Plenary Session

Prof. Peter Bühlmann: Robust, Generalizable and Causal-oriented Statistical Machine Learning

Abstract:
Reliable, robust and interpretable machine learning is a big emerging theme in data science and artificial intelligence, complementing the development of pure black box prediction algorithms. New connections between distributional robustness, external validity and causality provide methodological paths for improving the reliability and understanding of machine learning algorithms, with wide-ranging prospects for various applications.

First two parallel sessions

Statistics in Medicine (First session)

Session organizer: Micha Mandel

Shai Carmi, Hebrew U.

1. Title: Screening human IVF embryos with polygenic risk scores: evaluating the expected risk reduction

Abstract:
Polygenic risk scores (PRSs) are already offered commercially in the USA to screen IVF embryos for genetic liability to adult diseases, despite an incomplete understanding of the expected outcomes and in addition to social and ethical concerns. We used statistical genetics modeling to predict the expected reduction in complex disease risk when screening embryos for a single disease. We showed that selecting the embryo with the lowest PRS can lead to substantial relative risk reductions under realistic, though best-case, settings. I will discuss the impact of several factors on risk reduction, including the accuracy of the PRS, the number of available embryos, the disease prevalence, and the parental disease status. I will also discuss practical limitations and barriers to implementation.

2. Yael Travis-Lumer, Technion

Title: Effect size quantification for interrupted time series analysis with application to Covid-19 research

Abstract:
Interrupted time series (ITS) analysis is a time series regression model that aims to evaluate the effect of an intervention on an outcome of interest. ITS analysis is a quasi-experimental study design instrumental in situations where natural experiments occur, gaining popularity, particularly due to the Covid-19 pandemic. However, challenges, including the lack of a control group, have impeded the quantification of the effect size in ITS. We quantify the effect size of an ITS regression model for continuous and count outcomes, with or without seasonal adjustment. The effect size, together with its corresponding 95% confidence interval (CI) and P-value, is based on the ITS model-based fitted values and the predicted counterfactual (the exposed period had the intervention not occurred) values. We apply our method to national data to quantify the effect size of the Covid-19 period on several public health outcomes including suicide attempts, schizophrenia, and mortality. We found a statistically significant effect size for all three outcomes, with a reduced relative risk for both suicide attempts and schizophrenia, and an increased all-cause mortality rate.

---

3. Anat Reiner-Benaim, BGU.

Title: Time-course analysis of multiple endpoints in real-world medical data, with an example on hospitalized COVID-19 patients

Abstract:
Real world data, which is retrieved from continuously accumulating electronic databases, poses many statistical challenges, including selection bias, missing data and sparsity. Specifically, time course medical data is often characterised by heterogenous measurement times and measurement frequencies between patients, as well as repeated admissions of typically a small portion of the patients. Furthermore, when many endpoints are of concern, multiple testing of diversely distributed variables is involved. Post hoc analysis, aimed to focus the timing of measurement dynamics, is often required. I will discuss approaches to address these concerns, including a non-parametric mixed effect model, hierarchical multiple testing and a standardized heatmap. I will demonstrate an integrated analysis within a study on disease trajectory among hospitalized COVID-19 patients.

---

4. Yair Goldberg, Technion

Title: Comparing the protection of 2-dose vaccine and the booster: A quasi-experiment approach

Abstract:
In August 2021, Israel began administering the BNT162b2 booster dose to restore protection following the waning of the 2-dose vaccine. Biological studies have shown that a “fresh” booster

dose leads to increased antibody levels compared to a fresh 2-dose vaccine, which may suggest increased effectiveness. To compare the real-world effectiveness of a fresh (up to 60 days) booster dose with that of a fresh 2-dose vaccine, we took advantage of a quasi-experimental study that compares populations that were eligible to receive the vaccine at different times due to age-dependent policies. Specifically, we compared the confirmed infection rates in adolescents aged 12-14 who received the 2-dose vaccine and in adolescents aged 16-18 who received the booster dose. Our analysis shows that the confirmed infection rate was lower by a factor of 3.7 in the booster group.

This is a joint work with Ofra Amir, Micha Mandel, Yinon M. Bar-On, Omri Bodenheimer, Nachman Ash, Sharon Alroy-Preis, Amit Huppert, and Ron Milo

Out-of-domain generalization: when the train and test distributions don’t match (First Session)

Session organizer: Uri Shalit

1. Sara Magliacane, University of Amsterdam, MIT-IBM Watson AI Lab

Title: Causality-inspired ML: what can causality do for ML? The domain adaptation case

Abstract:
Applying machine learning to real-world cases often requires methods that are robust w.r.t. heterogeneity, missing not at random or corrupt data, selection bias, non i.i.d. data etc. and that can generalize across different domains. Moreover, many tasks are inherently trying to answer causal questions and gather actionable insights, a task for which correlations are usually not enough. Several of these issues are addressed in the rich causal inference literature. On the other hand, often classical causal inference methods require either a complete knowledge of a causal graph or enough experimental data (interventions) to estimate it accurately.

Recently, a new line of research has focused on causality-inspired machine learning, i.e. on the application ideas from causal inference to machine learning methods without necessarily knowing or even trying to estimate the complete causal graph. In this talk, I will present an example of this line of research in the unsupervised domain adaptation case, in which we have labelled data in a set of source domains and unlabelled data in a target domain ("zero-shot"), for which we want to predict the labels. In particular, given certain assumptions, our approach is able to select a set of provably "stable" features (a separating set), for which the generalization error can be bound, even in case of arbitrarily large distribution shifts. As opposed to other works, it also exploits the information in the unlabelled target data, allowing for some unseen shifts w.r.t. to the source domains. While using ideas from causal inference, our method never aims at reconstructing the causal graph or even the Markov equivalence class, showing that causal inference ideas can help machine learning even in this more relaxed setting.

2. Yair Carmon, School of Computer Science, Tel Aviv University

Title: When is out-of-distribution generalization predictable?

Abstract:
To make machine learning reliable, we must understand generalization to out-of-distribution environments (unseen during training) in addition to the in-distribution generalization measured by standard test sets. This talk will present empirical findings showing that, to very good approximation and across diverse distribution shifts and learning algorithms, out-of-distribution performance is a simple function of in-distribution performance. We will also discuss exceptions to this relationship and test hypotheses for their causes. Finally, we will review theoretical models attempting to explain this phenomenon and discuss its implications.

---

3. Yonatan Belinkov, Faculty of Computer Science, Technion

Title: Advances in Characterizing and Improving Out-of-Distribution Generalization in Natural Language Processing

Abstract:
Deep neural networks trained on large amounts of data achieve state-of-the-art results on many natural language processing (NLP) tasks. However, when they are evaluated outside of their training distribution (OOD), these models typically exhibit much worse performance. In this talk, I will describe recent work on characterizing situations that lead to poor OOD generalization, focusing on the role of spurious correlations found in common datasets. I will discuss how the generalization capabilities of models are reflected in their internal representations using information-theoretic measures. Finally, I'll show how a generative reformulation of a classification task leads to models that are less prone to capturing said spurious correlations and exhibit similar performance in and out of distribution.

---

4. Gal Chechik, Bar-Ilan University and director of AI research, NVIDIA,

Title: Adapting to new federated-learning clients using Hypernetworks

Abstract:

In Federated learning (FL), multiple clients collaborate to learn a shared model through a central server while they keep data decentralized. Personalized federated learning (PFL) further extends FL by learning personalized models per client. In both FL and PFL, all clients participate in the training process and their labeled data is used for training. However, in reality, novel clients may wish to join a prediction service after it has been deployed. These new clients may have data from a different distribution, and sometimes may not even have any labeled data. I will first

discuss an approach for personalized federated learning using Hypernetworks, and then discuss how they can be extended to produce a new model for the late-to-the-party client.

Second two parallel sessions

Model-free statistical inference (Second Session)

Session organizer: Yaniv Romano

1. Asaf Weinstein, *Assistant Professor at the Statistics and Data Science Department at the Hebrew University of Jerusalem*

Title: A Power Analysis for Model-X Knockoffs with l_p-Regularized Statistics

Abstract:
Variable selection properties of procedures utilizing penalized-likelihood estimates is a central topic in the study of high dimensional linear regression problems. Existing literature emphasizes the quality of ranking of the variables by such procedures as reflected in the receiver operating characteristic curve or in prediction performance. Specifically, recent works have harnessed modern theory of approximate message-passing (AMP) to obtain, in a particular setting, exact asymptotic predictions of the type I–type II error tradeoff for selection procedures that rely on l_p-regularized estimators.
In practice, effective ranking by itself is often not enough because some calibration for Type I error is required. In this work we study theoretically the power of selection procedures that similarly rank the features by the size of an l_p-regularized estimator, but further use Model-X knockoffs to control the false discovery rate in the realistic situation where no prior information about the signal is available. In analyzing the power of the resulting procedure, we extend existing results in AMP theory to handle the pairing between original variables and their knockoffs. This is used to derive exact asymptotic predictions for power. We apply the general results to compare the power of the knockoffs versions of Lasso and thresholded-Lasso selection, and show that in the i.i.d. covariates setting under consideration, tuning by cross-validation on the augmented design matrix is nearly optimal. We further demonstrate how the techniques allow to analyze also the Type S error when selections are supplemented with a decision on the sign of the coefficients.

---

2. Tzviel Frostig, *Ph.D. Candidate, Department of Statistics and Operation Research, Tel Aviv University*

Title: Statistically Testing for Out Of Distribution Examples in Deep Neural Networks

Abstract:

Building on the conformal predictors we frame Out Of Distribution (OOD) detection in Deep Neural Networks (DNN) as a statistical hypothesis testing problem. We suggest a novel OOD procedure based on low-order statistics and multiple-comparison based normalization. The results achieved by the method are better (or comparable) compared to the state-of-the-art methods on well accepted OOD benchmarks. The method does not require the retraining of the DNN and is computationally efficient.

---

3. Shalev Shaer, *Direct Ph.D. Candidate, Department of Electrical and Computer Engineering, Technion—Israel Institute of Technology*

Title: Learning to Increase the Power of Model-X Randomization Tests

Abstract:
This talk deals with the task of accurately identifying a subset of features associated with a response understudy while controlling the false discovery rate. We introduce model-fitting schemes designed to improve the power of the model-X randomization test, a general framework for conditional independence testing. Our key novelty is in formulating the 'risk discrepancy loss' that contrasts between the empirical risk (test statistic) of a model applied to the original data and its carefully designed artificial copy that mimics the null distribution. Using synthetic and real data sets, we demonstrate that the combination of our proposal with various base predictive models (lasso, elastic net, and deep neural networks) consistently improves the power of the underlying controlled feature selection procedure.

---

4. Nadav Trumer, *Istra Research*

Title: A non-parametric test for analyzing A/B testing results while controlling over two exogenous factors with applications for trials of trading strategy variations

A common randomized control trial in trading operations would independently randomly assign every day a stock to trade in either one of two possible strategies (control and treatment). At the end of the trial, we want to decide which one is better. Large variance of the observed results is explained by both the specific “date” and the specific “stock”. Thus, when analyzing any statistic (for example when calculating p-values) it is desirable (more power) that we control over those two exogenous factors when estimating its distribution under the null hypothesis. The distribution of our statistic under the proposed null hypothesis is not known analytically but we show how to sample from it using MCMC. In large settings, this might be too computationally heavy. We therefore suggest a parametric approximation to the null distribution, which can be fitted by solving a quadratic programming problem and in many cases boils down to solving a linear equation system. On our real trials data, we observed that the approximation and the MCMC produced very similar p-values.

Machine learning for the public good (Second Session)

Session organizer: David Wajnryt

1. Reut Tsarfaty, director of the ONLP lab on Text analysis in Hebrew

Title: Hebrew NLP: What How and Whither

The introduction of neural network models into natural language processing (NLP) has brought to unprecedented advances in NLP and AI. However, such dramatic advances are often reported for English, with Hebrew NLP lagging behind. In this talk I will present the key components of neural NLP pipelines and discuss the particular challenges pertaining to processing Hebrew in these pipelines. I will present solutions and recent advances we developed at the lab to address these limitations, leading to new state-of-the-art results on all the NLP tasks for which Hebrew benchmarks exist.

---

2. Yulia Nudelman, Bank of Israel,

Title: Knowledge Graph: New Data Sources for Better Decisions

Abstract:

בנק ישראל משתמש במקורות נתונים מסורתיים (structured data) לצורך בניית מודלים סטטיסטיים עבור ניטור וחיזוי פעילות כלכלית. בשנים אחרונות, אנחנו עדים להתפתחות טכנולוגית מהירה אשר חושפת בפנינו את הצורך במציאת מקורות מידע חדשים כדי לשפר את המודלים אלו. בדרך כלל מקורות אלו הם יותר פרטיים וזמינים יותר, כמו כתבות של חדשות, מדיה חברתית, דוחות פיננסים ועוד.

אני רוצה להראות איך אנחנו בבנק ישראל בונים גרף ידע ממקורות מידע שונים, מוצאים קשרים חדשים, מייצרים סדרות ומשפרים את המודלים סטטיסטים כדי שנוכל לקבל החלטות יותר טובות.

---

3. Michal Frenkel Head of Innovation branch at the Innovation & Combat Methods Division, J8, IDF

Title: From an enemy to a loved one - People Analytics in security organization - a case study in IDF's Intelligence Division

Abstract:

ארגונים שמבססים את עשייתם על אנשים מחויבים לפתח אסטרטגיה ארגונית בהתבסס על מערכות לניהול ופיתוח כוח אדם מתקדמות ומבוססות מידע (Data driven). תופעה זו אינה פוסחת גם על צה"ל כארגון הגדול ביותר במדינת ישראל. ארגון זה, מתמודד עם שינויים תכופים במודלי השירות של אנשיו. לכן, המעבר לשימוש בנתוני עתק במסגרת תהליכי קבלת החלטות על כוח אדם הינו מחויב המציאות. ההרצאה תציג מקרה מבחן לפרויקט בתחום הPeople -

Big Data שבוצע באגף המודיעין, החוזקות והקשיים בדרך למימוש במעבר מארגון הרגיל לאסוף ולעבד Analytics על "צד אדום", אל פרקטיקות דומות ב"צד הכחול."

---

4. Tal Galili, META

Title: Using the Facebook Platform to Create a Global Symptom and Pandemic Effects Monitoring Resource
Abstract:
Facebook is partnering with the University of Maryland (UMD) Joint Program in Survey Methodology (JPSM) and the Carnegie Mellon University (CMU) Delphi Research Center to support COVID-19 research. Facebook invites its users in more than 200 countries or territories globally to take a survey hosted and collected by either CMU Delphi (US-only) or UMD JPSM (globally). As part of this initiative, we apply best practices from survey statistics to design and execute two components: (1) sampling design and (2) survey weights, which make the sample more representative of the general population. This talk describes the methods we use in these efforts in order to allow data users to execute their analyses using the weights.

The talk will be based on the paper "Weights and methodology brief for the COVID-19 symptom survey by University of Maryland and Carnegie Mellon University, in partnership with Facebook"
Which is available here:
https://arxiv.org/abs/2009.14675

Third two parallel sessions

Statistical theory and methods (Third Session)

Session organizer: Boaz Nadler.

1. Tirza Routtenberg, BGU,

Title: estimation after model selection

Abstract:
In many practical parameter estimation problems, such as coefficient estimation of polynomial regression and direction-of-arrival (DOA) estimation, the exact model is unknown, and a model selection stage is performed prior to estimation. This data-based model selection stage affects the

subsequent estimation, e.g., by introducing a selection bias. Thus, new methodologies are needed for both frequentist and Bayesian estimation. This study considers the problem of estimating unknown parameters after a data-based model selection stage. In the considered setup, the selection of a model is equivalent to the recovery of the deterministic support of the unknown parameter vector. We assume that the data-based model selection criterion is given and analyze the consequent Bayesian and frequentist estimation properties for this specific criterion. For Bayesian parameter estimation after model selection, we develop the selective Bayesian Cramér-Rao bound (CRB) on the mean-squared-error (MSE) of coherent estimators that force unselected parameters to zero. Similarly, for the frequentist (non-Bayesian) estimation of deterministic unknown parameters, we derive the corresponding frequentist CRB on the MSE of any coherent estimator, which is also Lehmann-unbiased. To this end, the relevant Lehmann-unbiasedness is defined with respect to the model selection rule. We analyze the properties of the proposed selective CRBs, including the order relation with the oracle CRBs that assume knowledge of the model. The selective CRBs are evaluated in simulations and are shown as an informative lower bound on the performance of practical coherent estimators.

---

2. Roi Weiss, Ariel University.

Title: On Universal Bayes Consistency

Abstract:
A basic, minimum desideratum for a learning rule is consistency: its error should converge to the lowest possible error as the dataset grows. When data instances reside in Euclidean space, several algorithms are well known to be consistent for *any* data distribution, that is, they are universally consistent. However, those rules fail to be consistent for more general instance spaces. This raises two fundamental questions: (i) can one characterize all instance spaces for which a universally consistent learning rule exists? and (ii) is there a single such rule that is universally consistent whenever possible? Such a rule would have the appealing property that if it fails on a given problem, then any other algorithm will fail as well. In this talk I will present some recent positive developments in answering these two questions.

Joint work with Aryeh Kontorovitch, Sivan Sabato, Steve Hanneke, and László Györfi.

---

3. Tomer Levy, TLV,
Title: High Dimensional Classification by Sparse Multinomial Regression

Abstract:
In logistic multinomial regression, we connect the variables to class probabilities via a coefficient vector, in the binary case, or a coefficient matrix. Unlike binary classification, in the multiclass setup one can think about various notions of sparsity associated with different structural assumptions on the coefficients matrix. In this talk, we present three such notions, and

propose convex penalties capturing the specific type of sparsity at hand. In particular, we consider global sparsity, double row-wise sparsity, and low-rank sparsity, and show that with the properly chosen tuning parameters the derived plug-in classifiers attain the minimax generalization error bounds (in terms of misclassification excess risk) within the corresponding classes of multiclass sparse linear classifiers.

---

4. Ariel Jaffe, HUJI.

Title: Recovering tree models via spectral graph theory

Abstract:
Recovering a tree model that accurately represents the developmental process of high-dimensional data is a key challenge in multiple domains. A common setting is the latent tree model, where the task is to infer the tree structure given only observations of its terminal nodes. For example, in phylogenetics, the evolutionary history of a set of organisms is inferred by their nucleotide or protein sequences.
In our work, we incorporate results from spectral graph theory to develop novel methods for recovering latent tree models. We show that the tree structure is strongly related to the spectral properties of a fully connected graph defined over the terminal nodes of the tree. This relation forms the theoretical basis of two new methods: (i) spectral neighbor-joining, where subsets of nodes are iteratively merged to form the full tree, and (ii) spectral top-down recovery, where the terminal nodes are iteratively partitioned into smaller subsets. Comparing our approach to several competing methods, we show that in many settings, spectral methods have stronger theoretical guarantees and work better in practice.

Recent advances in high-dimensional statistics (Third Session)

Session organizer: Daniel Yekutieli

1. Amichai Painsky, Tel-Aviv University

Title: Confidence Intervals for Unobserved Events

Abstract:
Consider a finite sample from an unknown distribution over a countable alphabet. Unobserved events are alphabet symbols which do not appear in the sample. Estimating the probabilities of unobserved events is a basic problem in statistics and related fields, which was extensively studied in the context of point estimation. In this work we introduce a novel interval estimation scheme for unobserved events. Our proposed framework applies selective inference, as we construct confidence intervals (CIs) for the desired set of parameters. Interestingly, we show that obtained CIs are dimension-free, as they do not grow with the alphabet size. Further, we show

that our CIs are (almost) tight, in the sense that they cannot be further improved without violating the prescribed coverage rate. We demonstrate the performance of our proposed scheme in synthetic and real-world experiments, showing a significant improvement over alternatives.

---

2. Assaf Rabinowicz, Spotitearly

Title: Tree-Based Models for Correlated Data

Abstract:
This presentation introduces a new approach for regression tree-based models, such as simple regression tree, random forest and gradient boosting, in settings involving correlated data.
I will present the problems that arise when implementing standard regression tree-based models, which ignore the correlation structure. The new suggested approach explicitly takes the correlation
structure into account in the splitting criterion, stopping rules and fitted values in the leaves, which induces some major modifications of standard methodology. The superiority of this new approach over tree-based models that do not account for the correlation, and over previous work that integrated some aspects of our approach, is supported by simulation experiments and real data analyses."

---

3. Alon Kipnis, Reichman University (IDC Herzliya)

Title: Rare and Weak Detection Models using Moderate Deviations Analysis and Log-Chisquared P-values

Abstract:
Rare and weak models for multiple hypothesis testing assume that only a small proportion of the tested hypotheses concern non-null effects and that the individual effects are only moderately large so that they generally do not stand out individually, for example in a Bonferroni analysis. Such rare/weak models have been studied in quite a few settings, for example in some cases studies focused on the underlying Gaussian means model for the hypotheses being tested; in some others, Poisson. It seems not to have been noticed before that such seemingly different models have asymptotically the following common structure: Summarizing the evidence each test provides by the negative logarithm of its P-value, previous rare/weak model settings are asymptotically equivalent to detection where most negative log P-values have a standard exponential distribution but a small fraction of the P-values might possibly have an alternative distribution which is moderately larger; we do not know which individual tests those might be, or even if there are any such. Moreover, the alternative distribution is approximately noncentral chisquared on one degree of freedom.
We characterize the asymptotic performance of global tests combining these P-values in terms of the chisquared mixture parameters: the scaling parameters controlling heteroscedasticity, the

non-centrality parameter describing the effect size whenever it exists, and the parameter controlling the rarity of the non-null effects. Specifically, in a phase space involving the last two parameters, we derive a region where all tests are asymptotically powerless. Outside of this region, some tests like higher criticism have maximal power. Inference techniques based on the minimal P-value, false-discovery-rate controlling, and Fisher's combination test have sub-optimal asymptotic phase diagrams. We provide various examples for multiple testing problems of the said common structure.
Our log-chisquared approximation for P-values is different from Bahadur's log-normal approximation; the log-normal approximation is a large deviations phenomenon, while the effects we study appear instead on the moderate-deviations scale. The log-normal approximation would be unsuitable for understanding Rare/Weak multiple testing models.

Prof. Peter Bühlmann: Workshop on Robust, Generalizable and Causal-oriented Statistical Machine Learning

The workshop will provide more details on the plenary lecture. In particular, we will give a more in-depth discussion on causal statistical machine learning (invariant causal prediction, causal regularization and its distributional robustness), aspects of domain adaptation, and also explain some ideas to cope with situations when latent confounding variables are present. Illustrations on some real data sets are complementing the presentation on methodology and statistical theory.