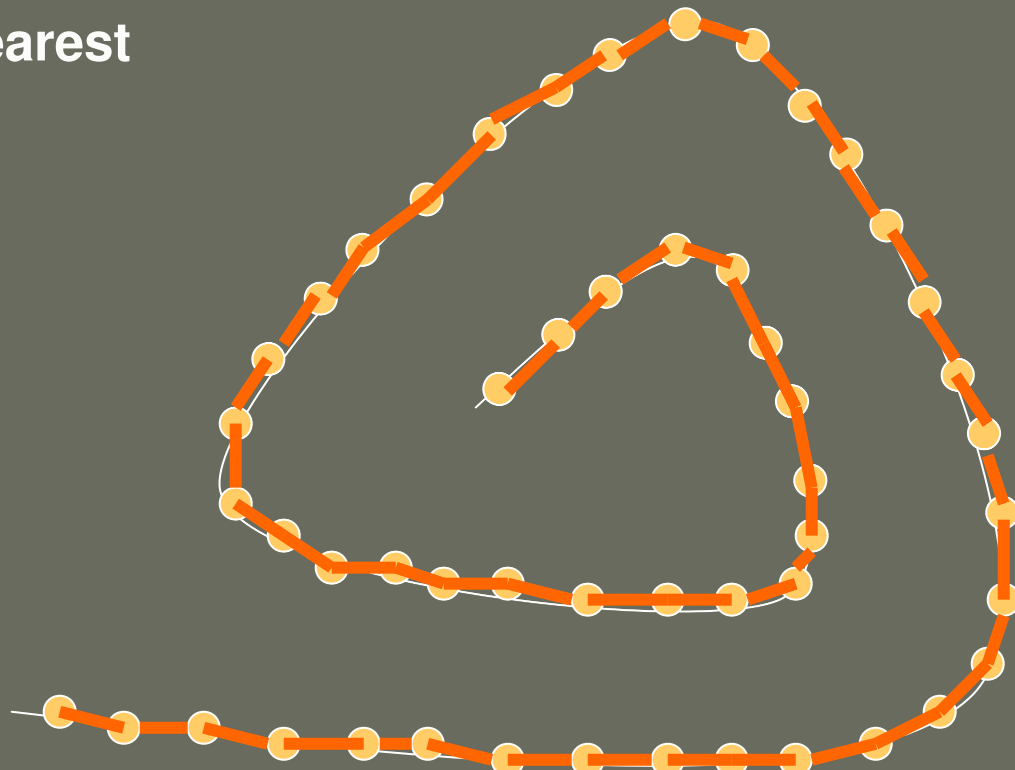
"Real" Distance

“Classic” Manifold Method: ISOMAP

1. **Link To Nearest Neighbors**

“Classic” Manifold Method: ISOMAP

1. **Link To Nearest Neighbors**



“Classic” Manifold Method: ISOMAP

1. **Link To Nearest Neighbors**

2. **Compute Distances THROUGH Graph**

“Classic” Manifold Method: ISOMAP

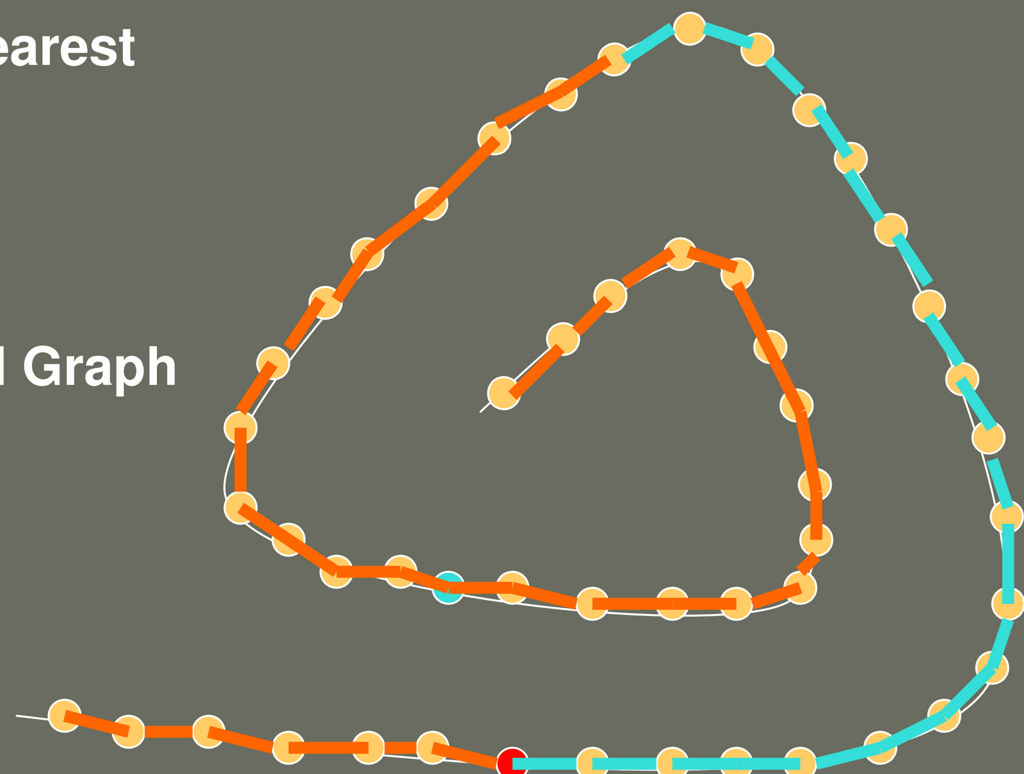
1. **Link To Nearest Neighbors**

2. **Compute Distances THROUGH Graph**

"Classic" Manifold Method: ISOMAP

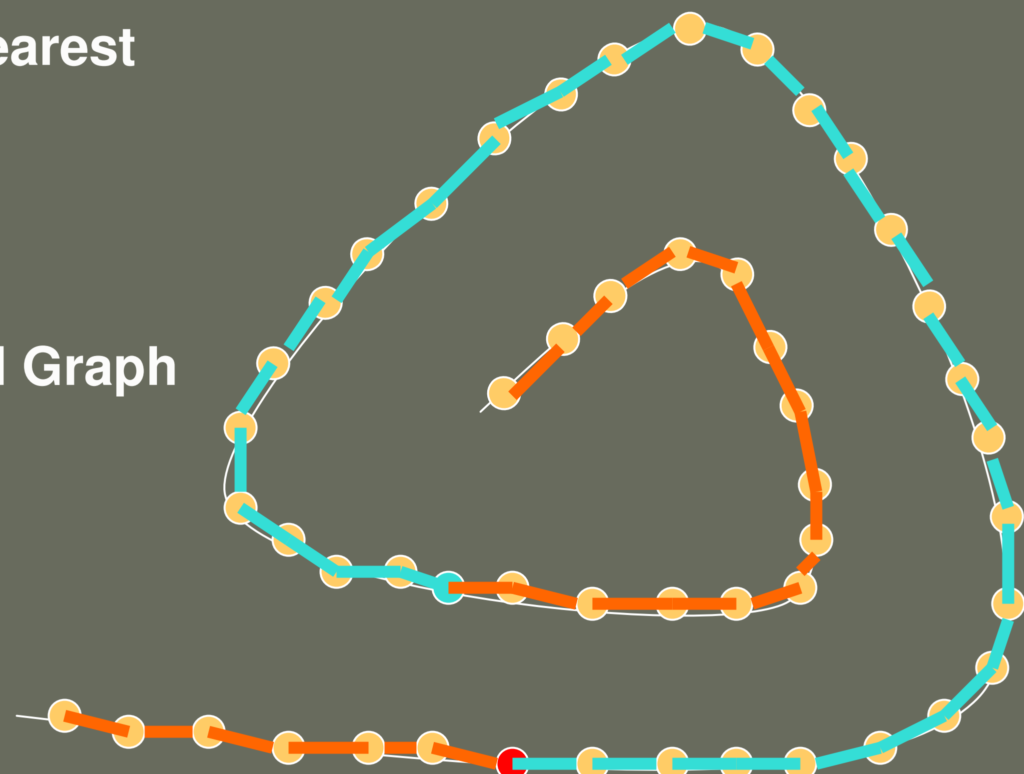
1. **Link To Nearest Neighbors**

2. **Compute Distances THROUGH Graph**



“Classic” Manifold Method: ISOMAP

1. **Link To Nearest Neighbors**
2. **Compute Distances THROUGH Graph**

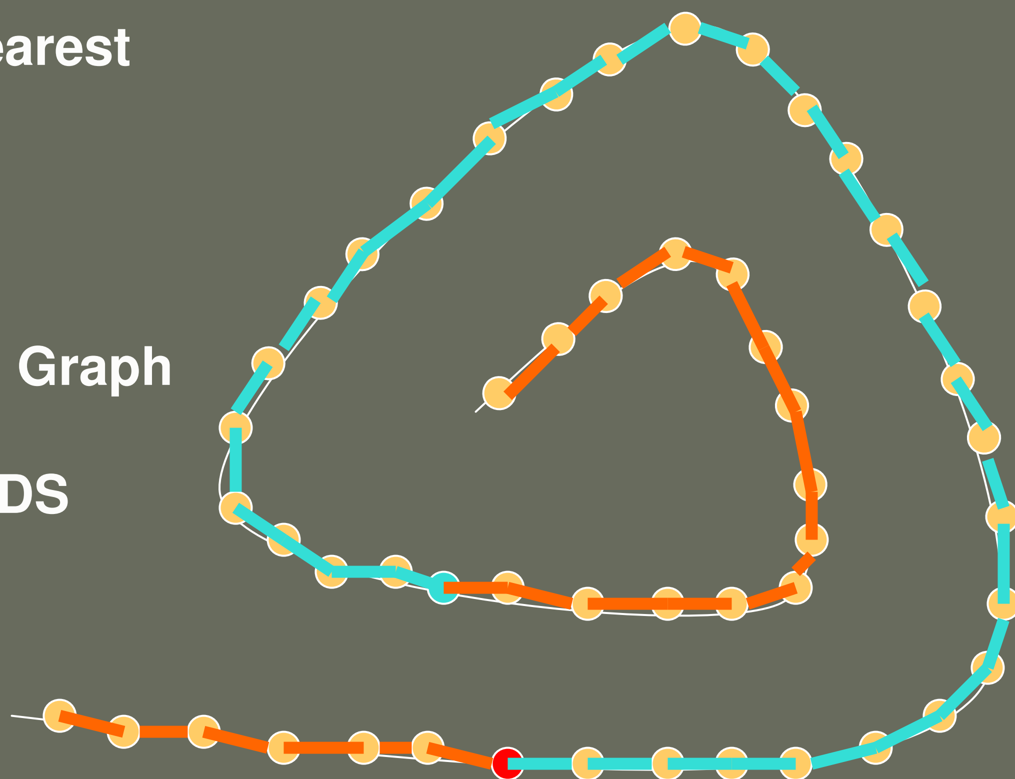


"Classic" Manifold Method: ISOMAP

1. **Link To Nearest Neighbors**

2. **Compute Distances THROUGH Graph**

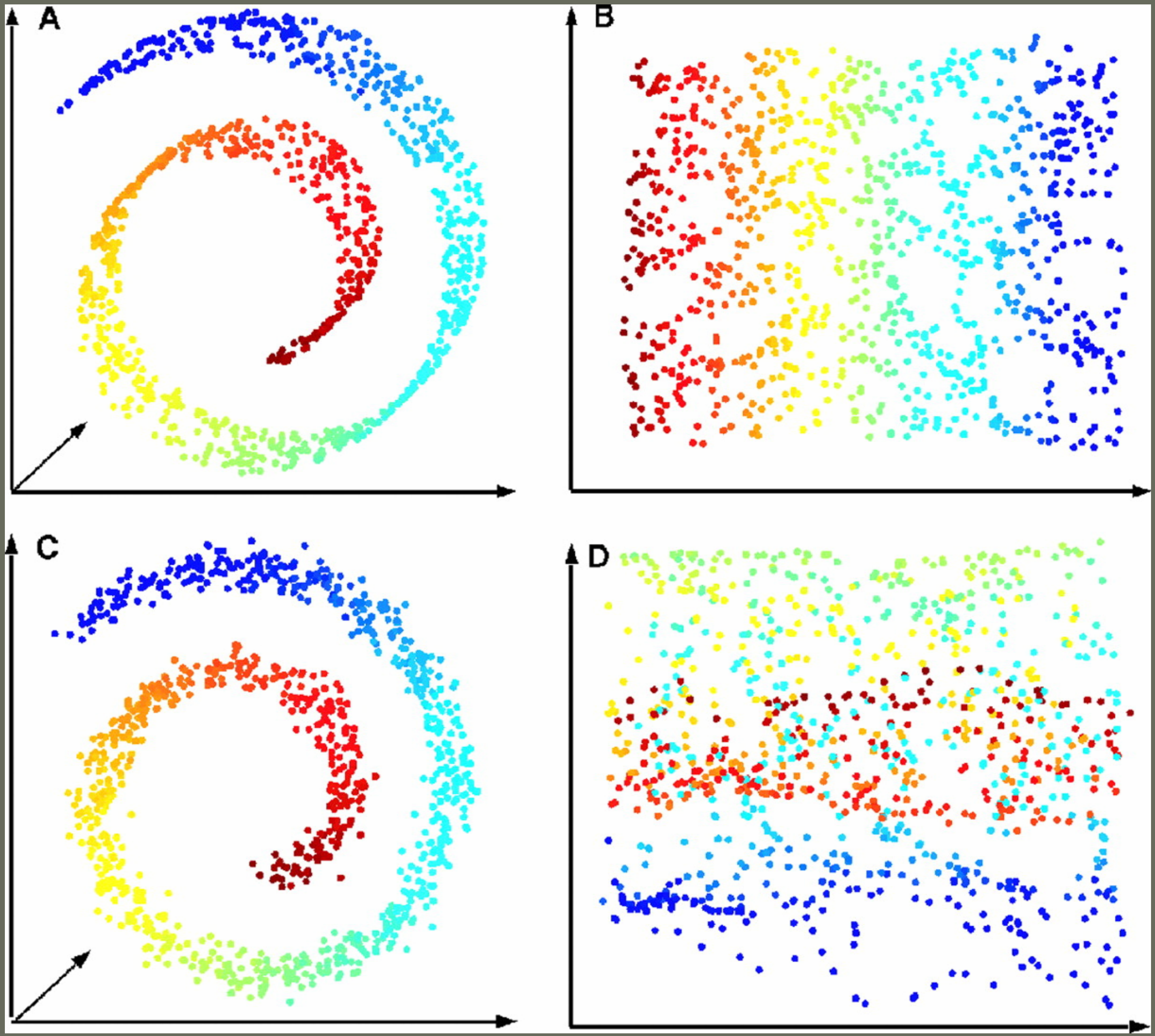
3. **Perform MDS**



Paper I: Charting a Manifold

Matthew Brand

Why Bother?



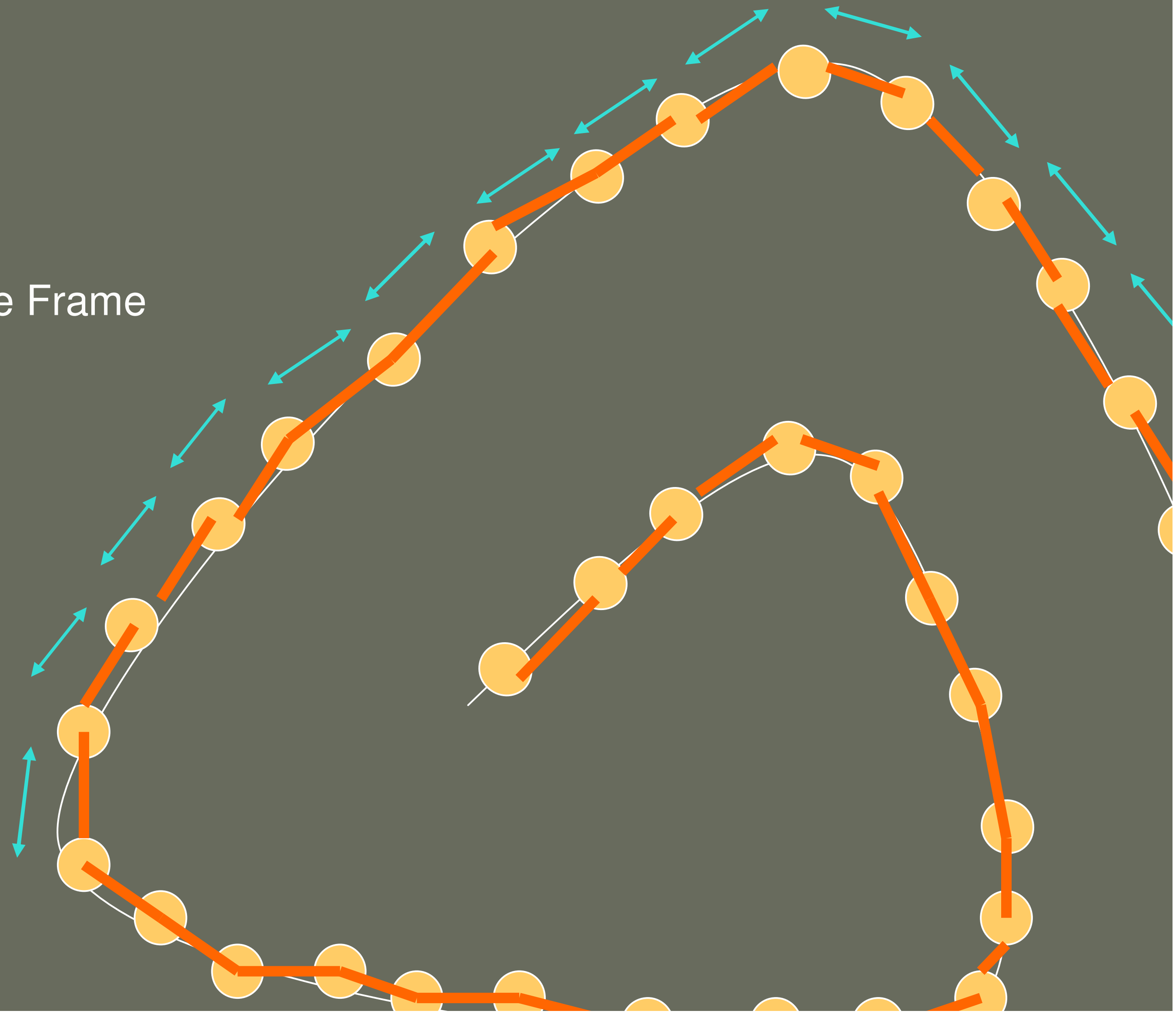
What's Going On?

- Isomap depends on the integrity of the local structure of the manifold
- Noise perturbs the structure leading to an incorrect embedding.

What do we do about it?

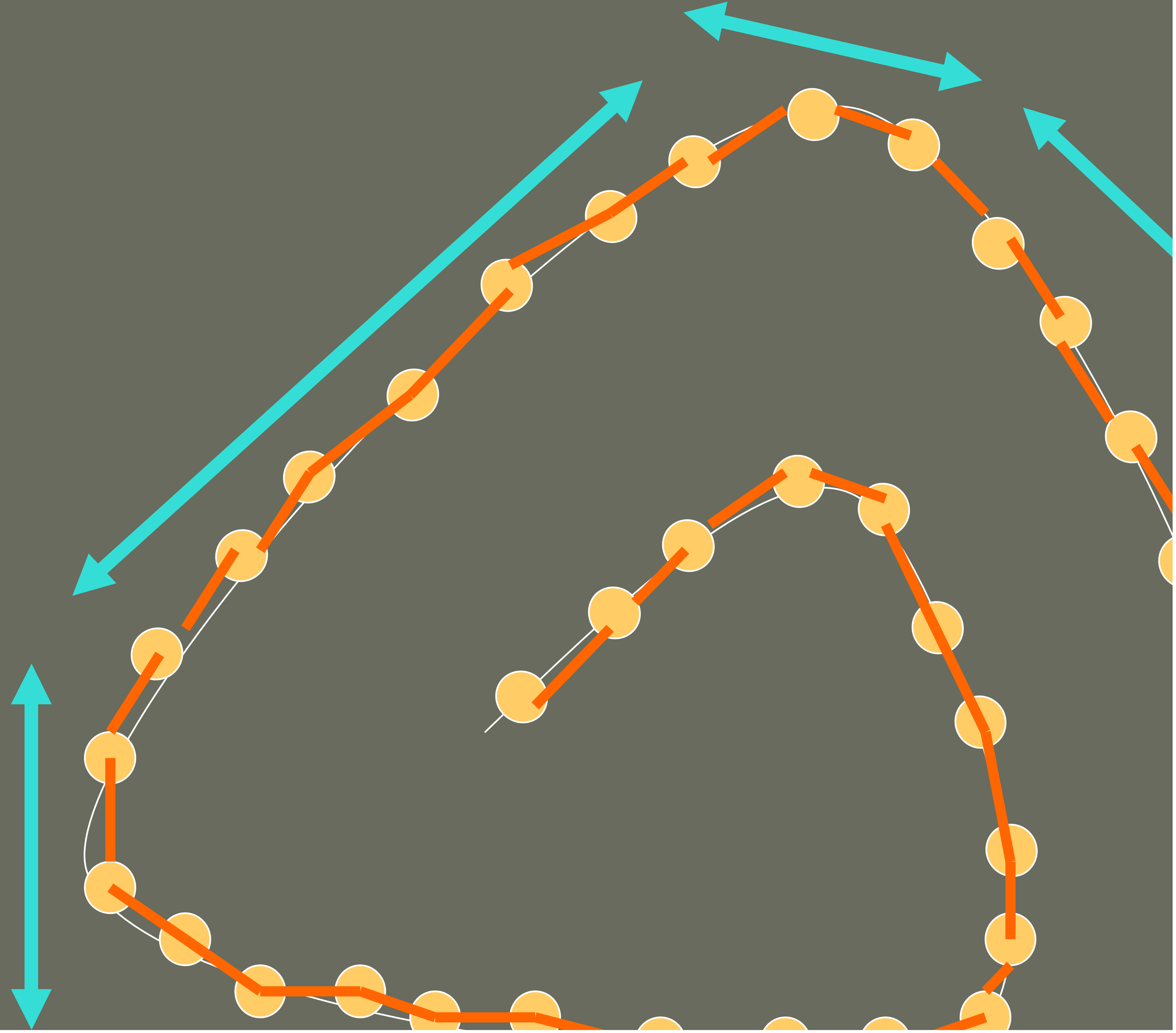
What do we do about it?

R^1 Coordinate Frame
At each link



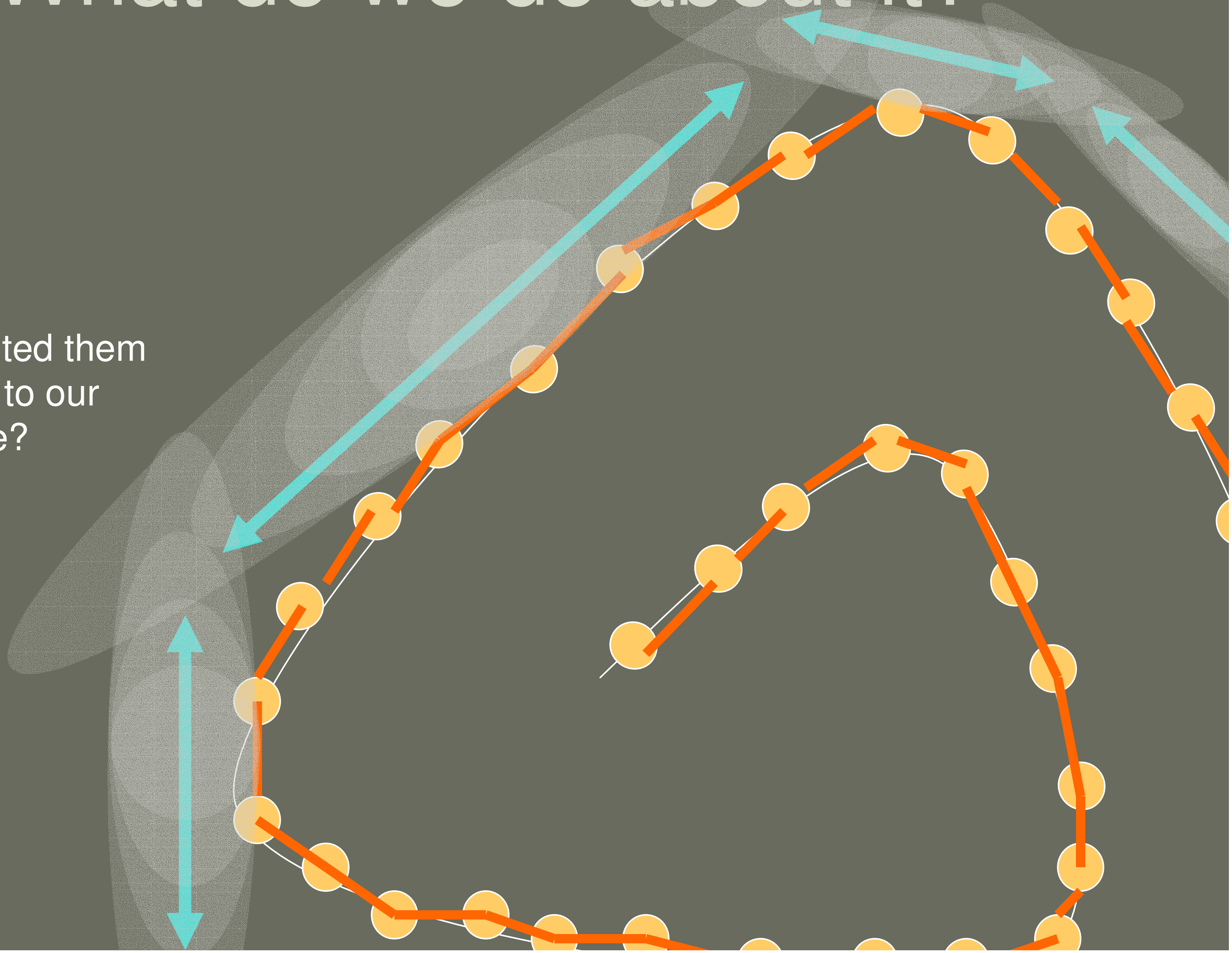
What do we do about it?

What if we *merged*
the similar frames?



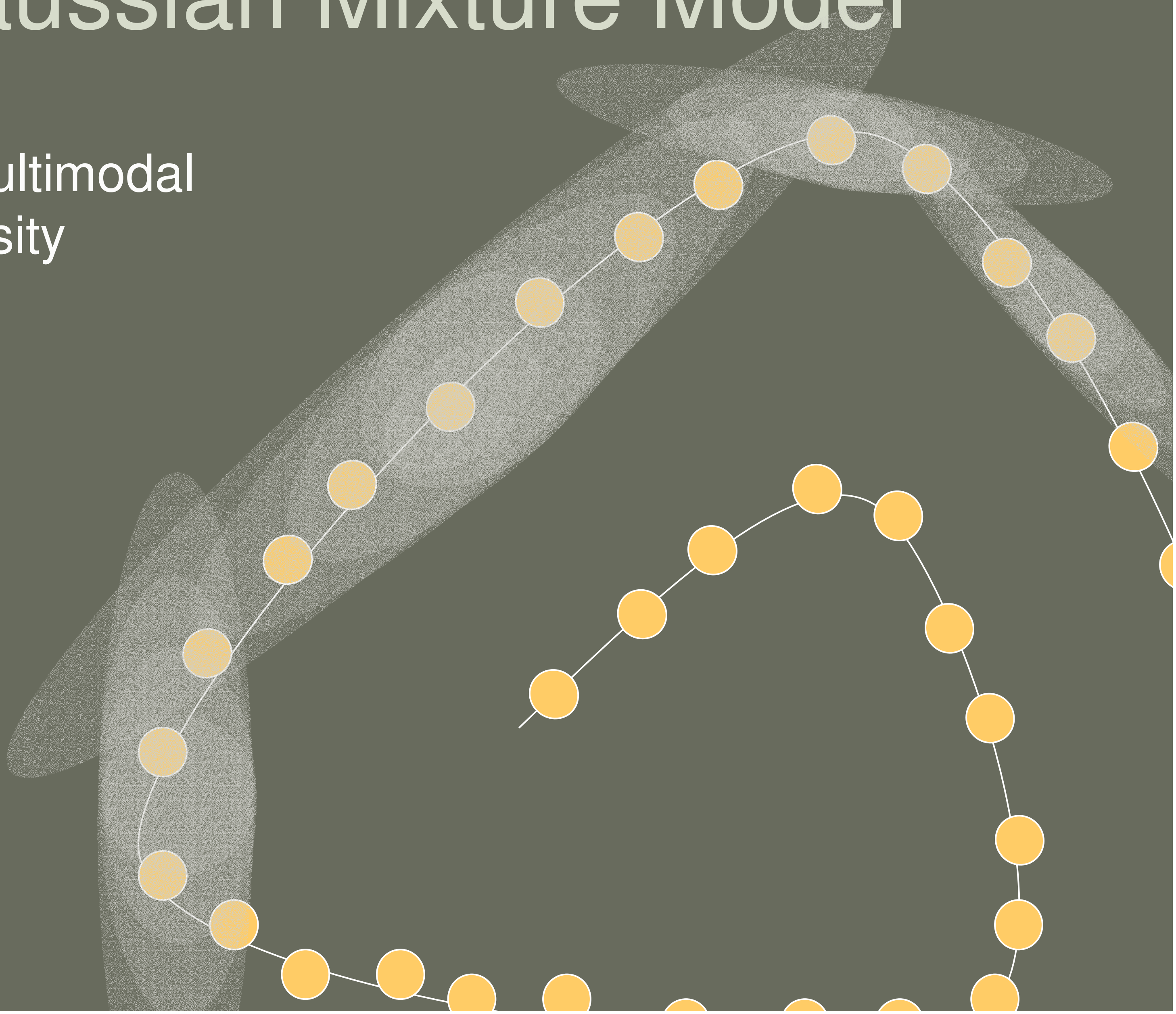
What do we do about it?

And weighted them
According to our
confidence?



Gaussian Mixture Model

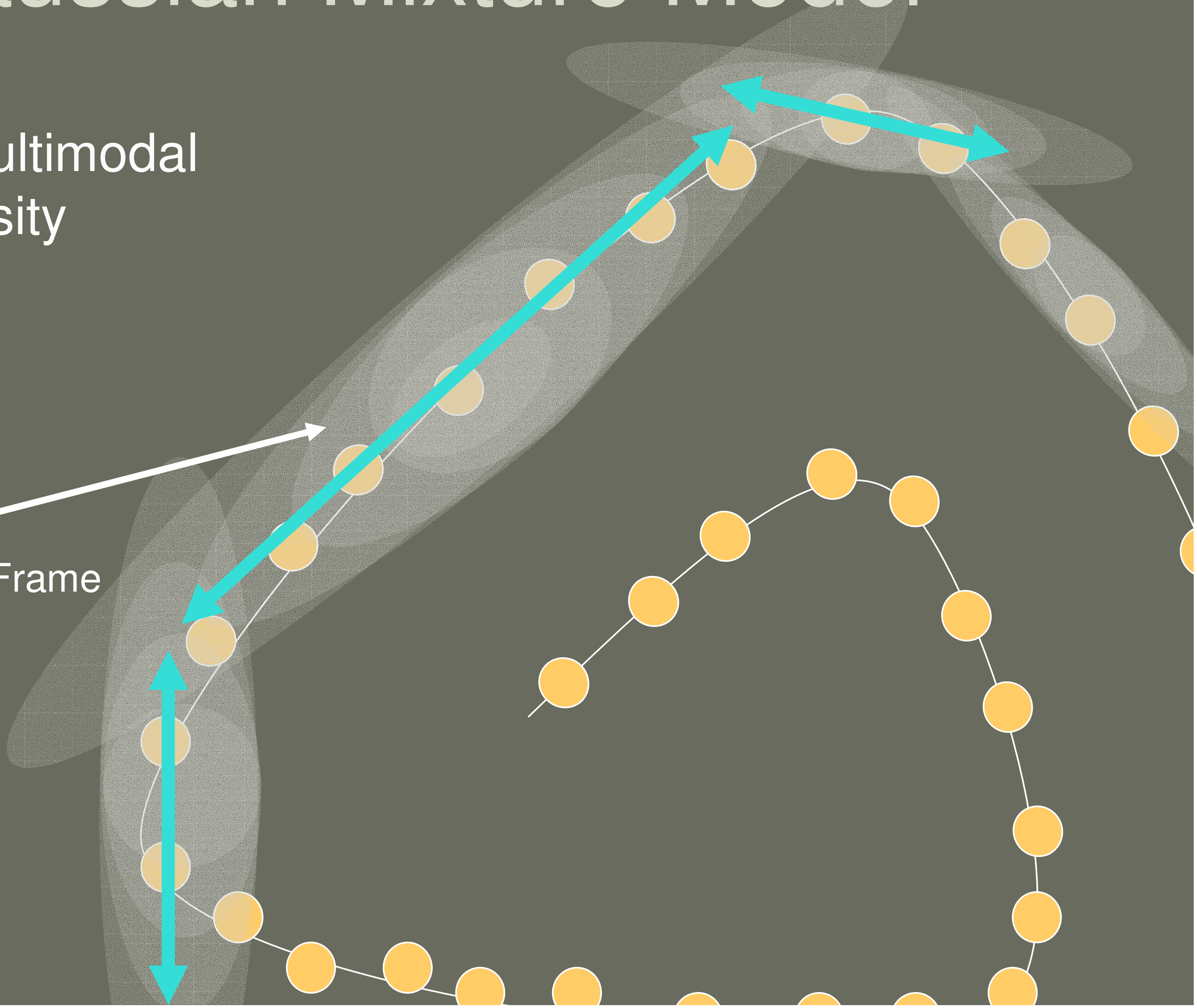
Estimates Multimodal
Sample Density



Gaussian Mixture Model

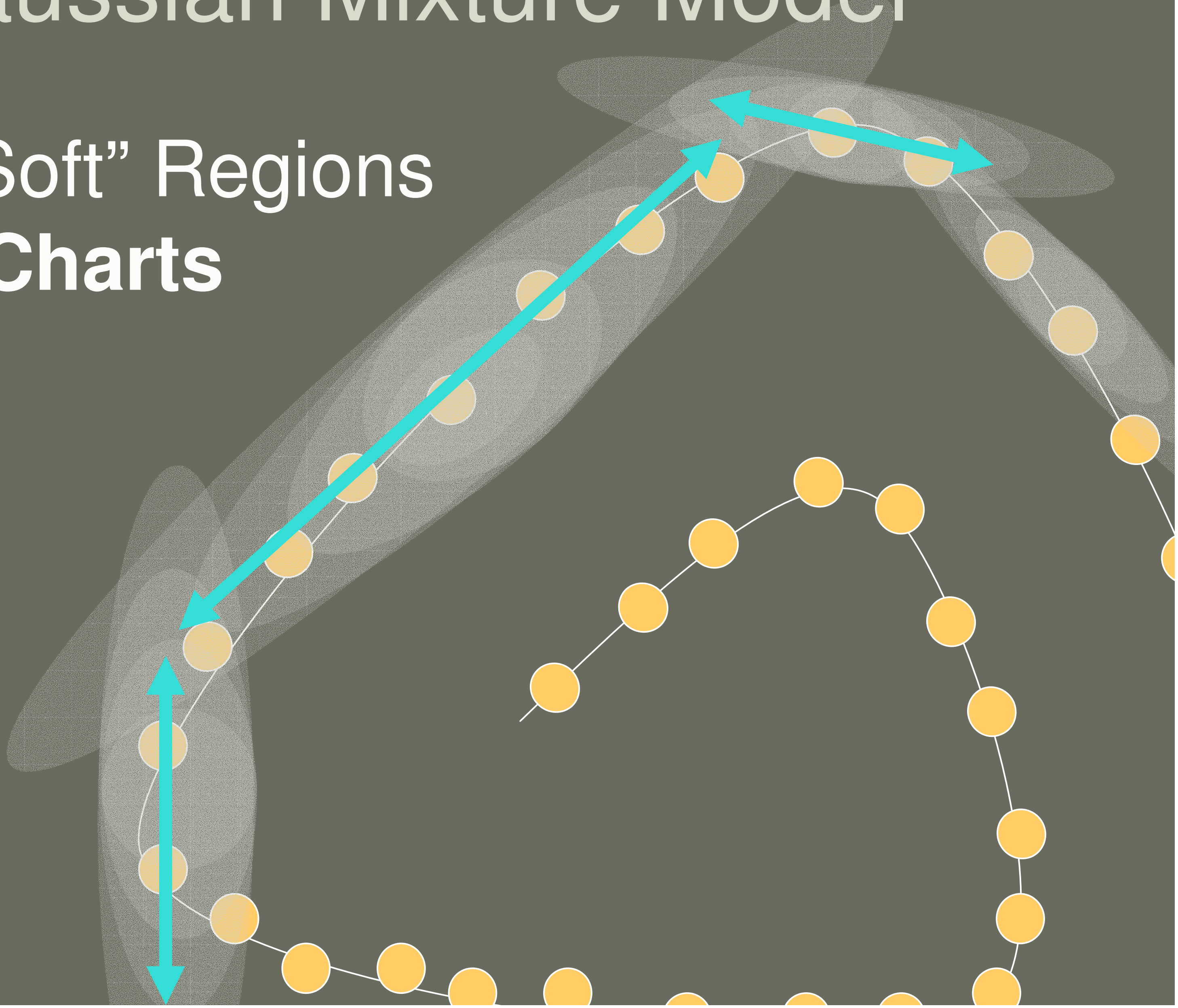
Estimates Multimodal
Sample Density

Derive Coordinate Frame
From Eigenvectors
Of Distribution

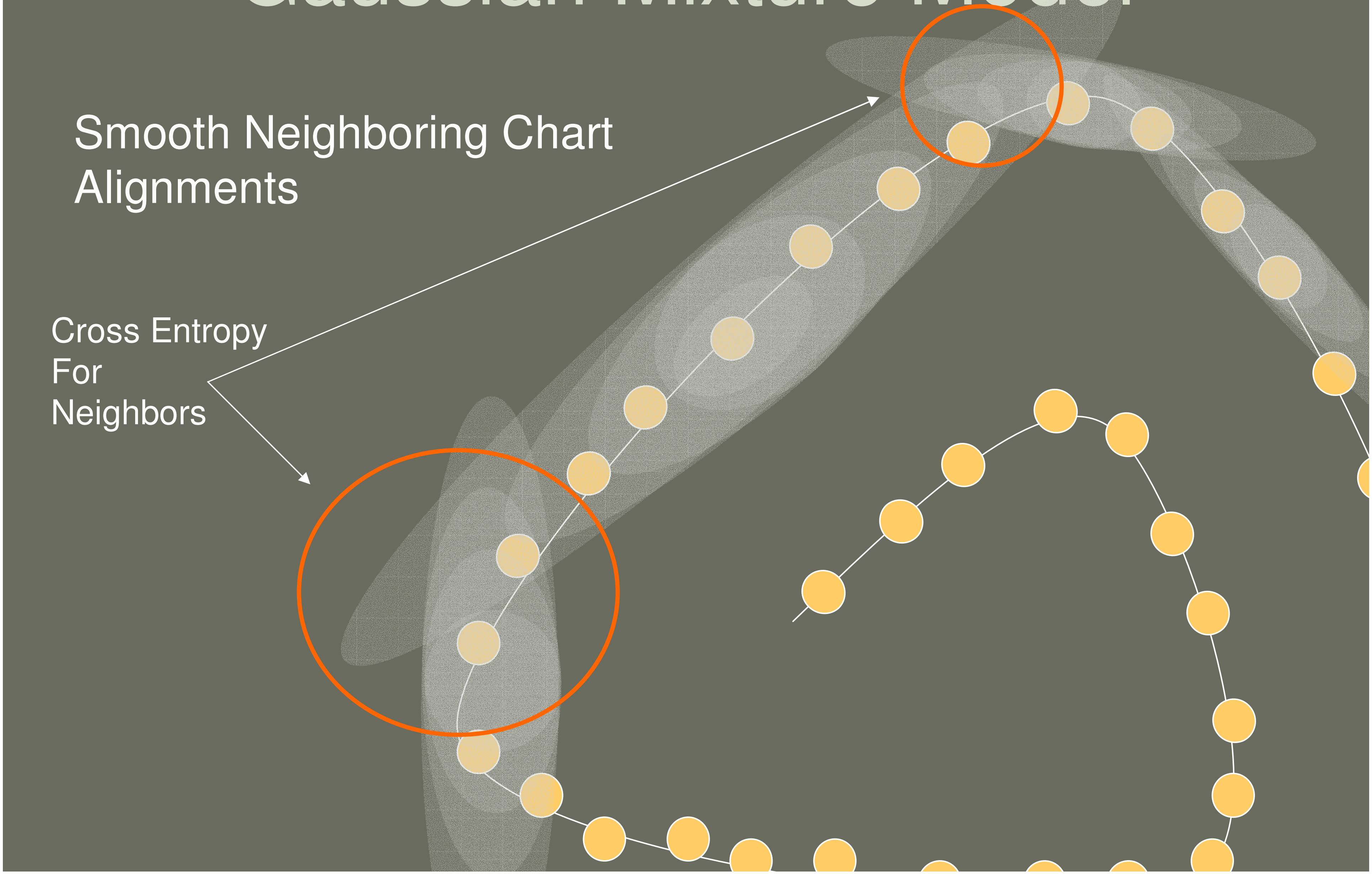


Gaussian Mixture Model

These "Soft" Regions
Are our **Charts**



Gaussian Mixture Model



Semi-Invertable Transform

- A transformation to and from the manifold

- Interpolate on the manifold and "backproject" to original sample space

Critique

- GOOD
  - Elegant, robust idea solves shortcomings of former methods
  - Lots of novel examples to prove utility
  - Backprojection provides visualization opportunities

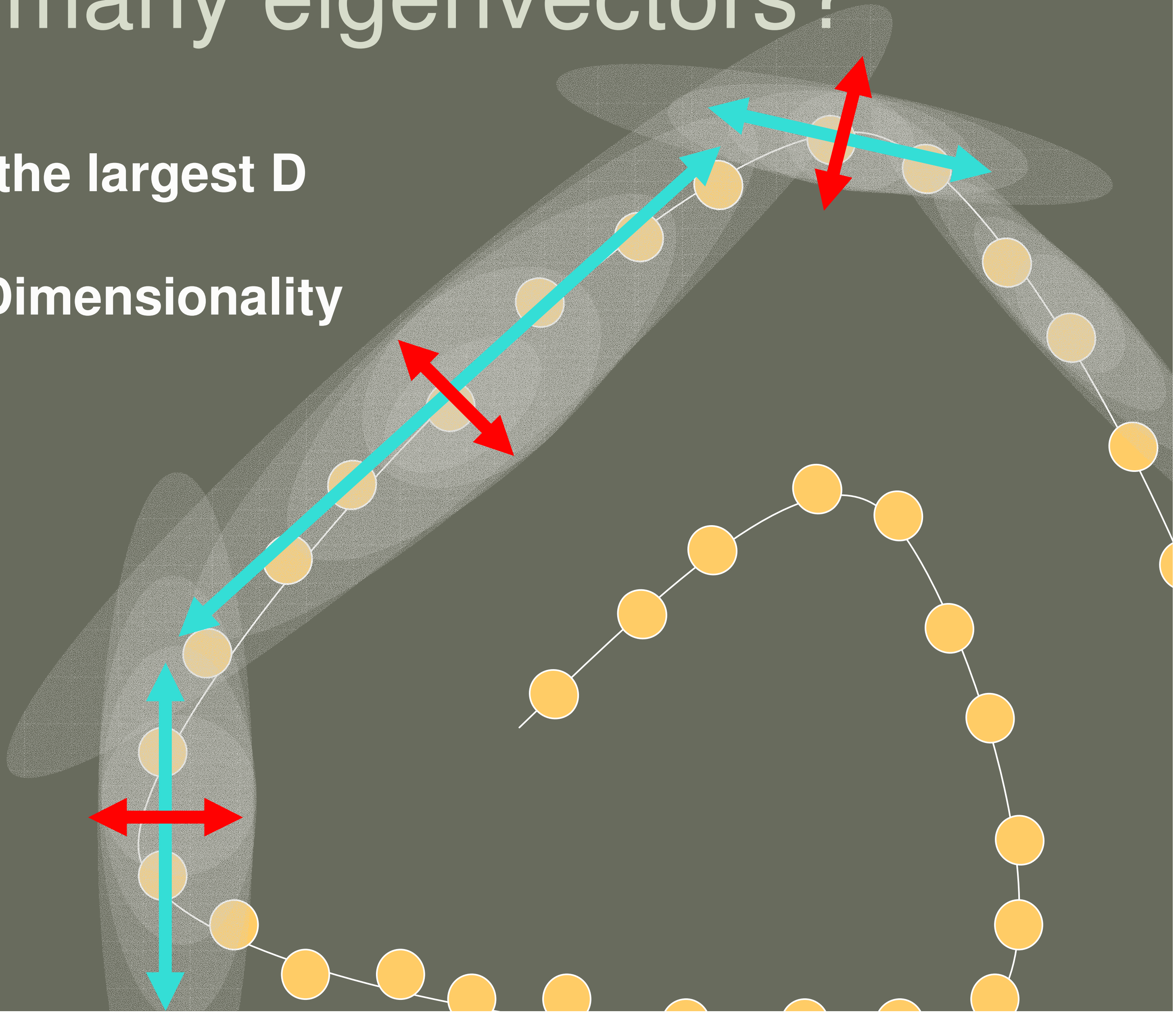
- BAD
  - Little appeal to intuition
  - No Code
  - Runtimes? How does it scale?

Paper II: Maximum Likelihood Estimation of Intrinsic Dimension

Elizaveta Levina and Peter J. Bickel

How many eigenvectors?

We use only the largest D where
D = Intrinsic Dimensionality

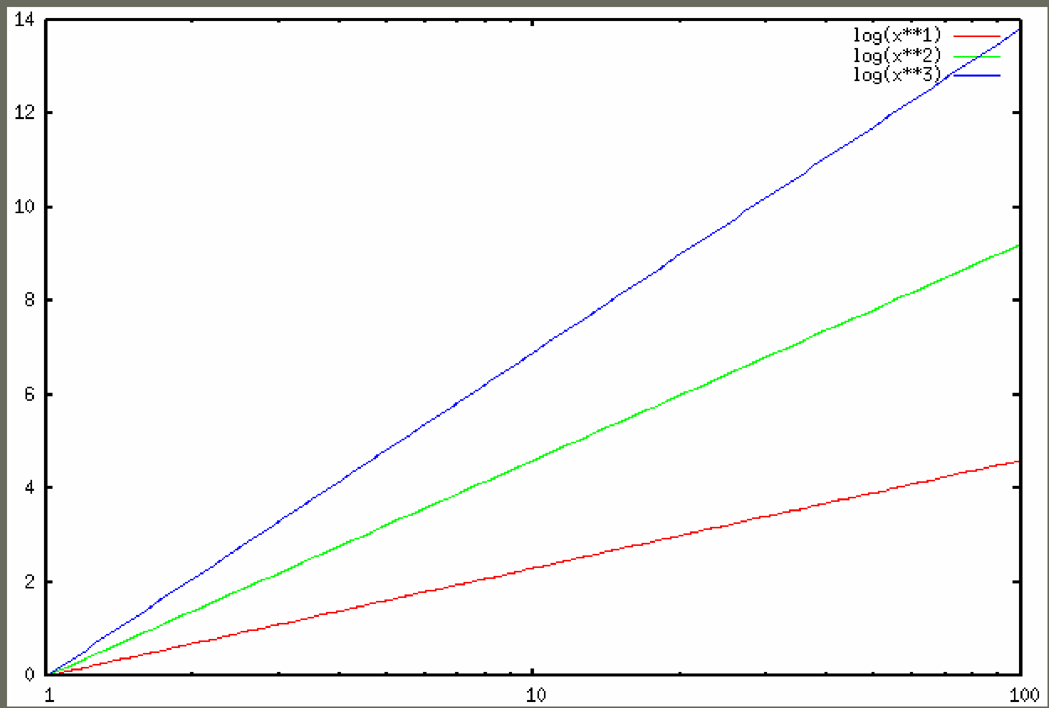


How do we get D?

- Most often = User makes a guess
- Use an estimation method
  - Projection Methods (PCA, local PCA)
  - Geometric Methods

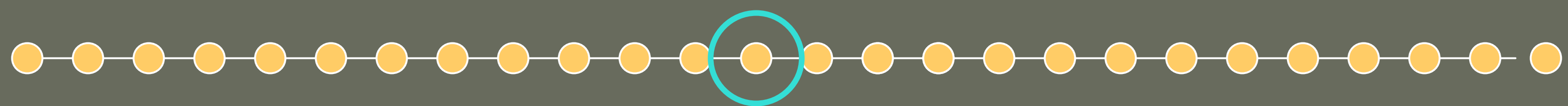
Geometric Methods

- C(r) = average number of points in radius **r** for each point in dataset
- Plot log(C(r)) against log(r)
- D = slope

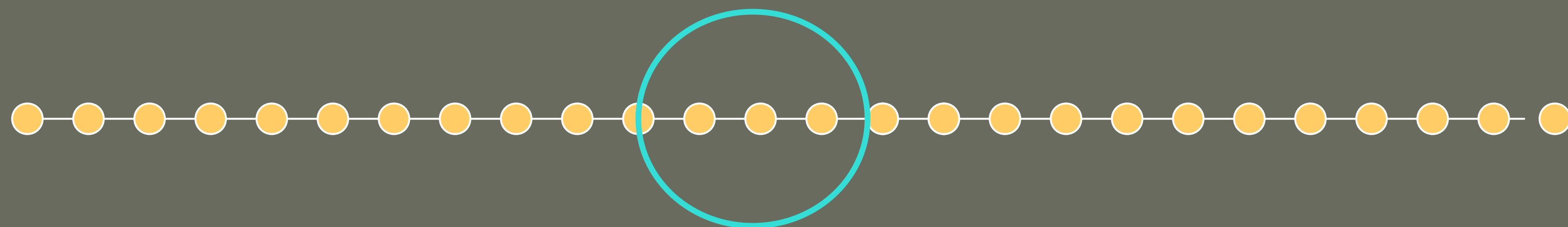


Why?

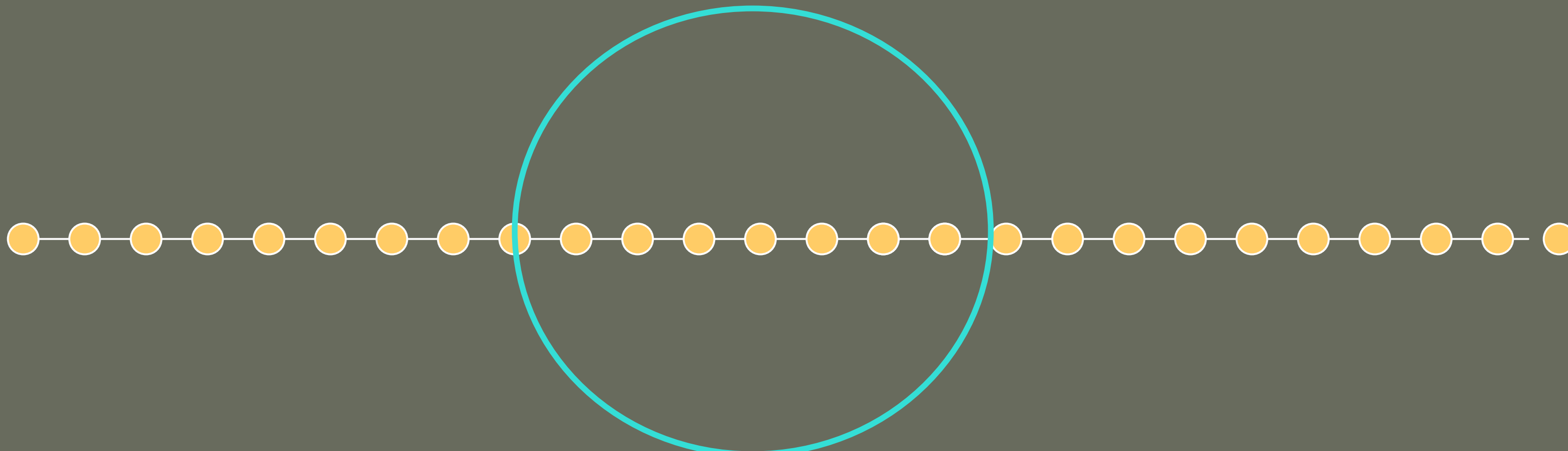
Why?



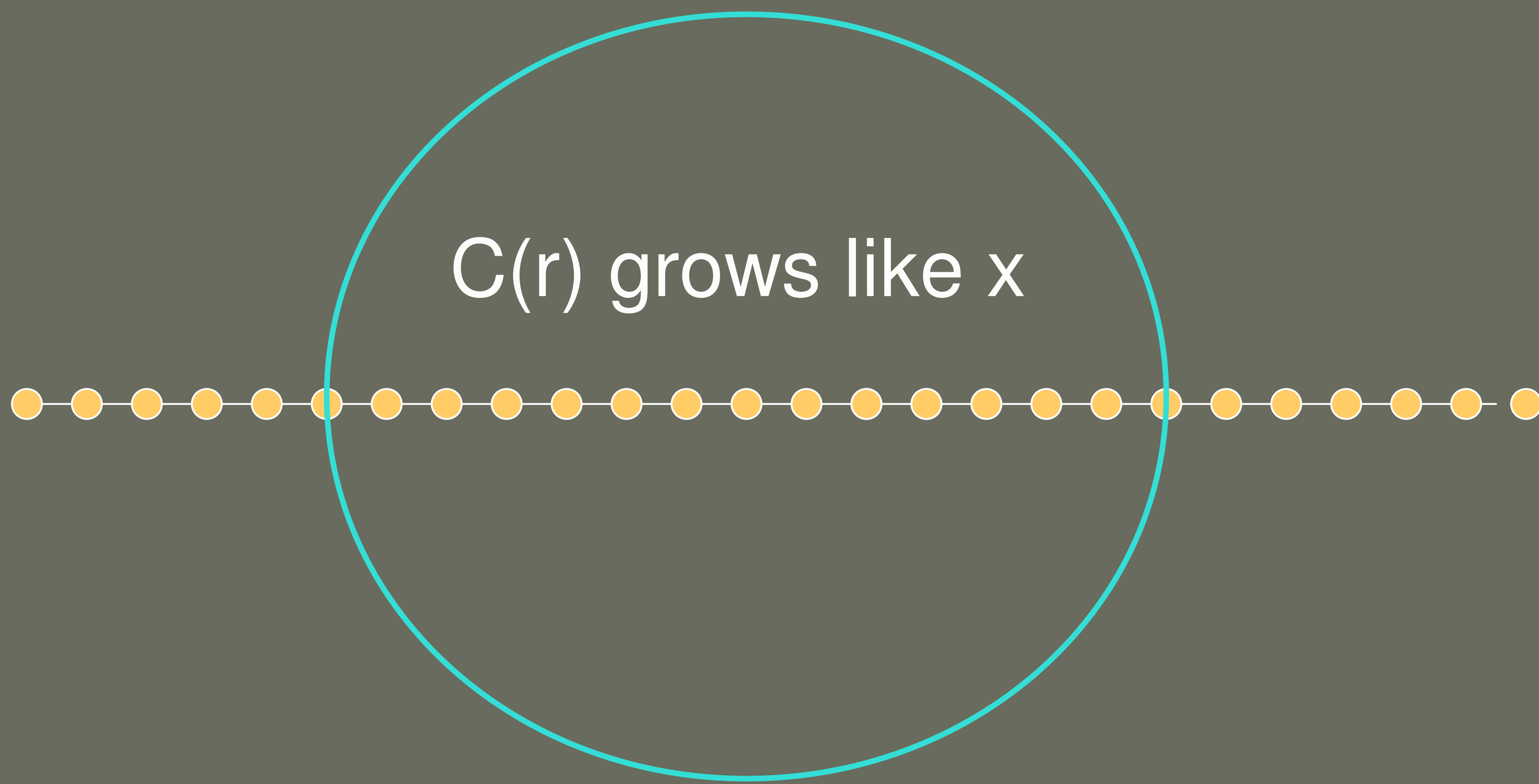
Why?



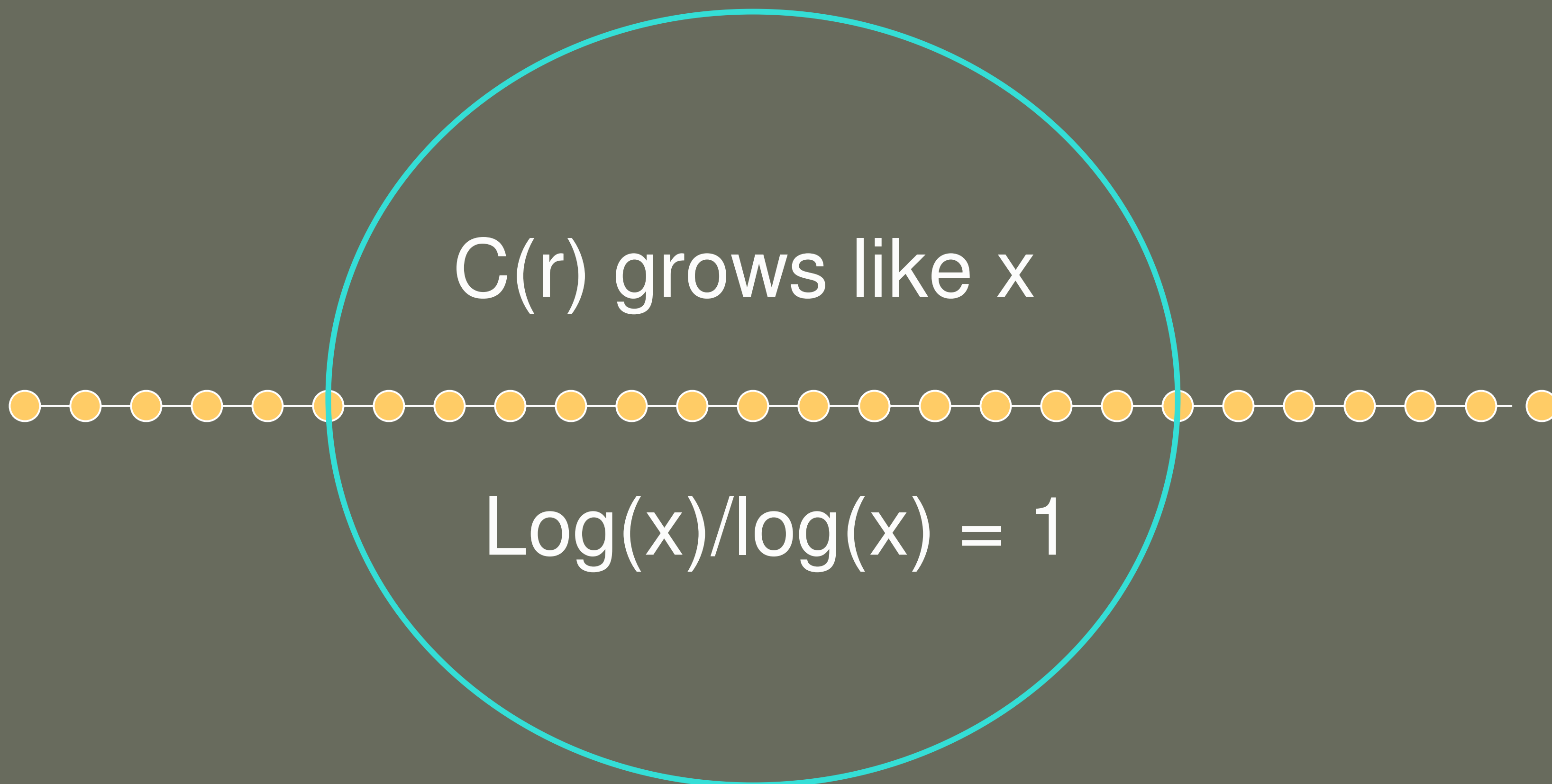
Why?

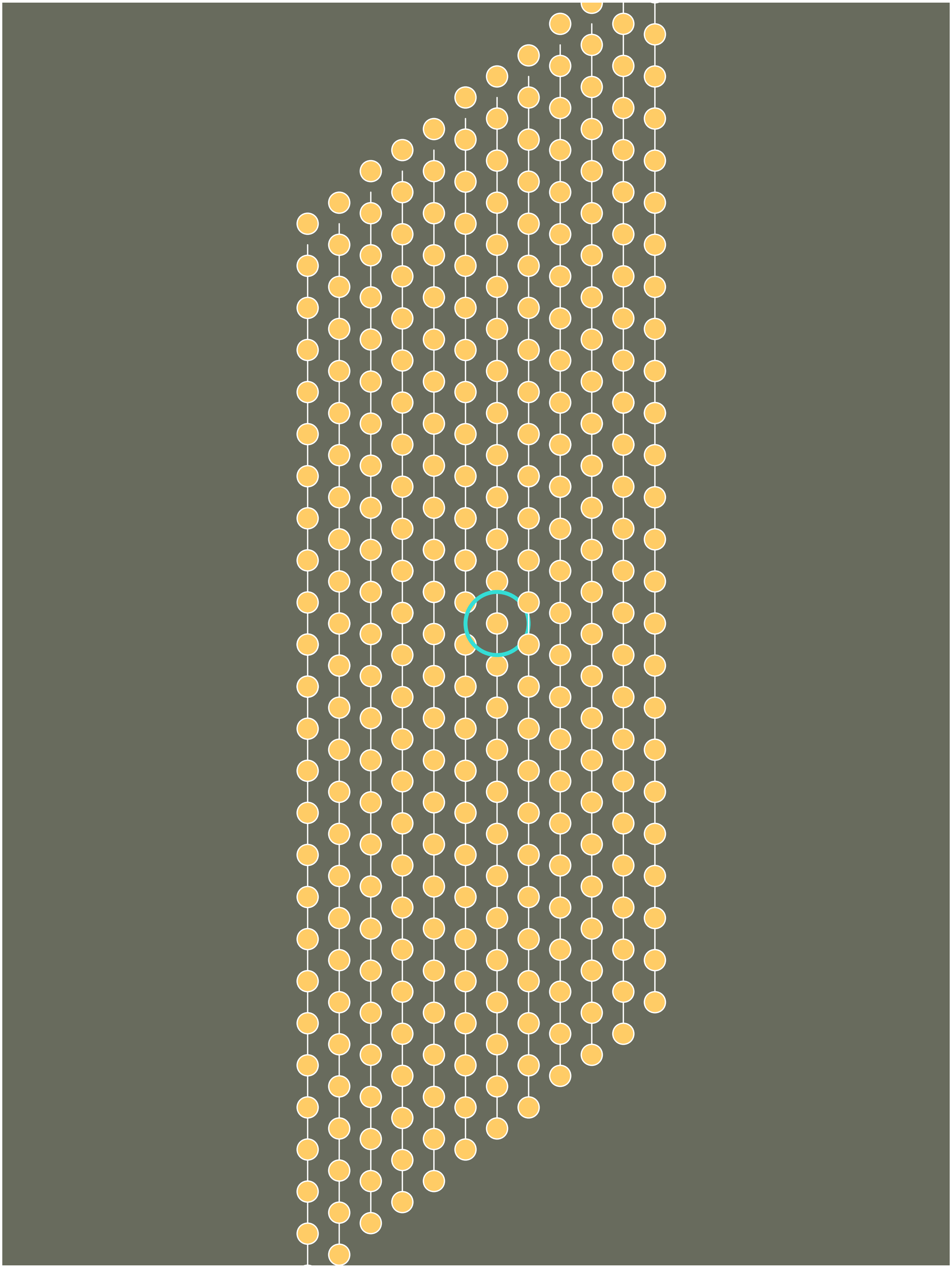


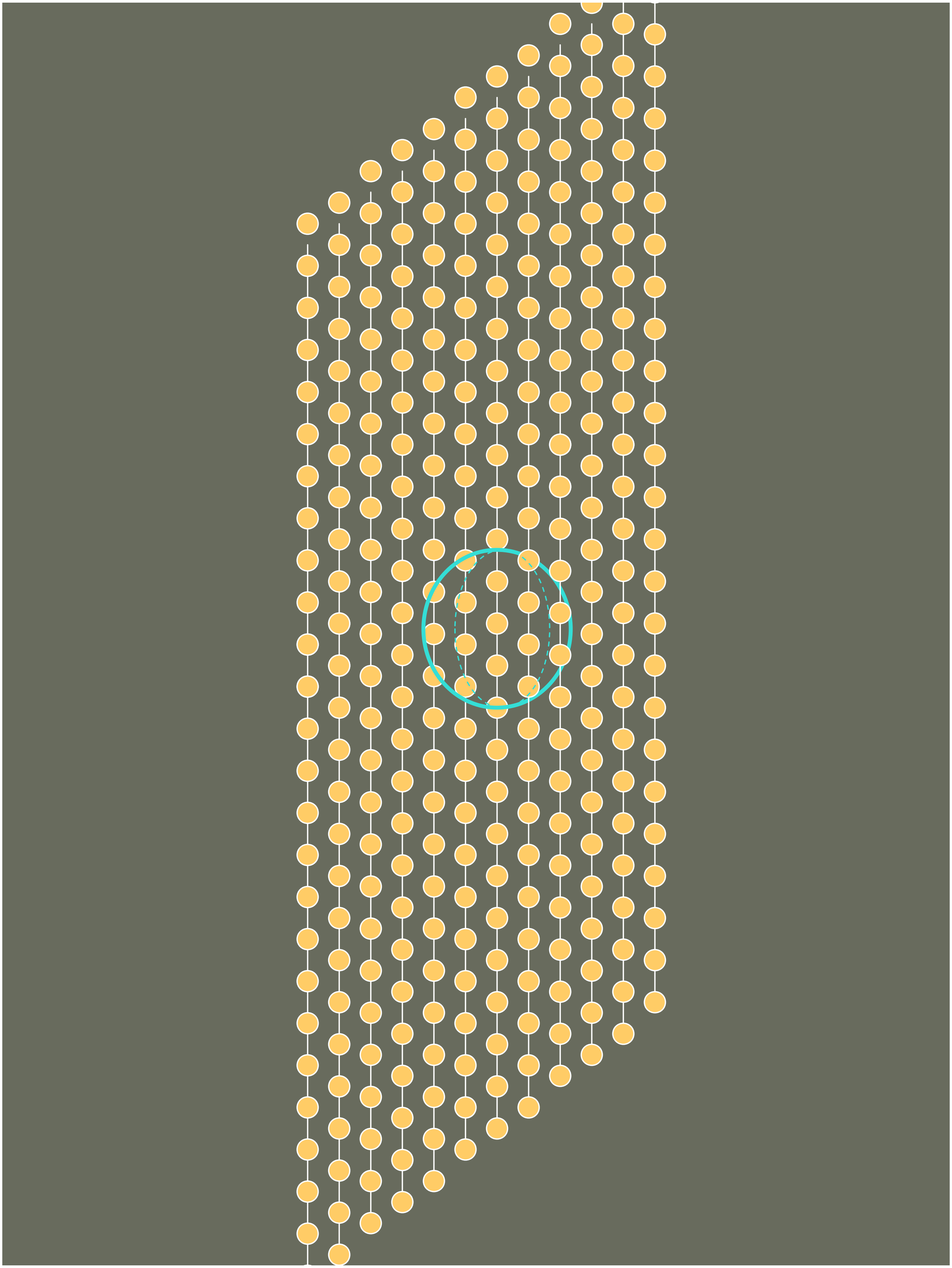
Why?

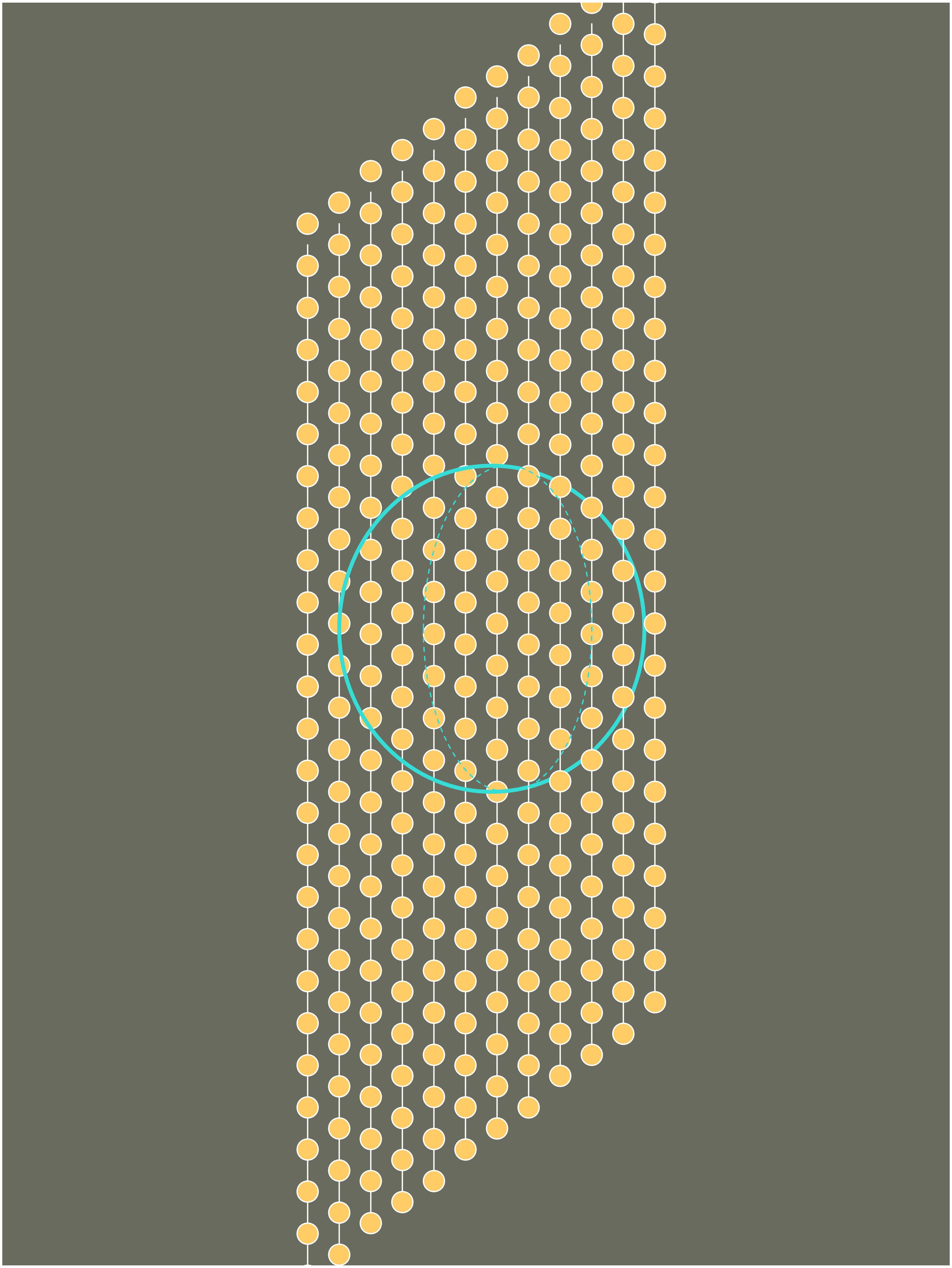


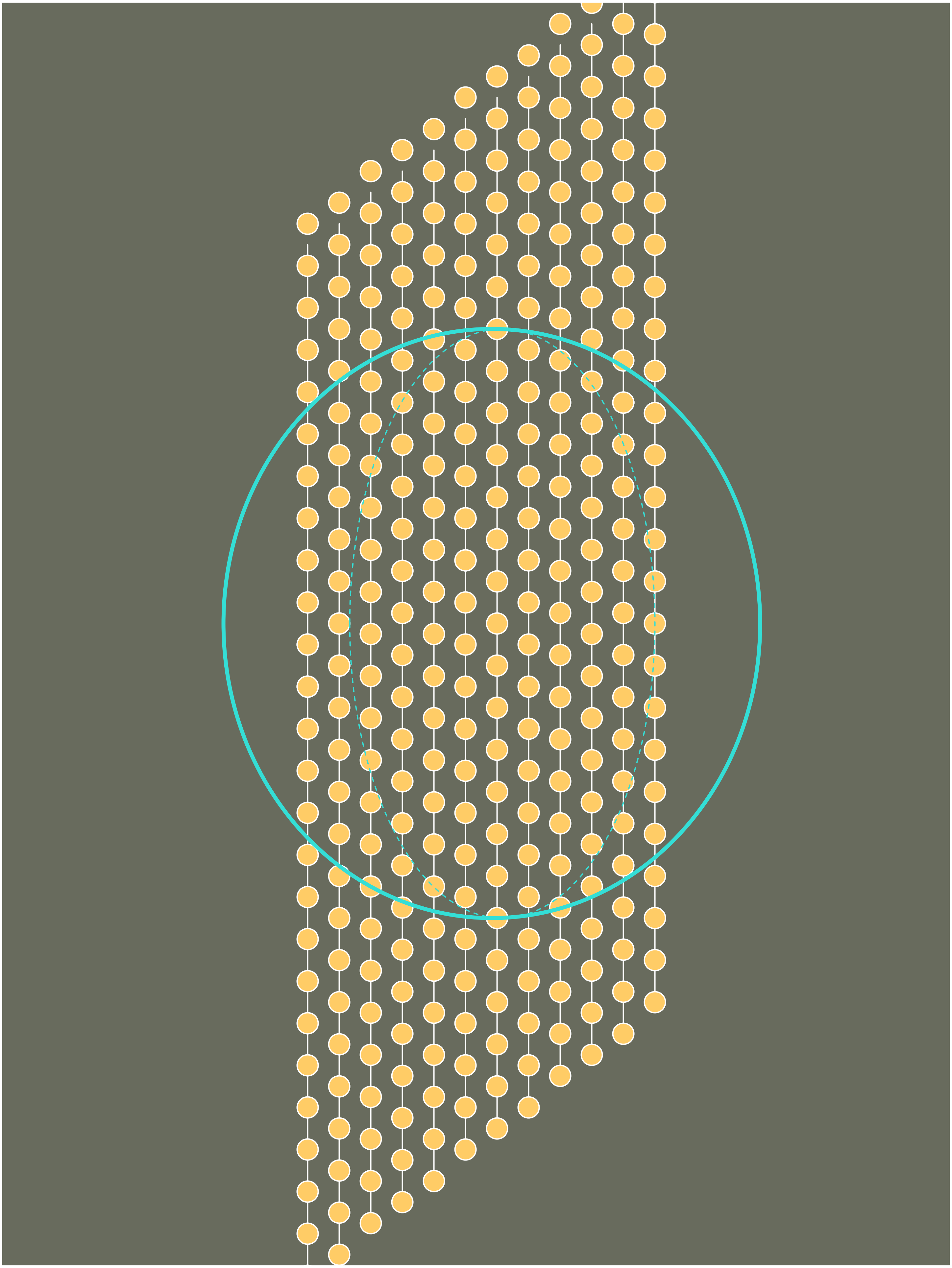
Why?



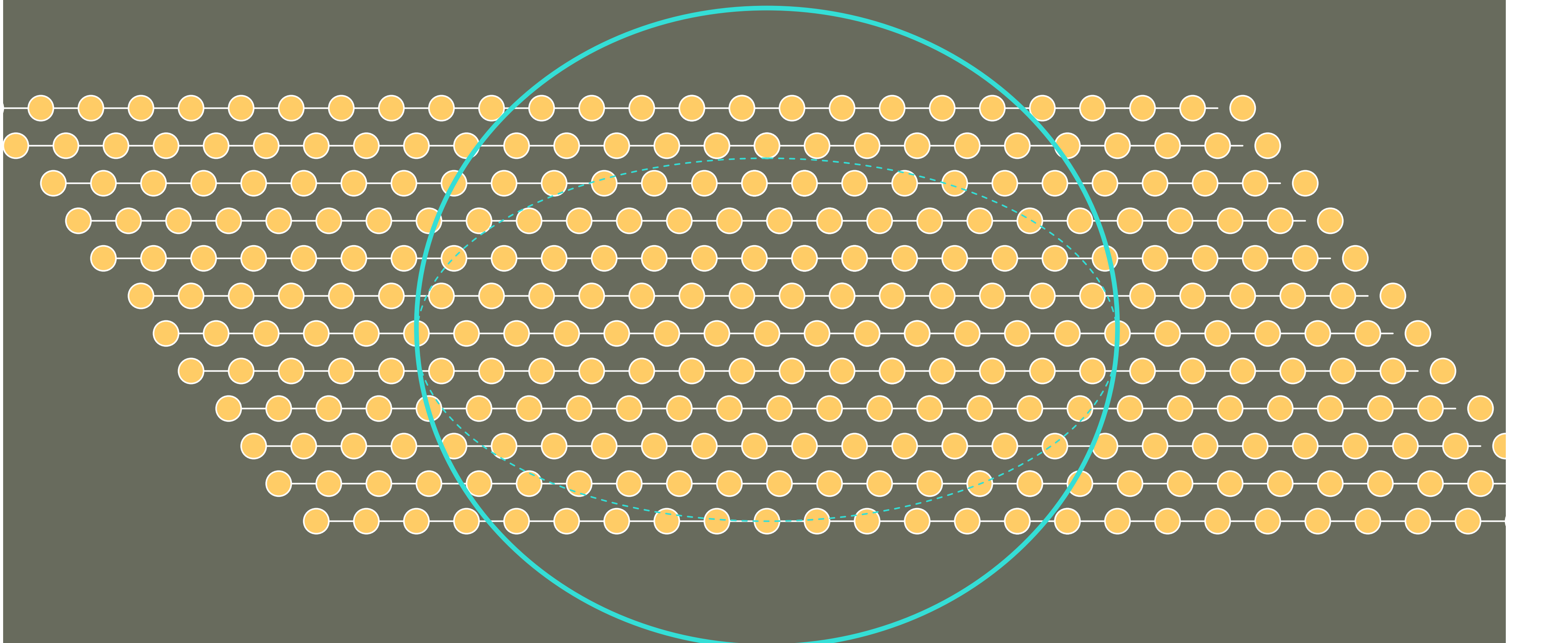








C(r) grows like x^2



$$\mathrm{Log}(x^2)/\log(x) = 2$$

This is called the correlation dimension

How well does this work?

Issues

- We don’t know the effect of
  - Sample Size
  - Dimension
- We also don’t understand bias or variance

Strategy of Paper II

- Define a stochastic process to model observations in sphere for some low dimensional density.
- Define a MLE for the dimension parameter of the process.
- Examine statistical properties of the estimator.

Step 1: Define the Process

- $N(t,x)$ = number of points in a sphere of radius t around point x
- We approximate this with a Poisson process
- The rate of this process depends on D!

Step 2: Define the MLE

- MLEs infer values of parameters of underlying process.
- Build an MLE for D

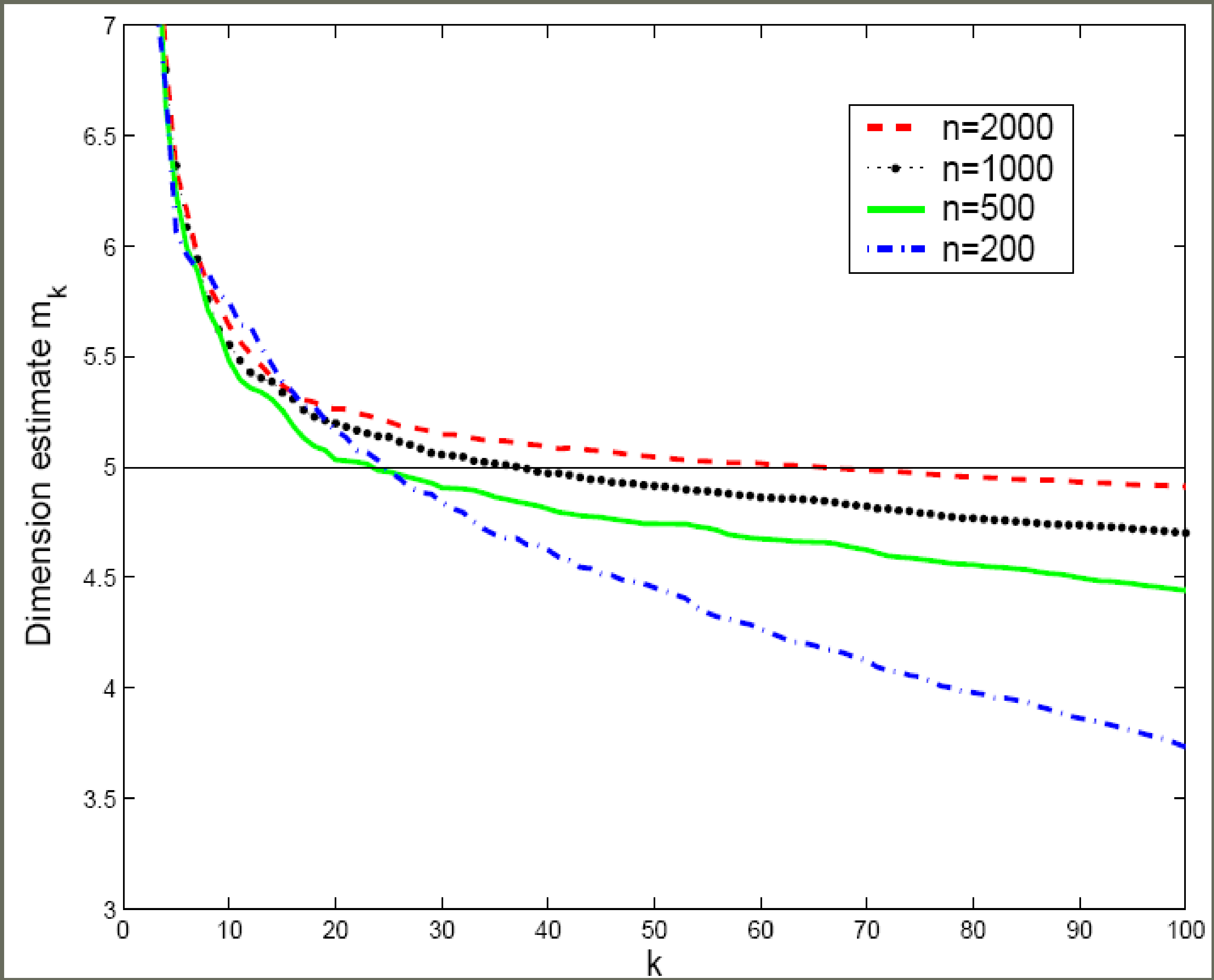
$$\left[\frac{1}{k-1}\sum_{j=1}^{k-1}\log\frac{T_k(x)}{T_j(x)}\right]^{-1}$$

- Average over all points
- Average over a range of k

Step 3: Discuss Properties of MLE

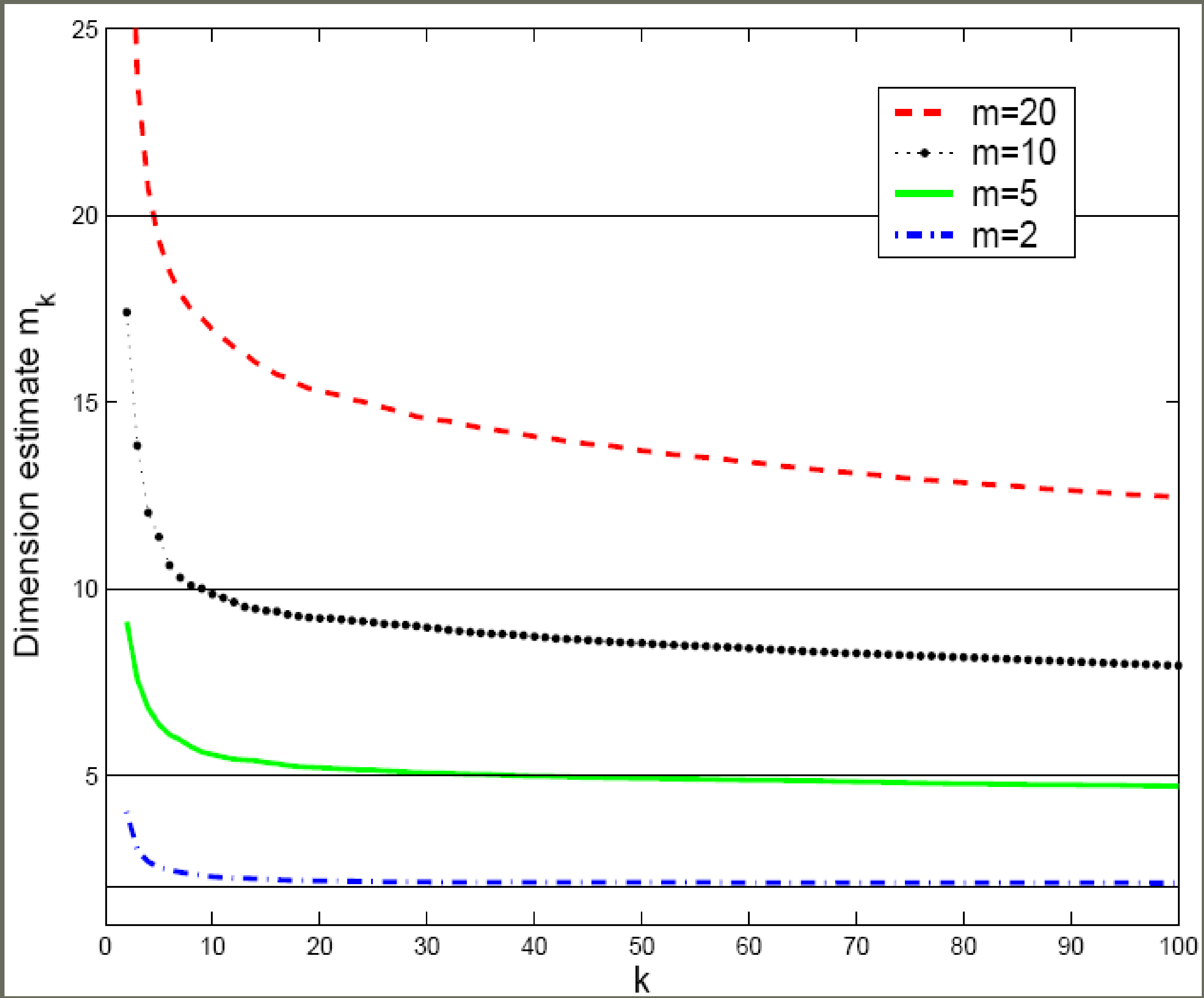
- Expected value of MLE = D
- Variance = $D^2/(k-3)$
- These are asymptotic for k and sample size

Results



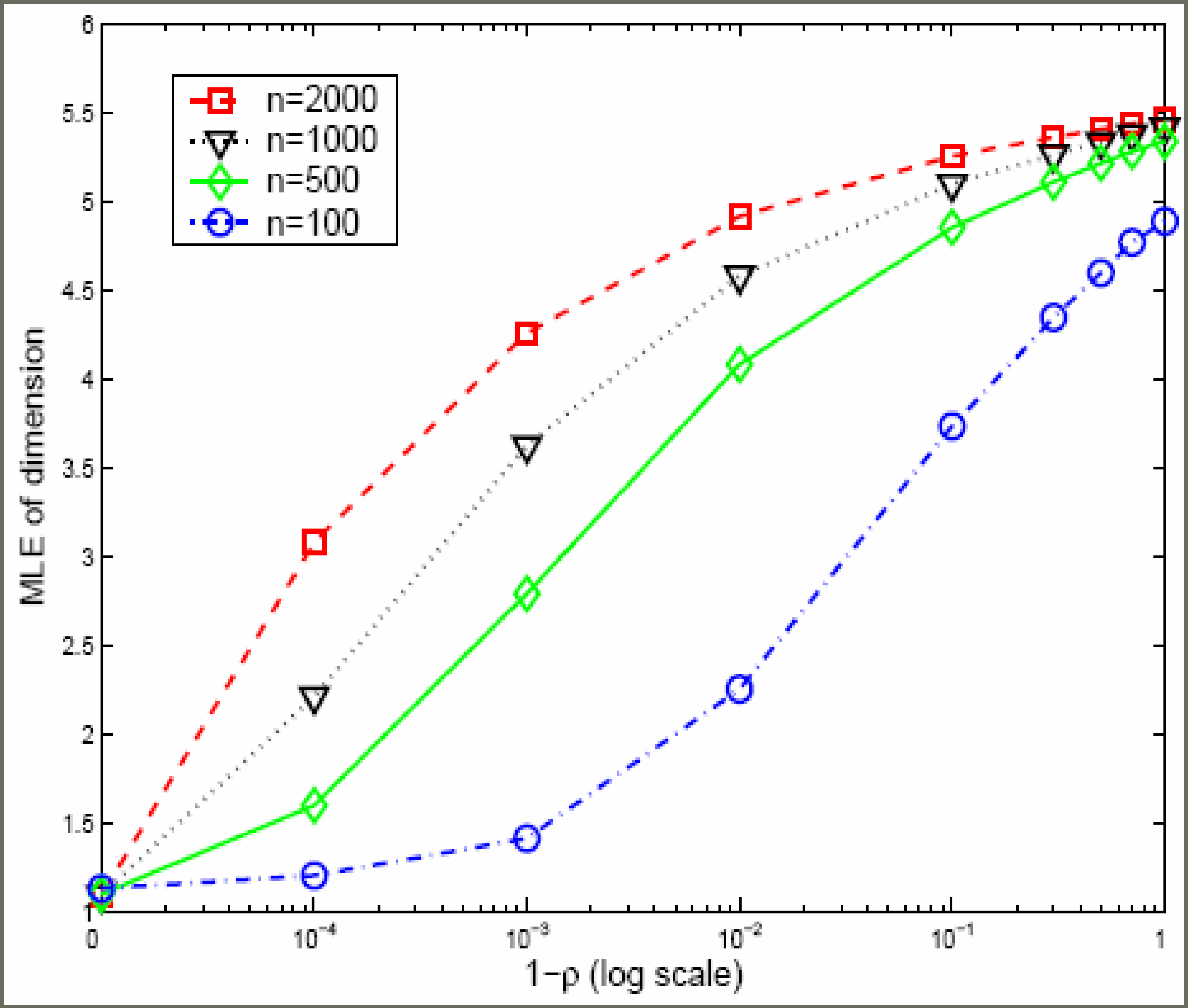
n=2000
n=1000
n=500
n=200
Dimension estimate m_k
k
7
6.5
6
5.5
5
4.5
4
3.5
3
0
10
20
30
40
50
60
70
80
90
100

Results

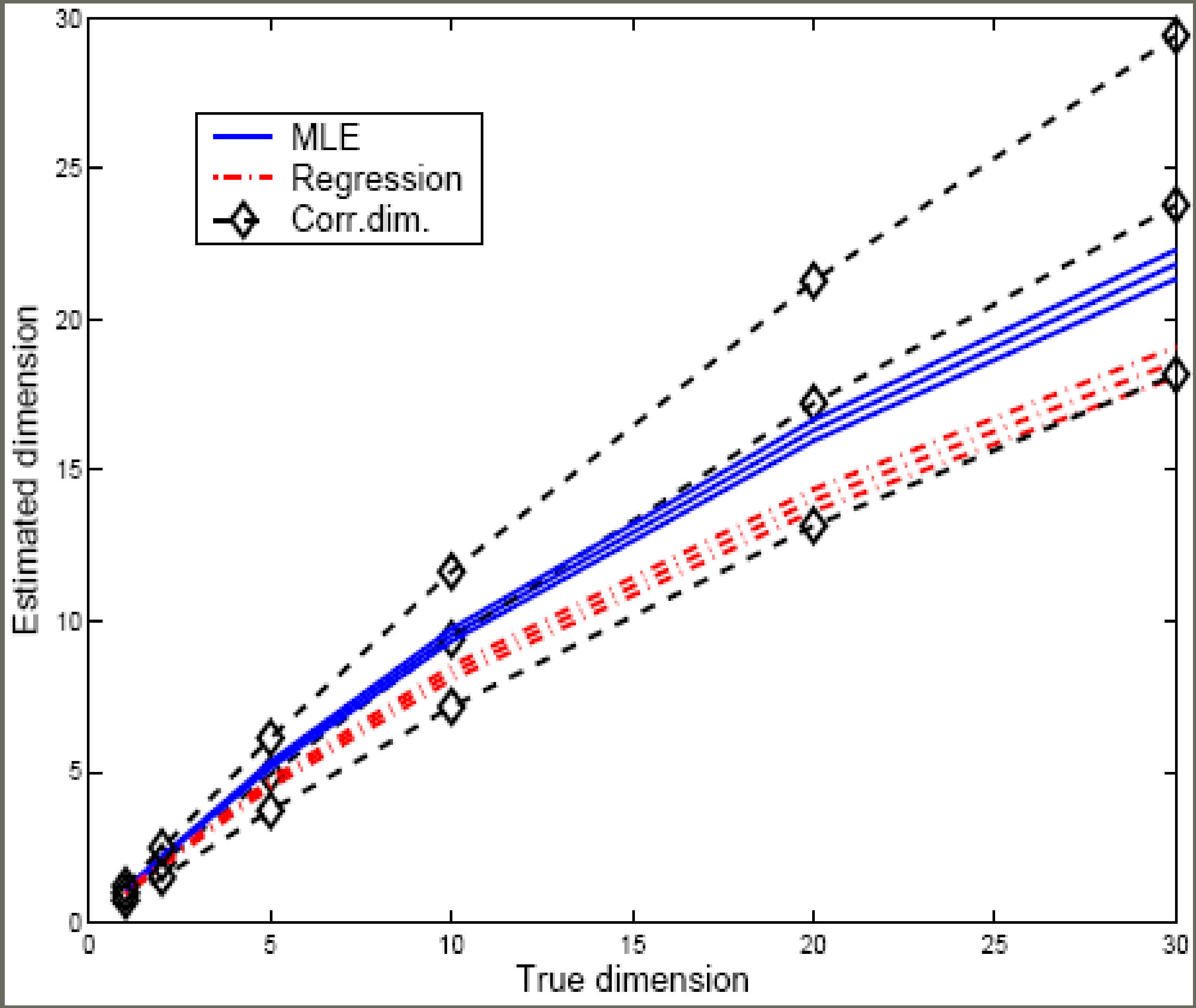


m=20
m=10
m=5
m=2
Dimension estimate m_k
k
25
20
15
10
5
0
10
20
30
40
50
60
70
80
90
100

Results



Results


MLE
Regression
Corr.dim.
Estimated dimension
True dimension
0
5
10
15
20
25
30
0
5
10
15
20
25
30

Critique

- GOOD
  - Provides a well-defined tool for estimating dimensionality
  - Suitable for dimensions appropriate for visualizing
- BAD
  - Written by Statisticians
  - Absolutely no appeal to intuition
  - No geometric description of Estimator!

Questions?