Université de Strasbourg

Recherche d'information
Introduction

Enzo Doyen

2025 - LGC6KM43 - M2

🕮 Plan

I. Recherche d'information : définition et développement
II. Principes fondamentaux de la recherche d'information

ⓘ Informations générales

- 6 séances en salle 4S04 le lundi de **15 h** à **18 h** ; 03/11, 10/11, 17/11, 24/11, 01/12 et 08/12.
- Nécessaire d'avoir un compte **Google Colab** ainsi qu'un compte **Hugging Face** pour l'évaluation et les notebooks d'exemple.

Page Moodle : **https://moodle.unistra.fr/course/view.php?id=9754**

Mot de passe : **RECHINF43**

Objectifs

1. Comprendre les concepts fondamentaux de la recherche d'information (RI).
2. Comprendre la conception des différents systèmes de RI (recherche booléenne, vectorielle, probabiliste).
3. Implémenter des systèmes de RI et mener une évaluation qualitative de ceux-ci.
4. Créer une interface permettant une utilisation simplifiée des systèmes de RI.

🎓 Évaluation

L'évaluation du cours se fait sur la base d'un **projet à rendre** à la fin du semestre. Informations détaillées à la fin de ce cours.

🕮 Plan

I. Recherche d'information : définition et développement
II. Principes fondamentaux de la recherche d'information

I. Recherche d'information : définition et développement

Définition de la RI

Définition :

"[...] finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)" **(Manning et al., 2008)**

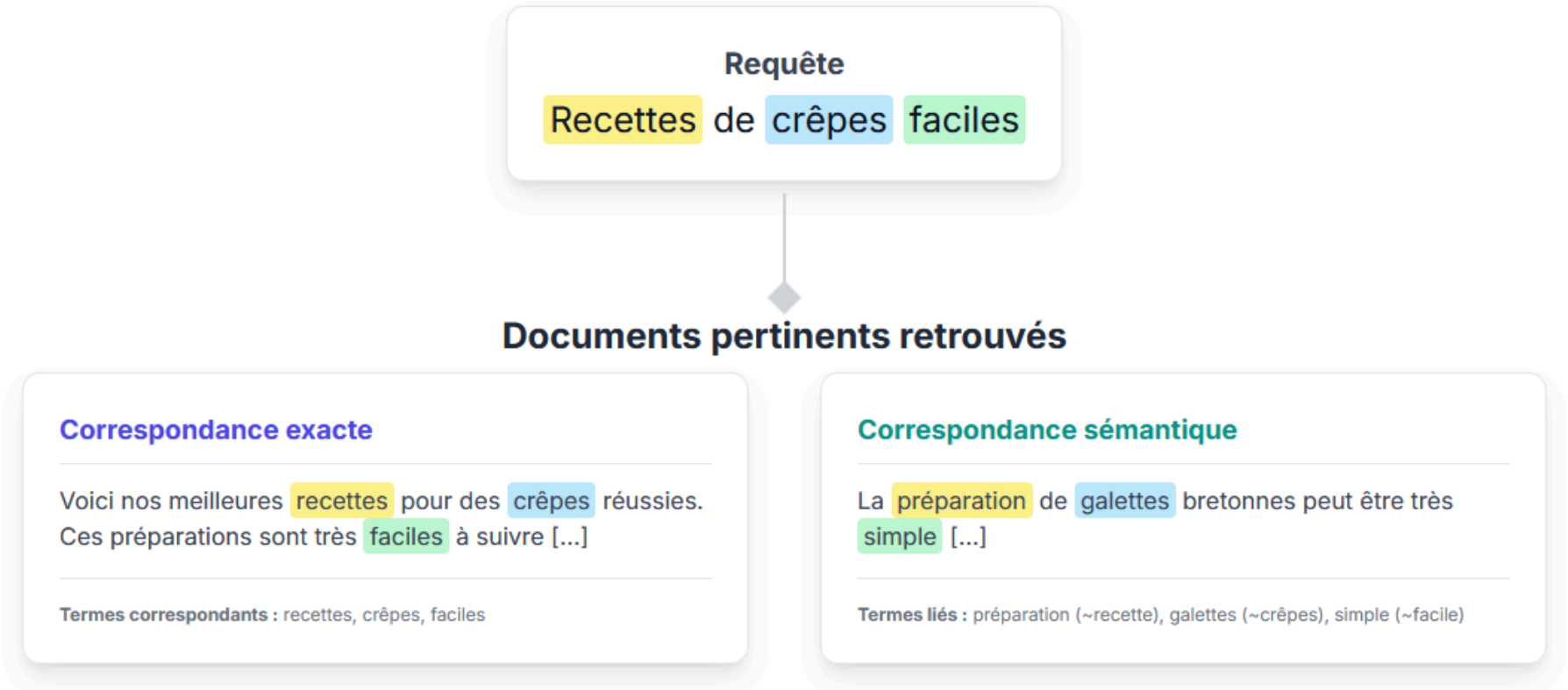
Définition de la RI

Définition :

"[...] finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)" **(Manning et al., 2008)**

Dans ce cours, nous nous concentrerons sur la recherche de documents textuels.

Exemple de RI



Recherche d'information ou extraction d'information ?

Termes parfois utilisés de manière interchangeable, mais concepts différents :

- **Recherche d'information** (RI) : processus de recherche de documents pertinents dans une collection de données à partir d'une requête.
- **Extraction d'information** (EI) : processus de récupération d'informations spécifiques à partir de documents (relations, entités nommées, etc.).

Exemples d'application de la RI

- recherche Web (Google, Bing, etc.) ;
- recherche d'e-mails ;
- recherche de documents dans des bases de données ;
- ...

Exemples d'application de la RI

Exemple de recherche d'information sur le site de la BU

Développement de la RI

> « Consider a future device … in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory. »
>
> **(Bush, 1945)**

- Appareil hypothétique (« Memex », pour « memory expansion ») imaginé par Vannevar Bush en 1945.
- Donnera plus tard naissance aux « personal knowledge bases » (Evernote, Obsidian…).

Développement de la RI

- En 1950, terme « information retrieval » utilisé pour la première fois par Calvin Mooers.
- Système de recherche à l'aide de cartes perforées développé par IBM en 1952 (décrit par **Luhn (1958)**).
  - Une des premières implémentations du **modèle booléen**.

Développement de la RI

- En 1976, implémentation d'un **modèle probabiliste** par **Robertson et Jones (1976)**.
- Application de « poids » aux termes en fonction de leur distribution dans les documents afin d'évaluer leur pertinence.

Développement de la RI

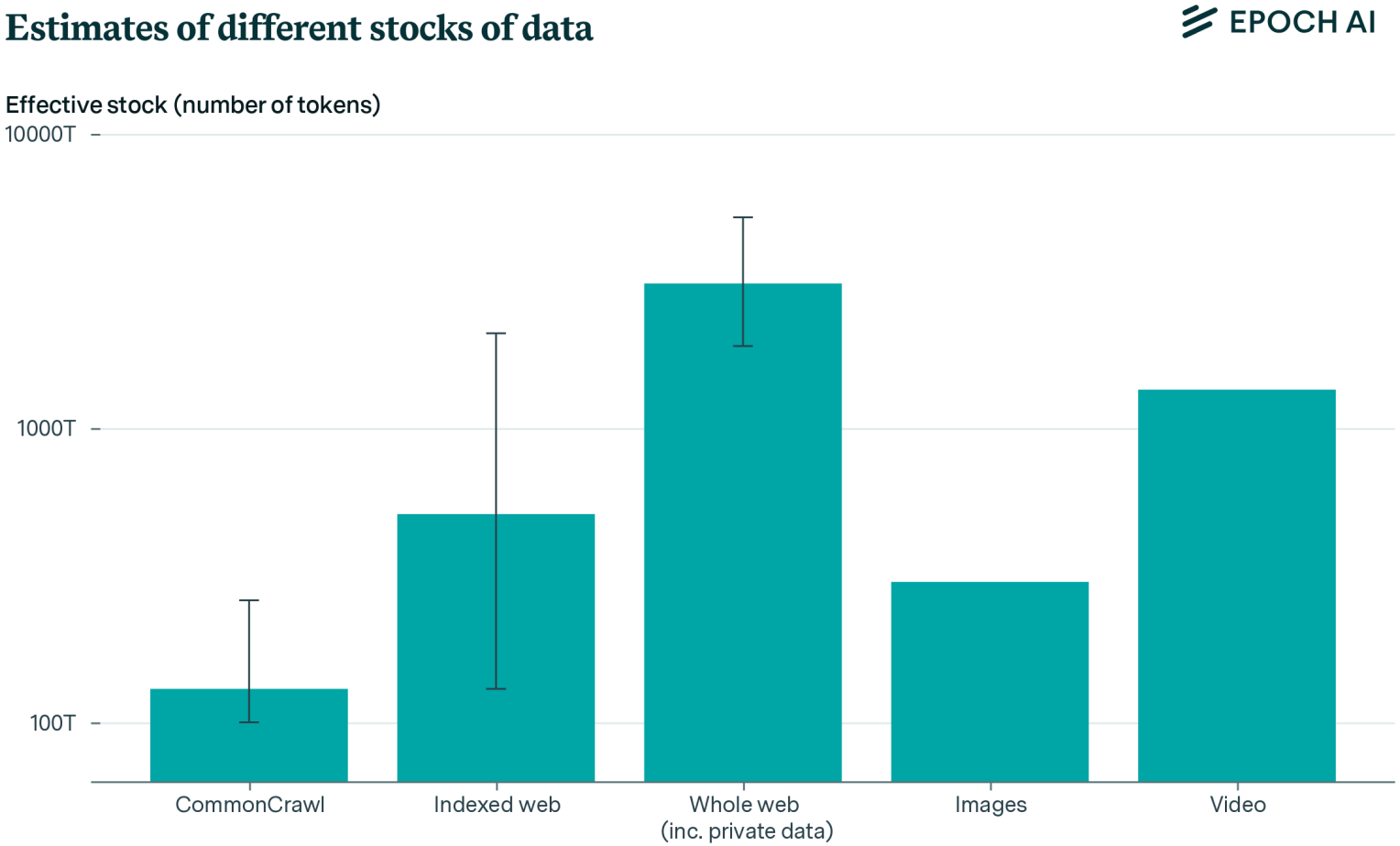
- En 1992, lancement de la première édition de la *Text Retrieval Conference* (TREC), qui incite à davantage de travaux de recherche dans le domaine.
- À partir de 1994, développement d'Internet pour le grand public et des moteurs de recherche (Lycos, Yahoo! Search, AltaVista…).
  - En 1998, naissance de Google, qui propose l'algorithme PageRank pour indexer et classer les pages Web en fonction de leur popularité et de leur pertinence **(Brin et Page, 1998)**.

Développement de la RI

- Depuis les années 2010, utilisation des avancements en *machine learning* et en traitement automatique des langues pour améliorer les systèmes de RI.
- Recherches contextuelles et personnalisées, suggestions de recherche…
- Représentation des requêtes de recherche et des documents sous forme vectorielle.

Recherche d'information et *big data*

- Phénomène de *big data* : explosion des données disponibles, notamment sur le Web.
- Comment récupérer les données les plus pertinentes dans un contexte où la quantité de données devient de plus en plus importante ?

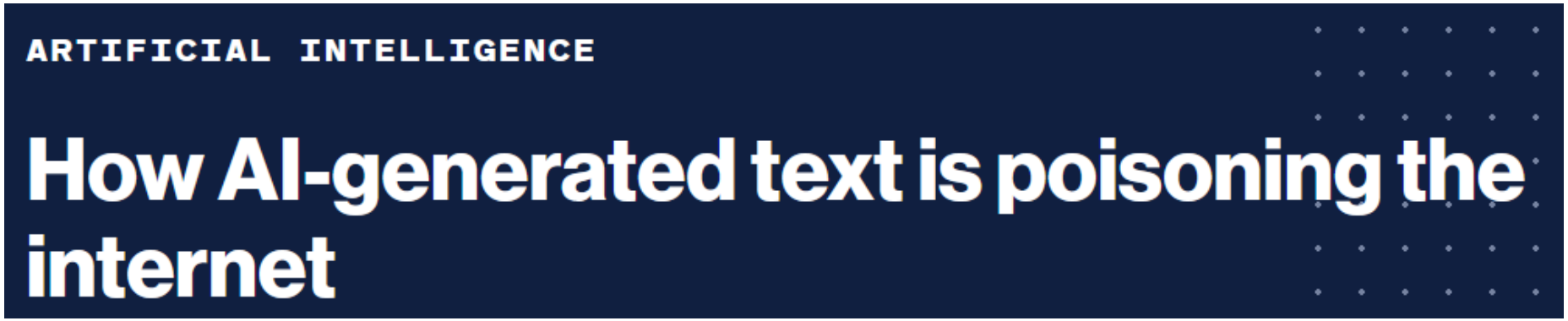


Estimation des données sur le Web | Source : https://epoch.ai/blog/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data

Recherche d'information et *big data*

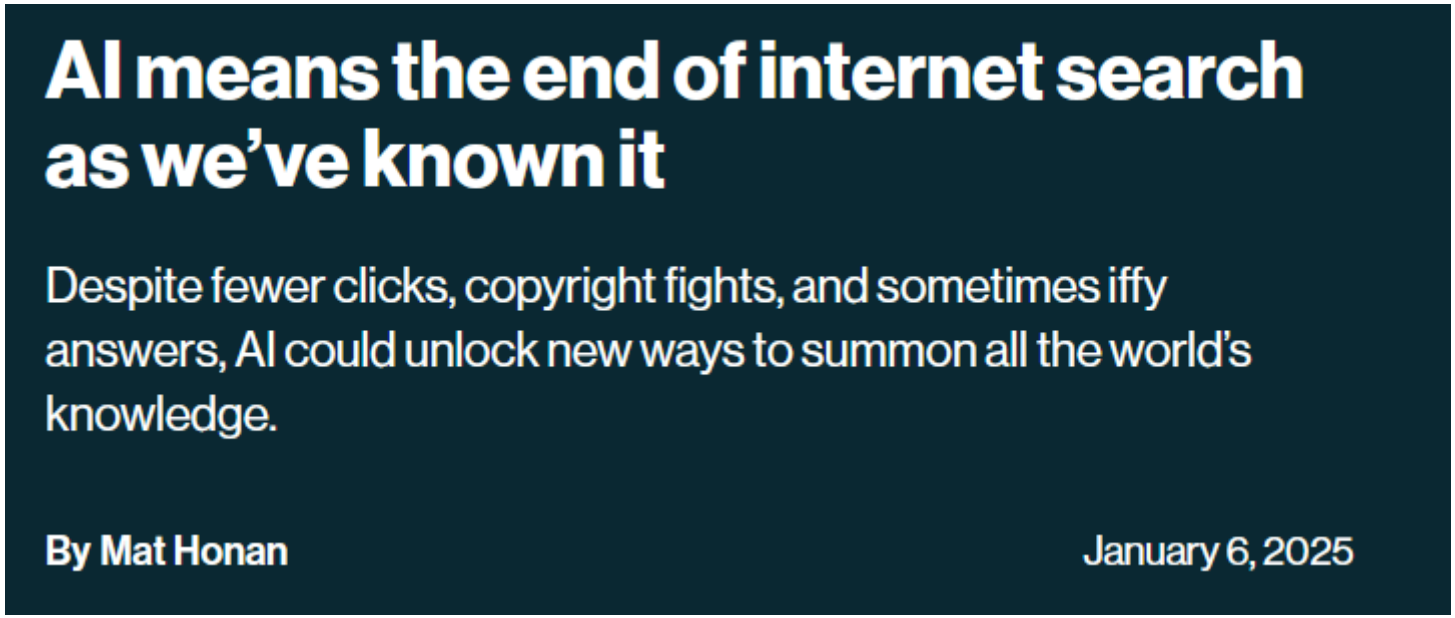
- Phénomène de *big data* : explosion des données disponibles, notamment sur le Web.
- Comment récupérer les données les plus pertinentes dans un contexte où la quantité de données devient de plus en plus importante ?
- Plus récemment, l'émergence de grands modèles de langue génératifs (LLM) a conduit à une publication de masse de contenus générés par IA sur Internet.
  - Elle pose aussi des questions sur l'**utilisation future des moteurs de recherche**.

Recherche Web à l'ère des LLM



Source : https://www.technologyreview.com/2022/12/20/1065667/how-ai-generated-text-is-poisoning-the-internet/

Recherche Web à l'ère des LLM



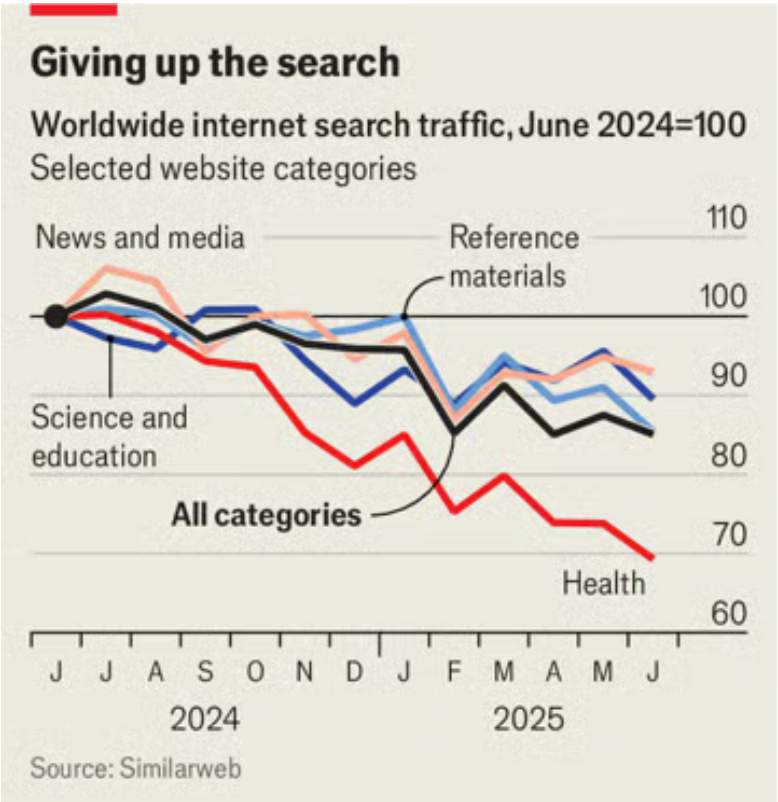
Source : https://www.technologyreview.com/2025/01/06/1108679/ai-generative-search-internet-breakthroughs/

Recherche Web à l'ère des LLM

« The biggest change to the way search engines have delivered information to us since the 1990s is happening right now. No more keyword searching. No more sorting through links to click. Instead, we're **entering an era of conversational search**. Which means **instead of keywords, you use real questions, expressed in natural language**. And **instead of links, you'll increasingly be met with answers, written by generative AI and based on live information from all across the internet**, delivered the same way. »

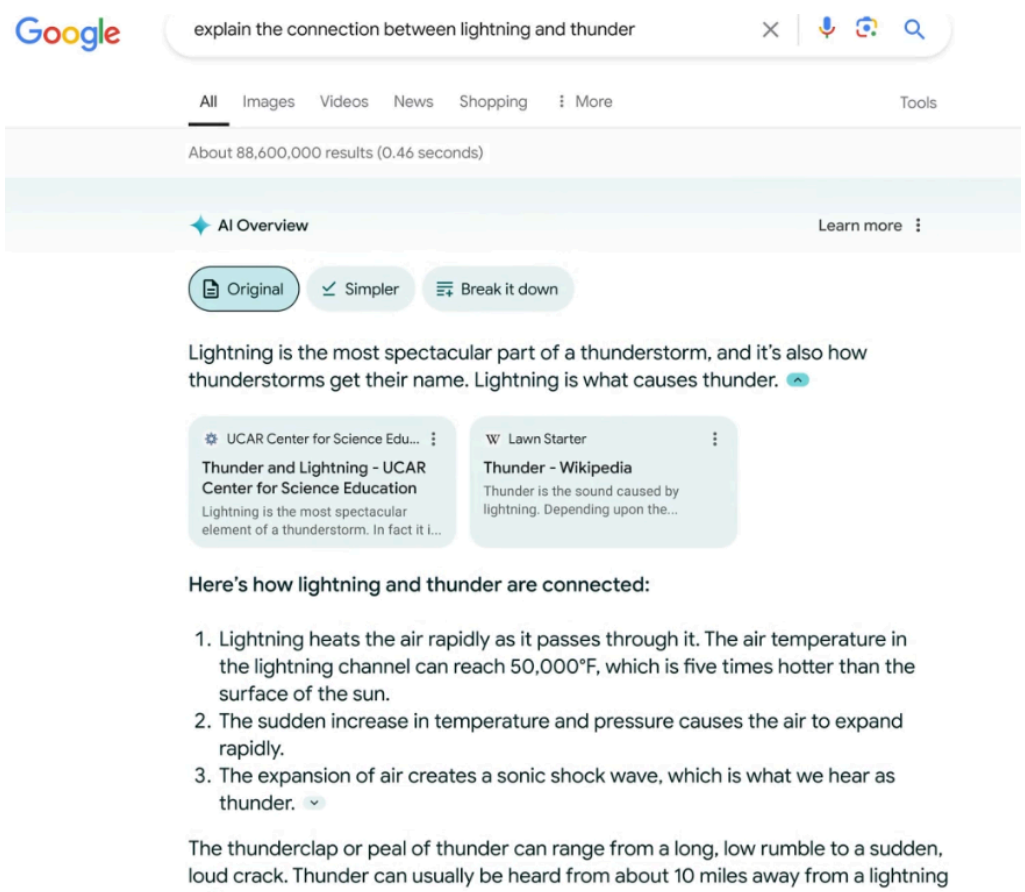
Mat Honan, AI means the end of internet search as we've known it, *MIT Technology Review*,
https://www.technologyreview.com/2025/01/06/1108679/ai-generative-search-internet-breakthroughs/

Recherche Web à l'ère des LLM



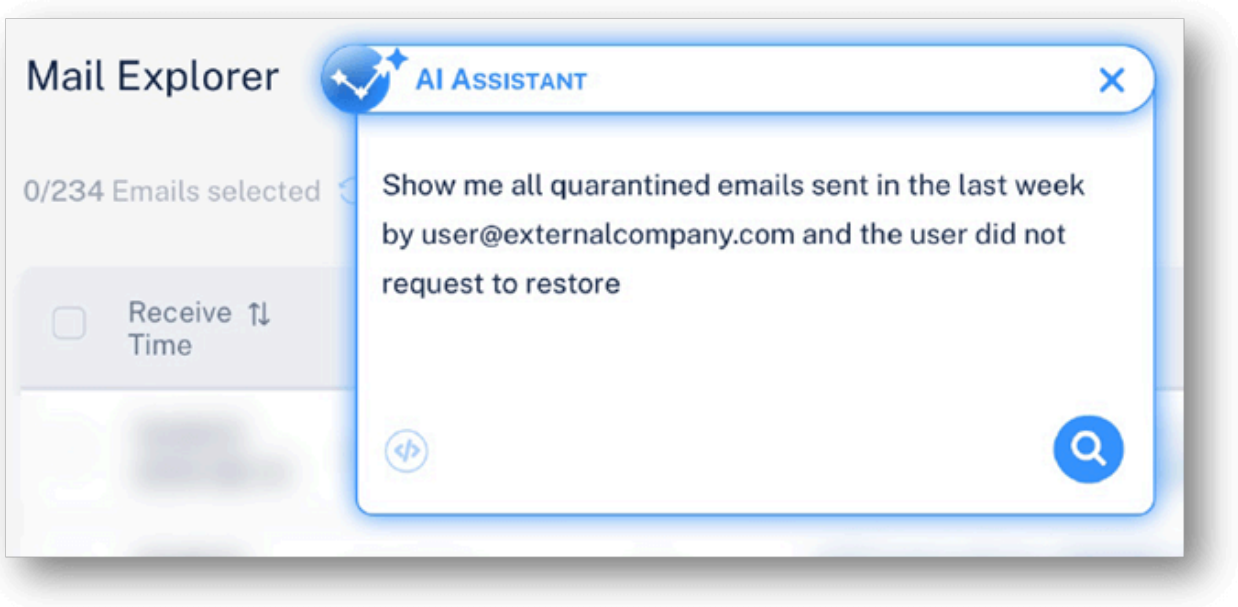
Source : https://www.economist.com/business/2025/07/14/ai-is-killing-the-web-can-anything-save-it

Recherche Web à l'ère des LLM : intégration



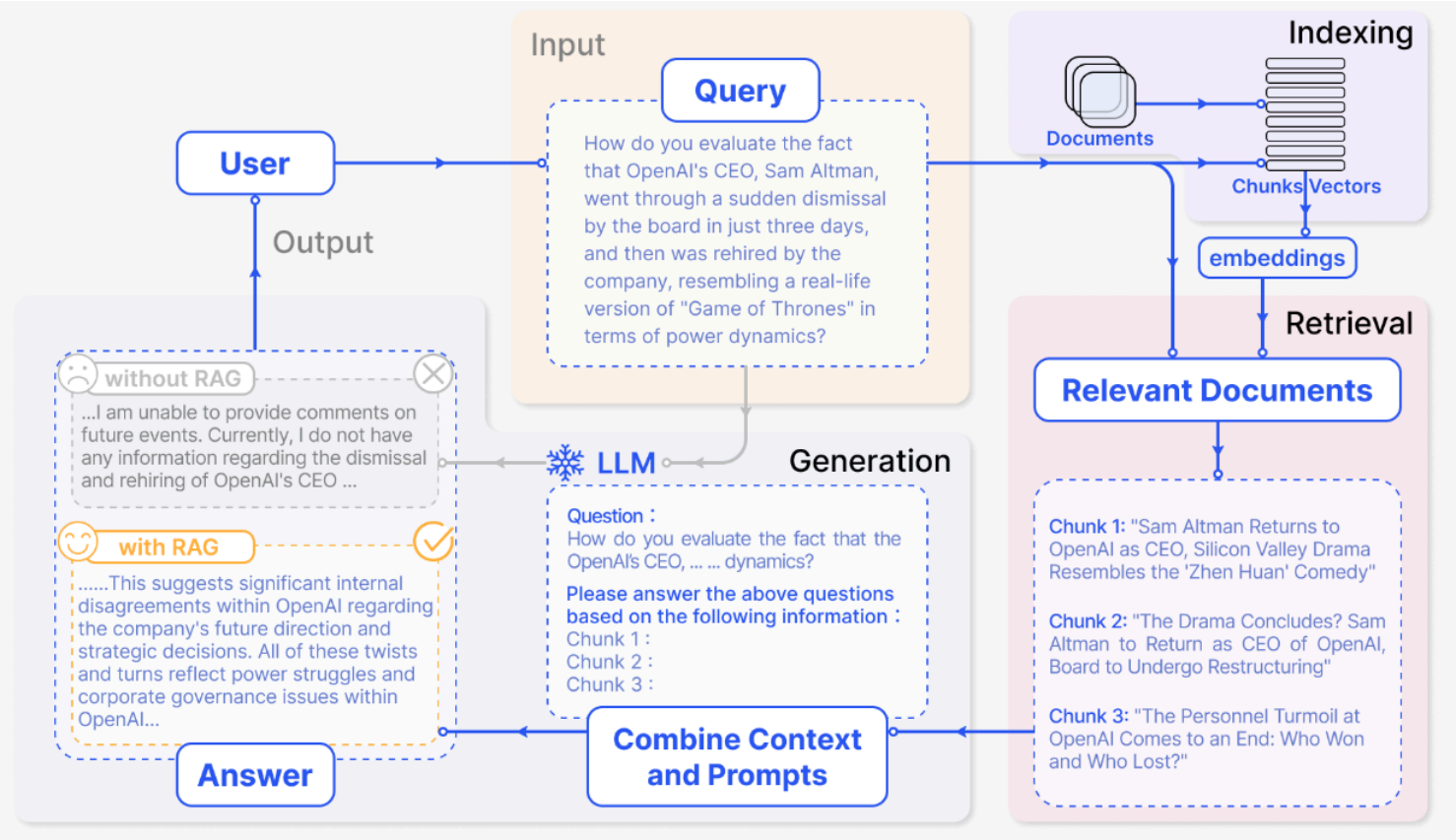
Source : https://blog.google/products/search/generative-ai-google-search-may-2024/

Recherche d'e-mails à l'ère des LLM

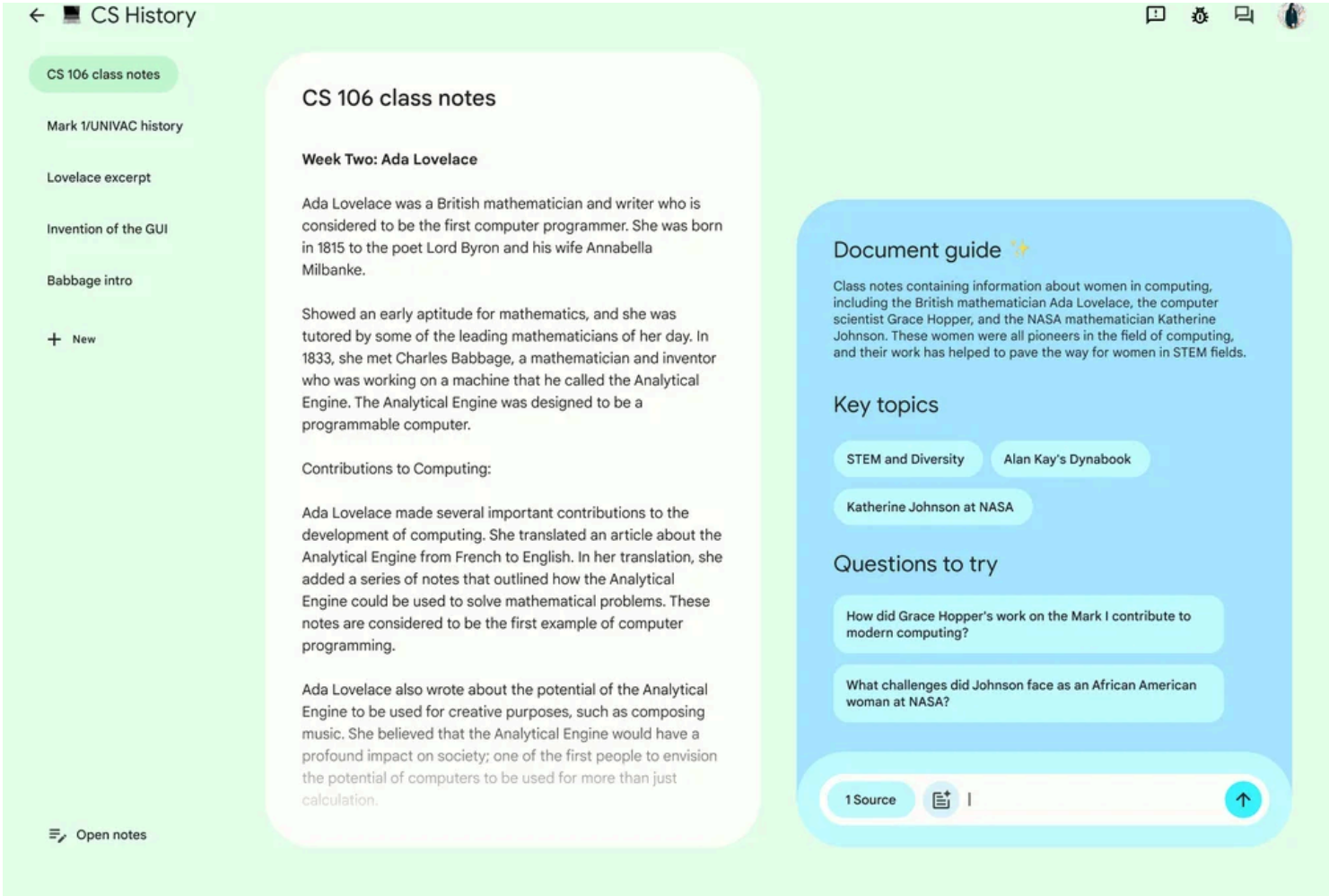


Source : https://emailsecurity.checkpoint.com/product-updates/ai-assistant-email-search-powered-by-generative-ai

Recherche d'information pour les LLM : *RAG*



Source : **Gao et al. (2024)**



Source : https://blog.google/technology/ai/notebooklm-google-ai/

II. Principes fondamentaux de la recherche d'information

RI : l'utilisateur ou l'utilisatrice avant tout

- La recherche d'information est centrée sur l'utilisateur ou l'utilisatrice.
- L'objectif est de répondre à un **besoin d'information** spécifique.
- Il faut alors penser à :
  - qui est le public visé (qui va utiliser le système de RI ?) ;
  - quel est le type de documents à rechercher et le domaine ;
  - comment il/elle est susceptible de formuler sa requête ;
  - quels outils de recherche mettre à sa disposition (système de recherche avancée ? si oui, quelles options ?) ;
  - comment lui proposer les résultats les plus pertinents possibles…
    - … quand bien même la requête n'est pas forcément formulée correctement, ou de manière exacte.

RI : pertinence et évaluation

- Les documents indexés sont-ils pertinents par rapport à la requête ? Répondent-ils au besoin d'information exprimé ?
  - Comment évaluer cette pertinence ?
- Les documents retournés sont-ils à jour ? Fiables ?
  - Flux d'informations en continu, *fake news*…

RI : pertinence et évaluation

La **précision** et le **rappel** sont deux mesures couramment utilisées pour évaluer la pertinence des résultats retournés par un système de RI.

	Pertinent	Non pertinent
Récupéré	TP	FP
Non récupéré	FN	TN

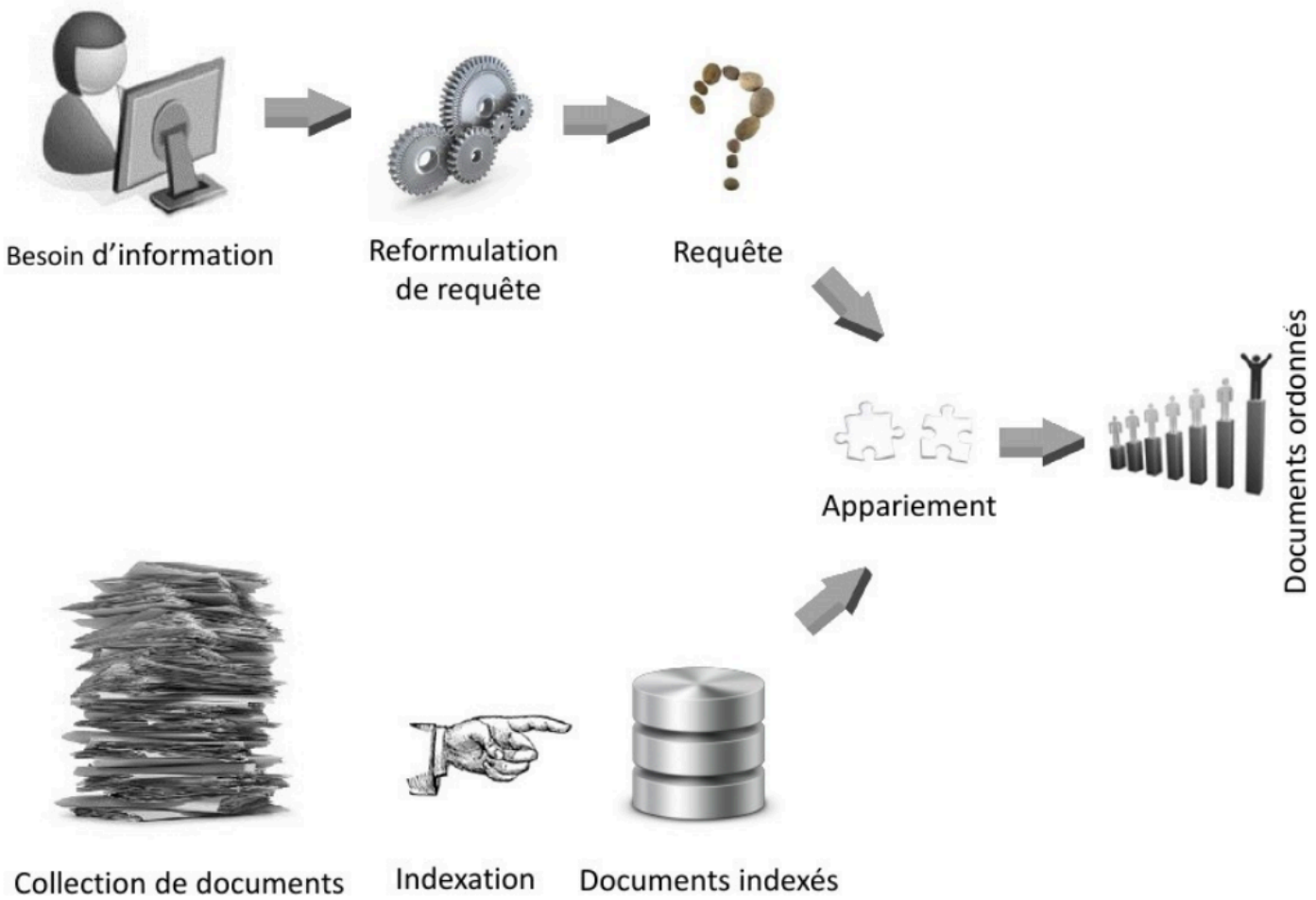
Tableau de contingence (adapté de **Manning et al. (2008:155)**)

$$\text{Précision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Rappel} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Étapes et processus d'un système de RI

1. Prétraitements des documents (nettoyage, normalisation…).
2. Indexation des documents :
   - représentation des documents sous un format exploitable par le système de RI.
3. Prétraitements de la requête.
4. Appariemment de la requête avec les documents indexés :
   - recherche des documents pertinents en fonction de la requête.
5. Classement des documents pertinents.



Étapes d'un système de recherche d'information **(Amini et al., 2013)**

Bibliographie

Amini, M.-R., Gaussier, É., et Péan, G. (2013). *Recherche d'information : applications, modèles et algorithmes*. Eyrolles.

Brin, S., et Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks*, *30*, 107-117.

Bush, V. (juillet 1945). As We May Think. *The Atlantic*.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., et Wang, H. (mars 2024). *Retrieval-Augmented Generation for Large Language Models: A Survey* (Numéro arXiv:2312.10997). arXiv. **10.48550/arXiv.2312.10997**

Luhn, H. P. (1958). *The IBM Electronic Information Searching System*. International Business Machines Corp., Research Center.

Manning, C. D., Raghavan, P., et Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Mitra, B. (2018). *An Introduction to Neural Information Retrieval* (Numéro v.41). Now Publishers.

Robertson, S. E., et Jones, K. S. (1976). Relevance Weighting of Search Terms. *Journal of the American Society for Information Science*, *27*(3), 129-146. **10.1002/asi.4630270302**

Zhai, C., et Massung, S. (juin 2016). *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. Association for Computing Machinery and Morgan & Claypool. **10.1145/2915031**

♡ Remerciements

- Pablo Ruiz Fabo pour le contenu de certaines diapositives.