COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

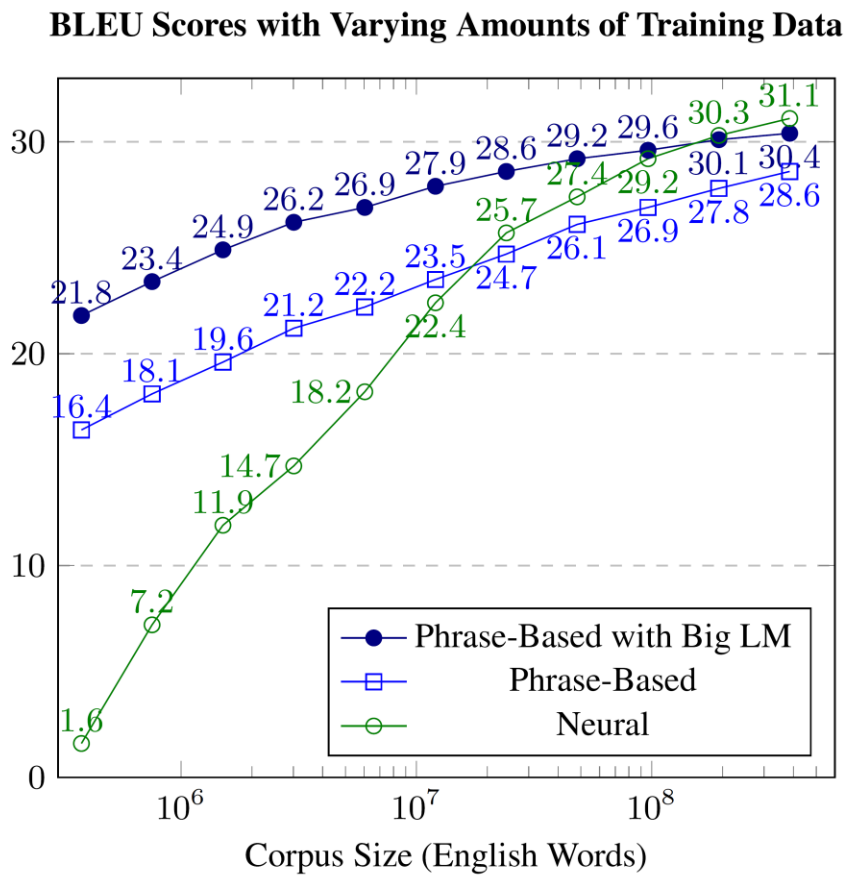
Multilingual and Multitask Learning in seq2seq Models

CMSC 470

Marine Carpuat

Multilingual Machine Translation

Neural MT only helps in high-resource settings



[Koehn & Knowles 2017]

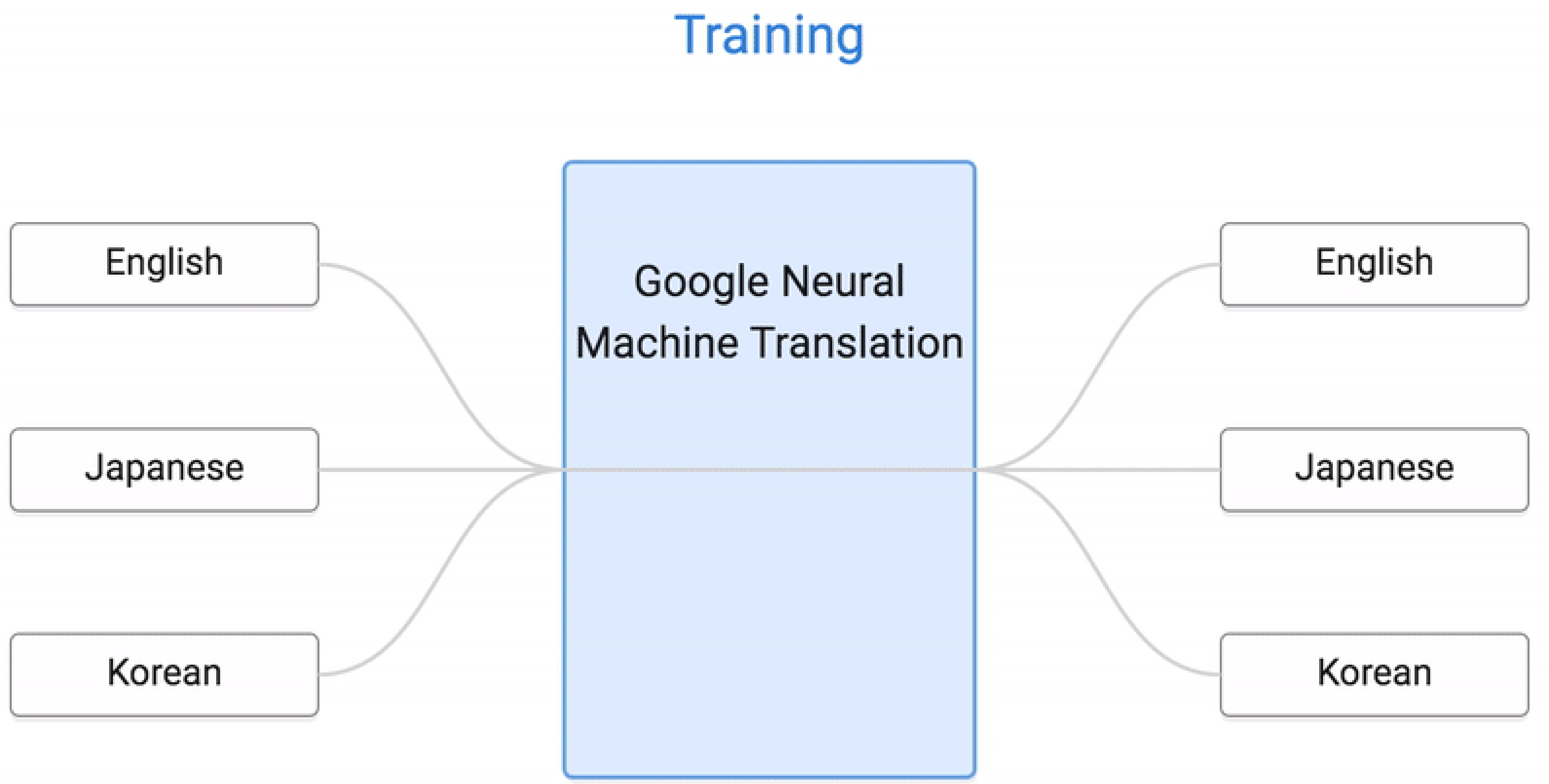
Ongoing research

- Learn from other sources of supervision than pairs (E,F)
  - Monolingual text
  - **Multiple languages**
- Incorporate linguistic knowledge
  - As additional embeddings
  - As prior on network structure or parameters
  - To make better use of training data

Multilingual Translation

- Goal: support translation between any N languages
- Naïve approach: build on translation system for each language pair and translation direction
  - Results in N^2 models
  - Impractical computation time
  - Some language pairs have more training data than others
- Can we train a single model instead?

The Google Multilingual NMT System
[Johnson et al. 2017]



The Google Multilingual NMT System
[Johnson et al. 2017]

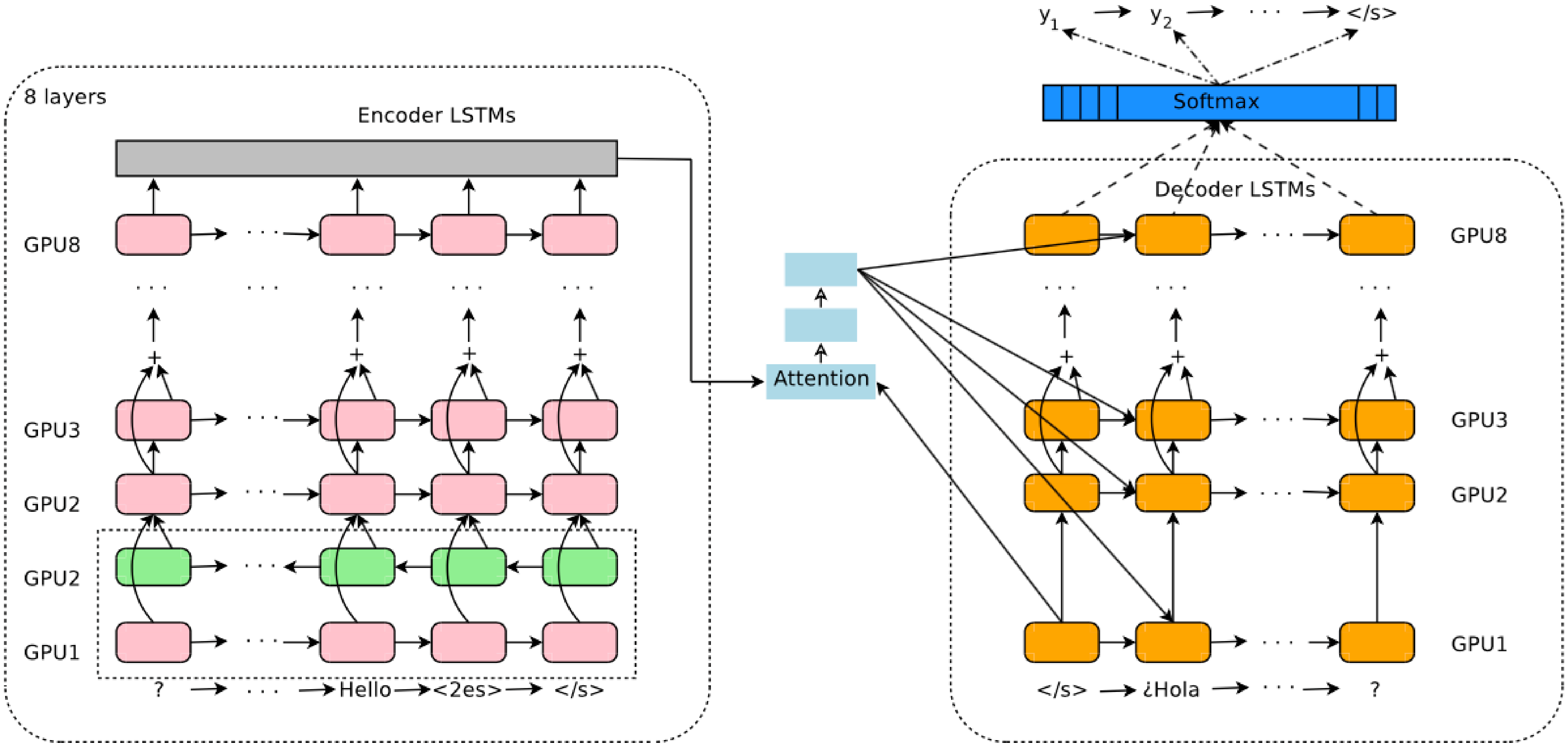
- Shared encoder, shared decoder for all languages

- Train on sentence pairs in all languages

- Add token to the input to mark target language

```
<2es> Hello, how are you? -> Hola, ¿cómo estás?
```

A standard encoder-decoder LSTM architecture, updated to enable parallelization/multi-GPU training



Pros and Cons?

Advantages

- Translation for low resource languages benefits from data for high resource languages
- Enables "zero shot" translation
    - Translation between language pairs which have not been seen (as a pair) during training
- Can handle code-switched input
    - Sequences that contain more than one language

Drawbacks/Issues

- Requires a single shared vocabulary for all languages
    - BPE, wordpiece
- Model size
- Opaque
- No direct control on output language
    - Bias toward high-resource languages?

How well does this work?
Evaluation Set Up

- WMT
  - Train
    - English↔French(Fr)
    - English↔German(De)
  - Test: newstest2014+15
- Google production
  - English↔Japanese(Ja)
  - English↔Korean(Ko)
  - English↔Spanish(Es)
  - English↔Portuguese(Pt)
- BLEU evaluation

BLEU scores in the “many to one” condition

Model	Single	Multi	Diff
WMT German→English (oversampling)	30.43	30.59	+0.16
WMT French→English (oversampling)	35.50	35.73	+0.23
WMT German→English (no oversampling)	30.43	30.54	+0.11
WMT French→English (no oversampling)	35.50	36.77	+1.27
Prod Japanese→English	23.41	23.87	+0.46
Prod Korean→English	25.42	25.47	+0.05
Prod Spanish→English	38.00	38.73	+0.73
Prod Portuguese→English	44.40	45.19	+0.79

Single language pair baseline

Multilingual model

BLEU scores in the "one to many" condition

Model	Single	Multi	Diff
WMT English→German (oversampling)	24.67	24.97	+0.30
WMT English→French (oversampling)	38.95	36.84	-2.11
WMT English→German (no oversampling)	24.67	22.61	-2.06
WMT English→French (no oversampling)	38.95	38.16	-0.79
Prod English→Japanese	23.66	23.73	+0.07
Prod English→Korean	19.75	19.58	-0.17
Prod English→Spanish	34.50	35.40	+0.90
Prod English→Portuguese	38.40	38.63	+0.23

Single language pair baseline

Multilingual model

BLEU scores in the “many to many” condition

Model	Single	Multi	Diff
WMT English→German (oversampling)	24.67	24.49	-0.18
WMT English→French (oversampling)	38.95	36.23	-2.72
WMT German→English (oversampling)	30.43	29.84	-0.59
WMT French→English (oversampling)	35.50	34.89	-0.61
WMT English→German (no oversampling)	24.67	21.92	-2.75
WMT English→French (no oversampling)	38.95	37.45	-1.50
WMT German→English (no oversampling)	30.43	29.22	-1.21
WMT French→English (no oversampling)	35.50	35.93	+0.43
Prod English→Japanese	23.66	23.12	-0.54
Prod English→Korean	19.75	19.73	-0.02
Prod Japanese→English	23.41	22.86	-0.55
Prod Korean→English	25.42	24.76	-0.66
Prod English→Spanish	34.50	34.69	+0.19
Prod English→Portuguese	38.40	37.25	-1.15
Prod Spanish→English	38.00	37.65	-0.35
Prod Portuguese→English	44.40	44.02	-0.38

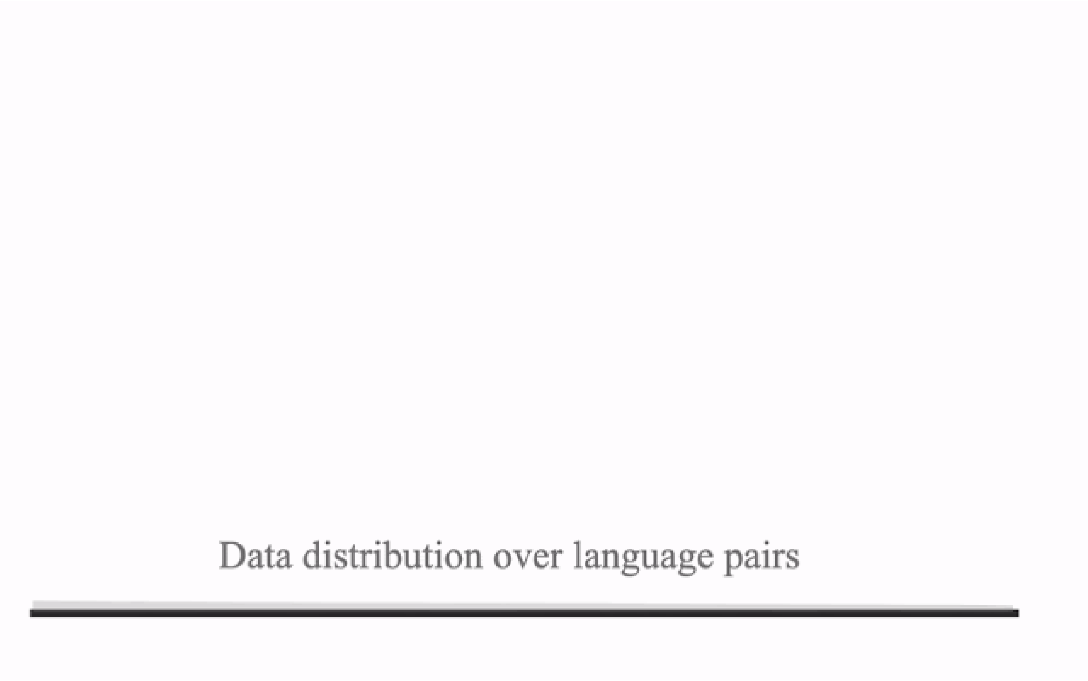
Impact of model size in “many to many” condition

Model	Single	Multi	Multi	Multi	Multi
#nodes	1024	1024	1280	1536	1792
#params	3B	255M	367M	499M	650M
Prod English→Japanese	23.66	21.10	21.17	21.72	21.70
Prod English→Korean	19.75	18.41	18.36	18.30	18.28
Prod Japanese→English	23.41	21.62	22.03	22.51	23.18
Prod Korean→English	25.42	22.87	23.46	24.00	24.67
Prod English→Spanish	34.50				
Prod English→Portuguese	38.40				
Prod Spanish→English	38.00				
Prod Portuguese→English	44.40				
Prod English→German	26.43				
Prod English→French	35.37				
Prod German→English	31.77				
Prod French→English	36.47				
ave diff	-				
vs single	-				

Findings so far: multilingual model

- can improve translation quality (BLEU) for low resource language pairs
- reduce training costs compared to training one model per language pair, at no (or little) loss in translation quality

Follow up work: evaluating multilingual models at scale

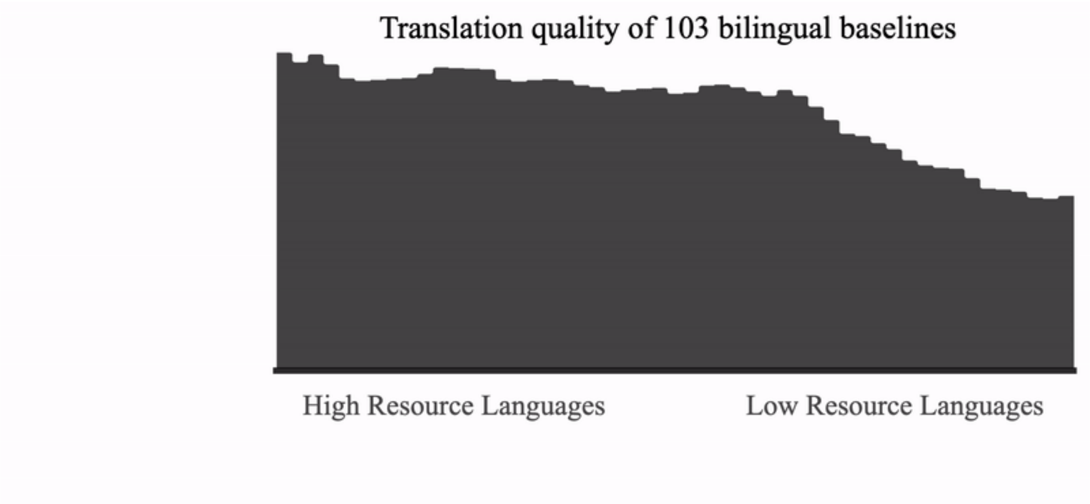


- 25+ billion sentence pairs
- from 100+ languages to and from English
- with 50+ billion parameters

- Comparing against strong bilingual baselines

https://ai.googleblog.com/2019/10/exploring-massively-multilingual.html

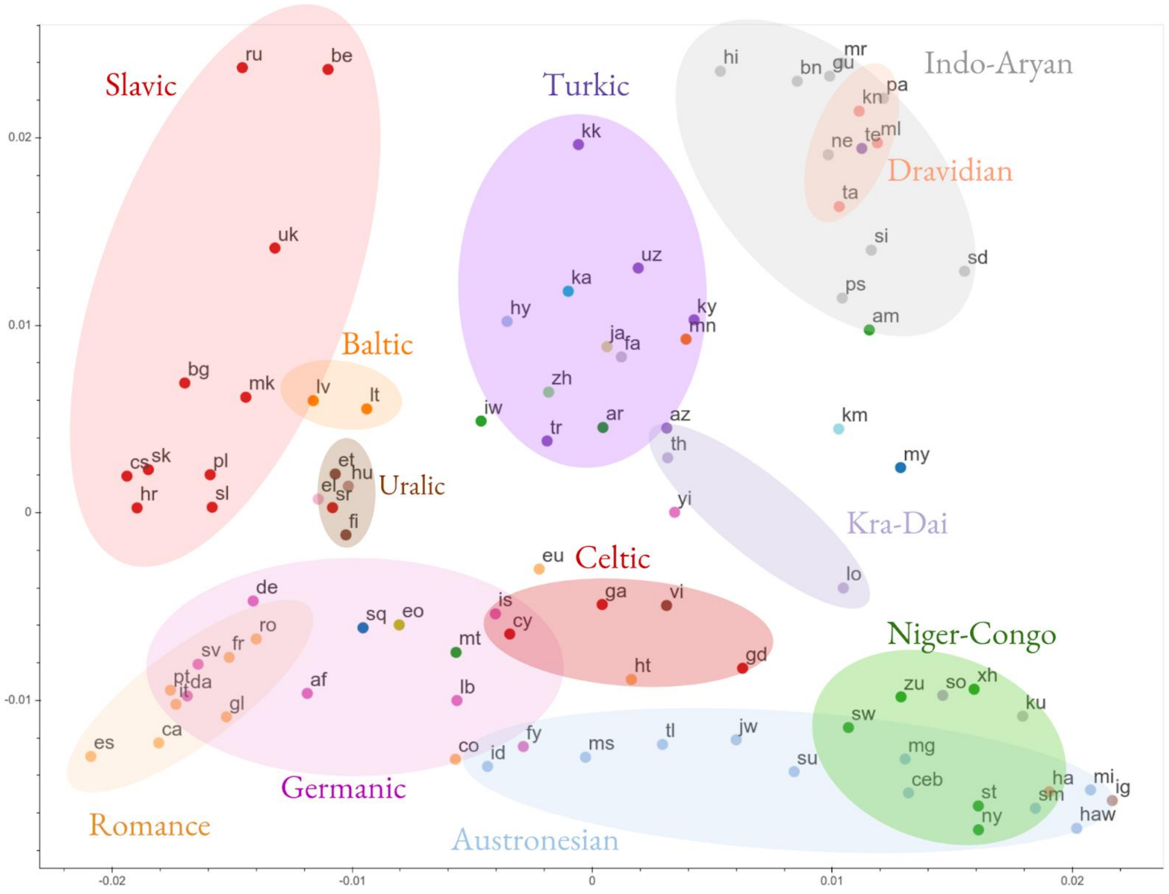
Follow up work: evaluating multilingual models at scale



Translation quality comparison of a single massively multilingual model against bilingual baselines that are trained for each one of the 103 language pairs.

- The multilingual model improves BLEU by 5 points (on average) for low-resource language pairs
- With multilingual and bilingual models of the same capacity (i.e. number of parameters)!
- Suggests that the multilingual model is able to transfer knowledge from high-resource to low-resource languages

Analysis: representations in multilingual model cluster by language family [Kudugunta et al. 2019]



Multilingual Machine Translation Summary

- A simple idea:
  - Shared model for all language pairs
  - Add a token to input to identify output language

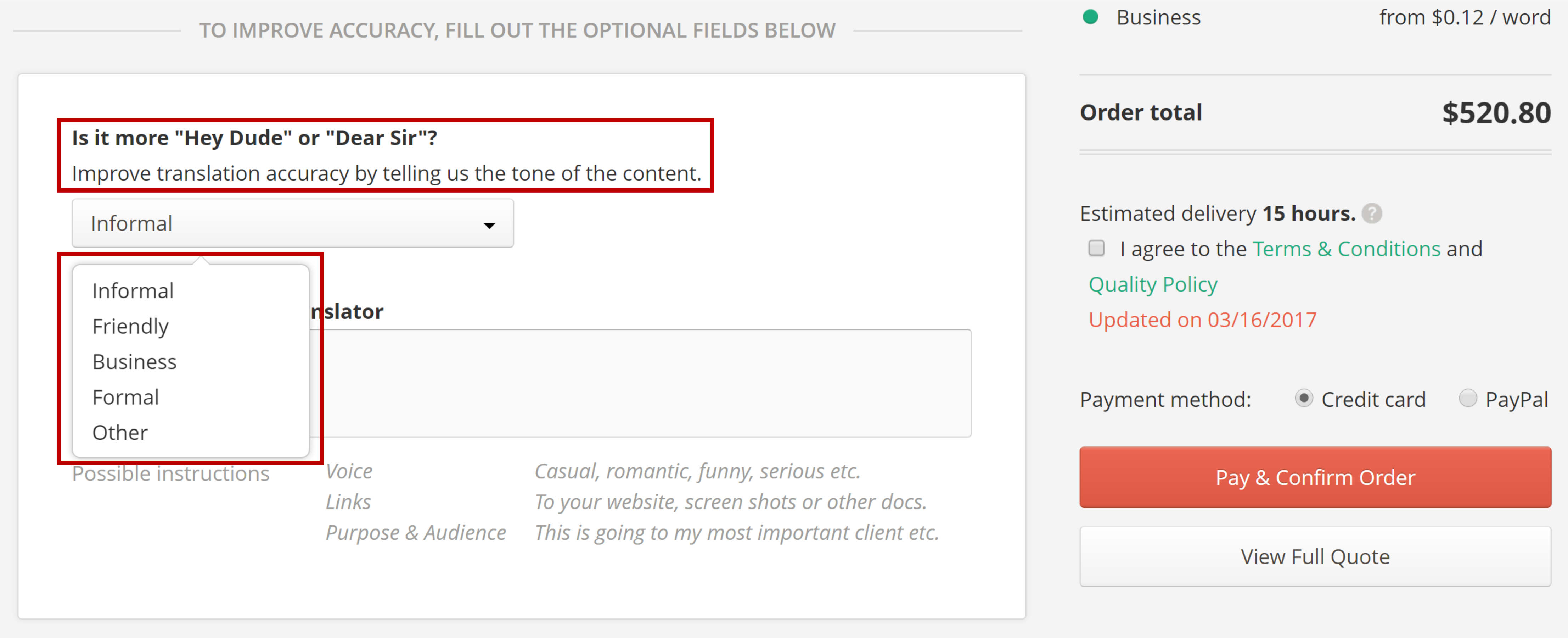
- Improves BLEU for low-resource language pairs

- But open questions remain
  - How to train massive models efficiently?
  - What properties are transferred from one language to another?
  - Are there unwanted effects on translation output? Bias toward high-resource languages / dominant language families?

Multitask Models for Controlling MT Output Style

Case Study I: formality

Style Matters for Translation



www.gengo.com

New Task: Formality-Sensitive Machine Translation (FSMT)

Comment ça va? Source ($\boldsymbol{f}$) → FSMT ($\boldsymbol{\theta}$) → Translation-1 ($\boldsymbol{e}_1$) How are you doing?

Desired formality level (ℓ) → or

Translation-2 ($\boldsymbol{e}_2$) What's up?

How to train?

$\boldsymbol{f}$	ℓ_1	$\boldsymbol{e}_1$
$\boldsymbol{f}$	ℓ_2	$\boldsymbol{e}_2$

Ideal training data doesn't occur naturally!

[Niu, Martindale & Carpuat, EMNLP 2017]

Formality in MT Corpora

Formal

Informal

Formal	delegates are kindly requested to bring their copies of documents to meetings .	[UN]
	in these centers , the children were fed , medically treated and rehabilitated on both a physical and mental level .	[OpenSubs]
	there can be no turning back the clock	[UN]
	I just wanted to introduce myself	[OpenSubs]
Informal	-yeah , bro , up top .	[OpenSubs]

Formality Transfer (FT)

What's up? EN Informal-Source → FT → Formal-Target EN How are you doing?

How are you doing? EN Formal-Source → FT → Informal-Target EN What's up?

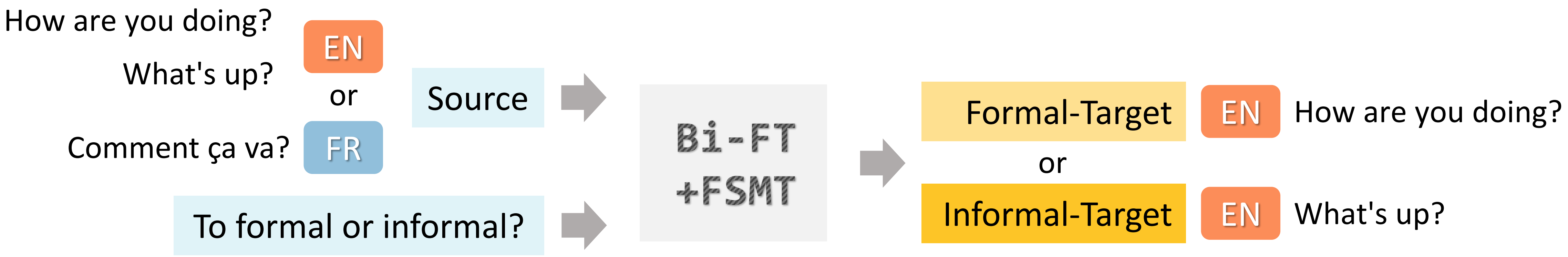
Given a large parallel formal-informal corpus

(e.g., Grammarly's Yahoo Answers Formality Corpus)

these are sequence-to-sequence tasks

[Rao and Tetreault, 2018]

Formality Sensitive MT as Multitask Formality Transfer + MT



Multitask Formality Transfer + MT

- Model: shared encoder, shared decoder as in multilingual NMT [Johnson et al. 2017]

- Training objective:
$$\mathcal{L}_{MT} + \mathcal{L}_{FT}$$

$$\mathcal{L}_{MT} = \sum_{(\boldsymbol{X},\boldsymbol{Y})} \log P(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\theta})$$ MT pairs

$$\mathcal{L}_{FT} = \sum_{(\boldsymbol{Y}_{\bar{\ell}},\boldsymbol{Y}_{\ell})} \log P(\boldsymbol{Y}_{\ell} \,|\, \boldsymbol{Y}_{\bar{\ell}}, \ell; \boldsymbol{\theta})$$ FT pairs

Formality Transfer MT
Human Evaluation

Model	**Formality Difference** Range =[0,2]	**Meaning Preservation** Range = [0,3]
MultiTask	0.35	2.95
Phrase-based MT + formality reranking [Niu & Carpuat 2017]	0.05	2.97

300 samples per model
3 judgments per sample
Protocol based on Rao & Tetreault

Multitask model makes more formality changes than re-ranking baseline

	Reference	Refrain from the commentary and respond to the question, Chief Toohey.
Formal	MultiTask	**You need to be quiet** and answer the question, Chief Toohey.
	Baseline	Please refrain from comment and **just** answer **the** question, **the** Tooheys's boss.
Informal	MultiTask	**Shut up** and answer the question, Chief Toohey.
	Baseline	Please refrain from comment and answer **my** question, Tooheys's boss.

Multitask model introduces more meaning errors than re-ranking baseline

	Reference	Try to file any additional motions as soon as you can.
Formal	MultiTask	You should try to introduce the **sharks** as soon as you can.
	Baseline	Try to introduce any additional requests as soon as you can.
Informal	MultiTask	Try to introduce **sharks** as soon as you can.
	Baseline	Try to introduce any additional requests as soon as you can.

Meaning errors can be addressed by introducing additional synthetic supervision
[Niu, PhD thesis 2019]

Controlling Machine Translation formality via multitask learning

- A multitask formality transfer + MT model
- Can produce distinct formal/informal translations of same input
- Introduces more formality rewrites, while roughly preserving meaning, esp. with synthetic supervision

Details:

- **Formality Style Transfer Within and Across Languages with Limited Supervision.** Xing Niu, PhD Thesis 2019.
- **Multi-task Neural Models for Translating Between Styles Within and Across Languages.** Xing Niu, Sudha Rao & Marine Carpuat. COLING 2018.
- **A Study of Style in Machine Translation: Controlling the Formality of Machine Translation Output.** Xing Niu, Marianna Martindale & Marine Carpuat. EMNLP 2017.

github.com/xingniu/multitask-ft-fsmt

Multitask Models for Controlling MT Output Style

Case Study II: Complexity

Our goal: control the complexity of MT output

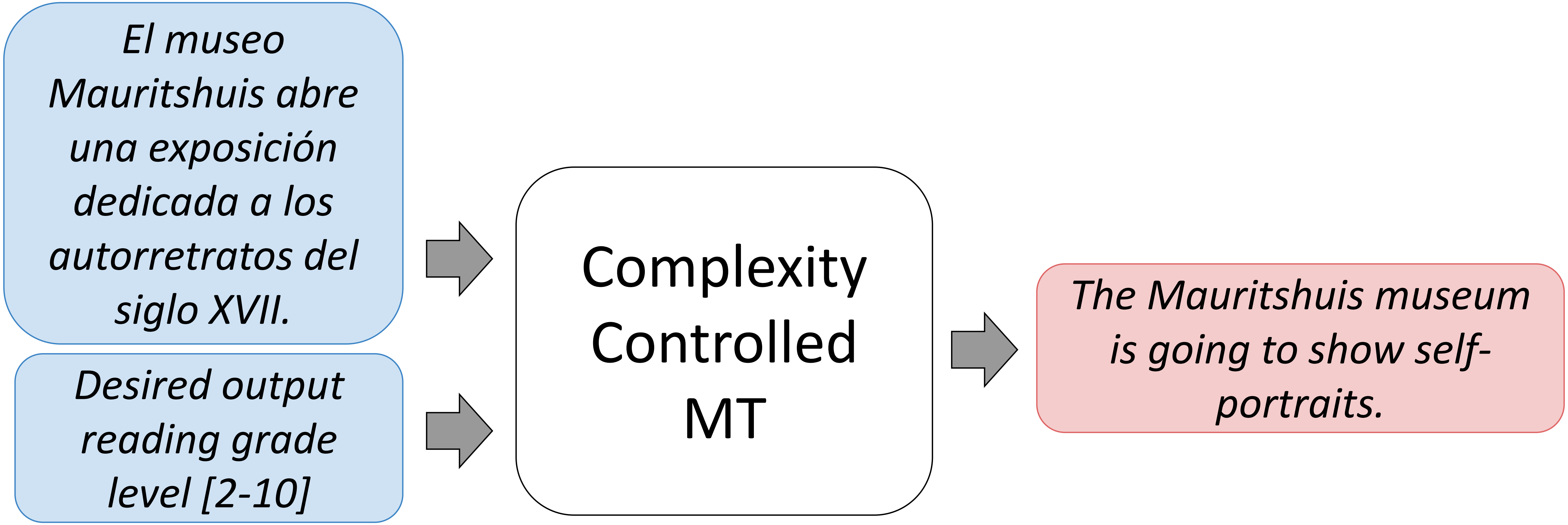
To make machine translation output accessible to broader audiences

Es: El museo Mauritshuis abre una exposición dedicada a los autorretratos del siglo XVII.

En (grade 8): The Mauritshuis museum is staging an exhibition focused solely on 17th century self-portraits.

En (grade 3): The Mauritshuis museum is going to show self-portraits.

Our goal: control the complexity of MT output



Summary

What you should know

- Multitask sequence-to-sequence models
    - How they are defined and trained (loss function)
- A simple yet powerful approach that can be applied to many translation and related sequence-to-sequence tasks
    - Can help improve performance by sharing data from multiple tasks
    - Has been applied to multilingual MT, style controlled MT, among other tasks

Also in discussing recent research papers, we illustrated:

- Pros and cons of automatic vs. manual evaluation
- Experiment design and result interpretation