LECTURE 4

Inter-Annotator Agreement

Cohen's Kappa *(κ)*
Fleiss' Kappa *(κ)*

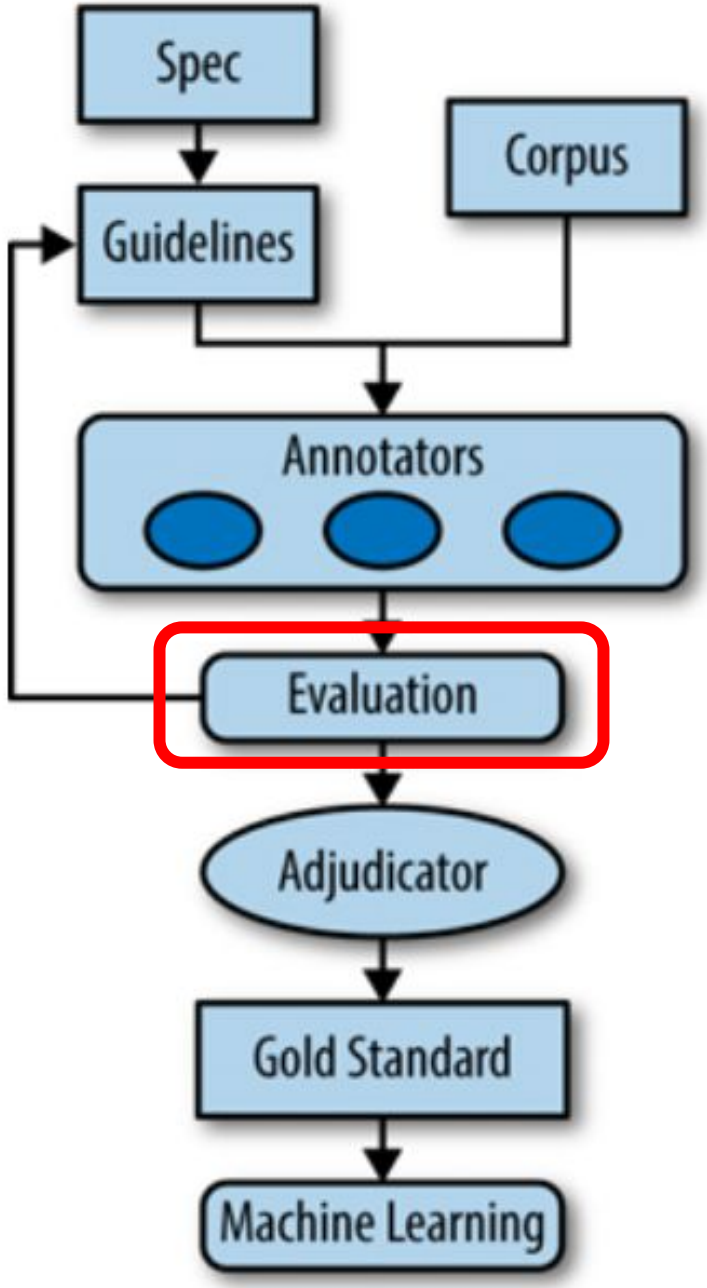
Evaluating Annotations

Inter-Annotator Agreement (IAA) shows:

- How clear your guidelines are
- How uniformly the annotators understood it
- How reproducible the annotation task is

A high IAA doesn't lead to the high performance of a machine learning algorithm

But a low IAA is improbable to produce high performing ML.



Accuracy as IAA

Where do you see more (trustworthy) agreement?

		B	
		positive	negative
A	positive	4	1
	negative	2	43

1

		B	
		positive	negative
A	positive	25	1
	negative	2	22

2

Accuracy as IAA

The same accuracy but a significant difference in chance agreement

Where do you see more (trustworthy) agreement?

47		B		
	50	positive	negative	
A	positive	4	1	5
	negative	2	43	45
		6	44	0.9400

0.12 0.88 0.10 0.90

.90 x .88 + .10 x .12 =.80*

47		B		
	50	positive	negative	
A	positive	25	1	26
	negative	2	22	24
		27	23	0.9400

0.54 0.46 0.52 0.48

.48 x .46 + .52 x .54 =.50*

*Numbers are rounded

F1 score as IAA

The same F1 score but a significant difference in chance agreement

.13 x .14 + .87 x .87 =.77

197		B		
	200	positive	negative	
A	positive	25	1	0.9615
	negative	2	172	
		0.9259		0.9434

.48 x .46 + .52 x .54 =.50

47		B		
	50	positive	negative	
A	positive	25	1	0.9615
	negative	2	22	
		0.9259		0.9434

Recall

Precision

F1

IAA measures: κ

Kappa

Accuracy and F1 score don't take into account expected chance agreements that are likely to occur when people annotate texts

The measures taking expected chance agreement into account:

- Cohen's κ: two annotators annotating each subject with a category
- Fleiss' κ: each subject was annotated n times with a category

Observed agreement

$$\kappa \equiv \frac{A_o - A_c}{1 - A_c}$$

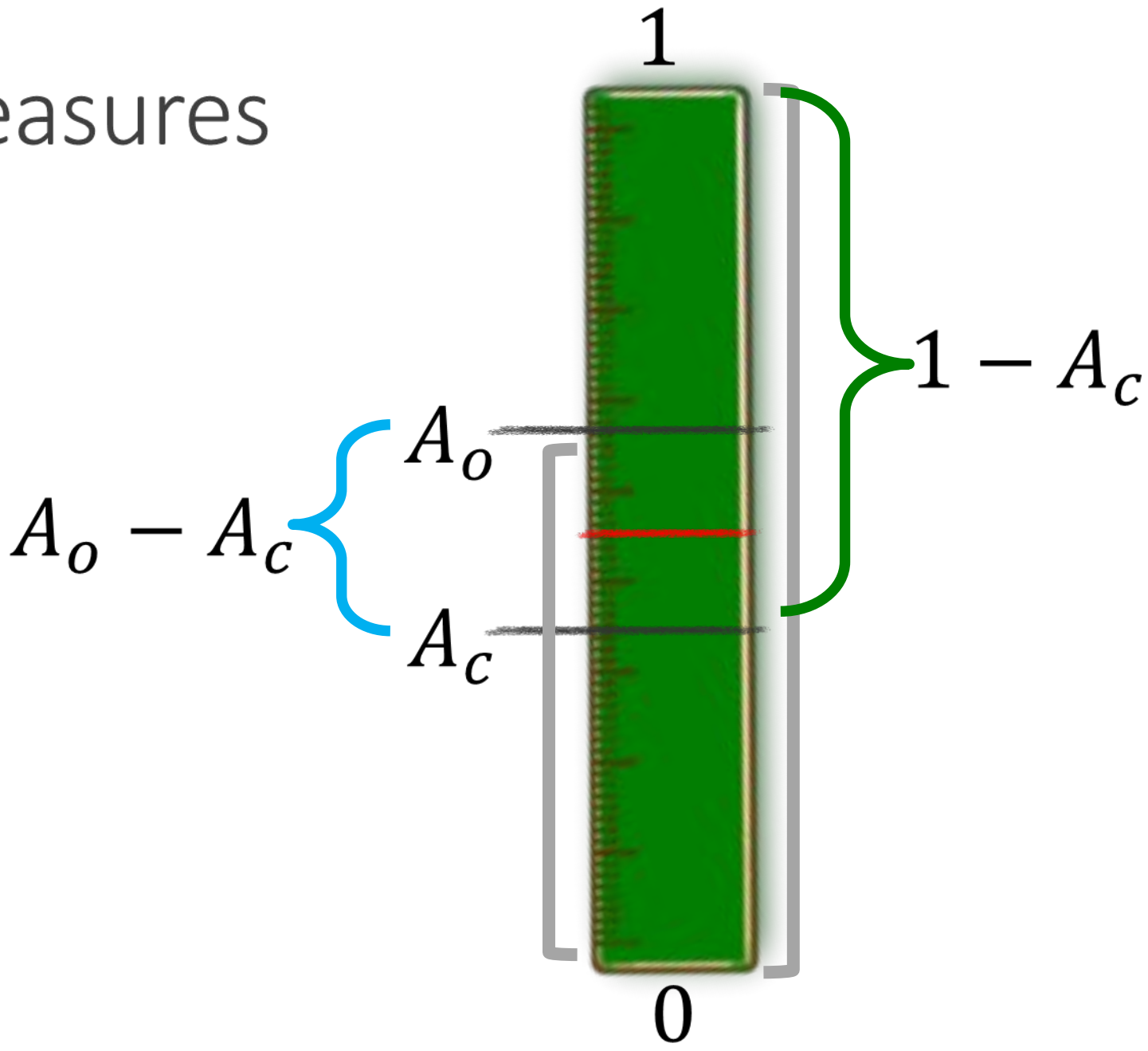
Chance agreement

Idea behind κ measures

Observed agreement

$$\kappa \equiv \frac{A_o - A_c}{1 - A_c}$$

Chance agreement

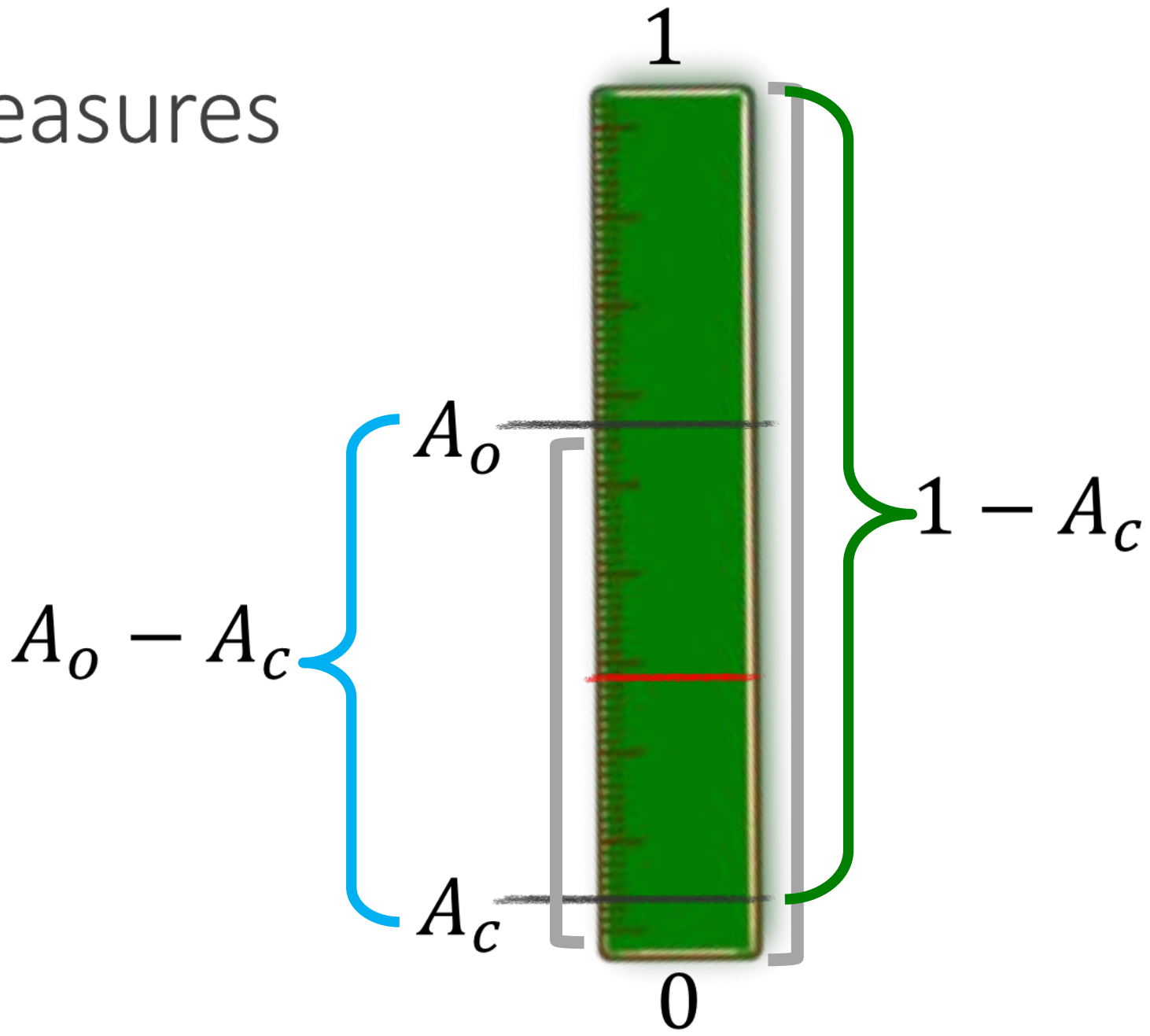


Idea behind κ measures

Observed agreement

$$\kappa \equiv \frac{A_o - A_c}{1 - A_c}$$

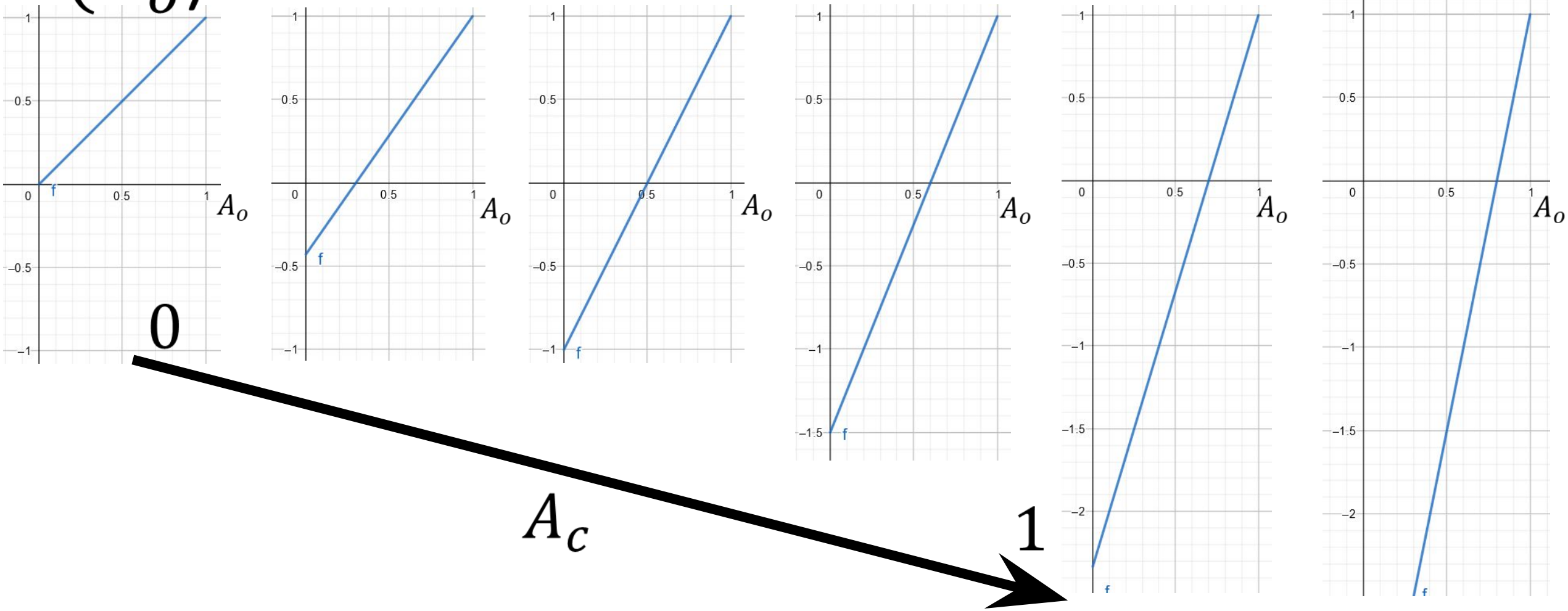
Chance agreement



Idea behind κ measures (2)

$$\kappa \equiv \frac{A_o - A_c}{1 - A_c}$$

$\kappa(A_o)$



Cohen's κ

Observed agreement

$$\kappa \equiv \frac{A_o - A_c}{1 - A_c}$$

Chance agreement

Example

250 movie reviews

Annotated by two coders:

- Ann and Bob

With three categories:

- Positive
- Neutral
- Negative

*The errata: https://www.oreilly.com/catalog/errata.csp?isbn=0636920020578

Cohen's κ: calculate p_o

Total items with annot. agreement

Total items

144		B		
	250	**positive**	**neutral**	**negative**
A	**positive**	54	28	3
	neutral	31	18	23
	negative	0	21	72
				0.5760

p_o

Accuracy: ratio of agreed annotations and total items

Cohen's κ: calculate p_c

Calculate distribution of categories per annotator

144		B				
	250	positive	neutral	negative		
A	positive	54	28	3	85	0.3400
	neutral	31	18	23	72	0.2880
	negative	0	21	72	93	0.3720
		85	67	98	0.5760	
		0.3400	0.2680	0.3920		

85/250

Cohen's κ: calculate p_c

Calculate distribution of categories per annotator

Calculate chance agreement from the distributions

144		B					
	250	positive	neutral	negative			
A	positive	54	28	3	85	0.3400	0.1156
	neutral	31	18	23	72	0.2880	0.0772
	negative	0	21	72	93	0.3720	0.1458
		85	67	98	0.5760		0.3386
		0.3400	0.2680	0.3920			

0.34 x 0.34

$\sum$

p_c

Cohen's κ: calculate

$$\kappa \equiv \frac{A_o - A_c}{1 - A_c}$$

0.5760 - 0.3386

1 - 0.3386

						0.3400	0.1156
						0.2880	0.0772
						0.3720	0.1458
		85	67	98	0.5760	0.6614	0.3386
		0.3400	0.2680	0.3920	0.2374	0.3589	

κ

Cohen's κ

$$A_m = \sum_{i=1}^{k} n_{mi} \qquad B_m = \sum_{i=1}^{k} n_{im} \qquad p_m^X = \frac{X_m}{N}$$

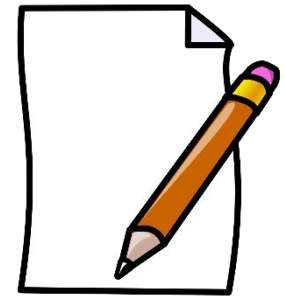
$$\kappa \equiv \frac{p_o - p_c}{1 - p_c}$$

Observed agreement

Chance agreement

		B					
A							

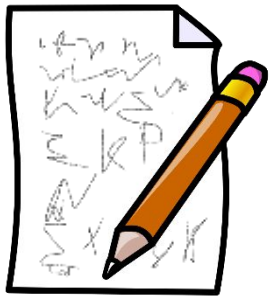
Cohen's κ



		B				
		positive	negative			
A	positive	3	2			
	negative	1	4			

Cohen's κ

$$\kappa \equiv \frac{A_o - A_c}{1 - A_c}$$



7		B				
	10	positive	negative			
A	positive	3	2	5	0.50	0.20
	negative	1	4	5	0.50	0.30
		4	6	0.70	0.50	0.50
		0.40	0.60	0.20	0.40	

Cohen's κ vs Accuracy & F-score

Cohen's κ

Accuracy

F1 score

Accuracy: **0.940**

.694

.727

		B	
		positive	negative
A	positive	4	1
	negative	2	43

.880

.943

		B	
		positive	negative
A	positive	25	1
	negative	2	22

.935

.985

		B	
		positive	negative
A	positive	25	1
	negative	2	172

Cohen's

$$\kappa \equiv \frac{A_o - A_c}{1 - A_c}$$

```
N, k = M.shape  # N is # of items, k is # of categories
n_annotators = np.sum(M[0, :])  # # of annotators
tot_annotations = N * n_annotators  # the total # of annotations

# observed agreement
p = (np.sum(M * M, axis=1) - n_annotators) / (n_annotators * (n_annotators - 1))
P_o = np.sum(P) / N # add all observed agreement

# chance agreement
p = np.sum(M, axis=0) / tot_annotations # the distribution of each category
P_e = np.sum(p * p)  # average chance agreement

k = (P_o - P_e) / (1 - P_e)
```

Fleiss' κ

Cohen's

$$\kappa \equiv \frac{\overline{\boldsymbol{P}}_o - \overline{\boldsymbol{P}}_c}{1 - \overline{\boldsymbol{P}}_c}$$

Fleiss'

$$\kappa \equiv \frac{A_o - A_c}{1 - A_c}$$

Fleiss' κ

Observed agreement

$$\kappa \equiv \frac{\bar{P}_o - \bar{P}_c}{1 - \bar{P}_c}$$

Chance agreement

Example

5 movie reviews

Annotated by 250 coders

With three categories:

- Positive
- Neutral
- Negative

Cohen's κ

What if an item is annotated by more than 2 annotators?

Fleiss' κ: compute $\overline{P}_c$

Total annotations

1250	Positive	Neutral	Negative
Review 1	85	72	93
Review 2	85	67	98
Review 3	68	99	83
Review 4	88	88	74
Review 5	58	120	72
	384	446	420
P_c	0.3072	0.3568	0.3360

No info per annotator

$0.3072^2 + 0.3568^2 + 0.3360^2$

0.3346

$\overline{P}_c$

Distribution of categories

Fleiss' κ: compute $\overline{P}_o$

1250	Positive	Neutral	Negative		P_o
Review 1	85	72	93	250	0.3343
Review 2	85	67	98		
Review 3	68	99	83		
Review 4	88	88	74		
Review 5	58	120	72		
	384	446	420		
P_c	0.3072	0.3568	0.3360		

$$\frac{85 \times 84 + 72 \times 71 + 93 \times 92}{250 \times 249}$$

$$\frac{85^2 + 72^2 + 93^2 - 250}{250 \times 250 - 250}$$

Fleiss' κ: compute $\overline{P}_o$

1250	Positive	Neutral	Negative		P_o
Review 1	85	72	93	250	0.3343
Review 2	85	67	98	250	0.3384
Review 3	68	99	83	250	0.3384
Review 4	88	88	74	250	0.3328
Review 5	58	120	72	250	0.3646
	384	446	420		**0.3417**
P_c	0.3072	0.3568	0.3360	0.3346	

$\frac{\sum}{5}$

$\overline{P}_o$

Fleiss' κ: compute $\kappa \equiv \frac{\bar{P}_o - \bar{P}_c}{1 - \bar{P}_c}$

1250	Positive	Neutral	Negative		P_o
Review 1	85	72	93	250	0.3343
Review 2	85	67	98	250	0.3384
Review 3	68	99	83	250	0.3384
Review 4	88	88		250	0.3328
Review 5	58	120	72	250	0.3646
	384	446	320	**0.0071**	**0.3417**
P_c	0.3072	0.3568	0.3360	**0.3346**	**0.0107**

$\bar{P}_o - \bar{P}_c$

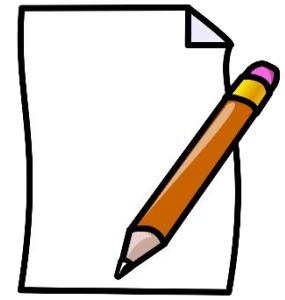
$\bar{P}_c$

$\bar{P}_o$

κ

Fleiss' κ for 2 coders

Scott's π



		B	
		positive	**negative**
A	**positive**	3	2
	negative	1	4

Fleiss' κ for 2 coders

		B	
		positive	**negative**
A	**positive**	3	2
	negative	1	4

20	Pos.	Neg.		
Review 1	**2**	**0**	2	1
Review 2	**2**	**0**	2	1
Review 3	**2**	**0**	2	1
Review 4	**1**	**1**	2	0
Review 5	**1**	**1**	2	0
Review 6	**0**	**2**	2	1
Review 7	**0**	**2**	2	1
Review 8	**0**	**2**	2	1
Review 9	**0**	**2**	2	1
Review 10	**1**	**1**	2	0
	9	11	**0.1950**	**0.70**
	0.45	0.55	**0.5050**	**0.3939**

```
for an1, an2 in zip(ann1, ann2):
    if an1 == an2:
        count += 1
A = count / len(ann1)   # observed agreement A (Po)

uniq = set(ann1 + ann2)
E = 0   # expected agreement E (Pe)
for item in uniq:
    cnt1 = ann1.count(item)
    cnt2 = ann2.count(item)
    count = ((cnt1 / len(ann1)) * (cnt2 / len(ann2)))
    E += count

k = (A - E) / (1 - E)
```

Cohen's κ vs Fleiss' κ

2 annotators

Each annotator annotates every item

Table: confusion matrix

- Annotation counts of Annotator 1
- Annotation counts of Annotator 2

X annotators

Every item is not necessarily annotated by each annotator

Table: annotation counts per items:

- Items
- Annotation counts per item

Fleiss' κ is an extension of Scott's π for two coders (not Cohen's κ) 

Cohen's κ is more informative than Scott's π due to the way the chance agreement is calculated

Interpreting κ coefficients

Values	Agreement level*
< 0	Poor
0.01 - 0.20	Slight
0.21 - 0.40	Fair
0.41 - 0.61	Moderate
0.61 - 0.80	Substantial
0.81 - 1.00	Almost perfect
1	perfect

By no means universally accepted

The number of categories

The number of annotators

Annotation task

Part-of-speech tagging
VS
Semantic role labelling

Compare κ coefficients of related tasks

Landis, J.R.; Koch, G.G. (1977). "The measurement of observer agreement for categorical data". Biometrics. **33 (1)*

Adjudication

Prerequisite: being happy with IAA

Deliverable: gold standard dataset

Who: those who helped create the guidelines

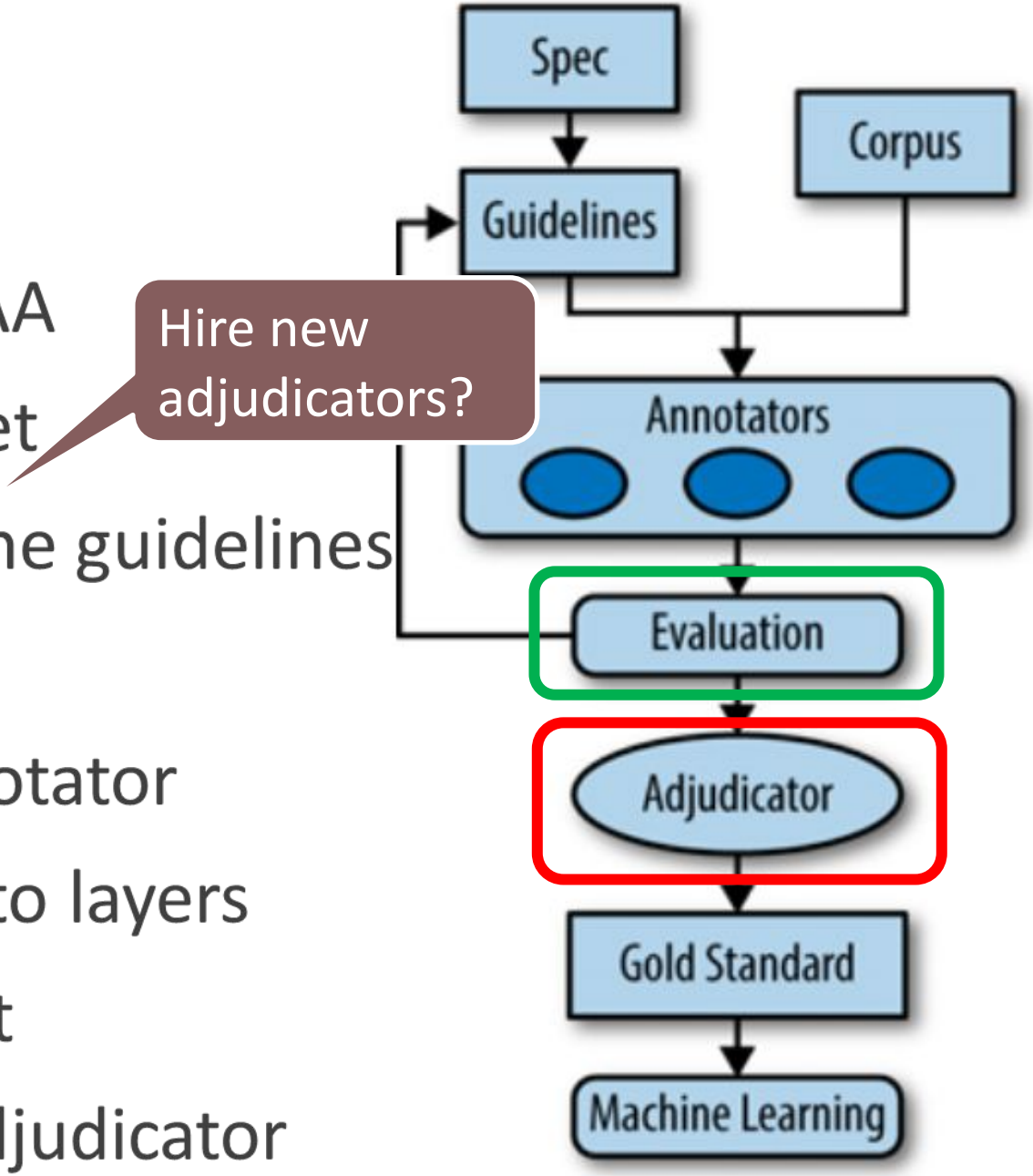
Tool: adjudication software

Time: ~ annotation time per annotator

How: break up the adjud. task into layers

Caution: same decision ↛ correct

Advice: IAA for more than one adjudicator



After annotation, rate IAA with:

- Cohen's κ
- Fleiss' κ

Most common in computational linguistics

Calculating κ is not always straightforward

Text extent annotations

Possible revisions (via the MAMA cycle) for satisfactory IAA

Interpreting IAA scores isn't an exact science

High IAA means the task is likely to be reproducible

Not necessarily suitable for feeding ML algorithms

After high IAA scores, set the annotators loose on the full data

Adjudicate disagreements (use IAA score for adjudicators)