Mining the Vatican
Secret Archives

In Codice Ratio

Something about me

Assistant Professor @ Sapienza, with a focus on deep learning.

Strong interest in promoting machine learning in Italy.

Co-organizer
ML/Data Science Meetup Rome

Co-founder / chairman
Italian Association for Machine Learning

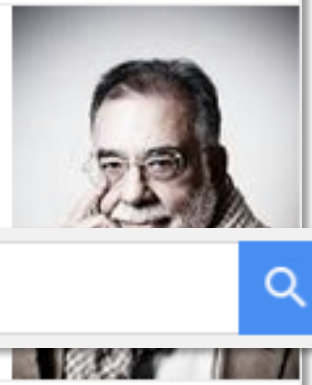
Google Developer Expert
Machine Learning

MACHINE LEARNING
DATA SCIENCE
Meetup

iaml
ITALIAN ASSOCIATION FOR MACHINE LEARNING

Apocalypse Now / Director

Francis Ford Coppola



Google who is the director of apocalypse now?

Francis Ford Coppola is an American screenwriter, film director, and producer. He was considered to be the central figure of the New Hollywood wave of filmmaking. Wikipedia

Born: April 7, 1939 (age 77), Detroit, MI

Spouse: Eleanor Coppola (m. 1963)

Children: Sofia Coppola, Roman Coppola, Gian-Carlo Coppola

Nephews: Nicolas Cage, Jason Schwartzman, More

Siblings: Talia Shire, August Coppola

Movies

View 40+ more

Sofia Coppola / **Place of birth**

Google what is the birthplace of the daughter of the director of apocalypse now

New York City, NY

Home to the Empire State Building, Times Square, Statue of Liberty and other iconic sites, New York City is a fast-paced, globally influential center of art, culture, fashion and finance. The city's 5 boroughs sit where the Hudson River meets the Atlantic Ocean, with the island borough of Manhattan at the "Big Apple's" core.

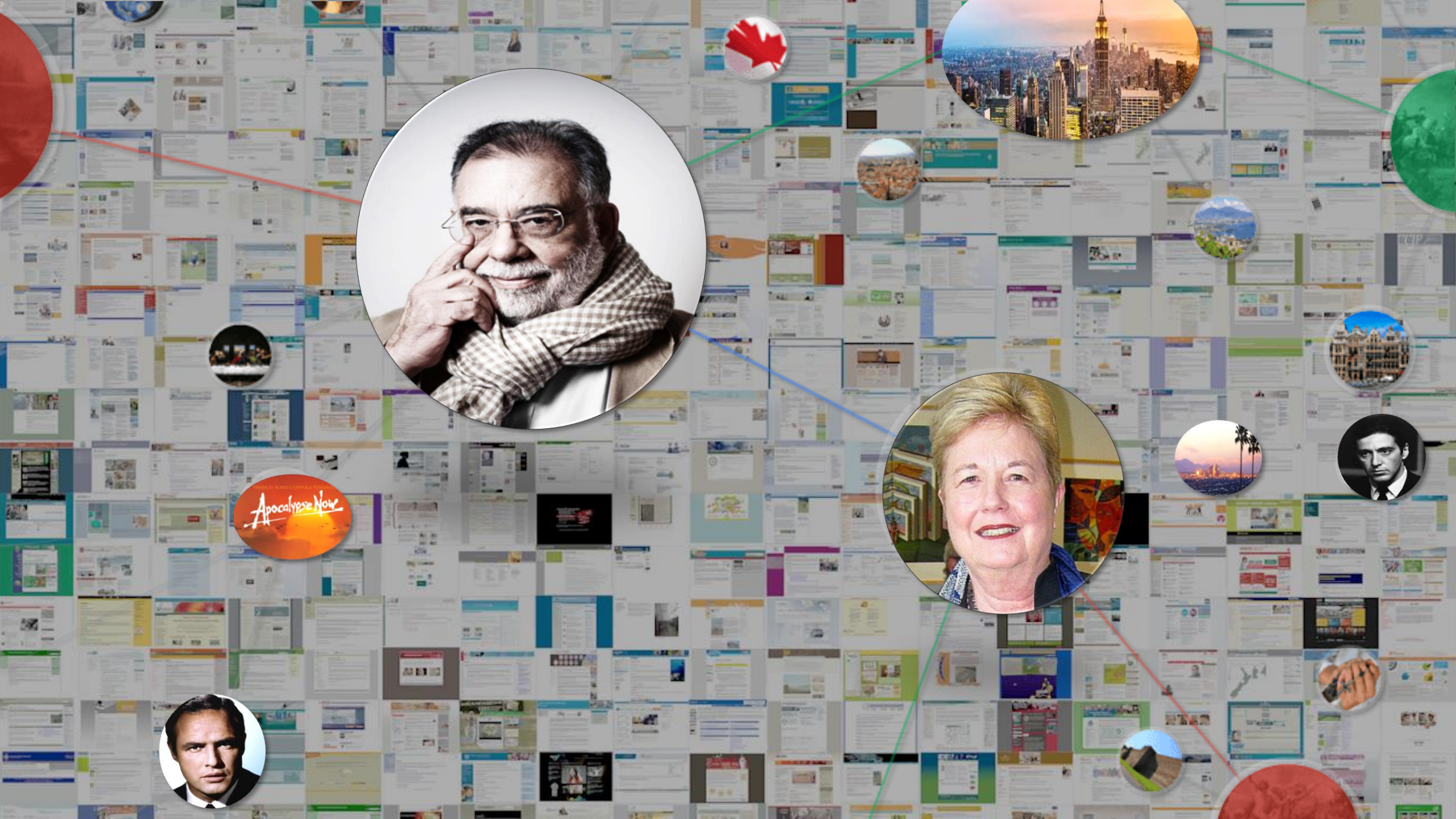
Land area: 304.6 mi²

Weather: 80°F (27°C), Wind S at 11 mph (18 km/h), 51% Humidity

Hotels: 3-star averaging $210, 5-star averaging $420. View hotels

Getting there: 5 h 19 min flight, around $405. View flights

Local time: Thursday 6:34 PM


Apocalypse Now

Lector

an ongoing research project at RM3
in collaboration with UofA

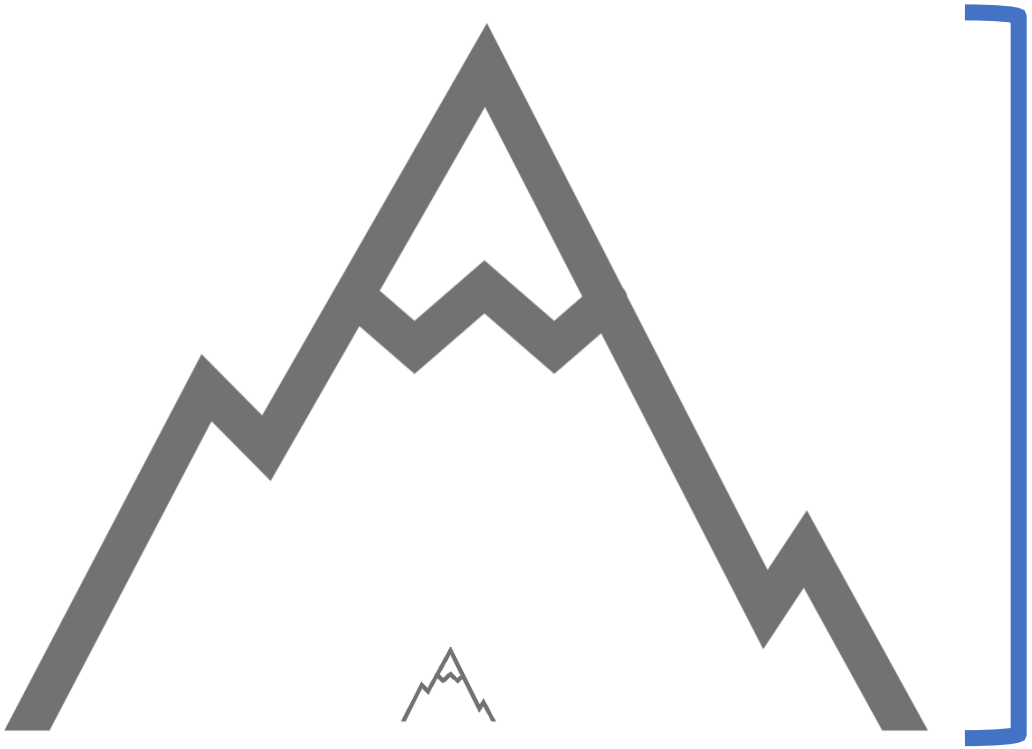
3M new facts

extracted from Wikipedia

Fun fact:

Rome hosts the largest historical archive in the world

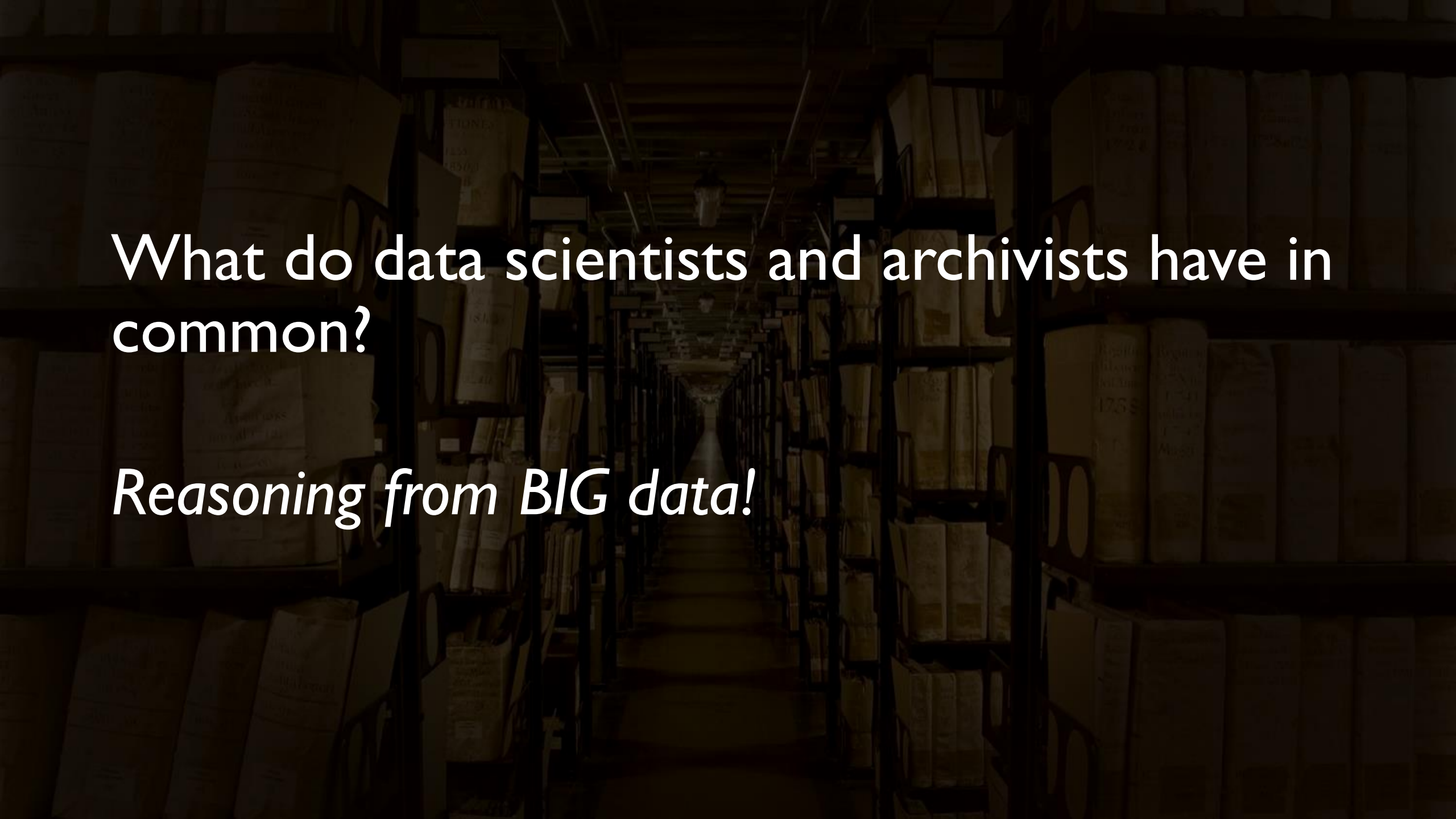
How large is the Vatican secret archive?



Mt. Everest
8.8 km

All shelvings of the Archive, vertically stacked
85 km

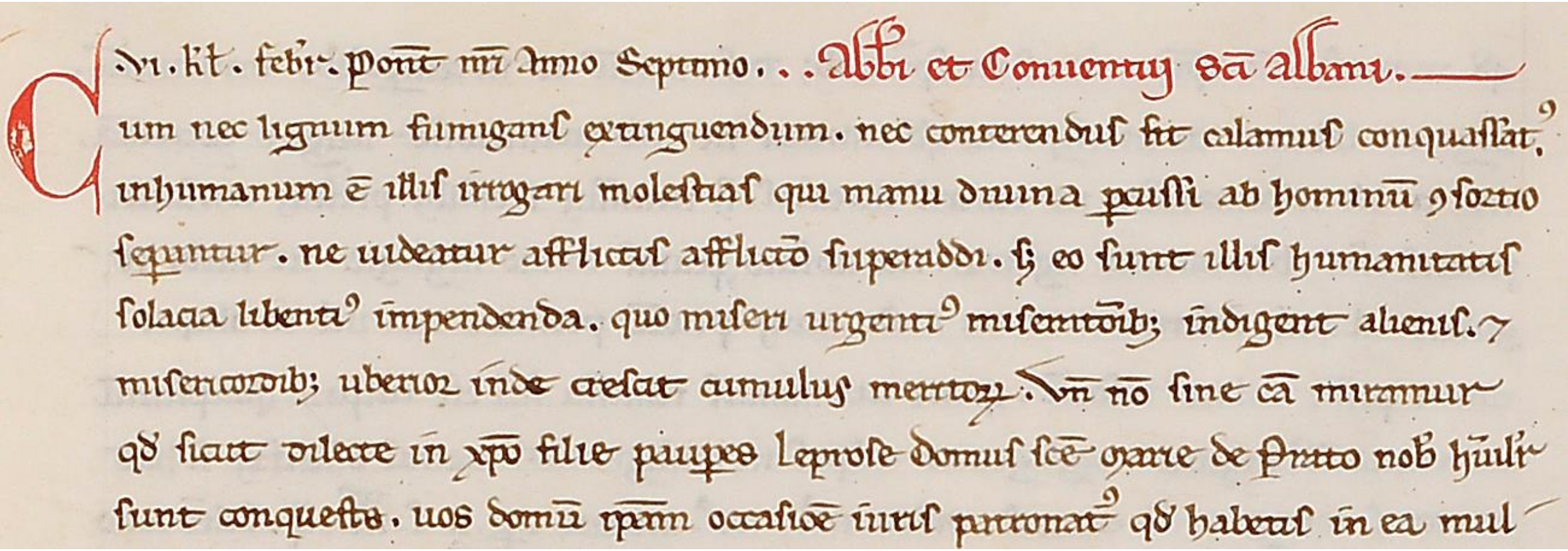
What is in the archive?



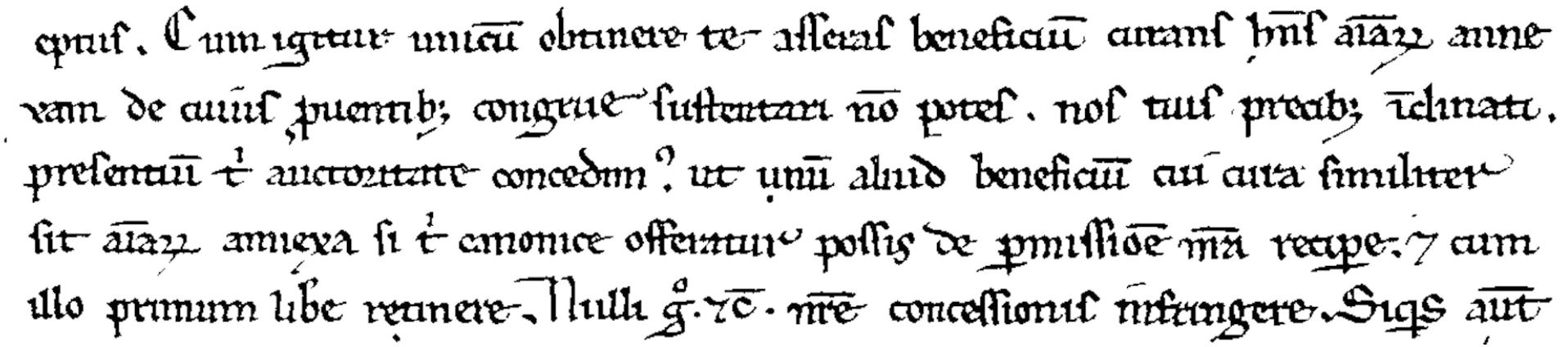
What do data scientists and archivists have in common?

Reasoning from BIG data!

What we are working on

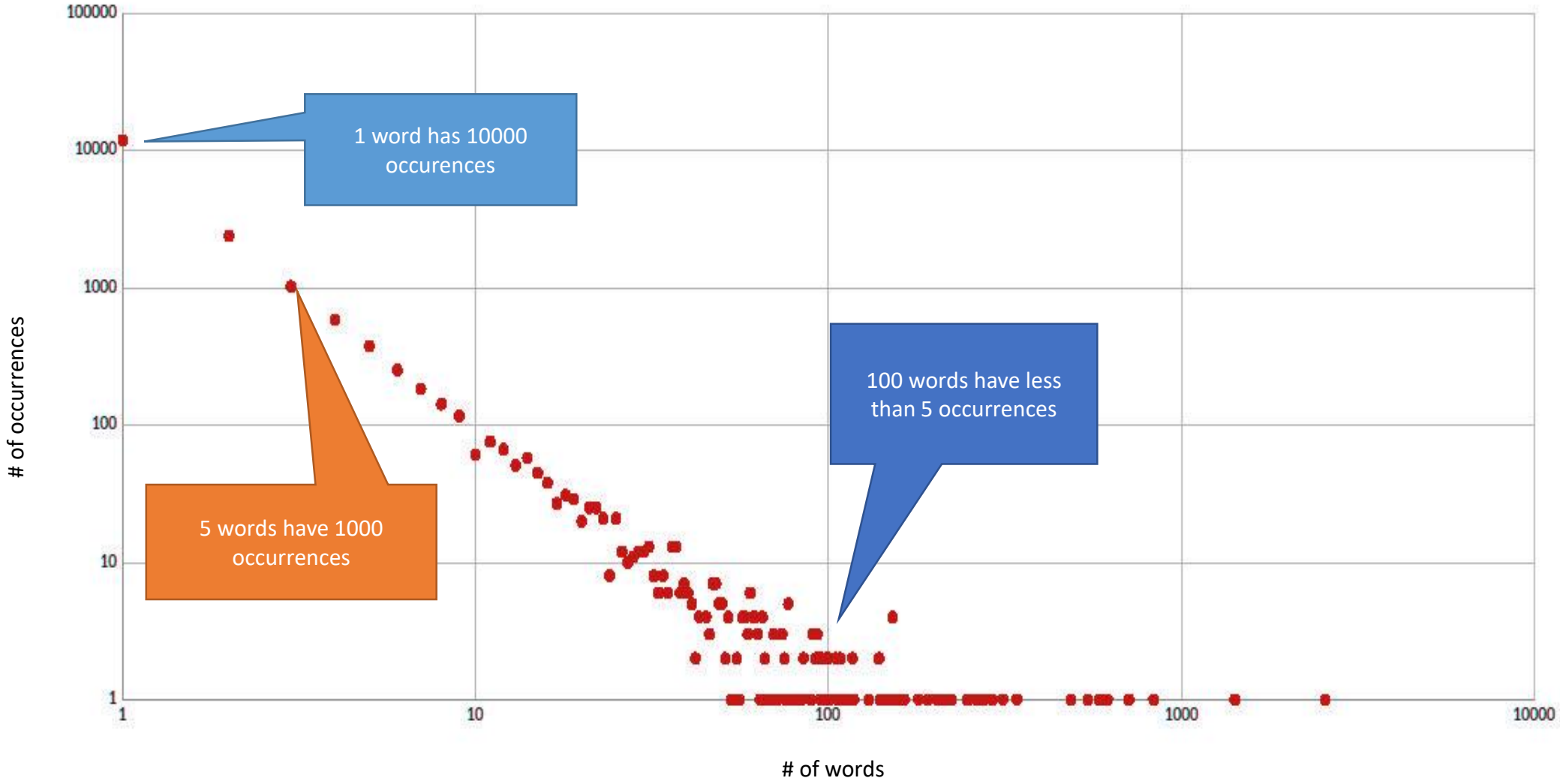


Challenge #1: transcription

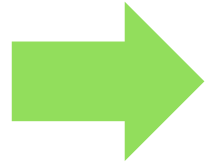


-eptus. Cum igitur unicu(m) obtinere te asseras beneficiu(m) curans h(abe)ns a(n)i(m)ar(um) anne-
xam de cuius p(ro)ventib(us) congrue sustentari no(n) potes nos tuis precib(us) i(n)clinati
presentiu(m) t(ib)i auctoritate concedim(us) ut unu(m) aliud beneficiu(m) cui cura similiter
sit a(n)i(m)ar(um) annexa si t(ib)i canonice offeratur possis de p(er)missio(n)e n(ost)ra
recip(er)e et cum
illo primum lib(er)e retinere. Nulli (er)go et c(etera) n(ost)re concessionis infringere. Siq(ui)s aut(em)

Challenge #2: scalability



Let's start from characters



Crowdsourcing the annotation process (2016 edition)

120 high school students

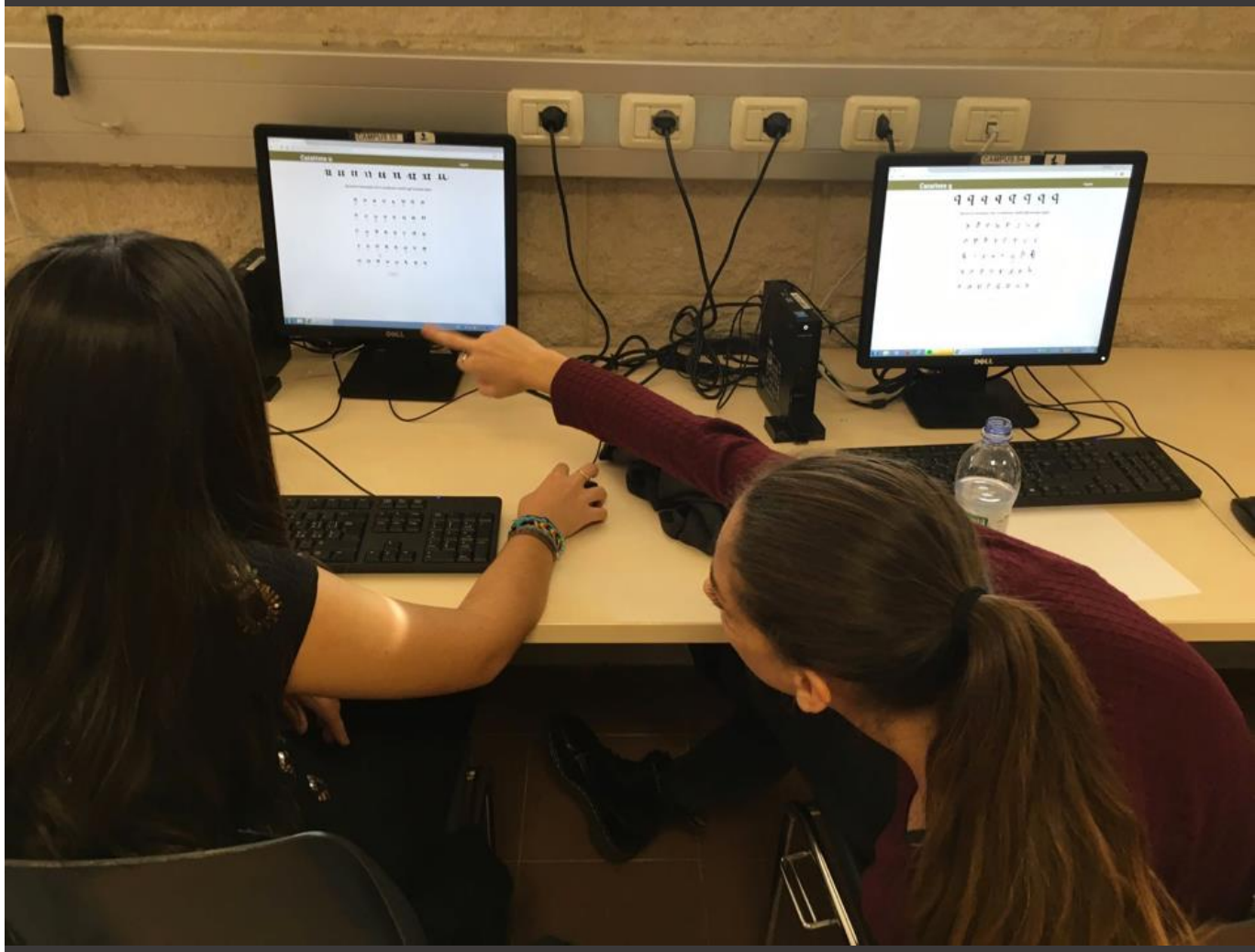
25 symbols

263 029 images

Every image labeled by 4 students

1M labelled images

13 008 positive examples



Crowdsourcing the annotation process (2018 edition)

500+ high school students

33 symbols

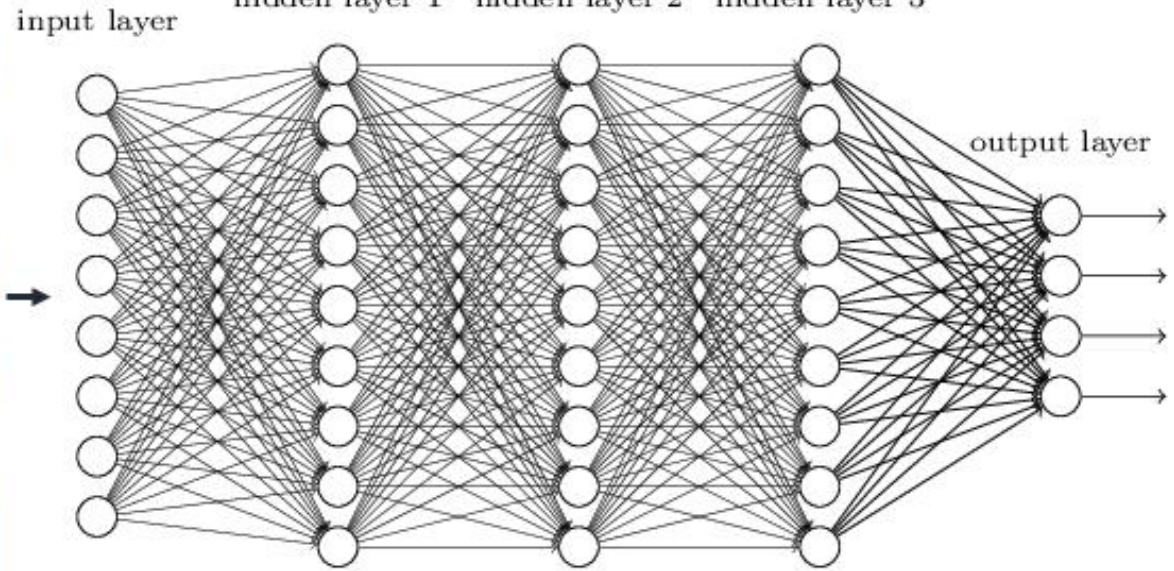
500k + images

Every image labeled by 5+ students

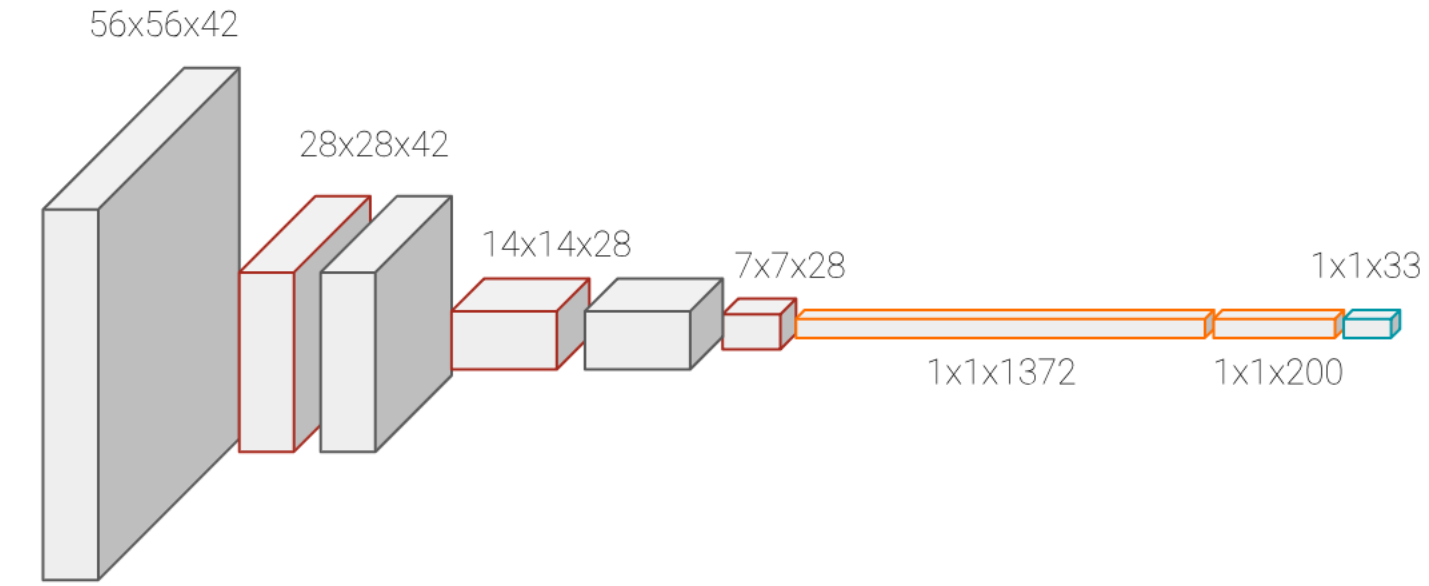
> 40k annotations

Convolutional neural networks

Deep neural networks learn hierarchical feature representations



Our model



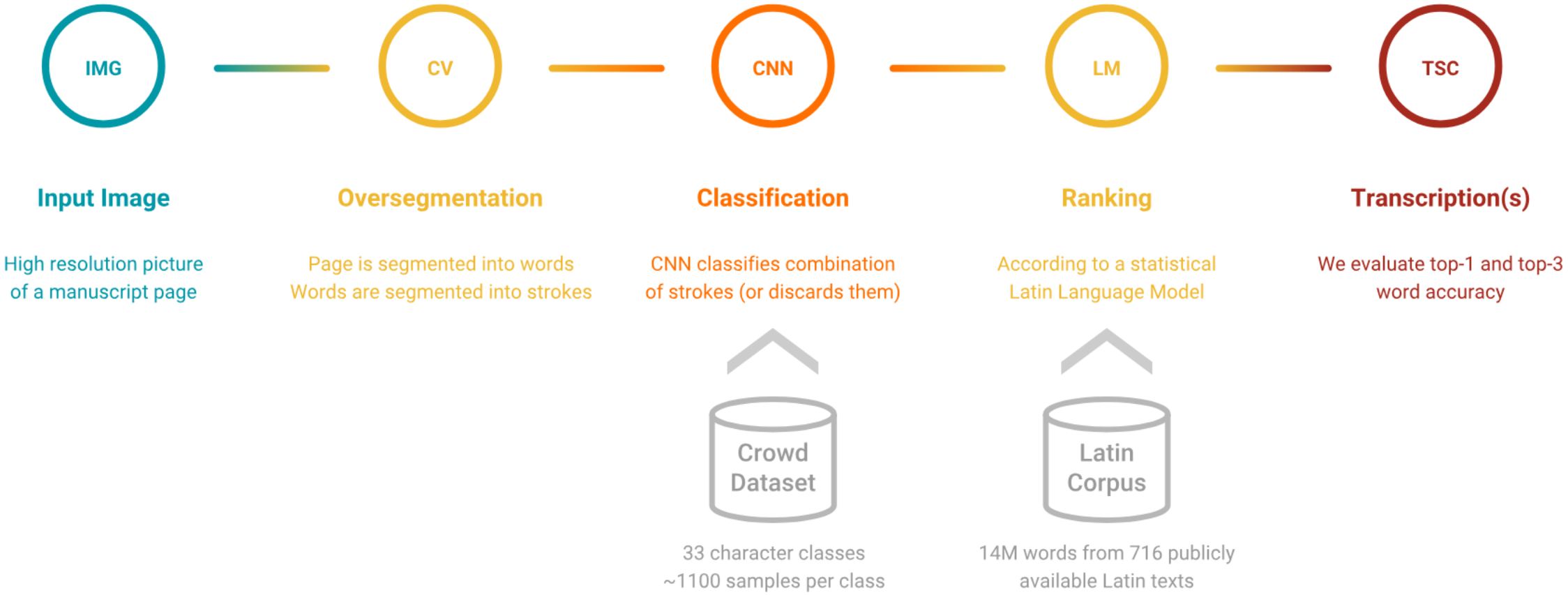
- **Optimizer:** Adam
- **Initializer:** Glorot normal
- **Regularizers:**
  - Dropout = 0.5
  - L2, λ = 0.001

94% avg accuracy

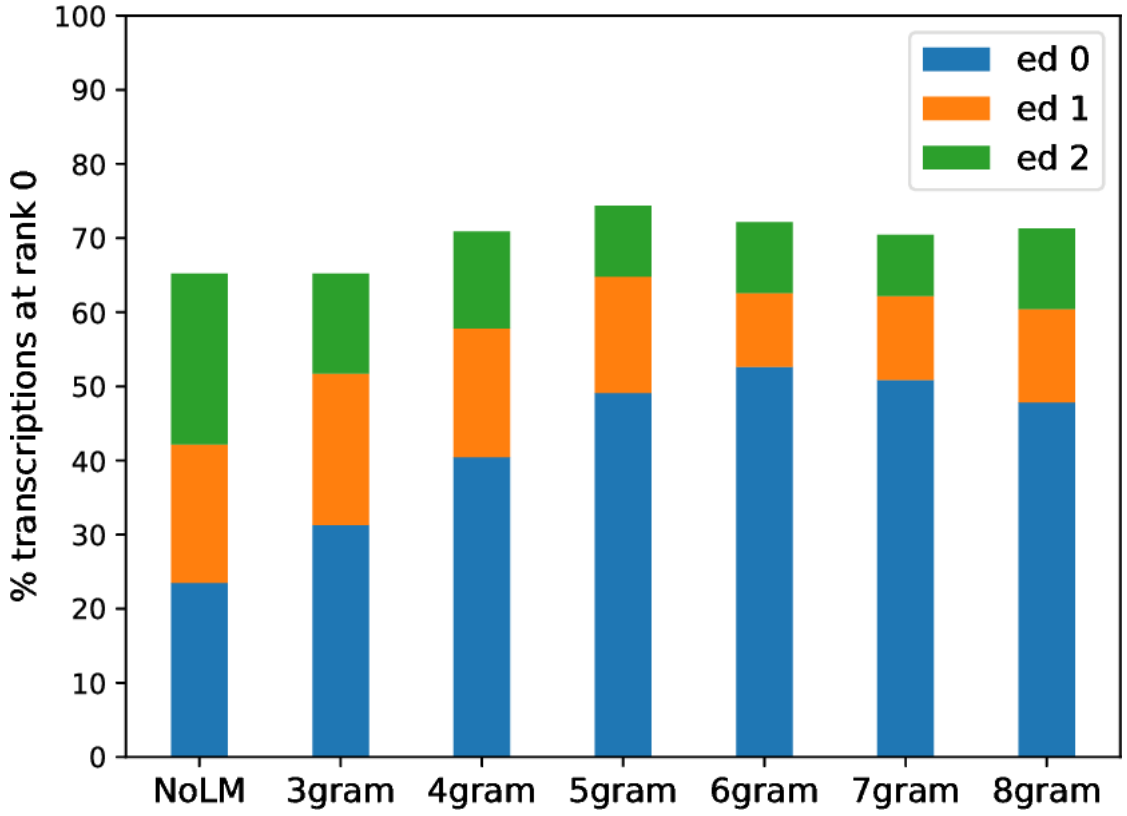
Convolution+ReLU Max Pooling Fully Connected+ReLU Fully Connected+Softmax

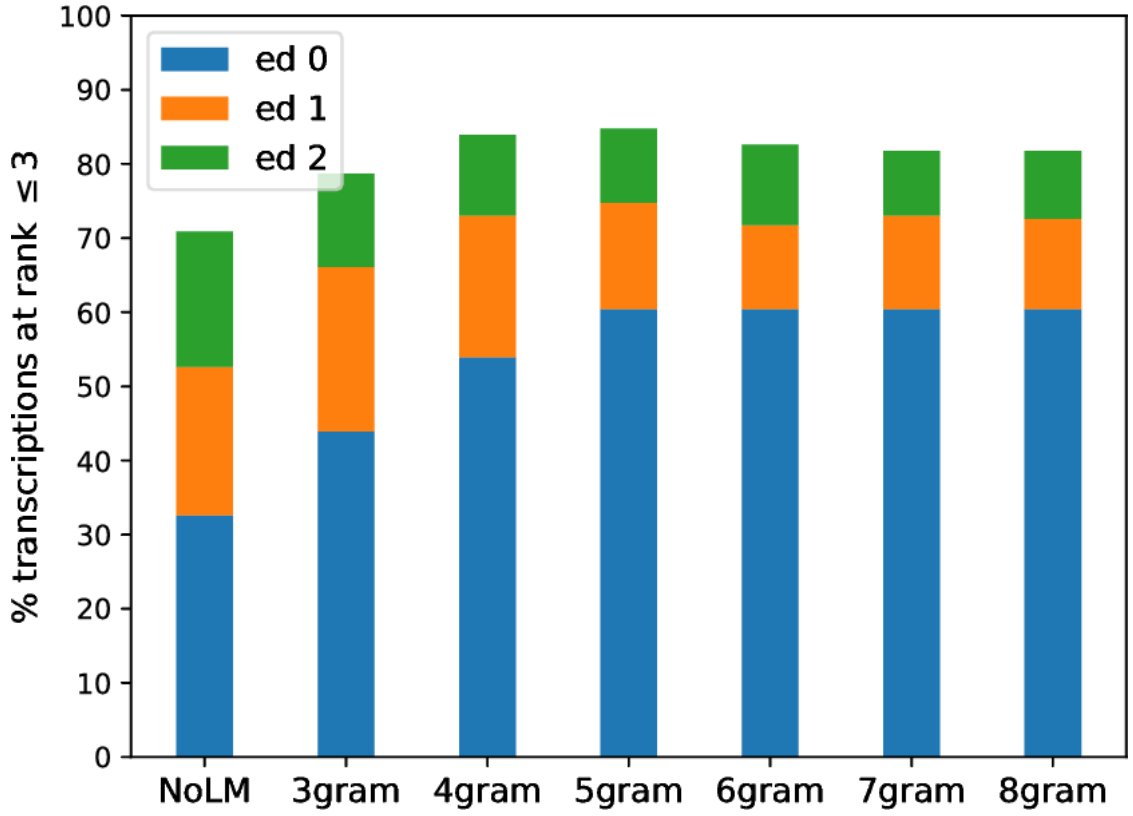
Firmani, D., Merialdo, P., Nieddu, E. and Scardapane, S., 2017. **In Codice Ratio: OCR of Handwritten Latin Documents using Deep Convolutional Networks**. In *AI* CH@ AI* IA* (pp. 9-16).

The complete pipeline

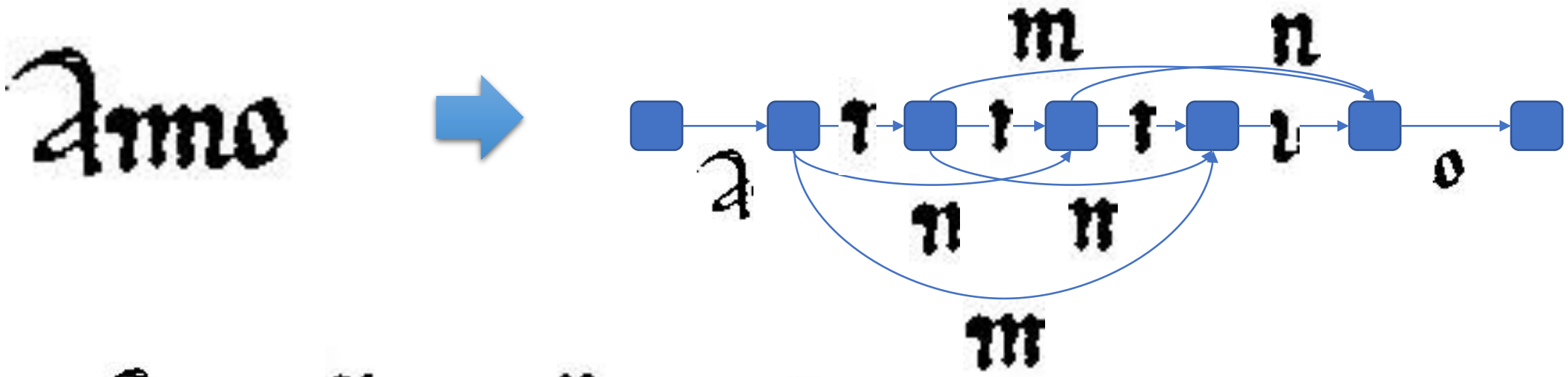


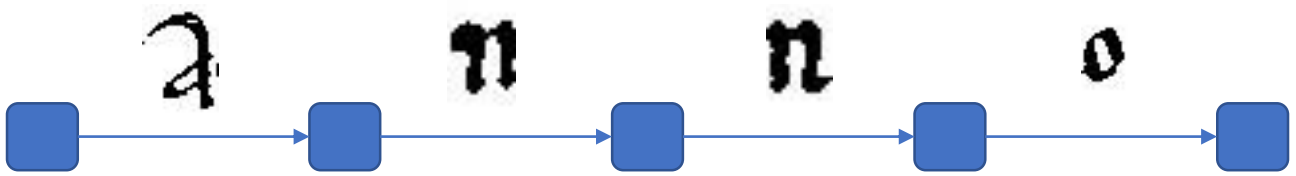
Current post-processing pipeline



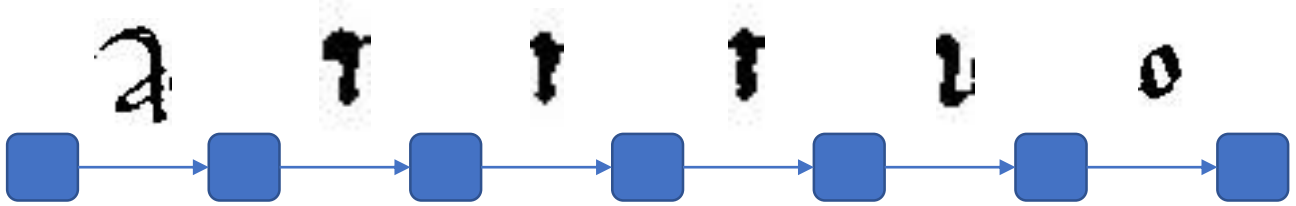


The post-processing phase





P(anno) = 1,7*10^{-7}

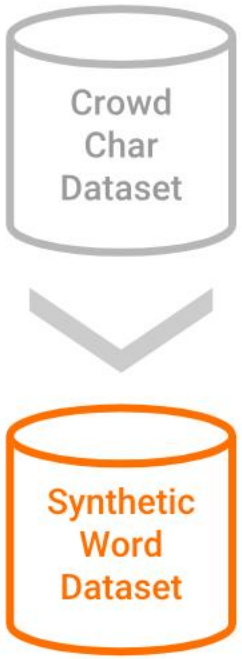


P(aiiiio) = 2,4*10^{-21}

What's next?

Image segmentation

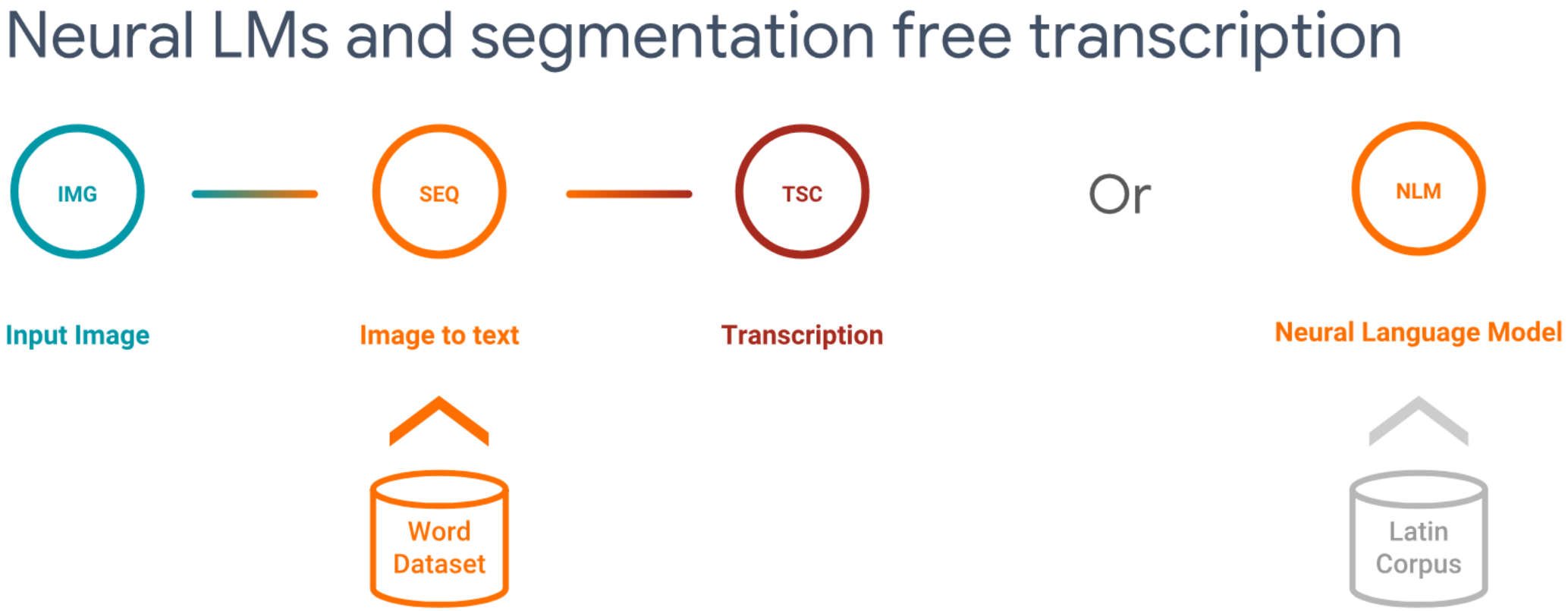
Syntethic data generation



Input:

Generated:

A full neural pipeline



The In Codice Ratio team

PAOLO MERIALDO
Professor
Università degli Studi Roma Tre
Dipartimento di Ingegneria
paolo.merialdo@uniroma3.it

SERENA AMMIRATI
Professor
Università degli Studi Roma Tre
Dipartimento Studi Umanistici
serena.ammirati@uniroma3.it

MARCO MAIORINO
Professor
Archivio Segreto Vaticano
Scuola Vaticana di Paleografia, Diplomatica e Archivistica
marco.maiorino2109@gmail.com

DONATELLA FIRMANI
Assistant Professor
Università degli Studi Roma Tre
Dipartimento di Ingegneria
donatella.firmani@uniroma3.it

SIMONE SCARDAPANE
Assistant Professor
Università degli Studi La Sapienza
Dipartimento di Ingegneria dell'Informazione, Elettronica e Telecomunicazioni
simone.scardapane@uniroma1.it

ELENA NIEDDU
PhD student
Università degli Studi Roma Tre
Dipartimento di Ingegneria
elena.nieddu@uniroma3.it

To learn more (publications)

Donatella Firmani, Marco Maiorino, Paolo Merialdo, Elena Nieddu. Towards Knowledge Discovery from the Vatican Secret Archives. In Codice Ratio - Episode 1: Machine Transcription of the Manuscripts. KDD 2018: 263-272.

Donatella Firmani, Paolo Merialdo, Elena Nieddu, Simone Scardapane. In Codice Ratio: OCR of Handwritten Latin Documents using Deep Convolutional Networks. AI*CH@AI*IA 2017: 9-16.

To learn more (media)

Artificial Intelligence Is Cracking Open the Vatican's Secret Archives - The Atlantic (Italian version, by Internazionale)

AI tackles the Vatican's secrets - MIT Technology Review

Focus-on: In Codice Ratio – Blog of the Italian Association for Machine Learning