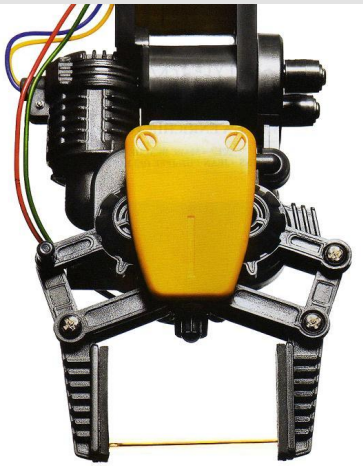
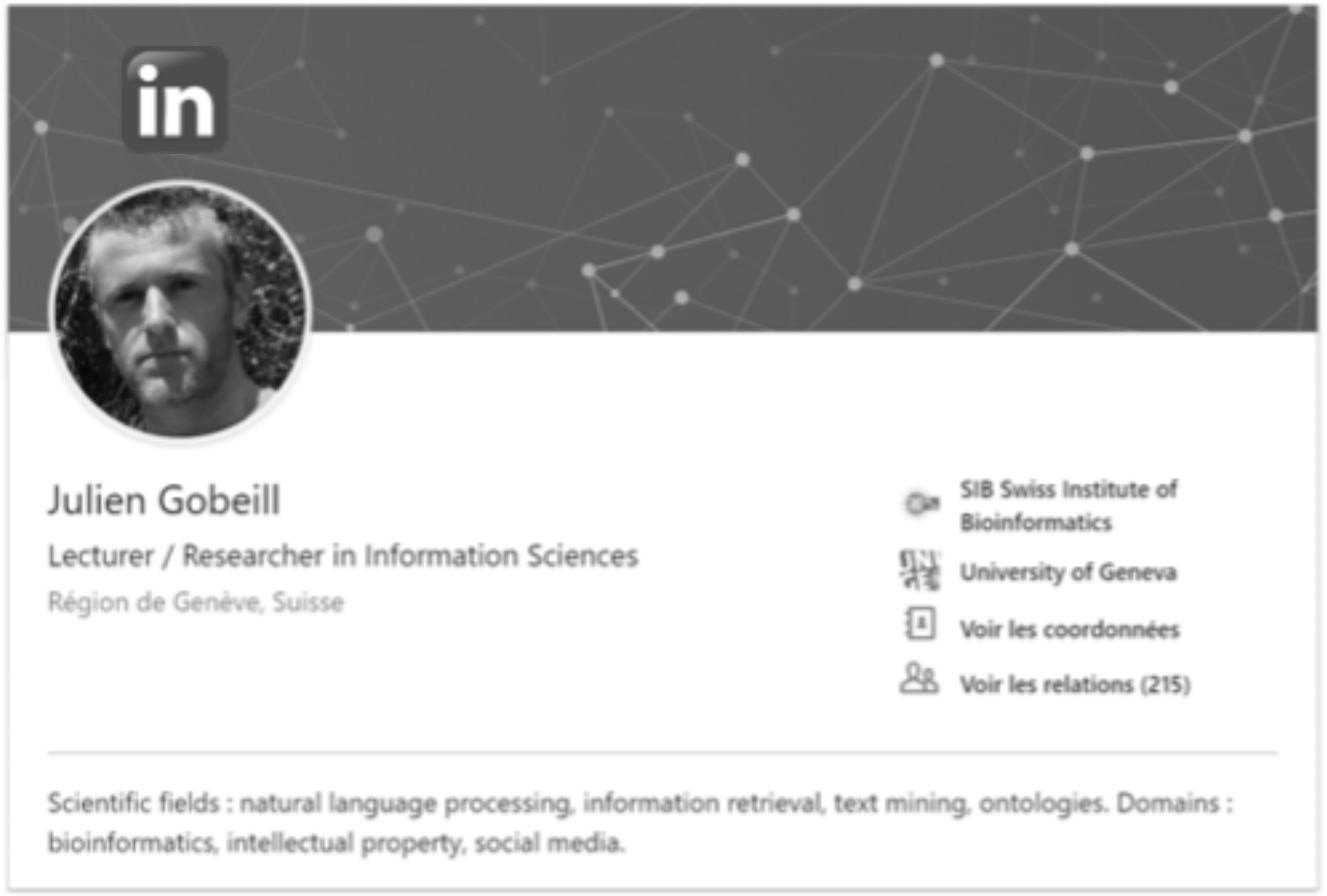
Data scientist

#100 ID – Data Professionals Day

Jeudi 21 juin 2018



Qui suis-je ?



Mon profil, mon domaine, mes activités

Julien Gobeill

Lecturer / Researcher in Information Sciences

Institut Suisse de Bioinformatique, **Haute École de Gestion**

Formation en **informatique**, doctorat en **sciences**

30 %

Chargé de cours en filière ID

Programmation, intro et avancée (data skills)

Moteurs de recherche

Bases de Données

70 %

R&D en bioinformatique

Fouille de données textuelles

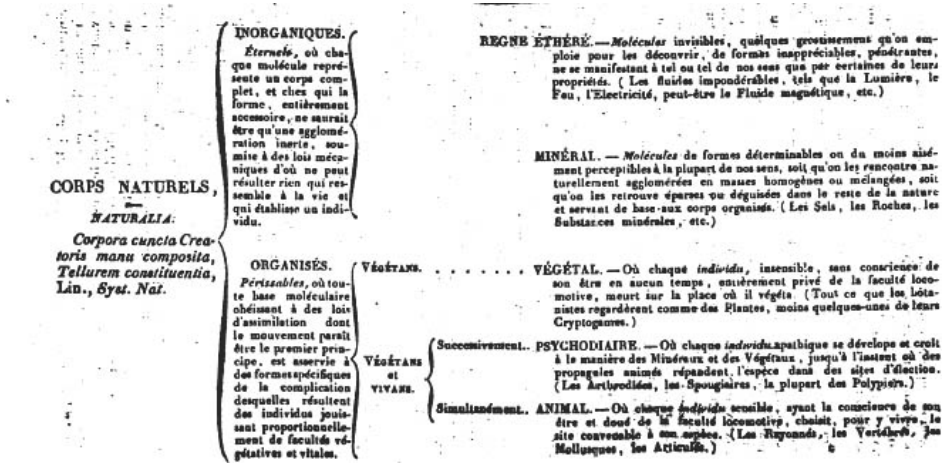
Moteurs de recherche

Bases de données

Les sciences de la vie (biologie / médecine)

- Sciences d'observation
  - Collecter des faits
  - Les décrire en langage naturel
- Sciences de classement
  - Taxonomies, ontologies…

Augustin Pyrame de Candolle



CORPS NATURELS,

NATURALIA:

Corpora cuncta Creatoris manu composita, Tellurem constituentia, Lin., *Syst. Nat.*

INORGANIQUES.

Éternels, où chaque molécule représente un corps complet, et chez qui la forme, entièrement accessoire, ne saurait être qu'une agglomération inerte, soumise à des lois mécaniques d'où ne peut résulter rien qui ressemble à la vie et qui établisse un individu.

ORGANISÉS.

Périssables, où toute base moléculaire obéissant à des lois d'assimilation dont le mouvement paraît être le premier principe, est asservie à des formes spécifiques de la complication desquelles résultent des individus jouissant proportionnellement de facultés végétatives et vitales.

VÉGÉTANS.

VÉGÉTANS et VIVANS.

Successivement.

Simultanément.

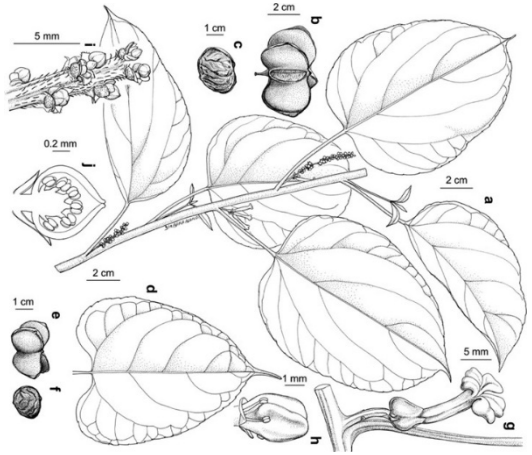
REGNE ÉTHÉRÉ. — *Molécules* invisibles, quelques grossissement qu'on emploie pour les découvrir, de formes inappréciables, pénétrantes, ne se manifestant à tel ou tel de nos sens que par certaines de leurs propriétés. (Les fluides impondérables, tels que la Lumière, le Feu, l'Electricité, peut-être le Fluide magnétique, etc.)

MINÉRAL. — *Molécules* de formes déterminables ou du moins aisément perceptibles à la plupart de nos sens, soit qu'on les rencontre naturellement agglomérées en masses homogènes ou mélangées, soit qu'on les retrouve éparses ou déguisées dans le reste de la nature et servant de base aux corps organisés. (Les Sels, les Roches, les Substances minérales, etc.)

VÉGÉTAL. — Où chaque *individu*, insensible, sans conscience de son être en aucun temps, entièrement privé de la faculté locomotive, meurt sur la place où il végète. (Tout ce que les botanistes regardèrent comme des Plantes, moins quelques-unes de leurs Cryptogames.)

PSYCHODIAIRE. — Où chaque *individu* apathique se développe et croît à la manière des Minéraux et des Végétaux, jusqu'à l'instant où des propagules animés répandent l'espèce dans des sites d'élection. (Les Arthrodiées, les Spongiaires, la plupart des Polypiers.)

ANIMAL. — Où chaque *individu* sensible, ayant la conscience de son être et doué de la faculté locomotive, choisit, pour y vivre, le site convenable à son espèce. (Les Rayonnés, les Vertébrés, les Mollusques, les Articulés.)



Salle informatique à la National Library of Medicine (1976)

MEDLINE (là où je passe beaucoup de temps)

- Base de données bibliographiques gérée par la NLM
  - 6 000 journaux scientifiques
  - 28M de notices bibliographiques (+1M par an)
  - 90% en anglais
- Catalogues en ligne :
  - 1971 : MEDLARS Online
  - 1997 : PubMed (Web)
- **Plateforme publique et ouverte**

Première recherche publique dans MEDLINE

Vice-Président Al Gore et MEDLINE (1997)

Une entrée dans MEDLINE

Journal, date, référence

PLoS One. 2017 Sep 19;12(9):e0183715. doi: 10.1371/journal.pone.0183715. eCollection 2017.

Titre

Assessment of common somatic mutations of EGFR, KRAS, BRAF, NRAS in pulmonary non-small cell carcinoma using iPLEX® HS, a new highly sensitive assay for the MassARRAY® System.

Sutton BC[1], Birse RT[2], Maggert K[1], Ray T[1], Hobbs J[1], Ezenekwe A[1], Kazmierczak J[2], Mosko M[2], Kish J[1], Bullock A[3], Shi Z[3], Stack MS[3], Irwin D[2].

Author information

Auteurs

Résumé

Abstract

Increased early detection and personalized therapy for lung cancer have coincided with greater use of minimally invasive sampling techniques such as endobronchial ultrasound-guided biopsy (EBUS), endoscopic ultrasound-guided biopsy (EUS), and navigational biopsy, as well as thin needle core biopsies. As many lung cancer patients have late stage disease and other comorbidities that make open surgical procedures hazardous, the least invasive biopsy technique with the highest potential specimen yield is now the preferred first diagnostic study. However, use of these less invasive procedures generates significant analytical challenges for the laboratory, such as a requirement for robust detection of low level somatic mutations, particularly when the starting sample is very small or demonstrates few intact tumor cells. In this study, we assessed 179 clinical cases of non-small cell lung carcinoma (NSCLC) that had been previously tested for EGFR, KRAS, NRAS, and BRAF mutations using a novel multiplexed analytic approach that reduces wild-type signal and allows for detection of low mutation load approaching 1%, iPLEX® HS panel for the MassARRAY® System (Agena Bioscience, San Diego, CA). This highly sensitive system identified approximately 10% more KRAS, NRAS, EGFR and BRAF mutations than were detected by the original test platform, which had a sensitivity range of 5-10% variant allele frequency (VAF).

PMID: 28926605 PMCID: PMC5604939 DOI: 10.1371/journal.pone.0183715

[Indexed for MEDLINE] **Free PMC Article**

Identifiants, lien vers l'article

Les mots-clés MeSH (Medical Subject Headings)

- Vocabulaire contrôlé (29 000 concepts)
  - Indexé par les bibliothécaires de la NLM
  - Facilitent la recherche

Le bibliothécaire Stanley Jabonski (1967)

PLoS One. 2017 Sep 19;12(9):e0183715. doi: 10.1371/journal.pone.0183715. eCollection 2017.

Assessment of common somatic mutations of EGFR, KRAS, BRAF, NRAS in pulmonary non-small cell carcinoma using iPLEX® HS, a new highly sensitive assay for the MassARRAY® System.

- Pour l'article précédent :
  - D005838: *Genotype*
  - D008175: *Lung Neoplasms*
  - D048493: *BRAF protein*

Respiratory Tract Diseases [C08]
Respiratory Tract Neoplasms [C08.785]
Laryngeal Neoplasms [C08.785.481]
Lung Neoplasms [C08.785.520]
Bronchial Neoplasms [C08.785.520.100]
Pancoast Syndrome [C08.785.520.734]
Pulmonary Sclerosing Hemangioma [C08.785.520.867]
Nose Neoplasms [C08.785.600]
Tracheal Neoplasms [C08.785.760]

Problème : le déluge de données

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed ▾ lung cancer Search

Help

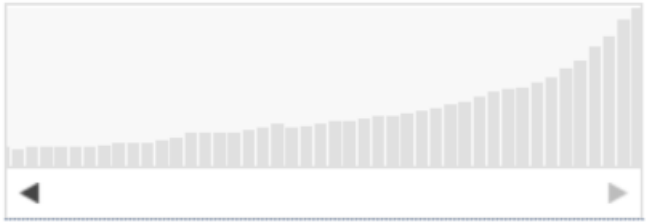
Search results

Items: 1 to 20 of 292591

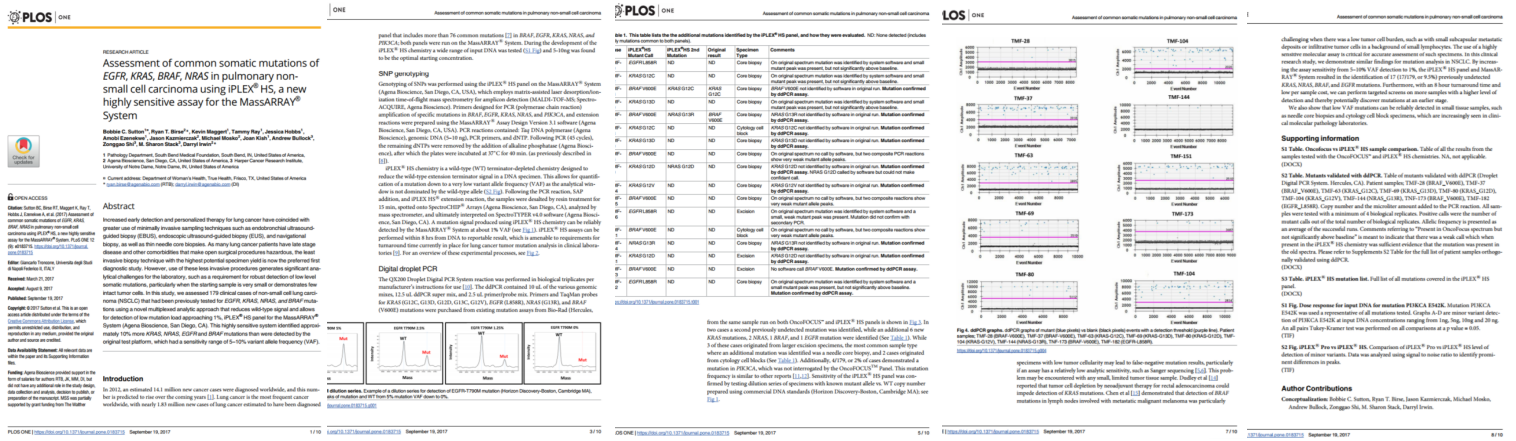
<< First < Prev Page 1 of 14630 Next > Last >>

1. Assessment of common somatic mutations of EGFR, KRAS, BRAF, NRAS in pulmonary non-small cell carcinoma using iPLEX® HS, a new highly sensitive assay for the MassARRAY® System.

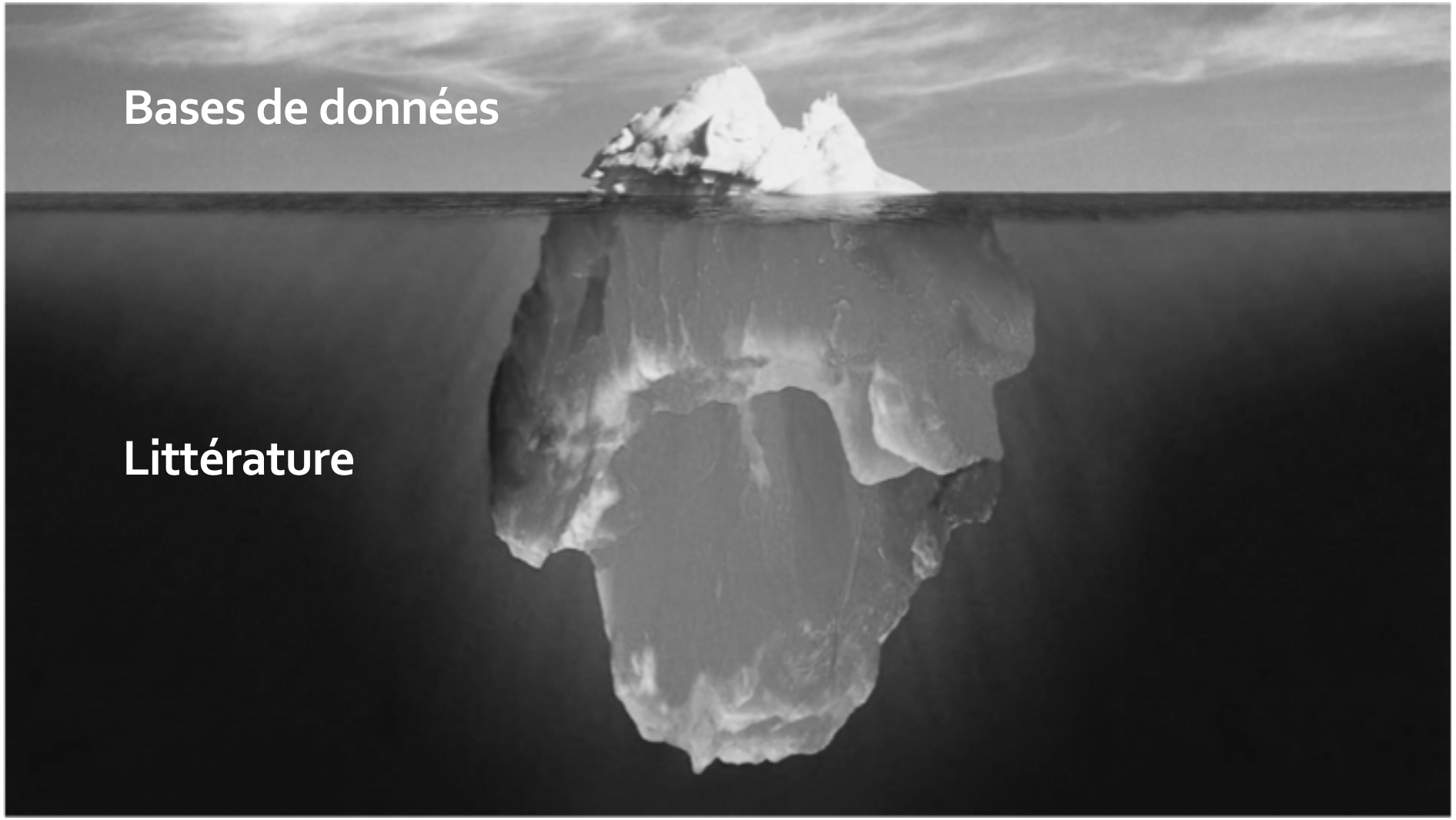
Results by year



44 articles par jour !



Les connaissances en biomédecine



La plate-forme BioMed : moteur de recherche

h e g

Haute école de gestion
Genève

BioMed Quick Start

Introduction

This page provides a description of BioMed, a platform for accessing MEDLINE and PubMed Central entries and annotations. BioMed is maintained by the Text Mining group from the SIB Swiss Institute of Bioinformatics.

MEDLINE search engine

This service provides search services in MEDLINE, either via traditional PubMed engine, or via local vectorial engines.

http://candy.hesge.ch/BioMed/html/MEDLINEsearch.html

MEDLINE citations fetcher

This service provides MEDLINE citations, given a list of PMIDs. Thanks to a local copy of MEDLINE, BioMed MEDLINE fetcher tends to limit access to e-utils and to be really fast.

http://candy.hesge.ch/BioMed/html/MEDLINEfetch.html

La plate-forme BioMed : détection d'entités

ANNOTATIONS FETCHER

This service provides concepts from different terminologies mapped in MEDLINE and PubMed Central entries.

http://candy.hesge.ch/BioMed/html/annotations.html


```
- {
      concept_id: "NX_P15056",
      concept_source: "neXtProt_Genes",
      concept_form: "BRAF"
      offset_begin: "54",
      offset_length: "4",
      passage: "Assessment of common somati
      BRAF, NRAS in pulmonary non-small cel
      a new highly sensitive assay for the
   },
```

```
- {
      concept_id: "D008175",
      concept_source: "MeSH",
      concept_form: "lung cancer",
      offset_begin: "55",
      offset_length: "11",
      passage: "Increased early detection
      lung cancer have coincided with gre
      sampling techniques such as endobro
      (EBUS), endoscopic ultrasound-guide
      biopsy, as well as thin needle core
   },
```


UniProt

Protein | **Serine/threonine-protein kinase B-raf**

Gene | **BRAF**

Organism | *Homo sapiens (Human)*

Pathology & Biotech

Involvement in disease

Defects in BRAF are found in a wide range of cancers. 1 Publication

Lung cancer (LNCR) 1 Publication

The gene represented in this entry is involved in disease pathogenesis.

Détecter automatiquement les maladies

- Une entrée MeSH : concept, synonymes, identifiants

```
*NEWRECORD
CONCEPT = Lung Neoplasms
SYNONYM = Lung Cancers
SYNONYM = Cancer of the Lung
SYNONYM = Pulmonary Cancers
SYNONYM = Pulmonary Neoplasm
DEFINITION = Tumors or cancer of the LUNG.
IDENTIFIER = D008175
```

Détecter automatiquement la protéomique

- Une entrée GO : concept, synonymes, identifiants

```
[Term]
name: positive regulation of the force of heart
contraction by epinephrine-norepinephrine
synonym: "increased strength of cardiac contraction
by epinephrine-norepinephrine"
synonym: "increased inotropy by
epinephrine-norepinephrine"
synonym: "positive regulation of heart contraction
by adrenaline-noradrenaline"
id: GO:0001997
```

GENEONTOLOGY
Unifying Biology

Outils de Text Mining pour prédire

40-year-old male

Lung tumor

Mutation on gene BRAF

Lung Tumor

- Lung Neoplasm
- Lung Cancer
- Pulmonary Tumor
- …

BRAF gene

- Serine/threonine-protein kinase B-raf
- BRAF1
- RAFB1

Trouver des documents pertinents

Trouver des nouveaux traitements pertinents

Qui suis-je vraiment ?

Qu’est-ce qu’un **data scientist** ?

Avant tout : suis-je un data librarian ?

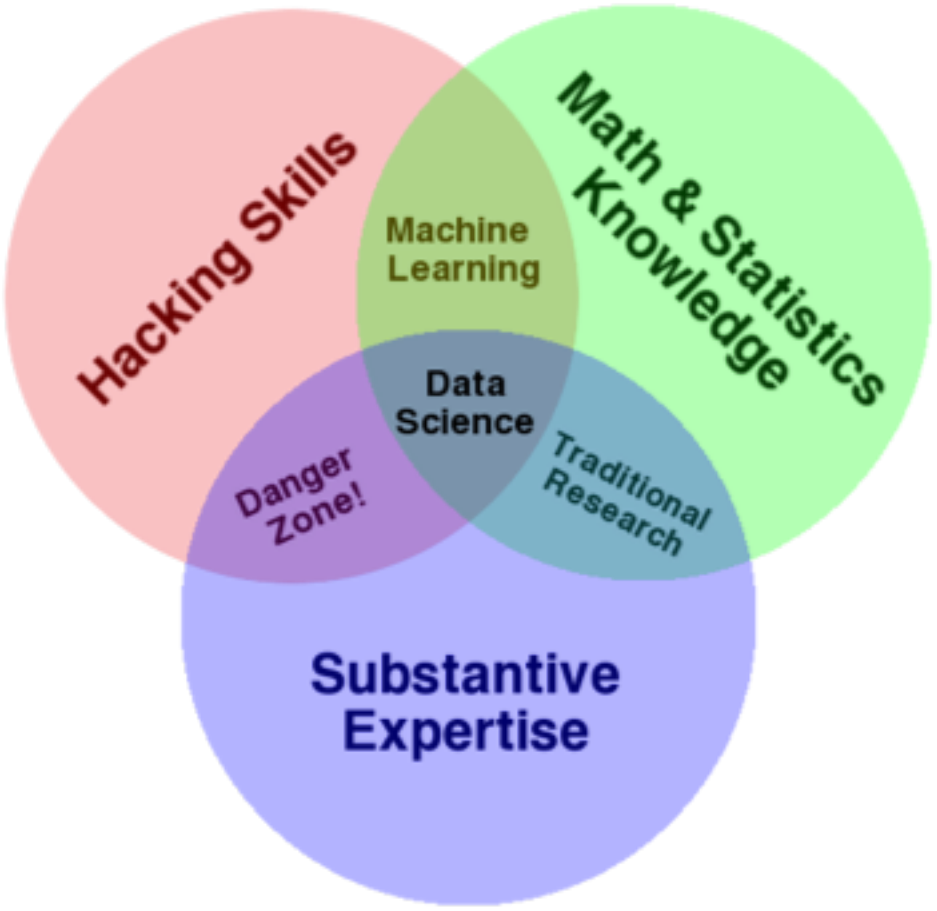
- Collecter / trier / proposer les documents
- Faciliter l'accès de l'utilisateur à la connaissance

Data science pour Wikipedia

- « Extraction de connaissances à partir des données. »
- Techniques et théories tirées de plusieurs domaines.

WIKIPÉDIA
L'encyclopédie libre

Data science Venn diagram (Drew conway)



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Un exemple de data science (Chaney et al.)

- Données :
  - Voyages des marchants de l'âge de Bronze
  - Depuis des tablettes d'argile !
- But :
  - **Détecter des régularités empiriques**

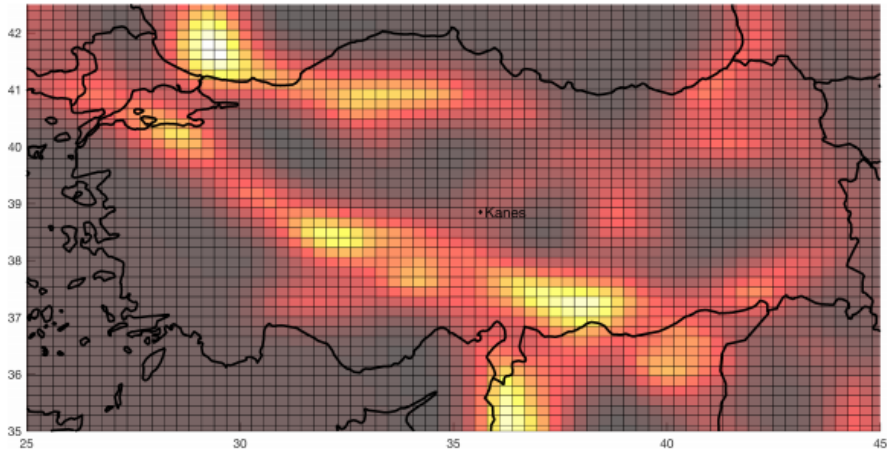


Figure 15: **Natural Roads Scores.**

- Résultats :
  - Villes perdues localisées
  - Hypothèse économique de gravité confirmée : « **les grandes villes émergent au croisement des routes naturelles** »

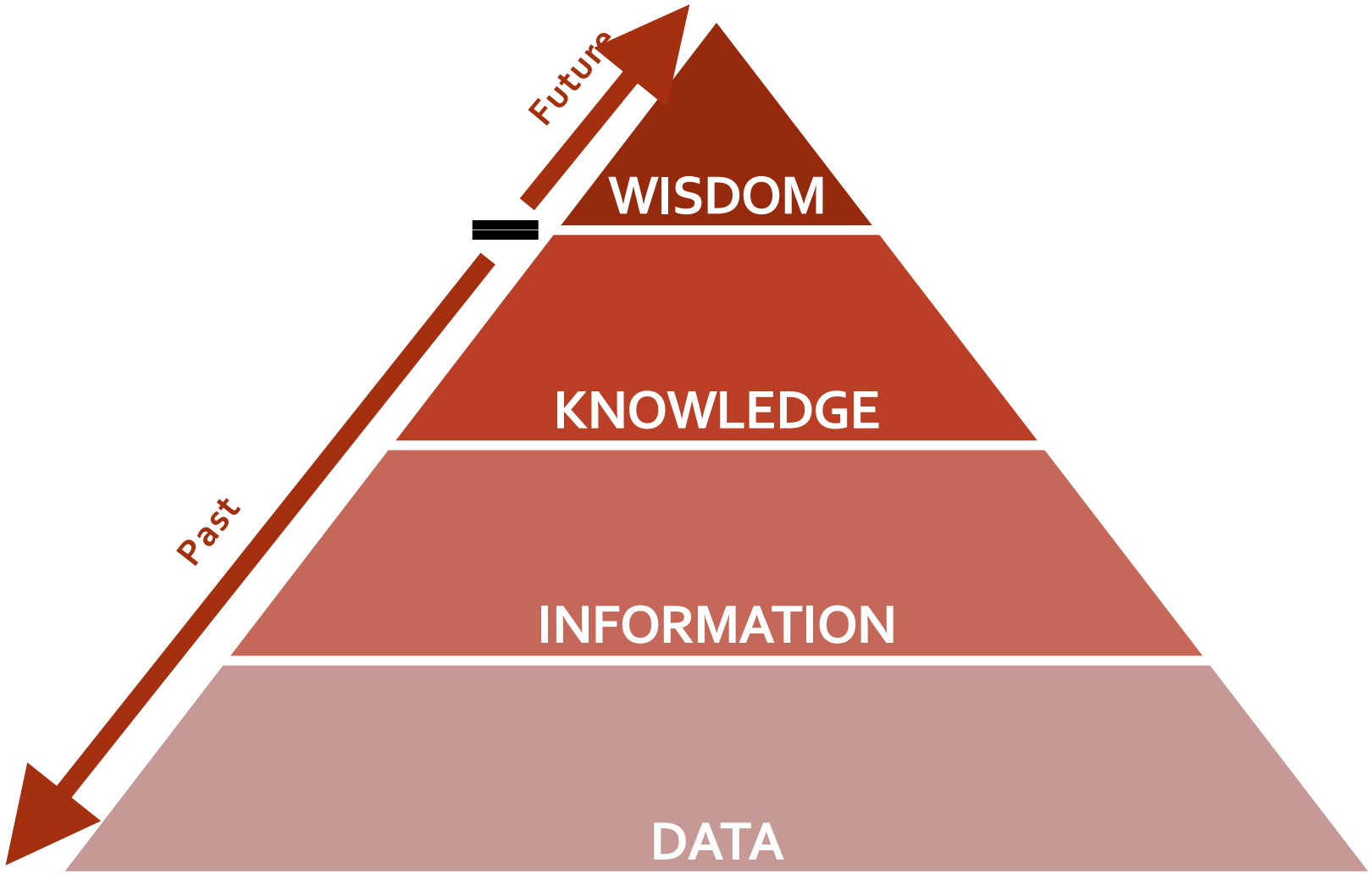
Barjamovic, G., Chaney, T., Coşar, K. A., & Hortaçsu, A. (2017). *Trade, Merchants, and the Lost Cities of the Bronze Age* (No. w23992). National Bureau of Economic Research.

Data Science

Détecter des régularités dans les données pour :

- Comprendre le passé
- Prédire le futur

Le modèle DIKW



Data Wrangling

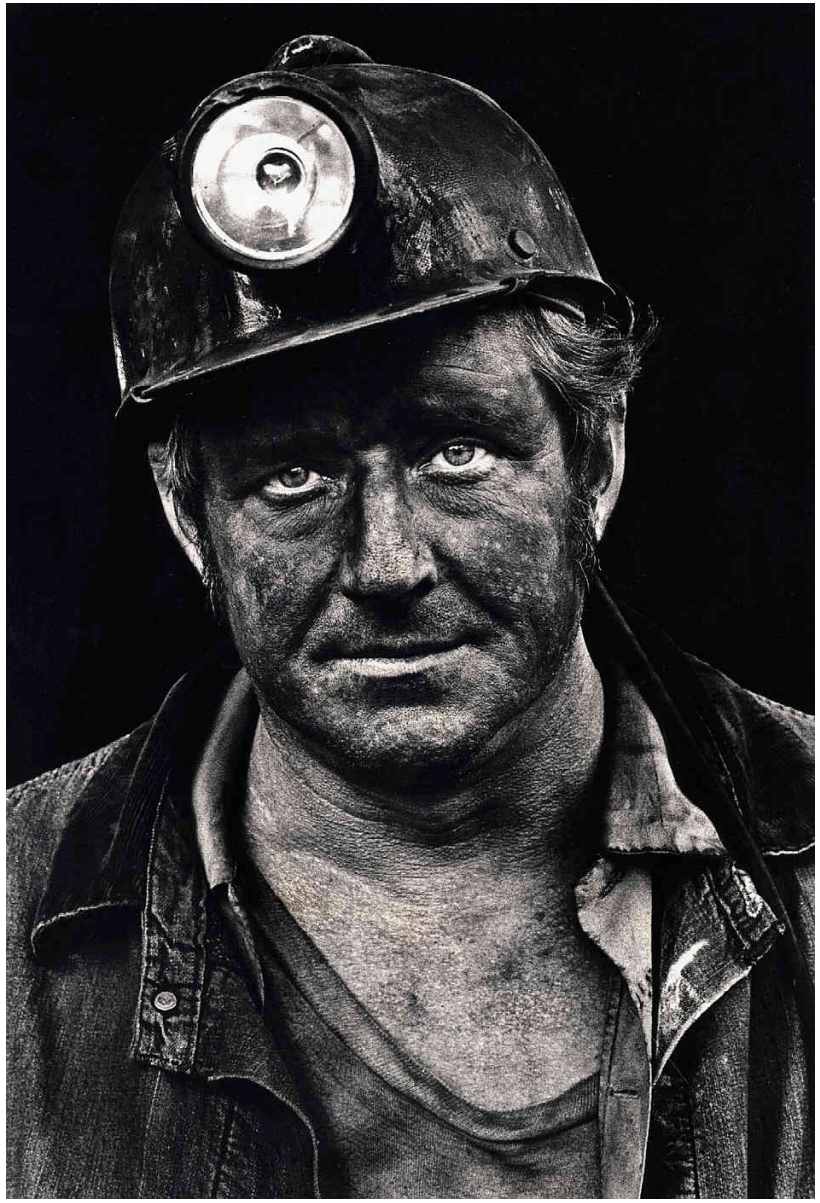
Le « cowboy » de la donnée

From raw to ready data

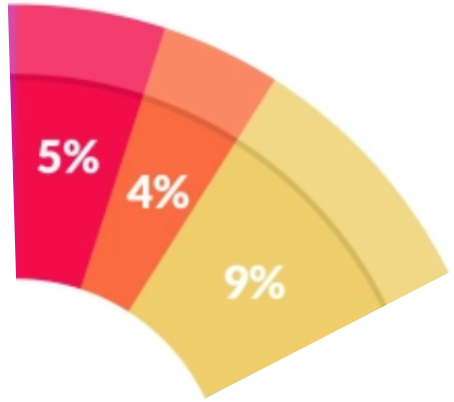
Data Mining

« Fouiller » les données

From data to knowledge



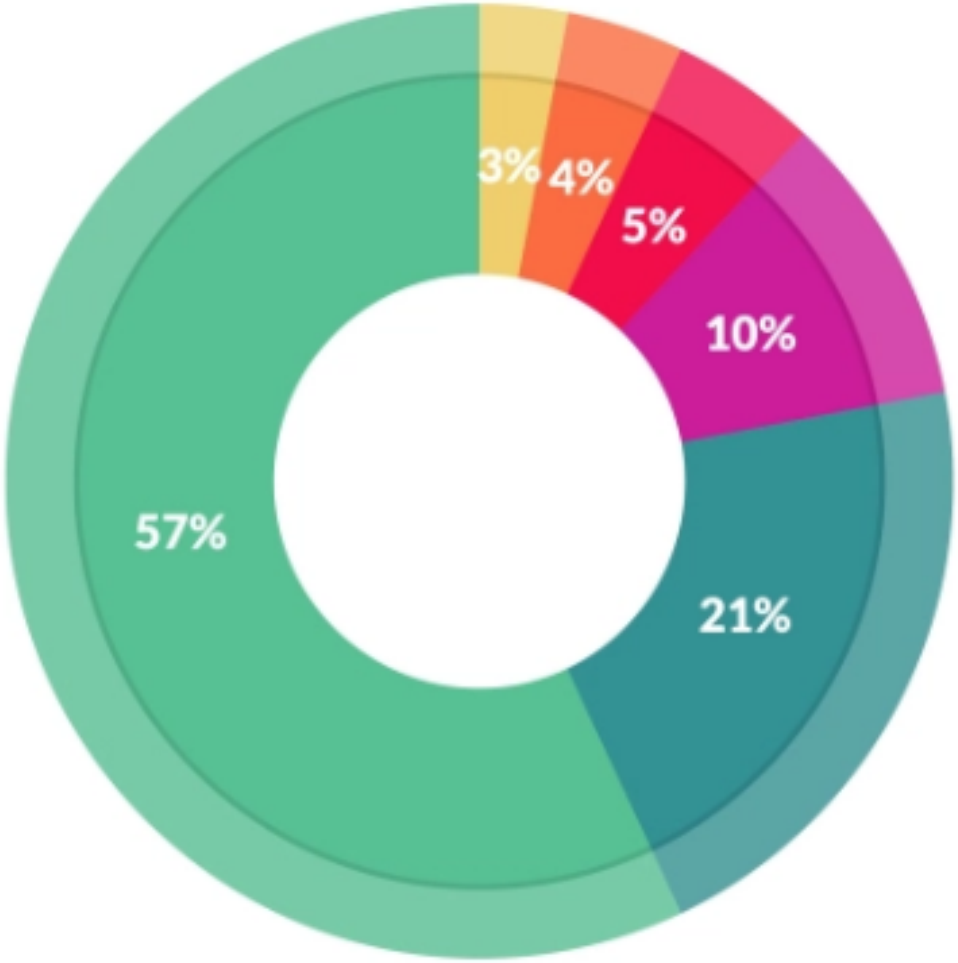
Quelle est la partie la plus chronophage du travail ?



- Fouiller les données: 9%
- Raffiner les algorithmes: 4%
- Autres: 5%
- **Collecter les données: 19%**
- **Nettoyer et organiser les données: 60%**
- **Répartir les données (entraînement/test): 3%**

https://whatsthebigdata.com/2016/05/01/data-scientists-spend-most-of-their-time-cleaning-data/

Quelle est la partie **la moins agréable** du travail ?



- Fouiller les données: 3%
- Raffiner les algorithmes: 4%
- Autres: 5%
- Collecter les données: 21%
- Nettoyer et organiser les données: 57%
- Répartir les données (entraînement/test): 10%

https://whatsthebigdata.com/2016/05/01/data-scientists-spend-most-of-their-time-cleaning-data/

Data Wrangling

Préparer les données

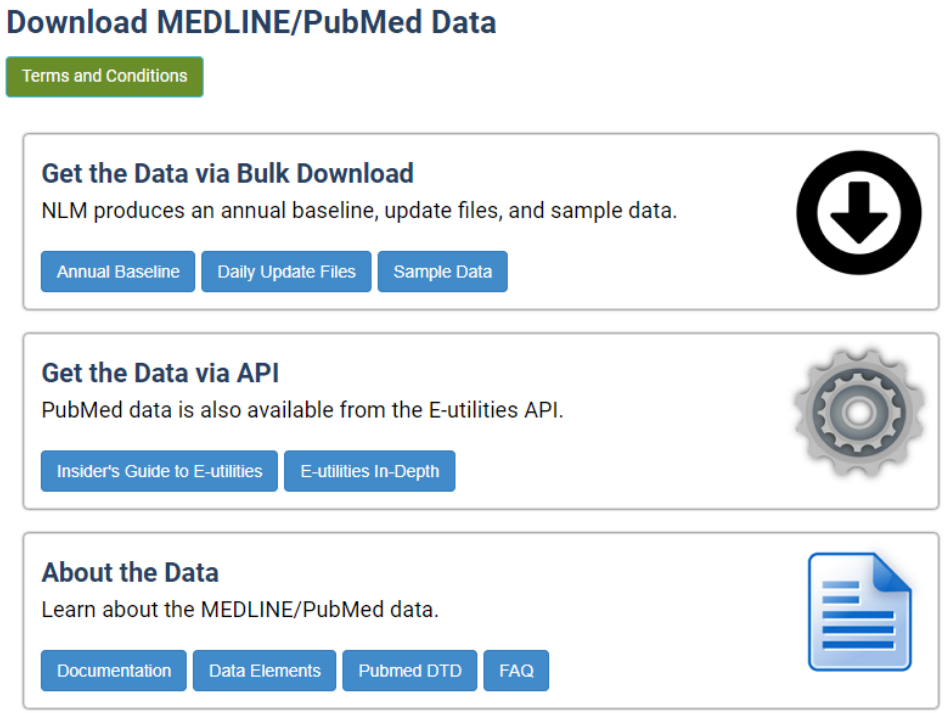
Les étapes du raffinement des données

1. Découvrir
2. Organiser
3. Nettoyer
4. Enrichir
5. Valider
6. Publier

https://www.trifacta.com/blog/quest-ce-que-le-data-wrangling/

Exemple parfait : données biomédicales

- Haut niveau d'interopérabilité (méta-données, identifiants)
- Des données pour la machine (Bulk files, APIs)



Une notice bibliographique de MEDLINE

```
<PubmedArticle>
  <MedlineCitation Status="MEDLINE" Owner="NLM">
    <PMID Version="1">12806098</PMID>
    <Article PubModel="Electronic">
      <Journal>
        <ISSN IssnType="Electronic">1537-744X</ISSN>
        <JournalIssue CitedMedium="Internet">
          <Volume>3</Volume>
          <PubDate>
            <Year>2003</Year>
            <Month>May</Month>
            <Day>05</Day>
          </PubDate>
        </JournalIssue>
        <Title>TheScientificWorldJournal</Title>
        <ISOAbbreviation>ScientificWorldJournal</ISOAbbreviation>
      </Journal>
      <ArticleTitle>Molluscan shell matrix characterization by preparative SDS-PAGE.</ArticleTitle>
      <Pagination>
        <MedlinePgn>342-7</MedlinePgn>
      </Pagination>
      <Abstract>
        <AbstractText>The glycoproteinaceous constituents of molluscan shell matrices usually resist chromatograph
        overcomes this difficulty and permits collection of a large amount of shell proteins for further in vitro
        mineral phase, the glycoproteins are fractionated "blind" on a preparative electrophoresis. They are subse
        raised against the whole matrix.</AbstractText>
      </Abstract>
  <ArticleIdList>
      <ArticleId IdType="pubmed">12806098</ArticleId>
      <ArticleId IdType="doi">10.1100/tsw.2003.30</ArticleId>
      <ArticleId IdType="pmc">PMC5974769</ArticleId>
  </ArticleIdList>
```

Date

Titre

Résumé

DOI

Autre exemple parfait : Wikipedia

Index of wikimedia.org/dumps/enwiki/20180401

Name	Last modified	Size
Parent Directory		-
dumpruninfo.json	2018-04-10 11:11	2.8K
dumpruninfo.txt	2018-04-10 11:11	2.4K
enwiki-20180401-pagelinks.sql.gz	2018-04-02 11:15	5.4G
enwiki-20180401-pages-articles.xml.bz2	2018-04-02 17:19	14G
enwiki-20180401-pages-meta-current.xml.bz2	2018-04-03 08:32	27G
enwiki-20180401-stub-meta-history.xml.gz	2018-04-02 08:36	54G

54 Go compressés !!

Exemple moins parfait : les archives du New York Times

- 2,8M de liens (1 mois de crawling)
- Des données pour l'humain (pages HTML)

Page HTML : 143 000 caractères

```
<!DOCTYPE html>
<!--[if (gt IE 9)|!(IE)]> <!--> <html lang="en" class="no-js section-business
format-medium tone-news app-article page-theme-standard  has-top-ad
type-size-small has-large-lede" itemid="
https://www.nytimes.com/2014/06/14/business/high-end-hermes-handbags-at-center-o
f-suit-against-christies.html" > <!--<![endif]-->
<!--[if IE 9]> <html lang="en" class="no-js ie9 lt-ie10 section-business
format-medium tone-news app-article page-theme-standard  has-top-ad
type-size-small has-large-lede" xmlns:og="http://opengraphprotocol.org/schema/
"> <![endif]-->
<head>
    <title>High-End Hermès Handbags at Center of Suit Against Christie's - The
    New York Times</title>
        <meta http-equiv="X-UA-Compatible" content="IE=edge,chrome=1" />
    <meta name="viewport" content="width=device-width, initial-scale=1,
    maximum-scale=1" />.setTimeout(function(){e(null)},0)},n}(this);
</script>
<script>NYTD.Abra.init([["www-story-reader-satisfaction",[[21474837,"Control",1
],[21474836,"VariantA",1],[4252017623,null,0]]],["www-signup-favor-test",[[
2147483648,"0",0],[2147483648,"1",0]]],'//et.nytimes.com/')</script>
```

Article réel : 1 000 caractères

https://www.nytimes.com/2014/06/14/business/high-end-hermes-handbags-at-center-of-suit-against-christies.html

High-End Hermès Handbags at Center of Suit Against Christie's

Locked behind glass, illuminated like a jewel, lies an Hermès Birkin — totem of wealth, bestower of status, and one seriously expensive handbag.This particular specimen is a coveted Hermès White Himalayan Birkin, dyed in brown and beige crocodile. Gently, if ever, used, it is offered at $115,000.Wait — used?That word is never spoken, not here inside the hushed Midtown Manhattan showroom of Heritage Auctions. The preferred term is "rare" or "vintage" — in this case, applied to a handbag made all the way back in 2013.No one, it is said, knows more about the buying and selling of pre-owned Hermès bags than Matthew Rubinger, Heritage's Birkin whisperer. But now Mr. Rubinger, 26, has left for another more famous auction house — Christie's International — and the battle of the Birkins has begun.Heritage on Friday filed a lawsuit against Christie's, claiming Mr. Rubinger breached his contract and stole trade secrets. Christie's, whose auctions tend to run more to Picasso than purses, is expanding its push into high-end accessories, Heritage's sweet spot. Heritage is seeking a Birkinesque $60 million in damages and lost profits from Christie's, Mr. Rubinger and two other former employees.A spokeswoman for Christie's said, "We have reviewed the complaint and find it to be wholly without merit. We are prepared to vigorously defend these claims and Christie's decision to expand our existing handbag department." Mr. Rubinger did not respond to an email.The brouhaha partly reflects how the world's wealthy are spending their money. But it also underscores how Christie's and its traditional rival, Sotheby's, are coming under pressure. While art prices keep rising, competition among auction houses has squeezed commissions. Daniel S. Loeb, a hedge fund manager who owns 9.6 percent of Sotheby's, got a seat on the company's board after attacking Sotheby's executives for, among other things, giving up profit For Heritage, the flourishing designer handbag

Exemple beaucoup moins parfait : la FDA

- Autorisations de mise sur le marché de médicaments
- Données pour… l'imprimante ☹ (10 000 fichiers **PDF**)

1 INDICATIONS AND USAGE

1.1 Epilepsy

Depakene (valproic acid) is indicated as monotherapy and adjunctive therapy in the treatment of patients with complex partial seizures that occur either in isolation or in association with other types of seizures. Depakene (valproic acid) is indicated for use as sole and adjunctive therapy in the treatment of simple and complex absence seizures, and adjunctively in patients with multiple seizure types which include absence seizures.

Simple absence is defined as very brief clouding of the sensorium or loss of consciousness accompanied by certain generalized epileptic discharges without other detectable clinical signs. Complex absence is the term used when other signs are also present.

Quand tout va bien : PDF textes

1 INDICATIONS AND USAGE

1.1 Epilepsy

Depakene (valproic acid) is indicated as monotherapy and adjunctive therapy in the treatment of patients with complex partial seizures that occur either in isolation or in association with other types of seizures. Depakene (valproic acid) is indicated for use as sole and adjunctive therapy in the treatment of simple and complex absence seizures, and adjunctively in patients with multiple seizure types which include absence seizures.

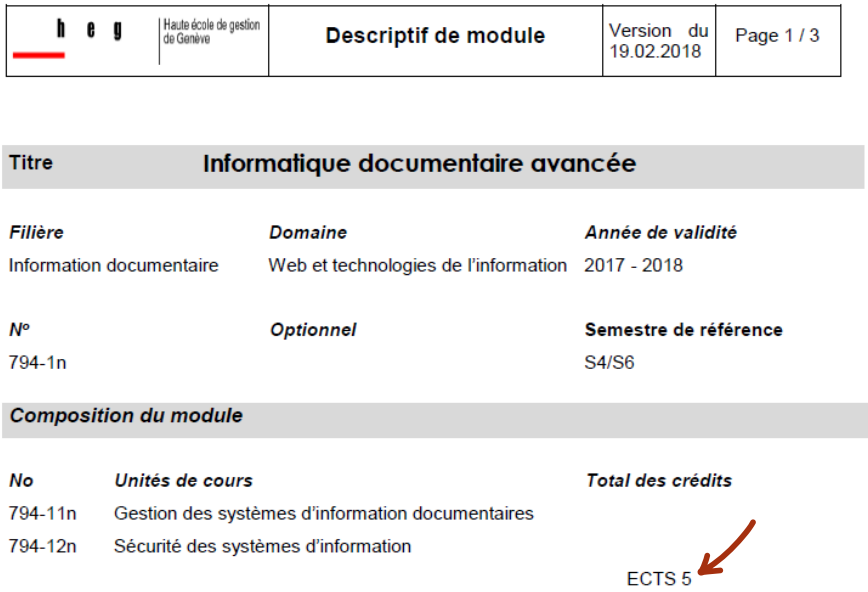
Simple absence is defined as very brief clouding of the sensorium or loss of consciousness accompanied by certain generalized epileptic discharges without other detectable clinical signs. Complex absence is the term used when other signs are also present.

- Extraction de texte (librairies Python, Elasticsearch…) :

```
1 INDICATIONS AND USAGE
1.1 Epilepsy
Depakene (valproic acid) is indicated as monotherapy and adjunctive therapy in
the treatment of patients with complex partial seizures that occur either in
isolation or in association with other types of seizures. Depakene (valproic
acid) is indicated for use as sole and adjunctive therapy in the treatment of
simple and complex absence seizures, and adjunctively in patients with multiple
seizure types which include absence seizures.
Simple absence is defined as very brief clouding of the sensorium or loss of
consciousness accompanied by certain generalized epileptic discharges without
other detectable clinical signs. Complex absence is the term used when other
signs are also present.
```

Plans de cours de la filière ID !

```
<li><a  href=
"/heg/sites/default/files/formation-base/ID/plan-modulaire/741n_descriptif_de_mo
dule_vfinale_2017-2018.pdf" target="_blank" title="document pdf" class=
"hesge-link-pdf">741n Programmation et gestion de projet</a></li>
<li><a  href=
"/heg/sites/default/files/formation-base/ID/plan-modulaire/712n_descriptif_de_mo
dule_v_finale_2017_2018.pdf" target="_blank" class="hesge-link-pdf">712n
Positionnement professionnel I </a></li>
```



heg Haute école de gestion de Genève	Descriptif de module	Version du 19.02.2018	Page 1 / 3

Titre **Informatique documentaire avancée**

Filière	*Domaine*	*Année de validité*
Information documentaire	Web et technologies de l'information	2017 - 2018

N°	*Optionnel*	Semestre de référence
794-1n		S4/S6

Composition du module

No	*Unités de cours*	*Total des crédits*
794-11n	Gestion des systèmes d'information documentaires	
794-12n	Sécurité des systèmes d'information	
		ECTS 5

Descriptif de module Version du 19.02.2018 Page 1 / 3 Module 794-1n Titre Informatique documentaire avancée Filière Domaine Année de validité Information documentaire 2017 - 2018 No Optionnel Semestre de référence 794-1n S4/S6 Prérequis Composition du module No Unités de cours Total des crédits 794-11n 794-12n ECTS 5 Règles de validation du module Module acquis si la moyenne des évaluations des unités de cours est supérieure ou égale à 4. Rattrapage possible si la moyenne est comprise entre 3.9 et 3.5 Module à refaire si la moyenne est inférieure à 3.5. Le résultat de chaque unité de cours doit être égal ou supérieur à 3. Unité de cours nformation documentaires No 794-11n Objectifs

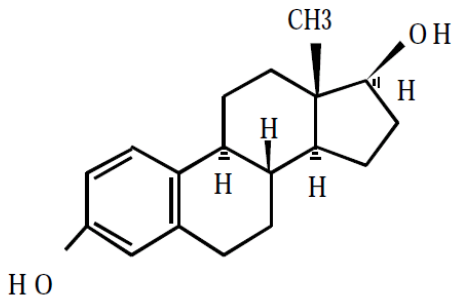
Tableaux, figures, encarts…

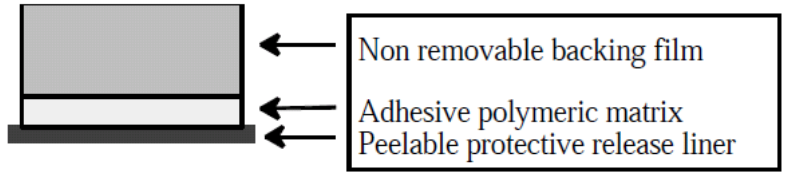
Table 2. Risk by indication for antiepileptic drugs in the pooled analysis

Indication	Placebo Patients with Events Per 1,000 Patients	Drug Patients with Events Per 1,000 Patients	Relative Risk: Incidence of Events in Drug Patients/Incidence in Placebo Patients	Risk Difference: Additional Drug Patients with Events Per 1,000 Patients
Epilepsy	1.0	3.4	3.5	2.4
Psychiatric	5.7	8.5	1.5	2.9
Other	1.0	1.8	1.9	0.9
Total	2.4	4.3	1.8	1.9

Estradiol USP (17ß–estradiol) is a white, crystalline powder, chemically described as estra–1, 3, 5 (10)

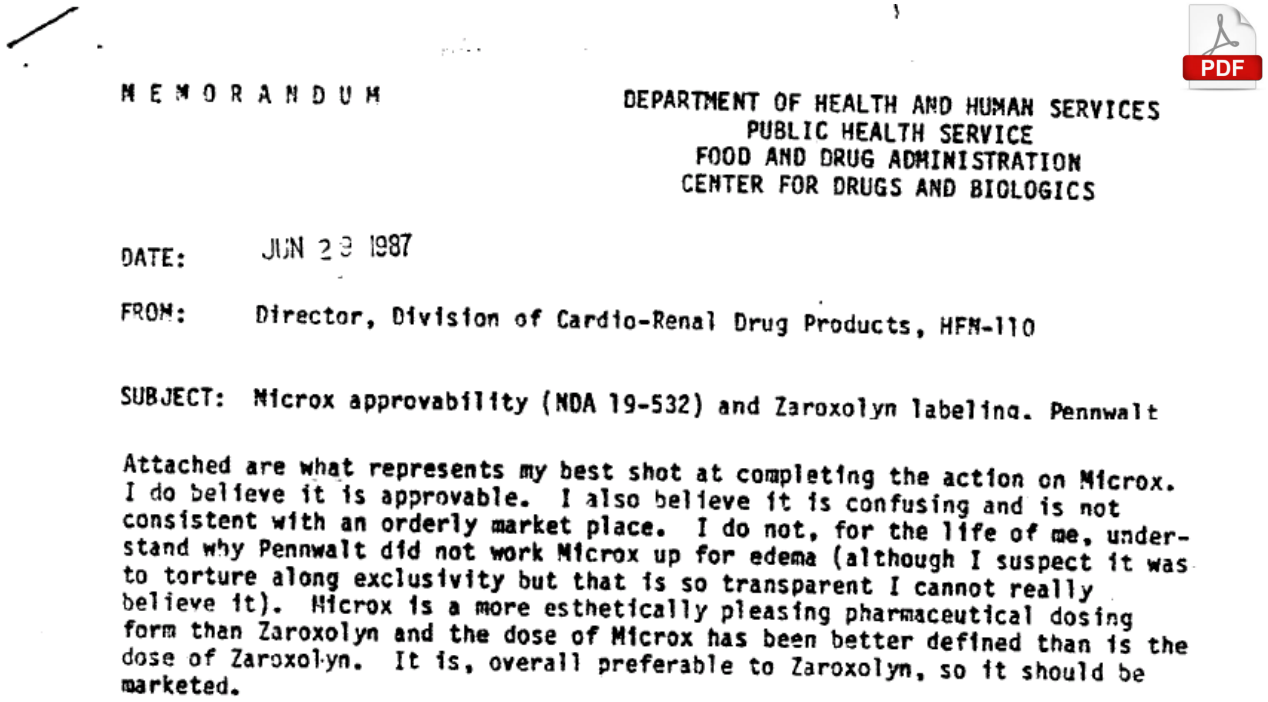
:





The molecular formula of estradiol is C_{18} H_{24} O_2. The molecular weight is 272.39.

PDF images : programmes d'océrisation



MEMORANDUM

DEPARTMENT OF HEALTH AND HUMAN SERVICES
PUBLIC HEALTH SERVICE
FOOD AND DRUG ADMINISTRATION
CENTER FOR DRUGS AND BIOLOGICS

DATE: JUN 2 3 1987

FROM: Director, Division of Cardio-Renal Drug Products, HFN-110

SUBJECT: Microx approvability (NDA 19-532) and Zaroxolyn labeling, Pennwalt

Attached are what represents my best shot at completing the action on Microx. I do believe it is approvable. I also believe it is confusing and is not consistent with an orderly market place. I do not, for the life of me, understand why Pennwalt did not work Microx up for edema (although I suspect it was to torture along exclusivity but that is so transparent I cannot really believe it). Microx is a more esthetically pleasing pharmaceutical dosing form than Zaroxolyn and the dose of Microx has been better defined than is the dose of Zaroxolyn. It is, overall preferable to Zaroxolyn, so it should be marketed.

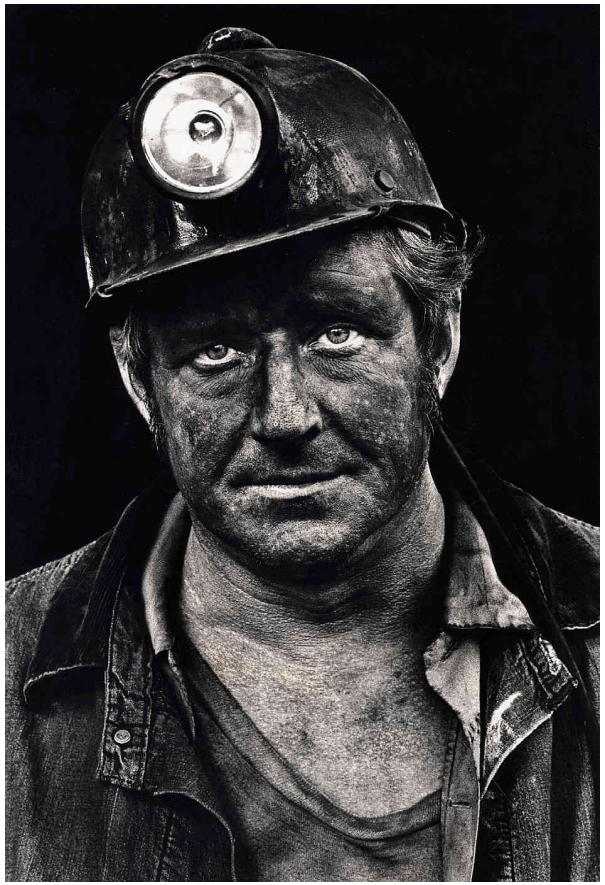
- Technologie mature, mais dépend de la qualité de la numérisation

B. Changes proposed in Pennwall September 26, 1986 letter.

Le pire du pire

WTF ??

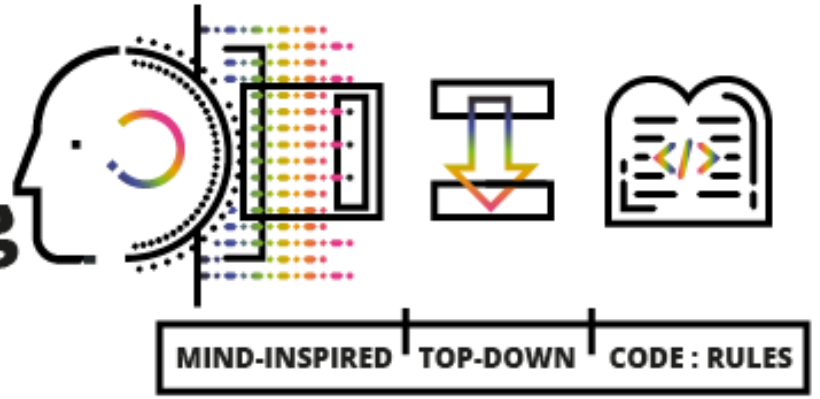
Data Mining



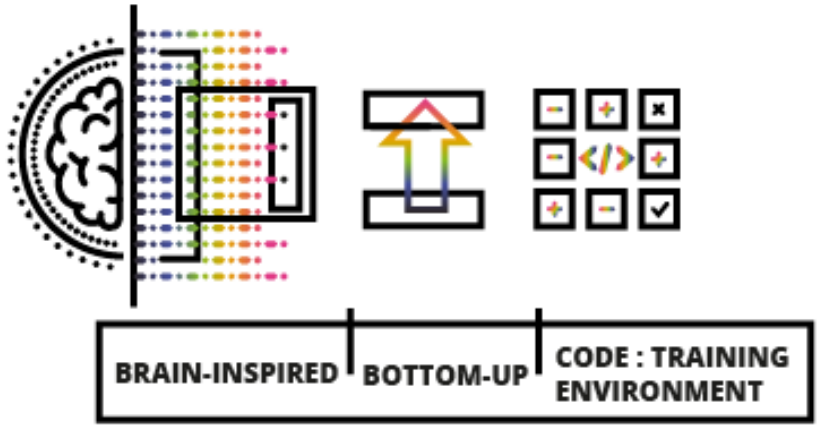
Fouiller les données, trouver un filon

Les deux approches de l'Intelligence Artificielle

1. **Symbolic Programming**

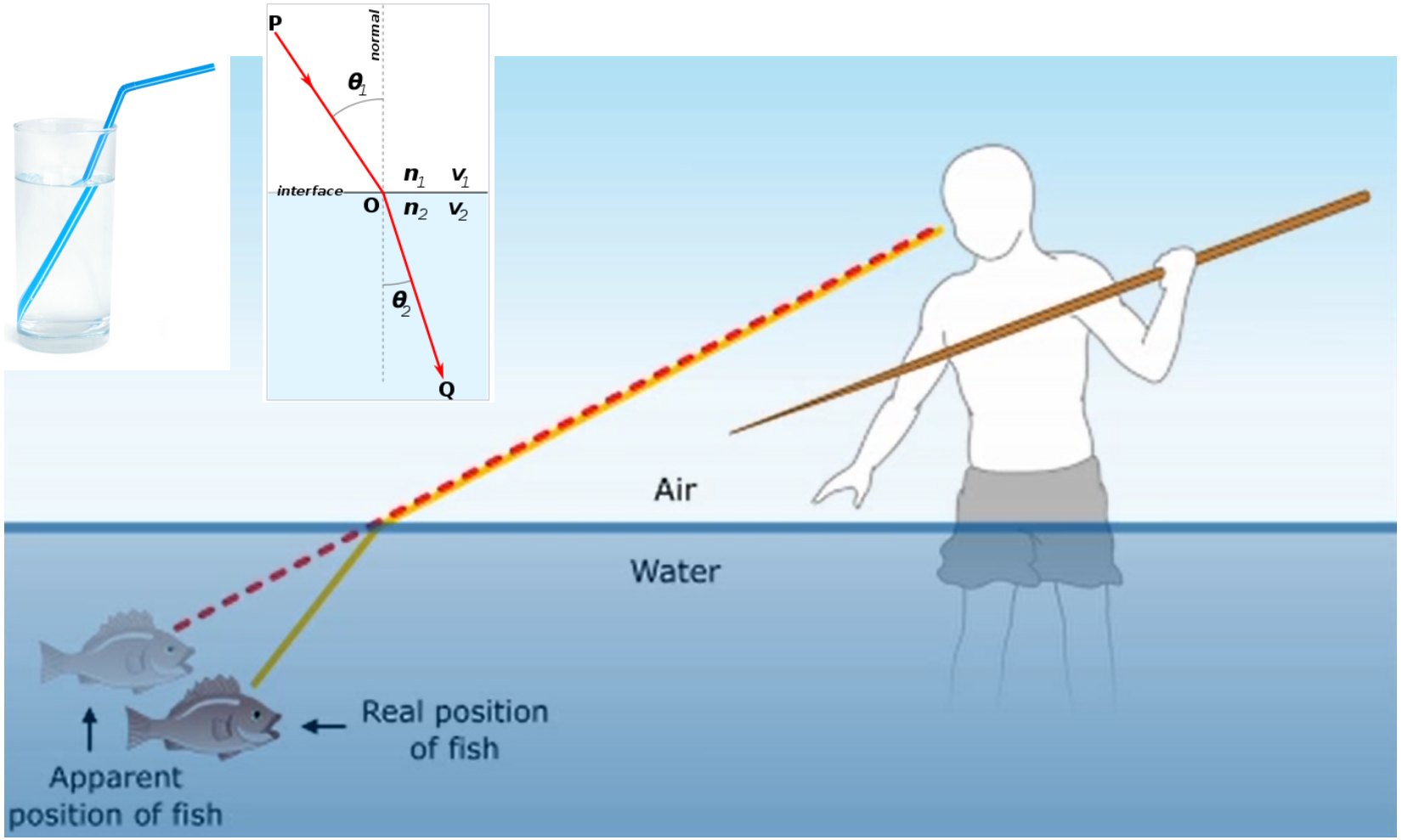


2. **Machine Learning**



https://en.fabernovel.com/insights/tech-en/ai-for-dummies

Exemple : la pêche au harpon



1. Symbolique

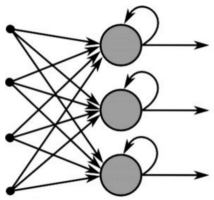
Règles

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{v_1}{v_2} = \frac{\lambda_1}{\lambda_2}$$

2. Apprentissage

Entraînement, données

Réussites, échecs



Évolution de l'Intelligence Artificielle

1. Symbolique

- Populaire dans les années 80
- Règles coûteuses à produire

2. Apprentissage

- A franchi plusieurs « hivers »
- Aujourd'hui : nouvelles méthodes
- Aujourd'hui : **data deluge**

AUJOURD'HUI :

QUALITÉ DES DONNÉES > QUALITÉ DES ALGORITHMES

Bigger data is better

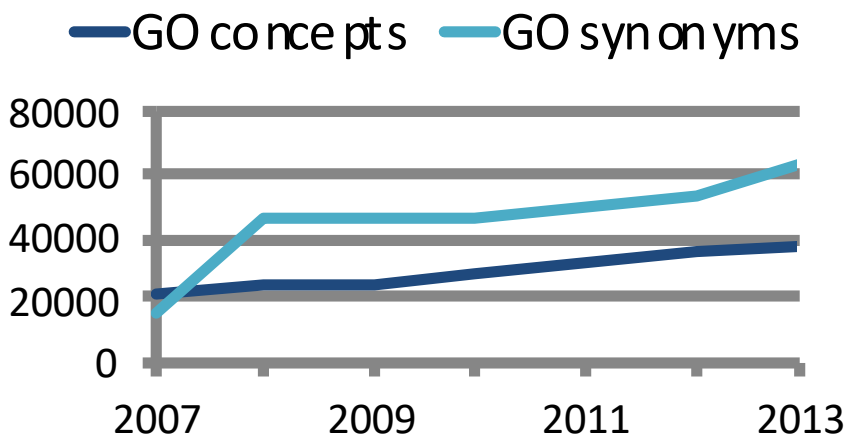
DATABASE
The Journal of Biological Databases and Curation

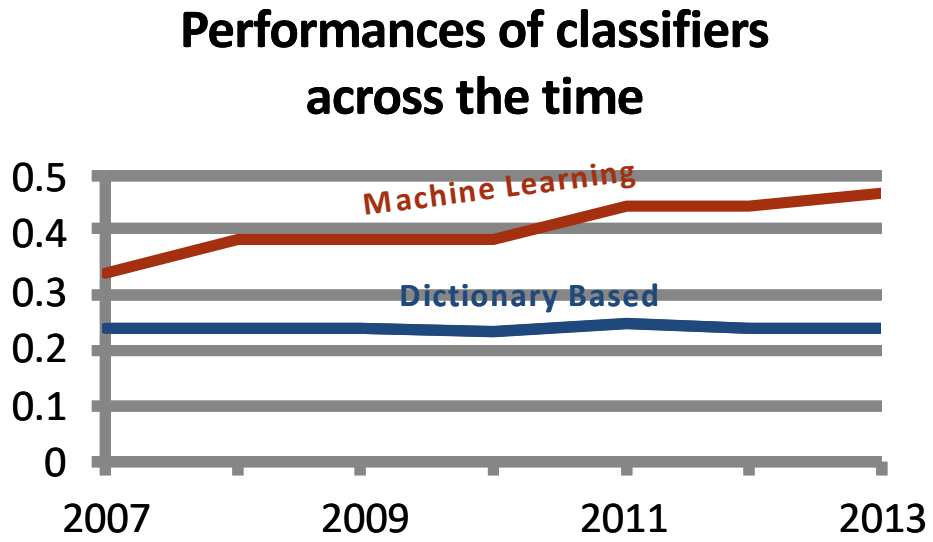
Database, Vol. 2013, Article ID bat041, doi:10.1093/database/bat041

Original article

Managing the data deluge: data-driven GO category assignment improves while complexity of functional annotation increases

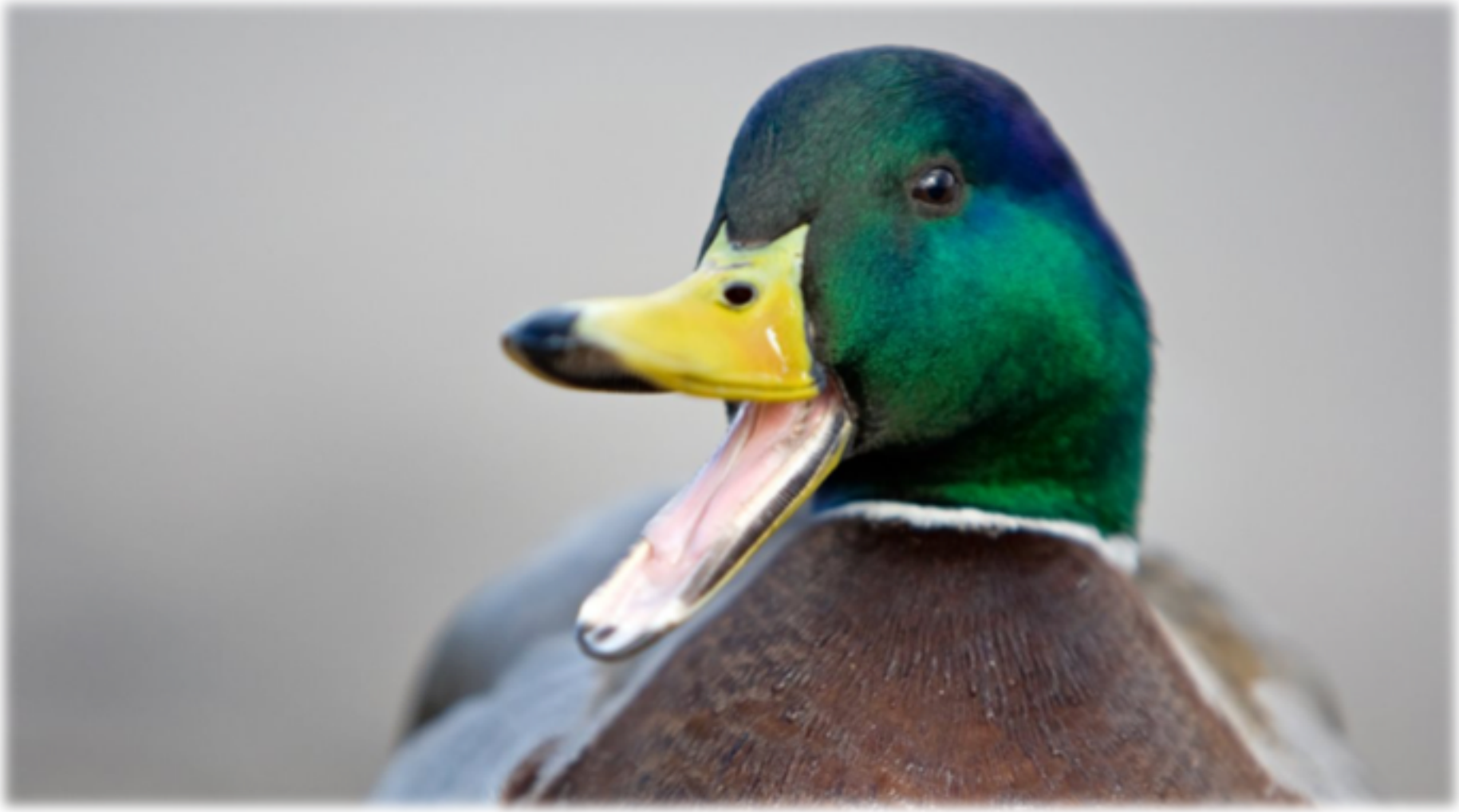
Julien Gobeill[1,*], Emilie Pasche[2], Dina Vishnyakova[2] and Patrick Ruch[1]





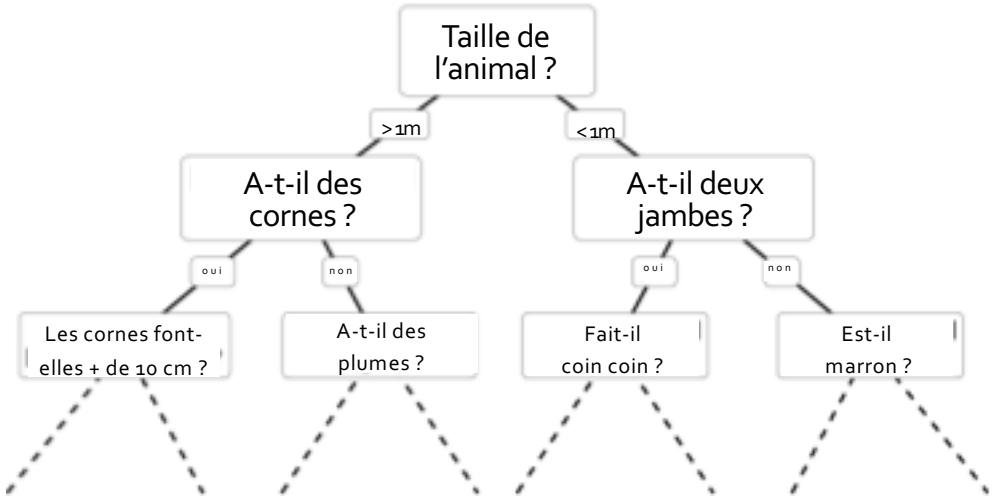
Gobeill, J., Pasche, E., Vishnyakova, D., & Ruch, P. (2013). Managing the data deluge: data-driven GO category assignment improves while complexity of functional annotation increases. *Database*, *2013*.

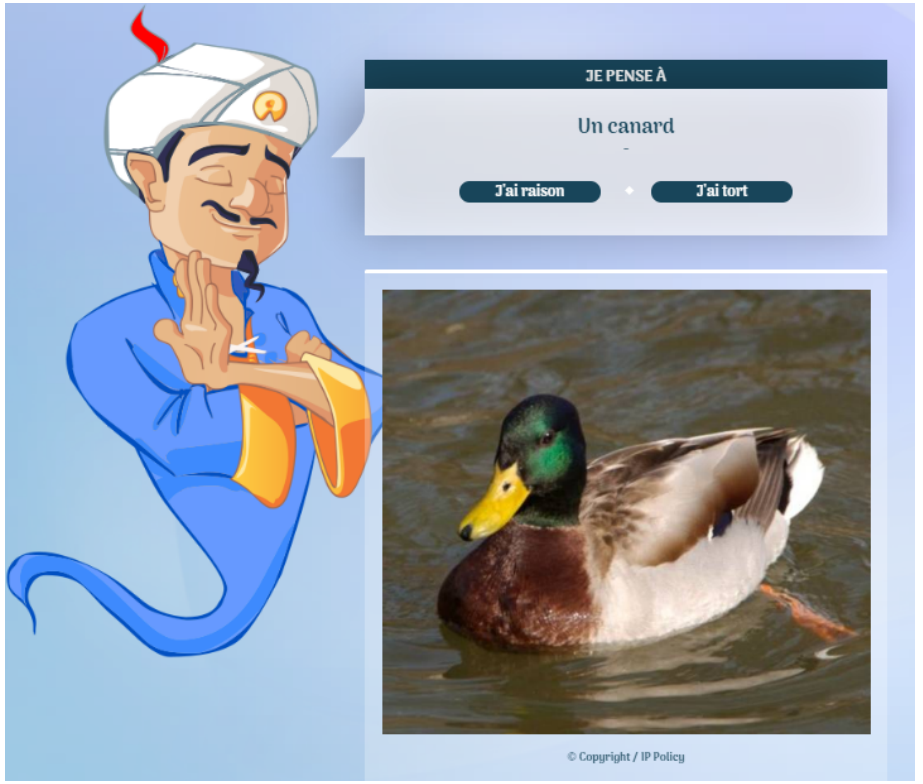
Exemple pratique : reconnaître un canard



1. Approche symbolique

- Règles faites par un expert :





2. Approche machine learning

- Des canards pour apprendre :

- Mais j'ai aussi besoin de non-canards ! (exemples négatifs)

Quels exemples négatifs ?

- Mauvaises données :
  - Des girafes

- Bonnes données :
  - Oiseaux du Léman

Mouettes

Poules d'eau

BONNES DONNÉES = REPRÉSENTATIVES DU MONDE / PROBLÈME RÉEL

Les données

- Capture / représentation du monde réel
  - Propriétés manquantes
  - Données incomplètes, bruit

Individu	Taille cm	Vole	Cou	Bec	Chant	Type
A	50	Oui	Vert	Orange	Coin	Canard
B	70	Oui	Blanc	Orange	Krrr	Cygne
C	40	Oui	Blanc	Orange		Cygne
D	50	Oui	Marron	Noir	Coin	Poule d'eau
E	40	Oui	Marron	Orange	Coin	Canard
F	60	Oui	Vert	Noir		Canard
G	30	Oui	Blanc	Orange	Amouah	Mouette

- Raffinement des données :
  - Nettoyer, filtrer
  - Enrichir

Algorithme d'apprentissage

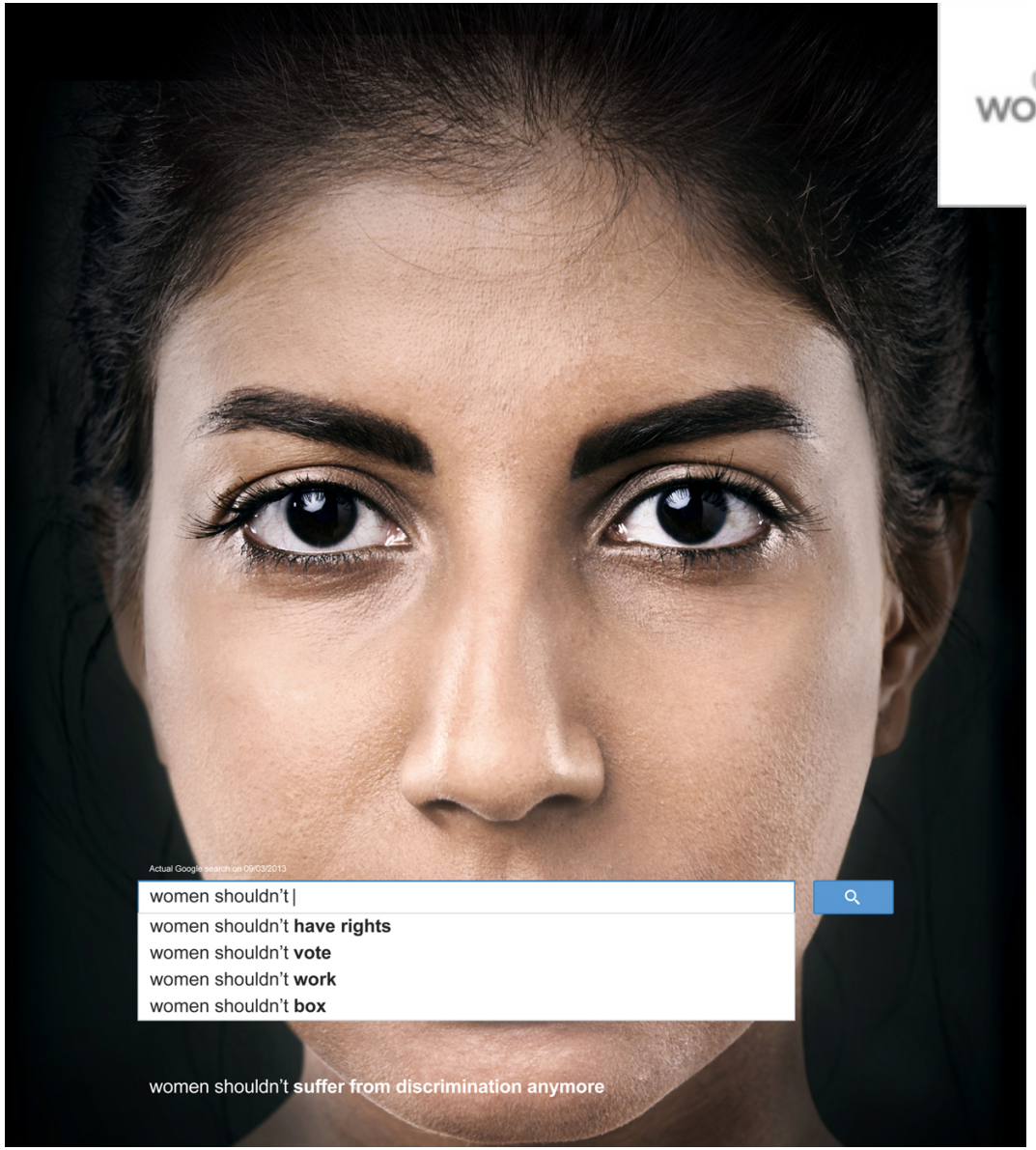
- Deux échantillons représentatifs : entraînement (90%) et test (10%)
- Évaluation scientifique, métriques :
  - Précision / bruit : combien de vrais canards dans les reconnus canards ?
  - Rappel / silence : combien de reconnus canards dans les vrais canards ?

Les limites du dataïsme

- Majorité favorisée : besoin de règles pour corriger
- Données biaisés = biais reproduit par la machine
- Quid de la nouveauté, de l'imprévu ?
- Compréhensibilité de la réponse
  - Prêts bancaires, aide à la décision médicale…

RÈGLES, GARDE-FOUS (APPROCHE HYBRIDE)

PRÉSERVER NOTRE LIBRE-ARBITRE


Actual Google search on 09/03/2013
women shouldn't |
women shouldn't **have rights**
women shouldn't **vote**
women shouldn't **work**
women shouldn't **box**
women shouldn't **suffer from discrimination anymore**

UN WOMEN

Actual Google search on 09/03/2013
women cannot |
women cannot **drive**
women cannot **be bishops**
women cannot **be trusted**
women cannot **speak in church**
women cannot **accept the way things are**

Data librarians / archivists

Comment collaborer ?

Quelle collaboration ?

- Collaborer pour mieux publier et partager les données futures
  - Données pour la recherche, et pour la machine, FAIR
  - Comment gérer les données personnelles ? (ex: génotype, dossier patient)
- Collecter les données ?
  - Qui fait le data crawling ? La conversion de PDF ?
- Raffiner les données ?
  - Qualité du jeu de données, de l'échantillonnage ?
  - Documentation des manipulations effectuées

Quelle place pour le data librarian dans les sciences ?

BMC Medical Informatics and Decision Making

BioMed Central

Proceedings

Open Access

Automatic medical encoding with SNOMED categories

Patrick Ruch*, Julien Gobeill, Christian Lovis and Antoine Geissbühler

Co-auteurs ?

Abstract

Background: In this paper, we describe the design and preliminary evaluation of a new type of tools to speed up the encoding of episodes of care using the SNOMED CT terminology.

Methods: The proposed system can be used either as a search tool to browse the terminology or as a categorization tool to support automatic annotation of textual contents with SNOMED concepts. The general strategy is similar for both tools and is based on the fusion of two complementary retrieval strategies with thesaural resources. The first classification module uses a traditional vector-space retrieval engine which has been fine-tuned for the task, while the second classifier is based on regular variations of the term list. For evaluating the system, we use a sample of MEDLINE. SNOMED CT categories have been restricted to Medical Subject Headings (MeSH) using the SNOMED-MeSH mapping provided by the UMLS (version 2006).

Results: Consistent with previous investigations applied on biomedical terminologies, our results show that performances of the hybrid system are significantly improved as compared to each single module. For top returned concepts, a precision at high ranks (P0) of more than 80% is observed. In addition, a manual and qualitative evaluation on a dozen of MEDLINE abstracts suggests that SNOMED CT could represent an improvement compared to existing medical terminologies such as MeSH.

Conclusion: Although the precision of the SNOMED categorizer seems sufficient to help professional encoders, it is concluded that clinical benchmarks as well as usability studies are needed to assess the impact of our SNOMED encoding method in real settings.

Availabilities: The system is available for research purposes on: http://eagl.unige.ch/SNOCat.

Remerciements ?

Data and metrics

Because MEDLINE is indexed with Medical Subject Headings (MeSH) rather than with SNOMED codes, we need to transform MeSH terms, which are used to index MEDLINE records, into SNOMED terms. This translation is done thanks to the MeSH-SNOMED mapping table provided by the Unified Medical Language System. However, the translation process is not objective: it means that several MeSH terms cannot be mapped appropriately to SNOMED codes and vice versa. Thus, non medically-specific MeSH terms have usually no equivalent in SNOMED. Thus, technical categories methods, such as *Storage, data,*

Authors' contributions

The first two authors: design and evaluation of the tools. The other two authors: determination of the user requirements, and test of the graphic user interface.

Acknowledgements

The study has been partially sponsored by the Swiss National Science Foundation (SNF Grant no 3252B0-105755 – EAGL: an Engine for question Answering in Genomics Literature). The online version of the tool has been made available thanks to the support of the EU DebugIT project (http://www.debugit.eu/); it uses dynamically generated links to the SNOMED

Remercie-ments

- René Schneider
- Les assistants et les enseignants ID
  - Matthieu et Élodie, les étudiants ID2
- L'équipe de recherche BiTeM / SIBtex
  - Patrick Ruch and co
- Vous le public