Université de Montréal

École de **bibliothéconomie** et des **sciences de l'information**

SCI6052 Information documentaire numérique

Cours 3
Recherche d'information, rappel, précision. Logique booléenne, arbres renversés, diagrammes de Venn

Fonction recherche de fichiers de Windows (Windows Search), recherche en texte intégral (début). Recherches séquentielle vs indexée. Fichiers inversés/index. Métadonnées : internes et externes

Au menu aujourd'hui

- Précisions sur le Quiz
- Recherche de fichiers en texte intégral (*début*)
  - Mise en situation
  - Notions préalables sur la recherche d'information
    - Bruit et silence / Précision et rappel
    - Logique booléenne, diagrammes de Venn et arbres renversés
    - Autres opérateurs de recherche
  - Recherche dans le contenu textuel des fichiers
    - Recherches séquentielle & indexée
    - Contenu textuel des fichiers
      - Métadonnées internes et externes
    - Exemple d'un outil de recherche de fichiers en texte intégral intégré au système d'exploitation (OS) : Windows Search (Windows 10)

Université de Montréal

Cours 3 – Objectifs visés, matériel associé et évaluation

Examen intra

OG1 Comprendre les principaux volets de la gestion d'information documentaire numérique

OS f) Expliciter les principes de base de la recherche d'information textuelle

Compétences à développer :

- Pouvoir expliquer les mesures d'efficacité (bruit et silence/rappel et précision) (acétates 14-15)
- Comprendre ce que font les opérateurs booléens (acétates 16, 18-28, trousse d'autoformation sur les opérateurs booléens (StudiUM), trousse sur la gymnastique booléenne (StudiUM))
- Être en mesure de résoudre une requête impliquant des opérateurs booléens en utilisant des diagrammes de Venn et/ou des arbres renversés (acétates 30-38, devoir, exerciseur de diagrammes de Venn)
- Comprendre l'utilisation du masque et de la troncature (acétates 39-40)
- Comprendre à quoi servent les opérateurs de distance (recherche d'expression) (acétate 41)
- Pouvoir expliquer le fonctionnement ainsi que les avantages/désavantages respectifs de la recherche séquentielle et de la recherche indexée (acétates 43-52, trousse sur la construction des index (StudiUM))

OG2 Mettre sur pied des systèmes de gestion d'information documentaire numérique au moyen d'outils représentatifs de la réalité des milieux

OS e) Rechercher efficacement de l'information dans un ensemble de fichiers selon différents critères

Compétences à développer :

- Être en mesure d'expliquer le fonctionnement de la recherche dans le contenu textuel (acétate 53)
- Pouvoir décrire ce que sont les métadonnées d'applications et les métadonnées système (acétates 54-56)
- Comprendre la recherche séquentielle et indexée dans Windows Search (acétates 57-62, TP3, Pratique sur Windows Search (StudiUM))

Cours 3 – Objectifs visés et activités associées

Mise en application

OG1 Comprendre les principaux volets de la gestion d'information documentaire numérique

OS f) Expliciter les principes de base de la recherche d'information textuelle

OG2 Mettre sur pied des systèmes de gestion d'information documentaire numérique au moyen d'outils représentatifs de la réalité des milieux

OS e) Rechercher efficacement de l'information dans un ensemble de fichiers selon différents critères

Activités (en ordre chronologique) :

- Trousse d'autoformation sur les opérateurs booléens (StudiUM) [https://studium.umontreal.ca/course/view.php?id=77216&section=3]
- Exerciseur de diagrammes de Venn [http://marcoux.ebsi.umontreal.ca/enseign/6052/VennEtc/exerciseur.htm]
- Trousse sur la construction des index (StudiUM) [https://studium.umontreal.ca/course/view.php?id=1332317&section=3#trousse_index]
- TP3
- Trousse sur la gymnastique booléenne (StudiUM) *(à faire avec le TP3)* [https://studium.umontreal.ca/course/view.php?id=158948&section=3#trousse_gymnastique]
- Pratique sur Windows Search (StudiUM) (*à faire après le TP3*) [https://studium.umontreal.ca/course/view.php?id=158948&section=3#windows_search]

Université de Montréal

Précisions sur le Quiz [1/2]

- **Moment** : Mardi 1er octobre, de 8h30 à 9h30, en classe
- **Durée** : 1 heure
- **Pondération** : 10%
- **Matière validée** : Compétences informatiques de base, TP1 et TP2 (voir Georges pour détails dans matériel des cours 1 et 2)
  - Ce ne sera pas des questions pointues du type « Dans quel menu retrouve-t-on... » ou « Quel est le raccourci pour... », mais des questions sur la compréhension des opérations effectuées (« À quoi sert... », etc.)
- **Questions de types variés** : vrai ou faux, choix multiples, phrases à compléter, développement court
- **Aucune documentation** ni échange avec collègues

Précisions sur le Quiz [2/2]

- **Rappel des règles**
  - Évitez de parler;
  - Si quelqu'un d'autre que le surveillant vous pose une question, même si ça ne concerne pas l'examen, évitez de répondre. La seule personne à laquelle les étudiants doivent s'adresser est le surveillant;
  - N'ayez en votre possession que le matériel autorisé;
  - Évitez d'emprunter ou de prêter des objets à votre voisin (efface, mouchoir, etc.);
  - Fermez votre téléphone cellulaire, téléavertisseur, radio portative et baladeur durant l'examen. En cas d'oubli de votre part, s'ils sonnent, vous ne pouvez y répondre;
  - Arrivez à l'heure; aucune période supplémentaire ne sera allouée aux retardataires et le surveillant pourra même vous refuser l'accès à la salle d'examen;
  - Aucune sortie n'est autorisée;
  - Ayez en main votre carte étudiante ou une pièce d'identité avec photo.

Recherche de fichiers en texte intégral

Mise en situation [1/3]

- **Contexte**
  - Grande quantité de documents en format numérique (sur le Web, au travail, sur notre ordinateur personnel)
- **Problème**
  - Comment retrouver de l'information dans ces documents sans relire chacun d'eux?
- **Solution**
  - Les outils de recherche de fichiers en texte intégral
    - *Recherche de fichiers :* résultats retournés sont des listes de fichiers et non des extraits des fichiers
    - *Recherche en texte intégral* : recherche se fait sur le contenu textuel des fichiers

Recherche de fichiers en texte intégral

Mise en situation [2/3]

- **Outils**
  - Permettent de faire des recherches dans des fichiers en format numérique
  - *Différentes problématiques* : multiples formats (pdf, doc, htm, txt, etc.), types de documents variés (procès-verbaux, articles, courriels, notes, etc.), fichiers stockés dans des endroits dispersés (poste local, serveur réseau, Web)
  - *Différents types d'outils*
    - *Outils intégrés au système d'exploitation* (par ex., la recherche de fichiers dans Mac OS avec Spotlight, Windows Search pour Windows 10)
    - *Outils indépendants* que l'on installe sur un ordinateur (*Desktop Search Tool*) (par exemple Google Desktop)
    - *Logiciels spécialisés en recherche en texte intégral* (LRTI) (par exemple NatQuest Pro (Agir), dtSearch (dtSearch))

Recherche de fichiers en texte intégral

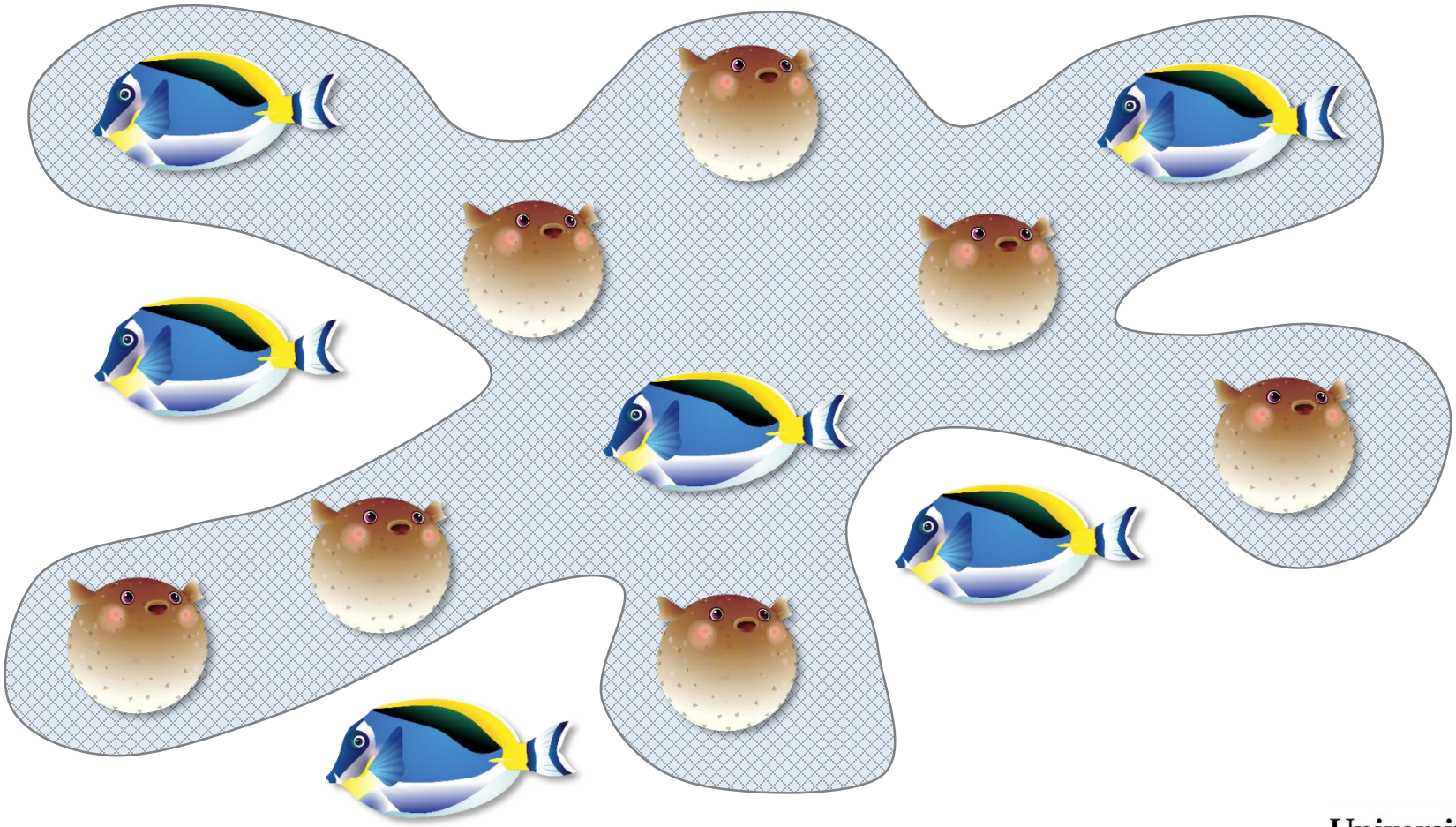
Mise en situation [3/3]

- *Différents types de recherche*
  - Recherche séquentielle & recherche indexée
- *Différents opérateurs de recherche*
  - Opérateurs booléens, opérateurs de distance, etc.

- **Défi pour le chercheur d'information?** Maîtriser les différents outils pour :
  - d'une part, choisir le **bon** outil en fonction des **besoins** et,
  - d'autre part, pouvoir l'utiliser afin de repérer le plus possible toute l'information pertinente et seulement celle-ci!

Notions préalables : Mesures d'efficacité des systèmes de repérage d'information

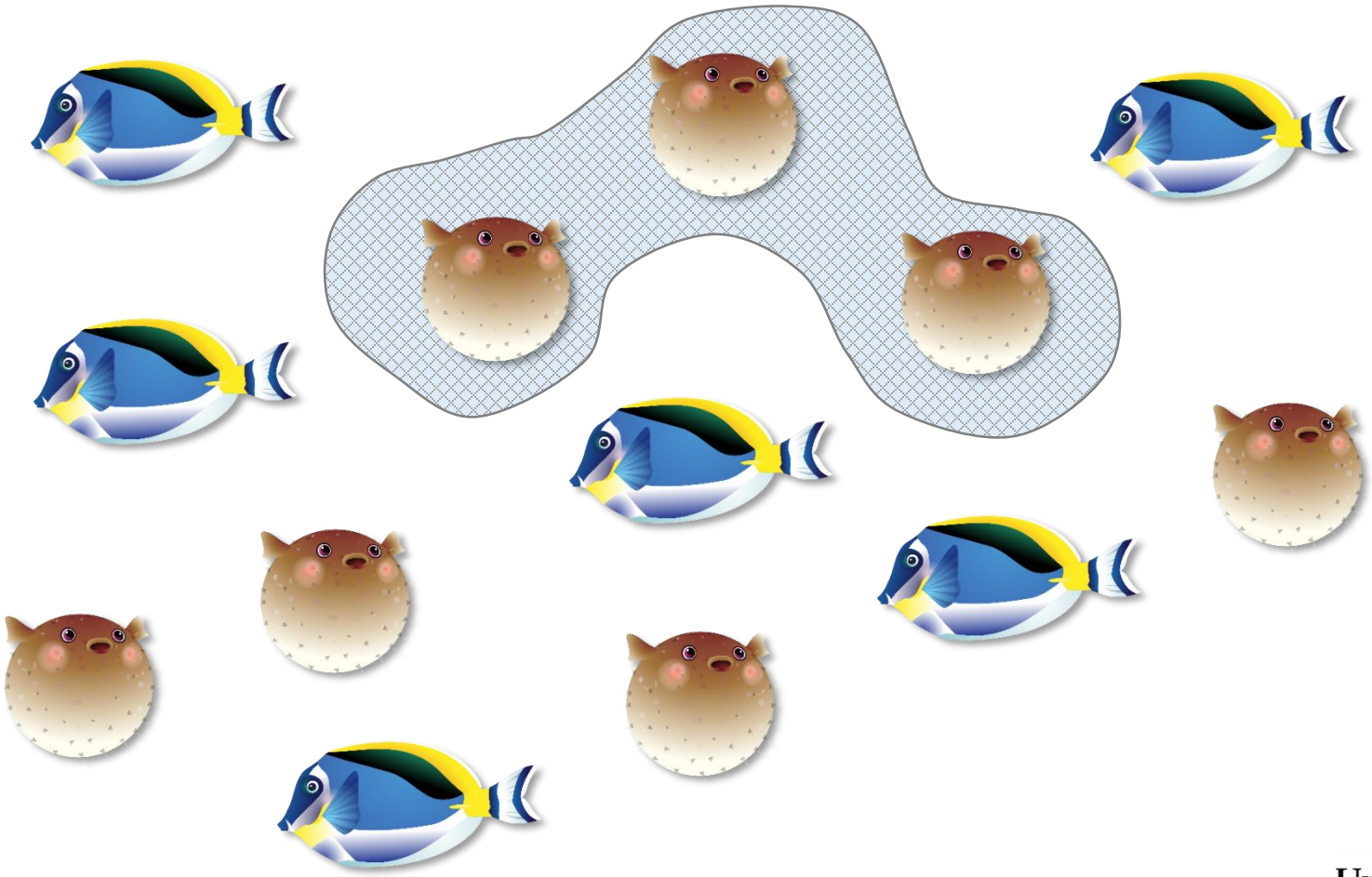
Précision et rappel / Bruit et silence [1/6]



Besoin d'information : les poissons ronds

Notions préalables : Mesures d'efficacité des systèmes de repérage d'information

Précision et rappel / Bruit et silence [2/6]

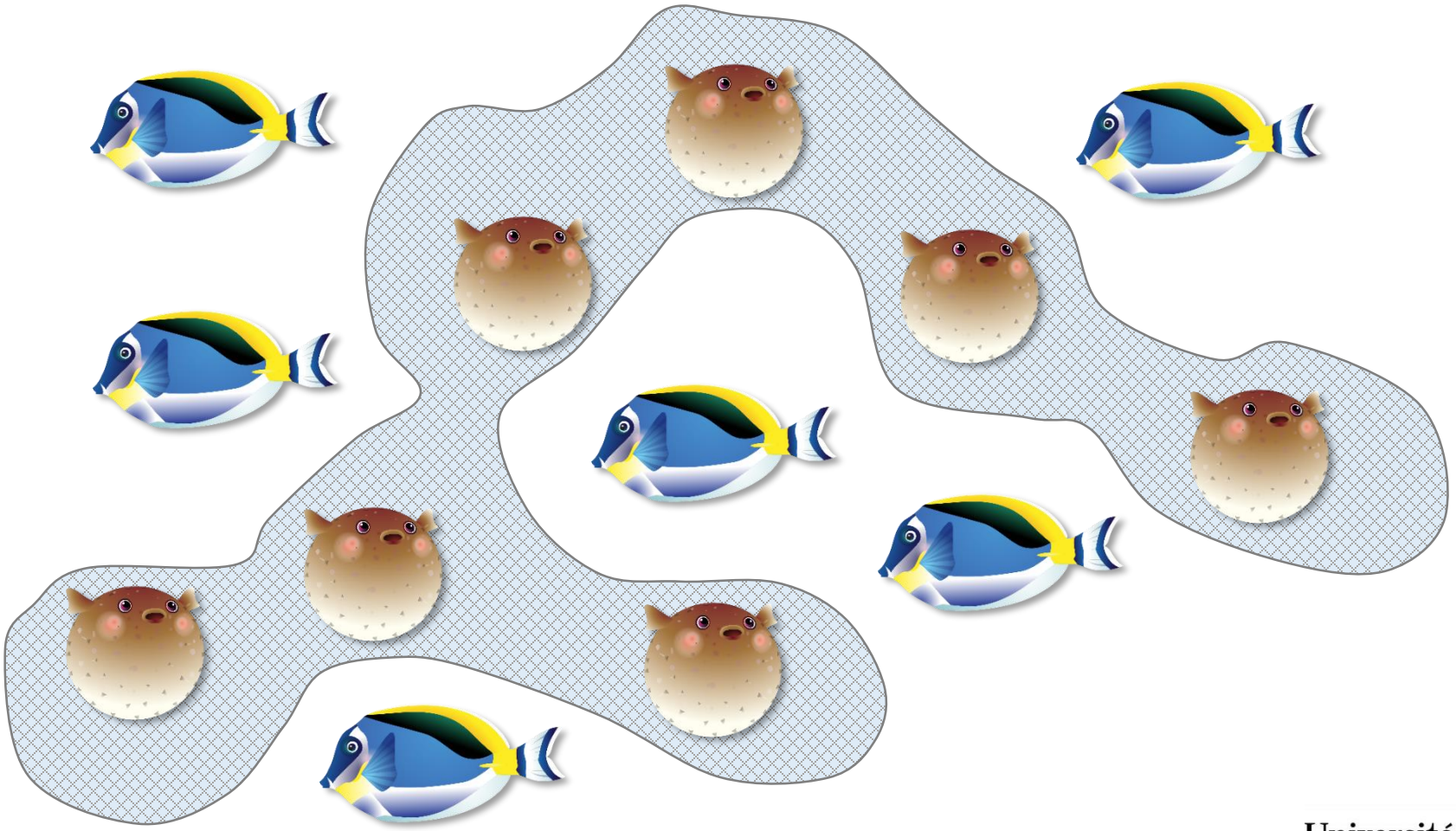


Besoin d'information : les poissons ronds

Université de Montréal

Notions préalables : Mesures d'efficacité des systèmes de repérage d'information

Précision et rappel / Bruit et silence [3/6]

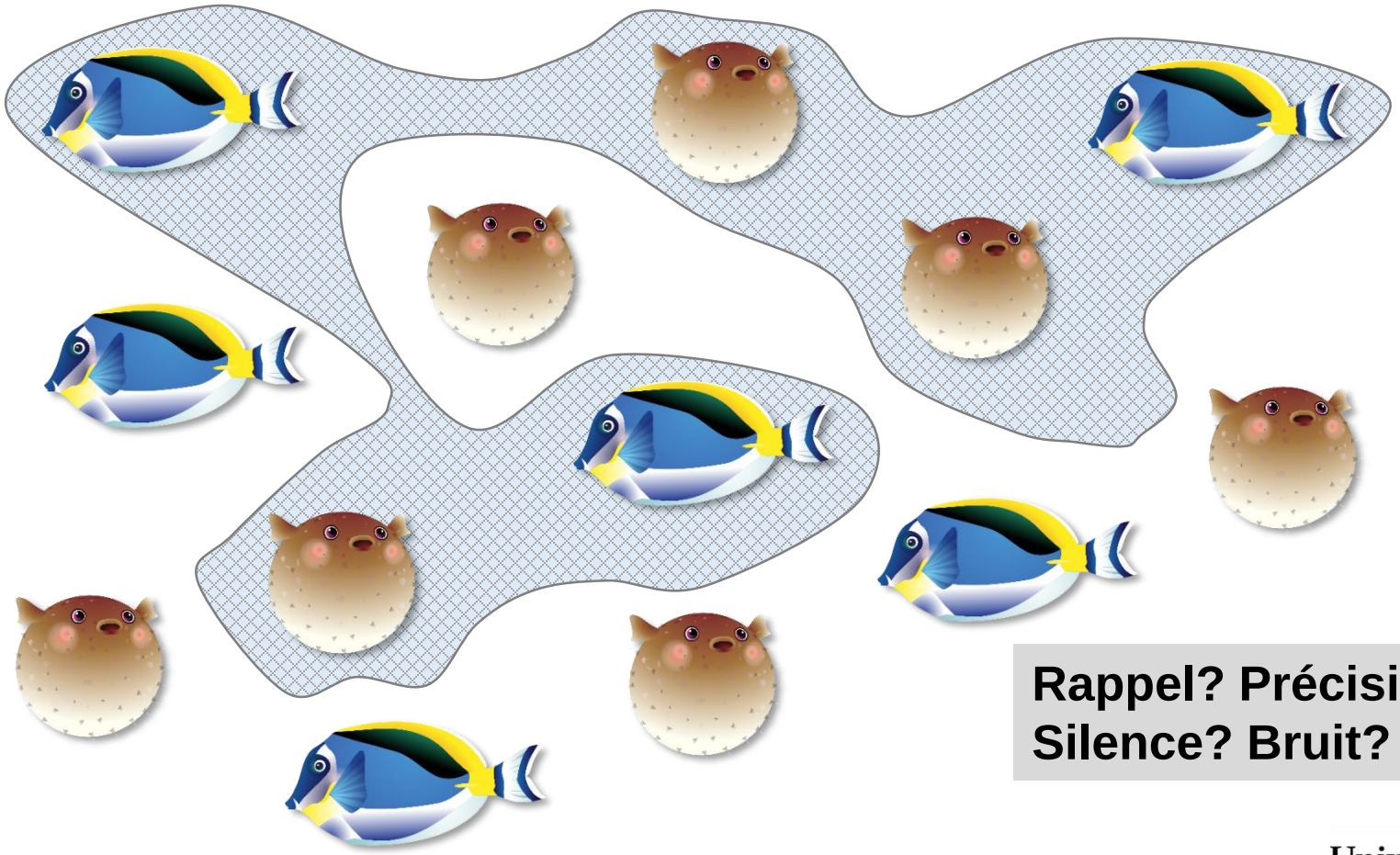


Besoin d'information : les poissons ronds

Université de Montréal

Notions préalables : Mesures d'efficacité des systèmes de repérage d'information

Précision et rappel / Bruit et silence [4/6]



Rappel? Précision?
Silence? Bruit?

Besoin d'information : les jolis poissons

Université de Montréal

Notions préalables : Mesures d'efficacité des systèmes de repérage d'information

Précision et rappel / Bruit et silence [5/6]

- *Objectif ultime de la recherche d'information* : repérer **tout** ce qui est pertinent et **seulement** ce qui est pertinent pour répondre à un besoin d'information
  - **Tout** ce qui est pertinent = notion de *rappel*
  - **Seulement** ce qui est pertinent = notion de *précision*

    ☞ Ainsi notre idéal est une précision maximale ainsi qu'un rappel maximal

- *Attention* : la notion de « pertinence » est **relative**! **Seule** la personne éprouvant le besoin d'information à l'origine de la démarche peut évaluer de manière exacte la pertinence d'un résultat de recherche par rapport à son besoin

Notions préalables : Mesures d'efficacité des systèmes de repérage d'information

Précision et rappel / Bruit et silence [6/6]

- ***Définitions*** : dans *toute* démarche de recherche d'information
  - Information **non pertinente** repérée = ***bruit***
  - Information **pertinente non repérée** = ***silence***
- **Liens** entre précision/rappel et bruit/silence
  - Une recherche ayant une *bonne précision* est une recherche ayant généré peu de bruit
  - Une recherche ayant un *bon rappel* est une recherche ayant généré peu de silence
- **Théoriquement**, il est possible qu'une modification à une démarche de recherche (ex. : modifier une requête de recherche) augmente à la fois le rappel *et* la précision
- **En pratique**, toutefois, une mesure pour augmenter le rappel (ex. : changer un opérateur de recherche par un autre) diminue souvent la précision, et vice-versa…
- La précision maximale et le rappel maximal sont ainsi un **idéal** qui n'est pas toujours atteignable

Université de Montréal

Notions préalables : Logique booléenne

Opérateurs booléens [1/3]

Opérateur & Exemple(s)	Définition	Utilité
ET violence ET films informatique ET documentaire	Résultats doivent contenir **tous** les termes de recherche	Permet de préciser une recherche (*Peut augmenter la précision*)
OU québec OU suisse arbre OU tree	Résultats doivent contenir **au moins un** des termes de recherche	Permet d'élargir une recherche (*Peut augmenter le rappel*)
SAUF kiwi SAUF oiseau montréal SAUF france	Résultats **ne doivent pas** contenir un terme de recherche	Permet de préciser une recherche (*Peut augmenter la précision*)

Notions préalables : Logique booléenne

Opérateurs booléens [2/3]

- Illustration des opérateurs booléens
  Levez la main…
  - Les personnes portant des lunettes rondes? 
  - Les gauchers? 
  - Les personnes portant des lunettes SAUF les gauchers?
  - Les personnes gauchères portant des lunettes?
  - Les personnes portant des lunettes OU les gauchers?
  - Les personnes portant des lunettes ET les gauchers?

Notions préalables : Logique booléenne

Opérateurs booléens [3/3]

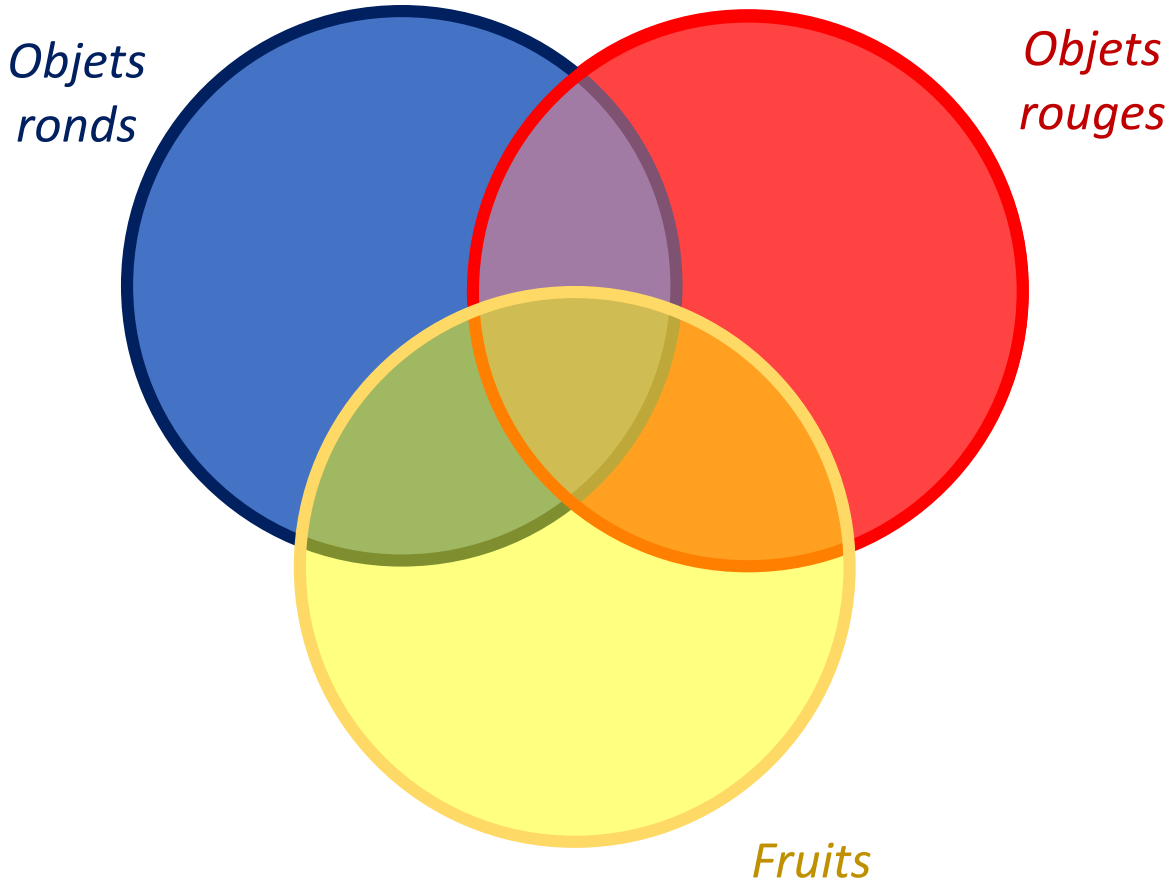
- ATTENTION! « ET » et « OU » dans la **langue courante** n'ont pas toujours la même signification qu'en **logique booléenne**. Par exemple :
  - Pour avoir "tout ce qui s'est publié au Québec *et* en France" il faut utiliser un « OU » booléen sur le lieu de publication

- En logique booléenne, le « OU » est inclusif (et non exclusif) : « chien OU chat » veut dire « chien », « chat » ou « chien ET chat »
  - *Exclusif* : l'un ou l'autre mais non les deux

Notions préalables : Logique booléenne

Diagrammes de Venn [1/7]

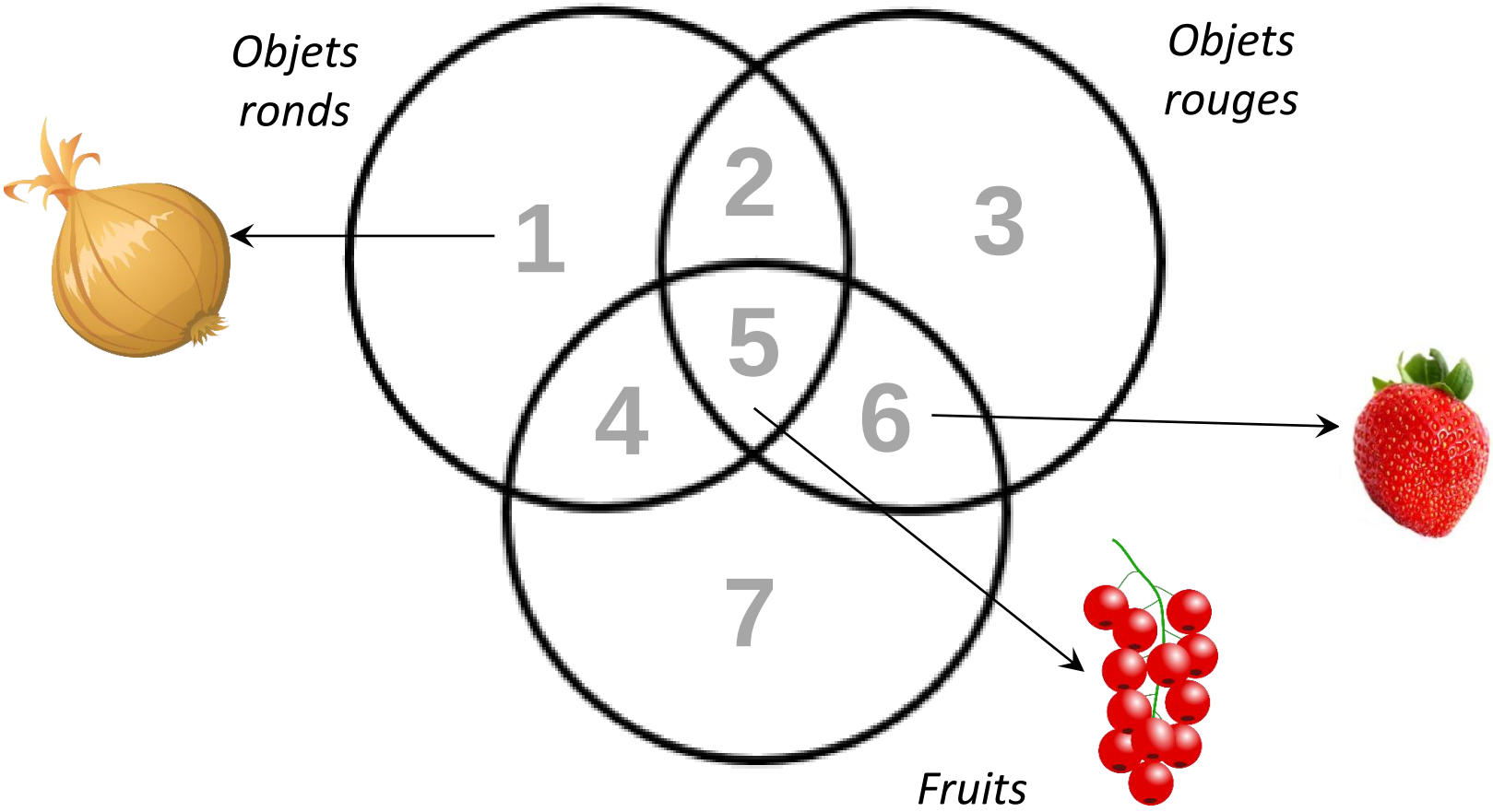
Diagramme de Venn **:**
Permet de représenter les opérateurs booléens



http://moebio.com/research/sevensets/

Notions préalables : Logique booléenne

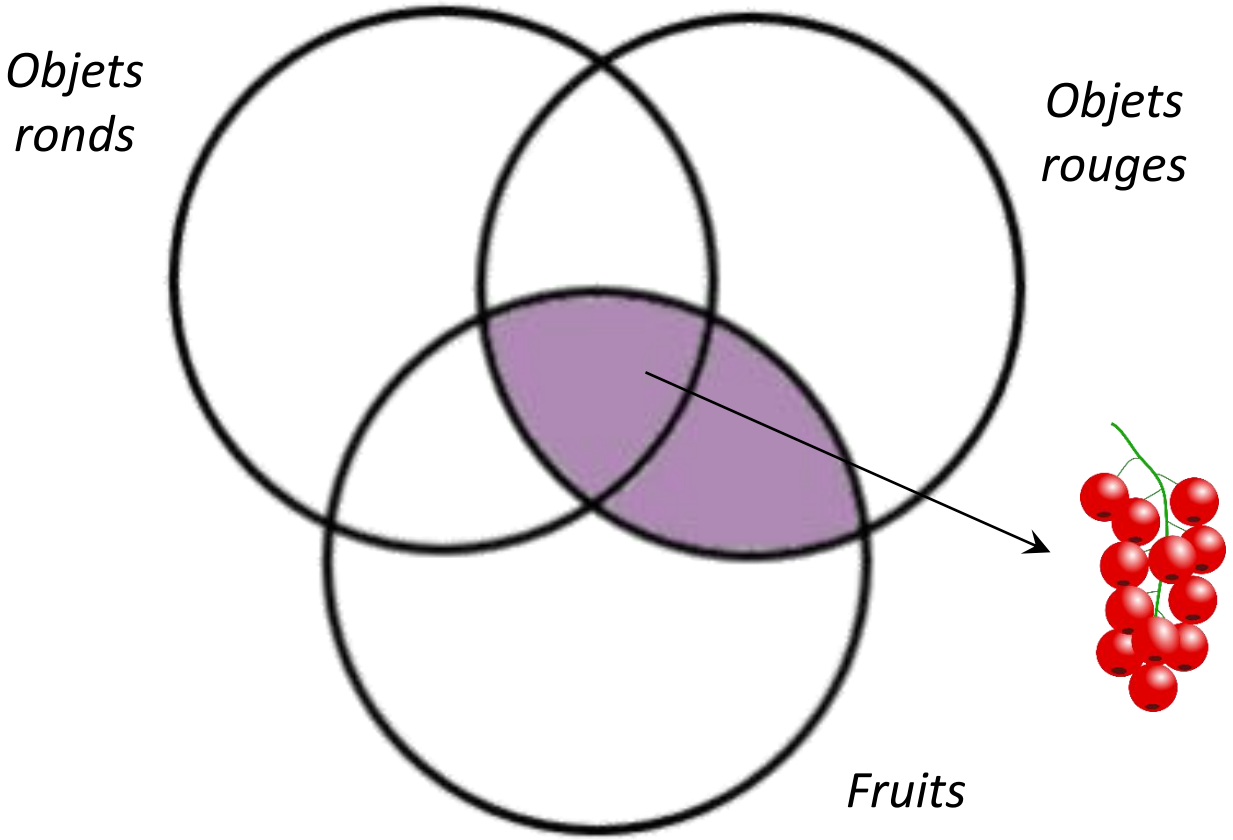
Diagrammes de Venn [2/7]



Version dynamique : http://cours.ebsi.umontreal.ca/sci6052/booleens_objet.php

Notions préalables : Logique booléenne

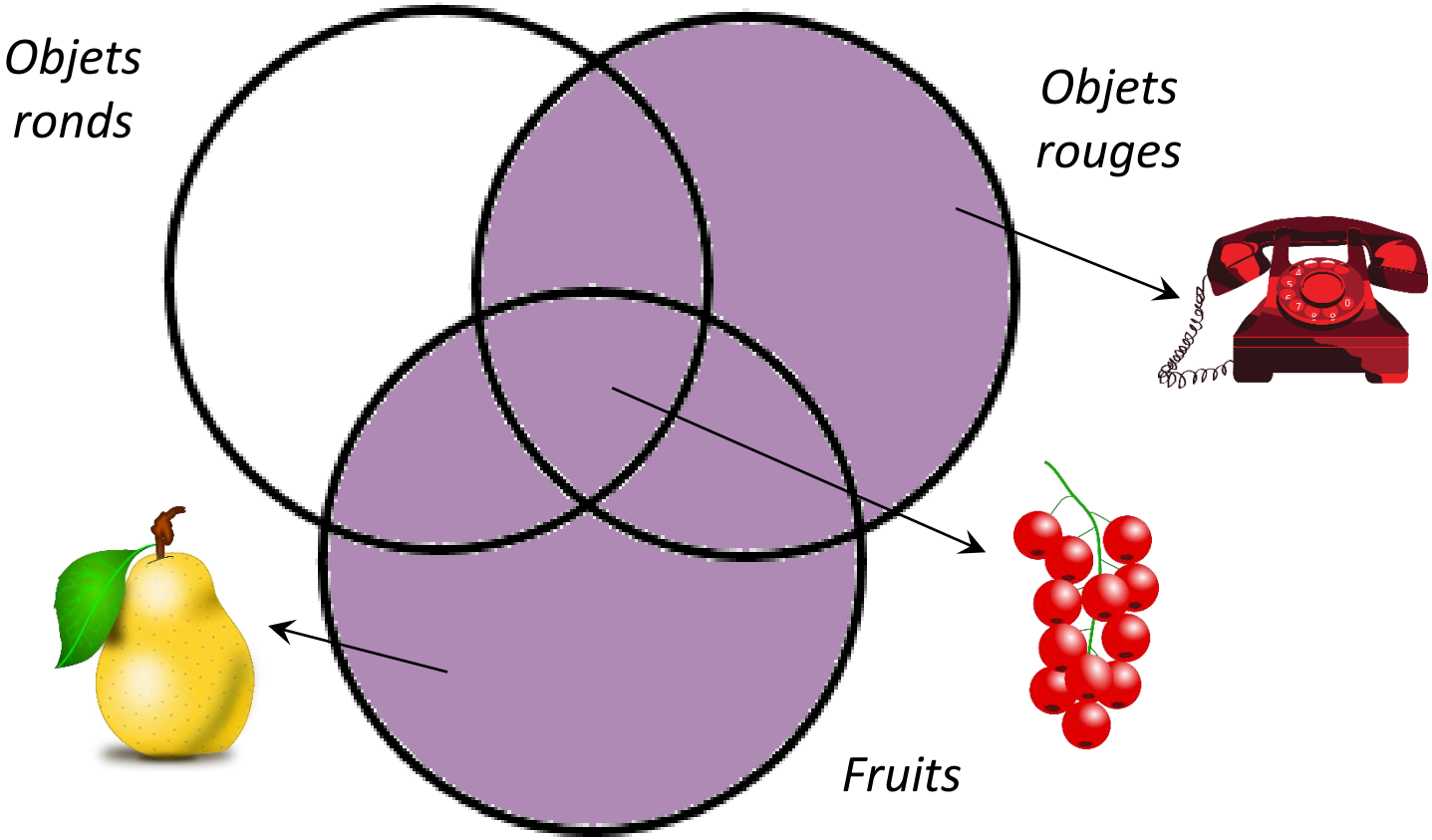
Diagrammes de Venn [3/7]



objets rouges **ET** fruits

Notions préalables : Logique booléenne

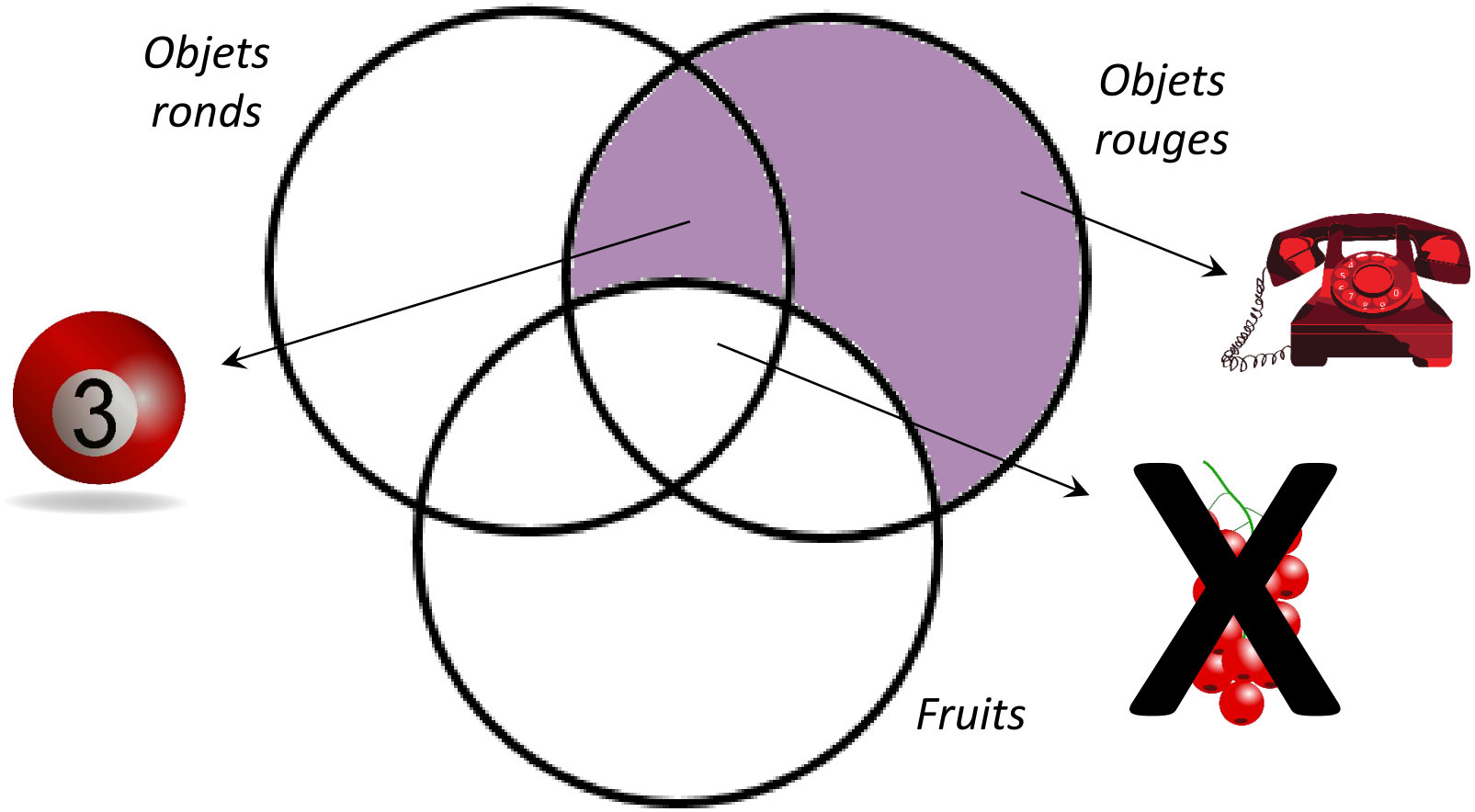
Diagrammes de Venn [4/7]



objets rouges **OU** fruits

Notions préalables : Logique booléenne

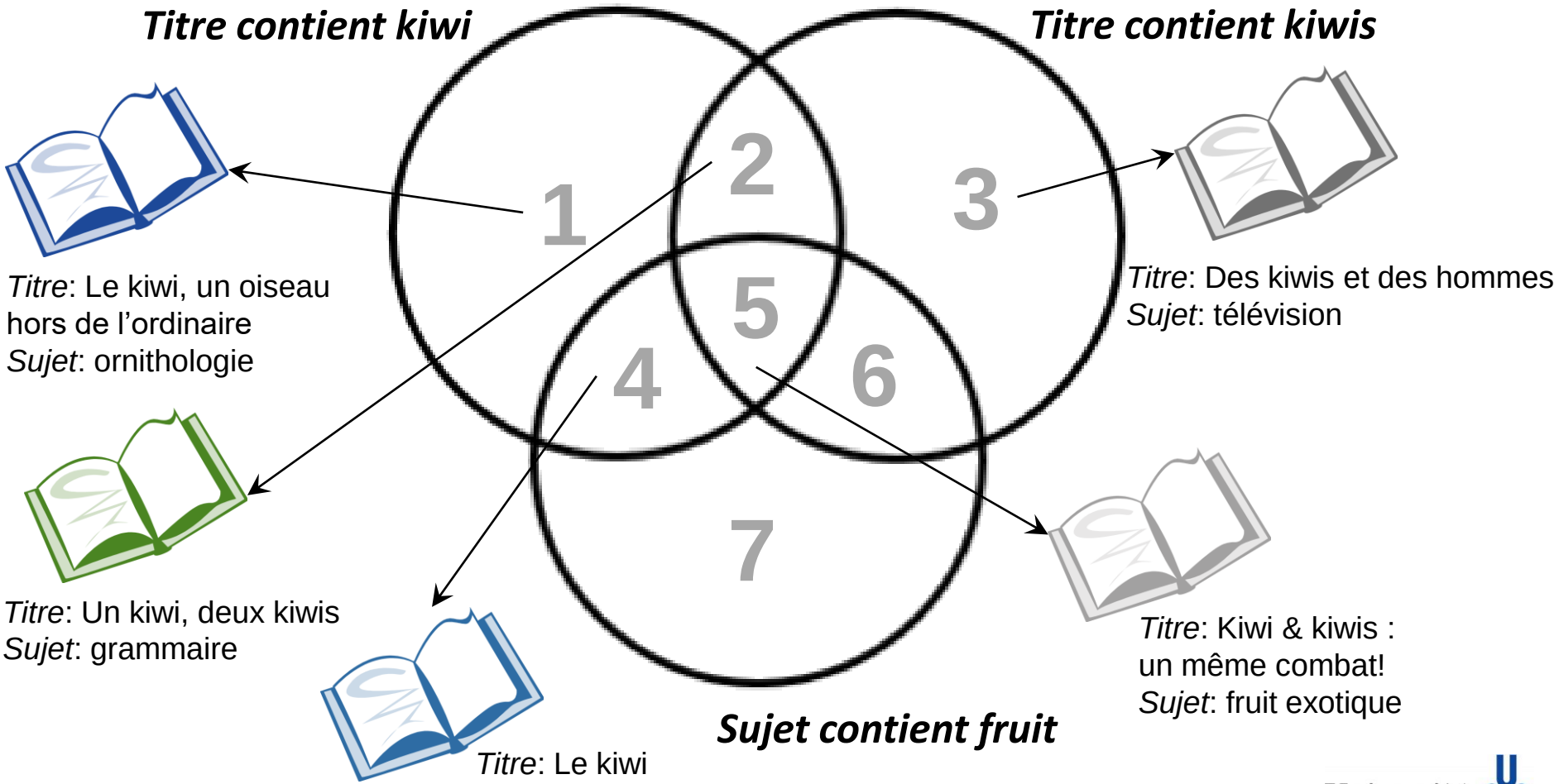
Diagrammes de Venn [5/7]



objets rouges **SAUF** fruits

Notions préalables : Logique booléenne

Diagrammes de Venn [6/7]

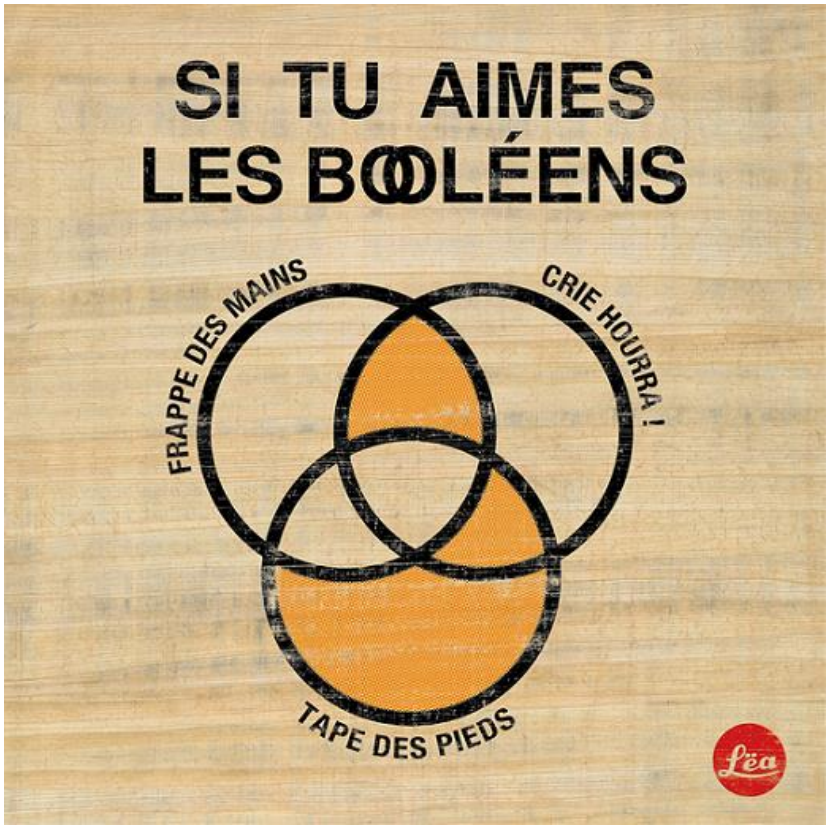


Université de Montréal

Version dynamique : http://cours.ebsi.umontreal.ca/sci6052/booleens_document.php

Notions préalables : Logique booléenne

Diagrammes de Venn [7/7]



Lëa-Kim Châteauneuf, 2014, https://www.flickr.com/photos/lea-kim/14959575554/

Devoir #1 pour la semaine prochaine :

Accéder à la version interactive à l'URL http://cours.ebsi.umontreal.ca/sci6052/booleens_danse.php et trouver, parmi les trois danses proposées, celle qui correspond à la danse des amateurs de booléens.

Université de Montréal

Notions préalables : Logique booléenne

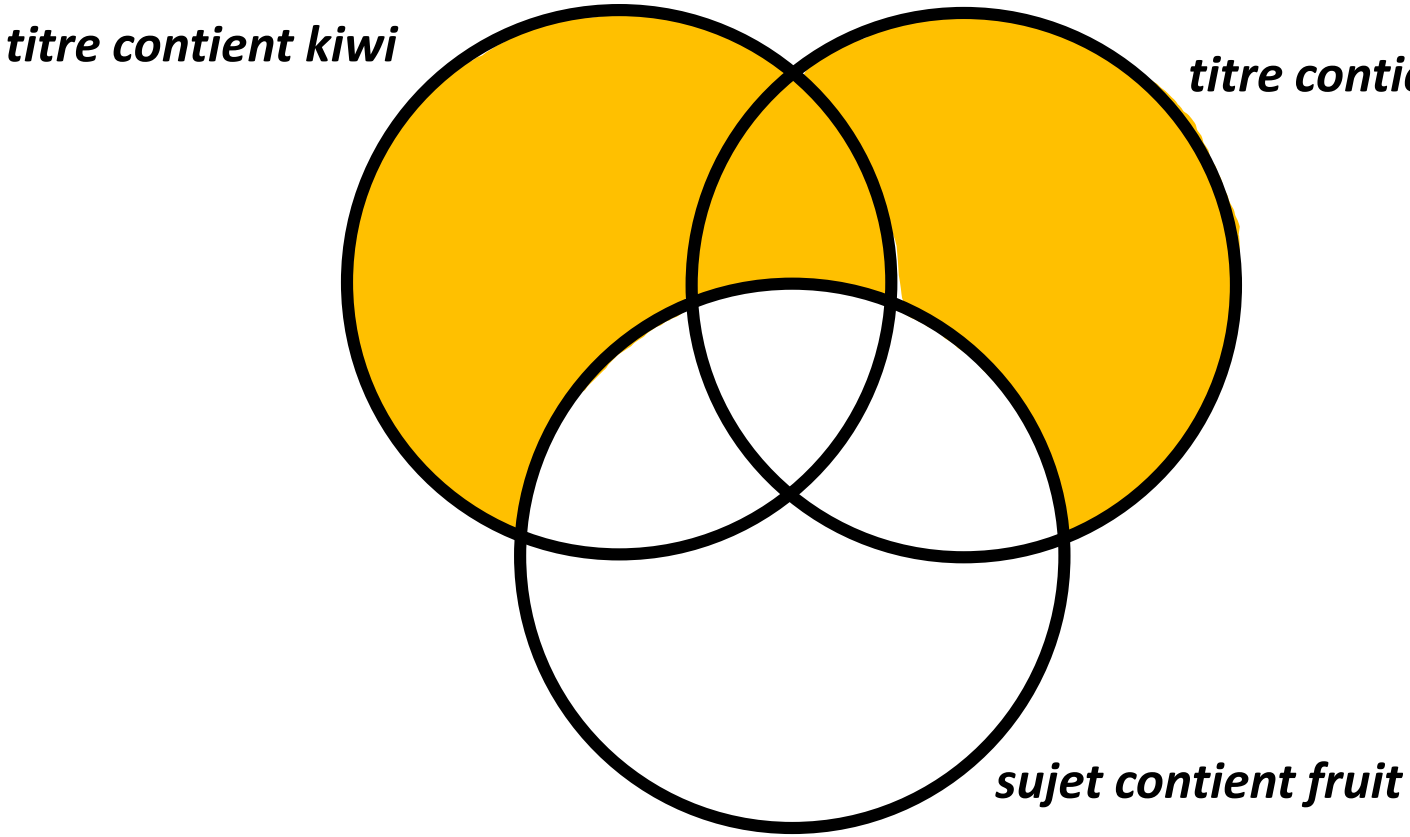
Ordre d'exécution des opérateurs booléens [1/3]

- S'il y a plus d'un opérateur booléen dans une même requête, il faut savoir dans quel ordre ils sont exécutés car l'ordre d'exécution **peut changer les résultats** selon les opérateurs utilisés
  - Par ex., la requête « objets rouges OU objets ronds SAUF fruits » donne des résultats différents si on priorise le OU, ou si on priorise le SAUF
- Ordre d'exécution **dépend** du système
  - Par exemple, certains systèmes exécutent la requête simplement de gauche à droite
  - D'autres systèmes vont prioriser certains opérateurs pour les exécuter en premier (par ex. ET)
- Généralement possible d'utiliser les **parenthèses** pour déterminer l'ordre d'exécution
  - Ce qui est entre parenthèses sera exécuté en premier

Notions préalables : Logique booléenne

Ordre d'exécution des opérateurs booléens [2/3]

(titre contient kiwi **OU** titre contient kiwis) **SAUF** sujet contient fruit

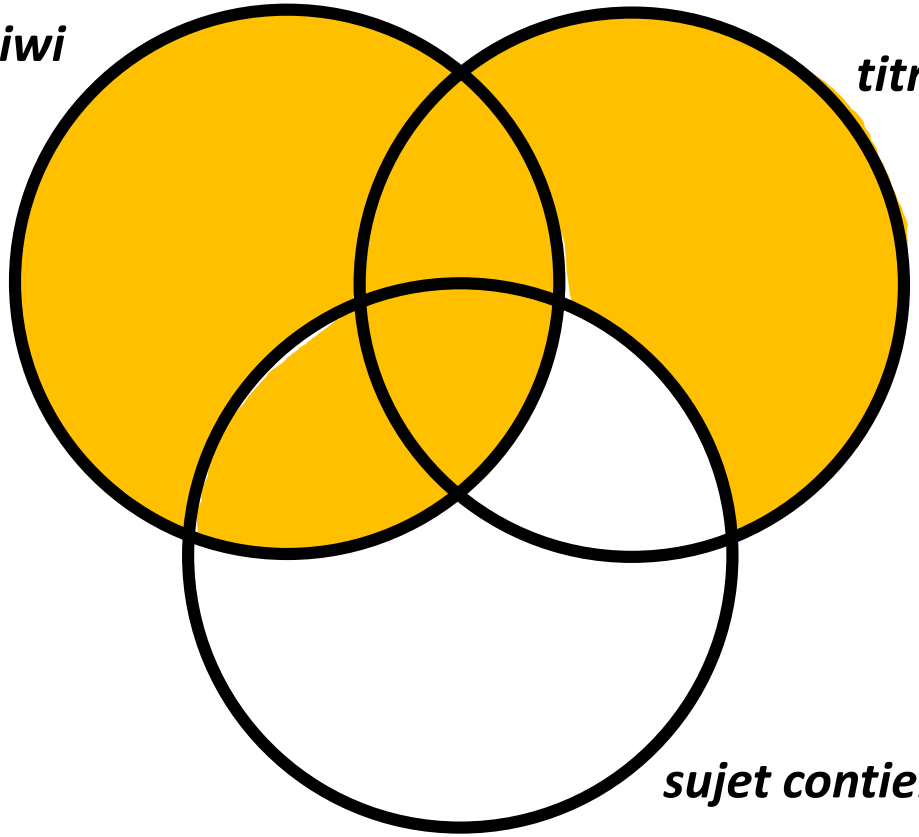


Titre: Le kiwi
Sujet: fruit

Notions préalables : Logique booléenne

Ordre d'exécution des opérateurs booléens [3/3]

titre contient kiwi **OU** (titre contient kiwis **SAUF** sujet contient fruit)



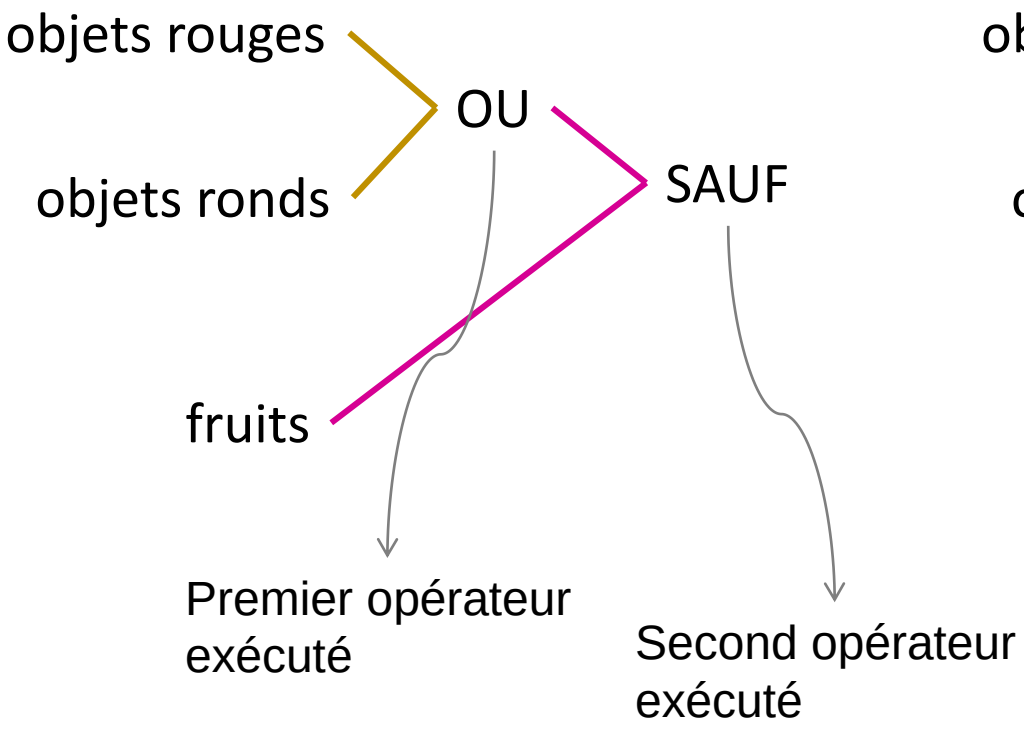
Titre: Le kiwi
Sujet: fruit

Notions préalables : Logique booléenne

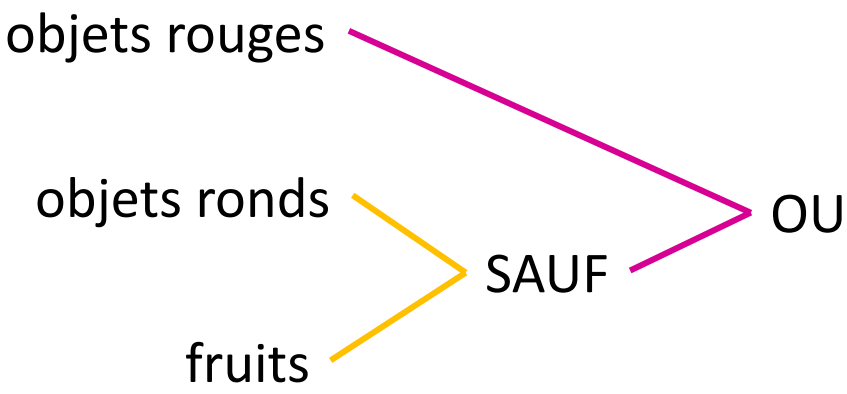
Arbres renversés

Arbres renversés :
Permet de comprendre comment une requête complexe est interprétée

(objets rouges OU objets ronds) SAUF fruits



objets rouges OU (objets ronds SAUF fruits)



Utilisation créative des arbres renversés!
http://lareference.ebsi.umontreal.ca/2010/01/01/janvier-2010-vol-27-n1/

Notions préalables : Logique booléenne

Arbres renversés & Diagrammes de Venn

- Il est possible de combiner les arbres renversés et les diagrammes de Venn pour décortiquer des requêtes de recherche complexes

- **Exemple**
  - Vous vous magasinez une voiture et décidez de chercher des articles sur trois marques qui vous intéressent particulièrement, Volvo, Toyota et Mazda avec les contraintes suivantes :
    1. Comme vous avez une légère préférence pour Volvo, vous voulez retrouver des articles qui en parlent sans toutefois parler de Toyota ou de Mazda (donc des articles qui ne parlent que de Volvo).
    2. De plus, vous aimeriez bien avoir des articles qui parlent des trois en espérant y trouver des comparaisons…
  - Vous faites la requête suivante dans votre base de données automobiles préférée :

    ```
    ((volvo SAUF toyota) OU mazda) OU ((volvo ET toyota) ET mazda)
    ```

  - Après avoir lancé la recherche vous remarquez des résultats qui ne parlent pas du tout de Volvo… Est-ce normal? Que se passe-t-il?

((volvo SAUF toyota) OU mazda) OU ((volvo ET toyota) ET mazda)

volvo

toyota

mazda

volvo

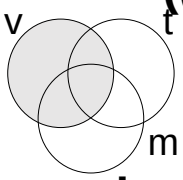
toyota

mazda

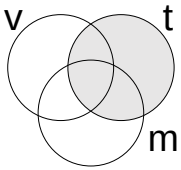
Première étape : définir les « feuilles » de l'arbre inversé.

Les feuilles correspondent aux **termes** présents dans la requête et reliés par des opérateurs. Il faut les écrire les uns sous les autres, **dans l'ordre où ils apparaissent dans la requête**.

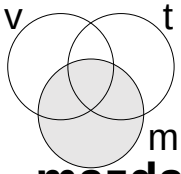
((volvo SAUF toyota) OU mazda) OU ((volvo ET toyota) ET mazda)



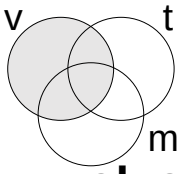
volvo



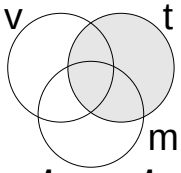
toyota



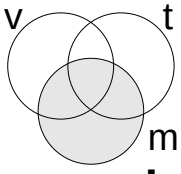
mazda



volvo



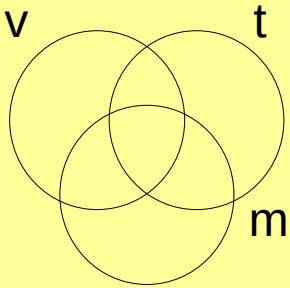
toyota



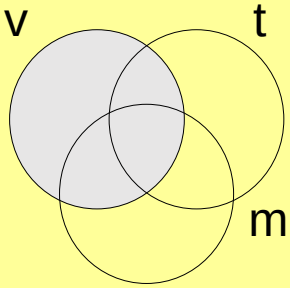
mazda

Deuxième étape : pour chacune des feuilles, dessiner sa représentation sous forme de diagramme de Venn.

Dans notre cas, le diagramme de base comportera trois cercles, un pour chacune des marques d'auto *(v=Volvo, t=Toyota, m=Mazda)*.



Par exemple, les documents où se retrouvent le mot « volvo » sont représentés ainsi par la zone grisée ci-dessous :



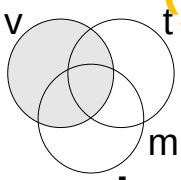
((volvo SAUF toyota) OU mazda) OU ((volvo ET toyota) ET mazda)

Premier niveau de parenthèses

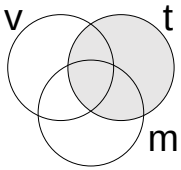
Deuxième niveau de parenthèses

Deuxième niveau de parenthèses

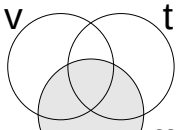
Premier niveau de parenthèses



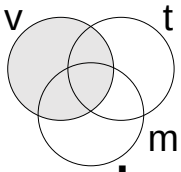
volvo



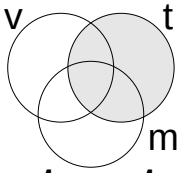
toyota



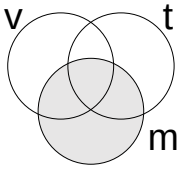
mazda



volvo



toyota



mazda

Troisième étape : identifier l'ordre d'exécution des opérateurs.

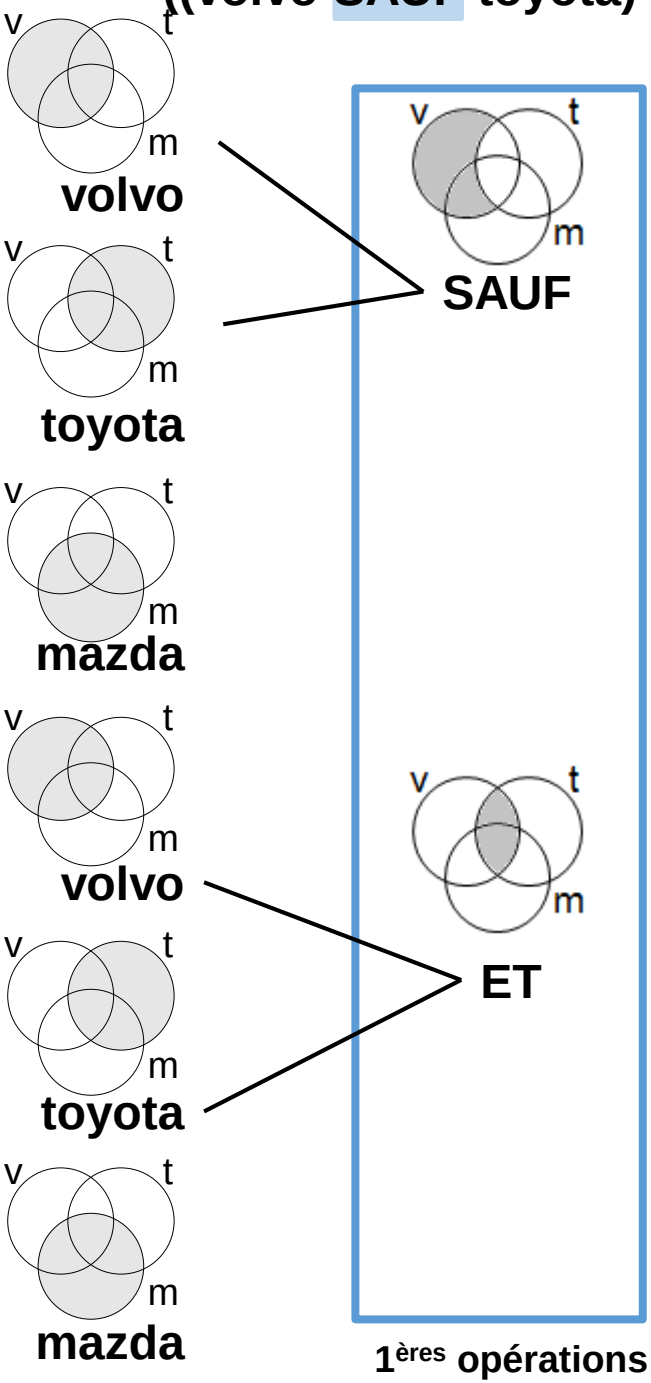
Le principe à se rappeler est que les **parenthèses imposent l'ordre d'exécution**. Les parenthèses les plus à l'intérieur (=niveau de profondeur le plus grand) seront celles qui seront exécutées en premier.

Dans la requête, nous retrouvons **deux niveaux de parenthèses** :

- Les orangées sont au premier niveau.
- Les bleues, qui se trouvent à l'intérieur des orangées, sont ainsi au deuxième niveau.

Ce sont les opérateurs à l'intérieur des parenthèses bleues qui seront exécutés en premier comme ils correspondent au niveau le plus profond. Par la suite, les opérateurs entre les parenthèses orangées seront exécutés. Finalement, le OU vert, qui n'est à l'intérieur d'aucune parenthèse sera exécuté.

((volvo SAUF toyota) OU mazda) OU ((volvo ET toyota) ET mazda)



Quatrième étape : représenter à l'aide de diagrammes de Venn le résultat des premières opérations.

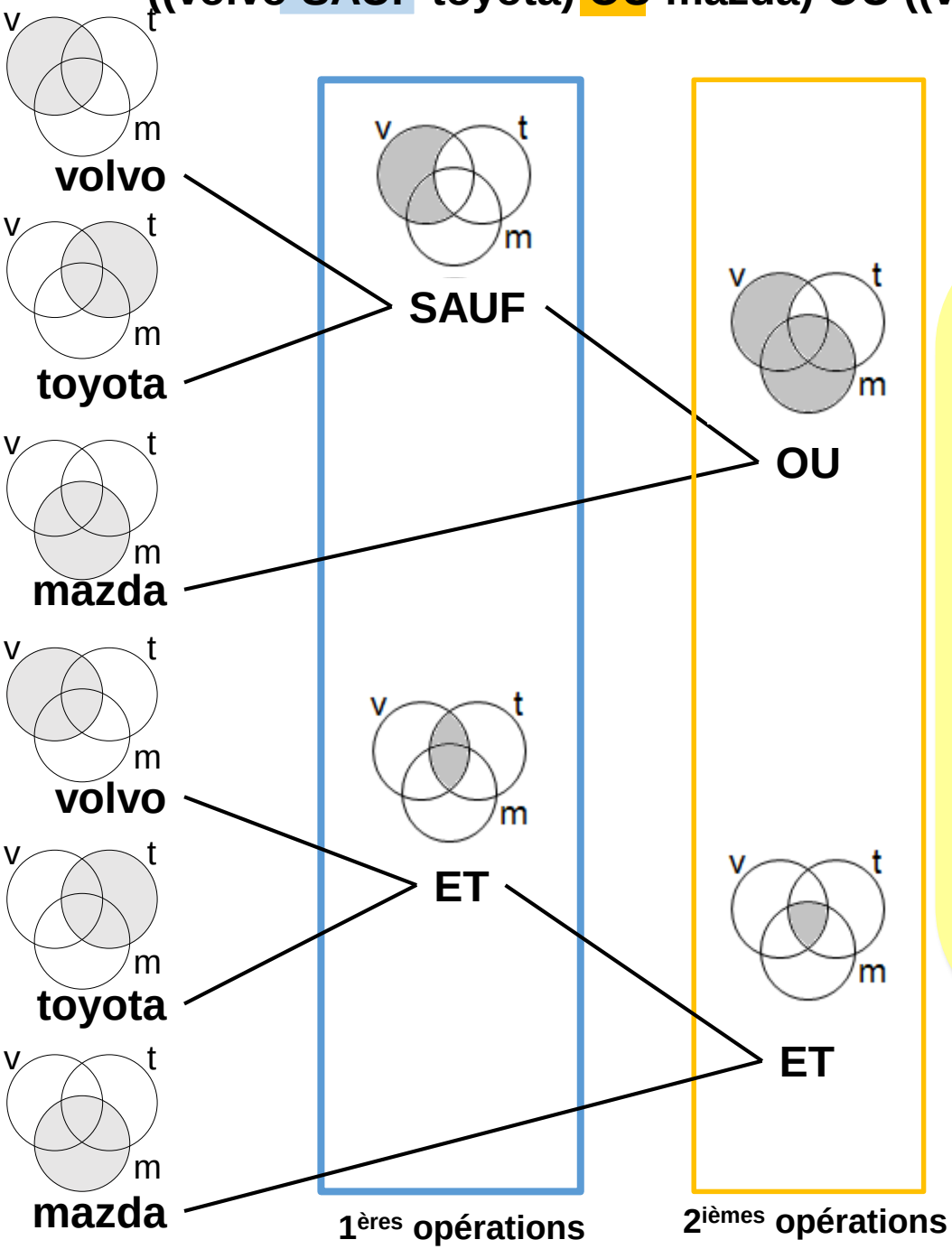
Il s'agit ici de **relier** les termes qui sont liés par les opérateurs exécutés en premier, soit le SAUF et le ET en bleu. Le SAUF relie les deux termes qui lui sont adjacents, soit volvo ainsi que toyota. Le ET relie volvo à toyota.

Une fois ces termes reliés et l'opérateur à appliquer indiqué, il faut **construire le diagramme de Venn** résultant de l'application de cet opérateur :

- Attardons nous au SAUF. Le résultat de l'application de cet opérateur booléen est **la suppression** des résultats du deuxième terme des résultats du premier terme. Ainsi **`volvo SAUF toyota`** est représenté par la soustraction du cercle grisé correspondant à volvo de la partie commune de celui de toyota (zone grisée). C'est en fait comme si vous passiez la gomme à effacer sur le diagramme de Venn de **`volvo`** pour enlever le gris de toyota!

- Le deuxième opérateur, le ET, donne comme résultat l'intersection (i.e. la zone commune) des résultats des deux termes. Ainsi **`volvo ET toyota`** correspond à la zone grisée qui est partagée par les deux cercles.

((volvo SAUF toyota) OU mazda) OU ((volvo ET toyota) ET mazda)

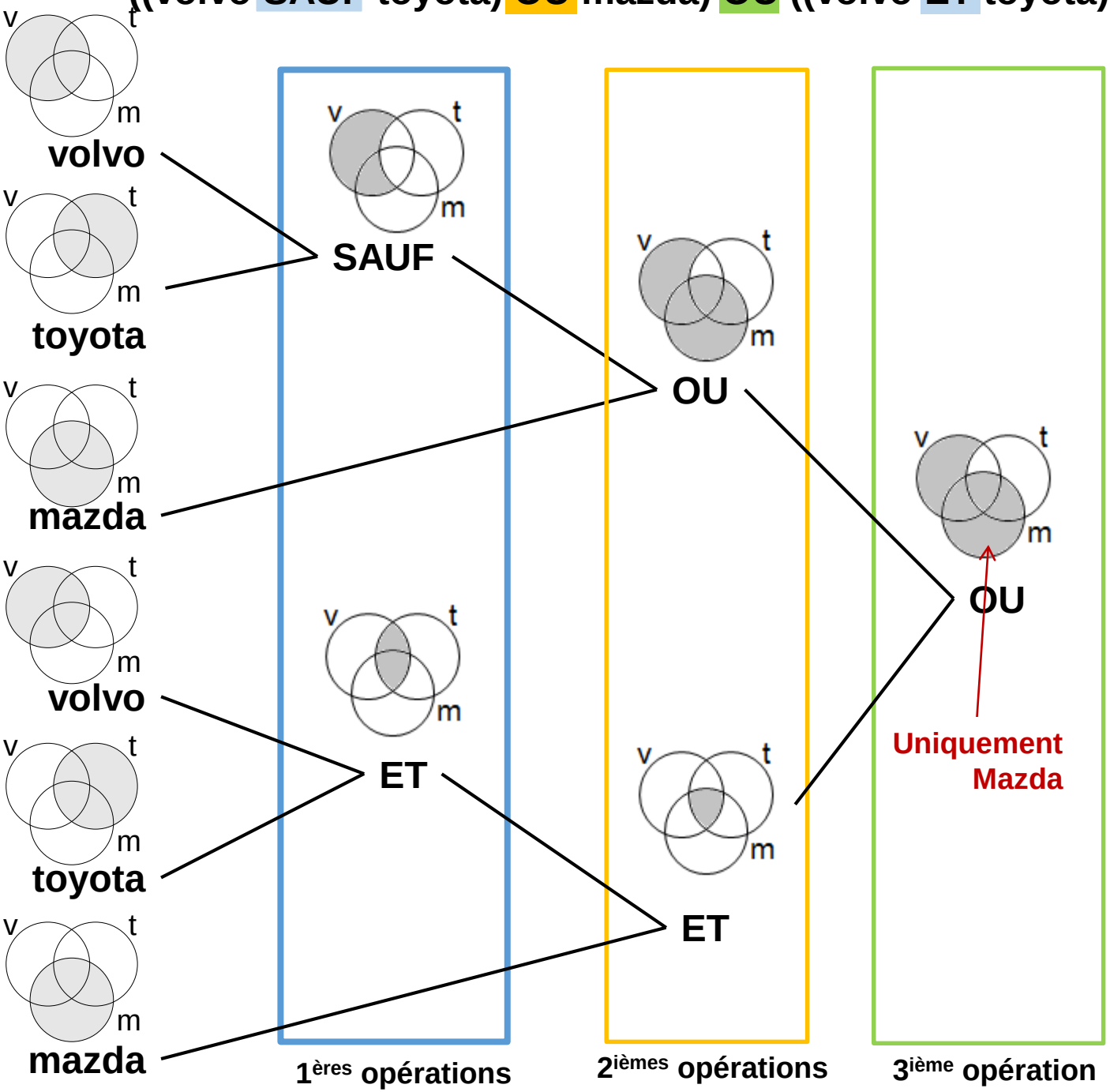


Cinquième étape : représenter à l'aide de diagrammes de Venn le résultat des opérateurs appliqués en deuxième.

Le OU et le ET orangés seront appliqués par la suite. Tout comme à l'étape précédente, il faut relier les éléments sur lesquels ils s'appliquent et dessiner, à l'aide du diagramme de Venn, les résultats de cette application :

- Ainsi le OU s'applique entre le résultat de (volvo SAUF toyota) et le terme masda. OU correspondant à une union (addition), il faut ainsi ajouter à zone grisée représentant (volvo SAUF toyota) celle représentant mazda.
- Le ET, quant à lui, fera l'intersection entre les résultats de `(volvo ET toyota)` et ceux de `mazda`. Vous retrouverez ainsi uniquement la zone centrale commune aux trois cercles.

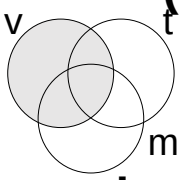
((volvo SAUF toyota) OU mazda) OU ((volvo ET toyota) ET mazda)



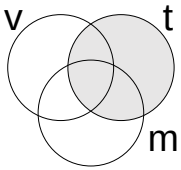
Étape finale : représenter à l'aide de diagrammes de Venn le résultat du dernier opérateur à appliquer.

Ne reste alors qu'à appliquer le dernier opérateur, le OU vert. Le OU étant une addition (union), son résultat sera le cumul des zones grisées correspondant à l'application des opérateurs précédents. Le résultat nous montre bien qu'il y a quelque chose qui cloche dans notre requête comme nous retrouvons des éléments qui ne parlent que de Mazda.

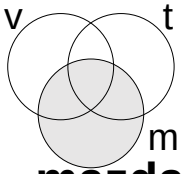
(volvo SAUF (toyota OU mazda)) OU ((volvo ET toyota) ET mazda)



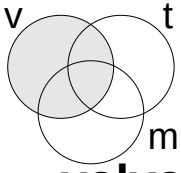
volvo



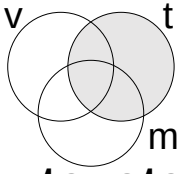
toyota



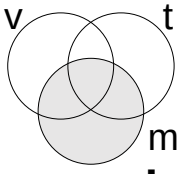
mazda



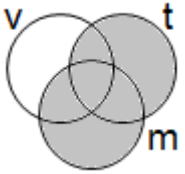
volvo



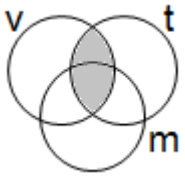
toyota



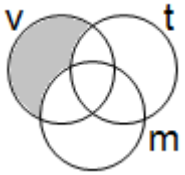
mazda



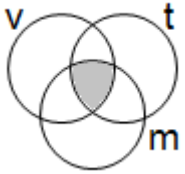
OU



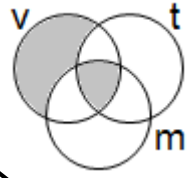
ET



SAUF



ET



OU

C’est bien ce que l’on cherche!

1ères opérations

2ièmes opérations

3ième opération

Notions préalables : Logique booléenne

Arbres renversés & Diagrammes de Venn

- **Devoir # 2 pour la semaine prochaine** :
  Utilisez la combinaison d'un arbre renversé et de diagrammes de Venn pour comprendre ce que fait la requête suivante :

```
begonia SAUF ((tulipe OU pensée) SAUF ((begonia ET tulipe) ET pensée))
```

  - Que représentent les résultats retrouvés?
  - Y a-t-il moyen d'écrire autrement cette requête?

Notions préalables : Autres opérateurs de recherche

Troncature *

- Remplace **0 à *n* caractères** (parfois 1 à *n* caractères)
- Dépendamment des outils de recherche, peut se mettre au début, au milieu et/ou à la fin d'un terme de recherche
  - Parfois *implicite* aussi dans certains outils (i.e. qu'ils font automatiquement une troncature à droite par exemple)
- Utile pour trouver, entre autres
  - Mots de même racine/famille
    - `cheva*` : cheval, chevaux, chevauchement, chevalier, etc.
    - `*mobil*` : immobile, mobilier, immobiliser, etc.
  - Variantes de langues / variantes orthographiques
    - `ent*prise` : entreprise, enterprise
  - Féminin et pluriel d'un mot
    - `professeur*` : professeure, professeurs, professeures
- Permet d'augmenter le rappel mais peut diminuer la précision
  - Par ex., `institut*` retrouve le féminin et le masculin *institutrice* et *instituteur* (et le pluriel) mais retrouve aussi *institutionnel*

Université de Montréal

Notions préalables : Autres opérateurs de recherche

Masque ?

- Remplace **un seul caractère** (parfois 0 à 1 caractère)
- Utile pour trouver, par exemple
  - variantes de langues
    - *organi?ation* : organisation, organization
  - féminin ou pluriel d'un mot
    - *haut?* : (haut) hauts, haute
  - mots de même famille
    - *journal???e* : journalisme, journaliste
  - variantes orthographiques
    - *cent??* : centre, center (variante d'orthographe)
- Comme la troncature, permet d'augmenter le rappel mais peut diminuer la précision
  - *journal???e* : journalisme, journaliste, journalière
    - *journalis?e* serait ici préférable!

Notions préalables : Autres opérateurs de recherche

Recherche d'expression " "

- Repère une suite de mots (les mots doivent être immédiatement juxtaposés et apparaître dans l'ordre donné)

  `"base de données"`

  → ne repêche pas "données de base"

  → Attention : les spécificités des outils peuvent varier quant à la recherche d'expression

- **Permet une recherche plus précise mais il faut bien connaître la terminologie du domaine et du corpus interrogé pour ne pas diminuer le rappel**
  - Par exemple, certains écrivent *sciences de l'information* et d'autres *science de l'information*. La recherche "sciences de l'information" passera ainsi à côté de la forme au singulier.

Notions préalables : Autres opérateurs de recherche

Parenthèses ( )

- Permet de déterminer l'ordre d'exécution des opérateurs pour s'assurer que la requête soit bien exécutée comme on le souhaite

```
(peau OU skin) ET cancer
```

  - *Besoin* : les documents traitant du cancer de la peau dans une base de données bilingue anglais-français

```
(tableur OU ("chiffrier électronique")) ET logiciel
```

  - *Besoin* :

Les documents parlant de logiciel de la famille des tableurs (ou chiffrier électronique). Il est à noter que la parenthèse autour de "chiffrier électronique" est un excès de zèle qui ne nuira pas à la recherche.

Recherche de fichiers en texte intégral

Types de recherche dans les fichiers

- Peut se faire de deux façons, selon les outils
  - Recherche <u>séquentielle</u>
    - Lecture séquentielle de chaque fichier pour trouver une séquence de caractères
      - Lecture de *tous* les fichiers
  - Recherche <u>indexée</u>
    - Recherche se fait par le biais d'un index (fichier inversé) construit, suite à un processus d'indexation, à partir de tous les fichiers
      - Lecture d'*un seul* fichier, l'index

Recherche de fichiers en texte intégral

Recherche séquentielle [1/2]

- Système lit **séquentiellement** le contenu de **chaque fichier** pour trouver une séquence de caractères
- **Temps d'exécution** d'une requête proportionnel au nombre et à la **longueur** des fichiers
- Utile pour rechercher de l'information dans un **dossier de petite taille** ou pour un **besoin ponctuel**
  - Ne demande aucune préparation
  - Outil de recherche séquentielle généralement disponible dans les systèmes d'exploitation (par ex. approche par défaut de la recherche dans Windows 8 lorsqu'un fichier n'est pas indexé)

Recherche de fichiers en texte intégral

Recherche séquentielle [2/2]

Requête : guimauves

fichier 1

J'aime les oignons et les guimauves. → guimauves = **oui**

fichier 2

Ma voisine m'a donné trois tomates et deux oignons. → guimauves = **non**

fichier 3

J'ai mangé quinze guimauves pour déjeuner. → guimauves = **oui**

Réponse :

fichier 1 et fichier 3

Recherche de fichiers en texte intégral

Recherche indexée

- Recherche se fait à l'intérieur d'un fichier **d'index** (***fichier inversé***), résultat du processus d'indexation
- Temps d'exécution de la recherche beaucoup **plus rapide** : système n'a qu'à consulter le fichier inversé pour repérer tous les fichiers dans lesquels apparait le terme de recherche
- Requiert une **préparation** : système doit indexer le corpus avant qu'on puisse faire des recherches
  - Peut prendre un certain temps la première fois
  - Doit être mis à jour lorsqu'il y a des modifications des fichiers indexés pour que l'index reflète bien la réalité
- Rentable si on prévoit effectuer **plusieurs recherches** dans un ensemble de fichiers donné

Recherche de fichiers en texte intégral

Recherche indexée : Processus d'indexation

- Système lit séquentiellement le contenu de tous les fichiers pour créer un fichier inversé (*index*) où seront consignés chaque mot (retenu suite à d'éventuels traitements) du fichier et leur position dans le fichier
  - Système peut « **traiter** » les mots avant de les verser dans le fichier inversé. Par exemple
    - Retirer accents et signes de ponctuation
    - Mettre tout en minuscules
  - Système peut **ne pas verser** les mots considérés comme **vides de sens** comme, par exemple, des articles ou des pronoms (consignés dans un fichier appelé « antidictionnaire ») dans le fichier inversé

Recherche de fichiers en texte intégral

Recherche indexée : Ex. d'indexation [1/4]

fichier 1

J'aime les oignons et les guimauves.

fichier 2

Ma voisine m'a donné trois tomates et deux oignons.

fichier 3

J'ai mangé quinze guimauves pour déjeuner.

Fichiers originaux

Recherche de fichiers en texte intégral

Recherche indexée : Ex. d'indexation [2/4]

fichier 1

j aime les oignons et les guimauves

fichier 2

ma voisine m a donne trois tomates et deux oignons

fichier 3

j ai mange quinze guimauves pour dejeuner

Première étape :
retirer majuscules, accents et ponctuation

Recherche de fichiers en texte intégral

Recherche indexée : Ex. d'indexation [3/4]

fichier 1

~~j~~ aime ~~les~~ oignons ~~et les~~ guimauves

fichier 2

~~ma~~ voisine ~~m a~~ donne trois tomates ~~et~~ deux oignons

fichier 3

~~j ai~~ mange quinze guimauves ~~pour~~ dejeuner

Deuxième étape :
retirer les mots de l'antidictionnaire

Recherche de fichiers en texte intégral

Recherche indexée : Ex. d'indexation [4/4]

Troisième étape :
verser les mots et leur position dans fichier inversé

fichier 1

j aime ~~les~~ oignons ~~et les~~ guimauves

fichier 2

~~ma~~ voisine ~~m a~~ donne trois tomates ~~et~~ deux oignons

fichier 3

~~j ai~~ mange quinze guimauves ~~pour~~ dejeuner

mot	*position*
aime	fichier 1 (1)
dejeuner	fichier 3 (4)
deux	fichier 2 (5)
donne	fichier 2 (2)
guimauves	fichier 1 (3), fichier 3 (3)
mange	fichier 3 (1)
oignons	fichier 1 (2), fichier 2 (6)
quinze	fichier 3 (2)
tomates	fichier 2 (4)
trois	fichier 2 (3)
voisine	fichier 2 (1)

Recherche de fichiers en texte intégral

Recherche indexée : Ex. de recherche

mot	*position*
aime	fichier 1 (1)
dejeuner	fichier 3 (4)
deux	fichier 2 (5)
donne	fichier 2 (2)
guimauves	fichier 1 (3), fichier 3 (3)
mange	fichier 3 (1)
oignons	fichier 1 (2), fichier 2 (6)
quinze	fichier 3 (2)
tomates	fichier 2 (4)
trois	fichier 2 (3)
voisine	fichier 2 (1)

Requête : guimauves

→ *Réponse :*

fichier 1 et fichier 3

Recherche de fichiers en texte intégral

Contenu textuel des fichiers

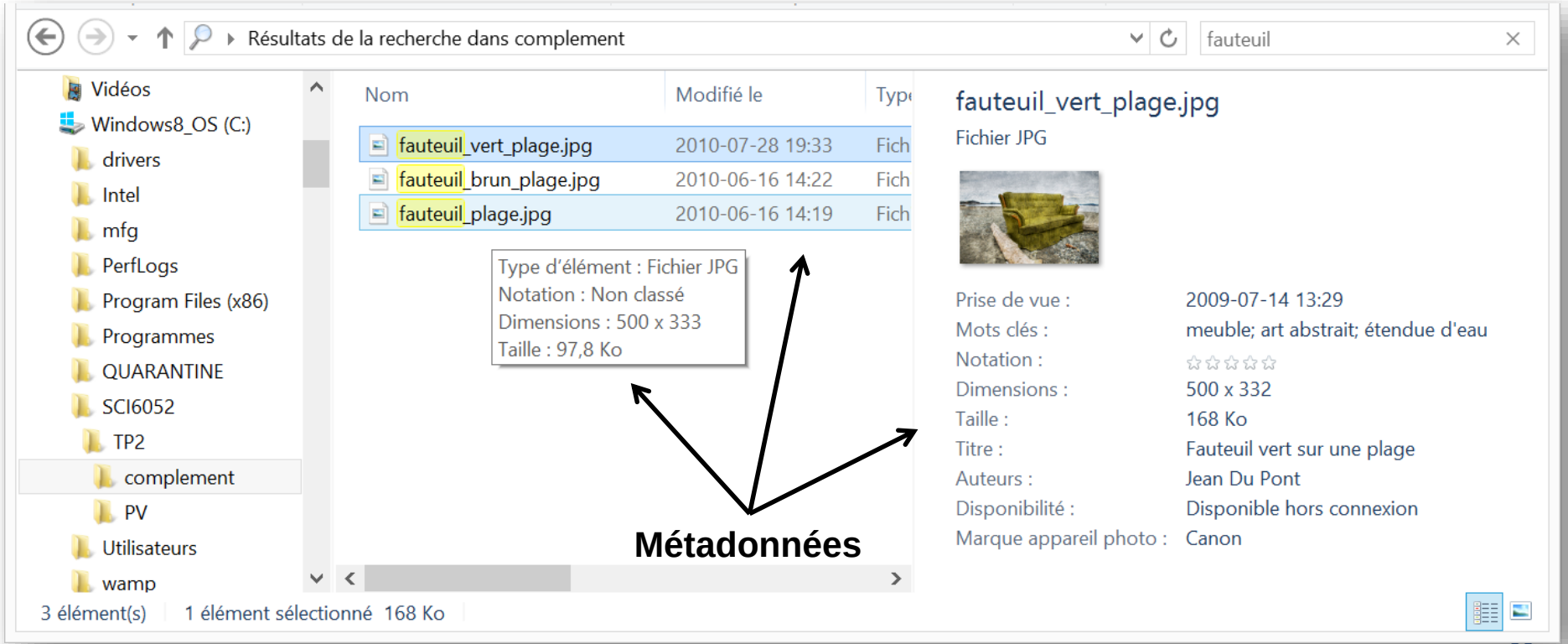
- Rappel : *Recherche en texte intégral* = recherche se fait sur le contenu textuel des fichiers

- Qu'est-ce que le contenu textuel d'un fichier?
  - Son **contenu comme tel** (par ex. le texte d'un document Word ou d'un document PDF)
    - Accessible aux outils de recherche en fonction des filtres qu'ils possèdent pour accéder à ce contenu
  - Certains outils considèrent aussi comme du contenu textuel certaines de ses **métadonnées**

Recherche de fichiers en texte intégral

Métadonnées internes et externes [1/3]

Métadonnées = données « contextuelles » concernant un fichier (par ex. Titre, Auteur, Date) de deux types : métadonnées d'application et métadonnées système



Visualisation des métadonnées dans l'Explorateur Windows

Recherche de fichiers en texte intégral

Métadonnées internes et externes [2/3]

- **Métadonnées *d'application* (= *interne*) : par ex. *mots-clés***
  - Pour des fichiers générés par des **logiciels d'application** (documents bureautiques, images, …)
  - Métadonnées stockées ***dans*** les fichiers (*internes*) et gérées par l'application concernée
    - Incluent : titre, objet, auteur, mots-clés, commentaires, etc.
    - Varient en fonction du type de fichiers (par ex. pour les fichiers sonores: genre, durée)
  - Comment les visualiser?
    - Dans le logiciel d'application (*complet*)
    - Dans l'explorateur de fichiers du système d'exploitation (*partiel*)
    - Dans la fenêtre « Propriétés » des fichiers accessible via son menu contextuel (*partiel*)
      - Certaines sont en lecture seule; d'autres sont modifiables par l'utilisateur

Recherche de fichiers en texte intégral

Métadonnées internes et externes [3/3]

- **Métadonnées *<u>système</u> (= externe)*** : par ex. nom du fichier, taille
  - **Pas stockées dans le fichier lui-même (*<u>externes</u>*)**, mais dans le *dossier* qui le contient. Incluent, pour tous les fichiers :
    - Nom du fichier
    - Dates de création, dernière modification
    - Nom d'usager du créateur
    - Taille (en octets)
    - Propriétés "caché", "lecture seulement", etc.
  - Visibles directement dans **l'Explorateur Windows** (colonne dédiée ou infobulle), et/ou via la fenêtre « propriétés » du fichier
    - Parfois en lecture seule; d'autres fois modifiables

Recherche de fichiers en texte intégral

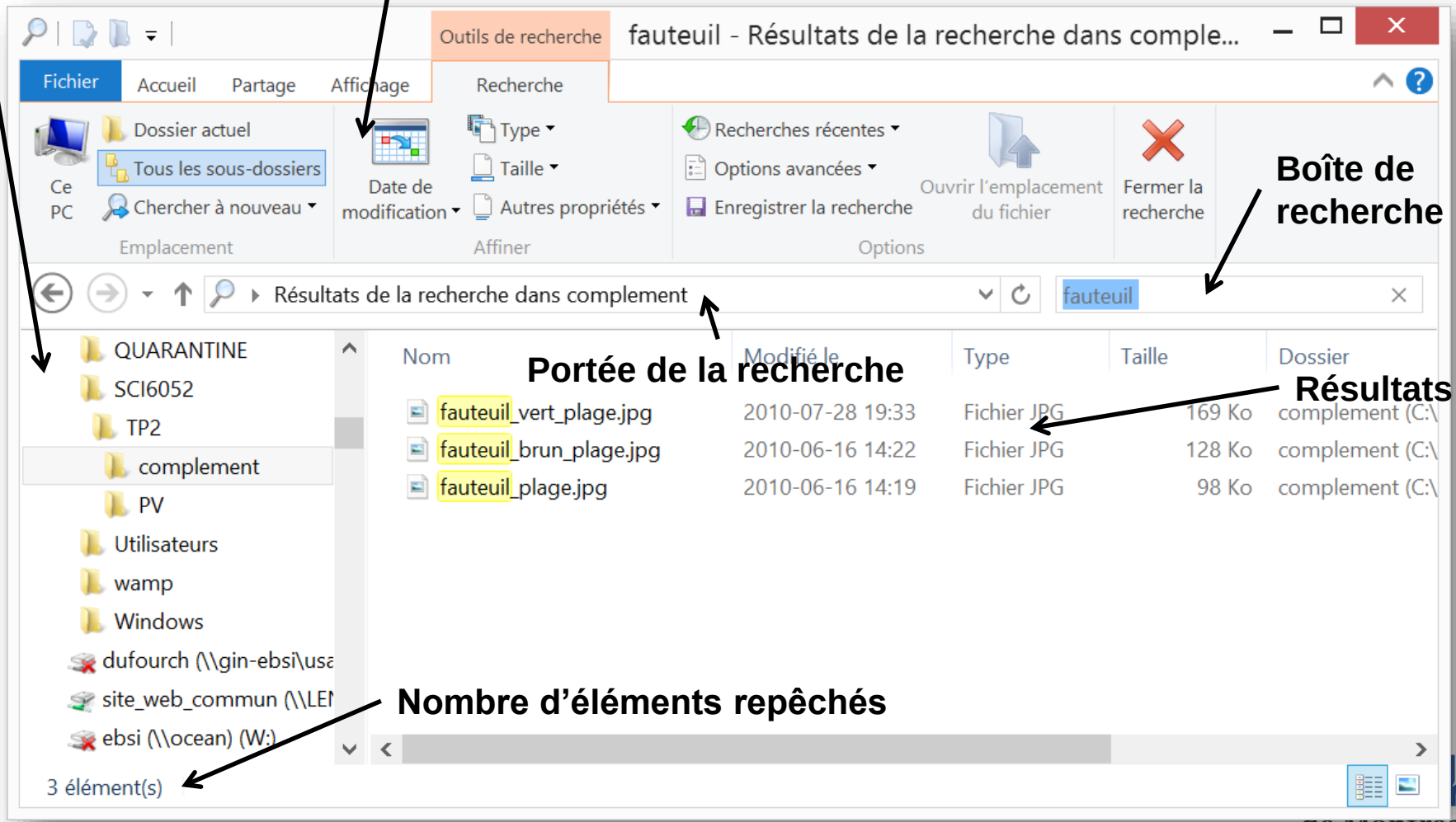
Outil intégré à l'OS : Windows Search [1/6]

- Logiciel de recherche de fichiers en texte intégral permettant la recherche séquentielle et la recherche indexée de fichiers
  - Intégré de base sous Windows 10 *
  - Recherche séquentielle par défaut sur les dossiers qui ne sont pas indexés
  - Recherche indexée sur les dossiers indexés
    - Indexe plus de 200 types de fichiers courants (courrier électronique, documents bureautiques, images, sons, etc.)
    - Peut être tout le disque dur ou un sous-ensemble de dossiers
    - Indexation doit se faire au fur et à mesure des modifications sur les dossiers indexés pour offrir un reflet exact de ces dossiers
    - Si indexation de tout le disque dur avec mise à jour en temps réel, peut causer un ralentissement de l'ordinateur

*** Dans la version actuelle de Windows 10, il a été observé que Windows Search n'est pleinement fonctionnel que pour ce qui se retrouve dans le dossier Documents.**

Recherche de fichiers en texte intégral

Outil intégré à l'OS : Windows Search [2/6]



Interface Windows Search

de Montréal

Recherche de fichiers en texte intégral

Outil intégré à l'OS : Windows Search [3/6]

- Principales caractéristiques de la recherche [1/3]
  - Recherche dans le **contenu textuel** des fichiers (c'est-à-dire contenu des fichiers (texte comme tel) + certaines métadonnées d'application & système)
    - Pour la recherche séquentielle, la recherche est faite dans un premier temps uniquement dans les noms des fichiers. Pour « l'étendre » au contenu textuel des fichiers, il faut cocher l'option « Contenu du fichier » dans les options avancées du groupe « Options » de l'onglet « Recherche »
  - **Troncature implicite** à la fin des mots (* peut aussi être utilisé)
  - **Recherche d'expression** avec les guillemets ("")
    - Attention, la troncature peut être utilisée à la fin d'une expression mais elle « s'étendra » à tous les mots!
  - Opérateurs booléens **AND OR NOT** supportés
    - *article AND pdf* (recherche indexée) : fichiers ayant « article » et « pdf » dans leur contenu textuel
    - *article NOT breeding* (recherche séquentielle) : dans un premier temps, fichiers dont le nom contient « article » mais pas « breeding »
    - *word OR pdf* (recherche indexée) : fichiers ayant « word » ou « pdf » ou les deux dans leur contenu textuel

Université de Montréal

Recherche de fichiers en texte intégral

Outil intégré à l'OS : Windows Search [4/6]

- Principales caractéristiques de la recherche [2/3]
  - **Signes diacritiques** et **casse ignorés** par défaut
  - **Restriction possible à une propriété** (=*métadonnée*) [1/2]
    - Plusieurs **métadonnées d'application** ainsi que des **métadonnées système** sont considérées comme du contenu textuel, par exemple:
      - **`titre`**, **`auteur`**, **`mots-clés`** (métadonnées d'application pour plusieurs types de fichiers);
      - **`nom`** (métadonnée système pour le nom de fichier);
      - **`datedemodification`** (métadonnée système pour la date de dernière modification);
      - **`type`** (métadonnée système pour le type de fichier, en texte);
      - **`cheminudossier`** (métadonnée système pour le chemin d'accès);
      - **`album`**, **`artistes`**, **`genre`**, **`longueur`** (durée), **`année`** (métadonnées d'application pour les fichiers musicaux);
      - **`prisedevue`** (métadonnée d'application pour la date de la prise de vue) pour les fichiers image.

Recherche de fichiers en texte intégral

Outil intégré à l'OS : Windows Search [5/6]

- Principales caractéristiques de la recherche [3/3]
  - **Restriction possible à une propriété** (=*métadonnée*) [2/2]
    - Par exemple : « nom:info » retrouvera « info » uniquement dans le nom des fichiers
    - *À noter* : on peut utiliser le « pseudo » nom de métadonnées « contenu: » pour restreindre la recherche uniquement dans le texte intégral des fichiers.
      « contenu:article » retrouverait un document où « article » se retrouve dans son contenu comme tel mais pas un document où « article » se retrouve uniquement dans le nom du fichier par exemple
  - **Parenthèses** peuvent être utilisées pour s'assurer de l'ordre d'exécution dans les requêtes ou de la portée de la restriction à une propriété
  - **Combinaison** possible, au moyen des opérateurs booléens, de critères sur le texte intégral avec des critères sur différentes métadonnées
    - *article AND type:pdf* (recherche indexée) : fichiers PDF ayant « article » dans leur contenu textuel

Recherche de fichiers en texte intégral

Outil intégré à l'OS : Windows Search [6/6]

- Conclusions sur la recherche avec Windows Search
  - Point fort : possibilité de chercher à la fois, en recherche séquentielle ou indexée, le contenu comme tel des fichiers ainsi que certaines de leurs métadonnées systèmes et d'application
  - Limité sur plusieurs aspects
    - Résultats = liste de *fichiers* (et non de passages à l'intérieur de fichiers)
    - Pas de contrôle sur l'index (par ex. pas d'antidictionnaire, non visualisable)
    - Langage d'interrogation limité (seuls les opérateurs de base comme les booléens et la troncature)
  - Pour des recherches plus performantes: les outils spécialisés comme, par exemple, NatQuest Pro