

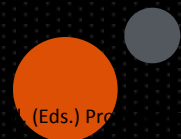
AMTA

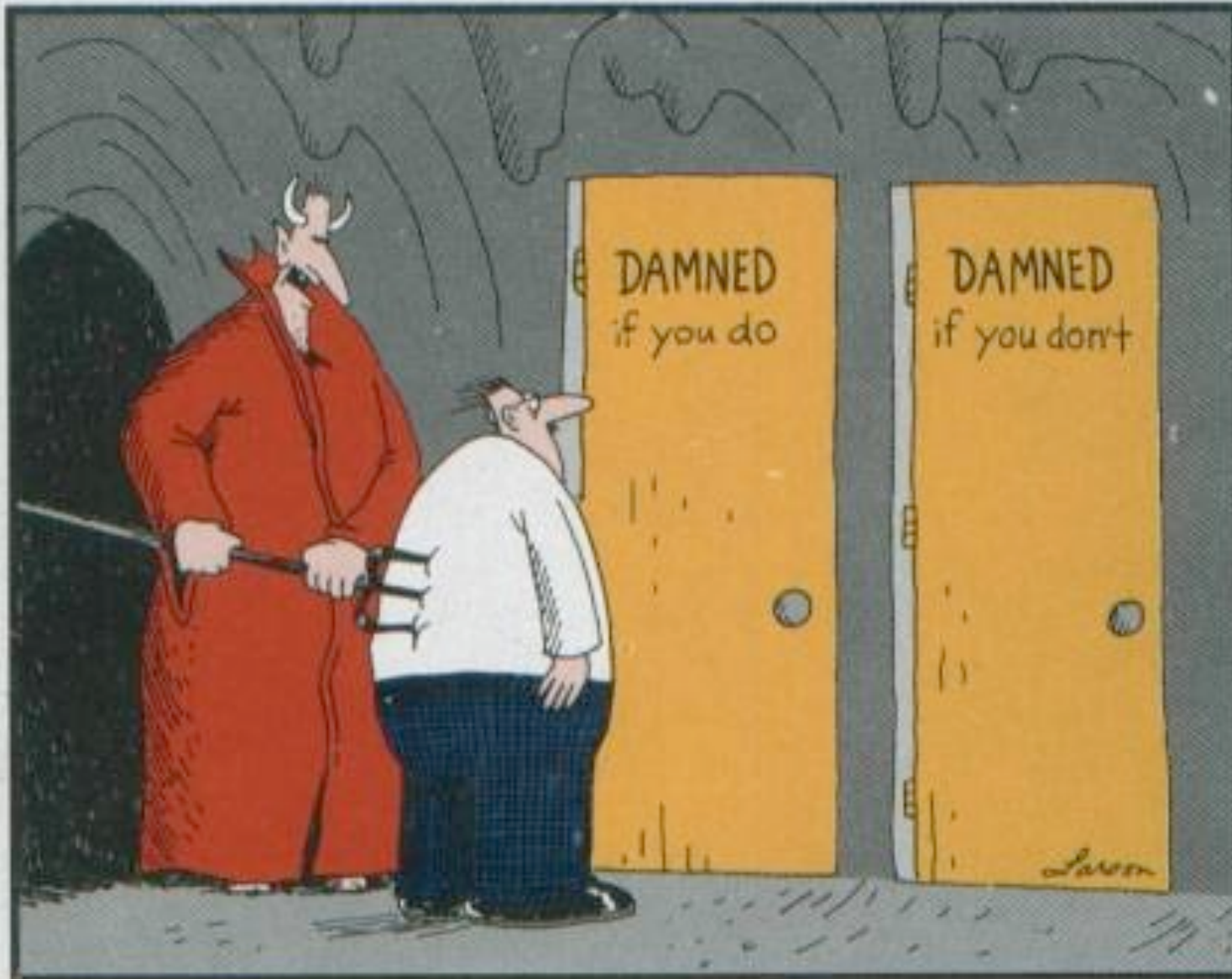
PRESENTATION

Tools-Driven Content Curation & Engine Training

———— ALEXYANISHEVSKY ————

Welocalize
October 2014





"C'mon, c'mon — it's either one or the other."

"C'mon, c'mon — it's either one or the other."

OVERVIEW

Sparsity of data is bad. Abundance of data is bad.

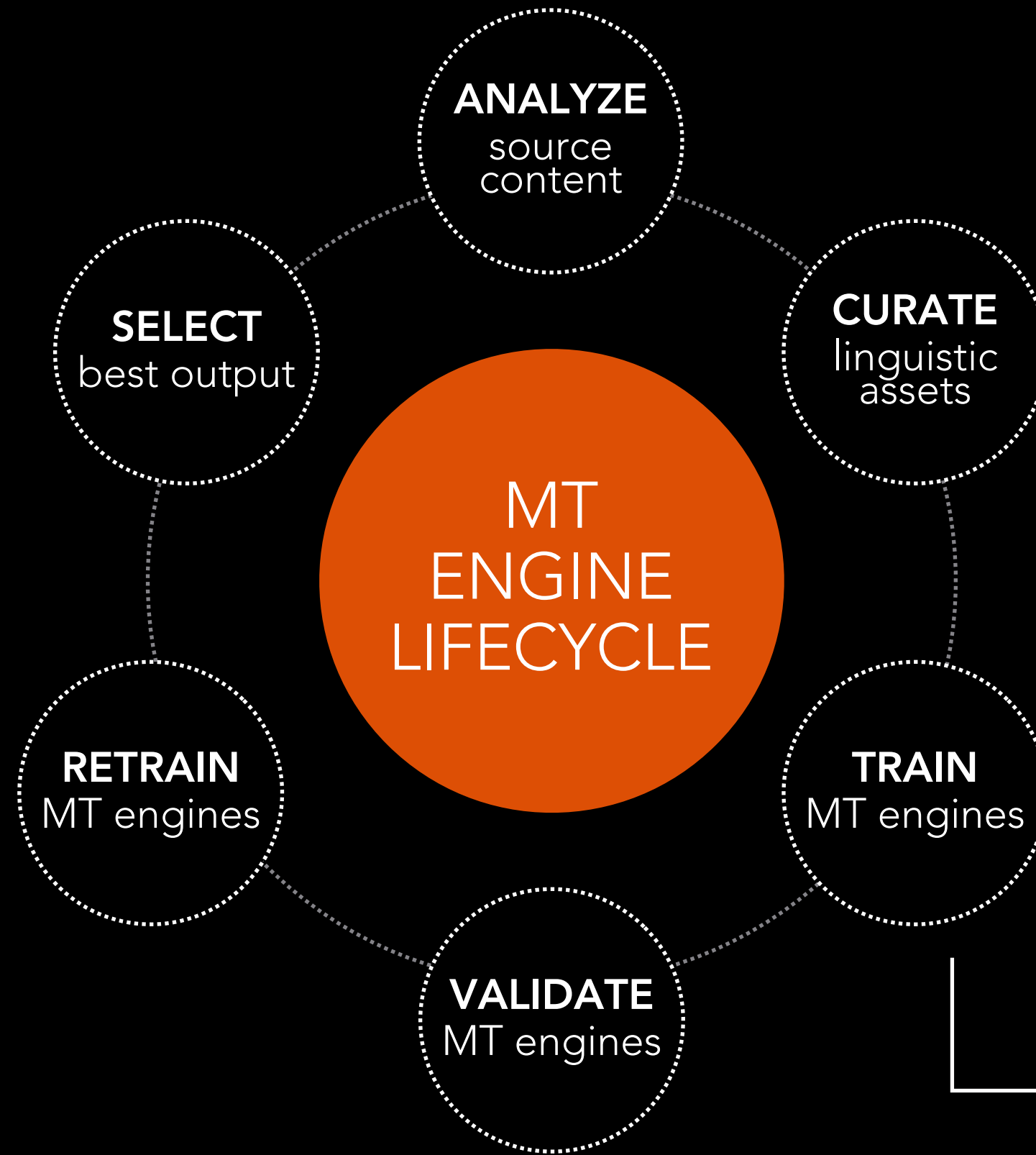
Let's talk about the proper way to curate data.



CLIENT CASE STUDY: The Lifecycle of Content Curation Using weMT



ANALYTICS
AT EACH STEP



THE
LIFECYCLE
DIAGRAM



ANALYZE

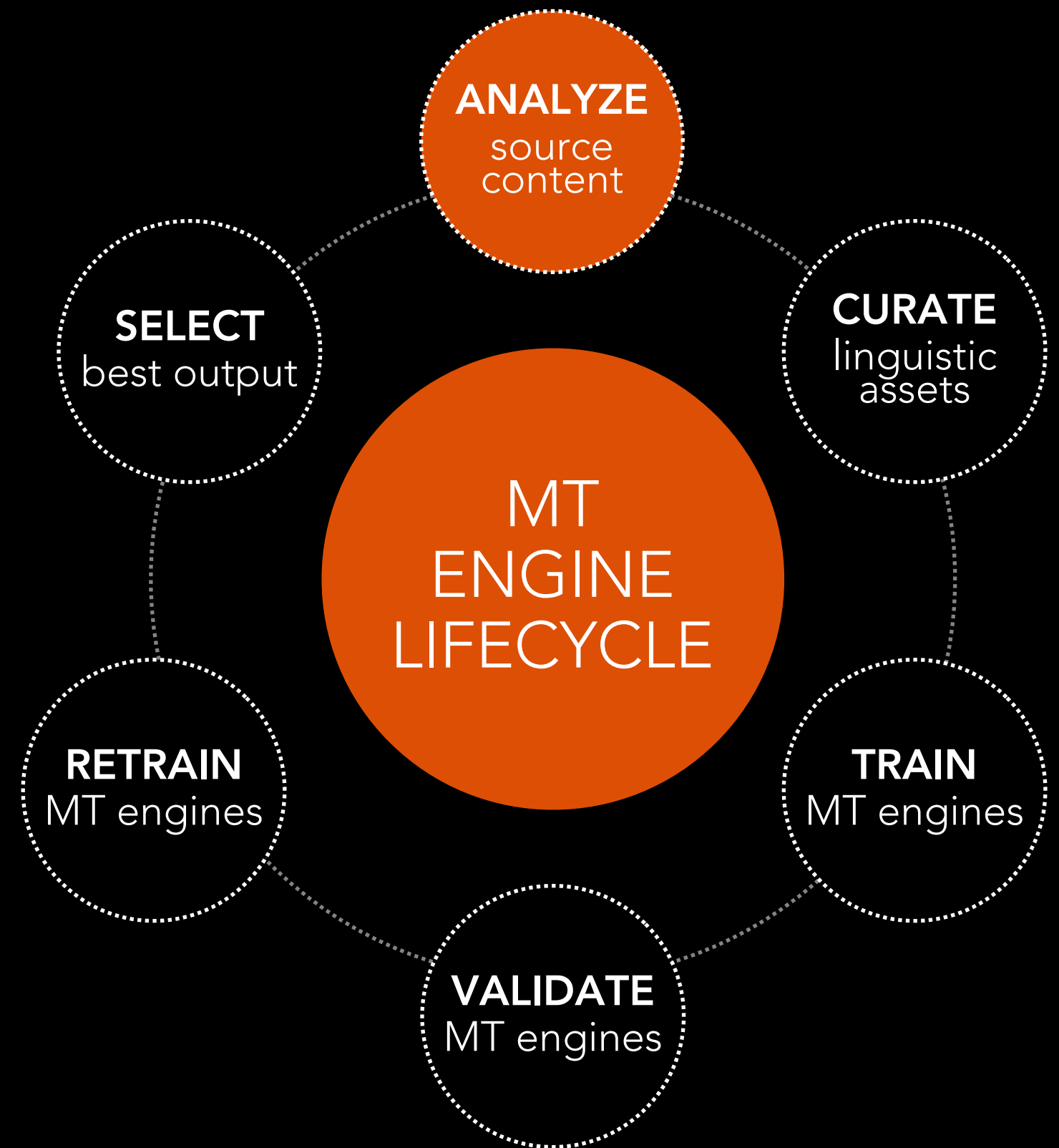
TWO STEPS: CLEAN & ANALYZE

CLEAN

- Remove Markup
- Remove Bullets
- Dedupe
- Remove Misaligned Tus
- Report on Discards (Potential Use as LM)

ANALYZE: Tools Used

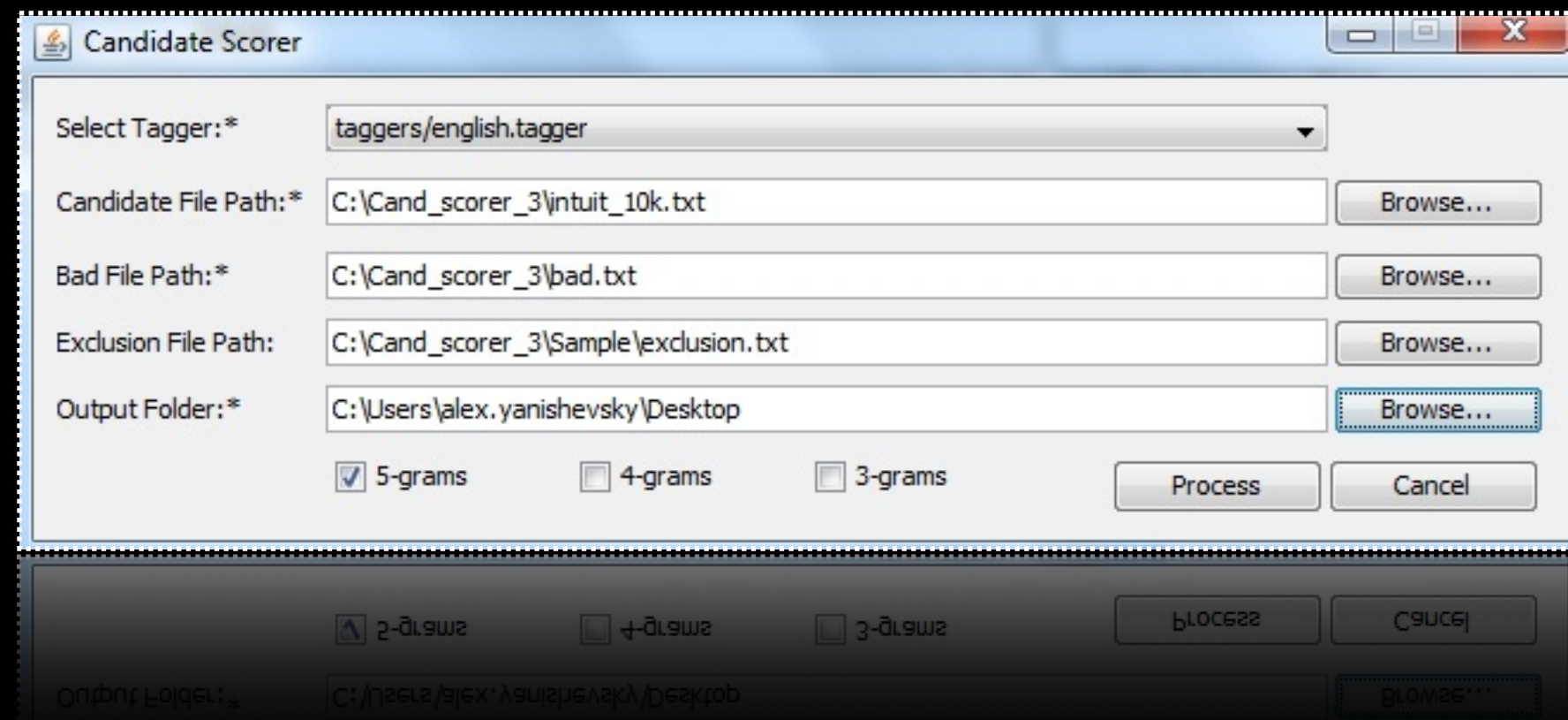
- Candadite Scorer — Analysis of Content Suitability & Domain
- Style Scorer — Adherence to Client's Style



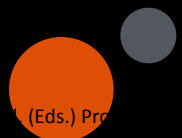
ANALYZE: CANDIDATE SCORER

How Good is Candidate Text?

- ✓ Take Historically "Bad" Text
- ✓ POS Tagger on "Bad" Text & Candidate Text
- ✓ Exclude List to Reduce False Positives
- ✓ Lower Score is Better (Candidate Text Does NOT Match "Bad" Text)



Compare to Historically "Bad" Text
LOWER SCORE IS BETTER



ANALYZE: CANDIDATE SCORER 2

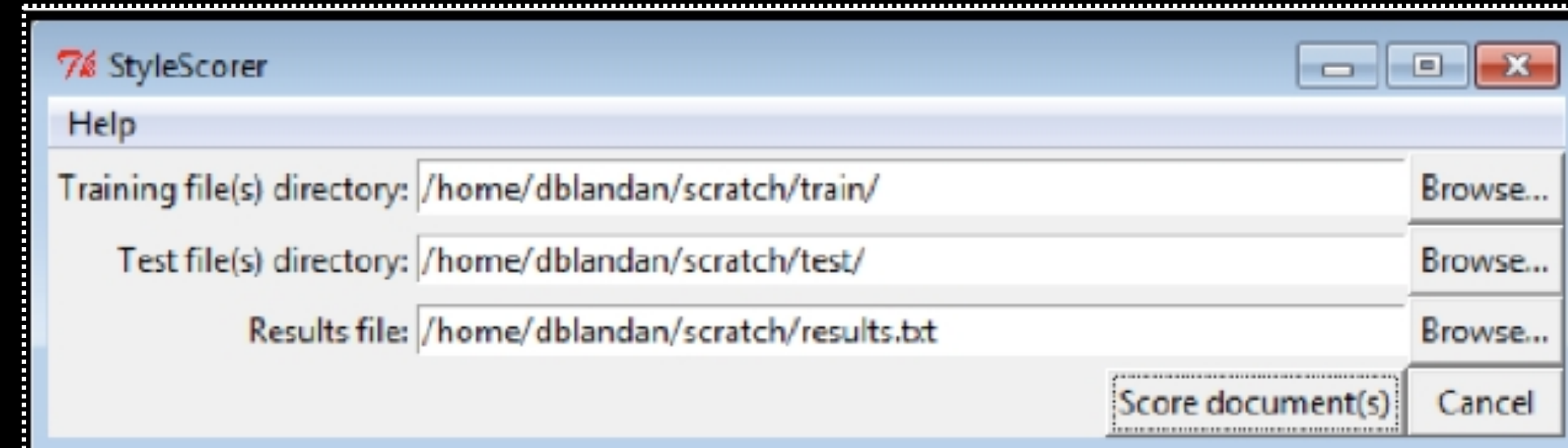
	A	B	C	D	E
1	Bad Construct	# Words	# of Occur	Candidate Phrase 1 (Pos Tagged)	Candidate Segment 1 (Pos Tagged)
2	NN IN DT NN NNS CC NNS	7	1	notification_NN of_IN any_DT printer_NN problems_NNS or_CC errors_NNS	You_PRP can_MD also_RB set_VB up_FP automatic_JJ email_NN notification_NN of_IN any_DT printer_NN problems_NNS or_CC errors_NNS to_TO email_JJ addresses_NNS you_PRP designate_VBP in_IN advance_NN . You_PRP can_MD use_VB this_DT procedure_NN to_TO cancel_VB the_DT regular_JJ acquisition_NN of_IN print_NN job_NN logs_NNS from_IN printers_NNS .
3	PRP MD VB DT NN TO VB	7	1	You_PRP can_MD use_VB this_DT procedure_NN to_TO cancel_VB	If_IN you_PRP are_VBP viewing_VBG regularly_RB acquired_VBN print_NN jobs_NNS ., selecting_VBG a_DT period_NN on_IN the_DT left_JJ side_NN of_IN the_DT window_NN displays_VBZ the_DT total_JJ costs_NNS for_IN that_DT period_NN and_CC the_DT amounts_NNS of_IN paper_NN and_CC ink_NN consumed_VBN .
4	DT NN IN DT JJ NN IN	7	1	a_DT period_NN on_IN the_DT left_JJ side_NN of_IN	If_IN you_PRP are_VBP viewing_VBG regularly_RB acquired_VBN print_NN jobs_NNS ., selecting_VBG a_DT period_NN on_IN the_DT left_JJ side_NN of_IN the_DT window_NN displays_VBZ the_DT total_JJ costs_NNS for_IN that_DT period_NN and_CC the_DT amounts_NNS of_IN paper_NN and_CC ink_NN consumed_VBN .
5	NN IN DT JJ NN IN DT	7	1	period_NN on_IN the_DT left_JJ side_NN of_IN the_DT	If_IN you_PRP are_VBP viewing_VBG regularly_RB acquired_VBN print_NN jobs_NNS ., selecting_VBG a_DT period_NN on_IN the_DT left_JJ side_NN of_IN the_DT window_NN displays_VBZ the_DT total_JJ costs_NNS for_IN that_DT period_NN and_CC the_DT amounts_NNS of_IN paper_NN and_CC ink_NN consumed_VBN .
6	IN DT JJ NN IN DT NN	7	1	on_IN the_DT left_JJ side_NN of_IN the_DT window_NN	Replace_VB the_DT paper_NN in_IN the_DT printer_NN with_IN the_DT correct_JJ paper_NN .
7	DT NN IN DT NN IN DT	7	1	the_DT paper_NN in_IN the_DT printer_NN with_IN the_DT	Replace_VB the_DT paper_NN in_IN the_DT printer_NN with_IN the_DT correct_JJ paper_NN .
8	DT NN IN DT NN IN DT	7	1	the_DT paper_NN in_IN the_DT printer_NN with_IN the_DT	Replace_VB the_DT paper_NN in_IN the_DT printer_NN with_IN the_DT correct_JJ paper_NN .
9	IN DT JJ NN IN DT NN	7	1	on_IN the_DT left_JJ side_NN of_IN the_DT window_NN	Replace_VB the_DT paper_NN in_IN the_DT printer_NN with_IN the_DT correct_JJ paper_NN .

	A	B	C
1	Candidate Segment		# of Occurrences
2	If you are viewing regularly acquired print jobs, selecting a period on the left side of the window displays the total costs for that period and the amounts of paper and ink consumed.	7	
3	You can use this procedure to cancel the regular acquisition of print job logs from printers.	4	
4	You can also set up automatic email notification of any printer problems or errors to email addresses you designate in advance.	3	
5	Replace the paper in the printer with the correct paper.	3	
6	Displays the ink level of every color in the printer.	2	
7	You can check the status of the hard disk and the documents saved on the hard disk.	1	
8	If the job has multiple pages, the size of the last page is displayed.	1	
9	If you do not set the standard prices, you cannot set the price for each roll paper.	1	
10	Enter the Width and Length of the paper, as well as the Price for the area of the configured width x height.	1	
11	The job items and details that can be displayed in the job list area are as follows.	1	
12	This feature collects print job logs from the printer at regular intervals and saves the logs on your computer.	1	
13	You can use the Job sheet in imagePROGRAF Status Monitor for operations such as pausing and canceling print jobs.	1	
52	Total	26	

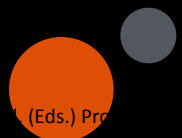
ANALYZE: STYLE SCORER

Combines PPL Ration, Dissimilarity Score & Classification Score

- ✓ Each Document Receives a Score from 0-4
- ✓ Higher Score Indicates Better Match to Style Established by Client's Documents
- ✓ Does NOT Require Parallel Data
- ✓ Source Scored for Training/Tuning Suitability

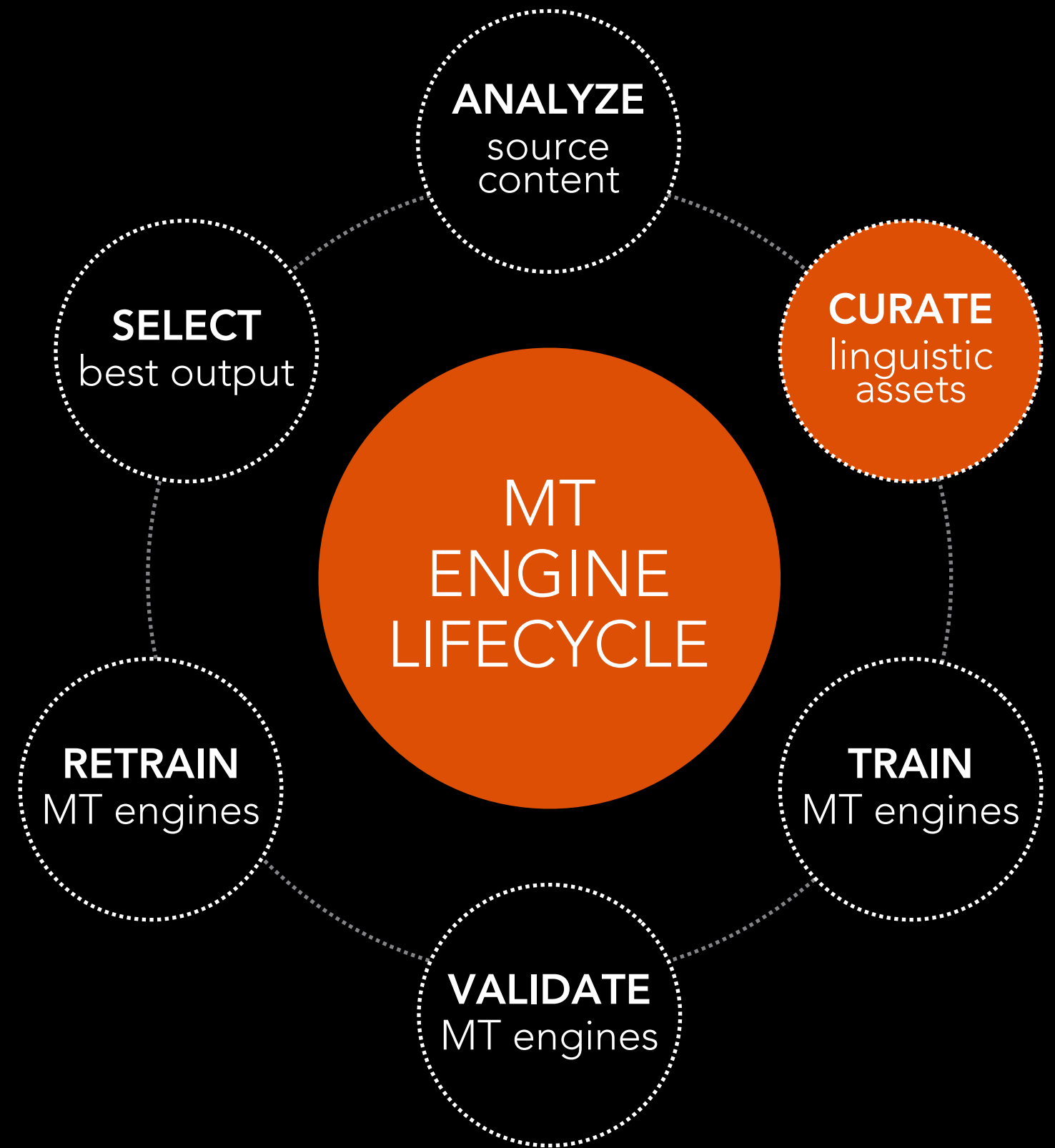


Filename	Score
/home/dblandan/scratch/test/test001.txt	2.73
/home/dblandan/scratch/test/test002.txt	0.20
/home/dblandan/scratch/test/test003.txt	3.32
/home/dblandan/scratch/test/test004.txt	1.19
/home/dblandan/scratch/test/test005.txt	3.98
/home/dblandan/scratch/test/test006.txt	3.89
/home/dblandan/scratch/test/test007.txt	0.38
/home/dblandan/scratch/test/test008.txt	0.73
/home/dblandan/scratch/test/test009.txt	3.72
/home/dblandan/scratch/test/test010.txt	3.43
/home/dblandan/scratch/test/test011.txt	2.84
/home/dblandan/scratch/test/test012.txt	2.60



CURATE

Tools Used:
Perplexity Evaluator
Source Content Profiler



CURATE: PERPLEXITY EVALUATOR

- ✓ Build Language Model (LM) of Historically “Bad” Text, Historically “Good” Text & Domain-Specific Text
- ✓ Is Candidate Text Closer to Historically “Bad” Text, Historically “Good” Text & Domain-Specific Text
- ✓ What Domain is Candidate Text Closest To?

LOWER SCORE IS BETTER



CURATE: PERPLEXITY EVALUATOR 2

FILE LEVEL

		LM		
document		good	bad	client
	good	7.78	1004.34	924.53
	bad	2910.85	33.46	511.04
	client	144.21	232.00	49.42

TU LEVEL

```
<tu srclang="EN-US" tuid="75438"> <prop type="x-ppl:train2">208</prop><prop type="x-
ppl:techdoc6">191.025</prop><prop type="x-ppl:support2">325.983</prop><prop type="x-
ppl:sales1">97.0736</prop><prop type="x-ppl:productLoc1">396.398</prop><prop type="x-
ppl:legal1">617.876</prop><tuv xml:lang="EN-US"> <seg>Consistent feature set across multiple
platforms (Windows, Mac, iOS, Android).</seg> </tuv> <tuv changedate="20140325T122530Z"
changeid="serviceaaa" creationdate="20140325T122530Z" creationid="serviceaaa"
lastusedate="20140325T122530Z" usagecount="0" xml:lang="ES-XL"> <prop type="x-
ALS:Context">TEXT</prop> <prop type="x-ALS:Source File">\\DATA\TC\39720\SRC\EN-US
\co-02_battle-card_en\co-02_battle-card_en.inx</prop> <seg>Conjunto de características
coherente en varias plataformas (Windows, Mac, iOS, Android)</seg> </tuv> </tu>
```



CURATE: SOURCE CONTENT PROFILER

- ✓ A Collaborative Project Between CNGl & Several Industry Partners
- ✓ Analyses Source Content & Group Into “Profiles” Based on Set of “Features” — Syntax, Grammar, Readability Score, Terminology, Engineering Characteristics — DNT, Glossary Hits & Tag Ratio

The screenshot shows the 'Source Content Profiler' web application. At the top, there is a navigation menu with links for 'ABOUT US', 'RESEARCH', 'INDUSTRY', 'PEOPLE', 'NEWS', 'OUTREACH', and 'CONTACT'. The CNGl logo (Centre for Global Intelligent Content) is in the top left. A search bar is located in the top right. Below the header, there is a table with columns for 'Filename' and 'Progress'. The table lists several files, all with a 'complete' status and a 'view' link. Below the table, there is a 'Results' section with a 'SCP Score' of 81. The results are presented as a list of metrics with corresponding values and icons:

Metric	Value
Words	1,003
Sentences	52
Characters per word (AVG)	4
Words per sentence (WPS)	29
Readability score	101.54
Sentences with unusual POS sequences	21
Grammar issues	0
Spelling issues	59
Language model issues	52
Passive voice issues	22

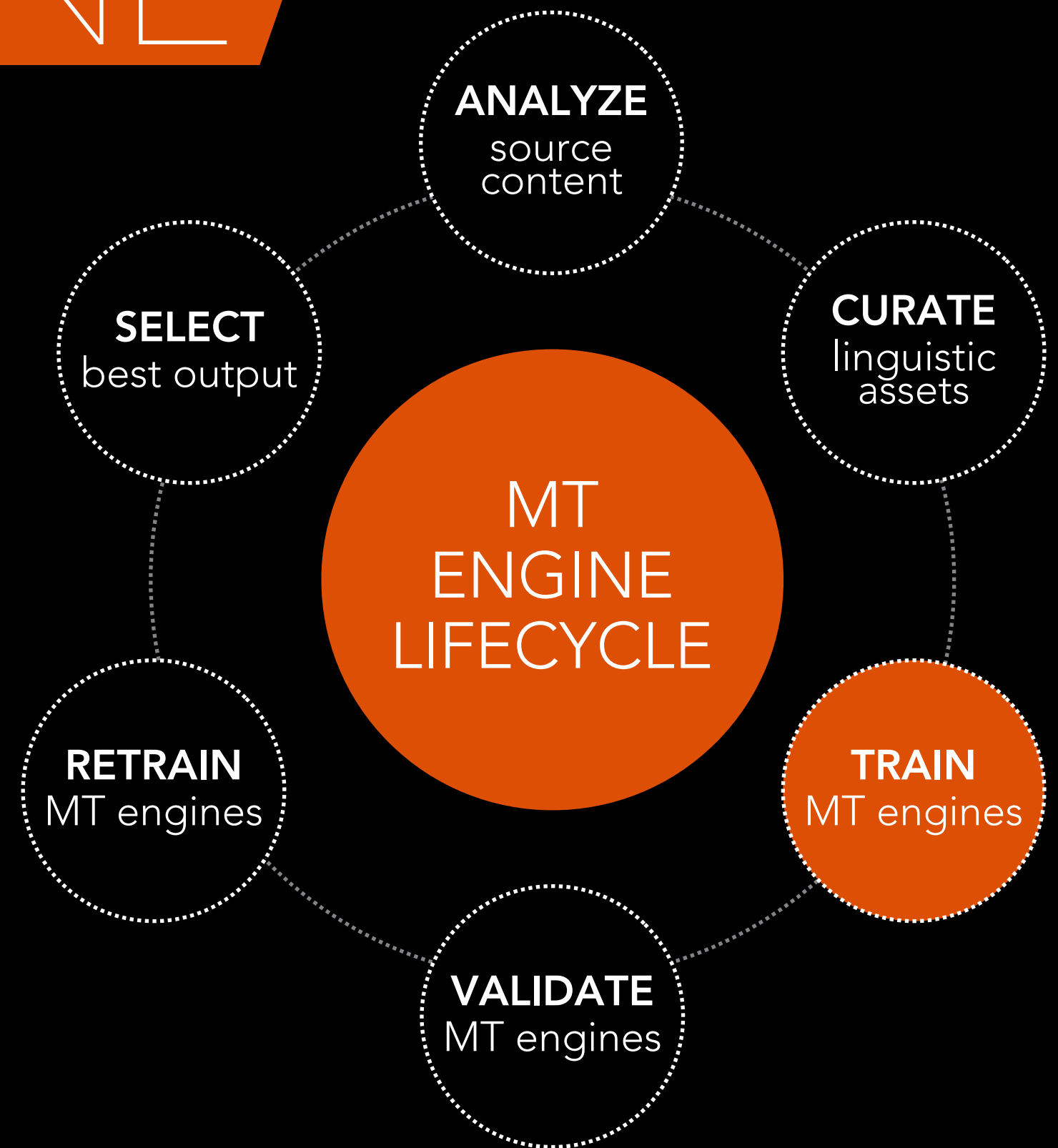
At the bottom of the results section, there is a button labeled 'View Source Content Issues Highlighter'.

TRAIN MT ENGINE

Target Train, Tune, Test Data Selection:

- Random
- Frequency
- Length
- Product or Domain

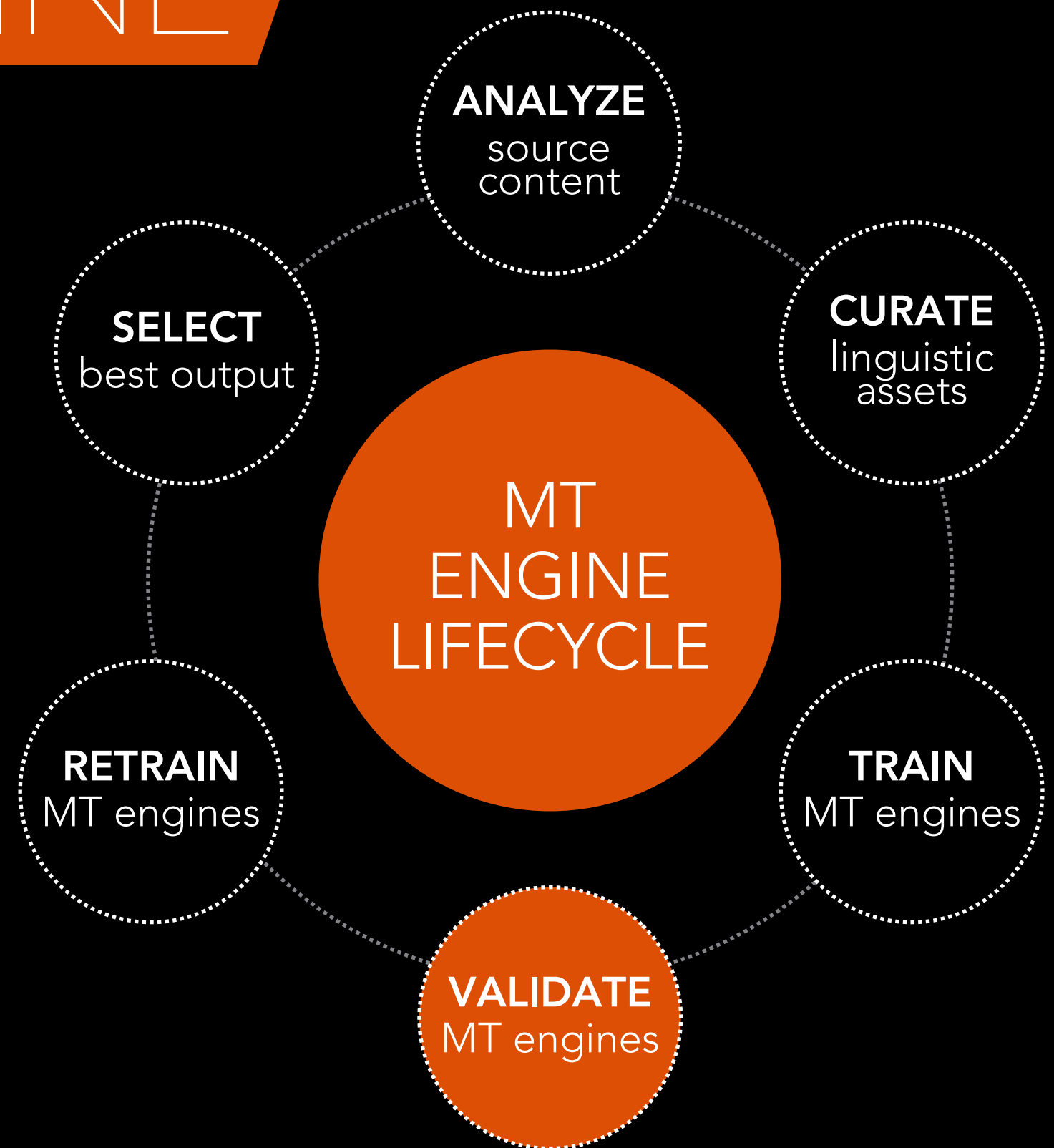
Check for "Cheating"



VALIDATE MT ENGINE

Targeted Error Correction of OOVs Based On:

Ratio Per Sentence
Token Count
Ratio Per File



VALIDATE MT ENGINE 2: weSCORE

Dashboard for Viewing MT Metrics

- ✓ Tokenizes Input from Variety of Formats & Runs Several Scoring Algorithms in Parallel
- ✓ Exports Detailed Analysis to Spreadsheet for Sentence-By-Sentence Review

Root Folder Locations & Options

Source root folder: Browse...

Hypothesis (MT) root folder: C:\Projects_Local\Test\00_wescore_test_jp\hyp Browse...

Reference (PE) root folders: C:\Projects_Local\Test\00_wescore_test_jp\ref1 C:\Projects_Local\Test\00_wescore_test_jp\ref2 Add - Browse... Add/Edit Path Remove Selected Remove All

Format Selection

TMX 1.4 SDL Trados TTX WorldServer XLIFF Extract XLF from XLZ Plain Text

Strict translation unit ID Matching Ignore duplicate segments Ignore source inline tags for string equality comparisons

Required Settings

Locale: JA BLEU & NIST 13m METEOR 1.4 GTM 1.4 Levenshtein PE Dist. Lowercase scoring

Process Scores View Job Log Close

Results Summary

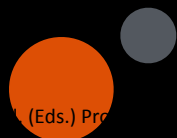
Automatic scores:

BLEU	NIST	METEOR	GTM	Avg. PE	TERp
22.61	4.67	39.13	59.49	87.08 %	Calc. Err.

Export Results

Levenshtein PE distance:

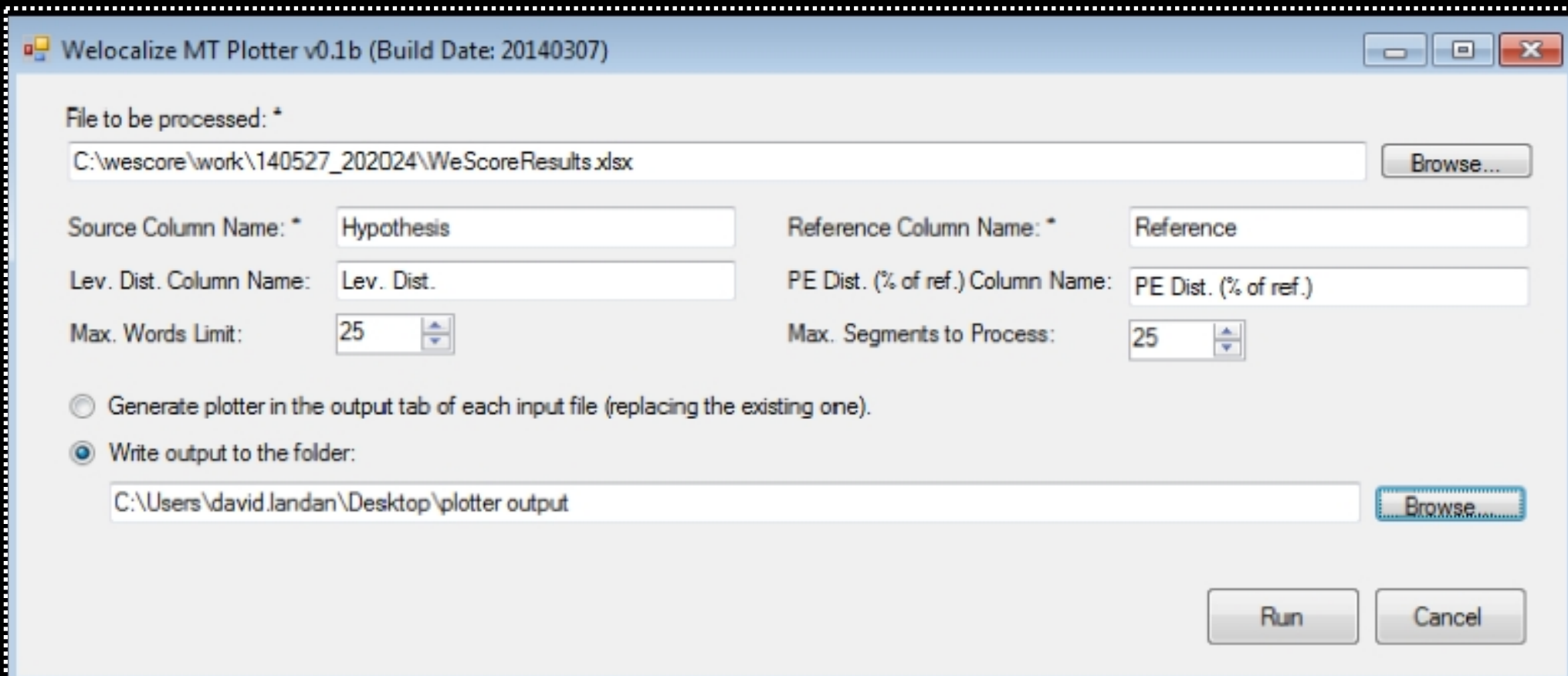
Hypothesis	Reference	Ref. Set	Lev. Dist.	PE Dist. (% of ref. len.)
安全性とセキュリティ	安全性とセキュリティ	1	0	0.00 %
クリーニングとメンテナンス	クリーニングとメンテナンス	1	0	0.00 %
カウンタ、そして、の重要性	カウンタマネージャの重要性	1	6	46.15 %



VALIDATE MT ENGINE 3: wePLOTTER

Visualization of Individual Segments Based on Candidate & Reference

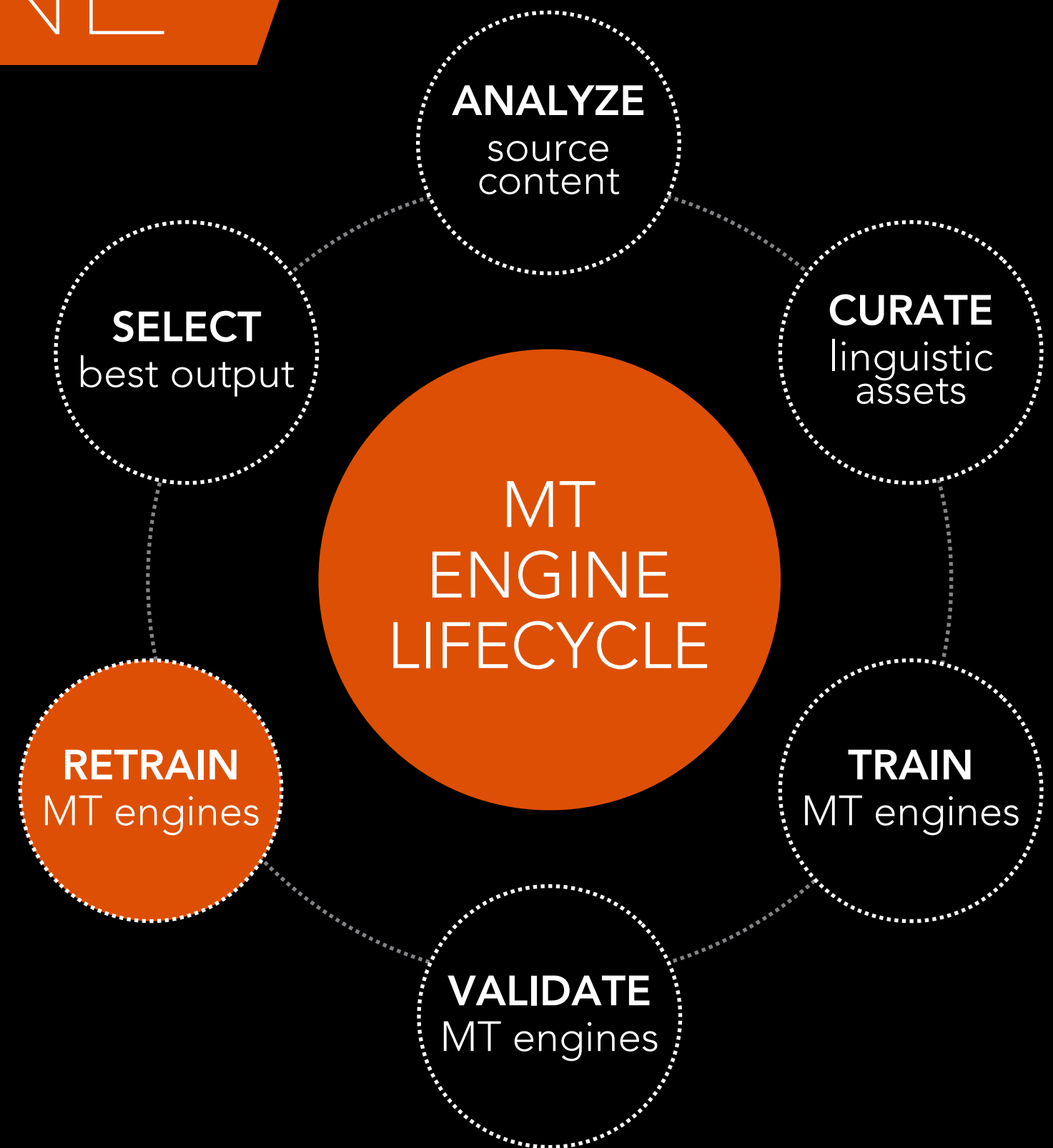
- ✓ Precision
- ✓ Recall



	Принудительное	завершение	этого	достижения	после	определенной	даты.	Если	дата	не	выбрана,	вы	сможете	завершить	это	достижение	в	любое	время.	
принудительное	X																			
Завершение		X																		
после					X															
определенного																				
этого			X																	
достижения.																				
Если								X												
даты																				
не										X										
выбраны,																				
Дата									X											
завершения																				
этого			X																	
достижения				X																
можно																				
будет																				
в																	X			
любое																		X		
время.																			X	

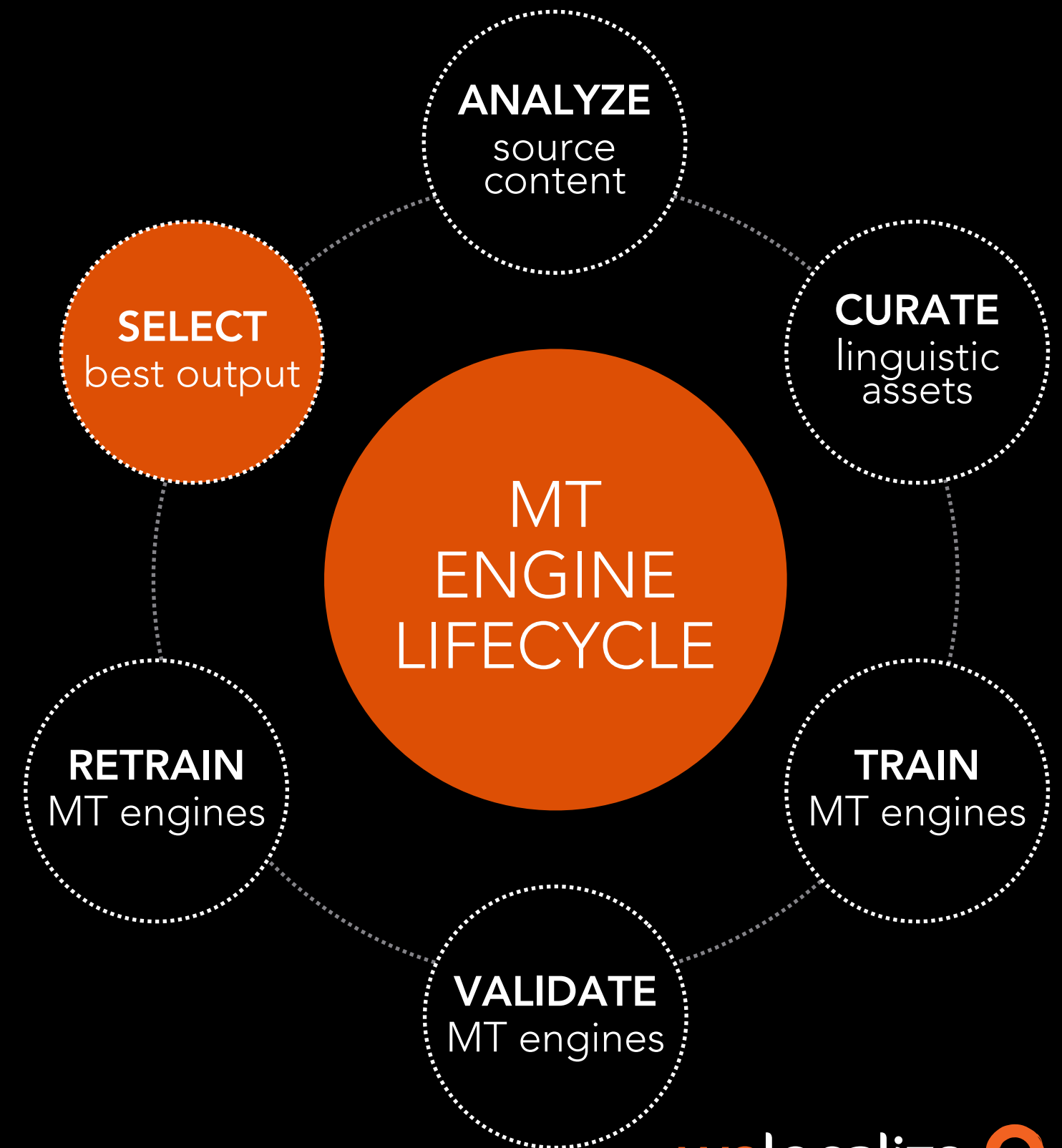
RETRAIN MT ENGINE

Correct OOV
Implement Feedback From Production & LQA

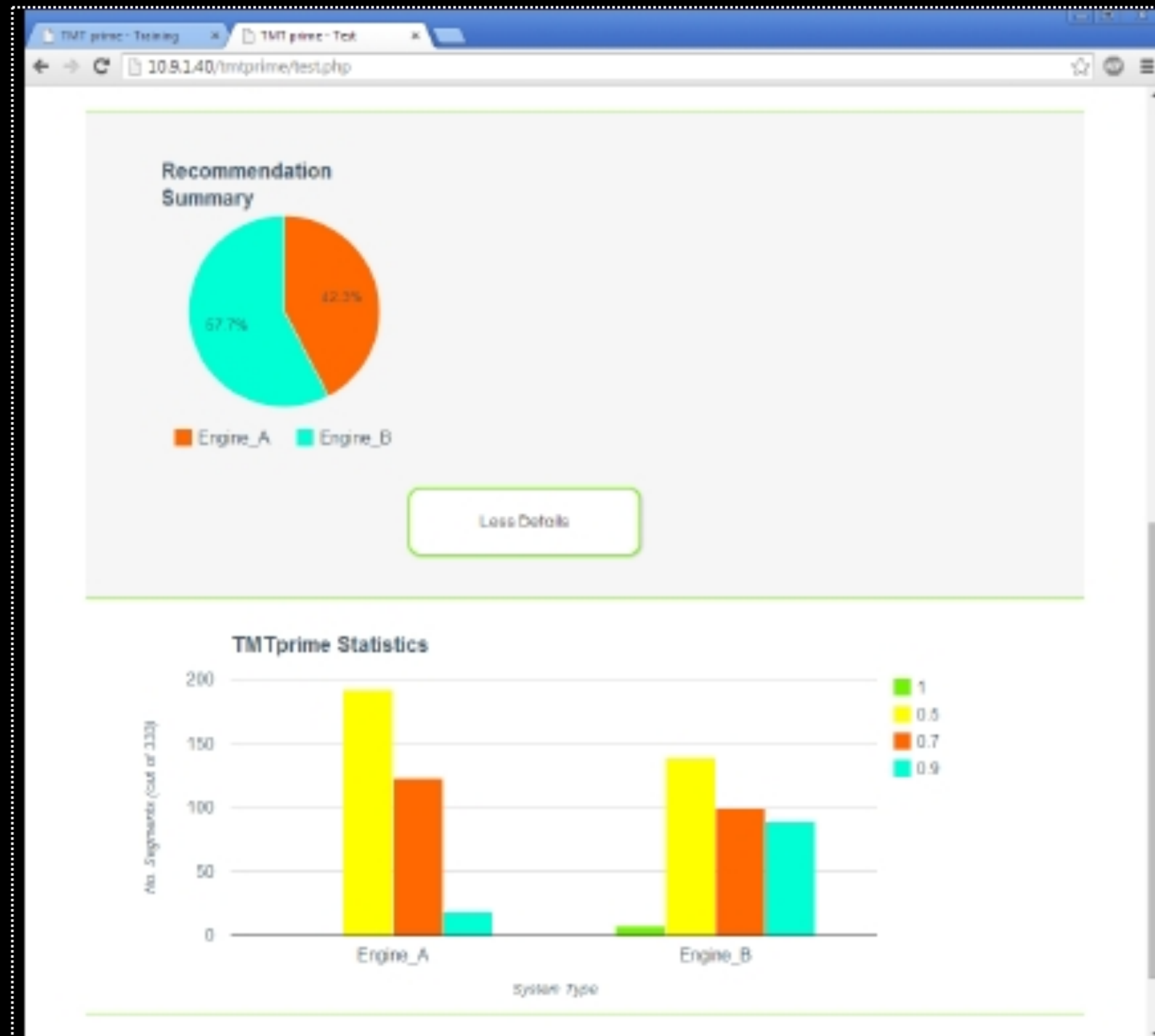


Select Best Output: TMT PRIME

- TMT Prime Provides a Commercially Viable Platform that Allows for the Seamless Integration of MT/TM Systems
 - MT/TM System Recommender
- Recommendation Based on **TMT Prime Score**
 - Confidence Score
 - “Intelligent” Fuzzy Match
- Translation Can Be Obtained from Either In-House or Third Party Systems (Bing or Google)
- A Cloud-Based Framework that Effectively Selects Translation Outputs by Using Different MT/TM Systems



Select Best Output: TMT PRIME 2



- ✓ All TMs for All Languages Have Been Profiled
- ✓ Implemented Tool to Query Most Relevant TUs by Domain
- ✓ BLEU Scores Increased by 10 Plus Points on Domain-Specific Engines for Both Asian & European Languages
- ✓ Productivity Increased by 5-7% on Domain-Specific Engines for Both Asian & European Languages

CLIENT RESULTS



- ✓ Develop Domain-Specific Engines for All Languages Localized by the Client (50-60 Engines)
- ✓ Assess Productivity Increase for All Languages & Engines as Related to BLEU Score Improvement
- ✓ Use TMT Prime to Recommend & Get Best Engine Translation on a Per Segment Bases as Opposed to Current File Level

NEXT STEPS



THANK YOU

———— ALEXYANISHEVSKY ————

Welocalize
October 2014

