10 Más temas y más métodos empiricos en las ciencias del lenguaje

Métodos empíricos 2

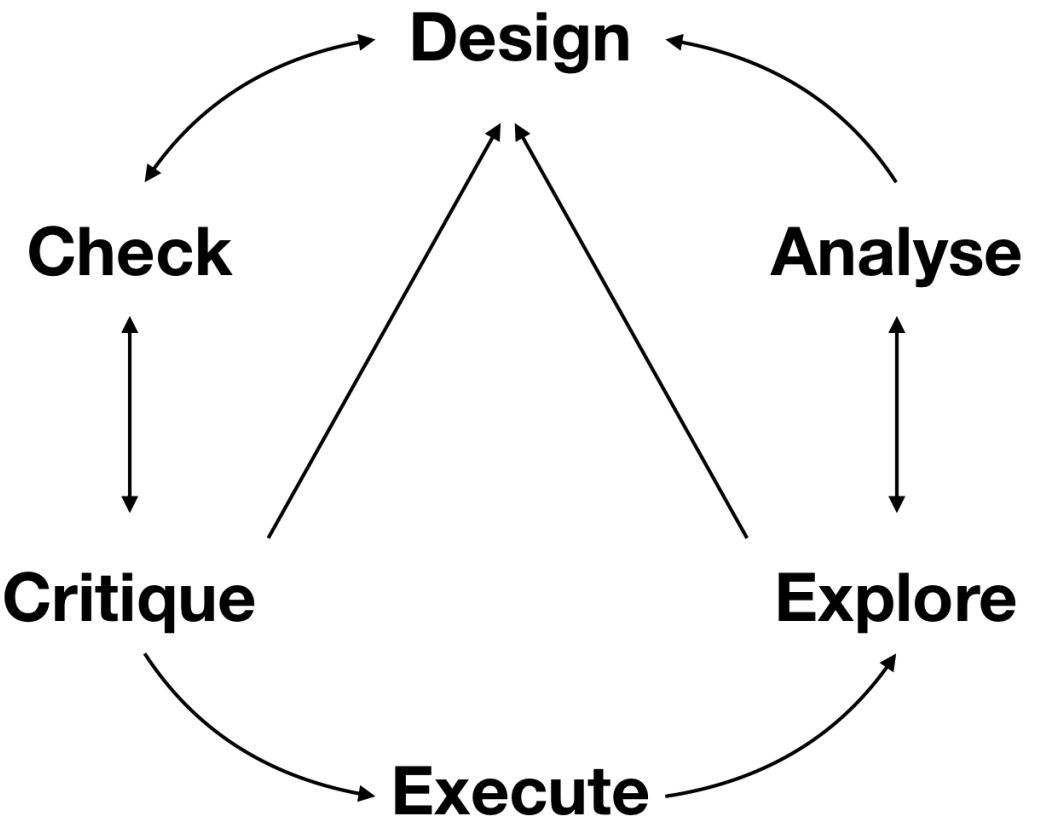
14/06/2022

Hoy

- Recapitulación
- Más allá: Métodos empiricos inferenciales
- Más allá: Ciencias del lenguaje empiricas

Recapitulación

Ciclo de análisis



Análisis inferencial vs. descriptivo

- Inferencia de propiedades (más allá de la muestra)
- Predicción
- Comparación
- Causa-efecto

x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89
Mean	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
Stdev	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03
R	0.82		**R**	0.82	**R**	0.82	**R**	0.82

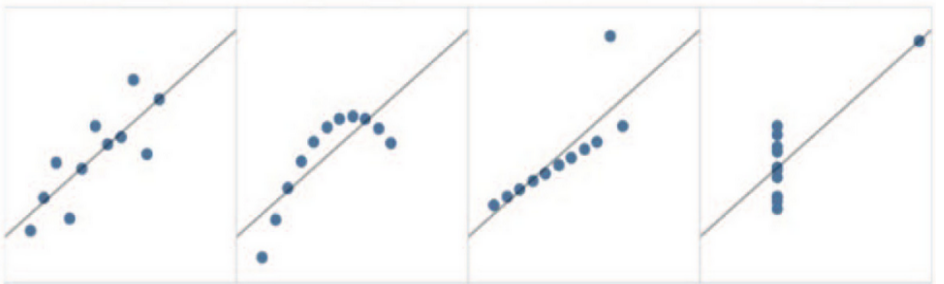


Fig. 1 de Franconeri et al. 2021 The Science of Visual Data Communication: What Works

Replicabilidad

Que se puedan obtener resultados consistentes con los mismo datos de entrada; pasos computacionales; métodos; código; y condiciones de análisis

---

Reproducibilidad

Que se puedan obtener resultados consistentes en diferentes análisis que buscan responder la misma pregunta, cada cual con sus propios datos

Diseño de análisis (componentes)

- Pregunta(s) del análisis
- Plan de diseño
- Plan de muestreo (sampling plan)
- Especificación de variables
- Plan de análisis

Terminología: Tipos de variables

- **Nominales**
- **Ordinales**
- **Binarias**
- **Booleanas**
- **Métricas**

Terminología: Tipos de muestras

Muestra completa: toda la población de interés

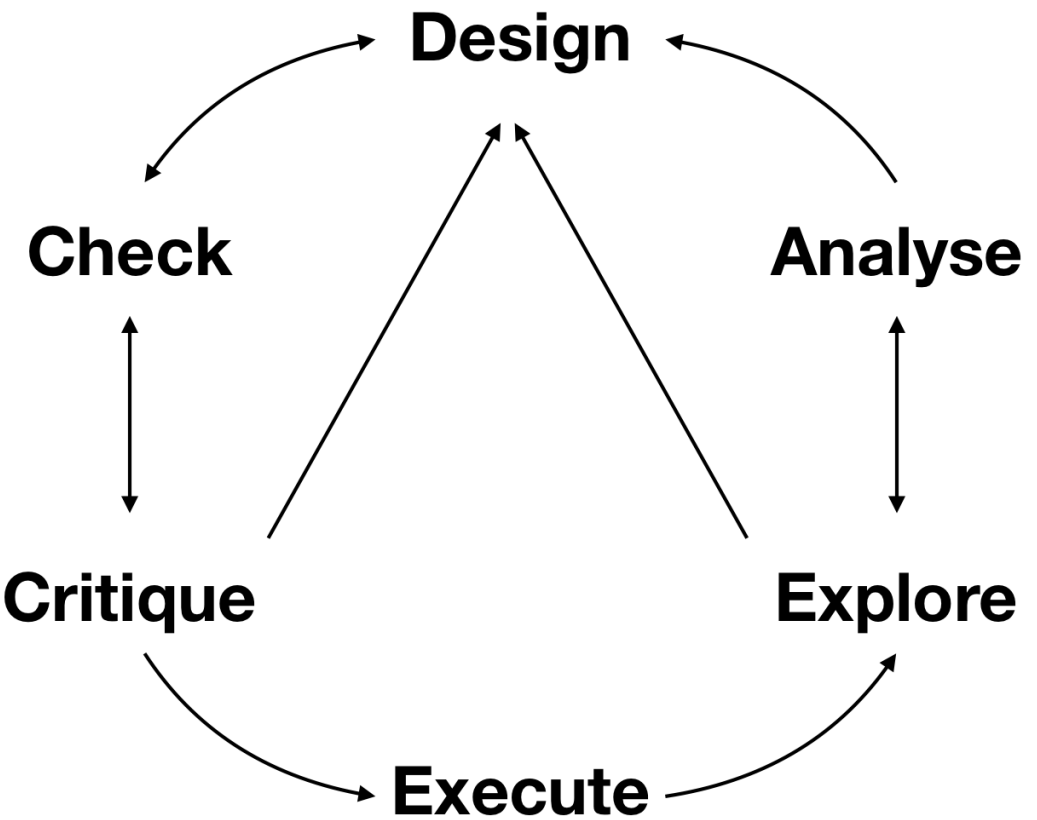
Muestra representativa/sin sesgo: tomada de la muestra completa con un método que no depende de la muestra que se está tomando

Muestra no representativa/con sesgo: los datos son influenciados por el método de toma

Terminología: Distribuciones

- Gaussiana/Normal
- Poisson
- Bernoulli (Binomial)

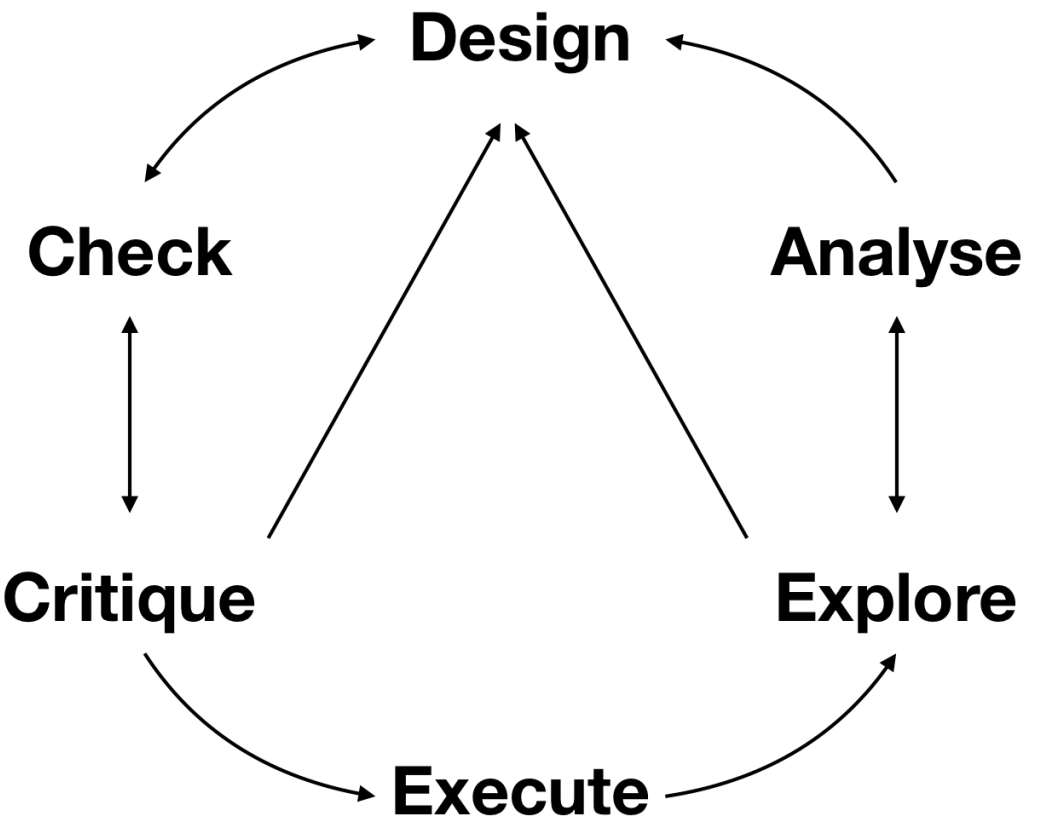
Ciclo de análisis



Control y crítica

- Estudios piloto
- Simulaciones

Ciclo de análisis



Exploración

- (Manipulación)
- Estadística descriptiva
- Visualización

Análisis

- Regresión lineal generalizada con uno o más predictores
  - Normal/Gaussiana
  - Poisson
  - Bernoulli
- Visualización
- Análisis de corpus

Fenómenos

- Tono
- Gestos
- Ambigüedad temporal
- Resolución de pronómbres
- Leyes de Zipf (laboratorio y gran escala)
- …

Kahoot!

www.kahoot.it

Más allá: Métodos empiricos inferenciales

Paramétricos

- Modelos lineales (hierárquicos)
- k-means
- ...

No-paramétricos

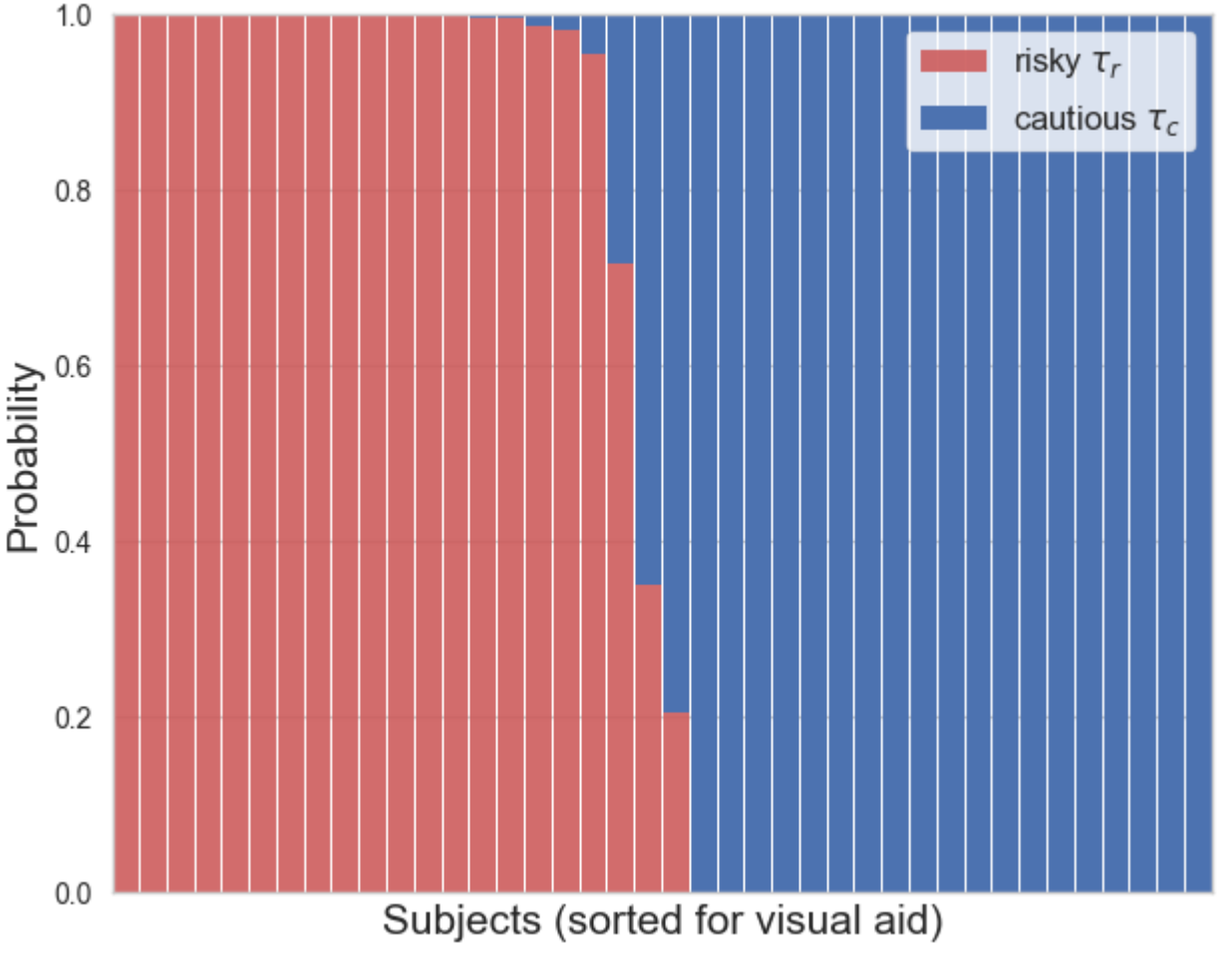
- Modelos generalizados aditivos
- Redes neuronales (en realidad no, pero en términos prácticos sí)
- ...

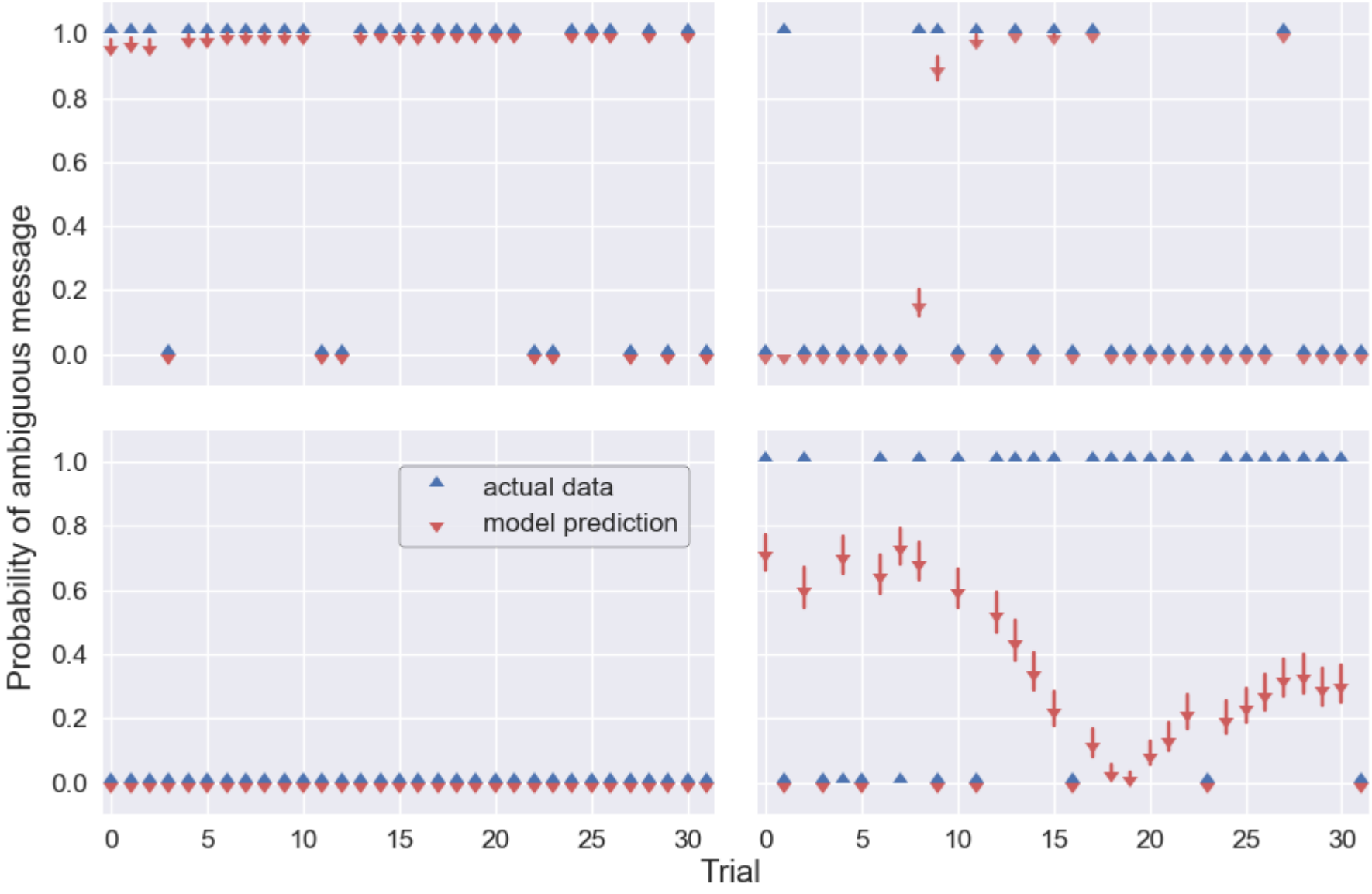
Modelo hierarquico para Kanwal et al. 2017

Hay dos tipos "latentes" de hablantes:

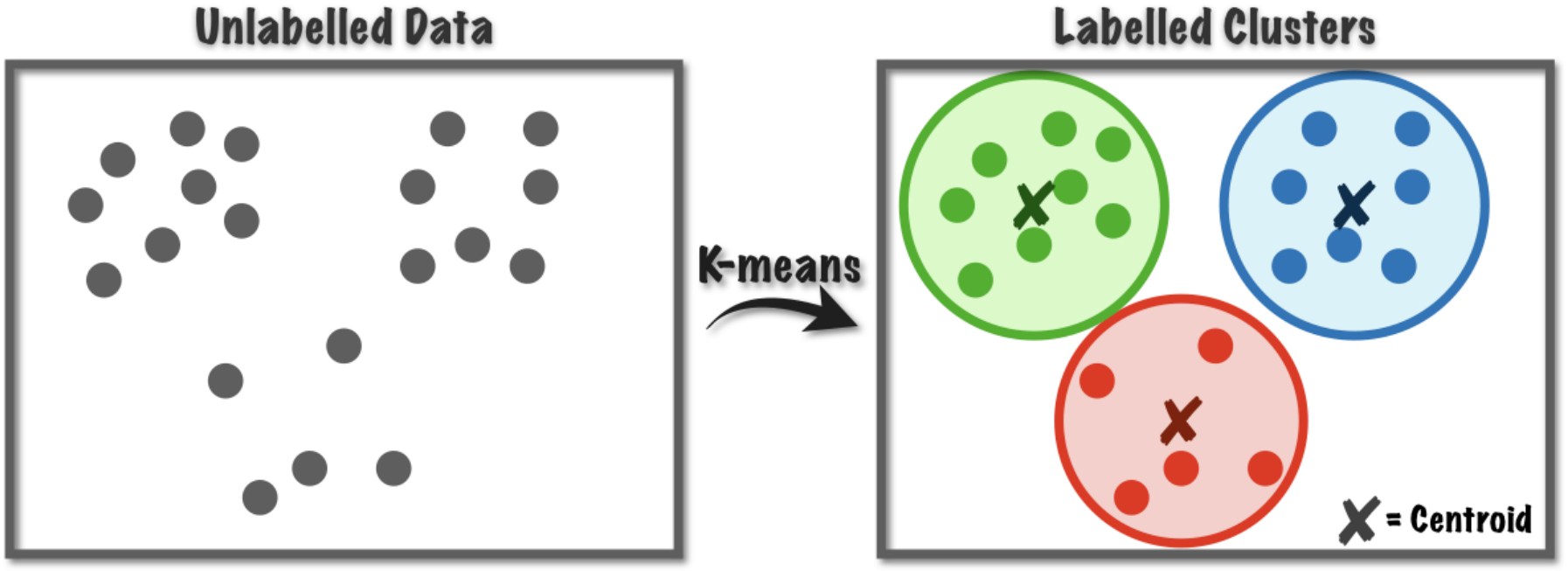
1. Arriesgados: Piensan que el interlocutor piensa que el significado frequente es más esperado
2. Cuidadosos: Tienen incertidumbre sobre cual es el significado esperado

Hablantes usan el mensaje (que consideran) más probable a ser entendido, minimizando longitud de mensaje

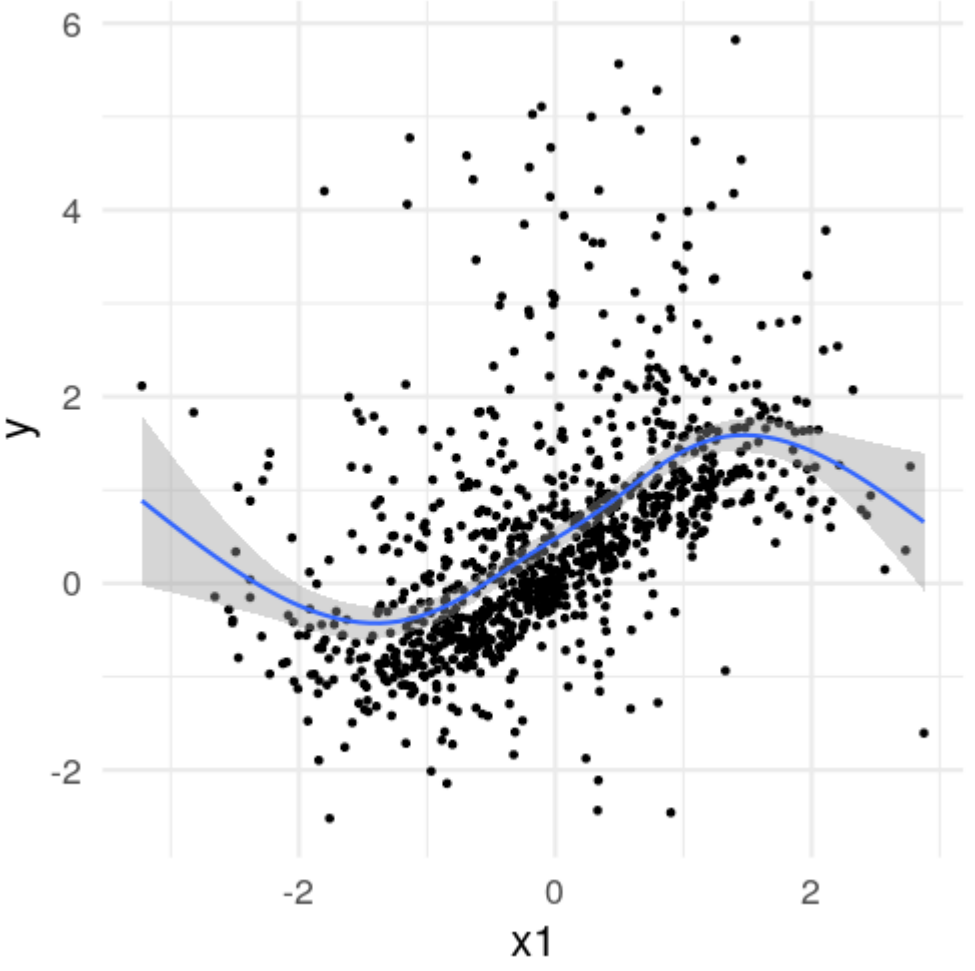

1.0
0.8
0.6
0.4
0.2
0.0
Probability
risky τ_r
cautious τ_c
Subjects (sorted for visual aid)


Probability of ambiguous message
1.0
0.8
0.6
0.4
0.2
0.0
actual data
model prediction
0
5
10
15
20
25
30
Trial

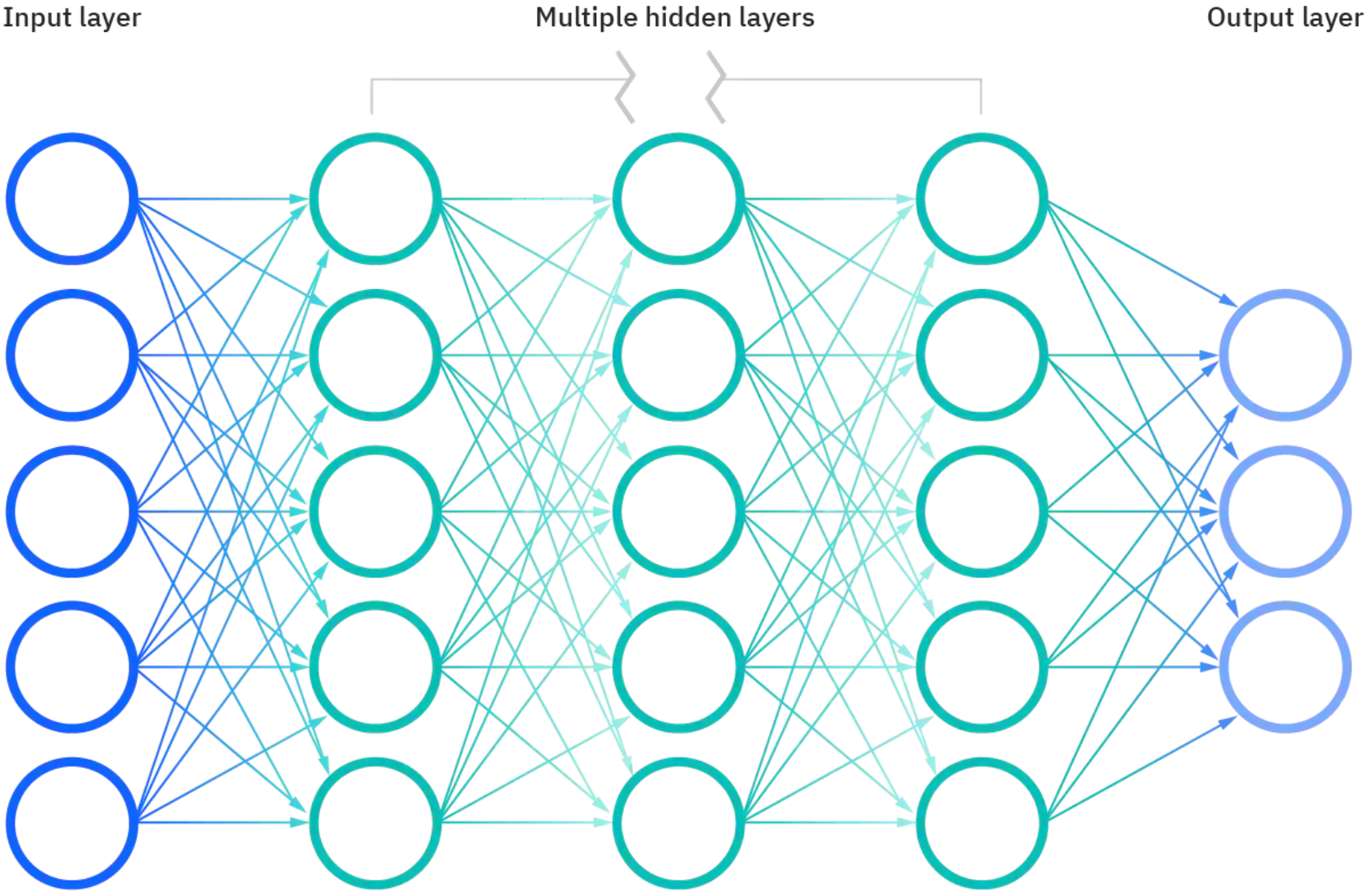
K-means



Generalized Additive Models



Deep neural network



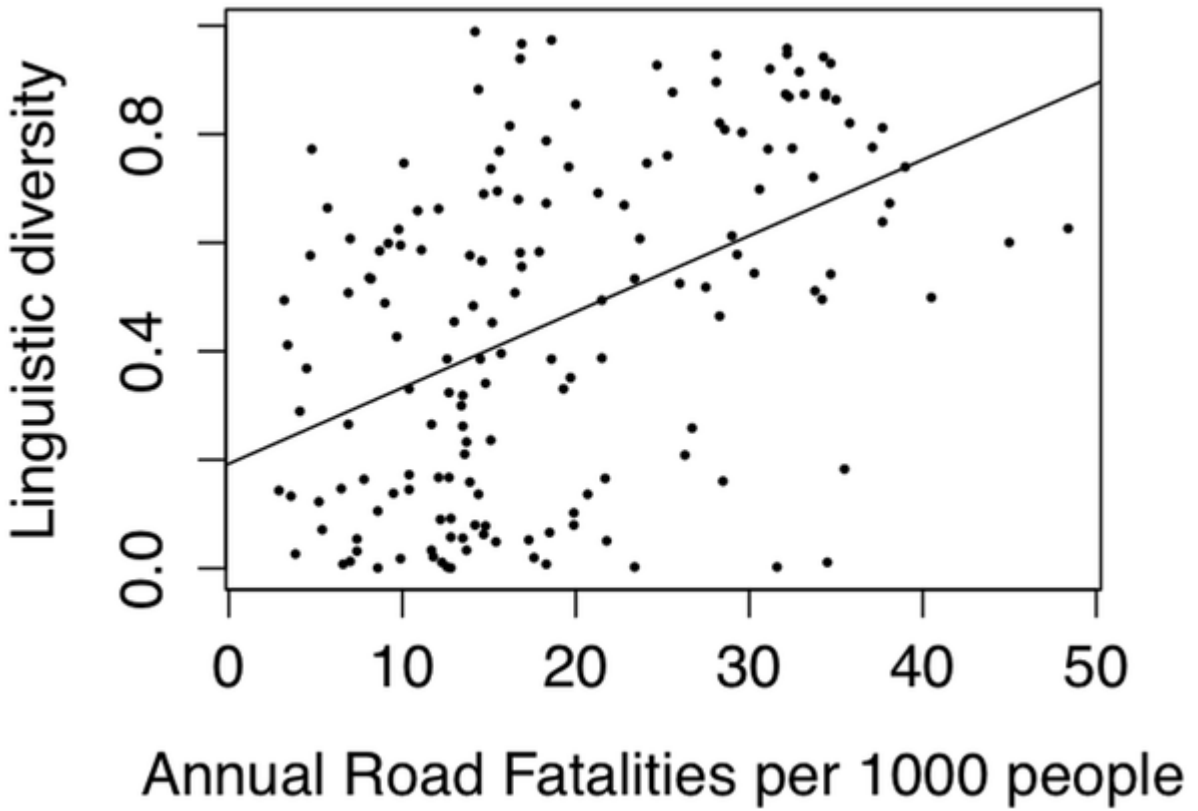
Todos estos métodos siguen el mismo proceso:

- Define objectivo ("función de perdida")
- "Aprende" de los datos
- Encuentra el parámetro, o combinación de parámetoros, que minimizan la perdida (maximizan el objectivo)

Todos estos métodos siguen el mismo proceso:

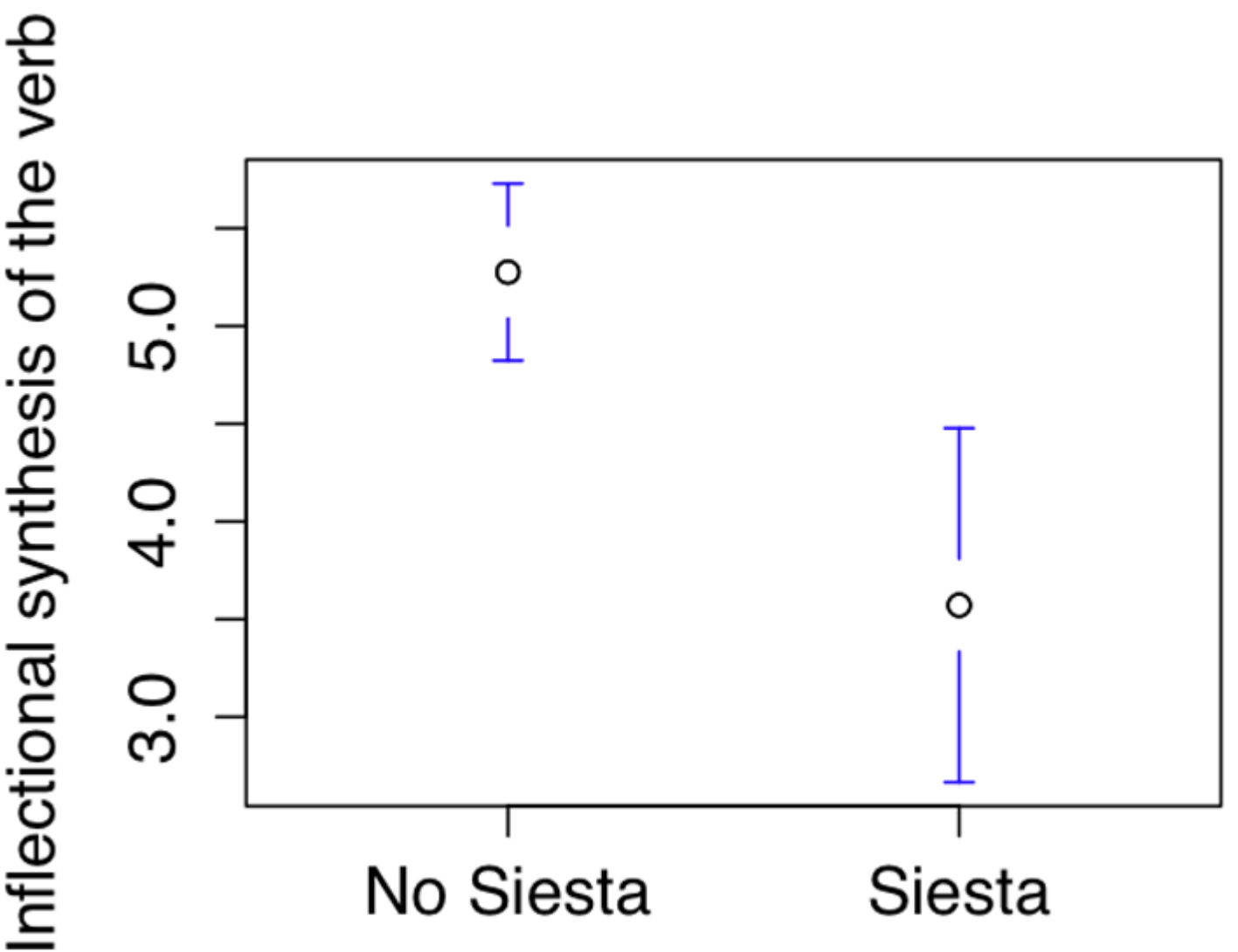
- Define objectivo ("función de perdida")
- "Aprende" de los **datos**
  - Calidad de datos
  - Razón por qué una o más variables podrían solucionar el problema
- Encuentra el parámetro, o combinación de parámetoros, que minimizan la perdida (maximizan el objectivo)

- 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: probabilidad que dos personas de una población tengan la misma lengua nativa



y = Promedio de categorias gramáticales que puede tomar el verbo

Más allá: Ciencias del lenguaje empiricas

De momento: Saliendo lentamente de la crísis de replicabilidad

"Nuevas" áreas de investigación/aplicación:

- Traducción asistida
- Tipología computacional
- Human in the loop QA
- PLN
- Metodologías de aprendizaje asistidas / automatizadas
- ...

Coda

- Avaldo (20/06)
- Informe final: 28/06

Gracias!

https://brochhagen.github.io/

thomas.brochhagen@upf.edu

52.631