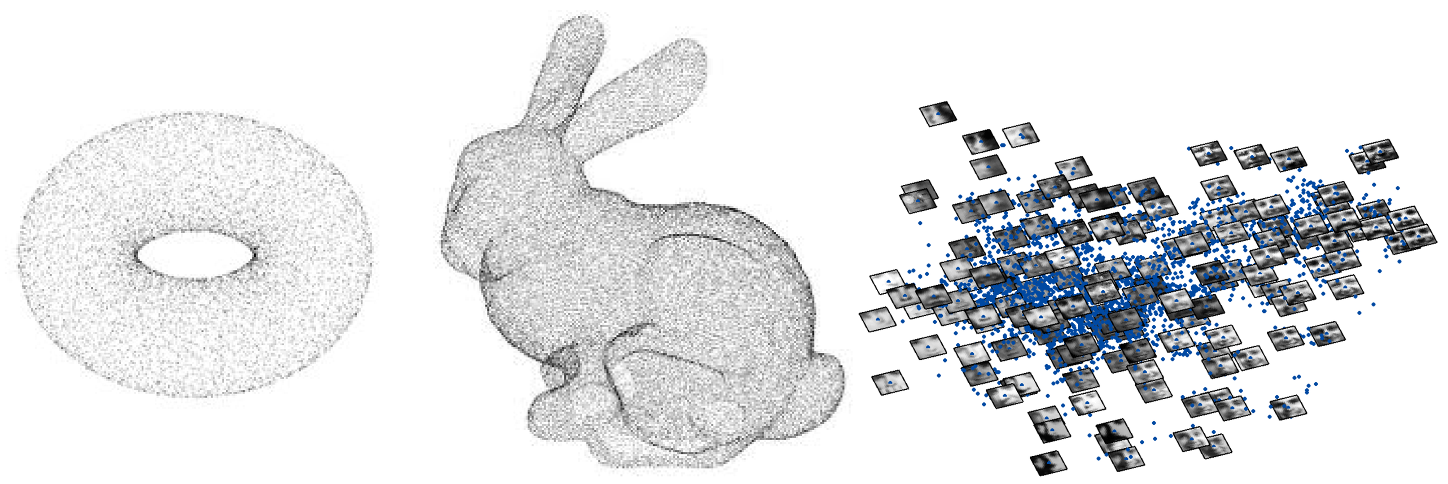
A Prototype for Graph-based Geometric Data Analysis

Griffin Smith, Xingzi Xu, Dr. Dominique Zosso

Department of Mathematical Sciences, Montana State University

Introduction

In our current age, we are bombarded with information. Interpreting and making predictions from gathered data is crucial.
Here, we see a single data point as a sample in a high-dimensional feature space; an individual document, person, image, or record, for example:



Think of linear regression models: Once we assume a linear model, we can find optimal parameters that help us make predictions. How do we know that a given high-dimensional and noisy data set stems from such a simple manifold, to begin with?

Overall Goal: understand the underlying geometry from samples.
This Project: develop a low-dimensional prototype.

Background: Convex Geometry

Consider K **convex** and $\varepsilon > 0$. Then **thicken** with ball B_ε of radius ε:

$$K_\varepsilon = K \oplus B_\varepsilon = \bigcup_{x \in K} B_\varepsilon(x) = \{x \in \mathbb{R}^n \mid d(K,x) \leq \varepsilon\}$$

The volume of the thickened body grows polynomially in ε,

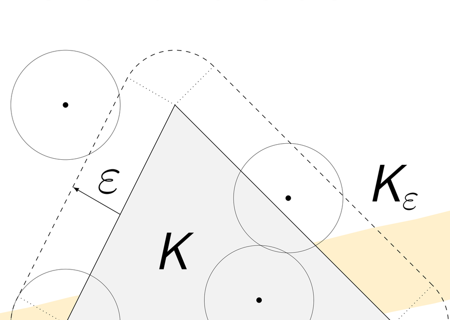
$$|K_\varepsilon| = \sum_{\nu=0}^{n} \binom{n}{\nu} W_\nu(K)\varepsilon^\nu,$$

where the coefficients $W_\nu(K)$ are the **Minkowski functionals or Quermassintegrals**, characterizing the body K: volume, perimeter, ...

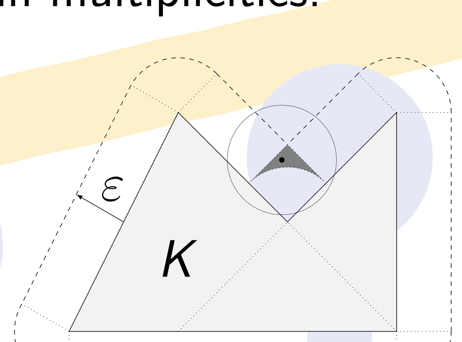
One can extend this into the **convex ring**: For $K = \bigcup_i K_i$ (i.e., K is a finite union of convex bodies K_i):

$$\int \chi(K \cap B_\varepsilon(x))dx = \sum_{\nu=0}^{n} \binom{n}{\nu} W_\nu(K)\varepsilon^\nu$$

Left hand side = volume of thickening with certain multiplicities:



$$|K_\varepsilon| = A_K + P_K\varepsilon + \chi_K \pi \varepsilon^2$$



Convex: simple area of thickening $\qquad \bigcup$ **convex**: area with multiplicity

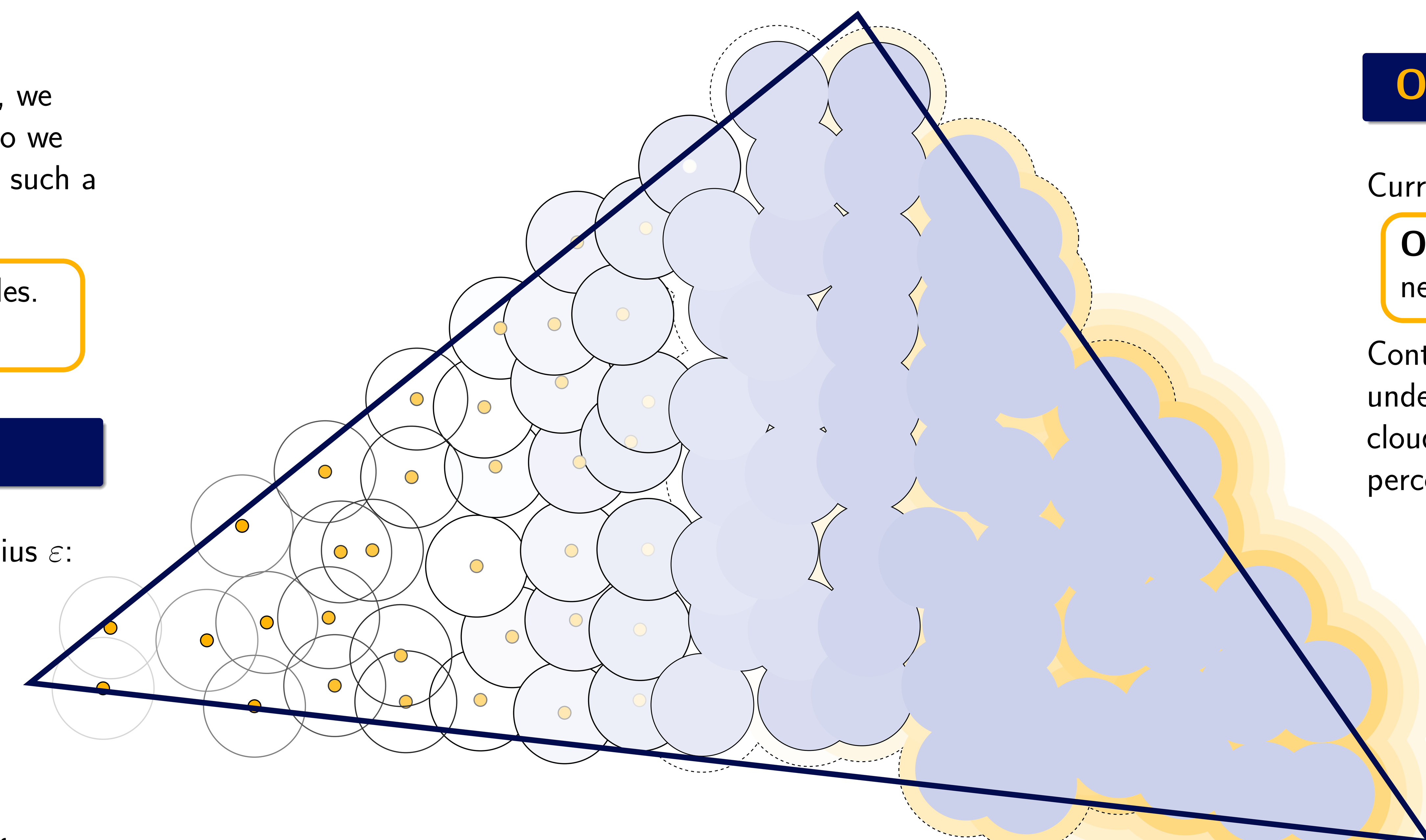
Data Science

Data: N data points in a cloud $\{x_i\}_{i=1}^N$, $x_i \in \mathbb{R}^n$. Since these points have no volume, surface, perimeter, ..., themselves, we create a one-parameter family (r) of polyconvex sets for each point cloud:

$$K_r^N := \bigcup_{i=1}^{N} B_r(x_i).$$

This family interpolates between a dust cloud and a single giant mass.

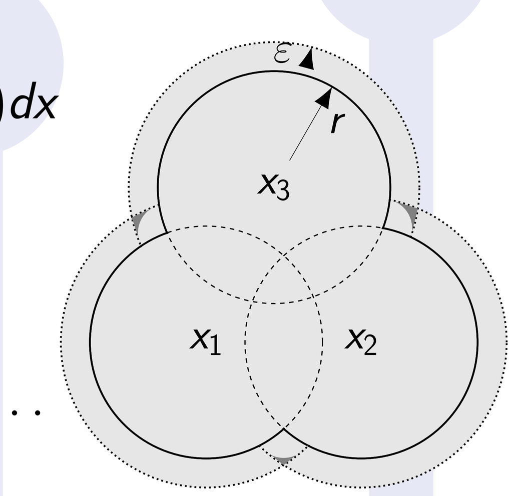
Idea: Compute the volume with multiplicities for various $r + \varepsilon$, then fit a polynomial in ε to estimate geometry of the underlying structure!



"Howto" in Theory

We break-up the area with multiplicity into a collection of simple areas (thickening of balls and their intersections), using the inclusion-exclusion principle $\chi(A \cup B) = \chi(A) + \chi(B) - \chi(A \cap B)$:

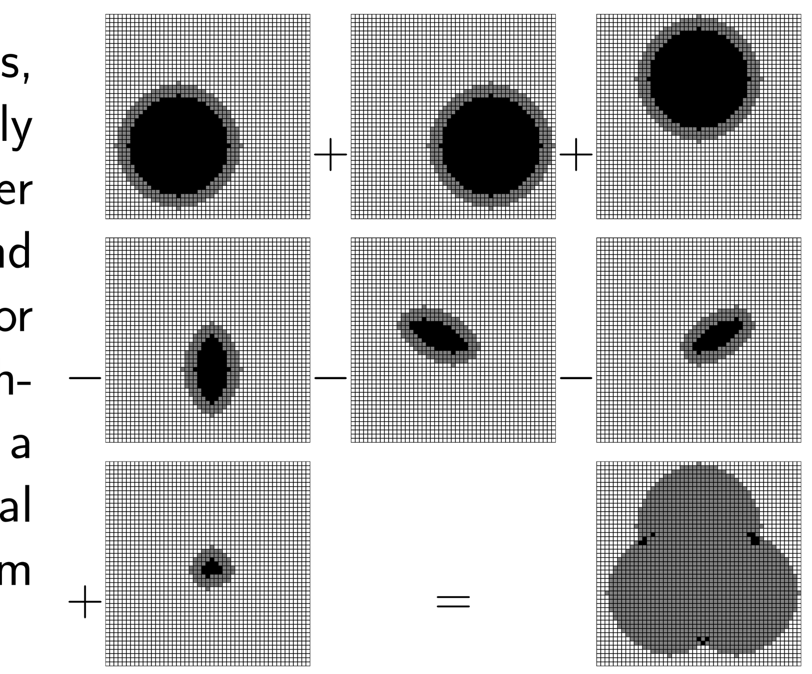
$$\int \chi(K_r^N \cap B_\varepsilon^n(x))dx = \sum_i \int \chi(B_r(x_i) \cap B_\varepsilon^n(x))dx$$
$$- \sum_{ij} \int \chi(B_r(x_i) \cap B_r(x_j) \cap B_\varepsilon^n(x))dx$$
$$+ \sum_{ijk} \int \chi(B_r(x_i) \cap B_r(x_j) \cap B_r(x_k) \cap B_\varepsilon^n(x))dx - \dots$$



To get the thickened area with multiplicity of our union of disks: add all the thickened circle areas; subtract the thickened pairwise overlaps (lenses); add back in the thickened triple intersects, etc.

Naïve MATLAB Implementation

Computing area of spheres, lenses, etc. sounds simple, but is actually hard. In the prototype, we raster thickenings of circles, lenses, etc. and estimate areas by counting pixels, for various ε. The resulting volume-with-multiplicity estimates are fit with a polynomial whose coefficients reveal the geometry of the structure from which data-points were sampled.

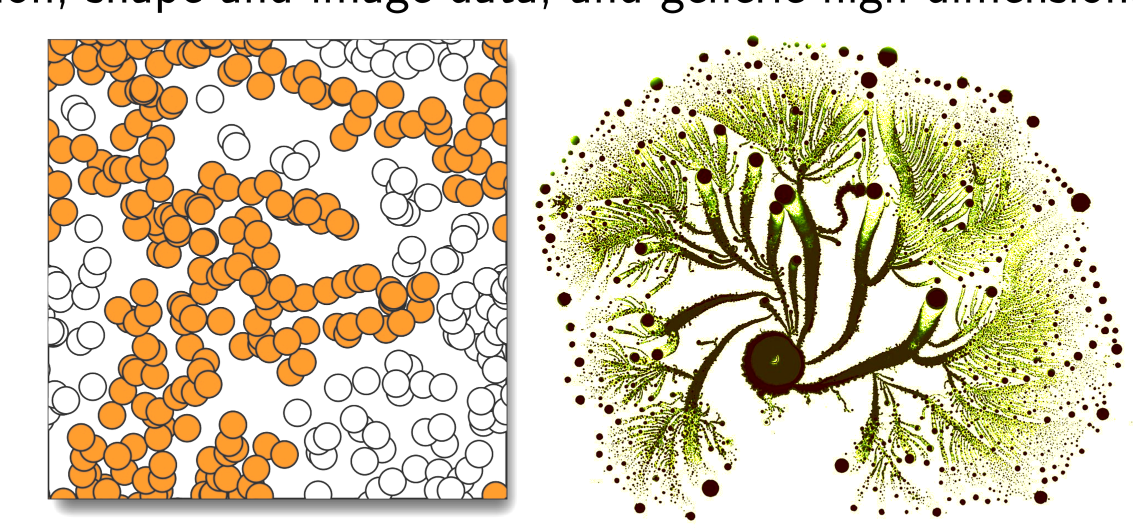


Ongoing and Future Work

Current work focuses on making the prototype faster:

Outlook: Data from the prototype will serve to train artificial neural networks for fast estimates, even with high-dimensional point clouds.

Continuing to explore, learn, and look to the future is pivotal in us understanding all the data that surrounds us. Applications involving data clouds the way we consider them range from bacterial colonies, percolation, shape and image data, and generic high-dimensional data.



Other

- Hugo Hadwiger.
  Vorlesungen über Inhalt, Oberfläche und Isoperimetrie.
  Springer-Verlag, Berlin, Göttingen, Heidelberg, 1957.
- Daniel A Klain and Gian-Carlo Rota.
  Introduction to Geometric Probability.
  Cambridge University Press, 1997.
- Rolf Schneider.
  Convex bodies: the Brunn-Minkowski theory.
  Cambridge University Press, 2014.

This work supported by: **Simons Foundation** (Collab. Grant for Mathematicians)

Undergraduate Scholars Program