# CS 6170: Computational Topology, Spring 2019
# Project 1
## Topological Data Analysis for Data Scientists

Dr. Bei Wang

School of Computing
Scientific Computing and Imaging Institute (SCI)
University of Utah
www.sci.utah.edu/~beiwang
beiwang@sci.utah.edu

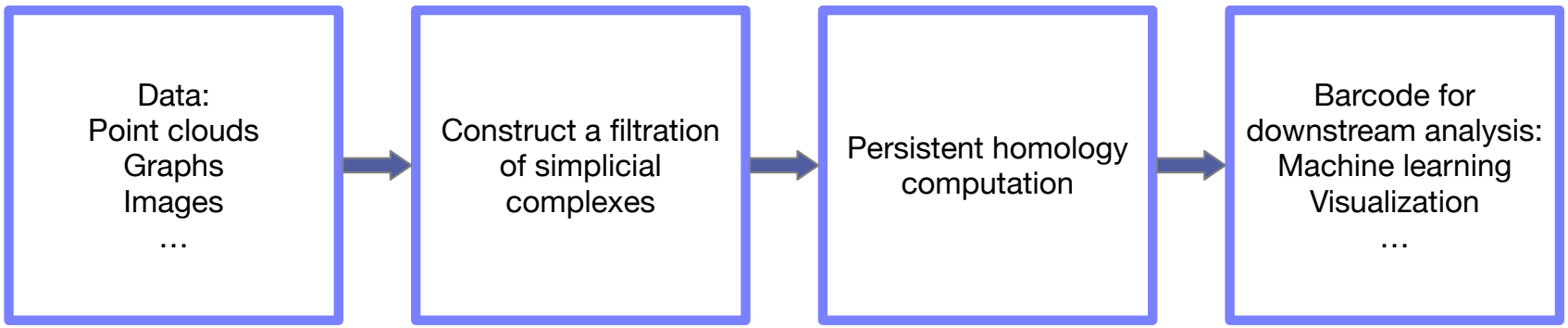
Feb 5, 2019

# Project 1: Overview

# Compute barcodes for synthetic and complex datasets

- Part 1: Compute barcodes of synthetic datasets using Ripser (5 points)
  - `https://github.com/Ripser/ripser`
- Part 2: Compute barcodes of image datasets (15 points)
- Minimal guidelines are provided to simulate a real-world situation faced by a data scientist: there is no user manual for a new dataset.
- Think about this project as a "hack" using existing packages and toolsets, plus some programming.
- The barcodes should be recorded in the form of TXT file, each line contains [birth, death) pair.
- Screen shots (PDF) and source code should be included as part of the submission as well.
- Barcodes (TXT), Screen shots (PDF), report (PDF) and source code should be submitted in a single ZIP file.

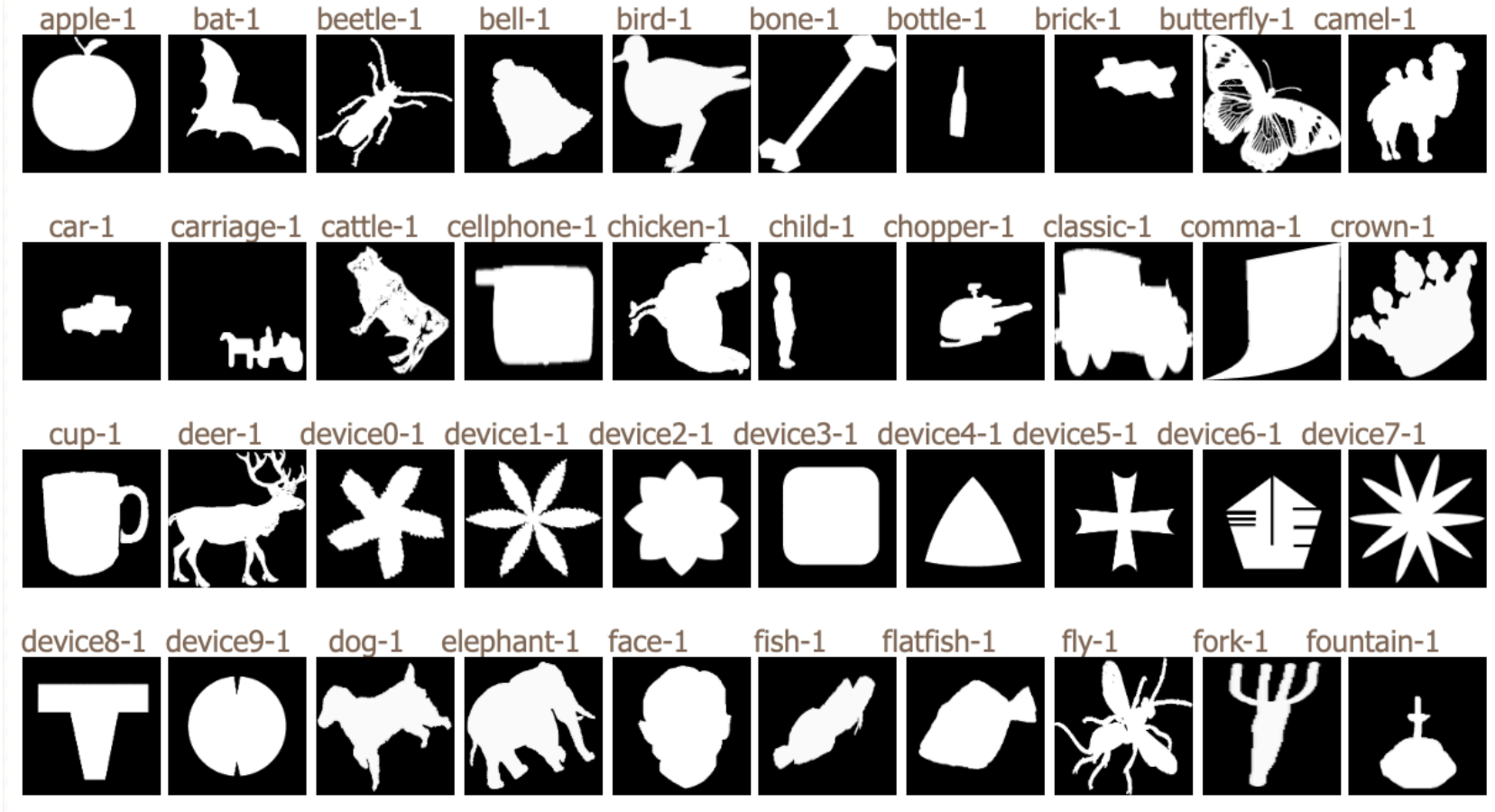
# A TDA pipeline with variations



# Project 1: Data

# Compute barcodes for synthetic and complex datasets

- Part 1: synthetic data sets: octa.txt, cylinder.txt
- Part 2: MPEG-7 image data set
  `http://www.dabi.temple.edu/~shape/MPEG7/dataset.html`



# Project 1 Part 1

# Compute barcodes for datasets

- Using Ripser to compute the dimension 1 persistent homology of *octa* data set; return the first 8 longest bars in the form of [birth, death) pairs (2 points). Take an educated guess as what type of topological space the data may be sampled from given the barcode (1 point).
- Hint: the space should have 4 very large loops with the same size; and 4 slightly smaller loops with the same size.
- Using Ripser to compute the dimension 1 persistent homology of *shape* data set; give an explanation for the longest bar obtained in the barcode (2 points).
- Hint: you can plot these 3D point clouds.

# Project 1 Part 2

# Compute barcodes for MPEG-7 image data set

1. Take 10 sample images from the MPEG-7 image data set (1 from each class) (3 points)
2. Design a filtration for each image
3. Compute barcode in dimension 0 and return the barcode as TXT file for each image (12 points)

# Some pointers

- There are many ways to construct a filtration
- Option 1: recover the 2D points on the boundary of each shape and apply persistent homology computation
- Option 2: a different filtration (persistent homology transform) is used in the paper Hofer et al. (2017): *Deep Learning with Topological Signatures*.
- Hint: read both the paper and its code repo `https://github.com/c-hofer/nips2017`

# References I

Hofer, C., Kwitt, R., Niethammer, M., and Uhl, A. (2017). Deep learning with topological signatures. *Neural Information Processing Systems Conference (NIPS)*.