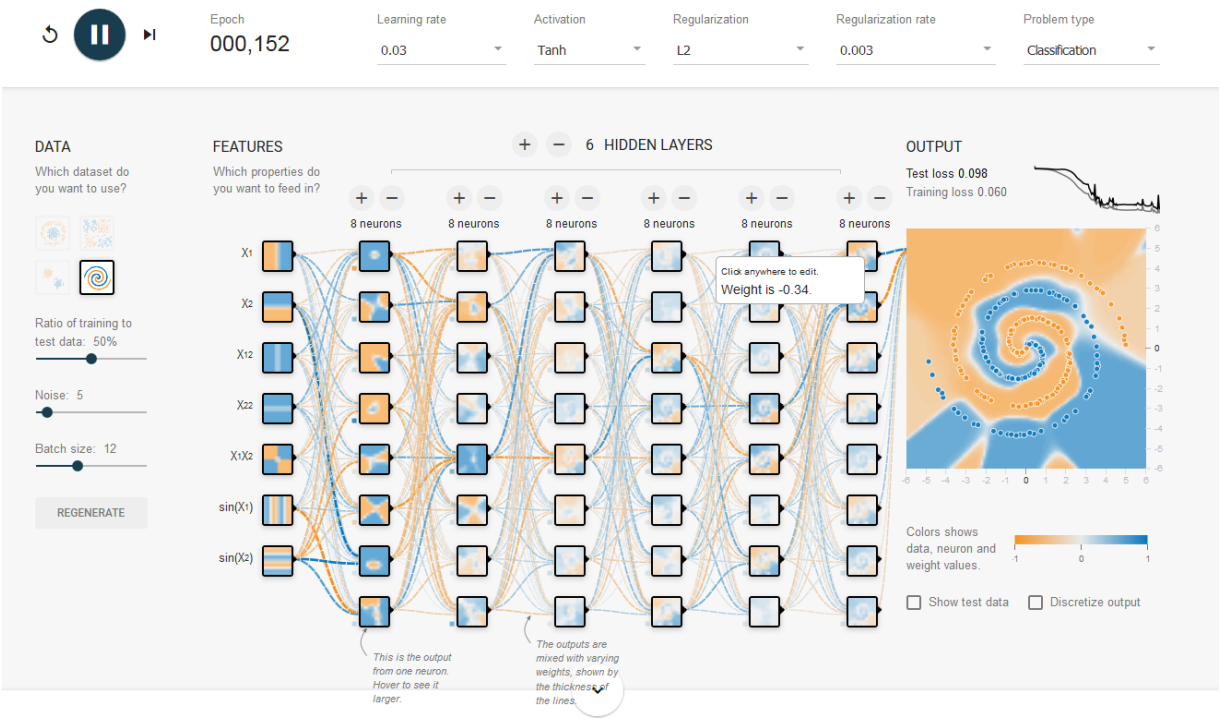
Wage Imputation With Deep Learning in the French Labor Force Survey

Damien Babet, Insee

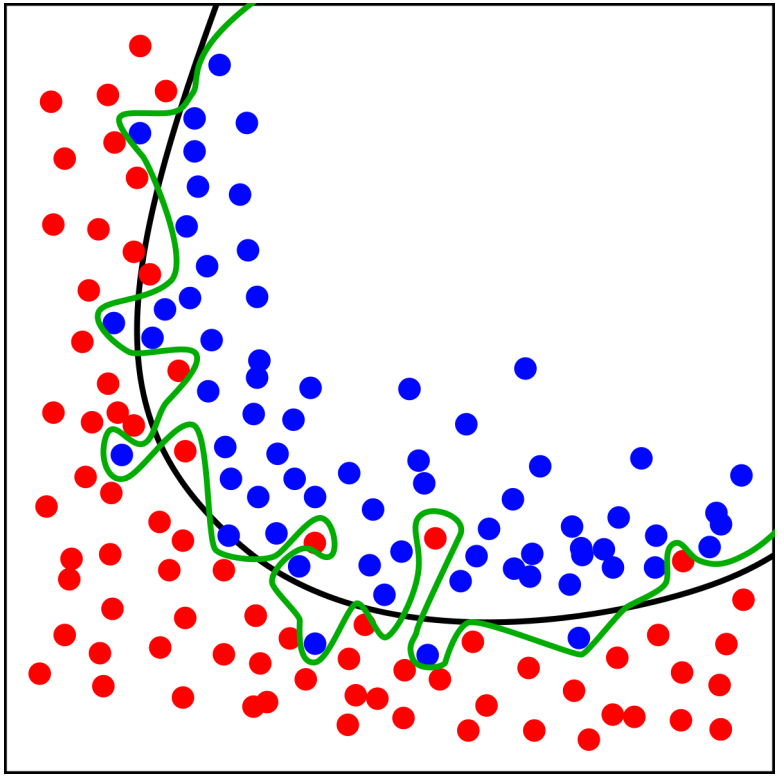
UNECE Conference of European Statisticians
Workshop on statistical data editing
31 August – 5 September 2020

Insee
Mesurer pour comprendre

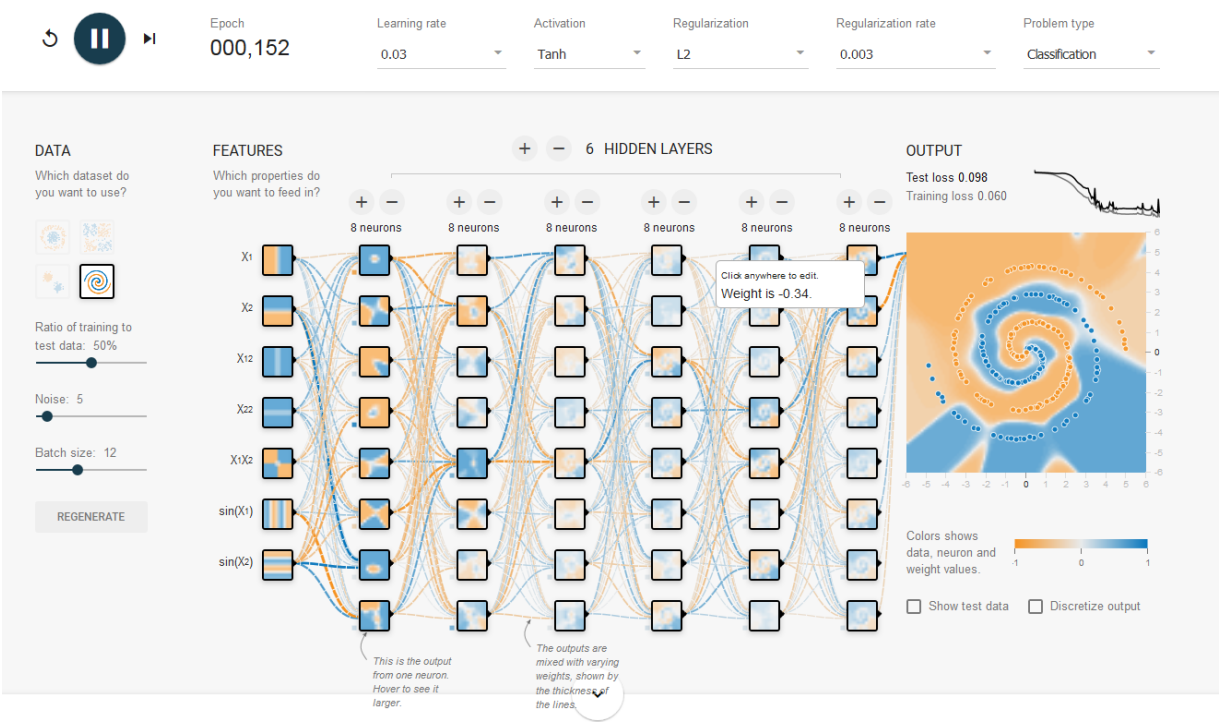
Is deep learning a good tool for imputation?



https://playground.tensorflow.org/



Is deep learning a good tool for imputation?



https://playground.tensorflow.org/

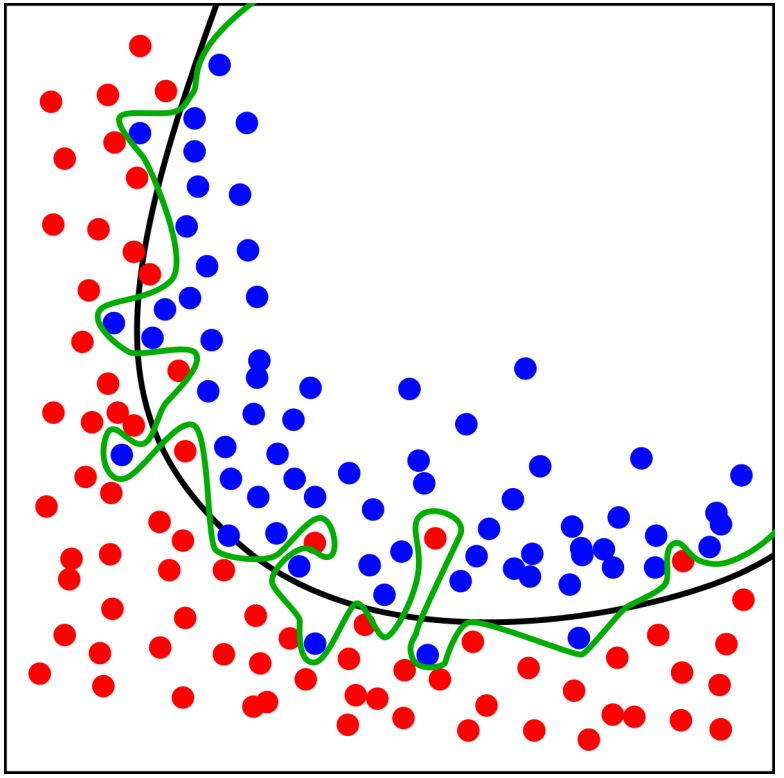
Black box?

- There might be solutions (but don't hold your breath)
- We don't necessarily need interpretability for imputation

Is deep learning a good tool for imputation?

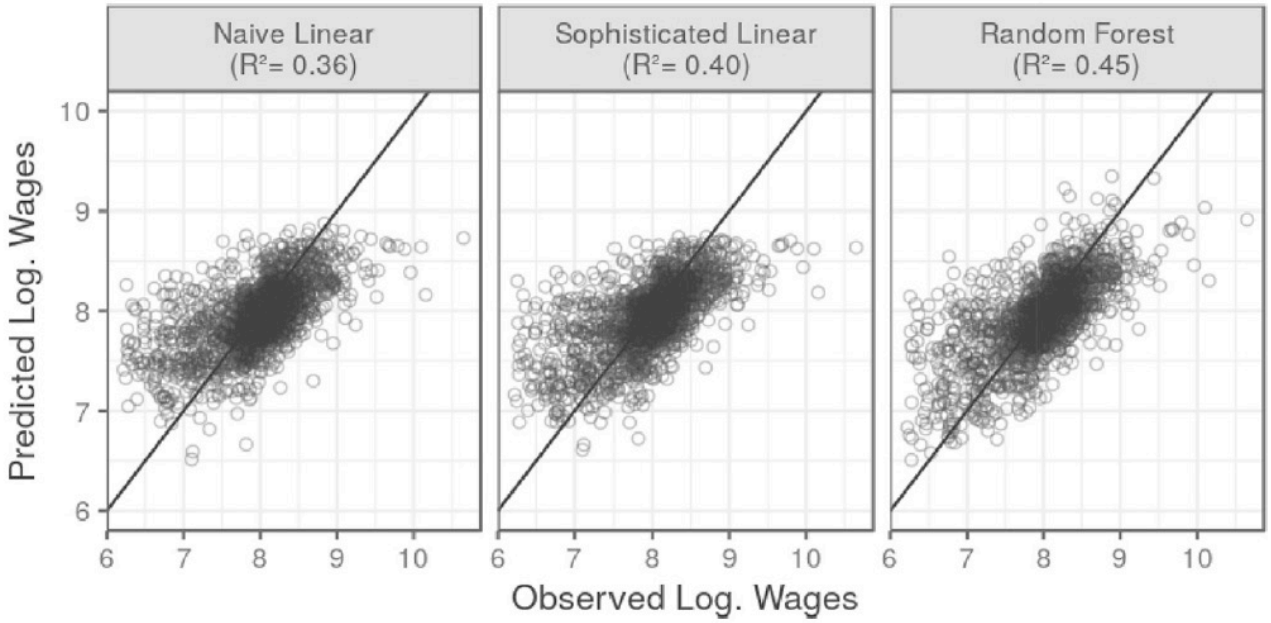
Overfitting?

- It is often assumed that DL needs loads of data. Not really true. Data is the best regularizer, not the only one. And DL seems to self-regularize
- Overfitting and underfitting diagnostic is done through validation and test set. A good methodology anyway



Are wages a good match for deep learning ?

FIGURE 1. – *Diagnostics plots, observed vs. predicted wages*



Boelaert, Julien, and Étienne Ollion. "The Great Regression." *Revue française de sociologie* 59.3 (2018): 475-506.

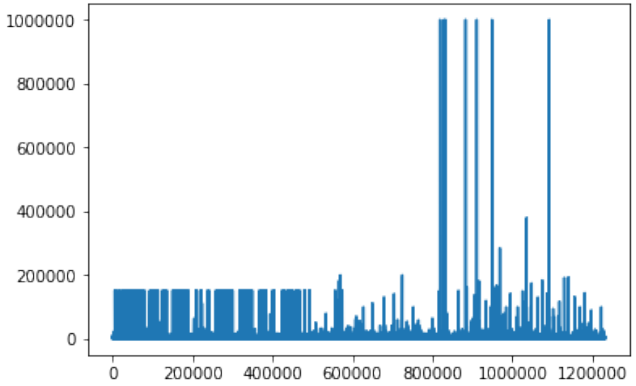
Well studied problem with a good traditional model: the Mincer equation

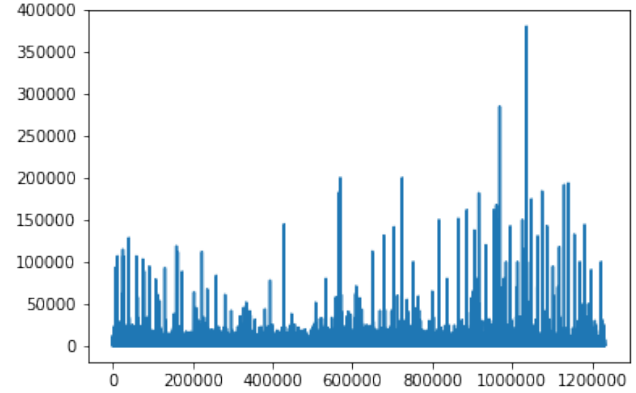
Still lots of unexplained variation

Interactions and non-linearities

Good POC opportunity

Data : French LFS Wage questions





Pour les salaries ou salaries chefs d'entreprise : STCR=2, 3

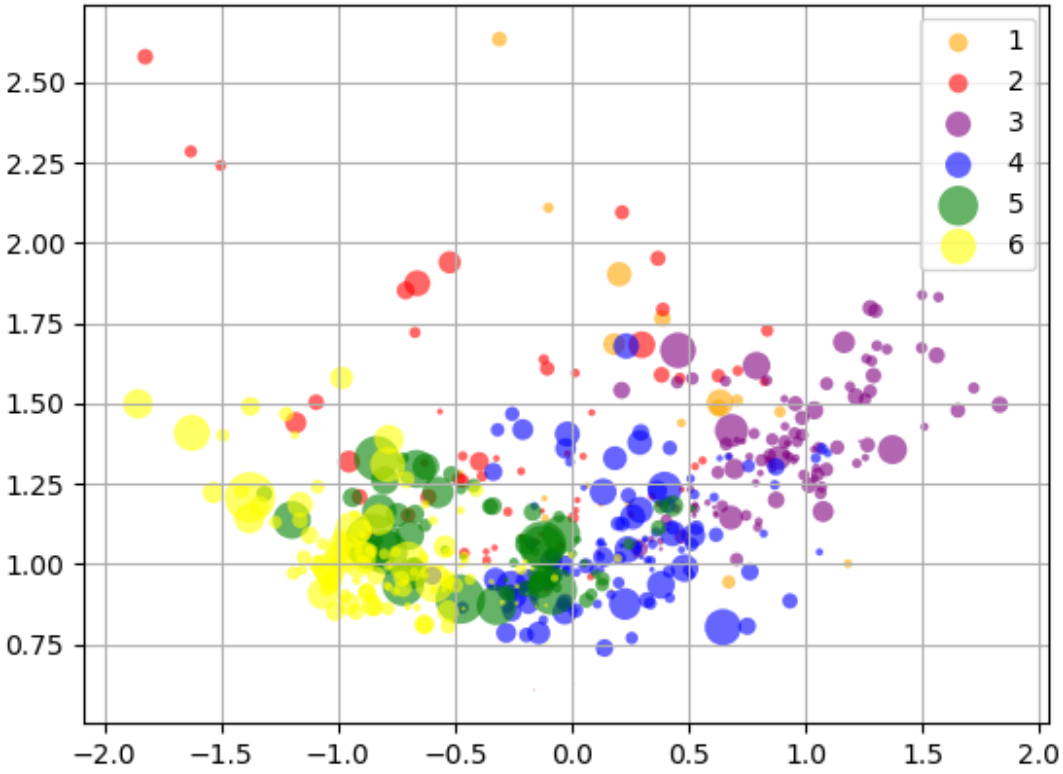
BD1 (Si NBTEMPR=3) **Quelle remuneration totale mensuelle retirez-vous** (si UNEPROF≠1) **de votre profession principale** (si UNEPROF=1) **de votre profession ?** (si NBTEMPR≠3) **quelle remuneration totale mensuelle retirez-vous de votre emploi** (si PLRACTR=1) **principal de¨*profession* pour *etablissement*** (sinon) **de *profession* pour *etablissement* ?**
SALMEE

Pour les salaries ou salaries chefs d'entreprise STCR=2, 3 qui ne peuvent pas ou ne veulent pas repondre :
SALMEE=≪ ne sait pas ≫
ou SALMEE=≪ refus ≫

BD2 Si vous ne savez pas ou si vous n'avez pas de montant exact, pouvez-vous indiquer la tranche mensuelle ? ou : **seriez-vous d'accord pour indiquer la tranche mensuelle ?**
SALMET

BD3 S'agit-il d'une remuneration nette ou brute ? TYPSAL
BD4 Cette remuneration mensuelle inclut-elle des primes ou complements mensuels ? PRIM
BD5 Quel est le montant mensuel de ces primes ou complements mensuels ? VALPRIE
BD6 Touchez-vous d'autres primes ou complements mensuels au cours de l'annee ?PRIMS
BD7 Quel est le montant de ces autres paiements sur l'annee ? VALPRE

What to make of high cardinality categorical variables?



486 occupations projected on dimensions 1 and 4 of a 4-dimensions encoding

Professional groups:

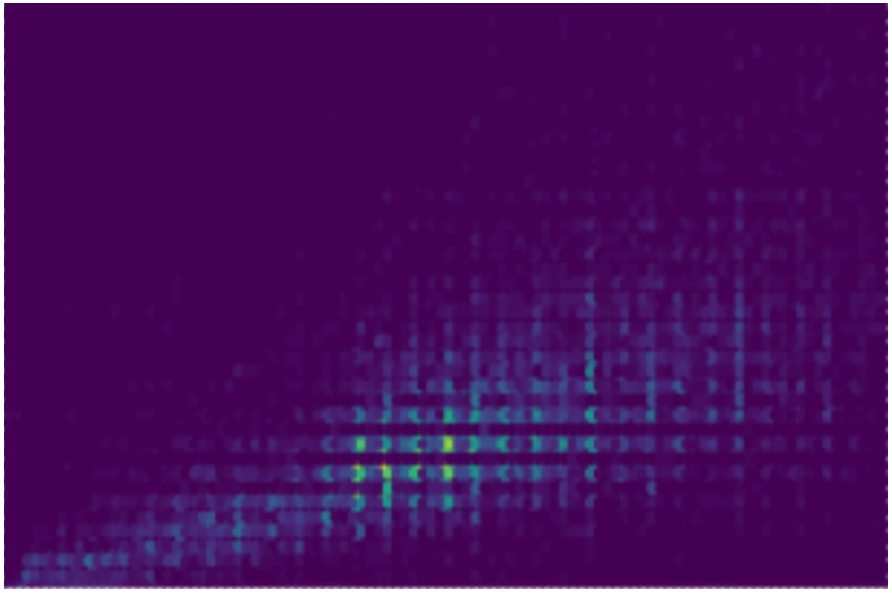
1. Farmers
2. Craftsmen, tradesmen, entrepreneurs
3. Managers and intellectual occupations
4. Intermediate occupations
5. Shop and service employees
6. Manual workers

Dimensions have no explicit meaning, the two dimensions shown are an arbitrary choice.

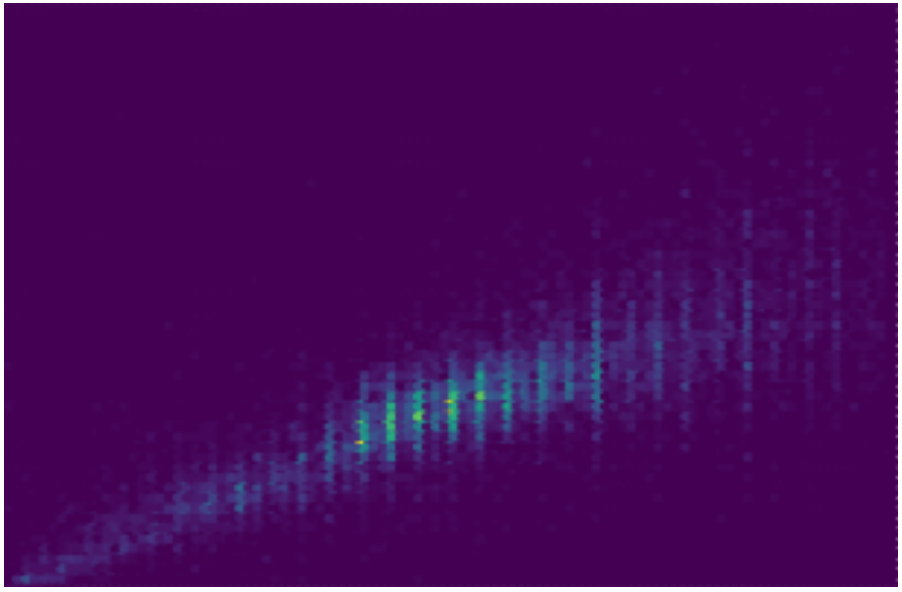
Area is proportional to the number of observations for each occupation

Results: improvement on benchmark

True versus predicted wage, in-production imputation model and deep learning model:



R2 Benchmark: 0.185



R2 DL: 0.381

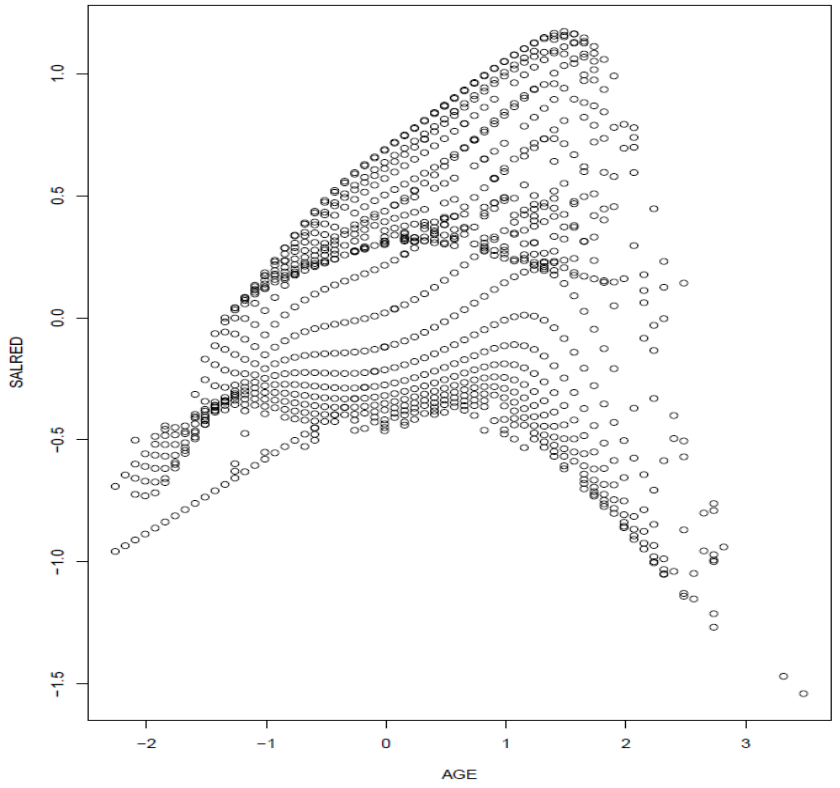
Results: better than Mincer

		Train set, log wage	Test set: Q1 2018		Test set: people born in march	
			Log wage	wage	Log wage	wage
Classic models	SALRED with no outlier correction	///	0.344	0.103	///	///
	SALRED	///	0.432	0.186	///	///
	Mincer equation with SALRED variables	0.545	0.597	0.327	0.549	0.217
	Mincer equation with additional variables	0.548	0.603	0.320	0.552	0.218
DL models	Mincer equation with vectorised nomenclatures	0.557	0.608	0.333	0.561	0.228
	Deep Learning model		0.696	0.381	0.662	0.270

So… when will it be in production ?

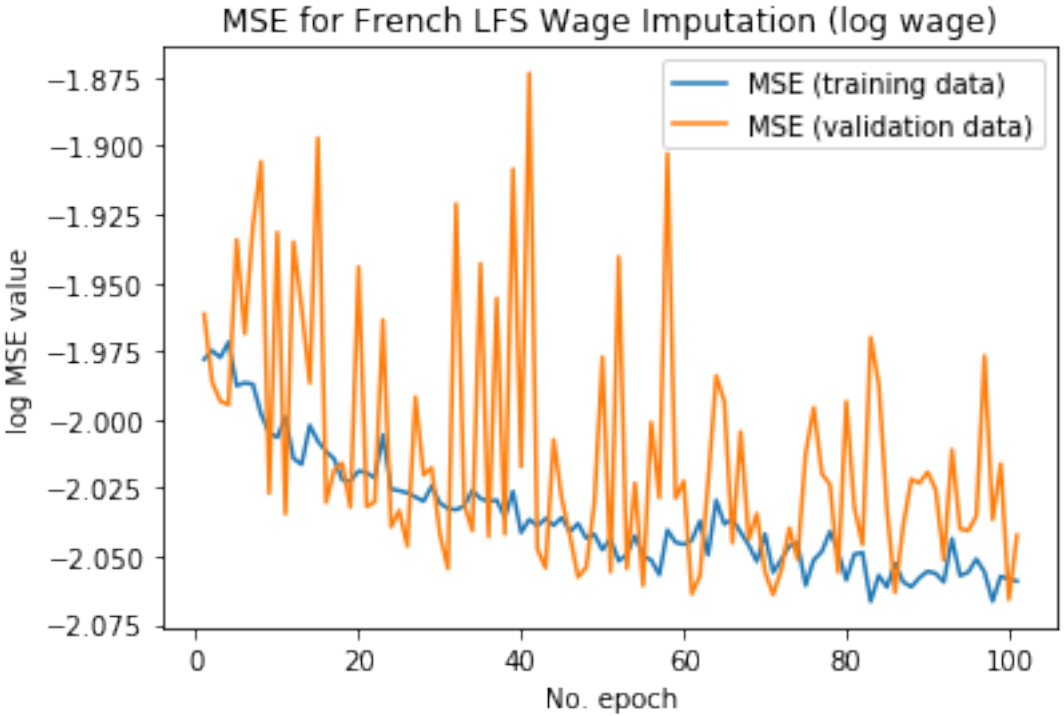
Several obstacles:

- Cultural
  - Fear of black boxes
  - Unease with train / test split
  - No knowledge of languages
- Computational
  - Python not readily available
  - Limited computational power
- Ease of use
  - Somewhat complex coding
  - Computation time
  - Randomness of performance
- Other priorities…



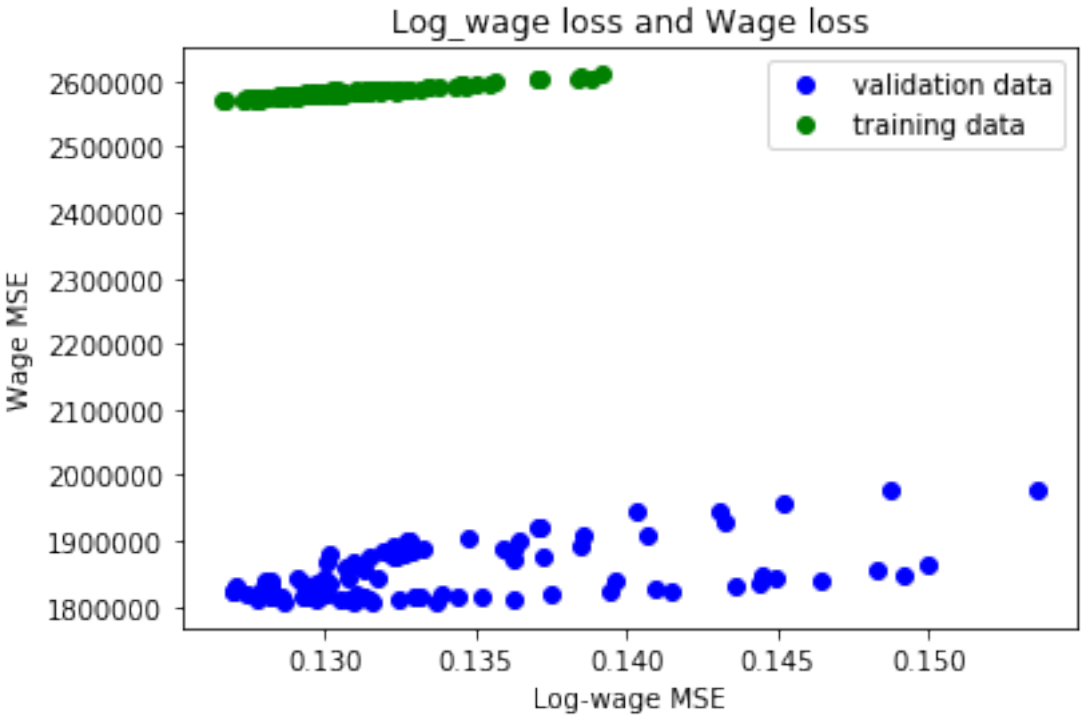
Neural networks with R: easy access at Insee, low quality results

Discussion



- Difficulty of hyper-parameters tuning
- Performance instability

Discussion



Should we optimize for wage or log-wage?

- Log-wage trained models are better predictors all over
- But relation between log-wage and wage performances another source of instability

Find us on:

insee.fr

Insee

Mesurer pour comprendre