# Modelowanie świadomości politycznej

Mikołaj Płóciniczak

18 marca 2026

# Cel i temat pracy

Analiza zależności między świadomością polityczną a wybranymi czynnikami społecznymi z wykorzystaniem metod uczenia maszynowego

# Co pokażę w prezentacji

1. Dane o nastawieniu do zmian klimatu
2. Dane o nastawieniu do innych kluczowych spraw politycznych
3. Pojęcie ideal point i bridging questions
4. Item response model
5. Multilevel regression and poststratification
6. Przetwarzanie danych

# Dlaczego Stany Zjednoczone?

**Dostępność danych:**

- Duże, ogólnokrajowe badania ankietowe
- Setki tysięcy respondentów

**Zróżnicowanie:**

- Duże różnice regionalne i społeczne
- Silna polaryzacja polityczna

**Znaczenie:**

- USA jako „laboratorium” badań opinii publicznej
- Dobrze rozwinięta infrastruktura danych

# Jednostki geograficzne w USA

**Hrabstwa (counties):**

- Podstawowy poziom administracyjny poniżej stanów
- 3 143 hrabstwa i jednostki równoważne

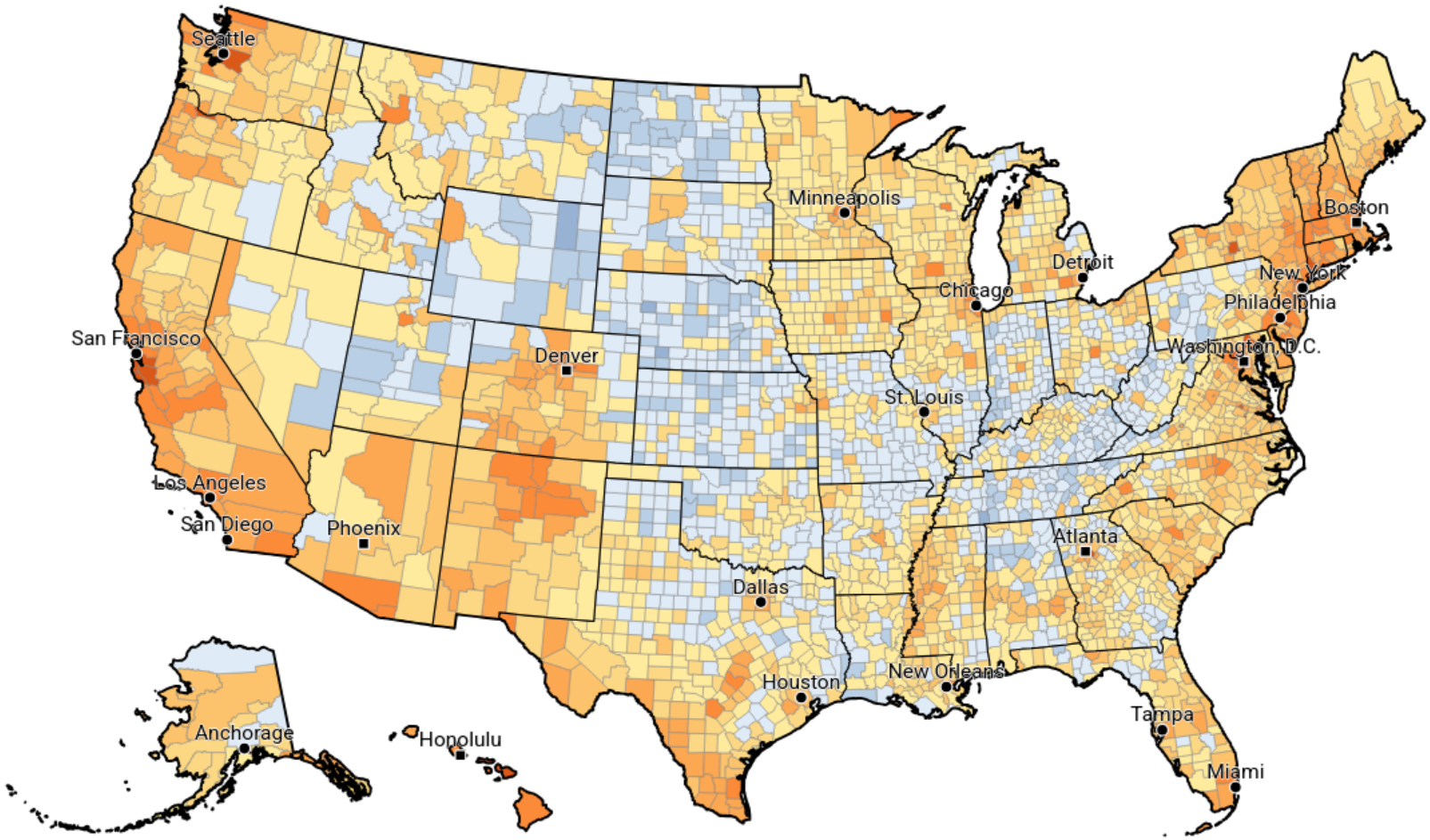
**Dlaczego counties?**

- Pozwalają analizować dane na poziomie lokalnym
- Wystarczająco duża liczba jednostek

**Ważne rozróżnienie:**

- Hrabstwa $\neq$ okręgi wyborcze
- Okręgi służą reprezentacji politycznej, hrabstwa administracji

# Yale CLimate Opinion Maps




Rysunek: *Czy jesteś zaniepokojona/zaniepokojony zmianami klimatu? (2024)*

# Yale Climate Opinion Maps (YCOM)

**Czym jest YCOM?**

- Projekt Uniwersytetu Yale
- Szacuje opinie Amerykanów o zmianach klimatu
- $n > 35,000$ respondentów w latach 2008-2024

**Co pokazuje:**

- Wyniki dla małych jednostek geograficznych:
  - hrabstwa (counties)
  - okręgi wyborcze

**Forma:**

- Interaktywne mapy (np. poziom poparcia w %)
- plik `.csv`

# YCOM: pytania ankietowe

**Charakter pytań:**

- Dotyczą konkretnych aspektów zmian klimatu
- $2 \cdot 44 = 88$ pytań/zmiennych

**Typy pytań:**

- Czy globalne ocieplenie zachodzi?
- Czy jest spowodowane przez człowieka?
- Czy jest groźne?
- Czy popierasz polityki klimatyczne?

**Odpowiedzi:**

- Binarne (zgadzam się/popieram vs. nie zgadzam się/nie popieram)
- Wyniki prezentowane jako odsetki w populacji

# Jak powstają estymaty w YCOM

**Problem:**

- Brak danych ankietowych na poziomie lokalnym

**Rozwiązanie:**

- Połączenie:
  - danych ankietowych
  - danych demograficznych (np. spis ludności)

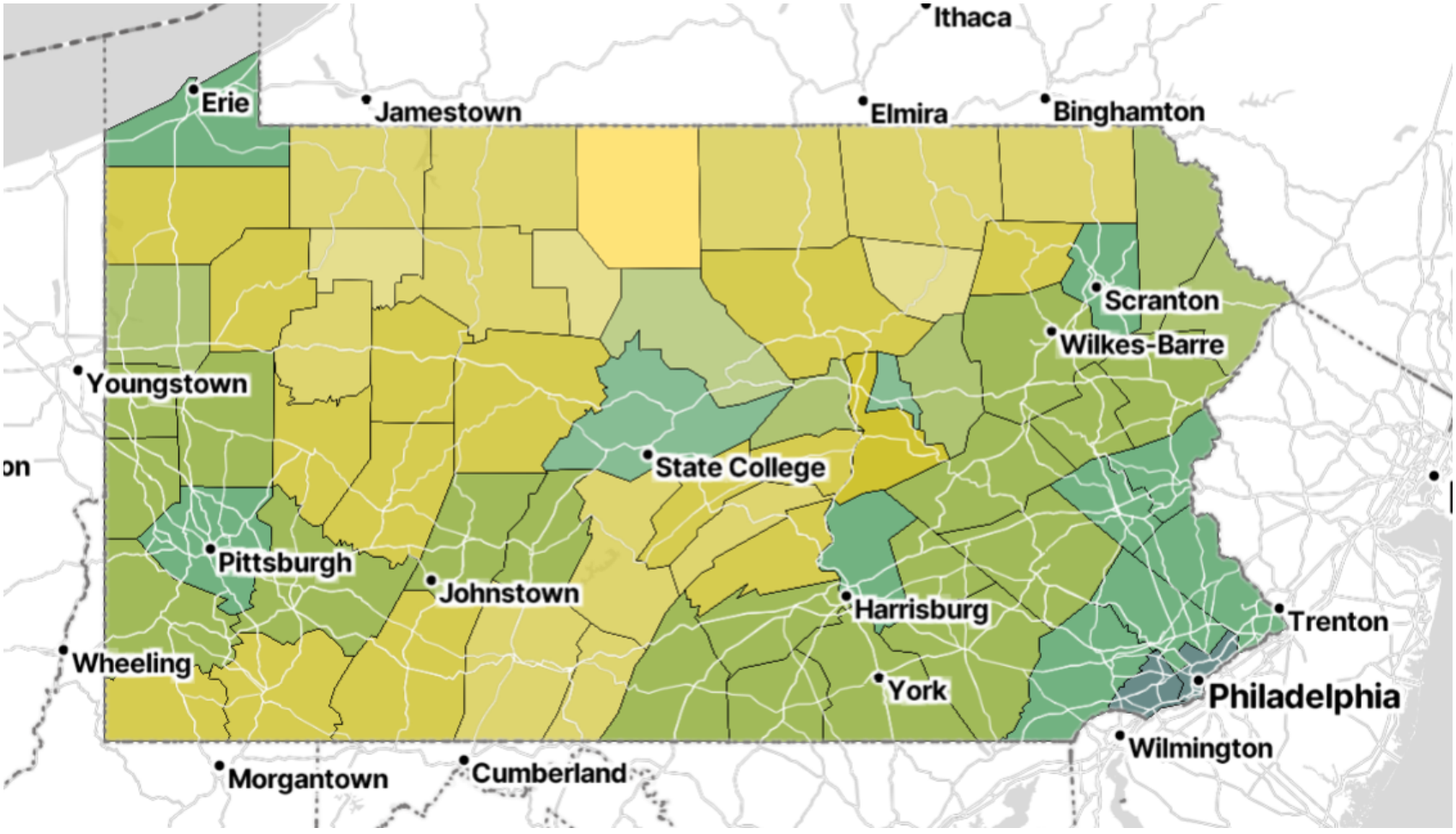
**Metoda:**

- Modelowanie statystyczne (*Multilevel Regression and Poststratification*)

**Efekt:**

- Estymaty opinii dla małych obszarów

# TrueViews




Rysunek: *Czy jesteś za czy przeciw zakazowi karabinów szturmowych* – w różnych hrabstwach stanu Pensylwania

# Źródła danych: duże badania ankietowe

**Główne źródła:**

- Cooperative Congressional Election Study (CCES)
- Annenberg National Election Survey (NAES)

**Charakterystyka:**

- Duże, ogólnokrajowe badania opinii publicznej w USA
- Przeprowadzane w wielu latach (2000–2011)

**Skala:**

- Ponad 275,000 respondentów
- Dziesiątki tysięcy osób w pojedynczym badaniu

# Pytania w ankietach

**Zakres tematyczny:**

- Podatki i wydatki publiczne
- Opieka zdrowotna
- Polityka społeczna
- Polityka zagraniczna

**Forma pytań:**

- Najczęściej pytania binarne (tak/nie)
- Czasem skale odpowiedzi $\rightarrow$ przekształcane do 0/1

**Przykład:**

- Czy rząd powinien zwiększyć wydatki na opiekę zdrowotną?

# Struktura danych

**Macierz danych:**

- Wiersze: respondenci
- Kolumny: pytania (policy items)

**Obserwacje:**

- $y_{ij} = 1 \rightarrow$ odpowiedź „tak”
- $y_{ij} = 0 \rightarrow$ odpowiedź „nie”

**Cechy:**

- Różne osoby odpowiadają na różne zestawy pytań
- Dane są częściowo niekompletne (missing data)

# Zbieranie danych i wyzwania

**Sposób zbierania:**

- Ankiety internetowe i/lub telefoniczne
- Próby reprezentatywne dla populacji USA

**Wyzwania:**

- Mała liczba respondentów na poziomie lokalnym
- Różne pytania w różnych badaniach
- Trudność porównywania wyników

**Konsekwencja:**

- Potrzeba metod takich jak IRT i MRP

# YCOM vs TrueViews

**YCOM:**

- Analiza pojedynczych pytań dot. klimatu
- Każda mapa = jedno zagadnienie

**TrueViews:**

- Łączy wiele pytań
- Szacuje jedną zmienną: **ideal point**

**Stąd:**

- YCOM $\rightarrow$ opinie na konkretne tematy "klimatyczne"
- TrueViews $\rightarrow$ ogólna ideologia polityczna

# Czym jest ideal point?

**Definicja formalna:**

- Każda jednostka $i$ ma preferencje $u_i(x)$ nad politykami $x \in \mathbb{R}^d$
- Ideal point:

$$x_i^* = \arg\max_x \; u_i(x)$$

**Założenie (single-peaked):**

- Użyteczność maleje wraz z odległością od $x_i^*$

$$u_i(x) = -\|x - x_i^*\|^2$$

**Wniosek:**

- $x_i^*$ jest zmienną latentną, estymowaną na podstawie wyborów

# Geneza koncepcji ideal point

**Korzenie w ekonomii i teorii wyboru:**

- Hotelling (1929): preferencje w przestrzeni
- Black (1948): single-peaked preferences

**Model polityczny:**

- Downs (1957): wyborcy mają idealny punkt ideologiczny

**Późniejsze rozwinięcia:**

- IRT i modele statystyczne (XX w.)

# Model przestrzenny preferencji

**Założenie:**

- Każda polityka ma pozycję $x_j \in \mathbb{R}^d$

**Model wyboru:**

- Jednostka $i$ wybiera opcję maksymalizującą użyteczność:

$$u_i(x_j) = -\|x_j - x_i^*\|^2$$

**Ujęcie probabilistyczne:**

- $$P(y_{ij} = 1) = f\left(u_i(x_j)\right)$$

**Wniosek:**

- Preferencje zależą od odległości $\|x_j - x_i^*\|$

# Własności i ograniczenia

**Własności:**

- Redukuje złożone poglądy do jednej (lub kilku) osi
- Umożliwia porównywanie aktorów politycznych

**Ograniczenia:**

- Zależny od danych użytych do estymacji
- Ideal points są relatywne, nie absolutne
- Trudno porównywać wyniki z różnych badań

**Implikacja:**

- Interpretacja zależy od kontekstu badania

# Zastosowania w badaniach

**W praktyce:**

- Analiza preferencji wyborców
- Szacowanie ideologii polityków

**Metody:**

- Modele IRT (Item Response Theory)
- Analiza roll-call votes (Jessee 2009)

**Dlaczego to ważne:**

- Pozwala badać reprezentację polityczną
- Łączy dane indywidualne z wynikami politycznymi

# Bridging questions (Tausanovitch i Warshaw 2013)

**Problem:**

- Ideal points zależą od zestawu pytań użytych w badaniu
- Nie można bezpośrednio porównywać wyników z różnych ankiet

**Rozwiązanie: bridging questions**

- Te same pytania zadawane różnym respondentom w różnych badaniach
- Tworzą wspólną skalę ideologiczną

**W tej pracy:**

- Autorzy używają tzw. "supersurvey"
- Zawiera pytania wspólne dla wielu badań
- Pozwala połączyć dane i estymować ideal points w jednej przestrzeni

**Efekt:**

- Porównywalne estymaty preferencji dla wszystkich respondentów

# Model IRT (Item Response Theory)

**Definicja:**

- Model statystyczny do estymacji **ukrytych cech** (latent traits)
- W tej pracy: ukrytą cechą jest **ideal point**

**Dane:**

- Odpowiedzi na pytania polityczne (tak/nie)

**Intuicja:**

- Odpowiedzi respondentów ujawniają ich pozycję ideologiczną

# Jak działa model IRT?

**Założenie:**

- Prawdopodobieństwo odpowiedzi zależy od:
  - pozycji respondenta
  - charakterystyki pytania

**Model:**

$$Pr(y_{ij} = 1) = \Phi(x_i b_j - a_j)$$

**Interpretacja:**

- Im bliżej ideal point, tym większa szansa odpowiedzi „tak”

# Parametry modelu IRT

**Ideal point ($x_i$):**

- Pozycja ideologiczna jednostki (lewo–prawo)

**Discrimination ($b_j$):**

- Jak dobrze pytanie rozróżnia respondentów

**Difficulty ($a_j$):**

- Pozycja pytania na osi ideologicznej

**Funkcja linkująca:**

- $\Phi(\cdot)$ – dystrybuanta rozkładu normalnego (model probitowy)

# Estymacja modelu IRT

**Problem:**

- Parametry $x_i$, $a_j$, $b_j$ są nieobserwowalne

**Rozwiązanie:**

- Estymacja **joint** (wszystko naraz)

**W tej pracy:**

- Podejście bayesowskie
- Estymacja przez MCMC

**Dodatkowo:**

- Normalizacja skali (identyfikacja modelu)

# Algorytm

**Inicjalizacja – rozkłady a priori:**

$$x_i^{(0)} \sim \mathcal{N}(0, \sigma_{x_i}^2), \quad a_j^{(0)} \sim \mathcal{N}(0, \sigma_{a_j}^2), \quad b_j^{(0)} \sim \mathcal{N}(0, \sigma_{b_j}^2)$$

**MCMC (Gibbs / iteracyjnie):**
dla $t = 1, \ldots, T$

1. **Aktualizacja ideal points:**

$$x_i^{(t)} \sim p(x_i \mid \mathrm{Y}, a^{(t-1)}, b^{(t-1)})$$

2. **Aktualizacja trudności:**

$$a_j^{(t)} \sim p(a_j \mid \mathrm{Y}, x^{(t)}, b^{(t-1)})$$

3. **Aktualizacja dyskryminacji:**

$$b_j^{(t)} \sim p(b_j \mid \mathrm{Y}, x^{(t)}, a^{(t)})$$

**Wynik:** Otrzymujemy $\hat{x}_i, \hat{a}_j, \hat{b}_j$ jako średnie próbkowe z rozkładów a posteriori.

# MRP: Multilevel Regression and Poststratification

**Problem:**

- Małe próby w badaniach dla okręgów / regionów
- Trudno oszacować preferencje lokalne

**Rozwiązanie: MRP**

- Łączy dane indywidualne z informacją o populacji
- Pozwala estymować preferencje dla małych jednostek

**Idea:**

- Modelujemy preferencje, a następnie dopasowujemy je do struktury populacji

# Etap 1: Multilevel regression

**Model:**

- Preferencje jako funkcja:
  - cech demograficznych
  - cech geograficznych

**Cechy podejścia:**

- Hierarchiczna struktura (multilevel)
- Częściowe "pooling" między jednostkami

**Efekt:**

- Wszystkie obserwacje pomagają estymować każdą jednostkę

# Etap 2: Poststratyfikacja

**Krok 1:**

- Przewidujemy preferencje dla każdej grupy:
  - np. wiek $\times$ płeć $\times$ region

**Krok 2:**

- Ważymy wyniki udziałem grup w populacji

**Krok 3:**

- Sumujemy $\rightarrow$ estymata dla okręgu

**Efekt:**

- Dokładne estymaty nawet przy małych próbach

# Pozyskiwanie danych z TrueViews

**Problem:**

- TrueViews nie udostępnia gotowych plików `.csv`
- Dane są prezentowane bezpośrednio na stronach jako tabele

**Konsekwencja:**

- Konieczne było automatyczne pobieranie i parsowanie stron

**Podejście:**

- Scrapowanie danych z poziomu HTML
- Ekstrakcja tabel i konwersja do formatu tabelarycznego

# Automatyzacja scrapowania

**Kluczowa obserwacja:**

- Adresy URL mają regularną strukturę
- Zawierają parametry, m.in.:
  - `questioncode`
  - `geotype`
  - identyfikator stanu

**Dzięki temu można było:**

- generować adresy stron automatycznie
- wczytywać strony hurtowo
- lokalizować tabelę w kodzie strony
- zamieniać tabelę na obiekt `pandas.DataFrame`
- eksportować dane do `.csv`

# Hipoteza badawcza

**Punkt wyjścia:**

- Standardowo: ideal point jednowymiarowy

$$x_i \in \mathbb{R}$$

- Interpretacja: oś lewica–prawica

**Hipoteza:**

- W obecności danych klimatycznych:

$$x_i \in \mathbb{R}^d, \quad d > 1$$

- Preferencje mogą być wielowymiarowe

**Uzasadnienie:**

- Postawy klimatyczne nie zawsze pokrywają się z ideologią ekonomiczną
- Dane YCOM dotyczą specyficznego obszaru polityki (klimat)
- Różnice lokalne mogą generować dodatkowy wymiar preferencji

**Implikacja:**

- Podział ideologiczny nie musi być redukowalny do jednej osi

# Ideal point na poziomie hrabstwa

**Punkt wyjścia:**

- Standardowo: ideal point dla jednostki

$$x_i \in \mathbb{R}^d$$

**Proponowane założenie:**

- Każdemu hrabstwu $c$ przypisujemy ideal point:

$$x_c \in \mathbb{R}^d$$

**Interpretacja:**

- $x_c$ reprezentuje „średnie" preferencje mieszkańców hrabstwa

**Możliwe podejścia:**

- Agregacja:

$$x_c = \frac{1}{N_c} \sum_{i \in c} x_i$$

- Estymacja bezpośrednia (MRP / modele hierarchiczne)

**Cel:**

- Porównywanie regionów w jednej przestrzeni preferencji

# Problemy integracji danych

**Cel:**

- Połączenie danych z YCOM i TrueViews

**Wyzwania:**

- Konieczność wspólnego klucza (np. FIPS)
- Niespójne nazwy jednostek geograficznych

**Przykłady błędów:**

- Bedford City vs Bedford County
- DuBois County vs Dubois County

**Rozwiązanie:**

- Standaryzacja nazw
- Mapowanie do wspólnego identyfikatora

# Problemy administracyjne

**Connecticut:**

- Zmiana podziału administracyjnego (2013)
- Dane w różnych wersjach podziału

**Alaska:**

- Brak klasycznych counties (boroughs, census areas)

**Konsekwencje:**

- Trudności w dopasowaniu danych między źródłami
- Konieczność ręcznego ujednolicenia struktur

# Klasyfikacja i grupowanie

**Cel:**

- Identyfikacja grup regionów o podobnych preferencjach

**Metody:**

- K-means:

$$\min \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - \mu_k\|^2$$

- Hierarchiczne grupowanie

**Pytania badawcze:**

- Czy istnieją wyraźne klastry regionów?
- Czy odpowiadają podziałom politycznym?

# Bibliografia i źródła danych

**Artykuły naukowe:**


- Jessee, S. A. (2009). *Spatial Voting in the 2004 Presidential Election*. American Political Science Review.
- Clinton, J., Jackman, S., Rivers, D. (2004). *The Statistical Analysis of Roll Call Data*. American Political Science Review.
- Bafumi, J., Herron, M. C. (2010). *Leapfrog Representation and Extremism*. American Political Science Review.
- Tausanovitch, C., Warshaw, C. (2013). *Measuring Constituent Policy Preferences*. Journal of Politics.
- Howe, P. D. et al. (2015). *Geographic variation in opinions on climate change*. Nature Climate Change.


**Dodatkowe źródła danych:**

- `https://trueviews.org`
- `https://climatecommunication.yale.edu/visualizations-data/ycom-us/`

# Koniec

Dziękuję za uwagę