Online Collaborative Prediction of Regional Vote Results

Vincent Etter, Emtiyaz Khan, Mattias Grossglauser, Patrick Thiran

DSAA — October 17, 2016 — Montréal, Canada

Data Opportunity

- Many countries adopt **open government** initiatives
- Several **datasets** published
  - Demographics
  - State affairs
  - Votes and elections
- **Unique opportunity**
  - Get a better **understanding**
  - Build **tools** useful to others

Voting Data

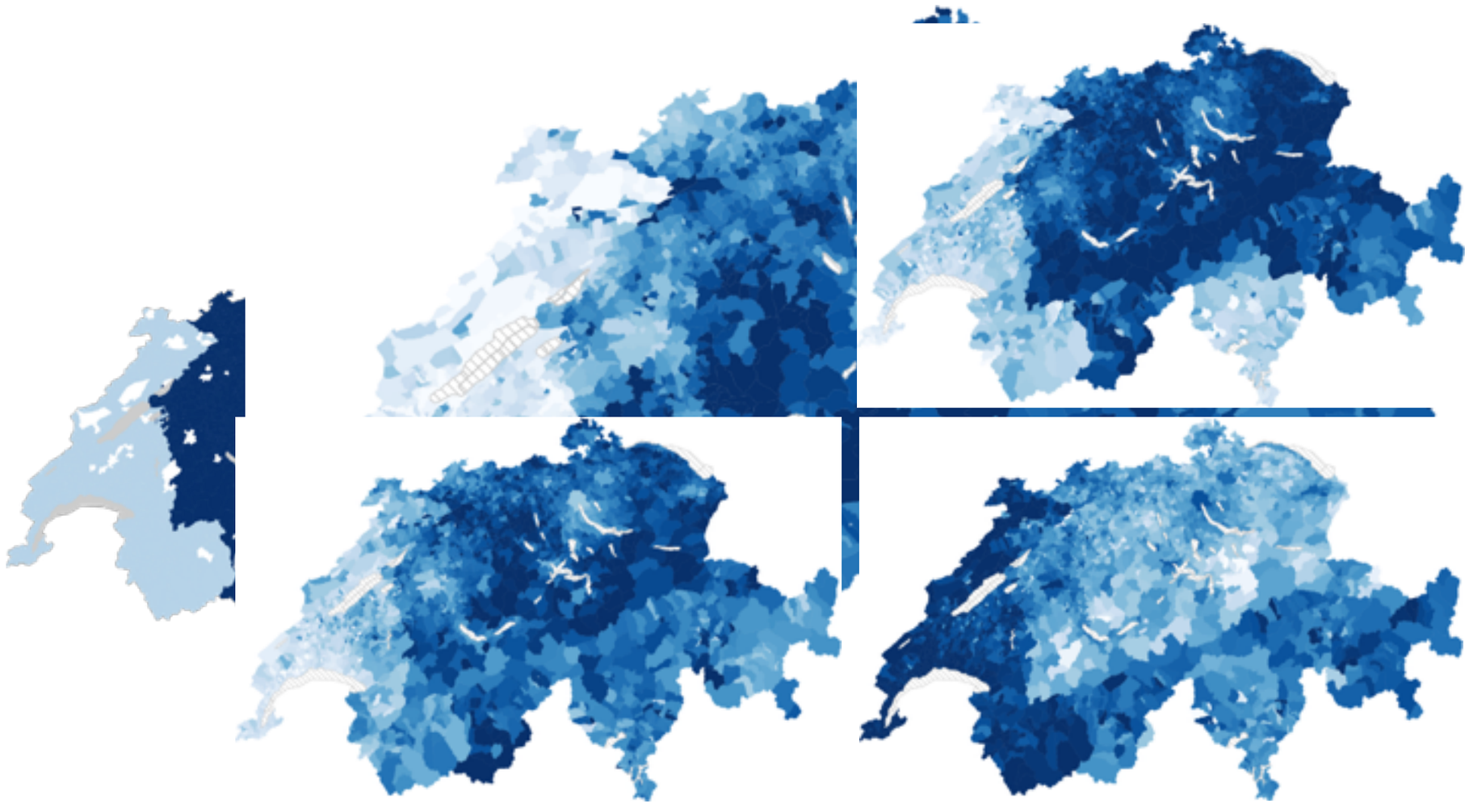
- News agencies, political parties, and polling institutes are all interested in **understanding** voting behaviors
- Will the next vote pass easily?
- What makes two regions vote similarly?
- Where should we focus our efforts?

Dataset

- Vote results from **Switzerland**
  - Issue votes between **1981** and **2014**
  - Outcome (% of “yes”) at the **municipality** level
- **281 votes**
  - 13 features: voting recommendation of the main parties
- **2352 regions**
  - 25 features: languages spoken, demographics, etc.

Data available at http://vincent.etter.io/dsaa16

Similarities Between Results



Online Predictions

- On the day of the vote, regional results are released in **sequence**
- Use **published results** to predict others
- … and **refine** the prediction as more results are published?

Our Approach

- Use a matrix-factorization model to capture the **bi-clustering**
- Add region and vote **features**
  - Reduce the **cold-start** problem
  - More **interpretable**
- Build the model **incrementally** to assess the effect of each component

Our Model

$$y_{dn} = z_{dn} + \epsilon$$

$$z_{dn} = \mu_n + f_n(\boldsymbol{x}_d) + f_d(\boldsymbol{w}_n) + \boldsymbol{v}_d^T \boldsymbol{u}_n$$

bias

regression on region

regression on vote

matrix factorization

Our Models

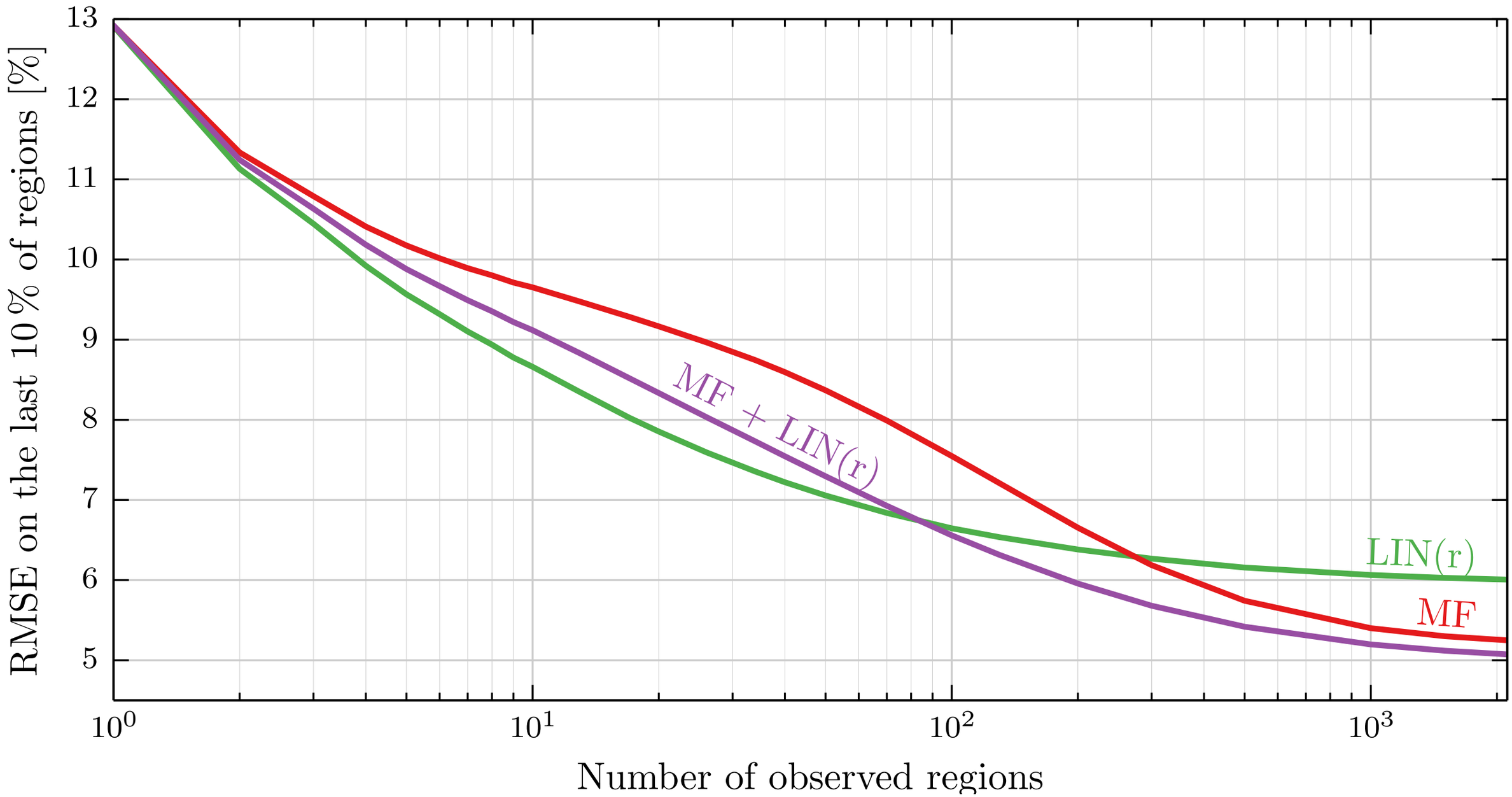
$$z_{dn} = \mu_n + \quad f_n(\boldsymbol{x}_d) \quad + \quad f_d(\boldsymbol{w}_n) \quad + \quad \boldsymbol{v}_d^T \boldsymbol{u}_n$$

LIN(r)	$z_{dn} = \mu_n +$	$\boldsymbol{\beta}_n^T \boldsymbol{x}_d$					
LIN(v)	$z_{dn} = \mu_n +$			$\boldsymbol{\gamma}_d^T \boldsymbol{w}_n$			
LIN(r) + LIN(v)	$z_{dn} = \mu_n +$	$\boldsymbol{\beta}_n^T \boldsymbol{x}_d$	$+$	$\boldsymbol{\gamma}_d^T \boldsymbol{w}_n$			$\lambda_\beta, \lambda_\gamma, \lambda_u, \lambda_v$
MF	$z_{dn} = \mu_n +$				$+$	$\boldsymbol{v}_d^T \boldsymbol{u}_n$	
MF + LIN(r)	$z_{dn} = \mu_n +$	$\boldsymbol{\beta}_n^T \boldsymbol{x}_d$			$+$	$\boldsymbol{v}_d^T \boldsymbol{u}_n$	
MF + GP(r)	$z_{dn} = \mu_n +$	$\mathrm{GP}(\boldsymbol{x}_d)$			$+$	$\boldsymbol{v}_d^T \boldsymbol{u}_n$	$\boldsymbol{\theta}, \sigma_s, \lambda_\gamma$
MF + GP(r) + LIN(v)	$z_{dn} = \mu_n +$	$\mathrm{GP}(\boldsymbol{x}_d)$	$+$	$\boldsymbol{\gamma}_d^T \boldsymbol{w}_n$	$+$	$\boldsymbol{v}_d^T \boldsymbol{u}_n$	

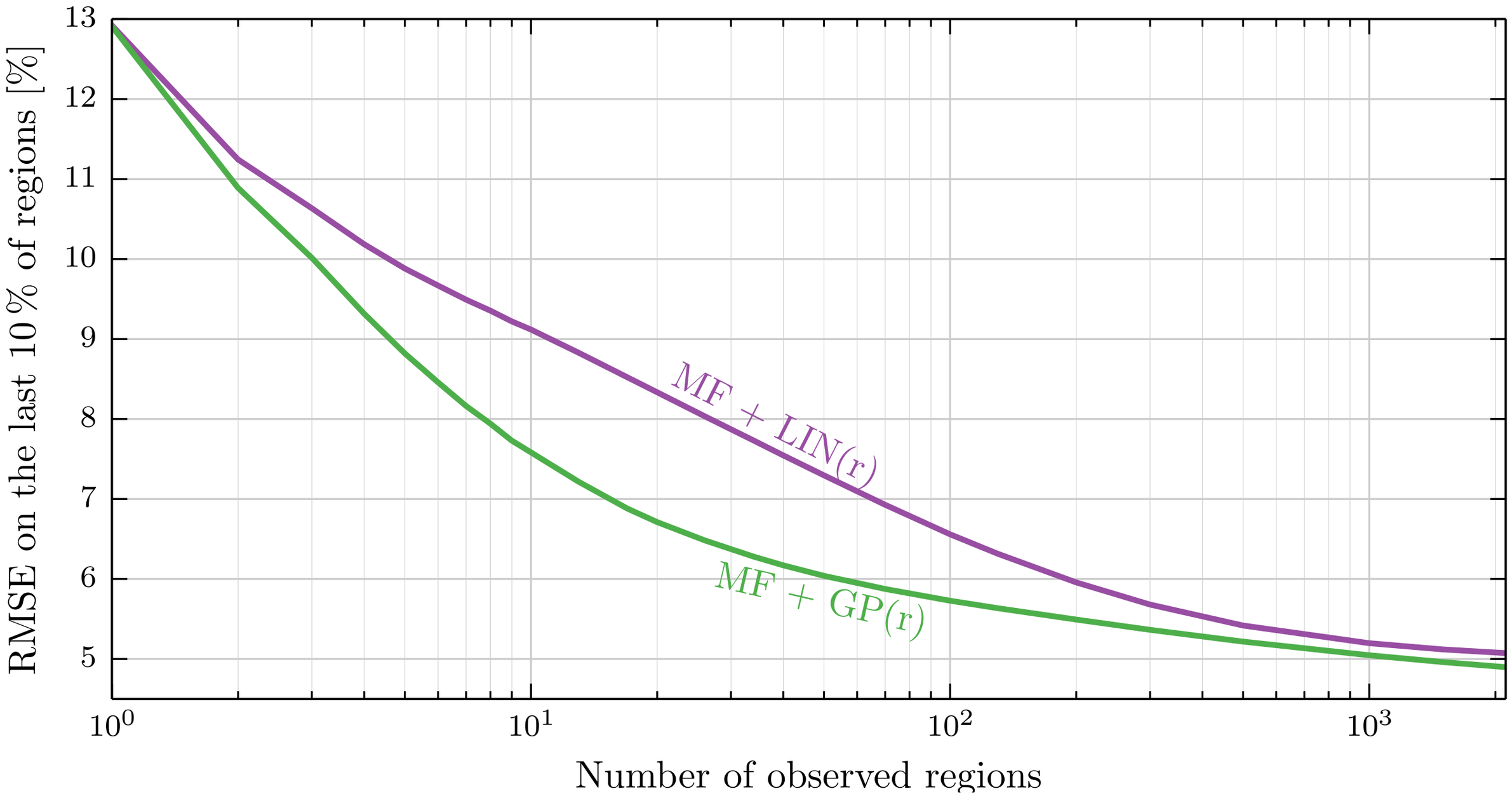
Performance Evaluation

- **Last 50** votes as test data
- Simulate 500 random **reveal order**
- Last 10% of regions as **test regions**
- **Observe** increasing number of regions
- **Predict** result of test regions

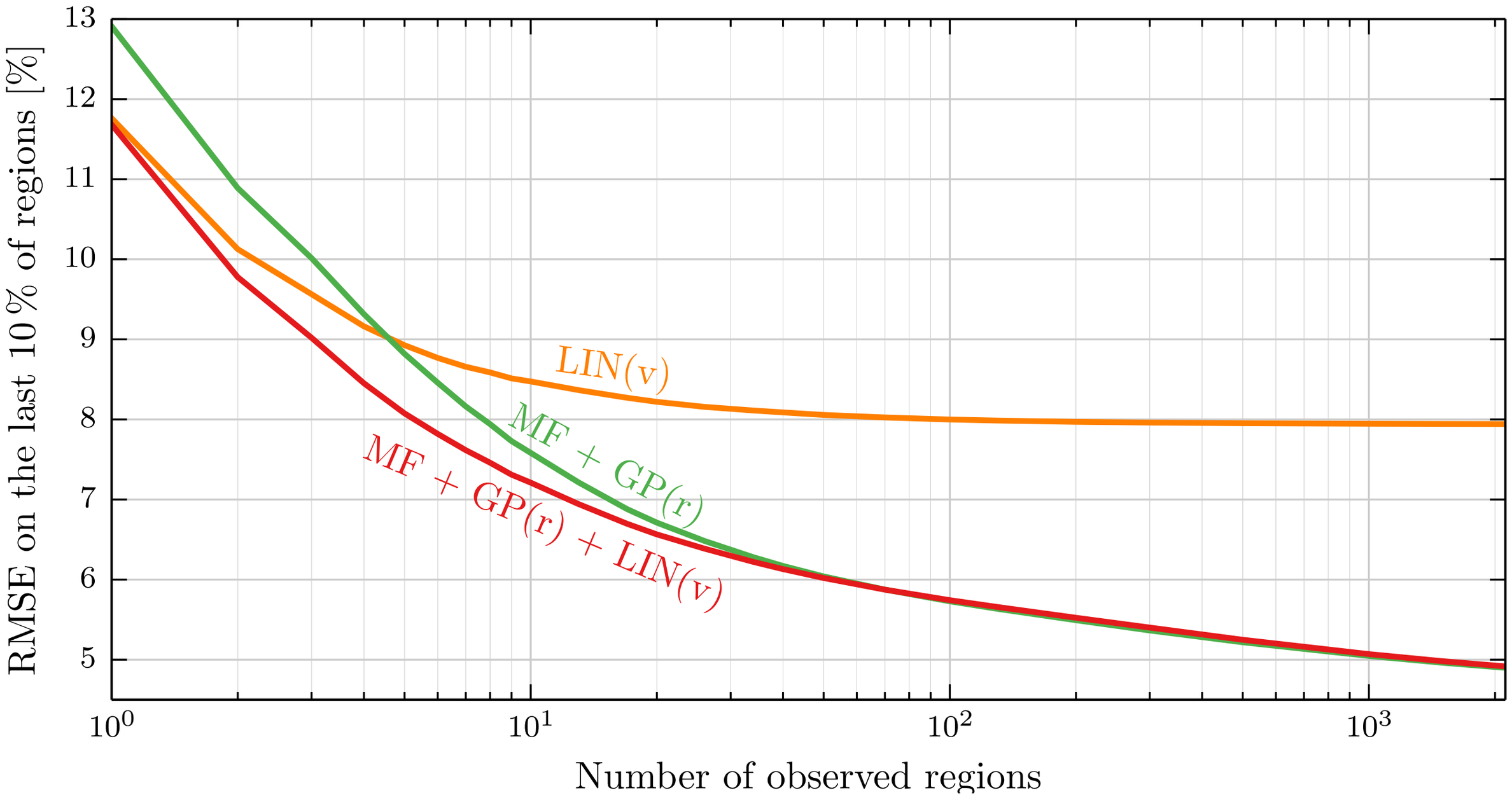
Results


RMSE on the last 10 % of regions [%]
13
12
11
10
9
8
7
6
5
10^0
10^1
10^2
10^3
Number of observed regions
MF + LIN(r)
LIN(r)
MF

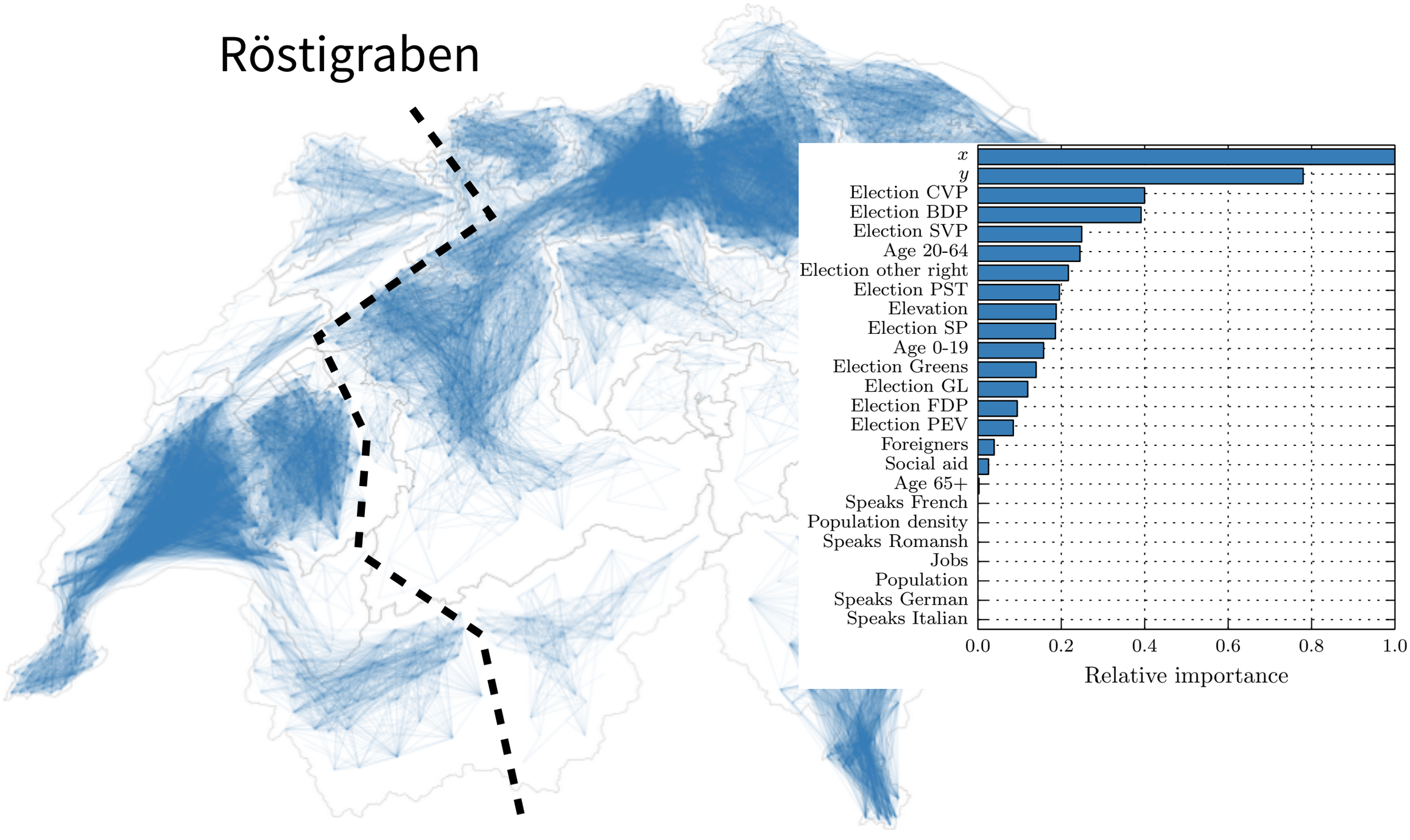
Bayesian VS Non-Bayesian



Final Model


RMSE on the last 10 % of regions [%]
13
12
11
10
9
8
7
6
5
10^0
10^1
10^2
10^3
Number of observed regions
LIN(v)
MF + GP(r)
MF + GP(r) + LIN(v)

Interpretation



Summary

- Individual models have different **strengths**
  - Vote features regression for **cold start**
  - Region features and bi-clustering when **more** observations
- **Bayesian** methods are useful
  - Proper **hyperparameters** setting
  - Accurate and interpretable **results**

Thank you!

Code and data available at

http://vincent.etter.io/dsaa16

Any questions?