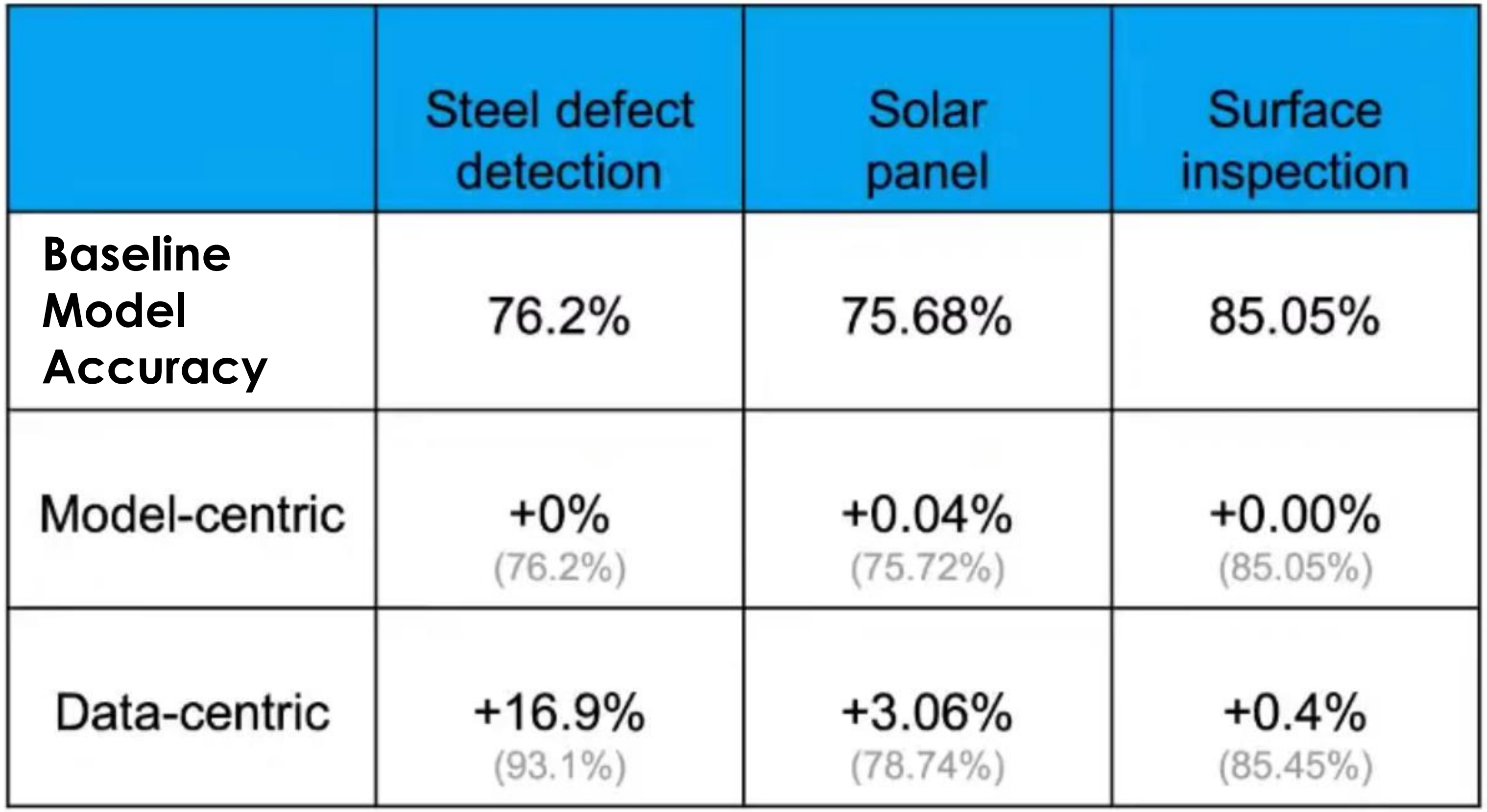
DATA QUALITY IN DATA SCIENCE

Stephanie Eckman, PhD

“Everyone wants to do the model work, not the data work”

Sambasivan et al, 2021 doi:10.1145/3411764.3445518

Data Focus Improves Model Accuracy



	Steel defect detection	Solar panel	Surface inspection
Baseline Model Accuracy	76.2%	75.68%	85.05%
Model-centric	+0% (76.2%)	+0.04% (75.72%)	+0.00% (85.05%)
Data-centric	+16.9% (93.1%)	+3.06% (78.74%)	+0.4% (85.45%)

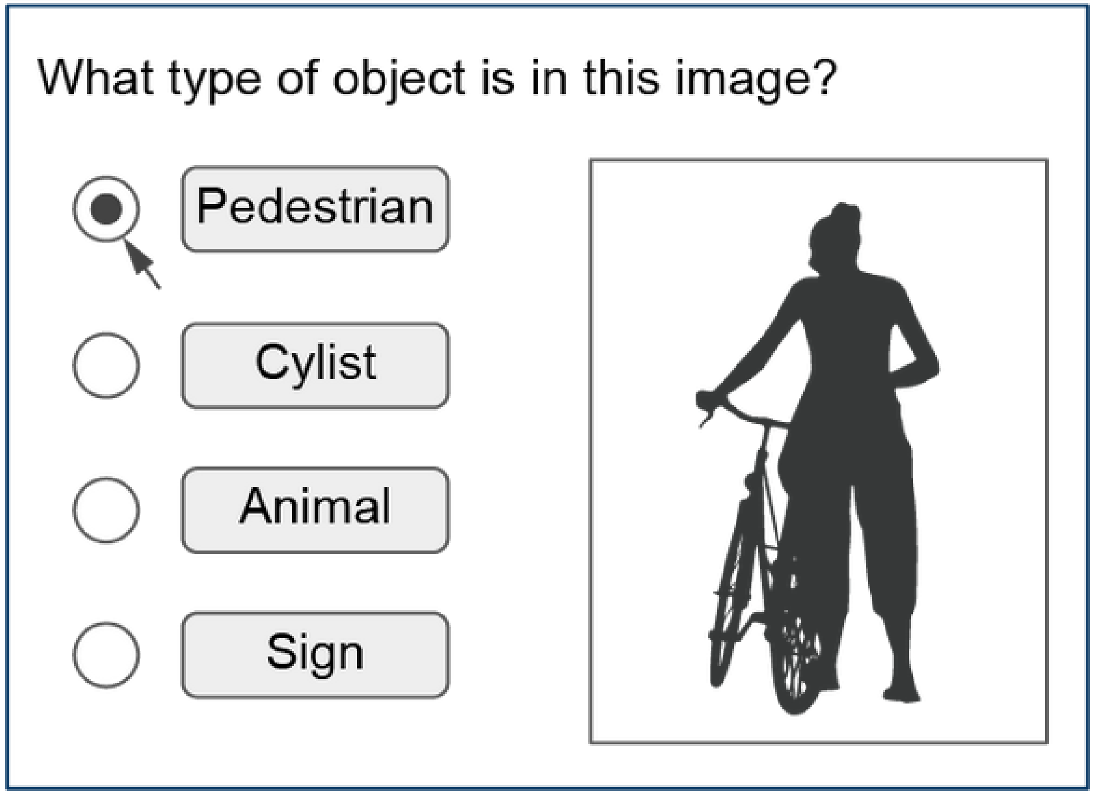
Andrew Ng, Mar 24, 2021 https://youtu.be/06-AZXmwHjo

Measurement Error

- Values in data set are wrong

- Labels assigned by annotators

- Forms look like *surveys*



What type of object is in this image?

● Pedestrian

○ Cylist

○ Animal

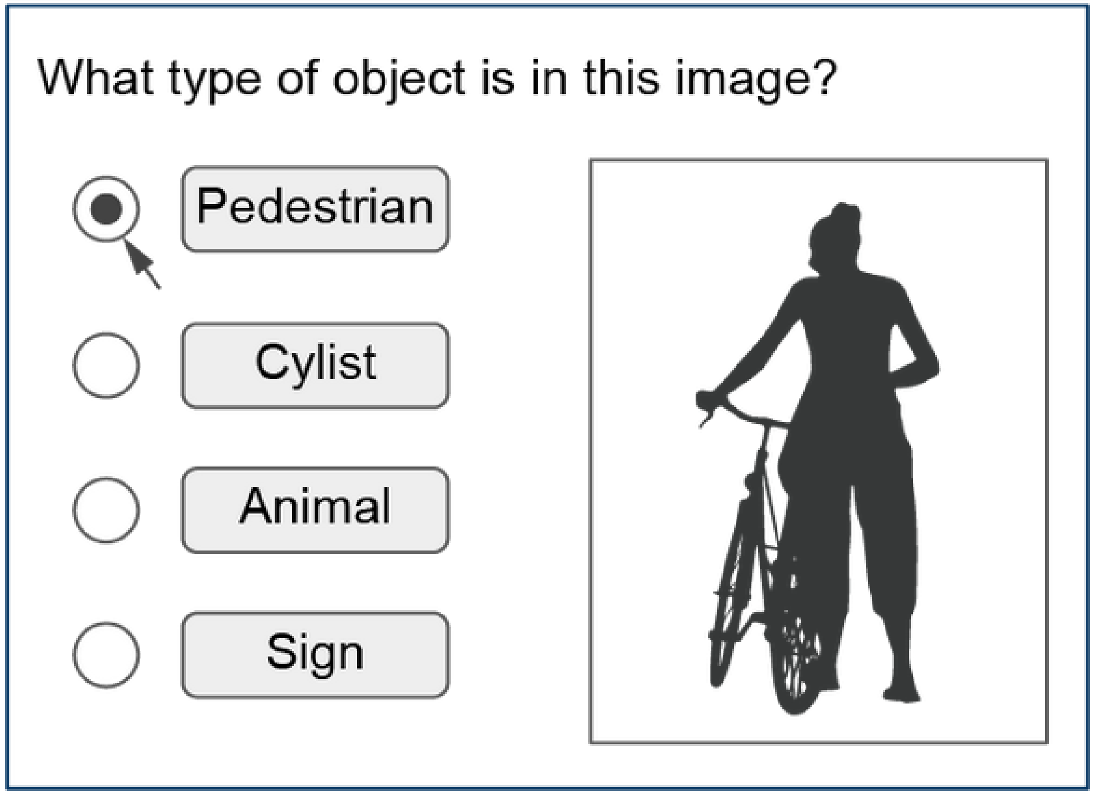
○ Sign

From Monarch (2021) *Human in the Loop*

Wording & Order Effects

- Question Wording
- Question Order
- Response Scales
- Response Order

- Not all findings will carry over
  - Social desirability effects



What type of object is in this image?

● Pedestrian

○ Cylist

○ Animal

○ Sign

From Monarch (2021) *Human in the Loop*

Motivated Misreporting

Has anyone in this household purchased ***pants*** in the last 3 months?

Yes

How much did they cost?
Does that price include tax?
Did you buy them online?

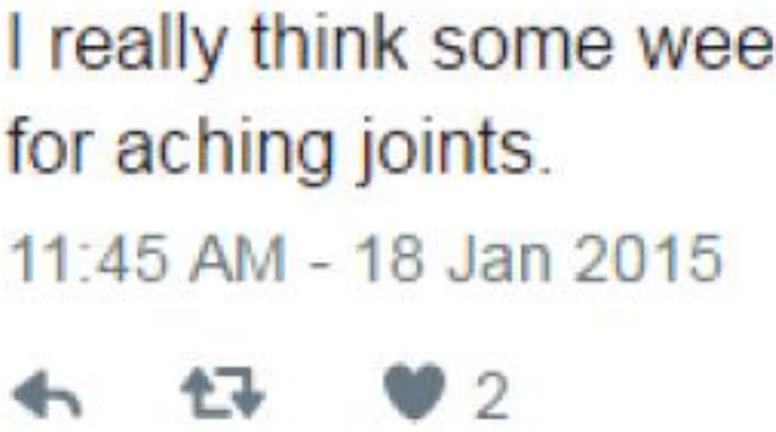
……………

Has anyone in this household purchased ***shoes*** in the last 3 months?

Yes…

Questions adapted from Consumer Expenditure Survey

In Annotation Context



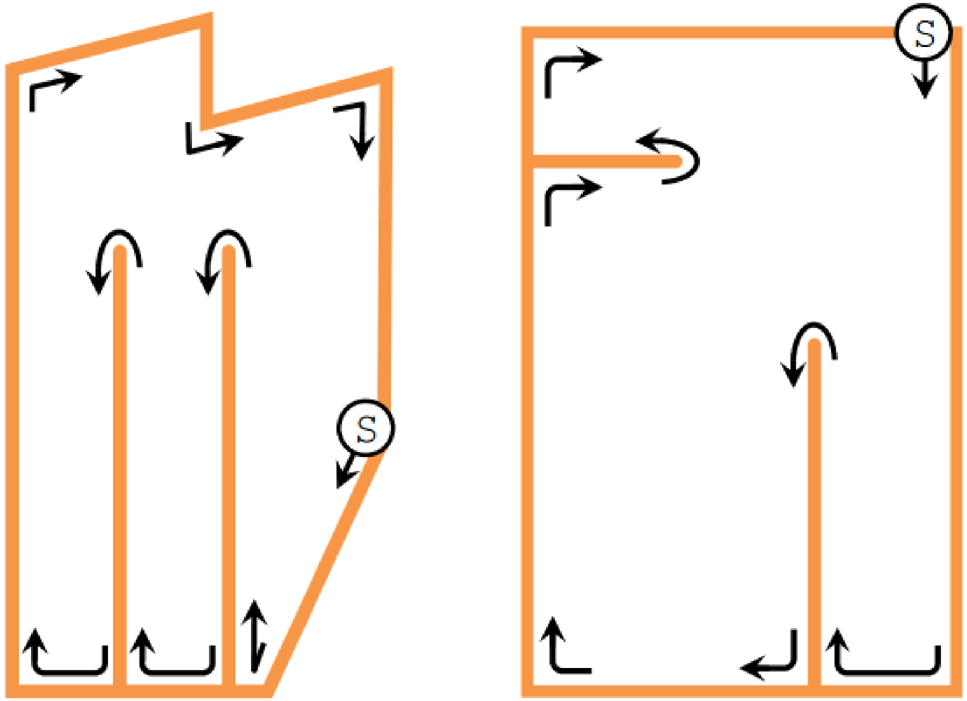
I really think some weed brownies would do my grandma good for aching joints.

11:45 AM - 18 Jan 2015

♥ 2

- **Is this tweet about marijuana?**
  - Is this tweet about sleep (as it relates to marijuana)?
  - …physical or emotional pain…?
  - …nausea…?
  - What is this tweet’s tone or opinion regarding using marijuana?

Anchoring Effects (Confirmation Bias)

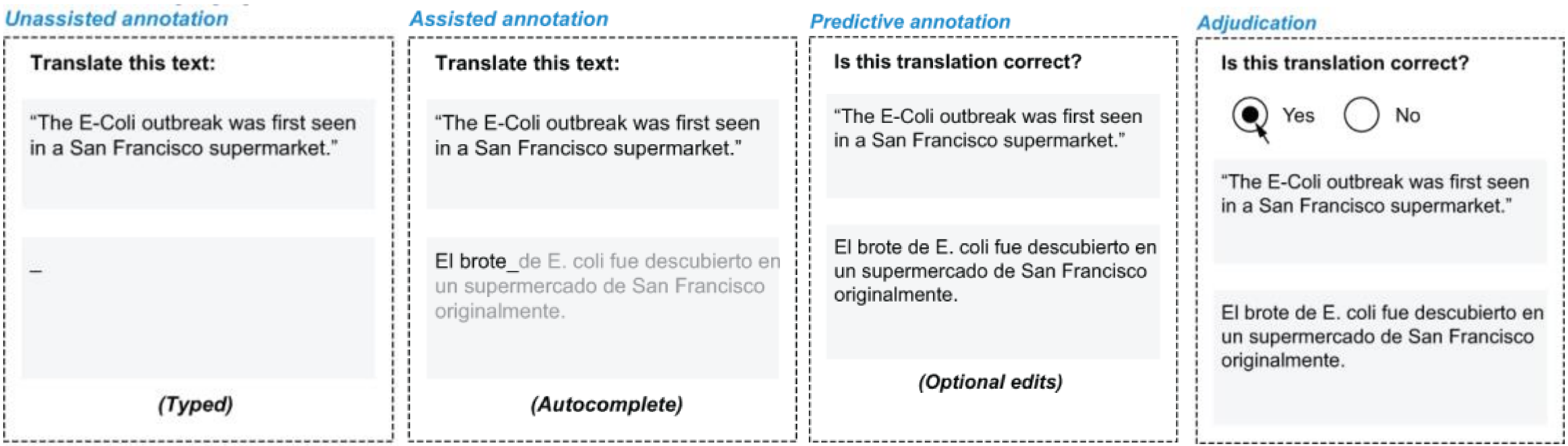


https://www.who.int/tobacco/surveillance/en_tfi_gats_mappingandlistingmanual_v2_final_15dec2010.pdf

- Updating address lists in the field
- Too trusting of list
  - Missing addresses not added
  - Incorrect addresses not deleted

Eckman & Kreuter, 2011

Anchoring Effects (Confirmation Bias)



Unassisted annotation

Translate this text:

"The E-Coli outbreak was first seen in a San Francisco supermarket."

_

(Typed)

Assisted annotation

Translate this text:

"The E-Coli outbreak was first seen in a San Francisco supermarket."

El brote_de E. coli fue descubierto en un supermercado de San Francisco originalmente.

(Autocomplete)

Predictive annotation

Is this translation correct?

"The E-Coli outbreak was first seen in a San Francisco supermarket."

El brote de E. coli fue descubierto en un supermercado de San Francisco originalmente.

(Optional edits)

Adjudication

Is this translation correct?

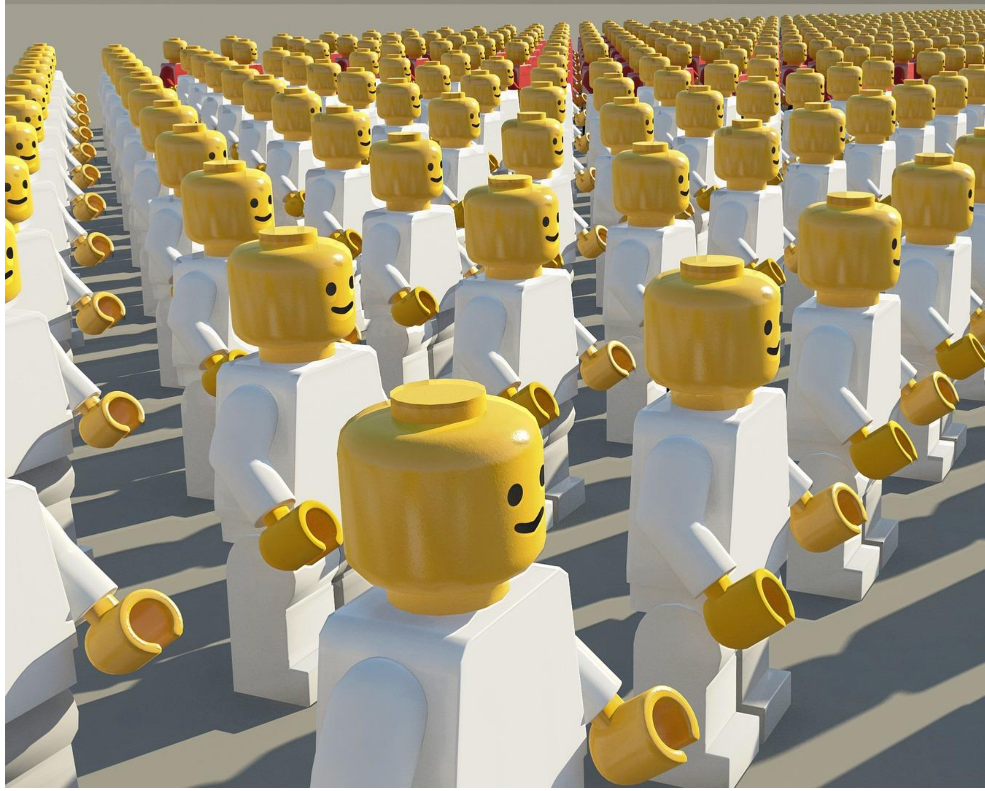
● Yes ○ No

"The E-Coli outbreak was first seen in a San Francisco supermarket."

El brote de E. coli fue descubierto en un supermercado de San Francisco originalmente.

From Monarch (2021) *Human in the Loop*

Measurement Error & Noisy Labels



- More sophistication on ***causes*** of error

- Statistical forms of error: $y = y^{*} + \epsilon$

$$\epsilon \sim N(\mu, \sigma^2)$$

- Majority-rule
  - 90% probability correct
  - 2 of 2 agree: 99% CORRECT
  - 3 of 3 agree: 99.9% CORRECT

Errors may not be *independent*

- **Order effects**
- **Motivated misreporting**
- **Anchoring effects**

Take Aways

- Lots of approaches in DS literature
  - Ordering from easy to hard
  - Ask annotators to code certainty (%)
  - Build model to predict label accuracy

- Make it easy for annotators to give correct answer
  - Incorporate findings on how to collect high quality data

- Be as careful about your data as you are with the models

Stephanie Eckman, PhD

Fellow, RTI International

@stephnie

stepheckman.com