Institut Interdisciplinaire
Données Massives

Préambule

Les données digitales connaissent une croissance explosive et ouvrent des possibilités nouvelles pour les activités économiques, industrielles et académiques. Le domaine des données massives est souvent caractérisé par des superlatifs liés aux paramètres des données digitales : vitesse, volume, valeur, variété, véridicité (les « V »). Les « big data » (et leur variantes souvent rebaptisées pour marquer des différences d'approche, comme « smart data », « small data « etc.) sont des données dont une des caractéristiques (V) dépasse les moyens de traitement classiques, et impose des nouveaux paradigmes. Par exemple les échantillons de données collectées au Large Hadron Collider sont tellement

massifs (de l'ordre de Exabyte) que l'utilisation d'une grille de calcul mondiale est nécessaire, avec son corollaire dans un futur proche via une approche de type « cloud computing ». Dans d'autres domaines, comme par exemple l'écologie, ce n'est pas la taille, mais plutôt la diversité et l'inhomogénéité des données qui représentent un véritable challenge.

Les sciences dites « big data » sont *interdisciplinaires* par excellence. En effet, non seulement les approches fondamentales en sciences des données se nourrissent et s'améliorent seulement en présence de problèmes nouveaux issus de champs disciplinaires divers, mais les expertises dans des domaines différents sont complémentaires, ce qui ouvre la voie à des projets interdisciplinaires à fort potentiel scientifique. De plus, certains sujets de recherche liées aux grandes masses de données (comme par exemple les techniques de fouilles dans des données massives ou les infrastructures de calcul de haute performance pour des simulations) touchent souvent aux mêmes problématiques que certaines applications industrielles ou économiques. Il existe par conséquent un potentiel interdisciplinaire énorme et des opportunités de cross fertilisation importantes avec le monde économique.

Il est donc nécessaire que le monde académique se saisisse complètement de l'ensemble des activités liées aux données massives : recherche fondamentale et appliquée (de la production des données jusqu'à leur exploitation), valorisation des méthodologies et des nouveaux usages et formation d'experts sensibilisés aux différents champs d'application. Un Institut Interdisciplinaire « Big Data » représente donc une réelle opportunité de fédérer ces différents efforts en un site unique, reconnu mondialement, à l'instar du « Data Science Institute » de l'université de Virginie (USA, https://dsi.virginia.edu/) ou du « Cambridge Big Data Strategic Research Initiative » (UK, http://www.bigdata.cam.ac.uk/). Au niveau national , un institut pluridisciplinaire de cette envergure, avec implantation locale très diverse et une forte composante enseignement serait une nouveauté.

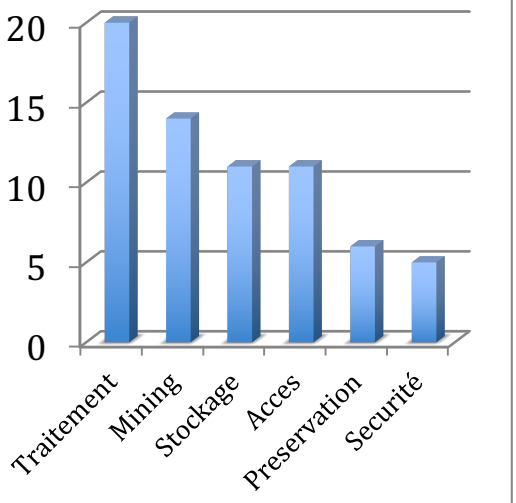
Big Data à l'AMU

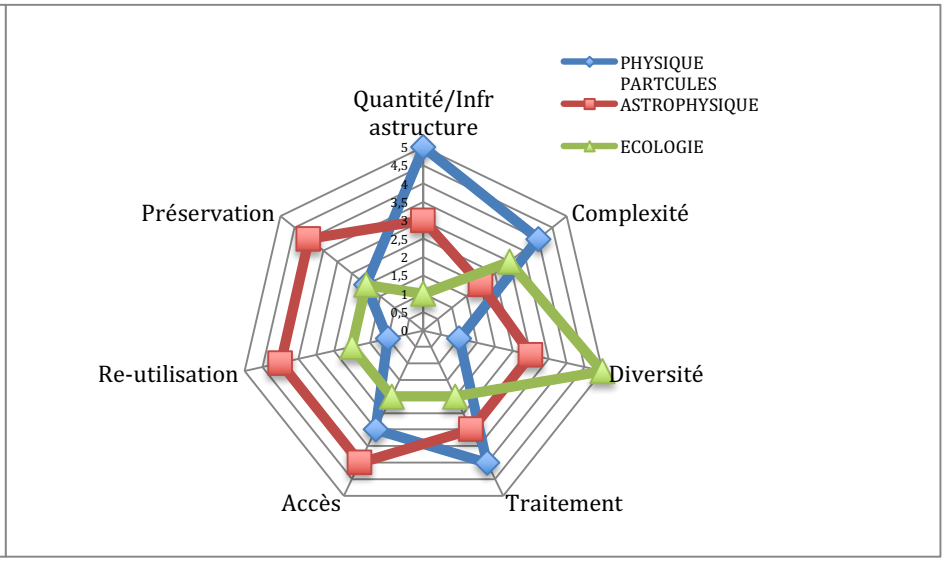
Les atouts de l'AMU dans ce domaine sont très importants :

- Taille importante, avec un spectre très large de domaines d'enseignement et de recherche. Par exemple, AMU étant la plus grande université francophone, elle possède une légitimité certaine à mener des recherches de pointe en linguistique et sciences du langage basés sur l'analyse de corpus massifs en français.
- Recherche de pointe avec un rayonnement international. Par exemple, les données analysées au CPPM font d'AMU un nœud essentiel dans l'effort international autour du LHC.
- Diversité des thématiques de recherche : sciences et technologies, environnement, énergie, humanités, santé, …
- Capacité de formation exceptionnelle dans les aspects liés aux données digitales. La formation de spécialistes Big Data par l'AMSE.
- Connexions avec des entreprises du domaine dans la région. Par exemple, les contact noués avec le pôle mer à Toulon ou TheCamp à Aix.

Axes de recherche

Les laboratoires d'AMU possèdent une recherche de pointe dans plusieurs domaines scientifiques (voir rapport COS, excellence en publications scientifiques etc.) La plupart de ces domaines d'excellence font face à une montée exponentielle des données digitales, et en profitent. Si certains domaines bénéficient d'une expertise ancienne dans le traitement des données massives (comme la physique des particules, l'astrophysique, la génétique etc.), d'autres domaines découvrent de nouvelles méthodes et paradigmes d'analyse de données (SHS, immunologie etc.), tandis que la recherché théorique sur les données massives connaît un essor important. Un atout important de l'AMU est aussi l'existence des infrastructures de calcul scientifiques complémentaires, déjà en cours d'unification (projet M3AMU labellisé CPER).





Une animation interdisciplinaire au sein des PR2Is a identifié 26 laboratoires avec des compétences dans ce domaine (liste en annexe). Les réponses par mot clé (figure de gauche) illustrent les préoccupations dans ces laboratoires. La complémentarité des domaines scientifiques est illustrée dans la figure de droite.

L'institut Big Data pourra consolider des actions interdisciplinaires suivant quatre piliers :

- Fouille des données
- Accès, mise à disposition et préservation
- Usages, éthique et droit
- Infrastructures et technologies pour des données massives

Plus de détails sur les compétences en places, les besoin et les opportunités dans chacun de ces axes se trouvent dans l'annexe de ce document.

Enseignement

Une étude Européenne fait état d'une croissance impressionnante du domaine des « big data ». Par exemple, dans un rapport[1] de la Comission Européenne du 2014 : *« Big data*

[1] https://ec.europa.eu/digital-single-market/en/news/communication-data-driven-economy

technology and services are expected to grow worldwide to USD 16.9 billion in 2015 at a compound annual growth rate of 40% – about seven times that of the information andcommunications technology (ICT) market overall. A recent study predicts that in the UK alone, the number of specialist big data staff working in larger firms will increase by more than 240% over the next five years. ». Dans une autre étude du groupe Axa mentionne un besoin de 200000 nouveau postes est évoquée à l'horizon 2020 en France. Il est clair que le marché du travail ainsi que les compétences en place à l'AMU offrent une excellente perspective pour la création de structures d 'enseignements dédiées.

L'Institut sera l'hôte idéal pour la création d 'un démarche enseignement-formation « Big Data » avec des facette multiples, mais avec un fort caractère interdisciplinaire (cette recommandation est explicite dans le rapport du COS). Tandis que la création d'un master dédié pourrait faire face à court terme à des problèmes administratifs (« labelisation », doublons dans les master déjà existants), l'institut pourra accueillir des cours inter-master sur la science des données, en profitant de manière coordonnées des infrastructures commune de calcul haute performance de l'AMU. La mise en place d'une telle démarche profitera de la proximité d'unités d'enseignement réputées au sein de l'AMU et de ses partenaires (Polytech, Centrale, etc.), des avancées les plus récentes obtenues par les unités de recherche de pointe et de la présence des entreprises concernées pouvant accueillir nos étudiants et les initier aux enjeux qui leur sont spécifiques. Ce « cocktail » sera à même d'offrir une formation de qualité, avec une vision pluri- et interdisciplinaire ce qui constituera un atout important par rapport aux enseignements plus classiquement attachées à des unités d'enseignement mono-disciplinaires. L'institut pourra s'impliquer dans les formations doctorales existantes ou en création afin d'y intégrer des compétences en données massives, valorisables par le docteur. Enfin, il existe une forte demande de formation de la part des chercheurs et des enseignants-chercheurs qui souhaitent se former dans des techniques et méthodes liées à l'utilisation des grandes masses de données. Cette démarche inter-laboratoire pourrait avoir un fort effet fédérateur et pourrait stimuler également les projets de recherche joints. Les initiatives d'enseignement et formation pourraient s 'appuyer sur l'infrastructure de calcul en développement sur le site d'AMU, qui est possède la diversité des moyens et la taille idéales pour mettre à l'échelle les méthodes « big data », dans des projets de recherche et dans des projets d'enseignements qui pourrait être combinés dans certains cas avec des projets industriels.

Coopération avec les entreprises

Certaines thématiques de recherche impliquent naturellement des entreprises de grande taille ou des start-ups technologiques. La proximité des campus technologiques (St. Jerome, Arbois) ainsi que les coopérations déjà existantes dans des projets communs constituent un atout important. L'Institut favorisera la création de projets communs avec des entreprises. On peut citer par exemple le traitement des données issues de simulations « grandeur nature » de ville intelligente par TheCamp sur le plateau de l'Arbois. La coopération se situerait sur le terrain du démarrage des start-ups de haute technologie. L'institut pourrait abriter et soutenir des projets triangulaires : chercher-start-up-infrastructure avec la possibilité forte d'impliquer des étudiants et ingénieurs des laboratoires. Cette démarche pourrait améliorer la visibilité de l'AMU auprès du monde économique régional et national.

Proposition Institut des Données Massives

Cadrage

La proposition a été discutée au sein du comité de pilotage de l'animation "Big Data" des PR2I. Ensuite une réunion plénière avec une large participation (30 personnes) a eu lieu le 3 juin, ou les aspects recherche, formation et valorisation ont été abordés avec des exemple concrets. Récemment (le 7 Octobre 2016), ces propositions ont été discutés dans le comité de pilotage.
Concernant la forme, quatre modèles ont été discutés:

1) "Institut convergence" (taille importante)
2) Graduate School (formation)
3) Recherche pure
4) Integrateur recherche/applications.

En tenant compte de l'information et des échanges des 2 dernières années, les modèles 1 ou 4 sont pertinents. Le comité souhaite proposer un modèle qui exploite la diversité et la flexibilité potentielle des sujets, tout en étant identifiable en terme d'excellence et de retour scientifique. S'il est clair que la valeur rajouté en termes de recherche, formation et valorisation doit être soulignée avec des exemple concrets, la spécialisation de la proposition sur des axes prédéfinies (comme par exemple des coopérations bi- ou tri-disciplinaires) est en général ressentie comme réductrice vis-à-vis du potentiel multi-disciplinaire de l'AMU, ainsi que de la capacité à stimuler de nouveau projets a moyen-long terme.

Missions et actions

I. **Mission recherche:** Sans doubler les initiatives interdisciplinaire existantes, l'Institut pourrait créer un cadre favorable pour les coopérations autour des grandes masses de données, en agissant comme pépinière de projets de recherche entre le laboratoires de l'AMU. Sur l'exemple de plusieurs initiatives de ce type dans le monde, l'Institut pourrait recevoir en résidence des projets d'excellence multi-polaires afin de fédérer et exploiter les atouts d'AMU dans les domaines s'appuyant sur des grandes masses de données (exemple: http://www.bigdata.cam.ac.uk/ *Cambridge Big Data addresses this multidisciplinary research challenge by bringing together expertise from across the University, both in the applications of Big Data and in underpinning technologies and concepts. Cambridge Big Data represents research in over 50 departments*.)
    - Actions possible :
        - Chercheurs/doctorants en « résidence » sur des projets de 1-3 ans.
            - Projets interdisciplinaires avec une composante « données massives » en utilisant des techniques transdisciplinaires
            - Projets collaboratifs avec au moins 2 laboratoire AMU et une composante internationale.
            - Exemples discuté pendant les worshops : utilisation des grilles de calcul pour l'analyse des données économiques ; le stockage/collecte de données dans le cadre de projets de science participative, utilisation des algorithmes de « machine learning » dans les expériences du LHC – thèse inter-ED en cours.

- A noter qu'il ne s'agit pas « d'expérimenter » pour la première fois des idées, mais bien d'armer avec une composante interdisciplinaire des projets dont l'excellence a été validée au niveau international.

II. **Mission formation** L'Institut pourrait créer une dynamique inter-master, inter-ED et pourrait proposer des séries de cours/séminaires au niveau ITA/chercheurs, s'appuyant sur l'expertise AMU (répertorié dans le document de synthèse).
    - Actions possibles :
        - Labellisation enseignements master, création de cours combinés.
        - Cours au niveau doctoral co-supervisé par plusieurs composantes sur la thématique « Big Data ».
        - Formation chercheurs (exemple : Utilisation des grandes structures de calcul à l'AMU, Nouveau algorithmes
        - Série séminaires invités recherche-industrie.
        - Enseignement à distance, mooc etc. (combinés avec des TP à distance sur l'infrastructure dédié AMU)

III. **Mission valorisation et rayonnement:** en s'appuiant sur les infrastructures AMU (mésocentre, M3AMU, Data center) l'institut pourrait créer un cadre favorable pour des projet tri-angulaires: chercheurs/start-ups/data-meso-grid-centre AMU, et donc abonder en support académique les poles d'innovation régionaux. L'institut pourrait (en connexion étroite avec Protis-valor et avec les Pôles) stimuler et accompagner des candidature aux appels à projets pertinents pour AMU. Mission communication à prendre en compte également.

Organisation et fonctionnement de l'Institut

Gouvernance

- Direction
- Comité de Pilotage
- Conseil scientifique

Départements

- Recherche
- Enseignement-Formation
- Relation avec le monde économique et les programmes de financement
- Infrastructure de calcul

Personnel de l'Institut

- Personnel coordination et support associé/dédié.
- Equipes des laboratoires associés (participation collaborative).
- Equipes « en résidence » : projets soutenus par l'Institut, accès aux unfrastrctures et au support scientifique et technique de l'Institut.
- Doctorants/Chercheurs/EC invités (recherche et enseignement) (financement dédié AMU, projets externes de l'Institut)

Fonctionnement

- Localisation : Institut virtuel
- Campagnes appels à projets sur des fonds propres
- Cellule programmes, aide au dépôt de projets ANR, Europe etc..
- Organisation événements : ateliers, séminaires
- Agenda cours/formations labellisés
- Communication
- Instance d'évaluations (conseil scientifique)
- Représentation dans les groupes de travail au niveau national (GDR BIg Data) et international (EUDATA, etc.)

Annexe A: Comité de pilotage de l'animation "Big Data" PR2I AMU

- **Sciences et Technologies**
  - Cristinel Diaconu (CPPM) : diaconu@cppm.in2p3.fr (coordonnateur)
  - Christian Surace (LAM): christian.surace@lam.fr
  - Thierry Artières (LIF) : thierry.artieres@lif.univ-mrs.fr
  - Nicolas Ferré (Mesocentre) nicolas.ferre@univ-amu.fr
  - Sebastien Fournier (LSIS) sebastien.fournier@lsis.org
- **Environnement**
  - Joel Guiot (ECCOREV) : guiot@eccorev.fr
  - Jean-Pierre Bracco (LAMPEA): jean-pierre.bracco@univ-amu.fr
- **Energie**
  - Mohamed Quafafou (LSIS) (mohamed.QUAFAFOU@univ-amu.fr)
- **Humanités**
  - Patrice Bellot (LSIS) (patrice.bellot@lsis.org)
  - Bernard Bel (LPL) (bernard.bel@lpl-aix.fr)
  - Dominique Augey (Faculté de Droit, AMU) dominique.augey@univ-amu.fr
  - Fidelia Ibekwe-SanJuan IRSIC fidelia.ibekwe-sanjuan@univ-amu.fr
- **Santé**
  - Frédéric Barras (LCB): barras@imm.cnrs.fr
  - Jean-Charles Dufour (SESSTIM) : jean-charles.dufour@univ-amu.fr
  - Christophe Béroud (Genetique) christophe.beroud@inserm.fr
- **AMU Recherche**
  - Mossadek Talby (CPPM): talby@cppm.in2p3.fr

Annexe B : Animation Big Data au sein de PR2I

Dans le cadre des pôles de recherche interdisciplinaires et intersectoriels de l'AMU, une animation sur les grandes masses de données a vu le jour dans le courant de l'année 2014 (le comité de pilotage en Annexe A).

Les objectifs de cette animations sont :

- Identifier les acteurs AMU dans ce domaine
- Identifier les besoins
- Suivre les appels à projets, documenter et diffuser
- Stimuler les échanges et la construction des consortia
- Faire émerger des nouvelles collaborations
- Connexions au niveau national et international
- Connexion avec les PR2I, Amidex, AMU

Une première prise de contact a été initiée à travers un formulaire distribué dans tous les laboratoire AMU ; 26 laboratoires ont nommé une personne de contact (Annexe B) et envoyé un formulaire rempli (Annexe C). Les laboratoires ont présenté brièvement leurs activités dans le domaine pendant une journée d'animation organisé le 24 Novembre 2015. Cette première réunion a permit une mise en perspective des compétences et des besoins des laboratoires de l'AMU dans le domaine des « Big Data ». Une distribution des mots-clefs mentionnés dans les réponses est présenté plus bas et illustre bien l'envergure et le potentiel très importants des laboratoires AMU dans ce domaine.

Compétences, besoins et projets dans les laboratoires AMU

Méthodologies de traitement et fouille (« data mining »)

Compétences et points forts

- Compétences dans les bases théoriques dans les méthodes de fouille, algorithmes, détection de corrélations, visualisation, classification de grande taille représentation, parallélisassions.
- Position de leadership au niveau national et international dans plusieurs domaines liés aux grandes masses de données (astrophysique, physique des particules, génomique)
- Développement d'outils et programmes (ex. bioinformatique), analyse de grandes masses de données.
- Panel de champs disciplinaires très complémentaire et favorisant les coopérations au niveau d'AMU avec un excellent potentiel au niveau international.

Besoins

- Collaborations souhaitées (recherche) entre fournisseurs de Big Data (biologistes, physiciens, chercheurs en sciences sociales, etc.), curateurs de données, statisticiens
- Formation et enseignement : master/licence, et aussi des formation pointues pour des chercheurs avancées.
- Plateformes d'échange données et logiciels, accès à des ressources communes.

Opportunités et projets

- Collaborations sur les méthodes et algorithmes : dimension spatiale des données (mobilités), représentations des données (sémiologie graphique), textes/ontologies.
- Montage des projets thématiques sur le Big Data incluant des domaines des sciences sociales, l'histoire, l'art, les biotechnologies, l'astronomie etc. Exemples de thématiques possibles/engagées : influence de l'e-publicité, traitement du langage, indexation des données sur la biodiversité méditerranéenne, biodiversité des habitats coralligènes, risque en finances et marchés, bio-diversité acoustique, données santé et gestion hospitalière, réseaux sociaux, géosciences, urbain analytics, renforcer nos relations avec des acteurs gestionnaires des territoires.

Mise à disposition et préservation

Compétences et point forts

- Projets d'envergure internationale dans la manipulation et l'accès aux grandes masses de données (Physique des particules, astrophysique, biologie).
- Participation à des projets de bases de données en biologie, pathologie, ressources pour la linguistique, microscopie, documents/Web, réseaux sociaux, données patrimoniales, géosciences ; bases de données relationnelles.
- Traitement systématique des données pluri-disciplinaires, compétences en stockage massif et pérenne. Participation à des projets dédiés à la préservation des données scientifiques au niveau national et international.

Besoins

- Plateformes de mise en commune des données des différentes disciplines qui pourrait stimuler l'échange et l'émergence de projets pluri- et inter-disciplinaires.
- Pour cela, une infrastructure matérielle importante (grappes de calcul, serveurs spécialisés, stockage) permettant de mettre en œuvre ces activités est nécessaire.

Opportunités et projets

- Traitement du langage et de ses bases cérébrales (LPL)
- Banques de données et des bases de données généralistes ou spécialisées, ainsi que des plates-formes logicielles donnant accès à des outils de bioinformatique et d'analyses d'images, des plus génériques aux plus pointus. Préservation pérenne de données scientifiques.
- Intégration de données hétérogènes, la mise en place de chaînes de traitement automatisées, la définition d'ontologies, la fouille de texte.

Usages, droit et éthique

Compétences et points forts

- Enjeux sociétaux du Big Data et de l'Open Data et les dimensions épistémologiques et éthiques de leurs usages: collecte, du traitement et d'usages de grandes masses de données dans différents domaines tels que la santé, les médias et le journalisme, la publicité et le e-marketing, la recherche scientifique, etc. .
- Communication engageante numérique, économie numérique, intelligence économique et veille informationnelle.

- Usages psychologiques et sociaux des medias sociaux en tant que « conscience collective virtuelle ».

Besoins

- La mise en commun des compétences dans les SHS dans ce domaine.
- Collaboration avec des laboratoires en sciences (médecine, neuro-sciences, physiques, etc) qui disposent des données massives pour étudier les modes d'usages sur le terrain.
- Formation et enseignement en éthique des données et préservation/archivage à long terme.
- Enjeux sur la sécurité, compétences sur le psycho-social et l'impact sur les usages

Opportunités et projets

- Etudes des transformations organisationnelles liées au Big Data.
- Enjeux éthiques et sociétaux de la génomique personnelle
- Enjeux psycho-social et impact sur les usages
- Propriété intellectuelle sur le code, droit et médecine (protection des données médicales), sécurité informatique, philosophie

Infrastructure et technologies

Competences

- Infrastructures de calcul HPC (méso-centre) et Grille (Tier2 LHC, CPPM)
- Expertise dans l'utilisation massive de grand centres de calcul grille et HPC au niveau national et international.
- Accès à des technologies de pointe ; outils d'accès aux ressources, virtualisation, techniques de calcul sur le « cloud ».

Besoins

- Inter-connectivité, accès aux ressources distribuées, amélioration des bases d'infrastructure informatique dans certains laboratoires.
- Formation pour l'utilisation massive des grandes ressources informatiques.
- Besoins génériques en experts « data management », systèmes/infrastructures (réseau, Cloud, langages), expertises technologiques concernant les outils permettant de traiter le Big Data et d'intégrer des données issues de différents canaux et sources

Opportunités et Projets

- Projet de rapprochement entre le méso-centre et la grille du CPPM est en cours de développement. Lorsqu'elles sont disponibles, les ressources du méso-centre pourront être intégrées à Dirac, le logiciel de gestion de la grille du CPPM et participer ponctuellement au traitement très efficace de données massives.

Annexe B: Contacts dans les laboratoires

Nom	Prénom	Institut	e-mail
Fossati	Caroline	Institut Fresnel	caroline.fossati@fresnel.fr
Novelli	Noël	LIF (info)	noel.novelli@lif.univ-mrs.fr
Libes	Maurice	OSU Pytheas	maurice.libes@univ-amu.fr
Humbel	Stéphane	iSm2	stephane.humbel@univ-amu.fr
Takerkart	Sylvain	INT	sylvain.takerkart@univ-amu.fr
Torresani	Bruno	i2M	bruno.torresani@univ-amu.fr
Gonçalves	Bruno	CPT	bgoncalves@gmail.com
Ferré	Nicolas	Mesocentre	nicolas.ferre@univ-amu.fr
Kadoch	Benjamin	IUSTI	benjamin.kadoch@univ-amu.fr
Vicente	Jérôme	IUSTI	jerome.vicente@univ-amu.fr
Ibekwe-Sansuan	Fidelia	IRSIC	fidelia.ibekwe-sanjuan@univ-amu.fr
Ghattas	Badih	I2M	badih.ghattas@univ-amu.fr
Lancini	Agnès	CRET-Log	lanciniagnes@hotmail.com
Blanpain	Cyril	OSU Pytheas	blanpain@osupytheas.fr
David	Romain	IMBE	romain.david@imbe.fr
Guiot	Joël	CEREGE	guiot@cerege.fr
Surace	Christian	LAM	christian.surace@lam.fr
Allard	Paul	ESPACE	paul.allard@univ-amu.fr
Talla	Emmanuel	LCB	talla@imm.cnrs.fr
Lugiez	Denis	LIF	denis.lugiez@univ-amu.fr
Llari	Maxime	LBA	maxime.llari@ifsttar.fr
Beroud	Christophe	INSERM	christophe.beroud@inserm.fr
Laporte	Cathy	AMU/Cellule Europe	cathy.laporte@univ-amu.fr
Laurent	Sébastien	GREQAM	sebastien.laurent@univ-amu.fr
Gingold	Arnaud	LPL	arnaud.gingold@lpl-aix.fr
Bel	Bernard	LPL	bernard.bel@lpl-aix.fr
Spinelli	Lionel	CIML/TAGC	spinelli@ciml.univ-mrs.fr
Chetrit	Bernard	CRCM	bernard.chetrit@inserm.fr
Jaeger	Sébastien	CIML	jaeger@ciml.univ-mrs.fr
Augey	Dominique	LID2MS/GRECAM	dominique.augey@univ-amu.fr
Tsaregorodstev	Andrei	CPPM	atsareg@in2p3.fr
Talby	Mossadek	CPPM	talby@cppm.in2p3.fr
Tadrist	Lounes	IUSTI	lounes.tadrist@univ-amu.fr
Quafafou	Mohamed	LSIS	mohamed.quafafou@univ-amu.fr
Bellot	Patrice	LSIS	patrice.bellot@univ-amu.fr
Dufour	Jean-Charles	SESSTIM	jean-charles.dufour@univ-amu.fr
Diaconu	Cristinel	CPPM	diaconu@cppm.in2p3.fr
van Helden	Jacques	TAGC	Jacques.van-Helden@univ-amu.fr
Martin	Jean-Charles		jean-charles.martin@univ-amu.fr
Damon	Céline	ProtisValor	celine.damon@univ-amu.fr
Bartoli	Jonathan	ProtisValor	jonathan.bartoli@univ-amu.fr

Annexe C: Compétences dans les laboratoires AMU (en bref)

IRSIC
Enjeux sociétaux du Big Data et de l'Open Data et les dimensions épistémologiques et éthiques de leurs usages: collecte, du traitement et d'usages de grandes masses de données dans différents domaines tels que la santé, les médias et le journalisme, la publicité et le e-marketing, la recherche scientifique, etc

LPL
Partenaire du réseau européen CLARIN dédié aux données numériques de la linguistique.
Centre de ressources SLDR (http://sldr.org) pour la linguistique.
Archivage pérenne des données à valeur patrimoniale.
Bases de données de parole pathologique très spécifiques (et uniques en France).

LAM
Datamining dans les données astrophysiques, (cosmologie, planètes extra solaires).
Visualisation de grands échantillons de données, visualisation 3D, Base de données astrophysiques, Développement d'application de traitement de données.

IMBE Données en écologie et sur la biodiversité, développement de graphes à partir de ces données, fouille de donnée, systèmes d'observations et d'informations répartis
Données génomiques, métagénomiques, transcriptomiques, RNASeq, RADSeq
Analyses de polymorphisme, d'expression de gènes, de phylogénie Après publication les données génétiques sont soumises à des bases de données publiques (Genbank par example) mais il y a un intérêt pour les stocker localement afin de poursuivre les analyses.

GREQAM
Estimation et prévision du risque sur les marchés financiers. Utilisation et développement d'outils statistiques et économétriques pour extraire l'information de données à très haute fréquence sur de très grands portefeuilles d'actifs.
Méthodes économétriques/statistiques permettant d'extraire des mesures de risque (volatilités, corrélations) très fines, robustes (à la présence de bruit dans les données observées, à des sauts abruptes dans les prix, …) et qui sont rapides à obtenir malgré le grand très nombre d'observations.

IBDM
Gestion de grandes masses de données de microscopie (images, films) pour l'exploration quantitative et intégrée de systèmes vivants à différentes échelles spatiales : moléculaires, cellulaires, tissulaires, organismes entiers.

LSIS
Recherche et extraction d'information dans de grandes masses de documents (Web, bibliothèques numériques)
Suivi d'opinion, contextualisation et détection de nouveauté sur les réseaux sociaux
Fouille et Intégration de Données
Gestion de données : élaboration de systèmes et langages, intégration (interopérabilité), qualité des données, données spatiales et biologiques

Cloud computing : Provenance, cloud/web services
Big data analytics : Algorithmes probabilistes et hypergraphes
Architecture orientée services : conception d'architectures offrant différents services distribués pour l'acquisition, le stockage et l'analyse de données massives (Big Data).
Structuration non supervisée et supervisée de masses de données multimodales, vidéo (les collections d'images du web), audiovisuelles (ex : les challenges NIST TREC VIDEO) ou audio (parole et bioacoustique).
Modèles neuronaux profonds, les plus performants en indexation de masses de données d'images et sons
Analyse de données pour la surveillance, pilotage et aide à la décision pour les systèmes Complexes

ESPACE
Traitement systématique des données pluridisciplinaires (sociologie, histoire, écologie…) accumulées depuis 20 ans sur les inondations, les territoires fluviaux (en particulier la Camargue) et les zones humides dans le cadre de programmes de recherche, de thèses, de mémoires terminés ou en cours.

ESPACE
La recherche effectuée par les membres l'UMR ESPACE trouve son unité dans l'attention particulière accordée à l'information géographique. Cette réflexion repose sur des avancées théoriques récentes (théorie de la complexité par exemple), une forte dimension épistémologique, et bien entendu par la pratique quotidienne de cette information et des outils nécessaires à son traitement (Systèmes multi-agents, SIG, Géostatistique, etc.).

CPPM
Traitement de grandes masses de données
Infrastructure Grille (LHC)
Système de calcul distribué DIRAC
Projet de préservation de données

CEREGE
Géosciences de l'Environnement : Physique et chimie de la Terre, Biologie et écologie : passé/présent. Ressources et Risques en Environnement.

SESSTIM
Utilisation de données complexes dans les domains: Médecine personnalisé : cancer, autres pathologies, Veille Sanitaire, Méthodes quantitatives, traitement de l'information médicales
Méthodologies d'enquêtes et recherches axées sur l'apport des média sociaux et technologies du web 3.0 (objets connecté) pour la pratique médicale et la meilleure connaissance des problèmes de Santé en population (veille sanitaire notamment).

LCB
Génomique Comparative / Analyse des génomes
Développement de bases de données relationnelles
Développement d'outils et de programmes bioinformatiques
Analyse des données de séquençage à haut débit

Le laboratoire dispose d'un service de criblage phénotypique des bactéries basé sur la microscopie. Ce service produit de dizaines de milliers d'images lors d'une campagne de criblage. Actuellement, nous travaillons sur les dispositifs nécessaires à l'exploitation de ces grandes masses de données.

AMU mesocentre
Le mésocentre d'AMU est une structure dépendant de la DRV et basée sur un financement Equipex obtenu en 2010. Il donne accès à des ressources de calcul hautes performances (HPC), susceptibles de générer de grandes masses de données et de les post-traiter (visualisation). Ses utilisateurs sont représentatifs de tous les champs disciplinaires d'AMU présentant un besoin en HPC

TAGC
Différentes thématiques liées à la santé (sepsis, malaria, cancer, développement cardiaque) ;
Approches bioinformatiques pour le traitement des données biologique à haut débit (génomes, transcriptomes, interactomes) ;
Plate-forme technologique spécialisée dans la production de données massives (« Next Generation Sequencing », biopuces).

INT
Acquisition et traitement de grande bases de données issues d'expériences de neurosciences fondamentales et cliniques

LIF
Bases de Données Avancées (BDA) et éQuipe AppRentissage et MultimediA (QARMA)
Fouille de données, détection de corrélations et de dépendances, approximation avec modèle coût (contrôle des ressources), visualisation de données massives, problème de classification de grande taille (nombres d'exemples, nombre de classes, dimensionnalité), apprentissage budgétisé, apprentissage de représentations pour le texte et l'image, parallélisation."

CPT
Mining et l'analyse à grande échelle des données de réseau social pour l'étude du comportement social et humain.

INSERM_UMRS910
Etude génétique des maladies rares.
Gestion des grands volumes de données générées par le séquençage des exomes et génomes.
Compétences dans la création de bases de données génétiques et cliniques, dans le développements d'algorithmes permettant l'analyse de ces données, développement de systèmes de prédiction du caractère pathogène des mutations et de validation d'hypothèses thérapeutiques.

CRET-LOG
"Pas de compétences techniques en matière de traitement des big data
Intérêt au traitement des big data en vue d'aide à la décision dans plusieurs domaines
- décision marketing (compréhension du comportement du consommateur)
- décision logistique (liée aux données de commandes clients, à la traçabilité totale des supply chains, à la maîtrise des risques dans les supply chains)
Intérêt pour les transformations organisationnelles et technologiques de l'entreprise nécessaires pour absorber, traiter et analyser les Big Data
- notamment les big data client, multi-canal, logistique
Intérêt porté au lien entre Big Data et Connaissance
"

iSm2
(pas d'info spécifique fournie)
Calculs en chimie quantique, conservation de données.

IFSTTAR
(pas d'info spécifique fournie)

IMM
Traitement, mining, stockage, signal, génomique
(pas d'info spécifique fournie)

FRESNEL
Traitement tensoriel du signal pour les « Big-Data », Acquisition de données
(pas d'info spécifique fournie, à compléter)

CIML
Utilisation de diverses techniques expérimentales générant de grandes quantités de données dans le domaine de l'immunologie et s'articulant essentiellement autour de 2 axes : (i) Traitement de données d'expériences à haut débit (Biopuces, « Deep Sequencing »…) y compris au niveau unicellulaire par association avec des techniques de microfuidique., (ii) Analyses de données d'imagerie (« Cell Tracking, imagerie bi-photonique, cinétique 3D…).