



Un projet de recherche collaboratif entre

Historiens, économistes, démographes

Archivistes

Experts en intelligence artificielle



Et les services d'archives départementales et municipales

Bénéficie d'un financement sur 3 ans ½



SocFace:

The local face of social change: one century of French social structure seen from the ground, 1836–1936

Collecter, traiter, retranscrire, organiser et analyser l'ensemble des listes nominatives du recensement de 1836 à 1936 (20 recensements).

SocFace produira une base de données complète des individus ayant vécu en France entre 1836 et 1936 et l'utilisera pour analyser le changement social dans la longue durée.



De multiples objectifs et enjeux



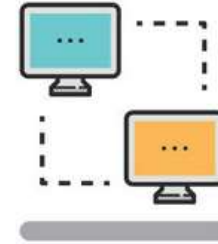
Science

Etudes en économie, en histoire et en démographie : évolution du marché du travail, des inégalités, de la structure sociale, des mobilités.



Technologie

Reconnaissance automatique d'écriture manuscrite, analyse de tableau, traitement de plusieurs million d'images, accès à de multiples sources.



Valorisation

Mise à disposition des données extraites en accès libre, versement aux propriétaires des fonds, valorisation par les portails.

Pourquoi Socface?

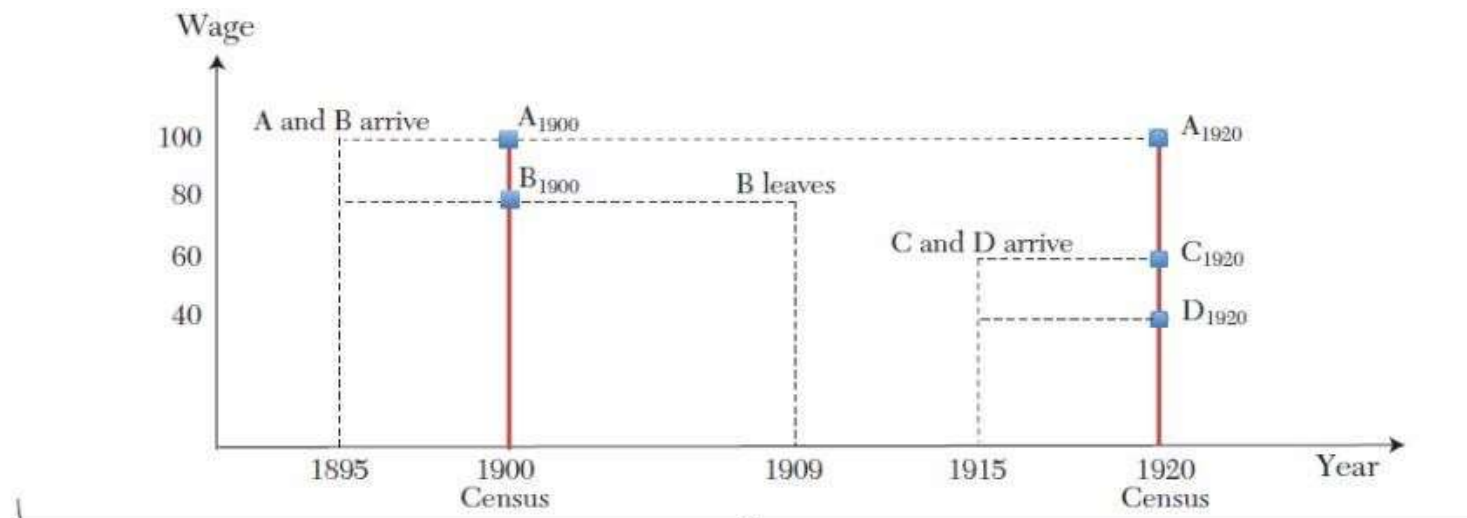
- Importance croissante des données micro (individuelles) dans la recherche quantitative en sciences sociales: économie, histoire, sociologie, démographie...
- Progrès considérables dans les technologies de reconnaissance d'écriture manuscrite et de traitement d'images.

**Produire des données
individuelles en masse à
l'échelle nationale**

**Développer les technologies de
traitement de vaste ensemble
de sources historiques**

**Diffuser les informations
obtenues au grand public
comme aux chercheurs**

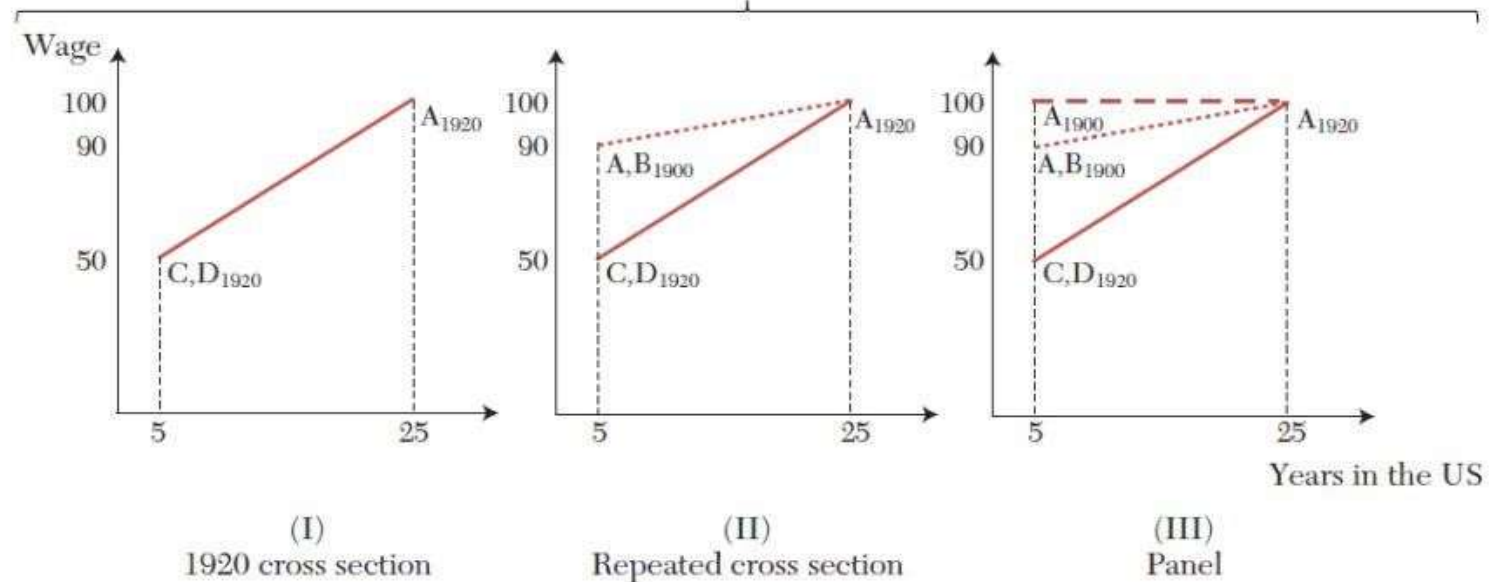
Un exemple d'apport
des données
individuelles
longitudinales:



Data

Estimated assimilation

Situation des
immigrés sur le
marché du travail
américain



Pourquoi Socface?

- Importance croissante des données micro (individuelles) dans la recherche quantitative en sciences sociales: économie, histoire, sociologie, démographie...
- Progrès considérables dans les technologies de reconnaissance d'écriture manuscrite et de traitement d'images.

**Produire des données
individuelles en masse à
l'échelle nationale**

**Développer les technologies de
traitement de vaste ensemble
de sources historiques**

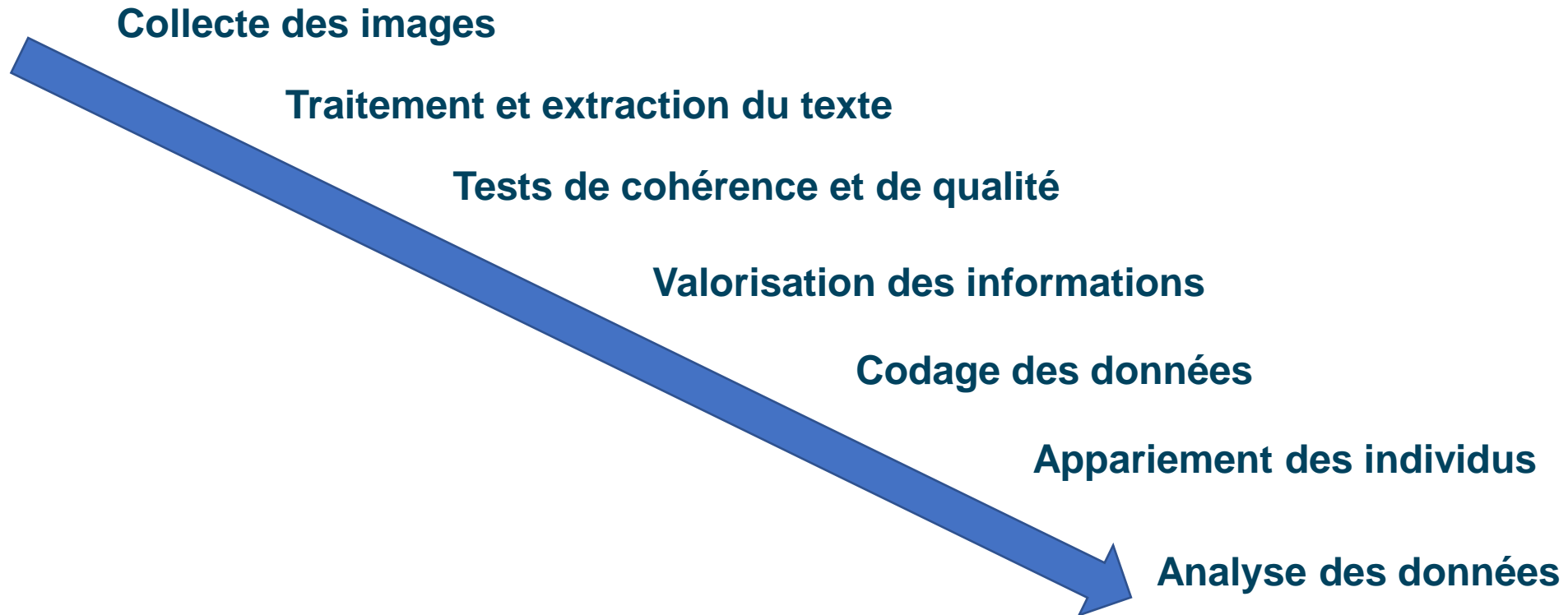
**Diffuser les informations
obtenues au grand public
comme aux chercheurs**

Pourquoi les listes nominatives?

- Une source abondante, simple, stable et standardisée.
 - Une source régulière dans le temps.
 - Déjà photographiée par beaucoup de services d'archives.
 - Source nationale.
 - Permet de construire un échantillon de toute la France (ou presque...).
- Une source idéale pour le passage à l'échelle de la reconnaissance de texte.
 - Une source qui n'a de sens que étudiée à grande échelle.



Etapes et objectifs



Etude pilote: l'ensemble des étapes sur une sélection d'archives – Objectif: un an

Puis montée en charge progressive à partir d'octobre 2022

Incorporation des services d'archives peu à peu et retours au fur et à mesure

Spécificités du projet Socface

➤ Traitement automatisé de vastes ensembles d'images

- ❖ Plusieurs millions d'images, de provenances variées (différents éditeurs, formats, etc.).
- ❖ Ouvre des perspectives considérables pour la recherche en intelligence artificielle.

➤ Une coproduction entre historiens et informaticiens

- ❖ Utiliser la structure des listes pour contrôler l'information (ordre des individus, ménages, etc.).
- ❖ Utiliser les informations agrégées (externes ou non) pour contrôler le traitement.
- ❖ Codage des informations obtenues pour les rendre accessibles aux chercheurs (lieux, professions...).

➤ Appariement des individus pour les suivre au cours du temps

- ❖ Première réalisation à l'échelle de la France entière, sur 100 ans.
- ❖ Ouvre des perspectives considérables pour la recherche en sciences sociales.

Participer à Socface

➤ Un projet en cours d'élaboration

- ❖ Importance de la phase pilote (un an).
- ❖ Des contours à préciser.

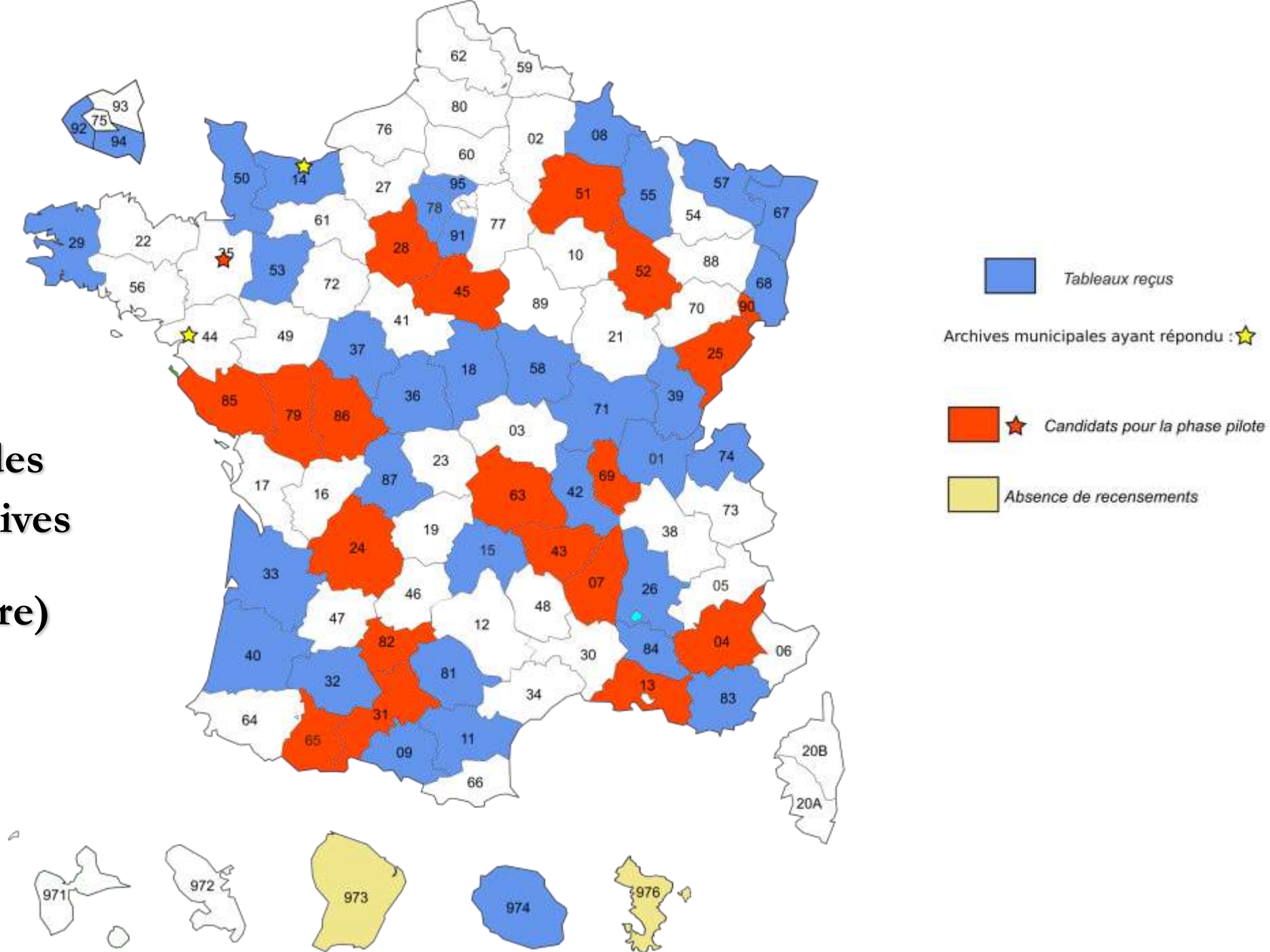
➤ Le rôle des services d'archives (ADs & AMs)

- ❖ Autorisations d'accès aux images.
- ❖ Pas de travail supplémentaire: pas de renumérisation ou de traitement des images à faire.

➤ Les retombées du projet pour les services d'archives

- ❖ Reversement des données extraites des listes nominatives.

L'implication des services d'archives (au 20 septembre)



SOCFACE : Quels défis pour la reconnaissance automatique d'écriture ?

Séminaire annuel des archives de France

Christopher Kermorvant

23 septembre 2021

T E K L I A

Reconnaissance d'écriture: un des plus vieux défis de l'IA



RAND corporation, 1960



The MNIST database

“The drosophila of machine learning”

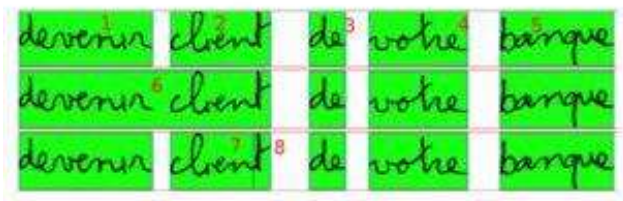
Geoffrey Hinton

Mais les machines ne rivalisent toujours pas avec l'humain pour la lecture de documents manuscrits

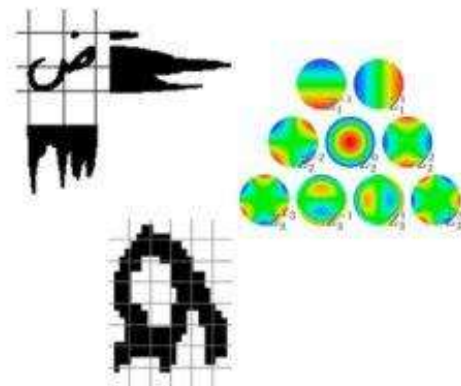
HTR: 30 ans de recherche en 3 minutes

1970 – 1990: reconnaissance des formes (approches analytiques)

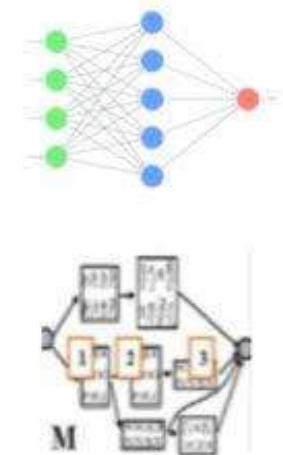
Segmentation



Mesure/classification



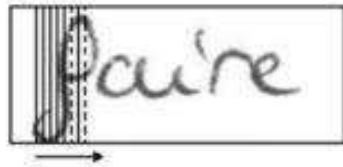
Interprétation



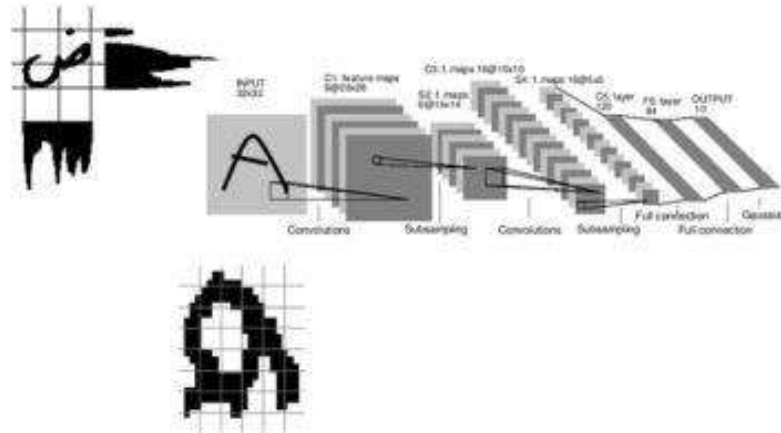
HTR: 30 ans de recherche en 3 minutes

1990 – 2008: apprentissage automatique statistique

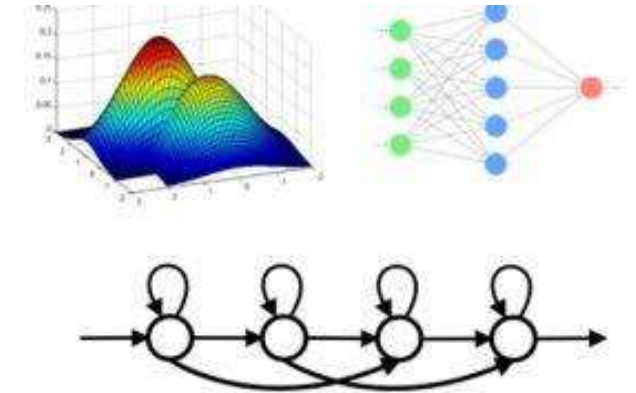
Segmentation



Mesure/classification



Interprétation



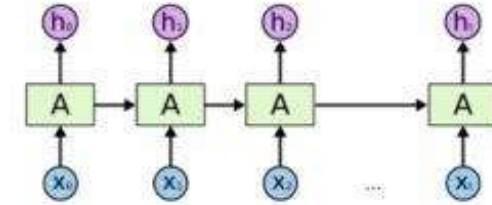
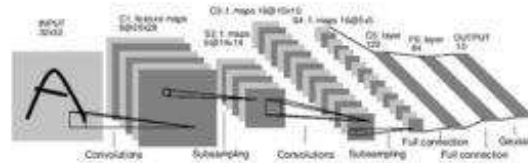
HTR: 30 ans de recherche en 3 minutes

2008: Deep Learning

Segmentation

Mesure/classification

prédictions



L'ère des reseaux de neurones

HTR: 30 ans de recherche en 3 minutes

Even the bone cuff-links found beside the body, which had at first been considered as belonging to the killer, proved yet another red herring, for it was learned that they had been borrowed by Elizabeth Camp from one of her sisters. A young man from Reading named Marshall had an uncomfortable time in the presence of the coroner.

Théodore Bluche, Jérôme Louradour, Ronaldo Messina (2017) Scan, Attend and Read: End-to-End Handwritten Paragraph Recognition with MDLSTM Attention. In 14th International Conference on Document Analysis and Recognition (ICDAR), 2017,

Est-ce que ça fonctionne ?

The image shows a digital interface for document analysis. On the left, a handwritten document is displayed with green highlights and labels. The text is in French and appears to be a church record. On the right, a transcription and annotation interface is shown. It includes a title 'Paragraphe 3', a 'CLASSIFICATIONS' section, and a 'TRANSCRIPTIONS' section. The transcription is a typed version of the handwritten text, with entities like dates, names, and professions highlighted and labeled. A list of 'ALL ENTITIES' is shown at the bottom right, indicating 9 entities were detected.

Paragraphe 3

CLASSIFICATIONS

TRANSCRIPTIONS

filter entities by worker version

Created by Kaldi Balsac

Le **DATE** **sept janvier dix neuf cent un** nous prêtre curé sousigné, avons bap-
tisé **PRENOM** Marie Juliette **ARMOIR** née **DATE** ce jour **PERSONNE** Joseph Gravel
PROFESSION Cultivateur et de **PERSONNE** Anselmé Brault
de cette paroisse, le parrain a été **PERSONNE** Armand Gravel
PROFESSION Cultivateur de cette paroisse, et
la marraine a été **PERSONNE** Alphonsine Lessard
épouse du parrain, oncle et tante de l'en-
fant, lesquels ont signé avec nous
ainsi que le père, après lecture faite
Alphonsine Lessard
Omer Frant le
Joseph Gravel
S. Turcotte Ptre

METADATA

ALL ENTITIES

9 entities

Any worker version

- 2.7M d'images de registres paroissiaux de 1850 à 1920
- HTR, segmentation en acte, détection et typage des entités
- Taux d'erreur caractère moyen : 6.4%

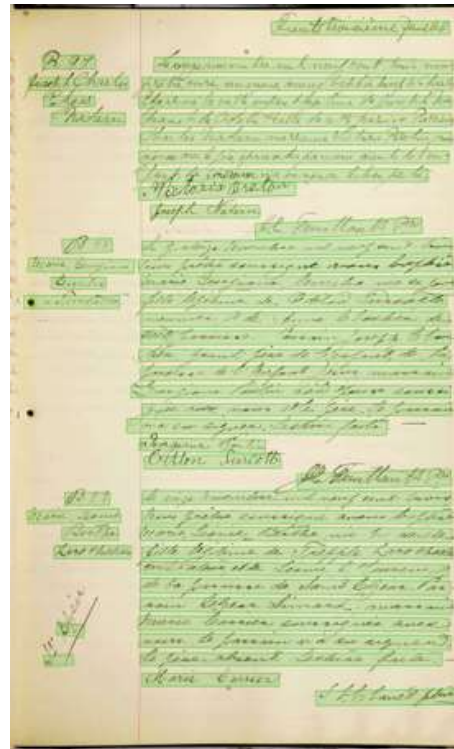
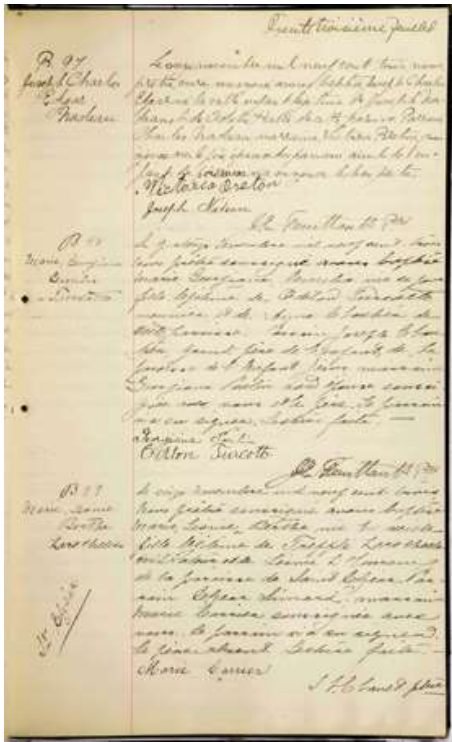
Chaîne de traitement automatique

Image

Détection des lignes

Reconnaissance

Extraction d'entités



Le onze novembre mil neuf cent trois , nous prêtre , curé soussigné , avons baptisé Joseph Charles Edgar né la veille , enfant légitime de Joseph na deau et de Odibs Hallé , de cette paroisse . Parrain Charles Nadeau , marraine Victoria Bleton , vu n ' a signé avec le père époux du parrain , Etaient de l ' en - Vant . Le parrain n ' a su signer . Lechoe faite . J . Victoria Arston Joseph Nadeau J . E . Feuilteault Ptre Le quatorze Novembre mil neuf cent trois nous prêtre soussigné avons baptisé Marie Georgiana , Emilia née ce jour fille légitime de Odilon Turcotte meunier , et de Anna Cloutier de cette paroisse . Parrain Joseph Clau tier , grand père de l ' enfant de la paroisse de l ' Enfant Jésus ; marraine Georgianna Pinaulin son épouse , soussi gné avec nous . et le père . Le parrain n ' a su signer . Lecture faite



Le [onze novembre mil neuf cent trois](#) , nous prêtre , curé soussigné , avons baptisé [Joseph Charles Edgar](#) né [la veille](#) , enfant légitime de [Joseph na deau](#) et de [Odibs Hallé](#) , de cette paroisse . Parrain [Charles Nadeau](#) , marraine [Victoria Bleton](#) , vu n ' a signé avec le père époux du parrain , Etaient de l ' en - Vant . Le parrain n ' a su signer . Lechoe faite . J . Victoria Arston Joseph Nadeau J . E . Feuilteault Ptre Le [quatorze Novembre mil neuf cent trois](#) nous prêtre soussigné avons baptisé [Marie Georgiana](#) , Emilia née [ce jour](#) fille légitime de [Odilon Turcotte meunier](#) , et de [Anna Cloutier](#) de cette paroisse . Parrain [Joseph Clau tier](#) , grand père de l ' enfant de la paroisse de l ' Enfant Jésus ; marraine [Georgianna Pinaulin](#) son épouse , soussi gné avec nous . et le père . Le parrain n ' a su signer . Lecture faite

Tous les modèles sont entraînés à partir d'exemples

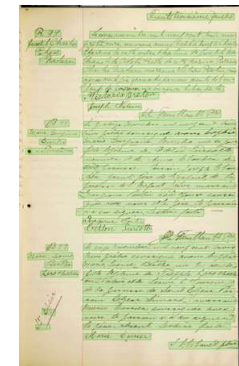
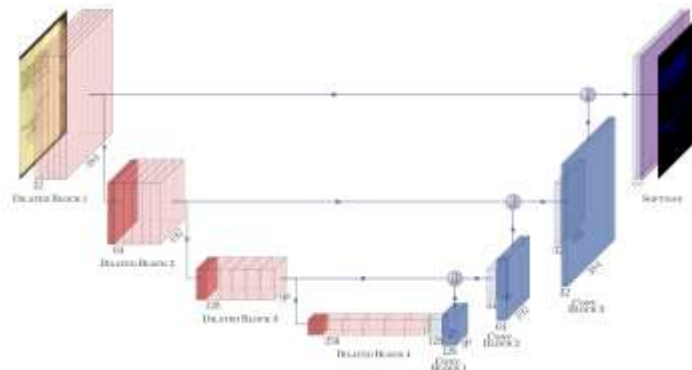
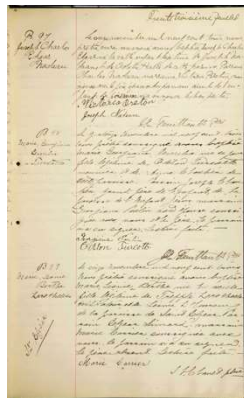
Forces et faiblesses des reseaux de neurones



- Peuvent être entraînés sur de très grandes quantités de données
- Très bons pour restituer ce qu'ils ont appris



- Nécessitent des données annotées manuellement
- Mauvais pour prédire dans des situations inconnues



Listes nominatives de recensement

Points favorables au projet:

- Documents normalisés au niveau national par année de recensement
- Structuration constante par ligne et colonne
- Séries de même main
- Statistiques récapitulatives pour validation croisée

Listes nominatives de recensement

Défis du projet:

- Documents décentralisés
- Volume inconnu mais important (entre 5 et 15 M d'images ?)
- Conditions de conservation/numérisation variables
- Lecture en 2 dimensions
- Evaluer l'erreur
- Structurer, consolider, croiser l'information

Lecture en 2 dimensions

		3	2	49	186	nom Beulot ni Beusgat	anne	indigente					1	76	1
				50	187	Maurin	jeune	G ^e Beulanger	1					54	1
					188	Gatobert fem Maurin	Marié	vivant en commun avec son mari				1		52	1
					189	Leur fils	organe	vivant en commun avec son père	1					16	1
					190	Leur fille	Lise	"			1			11	1
				51	191	Colrat	Parce	Carrossier a façon	1					45	1
				4	192	Graven fem Colrat	Acadie	vivant en commun avec son mari				1		45	1
				52	5	Van Privat	Ernest	Pensionné					1	36	1
1	2				194	Leur fils	Louis	"	1					8	1
					195	Banal	Justine	Comestique			1			25	1
				53	196	nom Martin	Marié	Pensionné					1	75	1
					197	La fille	Marié	vivant en commun avec son mari			1			32	1
					198	Marié	"	Comestique			1			48	1

1. Quartier
2. Rue
3. Maison
4. Ménage
5. Individu

Appel à manifestation d'intérêt !

Contact:

Christopher Kermorvant

Lionel Kesztenbaum

contact@socface.org