# APSEM2018
# Apprentissage et SEMantique



http://devlog.cnrs.fr/apsem2018
pascal.dayre@enseeiht.fr

# Remerciements


- CNRS
  - MITI, la Mission pour les Initiatives Transverses et Interdisciplinaires
  - DEVLOG
  - DR20
- INRA
- ENSEEIHT, pour la mise à disposition des locaux
- IRIT
- Comité de programme
- Les intervenants

# Constat

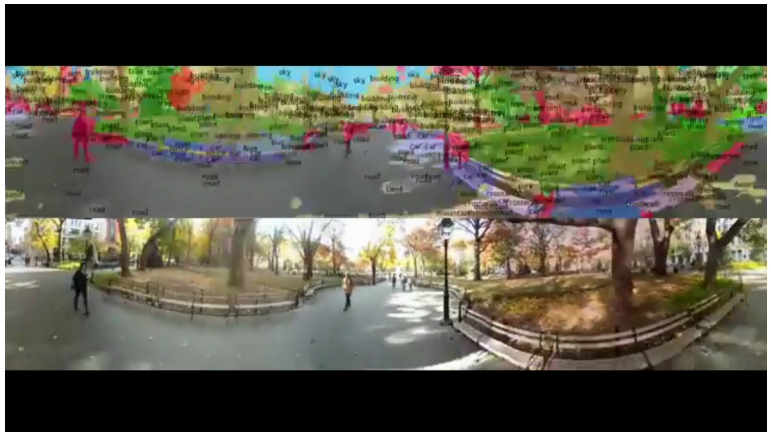
- Révolution des données ou révolution de la connaissance ?
- Recherche  par les données : en science expérimentale, le paradigme de construction de la connaissance est inversé avec l'afflux des données, les nouveaux moyens de calcul et les nouvelles techniques d'apprentissage.
- L'apprentissage automatique et la statistique sont au cœur de la production des connaissances.
- L'étape initiale de modélisation du phénomène est maintenant remplacée par une exploration des données qui aboutit à une émergence du modèle. Nous sommes à l'**ère de la science par les données**.
- Etude de systèmes complexes.

=> nécessité de faire se rencontrer la communauté des développeurs (DEVLOG), des statisticiens, de l'apprentissage, du web sémantique, des producteurs de données et des expérimentateurs (observatoires, INRA, ...)

# Objectifs

- Etude de l'apport croisé et des nouvelles perspectives des technologies du web des données et de la recherche par les données.
  - Apprentissage pour apprendre des représentations
  - Représentations pour supporter l'apprentissage.

Apprentissage ⇄ Représentation de connaissance



Clement Farabet, Camille Couprie, Laurent Najman and Yann LeCun: Learning Hierarchical Features for Scene Labeling, IEEE Transactions on Pattern Analysis and Machine Intelligence, August 2013

https://www.youtube.com/watch?v=aqt2PPZqldk

http://devlog.cnrs.fr/apsem2018
pascal.dayre@enseeiht.fr

# Agenda

- Jour 1 : Apprentissage  et science des données
- Jour 2 : Ingénierie des connaissances et web des données
- Jour 3 : Convergence apprentissage et sémantique
- Jour 4 : Retours d’expérience et ateliers

# AGENDA.J1 - Apprentissage

- **09h20-10h00** : - Accueil
- **10h00-10h15** : - Présentation et objectif de l'action - Quels apports croisés de l'apprentissage et du web sémantique? - **Pascal Dayre / CNRS/IRIT**
- **10h15-10h45** : - Recherche par les données : des données aux représentations des connaissances exploration, préparation des données d'apprentissage pour éviter les biais (données manquantes, classes surepresentées, bonne distribution, données erronnées), mise en forme et structuration des données - **Sébastien Déjean** / IMT
- **10h45-11h15** : - Pause
- **11h15-12h15** : Synthèse de la science des données et de l'apprentissage automatique. Les points de vue maths/info de l'ingénieur - **Laurent Risser / IMT**
- Exemple introductif qui pose le vocabulaire (observation/variable/label/apprentissage supervisé ou non).
- Evolution des tendances en science des données (de la statistique classique à l'apprentissage machine).
- Présentation illustrée d'algorithmes classiques (arbre de classification, random forest, K-means, SVM).
- Présentation illustrée de méthodes basées sur le calcul GPU (Deep-learning, XGBoost).
- Méthodes standard d'évaluation de l'efficacité d'un algorithme d'apprentissage (LOO, K-fold).
- Problématique actuelle de la réduction de dimension.
- Problématique montante d'explicabilité des choix d'un algorithme l'apprentissage. -
- **12h15-13h30** : - Pause repas
- **13h30-14h30** : - Synthèse de la science des données et de l'apprentissage automatique. Les points de vue maths/info de l'ingénieur - **Laurent Risser / IMT** (suite)
- **14h30-15h15** : - Un point sur l'explicabilité et l'interprétabilité en machine learning - **Mathieu Serrurier / IRIT**
- **15h15-15h45** : - Pause café
- **15h45-16h30** : - Comment faire émerger un graphe du décodage de vos données. Mise en oeuvre pour l'analyse de la structure du discours dans les tchats. Méthode, approches classiques et extraction automatique de représentation avec le deep learning. - **Stergos Afantenos / IRIT**
- **16h30-17h15** : - Apprentissage et représentation jointe dans une base connaissance pour la désambiguation d'entités. Application à une collection de texte. - **Jose Moreno / IRIT**
- **20h00-22h00** : - Evénement social dinatoîre

# Agenda.J2 - Ingénierie des connaissances et web des données

- **09h00-10h00** : Introduction à l'Ingénierie des Connaissances, ses usages, ses intérêts : web des données, données liées, ontologies, aperçu des standards du web sémantique (RDF/RDFS/OWL/SPARQL). **Franck Michel / CNRS**
- **10h00-10h45** : Réutiliser/créer des vocabulaires contrôlés, des ontologies de domaine: LOV, BioPortal… **Nathalie Hernandez/IRIT, Alban Gaignard/ Université de Nantes**
- **10h45-11h15** : Pause café
- **11h15-11h40** : Comment annoter sémantiquement des données existantes (Web Annotation, CSV on the Web, JSON-LD…). **Nathalie Hernandez/IRIT**
- **11h40-12h00** : Décrire et Publier des jeux de données sur le web: vocabulaires, catalogues et portails. **Franck Michel/CNRS**
- **12h00-12h30** : Vocabulaire liés aux statistiques, formaliser les activités d'analyse. **Franck Cotton/INSEE**
- **12h30-14h00** : Déjeuner
- **14h00-14h30** : La mise en oeuvre du machine learning à partir d'un problème, de son modèle et du jeu de données. Quel choix de workflow pour quel explicabilité des paramètres de l'apprentissage - **Gabriel Ferrettini / IRIT**
- **14h30-15h15** : Exploration et visualisation des données (définition du jeu de données/mise en oeuvre du web sémantique) - **Franck Cotton / INSEE** (ESAN - statistiques des entreprises européennes - cas avec des entreprises de l'agroalimentaire )
- **15h15-15h45** : Pause café
- **15h45-16h30** : Description sémantique d'un service de traitement/analyse/apprentissage et comment composer les services ? (SOA sémantique). Traçabilité/provenance des données avec PROV-O, actions schema.org. **Alban Gaignard/Univ Nantes**
- **16h30-17h00** : Table ronde “Quel apport du web des données pour l'usage des données dans un processus d'apprentissage?”

  Quel apport du web des données pour la préparation, la structuration et l'usage des données dans un processus d'apprentissage? Quels intérêts pour les infrastructures de recherche et les ENTC? ou comment les e-infrastructures se saisissent de la problématique de l'ouverture des données, de l'apprentissage et de l'IA.

# Agenda.J3 - Convergence apprentissage et sémantique

- **09h00-10h30** : Construction de graphes valués à partir des données
- Méthodes pour la construction de graphes valués: aperçu des méthodes et illustration par l'approche PLS-PM sur des données reliant agriculture et environnement - **Dominique Desbois (INRA/Versailles)**
- Construction de graphes à partir des variables décrivant l'environnement et la biodiversité - Romain David / IMBE / IndexMEED
- **10h30-11h00** : Pause
- **11h00-12h30** : Gestion et intégration de connaissance -Sémantique des Données génomiques des plantes et phénotypage-. Utilisation de graphes pour l'apprentissage “classique” - **Pascal Neveu / UMR INSTA / INRA Montpellier**
- **12h30-14h00** : Déjeuner
- **14h00-15h30** : Table ronde sur la convergence Apprentissage/Représentation des connaissances- Pascal Neveu +
- **14h00-15h00** : Les panélistes : donnez votre point du vue sur la convergence.
  - Le développement d'un algorithme d'apprentissage pour définir une ontologie et une cartographie sémantique - **Frédéric Assié / MSHSUD**
  - Fusion de données d'imagerie médicale et réduction de dimensionnalité par apprentissage à noyaux multiples - **Nicolas Duchateau / CREATIS**
  - La composition de services Web sémantiques et l'interopérabilité - **Thierry Louge / OMP/IRAP**
  - L’apprentissage symbolique **– Bernard Espinasse / LIS-Lab**
- **15h00-15h30** : Les questions
- **15h30-16h00** : Pause café
- **16h00-16h30** : * Quel apport de l'Approche bio-div pour l'apprentissage : mélanger des objets de différentes natures dans le même graphe. - **Romain David / IMBE** et GDR Madics (curation et fouille en fonction des différents contextes) Génération de graphes de décision. ou autre?
- **16h30-18h** - Atelier Graminé / GDR Madics – **Stéphane Pérennes / CNRS**

  Atelier de programmation GRAMINEES (GRAphe data Mining In Natural, Ecological and Environnemental Sciences, Responsables Romain David, IMBE, INEE, Nathan Cohen, I3S, INS2i)

# Agenda.J4 – REX et ateliers

- **09h00-10h00** - Apprentissage automatique / web sémantique/ retours d'expérience. Présentation d'une méthodologie d'alignement automatique avec Geonames et de désambiguisation d'entités géographiques en utilisant une méthode par apprentissage automatique (words embeddings avec AdaGram en Julia). Dans le réservoir ISTEX des entité nommées on été extraites. Nous nous focalisons sur les entités géographiques (de type place name) que nous cherchons à aligner automatiquement avec Geonames. La désambiguisation des entités est alors une étape importante qui peut être résolue grace à des méthodes d'apprentissage automatique et de vectorisation de mots.Nous nous basons sur l'algorithme AdaGram développé en Julia. Nous présenterons la problématique, la méthodologie et illustrerons avec quelques exemples. - **Pascal Cuxac / INIST / CNRS**
- **10h00-11h00** - Le langage Julia - **Dennis Wilson / IRIT**
- **11h00-11h15** - Pause
- **11h15-12h15** - DATANOOS: from DATA to a NOOSPHERE, une alliance académique transdisciplinaire sur les ressources numériques et les pratiques de connaissance - **Pascal Dayre / IRIT / CNRS**
- **12h15-13h30** - Repas
- **13h30-15h00** - Ateliers:
- A1 - Extraction de la sémantique et indexation de documents textes selon un modèle métier. Apport croisé de l'apprentissage et de la sémantique - **Pascal Cuxac / INIST / CNRS**
- A2 - Méthodes et mise en oeuvre du machine learning par les non spécialistes. Choix du modèle, des données et du workflow. Evaluation du résultat - **Gabriel Ferrettini / IRIT**
- A3 - Mise en oeuvre d'une infrastructure de données et de ses usages (Problématiques des infrastructures ouvertes / FAIR / Interopérabilité horizontale / Apprentissage pour apprendre des représentations / Représentations pour supporter l'apprentissage / apprentissage et IA) - **Pascal Dayre / IRIT / CNRS**
- **15h00-15h30** - Restitution
- **15h30-16h00** - Bilan des journées