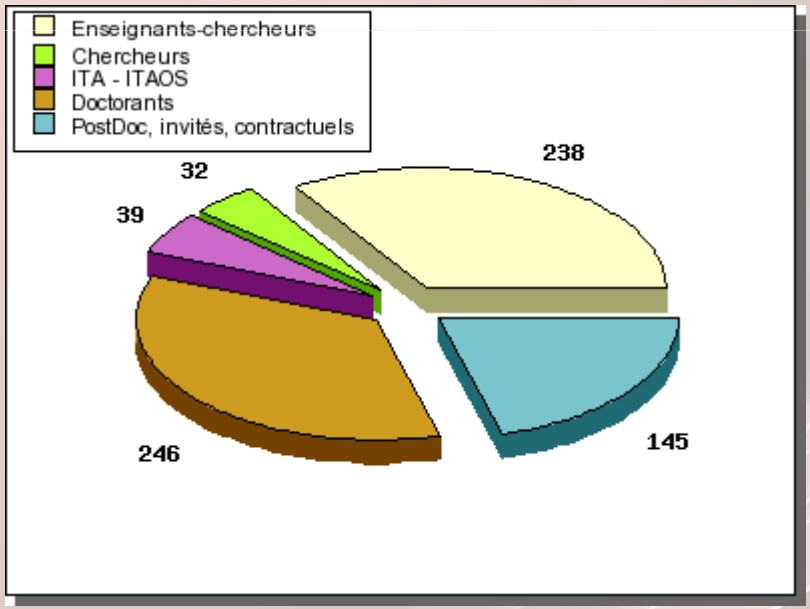
Masses de données et calcul : la recherche en lien avec les Big Data à l'IRIT

8 octobre 2013

L'IRIT en qq chiffres

- 700 personnes sur tous les sites toulousains
- 5 tutelles
- 7 thèmes et 21 équipes
- Nombreuses collaborations et contrats



Thème 1 - Analyse et synthèse de l'information
- **Equipe SAMoVA**
- **Equipe SC**
- **Equipe TCI**
- **Equipe VORTEX**

Thème 2 - Indexation et recherche d'informations
- **Equipe PYRAMIDE**
- **Equipe SIG**

Thème 3 - Interaction, autonomie, dialogue et coopération
- **Equipe ELIPSE**
- **Equipe SMAC**

Thème 4 - Raisonnement et décision
- **Equipe ADRIA**
- **Equipe LILaC**
- **Equipe MELODI**

Thème 5 - Modélisation, algorithmes et calcul haute performance
- **Equipe APO**

Thème 6 - Architecture, systèmes et réseaux
- **Equipe IRT**
- **Equipe SEPIA**
- **Equipe SIERA**
- **Equipe T2RS**
- **Equipe TRACES**

Thème 7 - Sûreté de développement du logiciel
- **Equipe ACADIE**
- **Equipe ICS**
- **Equipe MACAO**

Masse de données et Calcul à l'IRIT

- Des problématiques de recherche établies
- Des infrastructures
- Zoom sur quelques projets
- Une volonté de collaboration et visibilité

Big Data continuité et renouvellement

- Les équipes IRIT traitent de gros volumes de données
  - Infrastructure, sytèmes d'information, imagerie, traitements numériques, simulation, analyse du langage, etc.
- Contexte R&D toulousain : des problèmes "big data"
  - MétéoFrance, aéronautique, chercheurs en astronomie, en biologie …
  - Centres de calcul météo, CICT et CERFACS, CALMIP
- Les verrous liés aux "big data"
  - Données plus complexes, nombreuses, structurées
  - Traitements répétés, temps réel, au fur et à mesure de la production des données
  - Problématiser le passage à l'échelle, les performances

Masses de données et calcul à l'IRIT

- Contexte
  - Données complexes : Hétérogénéité, Qualité, Volume
  - Processus : Rechercher, analyser, indexer, fouiller, explorer, simuler, …
  - Ressources matérielles en constante évolution
    - Multithreading, processeurs vectoriels et calculateurs Petaflopiques
- Problématiques
  - Accessibilité, évaluation et Intelligibilité des données
    - Accès transparent, caractérisation sémantique, agrégation de réponses
  - Interactions Utilisateurs-Processus-Données
    - Prise en compte des contextes : utilisateur, données, environnement, traitement temps réel, gestion de flux
  - Souplesse et robustesse d'accés aux ressources
    - Adapter ressources matérielles et logicielles
    - Intégrer l'évolution des matériels : noeuds hybrides (processeurs cell et multicoeurs) des architectures Petaflops

Compétences IRIT

- Recherche d'information / systèmes d'information
  - RI contextuelle; RI sémantique, RI agrégée
  - Fouille, navigation dans les données, visualisation
- Documents multimédia
  - Annotation de vidéo/son/image
  - Indexation, fusion d'index audio et vidéo,
  - Macro-segmentation en émissions de télévision
- Web sémantique et sémantique de corpus
  - Extraction d'information et de données liées à partir de textes, alignement d'ontologies, annotation sémantique
  - Modèles distributionnels de sémantique lexicale sur de très grands corpus (To)
- Traitements massifs et distribués, grilles
  - Simulation et expérience ssur architectures massivement parallèles
  - Optimisation de requêtes à grande échelle
- Outils théoriques pour le traitement de l'information

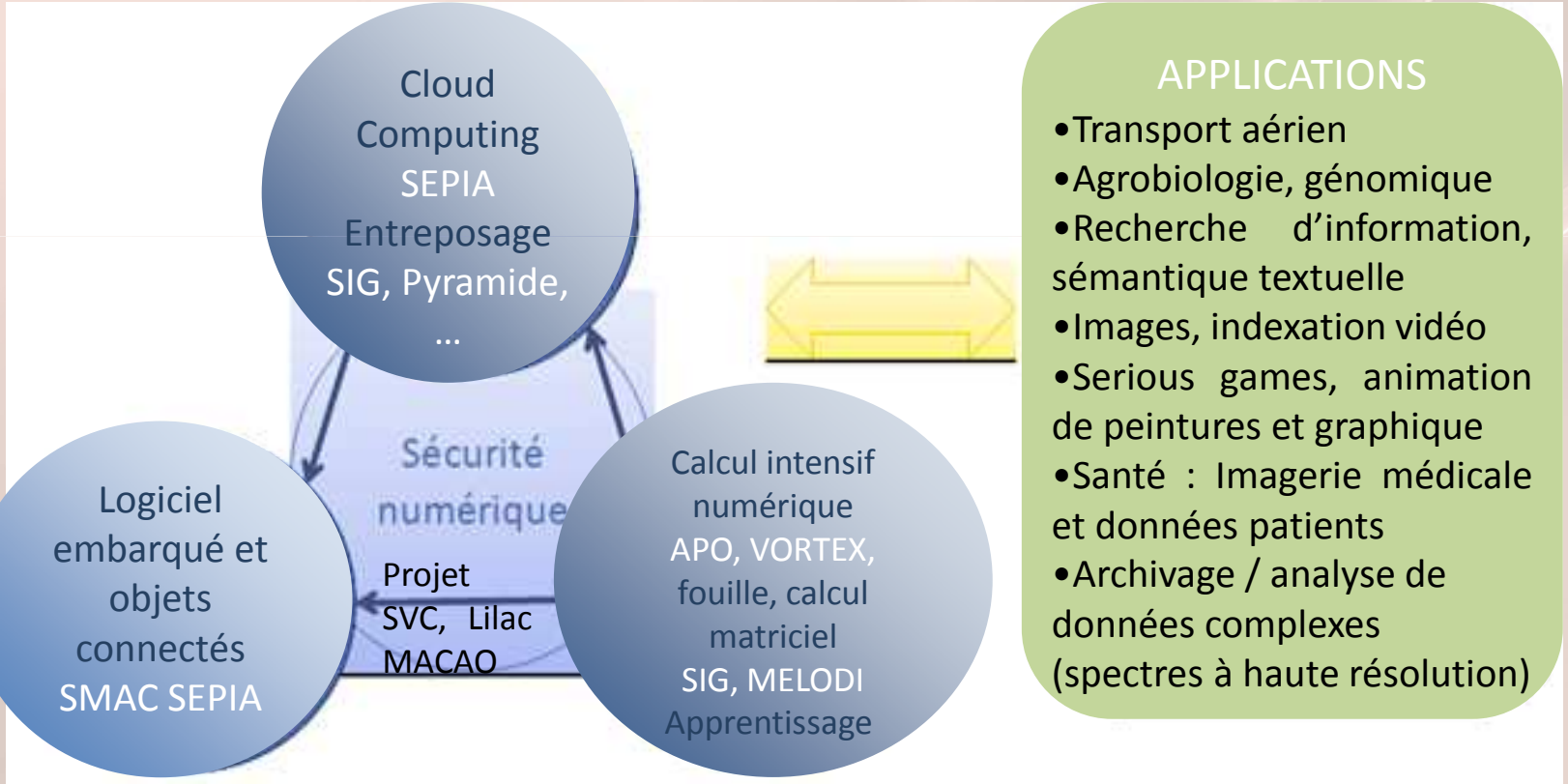
Infrastructures IRIT

- **Plate-forme OSIRIM** (CPER 2007-2013) http://osirim.irit.fr/
  - Matériel (baie de stockage ~407 To, cluster de calcul ~10 nœuds soit 64 coeurs)
  - données : 30 To de textes + contenus TV (plusieurs 10M heures)
  - Services : indexation, recherche d'info., résultats d'expérimentations
  - Gestion des campagnes d'évaluation (Quaero)

- **CLOUDMIP / GRIDMIP / GRID5000**
  - infrastructures de calculs sur grille et clouds
  - **GRID-TLSE** : Services logiciels (calcul) déployés sur la grille : calcul haute performance sur données en bio-informatique, données aéronautiques etc.
  - http://www.irit.fr/-GRID-5000-?lang=fr

Une bonne couverture des thématiques identifiées au niveau national

- http://www.dgcis.gouv.fr/secteurs-professionnels/coeur-filiere-numerique



Zoom : données en recherche / extraction d'information textuelle

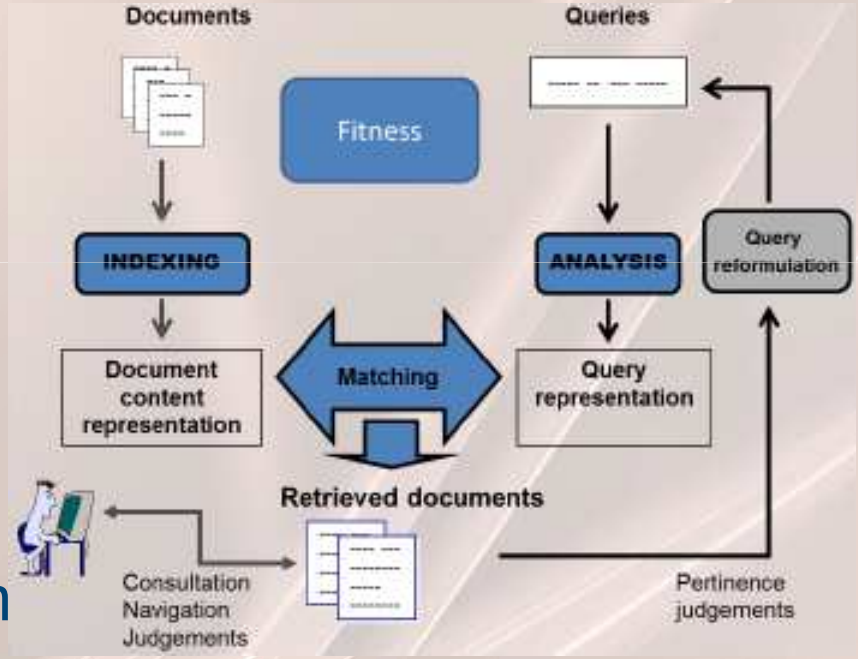
- Recherche d'information textuelle (SIG, MELODI, SAMOVA)
  - Indexation de gros volumes de documents

1996 : 500 Mb

1998 : 2 Gb

Puis : 200 Gb

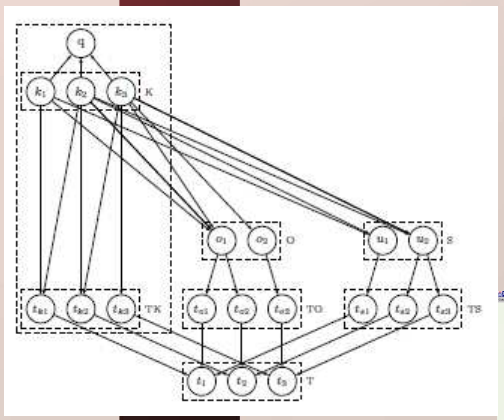
2014 : 25 Tb

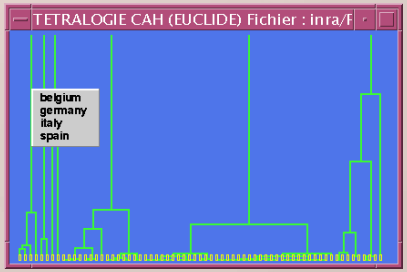


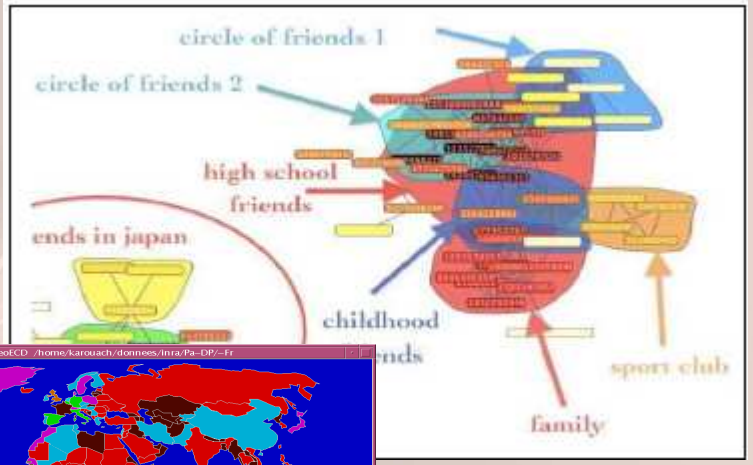
- Extraction d'information
  - Elicitation de structure ; granularité de l'information
  - Extraction de méta-données

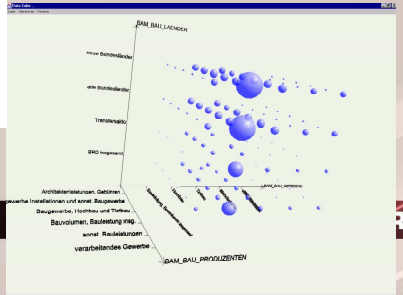
Zoom : information et données

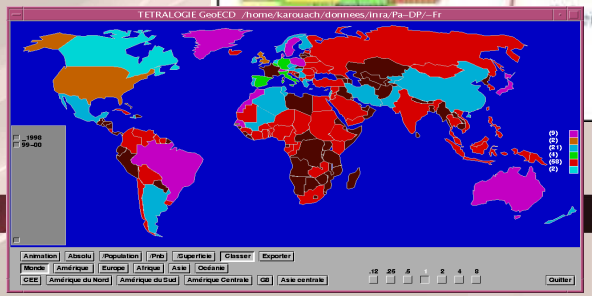
- Exploration et visualisation d'information (SIG, SAMOVA)
  - Entrepôts documentaires, structure de data Warehouse
    - Thèse « masse de données » : automatiser l'entreposage multidimensionnel de données ouvertes
  - Fouille de données
  - Information sociale, médicale, journaux, web …
  - Images de synthèse et visualisation de gros3D









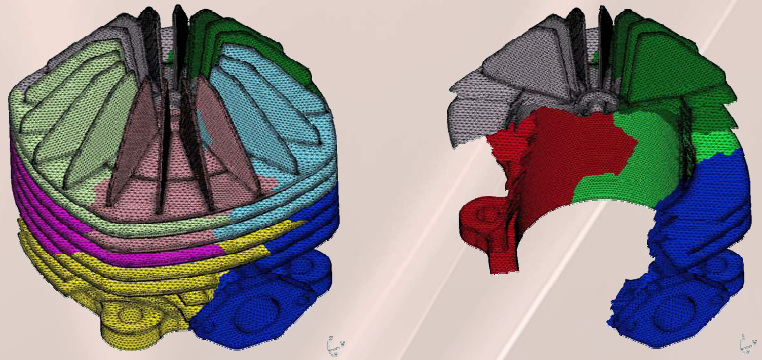


Zoom : analyse numérique (APO)
Problèmes « faciles » et « difficiles »

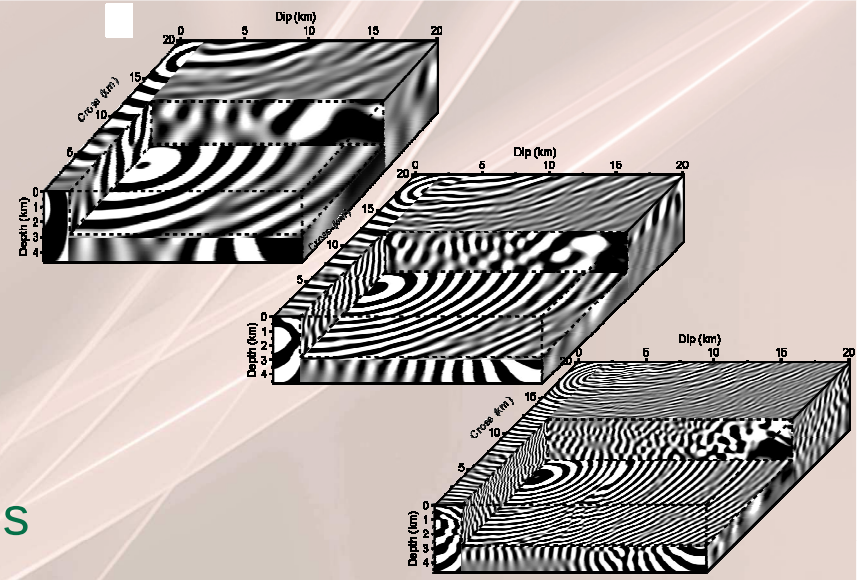
- Même des problèmes « simples » et « réguliers » peuvent devenir complexes quand le nombre de variables devient très grand
  - Ex: système à *n* équations et *n* inconnues
  - Facile : Résoudre le système:
    - 2x+3y=8
    - 3x+2y=7
  - Difficile : traiter plusieurs millions d'équations et de variables
  - Complexité mathématiques : étude de la stabilité numérique, de la précision numérique,
  - Complexité informatique : implantation du code (distribuée)

Zoom : analyse numérique (APO)
Problèmes « faciles » et « difficiles »

- **Maillage et partitionnement d'un carter de moteurs**
  - Objectif : simulation sur calculateurs multiprocesseurs (courtesy of EDF Code_Aster)



- **Détection de champs pétrolifères**
  - consortium SEISCOPE (TOTAL, BP, SHELL, …)
  - déduire la structure du sol à partir de mesures sismiques
  - Problème = résolution d'un système d'algèbre linéaire
  - Verrous: algèbre linéaire creuse impliquant des millions de variables.



Zoom : synthèse d'images (VORTEX)

- Réalité virtuelle
  - Données : images animées pour simulation + prise en compte du réel (co-animation)
  - Traitements : calcul en temps réel
- Edition interactive du rendu en synthèse d'image
  - Données : contexte du rendu hors-ligne de haute qualité
  - Traitements : échantillonage des fonctions de transport lumineux ; utilisation de rayons
  - 1 à 96 Go de données/image ; éditer des images fixes puis propager sur l'animation.
- Peinture animée
  - Données : photographies d'images peintes sur des vitres ;
  - sélectionner des zones du tableau (coups de pinceaux à l'origine de la couler d'un pixel) ; 30 Go d'identifiants des coups de pinceaux par image
  - passer d'une méthode manuelle (peindre des vitres et photographier puis animer les photos) à une génération de vidéo
  - Traitements : animation stylisée à partir d'une vidéo

Zoom : Secure Virtual Cloud (VORTEX et SEPIA)

- But :
  - Créer une infrastructure virtuelle en environnement cloud
  - Garantir la sécurité d'exécution et la confidentialité des données hébergées.
- Moyen :
  - adapter les compétences en simulation au traitement comportemental de remontées de sondes situées dans l'infrastructure afin de déclencher des alertes de sécurité.
- distribution et confidentialité des données en créant un filesystem distribué.
- labellisé dans le cadre des investissements d'avenir lors de l'appel e-cloud n2.

Vers de nouvelles collaborations ?

Au delà du phénomène de mode …

… problématiser le changement d'échelle

… identifier de nouvelles opportunités

- Nationales :
  - appel Mastodons (1 éq. IRIT),
  - Investissement d'avenir (appel 2012, BioDataCloud avec 2 eq. IRIT)
- Locales :
  - volonté de l'UPS -> contact avec des producteurs de données, identification de besoins
  - Collaborations renforcées grâce à Fremit et au Labex CIMI
- Coordination au sein de l'IRIT : axe stratégique depuis 2010

  http://www.irit.fr/-Masses-de-donnees-et-calcul,677-?lang=fr

Contacts

- Infrastructure
  - OSIRIM : Ph Joly, eq. SAMOVA
  - CLOUD : JM Pierson, eq SEPIA
- Données textuelles, archivage, visualisation,confiance et RI
  - SIG (resp. J. Mothe, G. Hubert, Dousset, M. Chevalier, O. Teste, …)
  - MELODI (N. Aussenac, Ph. Muller, T. Van de Cruyse, …)
- Web sémantique, données liées
  - MELODI (C. Trojhan, L. Vieu, O Haemmerlé, N. Hernandez …)
  - SIG (O. Teste, L. Lechani)
- Images, animation, réalité virtuelle, visualisation
  - VORTEX (H. Luga)
- Calcul numérique, optimisation
  - APO (J.M. Alliot)
- Sécurité, fiabilité
  - ACADIE

- Axe masse de données et calcul
  - M. Boughanem (SIG) , N. Aussenac-Gilles (MELODI)
- Thématique Machine Learning dans CIMI
  - S. Afantenos (MELODI), M. Serrurier (Adria) S Mouysset. (APO)