

10 Més temes i més mètodes empírics en les ciències del llenguatge

Mètodes empírics 2

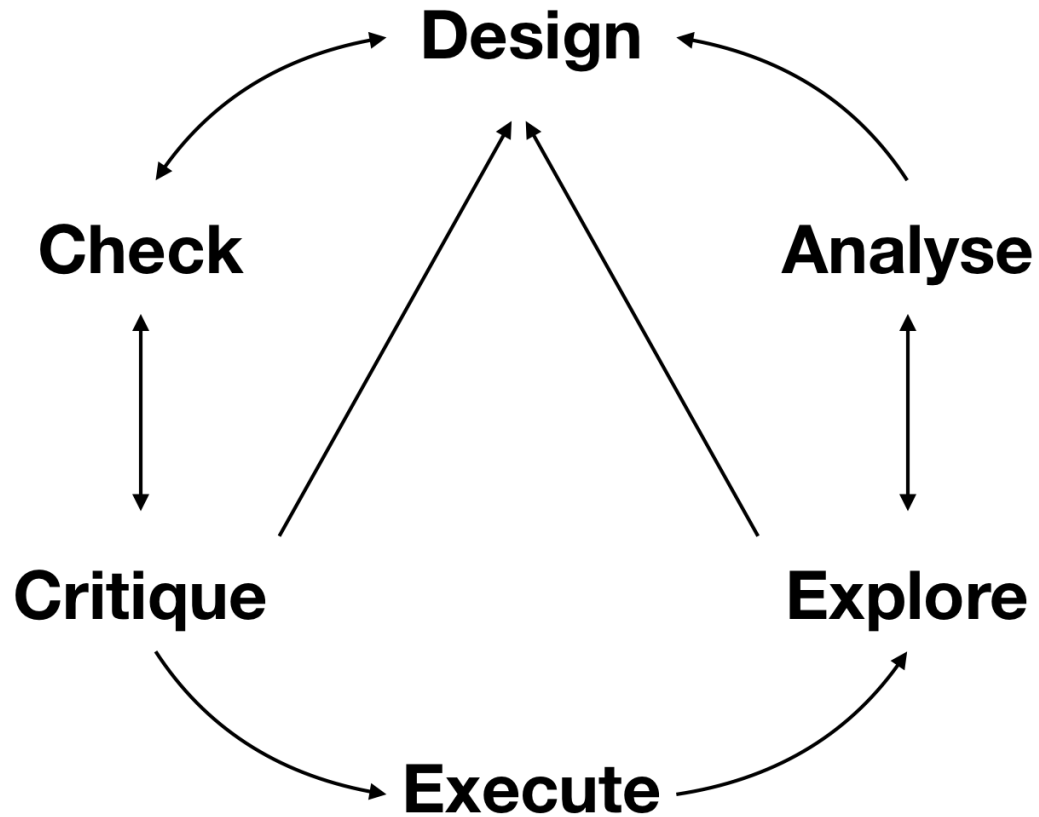
13/06/2023

Avui

- Recapitulació
- Més enllà: Mètodes empírics inferencials
- Més enllà: Ciències del llenguatge empíriques

Recapitulació

Cycle d'analyse



Anàlisi inferencial vs. descriptiu

- Inferència de propietats (més enllà de la mostra)
- Predicció
- Comparació
- Causa-efecte

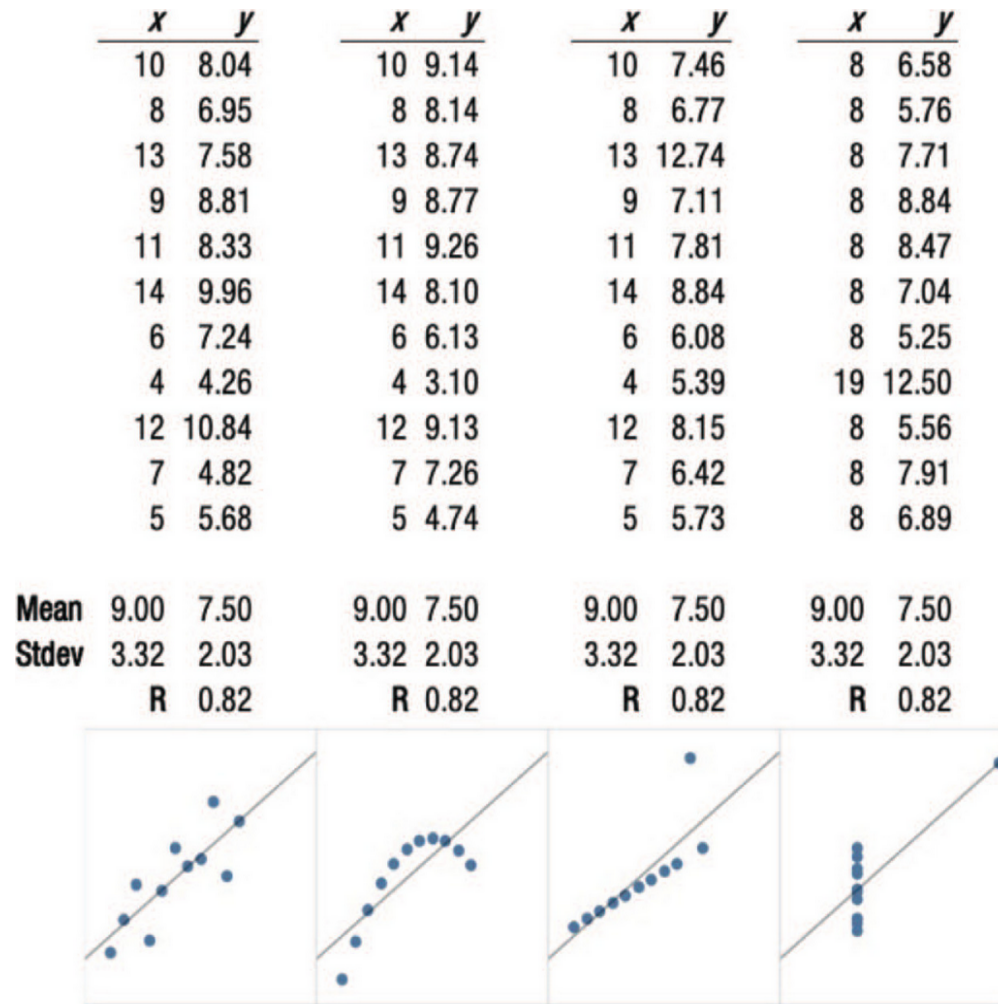


Fig. 1 de Franconeri et al. 2021 [The Science of Visual Data Communication: What Works](#)

Replicabilitat

Que es puguin obtenir resultats consistents amb les mateixes dades d'entrada; passos computacionals; mètodes; codi; i condicions d'anàlisi

Reproduïbilitat

Que es puguin obtenir resultats consistents en diferents anàlisis que busquen respondre la mateixa pregunta, cadascú amb les seves pròpies dades

Disseny d'anàlisi (components)

- Pregunta(es) de l'anàlisi
- Pla de disseny
- Pla de mostreig (sampling pla)
- Especificació de variables
- Pla d'anàlisi

Terminologia: Tipus de variables

- **Nominals**
- **Ordinals**
- **Binàries**
- **Booleans**
- **Mètriques**

Terminologia: Tipus de mostres

Mostra completa: tota la població d'interès

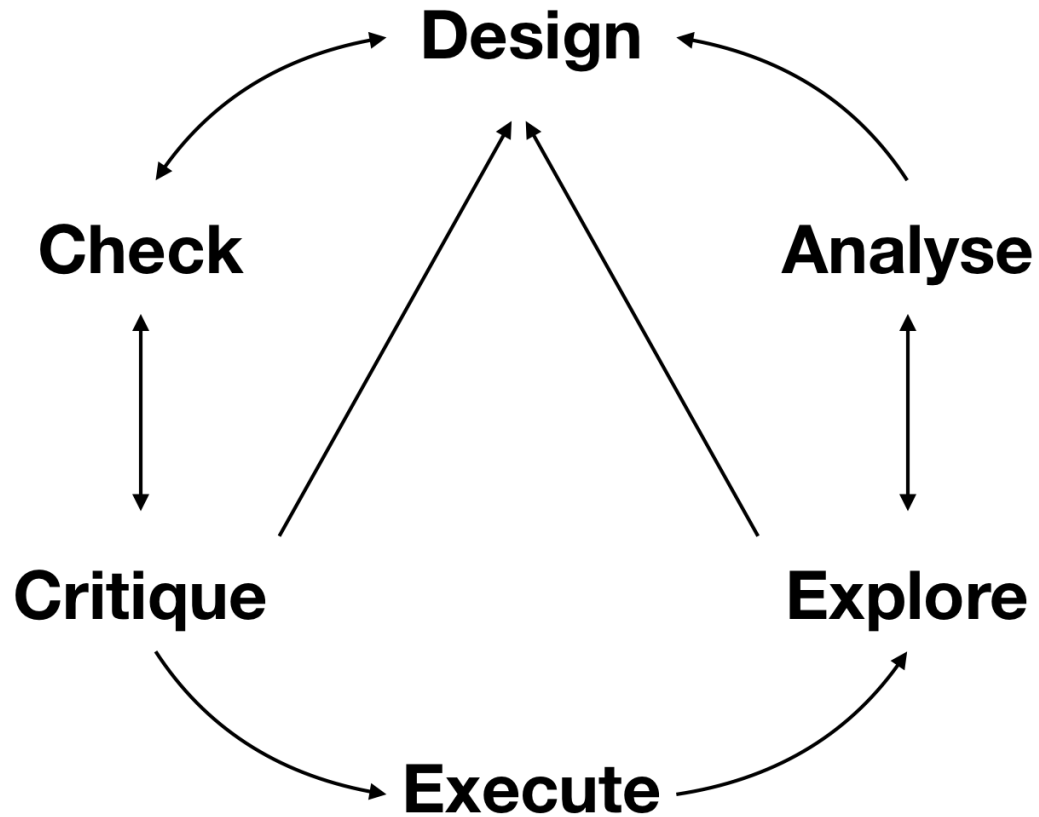
Mostra representativa/sense biaix: recollida de la mostra completa amb un mètode que no depèn de la mostra que s'està prenent

Mostra no representativa/amb biaix: les dades són influenciades pel mètode de recollida

Terminologia: Distribuciones

- Gaussiana/Normal
- Poisson
- Bernoulli (Binomial)

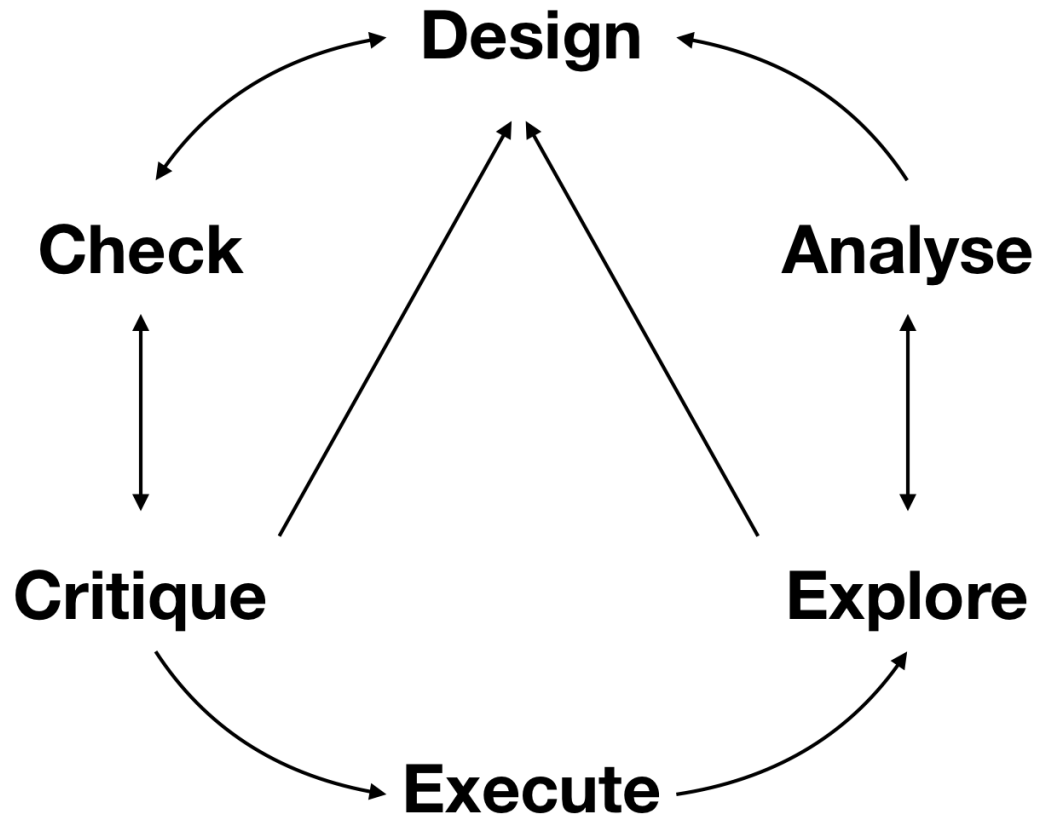
Cycle d'analyse



Control i crítica

- Estudis pilot
- Simulacions

Cycle d'analyse



Exploració

- (Manipulació)
- Estadística descriptiva
- Visualització

Anàlisi

- Regressió lineal generalitzada amb un o més predictors
 - Normal/Gaussiana
 - Poisson
 - Bernoulli
- Visualització
- Anàlisi de corpus

Fenòmens

- To
- Gestos
- Ambigüitat temporal
- Lleis de Zipf (laboratori i gran escala)
- ...

Kahoot!

www.kahoot.it

Més enllà: Mètodes empírics inferencials

Paramètrics

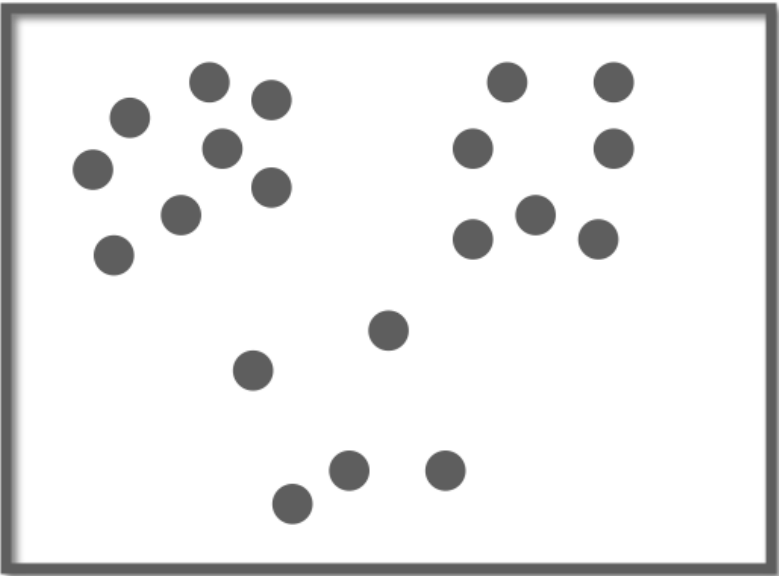
- Models lineals (heràrquics)
 - k-means
 - ...
-

No-paramètrics

- Models generalitzats additius
- Xarxes neuronals (en realitat no, però en termes pràctics sí)
- ...

K-means

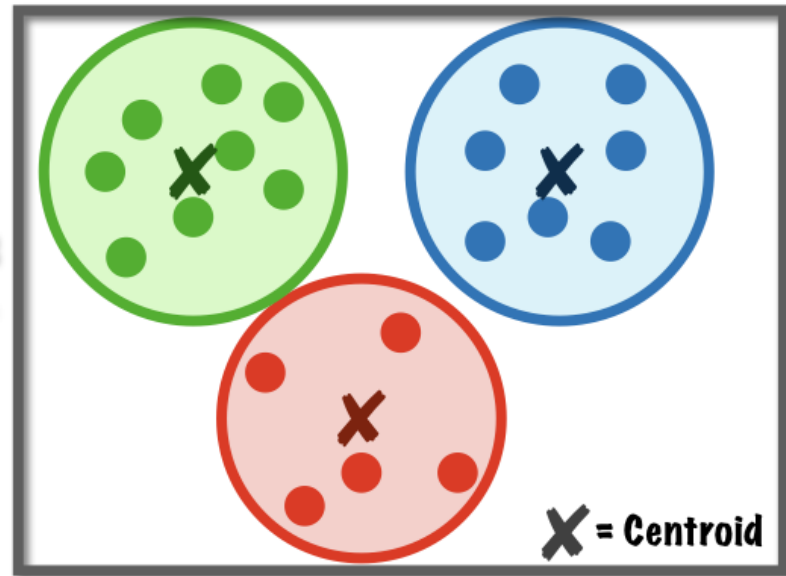
Unlabelled Data



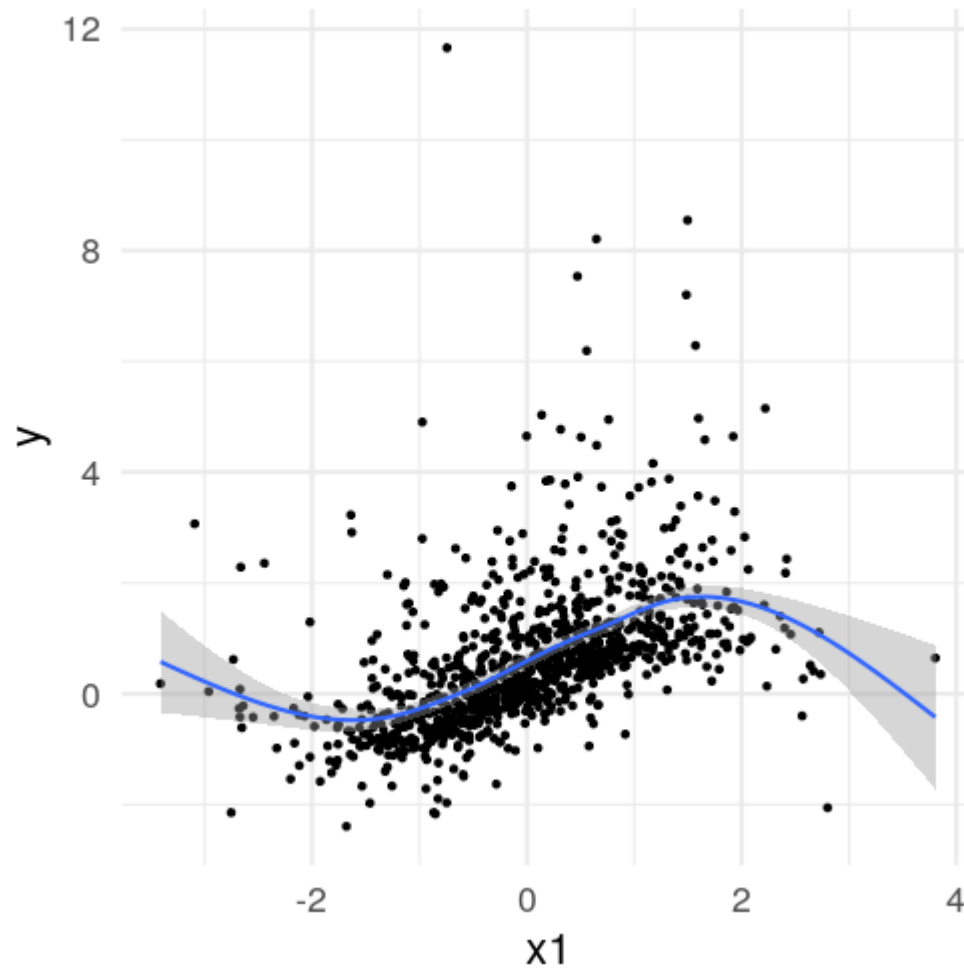
K-means



Labelled Clusters



Generalized Additive Models

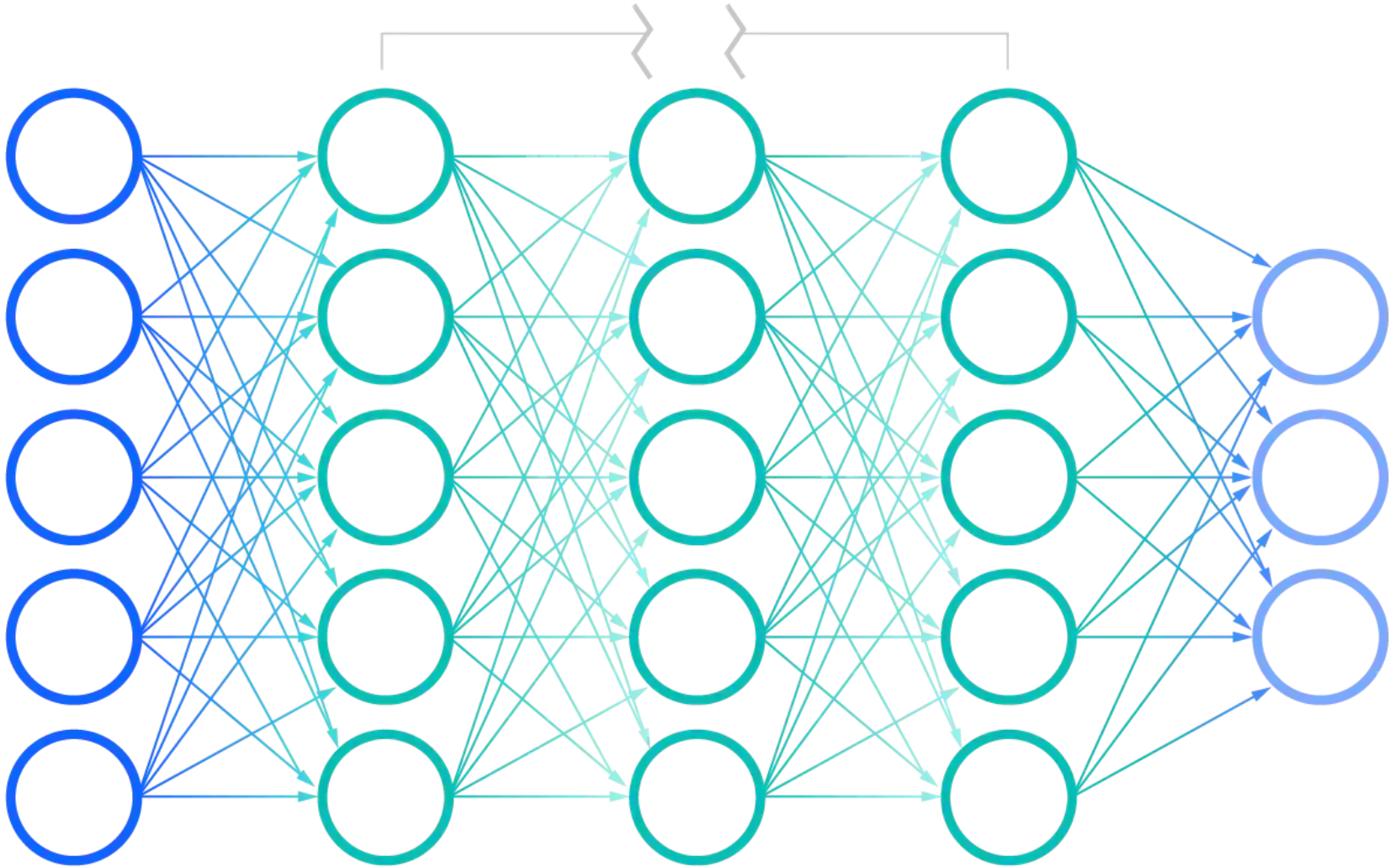


Deep neural network

Input layer

Multiple hidden layers

Output layer



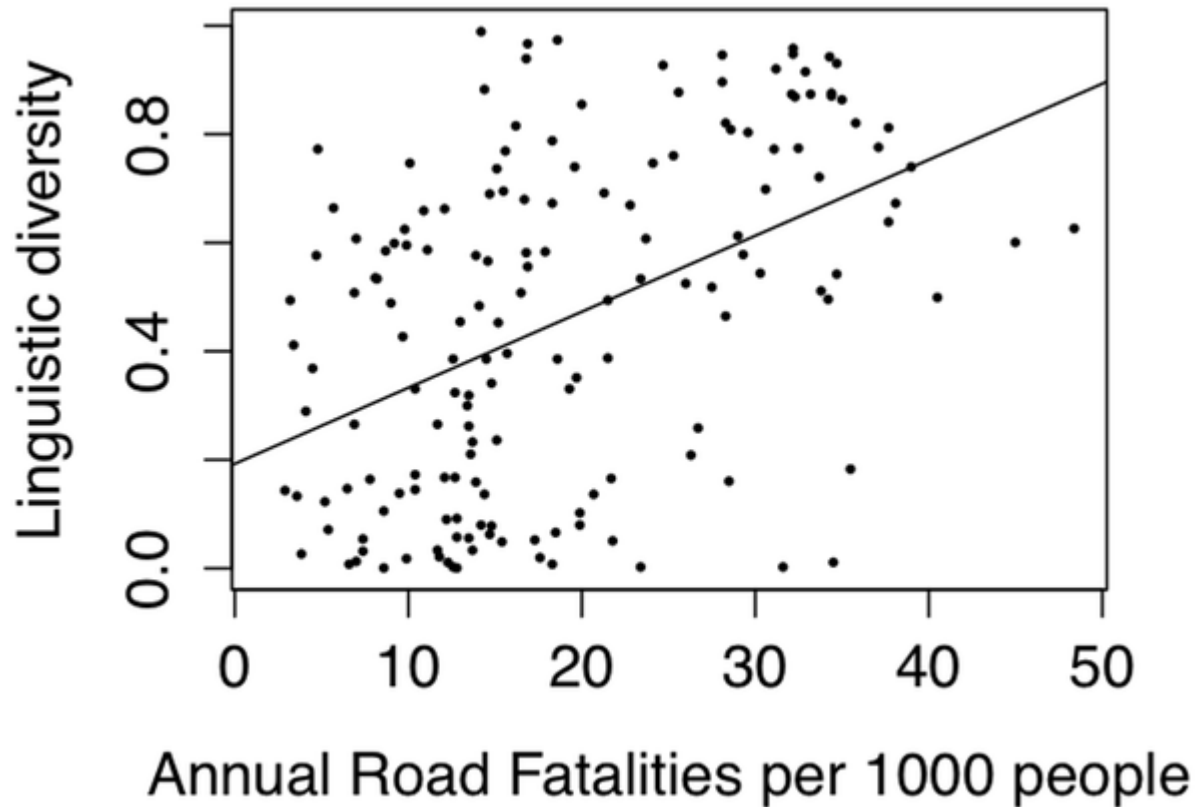
Tots aquests mètodes segueixen el mateix procés:

- Defineix objectiu ("funció de pèrdua")
- "Aprèn" de les dades
- Troba el paràmetre, o combinació de paràmetres, que minimitzen la pèrdua (maximitzen l'objectiu)

Tots aquests mètodes segueixen el mateix procés:

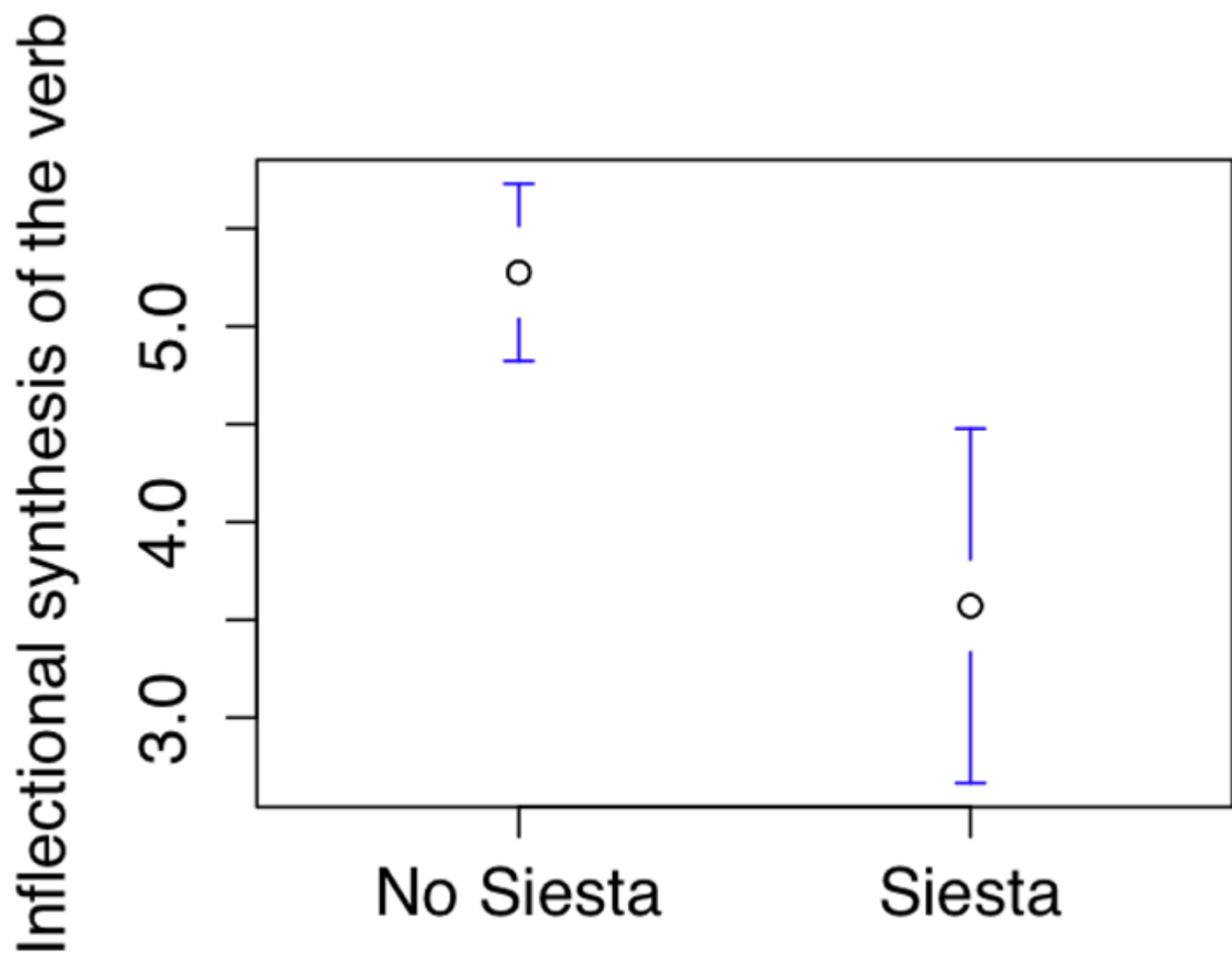
- Defineix objectiu ("funció de pèrdua")
- "Aprèn" dels **dades**
 - Qualitat de dades
 - Qualitat de (les suposicions) del model
- Troba el paràmetre, o combinació de paràmetres, que minimitzen la pèrdua (maximitzen l'objectiu)

-
- **Human in the Loop: Active learning and annotation for human-centered AI**
 - **Linguistic Diversity and Traffic Accidents: Lessons from Statistical Studies of Cultural Traits**



Linguistic Diversity and Traffic Accidents: Lessons from Statistical Studies of Cultural Traits

y = Linguistic diversity index: probabilitat que dues persones d'una població tinguin la mateixa llengua nativa



y = Mitjana de categories gramàtiques que pot prendre el verb

Més enllà: Ciències del llenguatge empíriques

De moment: Sortint lentament de la crisi de replicabilitat

"Noves" àrees de recerca/aplicació:

- Traducció assistida
- Tipologia computacional
- Human in the loop QA
- PLN
- Metodologies d'aprenentatge assistides/automatitzades
- ...

Is it on?

Preguntas?

Coda

- Avaldo
- Informe final: 25/06

Gràcies!

<https://brochhagen.github.io/>

thomas.brochhagen@upf.edu

52.609