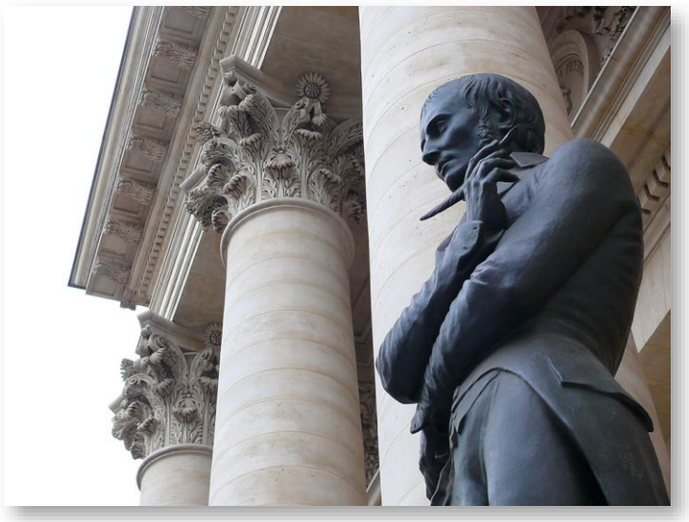


Les impacts du Big Data à l’université

Université Paris Descartes
“L’Université de l’Homme et de la Santé”

4 octobre 2013

Les données massives à l'université

Big data, pour quoi faire ?

- La recherche
- L'enseignement
- Le pilotage

Les données massives

- Explosion du volume des données numérique

Google = 26 Po / jour dès 2008

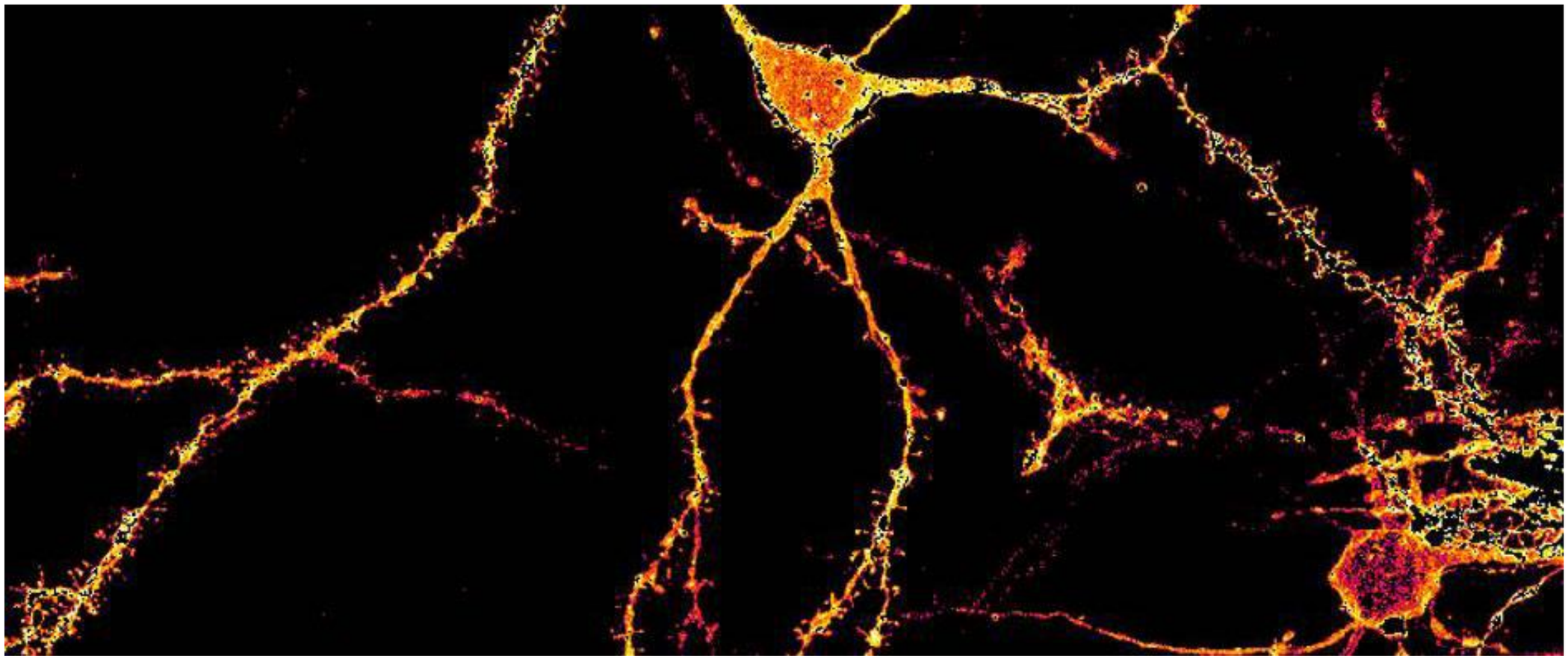
= 625 bibliothèques de 15 000 livres chacune…

…par heure !

Les données massives

Une lettre : 1 octet	1
Une page A4 : 1,5 kilo-octets	1500
Un livre de 200 pages : 300 kilo-octets	300 000
Un film d’une heure de cours : 1 giga-octet	1 000 000 000
L’ensemble des données de Descartes : 200 To	200 000 000 000 000
Les données quotidiennes de Google : 26 Po	26 000 000 000 000 000
Les données gérables par la NSA : 10 Zo	10 000 000 000 000 000 000

Recherche et données massives



Source : phototheque.parisdescartes.fr

La recherche : les enjeux

- Traiter des échantillons de données très massifs
- Résoudre des problèmes impossibles
- Réduire les coûts en partageant les données
- Rapprocher le grand public de la recherche

Crédit photo : CERN

La recherche : Solar StormWatch, un exemple de Crowdsourcing

- Analyser les tempêtes solaires
- Apprendre à les prévoir

- 100 000 images en deux ans
- 25 Téra-octets
- Impossible à analyser automatiquement

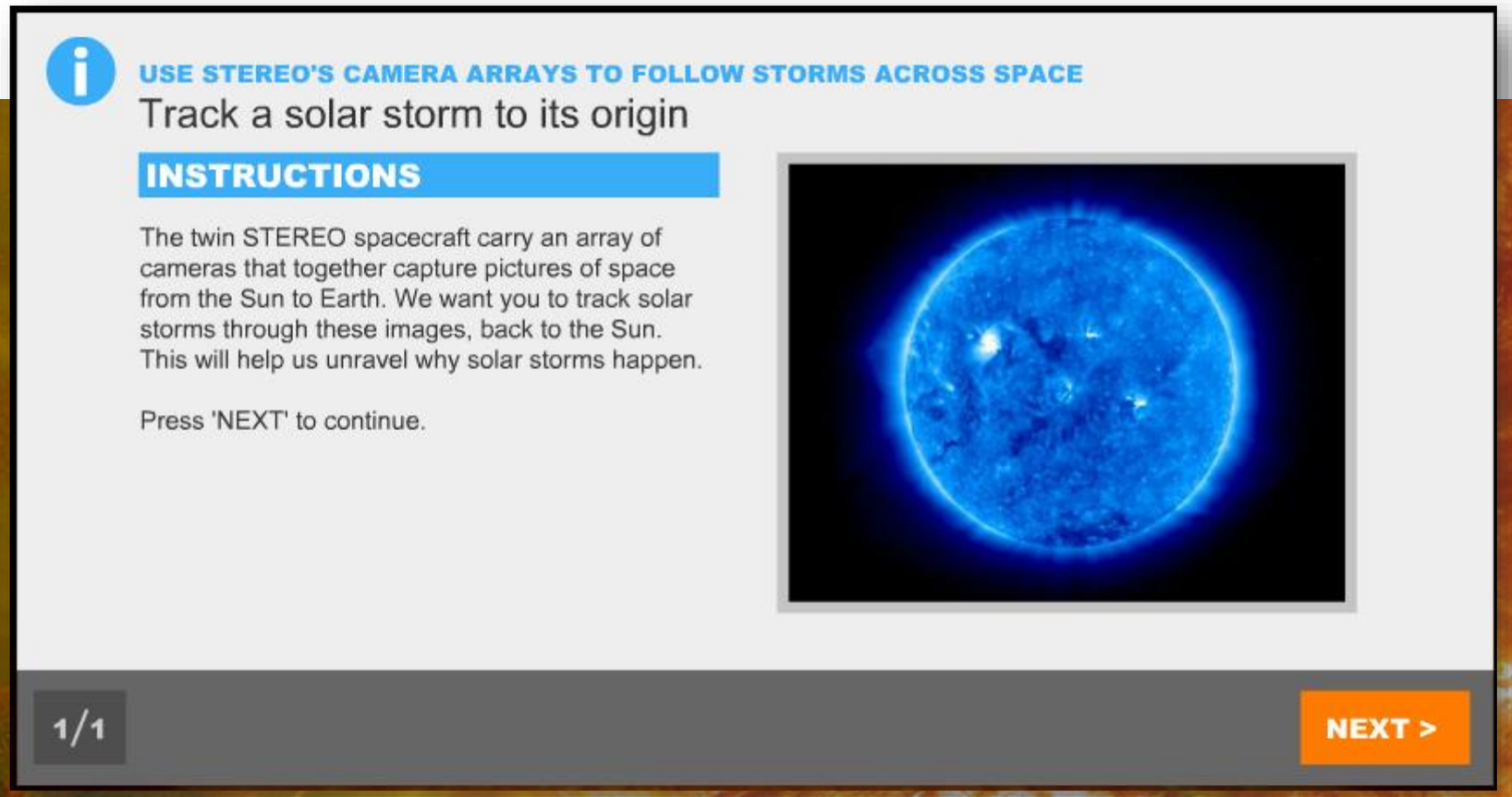
La recherche : Solar StormWatch

- 868 962 participants bénévoles
- Apprentissage humain
- Redondance des analyses

La recherche : Solar StormWatch

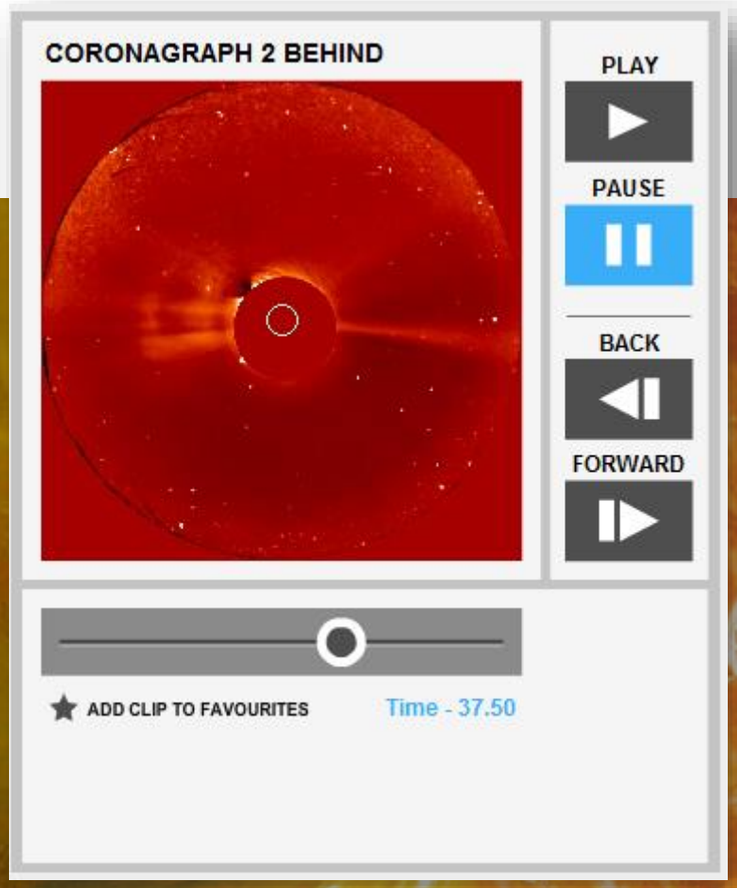
Comment ça marche ?

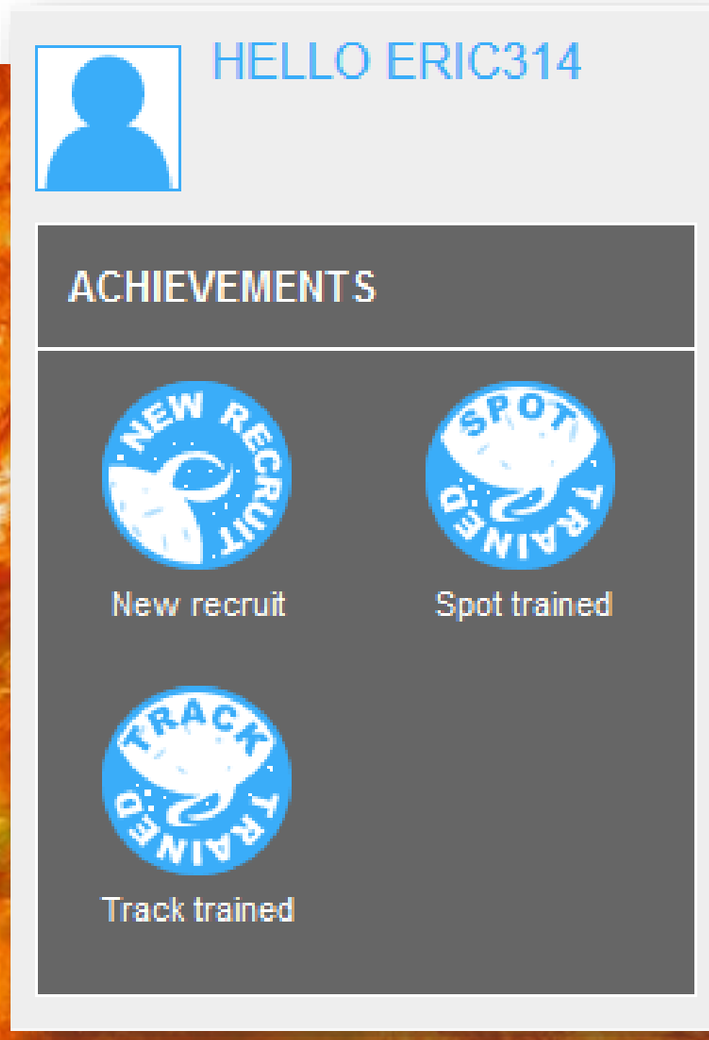
La recherche : Solar StormWatch



1. On s’entraine

La recherche : Solar StormWatch





2. On analyse

3. On gagne des récompenses

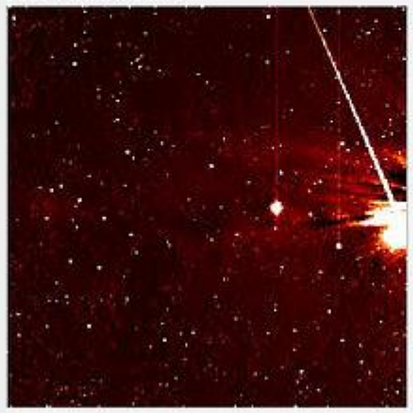
La recherche : Solar StormWatch

jules
Global Moderator
Hero Member

Posts: 1307

Re: Particle Strike Collection
« **Reply #68 on:** March 06, 2010, 10:54:26 pm »

http://solarstormwatch.com/favourites/764
@12.90
STEREO AHEAD



Classified it as a particle strike. Not seen a single one this bright before.

ChrisDavis
Science Team
Sr. Member

Posts: 325

Re: Particle Strike Collection
« **Reply #73 on:** March 07, 2010, 09:07:33 am »

Nice one!

Yes, I'd say that was a particle. On its own, it may have just been drifting past the camera. Amazing stuff.

Chris.

Logged

4. On échange

Photo by NASA

La recherche : OldWeather

Enter weather values

forum guides show help close

Date | Location | Weather Observation | Animals | Refueling | Mentions | Sea Ice | Events

Hour	Wind Dir	Force	Bar Height	Ther Attached	Dry	Wet	Water	Weather Code	Cloud Code	Clear Sky
--	--	--	--	--	--	--	--	--	--	--

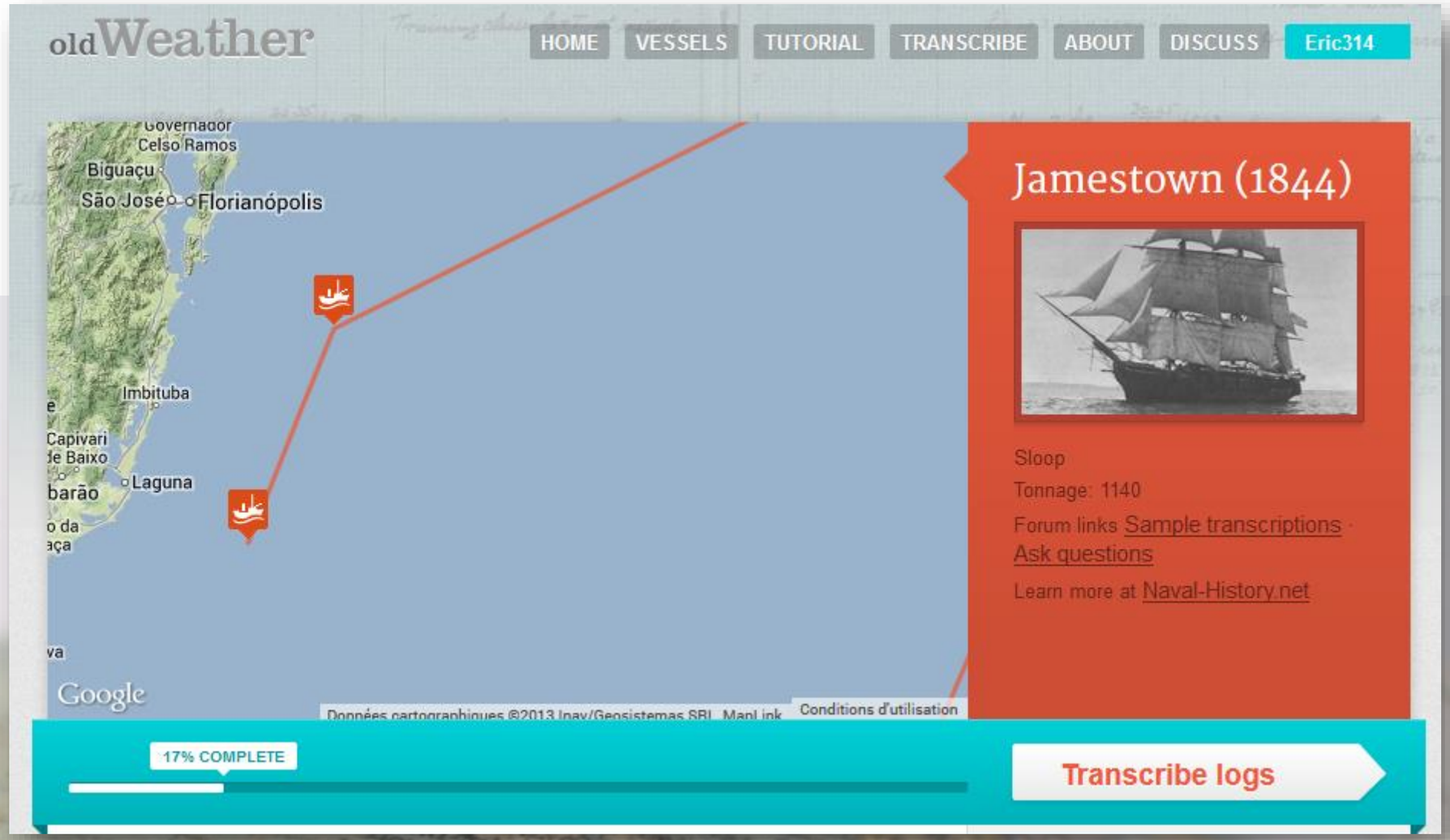
OK

LOG of the United States F. C. St. Albaiosa , Rate,

Direction by Standard Compass.	Force.	Heel.	Leeway.	Height in inches.	Ther. att'd.	Air, Dry Bulb.	Air, Wet Bulb.	Water at Surface.	State of the Weather, by symbols.	Forms of Clouds, by symbols.	Prop. of Cl. Sky, in 10ths.	State of the Sea.
Calm	0			30 28		61	57	57	Clear Sun Set		2	

Water expended during the preceding 24 hours, 10 Gallons.
Water during the preceding 24 hours, -
Water remaining on hand fit for use at Noon, 630 -
Coal consumed during the preceding 24 hours, tons 1570 lbs.
Coal remaining on hand at Noon, 21 - 610 -

La recherche : OldWeather



Données de route

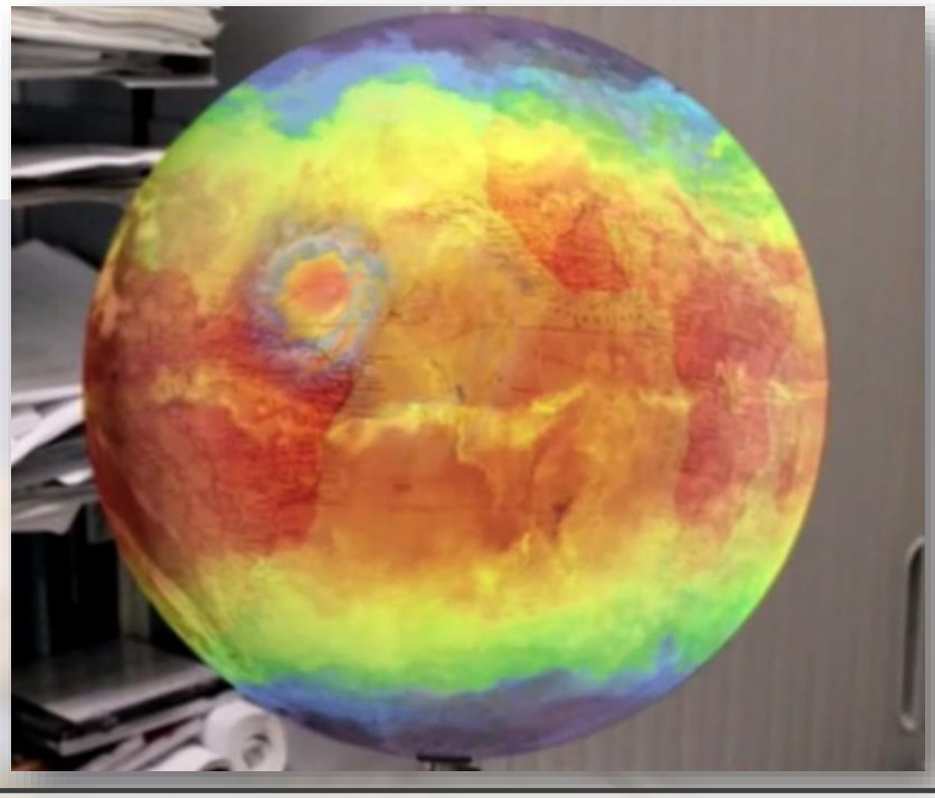
La recherche : OldWeather

Jeannette's crew

The crew of this ship for the current voyage

gastcra Captain 12228			
Clewi Lieutenant 11992		**jil** Lieutenant 10499	
okopho Lieutenant 2880		**Gunney** Lieutenant 775	

Récompenses



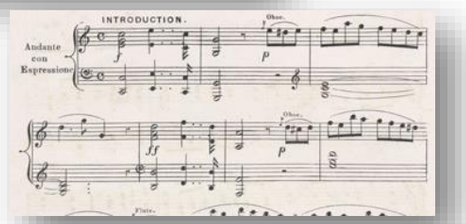
Modèle climatique

La recherche : OldWeather

- Données très riches : météo, économiques…
- Traitement manuel
- Construction de modèles climatiques fiables sur plusieurs siècles !
- Données publiques accessibles à tous

La recherche : crowdsourcing

Galaxy Zoo



What's the score?

Solar StormWatch

WhaleFM

OldWeather

AncientLives

Intérêt du crowdsourcing

- Traiter des données massives
- Le cerveau humain est plus efficace que les ordinateurs
- Améliorer la qualité des résultats
- Engager le grand public dans la recherche
- Utiliser des données de multiples fois à des fins non prévues (openData)

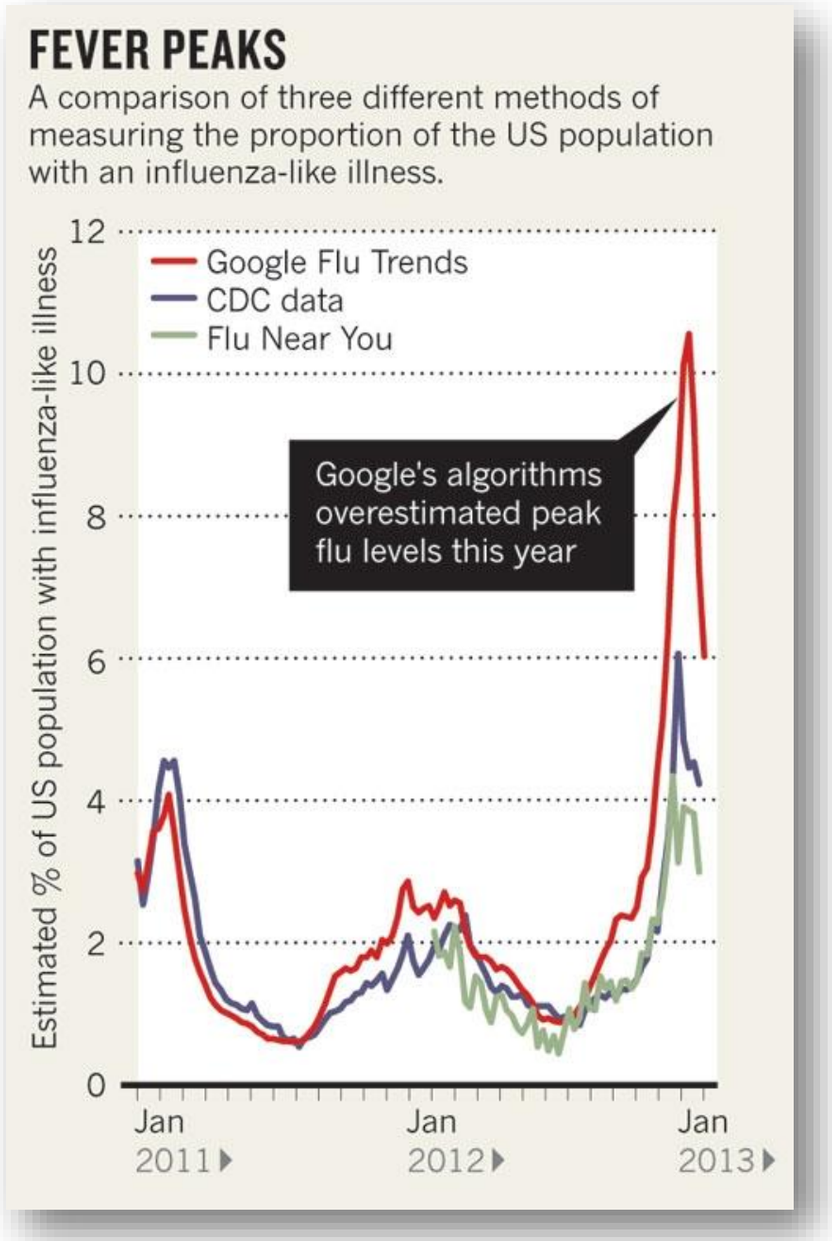
La recherche : le cas de Google trends

- Analyse des épidémies par la fréquence des requêtes sur des mots-clés dans Google

Données massives : Google Flu Trends

- Source de données scientifique nouvelle et majeure, en santé publique, en sciences sociales…
- **Mais…**
- Résultats fragiles :
  - Biais d'échantillonnage : jeunes
  - Corrélations imprévues : symptômes grippaux plutôt que grippe

Big data : analyse riche mais nouveaux risques



Sources: Google Flu Trends; CDC; Flu Near You

Données massives : Google Flu Trends

- **Croiser les sources de données est essentiel :**
  - Associations de malades
  - Data brokers (comme IMS ou Celtipharm)
  - Caisses d'assurance maladie, les HMO aux USA

> "*En plus des systèmes classiques de surveillance nord-américains, je cherche toujours à voir [*à travers Google Flu Trends*] ce qui se passe et si nous sommes en train de passer à côté de quelque chose, ou si un signal semble porter une information différente de celles de nos systèmes, une information potentiellement utile*".

Dr Finelli, responsable du programme grippe
des *Centers for Disease Control and Prevention* à Atlanta

Données massives en santé publique

- **Production massive de données géolocalisées**
  - Echographies cardiaques avec le smartphone des patients
  - Nanocapteurs pour la surveillance de la glycémie des diabétiques
  - Les personnes à risques d'infarctus du myocarde auront des systèmes implantables assurant une surveillance permanente et une alerte précoce d'infarctus en temps-réel.

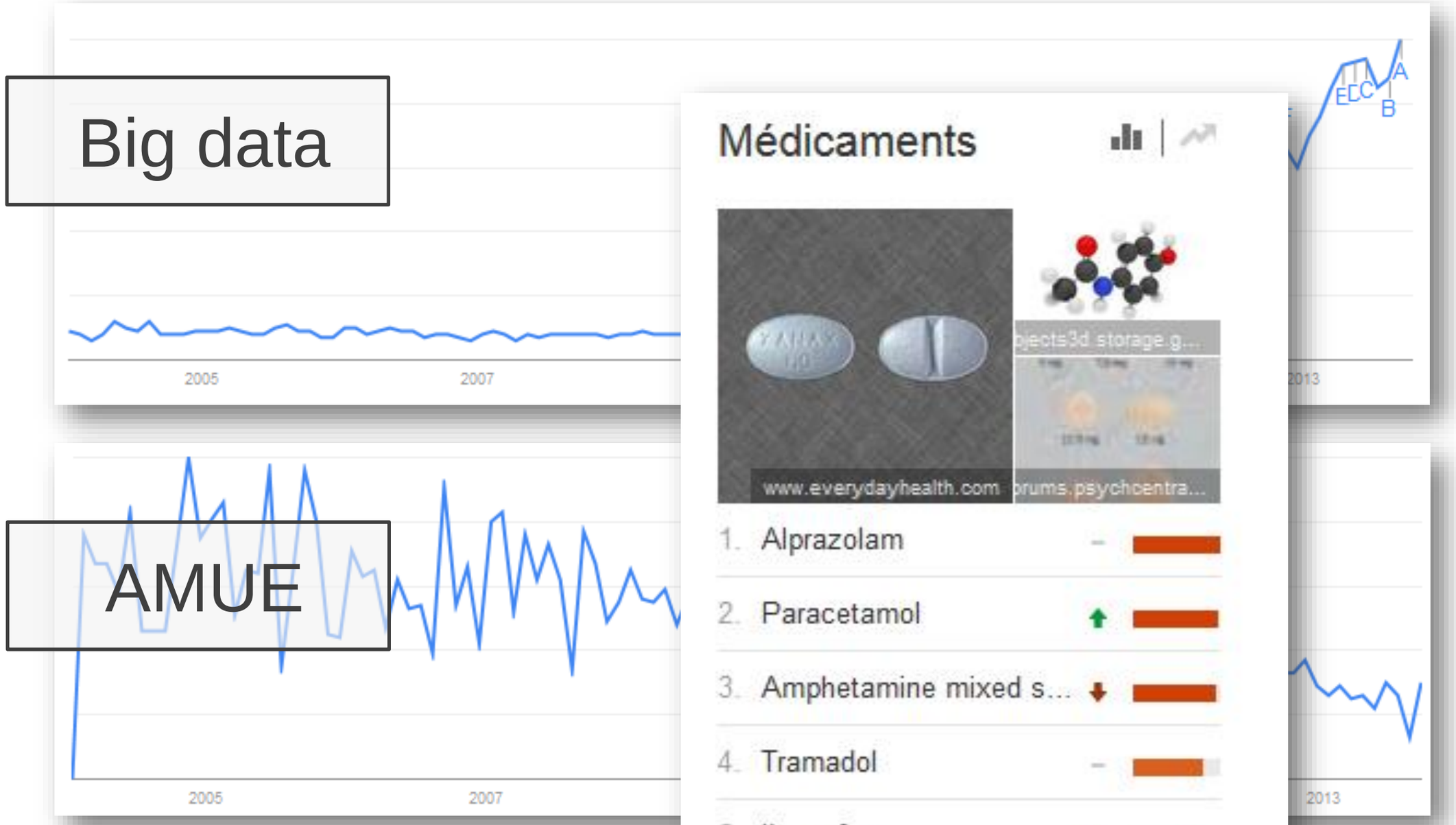
Le mouvement « d'analyse personnelle » va potentiellement révolutionner la santé publique et impacter fortement la recherche en associant le grand public à la production massive de données scientifiques.

Sources : E. Topol, cardiologiste à San Diego ; A. Flahaut, professeur à Paris Descartes

La recherche : le cas de Google trends

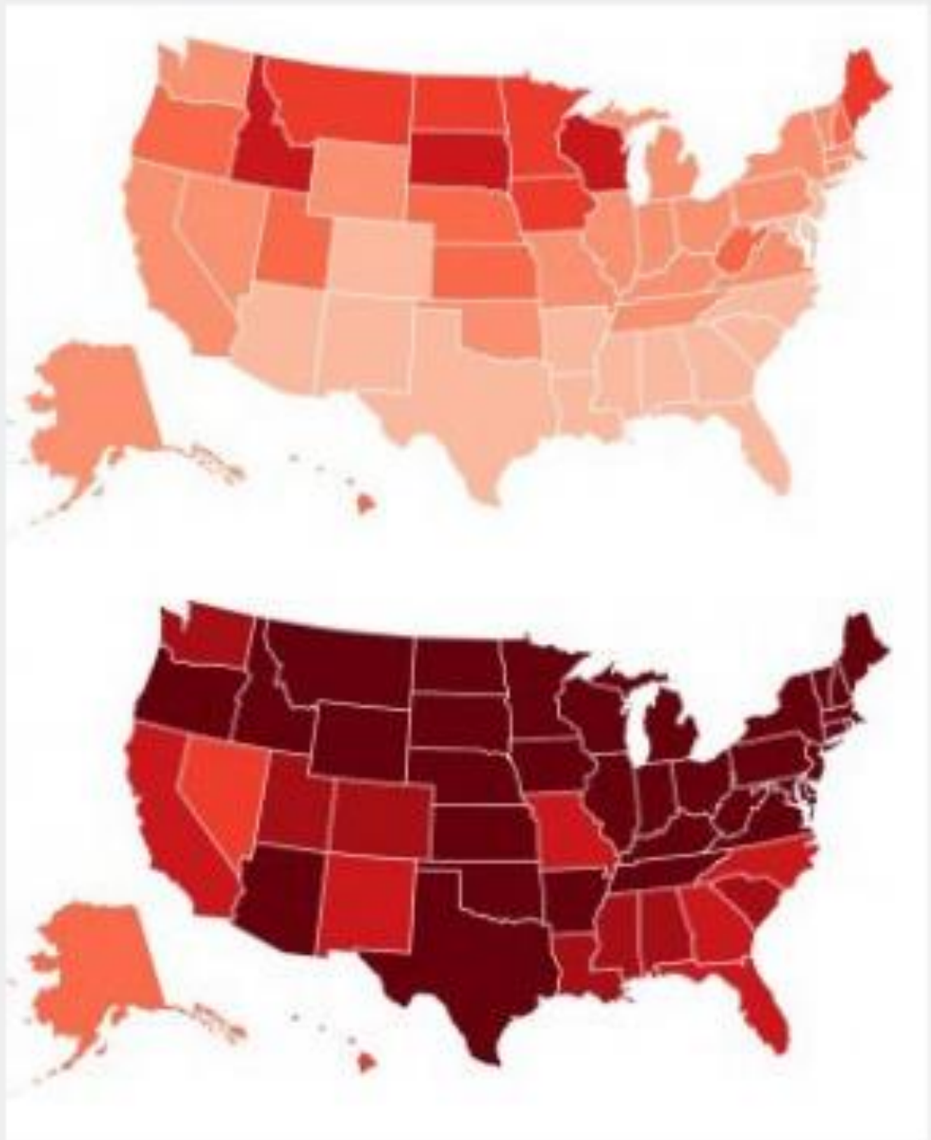
Big data

AMUE



Sources: Google Trends

Santé publique : Twitter

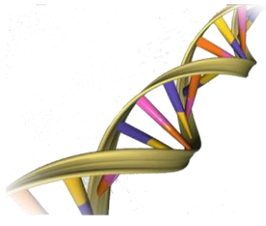


These U.S. maps indicate the Twitter system's rate of influenza in each state in the first week of January (higher flu rates are marked with darker red). They show that the country is awash in a high flu rate in 2013 (the bottom map), yet was relatively unscathed during the same week in 2012 (the top map).

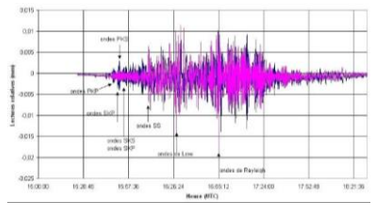
Sources: Health, Science+Technology

La recherche et les données massives

Imagerie médicale



Génomique



Sismologie

Zoologie

La recherche et les données massives

L'imagerie médicale

- Odontologie
- Radiographie
- Anatomie

- Enjeu : réduction des coûts d'acquisition et de stockage
- Usages pédagogiques et scientifiques

Source : photothèque Paris Descartes - Jean-Sébastien Leblond-Duniach

La recherche et les données massives

Génomique

- Enjeu : réduction des coûts de stockage et de traitement

Séquençage accéléré du génome :

- En 2007 : 3 milliards d'euros, 15 ans
- En 2014 : 400€, 3 jours

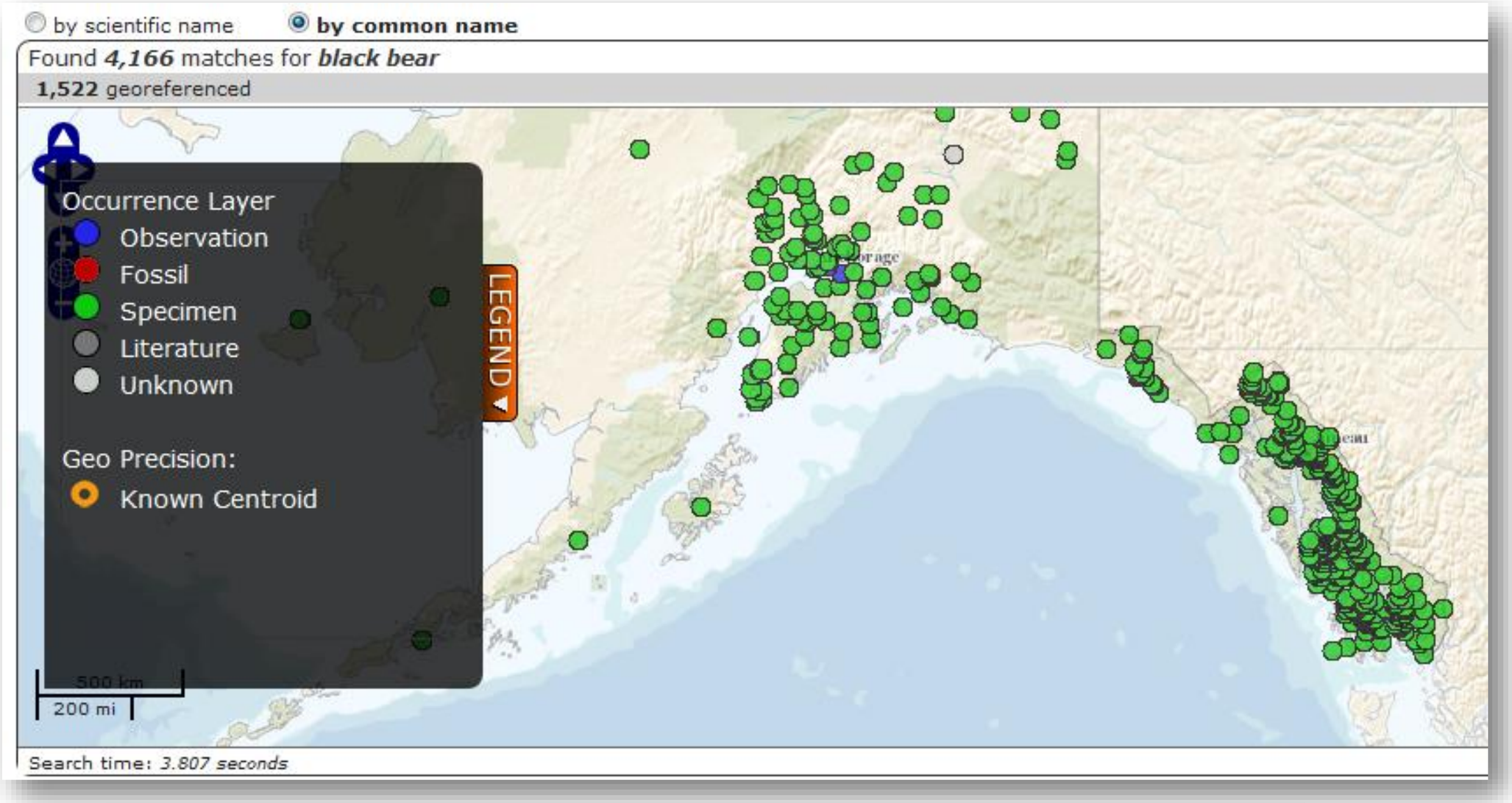
- Un génome = 6 milliards de nucléotides
- Environ 2 To par mois de données générées par un laboratoire à Cochin

Source photo : www. mentalactif.com

La recherche et les données massives

Zoologie

- Enjeu : récupérer des données détaillées à l'aide du grand public



Projet BISON (Biodiversity Information Serving Our Nation)

Source photo : http://bison.usgs.ornl.gov/

Enseignement et données massives

Source : phototheque.parisdescartes.fr

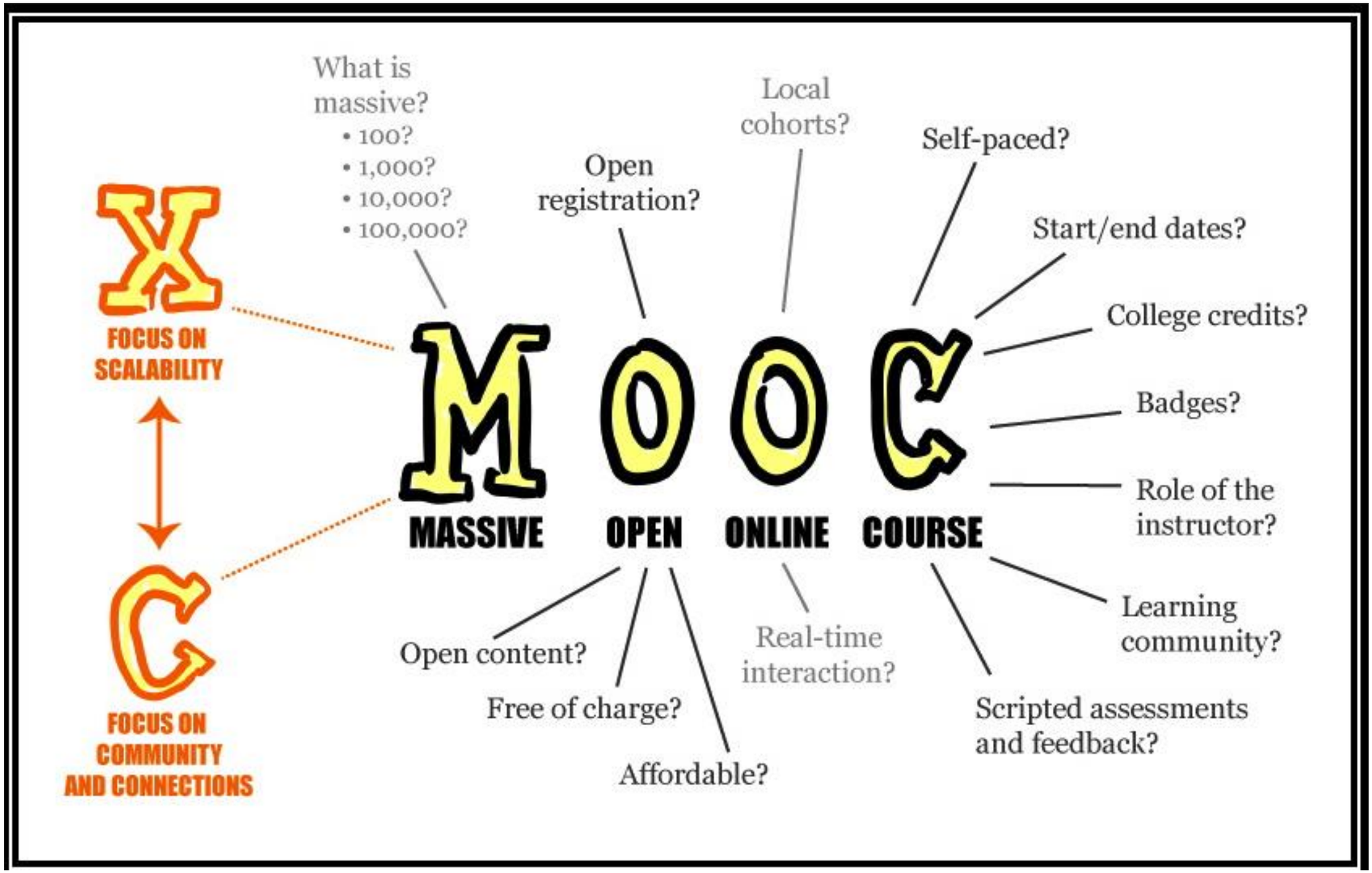
L’enseignement

- Les plate-formes d’enseignement produisent déjà des données massives

- L’arrivée des MOOC va amplifier ce phénomène

Coursera affiche déjà 3,5 millions d’inscrits et 400 cours dispensés par plus de 70 universités

Qu’est ce que le MOOC ?



MOOC ne veut pas dire « à la demande »!

Duke University

Healthcare Innovation and Entrepreneurship

with Marilyn M. Lombardi & Bob Barnes

Apr 15th 2013
6 weeks long
Signature Track

University of Michigan

Internet History, Technology, and Security

with Charles Severance

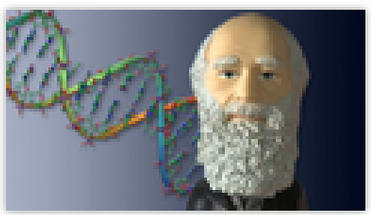
Jun 3rd 2013
11 weeks long
Signature Track

University of Michigan

Model Thinking

with Scott E. Page

Jun 3rd 2013
10 weeks long
Signature Track



Duke University

Introduction to Genetics and Evolution

with Mohamed Noor

Jan 3rd 2014
12 weeks long
Signature Track

Cas du centre Virchow-Villermé

- Une plateforme d'enseignement massif à distance, gratuit et entièrement dédié à la santé publique et la santé globale.
- Une collaboration technique avec l'INRIA
- Un couplage enseignement / recherche

http://virchowvillerme.eu/

Big data et enseignement

- Les universités couplent leurs MOOC avec une recherche scientifique sur les mécanismes d’apprentissage
- **Tout est tracé sur Coursera !**

Vos étudiants ne le réalisent pas, mais tout est déjà tracé dans votre LMS actuel !

Big data et MOOC

Les données comportementales permettent :

- L’identification des hauts potentiels
- La détection des risques de décrochage
- L’adaptation personnalisée du cours

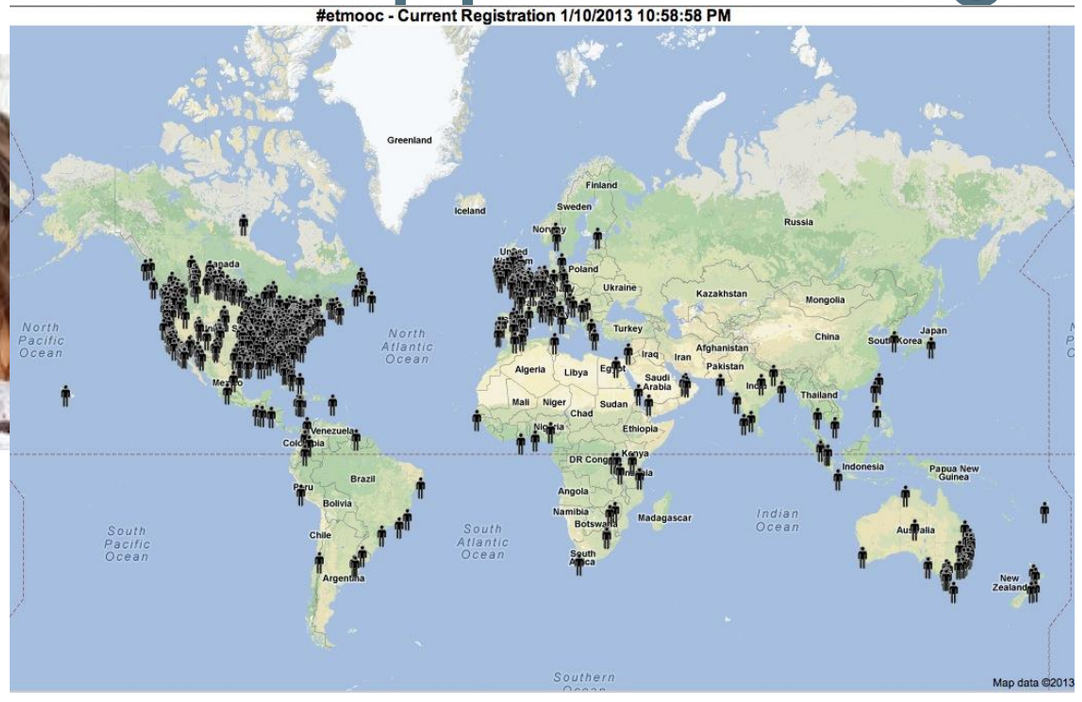
Big data et enseignement

- **Prédire les progrès** d'un étudiant
- Développer des **stratégies d'apprentissage adaptatives** (constituer des séquences d'activités pédagogiques ou recommander des interventions humaines)
- Dresser un **portrait des habiletés** actuelles d'un étudiant
- **Regrouper des étudiants** automatiquement, selon leurs besoins en matière d'apprentissage
- Identifier les **stratégies d'apprentissage** les plus efficaces selon diverses situations
- Organiser et présenter des données complexes de manière à aider les administrateurs, enseignants et étudiants à **gérer des processus de formation**.

On travaille sur 10 000 étudiants
Une étude classique porte sur 20 étudiants

Source : http://www.usherbrooke.ca/ Weston, 2012; traduction libre

Big data et apprentissage social



L'apprentissage social a une place essentielle sur le MOOC

- L'évaluation par les pairs
- Les projets collectifs évolutifs
- La remise en cause permanente du support d'enseignement

On rentre dans la logique de la recherche

Big data et enseignement

Une classe de 10 000 étudiants

- Les tâches assurées par des professeurs et leurs assistants ne sont plus gérables
- Tutorat, notation, modération des discussions doivent être automatisés
- Ces tâches peuvent être en partie gérées par les co-apprenants eux-mêmes.
- Les étudiants participent à l’enrichissement du cours.

Source : http://www.usherbrooke.ca/ Weston, 2012; traduction libre

Pilotage et données massives

Salle Broca, le datacenter de l'université Descartes
Source : phototheque.parisdescartes.fr

Où sont nos données massives ?

Plates-formes d'enseignement

Usage services numériques

ENT

Ressources multimédia

Données métier

sifac apogee harpège

Réseaux sociaux

Carnets

Données de recherche

Bibliothèques

Partenaires…

RATP velib Google

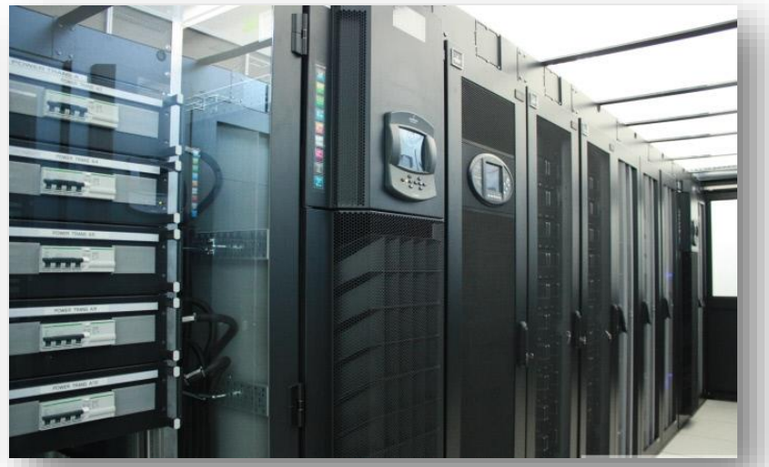
Big data : comment faire ?

- **Objectifs : mieux piloter l'université**
  - Croiser les données de sources multiples
  - Diffuser l'information à tous les acteurs
  - Permettre l'inférence, favoriser l'intuition
  - Permettre le feedback
  - Systématiser l'auto-analyse des usages

- **Compétences :**
  - Forage de données
  - Analyse des réseaux sociaux

Big data : comment faire ?

- **Moyens techniques :**
  - Le stockage : cloud
  - La géolocalisation : mobiles
  - Outils d’analyse : dataviz, hadoop
  - Les réseaux sociaux



- **Moyens humains :**
  - Chief Data Officer
  - En 2020, 190 000 data scientists aux USA (McKinsey)
  - Equipe de pilotage
  - Utiliser les ressources scientifiques de l’université

- **Choix fonctionnels :**
  - L’ouverture et le partage : open data
  - L’anonymisation des données
  - Accepter les données floues
  - Partenariats systématiques

Big data : que retenir ?

- Le cerveau humain est plus efficace que les ordinateurs
- Engager le grand public dans la recherche
- Utiliser des données de multiples fois à des fins non prévues
- Mélange de sources publiques et privées de données
- Résultats inattendus
- Nouvelle philosophie
- Nouveaux métiers
- Nouveaux moyens techniques

Les services centraux et les SI sont au service de l’enseignement et de la recherche

Les données massives doivent être un outil pour progresser.