

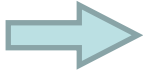
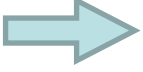
# A Case Study on Start-up of Dataset Construction: In Case of Recipe Named Entity Corpus

Yoko Yamakata, *Keishi Tajima* and Shinsuke Mori  
Kyoto University, Japan

# How We Start Dataset Construction?

- We have a new problem! Machine learning?
    - Need a new dataset! But it is an exhausting task...
    - Hiring annotators or crowdsourcing?
    - We need to give them a clear **annotation guideline!**
  - Guideline creation typically goes like this...
    - Find similar datasets and learn their guidelines
    - Adapt them to our own tasks
    - Try annotation by ourselves to see if it is OK
    - Now let the annotators start to give annotation!
- Even if you design the guideline very carefully, a lot of unexpected cases and ambiguity in rules will be found!**

# Three Main Problems

1. We need a system supporting the management of versions.
  - multiple versions of guidelines
  - also multiple versions of annotations under different guideline versions
2. How often should we update the guideline?
  - If very frequent  many versions of the same annotation
  - If less frequent  more annotation under old guidelines
3. When we have updated the guideline, which is better:
  - revising the old annotations under the new guideline, or
  - adding more data instead?

# Our Task: Recipe Named Entity Corpus

Give tags of Recipe Named Entity (r-NE) to a cooking procedural text

## 10 types of r-NE

Tag	Meaning
F	Food
T	Tool
D	Duration
Q	Quantity
Ac	Action by chef
Ac2	Discontinuous Ac
Af	Action by food
At	Action by tool
Sf	Food state
St	Tool state

## Example of annotation

Original text

Preheat oven to 200 C / Gas mark 6 .



24 min./recipe

Annotation result

Preheat/Ac-B oven/T-B to/O 200/St-B C/St-I  
//O Gas/St-B mark/St-I 6/St-I ./O

# Recipe data collection

Recipes were crawled  
at Allrecipes.co.uk

dish type	#recipe	propotion	#corpus
Bread	953	3.1%	3
Pies and tar	1251	4.0%	4
Soup	2046	6.6%	7
Salad	1755	5.7%	6
Main course	11523	37.2%	37
Dessert	3366	10.9%	11
Biscuits and	1655	5.3%	6
Pancakes	364	1.2%	1
Breakfast	1078	3.5%	3
Sandwiches	377	1.2%	1
Starters	2331	7.5%	8
Side dish	2166	7.0%	7
Sweets	416	1.3%	1
Preserves	423	1.4%	1
Drink	1231	4.0%	4
Cake	(4284)	-	-
<b>Total</b>	<b>30935</b>	<b>100.0%</b>	<b>100</b>

# Preparing the Guideline

1. We first defined tags and a guideline for Japanese recipe.  
<http://www.ar.media.kyoto-u.ac.jp/how-to/recipe-NLP/>
  2. We translated them in order to adapt it to English recipe.
  3. We hired a British doctor course student who had computational linguistics experience.
  4. He soon sent us many questions!
- Even when we adopt an existing guideline, we have many unexpected cases!
  - We had both
    - questions that do not require the revision of the guideline, and
    - **many questions requiring discussions and guideline revisions!**

# Example of Rules in the Original Guideline

**P1:** Prepositions and conjunctions are tagged O (i.e. outside an r-NE), except when they are part of a collocation.

**P2:** Adverbs and adverbial phrases are tagged O except when they are part of a phrasal verb.

- *throw/Ac-B away/Ac-I*
- *mix/Ac-B in/O the/O bowl/T*

**P3:** A sequence of words denoting a single action/food/tool in the cooking process is annotated as a single r-NE.

- *frying/T-B pan/T-I*
- *bring/Ac-B to/Ac-I the/Ac-I boil/Ac-I*

**P4:** Auxiliary and modal verbs are tagged O.

See detail in [13] Y. Yamakata, J. Carroll, and S. Mori, “A comparison of cooking recipe named entities between Japanese and English,” in *CEA*, pp. 7–12, 2017.

# Examples of Questions

- “Pour/Ac-B into/O the/O **digestive**/F-B biscuit/F-I base F-I”  
Q. "digestive biscuit base" still part of the food?  
A. Yes
- “**Butter**/Ac-B 5/Q-B slices/F-B of/O bread/F-B  
Q. Is “Butter” Ac (Action by chef) or F (Food)?  
A. Ac
- “Repeat/Ac-B with/O the/O **other**/Q-B dough/F-B balls/F-I ./O”  
Q. What’s the tag of “other”?  
A. **Q (Quantity)**      **revision of the guideline**
- “fry/Ac-B diced/Ac-B bacon/F-B in/O a/O **separate**/St-B pan/T-B”  
Q. What’s the tag of “separate”?  
A. **St (State of tool)**      **revision of the guideline**
- “**continue**/Ac-B **cooking**/Ac-I”  
Q. "continue cooking" a single NE?  
A. **Yes**      **revision of the guideline**

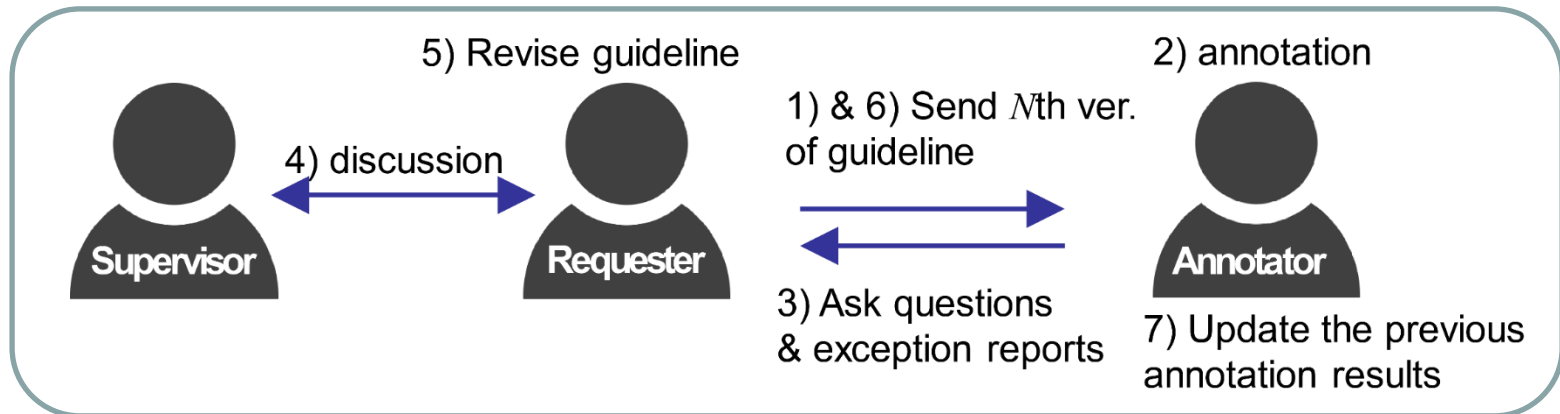


# Examples of Revisions Adding New Tags

Tag	Meaning	Remarks
F	Food	Eatables including intermediate products,
T	Tool	Knife, container, etc.
D	Duration	Duration of cooking
Q	Quantity	Quantity of food
Ac	Action by chef	Verbs representing of a chef's actions
Ac2	Discontinuous Ac	words that consists single "Ac" with adjacent but not contiguous "Ac". <b>English only</b>
Af	Action by food	Verbs representing food's actions
At	Action by tool	Verbs representing tool's actions. <b>English only</b>
Sf	Food state	Food's initial or intermediate states
St	Tool state	Tool's initial or intermediate states

Addition of these tags needed deep discussions with experts in NLP and ML.

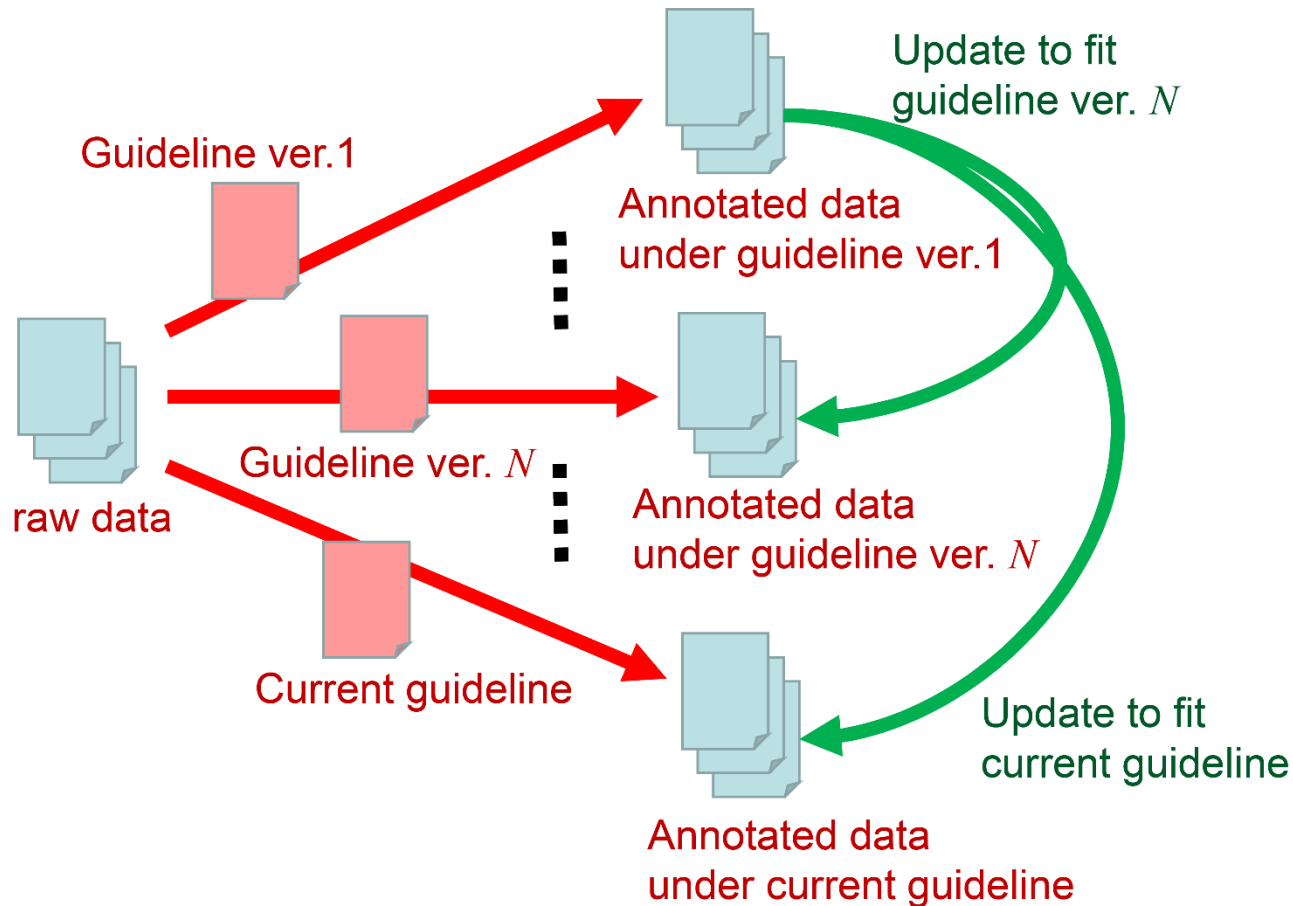
# Supervisor and Requester



- 1) The **requester** describes the annotation guideline and **sends** it to **annotators**.
- 2) The **annotators** **annotate** data according to it and
- 3) **return questions and exceptional cases** that are not clearly specified in the guideline.
- 4) The **requester** **discusses** with **supervisors** if required,
- 5) **revises the guideline** from ver.  $N$ th to  $(N + 1)$ th, and
- 6) **sends** the revised guideline to the **annotators**.
- 7) The **annotators** **update the previous annotation results** to fit the current guideline.

# Problem 1: Supporting Version Management

Guideline update brings different status of data which was annotated under different version of the guideline



## Problem 2: How often updating the guideline?

- We updated the guideline only twice.
  - We did not know which strategy (frequently or not) is better.
  - We were afraid of having repeated updates of the same annotation, which is inefficient.
- We cannot know if repeated updates could occur if we updated the guideline more often.
- Let us guess by looking at what types of revisions of annotations we needed.

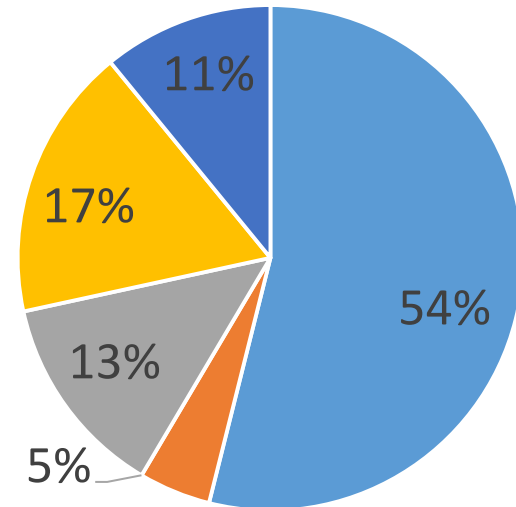
# Problem 2: How often updating the guideline?

#Recipe: 100

#Total word : 13,820

#Tagged word: 7107 (51.4)

#wrong tagged word: 2584 (18.7%)



- Tagged for outside of r-NE
- r-NE were not tagged (tagged as O)
- Correct tag but wrong B/I
- Wrong tag but correct B/I
- Wrong tag and wrong B/I

See next slide

# Problem 3: Revising or Adding Annotations?

## When we have updated the guideline, which is better?

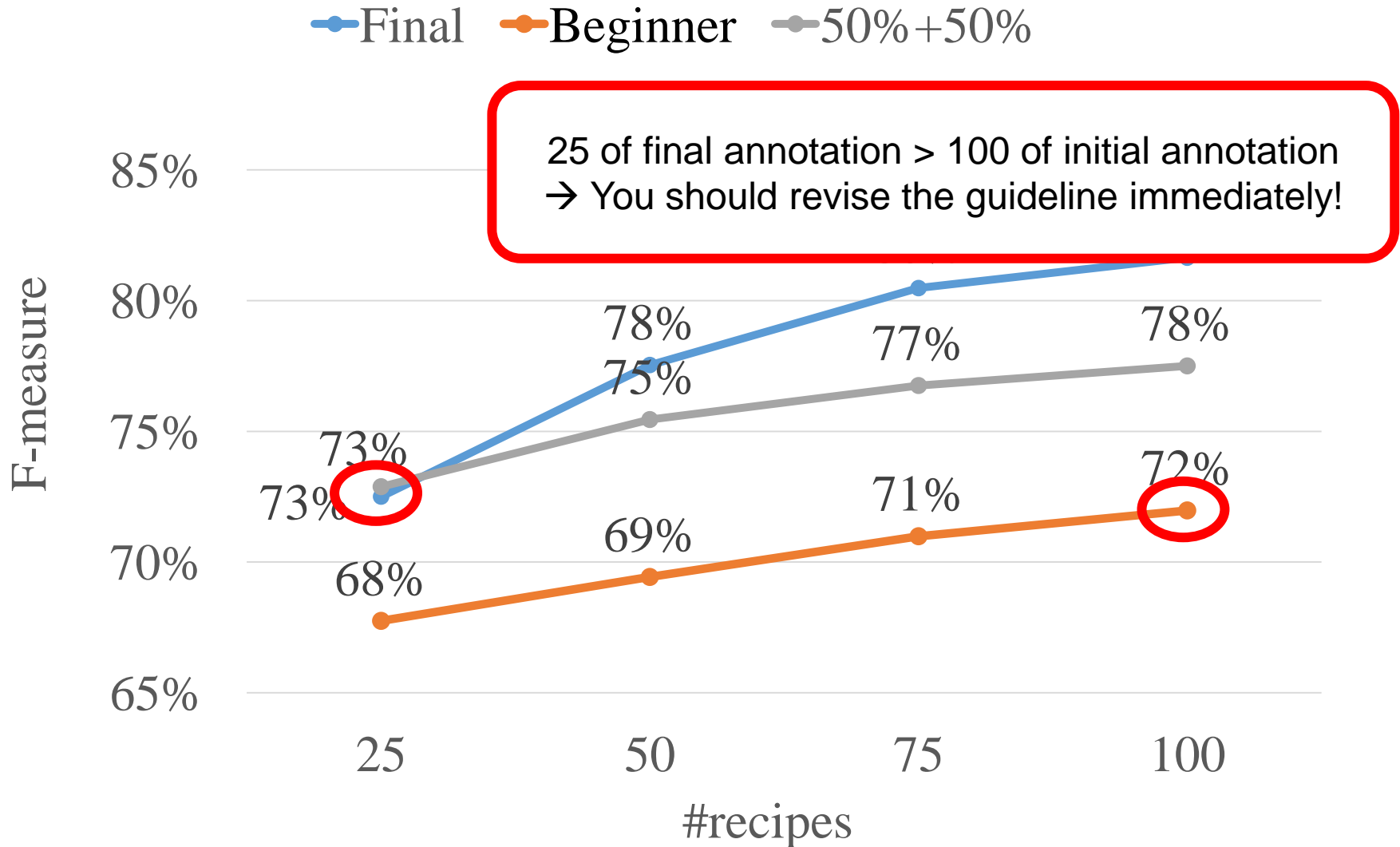
- Revising old annotations under the old guideline (consistency of data)
- Adding more annotations by using the human resource for it (size of data)

## Experiment:

Compare annotation accuracies by there types of training data

- 1. First annotation** based on the initial guideline
  - 2. Final annotation** based on the final guideline
  - 3. 50%-50% mixture** of these two
- Training data size (#recipes): 25, 50, 75, 100
  - Test data: another 120 recipes
  - Accuracies were evaluated with the named entity recognizer PWNER [Sasada et al. 2015]

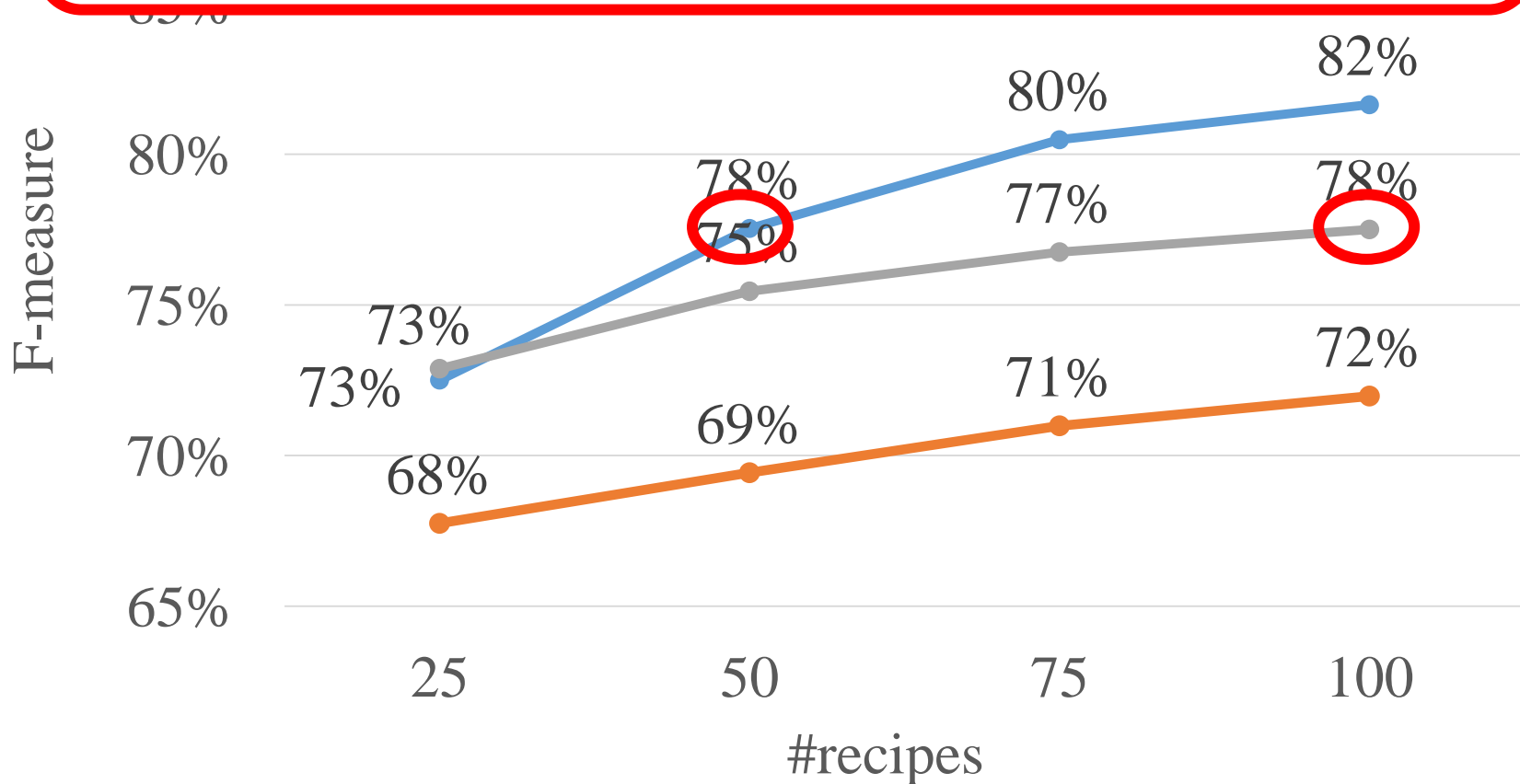
# Classification accuracy with different size of first and final annotation.



# Classification accuracy with different size of first and final annotation.

50 of final = 100 of 50%-50% mixture

If you have already 50 annotation results and have revised the guideline, you should update the old annotations rather than obtaining another 50 annotations.





# Conclusion

- We need a system supporting management of versions of guidelines and versions of annotations under them.
- We should update the guideline frequently.
  - Repeated updates of the same annotations may not occur
  - Annotations under immature guidelines are quite unreliable
- When we have updated the guideline, we should revise the old annotations rather than adding more data.