ENS de Lyon

JOURNÉE D'ÉTUDE
24 MAI 2019

SITE DESCARTES
SALLE D2 117

BIG DATA ET MACHINE LEARNING : **ENJEUX POLITIQUES, ENJEUX SOCIOLOGIQUES**

ENS DE LYON

triangle UMR 5206

Centre Max Weber UMR 5283

IDEX LYON
UNIVERSITÉ DE LYON

PROGRAMME

VENDREDI 24 MAI 2019

Accueil, petit-déjeuner **8H30-9H** SALLE D2 117

MATINÉE :

L'humain derrière la machine : questionnements politiques et épistémologiques à l'ère de la "révolution" du big data"

Emmanuel Didier (CMH, Paris), introduction de la journée : De la sociologie des statistiques à la sociologie des big data. **9H-9H30**

Cécile Favre (ERIC, Lyon), Eclairage au prisme du genre sur les big data et l'apprentissage machine. **9H30-10H**

Jean-Sébastien Vayre (LITEM, Évry), Prédire ou produire des futurs économiques : des machines apprenantes pour travailler l'organisation des marchés. **10H-10H30**

Bilel Benbouzid (LISIS, Paris), Enquête sur les machines prédictives. Le cas de la police aux États-Unis. **10H30-11H**

Table ronde, animée par Antoine Larribeau (ENS de Lyon), **Promesses et ambitions des algorithmes : révolution ou pratique sociale ordinaire ?** **11H15-12H30**

12h30-14h : PAUSE DEJEUNER

APRÈS-MIDI :

La science derrière les machineries : quels défis théoriques et méthodologiques pour les sciences sociales ?

Pablo Jensen (ENS de Lyon), La société se laisse-t-elle mettre en équation ? **14H-14H30**

Gabriel Alcaras (CMH, Paris), Les messages électroniques comme traces du travail : promesses empiriques et enjeux méthodologiques de la collecte et de l'analyse des emails. **14H30-15H**

Marta Severo (DICEN, Nanterre), L'analyse de l'opinion politique sur Twitter : entre théorie et méthode. **15H-15H30**

Samuel Coavoux (SENSE, Paris), Vieux débats et nouvelles opportunités. Apports et limites des données d'usage à la sociologie des consommations culturelles. **15H30-16H**

Table ronde, animée par Corentin Roquebert (CMW, Lyon) : **Big data et machine learning sont-ils solubles dans la sociologie ?** **16H15-17H30**

Journée d'étude
« Big data et apprentissage machine : enjeux sociologiques, enjeux politiques »

Programme détaillé

ENS de Lyon, vendredi 24 mai, D4-117, 9h-17h30

Journée organisée par Antoine Larribeau et Corentin Roquebert

Présentation générale

La journée d'étude s'organise autour du thème « Big data et apprentissage machine : enjeux sociologiques, enjeux politiques », et privilégie deux axes thématiques de réflexion au centre des questions que posent ces nouvelles méthodes. La matinée - consacrée à la mise au jour de "l'homme derrière la machine" - aura pour enjeu de mieux comprendre ce qui se joue dans les différents usages de ces techniques, sur un plan empirique, éthique et politique. L'après-midi - intitulé "la science derrière les machineries ?" - posera le débat sur un plan plus épistémologique, en interrogeant les conditions sous lesquelles la sociologie peut s'approprier ces techniques pour produire de la connaissance.

Big data et machine learning sont en effet l'objet de discours nombreux qu'il s'agit de clarifier : entre paniques morales et annonces prophétiques sur la « fin de la théorie » ou avènement d'une nouvelle « rupture épistémologique », les chercheurs sont invités ici à prendre du recul pour prendre au sérieux les enjeux de pouvoir et de production du savoir qu'induisent ces nouvelles données et méthodes statistiques. De quoi parle-t-on lorsque l'on évoque le Big Data ? Quel peut être le rôle du sociologue à l'ère de ces nouvelles données ? Leur accès, récolte et traitement sont-ils le signe d'une plus grande « disparité de pouvoir entre entreprises privés et recherche publique » (Bastin & Tubaro, 2018) ?

Cette journée voudrait défendre une approche de l'étude des Big Data en train de se faire : tout autant que les données classiques en sociologie (par exemple, issus d'un questionnaire), elles sont loin d'être des données déjà données, c'est à dire produite en dehors de tout contexte social. Pour mieux saisir les transformations des manières de connaître et d'agir sur le monde social induites par ces techniques, il importe alors de s'interroger à la fois sur les modes de production de celles-ci, mais aussi sur leurs utilisations par des acteurs sociaux, que ce soit dans le champ scientifique ou dans d'autres sphères de la vie sociale (entreprise, institutions, etc.).

Format des présentations : 20 minutes de temps de parole + 10 minutes de question.

Programme

8h30-9h : accueil

Mots des organisateurs.

9h-12h30 : « L'humain derrière la machine : questionnements politiques et épistémologiques à l'ère de la "révolution" du Big Data »

Comment les Big data transforment-elles les sociétés ? Le Big Data se développe et s'étend dans de nombreux domaines de la vie économique et sociale, comme dans les entreprises ou dans les administrations publiques, mais également dans la sphère privée. Comment étudier et interpréter ce phénomène ? Faut-il insister sur les ruptures dans les manières de connaître ou de gérer des populations induites par ces techniques, ou n'y voir qu'une continuité par rapport aux modes traditionnels de gestion et de connaissance ?

En effet, les promesses et ambitions du Big Data, telles que les corrélations prédictives, sont souvent présentés comme des "révolutions". Entre-t-on dans un monde de "traces numériques", où tout serait plus facilement mesurable ? Que signifie cette multiplication des sources de mesure ('quantified self', objets connectés), rendant le "réel" plus quantifiable, ainsi que des possibilités de calcul et d'analyse ? Faut-il voir cette évolution comme une "chance" pour parvenir à une meilleure objectivation du monde ? Au contraire ne faut-il pas considérer ces nouvelles sources de données avec la même prudence que celle attribuer aux artefacts statistiques "traditionnels" ? Dès lors, il faut se méfier d'une vision discontinuiste, portée par les emphases d'une "révolution Big Data". Prendre du recul sur ces transformations et faire un objet d'étude des nouvelles techniques de calcul permet à la fois d'ouvrir la "boite noire" des algorithmes, en montrant les actes sociaux qui les produisent, mais aussi de questionner le "projet politique" (Cardon, 2015) que sous-tendent les discours de promotion de la prétendue "neutralité" axiologique de ces outils.

Par exemple, l'échec du projet d'algorithme de recrutement chez Amazon est éclairant. Entraîné sur des CV d'hommes, il reproduit les discriminations déjà existantes dans la société. Ainsi, initier un projet d'extension des opérations de dénaturalisation du monde social à la déconstruction du Big Data, n'est-ce pas l'occasion pour les sociologues de dévoiler l'existence et les ressorts d'une rationalité, voir d'un "sens pratique" algorithmique ? En effet, comme dans toute pratique de quantification (Desrosières, 1993), la technique est toujours produite par un entrelacement de croyances, de manière de penser et d'agir sur le monde social : il importe donc d'étudier les pratiques humaines derrière une apparente automatisation et d'interroger les schèmes de perception qui participe de la constitution d'une révolution quantitative comme représentation.

***

9h-9h30 **Emmanuel Didier**, introduction de la journée, **De la sociologie des statistiques à la sociologie des big data.**

Les big data ouvrent aujourd'hui de gigantesques perspectives et en même temps engendrent bien des peurs. Il est important que les sociologues s'en saisissent pour mieux comprendre comment ces nouveaux gisements de données peuvent transformer la société., et inversement comment la société peut s'en assurer la maîtrise. Nous voudrions montrer qu'il existe déjà de nombreuses ressources qui le permettent. Ces ressources proviennent de la sociologie de la statistique, laquelle a une longue histoire derrière elle, et permet de comprendre l'entrelacs de technique et de politique qui caractérise toute entreprise de quantification. Nous voudrions proposer des pistes et

des prolongements permettant de les appliquer aux big data.

9h30-10h **Cécile Favre, Éclairage au prisme du genre sur les Big Data et l'apprentissage machine**

À l'ère où le numérique a imprégné notre société jusque dans de nombreux recoins de notre quotidien, au moment où les questions d'égalité femmes-hommes sont toujours bien vivantes, à l'instant où l'informatique ouvre de nombreuses perspectives d'emplois et ces derniers restant finalement majoritairement occupés par des hommes, cette présentation vise à poser un regard dans une perspective de genre sur les Big Data et l'apprentissage machine. Il s'agit de s'intéresser à la fois à ce qu'il se joue au niveau des usages, mais également du côté des personnes participant au développement de ces champs. En s'appuyant sur des exemples concrets, il s'agira de notamment mettre en lumière les enjeux, les points de vigilance, les questions qui se posent et les défis qu'il reste à relever, du point de vue du genre.

10h-10h30 **Jean-Sébastien Vayre, Prédire ou produire des futurs économiques : des machines apprenantes pour travailler l'organisation des marchés.**

Il n'est pas rare d'entendre les spécialistes de l'informatique vanter les mérites des performances prédictives des technologies d'apprentissage artificiel. Cela est particulièrement frappant dans le cas des applications marchandes de ces dispositifs. Dans cette communication, nous proposons de démystifier le caractère prédictif de ces algorithmes. Pour ce faire, nous nous appuierons sur différents éléments de description de leur conception et de leur fonctionnement afin de rendre compte de la façon dont ces machines produisent – plus qu'elles prédisent – les futurs économiques. Nous verrons de cette manière qu'un peu à la façon des prévisions que fabriquent les statistiques publiques, les prédictions que réalisent ces systèmes sont le produit d'un chaînage algorithmique qui a pour principale fonction de consolider un mode d'intervention marchand qui n'est pas indiscutable.

10h30-11h **Bilel Benbouzid, Enquête sur les machines prédictives. Le cas de la police aux Etats-Unis.**

Dans cette présentation, nous observons comment la prédiction policière (predictive policing) en vient à exister dans la physique statistique, l'administration et la jurisprudence. Elle apparaît ainsi comme une technologie morale de gouvernement jugée selon des logiques propres à chacun de ces trois domaines : dans la science, elle rompt avec l'exigence d'exactitude des modèles, privilégiant la précision des scores de risque ; dans l'organisation policière, elle intègre les enjeux de réforme managériale, intégrant des métriques de pondération par une mise en équivalence monétaire des ressources policières; dans le droit, si elle n'est pas proscrite, elle existe sous la forme d'une métrique des nuisances policières acceptables, favorisant le calcul au principe juridique et à la règle de droit. Par ces trois trajectoires d'algorithme, nous suivrons les économies morales en tension de la police prédictive. Ce sera aussi une manière de comprendre le monde auquel et par lequel les humains tiennent lorsqu'ils délèguent du pouvoir aux machines

11h-11h15	Pause
11h15-12h30	Table ronde, animée par Antoine Larribeau **Promesses et ambitions des algorithmes : révolution ou pratique sociale ordinaire ?**

***

12h30-14h : Déjeuner

***

14h-17h30 : La science derrière les machineries ? Quels défis théoriques et méthodologiques pour les sciences sociales ?

La profusion de données disponibles, associée à des nouvelles modélisations mathématiques et à des méthodes d'apprentissage algorithmique permet de renouveler les manières de faire de la statistique. Les possibilités offertes par les algorithmes (supervisés ou non) de machine learning posent donc un ensemble de « défis » pratiques et méthodologiques aux chercheurs. Les sociologues doivent-il amender leurs pratiques de recherche pour se convertir au machinisme ambiant ?

On se demandera donc la mesure dans laquelle les travaux actuels en sciences sociales s'approprient ces objets. Qu'apportent ces recherches et comment peut-on concrètement employer ces nouveaux outils ? En quoi permettent-ils d'améliorer, de compléter, voire de renouveler le regard des sciences sociales sur certains objets ?

Ces enjeux se retrouvent à un double niveau : celui de la construction des bases de données et celui de leurs traitements. Dans les deux cas, les procédures techniques s'éloignent des prérequis standards (quoique discutés) en sociologie quantitative (questionnaires posés à un échantillon représentatif d'une population, démarche hypothético-déductive plutôt qu'exploration à l'aveugle des corrélations, etc.), dont le fondement est une soumission des données à une question théorique, quand Big Data et machine learning semble plutôt imposer l'inverse. Pourtant, plutôt qu'un rapport de concurrence, on espère souligner ici qu'une complémentarité entre ces approches est possible, gouverné par une inventivité méthodologique qui est toujours le fait du chercheur.

Ainsi, en sociologie, il semble important de défendre un usage problématisé, raisonné et réflexif de ces méthodes, que ce soit une technique de collecte comme le webscraping, ou des techniques d'analyse modélisatrice comme les random forest ou le topic modeling : il s'agit alors de les utiliser sans se rendre aveugle aux effets de connaissance qu'elles produisent.

***

14h-14h30	**Pablo Jensen, La société se laisse-t-elle mettre en équation ?** Croissance économique, classements des lycées, publicités sur le Web: de plus en plus, nos actions sont mises en chiffres, en équations, pour aiguiller ou prédire nos comportements. Les big data, ces abondantes traces numériques que nous produisons

constamment, nous permettront-elles de créer une nouvelle science de la société, aussi performante que les sciences de la nature? Peut-on s'inspirer des techniques de modélisation mathématique et de simulation informatique élaborées dans les sciences naturelles pour comprendre enfin la société et l'améliorer?

14h30-15h **Gabriel Alcaras, Les messages électroniques comme traces du travail : promesses empiriques et enjeux méthodologiques de la collecte et de l'analyse des emails.**

La messagerie électronique occupe une place centrale dans un grand nombre d'activités, qu'elles soient professionnelles ou non. Sans surprise, nombre d'historiens, de sociologues et d'anthropologues ont donc cherché à utiliser ces messages comme matériau empirique. Nous considérerons l'étude des emails depuis la collecte jusqu'à leur analyse, en soulignant les possibilités qu'elle ouvre, mais aussi les précautions méthodologiques qu'elle soulève et les obstacles concrets qu'elle pose. Nous concentrerons notre exposé sur l'analyse des listes de discussion, en nous appuyant d'une part sur un état de l'art de l'utilisation des emails comme matériau en sciences sociales, et d'autre part sur notre expérience d'une enquête empirique en cours sur un corpus d'environ 3 millions d'emails rédigés par des ingénieurs informatiques.

15h-15h30 **Marta Severo, L'analyse de l'opinion politique sur Twitter : entre théorie et méthode.**

L'étude de l'opinion, champ traditionnel de la recherche en sciences sociales, a été profondément bousculée par la disponibilité de nouvelles données numériques, notamment des médias sociaux. Cette communication vise à étudier l'impact de l'utilisation des données Twitter sur analyse des opinions politiques. En nous appuyant sur l'état de l'art des études de l'opinion employant ces données, nous visons à mettre en relation les méthodes d'analyse utilisées dans ces études et les définitions de l'opinion politique qui y sont suggérées et à vérifier si une nouvelle approche pour la recherche sur l'opinion en sciences sociales pourrait être fondée sur des techniques multi-échelles.

15h30-16h **Samuel Coavoux, Vieux débats et nouvelles opportunités. Apports et limites des données d'usage à la sociologie des consommations culturelles.**

Les recherches sociologiques sur les consommations culturelles mobilisent habituellement des données issues de questionnaires. Outre leur caractère déclaratif, celles-ci ne permettent de mesurer que grossièrement la diversité des consommations, en s'appuyant sur des catégorisations génériques. En présentant quelques résultats d'une enquête mobilisant des données d'usages d'un service de streaming musical, je montrerai l'intérêt de telles observations pour renouveller les débats de la discipline, et plaiderai pour un usage raisonné de ces données, soucieux de leurs conditions de production et ancré dans le savoir cumulatif produit par les sciences sociales.

16h-16h15 Pause

16h15-17h30 Table ronde, animée par Corentin Roquebert :

Big data et machine learning sont-elles solubles dans la sociologie ?