

Introduction

Daniel Aloise <daniel.aloise@polymtl.ca>

Fouille de données

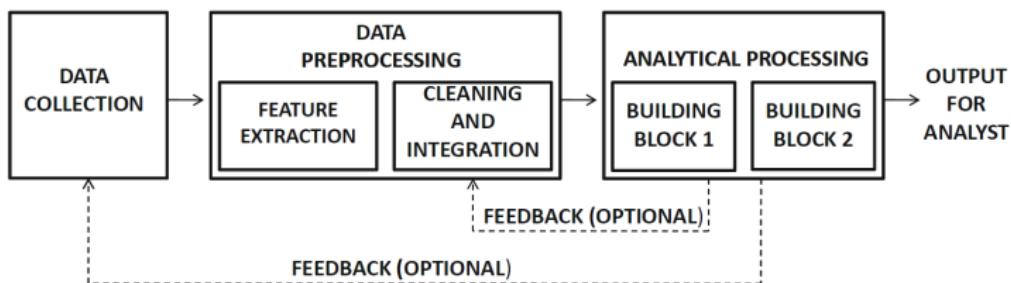
- La **fouille de données** est l'étude de la collecte, du nettoyage, du traitement, de l'analyse et des connaissances issus des données
- À présent, presque tous les systèmes automatisés génèrent des données à des fins de diagnostic ou d'analyse.
 - e.g. web, transactions financières, *smartphones*, IoT, etc.



Ertel, 2011

Fouille de données

- Les données brutes sont très souvent non structurées
- Le *pipeline* de la fouille de données :



C.Aggarwal, 2010

Exemple de la pipeline



S'INSCRIRE

Mon magasin: ST-LEONARD

Chercher à 10 km

Rechercher plus de 300 000 articles

Magasiner par rayon Magasiner par place Idées et instructions Services à domicile Promotions et offres Circular hebdomadaire

En raison d'une forte augmentation de la demande, les délais peuvent être plus longs pour certaines ventes en ligne et pour les livraisons. Veuillez adhérer à nos règles de sécurité du conducteur ou faire vérifier l'état de votre automobile.

Mon compte Il Parler

Accès Matériaux de construction Quincaillerie Électroménager Aménagement Accessoires pour cloisons séparées et cloisons

EZ Anchor Solvair de montant #8 (R) Ancrez au poteau en zinc pour cloison sèche avec vis - 275pc

Modèle: 545-481 | Réf.: 1001003072

(1) Vérifiez une évaluation : Q et R (2)

83,52 \$ / chaque

Livraison gratuite pour cette dépense de 49\$

En rupture de stock à St-Léonard

Quantité: 1 Ajouter au panier

Magasin de remplacement: St-Léonard Options de livraison pour le code postal: 9261, QC

RAMASSEZ EN MAGASIN SERVICE D'EXPÉDITION SERVICE DE LIVRAISON EXPRESS

Envoi de déchets Envoi de déchets

Cet article est en rupture de stock et ne peut pas être expédié immédiatement par Livraison Express à ce code postal.

Exemple de la pipeline



Saint-Hubert • Chaudière-Appalaches • Québec • Canada

Rechercher plus de 300 000 articles

Magasin par rayon • Magasins par place • Idées et instructions • Services à domicile • Promotions et offres • Circular hebdomadaire

En raison d'une forte augmentation de la demande, les délais peuvent être plus longs pour certaines ventes en ligne et pour la livraison. Veuillez adhérer à nos règles de sécurité du travail et de prévention des risques pour nous aider à maintenir une vente en ligne sûre et sûre.

[Accès](#) • [Matières de construction](#) • [Quincaillerie](#) • [Fournitures](#) • [Aérage](#) • [Accessoires pour cloisons séparées et cloisons](#)

EZ Anchors Solvair de montant #8 (R) Ancrez au poteau en zinc pour cloison sèche avec vis - 275pc

Modèle n° 545-401 • RPU02-1001003072

(1) Achetez une évaluation : Q et R (R)

83,52 \$ / chaque

Expédition gratuite pour cette dépense de 49\$

En rupture de stock + 3 semaines

Quantité : 1

Ajouter au panier

Magasin de remise : [St-Hubert](#)

Options de livraison pour le code postal H4L 1C2

RAMASSEZ EN MAGASIN

Service d'expédition

ENTREPOSAGE

Service de livraison express

En rupture de stock

Cet article est en rupture de stock et ne peut pas être expédié immédiatement par

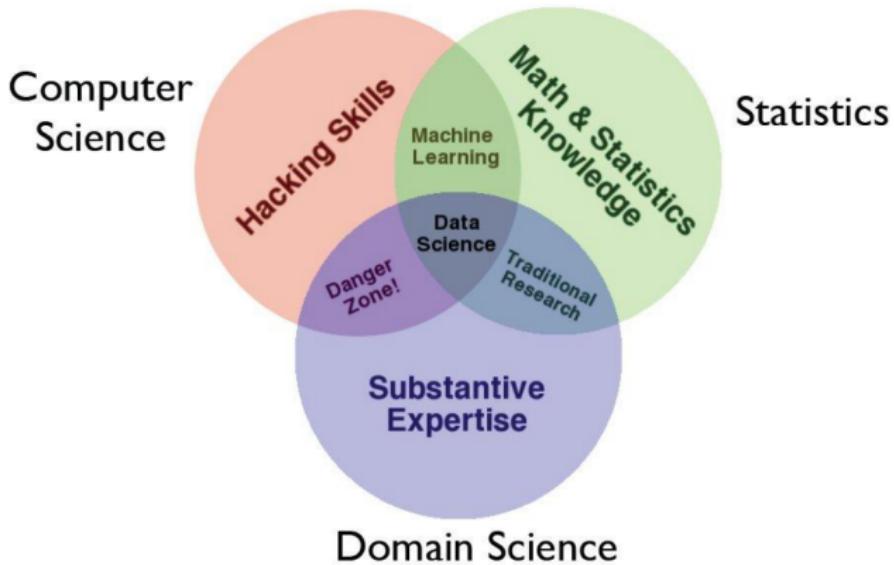
Livraison express à ce code postal.

Exemple d'un log d'accès :

```
98.206.207.157 - - [31/Jul/2013:18:09:38 -0700] "GET /productA.htm
HTTP/1.1" 200 328177 "-" "Mozilla/5.0 (Mac OS X) AppleWebKit/536.26
(KHTML, like Gecko) Version/6.0 Mobile/10B329 Safari/8536.25"
"retailer.net"
```

- **Nettoyage** : les logs contiennent beaucoup d'informations supplémentaires qui ne sont pas nécessairement utiles pour le détaillant.
- **Extraction d'attributs** : le détaillant décide de créer un enregistrement pour chaque client, avec un choix spécifique d'attributs.
- **Intégration des données** : ces enregistrements sont combinés aux données démographiques des clients stockées chez le détaillant.
- **Analyse des données** : maintenant, l'analyste doit décider comment utiliser cet ensemble de données nettoyé pour faire des recommandations.

Science de données



Drew Conway

Data science vs. Computer Science

- Généralement, les informaticiens n'apprécient pas les données :
ce sont juste des choses à exécuter dans un programme.
- La façon habituelle de tester les performances d'un algorithme consiste à exécuter l'implémentation sur des données aléatoires.
- Pourtant, les ensembles de données intéressantes sont rares, ce qui demande du travail et de l'imagination pour les obtenir.

Data science vs. Computer Science

Data scientists	Computer scientists
data-driven	algorithm-driven
bruit (ok)	monde structuré
§ pour des <i>insights</i>	§ pour code
focused on discovering	focused on inventing

Nouvelle science

- La science traditionnelle **hypothesis-driven** fait des questions et en suite génère les données spécifiques qui sont nécessaires pour confirmer ou réfuter l'hypothèse.
- La nouvelle science **data-driven** est axée sur la génération de données à une échelle inédite, en croyant que de nouvelles découvertes seront faites aussitôt que l'on sera capable de les explorer.

Les questions d'un(e) *data scientist*

- Qu'est-ce qu'on peut apprendre à partir d'un jeu de données ?

Les questions d'un(e) *data scientist*

- Qu'est-ce qu'on peut apprendre à partir d'un jeu de données ?
- Que veut-on vraiment savoir ?

Les questions d'un(e) *data scientist*

- Qu'est-ce qu'on peut apprendre à partir d'un jeu de données ?
- Que veut-on vraiment savoir ?
- Quels jeux de données pourraient l'aider à savoir ces choses ?

Pratiquons !

Sports Reference | Baseball | Football (college) | Basketball (college) | Hockey | Soccer | Blog | Stathead | Widgets | Create Account | Login | Questions or Comments?

**HOCKEY
REFERENCE**

Players Teams Seasons Leaders NHL Scores 1 Playoffs Play Index Full Site Menu Below ▾

NHL Stats and History The complete source for current and historical NHL players, teams, scores and leaders.

Every NHL Player



View any Active Player:

Choose a Team

Then a player

Go!

Select a Hall of Famer:

Every NHL Team

2017-18 NHL Standings

[Summary](#) · [Leaders](#) · [Schedule](#) · [Standings Prediction](#)

Click conference name for conference standings

	Eastern W	L	OL	PTS
Atlantic				
TBL	50	19	4	104
BOS	45	17	10	100
TOB	43	23	7	93
FLA	37	27	7	81
DET	27	35	11	65
MTL	26	36	12	64
OTT	26	35	11	63
BUF	23	38	12	58

	Western W	L	OL	PTS
Central				
NSH	48	14	10	106
WPG	44	19	10	98
MIN	41	24	8	90
COL	40	25	8	88
STL	40	28	5	85
DAL	38	28	8	84
CHI	30	35	9	69
CBU				

	Metropolitan			
VEG	47	21	5	99
WSH	42	24	7	91
PIT	42	27	5	89
CBJ	41	28	5	87
PHL	37	25	12	86

What's Happening

Site News

[2017-18 NHL Schedule](#)
[Every Hat Trick in NHL History Added to Hockey Reference](#)
[All-Time Hat Tricks Leaders](#)
[Single-Season Hat Tricks Leaders](#)
[Active Hat Tricks Leaders](#)
[Progressive Hat Tricks Leaders](#)
[Yearly Top 10 in Hat Tricks](#)
[Postseason Hat Tricks Leaders](#)

Upcoming Dates

[April 5-7: 2018 NCAA Frozen Four in St. Paul](#)
[April 7: Final Day of Regular Season](#)
[April 9: NHL Central Scouting Final Rankings release](#)
[View More Items ▾](#)

Trending Player Pages

[Sidney Crosby, Alex Ovechkin, Wayne Gretzky, Evgeni Malkin, Nathan MacKinnon, Jaromír Jagr, Connor McDavid, Artemi Panarin, Bobby Orr, Marc-André Fleury](#)

Questions hockey

- Comment mesurer au mieux les compétences, la valeur, ou la performance d'un joueur ?
- Dans quelle mesure les échanges entre les équipes marchent-ils ?
- Quelle est la trajectoire des performances du joueur à mesure qu'il vieillit ?
- Dans quelle mesure la performance d'un joueur est-elle en corrélation avec la position qu'il a été *drafté* ?

Think outside the box

Questions démographiques

- Les gauchers ont-ils une durée de vie plus courte que les droitiers ?

Questions démographiques

- Les gauchers ont-ils une durée de vie plus courte que les droitiers ?
- À quelle fréquence les gens retournent-ils là où ils sont nés ?

Questions démographiques

- Les gauchers ont-ils une durée de vie plus courte que les droitiers ?
- À quelle fréquence les gens retournent-ils là où ils sont nés ?
- La taille et le poids augmentent-ils dans la population le long du temps ?

Google Ngram

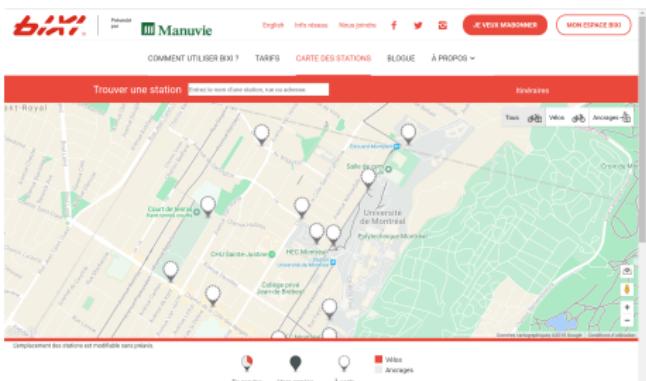
Google Books Ngram Viewer

Questions

- Comment le montant de “gros mots” a-t-il changé avec le temps ?
- Quelle est la durée de vie de la renommée et des technologies ? Est-ce croissant / décroissant ?
- Peut-on détecter quand les mots changent de sens au fil du temps ?
- À quelle fréquence les nouveaux mots apparaissent-ils ? Est-ce qu’ils restent dans l’usage commun ?



- Trip data : pour chaque heure le nombre de départs et d'arrivés dans chaque station.
- Inventory data : pour chaque station les heures où il n'y avait pas de vélos à louer
- Données Canada : les conditions climatiques à chaque heure



Questions

- Comment les conditions climatiques influencent-elles le trafic dans le système ?
- Comment les événements de grande taille influencent-ils le trafic dans le système ?
- Est-ce que la topologie de la ville influence le balancement du système ?
- Comment pouvons-nous diminuer les chances qu'un usager ne réussisse pas à avoir un vélo s'il en a besoin ?

IMDB : Movie Data

IMDb Find Movies, TV shows, Celebrities and more... All

Movies, TV & Showtimes Celebs, Events & Photos News & Community Watchlist

FULL CAST AND CREW | TRIVIA | USER REVIEWS | IMDbPro | MORE | SHARE

The Karate Kid (1984) ★ 7.2 146,087 Rate This

PG | 2h 6min | Action, Drama, Family | 22 June 1984 (USA)

Watch Now With Prime Video

0:17 | Trailer | 1 VIDEO | 71 IMAGES

A martial arts master agrees to teach karate to a bullied teenager.

Director: John G. Avildsen
Writer: Robert Mark Kamen
Stars: Ralph Macchio, Pat Morita, Elisabeth Shue | See full cast & crew »

59 Metascore From metacritic.com | 248 user | 89 critic | Popularity 409 (↑ 173)

IMDb Find Movies, TV shows, Celebrities and more... All

Movies, TV & Showtimes Celebs, Events & Photos News & Community Watchlist

Elisabeth Shue Top 500

Actress | Soundtrack | Producer

Elisabeth Shue was born in Wilmington, Delaware, to Anne Brewster (Wells), who worked for the Chemical Banking Corporation, and James William Shue, a lawyer and real estate developer. She is of German and English ancestry, including descent from Mayflower passengers. Shue's parents divorced while she was in the fourth grade. Owing to the ... See full bio »

Born: October 6, 1963 in Wilmington, Delaware, USA

More at IMDbPro » Contact Info: View agent

454 photos | 72 videos »

Questions

- Peut-on prédire à quel point les gens vont aimer un film ?
Qu'en est-il de ses revenus ?
- À quoi ressemble le réseau social des acteurs ?
- Quelle est la répartition par âge des acteurs et des actrices dans le cinéma ?
- Les stars vivent-elles plus ou moins longtemps que les joueurs de hockey ou le public en général ?