

Master IASD

Optional courses & internship

masteriasd.eu

Tristan Cazenave, Benjamin Negrevergne, Yasmine Abdi
Pierre Senellart, François Goulette



Program

- General information about optional courses
- Presentation of optional courses
- How to choose a good internship

General information

- Strict minimum: 6 optional courses (18 ECTS).
- You are strongly encouraged to choose more (e.g., 8 optional courses).
- Email your option choices to Yasmine before December 10, 16:00
To: `yasmine.abdi@dauphine.psl.eu`
Cc: `tristan.cazenave@dauphine.psl.eu`
Cc: `benjamin.negrevergne@dauphine.psl.eu`
- Check for potential schedule conflicts before selecting two courses
- Some courses may not be opened if attendance is too low (≤ 4)
► We will do a second round in case too many courses have been closed
- Mines students: at least 38 IASD ECTS (mandatory & optional courses)

Option conflicts

- **Block 1:** Advanced machine learning
- **Block 2:** Introduction to reinforcement learning
- **Block 3:** Natural language processing
- **Block 4:** Knowledge graphs, description logics, rod
Monte-Carlo search and games Mon. a.m.
- **Block 5:** Anonymization, privacy
Incremental learning, game theory, and applications Mon. p.m.
- **Block 6:** Systems, paradigms, languages for Big Data Wed. a.m. + evenings
- **Block 7:** Machine learning on Big Data Wed. a.m. + evenings
- **Block 8:** Data wrangling, data quality
Ethics and artificial intelligence Wed. p.m.
- **Block 9:** Point clouds and 3D modeling
Computational social choice Thu p.m.
- **Block 10:** Graph mining
Deep learning for image analysis Fri. a.m.
Fri. (all day)
- **Block 11:** Deep learning for image analysis
Foundations of graph and RDF query languages Fri. (all day)
Fri. p.m.

Who is in charge?

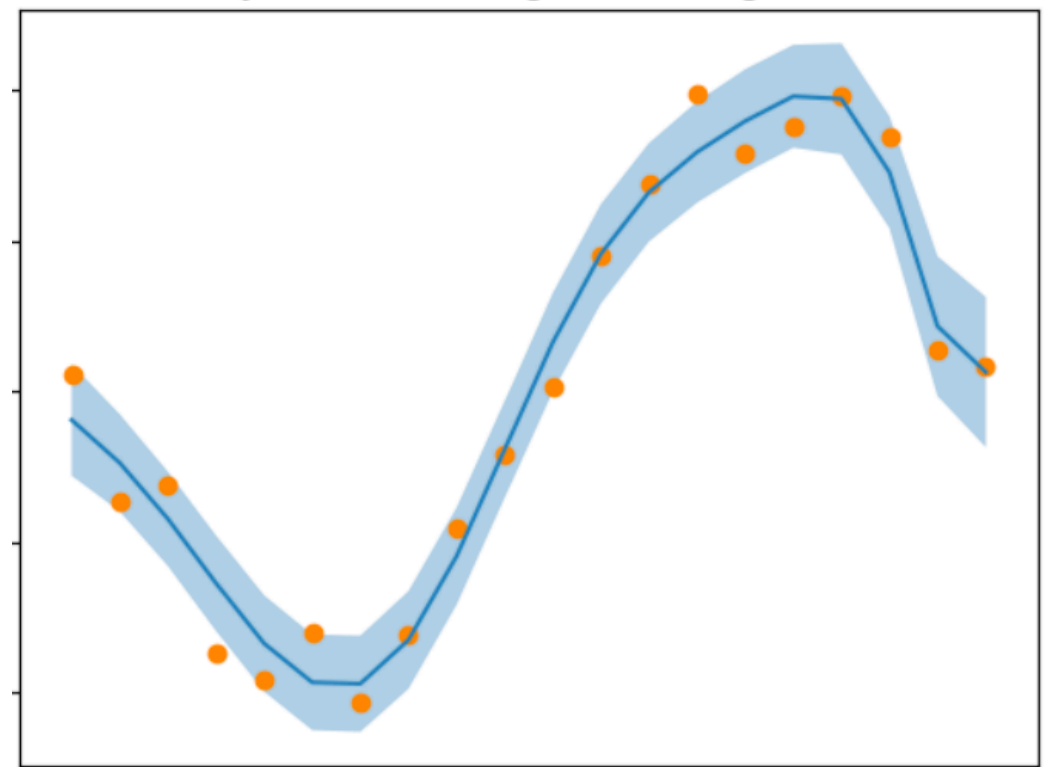
- Advanced machine learning Yann Chevalleyre
- Introduction to reinforcement learning Yann Chevalleyre
- Anonymization, privacy Pierre Senellart
- Computational social choice Jérôme Lang
- Data wrangling, data quality Pierre Senellart
- Deep learning for image analysis Etienne Decencière
- Ethics and artificial intelligence François Goulette
- Point clouds and 3D modeling François Goulette
- Foundations of graph and RDF query languages Leonid Libkin
- Graph mining Daniela Grigori
- Incremental learning, game theory, and applications Rida Laraki
- Knowledge graphs, description logics, reasoning on data Michaël Thomazo and Camille Bourgaux
- Monte-Carlo search and games Tristan Cazenave
- Natural language processing Alexandre Allauzen
- Machine learning on Big Data Dario Colazzo
- Systems, paradigms, and languages for Big Data Dario Colazzo

Apprentissage Automatique Avancé

Objectif du cours

- **Partie 1: Apprentissage Bayésien:**

- Dans le paradigme dominant de l'apprentissage artificiel vu en cours est le « max vraisemblance »
- L'apprentissage Bayésien permet d'aggréger l'ensemble des modèles probables (plutôt que de n'en prendre qu'un) et fournit des intervalles de confiance



- **Partie 2 : Recommandation et Ranking**

- Apprentissage de moteur de recherche, apprendre à partir de labels ordinaux...

Intervenants/Contenu

- **Moez Draif** (VP & chief scientist chez Capgemini, prof à Imperial college).

Methodes bayésiennes en Apprentissage.

- Intro : présentation de la structure du cours
- Approche bayésienne, différences avec la statistique fréquentiste
- Régression linéaire bayésienne
- Lois conjuguées
- Qu'est ce que le Topic Modeling ?
- Algorithme LDA (latent dirichelet allocation)
- Processus Gaussiens
- Optimisation bayésienne
- Exemples d'applications
- Probabilistic Inference for Learning Control (PILCO)

- **Calauzene (Criteo).**

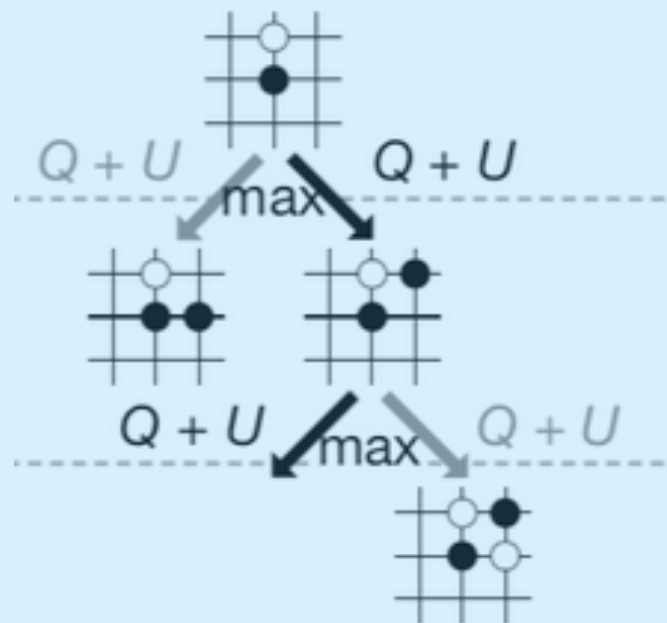
Learning to Rank and Recommender Systems. 12h

- Ce cours va explorer les architectures possibles pour construire un système de recommandation. En particulier,
- Comment découpler la phase de pre-selection (retrieval) de la recommandation elle-même (ranking) pour des raisons de passage à l'échelle ?
- Comment construire, apprendre et évaluer un modèle ranking ?
- Comment intégrer la nature implicite du feedback observé (click, rating...) qui n'est pas aussi riche qu'une supervision ?

Apprentissage par Renforcement

Objectif du cours

- Les techniques d'apprentissage par Renforcement modernes sont divisées en trois catégories:



MCTS

(valable quand l'environnement est connu)

**Apprentissage
direct de politique
(ex: PPO)**

**Apprentissage de
fonction de valeur Q
(Q-learning, DQN...)**

Contenu/Intervenants

- **Stéphane Airiaut. MCF Dauphine. 6h. Introduction.**
 - Processus de Markov, planification, renforcement (q-learning)
- **Olivier Cappé (ENS). 6h. Bandits stochastiques.**
 - UCB, exploration epsilon-greedy.
- **Jérémie Mary. (CRITEO) 12h. Deep RL**
 - Imitation Learning and Cross-entropy method
 - DQN and Classical Extensions
 - Policy gradients
 - Asynchroneous Actor Critic and parallelization of learning
 - Advanced policy gradients (and natural versions)
 - Continuous action space
 - Trust region methods TRPO/PPO/ACKTR
 - Inverse Reinforcement Learning
 - Transfer Learning and multitask
 - Meta Learning (MAML)
 - Distributional RL
 - Transformers in RL
 - Modern analysis of trust regions (Scherrer & all) and Safe RL

Anonymity, privacy

Pierre Senellart



3 December 2019

Need for anonymization

- Many datasets include personally identifying information (names, IDs, IPs, etc.) and sensible data (clinical data, sexual orientation, religion, etc.)
- EU General Data Protection Regulation imposes constraints on what can be done with personal data
- Machine learning models often based on personal data, potentially leaking sensitive information
- Many technical challenges in dealing with private data without disclosing private information

Curriculum

- Classes on **Mondays afternoon** (in conflict with the course on *Incremental learning, game theory, and applications* ☹)
- 8 sessions + project defenses
- Topics covered:
 - Basics of data privacy and anonymization
 - Measuring anonymity: k -anonymity, l -diversity, m -closeness
 - A probabilistic framework: differential privacy
 - Differentially private mechanisms
 - Local differential privacy
 - Differential privacy and federated learning
 - Processing data anonymously, homomorphic encryption

Evaluation

- One homework (40% of the total grade): take one research paper, summarize it, explain it in your own words, comment on its strengths and limitations
- One project (60% of the total grade): take one (or more) research paper, build something cool from it (implement it, improve the algorithm, test it on some interesting dataset, etc.), present it in a defense

Computational social choice (Master IASD, 2019-20)

- ▶ Social choice: *designing and analysing methods for collective decision making*
- ▶ Some examples of social choice problems:
 - ▶ political elections. **Voting**
 - ▶ finding a date for a meeting. **Voting**
 - ▶ deciding where and when to have dinner altogether tonight. **Voting**
 - ▶ in a high school: deciding who gets which class and who teaches when. **Fair division**
 - ▶ in a company: finding a partition of employees in groups of people who will work together. **Coalition structure formation**
 - ▶ assigning students to universities. **Matching**

Social choice rules

- ▶ **input**: agents have preferences over possible alternatives
- ▶ **output**: an alternative

What are alternatives?

- ▶ in **voting**: candidates, or sets of candidates
- ▶ in **fair division** : assignments from resources to agents
- ▶ in **matching**: assignment of agents of class 1 (students) to agents of class 2 (universities)
- ▶ in **coalition formation**: partitions of agents into groups

Social rules must be

- ▶ **designed**
- ▶ **studied** axiomatically (which properties do they satisfy?)
- ▶ **computed** (communication protocols + algorithms)

Plan of the course

1. Introduction to social choice
2. Computing hard voting rules.
3. Strategic issues and game-theoretic analyses of voting
4. Voting on combinatorial domains
5. Communication issues in voting
6. Fair division
7. Matching
8. Coalition formation
9. Specific applications: rent division, kidney exchange, school assignment etc.

No prerequisites (the course is self-contained).

Contact: Jérôme Lang (lang@lamsade.dauphine.fr)

Data wrangling, data quality

Leonid Libkin *Pierre Senellart*



3 December 2019

Ideal vs actual world

Ideal world for a data scientist:

- A single dataset, with a fixed, simple structure (e.g., one table with features and label)
- Structured data
- Exact, complete information
- Precise values, certain data

Ideal vs actual world

Ideal world for a data scientist:

- A single dataset, with a fixed, simple structure (e.g., one table with features and label)
- Structured data
- Exact, complete information
- Precise values, certain data

Actual world:

- Many datasets to be combined, with different structures and schemas
- Plain text, semi-structured data
- Duplicated information, missing information
- Imprecise values, uncertain data

Data wrangling, data quality

- How to wrangle real-world data and turn it into a nice structured form?
- How to assess the quality of data?
- How to deal with missing, imprecise, duplicated data?
- How to keep track of where data comes from?
- How to do all of this **efficiently**?

Curriculum and provisional schedule

- Classes on **Wednesdays afternoon** (in conflict with the course on *Ethics and AI* ☹)
- 8 sessions + project defenses
 - 22/01 Information extraction ([Pierre](#))
 - 29/01 Data cleaning, data deduplication ([Pierre](#))
 - 05/02 Data integration, view-based query answering ([Leonid](#))
 - 12/02 Provenance management ([Pierre](#))
 - 26/02 Data exchange ([Leonid](#))
 - 04/03 Probabilistic databases ([Pierre](#))
 - 11/03 Incomplete information in databases ([Leonid](#))
 - 18/03 Approximate query answering ([Leonid](#))
 - 25/03 Project defenses

Evaluation

- One homework (40% of the total grade): take one research paper, summarize it, explain it in your own words, comment on its strengths and limitations
- One project (60% of the total grade): take one (or more) research paper, build something cool from it (implement it, improve the algorithm, test it on some interesting dataset, etc.), present it in a defense

Deep Learning for Image Analysis

Course Introduction

E. Decencière, Thomas Walter, Santiago Velasco-Forero

MINES ParisTech
PSL Research University



About the lecturers



Thomas Walter

<http://members.cbio.mines-paristech.fr/~twalter>

- Research: bioimage informatics
- Main application fields: High Content Screening (HCS), as a method to systematically study biological processes by analyzing cellular phenotypes



Santiago Velasco-Forero

<http://cmm.mines-paristech.fr/~velasco>

- Research: image processing, pattern recognition, multivariate statistics, graph-based data/image analysis
- Main application fields: Remote Sensing, cosmetology, astronomy, hyperspectral imaging.



Etienne Decencière

<http://cmm.mines-paristech.fr/~decenciere>

- Research: mathematical morphology and image analysis
- Main application fields: Ophthalmology, dermatology, cosmetology, astronomy

Objectif et pré-requis

Objectif

Introduction à la théorie et à la pratique de l'apprentissage profond pour l'analyse d'images.

Connaissances préalables

- Algèbre et probabilités de base
- Programmation: python

Langue

- Slides: anglais
- Oral: anglais ou français (en fonction de l'auditoire)

Contenu

Approche pédagogique

- 4 journées entières
- Cours magistraux le matin
- Travaux pratiques l'après-midi (python, keras)

Sujets abordés

- Rappels: machine learning et réseaux de neurones
- De la classification à la transformation d'images
- Optimisation et visualisation des réseaux profonds
- Autoencodeurs et réseaux antagonistes génératifs (GANs)

Evaluation

Examen

Foundations of graph and RDF query languages

Leonid Libkin

Graph Data: background

- A new rapidly growing and popular data model
- Suitable for many applications: social networks, crime detection, route planning, investigative journalism, bioinformatics etc
- Many products, systems; SQL extension to cover graphs to become standard in 2021; a graph query language (GQL) standard in
- RDF: graph-based data model for data interchange on the Web.

Outline

- Review: other data models (relational, tree-shaped), graphs as relations, bottlenecks of query evaluation
- Simple graph database models and query languages (RPQs, CRPQs and relatives); query evaluation and optimization
- Property graph model; theoretical abstraction and query languages (extensions of RPQs, Graph XPath etc)
- Query language Cypher: linear structure of queries, morphism-based pattern matching, complexity, graph modifications
- RDF model, RDF vs property graphs, basics of SPARQL, connection of SPARQL with relational querying
- Tools to be used in the course: Neo4j community edition/Cypher queries; Data.world free tool for RDF querying

Graph analytics

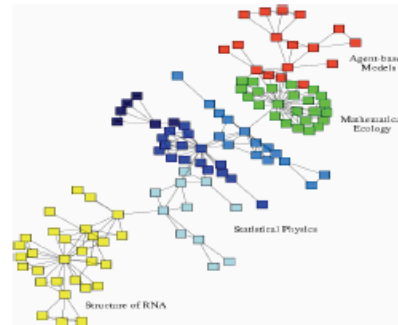
Daniela Grigori

Paris-Dauphine University, PSL

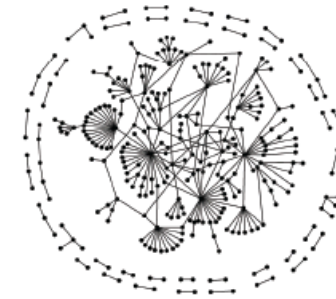
Many types of data are graphs



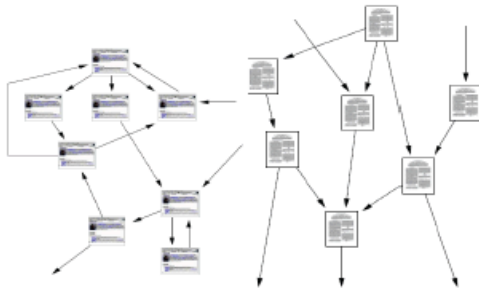
Social networks



Economic networks



Communication networks



Information networks:
Web & citations



Internet



Networks of neurons

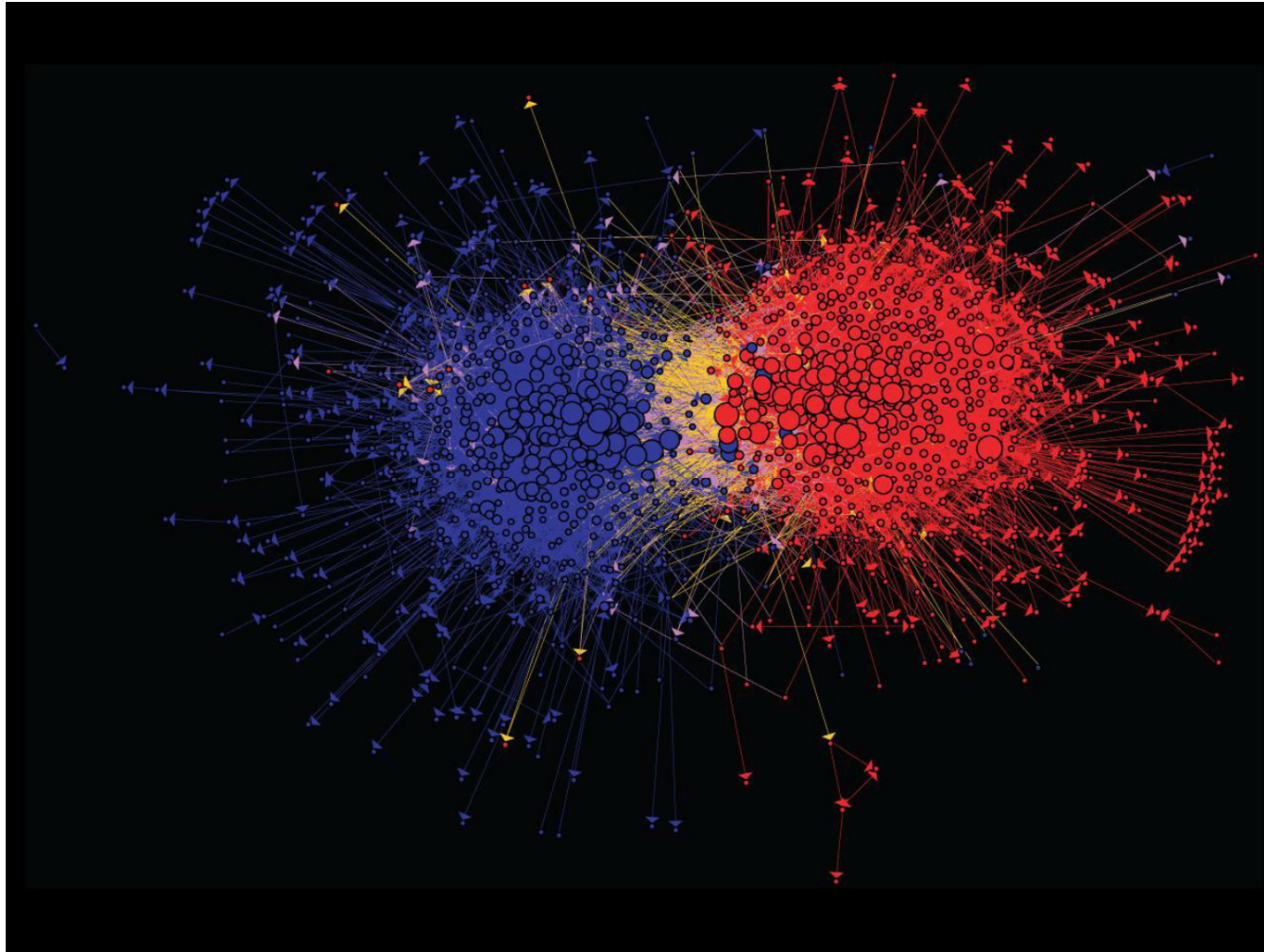
Graph Data: Social Networks



Facebook social graph

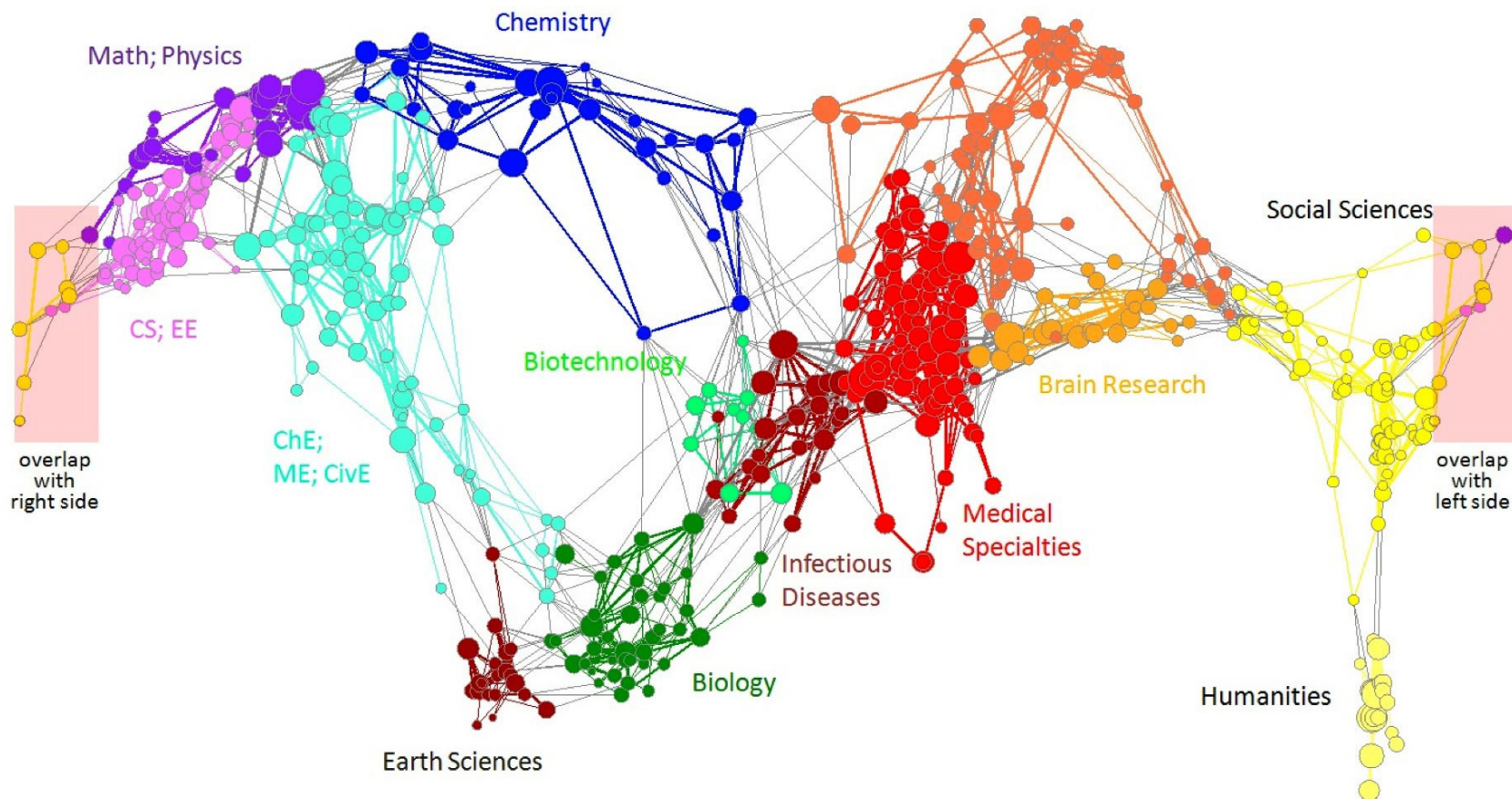
4-degrees of separation [Backstrom-Boldi-Rosa-Ugander-Vigna, 2011]

Graph Data: Media Networks



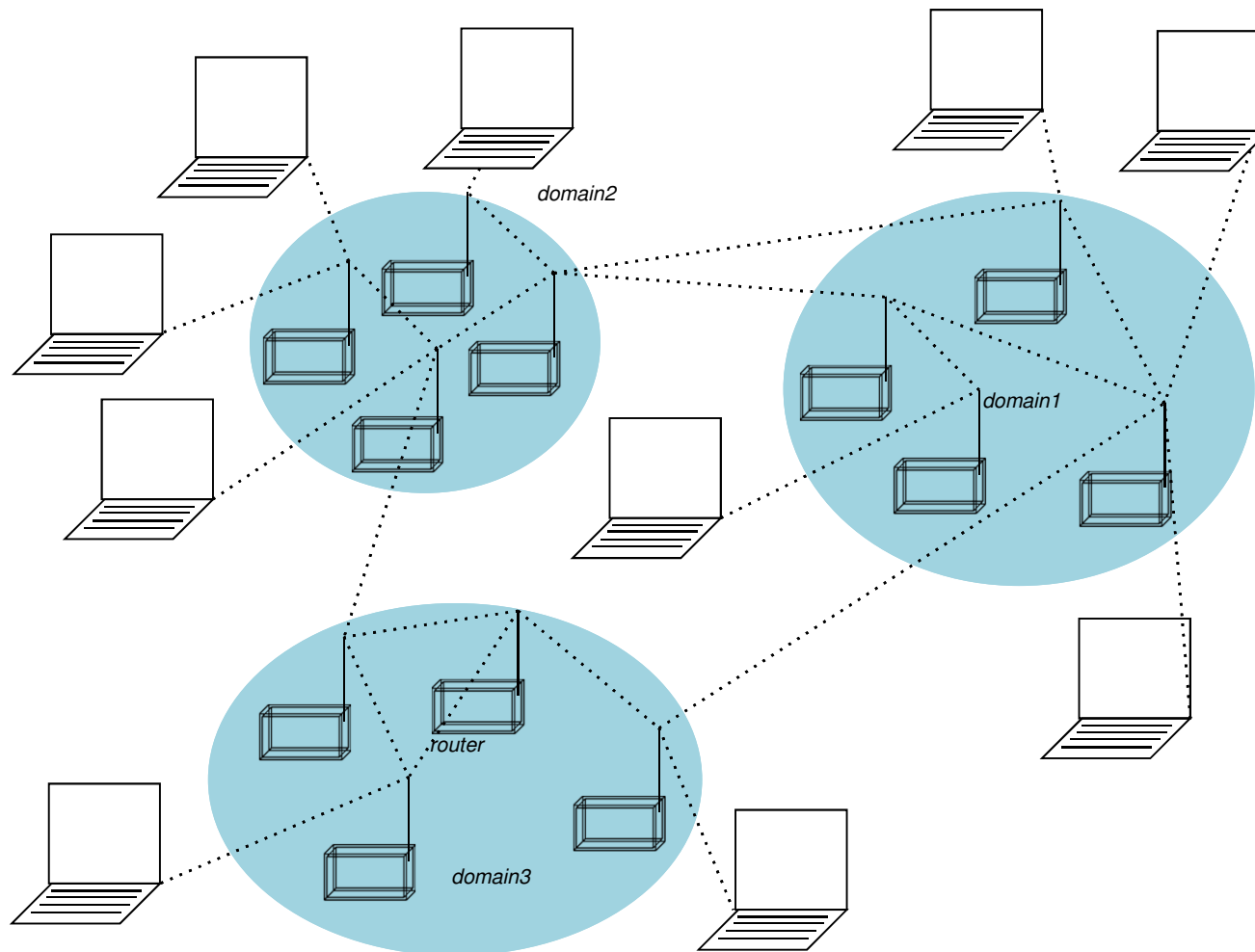
Connections between political blogs
Polarization of the network [Adamic-Glance, 2005]

Graph Data: Information Nets



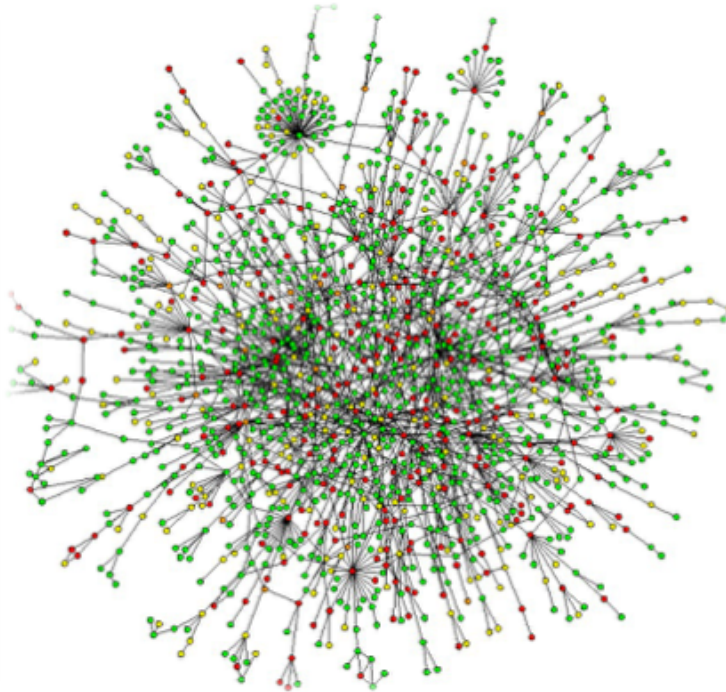
Citation networks and Maps of science
[Börner et al., 2012]

Graph Data: Communication Nets

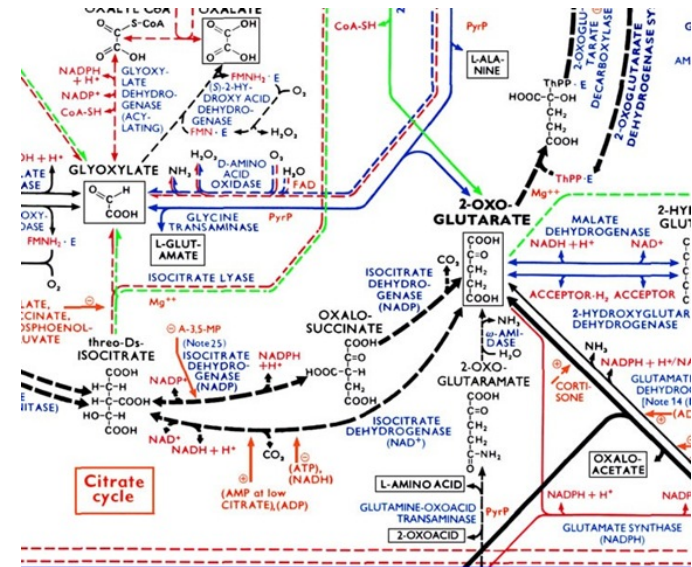


Internet

(5) Networks: Biomedicine



Protein-protein interaction (PPI) networks:
Nodes: Proteins
Edges: 'Physical' interactions



Metabolic networks:
Nodes: Metabolites and enzymes
Edges: Chemical reactions

Graphs : Machine learning

- Complex domains (knowledge, text, images, etc) have rich relational structure, which can be represented as a relational graph
- By explicitly modeling relationships we achieve better performance

Different kinds of graph analysis

- **Path analysis:** for route optimization that is particularly applicable to logistics, supply and distribution chains and traffic optimization for smart cities.
- **Connectivity analysis:** for determining weaknesses in networks such as a utility power grid, comparing connectivity across networks.
- **Community analysis:** Distance and density-based analysis is used to find groups of interacting people in a social network, for example, and identifying whether they are transient and predicting if the network will grow.
- **Centrality analysis:** find the most influential people in a social network, for example, or to find most highly accessed web pages—such as by using the PageRank algorithm.

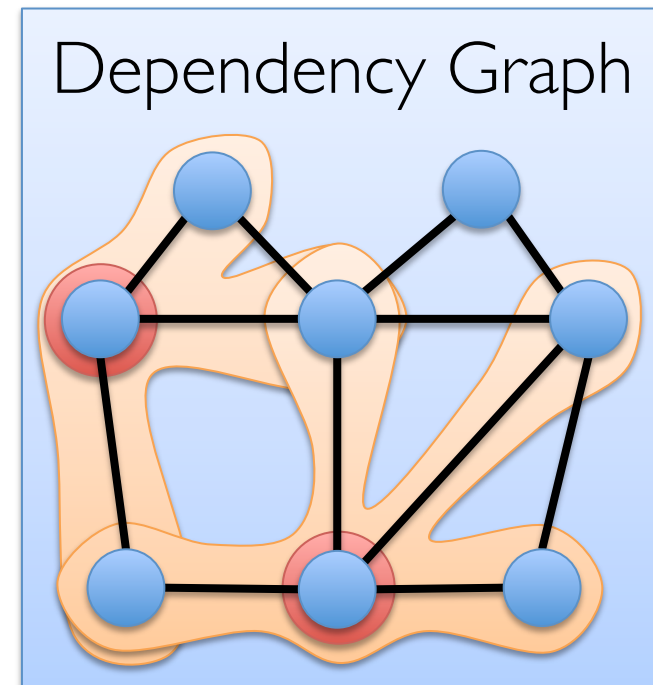
Graph-Parallel Systems

Pregel
oogle



GraphLab

*Exploit graph structure to
achieve **orders-of-
magnitude performance
gains** over more general
data-parallel systems.*



Agenda

- Graph analytics
 - Link Analysis
 - Community detection
 - Graph similarity, Graph clustering, ..
- Machine learning with graphs : Graph Neural Networks
- Frameworks for parallel graph analytics
 - Pregel – a model for parallel-graph computing
 - GraphX Spark – unifying graph- and data –parallel computing
- Practical work : graph analytics with GraphX

Incremental Learning, Game Theory and Applications

Rida Laraki

PSL University – Paris Dauphine



Contenu du cours

- **Introduction aux jeux** (1/3 du cours)

Jeux à somme nulle, valeur, stratégies optimales, théorème du min-max, équilibre de nash, équilibre corrélé, jeux de potentiels, jeux monotones

- **Algorithmes d'optimisation en ligne** (1/3 du cours)

regret learning, Blackwell approachability, online gradient descent

- **Apprentissage dans les jeux** (1/3 du cours)

- que se passe-t-il si tous les joueurs utilisent un algorithme online?
- quels algorithmes convergent vers un équilibre et quel équilibre?

Knowledge Graphs, Description Logics and Reasoning on Data

C. Bourgaux, M. Thomazo

Knowledge Graphs and Ontologies

Knowledge graphs : flexible tool to represent knowledge



Albert Einstein



Theoretical physicist

Albert Einstein was a German-born theoretical physicist who developed the theory of relativity, one of the two pillars of modern physics. His work is also known for its influence on the philosophy of science.

[Wikipedia](#)

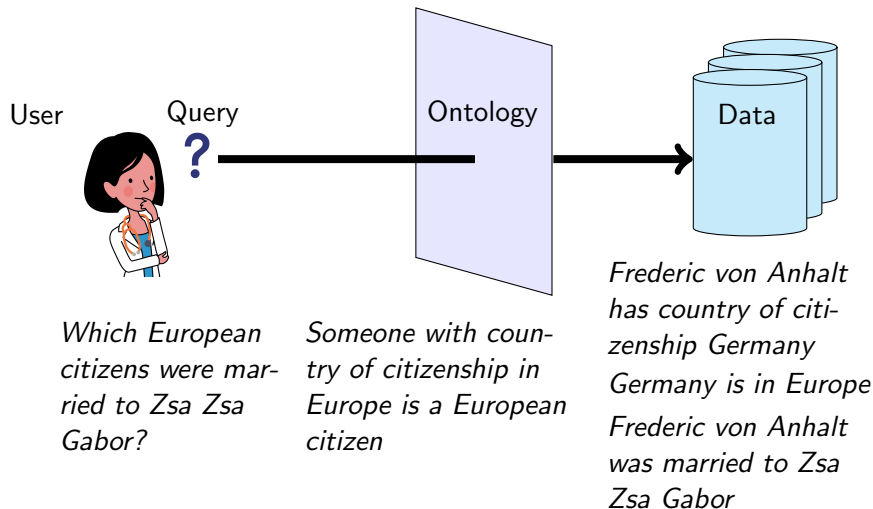
Born: March 14, 1879, [Ulm, Germany](#)

Died: April 18, 1955, [Princeton Medical Center, New Jersey, United States](#)

Education: [University of Zurich](#) (1905), [ETH Zürich](#) (1896–1900), [MORE](#)

Spouse: [Elsa Einstein](#) (m. 1919–1936), [Mileva Marić](#) (m. 1903–1919)

Knowledge Graphs and Ontologies



Why does it matter?

The screenshot shows the Data.GOV website with the search term 'earthquakes'. The results show 1,160 datasets found. The top result is '1988 Spitak Earthquake Database' from the National Oceanic and Atmospheric Administration. Below it is 'Utah FORGE Site Earthquake Animation' from the Department of Energy. A map of the United States is visible on the left side of the results.

The screenshot shows the Data.GOV website with the search term 'natural disaster'. The results show 149 datasets found. The top result is 'NASA Disasters Mapping Portal' from the National Aeronautics and Space Administration. Below it is 'Disaster Declarations USDA' from the Department of Agriculture. A map of the United States is visible on the left side of the results.

Outline of the Course

- ▶ how to represent domain knowledge for automated use?
- ▶ what are the associated reasoning tasks?
- ▶ how to use an ontology when querying data?

Both theory and practice will be covered.

Monte Carlo Search and Games

Tristan Cazenave

MCTS :

- Go, Chess, Social Choice, Chemistry, UCT, GRAVE, PPAF, PUCT, AlphaZero...

Nested Monte Carlo Search :

- Sudoku, Expression Discovery, Cooperative Pathfinding, Software testing, heuristic Model-Checking, the Pancake problem, Games, Cryptography, the RNA inverse folding problem...

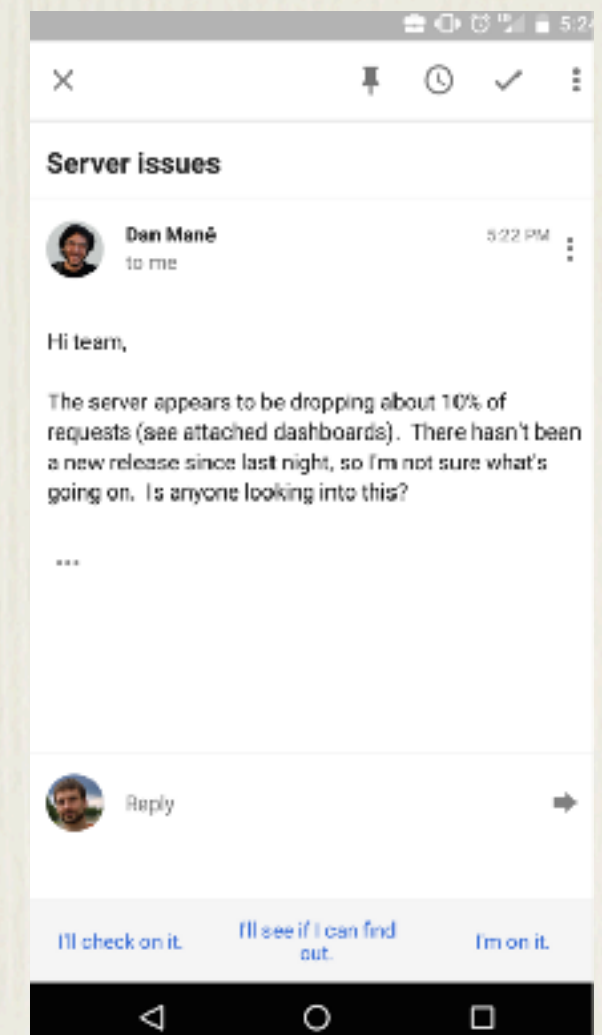
Nested Rollout Policy Adaptation :

- Morpion Solitaire, CrossWords, SameGame, TSPTW, MSA, Graph Coloring...

Natural Language Processing

Many successful applications:

- Analyse text/speech
- Extract the relevant information
- Decide or generate an answer



Challenges

- Ambiguity and diversity (usage and context)

GREEEAT job @justinbieber! ^^

U taught us 2 #neversaynever ! CUI8r

Iraqi Head Seeks Arms

A horse! a horse! my kingdom for a horse!

- World-Wide NLP

瓶矿泉水

бутылка минеральной воды

Mineralwasser flasche

Overview

Outline:

- The main tasks, issues and peculiarities
- Computational Semantics
- Morphology Syntax and Parsing
- Deep Learning for NLP:
 - Word embeddings, text representation
 - Machine Translation and Transformers
 - BERT, Elmo and all that jazz
- Large scale NLP & business perspective

Evaluation:

- 50% CC
- 50% project

Lab sessions:

python+pytorch

Team:

- Alex. Allauzen
(Dauphine Univ.)
- Benoit Crabbé
(Univ. Paris Diderot)
- Guillaume Pitel
(Exsenza)

Systems, paradigms and languages for Big Data

- Follow up of ADB course
- Focus on
 - large scale data processing via Spark and HIVE
 - Principles and techniques behind RDD, Dataframes, Datasets (in Scala)
 - Hive architecture, SQL-to-MapReduce compilation
- Lab sessions on:
 - text and graph processing, matrix manipulation in Spark
 - web-log analytics in Hive
- Evaluation : mini-project + written exam

ML over Big Data

- Some basic ML approaches, some of which are based on gradient descendant
 - Linear and logistic regression
 - Matrix completion for recommender systems
 - Clustering: k-means
 - Decision trees
- Spark implementation and focus on optimizations:
 - e.g., persistence and partitioning control, shuffle minimization
 - Optimizations via Dataframes and Datasets.
- Time permitting: introduction to Spark stream processing and application to decision trees
- Evaluation : mini-project + exam.