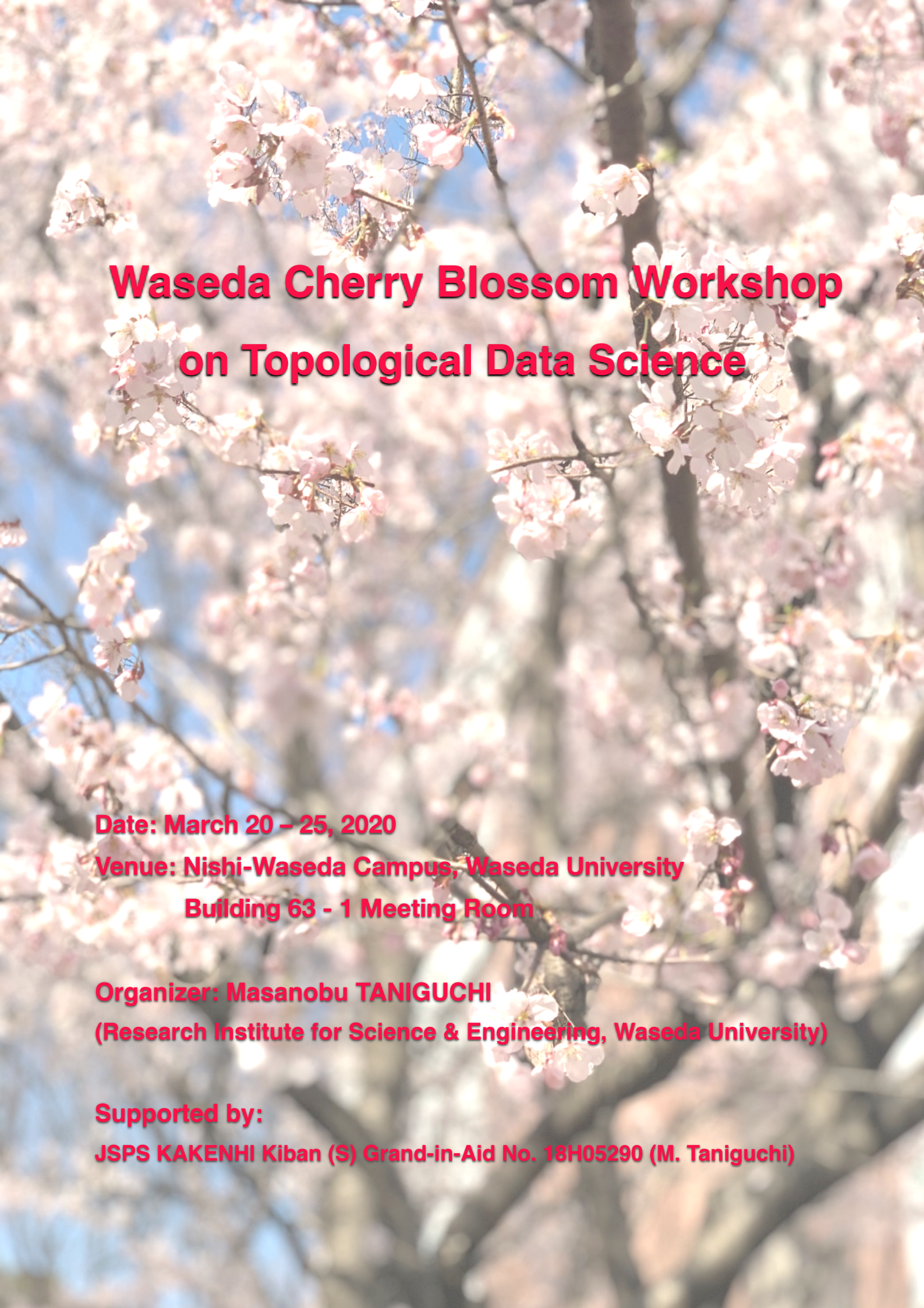

Waseda Cherry Blossom Workshop on Topological Data Science

Date: March 20 – 25, 2020

Venue: Nishi-Waseda Campus, Waseda University
Building 63 - 1 Meeting Room

Organizer: Masanobu TANIGUCHI
(Research Institute for Science & Engineering, Waseda University)

Supported by:
JSPS KAKENHI Kiban (S) Grand-in-Aid No. 18H05290 (M. Taniguchi)

Waseda Cherry Blossom Workshop on Topological Data Science

Date: March 20 – 25, 2020

Venue: Nishi-Waseda Campus, Waseda University

Building 63 - 1 Meeting Room

(Access map: https://www.waseda.jp/fsci/en/access/)

Organizer: Masanobu TANIGUCHI

(Research Institute for Science & Engineering, Waseda University)

This workshop is supported by:

JSPS KAKENHI Kiban (S) Grand-in-Aid No. 18H05290 (M. Taniguchi)

Program

March 20

09:50-10:00: Masanobu Taniguchi (Waseda University)
Opening

10:00-11:00: Yan Liu (Waseda University)
Statistical and Topological Inference of the Granger Causality

11:00-12:00: Takayuki Shiohama (Tokyo Univ. of Science)
Non-stationary Time Series Classification using Topological Data Analysis

12:00-13:30: Lunch Time

13:30-14:30: Momoko Hayamizu (The Institute of Statistical Mathematics, JST PRESTO)
A structure theorem for tree-based phylogenetic networks: from theory to algorithms

15:00-17:00: Discussion

March 21

10:00-12:00: Yasu Wang (Ohio University)
Topological Data Analysis: How it can help in modern data analysis

Lunch & Cherry Blossom Festival

March 23

10:30-12:00: Victor De Oliveira (University of Texas)
An Introduction to Geostatistics, Part I

12:00-13:30: Lunch Time

13:30-15:00: Victor De Oliveira (University of Texas)
An Introduction to Geostatistics, Part II

15:30-17:00: Victor De Oliveira (University of Texas)
Gaussian Copula Models for Geostatistical Count Data

March 24

10:30-12:00: Frederic Chazal (INRIA, France)
An introduction to Topological Data Analysis, Part I

12:00-13:30: Lunch Time

13:30-15:00: Frederic Chazal (INRIA, France)
An introduction to Topological Data Analysis. Part II

15:30-17:00: Frederic Chazal (INRIA, France)
Linearization of persistence and the density of expected persistence diagrams

Abstracts

March 20 (10:00–14:30)

Yan Liu

Title: Statistical and Topological Inference of the Granger Causality

Abstract: We propose a topological approach to statistically analyzing the Granger causality. Granger introduced his celebrated new measure of causality in the sense of prediction errors of multivariate time series 50 years ago. We localize his idea and construct a theory based on locally stationary processes for its alternative version, a natural refinement for stationary processes by Hosoya. To construct the theory, we provide a Gaussian approximation of the suprema of empirical spectral processes. Especially, the local extension of the theory serves for the statistical inference for the Granger causality curve. In addition, we provide a bootstrap procedure for the approximation to construct confidence bands. Finally, we discuss the persistence diagrams and persistence landscapes for the causality curves and numerically construct some examples of locally stationary processes for our simulations studies. (Joint work with Akitoshi Kimura, Masanobu Taniguchi and Hernando Ombao)

Takayuki Shiohama

Title: Non-stationary Time Series Classification using Topological Data Analysis

Abstract: Time series classification (TSC) is an important and challenging problem in data mining. There are hundreds of algorithms for TSC

available with the increase of time series data availability. For more details, we refer recent work of Bagnall et al. (2017) and review paper of Fawaz et al. (2019). Some of the machine learning algorithms are based on the bug of patterns, and learning patterns of similarity are key feature extraction for time series data. In this study, we employ feature extraction of time series using Topological Data Analysis (TDA). TDA refers to a collection of methods for finding topological structure in data. Until recently, topological inference relied on deterministic approaches, and it is well known that these inference is easily affected by outliers and/or noisy datasets. Moreover, the high computational costs are required for computing persistence homology with complex datasets in time and space. Some real data analysis is illustrated how TDA works well in TSC.

Momoko Hayamizu

Title: A structure theorem for tree-based phylogenetic networks: from theory to algorithms

Abstract: While phylogenetic networks are useful to visualise non-treelike data or complex evolutionary histories, there are many computationally hard problems regarding them. Therefore, it is important to define nice subclasses of phylogenetic networks that are mathematically tractable and biologically meaningful. In view of this, the concept of "tree-based" phylogenetic networks, which was originally introduced by Francis and Steel in 2015, has attracted great attention and given rise to various interesting research problems in combinatorial phylogenetics. In this talk, I provide the necessary background and explain how to solve those different problems in a unified manner. The talk is mainly based on arXiv:1811.05849 [math.CO]. I also mention more recent advancement that is joint work with Kazuhisa Makino (arXiv:1904.12432 [math.CO]).

March 21 (10:00–12:00)

Yasu Wang

Title: Topological Data Analysis: How it can help in modern data analysis

Abstract: In recent years, a new field for data, called Topological data analysis, has attracted much attention from researchers from diverse background, including computer science, applied mathematics and statistics. Leveraging various fundamental developments both in theoretical and algorithmic fronts in the past two decades, topological data analysis has been growing rapidly, and already applied in many applied domains, such as computational neuroscience, material science and bioinformatics.

In this time, I want to give some examples on where topological ideas could help with analyzing complex modern data. I will specifically focus on the following three aspects: (1) Topolgoical methods could provide flexibile yet generic framework for feature summarization / characterization. (2) Topological methods could help model, infer, and explore the hidden space behind data. (3) How to combine topological ideas with machine learning pipelines. I will use recent work from my research group to illustrate these points. Through the course, we will touch upon multiple topological objects, including persistent homology, discrete Morse theory, and contour trees.

March 23 (10:00–17:00)

Victor De Oliveira (10:00-12:30)

Title: An Introduction to Geostatistics, Part I

Abstract: In this talk I introduce some of the types of data and scientific problems for which geostatistics is used, the basic probabilistic tools needed to model geostatistical data, and the classical statistical methods of analysis. First, I describe the semivariogram function, the basic tool used in geostatistics to model the spatial association displayed by the quantity of interest, and then I describe the classical methods used for its estimation. These involve a two—step approach that is distribution-free as is based on moments and least squares. The pros and cons of these classical methods are discussed. Second, I describe several variants of the so--called 'kriging' prediction method, which are nothing other than applications of best linear unbiased prediction. I will review some of the properties of these predictors and their mean squared prediction errors, as well as the ability (or lack of) of the latter to properly account for the prediction uncertainty. The pros and cons of kriging predictors are discussed. The models and methods will be illustrated with several real data sets.

Victor De Oliveira (13:30-15:00)

Title: An Introduction to Geostatistics, Part II

Abstract: In this talk I introduce models for geostatistical data based on Gaussian random fields. First, I describe the frequentist methods of maximum likelihood and restricted maximum likelihood to estimate the model parameters, as well as some of the properties of these. I also describe the optimal predictors and their relation to kriging predictors. The two main asymptotic frameworks for this type of data are reviewed,

called increasing and fixed domain frameworks, and the dissimilar large--sample properties of maximum likelihood estimators under these two frameworks are discussed. Second, Bayesian methods for estimation and prediction are described as well as some basic Markov chain Monte Carlo algorithms currently used to make inference about these models. The issue of how to select `good priors' for these model is also briefly discussed. Finally, two classes of non--Gaussian models are introduced to describe continuous data with skewed distributions and geostatistical count data that use Gaussian random fields as building blocks: transformed Gaussian random fields and Poisson hierarchical models. The models and methods will be illustrated with several real data sets.

Victor De Oliveira (15:30-1700)

Title: Gaussian Copula Models for Geostatistical Count Data

Abstract: In this talk I describe a class of random field models for geostatistical count data based on Gaussian copulas. Unlike hierarchical Poisson models often used to describe this type of data, Gaussian copula models allow a more direct modeling of the marginal distributions and association structure of the count data. I describe in detail the correlation structure of these random fields when the family of marginal distributions is either negative binomial or zero--inflated Poisson; these represent two types of overdispersion often encountered in geostatistical count data. I also contrast the correlation structure of one of these Gaussian copula models with that of a hierarchical Poisson model having the same family of marginal distributions. I also describe the computation of maximum likelihood estimators which are a computationally challenging task. Finally, a data analysis of Lansing Woods tree counts is used to illustrate the methods.

March 24 (10:00–14:30)

Frederic Chazal (10:00-12:30)

Title: An introduction to Topological Data Analysis

Part I: persistent homology theory

Abstract: Topological Data Analysis (TDA) is a recent and fast growing field providing a set of new topological and geometric tools to infer relevant features of possibly complex data. Among these tools, persistent homology plays a central role. It provides a mathematically well-founded basis to design efficient and robust methods to estimate, analyze and exploit the topological and geometric structure of data. This first talk will be dedicated to a brief introduction to persistent homology and its usage in TDA. We will introduce persistent homology for functions and point cloud data and study its stability properties. The talk does not require any specific background in topology, the basic notions needed to introduce the persistent homology will be recalled or introduced during the talks.

Frederic Chazal (13:30-15:00)

Title: An introduction to Topological Data Analysis

Part II: statistical properties of persistent homology

Abstract: This second talk will be dedicated to the statistical study of persistent homology. We will show how the stability properties of persistence can be used to understand the behavior of persistence diagrams in various (selected) statistical settings. We will also illustrate how these statistical properties can be used to overcome some computational and noise issues encountered in practical TDA applications.

Frederic Chazal (15:30-17:00)

Title: Linearization of persistence and the density of expected persistence diagrams

Abstract: Persistence diagrams play a fundamental role in Topological Data Analysis (TDA) where they are used as topological descriptors of data represented as point cloud. They consist in discrete multisets of points in the plane that can equivalently be seen as discrete measures. When they are built on top of random data sets, persistence diagrams become random measures. In this talk, we will show that, in many cases, the expectation of these random discrete measures has a density with respect to the Lebesgue measure in the plane. We will discuss its estimation and show that various classical representations of persistence diagrams (persistence images, Betti curves,...) can be seen as kernel-based estimates of quantities deduced from it. This is a joint work with Vincent Divol (ENS Paris / Inria DataShape team).