Unfit for intended use:
Detecting emergent bias in machine learning systems

Kiri L. Wagstaff

Special Advisor on Artificial Intelligence

Oregon State University Libraries

Linköping University, Sweden – September 11, 2025

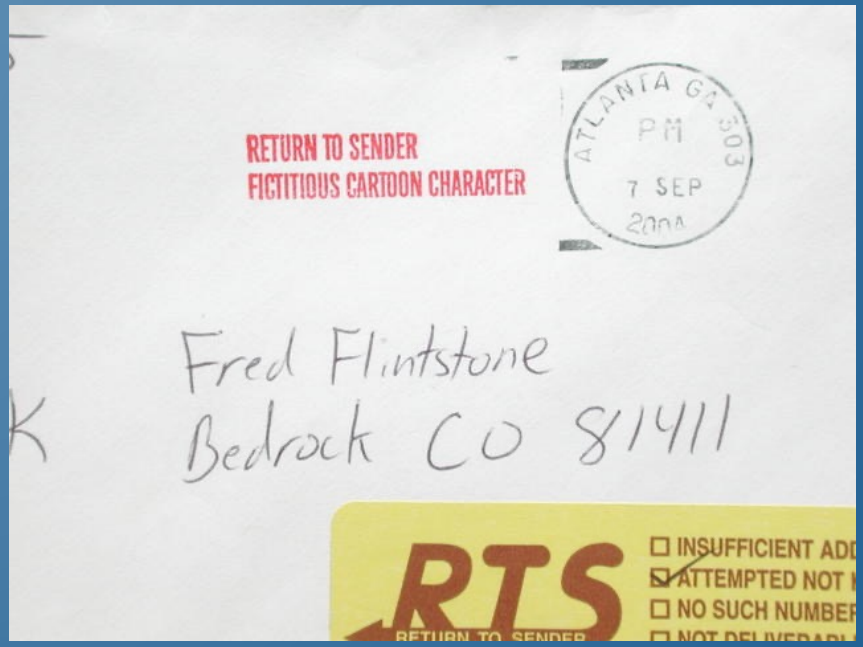
Background by Pixabay user harishs

Topics for today

1. What is emergent bias for machine learning systems?
2. When does emergent bias happen?
3. How can we detect it before using/deploying the ML model?

Goal: Recognize handwritten numbers

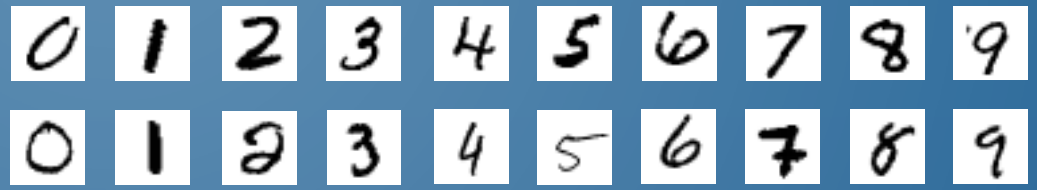
- Examples
  - U.S. ZIP Codes on postal mail
  - Phone numbers
  - Handwritten check amounts
  - Time of day on a handwritten note
  - Dosage on a medical prescription
  - (Etc.)



Multi-Digit Writer (MDW) U.S. ZIP Codes

- Synthetic handwritten multi-digit numbers written by a single writer
  - Assembled from MNIST digits (LeCun et al., 1998)


Yann LeCun




Writer 2445


Writer 489


Writer 480


Writer 1423


Writer 3856

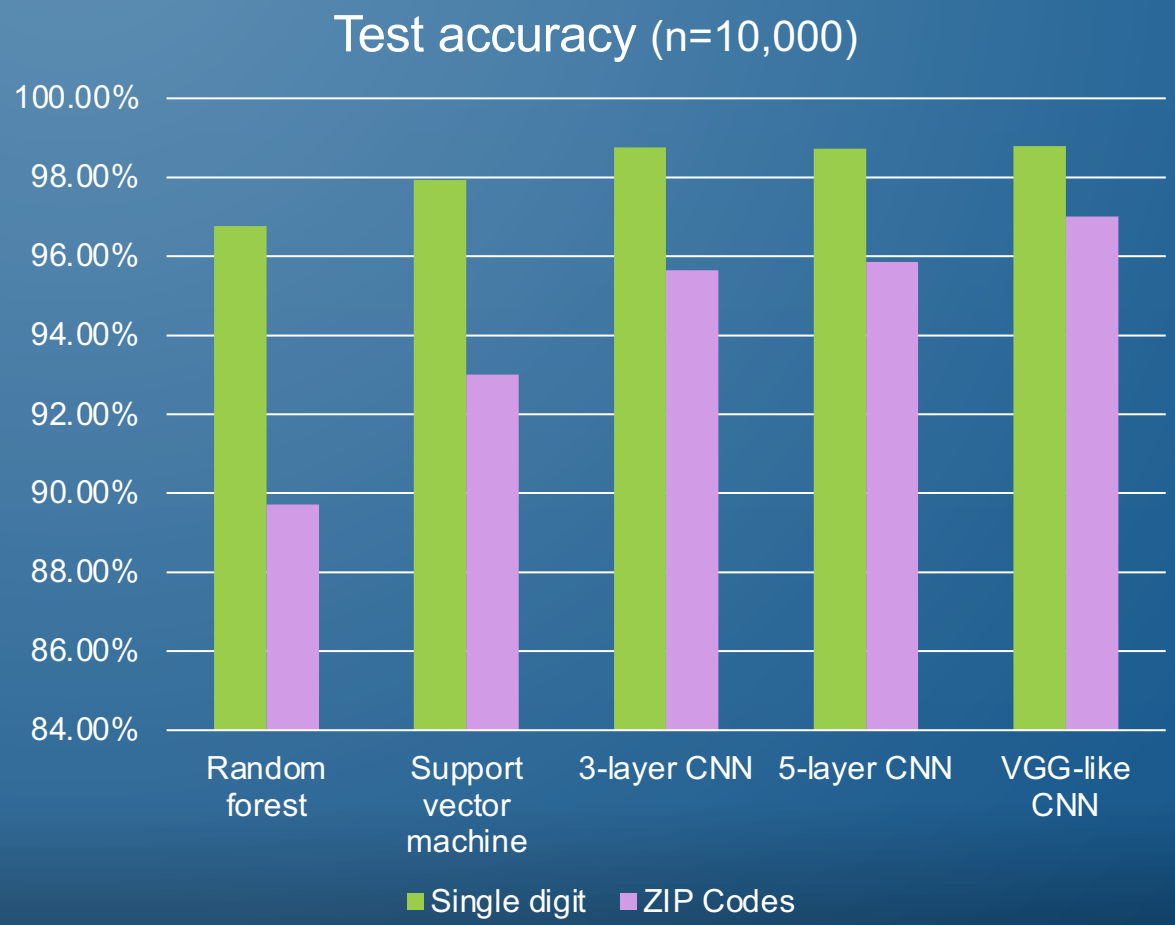

Writer 2389

- Limitations: artificially regular spacing, doesn't capture sequential writing behaviors, each digit is only 28x28 pixels
  - Prefer to collect new (sequential) handwriting samples if resources permit

Which ZIP Code model do you prefer?

- Recognizing 5-digit ZIP Codes is harder than single digits

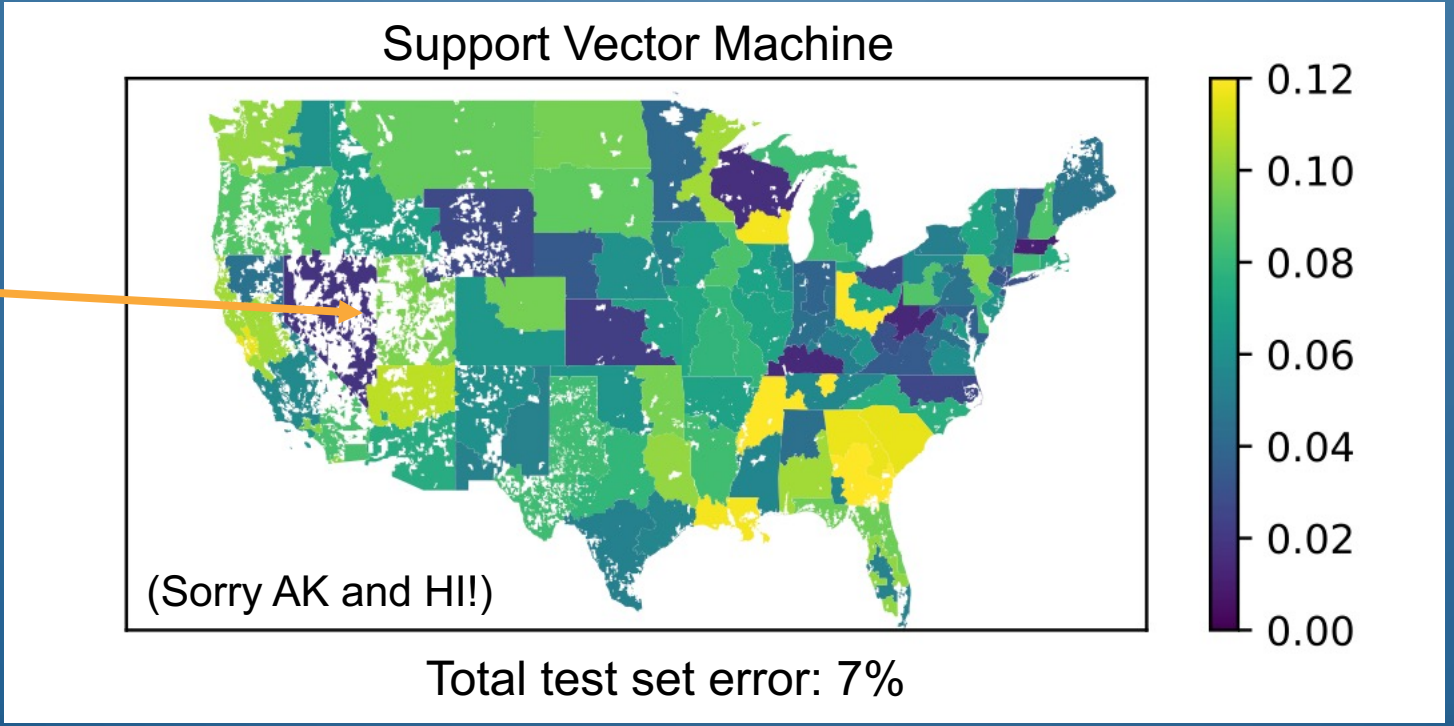
Classifier (n=60,000 training items)	Single digit: MNIST test accuracy (n=10,000)	5 digits: MDW ZIP Codes test accuracy (n=10,000)
Random forest	96.77%	89.71%
Support vector machine	97.93%	93.00%
3-layer CNN	98.75%	95.64%
5-layer CNN	98.73%	95.85%
VGG-like CNN	**98.79%**	**97.01%**



Error is not equally distributed (geographical bias)

- Error rates (lower is better) per U.S. ZIP Code sector
  - Sector: all ZIP Codes with the same first two digits (e.g., 00XXX, 01XXX…)

Some places do not belong to a ZIP Code!



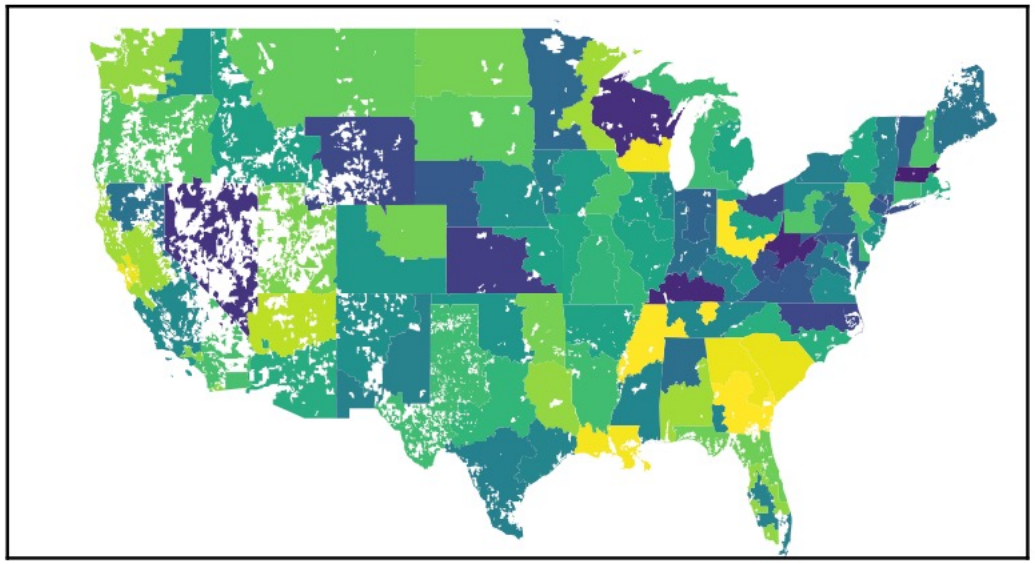
Emergent bias: performance differs by location

(Bias = differential performance for sub-groups)

Which ZIP Code model do you prefer?

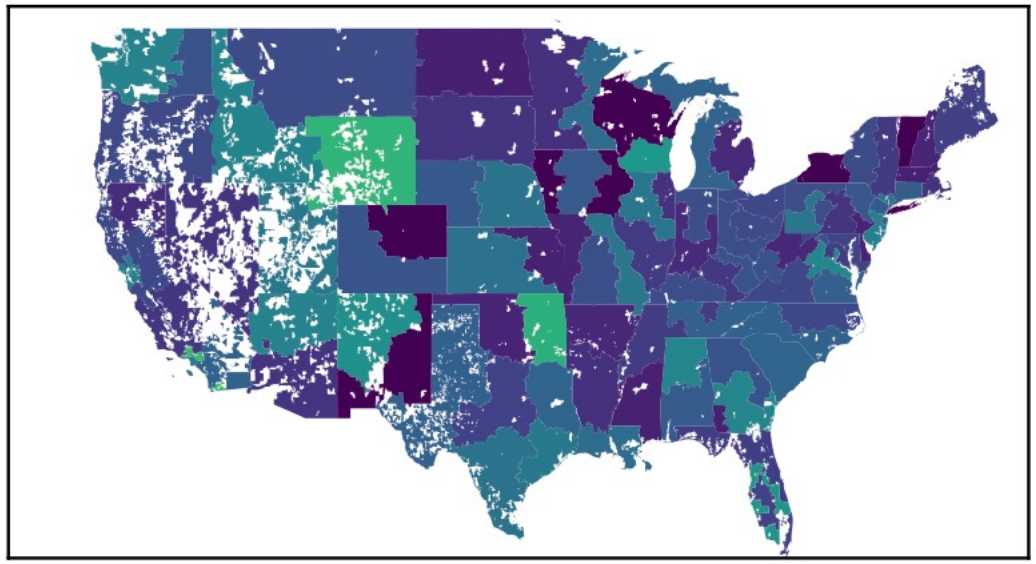
- VGG-like CNN has lower total error
  - But its emergent bias penalizes different areas than the SVM

Support Vector Machine



Total test set error: 7%

VGG-like Convolutional Neural Network

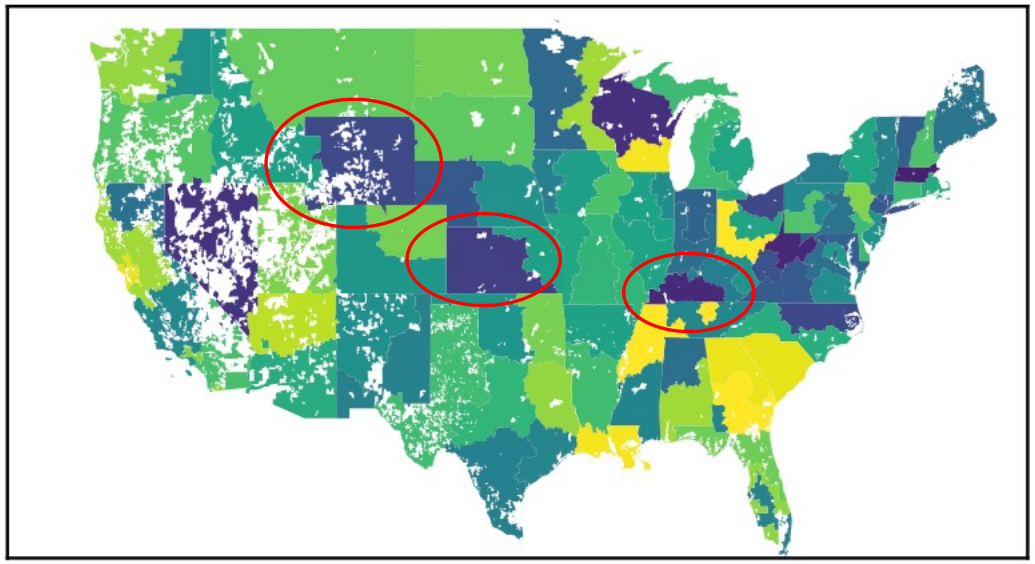


Total test set error: 3%

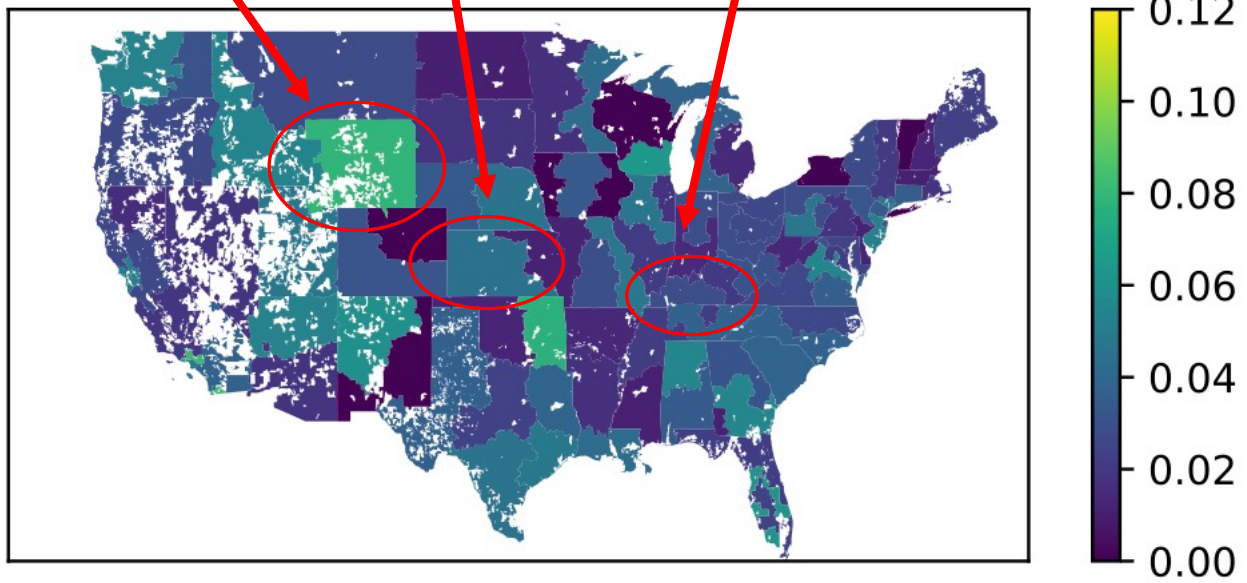
Different emergent bias in each model

What if you live in Wyoming, Kansas, or Kentucky?

Support Vector Machine

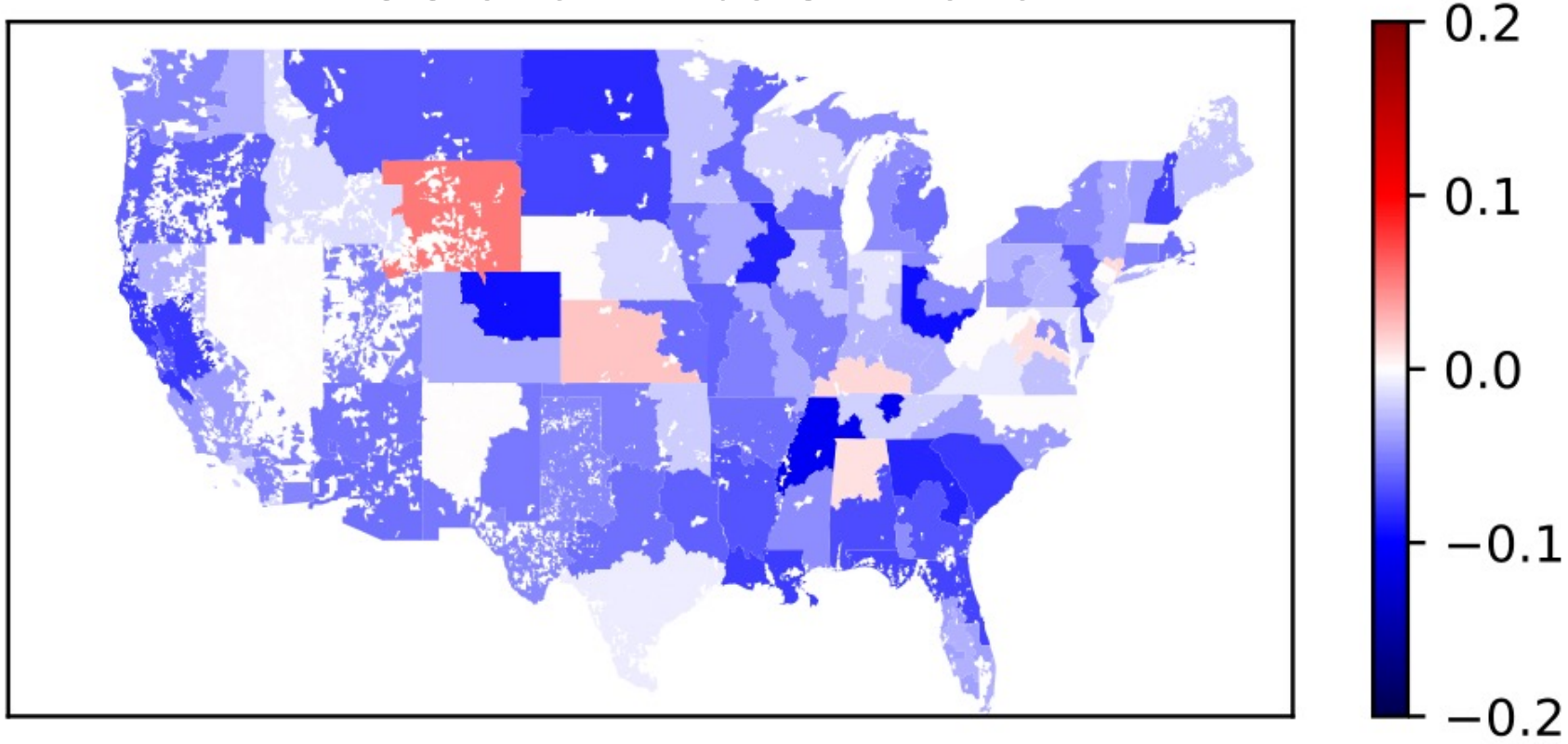


VGG-like Convolutional Neural Network



Different emergent bias in each model

Comparative geographical bias VGG and SVM models



Should we use a different model in each location?

We need to evaluate emergent bias carefully to inform deployment decisions

What is emergent bias?

- Emergent bias: Bias that manifests (only) when applying a system in contexts not originally envisioned
  - Geographical bias is not in the training data
  - Geographical bias is not even defined for isolated digit classifiers
- "While it is almost always possible to identify preexisting bias and technical bias in a system design at the time of creation or implementation, emergent bias arises only in a context of use."
  (Friedman & Nissenbaum, 1996)

Is the model unfit for intended use?

- What went wrong?

Digits

Generic model trained to recognize digits

Predictions

2 3 8

ZIP Code digits

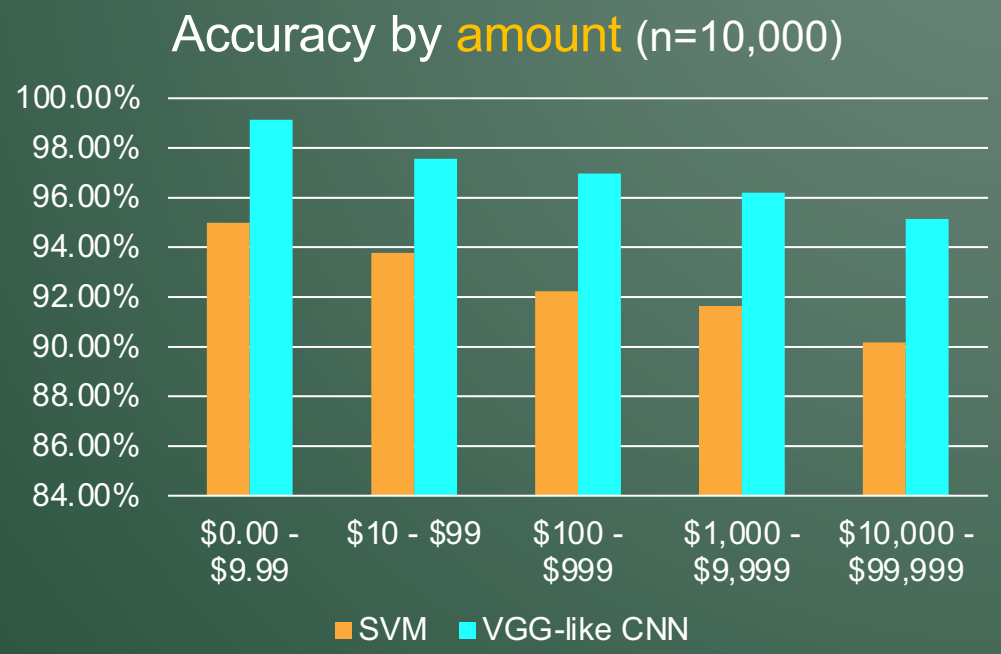
ZIP Code predictions

90210

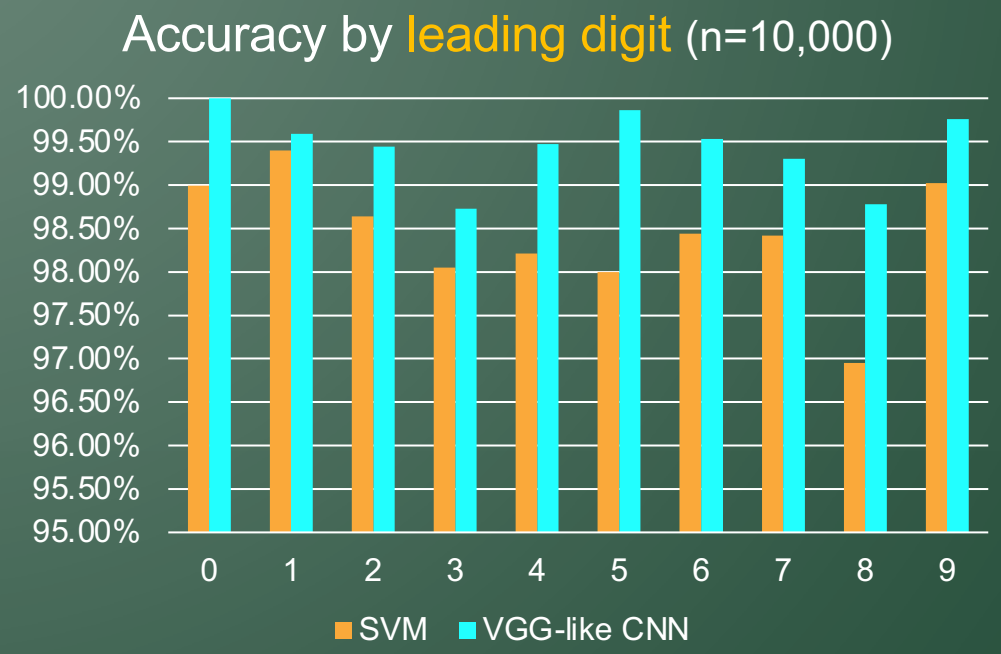
62989

More examples

- MDW handwritten check amounts: $0.00 to $99,999
  - The position of each digit matters



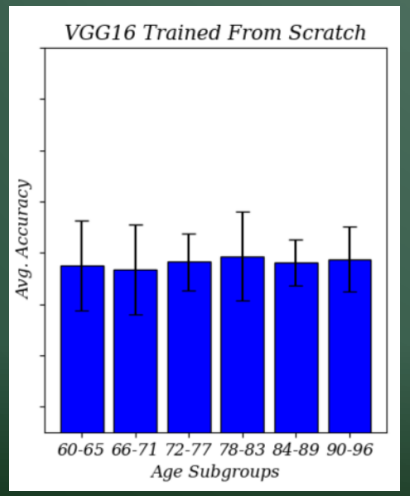
Longer numbers are harder for both models
(not surprising, but not what you want)

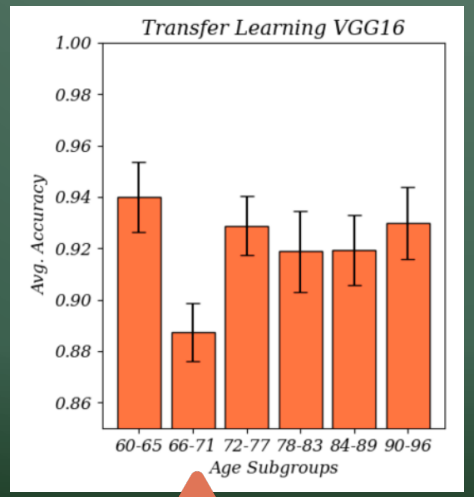


SVM has a weakness for amounts starting with $8
(surprising!)

Other causes of emergent bias

- Domain shift: Changes over time and/or space
- Transfer learning
  - Alzheimer's disease diagnosis from brain scan images (Salmani & Lewis, 2024)
  - VGG16 model trained on ImageNet and fine-tuned for brain scans exhibits 3x more bias (gender) and 5x more bias (age) compared to VGG16 model trained from scratch





Avg. performance for fine-tuned VGG16 is higher (92.3% vs. 91.7%), but it exhibits emergent bias

Detecting emergent bias before deployment

- The generic MNIST test set cannot reveal this bias to you
  - "emergent bias arises only in a context of use." (Friedman & Nissenbaum, 1996)
- Do not rely on generic metrics (accuracy, AUC, F1, etc.) to do model selection
- Ask: What performance metric do you actually care about?
  - Error rate across the entire sample
  - Error rate per region
  - Error rate for urban vs. rural
  - Error rate per region, weighted by population
  - Error rate per region, weighted by postal activity

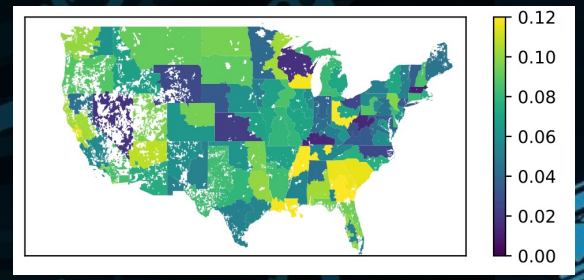
What are the sub-groups of interest?

What can you do to reduce emergent bias?

- Collect more data for sub-groups with lower performance and re-train
- Selectively deploy best classifier for each sub-group
  - But this means you are making decisions based on fewer examples per sub-group, so proceed with caution, especially for minority sub-groups

Emergent bias in machine learning systems

- Emergent bias: appears in a specific context of use
  - It is not measurable (or defined) in the generic context
  - Not a good surprise to get after deployment!
- Do not rely on abstract performance metrics for model selection
  - Choose and inspect domain-specific performance metrics
  - Compare performance between sub-groups of interest
- Consider selective use of models that work well for each sub-group



Thank you!

Contact: wkiri@wkiri.com