

# DATA DRIVEN SCIENCE

QUÉSACO ?

Microsoft®  
**Research**

**Data Driven Science**  
**Science dirigée par les données**

**4<sup>ème</sup> paradigme**      **Jim Gray**  
**de la science**      **1944-2007**

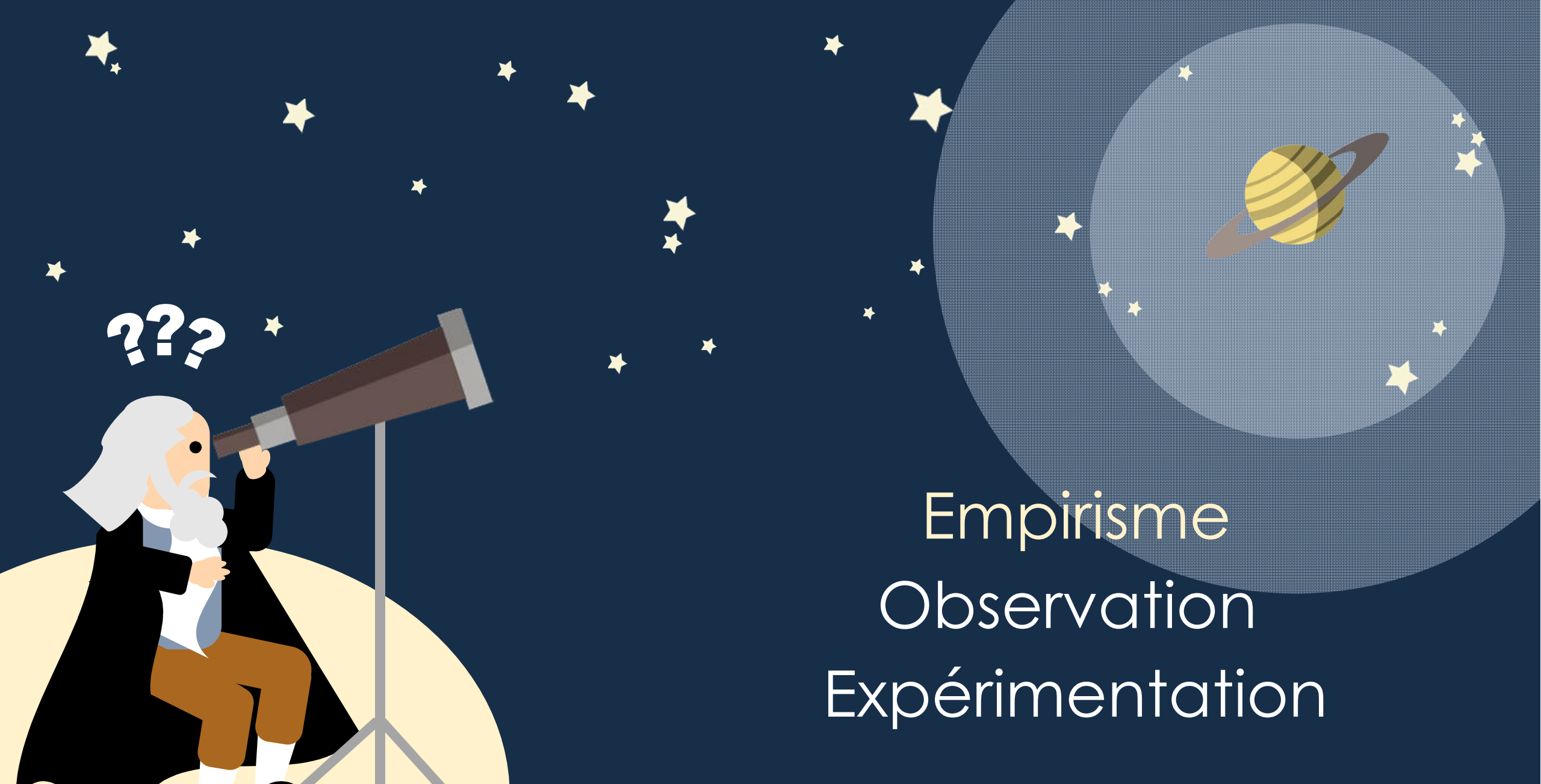
La data driven science, littéralement science dirigée par les données, a été considérée comme étant le 4<sup>ème</sup> paradigme de la science par Jim Gray, chercheur en informatique chez Microsoft. Pour comprendre comment nous en sommes arrivés là, passons rapidement en revue l'évolution de la science.

1<sup>er</sup>  
paradigme



Milliers d'années

Selon Jim Gray, le premier paradigme est apparu il y a des milliers d'années.



Empirisme  
Observation  
Expérimentation

C'est l'époque de l'empirisme, la méthodologie repose sur l'observation de phénomènes et l'expérimentation, conduisant les scientifiques à émettre des hypothèses qu'ils essayaient alors de vérifier pour les confirmer ou non.

1<sup>er</sup>  
paradigme

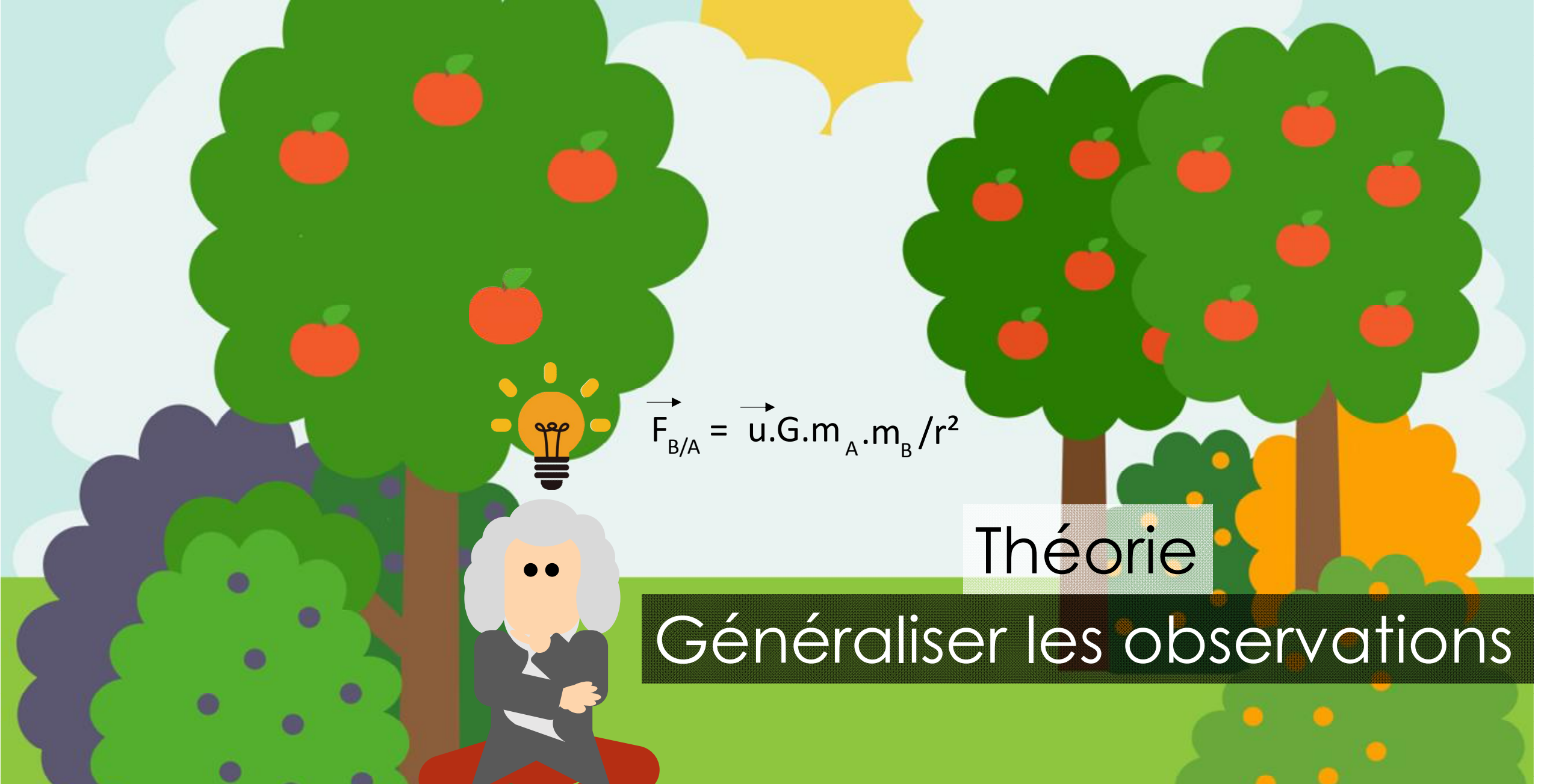
2<sup>ème</sup>  
paradigme



Milliers d'années

Centaines d'années

Il y a quelques centaines d'années, un second paradigme se dégage.



$$\vec{F}_{B/A} = \vec{u} \cdot G \cdot m_A \cdot m_B / r^2$$

Théorie

Généraliser les observations

Les scientifiques cherchent à émettre des théories pour expliquer et généraliser les observations empiriques. Ils établissent ainsi des lois scientifiques sous forme mathématique : par exemple la loi de la gravitation de Newton.

1<sup>er</sup>  
paradigme



Milliers d'années

2<sup>ème</sup>  
paradigme



Centaines d'année

3<sup>ème</sup>  
paradigme



Dizaines d'année

Puis, il y a quelques dizaines d'années, apparait le 3<sup>ème</sup> paradigme.



Simulation

Phénomènes complexes

Avec la puissance de calcul des ordinateurs, les scientifiques peuvent désormais recourir à des simulations pour vérifier des théories relevant de phénomènes complexes. C'est par exemple le cas en météorologie, où l'on simule l'évolution de l'atmosphère pour prédire le temps qu'il va faire.



1<sup>er</sup>  
paradigme



Milliers d'années

**Science  
empirique**

2<sup>ème</sup>  
paradigme



Centaines d'année

**Science  
théorique**

3<sup>ème</sup>  
paradigme



Dizaines d'année

**Science  
computationnelle**

Selon Jim Gray, 3 paradigmes se sont donc succédés : la science empirique, la science théorique, et la science computationnelle.

4<sup>ème</sup>  
paradigme



De nos jours  
**Science dirigée  
par les données**

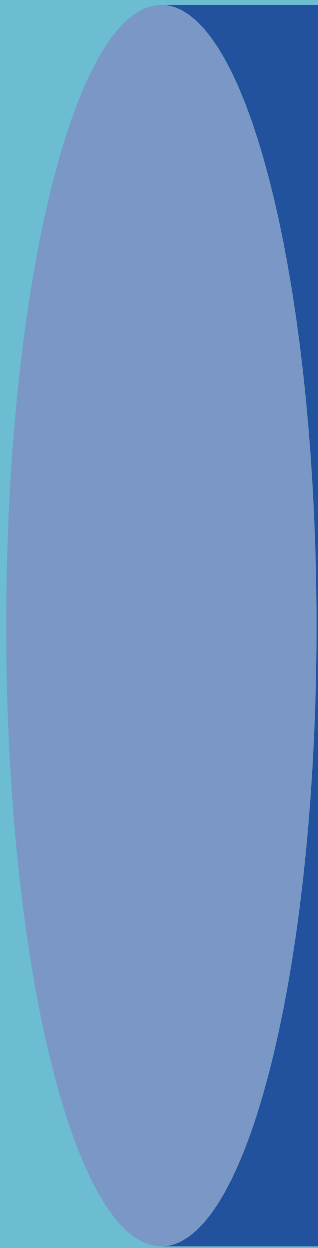
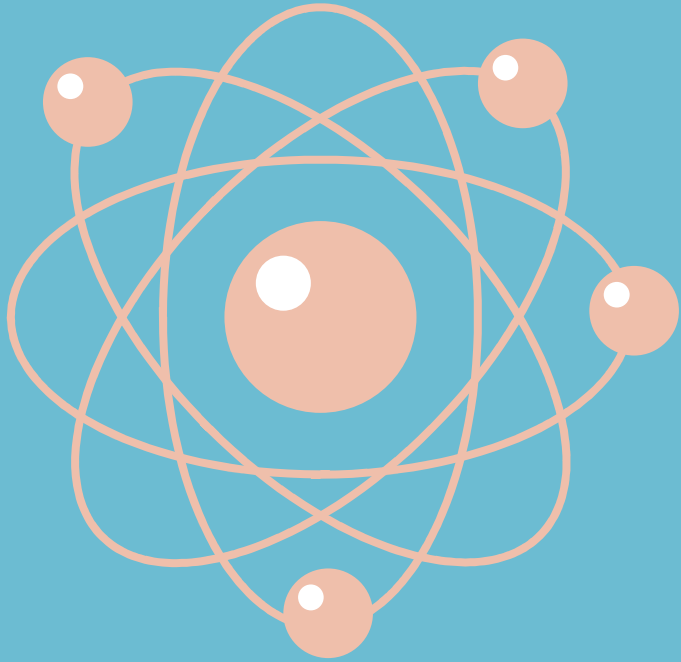
1010101010100101010010101010101010100101010101011010  
110110101010001011101010101010100101010101101001  
0101010100101011010101010100101010101001010010010101

Algorithmes



Objet  
Recherché

Depuis quelques années, certaines disciplines auraient adopté un 4<sup>ème</sup> paradigme, caractérisé par le *Big data* : la Science dirigée par les données. Que l'on travaille à partir de simulations, de capteurs de mesures, ou de données issues d'Internet, des recherches s'appuient sur de grandes masses de données qui sont, pour notre cerveau humain, bien difficiles à appréhender. On utilise alors des algorithmes pour faire parler les données et tenter de trouver un objet recherché, ou une hypothèse probable.



# LHC GRAND COLLISIONNEUR D'HADRONS



Prenons l'exemple du LHC, un grand anneau sous-terrain utilisé pour faire rentrer en collision des faisceaux de particules.

**1 000 000 000 000 000  
000 000 000  
octets / jour !**



Ces données sont donc traitées automatiquement par les ordinateurs et permettent de faire des découvertes.



Postulée par  
Peter Higgs  
1964

La plus célèbre est celle du boson de Higgs, une particule dont l'existence avait déjà été avancée en 1964 par Peter Higgs. Il y avait donc initialement une théorie scientifique, qui a pu être validée par le traitement algorithmique d'énormes quantités de données.

# 4<sup>ème</sup> paradigme ?

## Méthode complémentaire ?

Données



Science  
empirique

Science  
théorique

Science  
computationnelle

Science dirigée  
par les données

Le débat sur l'avènement d'un 4<sup>ème</sup> paradigme, tel qu'avancé par Jim Gray, reste ouvert. Pour certains, ce n'est pas tant un nouveau paradigme qu'une méthode complémentaire aux approches traditionnelles, rendue nécessaire par la présence de gros volumes de données. Quoi qu'il en soit, la science s'axe de plus en plus sur les données, qui de par leur ouverture et leur croissance exponentielle, sont désormais à prendre en compte dans le processus de recherche scientifique.