SOCFACE

Un projet de recherche collaboratif entre

Historiens, économistes, démographes

Archivistes

Experts en intelligence artificielle

ined
INSTITUT NATIONAL D'ÉTUDES DÉMOGRAPHIQUES

PARIS SCHOOL OF ECONOMICS
ÉCOLE D'ÉCONOMIE DE PARIS

FranceArchives
PORTAIL NATIONAL DES ARCHIVES

TEKLIA

Et les services d'archives départementales et municipales

Bénéficie d'un financement sur 3 ans ½

ANR AGENCE NATIONALE DE LA RECHERCHE

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.

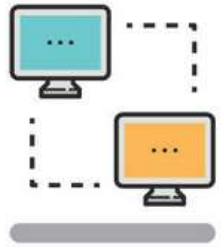
De multiples objectifs et enjeux

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.

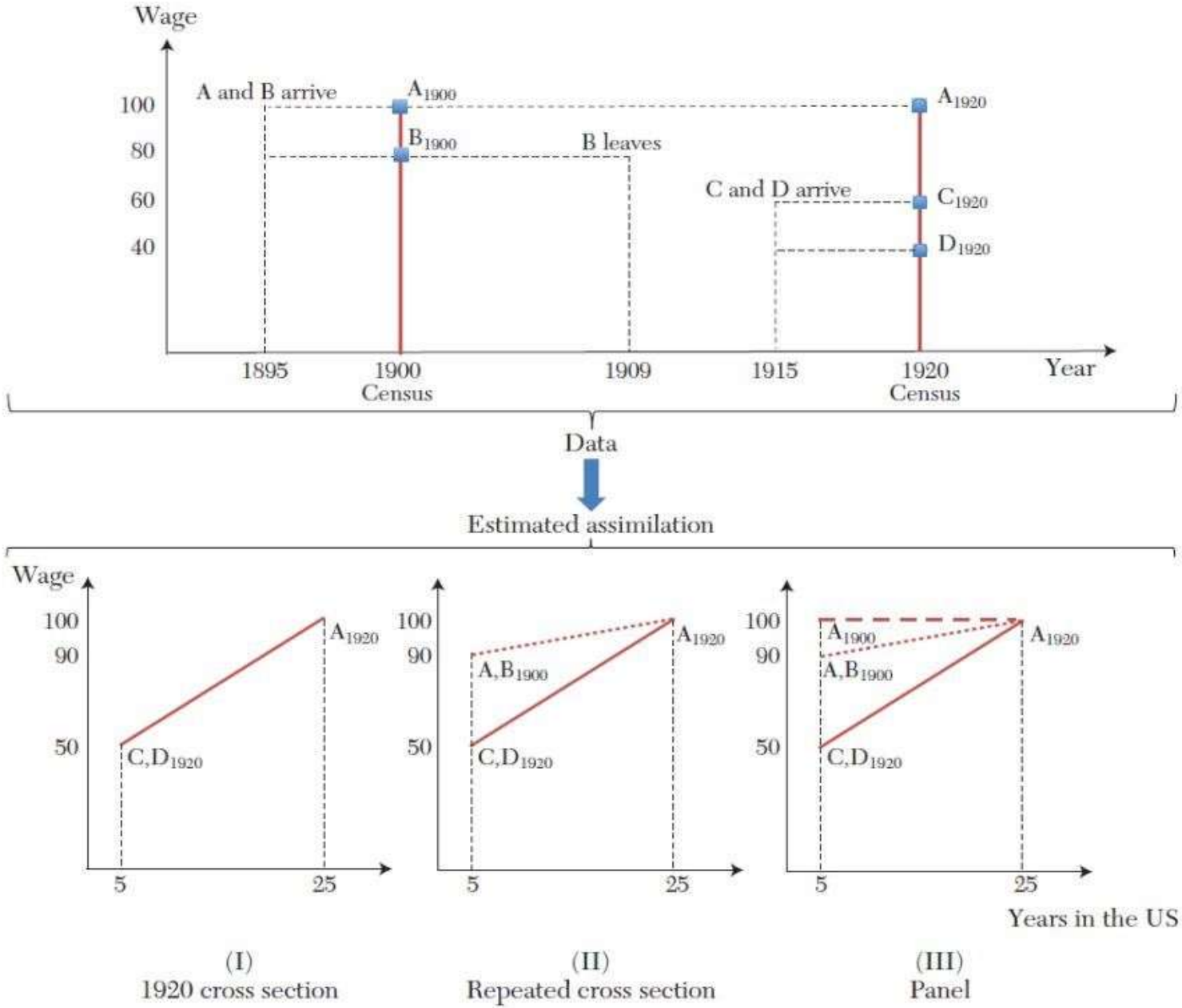
Produire des données individuelles en masse à l'échelle nationale

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:

Situation des immigrés sur le marché du travail américain



Abramitzky, Ran, et Leah Boustan. 2017. « Immigration in American Economic History ». Journal of Economic Literature 55 (4): 1311-45.

Pourquoi Socface?

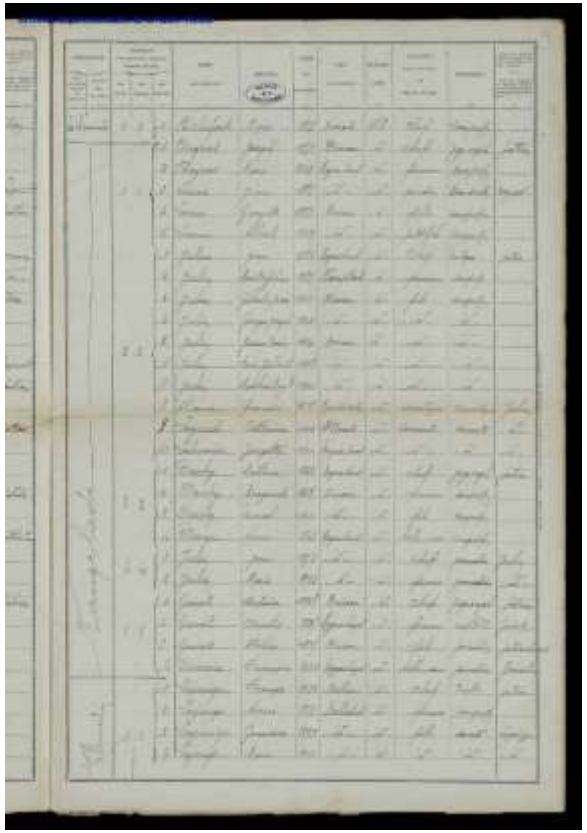
- Importance croissante des données micro (individuelles) dans la recherche quantitative en sciences sociales: économie, histoire, sociologie, démographie...
- Progrès considérables dans les technologies de reconnaissance d'écriture manuscrite et de traitement d'images.

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?

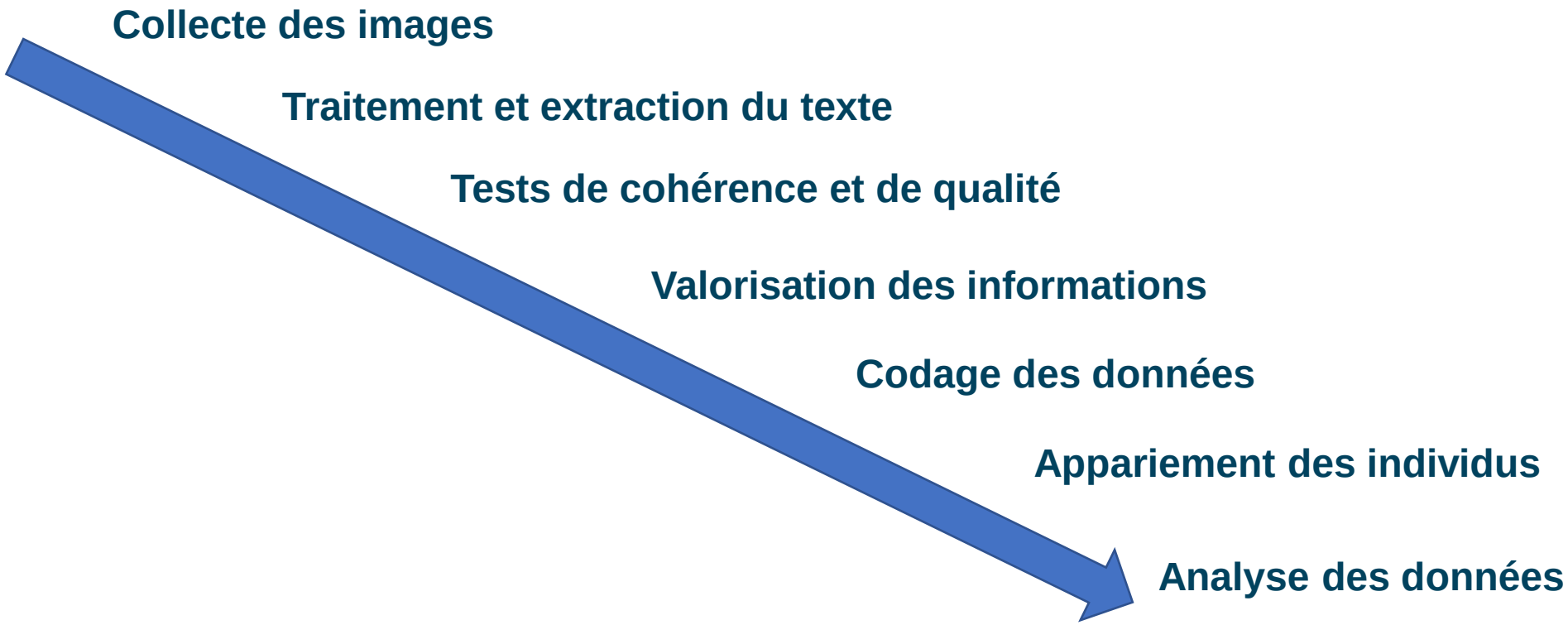


- Une source abondante, simple, stable et standardisée.
- Une source régulière dans le temps.
- Déjà photographiée par beaucoup de services d'archives.
- Source nationale.
- Permet de construire un échantillon de toute la France (ou presque…).

→ 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

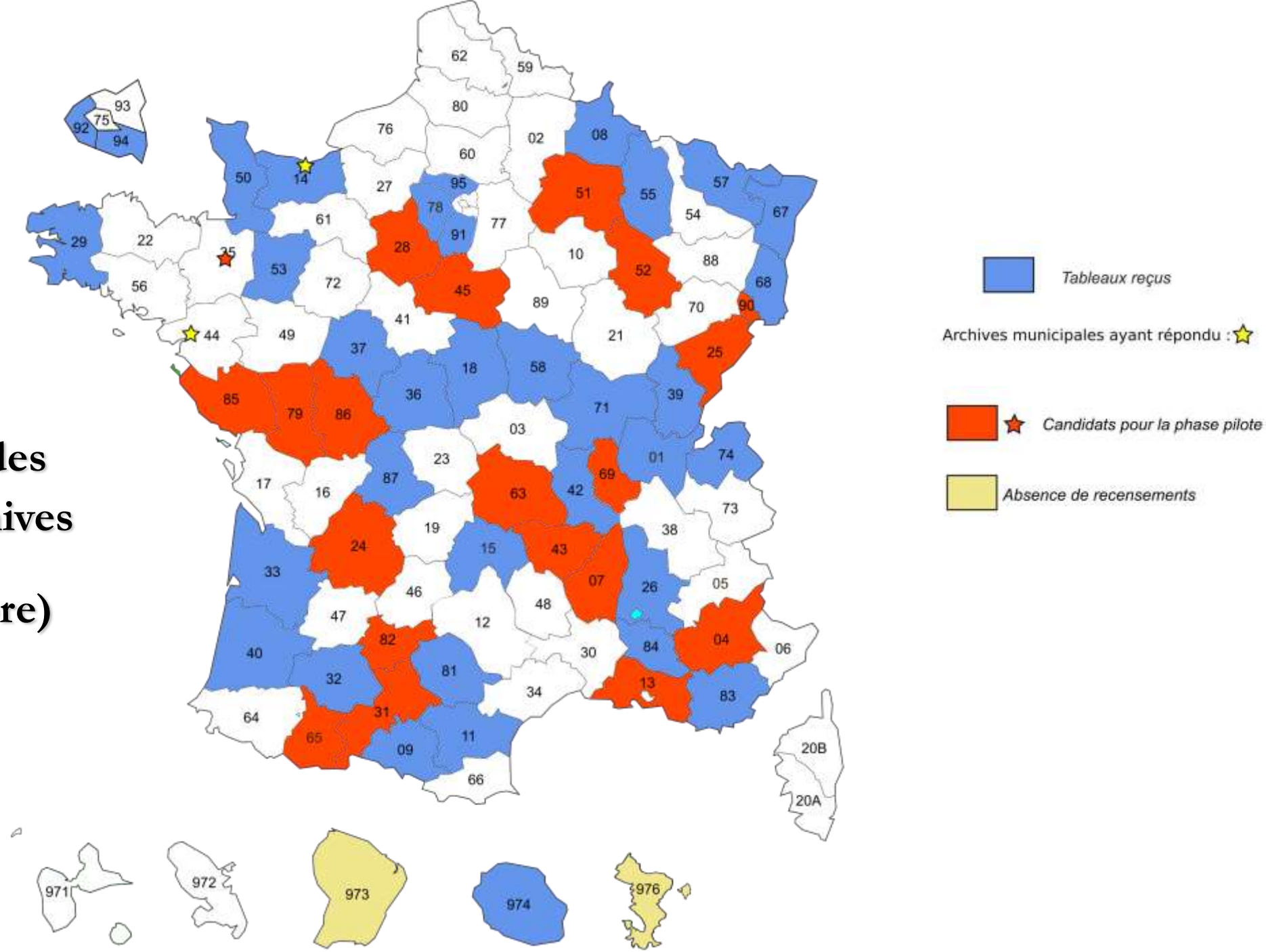
- Traitement automatisé de vastes ensembles d'images
  - Plusieurs millions d'images, de provenances variées (différents éditeurs, formats, etc.).
  - Ouvre des perspective considérables pour la recherche en intelligence artificielle.
- Une coproduction entre historiens et informaticiens
  - Utiliser 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 perspective 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 ?

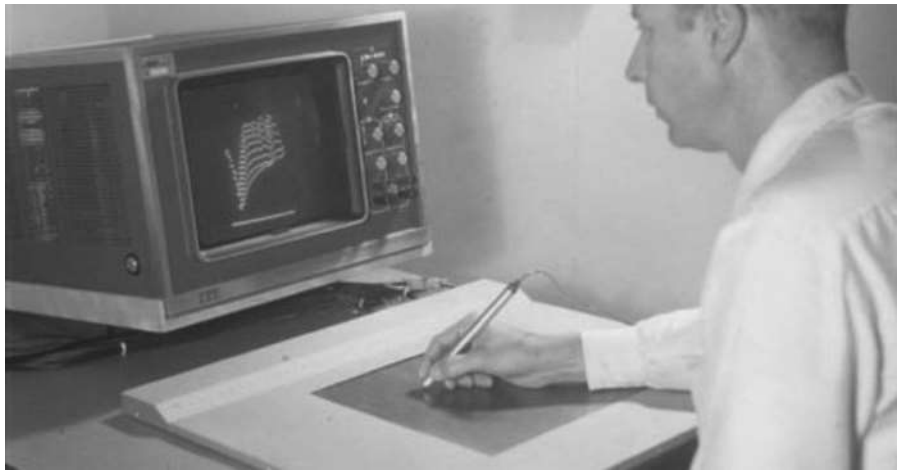
Séminaire annuel des archives de France

Christopher Kermorvant

23 septembre 2021

TEKLIA

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



RAND corporation, 1960

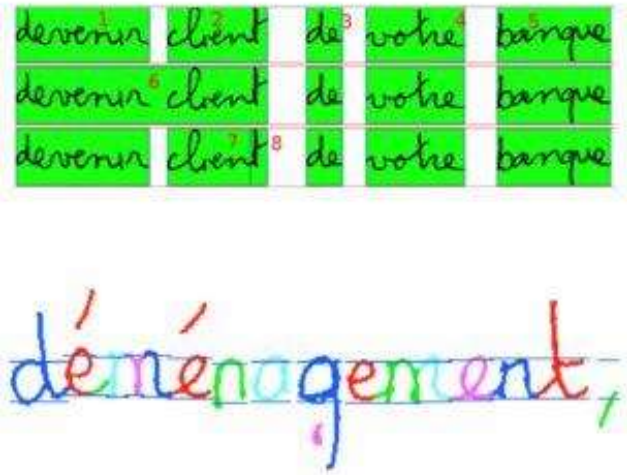
The MNIST database
“The drosophila of machine learning”
Geoffrey Hinton

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

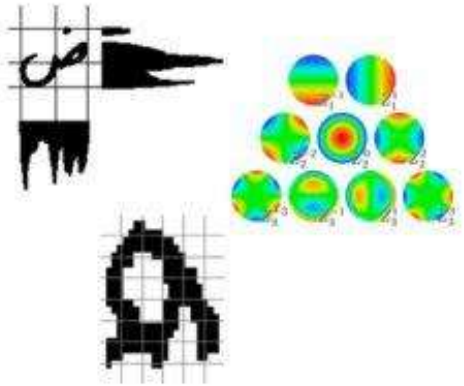
HTR: 30 ans de recherche en 3 minutes

1970 – 1990: reconnaissance des formes (approches analytiques)

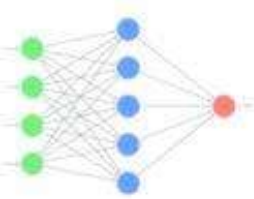
Segmentation

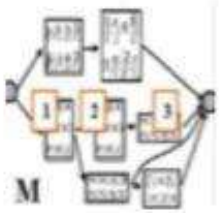


Mesure/classification



Interprétation





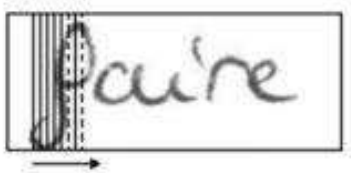
HTR: 30 ans de recherche en 3 minutes

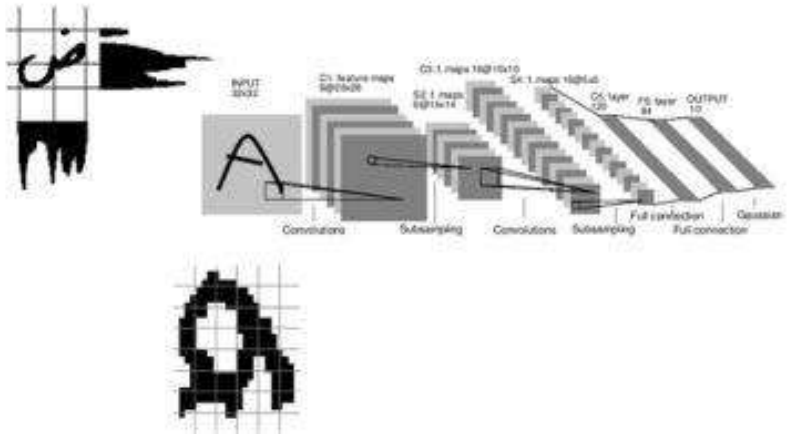
1990 – 2008: apprentissage automatique statistique

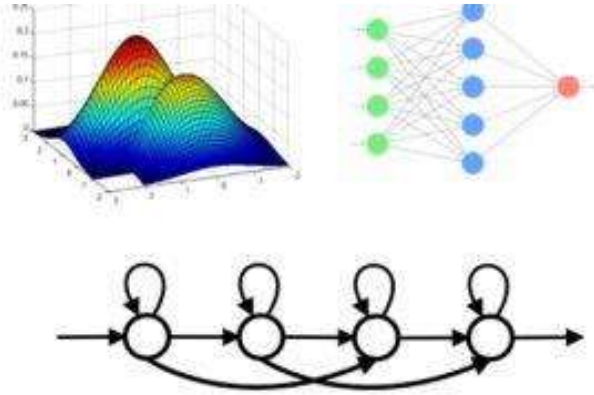
~~Segmentation~~

Mesure/classification

Interprétation



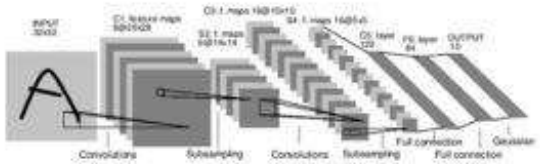


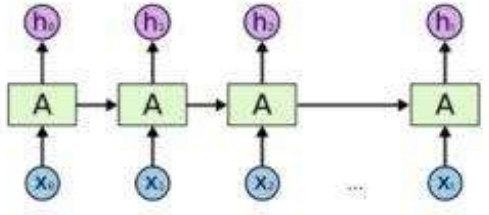


HTR: 30 ans de recherche en 3 minutes

2008: Deep Learning

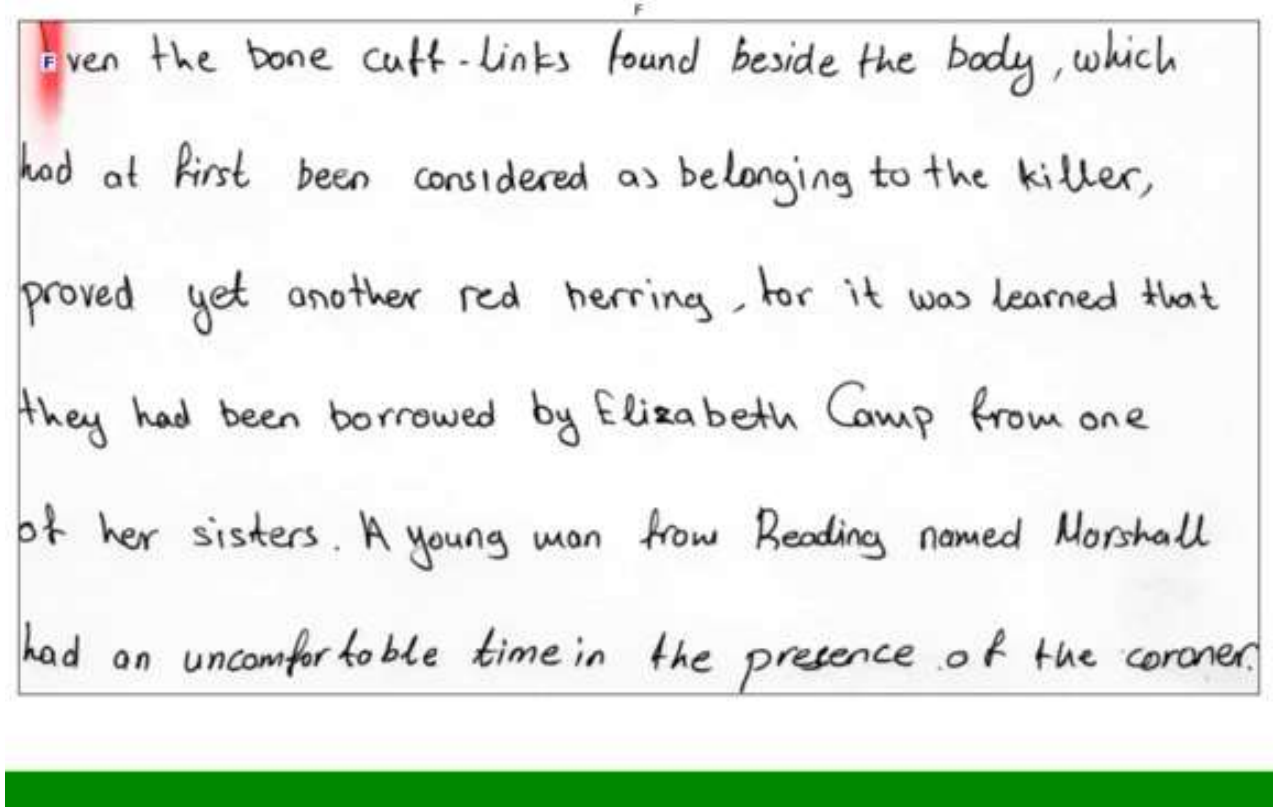
~~Segmentation~~ ~~Mesure/classificationc~~ prédictions





L’ère des reseaux de neurones

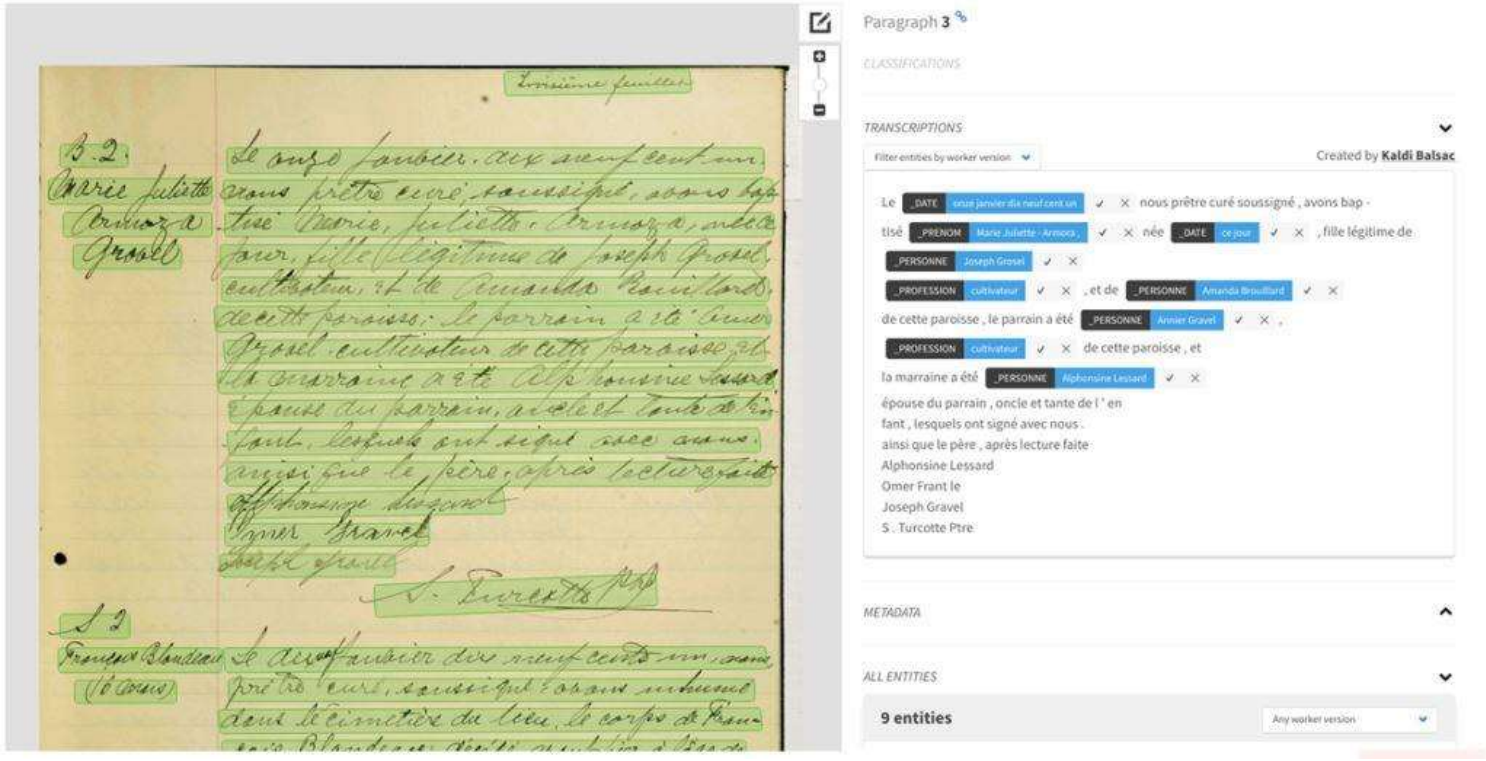
HTR: 30 ans de recherche en 3 minutes



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 ca fonctionne ?

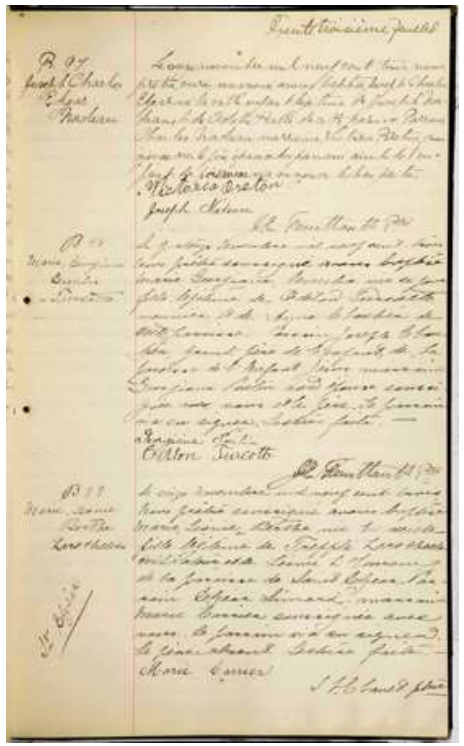
UQAC
Université du Québec
à Chicoutimi



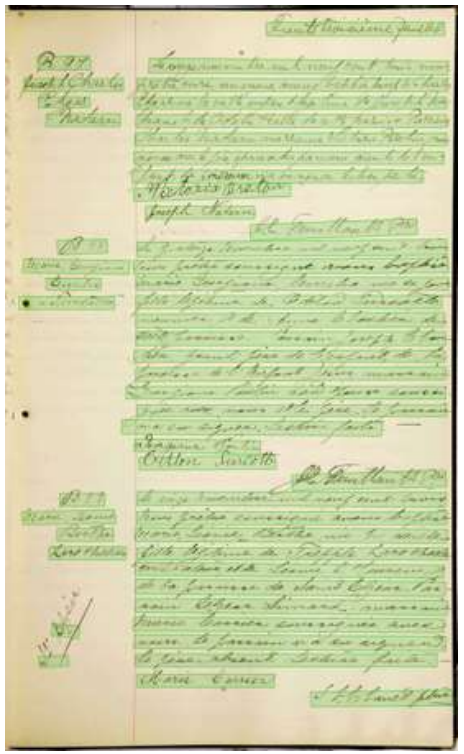
- 2.7M d'images de registres paroissiaux de 1850 à 1920
- HTR, segmentation en acte, detection et typage des entités
- Taux d'erreur caractère moyen : 6.4%

Chaîne de traitement automatique

Image



Détection des lignes



Reconnaisssance

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 , Etaients de l ' en - Vant . Le parrain n ' a su signer . Lechore 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

Extraction d'entités

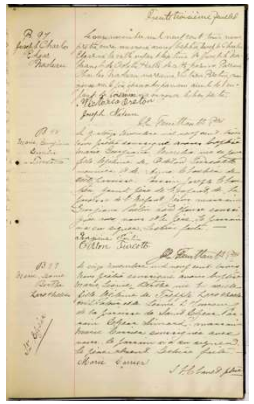
Le onze novembre mil neuf cent trois , nous prêtre , curé soussigné , avons baptisé Joseph Charles Edgar né la veille , enfant légitime de Joseph na deau et de Odibs Hallé , de cette paroisse . Parrain Charles Nadeau , marraine Victoria Bleton , vu n ' a signé avec le père époux du parrain , Etaients de l ' en - Vant . Le parrain n ' a su signer . Lechore 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

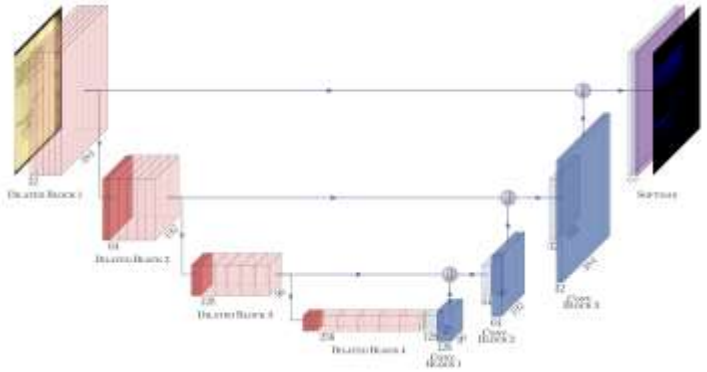
Tous les modèles sont entrainés à partir d'exemples

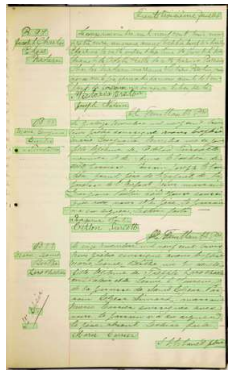
Forces et faiblesses des reseaux de neurones

- Peuvent être entrainés 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

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

Appel à manifestation d’intérêt !

Contact:

Christopher Kermorvant
Lionel Kesztenbaum

contact@socface.org