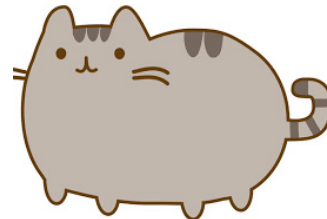# Covariate shift

Training set

At test time

Why would anyone be so stupid?

Covariate Shift

- **Web search**
  - Training - page relevance data for the US market
  - Testing - recommend pages for Canada (UK, Australia)
- **Speech recognition**
  - Training - West coast accent
  - Testing - Southern drawl, Texan, non-native speaker
- **Language**
  - Training - 'James, bring me a **soda**'
  - Testing - 'John, bring me a '**pop**' (or coke, etc.)

Covariate Shift

- **Medical**
  - Training - University students + old men with prostate cancer
  - Testing - Potentially sick old men
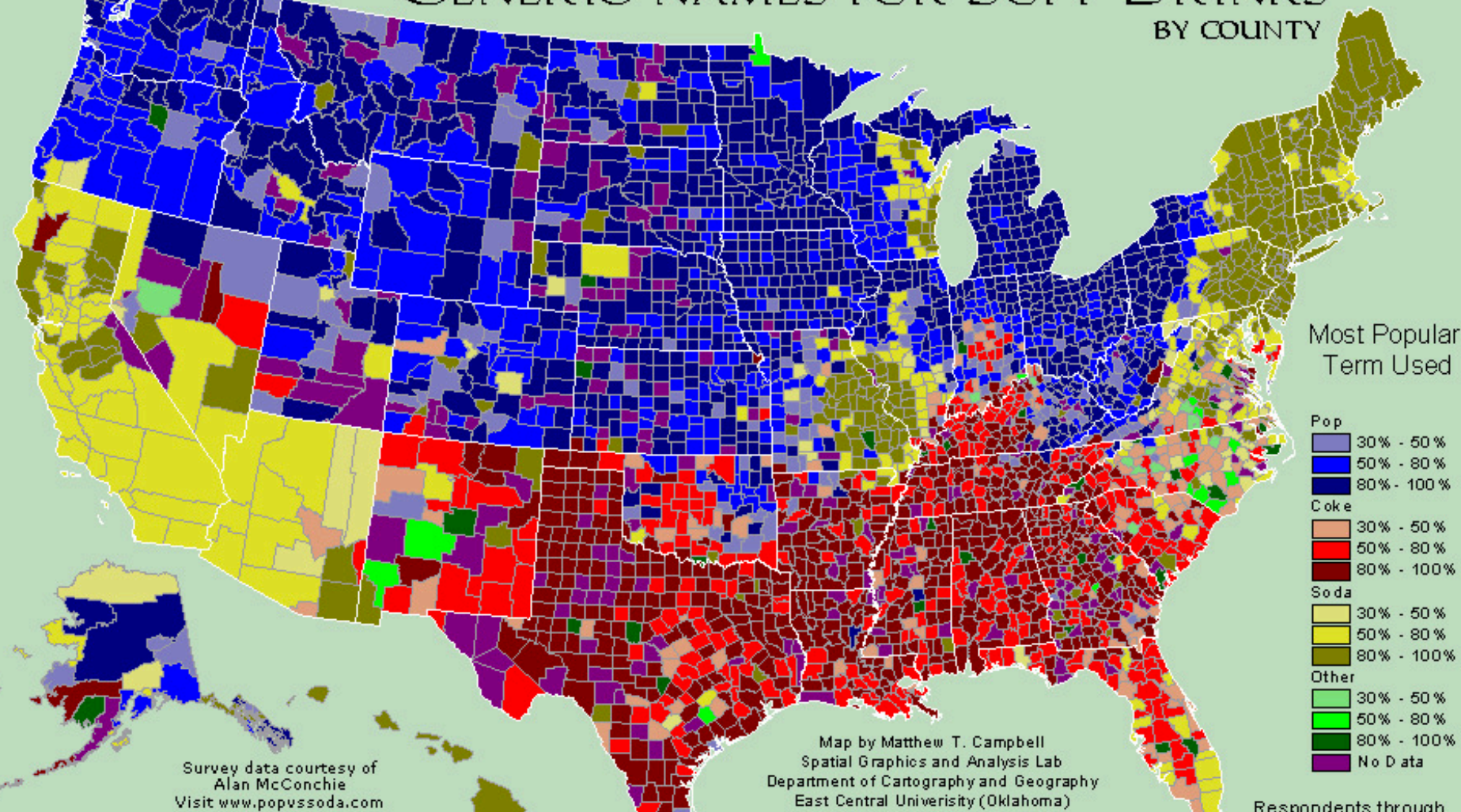- **Reinforcement Learning**
  - Training - Data gathered with current policy
  - Testing - Environment reacting to updated policy
- **Databases**
  - Training - DB tuned to 2017 usage pattern
  - Testing - DB deployed on AWS in 2018

Generic names for Soft Drinks
by county

Most Popular
Term Used

Pop
30% - 50%
50% - 80%
80% - 100%

Coke
30% - 50%
50% - 80%
80% - 100%

Soda
30% - 50%
50% - 80%
80% - 100%

Other
30% - 50%
50% - 80%
80% - 100%
No Data

Map by Matthew T. Campbell
Spatial Graphics and Analysis Lab
Department of Cartography and Geography
East Central Univeristy (Oklahoma)

Respondents through

Real Story from CES'19

- Startup with smart vending machine demo that identifies purchases via a camera

- Demo at CES failed
  - Different light temperature
  - Light reflection from table

- The fix
  - Collect new data
  - Buy tablecloth
  - Retrain all night

Source Source → Target

What is happening? $q(x, y) = q(x)p(y|x)$

- Training Risk

Training data

$$\underset{w}{\text{minimize}} \int dx\, p(x) \int dy\, p(y|x) l(f(x, w), y)$$

or rather $\underset{w}{\text{minimize}} \dfrac{1}{m} \sum_{I=1}^{m} l(f(x_i, w), y_i)$

- Test Risk is different

Test data

$$\int dx\, q(x) \int dy\, p(y|x) l(f(x, w), y)$$
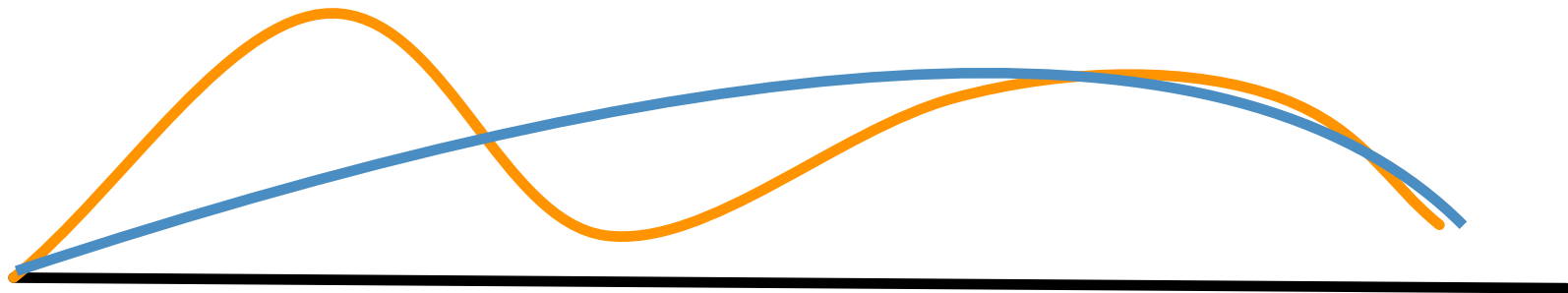
Fixing it (covariate shift correction)

- Density ratio

$$\int dx \, q(x) f(x) = \int dx \, p(x) \underbrace{\frac{q(x)}{p(x)}}_{\alpha(x)} f(x) = \int dx \, p(x) \alpha(x) f(x)$$

- Need to find density ratio, but we don't have p or q.

- Estimating densities is hard.
  E.g. what to do if density is very high (or very low)?

$$r(x, y) = \frac{1}{2} \left[p(x) \cdot \delta(y, 1) + q(x) \cdot \delta(y, 1) \right]$$

Density Ratio Estimation

- If $p(x)$ and $q(x)$ differ, we can train a classifier to distinguish between them.

- Classification $\quad r(x, y) = \frac{1}{2} \left[p(x) \cdot \delta(y, 1) + q(x) \cdot \delta(y, 1) \right]$

Fixing it (covariate shift correction)

$$r(x, y) = \tfrac{1}{2} \left[p(x) \cdot \delta(y, 1) + q(x) \cdot \delta(y, 1) \right]$$

- Conditional class probability

$$r(y = 1 | x) = \frac{p(x)}{p(x) + q(x)} \text{ and hence } \alpha = \frac{q(x)}{p(x)} = \frac{r(y = -1 | x)}{r(y = 1 | x)}$$

- Simple algorithm
  - Train classifier between training and test set
  - Reweight training data with $\dfrac{r(y = -1 | x)}{r(y = 1 | x)}$

- Can use Generalization Performance Estimate to decide whether we even have Covariate Shift.

Corrected Training Problem

- Original Problem

$$\underset{w}{\text{minimize}} \sum_{i=1}^{m} \frac{1}{m} l(y_i, f(x_i, w))$$

- Reweighted Problem

$$\underset{w}{\text{minimize}} \sum_{i=1}^{m} \alpha_i l(y_i, f(x_i, w)) \text{ where } \alpha_i = \frac{r(y = -1 \mid x_i)}{r(y = 1 \mid x_i)}$$

- Problems

  - Coefficients α_i are estimates. Adds variance & bias.

  - What if p and q are very different?

Problems with p and q

$$\sum_{i=1}^{m} \alpha_i l(y_i, f(x_i, w))$$

Problems with p and q

$$\sum_{i=1}^{m} \alpha_i l(y_i, f(x_i, w))$$

- Effective sample size

$$m* = \frac{\|\alpha\|_1^2}{\|\alpha\|_2^2}$$

- Unmodified weights

  - Less bias

  - High variance for small $m*$

- Clipped weights $\bar{\alpha}_i = \min(\alpha_i, C)$

  - Potentially some bias

  - Smaller variance
    (larger effective sample size)

Key Takeaways

Covariate Shift assumes

$$p(x, y) = p(x)p(y \mid x) \longrightarrow q(x, y) = q(x)p(y \mid x)$$

- Label dependency for training and test is the same
- Covariate distribution is different

Detection & Fix
- Train binary classifier between training and test set
- If nontrivial accuracy, use it to reweight training data via

$$\alpha(x) = r(y = 1 \mid x)/r(y = -1 \mid x)$$

- Fix effective sample size by clipping $\bar{\alpha}(x) = \min(\alpha(x), C)$