S’augmenter

Comment les LLMs transforment l’analyse de contenu, et les sciences sociales

Slides disponibles sur
www.ollion.cnrs.fr

Étienne Ollion
et le groupe CSS de l’Institut polytechnique de Paris

Étienne Ollion



Paola Tubaro



Samuel Coavoux



Julien Boelaert



Paul Coudry



Yacine Chitour



Emma Bonutti d'Agostini



Hesu Yoon



Émilien Schultz



Vincent Verger



Léo Labat



Axel Morin



Yasmine Houri



Arnault Chatelain



Annina Claesson



Senmiao Yang



Guilhem Sicard



Felix Lennert



www.css.cnrs.fr

Introduction

- **L’IA s’immisce partout… même en sciences sociales**

Introduction

- L'Intelligence artificielle s'immisce partout… même dans les sciences
- **Quid de l'analyse de contenu ?**
  - Un domaine ancien
  - Deux déclinaisons
  - Un réinvestissement depuis peu, porté par les LLMs

Introduction

- L'Intelligence artificielle s'immisce partout... même dans les sciences
- Quid de l'analyse de contenu ?
- **Que peuvent les LLMs ?**
  - Des avancées importantes
  - Différentes dimensions : exploration, annotation, extraction d'information
  - Des limites, et des critiques

Introduction

- **Cette présentation**
  - Une présentation critique des approches disponibles
  - Argument: sous conditions, il est possible à la chercheuse de **s'augmenter**
    - Annoter/extraire avec ~ *le même niveau de précision*
    - Le faire *à une échelle bien plus grande*
    - Le faire *sans se déqualifier*

Introduction

1. L’affaire n’est plus dans le sac
2. S’augmenter : méthodes
3. Des opportunités à différents niveaux
4. Des difficultés, selon les outils

Conclusion: S’augmenter?

L’affaire n’est plus dans le sac

- Le langage pour objet, une histoire ancienne (Weber, 1910; Berelson, 1952)

L’affaire n’est plus dans le sac

- Le langage pour objet, une histoire ancienne (Weber, 1910; Berelson, 1952)
- Différentes approches, mais un point commun : l’hypothèse du “sac de mots”

L’affaire n’est plus dans le sac

	Phèdre	Hippolyte	Mickey	Minnie	aime
Phèdre aime Hippolyte	1	1	0	0	1
Mickey aime Minnie	0	0	1	1	1

L’affaire n’est plus dans le sac

- Un point commun : l’hypothèse du “sac de mots”
  - “A shocking hypothesis” (Grimmer and Stewart, 2013)
  - D’indéniables succès
    - Décomptes (Moretti), co-occurences (Reinert), stat textuelle (Lebart)
    - Analyses thématiques [Blei & Di Maggio, 2013]
    - Stylométrie [Mosteller, 2012]
    - Réseaux sémantiques (Rule et al., 2015)
    - ...

L’affaire n’est plus dans le sac

Brève histoire de l’analyse de contenu (textuel)

- Un point commun : l’hypothèse du “sac de mots”
  - “A shocking hypothesis” (Grimmer and Stewart, 2013)
  - D’indéniables succès
  - **Mais des limites**
    - Ordre
    - Polysémie
    - Et plus généralement, granularité

L’affaire n’est plus dans le sac

Les Transformers, une représentation alternative du langage

L’affaire n’est plus dans le sac

Les Transformers, une représentation alternative du langage

- Une architecture, basée sur des réseaux de neurones profonds
- **Au coeur des principales technologies actuelles**
  - Generative Pretrained Transformers (GPTs), BERT
  - Utilisés pour le texte, l’image, le son

L’affaire n’est plus dans le sac

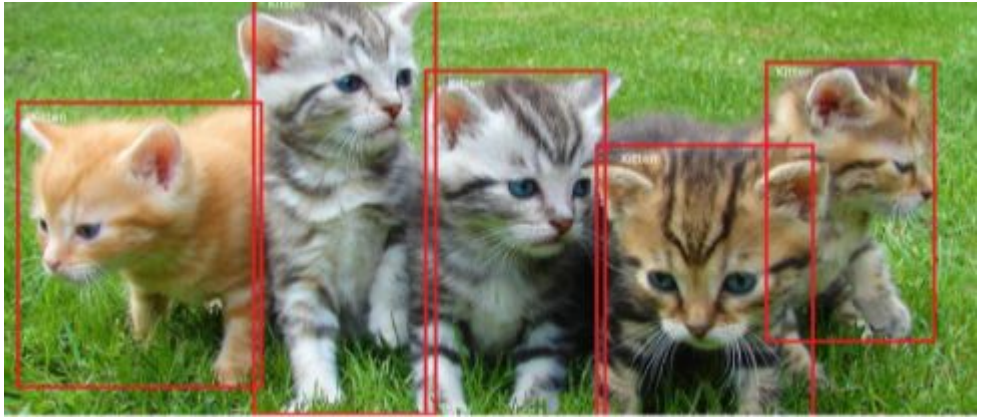
Les Transformers, une représentation alternative du langage

- Une architecture basée sur des réseaux de neurones profonds (Vaswani, 2017)
- Au coeur des principales technologies actuelles
- **Une nouvelle manière de représenter le langage**
  - Auto-attention : chaque mot a accès à tous les mots (précédents)
  - Nombreuses couches successives : contextualisation

L’affaire n’est plus dans le sac

Les Transformers, une représentation alternative du langage

- Une architecture basée sur des réseaux de neurones profonds (Vaswani, 2017)
- Une nouvelle manière de représenter le langage
- **D’une analyse lexicale à une sémantique du langage**



S’augmenter: méthodes

S’augmenter: méthodes

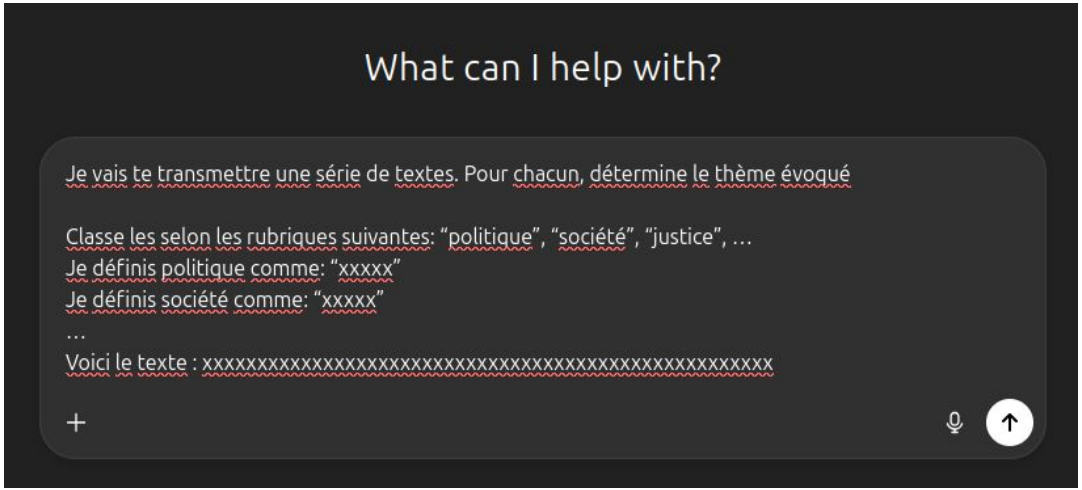
Deux grandes approches

S’augmenter: méthodes

Deux grandes approches

1. Prompter un LLM (decoders)

Utiliser des IA génératives



S’augmenter: méthodes

Deux grandes approches

1. Prompter un LLM (decoders)

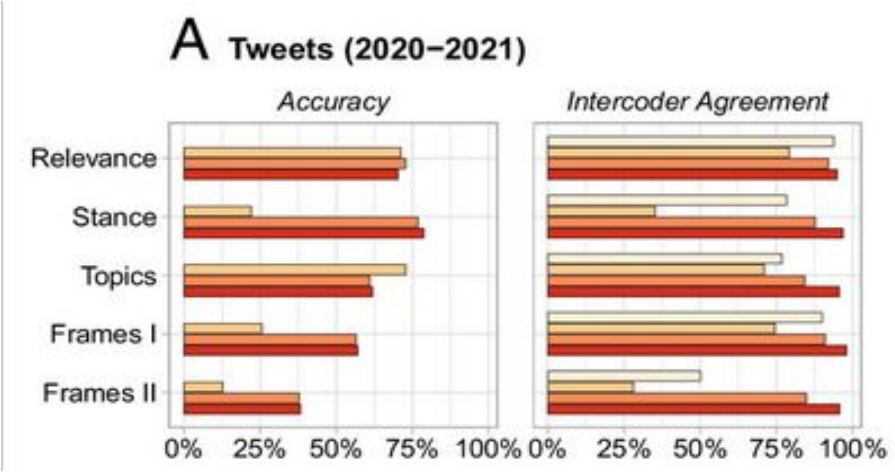
Utiliser des IA génératives

ChatGPT outperforms crowd workers for text-annotation tasks

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli Authors Info & Affiliations

Edited by Mary Waters, Harvard University, Cambridge, MA; received March 27, 2023; accepted June 2, 2023

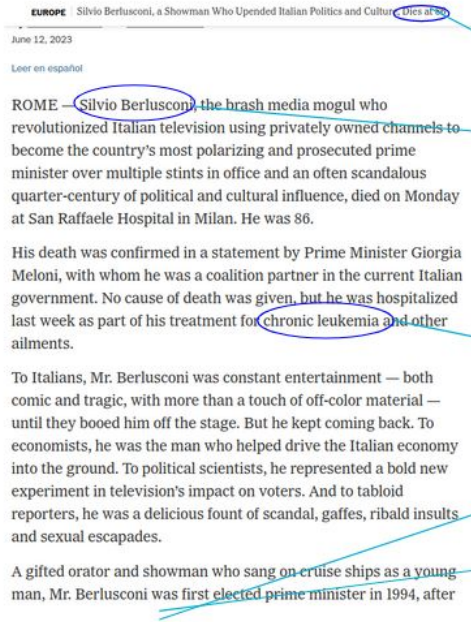
July 18, 2023 | 120 (30) e2305016120 | https://doi.org/10.1073/pnas.2305016120



S’augmenter: méthodes

1. Prompter un LLM (decoders)

Classification de texte, mais aussi extraction d’information



EUROPE Silvio Berlusconi, a Showman Who Upended Italian Politics and Culture, Dies at 86

June 12, 2023

Leer en español

ROME — Silvio Berlusconi, the brash media mogul who revolutionized Italian television using privately owned channels to become the country’s most polarizing and prosecuted prime minister over multiple stints in office and an often scandalous quarter-century of political and cultural influence, died on Monday at San Raffaele Hospital in Milan. He was 86.

His death was confirmed in a statement by Prime Minister Giorgia Meloni, with whom he was a coalition partner in the current Italian government. No cause of death was given, but he was hospitalized last week as part of his treatment for chronic leukemia and other ailments.

To Italians, Mr. Berlusconi was constant entertainment — both comic and tragic, with more than a touch of off-color material — until they booed him off the stage. But he kept coming back. To economists, he was the man who helped drive the Italian economy into the ground. To political scientists, he represented a bold new experiment in television’s impact on voters. And to tabloid reporters, he was a delicious fount of scandal, gaffes, ribald insults and sexual escapades.

A gifted orator and showman who sang on cruise ships as a young man, Mr. Berlusconi was first elected prime minister in 1994, after

Name	Silvio Berlusconi
Date of birth	Sept. 29, 1936
Age at Death	86
Born in	Milan
Cause of death	Chronic leukemia
Children	5 (3 daughters, 2 sons)
...	

(Stuhler, Dang Ton, Ollion, 2025)

S’augmenter: méthodes

2. Affiner un LLM (encoders)

Entraîner un modèle pour qu’il sache classifier comme le ferait la chercheuse

S’augmenter: méthodes

2. Affiner un LLM (encoders)

Entraîner un modèle pour qu’il sache classifier comme le ferait la chercheuse

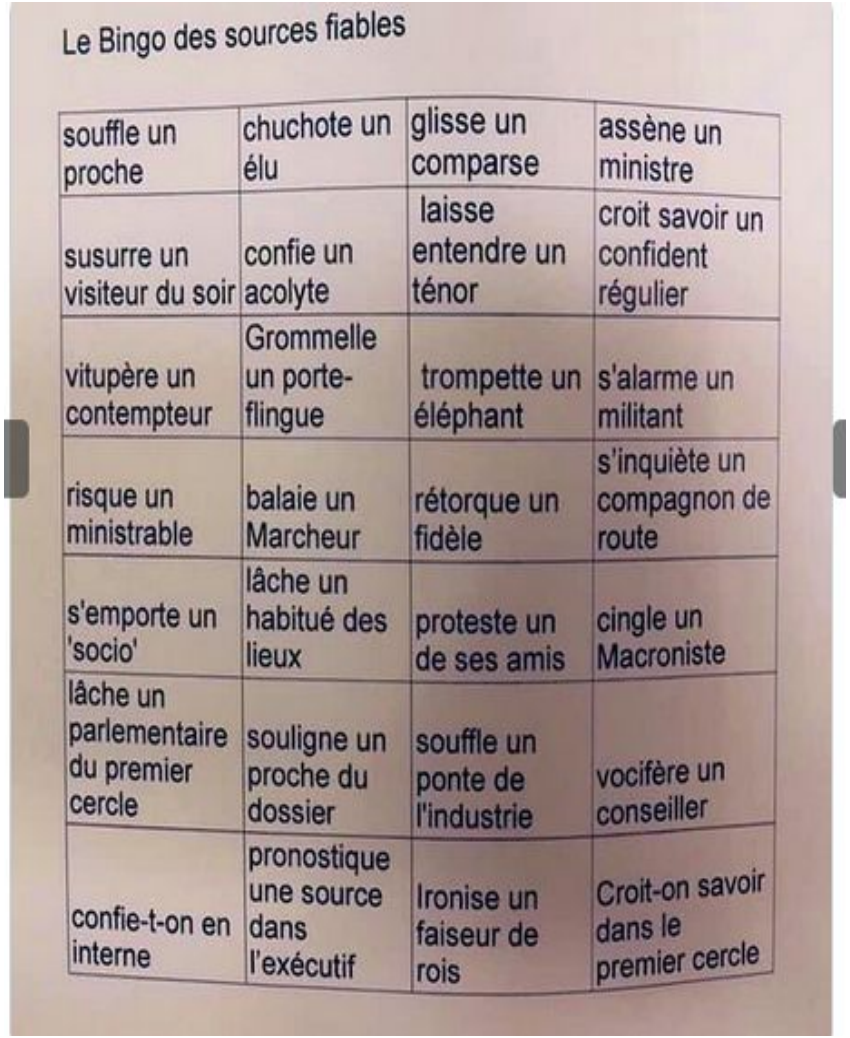
- Annotation manuelle d’exemples (positifs, négatifs), en quantité raisonnable
- Qu’on donne ensuite au modèle pour “l’affiner”
- Et auquel on demande de prédire

Preuve: (Do, Ollion et Shen, 2022)

S’augmenter: méthodes

2. Affiner un LLM (encoders)

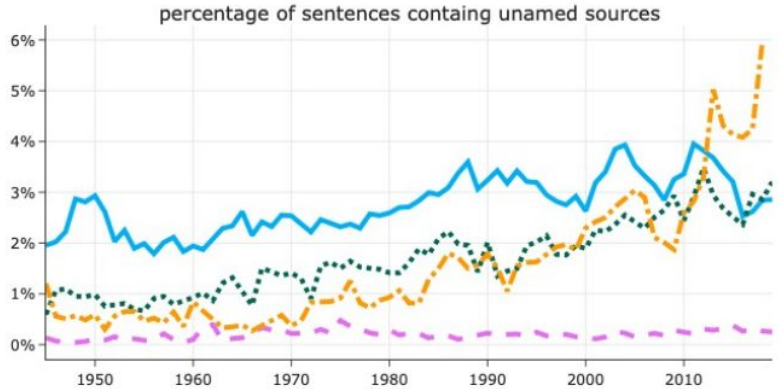
Ex: Déterminer les éléments d’un article de presse qui introduisent des propos “non attribués” (Shen, 2024)



S’augmenter: méthodes

2. Affiner un LLM (encoders)

Pour le D-Day , d ' accord , mais pour une remise de décoration , inutile ,
s ' agace un collaborateur.
off
Certains visiteurs du soir l ' ont vu éteindre lui-même la lumière au palais , alors qu ' il les raccompagnait. . . L ' obsession du détail est plutôt rassurante dans l ' organisation de l ' Etat , mais il ne faut pas faire rentrer les Français dans la cuisine ,
estime un communicant.
off
Beaucoup estiment ainsi qu
off
' un sommet fut atteint lors de l ' interview menée par Jean-Jacques Bourdin , pour le deuxième anniversaire de son élection , et surtout du dialogue organisé à cette occasion entre le président et les auditeurs de RMC .



Shen 2024

S’augmenter: méthodes

2. Affiner un LLM (encoders)

Des centaines d’exemples, et ce n’est que le début

- Le climato-scepticisme dans la presse (Luo et al., 2020)
- Le “populisme” dans les discours présidentiels (Bonikowski et al., 2022)
- Les groupes sociaux des partis politiques (Licht, Scepanzski, 2025)
- Le genre dans les sciences sociales (Boelaert et al., 2025)
- Ce qu’on dit aux élus (Claesson, 2025)
- ...

S’augmenter: méthodes

Dans les deux derniers cas, il faut toujours un protocole de validation

Phrase	Machine	Expert	
Une source proche du pouvoir rapporte que	Off	Off	✅
Selon un cadre de la majorité	Pas Off	Off	❌
L’histoire du parti faite de départs	Pas Off	Pas Off	✅
Un ministre, qui veut rester anonyme, affirme que	Off	Off	✅

Des opportunités à différents niveaux

Des opportunités à différents niveaux

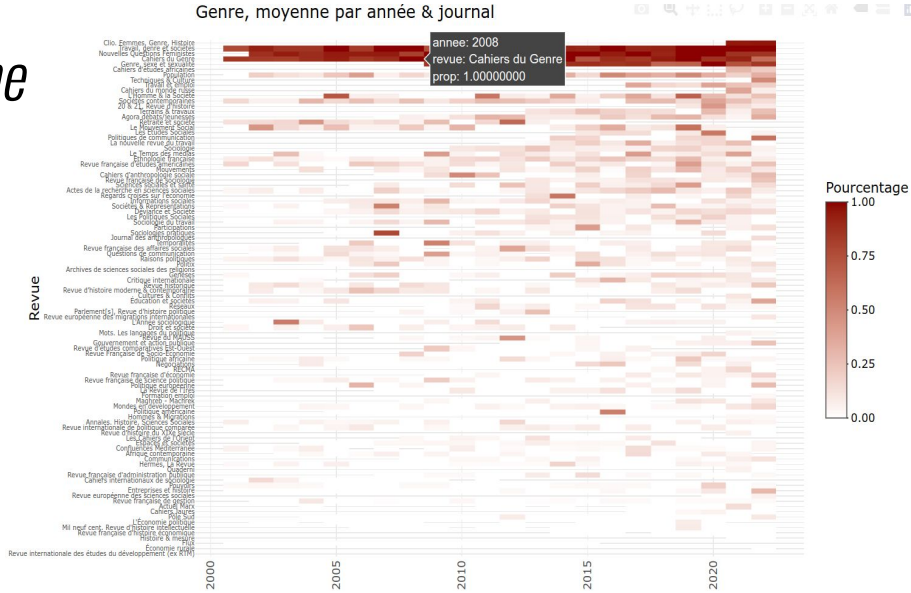
De manière rapide, simplifiée, et vérifiable, on obtient une annotation

Des opportunités à différents niveaux

1. Des données de choix

Des données préparées *pour la recherche*

- Choix des catégories
- Choix du périmètre



(Boelaert et al, 2025)

Des opportunités à différents niveaux

2. Ce que permet le nombre

- Des données plus granulaires, pour varier d'échelle
- Et si on peut tout analyser, la fin du problème de la représentativité?



Des opportunités à différents niveaux

3. L’atténuation d’oppositions classiques

- Internalisme et externalisme
- Quali, quanti, ... et ordi?

Des difficultés, qui varient selon les outils

Des difficultés, qui varient selon les outils

Des compétences techniques, malgré une barrière abaissée

- Savoir coder (un peu)
- Connaître les possibilités, les bonnes pratiques

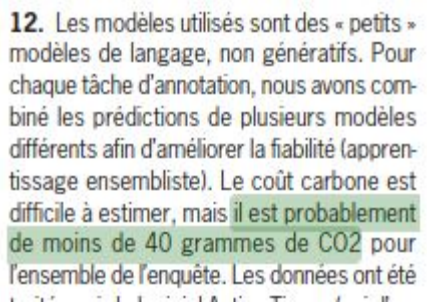
https://www.css.cnrs.fr/active-tigger/



Des difficultés qui varient selon les outils

Des coûts importants

- Puissance de calcul (besoin potentiel de GPU, et de spécialistes)
- Impact énergétique et écologique (Luccioni *et al.*, 2024)
- Qui varient radicalement d'un outil à l'autre



12. Les modèles utilisés sont des « petits » modèles de langage, non génératifs. Pour chaque tâche d'annotation, nous avons combiné les prédictions de plusieurs modèles différents afin d'améliorer la fiabilité (apprentissage ensembliste). Le coût carbone est difficile à estimer, mais il est probablement de moins de 40 grammes de CO_2 pour l'ensemble de l'enquête. Les données ont été

(Boelaert et al, 2025)

Des difficultés qui varient selon les outils

Protection des données

- Les données identifiantes ou propriétaires **ne peuvent être transférées à un service tiers**
- Dans bien des cas de SHS, les traitements doivent donc être faits en interne

Des difficultés qui varient selon les outils

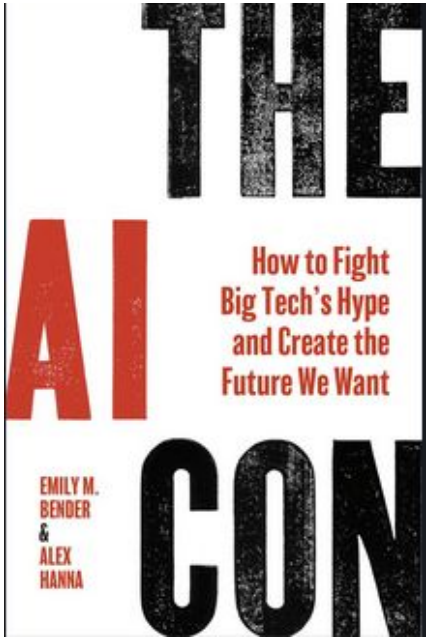
Réplicabilité des résultats

- Modèles fermés empêchent la réplicabilité (Barrie et al., 2024)
- Modèles ouverts sont stochastiques, donc pas de réplicabilité parfaite
- … mais si bien fait, une chance pour la réplication? (Stuhler et al., 2025)

Des difficultés qui varient selon les outils

Qui s'ajoutent aux problèmes bien connus des LLMs

- Quelles données, obtenues comment ?
- Transformations brutales du marché du travail



Des difficultés qui varient selon les outils

	Encoder	Décoder, ouvert	Décodeur, fermé
Exemples	BERT	Llama, Mixtral, Qwen, EuroLLM	chatGPT, Aya, Claude
Usage	Ordinateur personnel, serveur local	Serveur local, API	API
Tâches	Classification de texte	Classification de texte; extraction d'informations diverses	Classification de texte; extraction d'informations diverses
Fonctionnement	Annotation d'exemples	Prompt; prompt + exemples	Prompt; prompt + exemples
Facilité d'usage	Relativement simple	Simple	Simple
Performance	SOTA sur la classification, pas d'extraction d'information	Variable mais potentiellement très bon	Variable, souvent très bon
Coût énergétique	Limité	Important à très important	Important à très important
Protection des données	Bonne (locale)	Bonne si stockage local	Limitée à nulle
Réplication	Aisée	Possible, sous contraintes	Impossible

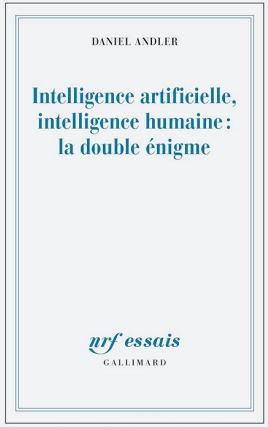
Conclusion: s’augmenter

Conclusion: s’augmenter

- Un débat ancien : les machines et le travail

Conclusion: s’augmenter

- Un débat ancien : les machines et le travail
- Remplacer, ou augmenter?
  - Doug Engelbart, the Internet and "The Augmentation Research Lab"



"*Increasing the capability of a man to approach a complex problem situation,to gain comprehension to suit his particular needs, and to derive solutions to problems.*" (Engelbart, *Augmenting Human Intellect*, 1962).

Conclusion: s’augmenter

- Un débat ancien : les machines et le travail
- Remplacer, ou augmenter?
- Les conditions de l’augmentation

Merci!

Étienne Ollion



Émilien Schultz



Julien Boelaert



Yacine Chitour



Slides disponibles sur demande
Tutoriels, matériel disponibles sur www.css.cnrs.fr

Annoter sans difficultés? ActiveTigger!