Evaluating machine translation systems

Evaluating systems

Evaluating systems

Someone gives you a program that computes Fibonacci numbers.

Evaluating systems

Someone gives you a program that computes Fibonacci numbers.

How would you decide if the implementation is correct?

Evaluating systems

Evaluating systems

Someone gives you a python interpreter.

Evaluating systems

Someone gives you a python interpreter.

How would you decide if the implementation is correct?

Evaluating systems

Evaluating systems

Someone gives you an automatic speech recognition system

Evaluating systems

Someone gives you an automatic speech recognition system

How would you decide if the implementation is correct?

Evaluating systems

Evaluating systems

Someone gives you a self-driving car.

Evaluating systems

Someone gives you a self-driving car.

How would you decide if the implementation is correct?

Evaluating systems

Someone gives you a self-driving car.

How would you decide if the implementation is correct?

What does it mean for an implementation to be correct?

Evaluating systems

What does it mean for an implementation to be correct?

Evaluating systems

Evaluating systems

Someone gives you a machine translation system.

Evaluating systems

Someone gives you a machine translation system.

How would you know if the implementation is correct?

Evaluating systems

Someone gives you a machine translation system.

How would you know if the implementation is correct?

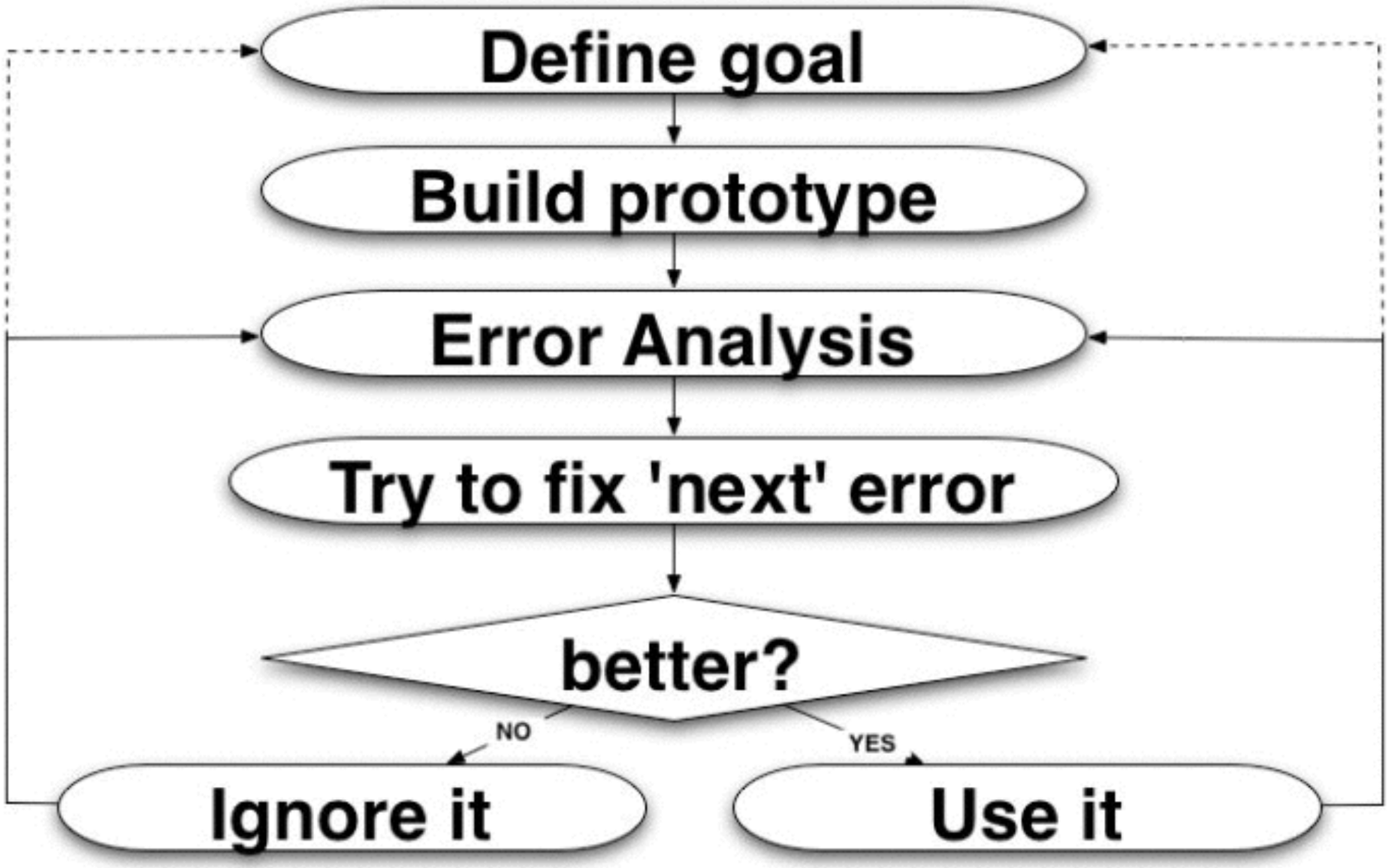
What does it mean for an implementation to be correct?

Yorick Wilks

More has been written about machine translation evaluation than about machine translation itself.

- Why evaluate?
  - Rank systems. Which one should I use?
  - Evaluate incremental changes. Does a new idea improve the results?
    - Should every idea be assessed the same way?
- Ideally, evaluation should be repeatable.
  - Is this possible?

Development Cycle for MT Research

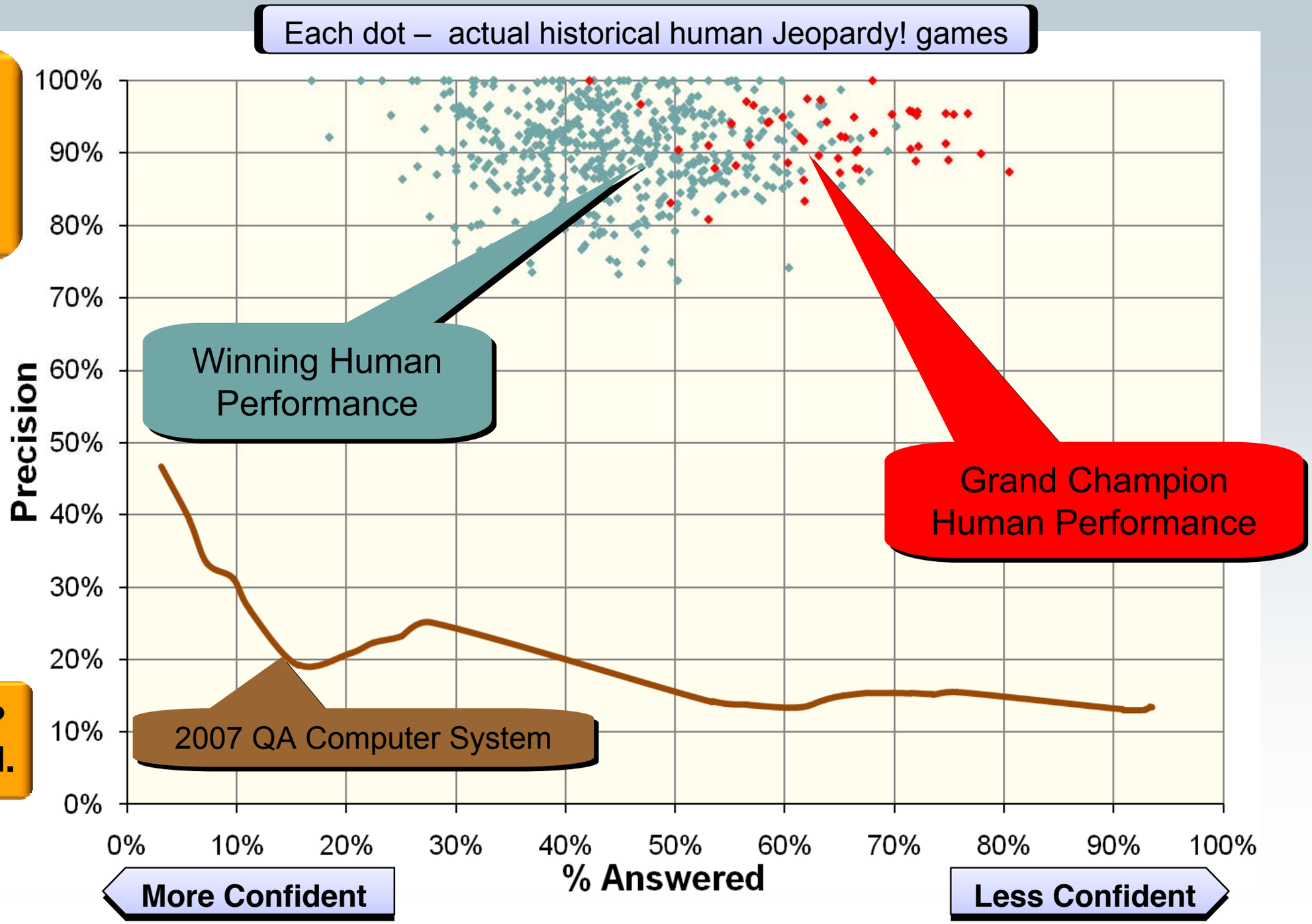


Google

What It Takes to compete against Top Human Jeopardy! Players

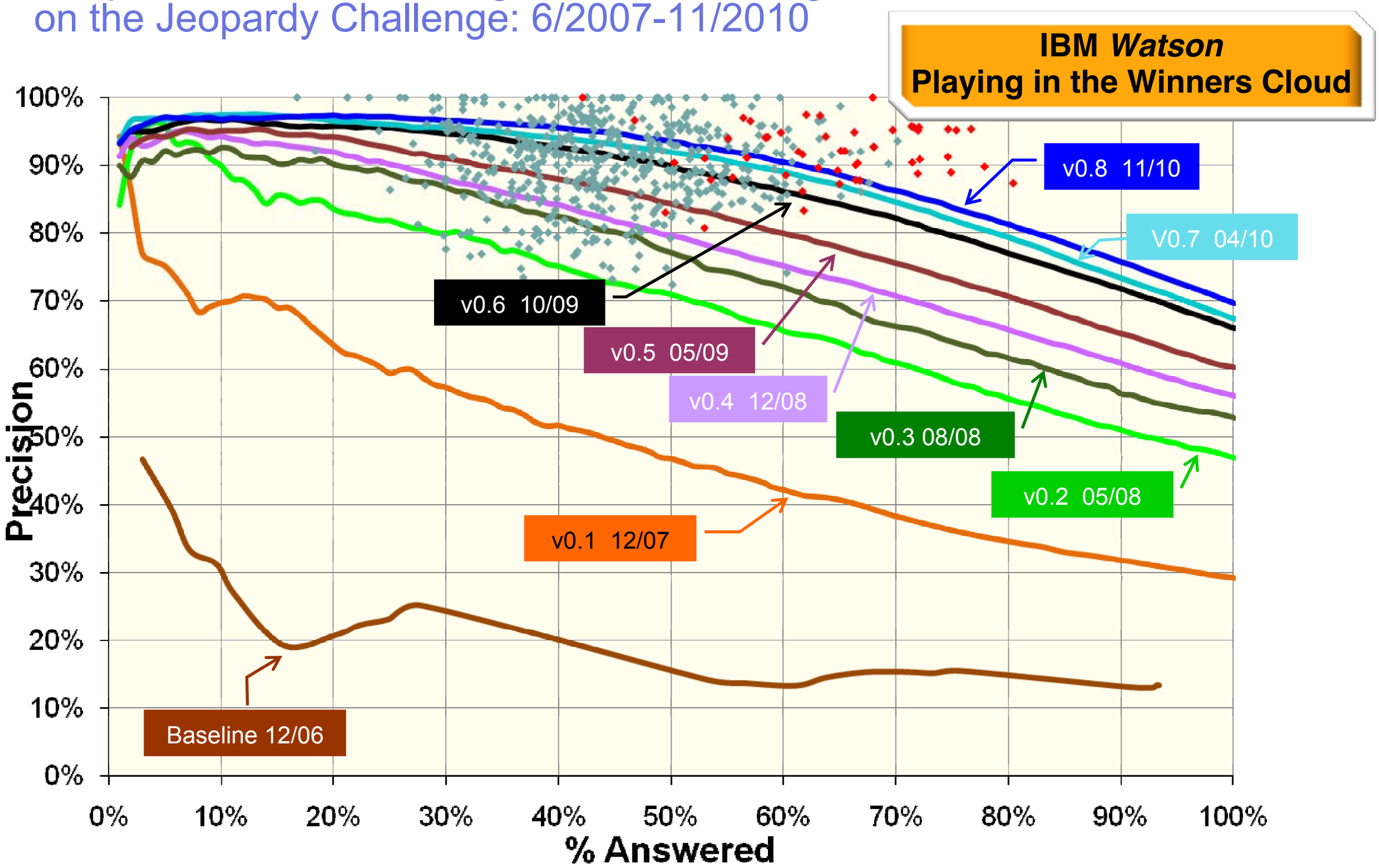
*Our Analysis Reveals the **Winner's Cloud***

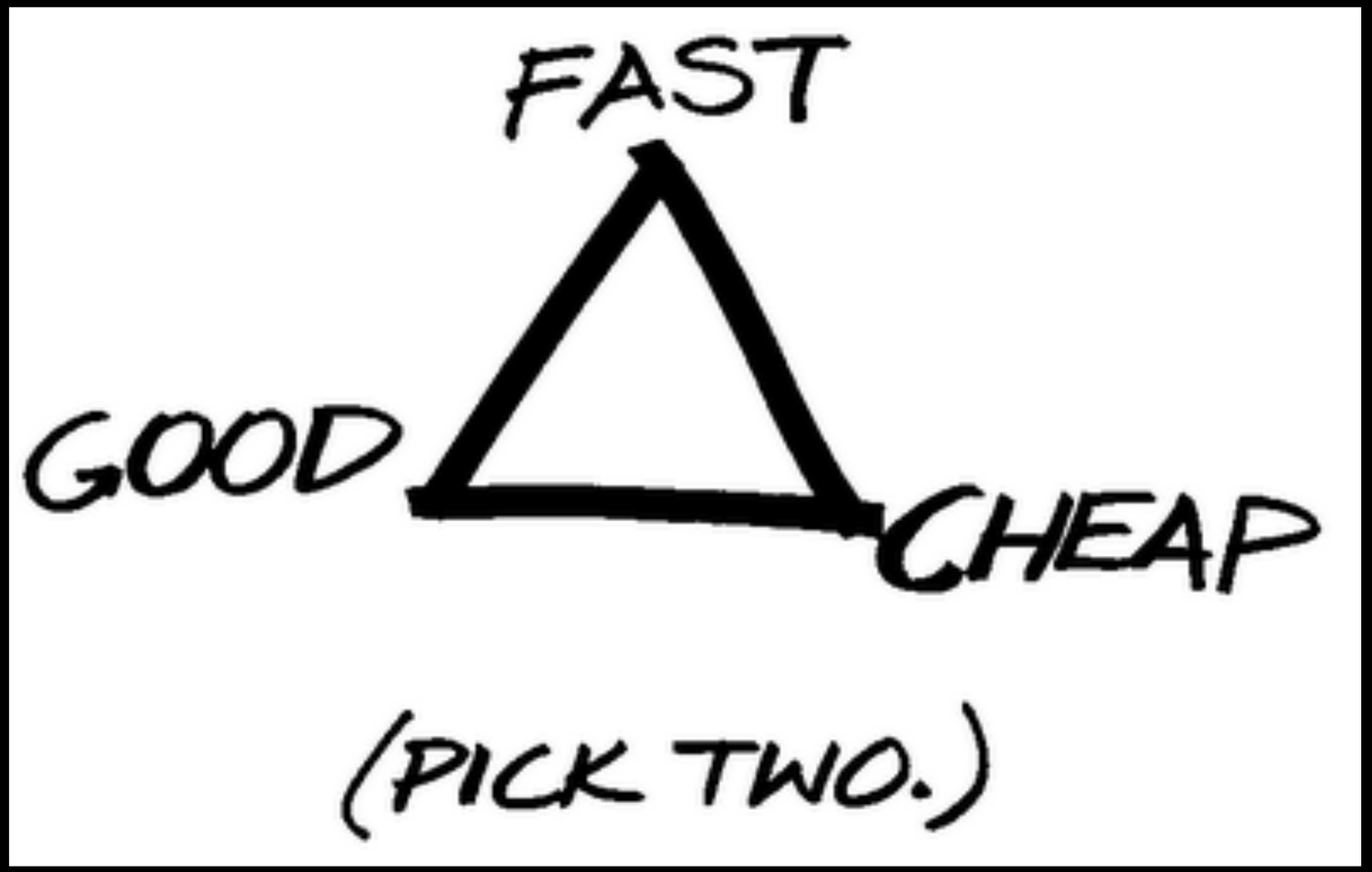
Top human players are *remarkably* good.

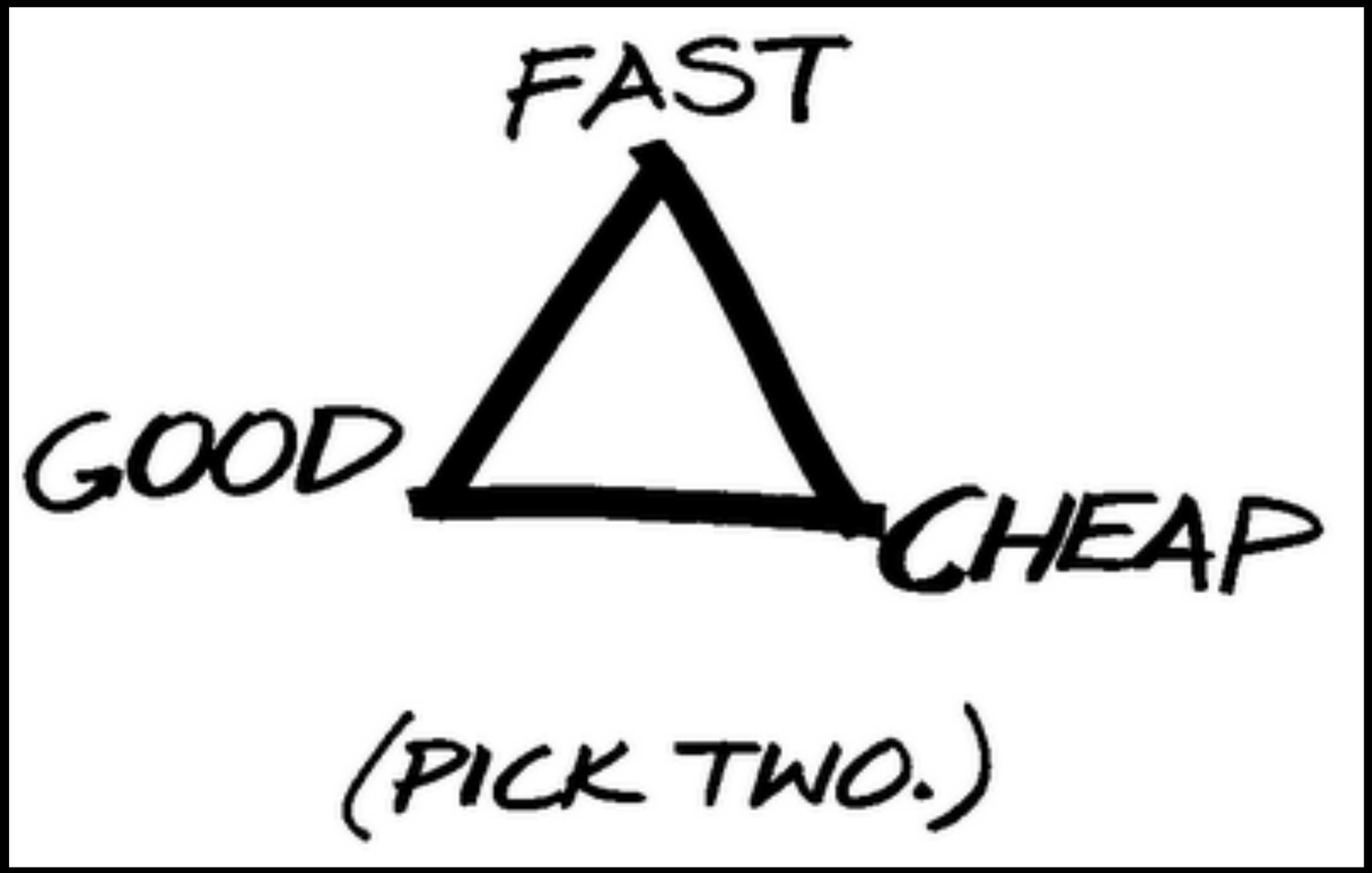


Computers? Not So Good.

DeepQA: Incremental Progress in Answering Precision on the Jeopardy Challenge: 6/2007-11/2010




FAST
GOOD
CHEAP
(PICK TWO.)



Claim: evaluation by humans is good and ???

Chinese people in the traditional Spring Festival is approaching, the CPC Central Committee this afternoon in Zhongnanhai on the 22nd non-Party personages to convene a forum in Spring Festival, invited the central committees of democratic parties, the leadership of the National Federation of Industry and Commerce and personages without party affiliation on behalf of comrades gathered together State yes, talked in length about the friendship, to greet the Chinese New Year. CPC Central Committee General Secretary and State President and Central Military Commission Chairman Hu Jintao on behalf of the CPC Central Committee, the State Council, to the central committees of democratic parties, leaders of the National Federation of Industry and Commerce and personages without party affiliation, to members of the united front, to extend my New Year's blessing.

Design of the WMT Evaluation (2008-2011)

Design of the WMT Evaluation (2008-2011)

Design of the WMT Evaluation (2008-2011)

system A

system B

system C

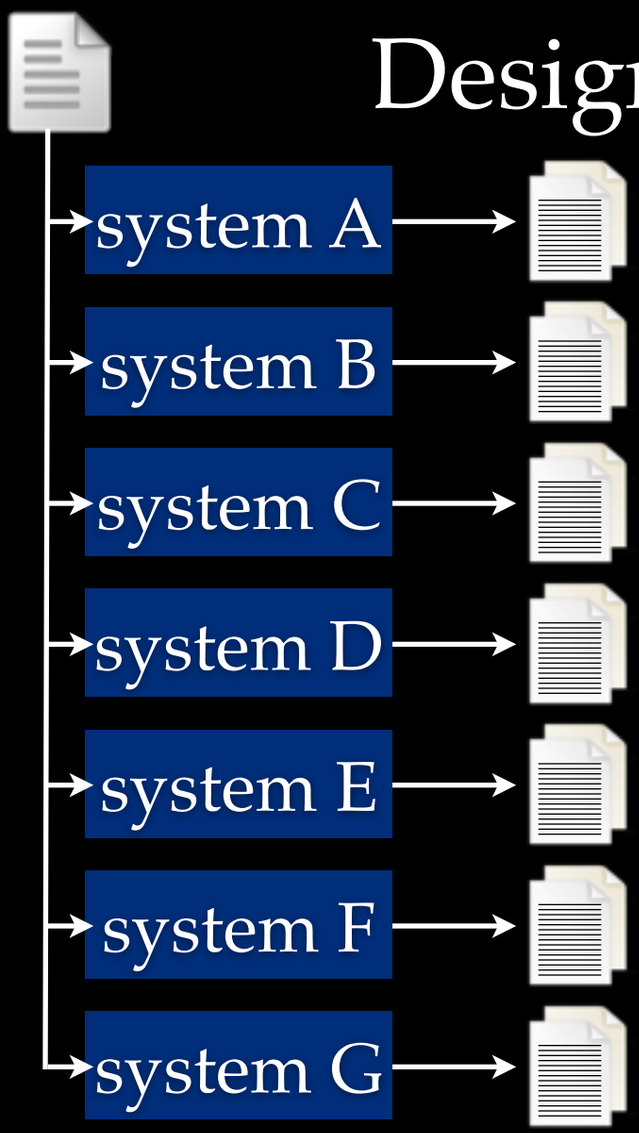
system D

system E

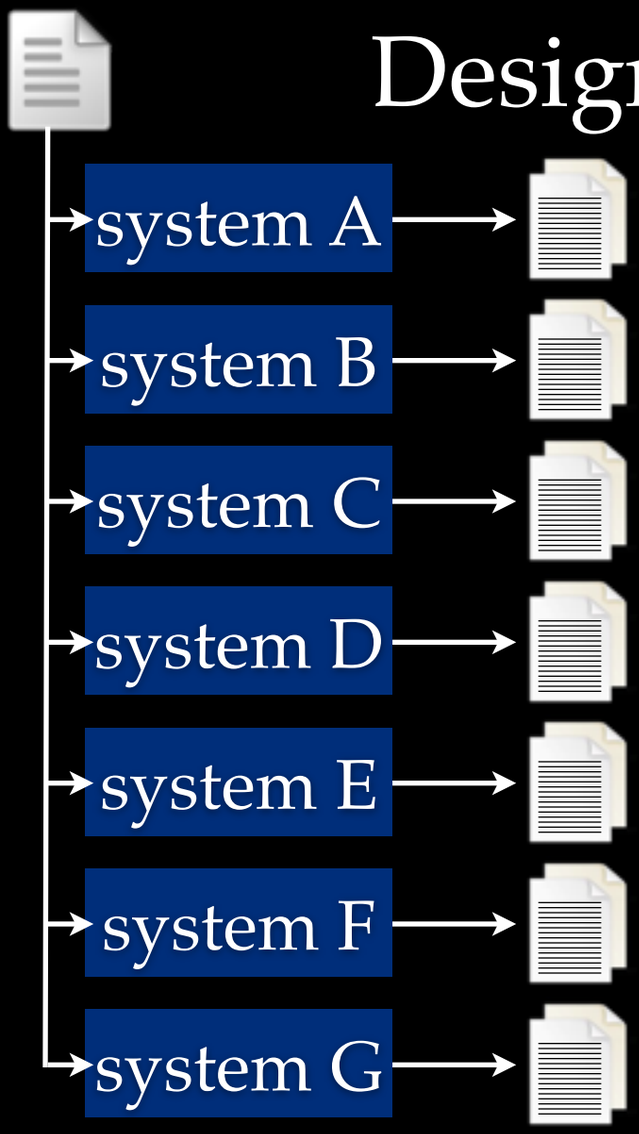
system F

system G

Design of the WMT Evaluation (2008-2011)

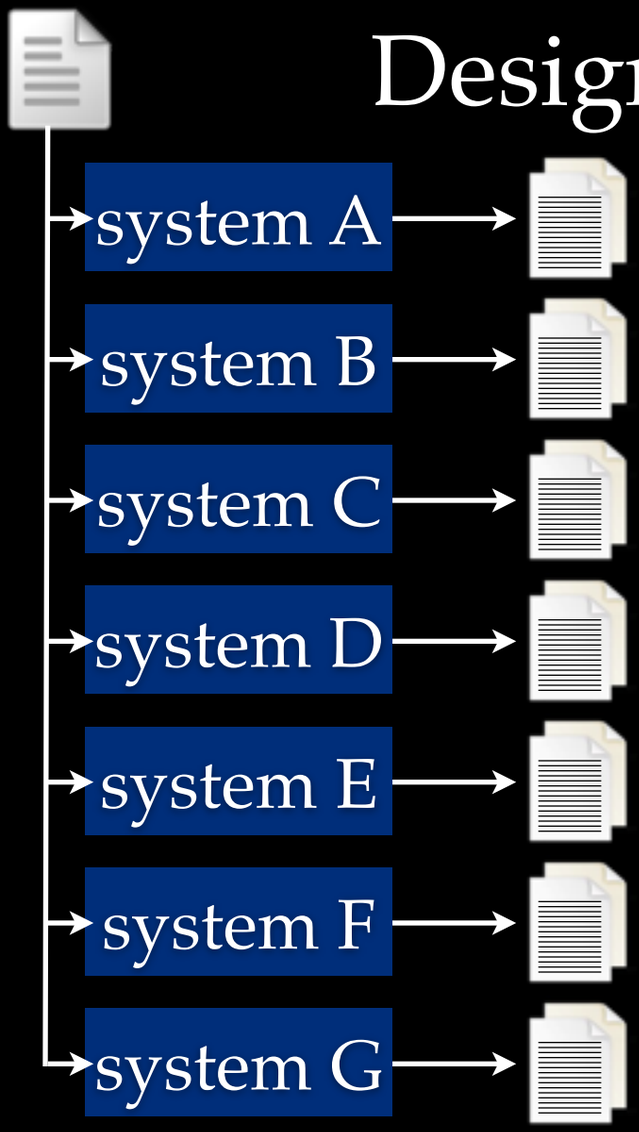


Design of the WMT Evaluation (2008-2011)



1. system C
2. system D
3. system A
4. system B
5. system G
6. system F
7. system E

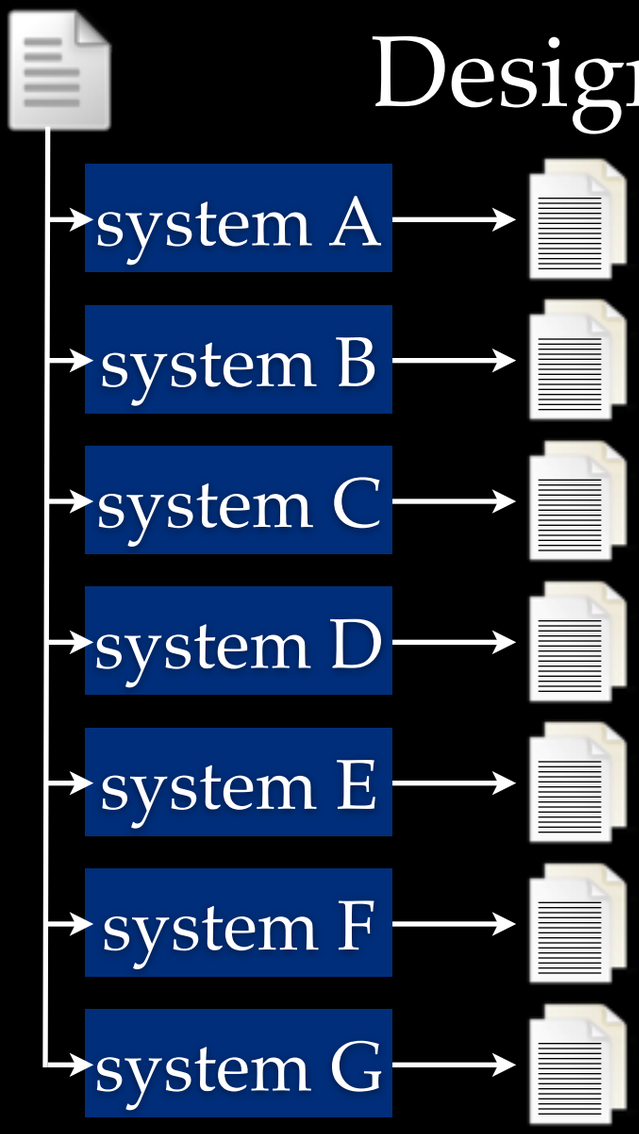
Design of the WMT Evaluation (2008-2011)



1. system C
2. system D
3. system A
4. system B
5. system G
6. system F
7. system E

➡ Costly: 361 hours of human effort in 2011.

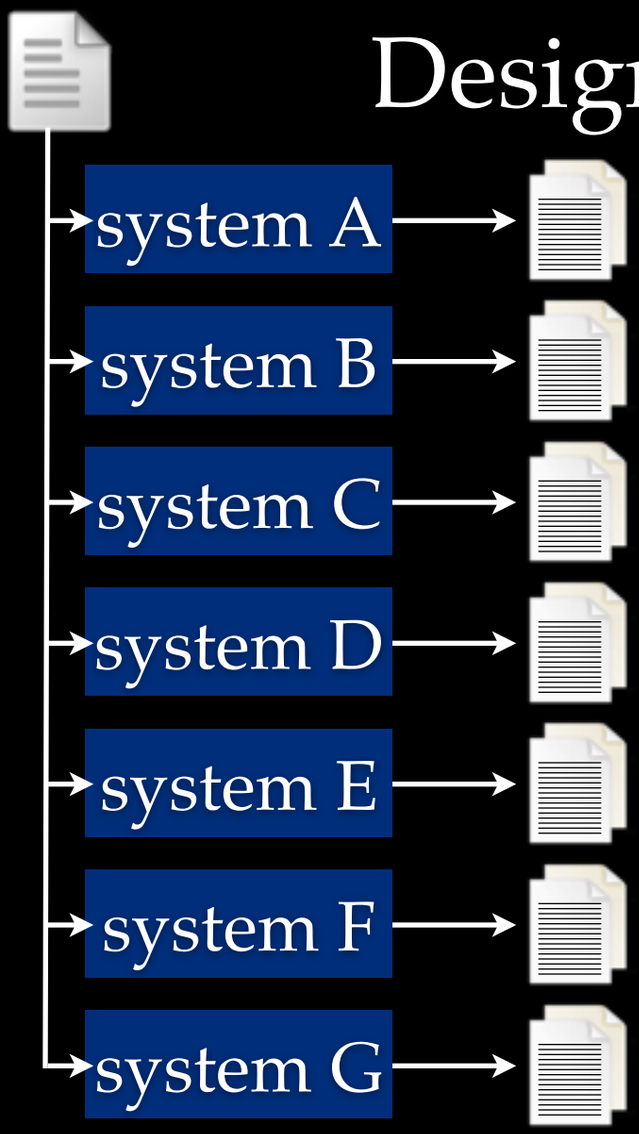
Design of the WMT Evaluation (2008-2011)



1. system C
2. system D
3. system A
4. system B
5. system G
6. system F
7. system E

Are you ***sure*** this is the correct ranking?

Design of the WMT Evaluation (2008-2011)

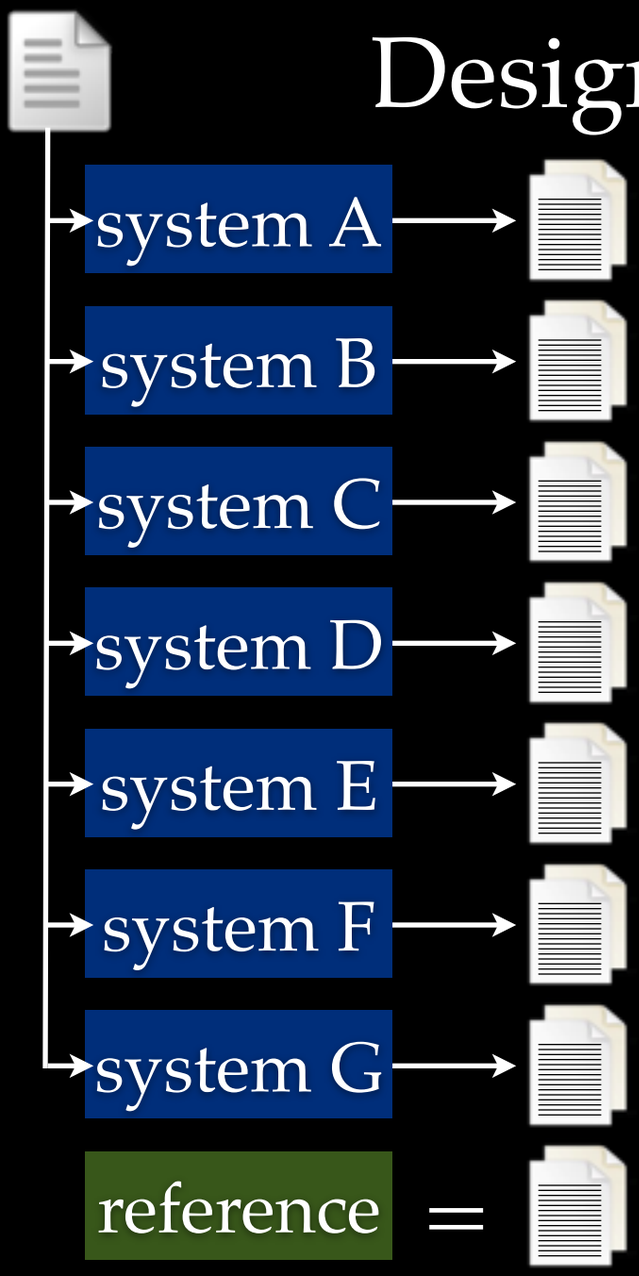


1. system C
2. system D
3. system A
4. system B
5. system G
6. system F
7. system E

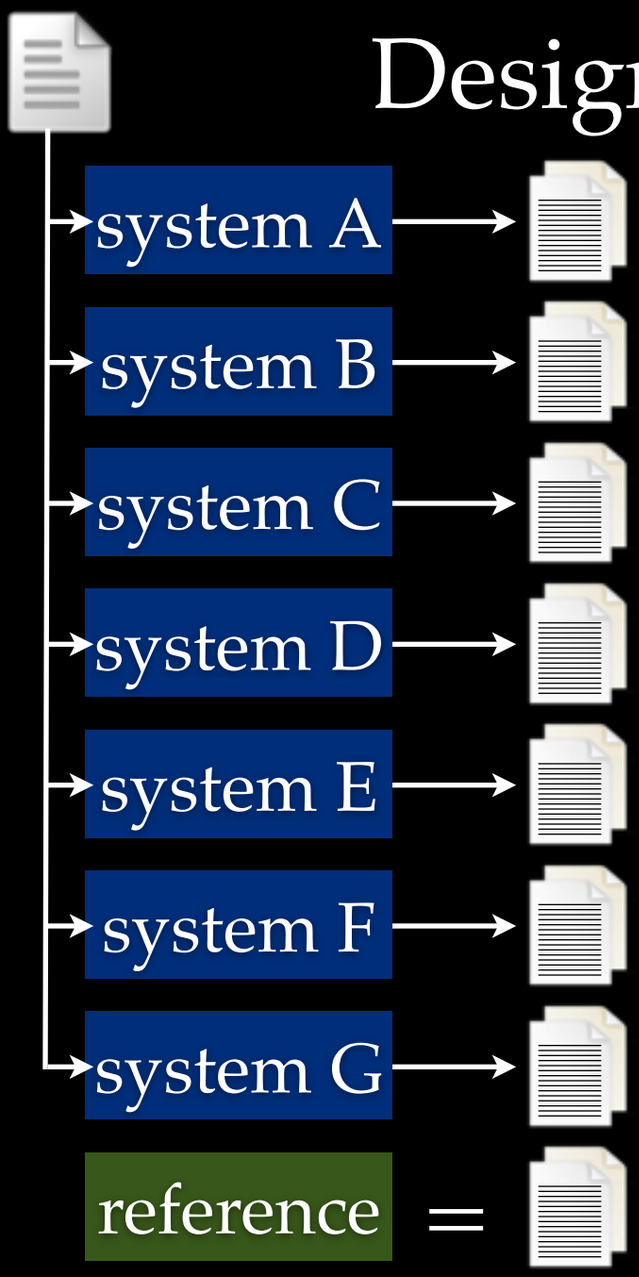
Are you ***sure*** this is the correct ranking?

- In above example, there are 5040 possible rankings.
- With 10 systems: 3 million possible rankings.
- With 20 systems: 2 quintillion possible rankings.

Design of the WMT Evaluation (2008-2011)

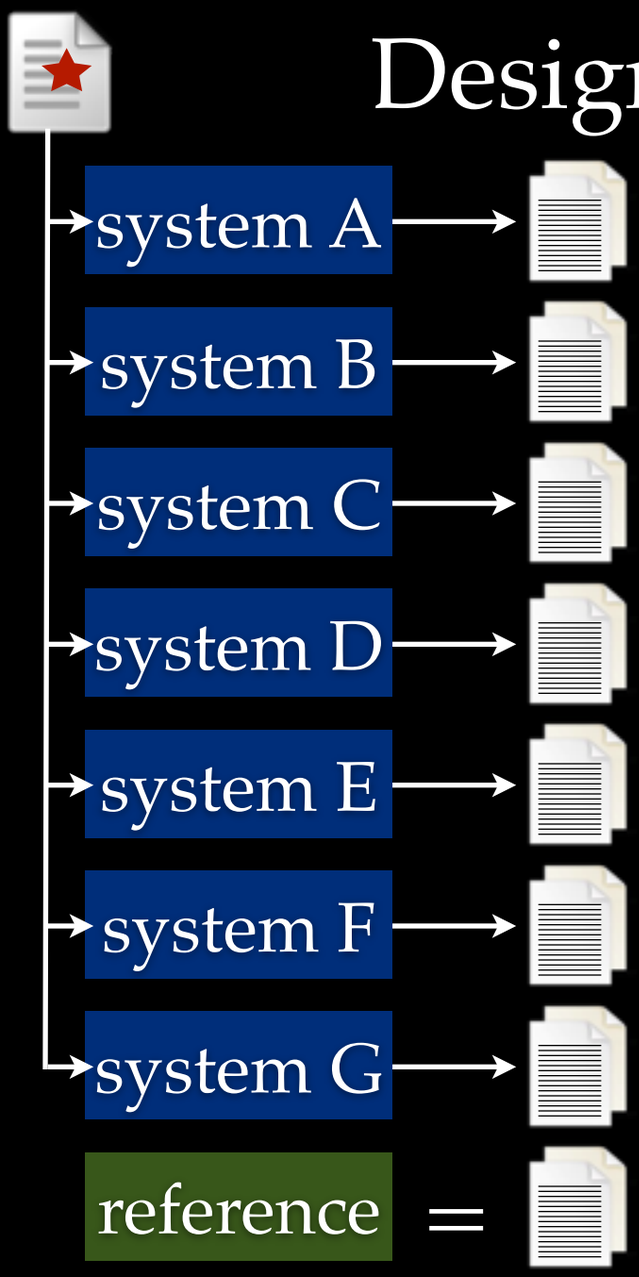


Design of the WMT Evaluation (2008-2011)



While (evaluation period is not over):

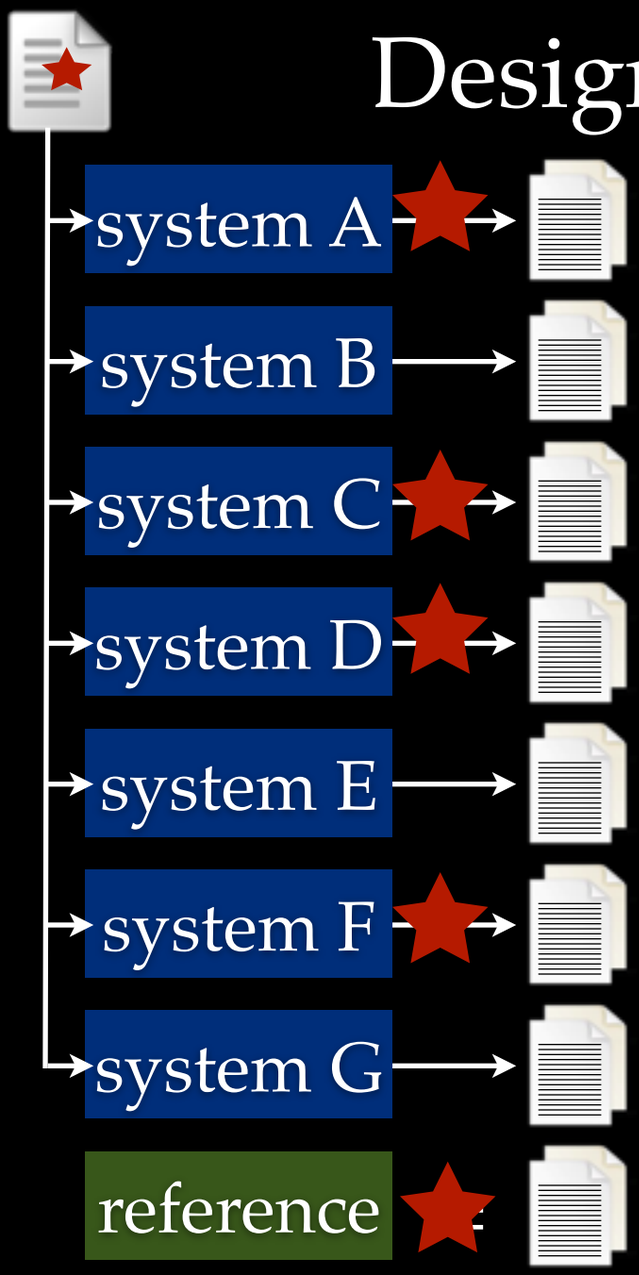
Design of the WMT Evaluation (2008-2011)



While (evaluation period is not over):

➡ Sample input sentence.

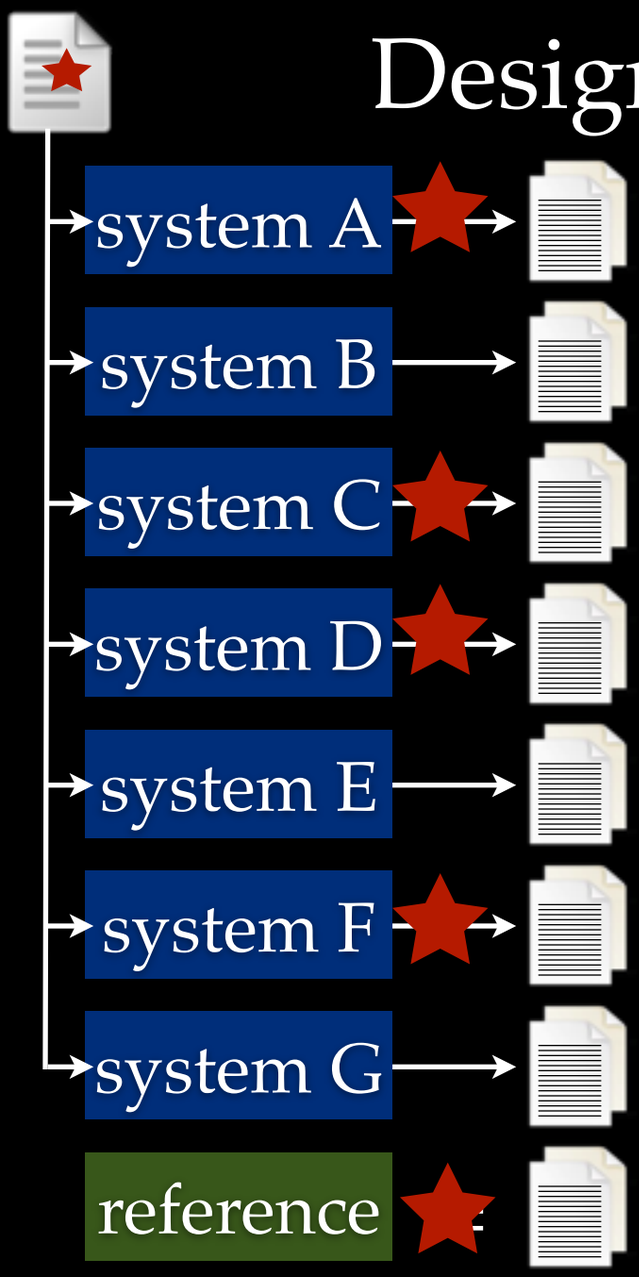
Design of the WMT Evaluation (2008-2011)



While (evaluation period is not over):

- ➡ Sample input sentence.
- ➡ Sample five translators of it from *Systems* ∪ *{Reference}*.

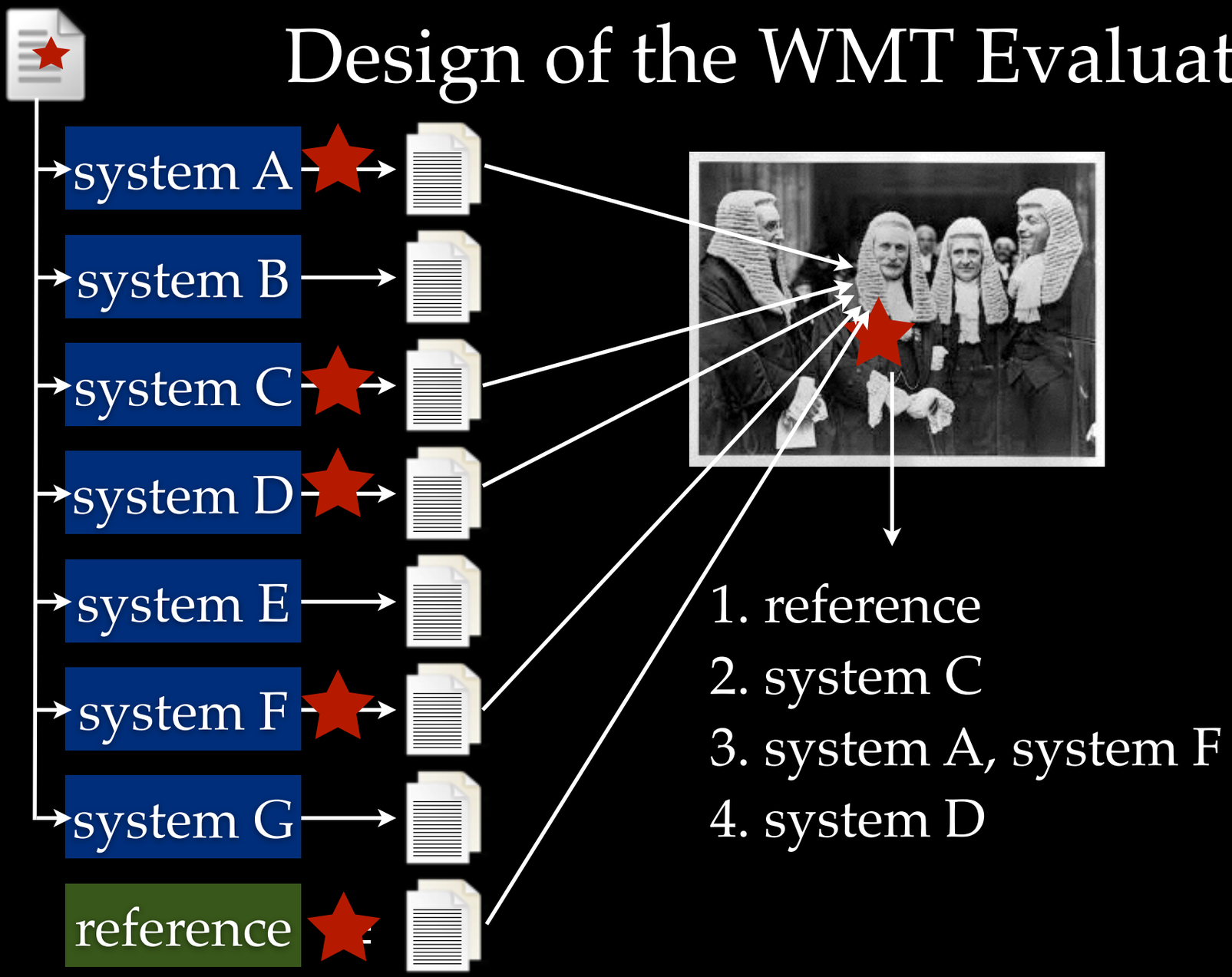
Design of the WMT Evaluation (2008-2011)



While (evaluation period is not over):

- ➡ Sample input sentence.
- ➡ Sample five translators of it from *Systems* ∪ *{Reference}*.
- ➡ Sample an assessor.

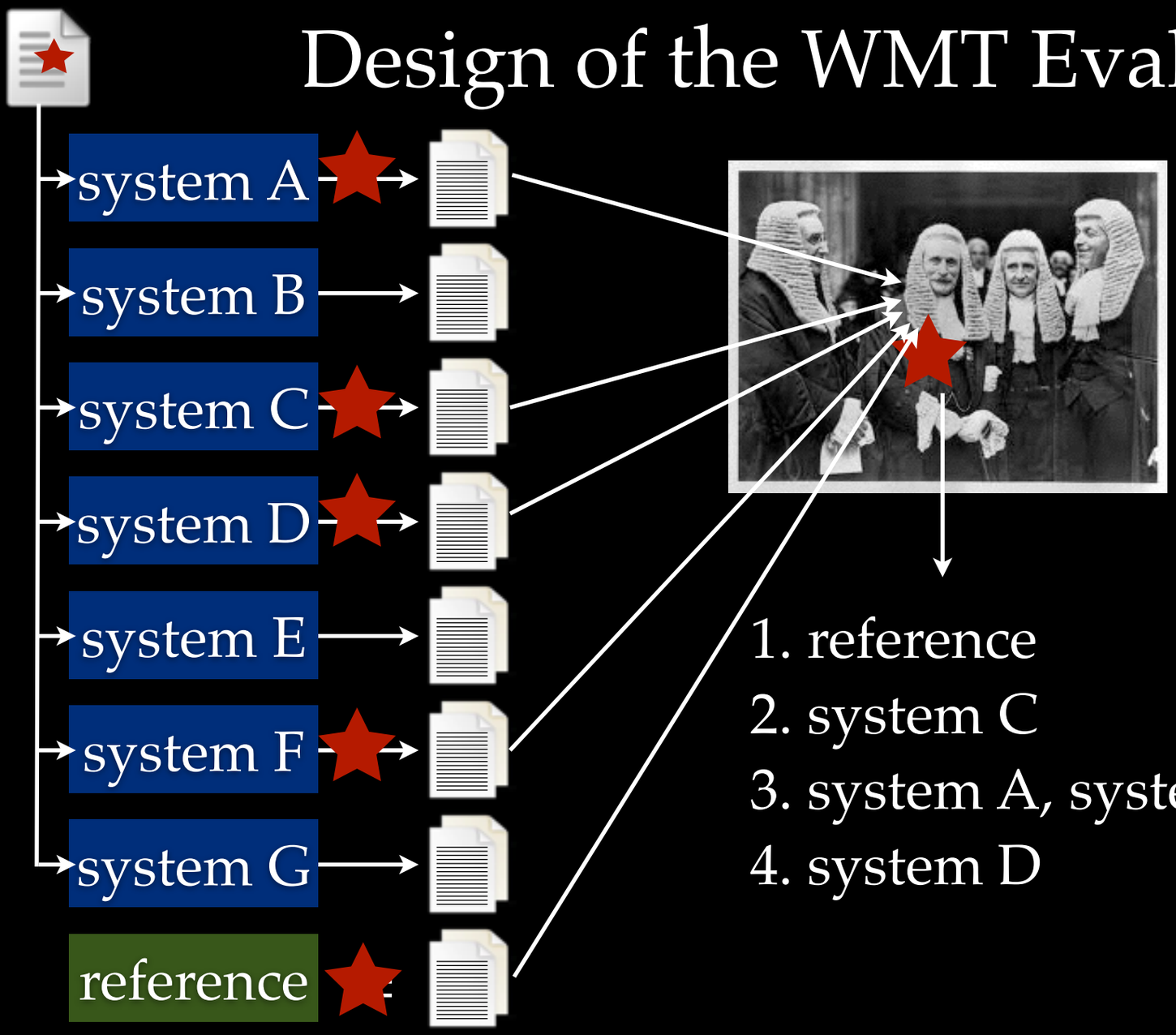
Design of the WMT Evaluation (2008-2011)



While (evaluation period is not over):

- ➡ Sample input sentence.
- ➡ Sample five translators of it from *Systems* ∪ *{Reference}*.
- ➡ Sample an assessor.
- ➡ Receive (partial) ranking of translations from assessor.

Design of the WMT Evaluation (2008-2011)



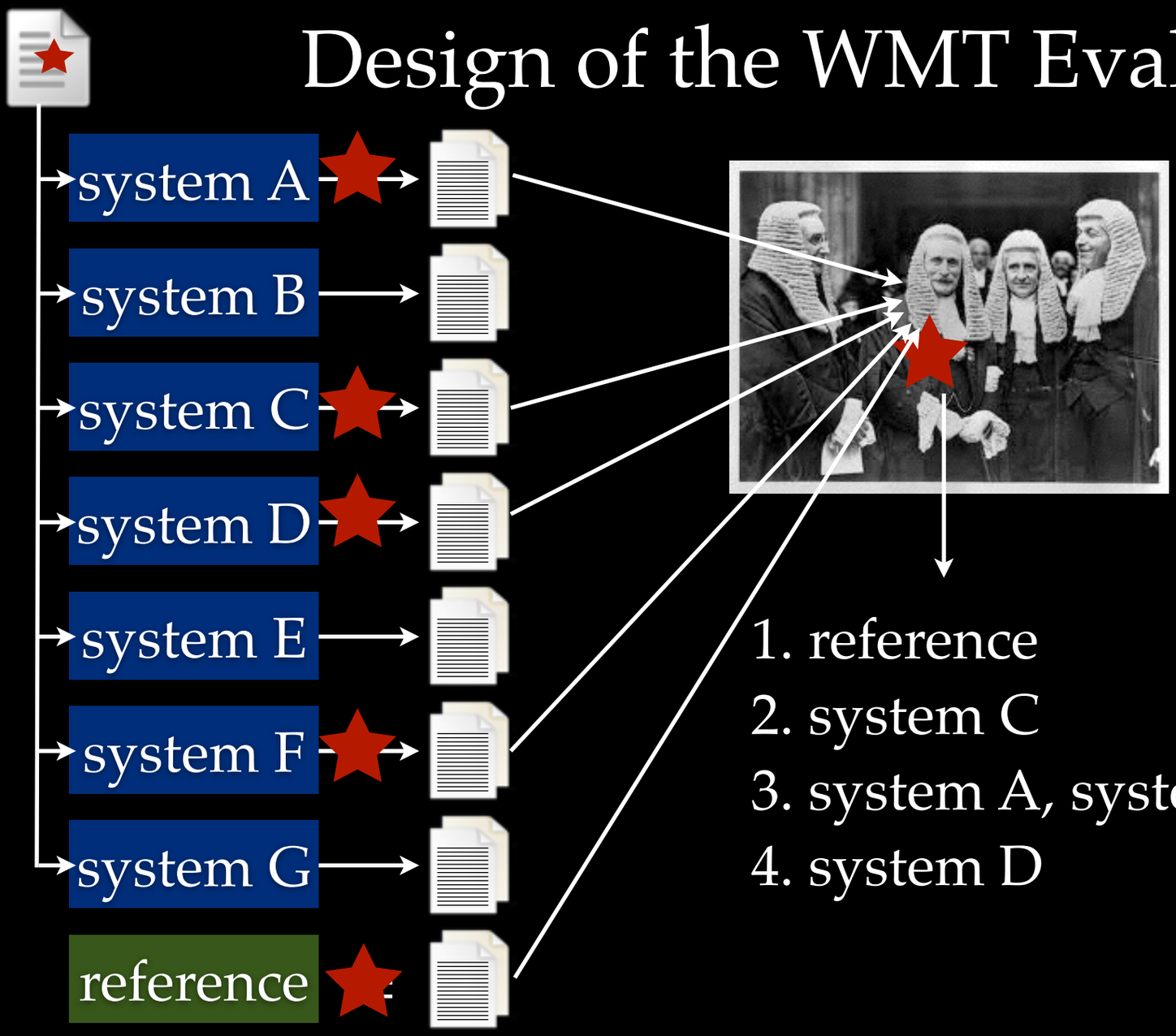
1. reference
2. system C
3. system A, system F
4. system D

$$\left\{\begin{array}{l}
\text{reference} \prec \text{system A} \\
\text{reference} \prec \text{system C} \\
\text{reference} \prec \text{system D} \\
\text{reference} \prec \text{system F} \\
\text{system A} \succ \text{system C} \\
\text{system A} \prec \text{system D} \\
\text{system A} \equiv \text{system F} \\
\text{system C} \prec \text{system D} \\
\text{system C} \prec \text{system F} \\
\text{system D} \prec \text{system F}
\end{array}\right.$$

While (evaluation period is not over):

- ➡ Sample input sentence.
- ➡ Sample five translators of it from *Systems* ∪ *{Reference}*.
- ➡ Sample an assessor.
- ➡ Receive (partial) ranking of translations from assessor.

Design of the WMT Evaluation (2008-2011)



1. reference
2. system C
3. system A, system F
4. system D

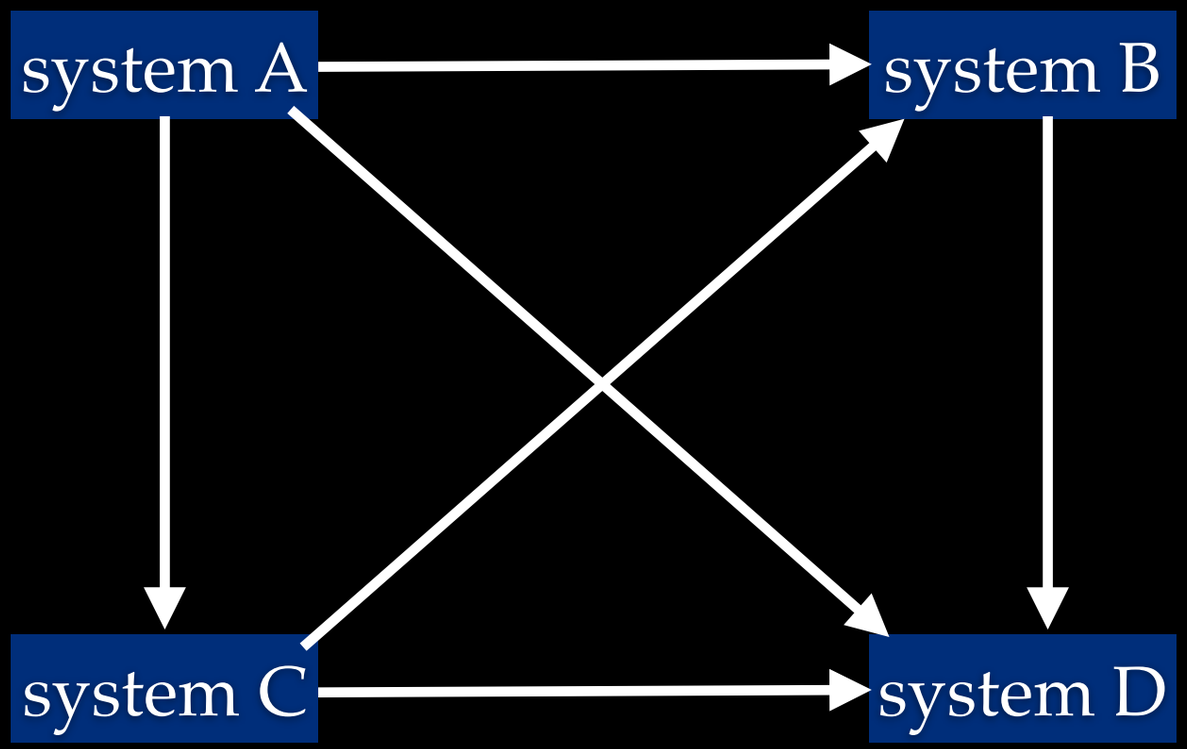
WMT Raw Data: pairwise rankings

$$
\begin{cases}
\text{reference} \prec \text{system A} \\
\text{reference} \prec \text{system C} \\
\text{reference} \prec \text{system D} \\
\text{reference} \prec \text{system F} \\
\text{system A} \succ \text{system C} \\
\text{system A} \prec \text{system D} \\
\text{system A} \equiv \text{system F} \\
\text{system C} \prec \text{system D} \\
\text{system C} \prec \text{system F} \\
\text{system D} \prec \text{system F}
\end{cases}
$$

While (evaluation period is not over):

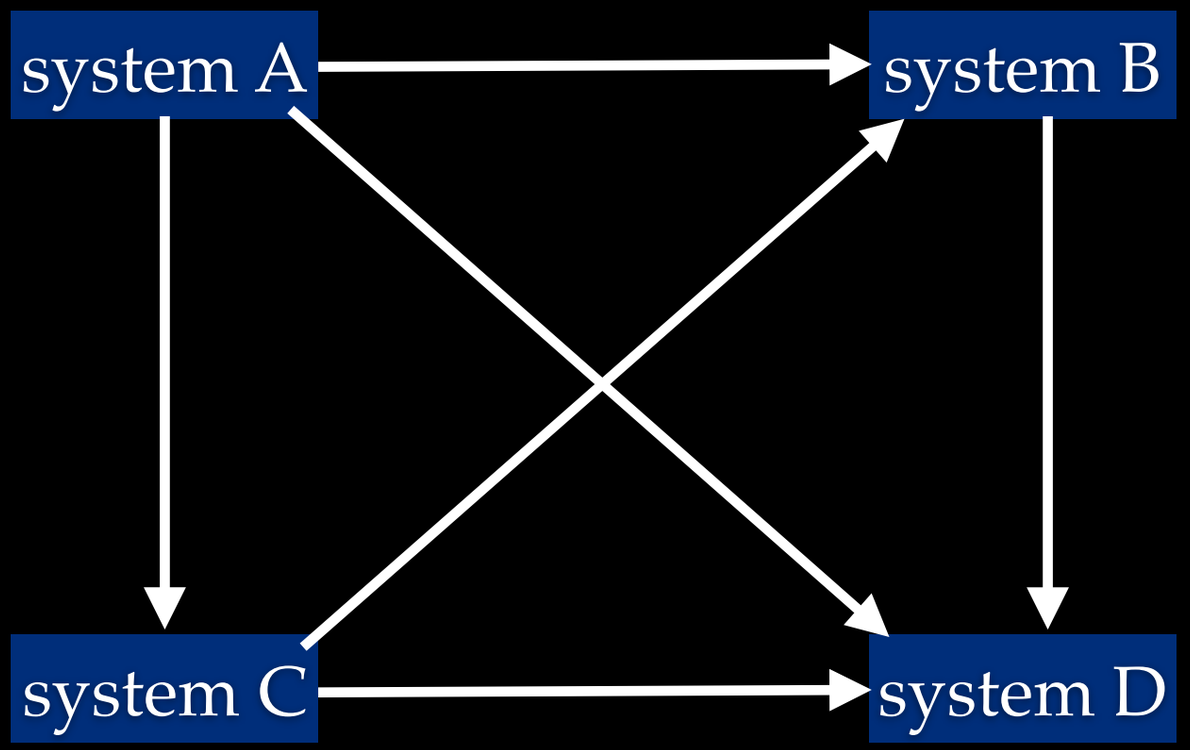
- ➡ Sample input sentence.
- ➡ Sample five translators of it from *Systems* ∪ *{Reference}*.
- ➡ Sample an assessor.
- ➡ Receive (partial) ranking of translations from assessor.

Tournaments



- Directed edge between every pair of vertices.
- Edge from A to B if A beats B in pairwise comparison.
- Widely used to model: sports, web results, elections.

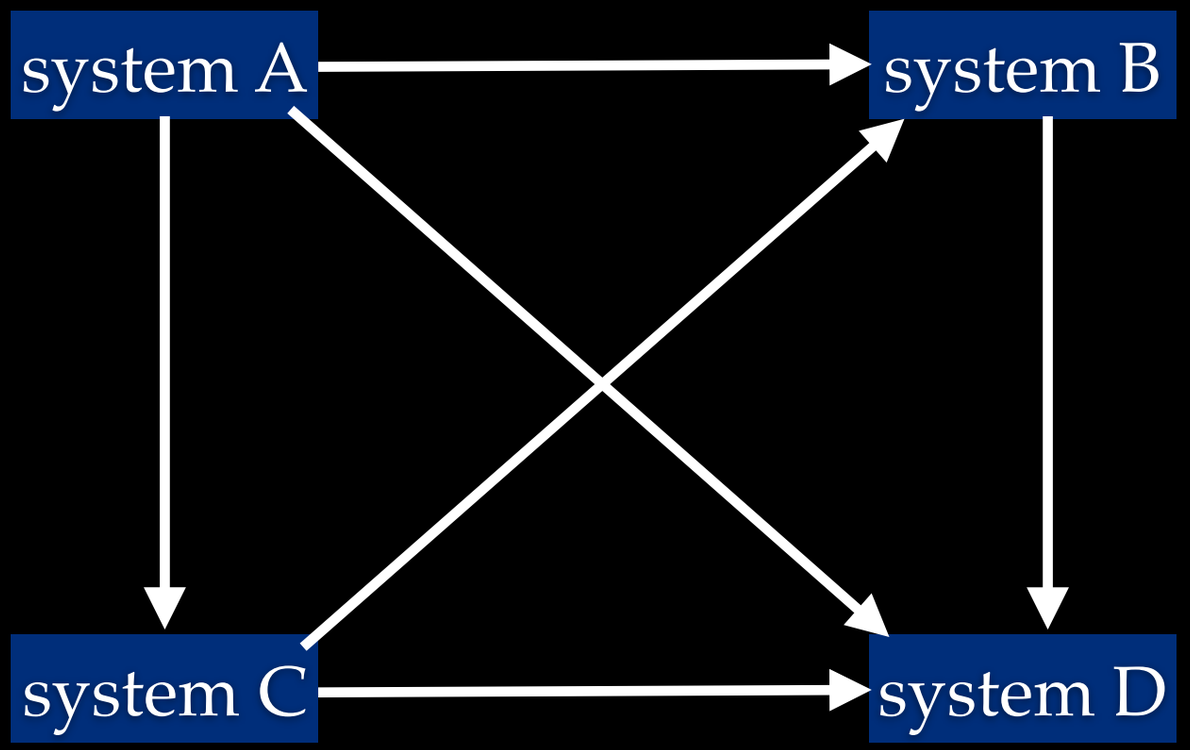
Tournaments



Landau, 1951. On dominance relations and the structure of animal societies

- Directed edge between every pair of vertices.
- Edge from A to B if A beats B in pairwise comparison.
- Widely used to model: sports, web results, elections.

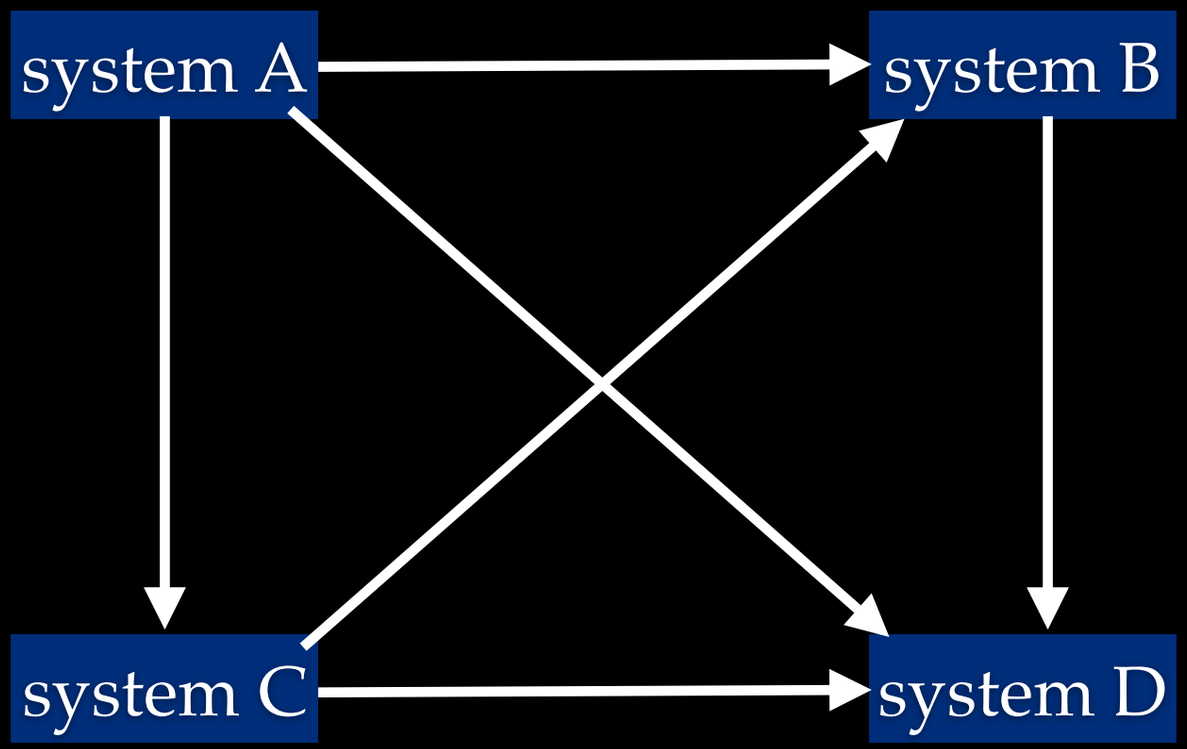
Tournaments



Landau, 1951. On dominance relations and the structure of animal societies

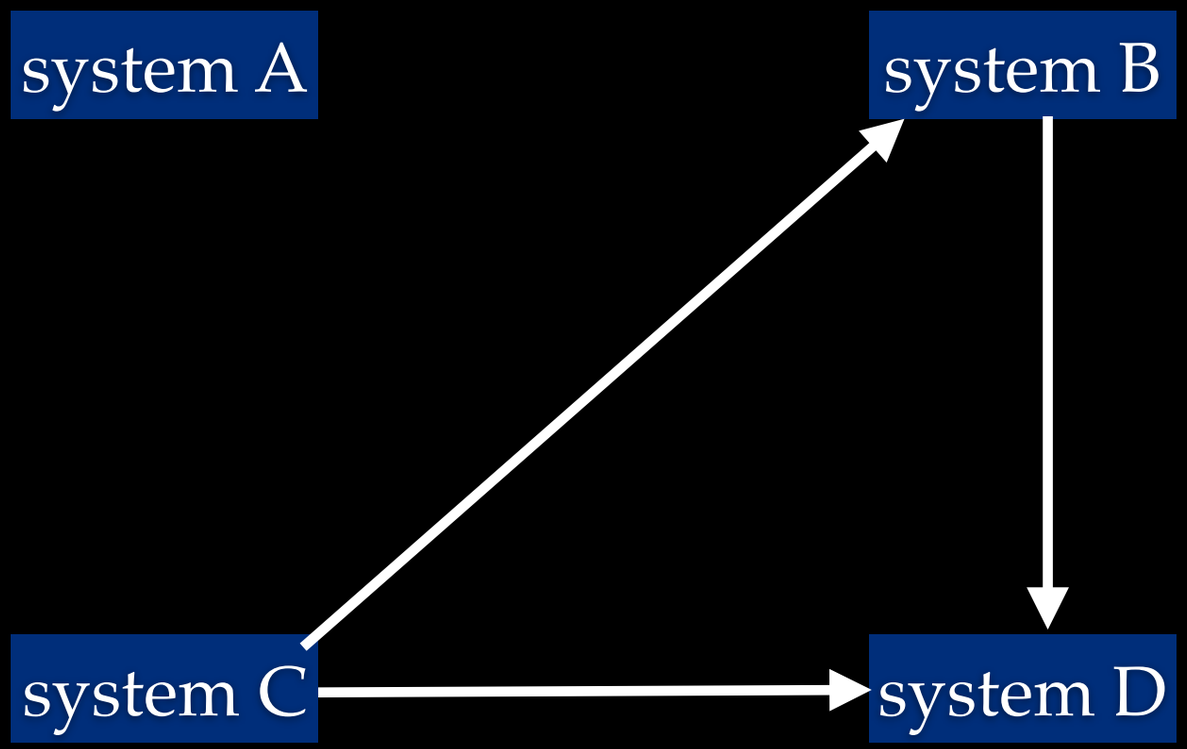
- Directed edge between every pair of vertices.
- Edge from A to B if A beats B in pairwise comparison.
- Widely used to model: sports, web results, elections.
- Used to model *all* WMT `10-`11 rankings (25 tasks).

Tournaments



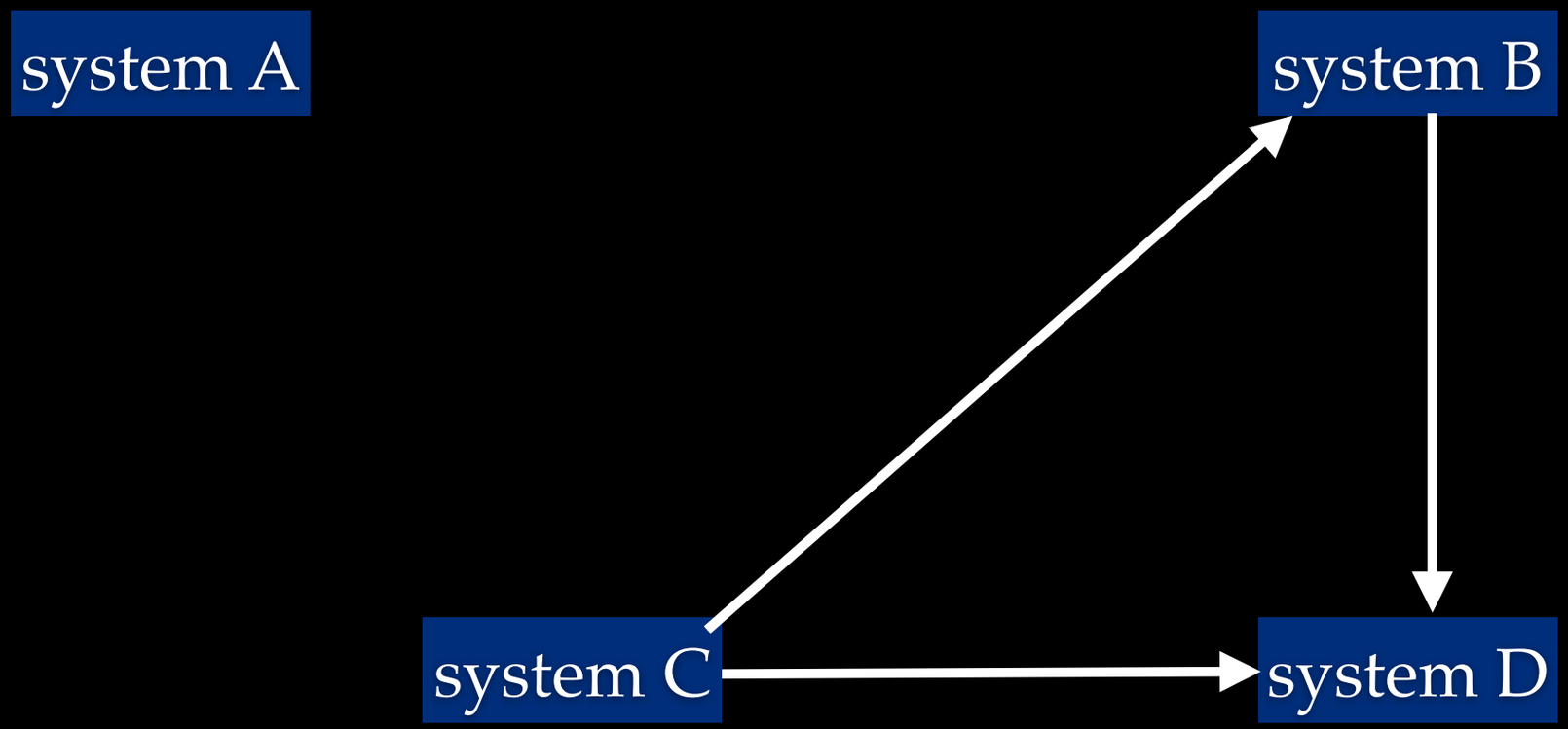
If tournament is acyclic: topological sort

Tournaments



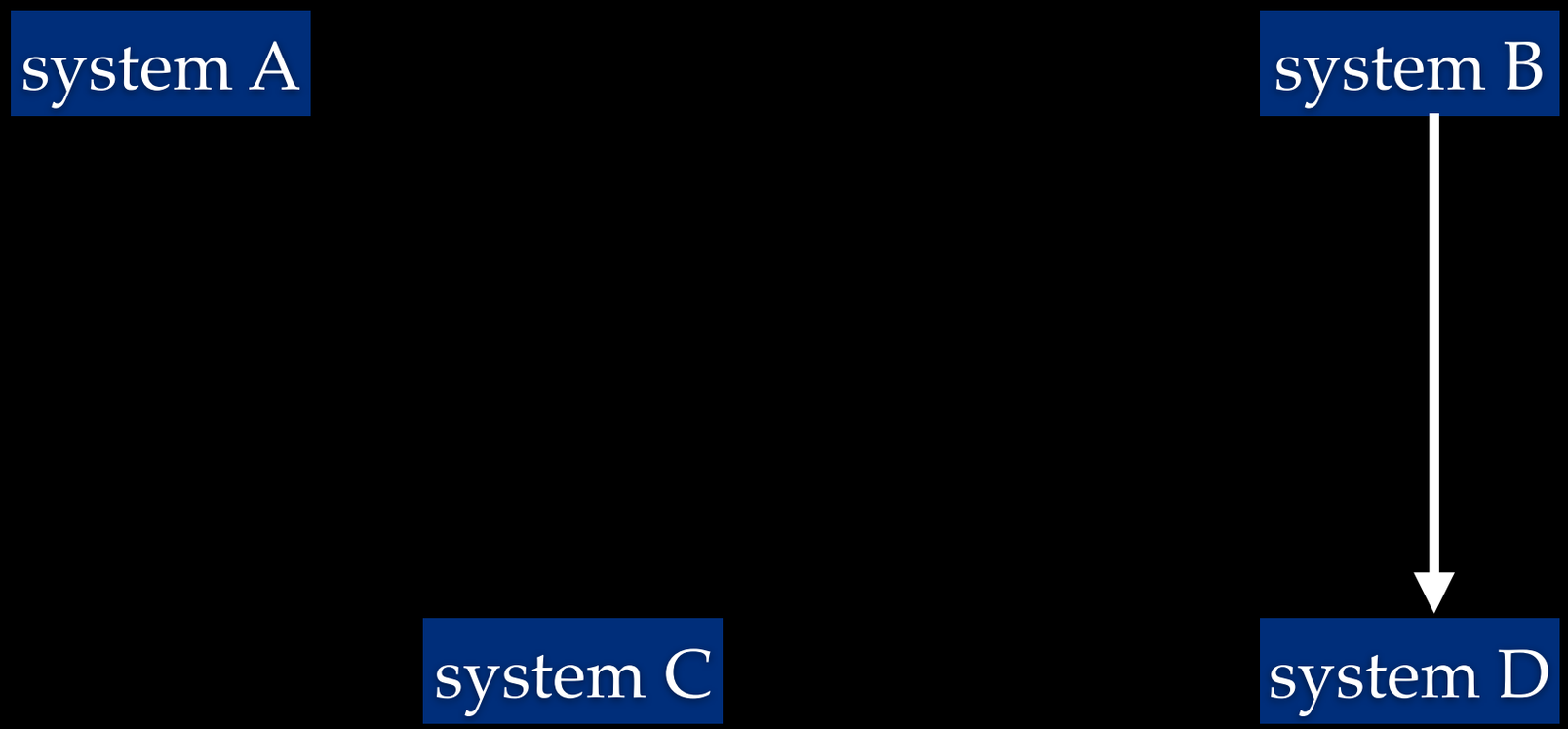
If tournament is acyclic: topological sort

Tournaments



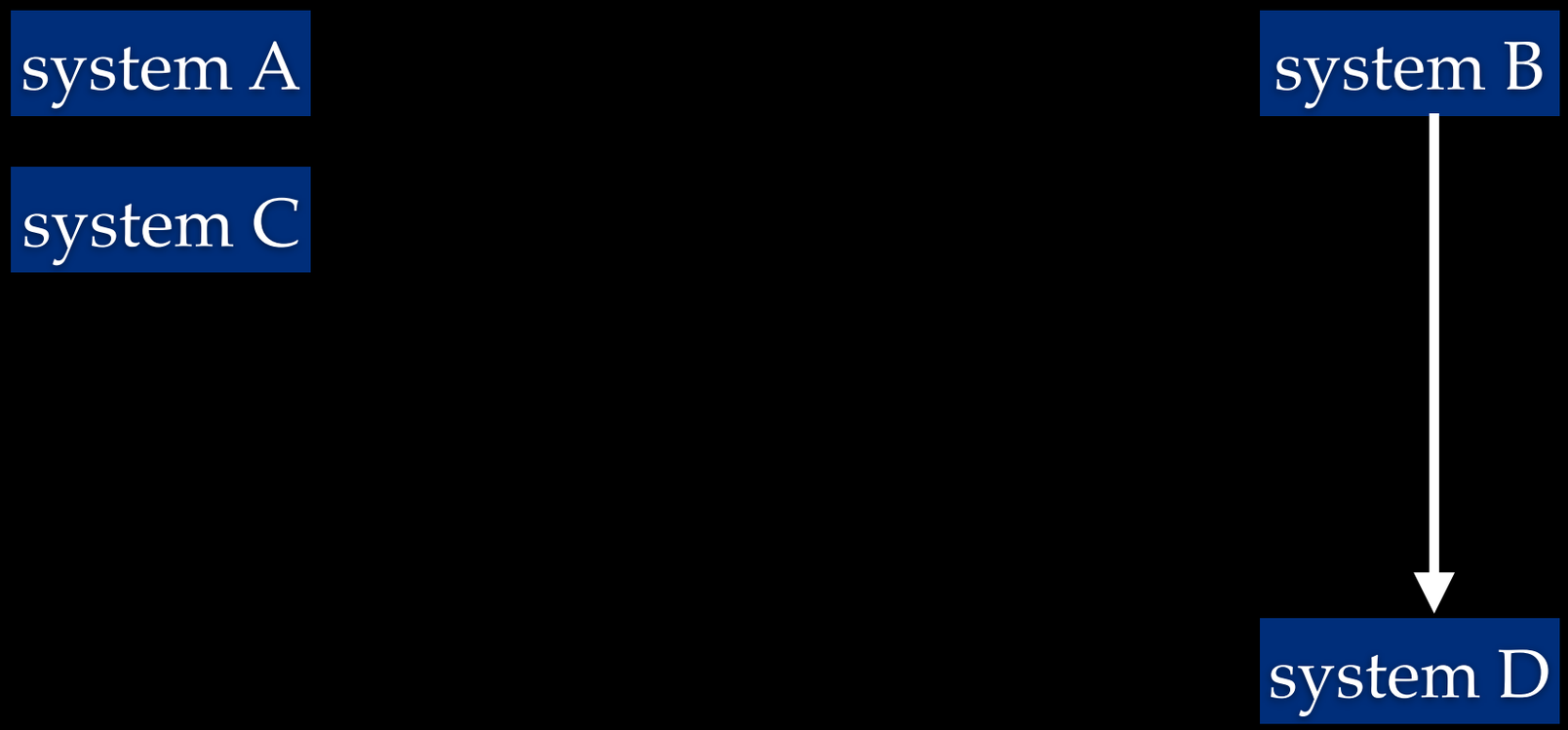
If tournament is acyclic: topological sort

Tournaments



If tournament is acyclic: topological sort

Tournaments



If tournament is acyclic: topological sort

Tournaments

system A

system B

system C

system D

If tournament is acyclic: topological sort

Tournaments

system A

system C

system B

system D

If tournament is acyclic: topological sort

Tournaments

system A

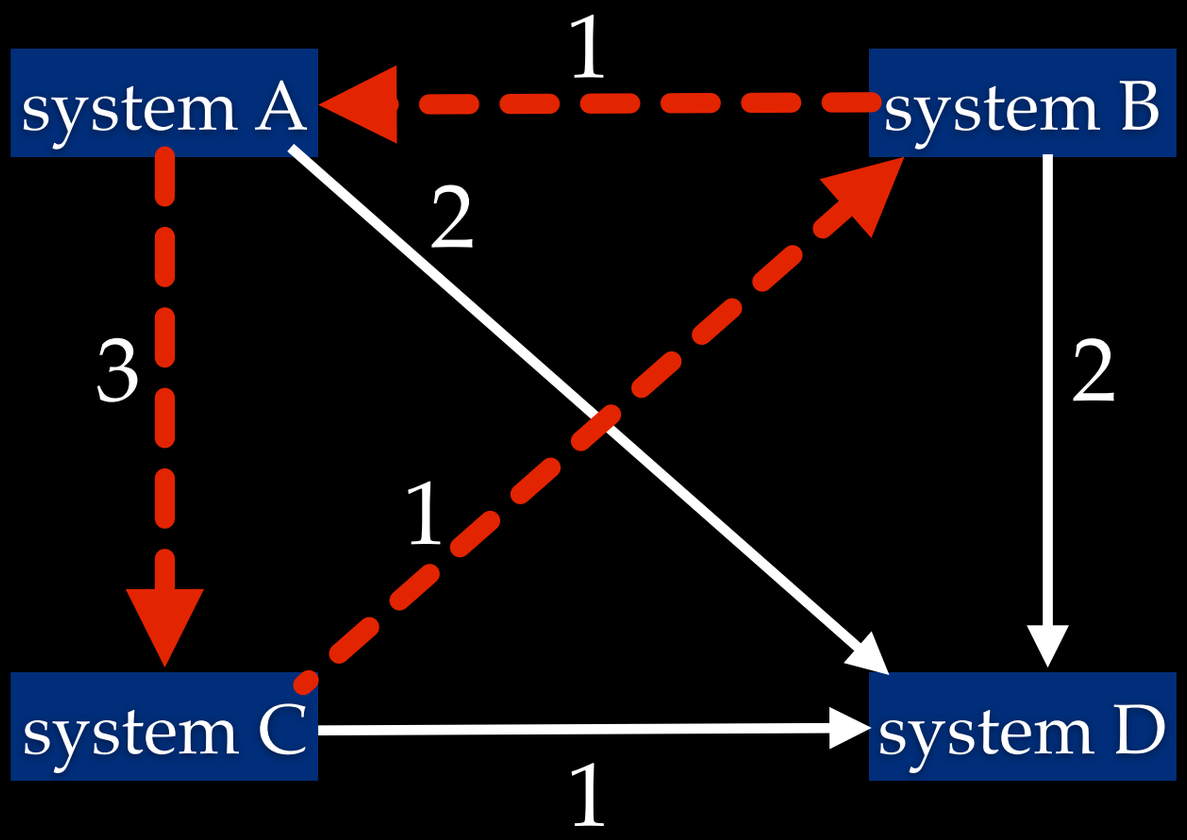
system C

system B

system D

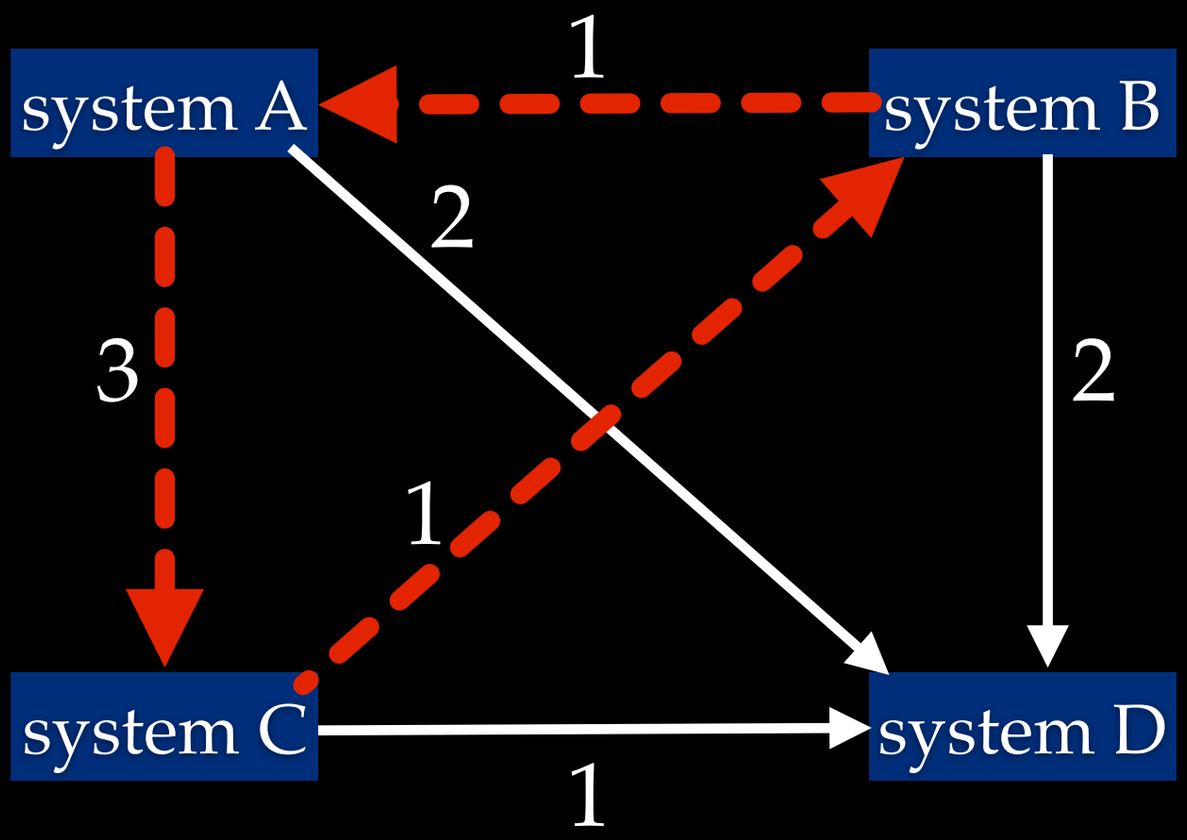
If tournament is acyclic: topological sort

Tournaments



What if tournament contains cycles?

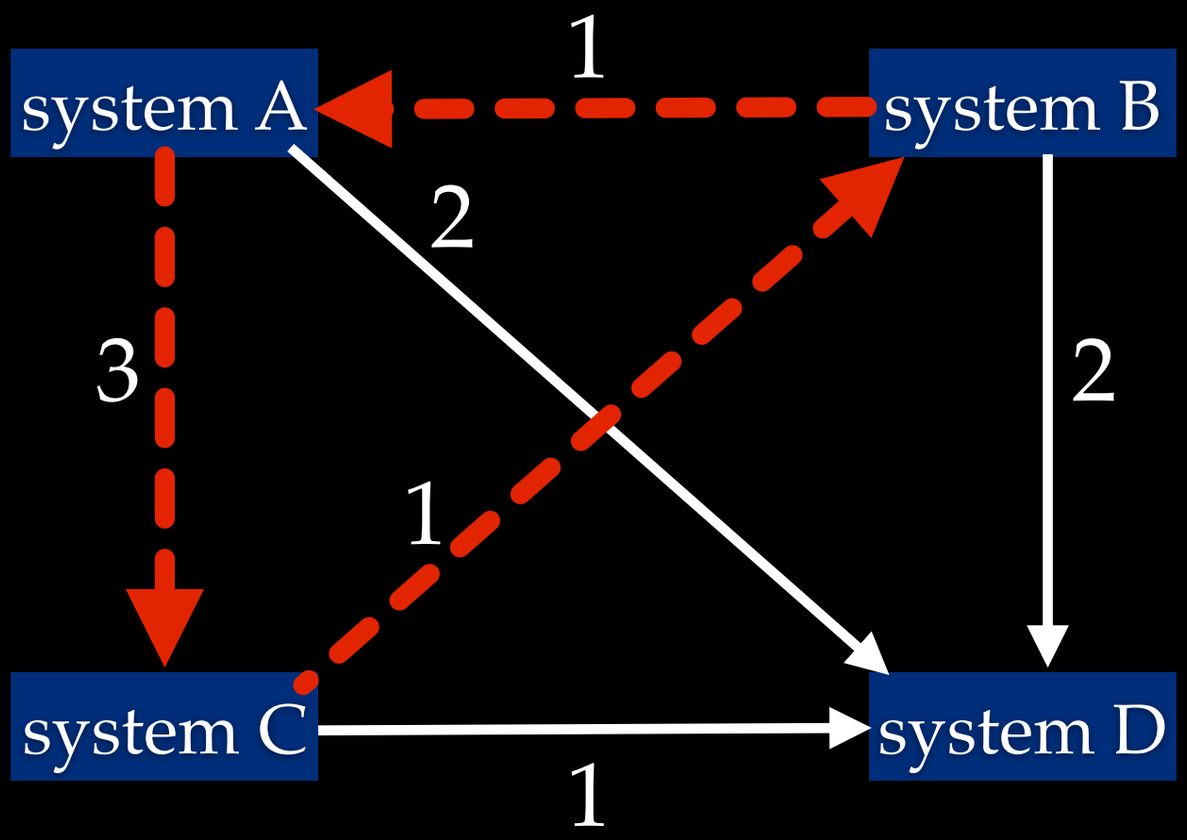
Tournaments



What if tournament contains cycles?

16 out of 25 tasks in WMT ’10-’11 contain cycles!

Tournaments



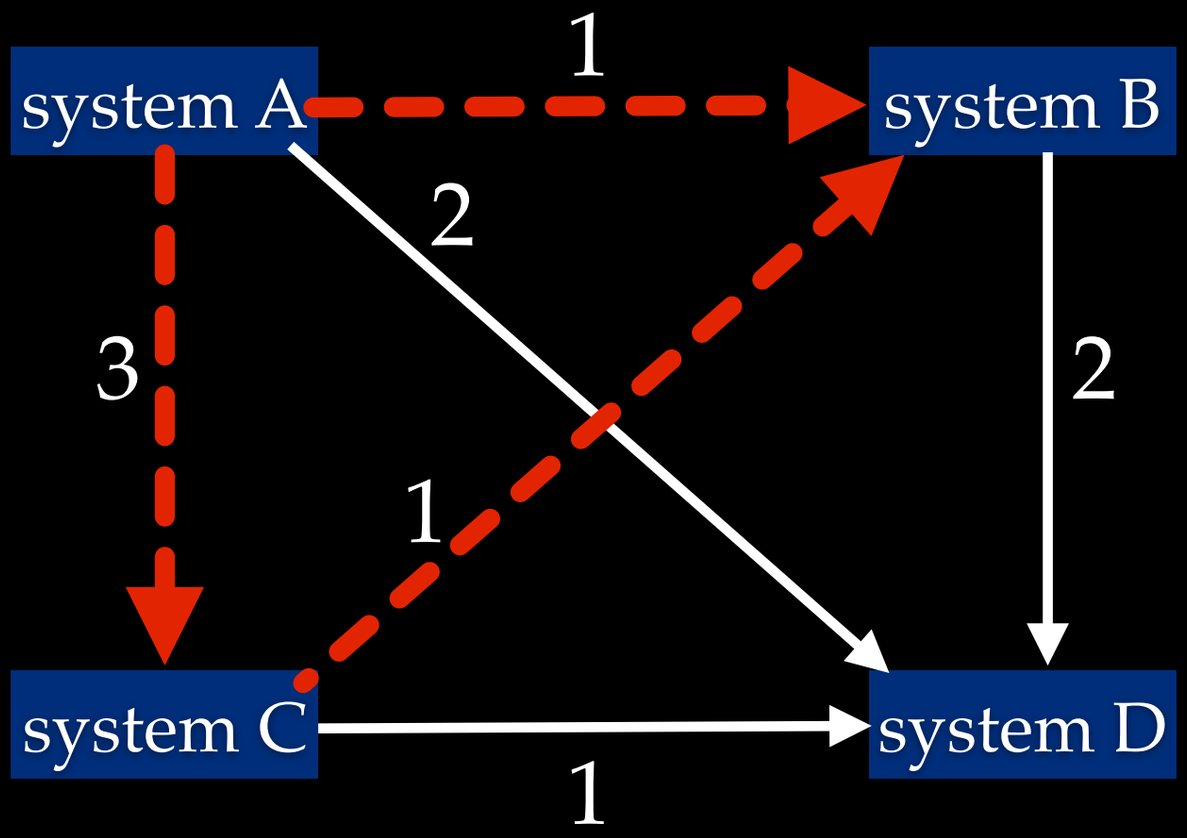
What if tournament contains cycles?

One solution: *Reverse* a set of edges such that:

(a) Resulting graph is acyclic.

(b) Sum of reversed edges weights is minimized.

Tournaments



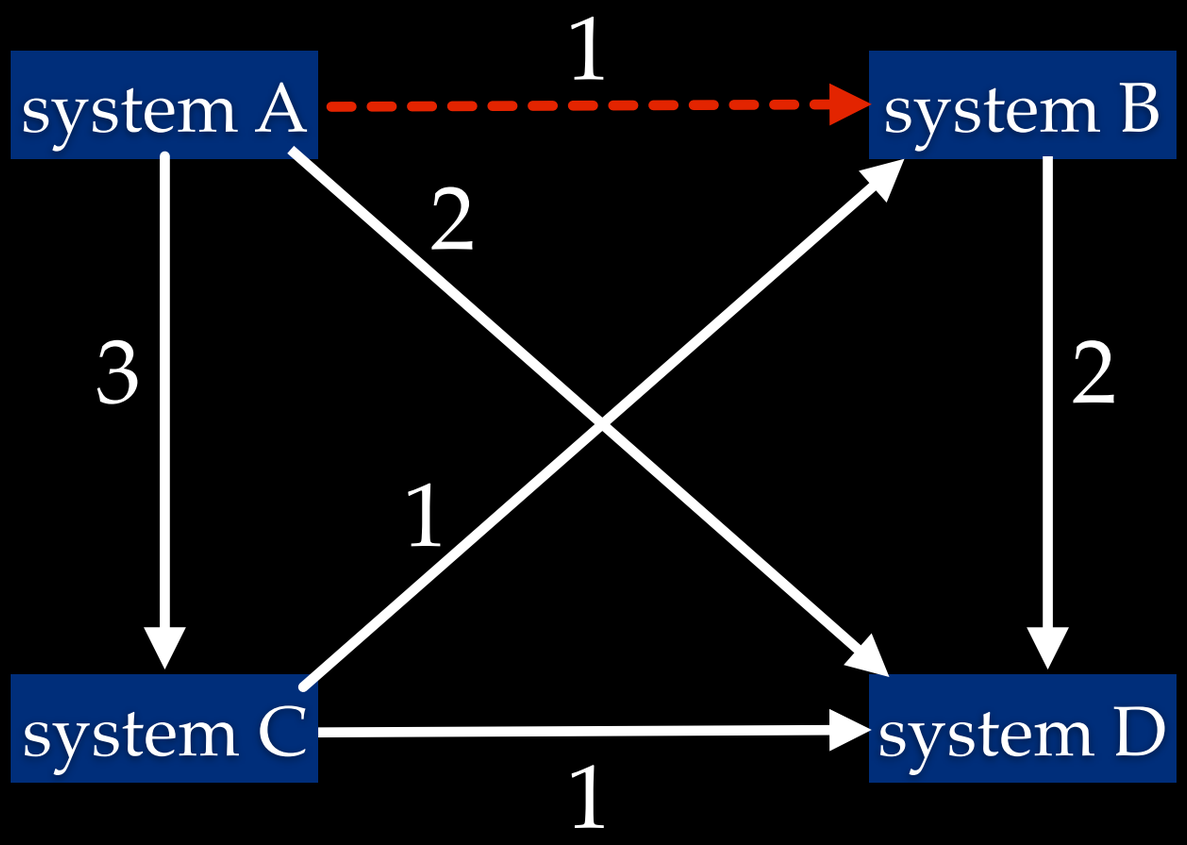
What if tournament contains cycles?

One solution: *Reverse* a set of edges such that:

(a) Resulting graph is acyclic.

(b) Sum of reversed edges weights is minimized.

Tournaments



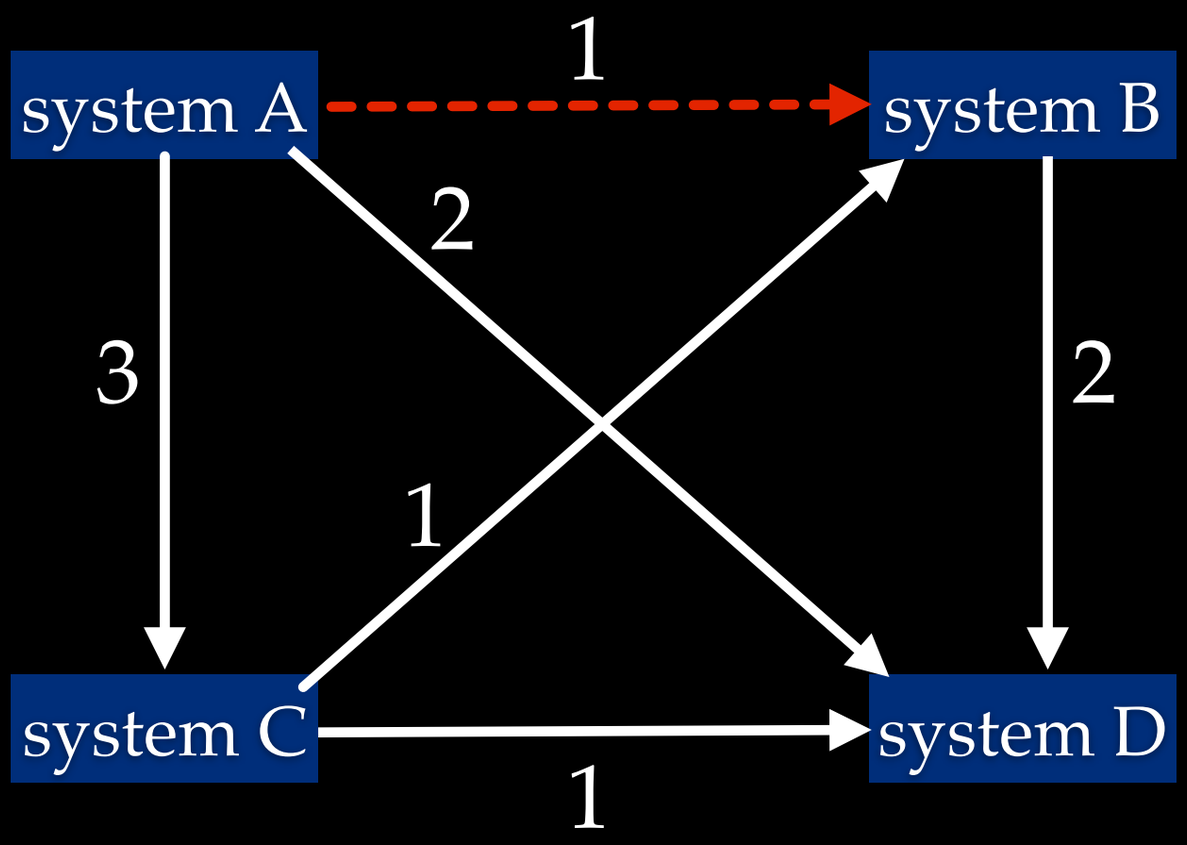
What if tournament contains cycles?

Solution: *Reverse* a set of edges such that:

(a) Graph is acyclic.

(b) Sum of reversed edges weights is minimized.

Tournaments



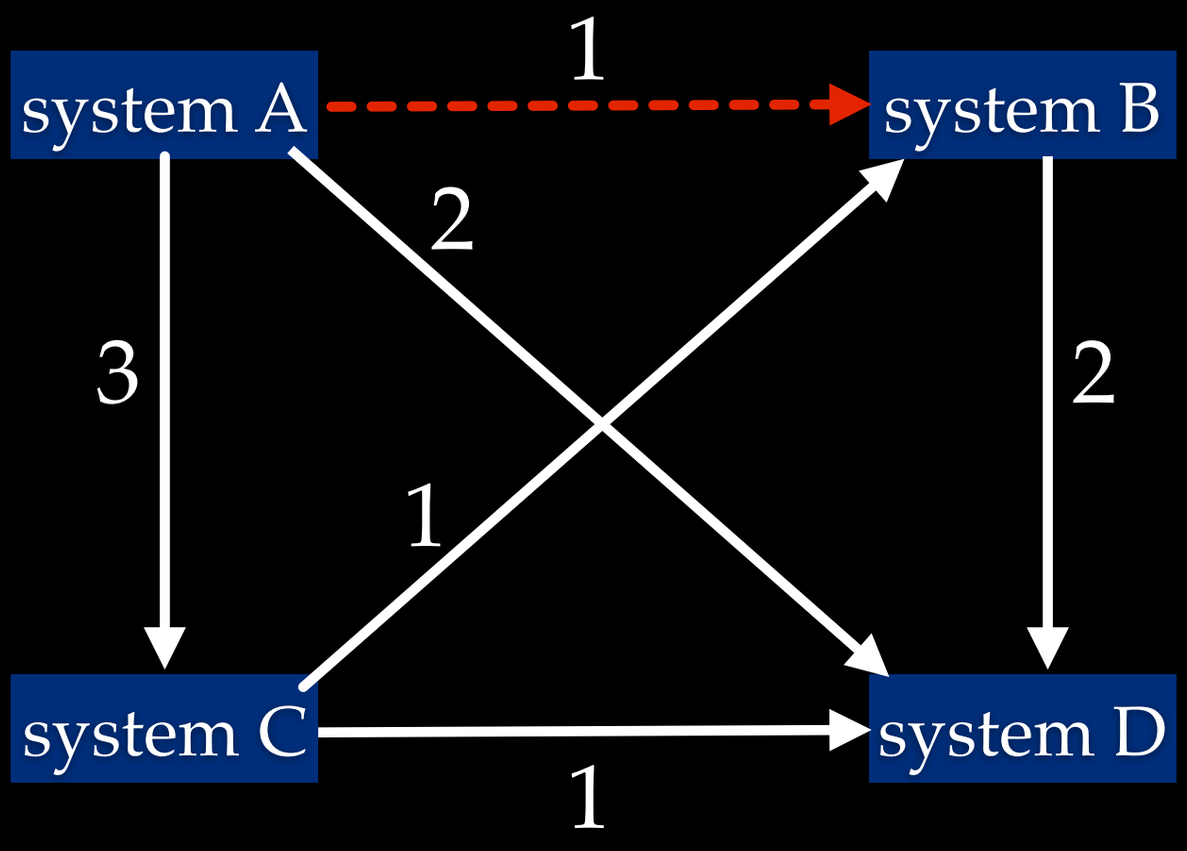
What if tournament contains cycles?

Set of reversed edges = *minimum feedback arc set (MFAS)*.

In theory, this optimization is NP-hard (Karp, 1972).

In practice, it's not too hard.

Tournaments



What if tournament contains cycles?

Important detail: What should the weight be?
Following analysis uses #(wins - losses).
Dumb, but counts each observation equally.

Example: French-English 2010 Task Rankings

MFAS

Example: French-English 2010 Task Rankings

MFAS

onlineB
rwth-combo
cmu-hyposel-combo
cambridge
lium
dcu-combo
cmu-heafield-combo
upv-combo
nrc
uedin
jhu
limsi
jhu-combo
lium-combo
rali
lig
bbn-combo
rwth
cmu-statxfer
onlineA
huicong
dfki
cu-zeman
geneva

Example: French-English 2010 Task Rankings

Has WMT solved these problems?

Human evaluation is too slow and expensive!

Human evaluation isn't reproducible!

Has WMT solved these problems?

Human evaluation is too slow and expensive!

With crowdsourcing, WMT has made a good dent in this problem.

Human evaluation isn't reproducible!

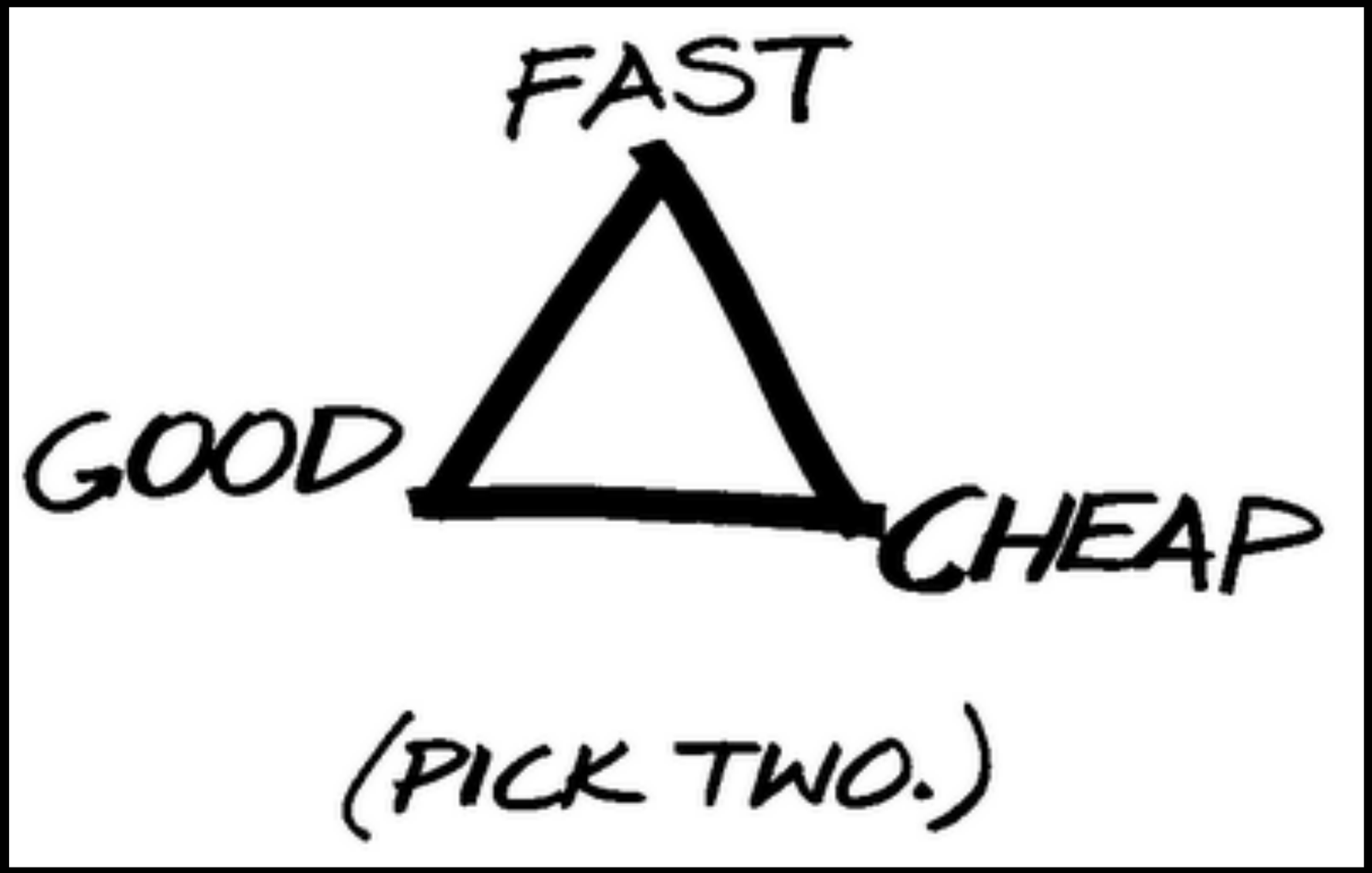
Has WMT solved these problems?

Human evaluation is too slow and expensive!

With crowdsourcing, WMT has made a good dent in this problem.

Human evaluation isn't reproducible!

Empirically true in the WMT data.



Human evaluation is fast and cheap!

美国愿和北韩谈判但拒绝再付出报酬

美国愿和北韩谈判但拒绝再付出报酬

US willing to negotiate with North Korea but not to pay more compensation.

美国愿和北韩谈判但拒绝再付出报酬

US willing to negotiate with North Korea but not to pay more compensation.

The United States is willing to hold talks with North Korea but refused to pay remuneration.

“奋进”号因机械手故障推迟到升空

“奋进”号因机械手故障推迟到升空

Launch of “Endeavour” delayed by robotic arm problems.

“奋进”号因机械手故障推迟到升空

Launch of “Endeavour” delayed by robotic arm problems.

“Progress” postponed because of mechanical hand into the sky.

Although the northern wind shrieked across the sky , it was still very clear .

Although the northern wind shrieked across the sky , it was still very clear .

However , the sky remained clear under the strong north wind .

Although the northern wind shrieked across the sky , it was still very clear .

Although the northern wind shrieked across the sky , it was still very clear .

However , the sky remained clear under the strong north wind .

Although the northern wind shrieked across the sky , it was still very clear .

However , the sky remained clear under the strong north wind .

Edit distance = 16

3 substitutions

8 deletions

5 insertions

Although the northern wind shrieked across the sky , it was still very clear .

However , the sky remained clear under the strong north wind .

$$\mathrm{a} = \mathrm{a}_1 ... a_n$$
$$\mathrm{a}_i = \langle w, \hat{w} \rangle \in \{\Sigma \cap \epsilon\}^2$$
$$\mathrm{e}_1 = a_{1,1} ... a_{n,1}$$
$$\mathrm{e}_2 = a_{1,2} ... a_{n,2}$$
$$\mathrm{cost}(\mathrm{a}_i) = 0 \text{ if } a_{i,1} = a_{i,2}, 1 \text{ otherwise}$$
$$\mathrm{edit\_distance}(\mathrm{e}_1, e_2) = \min_a \sum_{i=1}^{n} cost(a_i)$$

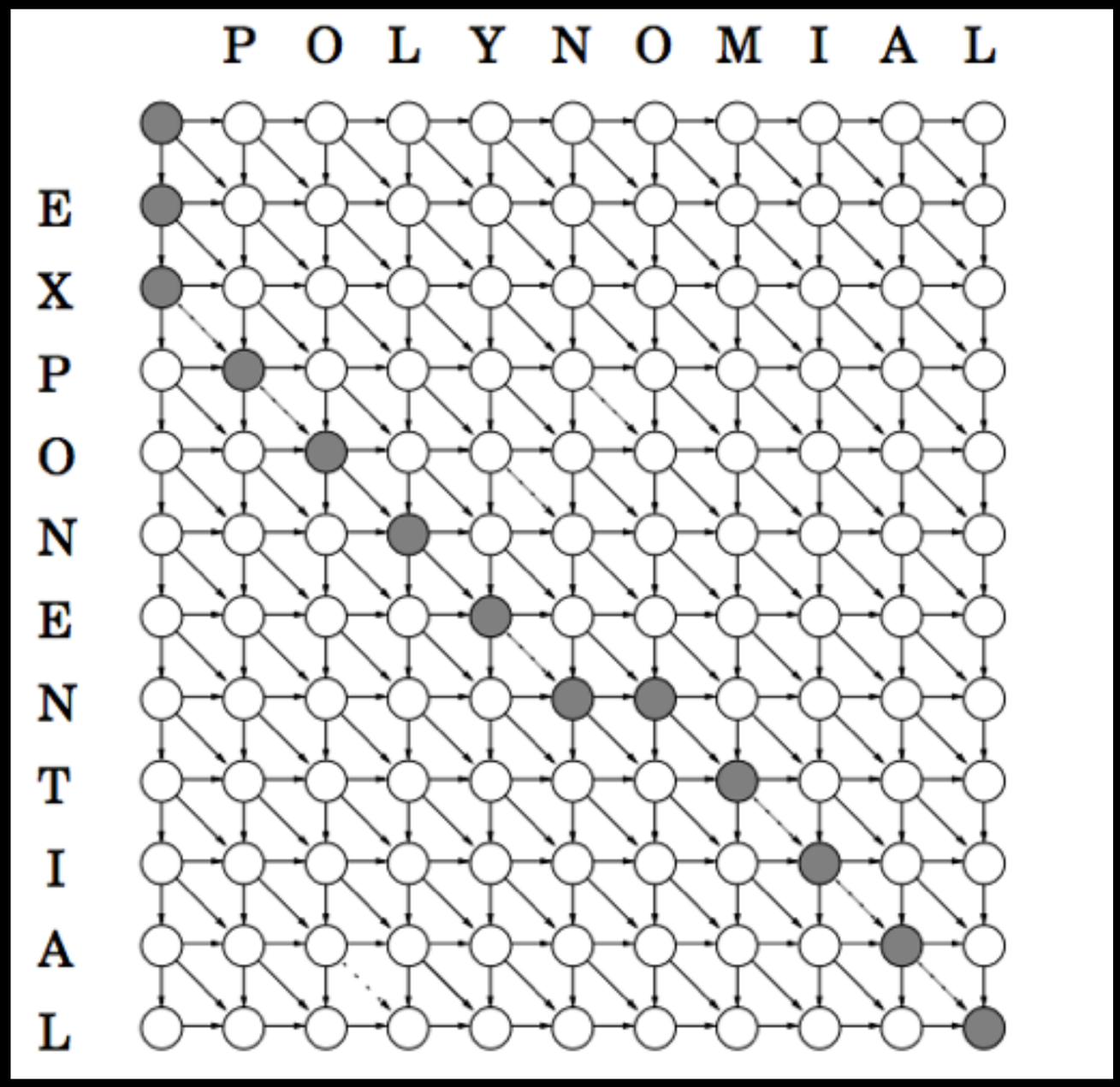
$$ed(i,j) = \min \begin{cases} ed(i-1,j) + del(w_i) \\ ed(i,j-1) + ins(w'_j) \\ ed(i-1,j-1) + sub(w_i, w'_j) \end{cases}$$

Although the northern wind shrieked across the sky , it was still very clear .

However , the sky remained clear under the strong north wind .

$$\mathrm{a} = \mathrm{a}_1 ... a_n$$
$$\mathrm{a}_i = \langle w, \hat{w} \rangle \in \{\Sigma \cap \epsilon\}^2$$
$$\mathrm{e}_1 = a_{1,1} ... a_{n,1}$$
$$\mathrm{e}_2 = a_{1,2} ... a_{n,2}$$
$$\mathrm{cost}(\mathrm{a}_i) = 0 \text{ if } a_{i,1} = a_{i,2}, 1 \text{ otherwise}$$
$$\mathrm{edit\_distance}(\mathrm{e}_1, e_2) = \min_a \sum_{i=1}^{n} cost(a_i)$$

$$ed(i,j) = \min \begin{cases} ed(i-1,j) + del(w_i) \\ ed(i,j-1) + ins(w'_j) \\ ed(i-1,j-1) + sub(w_i, w'_j) \end{cases}$$

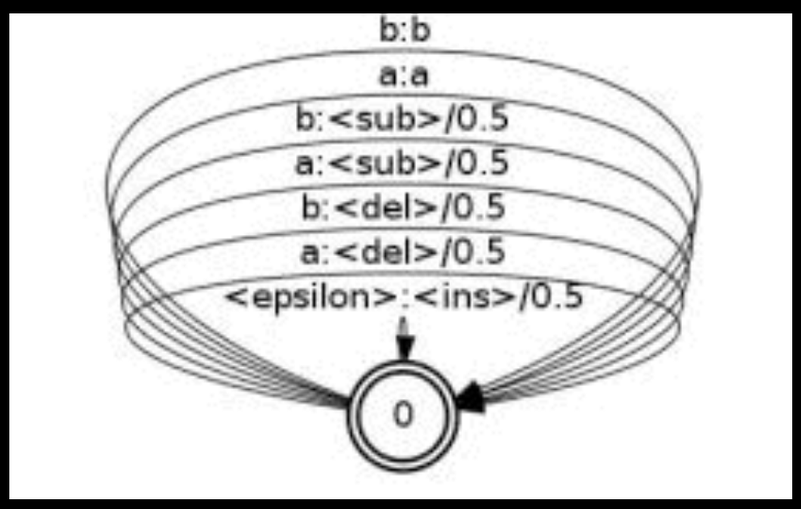


polynomial
exponen tial

$$ed(i,j) = \min \begin{cases} ed(i-1,j) + del(w_i) \\ ed(i,j-1) + ins(w'_j) \\ ed(i-1,j-1) + sub(w_i, w'_j) \end{cases}$$

Although the northern wind shrieked across the sky , it was still very clear .

However , the sky remained clear under the strong north wind .

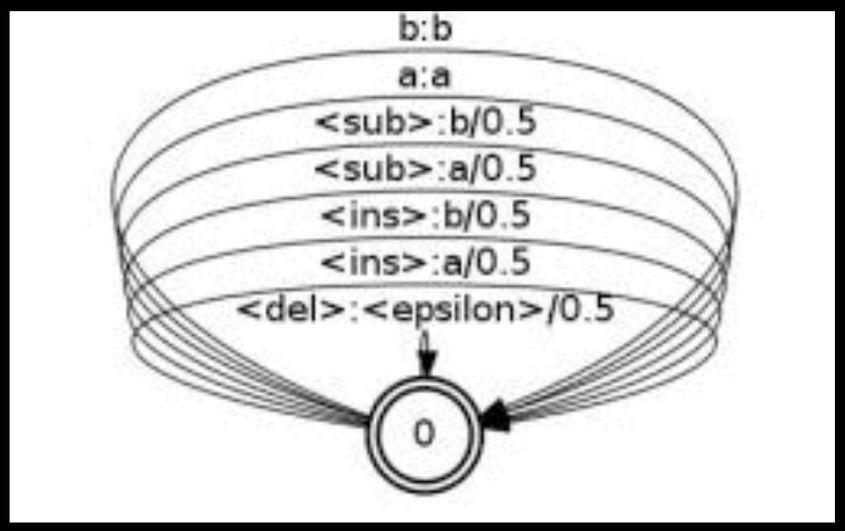


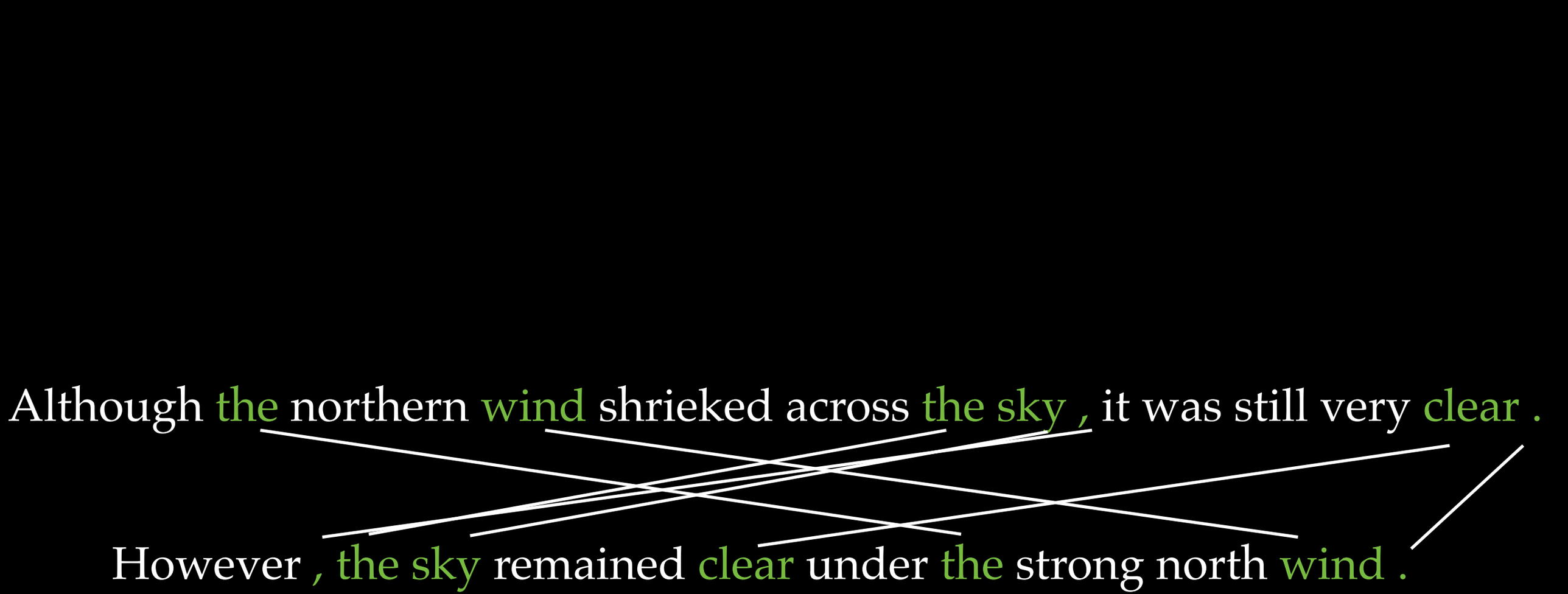
Edit distance = 16

3 substitutions

8 deletions

5 insertions





Precision:
$7/15$ tokens $= 47\%$

Recall:
$7/12$ tokens $= 58\%$

Precision: 11/15 tokens

Although the northern wind shrieked across the sky , it was still very clear .

However , the sky remained clear under the strong north wind .

Although a north wind was howling , the sky remained clear and blue .

The sky was still crystal clear , though the north wind was howling .

Despite the strong northerly winds , the sky remains very clear .

Precision: 11/15 tokens

sky very northern shrieked clear wind Although across the the , still was it .

However , the sky remained clear under the strong north wind .

Although a north wind was howling , the sky remained clear and blue .

The sky was still crystal clear , though the north wind was howling .

Despite the strong northerly winds , the sky remains very clear .

Precision: 11/15 tokens

4/14 bigrams

1/13 trigrams

Although the northern wind shrieked across the sky , it was still very clear .

However , the sky remained clear under the strong north wind .

Although a north wind was howling , the sky remained clear and blue .

The sky was still crystal clear , though the north wind was howling .

Despite the strong northerly winds , the sky remains very clear .

Precision: 11/15 tokens

0/14 bigrams

0/13 trigrams

sky very northern shrieked clear wind Although across the the , still was it .

However , the sky remained clear under the strong north wind .

Although a north wind was howling , the sky remained clear and blue .

The sky was still crystal clear , though the north wind was howling .

Despite the strong northerly winds , the sky remains very clear .

Precision: 3/1 tokens
2/2 bigrams
1/1 trigrams

very clear .

However , the sky remained clear under the strong north wind .

Although a north wind was howling , the sky remained clear and blue .

The sky was still crystal clear , though the north wind was howling .

Despite the strong northerly winds , the sky remains very clear .

Precision: 11/15 tokens

4/14 bigrams

1/13 trigrams

very clear . shrieked was still Although wind , across it northern the the sky

However , the sky remained clear under the strong north wind .

Although a north wind was howling , the sky remained clear and blue .

The sky was still crystal clear , though the north wind was howling .

Despite the strong northerly winds , the sky remains very clear .

Precision: 11/15 tokens

4/14 bigrams

1/13 trigrams

a north . the was and was the the the though the , the sky

However , the sky remained clear under the strong north wind .

Although a north wind was howling , the sky remained clear and blue .

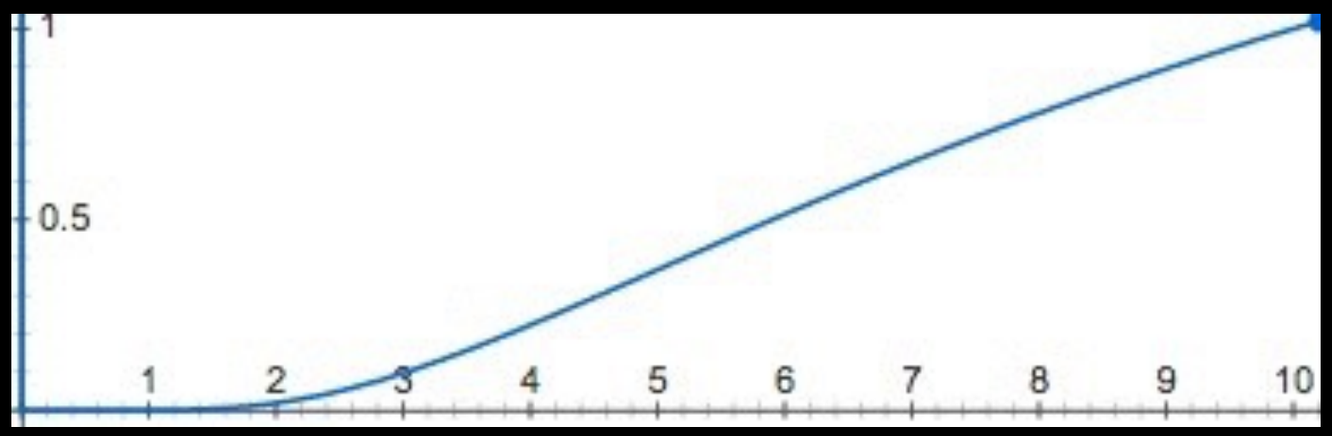
The sky was still crystal clear , though the north wind was howling .

Despite the strong northerly winds , the sky remains very clear .

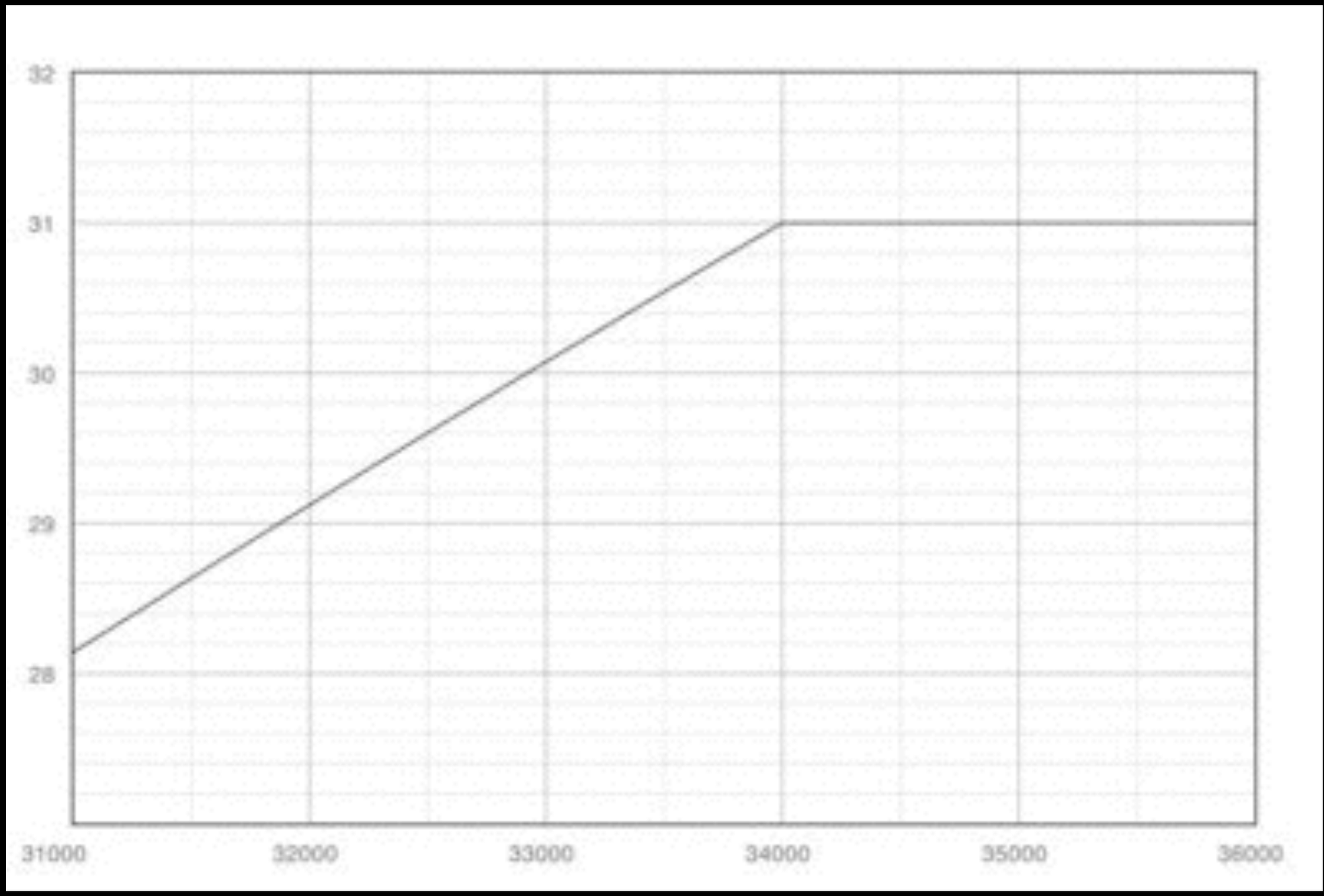
BLEU

$$BP = \begin{cases} 1 & if \quad c > r \\ e^{1-r/c} & if \quad c \leq r \end{cases}$$

$$Bleu = BP \cdot exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$



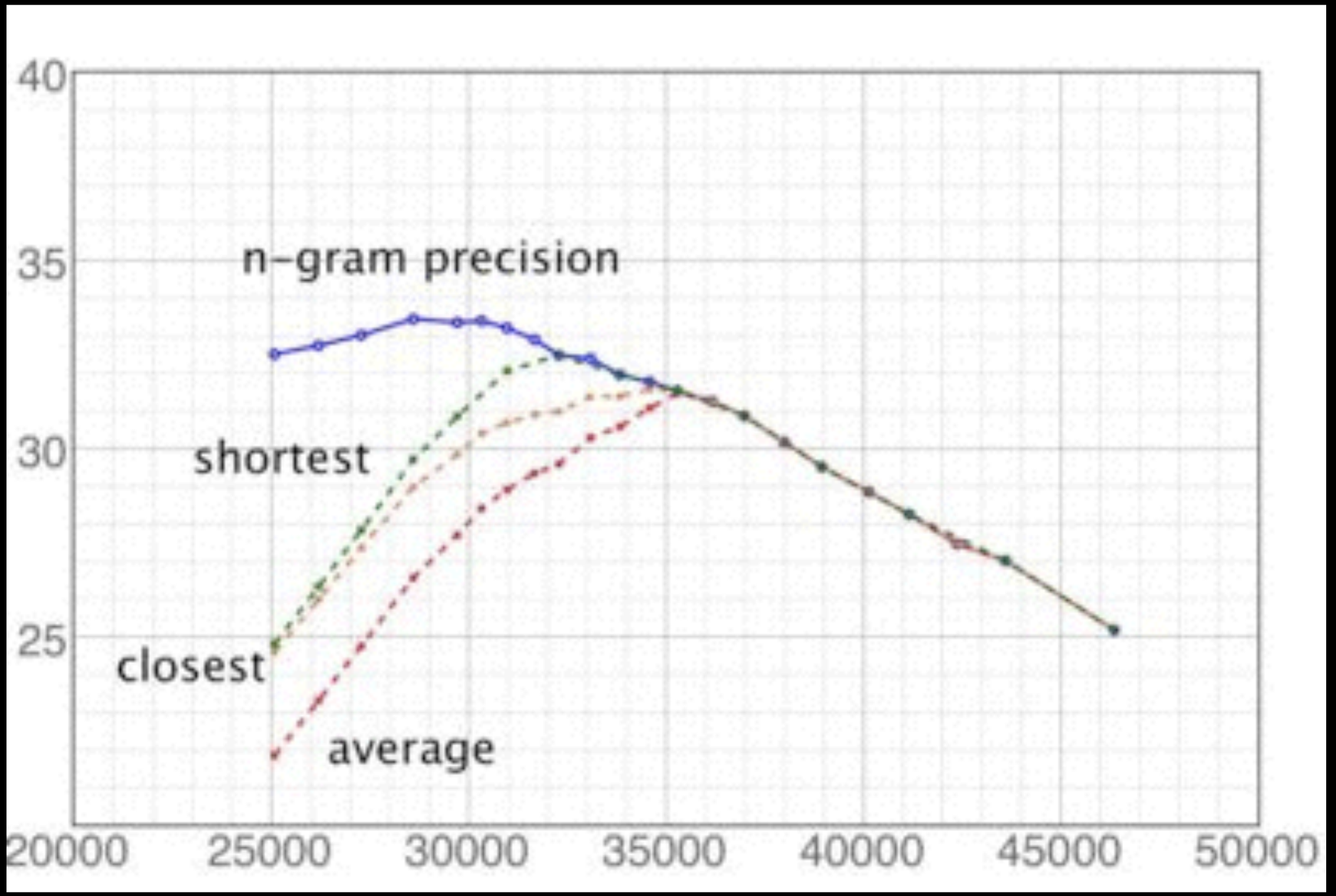
Details matter



BP

length

Details matter



Influence of BLEU

The mathematics of statistical machine translation: Parameter estimation
PF Brown, VJD Pietra, SAD Pietra… - Computational linguistics, 1993 - dl.acm.org
Abstract We describe a series of five statistical models of the translation process and give algorithms for estimating the parameters of these models given a set of pairs of sentences that are translations of one another. We define a concept of word-by-word alignment between such pairs of sentences. For any given pair of such sentences each of our models assigns a probability to each of the possible word-by-word alignments. We give an ...
Cited by 4481 Related articles All 46 versions Cite Save

BLEU: a **method** for **automatic evaluation** of machine translation
K Papineni, S Roukos, T Ward, WJ Zhu - … of the 40th annual meeting on …, 2002 - dl.acm.org
Abstract Human **evaluations** of machine translation are extensive but expensive. Human **evaluations** can take months to finish and involve human labor that can not be reused. We propose a **method** of **automatic** machine translation **evaluation** that is quick, inexpensive,
Cited by 6223 Related articles All 36 versions Cite Save

Influence of BLEU

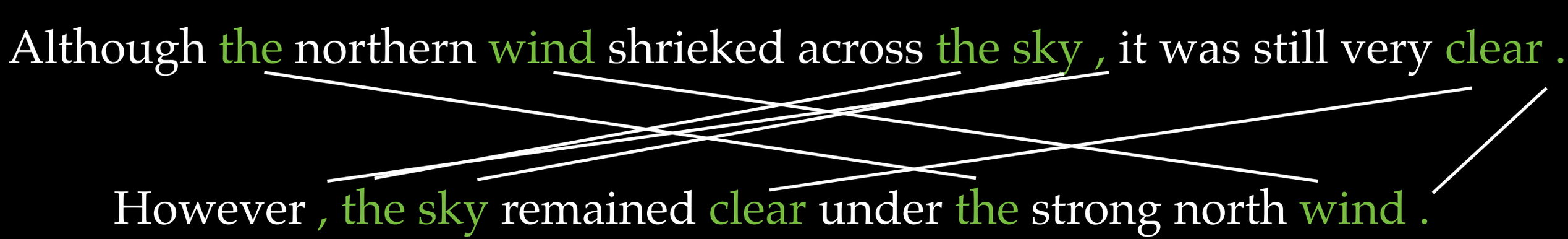
The mathematics of statistical machine translation: Parameter estimation
PF Brown, VJD Pietra, SAD Pietra… - Computational linguistics, 1993 - dl.acm.org
Abstract We describe a series of five statistical models of the translation process and give algorithms for estimating the parameters of these models given a set of pairs of sentences that are translations of one another. We define a concept of word-by-word alignment between such pairs of sentences. For any given pair of such sentences each of our models assigns a probability to each of the possible word-by-word alignments. We give an ...
Cited by 4481 Related articles All 46 versions Cite Save

BLEU: a **method** for **automatic evaluation** of machine translation
K Papineni, S Roukos, T Ward, WJ Zhu - … of the 40th annual meeting on …, 2002 - dl.acm.org
Abstract Human **evaluations** of machine translation are extensive but expensive. Human **evaluations** can take months to finish and involve human labor that can not be reused. We propose a **method** of **automatic** machine translation **evaluation** that is quick, inexpensive,
Cited by 6223 Related articles All 36 versions Cite Save

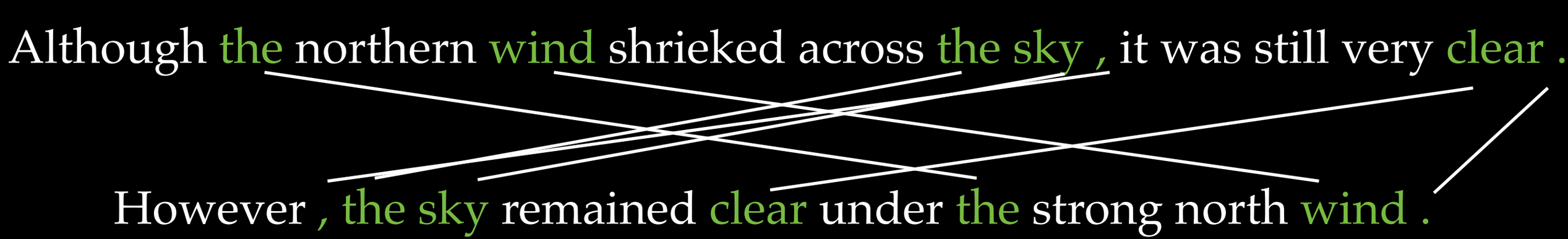
Goodhart's Law: When a measure becomes a target, it ceases to be a good measure.

BLEU-1
BLEU-4
BLEU-v11b
BLEU-v12
METEOR-v0.6
NIST-v11b
TER-v0.7.254-GRR
ATEC1
Amber
ATEC3
ATEC4
Meteor-v0.7
TerrorCat
BEwT-E
Badger
BadgerLite
Bleu-sbp
BleuSP
CDer
DP-Or
DP-Orp
DR-Or
EDPM
LET
METEOR-ranking
MaxSim
RTE
Rose
SEPIA1
SEPIA2
SNR
SR-Or
SVM-Rank
TERp
ULCh
ULCopt
invWer
mBLEU
mTER

TER: Translation (Error | Edit) Distance

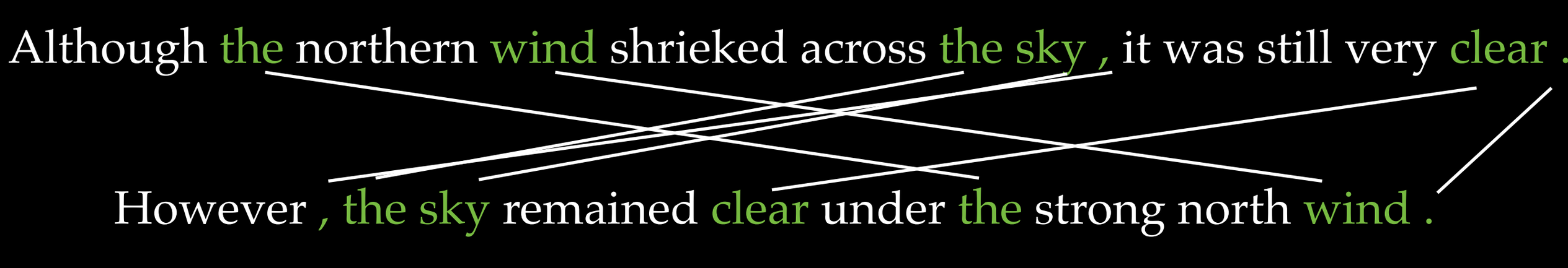


TER: Translation (Error | Edit) Distance



Basically edit distance with swaps

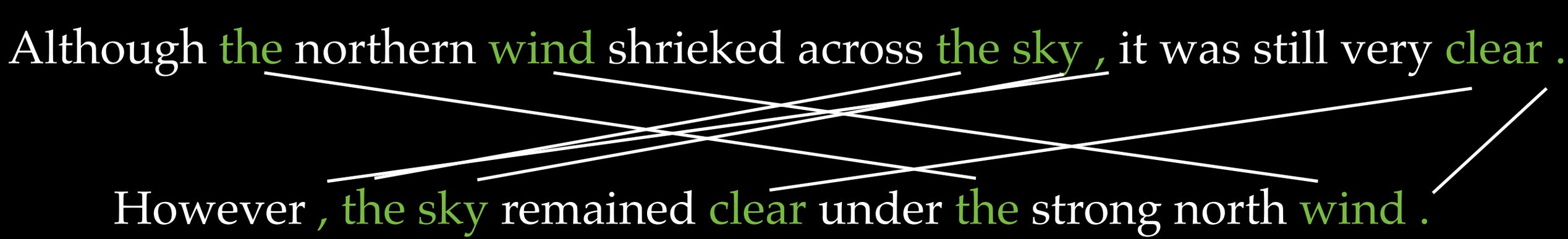
TER: Translation (Error | Edit) Distance



Basically edit distance with swaps

How hard is it to compute this?

TER: Translation (Error | Edit) Distance



Basically edit distance with swaps

How hard is it to compute this?

$$ter(i,j) = \min \begin{cases} ter(i-1,j) + del(w_i) \\ ter(i,j-1) + ins(w'_j) \\ ter(i-1,j-1) + sub(w_i, w'_j) \\ \max_k ter(i-1, [1, ...k-1, k+1, ...j]) + 1 \end{cases}$$

Why Not Use all Translations?

(Dreyer & Marcu ’12)

Why Not Use all Translations?

(Dreyer & Marcu ’12)

el primer ministro italiano Silvio Berlusconi

Why Not Use all Translations?

(Dreyer & Marcu ’12)

⟨PM⟩	⟨IT⟩	⟨SB⟩
el primer ministro	italiano	Silvio Berlusconi

Why Not Use all Translations?

(Dreyer & Marcu '12)

⟨PM⟩	⟨IT⟩	⟨SB⟩
el primer ministro	italiano	Silvio Berlusconi

⟨PM⟩ → prime-minister
⟨PM⟩ → PM
⟨PM⟩ → prime minister
⟨PM⟩ → head of government
⟨PM⟩ → premier

⟨IT⟩ → Italian

⟨SB⟩ → Silvio Berlusconi
⟨SB⟩ → Berlusconi

Why Not Use all Translations?

(Dreyer & Marcu '12)

⟨PM⟩	⟨IT⟩	⟨SB⟩
el primer ministro	italiano	Silvio Berlusconi

⟨PM⟩ → prime-minister
⟨PM⟩ → PM
⟨PM⟩ → prime minister
⟨PM⟩ → head of government
⟨PM⟩ → premier

⟨IT⟩ → Italian

⟨SB⟩ → Silvio Berlusconi
⟨SB⟩ → Berlusconi

⟨S⟩ → ⟨SB⟩ , ⟨IT⟩ ⟨PM⟩
⟨S⟩ → ⟨IT⟩ ⟨PM⟩ ⟨SB⟩
⟨S⟩ → the ⟨IT⟩ ⟨PM⟩ , ⟨SB⟩
⟨S⟩ → the ⟨PM⟩ of Italy

HyTER

- Entire set is exponential, but finite.
- Can be encoded as an FST.
- Then compute edit distance as FST composition!

HyTER statistics

- 3-4 annotators per sentence.
- 2-3 hours per annotator per sentence.
- >1M translations per annotator per sentence.
- >1B translations per sentence (combined).
- Shockingly low overlap between annotators (~10K).

Summary of evaluation

- Evaluating machine translation is really, really hard.
  - Human evaluation: expensive, slow, unreproducible. But arguably what we want.
  - Automatic evaluation: fast, cheap, consistent. But might not have anything to do with what we want.
- It's also really, really important.
  - It's easier to improve what you measure.
  - Research funding often driven by evaluation.
- What should we be measuring?

To think about

Some research in MT takes the form:

1. MT is poor at phenomenon X (e.g. agreement).
2. Build a model that handles X.
3. Measure BLEU score.
4. If BLEU score goes up, claim model is better at X.

To think about

Some research in MT takes the form:

1. MT is poor at phenomenon X (e.g. agreement).
2. Build a model that handles X.
3. Measure BLEU score.
4. If BLEU score goes up, claim model is better at X.

Question:

Is this good science?

To think about

Some research in MT takes the form:

1. MT is poor at phenomenon X (e.g. agreement).
2. Build a model that handles X.
3. Measure BLEU score.
4. If BLEU score goes up, claim model is better at X.

Question:

Is this good science?

Or is it cargo cult science?