SISSA: Data Science Offer for AY 2020-21

Final version – 30 September 2020
Roberto Trotta, Guido Sanguinetti & Sebastian Goldt

Introduction

This document sets out the structure and contents of the graduate offer in Data Science by the Data Science Excellence Department, starting in October 2020 (in-person or remote teaching) for the AY 2020-21.

This initial curriculum is intended to form the backbone of what will eventually become the PhD in Data Science for AY 2021-22.

Data Science modules will generally run in the afternoons, so that students from other PhD programmes (which are usually taught in the mornings) will be able to attend.

We kindly request that students from other PhD programmes who are interested in following our modules register their interest by filling out this form:

https://forms.gle/LosEvjrXyUaBhbdy7

This is for logistical (especially in view of COVID-19 restrictions to teaching spaces) and pedagogical reasons. **Deadline is Fri Oct 2nd 2020.**

Our formal learning opportunities will be flanked with a vigorous programme of online seminars (the "SISSA Data Science Seminar Series", or *SISSA DS³*), held approximately fortnightly from January 2021, with a focus on showcasing a young and diverse line-up of world-class speakers from all over the world.

Further details will be published on our webpage:
https://www.sissa.it/data-science-excellence-department-initiative-ds

Framework

- Students from the Data Science Excellence Department must take all core modules offered in Data Science and take at least 3 optional modules, which can also be chosen from the offering from other PhD programmes.
- To proceed to year 2, Data Science Excellence Department students must achieve an average of 27/30 in the the core modules.
- Credit size conversion: 1 credit = 6 hrs of lectures or labs.
- All of our Data Science modules are open to students from other PhD programmes. Credit can be accrued by such students with the written agreement of their PhD programme coordinator.

- Frontal lecturing generally takes place in 2 hrs slots, 2-4pm (but see timetable below for details, with some Labs, Ethics in AI and Journal Club taking place in the mornings due to timetabling constraints), while Labs are 3 hrs slots (2-5pm or 9-12am when necessary).

Teaching Staff

1. Guido Sanguinetti (SISSA)
2. Roberto Trotta (SISSA)
3. Sebastian Goldt (SISSA)
4. Alessandro Laio (SISSA)
5. Andrea de Simone (SISSA/Uni Camerino)
6. Jean Barbier (ICTP)
7. Luca Bortolussi (UniTS) + Guest lecturers
8. Guest lecturers for "Ethics in AI" module.

Modules on offer and timetable

Module name	Lecturer	Term	Type	Weeks	Dates	Lectu-res (hrs)	Labs (hrs)	Total (hrs)	Credits
Introduction to statistical modelling and inference	De Simone (delivered remotely)	1	Core	4	05/10 - 30/10	20	6	26	4.33
Bayesian inference I	Sanguinetti	1	Core	7	19/10 - 04/12	24	12	36	6.00
Information Theory and Inference	Barbier	1	Core	5	04/11 - 04/12	20	6	26	4.33
Unsupervised Learning and non-Parametric methods	Laio	2	Core	6	11/01 - 19/02	20	18	38	6.33
Neural Networks	Goldt	2	Core	6	25/01 - 05/03	24	12	36	6.00
Bayesian Inference II	Trotta	2	Core	6	11/01 - 24/02	24	12	36	6.00
Ethics in AI	Trotta + guests	year long	Core		9/11 onwards	20	0	20	3.33
Scientific Programming and Algorithms	Bortolussi + guests	2 or 3	Core	4	18/01-23/02	24	12	36	6.00
Monographics courses	Sanguinetti, Trotta, Goldt	3	Option	21	14/01 - 24/06	42	0	42	7.00

SISSA Data Science Offering 2020/21

FINAL v 1.0, RT, 21/09/2020

All modules will take place in Aula 128-129 (capacity: up to 28), except for Journal Club and monographic courses which will be in Aula 005 (capacity 16+9 floating)

2020

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Oct	Th	Fr	Sa	Su	Mo 2-4 (1)	Tu	We 2-4 (2)	Th	Fr 2-4 (3)	Sa	Su	Mo 2-4 (4)	Tu	We 2-5 Lab1	Th	Fr 2-4 (5)
Nov	Su	Mo	Tu	We 2-4 (5)	Th 2-5 Lab1	Fr 2-4 (1)	Sa	Su	Mo 2-4 (2)	Tu 2-4 (6)	We 2-5 Lab2	Th 2-4 (7)	Fr 2-4 (3)	Sa	Su	Mo 2-4 (4)
Dec	Tu 2-4 (12)	We 2-4 (9)	Th 2-5 Lab4	Fr 2-4 (10)	Sa	Su	Mo	Tu	We	Th	Fr	Sa	Su	Mo	Tu	We

2021

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Jan	Fr	Sa	Su	Mo	Tu	We	Th	Fr	Sa	Su	Mo 2-4 (1)	Tu 2-4 (1)	We 2-4 (2)	Th 2-4 (2)	Fr 9-12 Lab1	Sa
Feb	Mo 2-4 (7)	Tu 2-4 (7) Lab2 9-12	We 2-4 (8)	Th 2-4 (8)	Fr 9-12 Lab4	Sa	Su	Mo 2-4 (9)	Tu 2-4 (9) Lab3 9-12	We 2-4 (10)	Th 2-4 (10)	Fr 9-12 Lab5	Sa	Su	Mo	Tu
Mar	Mo Lab2 9-12	Tu Lab4 9-12	We 10-12 (5)	Th	Fr 2-4 (6)	Sa	Su	Mo 2-4 (7)	Tu	We 10-12 (8)	Th	Fr	Sa	Su	Mo	Tu Lab3 9-12
Apr	Th	Fr	Sa	Su	Mo	Tu	We	Th	Fr	Sa	Su	Mo	Tu	We	Th	Fr
May	Sa	Su	Mo	Tu	We	Th	Fr	Sa	Su	Mo	Tu	We	Th	Fr	Sa	Su
Jun	Tu	We	Th	Fr	Sa	Su	Mo	Tu	We	Th	Fr	Sa	Su	Mo	Tu	We
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

2020

	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
Oct	Sa	Su	Mo 2-4 (6)	Tu 2-4 (1)	We 2-4 (7)	Th 2-4 (2)	Fr 2-4 (8)	Sa	Su	Mo 2-4 (9)	Tu 2-4 (3)	We 2-5 Lab2	Th 2-4 (4)	Fr 2-4 (10)	Sa
Nov	Tu 2-4 (8) 9-12 Lab2	We 2-4 (5)	Th 2-4 (9)	Fr 2-5 Lab3	Sa	Su	Mo 2-4 (6)	Tu 2-4 (10)	We 2-4 (7)	Th 2-4 (11)	Fr 2-5 Lab2	Sa	Su	Mo 2-4 (8)	
Dec	Th	Fr	Sa	Su	Mo	Tu	We	Th	Fr	Sa	Su	Mo	Tu	We	Th

2021

	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
Jan	Su	Mo 2-4 (3)	Tu 2-4 (3) Lab1 9-12	We 2-4 (4)	Th 2-4 (4)	Fr 9-12 Lab2	Sa	Su	Mo 2-4 (5)	Tu 2-4 (5)	We 2-4 (6)	Th 2-4 (6)	Fr 9-12 Lab3	Sa	Su
Feb	We 10-12 (1)	Th	Fr 9-12 Lab6 2-4 (2)	Sa	Su	Mo 2-4 (11)	Tu Lab1 9-12	We 10-12 (3) 2-4 (12)	Th	Fr 2-4 (4)	Sa	Su			
Mar	We 10-12 (9)	Th	Fr 2-4 (10)	Sa	Su	Mo	Tu Lab4 9-12	We 10-12 (11)	Th	Fr 2-4 (12)	Sa	Su	Mo	Tu	We
Apr	Sa	Su	Mo	Tu	We	Th	Fr	Sa	Su	Mo	Tu	We	Th	Fr	
May	Mo	Tu	We	Th	Fr	Sa	Su	Mo	Tu	We	Th	Fr	Sa	Su	Mo
Jun	Th	Fr	Sa	Su	Mo	Tu	We	Th	Fr	Sa	Su	Mo	Tu	We	
	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31

Introduction to statistical modelling and inference (de Simone)

Bayesian inference I (Sanguinetti)

Information Theory and Inference (Barbier)

Ethics in AI (Trotta + Guests), 10-12 am

Unsupervised Learning and non-Parametric methods (Laio)

Neural Networks (RTDA)

Bayesian Inference II (Trotta)

Journal Club and monographic optional courses (21 2hrs meetings, 10-12am)

Scientific Computing and Algorithms (Bortolussi + Guests, dates TBC)

Syllabus

Term 1 (October-Dec)

1. Introduction to Statistical Modelling and Inference (Andrea de Simone)
   Each topic requires 4 hours = 20 hours + 2 x 3hrs Lab = 26 hours (4 weeks: 05/10-30/10)
   Pre-requisites: familiarity with Python and jupyter installed on students' computers.
   i. *Review of Probability and Statistics:* The language of probability. Random variables. Cumulative distribution function, probability density function. Conditional probability. Bayes theorem. Expected value and variance. Common distributions. Types of Convergence. Asymptotic theorems: Law of Large Numbers, Central Limit Theorem.
   ii. *Basics of Statistical Inference:* Inference: parametric vs non-parametric, frequentist vs Bayesian. Estimators (consistency, bias, variance). Likelihood. Maximum likelihood estimation and maximum a posteriori probability. Confidence and credible intervals. Nuisance parameters.
   iii. Lab 1*:* hands-on jupyter lab on probability and statistics.
   iv. *Hypothesis Testing I:* Significance, acceptance and Bayesian tests. Null hypothesis, p-value, Type I and II errors. Neyman-Pearson lemma. Multiple comparisons. Resampling methods.
   v. *Hypothesis Testing II:* Types of tests: location, independence, homogeneity. Order and rank statistics. Location: Z-test and Student's t-test. Independence: χ^2 test. Homogeneity: Kolmogorov-Smirnov and Mann-Whitney tests. Generalization error and overfitting. Cross-Validation and complexity penalization.
   vi. Lab 2: hands-on jupyter lab on hypothesis testing.

2. Bayesian Inference I (Guido Sanguinetti)
   Each lecture requires 2 hrs; 12 lectures + 4 labs = 32 hours (7 weeks: 19/10 - 4/12)
   i. The multivariate Gaussian distribution: conditionals, marginals, and conjugate prior (and its problems)
   ii. Laplace method and Fisher matrix
   iii. Linear/ Gaussian models: probabilistic PCA and linear regression. Basis function regression.
   iv. Gaussian processes for regression and Bayesian Optimization.
   v. Lab 1: linear regression and Gaussian Processes
   vi. Bayesian inference in non-conjugate models: Markov Chain Monte Carlo (MCMC), rejection and importance sampling, Metropolis-Hastings algorithm. Convergence diagnostics and rules of thumb.
   vii. Generalised linear models (GLMs) and inference; Gaussian processes for classification.
   viii. Lab 2: Bayesian GLMs.
   ix. Graphical models and hierarchical Bayesian models. Gibbs sampling.
   x. Mixture models and topic models.
   xi. Variable augmentation: probit and logistic regression with auxiliary variables
   xii. Lab 3: Gibbs sampling for mixture models.
   xiii. Variational inference: prelude, the EM algorithm
   xiv. Mean-field variational inference
   xv. Variational inference for general models: black-box variational inference and variational autoencoders, Stein variational inference.
   xvi. Lab 4: Variational mean field for mixture models.

3. <u>Information Theory and Inference</u> (Jean Barbier):
   10 x 2 hours + 2 x 3 hrs Labs = 26 hours (5 weeks: 04/11 - 04/12)
   i. Bayesian inference, information theory and statistical mechanics:
      i. Statistical inference, Bayes formula and decision theory
      ii. Surprise, Shannon entropy and mutual information
      iii. Statistical mechanics of disordered systems 101, and links with Bayesian inference
      iv. <u>Lab 1</u>
   ii. Information-theoretic limits
      i. Replica symmetric formula for the mutual information
      ii. A powerful (exact) heuristic: the replica method
      iii. Why ensembles matter? Concentration of the free energy
      iv. Replica symmetry in inference: overlap concentration
      v. Rigorous approach 1: the (adaptive) interpolation method
      vi. Rigorous approach 2: the cavity method
      vii. <u>Lab 2</u>
   iii. Algorithmic limits
      i. Message-passing
      ii. State evolution, and optimality of approximate message-passing

4. <u>Ethics in AI (Trotta as lead, plus guest lecturers):</u>
   10 x 2 hours throughout the year, 1 meeting fortnightly (starts 09/11).
   This interactive module will provide an introduction and overview to ethical issues in ML and AI, and illustrate them with contributions from guest speakers from a variety of fields. A detailed programme will be announced soon.

Term 2 (Jan-Apr)

5. <u>Unsupervised Learning and Non-parametric Methods</u> (Alessandro Laio & guest lecturer Alex Rodriguez, ICTP)
   10 x 2 hours + 6 x 3 hrs Labs = 38 hours (6 weeks: 11/01-19/02)
   i. Introduction: choosing the features and the metric.
   ii. <u>Lab 1</u>
   iii. Dimensional reduction and manifold learning
      i. Linear methods: principal component analysis and multidimensional scaling
      ii. Curved manifolds: ISOMAP, kernel PCA and Sketchmap
      iii. <u>Lab 2</u>
      iv. Diffusion Map and Stochastic Neighbor Embedding
      v. Characterizing the embedding manifold: the intrinsic dimension
      vi. <u>Lab 3</u>
   iv. Estimating the probability density
      i. Parametric density estimators
      ii. Non-parametric estimators: Histograms, Kernel density estimator and k-nearest neighbor estimator
      iii. Adaptive density estimators
      iv. <u>Lab 4</u>
   v. Clustering
      i. Partitioning schemes: k-means, k-medoids and k-centers.
      ii. hierarchical and spectral clustering
      iii. <u>Lab 5</u>
      iv. Density-based clustering

v. Clustering techniques exploiting kinetic information
vi. Lab 6

6. Neural Networks (Sebastian Goldt + guest lecturers)
   12x2h lectures + 4x3hrs labs = 36 hours (6 weeks: 25/01 - 05/03)
   i. Lab 0 (optional): fundamental programming tools and best practice
   ii. Introduction to learning (in high dimensions): Goals of learning; classification vs regression; training vs validation vs testing; linear regression, kernels; fully-connected feedforward networks: representational power; breaking the curse of dimensionality with neural networks?
   iii. Lab 1: neural networks from scratch
   iv. Computer Vision: analysing spatial correlations using convolutions; (the importance of) datasets (CIFAR10 / 100, ImageNet), basic training algorithm: mini-batch SGD; modern architectures (AlexNet, GoogLeNet, ResNet, DenseNet); acceleration techniques (Nesterov, Adam); dropout, batch normalisation.
   v. Lab 2: computer vision with pyTorch
   vi. Machine Learning for the sciences: solving quantum many-body problems with neural networks (case study); guest lectures (TBC)
   vii. Robustness in Deep Learning: adversarial examples and defences
   viii. Unsupervised learning: GANs and normalising flows; semi-supervised learning.
   ix. Recurrent neural networks: Hopfield networks (joint with guest lectures, TBC); vanishing and exploding gradients in recurrent nets; LSTM
   x. Lab 3: Generative models for images
   xi. Graph Neural Networks: introduction to GNNs and one application in science, e.g. protein-protein interactions.
   xii. Introduction to reinforcement learning
   xiii. Lab 4: Reinforcement learning
   xiv. Outlook: From the practice of deep learning to its science; surprises in high dimensions (failure of statistical learning theory bounds), the generalisation puzzle; open problems

7. Bayesian Inference II (Roberto Trotta)
   Each bloc requires 6 hrs; 24 lectures + 4 x 3 hrs labs = 36 hours (5 weeks: 11/01-24/02)
   i. Foundations of Bayesianism: Jaynes' robot; Cox theorem, objective and subjective Bayes; prior choice (maximum entropy, conjugancy, Jeffreys' prior, empirical Bayes, hyperpriors, etc); sensitivity analysis.
   ii. Lab 1: Sensitivity analysis and volume effects.
   iii. Advanced sampling methods: slice sampling, Langevin and Hamiltonian Monte Carlo, collapsed and augmented Gibbs sampling, reversible jump MC.
   iv. Lab 2: Writing an MCMC from scratch.
   v. Bayesian model comparison: stopping rule paradox, p-values, Lindley paradox; Bayesian evidence, Bayes factor and interpretation. Savage-Dickey Density Ratio, Laplace approximation, Approximate Bayesian Computation (ABC).
   vi. Lab 3: Applications of Bayesian model comparison.
   vii. Bayesian model averaging, Bayesian optimization and experiment design.
   viii. Lab 4: Bayesian optimization.

8. Scientific Programming and Algorithms (Luca Bortolussi (lead)+ Guest lecturers: Giulio Caravagna, Luca Manzoni, Lorenzo Castelli - UniTS)
   Each bloc requires 6 hrs of lectures and 3 hrs of Labs = 9 x 4 = 36 hours (18/01-23/02)
   i. Scientific computing in Python (Giulio Caravagna, UniTS)

ii. Programming methods and software development (Luca Manzoni, UniTS)
    i. Reproducibility and version management: virtual environments and git
    ii. How to make a library: Python modules and how to structure the code
    iii. Introduction to testing
iii. Introduction to algorithm and computational complexity (Luca Bortolussi)
    i. Introduction to algorithms, algorithmic design and complexity. Sorting.
    ii. Data structures. Lists, queues, stacks and hash tables. Binary search trees.
    iii. Algorithms on graphs. Connectivity and shortest paths.
iv. Introduction to mathematical optimization (Lorenzo Castelli, UniTS)
    i. Introduction to mathematical optimisation. Linear programming.
    ii. Integer programming
    iii. Exact, approximate and heuristic algorithms.

Term 3 (April-June)

Optional courses

A series of monographic courses will be offered in Term 3, introducing some key open problems in the chosen area and with the aim of taking students to the cutting edge of the current research in this area.

Duration and format of each module will be defined at a later point in time, with the aim of achieving a comparable student experience and workload across the different topics.

Each course will meet for 3 hours (1 afternoon, 2-5pm) weekly over 8 weeks, for the period 19/04-11/06.

Topics offered:

- Machine learning applications to single-cell genomics (Sanguinetti)
- Topics in Bayesian inference and modelling (Trotta)
- Current topics in the theory of neural networks: Dynamics and Data (Goldt)