Gen2Bio®

Jeudi 3 avril 2014
à Saint-Malo

E-BIOGENOUEST, VERS UN ENVIRONNEMENT VIRTUEL DE RECHERCHE (VRE) ORIENTÉ SCIENCES DE LA VIE ?

Intervenant(s) : Yvan Le Bras, Olivier Collin

E-BIOGENOUEST

Programme fédérateur Biogenouest co-financé par les Régions Bretagne et Pays de la Loire

- 24 mois
- Lancé depuis Mai 2012
- Porteur : Olivier Collin (IRISA) – Animateur : Yvan Le Bras (IRISA)

OBJECTIFS

Une démarche E-science

Objectifs

- Test d'une initiative e-Science
  - Impliquer les différents acteurs de Biogenouest
  - Intégrer les différents aspects du développement de la recherche dans le Grand Ouest et au niveau national

- Schéma directeur intégrateur
  - Mise en place d'une « plateforme » e-Science
  - Intégration de façon cohérente
    - Résultats du projet eBiogenouest
    - Différentes politiques (Régions, Instituts, Universités,…)
  - Soumission aux différentes tutelles et appels d'offres

POURQUOI ?

La Biologie (devient) une science numérique

Contexte

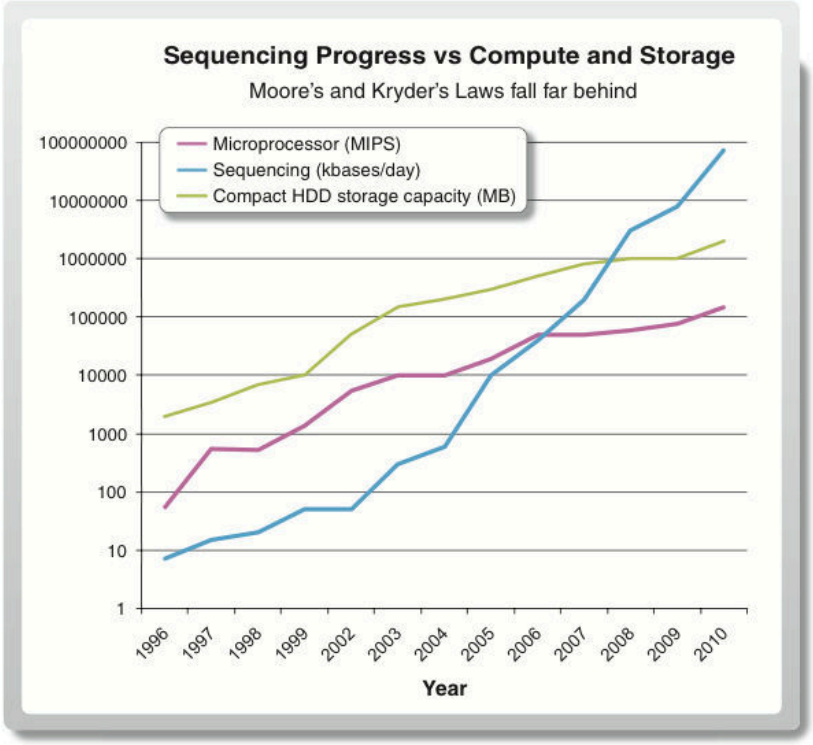
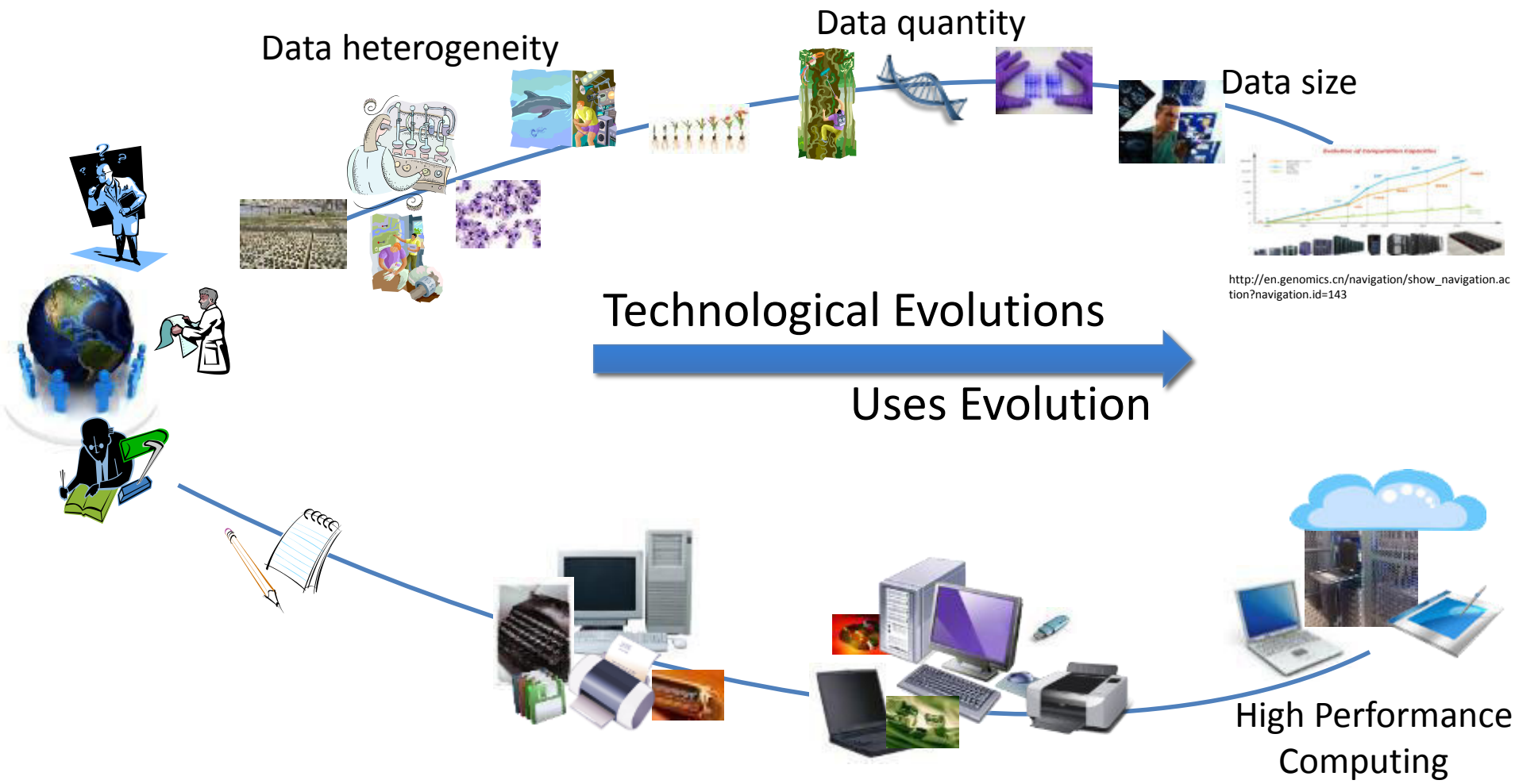


Fig. 1. A doubling of sequencing output every 9 months has outpaced and overtaken performance improvements within the disk storage and high-performance computation fields.

Kahn. On the future of genomic data. Science (2011) vol. 331 (6018) pp. 728-9

- « Déluge de données »
  - NGS première vague
  - *omics
  - Imagerie (ex: 16 To / jour)
- Evolution technologique
  - Nombre de capteurs ↗
  - Production par capteur ↗
- Evolution des infrastructures qui deviennent mutualisées
- Danger : ne pas être en mesure d'analyser les données
- Opportunité :
  - Plus d'information
  - Plus de précision
  - Plus de points de vue

Evolution de la recherche



Life sciences data= digital

Evolution de la recherche

Il y a mille an

Qq centaines d'années

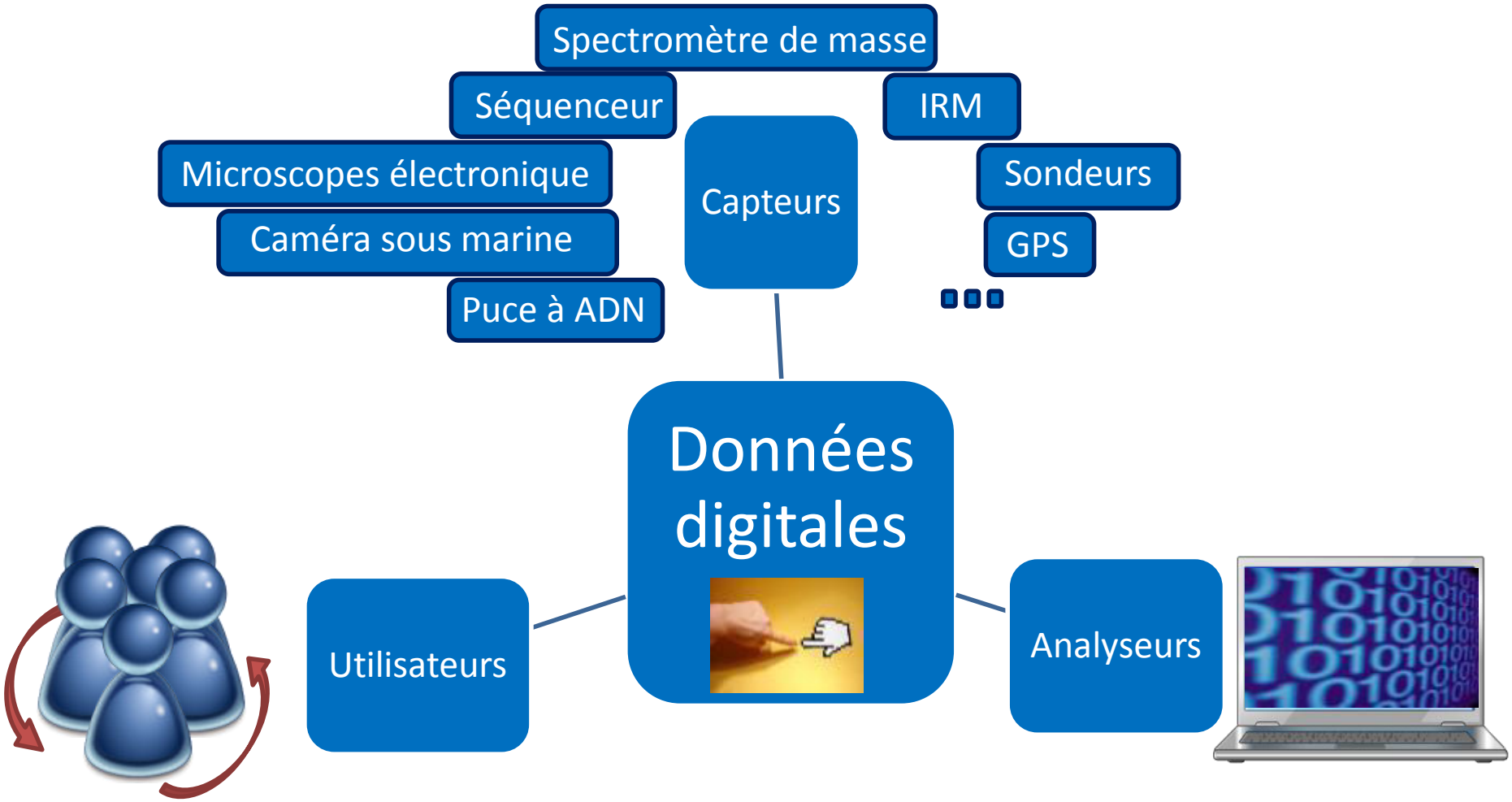
Qq dizaines d'années

maintenant

- Evolution de la Science : Le 4ième paradigme
  - Empirique : description, expérimentation
  - Théorique : modélisation, généralisation
  - Computationnelle : simulation
  - Exploration des données (unification de la théorie, de l'expérience et de la simulation), extraction de connaissances.

- Emergence des disciplines *info
  - Bioinformatique, Chemoinformatique, Astroinformatique, etc.

Evolution de la recherche



Evolution de la recherche

Biologie

-Trouver des biomarqueurs

-Comprendre la structure génétique de populations

-Modéliser le comportement d'un système

Bio-informatique

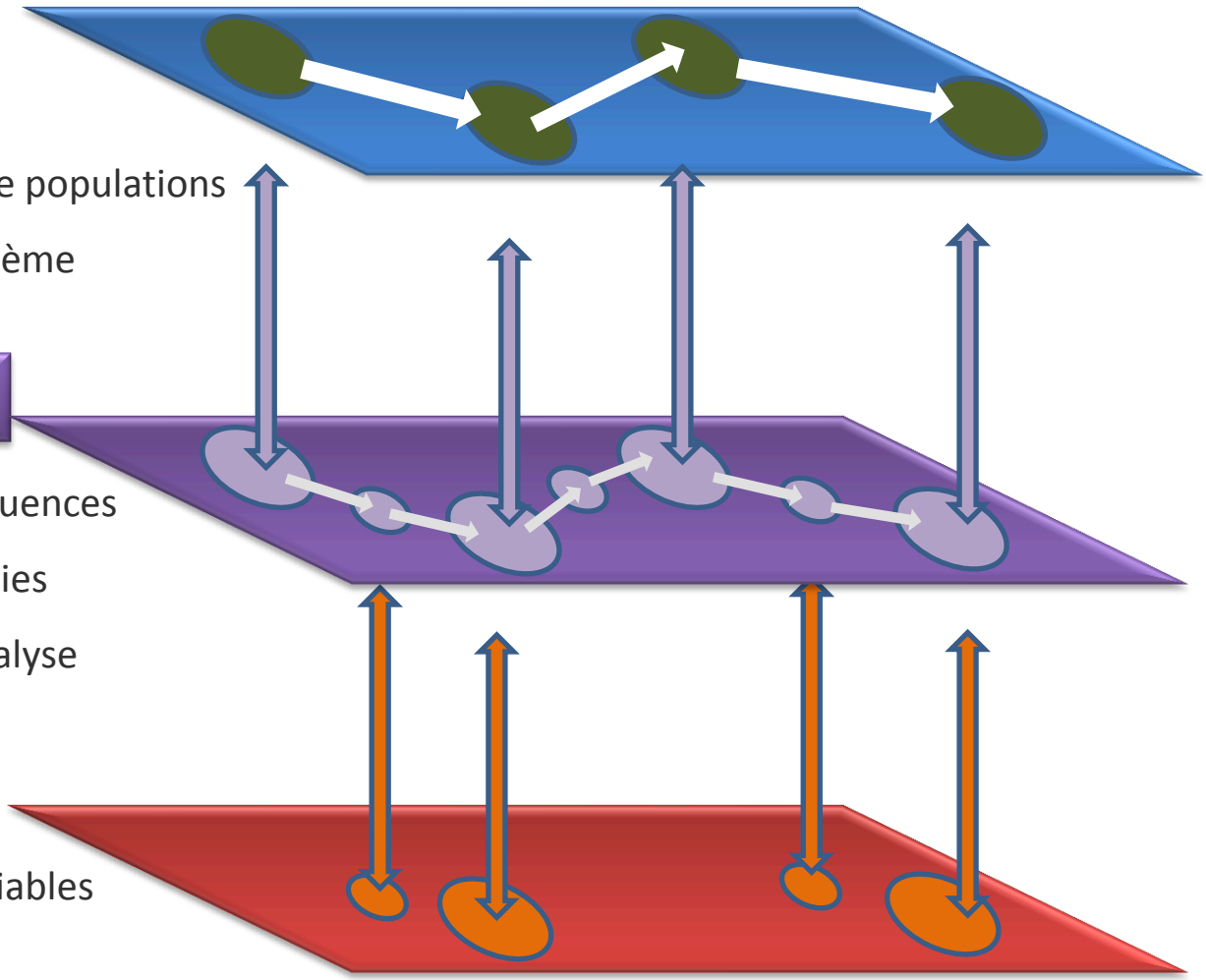
-Créer un outil de comparaison de séquences

-Développer de nouvelles méthodologies

-Concevoir un portail web dédié à l'analyse

Informatique

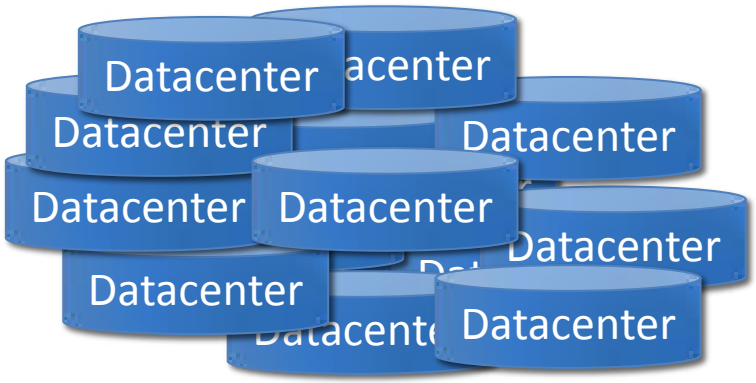
-Proposer des ressources techniques fiables et adaptées



Evolution des infrastructures

- Livre blanc du CNRS
  - Big Data = le 4ième pilier de la science moderne
    - Un enjeu majeur
    - Aspects pas bien pris en compte
  - Sciences de la vie non présentées!

… Ceci explique les demandes actuelles :



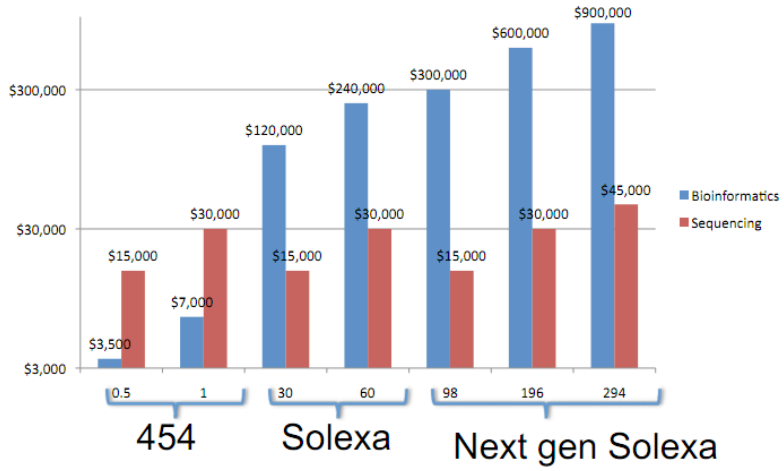
…Pourtant :

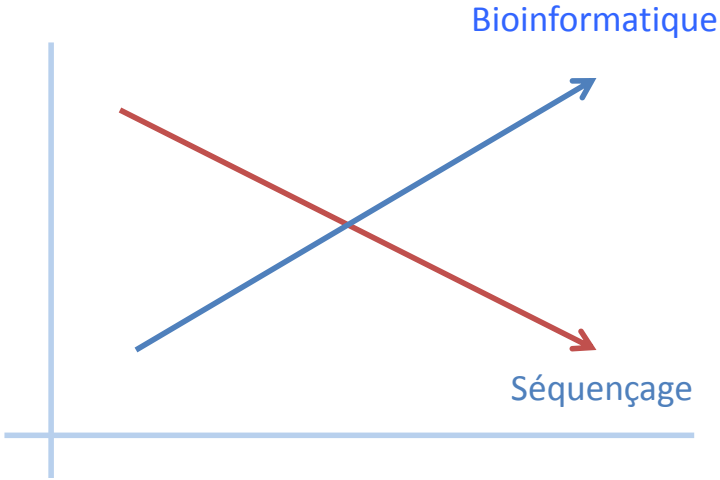
« les **avancées** scientifiques et **l'innovation** dans de multiples domaines (dont ceux des **sciences de la vie et de l'environnement en pleine explosion** actuellement et sans doute rapidement celui des sciences humaines et sociales) dépendent de manière cruciale d'une politique du CNRS, et d'un **renforcement des capacités des infrastructures pour le calcul et l'analyse de données** »

Evolution des usages

- Le passage au numérique nécessite de nouveaux **outils** et de nouveaux **usages** qu'il faut développer
  - Besoin de compétences techniques souvent absentes des groupes de recherche
  - Infrastructure : développer/adopter de nouveaux outils
    - Optimisation des ressources (duplication de la donnée, utilisation du réseau, scripts performants,…)
    - Collaboration scientifique (gestion de projet, échange de ressources, publication,…)
    - Gestion et analyse de données (métadonnées, partage, workflows, DMP…)
  - Infrastructure : développer/adopter de nouveaux usages
    - Centralisation / mutualisation
    - Mais **Proximité** (géographique et thématique)
- « People Paradox »
  - L'automatisation provoque des besoins croissants en ressources humaines
- Gestion des ressources humaines
  - Evolution des métiers
  - Nouvelles compétences

Bascule des coûts : facteur déstabilisant





Using Clouds for Metagenomics: A Case Study Wilkening et al.
IEEE cluster 2009

La **mutualisation des appareillages** a été réalisée (ex: Biogenouest)
Il faut maintenant **mutualiser les moyens humains** sur le versant « **TIC** » et analyse des données

Une discipline en évolution

- Mutualisation des ressources plates-formes
- Bascule vers le monde digital centré sur les données
- Besoin d'interdisciplinarité
  - problèmes devenant des thèmes de recherche TIC
    - Masse de données, intégration de données, etc.

L'APPROCHE E-SCIENCE

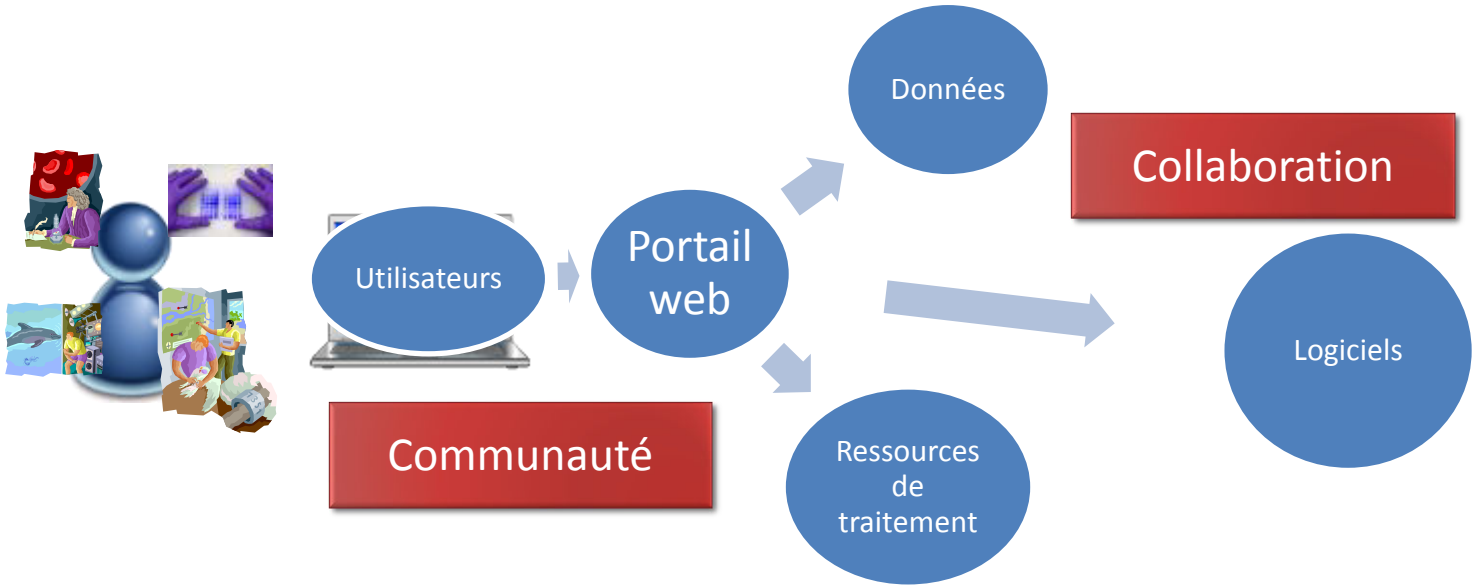
e-Science

- e-Science = Science + TIC
- « *Research done through distributed global collaboration enabled by the internet, using very large data collections, terascale computing resources and high performance visualization*» (Sir John Taylor - 2001)
- e-Science : trois volets
  - Calcul: grille, cloud
  - Stockage
  - Outils collaboratifs
- Pas d'initiative française purement e-Science

Caractéristiques e-Science

- Caractère distribué, pas de point central…
- … n'empêche pas un guichet unique!

Environnement de Recherche Virtuel



LE VRE

Un environnement virtuel de recherche en sciences de la vie

Un début de structuration e-Science?

Invite partners

Have an idea & create a group

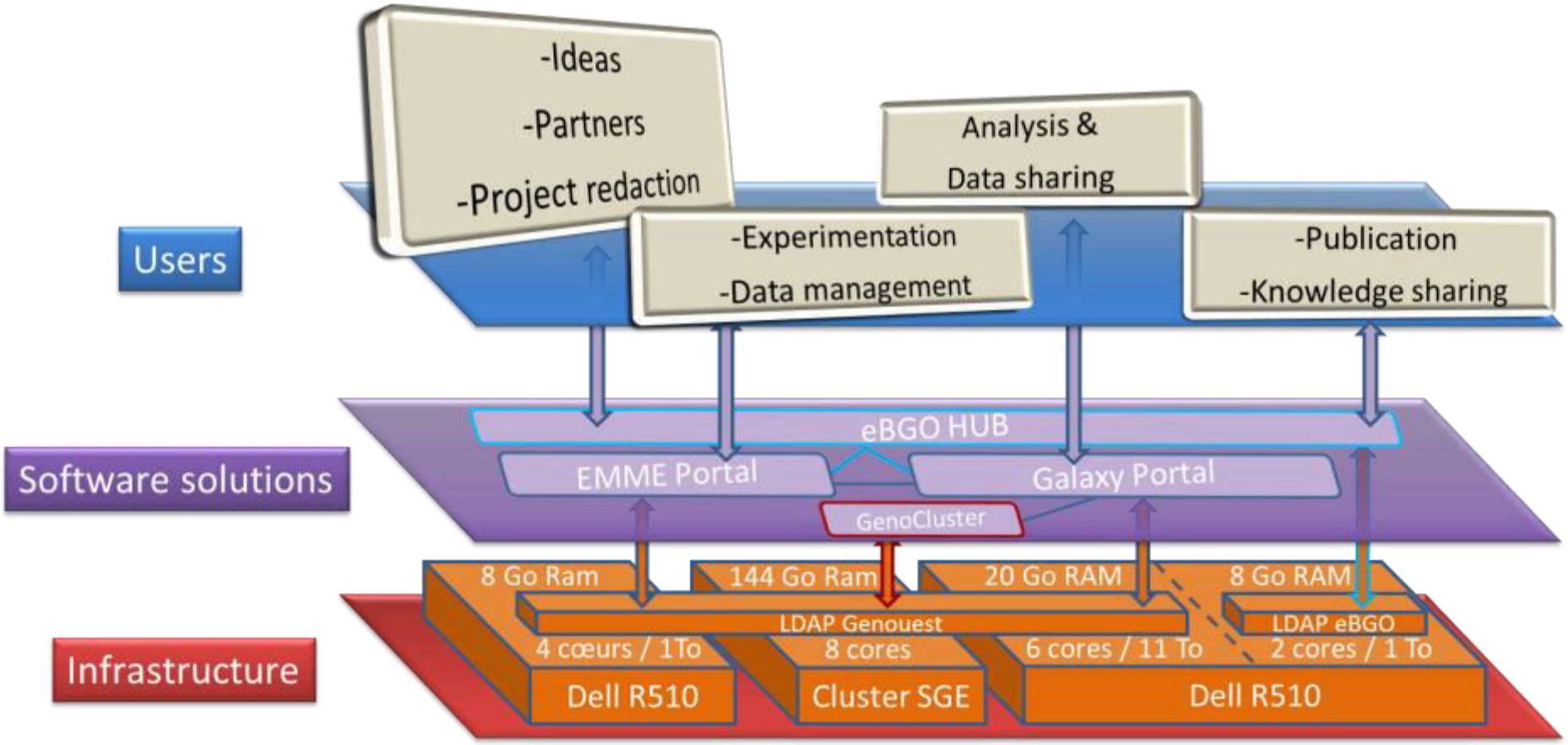
Training

record your data/metadata

Write & manage a new Project

analyse & share your data/metadata

Un début de structuration e-Science?



I have a dream…

- Pour les biologistes
  - ↘ perte de temps à apprendre à coder
  - ↘ compréhension des outils, leurs paramètres,
  - conserver l'historique des traitements appliqués aux données
  - accéder, partager et visualiser les jeux de données
  - avoir accès à un système de gestion de projet simplifiée…
- Pour le développement d'outils
  - ↘ perte de temps pour le graphique
  - faciliter l'utilisation des outils: Bioinfo **Recherche** → **Service**
  - ↘ temps entre développement, test et production….
- Pour la gestion des infrastructures
  - optimiser stockage, calcul et utilisation du réseau…
- Infrastructure **pour** la donnée → infra **de** données

Choisissez vous-même la suite…

Résultats du sondage

- Echanges sur les données, l'infrastructure et l'organisation sous forme de questions / sondages

- Présentation de l'organisation/structuration envisagée de la future initiative e-Science

ECHANGES SUR LES DONNEES

Echangeons autour de plusieurs questions

Données

- Durée de mise à disposition des données

- L'évolution technologique des appareillages scientifiques induit-elle une obsolescence des données antérieurement produites ?
  - Oui
  - Non

- Possédez vous une stratégie pour collecter et conserver vos métadonnées
  - Oui
  - Non

- Votre laboratoire/équipe a-t-il implémenté un plan de gestion des données ?
  - Oui
  - Non

Données

- Comment définissez vous vos données ?
  - Publiques (n'importe qui peut les consulter)
  - Ouverte (éutilisation par tous, sans restriction technique, juridique ou financière)
  - Sensibles (on ne préfère pas les diffuser)
  - Confidentielles (elles ne doivent pas être diffusées)
  - Propriétaires (leur spécification est contrôlée par une entité privée)

- Parmi les actions suivantes, identifiez-vous des besoins (aide, formation, développement de logicielles, ...)
  - Mise en place d'un plan de gestion des données
  - Sauvegarde des données
  - Partage des données
  - Restauration des données
  - Sécurité pour les données sensibles
  - Gestion des métadonnées
  - Maintenance des formats obsolètes
  - Numérisation de documents

Infrastructure

- Quelles ressources de calcul utilisez vous
  - Poste de travail
  - Cluster de laboratoire
  - Cluster mutualisé
  - Cloud
  - Grille de calcul

- Types de ressources
  - Machine standard (pc portable, pc de bureau)
  - Machine à forte capacité de mémoire (> 8 Go RAM)
  - Machine massivement paralllèle (> 8 CPU/coeurs)
  - Machine dédiée accès disques (rapidité d'accès ou grand espace de stockage)
  - Machine dédiée traitement image

- Environnement système
  - Windows
  - MacOS
  - Linux

Infrastructure

- Quels types de logiciels utilisez-vous ?
  - Surtout des logiciels propriétaires
  - Surtout des logiciels open source
  - Logiciels propriétaires et open source (~50/50)

- Durée de vie des logiciels utilisés
  - courte (< 6 mois (ex: scripts à façon))
  - moyenne (- de 6 mois à 5 ans (ex : logiciels rapidement obsolètes))
  - longue (5 - 10 ans (ex: logiciel de référence dans un domaine comme TMEV, CLUSTER, ...))
  - très longue (>10 ans (ex: logiciel basique et de référence comme BLAST, PHYLIP, EMBOSS, ...))

- Pensez-vous que votre connexion réseau limite votre activité ? (transfert de données, accès à des ressources de calcul distantes)
  - 1 = oui, tout à fait
  - 2 = non, pas du tout

Organisation et intérêts

- Rencontrez-vous actuellement des problèmes liés à l'organisation géographique de vos ressources ?
  - non
  - oui, problèmes de type politique (ouverture entre établissements de tutelles différentes, ...)
  - oui, problèmes de type technique (bande passante, sécurité, ...)

- Etes vous intéressé(e) par les services suivants ?
  - Mise à disposition de logiciels (via plateforme web par exemple)
  - Développement de logiciels
  - Développement de workflows
  - Accélération des programmes (parallélisation)
  - Développement web (site internet par exemple)
  - Développement de bases de données
  - Développement cloud (utiliser le cloud dans vos activités)
  - Développement grille de calcul (utiliser les grilles de calcul dans vos activités)
  - Service d'analyse de données

Organisation et intérêts

- Rencontrez-vous actuellement des problèmes liés à l'organisation géographique de vos ressources ?
  - non
  - oui, problèmes de type politique (ouverture entre établissements de tutelles différentes, ...)
  - oui, problèmes de type technique (bande passante, sécurité, ...)

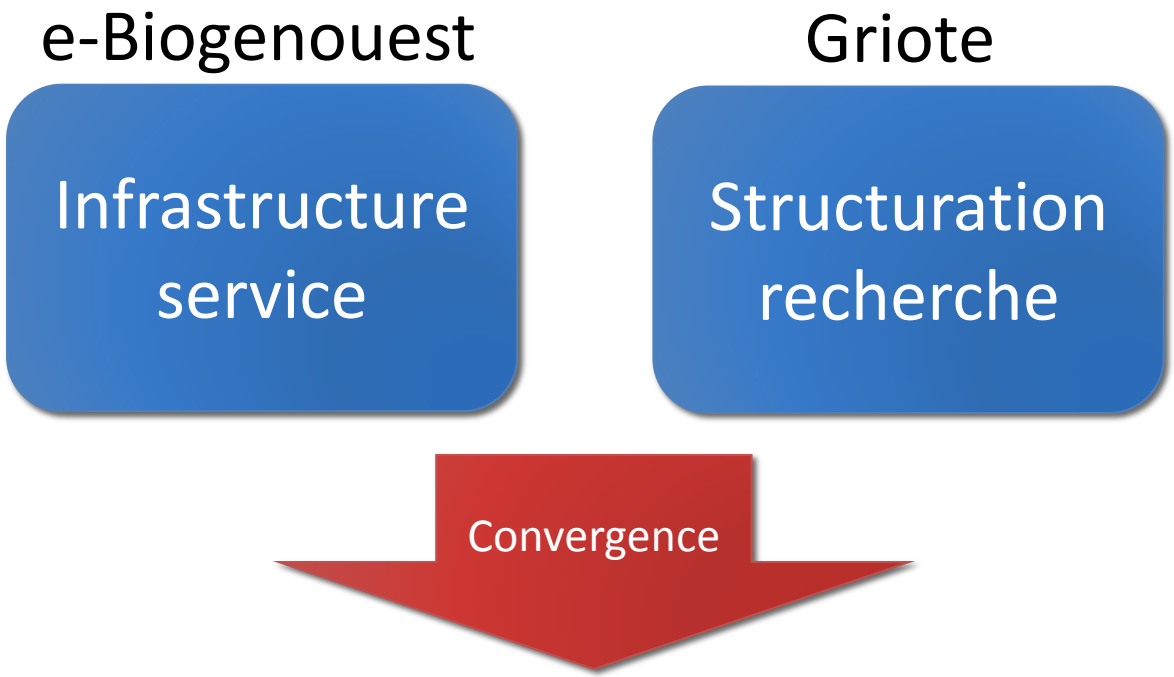
- Etes vous intéressé(e) par les services suivants ?
  - Mise à disposition de logiciels (via plateforme web par exemple)
  - Développement de logiciels
  - Développement de workflows
  - Accélération des programmes (parallélisation)
  - Développement web (site internet par exemple)
  - Développement de bases de données
  - Développement cloud (utiliser le cloud dans vos activités)
  - Développement grille de calcul (utiliser les grilles de calcul dans vos activités)
  - Service d'analyse de données

Aller à la conclusion

ORGANISATION

Une nouvelle plate-forme pour Biogenouest

Articulation Bretagne – Pays de la Loire



- Conclusions
  - Importance des **moyens humains** mutualisés
  - Lieu physique d'échange, nécessite **proximité**
  - I**nfrastructure** de calcul et de données
  - Importance de l'**intégration**

Perspective : Plateforme e-Science

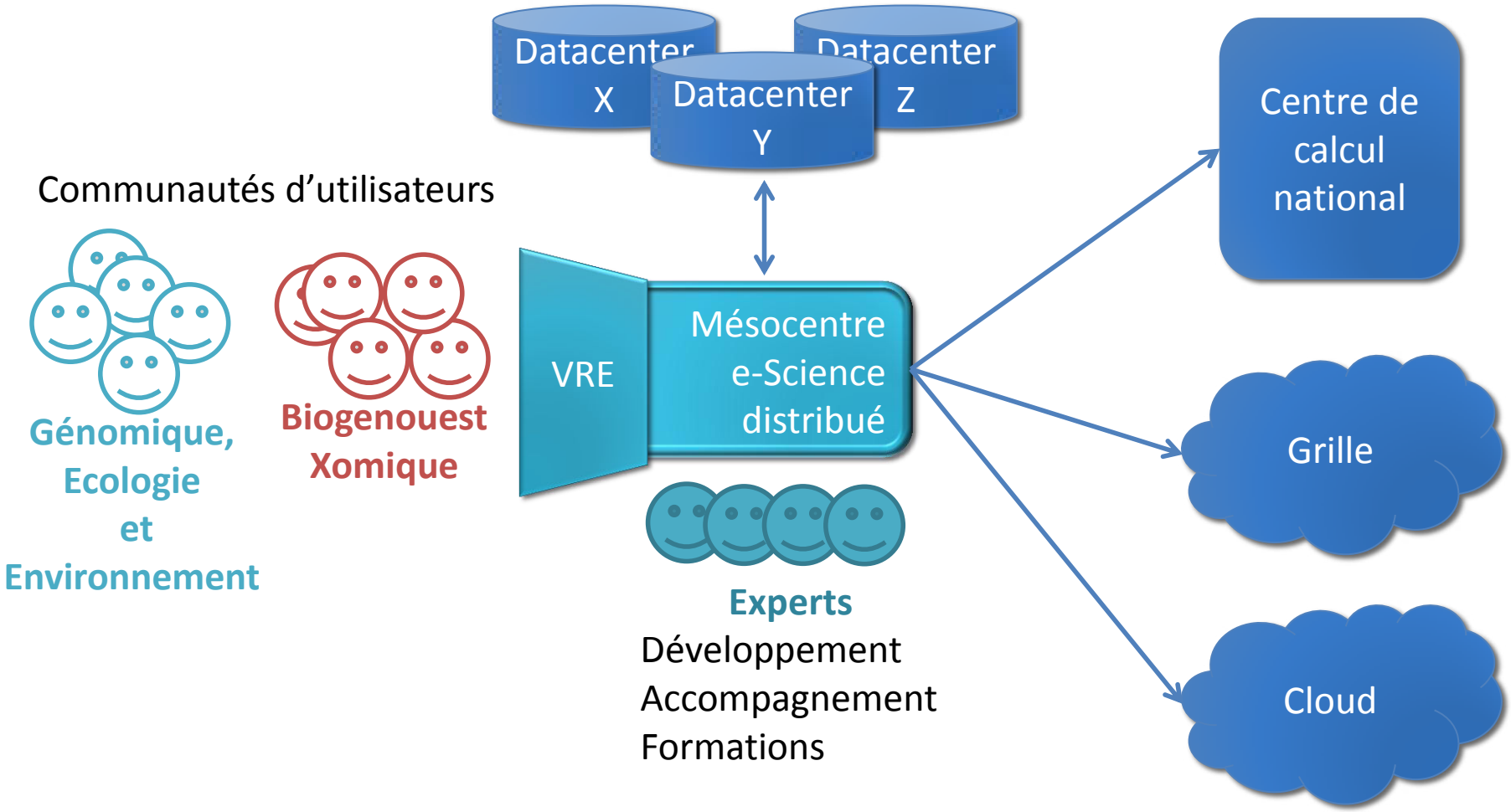
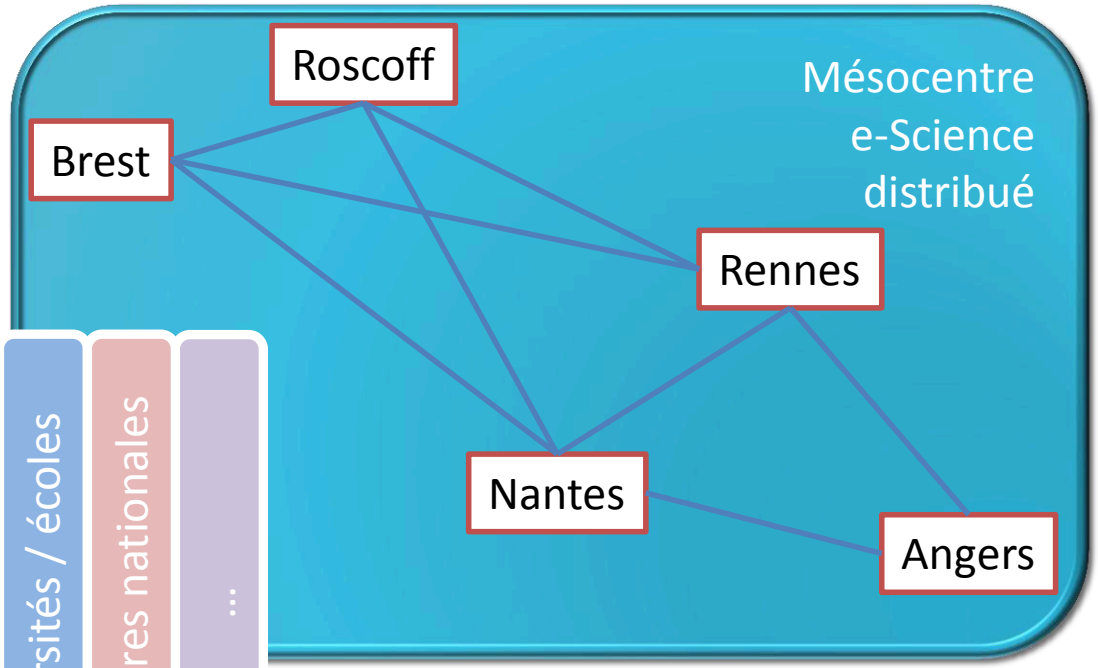
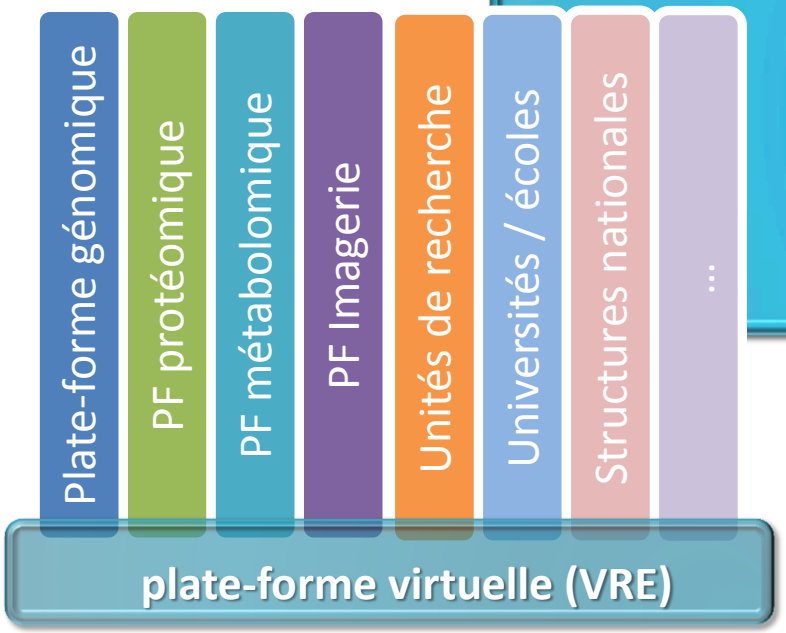


plate-forme e-Science : un réseau

Maillage du territoire Biogenouest

- Spécialisation thématique
- Infrastructure de données à l'échelle du territoire
- Proximité des plates-formes traditionnelles



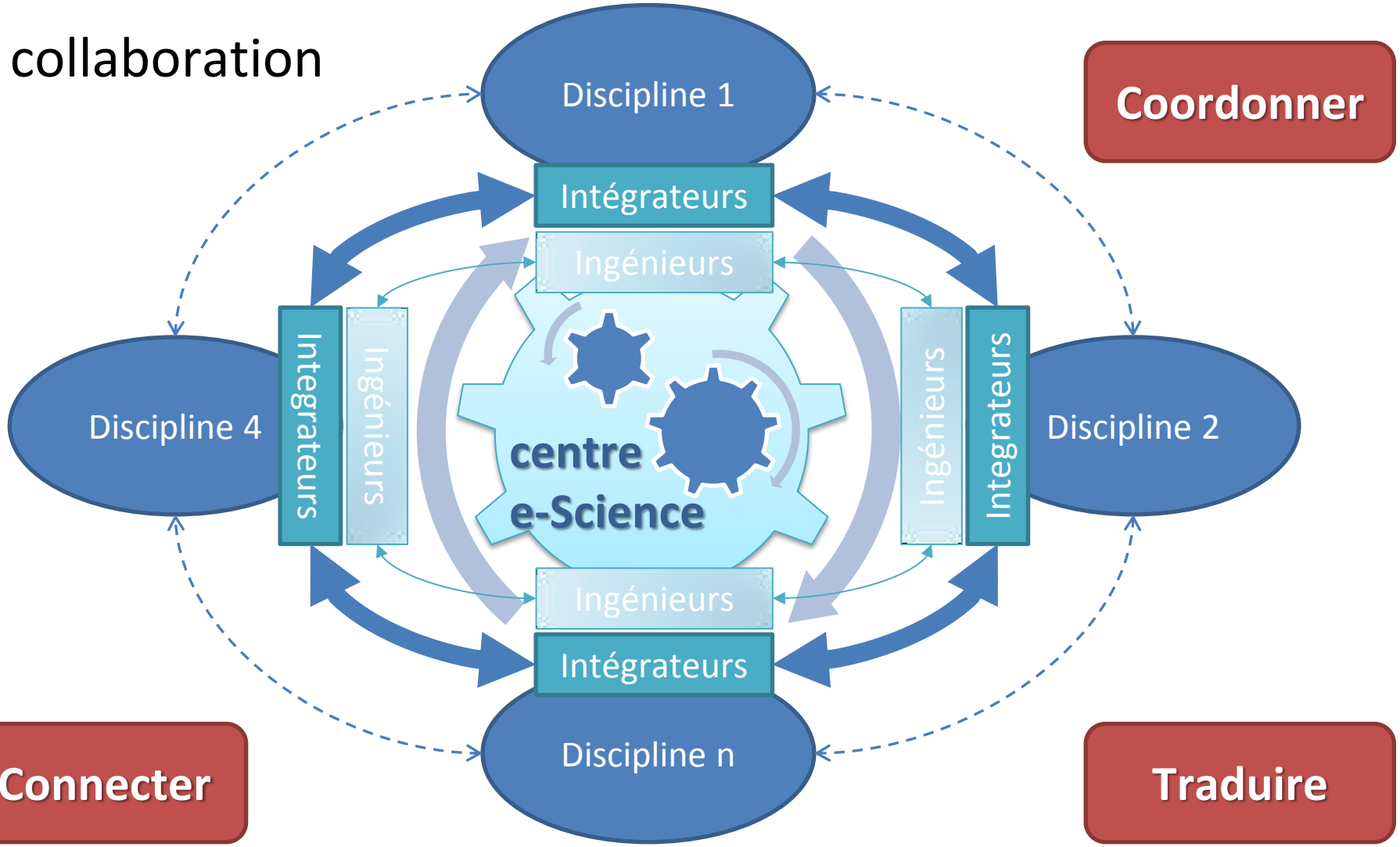


Biogenouest

e-Biogenouest

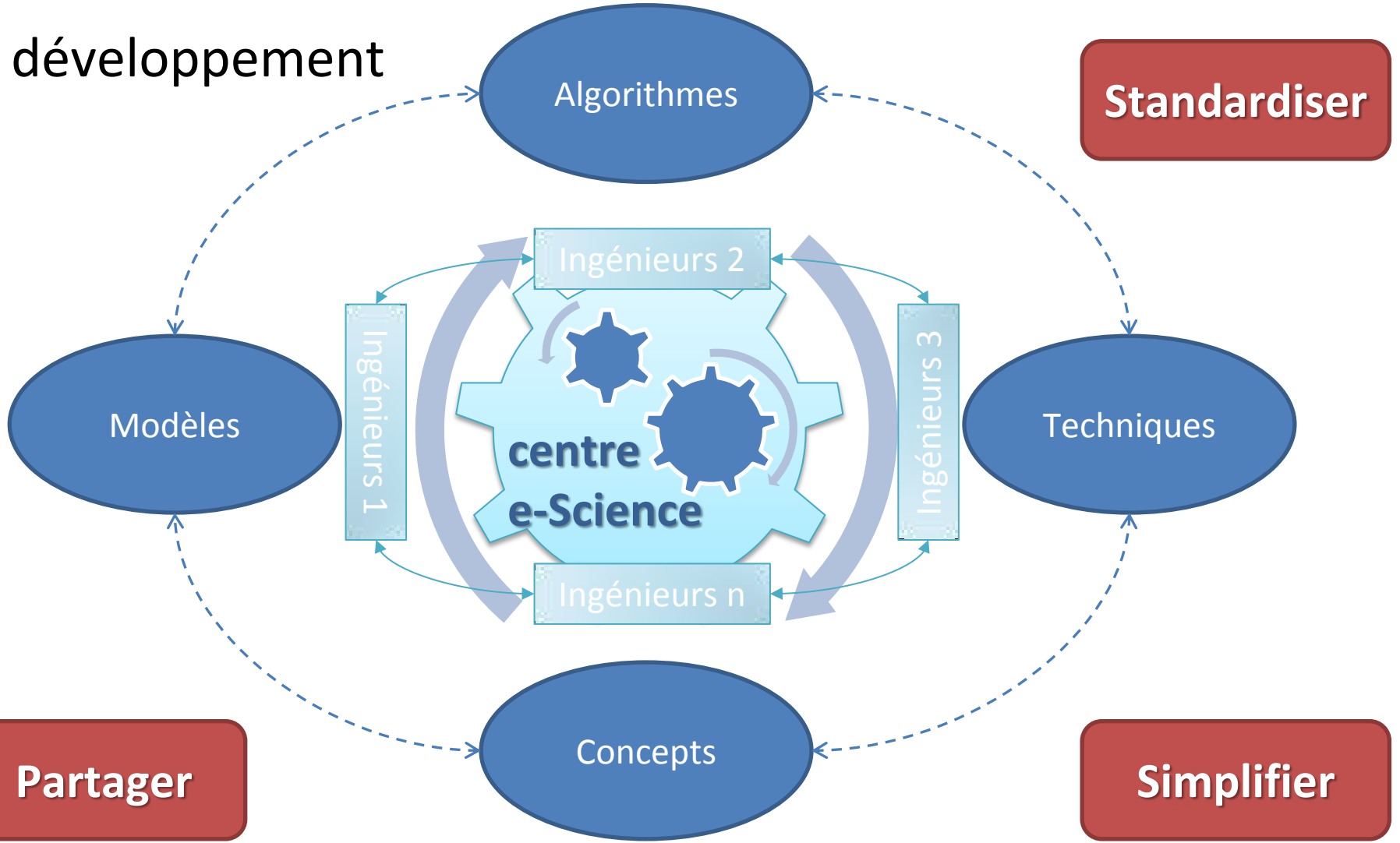
Plateforme e-Science

La collaboration



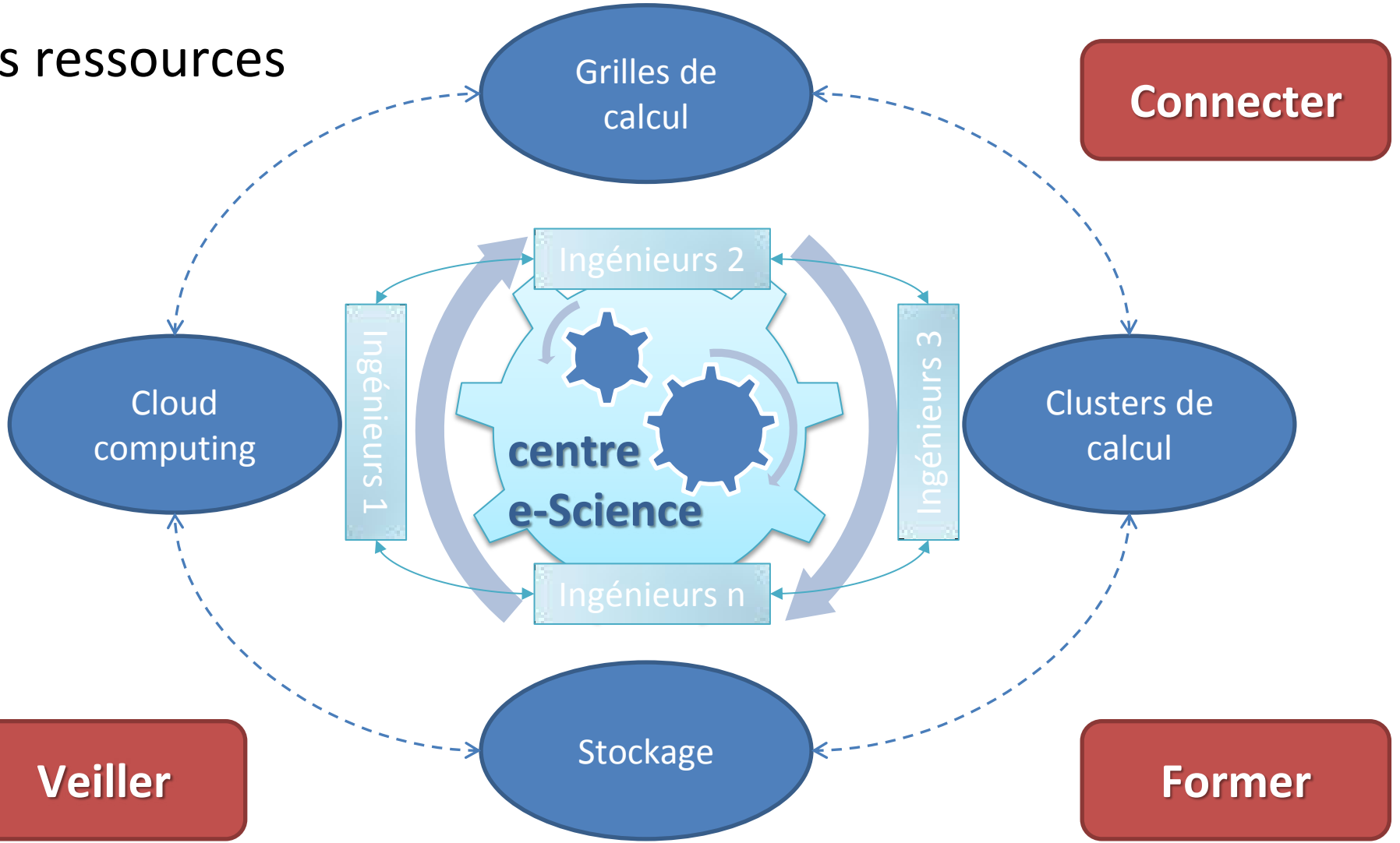
Plateforme e-Science

Le développement



Plateforme e-Science

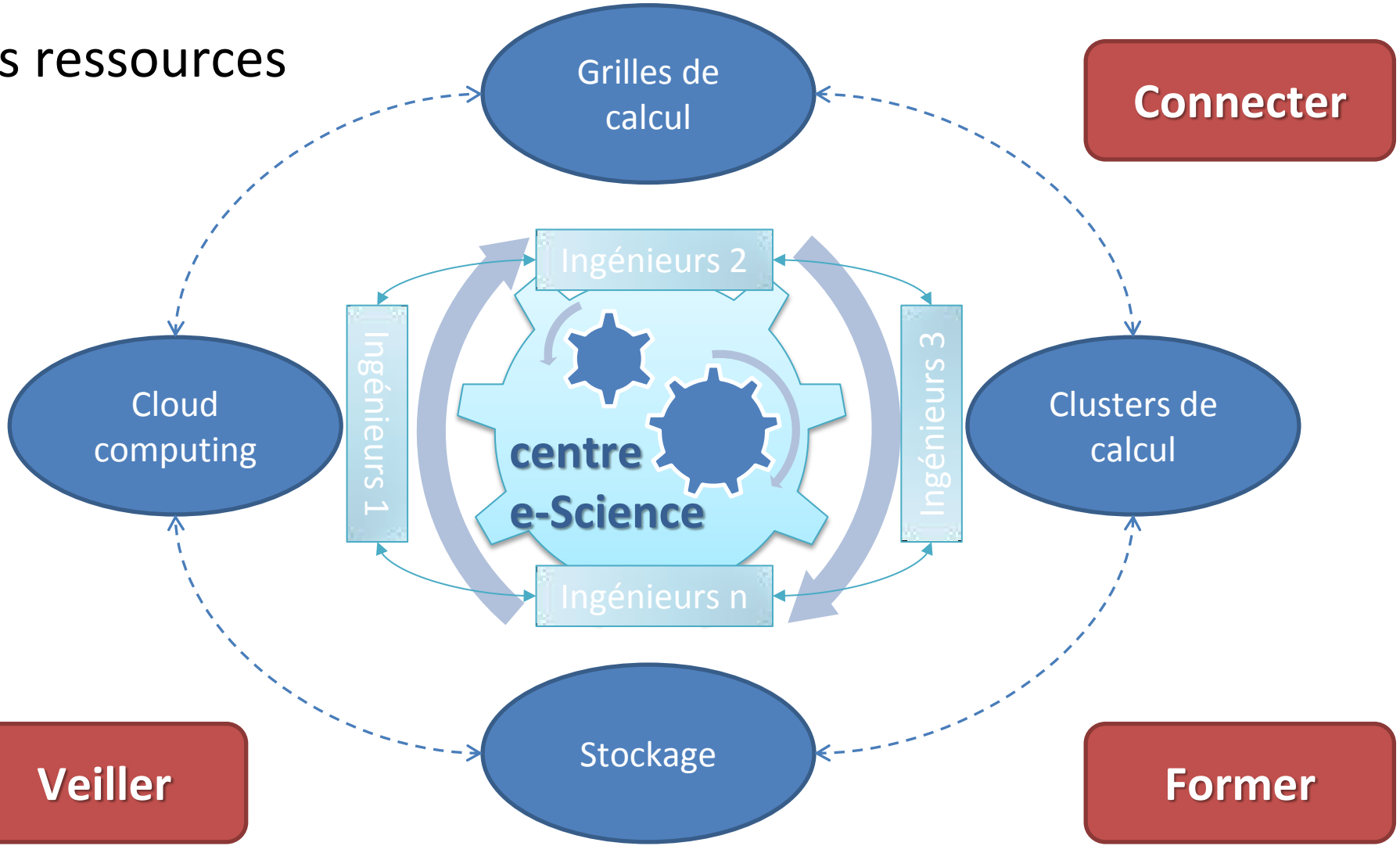
Les ressources



Plateforme e-Science

Aller à la conclusion

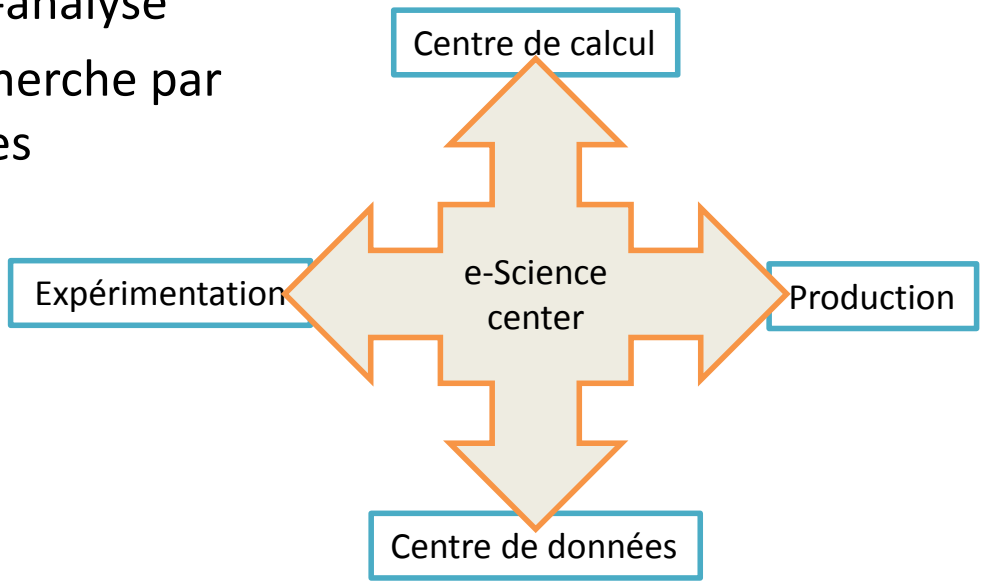
Les ressources



CONCLUSION

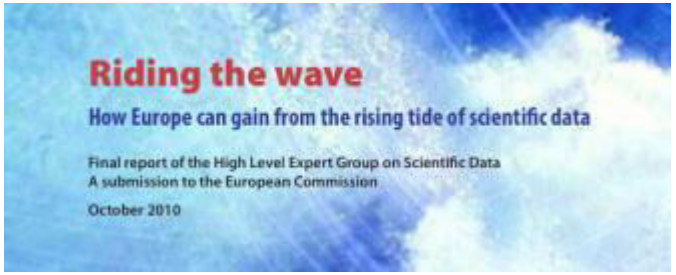
Infrastructure de / pour données

- Gestion : curation, stockage, récupération
- Bénéfices :
  - Données non « perdues »
  - Pas de doublonnage de la recherche
  - Qualité de la recherche grâce à la ré-analyse
  - Accroissement de la valeur de la recherche par combinaison/intégration des données
  - Permettre de nouvelles recherches
- C'est ce qui fera la différence



Enjeux

- Avantage compétitif pour la recherche et les réponses aux appels d'offre **Projets dimensionnants**
- Se structurer pour pouvoir répondre aux appels d'offres à venir
  - Roadmap à l'horizon 2020
- Préserver les données produites par les groupes de recherche
- Optimiser les infrastructures



"A fundamental characteristic of our age is the rising tide of data – global, diverse, valuable and complex. In the realm of science, this is both an opportunity and a challenge."

Riding the Wave report, High-Level Group on Data

Merci de votre attention

La plate-forme Bio-informatique GenOuest

Le groupe Symbiose IRISA/INRIA

GenOuest-Dyliss-Genscale

Cyril Monjeaud

Olivier Collin

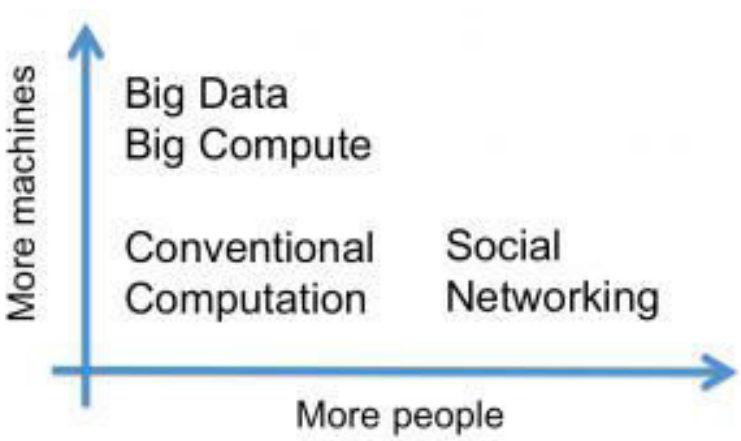
eBGO HUB *(collaboration)* ***http://www.e-biogenouest.org/***

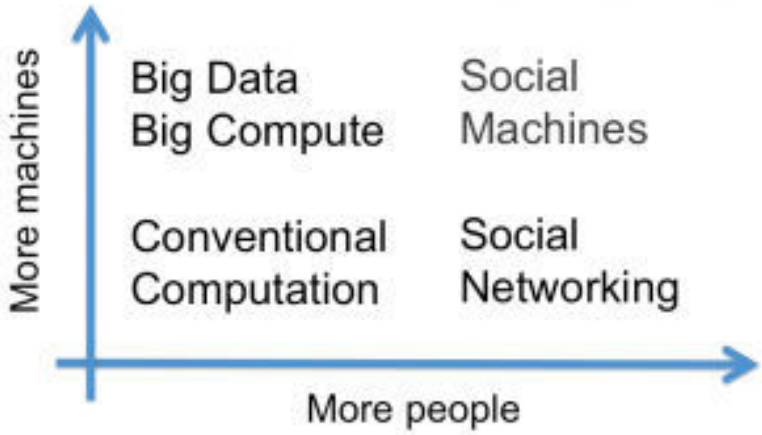
EMME portal *(data management)* ***http://emme.genouest.org/***

Galaxy instance *(data analysis)* ***http://galaxy.genouest.org/***

GO4Bioinformatics *(education)* ***http://go4bioinformatics.genouest.org/***

Prospective





"processes in which the people do the creative work and the machine does the administration"
Tim Berners-Lee

http://www.scilogs.com/eresearch/social-machines/