Towards Malleable & Data-Centric ML Systems

Karan Goel, Hazy Research Lab, Stanford University
kgoel@cs.stanford.edu

Data Quality Assessment for Machine Learning, KDD 2021
Data-Centric AI

In practice, “engineer’s intuition” and heuristics on data differentiates ML that “works”

Emerging engineering discipline within ML: tools, methods, study of data for ML
Data-Centric AI

In practice, “engineer’s intuition” and heuristics on data differentiates ML that “works”

Emerging engineering discipline within ML: tools, methods, study of data for ML

https://github.com/hazyresearch/data-centric-ai

Table of Contents

0. Background
1. Data Programming & Weak Supervision
2. Data Augmentation
3. Self-Supervision
4. The End of Modellitis
5. Fine-Grained Evaluation
6. Robustness
7. Data Cleaning
8. MLOps
9. Data Selection (Under construction)
10. Data Privacy (Under construction)
11. Data Flow (Under construction)
12. Multi-Task & Multi-Domain Learning (Under construction)
13. Emerging Trends
14. Applications
15. Case Studies
16. Awesome Lists

Lots to learn about this area together: contribute and use!
ML Paradigms

**Standard training**
Training data, deployment data

Data → Model

Karan Goel | @krandiash
ML Paradigms

**Standard training**
Training data, deployment data

**Pre-training**
Models learn embedding representations of the underlying training data *without* manual labels.

Embeddings become the data: used in downstream models e.g. BERT & sentiment analysis
ML Paradigms

**Standard training**
Training data, deployment data

**Pre-training**
Models learn embedding representations of the underlying training data without manual labels.

Embeddings become the data: used in downstream models e.g. BERT & sentiment analysis

Measuring quality across these different paradigms means dovetailing:

- "data-only" quality
- fine-grained model quality

Karan Goel | @krandiash
Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging

Error analysis

During error auditing, where examples of false positive and false negative predictions from the pretrained model were visually reviewed by a board certified radiologist [3], it was observed that pneumothorax cases without chest drains were highly prevalent in the set of false negatives. A chest (LOR) for each element of the test set. Due to higher prevalence of scans with chest drains in the dataset, clear discriminative features of a chest drain, and high label quality for the scans with chest drains, we hypothesize that a model trained on the CXR14 dataset will attain higher performance on the pneumothorax subclass with chest drains than that without chest drains.
How does the embedding ecosystem change ML pipeline management?
Preprint up and tutorial at VLDB next week!
Gives a broad perspective (academic + industry) on these new challenges
**Goal:** Design data-centric systems that are easy to modify e.g. to perform bug fixes, hit performance targets and meet ethical standards.

(Some) System Components

Stakeholders can interactively
1. (Measure) Understand fine-grained system capabilities & performance
2. (Monitor) Track data distribution shift & its effects over time
3. (Maintain) Perform atomic updates to data & then models
Plan: we’ll go over work in each of these buckets

**Karan Goel**, Nazneen Rajani*, Jesse Vig, Zachary Taschdjian, Mohit Bansal, Christopher Re. NAACL Demo 2021.

Mandoline: Model Evaluation under Distribution Shift
Mayee Chen*, **Karan Goel**, Nimit Sohoni*, Fait Poms, Kayvon Fatahalian, Christopher Re. ICML 2021.

Model Patching: Closing the Subgroup Performance Gap with Data Augmentation.
**Karan Goel**, Albert Gu, Yixuan Li, Christopher Re. ICLR 2021.

**Karan Goel**, Laurel Orr, Nazneen Rajani, Jesse Vig, Christopher Re. NAACL Industry 2021.
Big bottleneck to monitor and maintain: carefully measure fine-grained model performance across critical data slices.


Karan Goel*, Nazneen Rajani*, Jesse Vig, Zachary Taschdjian, Mohit Bansal, Christopher Re. NAACL Demo 2021.
Robustness Gym: Fine-Grained Evaluation Tools

github.com/robustness-gym/robustness-gym

Many Evaluation Strategies

- Critical data slices
- Bias / fairness concerns
- Sensitivity to perturbations
- Invariance to transformations
- and more!

Python toolkit for building evaluations and generating visualizations.

Karan Goel | @krandiash
# Load data & model

datapanel = rg.DataPanel.from_huggingface('imdb')
model = rg.Model.from_huggingface('bert_imdb')

# Run processing workflows
spacy = rg.ops.SpacyOp()
datapanel = spacy(datapanel, ['review'])

# Track evaluations and results
devbench = DevBench()

# Subpopulation of horror movie reviews
def is_horror(example):
tokens = set(rg.lookup(example, SpacyOp, ['review']))
return True if tokens.intersection(horror_movie_list) else False

# Add evaluations
devbench = devbench

devbench.add_aggregators('bert_imdb': {'accuracy': accuracy_fn})

# Generate reports
report = devbench.create_report()
report.figure()

# Save
devbench.write('sentiment.devbench')

# Reproduce!
rg.DevBench.read('sentiment.devbench').create_report().figure()
# Robustness Report for bert-base-uncased on Natural Language Inference

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>F1</th>
<th>Class Dist</th>
<th>Pred Dist</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Constituency Tree Overlap (McCoy, 2019)</td>
<td>90.2</td>
<td>89.7</td>
<td>20 39 41</td>
<td>20 39 41</td>
<td>2.1K</td>
</tr>
<tr>
<td>High Constituency Tree Overlap (McCoy, 2019)</td>
<td>93.2</td>
<td>92.2</td>
<td>53 24 23</td>
<td>51 24 25</td>
<td>1.99K</td>
</tr>
<tr>
<td>Negation @ hypothesis (Naik, 2018)</td>
<td>90.8</td>
<td>86.0</td>
<td>22 17 61</td>
<td>23 13 64</td>
<td>109</td>
</tr>
<tr>
<td>Negation @ premise (Naik, 2018)</td>
<td>79.5</td>
<td>79.5</td>
<td>31 38 31</td>
<td>38 26 36</td>
<td>39</td>
</tr>
<tr>
<td>Possessive Preposition @ hypothesis (Chen, 2020)</td>
<td>90.9</td>
<td>90.9</td>
<td>39 34 27</td>
<td>36 35 29</td>
<td>585</td>
</tr>
<tr>
<td>Quantifier @ hypothesis (Chen, 2020)</td>
<td>88.2</td>
<td>88.3</td>
<td>38 34 28</td>
<td>39 34 28</td>
<td>170</td>
</tr>
<tr>
<td>Temporal Preposition @ hypothesis (Chen, 2020)</td>
<td>87.7</td>
<td>86.0</td>
<td>13 61 25</td>
<td>13 61 25</td>
<td>106</td>
</tr>
<tr>
<td>Low Lexical Overlap (McCoy, 2019)</td>
<td>90.5</td>
<td>89.6</td>
<td>20 33 47</td>
<td>20 33 46</td>
<td>2.04K</td>
</tr>
<tr>
<td>High Lexical Overlap (McCoy, 2019)</td>
<td>92.7</td>
<td>91.9</td>
<td>52 29 19</td>
<td>51 30 20</td>
<td>1.98K</td>
</tr>
<tr>
<td>BAE (Garg, 2019)</td>
<td>80.3</td>
<td>78.4</td>
<td>13 58 29</td>
<td>12 48 40</td>
<td>2.92K</td>
</tr>
<tr>
<td>Easy Data Augmentation (Wei, 2019)</td>
<td>82.3</td>
<td>82.2</td>
<td>34 33 33</td>
<td>28 36 36</td>
<td>9.84K</td>
</tr>
<tr>
<td>Keyboard Character Errors (Ma, 2019)</td>
<td>65.8</td>
<td>65.4</td>
<td>34 33 33</td>
<td>24 33 44</td>
<td>9.14K</td>
</tr>
<tr>
<td>Synonym Substitution (Ma, 2019)</td>
<td>75.4</td>
<td>75.1</td>
<td>34 33 33</td>
<td>24 36 40</td>
<td>9.84K</td>
</tr>
<tr>
<td>SNLI (Bowman, 2015)</td>
<td>90.9</td>
<td>90.9</td>
<td>34 33 33</td>
<td>33 33 34</td>
<td>9.84K</td>
</tr>
</tbody>
</table>
Many Tasks

Ongoing: make this even easier and unified for many modalities!
Now we can measure: does performance degrade over time?

Need to understand data distribution shift - then calculate metrics on unlabeled data.

Mandoline: Model Evaluation under Distribution Shift

Mayee Chen*, Karan Goel*, Nimit Sohoni*, Fait Poms, Kayvon Fatahalian, Christopher Re. ICML 2021.
Mandoline: Evaluation under Distribution Shift
Available in RG

Mandoline: user guided framework to evaluate models using slices.
Slice: user-defined grouping of data

Mandoline: Evaluation under Distribution Shift
Available in RG

Source Accuracy: 91%
Target Accuracy: 84%

Slice: user-defined grouping of data

negation contains not, n't
male pronoun contains he, him
strong sentiment contains love, adore

\( g_1(x) \) \( g_2(x) \) \( g_3(x) \)

(Source) Labeled Validation Set
I love eating ice-cream.
He loved walking on the beach.
He didn’t like drinking coffee.

(Target) Unlabeled Test Set
He does not love eating scones.
He loves taking risks.
She likes drinking coffee.

Target performance estimate up to 5x more accurate than standard importance weighting baselines.

Karan Goel | @krandiash
Fix spurious correlations and data imbalance with learned data augmentations


Case study of targeted improvements from weak labeling

Make targeted fixes to model behavior by improving data quality

Fix spurious correlations and data imbalance with learned data augmentations

Model Patching: Closing the Subgroup Performance Gap with Data Augmentation.  
Karan Goel*, Albert Gu, Yixuan Li, Christopher Re. ICLR 2021.

Case study of targeted improvements from weak labeling

Karan Goel, Laurel Orr, Nazneen Rajani, Jesse Vig, Christopher Re. NAACL Industry 2021.
ideal: models succeed when deployed
reality: good on average but not robust
reality: good on average but not robust
Our contribution
improve robustness and mitigate the subgroup gap with **model patching**
class

subgroup 1

subgroup 2

subgroup 3

no cancer

bandages

no bandages

✅ ✅ ✅ ✅ ✗ ✗

each class consists of known subgroups
Learning subgroup invariant representations

Minimize

representation carries no subgroup information

\[ \text{Mutual Information}(\text{subgroup}; \text{representation} | \text{class}) \]
Coupled Sets

example

class subgroup 1

plain

class subgroup 2

counterpart in subgroup 2

class subgroup 3

counterpart in subgroup 3

example^2

example^3

...
Coupled Sets

example

class subgroup 1

counterpart in subgroup 2

counterpart in subgroup 3

coupled set

counterpart with no bandages

no cancer bandages

no cancer no bandages
Coupled Sets

\[ \text{Mutual Information}(\text{subgroup} ; \text{representation} | \text{class}) \geq \text{Mutual Information}(\text{subgroup} ; \text{representation} | \text{coupled set}) \]
Model Patching

\[ \text{Mutual Information}(\text{subgroup} ; \text{representation} \mid \text{coupled set}) \]
Model Patching

**Stage 1**
learn coupled set generation model (CycleGAN)

**Stage 2**
robust training with coupled set generation

Enforce consistency over coupled set
Experiments

Robust accuracy 5.3%↑  Subgroup performance gap 24x↓

1. **Other consistency losses** degrade robust accuracy by 2.5%
2. **Heuristic augmentations** degrade the subgroup gap by 3.35%
3. **Heuristic + learned augmentations** improve robust accuracy by 1.5%
Malleable & Interactive ML Systems

Make targeted fixes to model behavior by improving data quality

Fix spurious correlations and data imbalance with learned data augmentations

Model Patching: Closing the Subgroup Performance Gap with Data Augmentation.  
Karan Goel*, Albert Gu, Yixuan Li, Christopher Re. ICLR 2021.

Case study of targeted improvements from weak labeling

Karan Goel, Laurel Orr, Nazneen Rajani, Jesse Vig, Christopher Re. NAACL Industry 2021.
Background: Named Entity Linking

Named Entity Linking
map "strings" to "things" in a knowledge base like Wikipedia

When did England last win the football world cup?

England National Football Team
FIFA World Cup

Question Answering System

1966

Downstream System

A correct NEL is required by the downstream system!

Karan Goel | @krandiash

1. **Analyze** commercial & academic NEL systems “off-the-shelf” using RG.
2. **Repurpose** Bootleg system to guide behavior for sports QA.
1. **Analyze**

Results on a subset of Wikipedia

<table>
<thead>
<tr>
<th></th>
<th>Amazon</th>
<th>Google</th>
<th>Microsoft</th>
<th>Bootleg</th>
<th>Rel</th>
<th>Pop</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>everything</td>
<td>49.9</td>
<td>49.2</td>
<td>66.8</td>
<td>78.7</td>
<td>51.2</td>
<td>53.6</td>
<td>49.5K</td>
</tr>
<tr>
<td>popular</td>
<td>66.7</td>
<td>82.7</td>
<td>71.7</td>
<td>83.9</td>
<td>80.4</td>
<td>74.3</td>
<td>8.32K</td>
</tr>
<tr>
<td>everything</td>
<td>64.9</td>
<td>73.3</td>
<td>73.2</td>
<td>85.1</td>
<td>71.6</td>
<td>72.8</td>
<td>15.5K</td>
</tr>
<tr>
<td>kg-relation</td>
<td>61.6</td>
<td>80.3</td>
<td>69.9</td>
<td>83.8</td>
<td>77.3</td>
<td>74.7</td>
<td>5.07K</td>
</tr>
<tr>
<td>one-of-the-two</td>
<td>48.7</td>
<td>65.3</td>
<td>65.3</td>
<td>77.2</td>
<td>65.3</td>
<td>45.8</td>
<td>885</td>
</tr>
<tr>
<td>share-1-type</td>
<td>79.5</td>
<td>81.5</td>
<td>89.4</td>
<td>94.7</td>
<td>85.4</td>
<td>80.3</td>
<td>1.25K</td>
</tr>
<tr>
<td>strong affordance</td>
<td>65.1</td>
<td>73.3</td>
<td>73.1</td>
<td>85.5</td>
<td>71.2</td>
<td>72.9</td>
<td>14.4K</td>
</tr>
<tr>
<td>unpopular</td>
<td>39.3</td>
<td>37.8</td>
<td>49.5</td>
<td>79.2</td>
<td>33.7</td>
<td>650</td>
<td></td>
</tr>
<tr>
<td>everything</td>
<td>44.4</td>
<td>40.4</td>
<td>65.1</td>
<td>77.2</td>
<td>44.3</td>
<td>45.3</td>
<td>30K</td>
</tr>
<tr>
<td>kg-relation</td>
<td>55.5</td>
<td>53.4</td>
<td>77.5</td>
<td>86.9</td>
<td>50.3</td>
<td>52.5</td>
<td>6.22K</td>
</tr>
<tr>
<td>one-of-the-two</td>
<td>43.0</td>
<td>39.2</td>
<td>66.5</td>
<td>80.6</td>
<td>46.5</td>
<td>39.6</td>
<td>2.13K</td>
</tr>
<tr>
<td>share-1-type</td>
<td>64.4</td>
<td>44.8</td>
<td>77.4</td>
<td>88.2</td>
<td>57.9</td>
<td>53.5</td>
<td>2.78K</td>
</tr>
<tr>
<td>strong affordance</td>
<td>45.6</td>
<td>41.1</td>
<td>66.2</td>
<td>78.5</td>
<td>44.3</td>
<td>46.2</td>
<td>24.4K</td>
</tr>
<tr>
<td>unpopular</td>
<td>27.2</td>
<td>29.1</td>
<td>45.4</td>
<td>61.7</td>
<td>28.4</td>
<td>3.51K</td>
<td></td>
</tr>
<tr>
<td>everything</td>
<td>33.8</td>
<td>34.9</td>
<td>55.1</td>
<td>65.2</td>
<td>27.3</td>
<td>40.4K</td>
<td>4.04K</td>
</tr>
<tr>
<td>kg-relation</td>
<td>44.0</td>
<td>35.3</td>
<td>79.4</td>
<td>80.4</td>
<td>29.4</td>
<td>961</td>
<td>36</td>
</tr>
<tr>
<td>one-of-the-two</td>
<td>45.5</td>
<td>31.2</td>
<td>65.6</td>
<td>74.9</td>
<td>49.8</td>
<td>279</td>
<td>12</td>
</tr>
<tr>
<td>share-1-type</td>
<td>31.5</td>
<td>34.2</td>
<td>62.9</td>
<td>80.7</td>
<td>36.0</td>
<td>461</td>
<td>14</td>
</tr>
<tr>
<td>strong affordance</td>
<td>34.9</td>
<td>37.4</td>
<td>57.8</td>
<td>67.9</td>
<td>27.3</td>
<td>3.14K</td>
<td>104</td>
</tr>
<tr>
<td>unpopular</td>
<td>29.2</td>
<td>34.2</td>
<td>55.8</td>
<td>69.2</td>
<td>28.6</td>
<td>449</td>
<td></td>
</tr>
</tbody>
</table>

Recall

Results on a subset of Wikipedia
**Results on a subset of Wikipedia**

<table>
<thead>
<tr>
<th></th>
<th>Amazon</th>
<th>Google</th>
<th>Microsoft</th>
<th>Bootleg</th>
<th>Rel</th>
<th>Pop</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>everything</td>
<td>49.9</td>
<td>49.2</td>
<td>66.8</td>
<td>78.7</td>
<td>51.2</td>
<td>52.8</td>
<td>49.5K</td>
</tr>
<tr>
<td>popular</td>
<td>66.7</td>
<td>82.7</td>
<td>71.7</td>
<td>83.9</td>
<td>80.4</td>
<td>74.5</td>
<td>8.32K</td>
</tr>
<tr>
<td>everything</td>
<td>64.9</td>
<td>73.3</td>
<td>73.2</td>
<td>85.1</td>
<td>71.6</td>
<td>72.8</td>
<td>15.5K</td>
</tr>
<tr>
<td>kg-relation</td>
<td>61.6</td>
<td>80.3</td>
<td>69.9</td>
<td>83.8</td>
<td>77.3</td>
<td>74.7</td>
<td>5.07K</td>
</tr>
<tr>
<td>one-of-the-two</td>
<td>48.7</td>
<td>65.3</td>
<td>65.3</td>
<td>77.2</td>
<td>65.3</td>
<td>45.6</td>
<td>885</td>
</tr>
<tr>
<td>share-1-type</td>
<td>79.5</td>
<td>81.5</td>
<td>89.1</td>
<td>94.7</td>
<td>85.4</td>
<td>80.3</td>
<td>1.25K</td>
</tr>
<tr>
<td>strong affordance</td>
<td>65.1</td>
<td>73.3</td>
<td>73.1</td>
<td>85.5</td>
<td>71.2</td>
<td>72.9</td>
<td>14.4K</td>
</tr>
<tr>
<td>unpopular</td>
<td>30.3</td>
<td>37.8</td>
<td>49.5</td>
<td>79.2</td>
<td>33.7</td>
<td>650</td>
<td>650</td>
</tr>
<tr>
<td></td>
<td>44.4</td>
<td>40.4</td>
<td>65.1</td>
<td>77.2</td>
<td>44.3</td>
<td>45.3</td>
<td>30K</td>
</tr>
<tr>
<td>kg-relation</td>
<td>55.3</td>
<td>53.4</td>
<td>77.8</td>
<td>86.9</td>
<td>50.3</td>
<td>52.5</td>
<td>6.22K</td>
</tr>
<tr>
<td>one-of-the-two</td>
<td>43.0</td>
<td>39.2</td>
<td>66.5</td>
<td>80.6</td>
<td>46.5</td>
<td>39.6</td>
<td>2.13K</td>
</tr>
<tr>
<td>share-1-type</td>
<td>45.4</td>
<td>44.8</td>
<td>77.4</td>
<td>88.2</td>
<td>57.9</td>
<td>83.5</td>
<td>2.78K</td>
</tr>
<tr>
<td>strong affordance</td>
<td>45.6</td>
<td>41.1</td>
<td>66.2</td>
<td>78.5</td>
<td>44.3</td>
<td>46.2</td>
<td>24.4K</td>
</tr>
<tr>
<td>unpopular</td>
<td>32.7</td>
<td>39.5</td>
<td>45.4</td>
<td>61.7</td>
<td>33.6</td>
<td>3.51K</td>
<td></td>
</tr>
<tr>
<td></td>
<td>44.0</td>
<td>39.3</td>
<td>55.1</td>
<td>65.2</td>
<td>34.4</td>
<td>49.8</td>
<td>961</td>
</tr>
<tr>
<td>kg-relation</td>
<td>45.5</td>
<td>31.2</td>
<td>79.4</td>
<td>80.4</td>
<td>29.4</td>
<td>29.9</td>
<td>279</td>
</tr>
<tr>
<td>one-of-the-two</td>
<td>31.5</td>
<td>34.9</td>
<td>65.6</td>
<td>74.9</td>
<td>30.0</td>
<td>27.2</td>
<td>461</td>
</tr>
<tr>
<td>share-1-type</td>
<td>34.9</td>
<td>37.4</td>
<td>62.9</td>
<td>80.7</td>
<td>27.2</td>
<td>3.14K</td>
<td></td>
</tr>
<tr>
<td>strong affordance</td>
<td>34.9</td>
<td>37.4</td>
<td>37.4</td>
<td>44.1</td>
<td>27.2</td>
<td>449</td>
<td></td>
</tr>
<tr>
<td>unpopular</td>
<td>34.9</td>
<td>37.4</td>
<td>37.4</td>
<td>44.1</td>
<td>27.2</td>
<td>449</td>
<td></td>
</tr>
<tr>
<td></td>
<td>50.0</td>
<td>41.7</td>
<td>50.8</td>
<td>66.4</td>
<td>50.8</td>
<td>58.3</td>
<td>128</td>
</tr>
<tr>
<td>kg-relation</td>
<td>27.7</td>
<td>28.6</td>
<td>66.7</td>
<td>75.0</td>
<td>41.7</td>
<td>38.6</td>
<td>36</td>
</tr>
<tr>
<td>one-of-the-two</td>
<td>50.0</td>
<td>41.7</td>
<td>58.3</td>
<td>75.0</td>
<td>35.7</td>
<td>28.6</td>
<td>12</td>
</tr>
<tr>
<td>share-1-type</td>
<td>28.6</td>
<td>28.6</td>
<td>64.3</td>
<td>78.6</td>
<td>35.7</td>
<td>28.6</td>
<td>14</td>
</tr>
<tr>
<td>strong affordance</td>
<td>28.6</td>
<td>28.6</td>
<td>55.8</td>
<td>69.2</td>
<td>35.7</td>
<td>28.6</td>
<td>104</td>
</tr>
<tr>
<td>unpopular</td>
<td>30.9</td>
<td>40.0</td>
<td>40.0</td>
<td>40.0</td>
<td>35.7</td>
<td>28.6</td>
<td>10</td>
</tr>
</tbody>
</table>

**Data is head biased!**

**Most systems are poor on rare entities!**
1. **Analyze**

Results on a subset of Wikipedia

<table>
<thead>
<tr>
<th></th>
<th>Amazon</th>
<th>Google</th>
<th>Microsoft</th>
<th>Bootleg</th>
<th>Rel</th>
<th>Pop</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>everything</td>
<td>49.9</td>
<td>49.2</td>
<td>66.6</td>
<td>78.8</td>
<td>51.2</td>
<td>32.6</td>
<td>49.5K</td>
</tr>
<tr>
<td>popular</td>
<td>66.7</td>
<td>82.7</td>
<td>71.7</td>
<td>83.9</td>
<td>80.4</td>
<td>74.5</td>
<td>8.32K</td>
</tr>
<tr>
<td>everything</td>
<td>64.9</td>
<td>73.3</td>
<td>73.2</td>
<td>85.1</td>
<td>71.6</td>
<td>72.8</td>
<td>15.5K</td>
</tr>
<tr>
<td>kg-relation</td>
<td>61.6</td>
<td>80.3</td>
<td>69.9</td>
<td>83.8</td>
<td>77.3</td>
<td>74.7</td>
<td>5.07K</td>
</tr>
<tr>
<td>one-of-the-two</td>
<td>48.7</td>
<td>65.3</td>
<td>65.3</td>
<td>77.2</td>
<td>65.3</td>
<td>45.8</td>
<td>885</td>
</tr>
<tr>
<td>share-1-type</td>
<td>79.5</td>
<td>81.5</td>
<td>89.4</td>
<td>94.7</td>
<td>85.4</td>
<td>80.1</td>
<td>1.25K</td>
</tr>
<tr>
<td>strong affordance</td>
<td>65.1</td>
<td>73.3</td>
<td>73.1</td>
<td>85.5</td>
<td>71.2</td>
<td>72.9</td>
<td>14.4K</td>
</tr>
<tr>
<td>unpopular</td>
<td>30.3</td>
<td>37.8</td>
<td>49.5</td>
<td>79.2</td>
<td>23.7</td>
<td>650</td>
<td></td>
</tr>
<tr>
<td>everything</td>
<td>44.4</td>
<td>40.4</td>
<td>65.1</td>
<td>77.2</td>
<td>44.3</td>
<td>45.3</td>
<td>30K</td>
</tr>
<tr>
<td>kg-relation</td>
<td>55.3</td>
<td>53.4</td>
<td>77.5</td>
<td>86.9</td>
<td>50.3</td>
<td>52.6</td>
<td>6.22K</td>
</tr>
<tr>
<td>one-of-the-two</td>
<td>43.0</td>
<td>39.2</td>
<td>66.5</td>
<td>80.6</td>
<td>46.5</td>
<td>39.6</td>
<td>2.13K</td>
</tr>
<tr>
<td>share-1-type</td>
<td>45.4</td>
<td>44.8</td>
<td>77.4</td>
<td>88.2</td>
<td>57.9</td>
<td>53.5</td>
<td>2.78K</td>
</tr>
<tr>
<td>strong affordance</td>
<td>45.6</td>
<td>41.1</td>
<td>66.2</td>
<td>78.5</td>
<td>44.3</td>
<td>46.2</td>
<td>24.4K</td>
</tr>
<tr>
<td>unpopular</td>
<td>39.2</td>
<td>45.4</td>
<td>45.4</td>
<td>61.7</td>
<td>26.3</td>
<td>3.51K</td>
<td></td>
</tr>
<tr>
<td>everything</td>
<td>33.8</td>
<td>30.5</td>
<td>55.1</td>
<td>65.2</td>
<td>27.3</td>
<td>40.4K</td>
<td></td>
</tr>
<tr>
<td>kg-relation</td>
<td>44.0</td>
<td>31.2</td>
<td>79.4</td>
<td>80.4</td>
<td>29.4</td>
<td>961</td>
<td></td>
</tr>
<tr>
<td>one-of-the-two</td>
<td>45.5</td>
<td>31.5</td>
<td>65.6</td>
<td>74.9</td>
<td>49.8</td>
<td>279</td>
<td></td>
</tr>
<tr>
<td>share-1-type</td>
<td>34.9</td>
<td>34.4</td>
<td>62.9</td>
<td>80.7</td>
<td>36.0</td>
<td>461</td>
<td>3.14K</td>
</tr>
<tr>
<td>strong affordance</td>
<td>34.9</td>
<td>37.4</td>
<td>57.8</td>
<td>67.9</td>
<td>27.9</td>
<td>449</td>
<td></td>
</tr>
<tr>
<td>unpopular</td>
<td>29.9</td>
<td>37.4</td>
<td>40.0</td>
<td>44.1</td>
<td>27.4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Recall**

Bootleg degrades gracefully compared to other systems
Google strongly prefers the popular alternative when disambiguating.

Results on a subset of Wikipedia
And resembles the popularity heuristic’s errors.

Results on a subset of Wikipedia

<table>
<thead>
<tr>
<th></th>
<th>Amazon</th>
<th>Google</th>
<th>Microsoft</th>
<th>Bootleg</th>
<th>Rel</th>
<th>Pop</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>everything</td>
<td>49.9</td>
<td>49.2</td>
<td>66.6</td>
<td>78.7</td>
<td>51.2</td>
<td>52.6</td>
<td>49.5K</td>
</tr>
<tr>
<td>popular</td>
<td>66.7</td>
<td>82.7</td>
<td>71.7</td>
<td>83.9</td>
<td>80.4</td>
<td>74.5</td>
<td>8.32K</td>
</tr>
<tr>
<td>everything</td>
<td>64.9</td>
<td>73.3</td>
<td>73.2</td>
<td>85.1</td>
<td>71.6</td>
<td>72.8</td>
<td>15.5K</td>
</tr>
<tr>
<td>kg-relation</td>
<td>61.6</td>
<td>80.3</td>
<td>69.9</td>
<td>83.8</td>
<td>77.3</td>
<td>74.7</td>
<td>5.07K</td>
</tr>
<tr>
<td>one-of-the-two</td>
<td>48.7</td>
<td>65.3</td>
<td>65.3</td>
<td>77.2</td>
<td>65.3</td>
<td>45.8</td>
<td>885</td>
</tr>
<tr>
<td>share-1-type</td>
<td>79.5</td>
<td>81.5</td>
<td>89.4</td>
<td>94.7</td>
<td>85.4</td>
<td>80.1</td>
<td>1.25K</td>
</tr>
<tr>
<td>strong affordance</td>
<td>65.1</td>
<td>73.3</td>
<td>73.1</td>
<td>85.5</td>
<td>71.2</td>
<td>72.9</td>
<td>14.4K</td>
</tr>
<tr>
<td>unpopular</td>
<td>30.3</td>
<td>37.8</td>
<td>49.5</td>
<td>79.2</td>
<td>23.7</td>
<td>650</td>
<td>650</td>
</tr>
<tr>
<td>everything</td>
<td>44.4</td>
<td>40.4</td>
<td>65.1</td>
<td>77.2</td>
<td>44.3</td>
<td>45.5</td>
<td>30K</td>
</tr>
<tr>
<td>kg-relation</td>
<td>55.3</td>
<td>53.4</td>
<td>77.5</td>
<td>86.9</td>
<td>50.3</td>
<td>52.5</td>
<td>6.22K</td>
</tr>
<tr>
<td>one-of-the-two</td>
<td>43.0</td>
<td>39.2</td>
<td>66.5</td>
<td>80.6</td>
<td>46.5</td>
<td>39.6</td>
<td>2.13K</td>
</tr>
<tr>
<td>share-1-type</td>
<td>45.4</td>
<td>44.8</td>
<td>77.4</td>
<td>88.2</td>
<td>57.9</td>
<td>53.5</td>
<td>2.78K</td>
</tr>
<tr>
<td>strong affordance</td>
<td>45.6</td>
<td>41.1</td>
<td>66.2</td>
<td>78.5</td>
<td>44.3</td>
<td>46.2</td>
<td>24.4K</td>
</tr>
<tr>
<td>unpopular</td>
<td>11.1</td>
<td>5</td>
<td>45.4</td>
<td>61.7</td>
<td>2.8</td>
<td>3.51K</td>
<td></td>
</tr>
<tr>
<td>everything</td>
<td>33.8</td>
<td>10.5</td>
<td>55.1</td>
<td>65.2</td>
<td>27.3</td>
<td>34.4</td>
<td>4.04K</td>
</tr>
<tr>
<td>kg-relation</td>
<td>44.0</td>
<td>31.2</td>
<td>79.4</td>
<td>80.4</td>
<td>29.4</td>
<td>30.0</td>
<td>961</td>
</tr>
<tr>
<td>one-of-the-two</td>
<td>45.5</td>
<td>12.8</td>
<td>65.6</td>
<td>74.9</td>
<td>34.4</td>
<td>49.8</td>
<td>279</td>
</tr>
<tr>
<td>share-1-type</td>
<td>31.5</td>
<td>34.5</td>
<td>62.9</td>
<td>80.7</td>
<td>36.0</td>
<td>29.9</td>
<td>461</td>
</tr>
<tr>
<td>strong affordance</td>
<td>34.9</td>
<td>12.2</td>
<td>57.8</td>
<td>67.9</td>
<td>27.3</td>
<td>27.3</td>
<td>3.14K</td>
</tr>
<tr>
<td>unpopular</td>
<td>20.7</td>
<td>9.4</td>
<td>37.4</td>
<td>44.1</td>
<td>10.2</td>
<td>449</td>
<td>449</td>
</tr>
</tbody>
</table>

Recall

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>10</th>
<th>100</th>
<th>1000</th>
<th>0</th>
<th>10</th>
<th>100</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon</td>
<td>50.8</td>
<td>66.7</td>
<td>75.0</td>
<td>75.0</td>
<td>41.7</td>
<td>58.3</td>
<td>58.3</td>
<td>58.3</td>
</tr>
<tr>
<td>Google</td>
<td>50.8</td>
<td>66.7</td>
<td>75.0</td>
<td>75.0</td>
<td>41.7</td>
<td>58.3</td>
<td>58.3</td>
<td>58.3</td>
</tr>
<tr>
<td>Microsoft</td>
<td>50.8</td>
<td>66.7</td>
<td>75.0</td>
<td>75.0</td>
<td>41.7</td>
<td>58.3</td>
<td>58.3</td>
<td>58.3</td>
</tr>
<tr>
<td>Bootleg</td>
<td>50.8</td>
<td>66.7</td>
<td>75.0</td>
<td>75.0</td>
<td>41.7</td>
<td>58.3</td>
<td>58.3</td>
<td>58.3</td>
</tr>
<tr>
<td>Rel</td>
<td>50.8</td>
<td>66.7</td>
<td>75.0</td>
<td>75.0</td>
<td>41.7</td>
<td>58.3</td>
<td>58.3</td>
<td>58.3</td>
</tr>
<tr>
<td>Pop</td>
<td>50.8</td>
<td>66.7</td>
<td>75.0</td>
<td>75.0</td>
<td>41.7</td>
<td>58.3</td>
<td>58.3</td>
<td>58.3</td>
</tr>
<tr>
<td>Size</td>
<td>128</td>
<td>36</td>
<td>12</td>
<td>14</td>
<td>104</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>
1. Analyze

Upstream NEL will cause issues in downstream systems: e.g. sports QA application

Sports QA: prefer if the model predicted the national sports team instead of the country!
1. Analyze
2. Repurpose

Wikipedia data: noisy, incompletely labeled entities

Our data engineering approach: relabel the training set with a known reasoning pattern to guide model behavior
25% absolute accuracy improvement in sports-related errors

Wikipedia examples with mentions of countries and sports teams

<table>
<thead>
<tr>
<th>Subpop.</th>
<th>Gold Label</th>
<th>Pred. Label</th>
<th>Size (Off-The-Shelf $\rightarrow$ Patched)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Team</td>
<td>Country</td>
<td>Team</td>
<td>151 $\rightarrow$ 106</td>
</tr>
<tr>
<td></td>
<td>Team</td>
<td></td>
<td>3393 $\rightarrow$ 3447</td>
</tr>
</tbody>
</table>

Karan Goel | @krandiash
Meerkat: DataPanels for Interactive ML
github.com/robustness-gym/meerkat

Demo: https://tinyurl.com/meerkat-colab
Meerkat: DataPanels for Interactive ML

github.com/robustness-gym/meerkat

Measurement → Monitoring → Maintenance

Demo: https://tinyurl.com/meerkat-colab
Meerkat: DataPanels for Interactive ML

github.com/robustness-gym/meerkat

Measurement → Monitoring → Maintenance

Meerkat DataPanel

Demo: https://tinyurl.com/meerkat-colab
Meerkat: DataPanels for Interactive ML

github.com/robustness-gym/meerkat

dp = mk.DataPanel.from_csv("data.csv")
Meerkat: DataPanels for Interactive ML

github.com/robustness-gym/meerkat

dp = mk.DataPanel.from_csv("data.csv")
dp = mk.DataPanel.from_csv("data.csv")

dp["img"] = mk.ImageColumn(dp["filepath"])

Meerkat: DataPanels for Interactive ML

github.com/robustness-gym/meerkat
dp = mk.DataPanel.from_csv("data.csv")

dp["img"] = mk.ImageColumn(dp["filepath")

dp = dp.update(
    function=predict,
    is_batched_fn=True,
    batch_size=32,
    num_workers=4
)
dp = mk.DataPanel.from_csv("data.csv")

dp["img"] = mk.ImageColumn(dp["filepath"])

dp = dp.update(
    function=predict,
    is_batched_fn=True,
    batch_size=32,
    num_workers=4
)

error_dp = dp[dp["pred"] != dp["label"]]
error_dp["img"][0]
dp = mk.DataPanel.from_csv("data.csv")

dp["img"] = mk.ImageColumn(dp["filepath"])

dp = dp.update(
    function=predict,
    is_batched_fn=True,
    batch_size=32,
    num_workers=4
)

error_dp = dp[dp["pred"] != dp["label"]]
error_dp["img"][0]
**Goal:** Design data-centric systems that are easy to modify e.g. to perform bug fixes, hit performance targets and meet ethical standards.
*Karan Goel*, Nazneen Rajani*, Jesse Vig, Zachary Taschdjian, Mohit Bansal, Christopher Ré. NAACL Demo 2021.

Mandoline: Model Evaluation under Distribution Shift

Model Patching: Closing the Subgroup Performance Gap with Data Augmentation.
*Karan Goel*, Albert Gu, Yixuan Li, Christopher Ré. ICLR 2021.

*Karan Goel*, Laurel Orr, Nazneen Rajani, Jesse Vig, Christopher Ré. NAACL Industry 2021.