

TAL: de la technologie à l'industrie

Hugues de Mazancourt
Directeur Innovation, Meteojob
Fondateur, Mazancourt Conseil
hugues@mazancourt.com

Qui suis-je ?

- Ingénieur, débute en 1989 à Erli (Etudes et Réalisations en Linguistique Informatique), pionnier du TAL industriel en France et des moteurs de recherche intelligents
 - ENSTA, puis DEA d'IA à Paris 6
- Co-fondateur de Lingway en 2001 (spin-off d'Erli devenu LexiQuest), directeur technique.
- Directeur Innovation d'Eptica et CTO d'Eptica-Lingway 2013-2016
- Directeur Innovation à Meteojob et consultant indépendant
- Mais aussi:
 - Fondateur de l'APIL (Association des Professionnels des Industries de la Langue), membre du CA de l'ATALA
 - Co-rédacteur de la charte Ethique et Big Data, du blog ethique-et-tal.org
 - Expert à Cap Digital (Pôle de Compétitivité)
- Réalisations : moteurs de recherche sémantiques dédiés (3611, CVthèques...), text-mining, analyse d'opinion, réseaux sémantiques et outils de recrutement. Seulement dans l'écrit.

Et vous ?

Organisation du cours

- Plusieurs thématiques abordées
 - Le TAL dans la relation client
 - Le TAL dans les RH
 - Les métiers du TAL
 - Le TAL comme développement industriel
 - Utilisation des toolkits existants (+ word embeddings)
 - TAL et éthique
- Intervenants
 - François-Régis Chaumartin, Président de Proxem
 - Gaëlle Recourcé, CSO à Kwaga/Evercontact (à confirmer)
 - Enguerrand Spindler, CEO Alfstore Group (à confirmer)
- Un mini-projet
- Les supports sur <http://bit.ly/M2-Diderot>

Pour moi, tout a commencé avec ça...



... et Eliza (Weizenbaum, 1960)

- Le premier chatbot de l'histoire
- Simule un psychothérapeute rogérien en reformulant les questions du patient
 - fonctionne uniquement en anglais, écrit dans un ancêtre de LISP
- On trouve beaucoup de clones d'Eliza sur le net et ailleurs
 - Dans emacs, tapez : M-x doctor
- Beaucoup ont entretenu des discussions très émotionnelles avec le programme, ce qui veut dire qu'il a passé le test de Turing
 - cf. Weizenbaum : « *Je ne m'étais jamais rendu compte que ... de si courtes interactions avec un programme informatique relativement simple risquaient d'induire des pensées délirantes chez des personnes pourtant normales* »
- Joseph Weizenbaum a écrit en 1976 « Computer power and human reason »
 - Introduit la distinction entre les activités de **décision** (programmable) et de **choix** qui est le résultat d'un jugement humain (inclut en particulier les émotions)

Un dialogue avec Eliza

I am the psychotherapist. Please, describe your problems. Each time you are finished talking, type RET twice.

I talk to computers

ELIZA: You have your mind on computers, it seems.

i am worried about them

ELIZA: Are you worried often?

I am like any other person

ELIZA: Why do you say are like any other person?

I'm scared that artificial intelligence will take over the world

ELIZA: Why do you say you are afraid of something?

because it's true !

ELIZA: Can you elaborate on that?

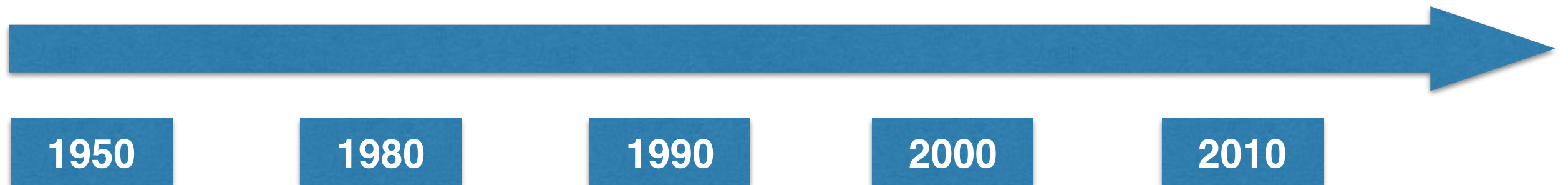
Comment ça marche

- Quelques mots-clés qui déclenchent des questions spécifiques pour relancer le dialogue
 - computer, scared, happy, unhappy, mother, father
- Des ré-écritures de la phrase lorsque le pattern est reconnu
 - « *I am ...* » -> « *are you talking to me because you are ...* »
 - (facile en anglais)
- Des relances standard
 - « *Can you elaborate on that?* »

Une brève histoire du TAL industriel

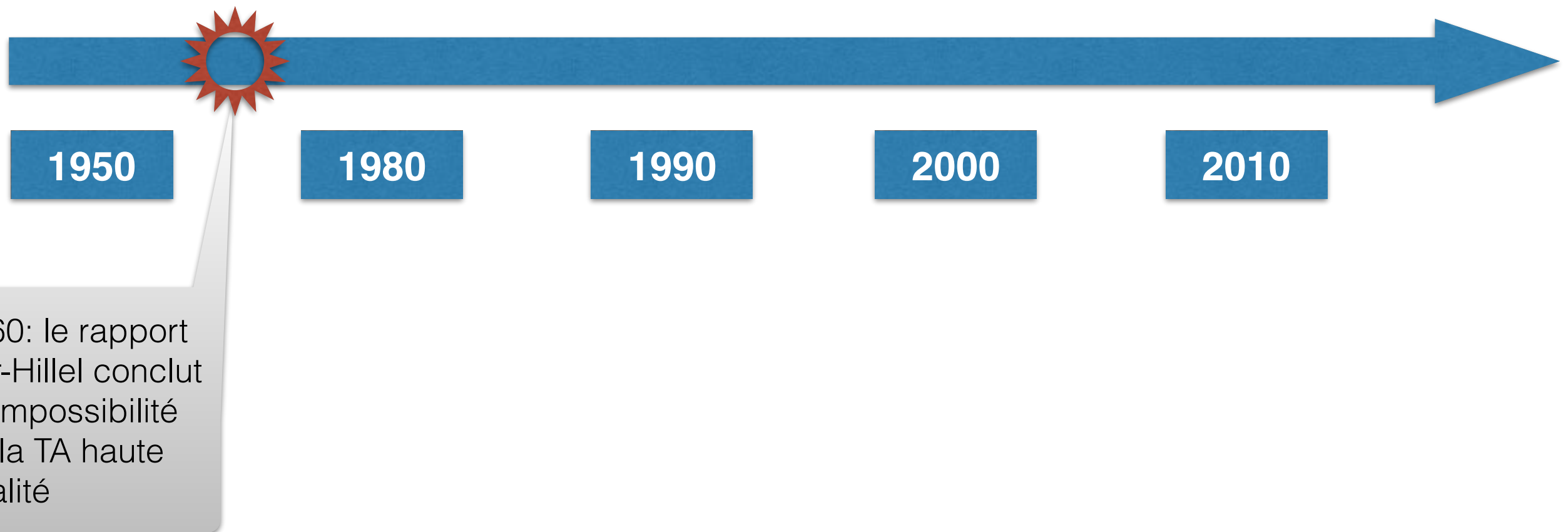
70 ans de TAL, 40 ans d'IL

Le TAL est
créé comme
une branche
de l'IA



70 ans de TAL, 40 ans d'IL

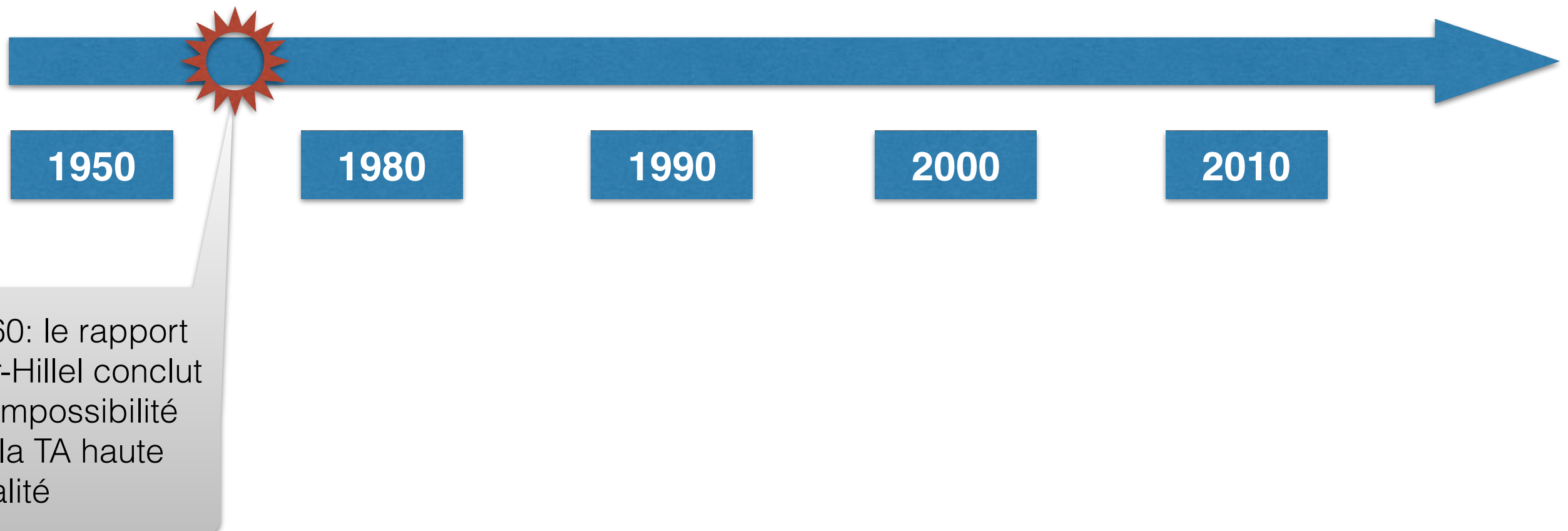
Le TAL est
créé comme
une branche
de l'IA



70 ans de TAL, 40 ans d'IL

Le TAL est
créé comme
une branche
de l'IA

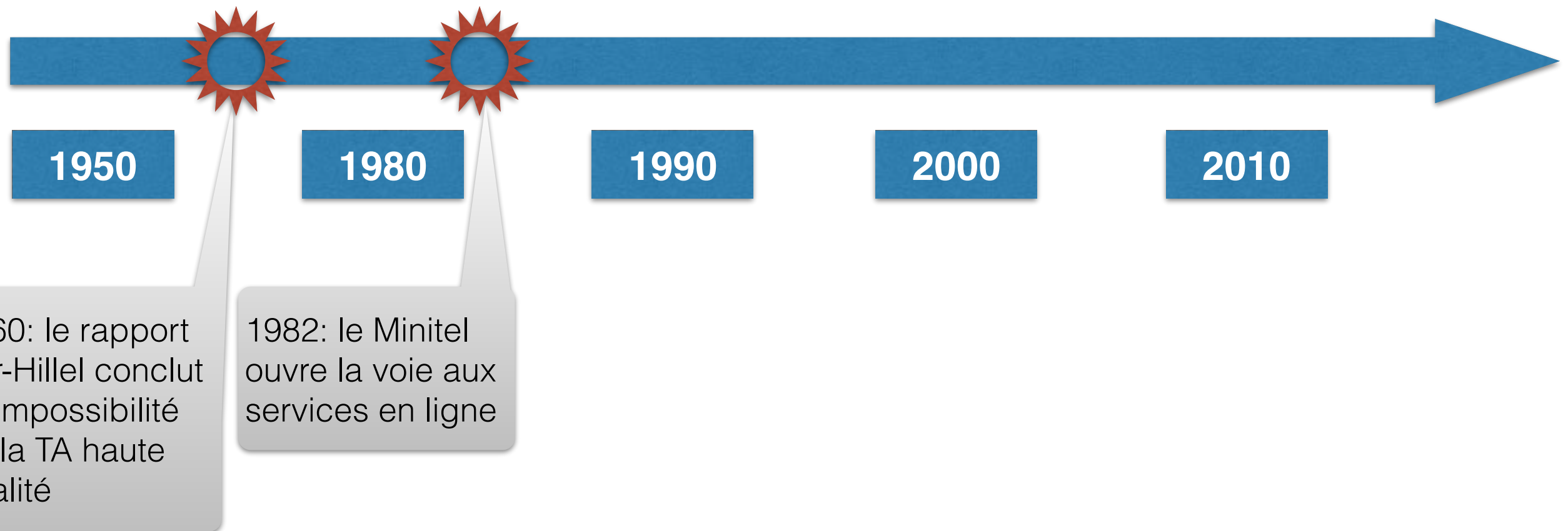
Naissance du
TAL industriel



70 ans de TAL, 40 ans d'IL

Le TAL est
créé comme
une branche
de l'IA

Naissance du
TAL industriel

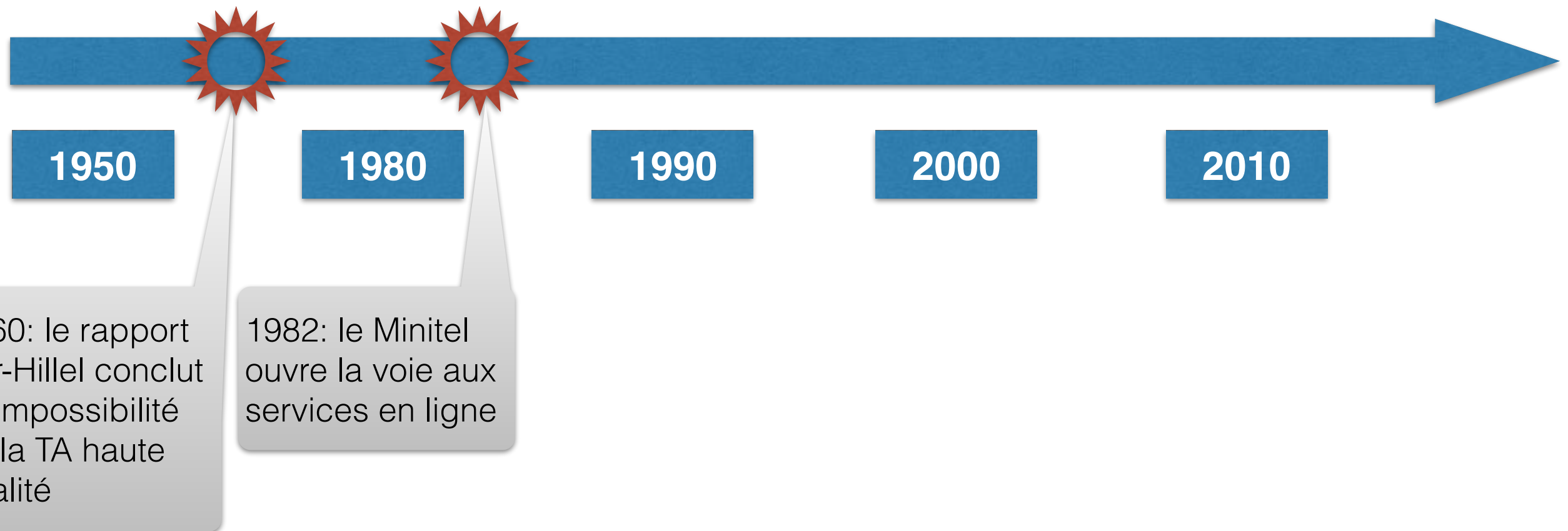


70 ans de TAL, 40 ans d'IL

Le TAL est
créé comme
une branche
de l'IA

Naissance du
TAL industriel

Explosion des
Industries de
la Langue

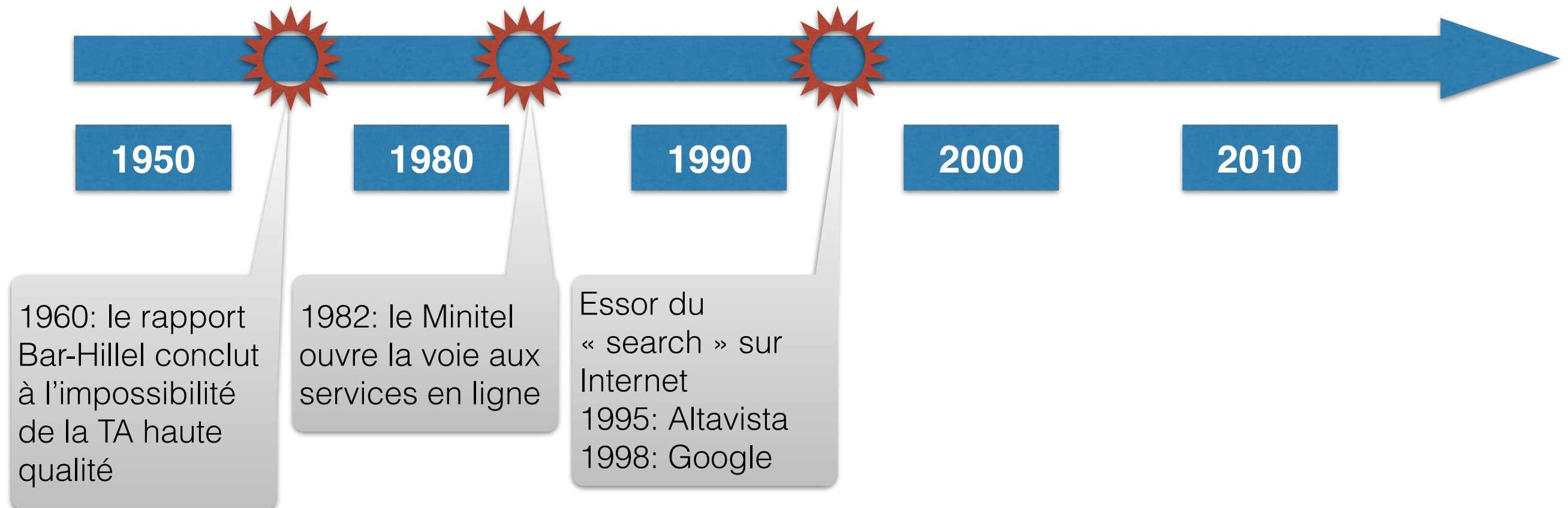


70 ans de TAL, 40 ans d'IL

Le TAL est
créé comme
une branche
de l'IA

Naissance du
TAL industriel

Explosion des
Industries de
la Langue



70 ans de TAL, 40 ans d'IL

Le TAL est
créé comme
une branche
de l'IA

Naissance du
TAL industriel

Explosion des
Industries de
la Langue

Le TAL
s'ouvre au
monde avec
Internet

1950

1980

1990

2000

2010

1960: le rapport
Bar-Hillel conclut
à l'impossibilité
de la TA haute
qualité

1982: le Minitel
ouvre la voie aux
services en ligne

Essor du
« search » sur
Internet
1995: Altavista
1998: Google

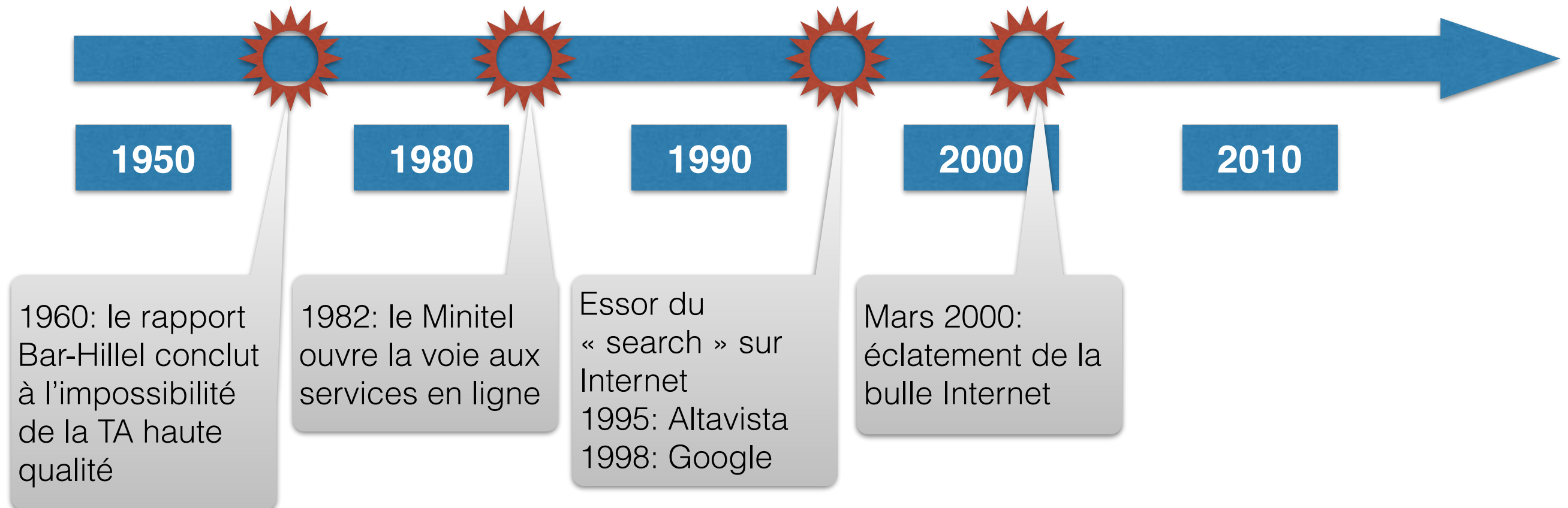
70 ans de TAL, 40 ans d'IL

Le TAL est
créé comme
une branche
de l'IA

Naissance du
TAL industriel

Explosion des
Industries de
la Langue

Le TAL
s'ouvre au
monde avec
Internet



70 ans de TAL, 40 ans d'IL

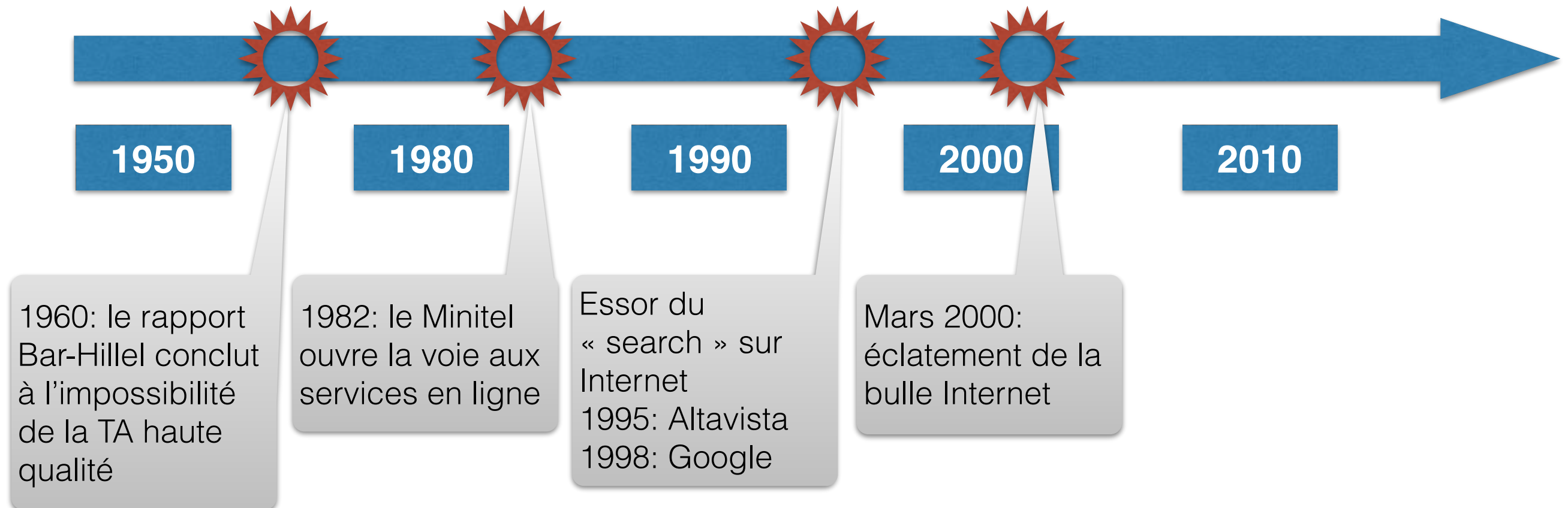
Le TAL est créé comme une branche de l'IA

Naissance du TAL industriel

Explosion des Industries de la Langue

Le TAL s'ouvre au monde avec Internet

Le NLP, un module de l'IA



70 ans de TAL, 40 ans d'IL

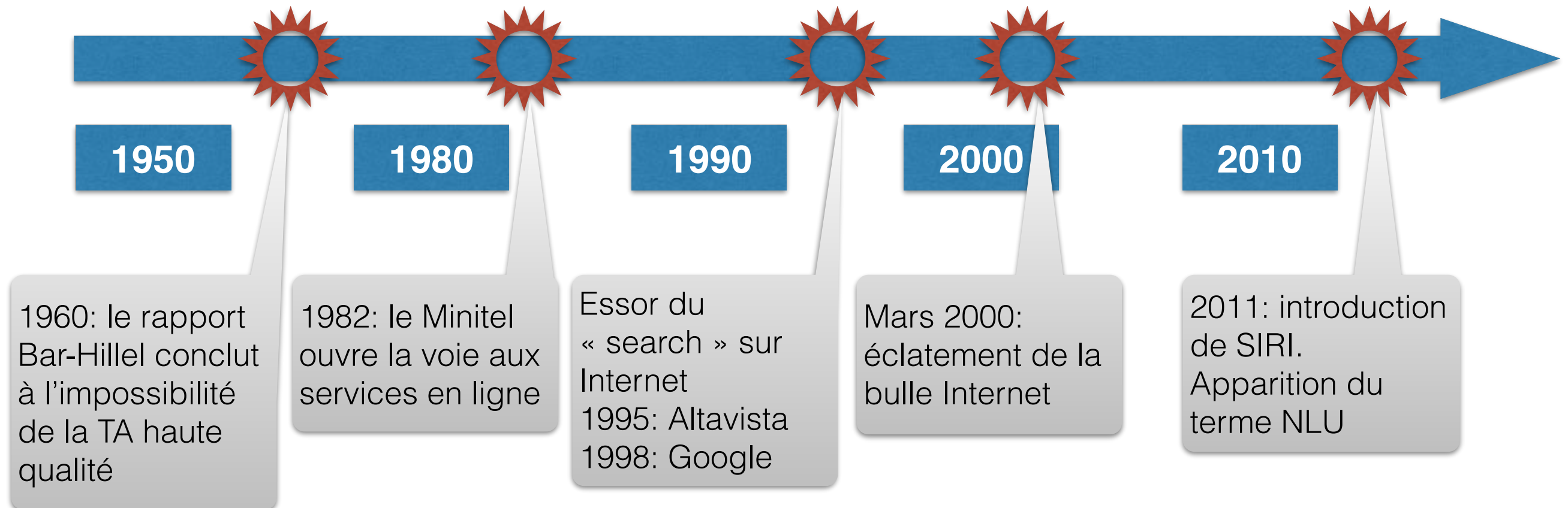
Le TAL est créé comme une branche de l'IA

Naissance du TAL industriel

Explosion des Industries de la Langue

Le TAL s'ouvre au monde avec Internet

Le NLP, un module de l'IA



Années 1980, naissance du TAL Industriel

- Le TAL « sort » de l'IA pour constituer une discipline nouvelle, la linguistique-informatique
- Quelques rares acteurs industriels proposent des solutions (en fait des projets outillés issus de la recherche universitaire)
 - Erli, Memodata, Vecsys, Synapse, CORA, Systran, ...
- Les grandes entreprises se dotent d'équipes de TAL et lancent de grands projets
 - Bull, Aérospatiale, France-Télécom, Dassault, Xerox, ...
- Les données sont rares et les technologies limitées
- La TA vit des heures difficiles : en 1985, Systran (créé en 1968 en Californie) est racheté par Gachot, industriel français spécialisé en robinetterie

D'un point de vue technique

- Des analyseurs linguistiques : phonétique, morphologie, syntaxe, sémantique mais pas de pragmatique
 - La phrase reste l'objet central des techniques
- Une analyse « hors contexte » qui omet sciemment le locuteur, le lieu, le moment....
- Des applications dans des domaines ciblés : correction orthographique, extraction de terminologie, construction de thesaurus, traduction automatique, accès à des nomenclatures
- L'objectif est un accès plus rapide/efficace au texte (textes traduits ou corrigés, résumés, thesaurus ...) et non des actions dans le monde

L'annuaire électronique sur Minitel (1992)



```
t 2 0 , 1 2 F p u i s 0 , 3 7 F / m i n
L E S P A G E S 3 6 1 1 ( c ) 2 0 0 0
J A U N E S f r a n c e t e l e c o m

N o m / M a r q u e : . . . . .
A c t i v i t é : . . . . .
L o c a l i t é : . . . . .
D é p a r t e m e n t : . . . . .
A d r e s s e : . . . . .
P r é n o m : . . . . .

V o u s p o u v e z p r é c i s e r

C o m m e n t r e m p l i r
I n f o s é d i t e u r , a u t r e s r e c h e r c h e s G u i d e
S o m m a i r e

E n r e c h e r c h e p a r a c t i v i t é , l ' o r d r e d e
p r é s e n t a t i o n d e s r é p o n s e s é v o l u e .
P o u r e n s a v o i r p l u s , t a p e z : * E n v o
```

Comment ça marchait

- Les rubriques de l'annuaire sont reliées à un couple <prédicat, objet>
 - il y a 13 prédicats, donc celui par défaut (PR01: acheter)
- Un réseau lexical groupant
 - des objets / des activités
 - des verbes / noms prédicatifs
 - ~30.000 mots dans le réseau, codés à la main
- Le réseau permet d'atteindre les objets et les prédicats à partir des mots entrés par l'utilisateur
- Séparation claire entre les objets linguistiques (le réseau lexical) et les objets métier (l'indexation des rubriques)
- Gros succès : passage de la taille moyenne des requête de 1 à un peu moins de 3 mots en mots en 3 ans

Un exemple

Restaurant de spécialités

Traiteur

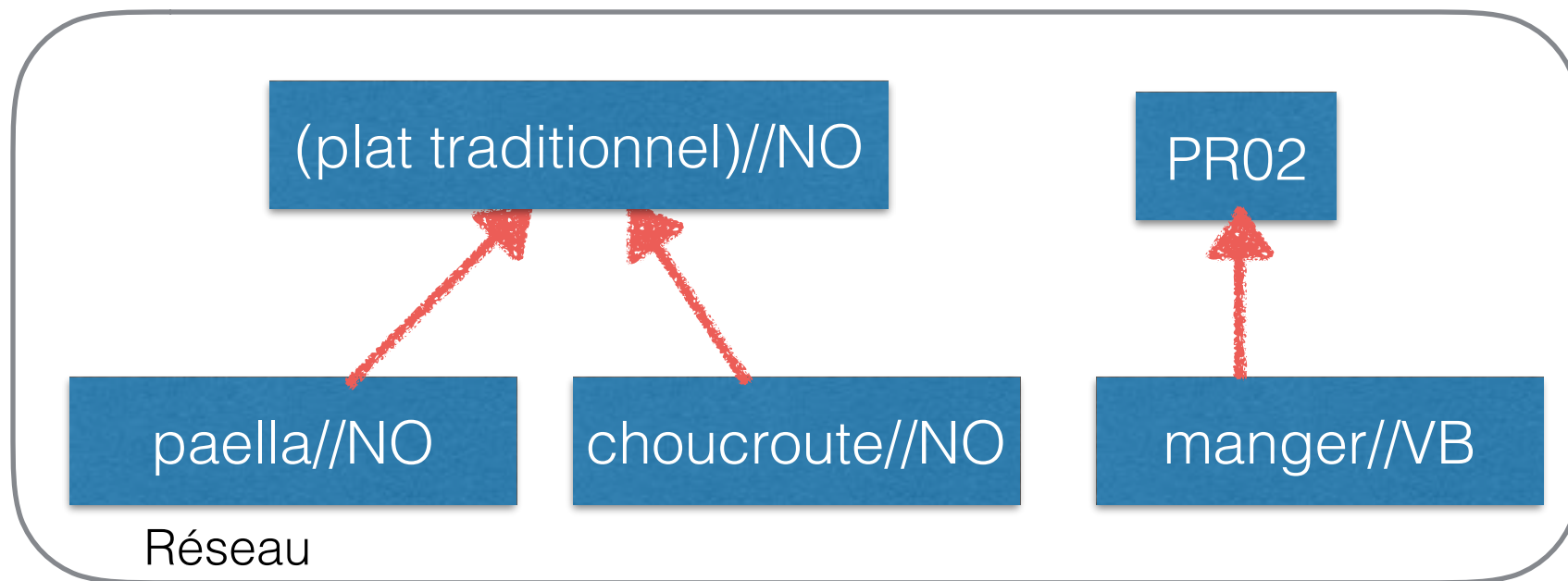
Taxi

Déménageur de piano

...

Rubriques de l'AE

Un exemple



Restaurant de spécialités

Traiteur

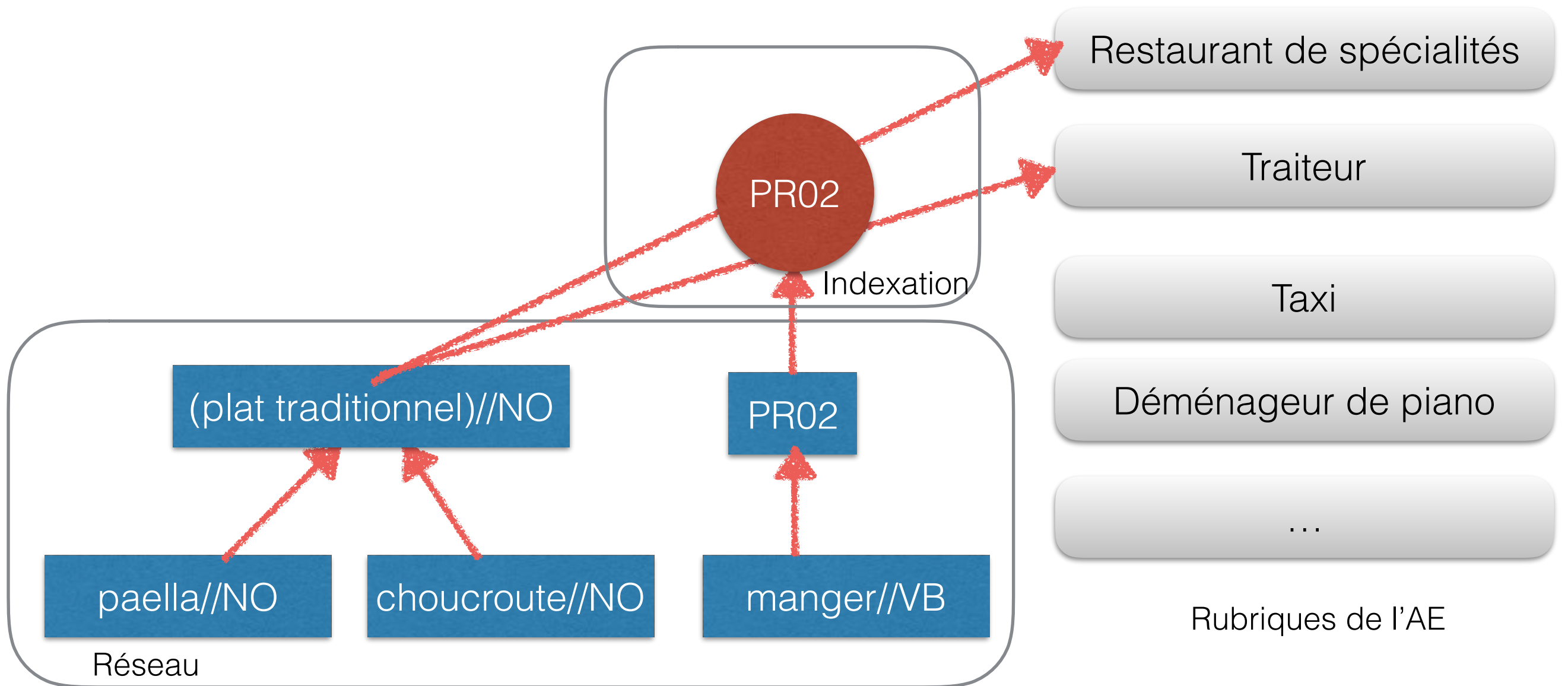
Taxi

Déménageur de piano

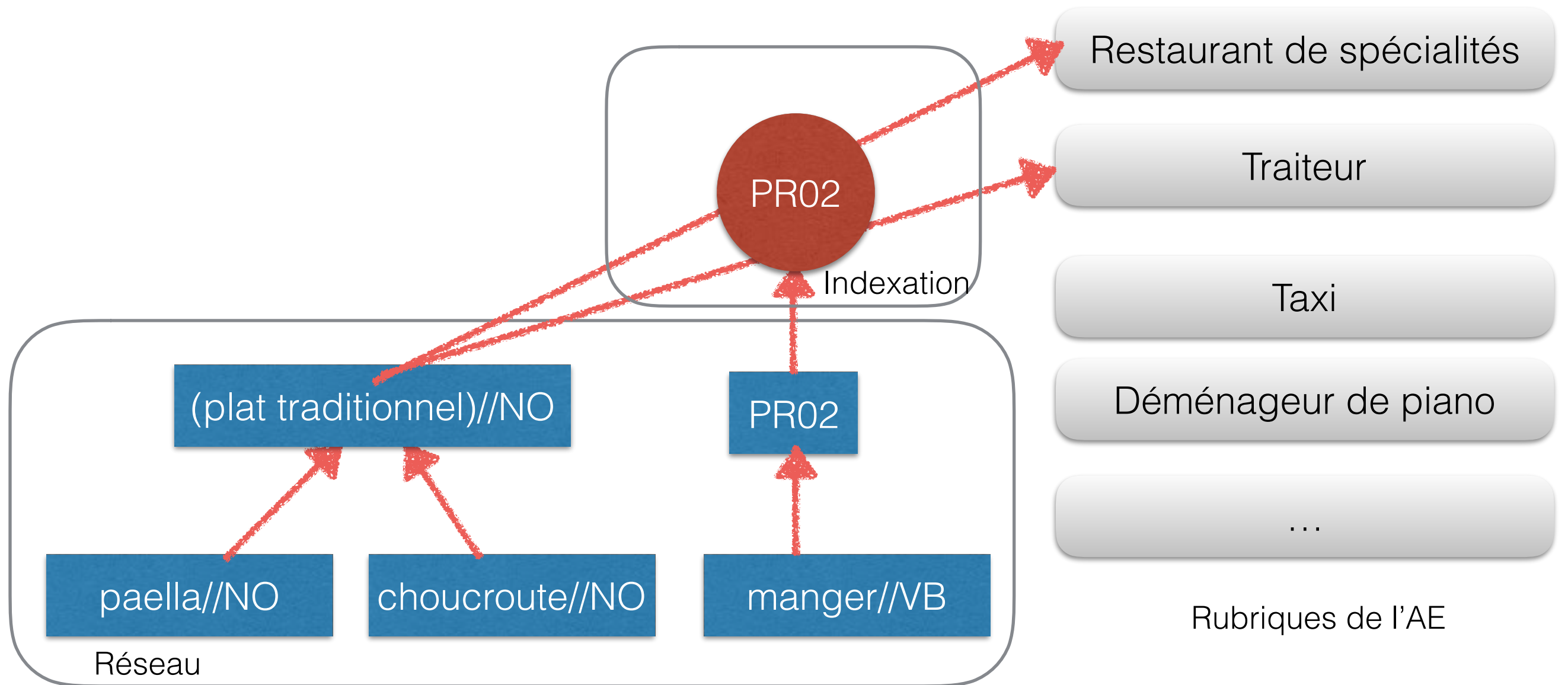
...

Rubriques de l'AE

Un exemple



Un exemple

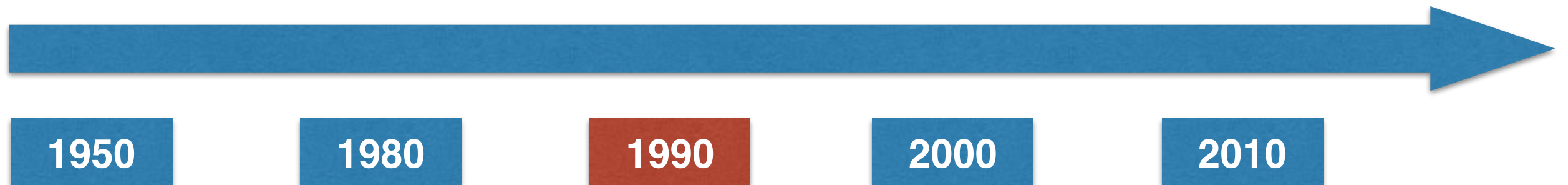


« ma femme veut manger de la choucroute »

Combien ça coûtait

- Une équipe de linguistes + informaticiens chez l'éditeur (Erli) pour maintenir le réseau et le moteur d'analyse (morphologie, syntaxe, appariement sémantique)
 - ~5 personnes, projet de 3 ans
 - = 15 années x homme
- Deux documentalistes à temps plein chez le client (France Télécom)
- L'intégration (télématique) était confiée à Cap Gemini

Explosion des Industries de la Langue



Explosion des Industries de la Langue



technolanguage 

Top » Directory » Répertoire des acteurs

Répertoire des acteurs-TEXT_TYPE_ENTREPRISE

Filtrage des résultats : Sélectionnez une typologie et/ou une lettre alphabétique.

TEXT_TOP ▾ alphab. ▾ OK

TEXT_ENTREPRISE

IBM France
Idylic
IES-ADICOR
Influo Software
Inforama International
INIST
Intelligence Process
Invention Machine Europe
Inxight Software Inc.
IRIS - France
Jériko
Language Weaver(Machine Translation)
LATRA
LCI
Learning Labs
Learning Labs(Learning Labs)
Libra Multimedia
Lingua Technologies(Québec)
Linguanet
Lingway

Afficher 41 à 60 sur 105 acteurs)

[<< Prev] Page 3 sur 6 [Next >>]

Informations

Presentation
Inscription
Contact Us

Directory

Private company
Research Organizations
Associations
Tools

Quick Find



Use keywords to find the product you are looking for.

Advanced Search

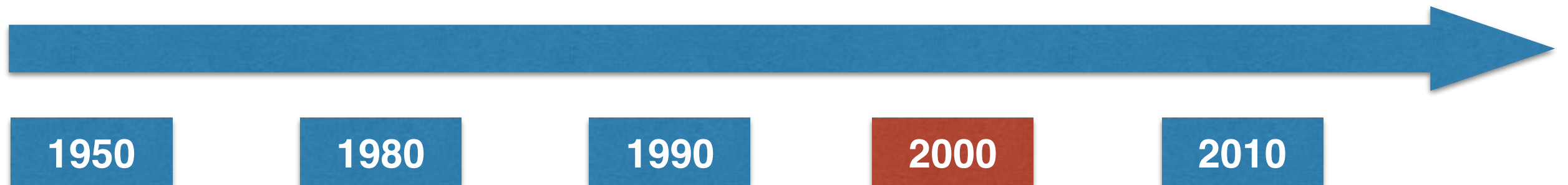
Les corpus deviennent disponibles

- En 1993, Le Monde sort son CD-ROM : une année d'articles disponibles de façon numérique
 - le plus grand corpus français contemporain jamais utilisé
- Ouvre la porte à :
 - l'apprentissage (technologies limitées)
 - l'évaluation (auparavant, les corpus de test ne dépassaient pas 100 phrases)
- Le défi naissant est le volume de données
 - D'où l'essor du search, mais aussi du text-mining
- La TA sort du marasme et se positionne comme « outil d'accès à l'information »

Point de vue technologique

- On cesse d'imiter l'humain
 - On s'éloigne de Chomsky (description analytique de la langue) pour des approches « corpus based »
- Les lexiques numériques commencent à apparaître en version ouverte
 - 1995 : création de WordNet, au départ uniquement un outil de psycho-linguistique
- et également des agences de diffusion des corpus et lexiques construits lors des projets subventionnés
 - aux US: LDC (Linguistic Data Consortium) sous l'instigation de la DARPA (1995)
 - en Europe, ELRA/ELDA sous l'instigation de l'UE (1995)

Internet et le TAL



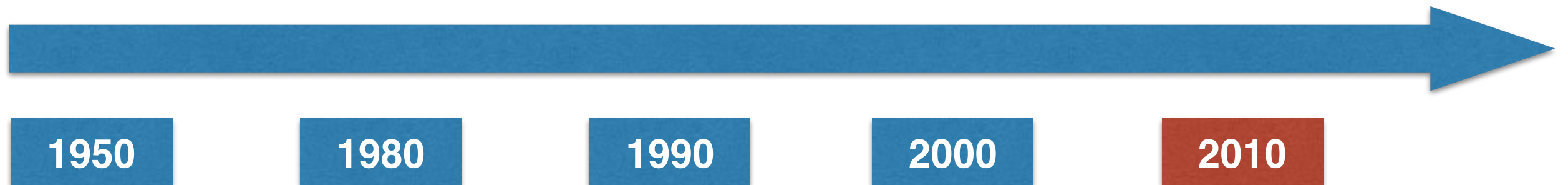
Explosion

- La tendance se confirme : de plus en plus de volumes de textes accessibles directement
 - Renforcement du search qui devient sémantique
- Mais l'explosion de la bulle Internet touche directement le TAL
 - les grandes entreprises ferment leurs laboratoires ou les revendent (IBM, Xerox, ...)
 - les petits acteurs dopés par le capital-risque se voient sommés de devenir rentables...
 - D'où un important renouvellement du paysage industriel
 - Avec quelques effets de bord bénéfiques :
 - Création de Lucene en 2001
 - Passage de UIMA d'IBM à la Fondation Apache en 2005

Point de vue technologique

- Au-delà du search, une tendance se fait jour avec l'évolution du nombre des avis consommateurs sur Internet, vers le début des années 2010 : l'analyse de sentiment
 - e-réputation
 - Voix du client (« voc »)
- Le TAL s'adresse désormais à de nombreux métiers
 - Marketing/support client
 - Recrutement (matching CV/offres)
 - Sécurité/défense

Retrouvailles entre TAL et IA



NLU, big-data

- NLU = Natural Language Understanding
- On se sert du langage, non plus comme outil pour accéder au langage (search, traduction, etc.) mais pour agir sur le monde
 - explosion des assistants personnels,
 - puis des chatbots : le langage est un élément de l'interface homme machine
- En même temps, le volume des données à traiter change de magnitude, on est à l'ère du big data, géré par les Data Scientists
- Il n'y a plus de start-ups spécialisées en TAL mais des start-ups spécialisées en robotique, vision, marketing, etc. qui utilisent des technologies de TAL (plus ou moins bien)

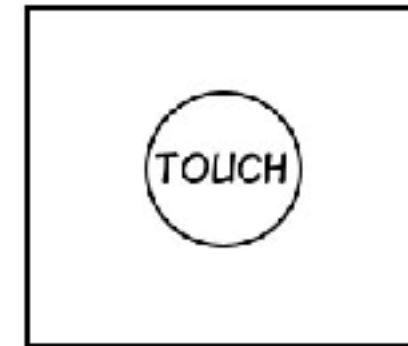
Point de vue technologique

- Le deep learning appliqué au TAL produit des résultats étonnants
 - Traduction, interprétation, compréhension (pas toujours...)
- Les techniques ne sont pas nouvelles mais la masse de données et la puissance de calcul permettent de les faire fonctionner efficacement
 - le premier réseau de neurones date de 1957 (perceptron)
 - le système SHRDLU (1970) permettait d'interagir en langage naturel avec un univers composé de cubes et de boules
 - les architectures de dialogue homme-machine des années 1980-1990 (cf. formalisation AIML en 1995)
- Multiplication des plateformes de TAL, utilisables par des non-spécialistes (cf. plateformes de chatbot)

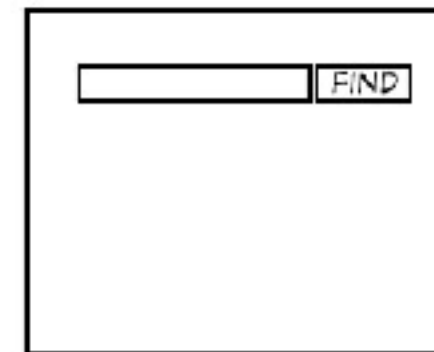
Le danger

- Le NLU devient une « facility », mise en oeuvre grâce à des API banalisées
- Il est mis en oeuvre par des développeurs qui n'ont pas de notions en étude de la langue
- Multiplication des applications de piètre qualité
- Cercle vicieux de l'appauvrissement du composant

TYPICAL APPLE PRODUCT...



A GOOGLE PRODUCT...



YOUR COMPANY'S APP...

FIRST NAME:	<input type="text"/>	TYPE CD:	<input type="text"/>	4 - K AA2- DK9B KKAP CNB AA-9
LAST NAME:	<input type="text"/>	TQP STAT:	<input type="checkbox"/>	
SSN:	<input type="text"/>	VER:	<input type="text"/>	
ID:	<input type="text"/>	FT/PT:	<input checked="" type="checkbox"/>	
PHONE 1:	<input type="text"/>	CAT CD:	<input type="text"/>	
PHONE 2:	<input type="text"/>	CITY:	<input type="text"/>	
ADDR 1:	<input type="text"/>	STATE:	<input type="text"/>	
ACCT #:	<input type="text"/>	ZIP:	<input type="text"/>	
		ORD #:	<input type="text"/>	
OKAY APPLY SAVE UNDO HELP DELETE EDIT				
SELECT BROWSE ERRORS				

STUFFTHATHAPPENS.COM BY ERIC BURKE

Petit retour en arrière

- Le Crédit Mutuel va sortir un chatbot basé sur Watson (IBM) pour ses clients qui a « *mobilisé 10 à 15 personnes côté IT et autant issues des métiers [...] pendant 10 mois* » (pas toutes à temps plein) [source silicon.fr]
 - soit 10 à 20 a.h pour le Crédit Mutuel (on ne connaît pas le coût de la licence Watson)
 - à comparer à l'annuaire électronique des années 1990 : 6 a.h chez le client France Télécom (et 15 a.h chez l'éditeur)
- Conclusion : on n'a pas beaucoup progressé en 30 ans en termes de productivité
 - Mais maintenant, c'est le client qui travaille, pas l'éditeur
 - ... ce qui implique aussi un déplacement de la responsabilité en cas de problèmes sur le projet
 - C'est pourquoi des spécialistes en TAL sont particulièrement utiles (à condition de savoir se vendre)

Et maintenant ?



quoi

où

nlp

Rechercher

métier, mots-clés ou entreprise



[Publiez votre CV](#) - Postulez à plus de 60 000 emplois depuis n'importe quel appareil

Emplois 1 à 1

Afficher : **tous les emplois** - [9 nouveaux emplois](#)

Lead développeur / PHP (H/F) - nouveau

Leankr - La Défense (92)

40 000 € - 55 000 € par an

Des connaissances en NLP, text mining ou machine learning sont un plus pour ce poste. Leankr est une start-up innovante dans le secteur de la vidéo sur Internet...

[Postuler directement](#)

Sponsorisée - il y a 2 jours - [sauvegarder](#)

Natural Language Processing (NLP) Engineer - France - nouveau

IC Resources - Paris (75)

Natural Language Processing (NLP) or nearby fields such as AI, pattern recognition, machine learning etc....

Sponsorisée par **Technojobs** - il y a 1 jour - [sauvegarder](#)

Python full stack developer at Chefclub (H/F)

ChefClub - Paris (75)

Chefclub is looking for a talented full stack developer who love working with Python. You will be working, not exclusively, on our food assistant chatbot

Sponsorisée par **RemixJobs** - [sauvegarder](#)

Ingénieur Data Science F/H

MERITIS - Paris 1er (75)

NLP. Vous intégrerez un Data & Intelligence Artificielle Lab, créé il y a un peu moins de 2 ans, et travaillant sur une multitude de sujets :....

Apec.fr - il y a 25 jours - [sauvegarder](#) - [plus...](#)

Data scientist NLP secteur RH

MP DATA - Paris (75)

Des connaissances en text-mining, NLP seront fortement appréciées. MP DATA est une société de Conseil tournée vers l'expertise qui accompagne les entreprises...

[Postuler directement](#)

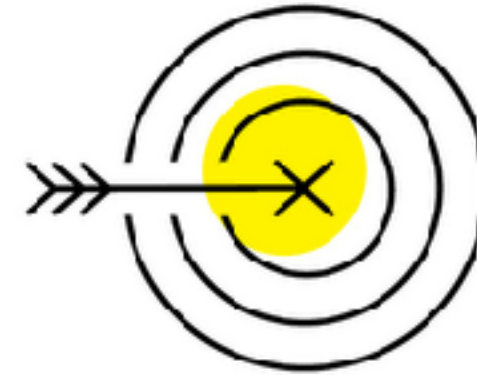
Perspectives pour le TAL

- De l'avis général, les interfaces conversationnelles tiennent la corde pour plusieurs années à venir
- Gartner:
 - « *les hommes parleront plus à leur assistant personnel qu'à leur femme en 2020* »
 - « *85% des relations entreprises-clients n'impliqueront aucune interaction humaine d'ici 2020* »
- Le e-commerce en est particulièrement friand, mais ses besoins dépassent le conversationnel :
 - analyse des retours client (« voice of the customer »),
 - personnalisation de l'expérience client, avec peut-être un retour en force de la génération automatique de textes
- Tout cela poussé par les avancées en deep-learning qui a encore beaucoup à progresser

Persp

- De l'avis général des années à venir
- Gartner:
 - « les hommes »
 - « 85% des »
 - « d'ici 2020 »
- Le e-commerce conversationnel
 - analyse de
 - personnalisation
 - génération
- Tout cela pour progresser

Top 10 Strategic Technology Trends for 2018



Intelligent



AI Foundations



Intelligent Apps and Analytics



Intelligent Things



Digital



Digital Twins



Cloud to the Edge



Conversational Platform



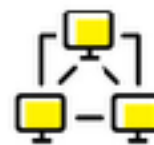
Immersive Experience



Mesh



Blockchain



Event-Driven



Continuous Adaptive Risk and Trust

gartner.com/SmarterWithGartner

Source: Gartner
© 2017 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. or its affiliates. FR_312654

Gartner

our plusieurs

omme en 2020 »

raction humaine

issent le

n force de la

aucoup à

Top 5 des besoins les plus critiques de la distribution et de l'e-commerce – France, 2017-2020

Besoins les plus critiques en logiciels et solutions applicatives des décideurs de la distribution et de l'e-commerce pour soutenir leurs stratégies clients digitales

Icons made by Vectors Market from www.flaticon.com



Echantillon : 30 décideurs de la distribution et de l'e-commerce interrogés entre février et avril 2017

Ce qui a changé en 25 ans

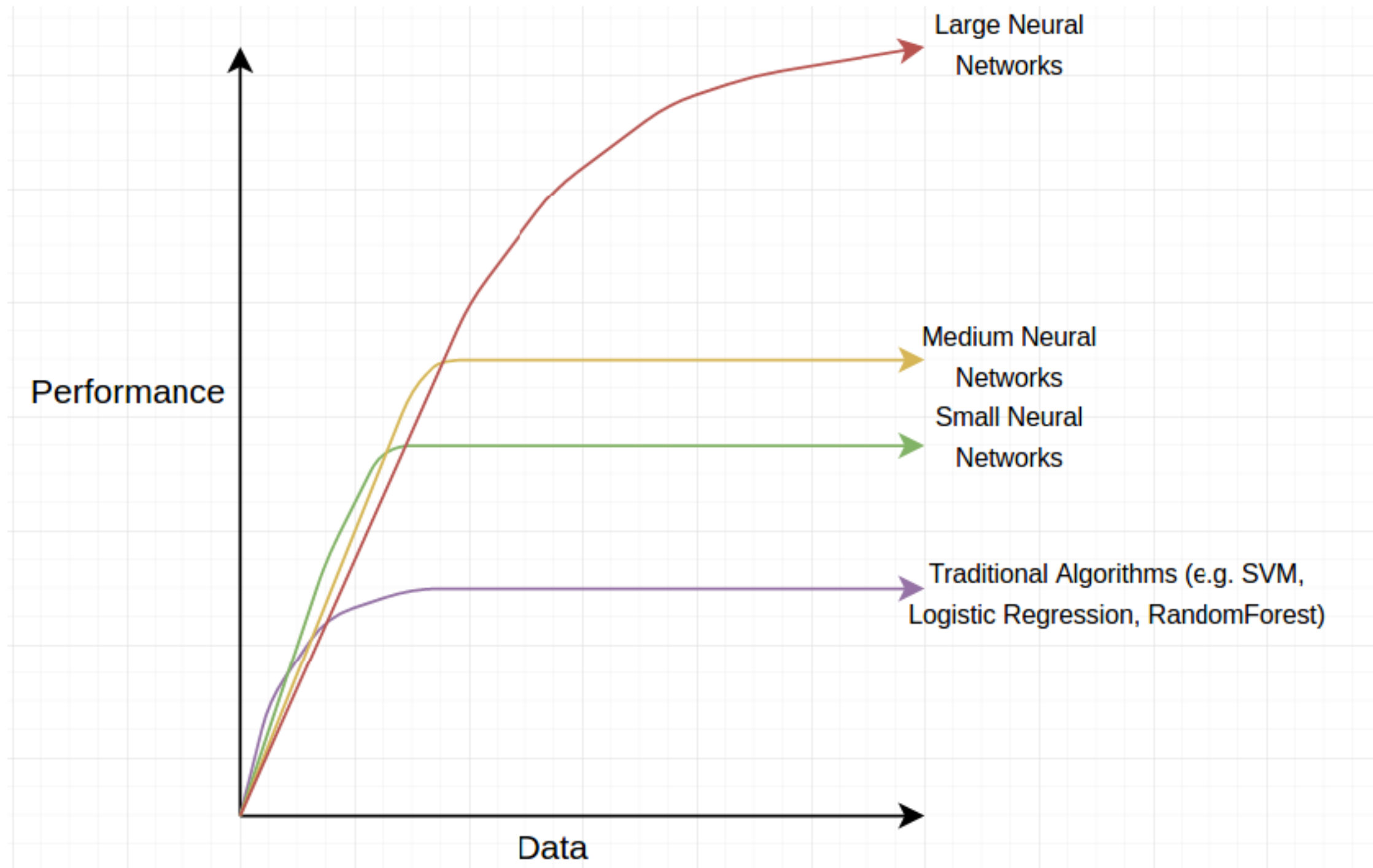
- La disponibilité de grandes masses de données textuelles
 - a permis l'essor de l'apprentissage automatique qui a largement pris le pas sur les approches symboliques
 - mais la linguistique formelle a encore des choses à apporter
- Le rôle central de l'évaluation dans la recherche
 - a facilité la mise en œuvre en industrie des systèmes réalisés par les chercheurs
 - même si des biais sont apparus par là-même : mesures de tout et n'importe quoi, course au gain sur des « gold-standards » définis hors sol
- La disponibilité d'outils (fréquemment en accès libre)
 - outils de haut niveau, comme les moteurs de recherche (ElasticSearch, SolR, Sphinx)
 - frameworks de développement linguistique (nltk, OpenNlp, UIMA ...)
 - services (API) linguistiques

Ce qui n'a pas changé

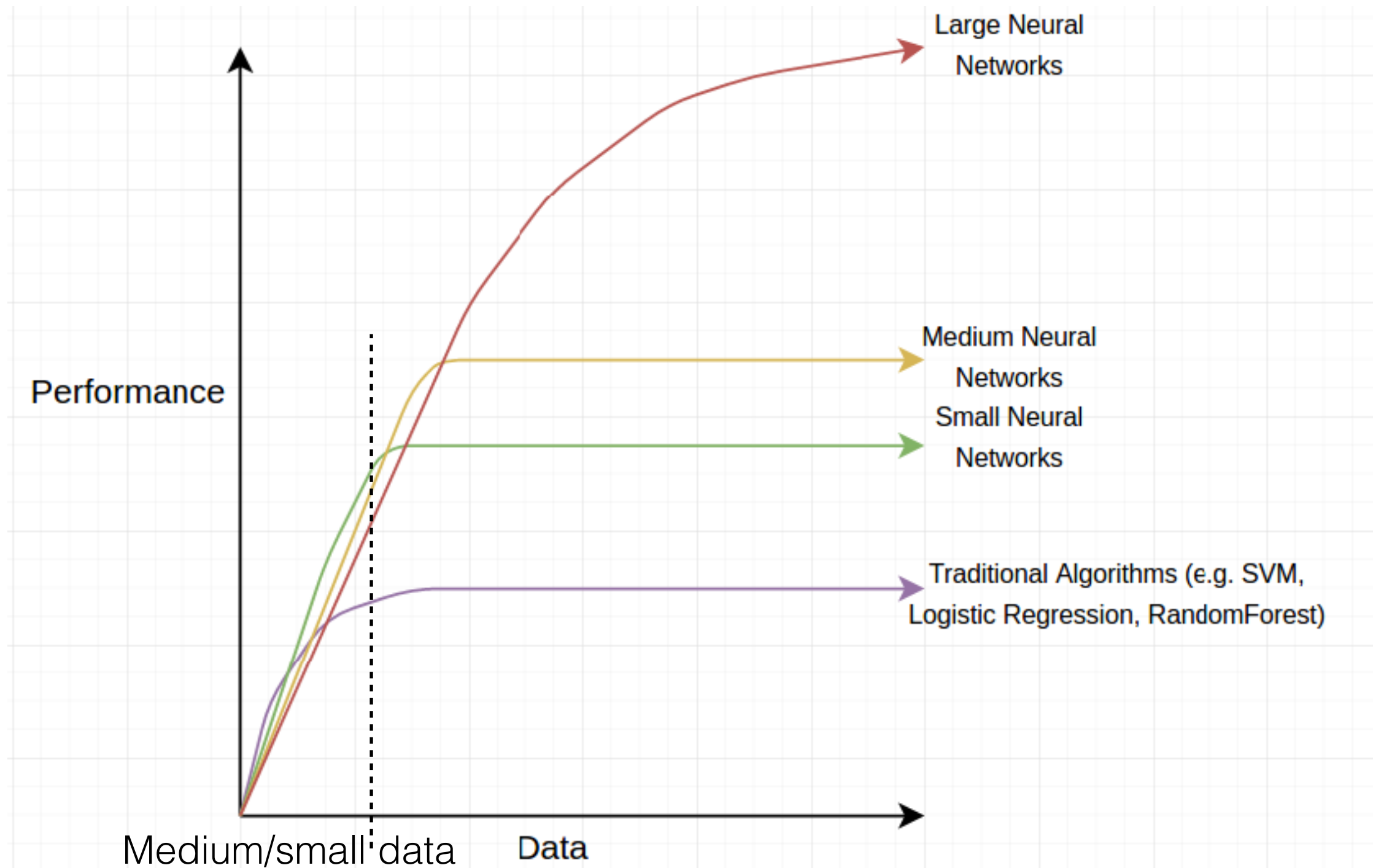
- Les informaticiens continuent à faire du TAL sans le savoir
 - cf. les listes de « stopwords », les regex, etc.
 - et de plus en plus avec les approches de type neuronal
 - mais toujours considérées comme des boîtes noires
- La langue est toujours un domaine complexe pour l'utilisation par une machine
- Et donc les systèmes restent limités, au grand dam des utilisateurs
 - ... qui croient toujours parler à HAL-9000 quand ils utilisent un système qui met en œuvre du TAL

TAL et deep-learning

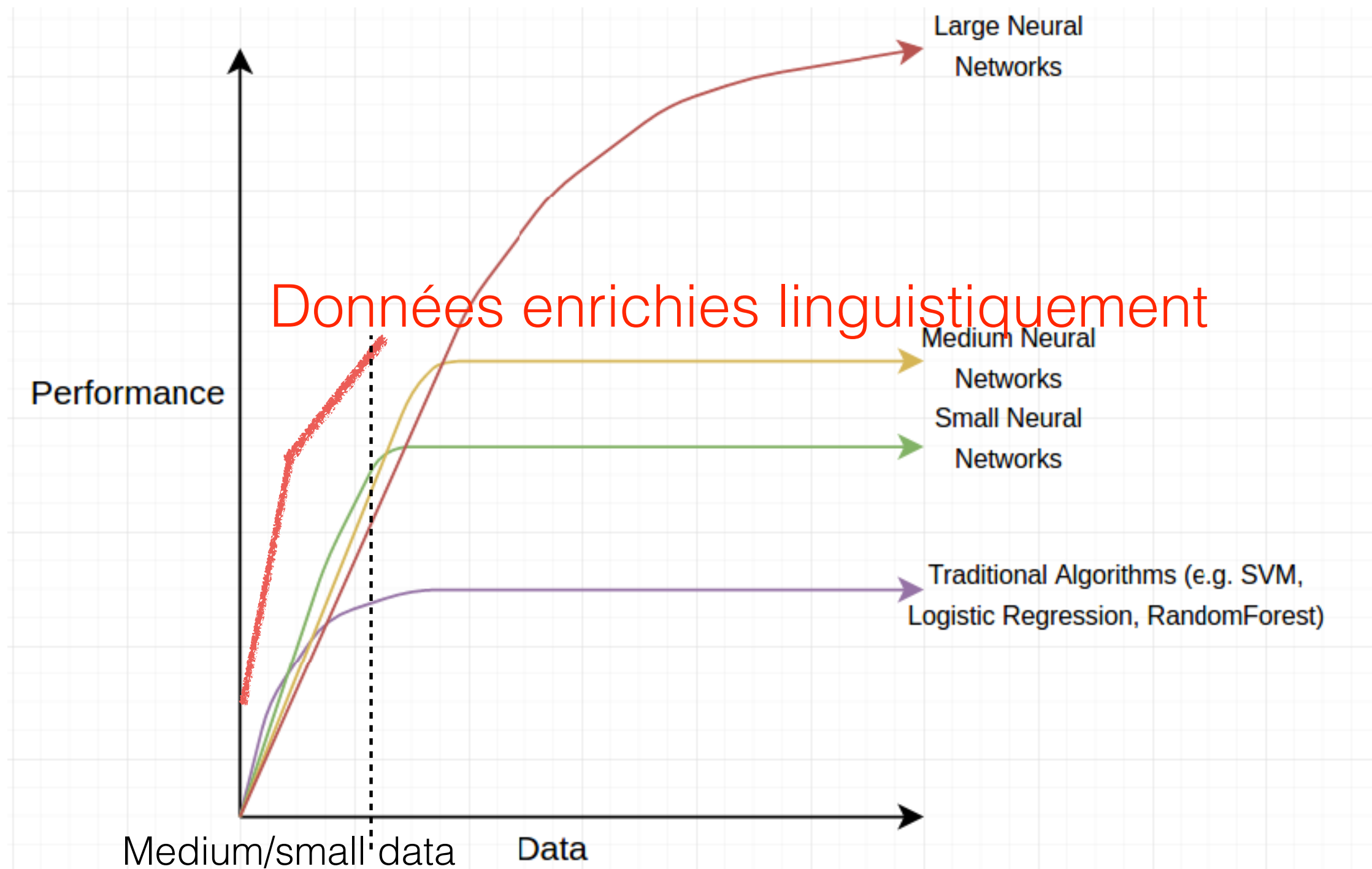
- Le deep-learning fonctionne bien quand il y a beaucoup de données d'entraînement
- Plus les données d'entraînement sont volumineuses, moins il est nécessaire de travailler sur ces données, le volume compensant la qualité
 - on peut se contenter de représentations en sacs de mots
- A contrario, s'il n'y a pas beaucoup de données d'apprentissage, la qualité de l'information donnée en entrée est cruciale
 - Or, le plus souvent, en contexte industriel, les données d'entraînement sont en faible volume
 - ... et leur constitution coûte cher



Qualité des algorithmes de deep-learning
(source: Andrew Ng)



Qualité des algorithmes de deep-learning
(source: Andrew Ng)



Qualité des algorithmes de deep-learning
(source: Andrew Ng)

Les données sont rarement au rendez-vous

- Sondages Internet : quelques centaines de verbatim
- Parfois des logs, mais les données ne sont pas qualifiées (donc pas d'apprentissage supervisé)
- De plus en plus de données publiques (open-data) mais qui ne correspondent pas vraiment au cadre applicatif visé
 - cf. offres Pôle-Emploi, classifiées dans la nomenclature ROME... que les recruteurs n'utilisent pas

Quelques exemples d'enrichissement

- Lemmatiser les entrées diminue le « vocabulaire » du réseau de neurones et donc diminue la taille du corpus nécessaire
 - cf. « Analysis of Morphology in Topic Modeling », (Chandler et al, 2016) : étude du russe sur une tâche de LDA.
 - Avec lemmatisation : 65% de taux de reconnaissance (vs. 50% sans)
- Analyse en morphèmes de l'allemand : +20% en qualité sur du clustering
 - « Topic Modelling with Morphologically Analyzed Vocabularies », M. Spies 2017
- Reconnaître les mots composés pour le français
 - les prépositions/conjonctions pour pouvoir les éliminer (« pour le compte de », « au vu de », ...)
 - les mots composés figés non compositionnels (« pomme de terre », ...)

Ne pas partir de zéro

- Analyse de sentiments : lexique de mots marqués en sentiment (cf. SentiWordnet, jeuxdemots, etc.)
 - complété par une analyse syntaxique
 - => prise en compte de la négation
 - => le dépendant du mot marqué (positif/négatif) est l'objet du sentiment (l'une des détections les plus complexes pour ce type d'analyse)
- NLU : un réseau sémantique + quelques règles
 - cf. chatscript (basé sur WordNet)
- etc.

Pourquoi l'hybridation est-elle nécessaire ?

- Parce qu'on a depuis longtemps les ressources (validées) qui permettent de trouver les variantes morphologiques des mots
 - idem pour les capitales des pays ou les dérivations des noms, etc.
 - il serait criminel d'ignorer des décennies de TAL face au mirage de « la machine qui apprend d'elle-même »
- Parce que les systèmes d'apprentissage sont d'autant meilleurs que leur entrée est riche
- Enfin parce que les systèmes basés sur l'apprentissage font peu de cas des phénomènes linguistiques qui ont occupé les TAListes pendant des générations : polysémie, composés terminologiques, etc.

Conclusion : les compétences pour être recruté en TAL

- Connaissance des outils de TAL et d'apprentissage, pratique de l'étude de corpus
 - fréquemment un rôle de prescripteur d'outils TAL à dimension industrielle (attention aux licences, aux performances, etc.)
- Maîtrise d'un langage de programmation
 - vous serez intégrés à une équipe de développement dont vous devrez apprendre les règles (agilité, intégration continue, etc.) sans partir de zéro
- Etre un « data-scientist » : notions de statistiques, des outils utilisés en « big data » (hadoop, spark, ...) et de préparation des données