



Big Graph Data Forum Teratec 2013

Julien Laugel
MFG Labs

MFG Labs
35 rue de Châteaudun
75009 Paris, France

www.mfglabs.com
twitter: @mfg_labs

julien.laugel@mfglabs.com
@roolio

SOMMAIRE

MFG Labs

Contexte : les 3 V aujourd'hui

Les bases de données graphe

Les solutions actuelles

Les challenges

The background of the slide is a dark blue gradient. It features a complex, abstract network of thin, light blue lines and dots. These lines and dots are interconnected in a way that suggests a global or digital network, with some areas being more densely connected than others. The overall effect is a sense of connectivity and technology.

MFG Labs

Un ADN mathématique



Jean-Michel Lasry (CEO)
Professor of university Paris Dauphine



Pierre-Louis Lions
Fields Medal 1994

- Fondé en 2010 par 2 mathématiciens
- MFG: "Mean Field Games"
- Initialement société de conseil dédié aux applications des MFG en économie

Les activités de MFG Labs



Recherche Mathématique



Big Data



Stratégie Digitale

Contexte : les 3 V aujourd'hui

L'ère post-Map/Reduce

Reste-t-il des challenges pour
la trinité BigData?

Volume
Velocity
Variety

VOLUME

bases de données MPP

Map/Reduce (Hadoop)

Spark

VÉLOCITÉ

Complex Event Processing

Bases de données orientée flux

Base de données In-memory

Storm, S4

VARIÉTÉ

Données non-structurées

Données semi-structurées

Données textuelles

Problématiques «maitrisées»

Stockage

=> Systèmes de Fichiers Distribués

Tâches fortement parallélisables

=> Map/Reduce

Bases de données distribuée à faible latence d'écriture

=> BigTable, Dynamo

Les question ouvertes

Nombre des sources de données

Volume des liaisons entre les données $\times$ Volume des données

Que faire avec les gros graphes?

Le constat

Les technologies Big Data traditionnelles ne savent pas gérer les gros graphes de données

- Traversée de graphe trop coûteuse (itérations)
- Partitionnement des données non trivial

BIG GRAPHS

à l'échelle sociale : Facebook

1 milliard de noeuds

144 milliards de relations

à l'échelle du web :

50 milliards de noeuds

1000 milliards de liens

à l'échelle du cerveau :

100 milliards de noeuds

100 000 milliards de liens

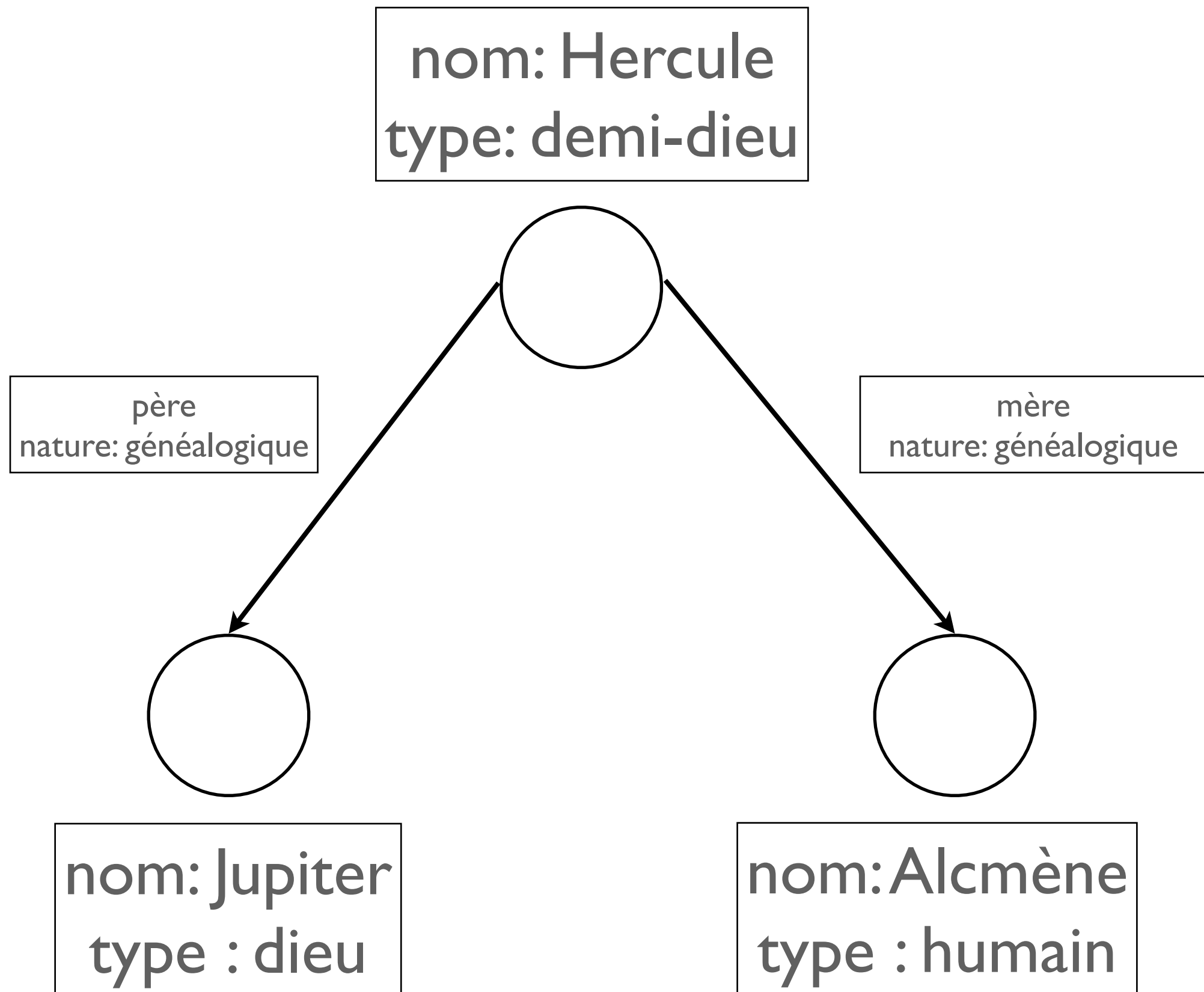
The background of the slide is a dark blue gradient. It features a complex, abstract network graph pattern. This pattern consists of numerous small, light blue dots (nodes) connected by thin, light blue lines (edges). The connections are dense in some areas, forming clusters, and more sparse in others, creating a web-like structure that fills the upper and middle portions of the slide.

Les bases de données graphe

The background of the slide is a dark blue gradient. It features a complex, abstract network of thin, light blue lines and dots. These lines and dots are interconnected in a way that suggests a global or digital network, with some areas being more densely connected than others. The overall effect is a sense of connectivity and data flow.

Définition

Le modèle property graph



The background of the slide is a dark blue gradient with a complex, abstract network pattern. This pattern consists of numerous small, light-colored dots connected by thin, curved lines, creating a sense of interconnectedness and data flow. The density of the network is higher in the upper left and center, with some clusters of dots and lines extending towards the right and bottom.

L'enjeu

L'importance du modèle graphe

Stockage avant usage : Flexibilité du modèle de données

Agilité : Nécessité d'un modèle de données dynamique

Machine Learning :

De très nombreux algorithmes s'expriment sous forme de graphe (PageRank, Filtrage collaboratif)

Analogie entre matrices sparses et les graphes

Les constituants

Un «moteur» basé sur la traversée de graphe

Un modèle de calcul permettant de développer des algorithmes

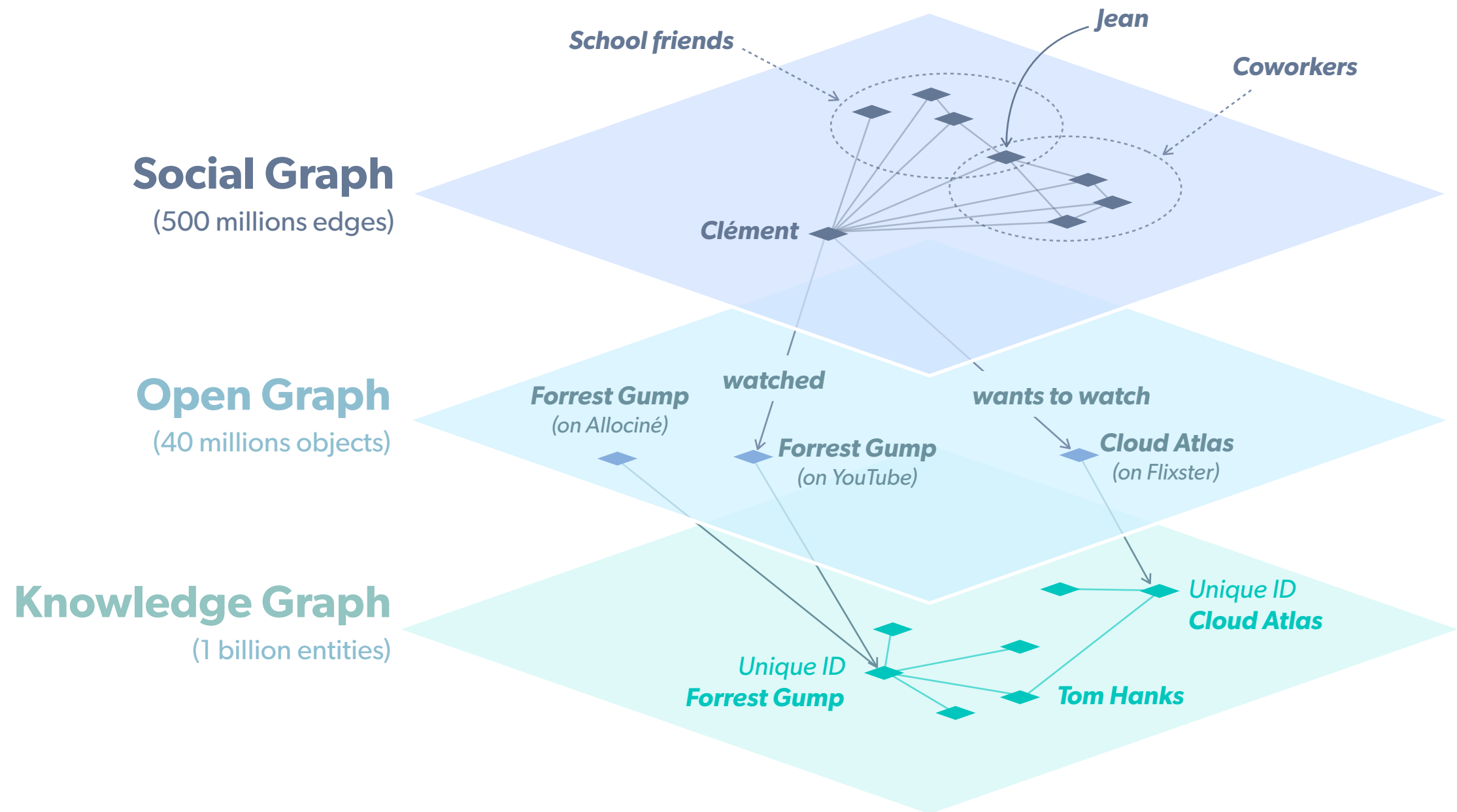
Des méthodes de partitionnement de données efficaces

Un langage de requêtes permettant la recherche de motifs

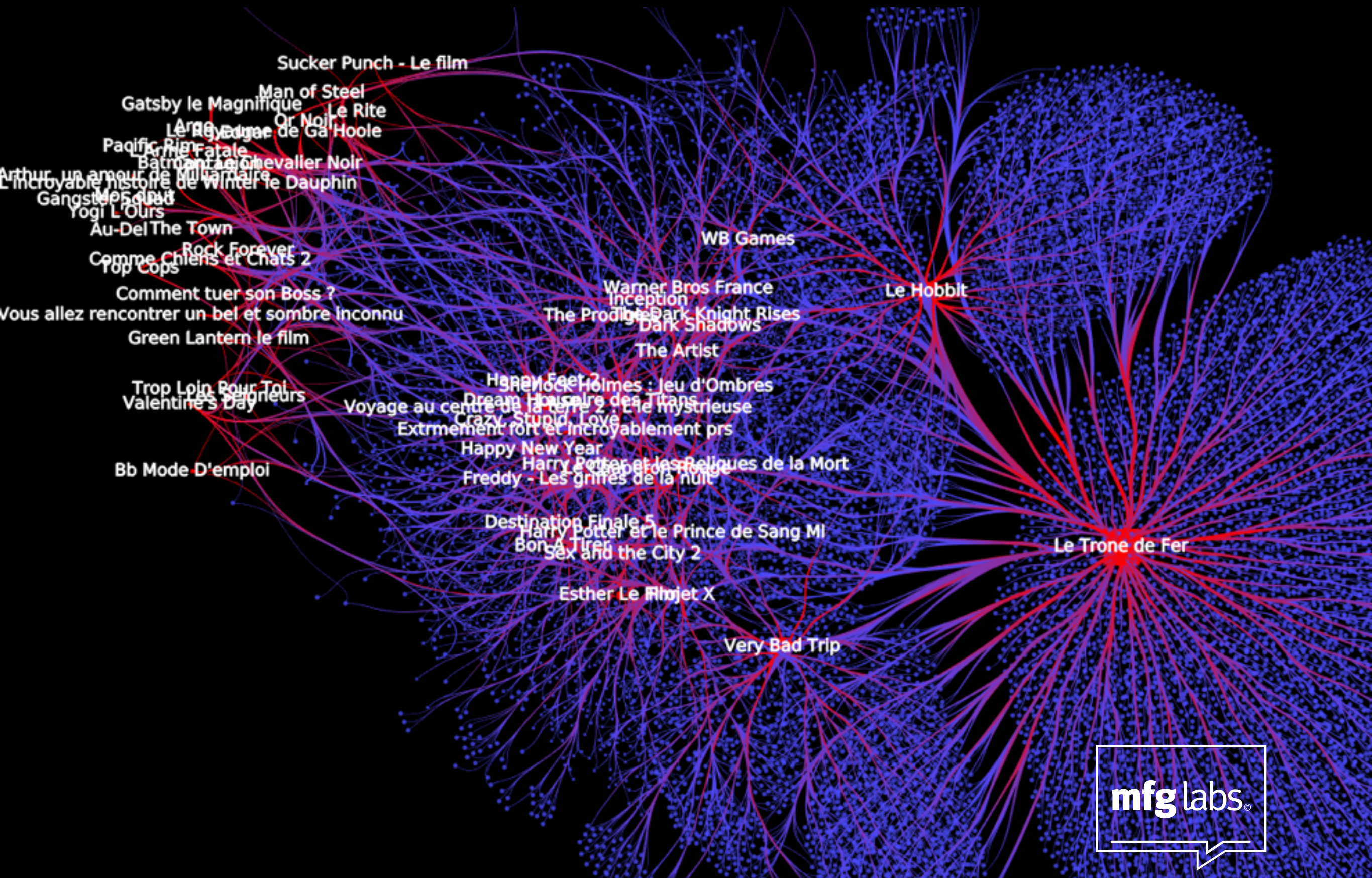
The background of the slide is a dark blue gradient. It features a complex, abstract network of thin, light blue lines and dots. These lines and dots are interconnected in a way that suggests a global or digital network, with some areas being more densely connected than others. The overall effect is a sense of connectivity and data flow.

Examples

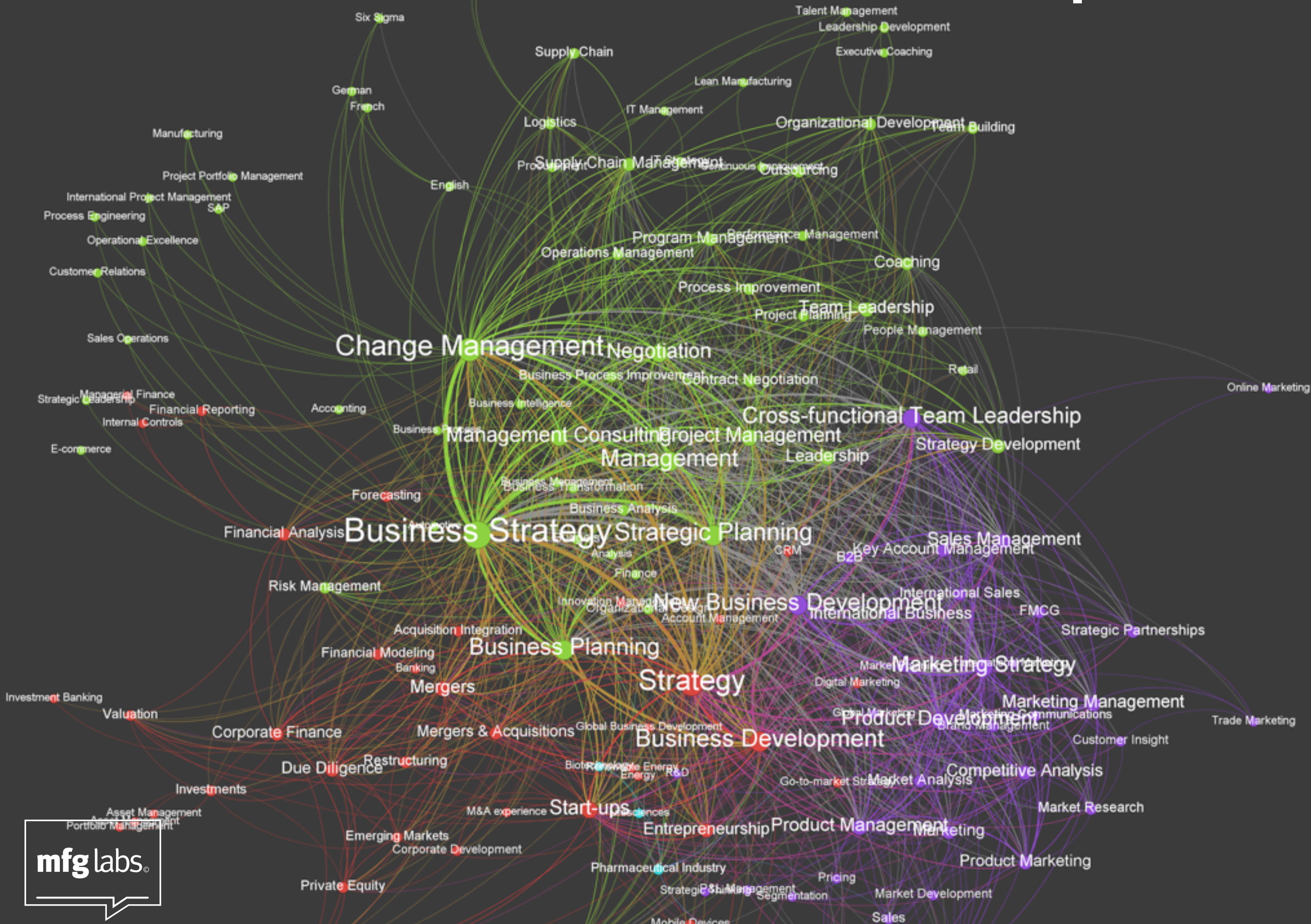
Interconnecter des sources



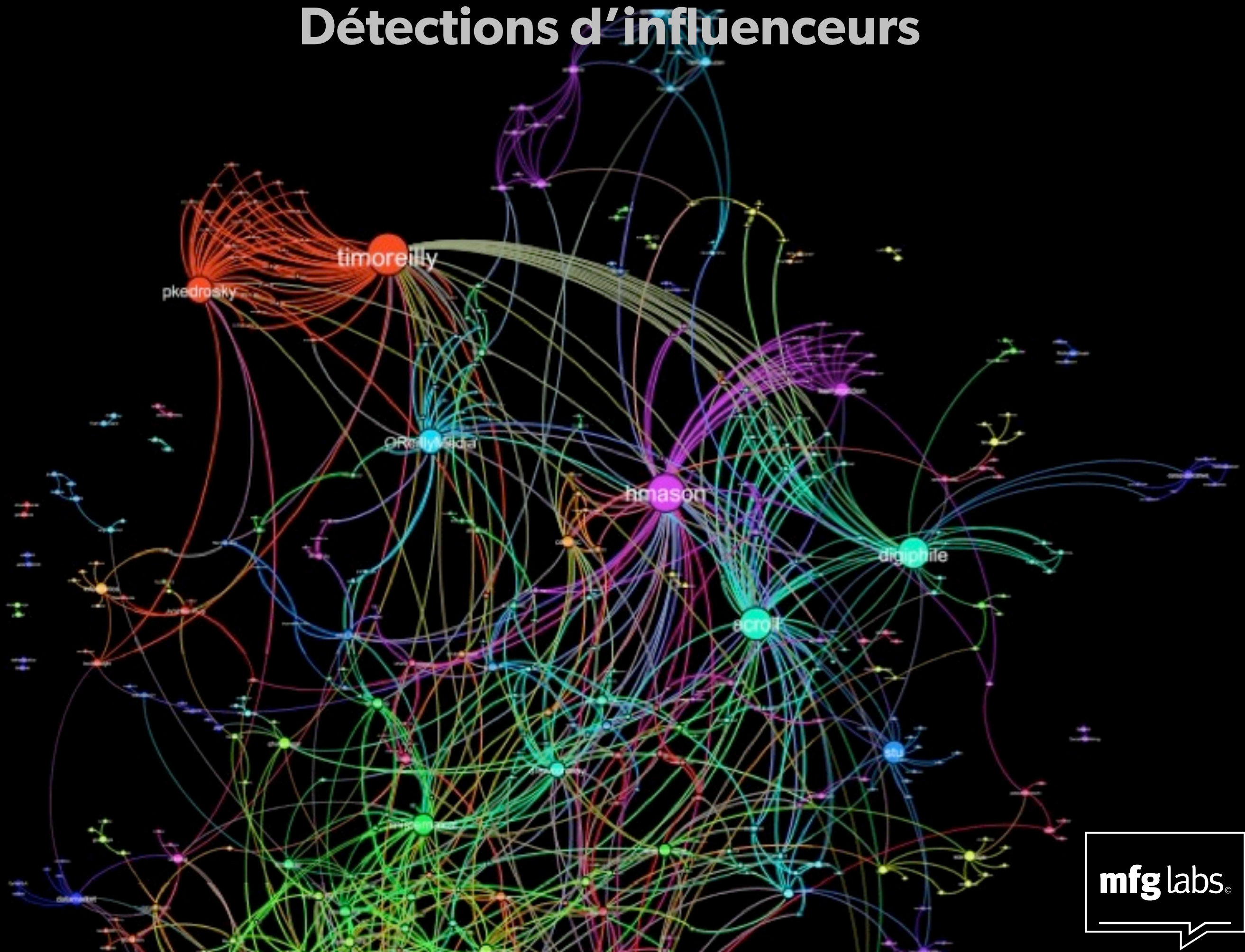
Recherche de similarités via les utilisateurs



Recherche de similarités entre concepts



Détections d'influenceurs



Diffusion dans un graphe

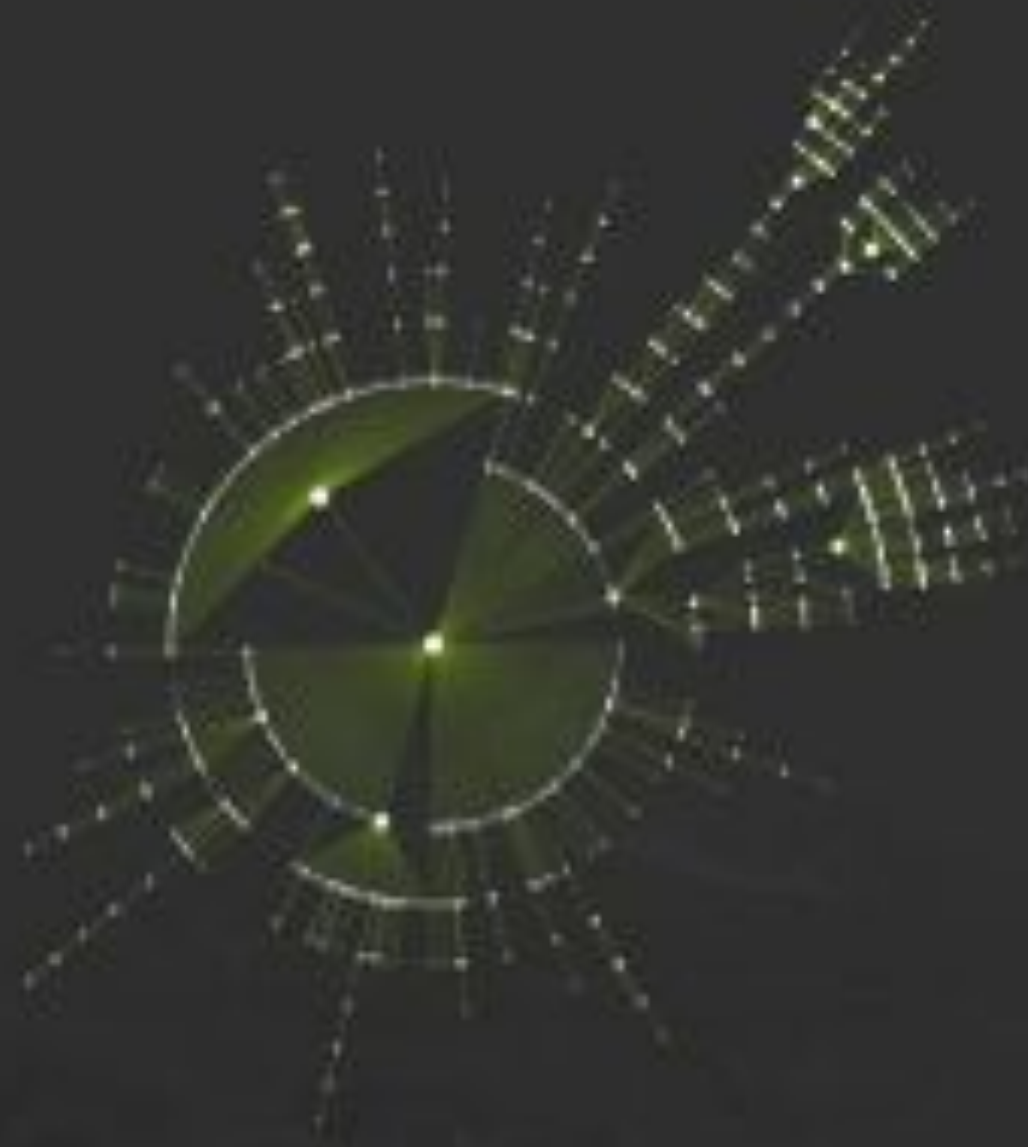
WHERE DOES
MY TWEET GO?

Measure the spread of
tweets over the world

The speed at which Twitter spreads
information is fascinating. Not looking
at your timeline it's not always easy to
guess how tweets get spread to the
show. To enable you to get a perfect
understanding of how tweets
spread, we created WORDS.

←

←



→



8 Hobbit = 1 Hobbyte?

@Roghayeh 1,555 Retweets 443,369 Reach



704 SpreadRank



See The graph

mfg labs

©2011 mfg labs. All rights reserved. Report a bug

«Penser» les données en graphe

Logs de serveur de demande d'itinéraires :

id | ville de départ | ville de destination

==>

ville de départ | ville de destination | occurrences

==>

noeuds = villes

liens = nombre de demandes de trajets entre ville de départ et ville de destination

Découvrir les structures endogènes



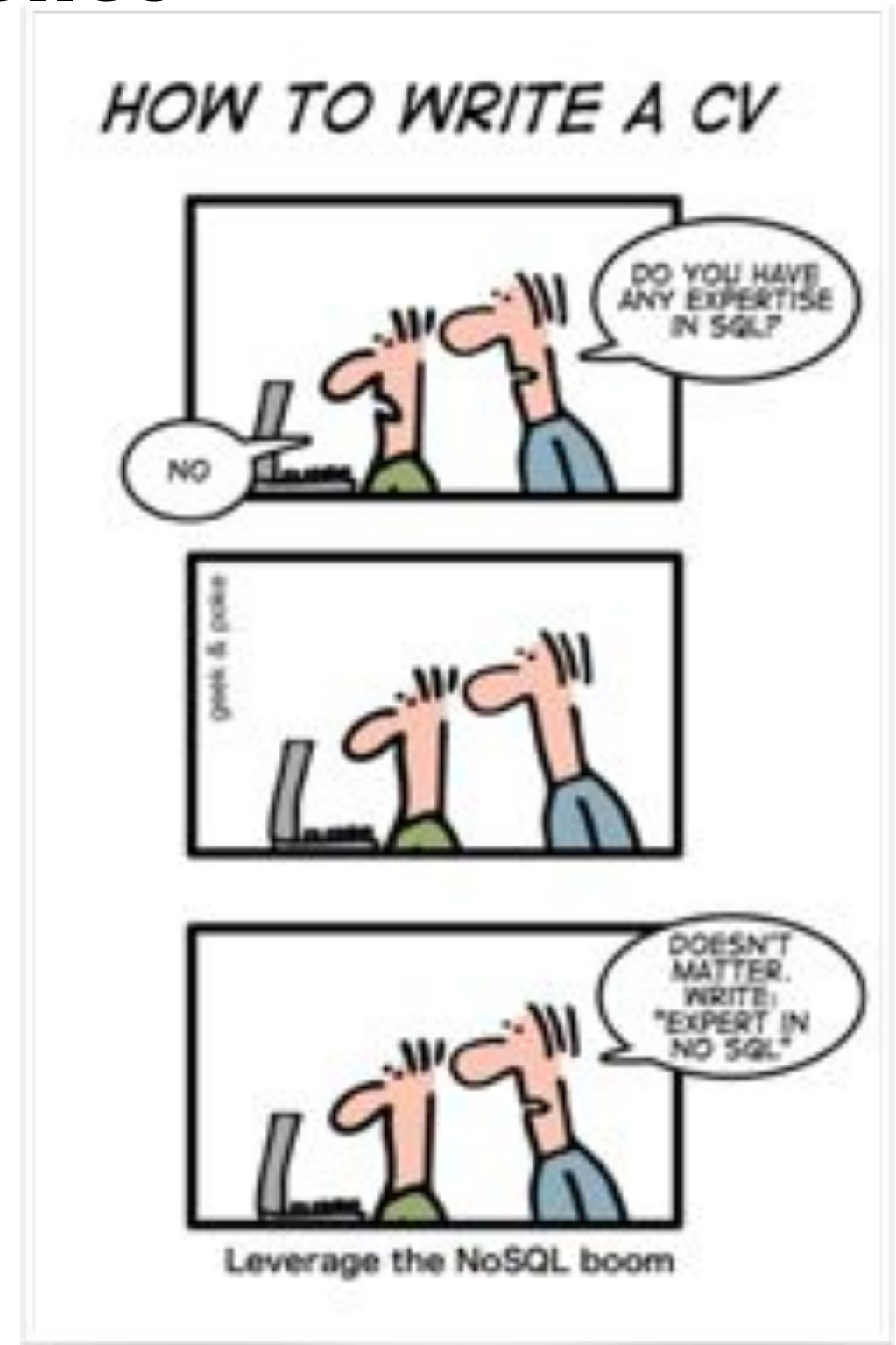
The background of the slide is a dark blue gradient with a complex, abstract network pattern. This pattern consists of numerous small, light-colored dots (nodes) connected by thin, light-colored lines (edges). The connections are dense in some areas, forming clusters, and more sparse in others, creating a sense of a vast, interconnected system. The overall effect is a modern, technological aesthetic.

Les solutions actuelles

The background of the slide is a dark blue gradient. Overlaid on this is a complex, abstract network of thin, light blue lines and dots. The lines connect various points, some of which are larger and more prominent than others, creating a sense of a vast, interconnected web or a fragmented landscape. The overall effect is one of complexity and connectivity.

Un paysage de plus en plus fragmenté

La fin de l'hégémonie des bases de données relationnelles



Map/Reduce



Bases de données orientées colonnes



Bases de données de flux



Stores Clé/Valeur



Bases de données orientées documents



Quid du processing des graphes ?

Frameworks

Bases de données

The background of the slide is a dark blue gradient with a complex, abstract network pattern. This pattern consists of numerous small, light-colored dots (nodes) connected by thin, light-colored lines (edges). The connections are dense in some areas, forming clusters, and more sparse in others, creating a sense of a vast, interconnected system. The overall effect is a modern, technological, and data-driven aesthetic.

Les frameworks

Les frameworks

Parallélisme de Données :

Map/Reduce : Systeme de fichier distribué

GraphX : Primitives de parallélisme de graphe
implémentées sur une abstraction data-parallèle

Parallélisme de Graphe («Think like a vertex»):

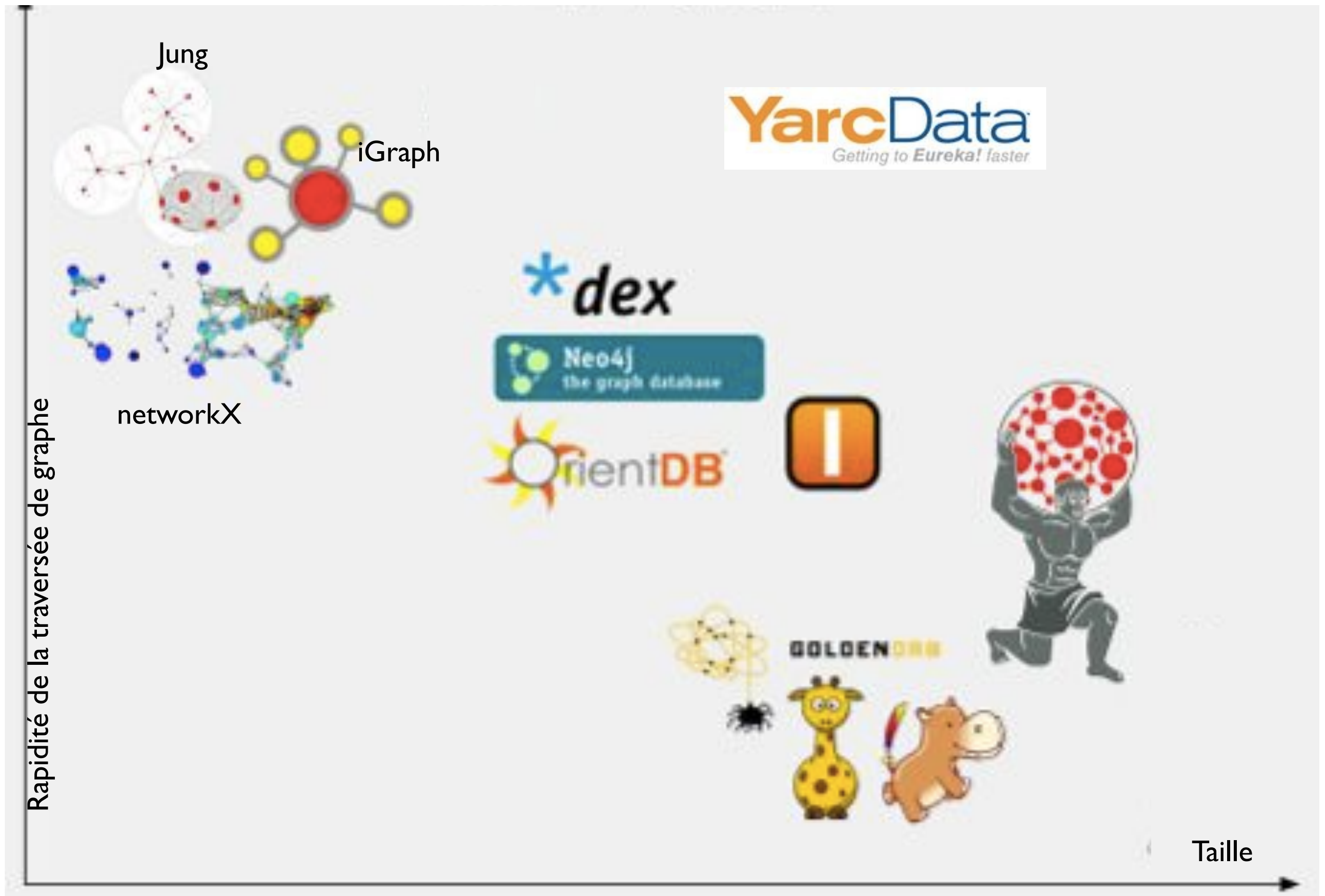
Bulk Synchronous Processing (BSP)
Modification des noeuds synchrone
Partitionnement des noeuds

Graphlab v2
Modification des noeuds asynchrone
Partitionnement des liens

The background of the slide is a dark blue gradient. Overlaid on this is a complex, abstract network of thin, light blue lines and dots. The lines connect various points, some of which are larger and more prominent than others, creating a sense of interconnectedness and data flow. The dots are scattered throughout, some forming small clusters.

Les bases de données

Les bases de données



Qui utilise quoi?

Neo4J en passe de devenir le PostgreSQL du graphe

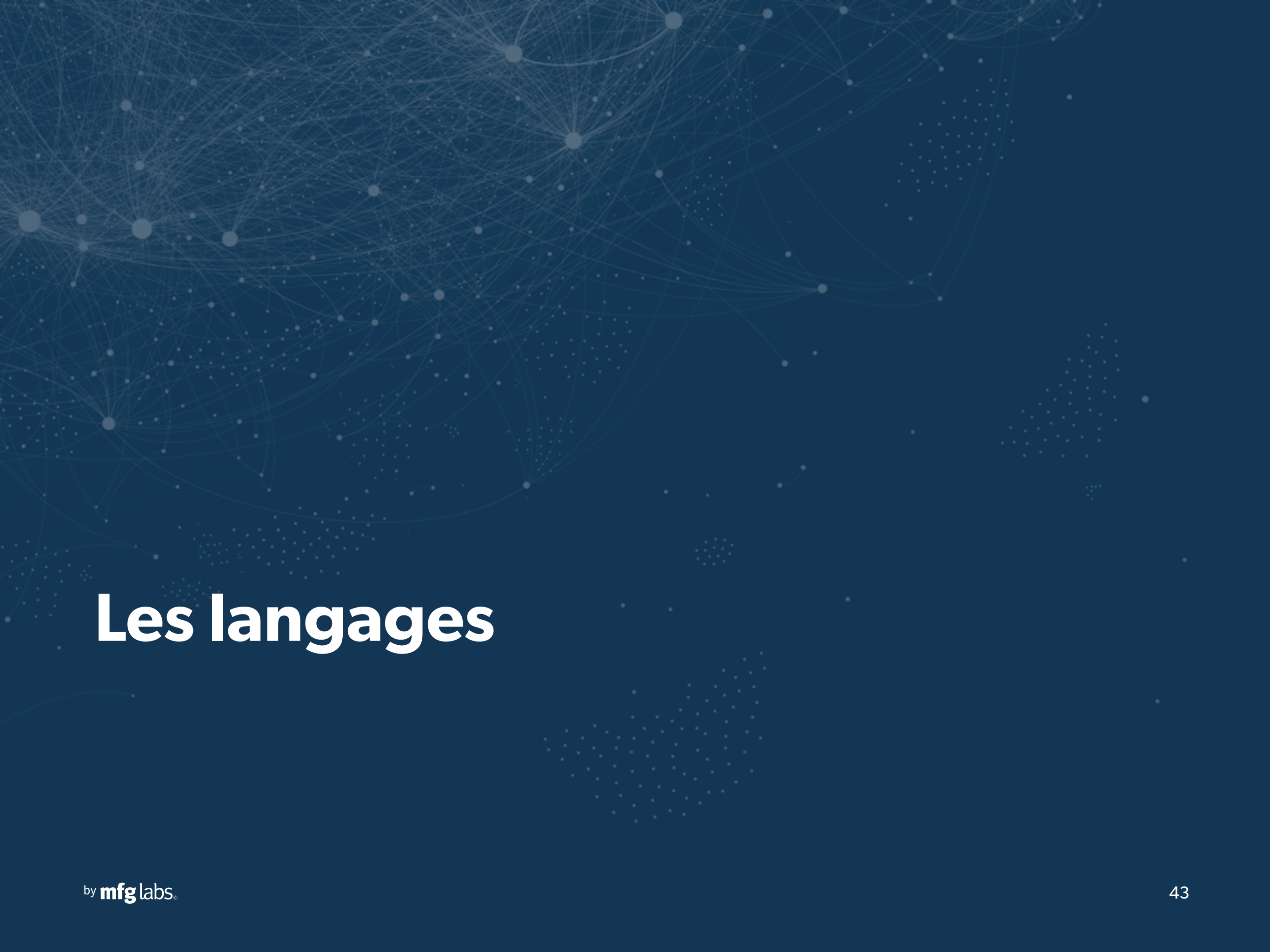
Les gros acteurs de la Silicon Valley utilisent Giraph, ou des systèmes spécifiques à leur besoin

Hadoop reste encore utilisé dans ce domaine

GraphX est une alternative prometteuse

Graphlab est à la pointe du développement au niveau algorithmique

Titan + Faunus aussi très intéressant à suivre pour un couple stockage+processing



Les langages

Les langages

- SPARQL, OWL (bases de données sémantiques)
- Impératif : Gremlin
«suivre» les relations
Expliciter les algorithmes (breadth-first , depth-first...)
- Déclaratif : Cypher
Description du pattern à rechercher

Le Langage Cypher

Simple description du pattern recherché

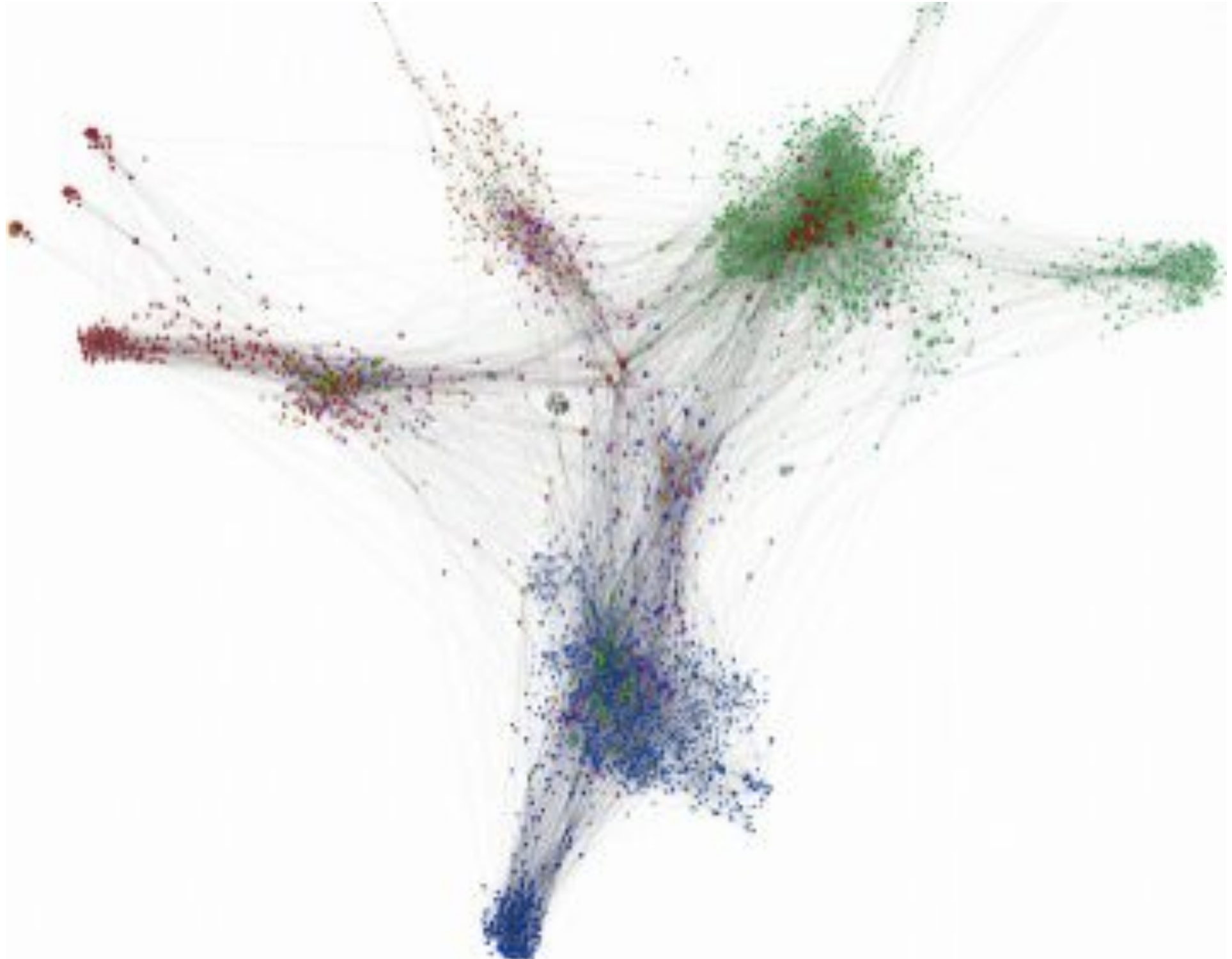
```
1
2
3 // Recommandation d'amis
4 start u = node:users(name="Julien Laugel")
5 match u -[:knows]-> v -[:knows]-> z
6 where not(u-[:knows]-z) and not(u = z)
7 return z, count(v) as cnt
8 order by cnt desc
9 limit 10;
10
11
12 // Quels sont mes amis les plus musicophiles?
13 start u = node:users(name="Julien Laugel")
14 match u -[:knows]-> v -[:likes]-> item -[:`type.object.type`]-> type
15 where has(type.freebase) and type.freebase = 'music.artist'
16 return v, count(item) as cnt
17 order by cnt desc
18 limit 10;
```

The background of the slide is a dark blue gradient. It features a complex, abstract network visualization consisting of numerous small, light blue dots (nodes) connected by thin, light blue lines (edges). The connections form a dense web in the upper left and center, with some lines curving and looping. There are also several smaller, more isolated clusters of dots scattered across the lower right portion of the slide.

Les outils de visualisation

Une spécialité française

Gephi: <http://gephi.org>



Une spécialité française

Tulip : <http://tulip.labri.fr>



Une spécialité française

Graphstream : <http://http://graphstream-project.org/>



The background of the slide is a dark blue gradient. It features a complex, abstract network of thin, light blue lines and dots. These lines and dots are interconnected, forming a web-like structure that is denser in some areas and sparser in others. The dots vary in size, with some being larger and more prominent than others. The overall effect is a sense of connectivity and complexity.

Les challenges

Les Challenges

1. Rendre les bases de données graphes «mainstream»
2. Emergence d'un standard pour les très gros graphes
3. Traitement des graphes en flux de données
4. Le «big graph processing» permettra l'accélération du «big machine learning»
5. Le sujet est complexe, et ouvre de grandes opportunités à qui saura gérer les aspects
 - infrastructure
 - algorithmes
 - langage de requête
 - visualisation



📍 MFG Labs
35 rue de Châteaudun
75009 Paris, France

www.mfglabs.com
twitter: @mfg_labs

julien.laugel@mfglabs.com
@roolio