Université de Montréal

How To Extract Good Knowledge From Bad Data

An Experiment With 18th-Century French Texts

François Dominic Laramée
PhD candidate in History, Université de Montréal
Presented at CSDH / SCHN 2017 Annual Conference
This work was supported by a FRQSC doctoral scholarship

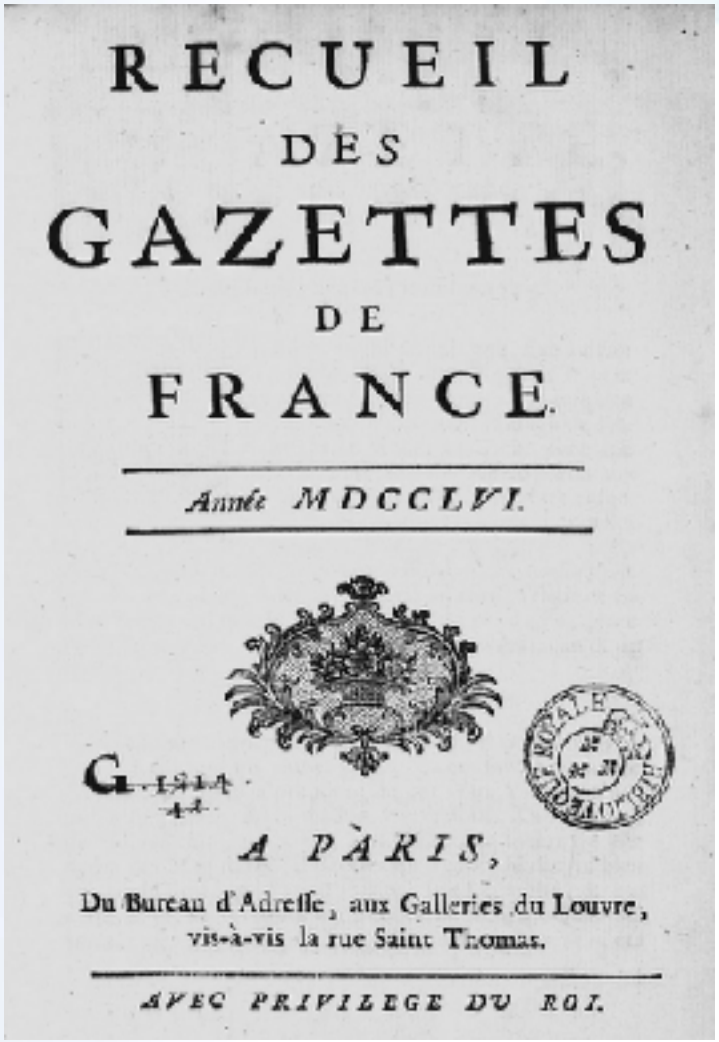
PRESENTATION STRUCTURE

Four parts

- 18th-century French print corpora
- The sins of Bad Data
- How to make Bad Data useful
- Some partial results

CORPORA TYPES — TRANSCRIBED

CORPORA TYPES — OCR-GENERATED



RECUEIL
DES
GAZETTES
DE
FRANCE.

Année MDCCLVI.

G 1914
44

A PARIS,

Du Bureau d'Adreſſe, aux Galleries du Louvre,
vis-à-vis la rue Saint Thomas.

AVEC PRIVILEGE DU ROI.

MERCURE
DE FRANCE
DÉDIÉ AU ROI.

NOVEMBRE 1745.

COLLIGIT UT SPARGAT

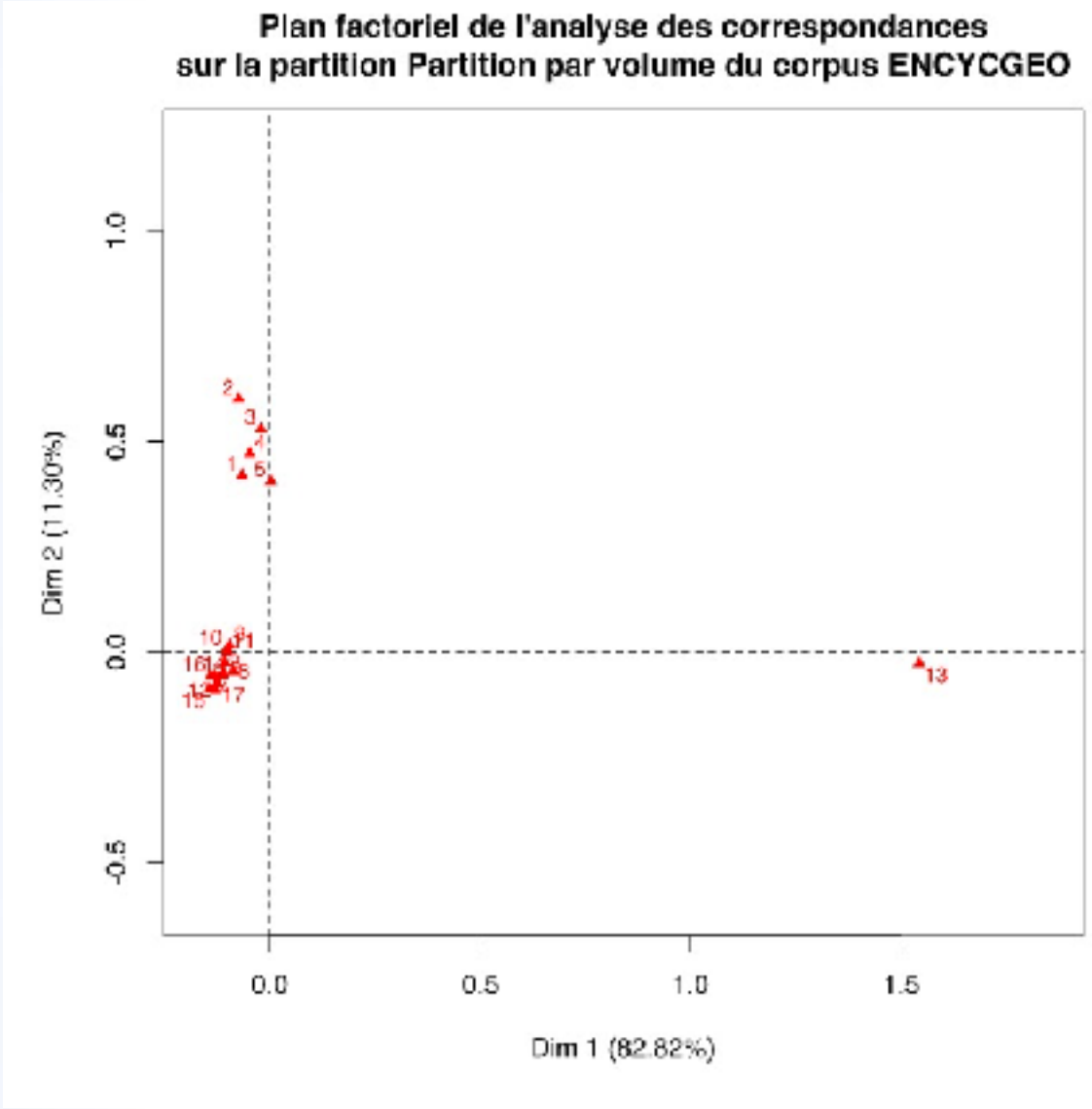
A PARIS,

Chés { GUILLAUME CAVELIER rue S. Jacques.
La Veuve PISSOT, Quai de Conty à la deſcente du Pont-Neuf.
JEAN DE NULLY, au Palais.

M. DCC. XLV.

Avec Approbation & Privilège du Roi

BAD DATA — HIDDEN OUTLIERS



BAD DATA — IRREGULAR SPELLING AND OCR ERRORS

Searching **Entire Database** for **LOU[YI]SIANE**

Your search found **53** occurrences

1. AMERIQUE, ou le Nouveau-monde, ou les Indes occidentales

en *septentrionale* & en *méridionale* par le golfe de Mexique & par l

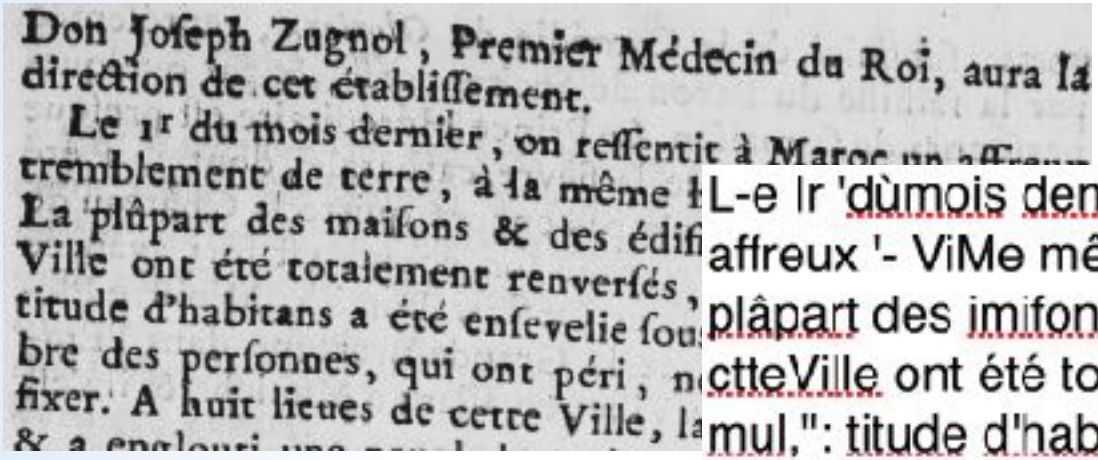
Loüisiane, la Virginie, le Canada, Terre- neuve

Pérou, le Paraguai, le

1. LOUYSIANE, la (page 9:707) [Géographie] Jaucourt

LOUYSIANE, la LOUYSIANE, la, (*Géog.*) grande contrée de l'

en dirai qu' un mot. Fernand de



Don Joſeph Zugnol, Premier Médecin du Roi, aura la
direction de cet établiſſement.

Le 1r du mois dernier, on reſſentit à Maroc un affreux
tremblement de terre, à la même h
La plûpart des maiſons & des édifi
Ville ont été totalement renverſés,
titude d'habitans a été enſevelie ſou
bre des perſonnes, qui ont péri, n
fixer. A huit lieues de cette Ville, la

L-e Ir 'dùmois demih, On reRèntic à Maroc un affreux '- ViMe même beure qu'en Efpa<>ne, La plâpart des imifons & des-édificies publics de cteVille ont été totalement renver * une grande mul.": titude d'habitans aéreenfevelie fous les ruines Le nom- PfS encore sebredes personnes, qui ne peut pas encore [e ftrer.

BAD DATA — ARTICLE ENDPOINTS

De Hanovre y le 20 Décembre 175 5»

:, On a rendu publics les deux Traités conclus cette année parle Roi , l'un avec l'Impératrice de Rume, l'autre avec le Landgrave de Hesse-Cassel. Le premier de ces Traités contient quinze Articles, **outre deux Articles réparés.'** **<— RECOGNIZABLE ENDING**

De RIltisGonne, le 19 Décembre 17 j S **<— RECOGNIZABLE BEGINNING**

LiCompagnie des Indes Orientales tint lé 6 de ce mois (u>L'Cof1:1pagme des In de ce tüiitt une igèrrtblëe ^^jc-EUe»éW'pottr!P«fi<fctt léBârbii a- plade du Cornce démission. 1 ,Il, '(- ", :-"I iié Trois Vaisseaux, arrivés dans un Port de Nonreee- c W"en«M ande dans lés-, d O&obre un violeur tremblement d.e terre. • - - '* - v "j ec

- \, J.,)' ,j .:,-\t¡,",,;- {t(.,j 'on I I i i ~, <— UNRECOGNIZABLE ENDING

e yiwMc y le 13 Décembre 17 5 y. <— NEXT ARTICLE BEGINS

BAD DATA — ERROR CLUSTERS

De Madrid, le 16 Décembre. 17 5 5 -

De Yeifailles) le ir Janvier 17^6.

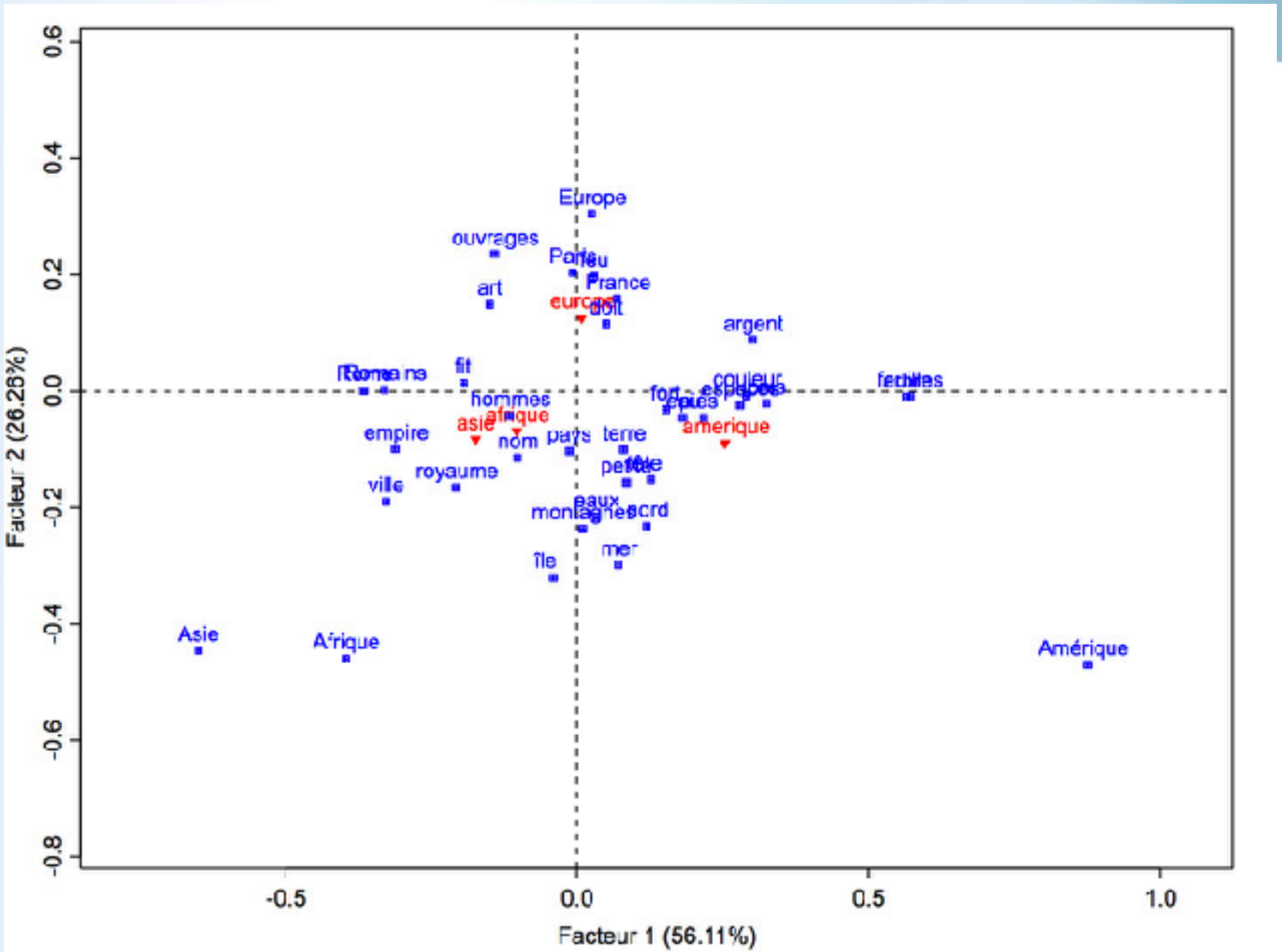
, 1 a.r * - 2 le 3 Janvier 17 j 6^

De Tœpl.litr., le 9 Décembre 17.5.J.l

De, ÇfJppellhgNe le 18 Décembre 17 t t.

De Vienne, le i o Décembre 17 5 y.

METHOD — HEURISTIC DESIGN



METHOD — HEURISTIC DESIGN

- Enhanced focus
- Filters out (much of) the noise
- Small keyword sets
- Manageable projects

LEVENSHTEIN'S ALGORITHM

Keyword	Candidate	Levenshtein Distance	Operations
Amérique	Amérique	0	—
Amérique	Amrique	1	Delete « é »
Amérique	Cmévrique	2	A —> C Insert « v »
Amérique	Musique	3	Delete « A » é —> u r —> s

LEVENSHTEIN'S PERFORMANCE

SOME OF THE RECOVERED INSTANCES

Keyword	Keyword occurrences	Alternate forms	Alternate form occurrences	Alternate forms as % of all occs	Examples
Amérique/ d'Amérique/ l'Amérique	485	36	128	20,9 %	l'amerique (59), d'amcrique (2), ramërique
Brésil	5	10	159	97,0 %	bresil (143), bretîl, brcfil
Canada/ Canadiens	139	3	3	2,1 %	canada*, en.canada
Colonie(s)	411	15	16	3,7 %	5lonie, coioniej
Domingue/ Saint-Domingue	88	11	13	12,9 %	jjomingue, saintdomingue
TOTAL	**1867**	**103**	**532**	**22,2%%**	**Increase: 29,5%**

COMMON ERROR TYPES

Lopresti (2009)

- Phantom blank space, e.g., ‘can ada’
- Phantom punctuation, e.g., ‘am.érique’

Underwood

- Common splits and merges: m -> in, in -> m, etc.

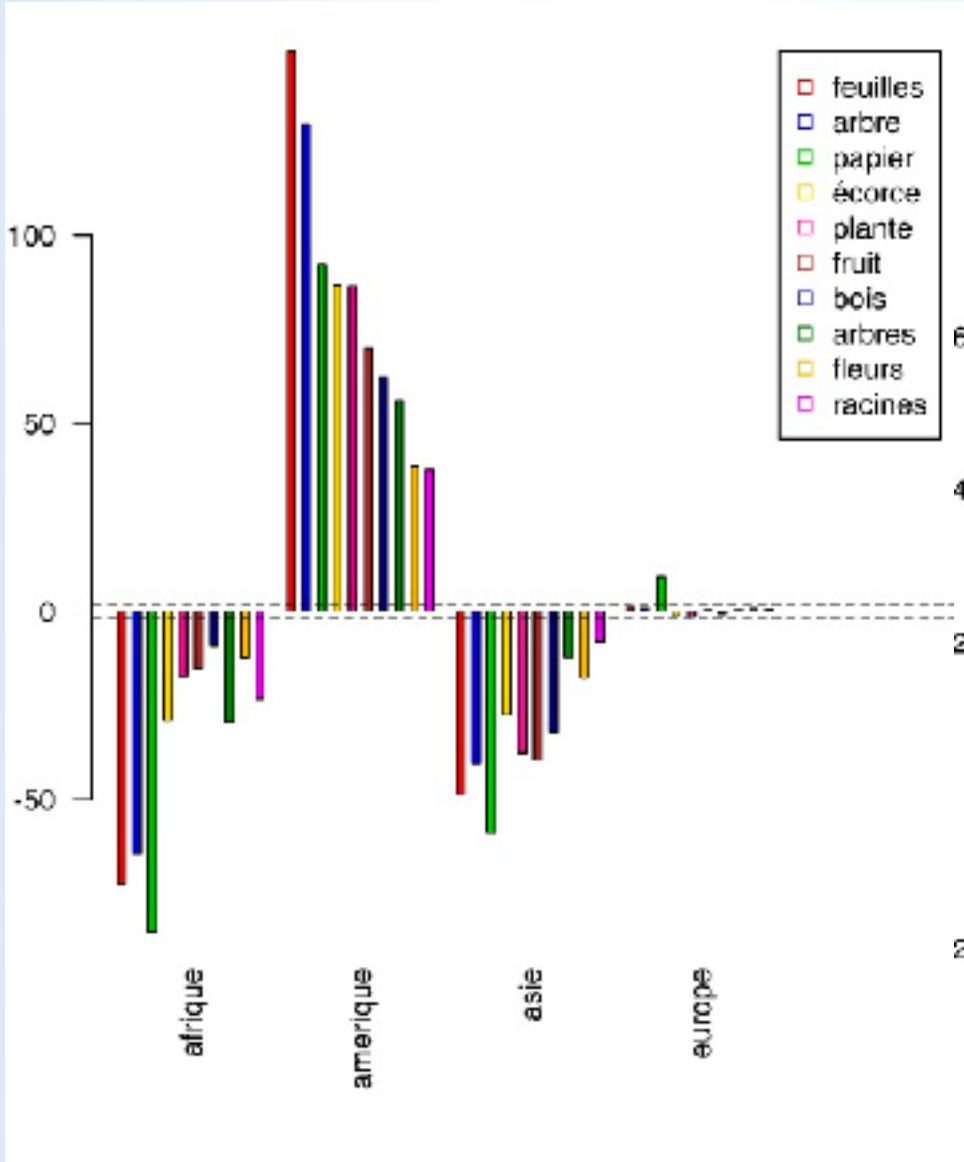
Ineffective with 18th-century French OCR:

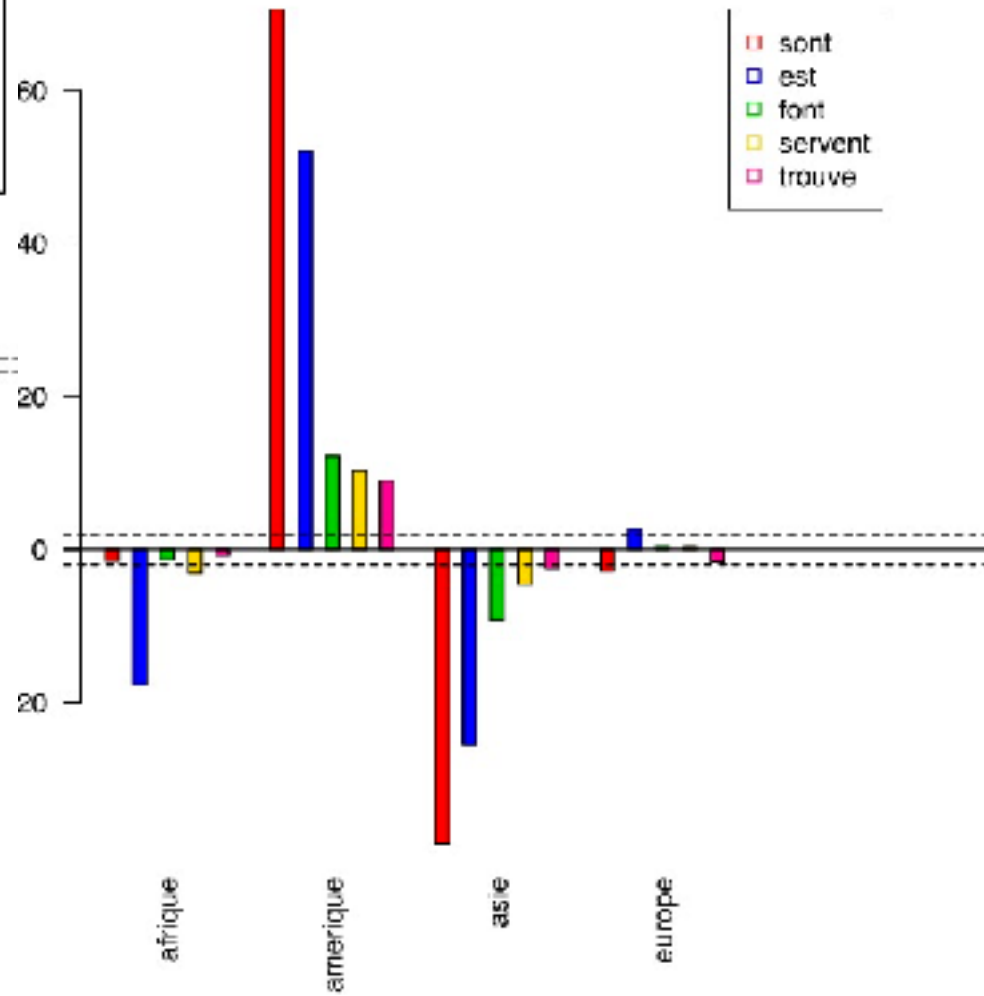
- Recovery: 322 tokens, 0.05% accuracy increase

METADATA FILES (EXTRACT)

id	"ville"	"annee_pub"	"amerique_yn"	"louisiane_yn"	"colonie_yn"	"canada_yn"
am1740_1	Londres	1740	1	0	0	0
am1740_10	Madrid	1740	1	0	0	0
am1740_11	Londres	1740	1	0	0	0
am1740_12	Londres	1740	1	0	0	0
am1740_13	Dresde	1740	1	0	0	0
am1740_14	Madrid	1740	1	0	0	0
am1740_15	Londres	1740	1	0	1	0
am1740_16	Londres	1740	1	0	0	0
am1740_17	Londres	1740	1	0	0	0
am1740_18	Madrid	1740	1	0	0	0
am1740_19	Londres	1740	0	0	1	0
am1740_2	Madrid	1740	1	0	1	0

RESULTS — AMERICA IN THE *ENCYCLOPÉDIE*





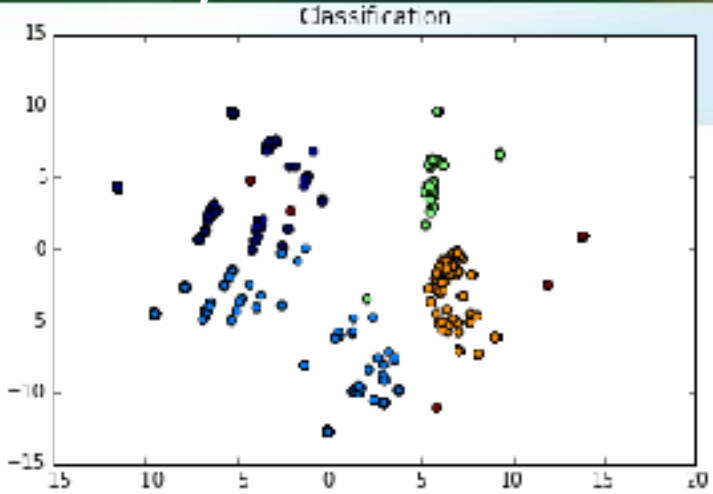
RESULTS — BIBLIOTHÈQUE BLEUE

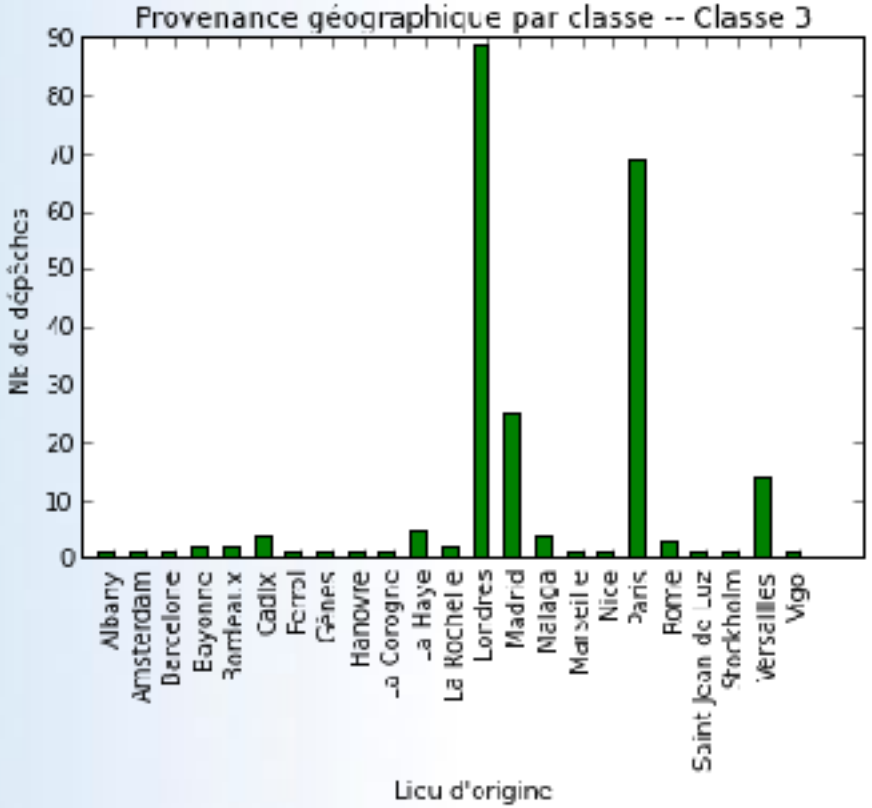
« … thus those who are born in Africa and in America, in places where it is hottest, are in good health, but strangers get sick there. »

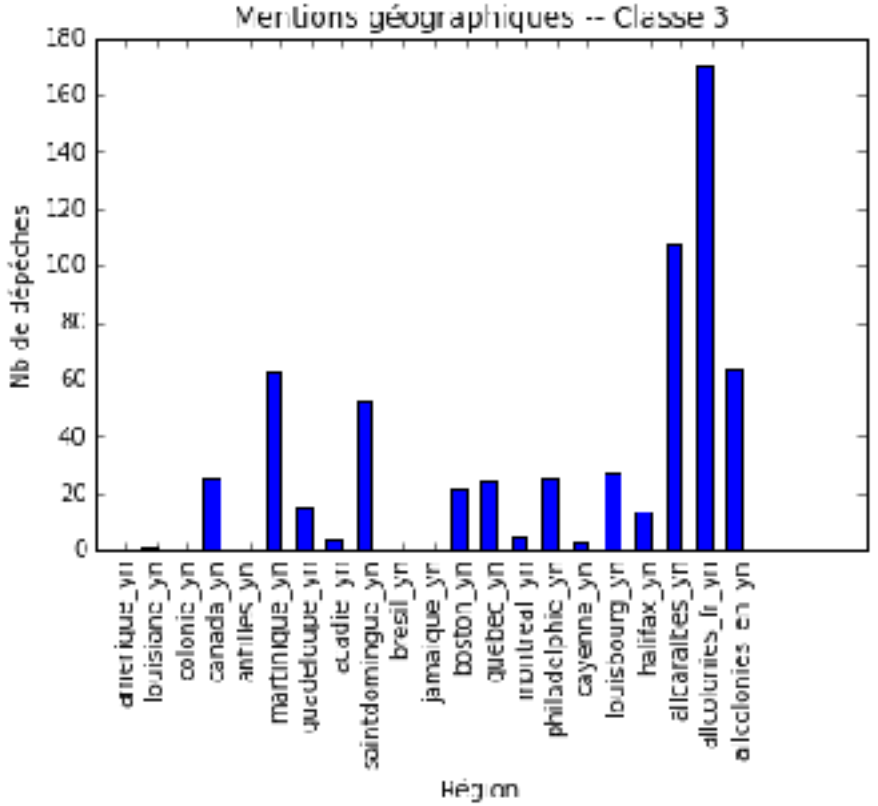
— *Instruction de la Jeunesse (« Instructing the Young »)*, 1782.

Source: Bibliothèque Bleue, ARTFL Project

RESULTS — GAZETTE, K-MEANS CLASSIFIER







CONCLUSIONS

Good science from Bad Data?

- Sleight of hand > technical wizardry
- Construct reliable metadata
- Distrust the text
- Limit the role of the digital

THANK YOU!

François Dominic Laramée
fdl@francoisdominiclaramee.com
www.francoisdominiclaramee.com
Twitter : @fdlaramee