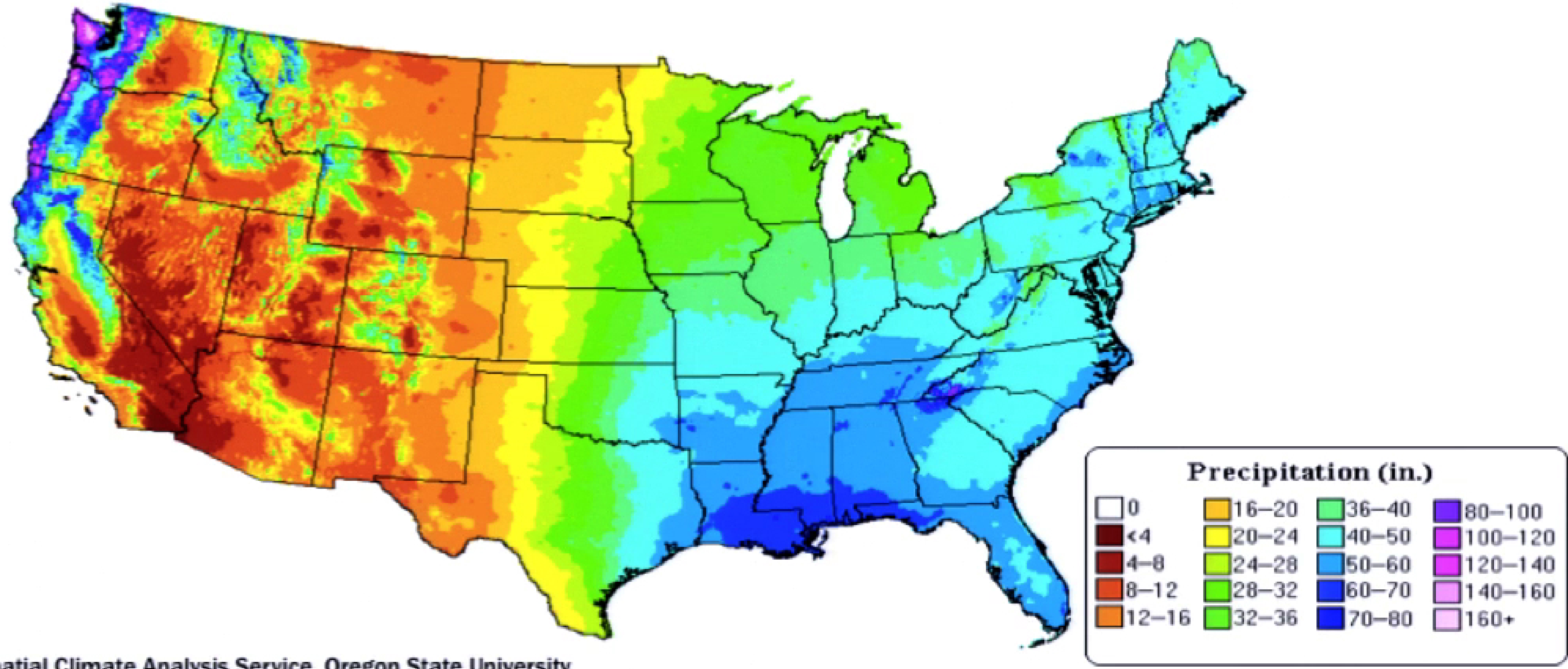
# Semi-Supervised Learning for Structured Regression on Partially Observed Attributed Graphs

Jelena Stojanovic (Temple University)
Milos Jovanovic (University of Belgrade)
Djordje Gligorijevic (Temple University)
Zoran Obradovic (Temple University)

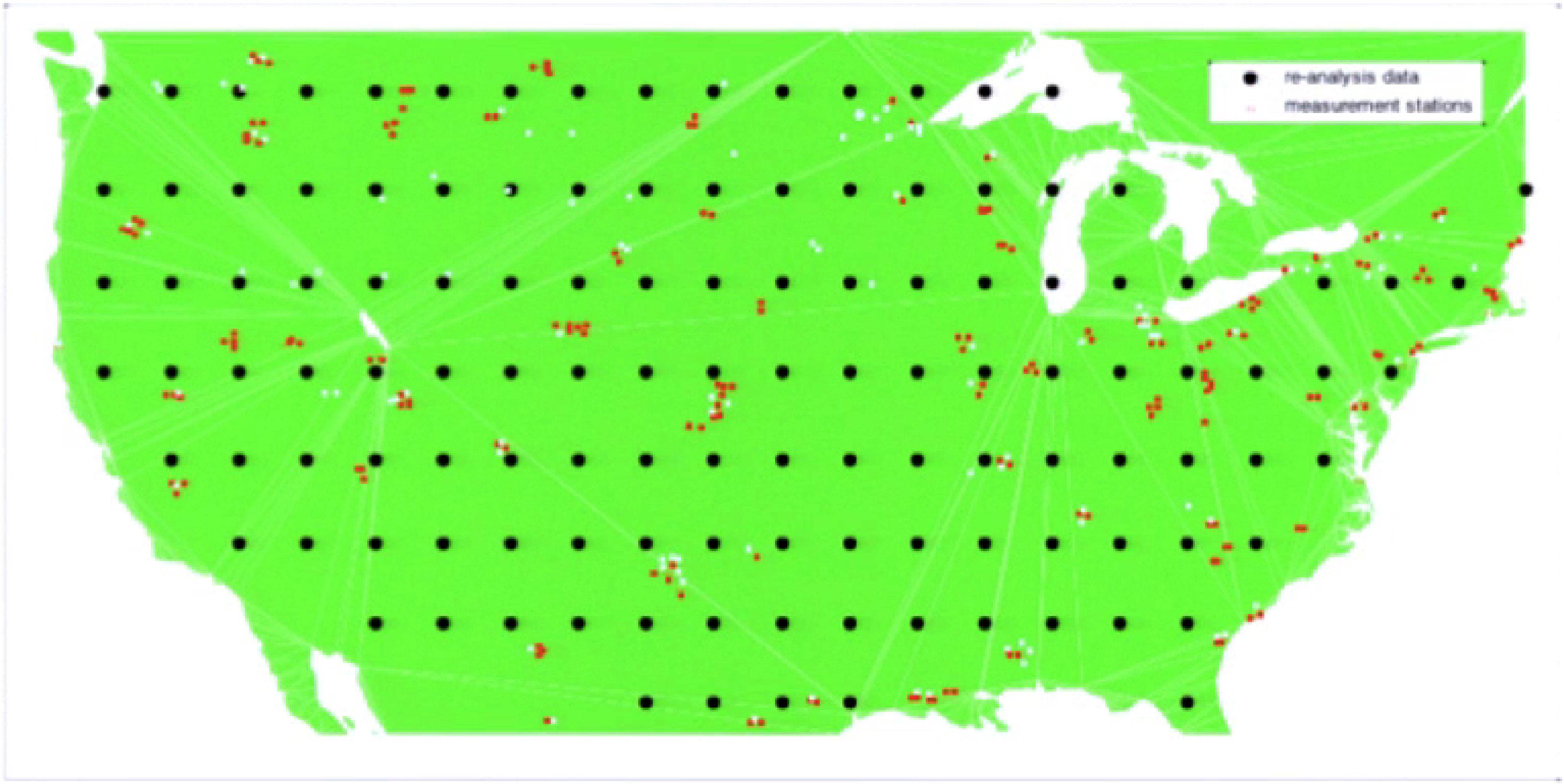
2015 SIAM International Conference on Data Mining, Vancouver, Canada

# Precipitation

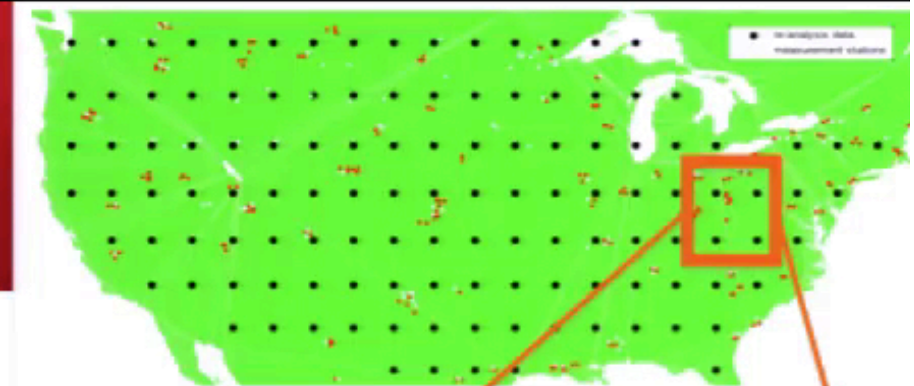



Spatial Climate Analysis Service, Oregon State University
http://prism.oregonstate.edu/ Map created Jul 10, 2012

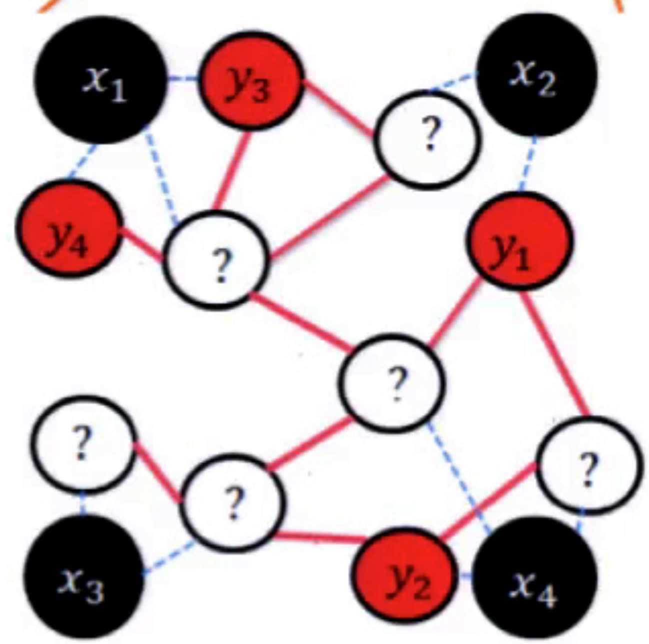
# Precipitation graph

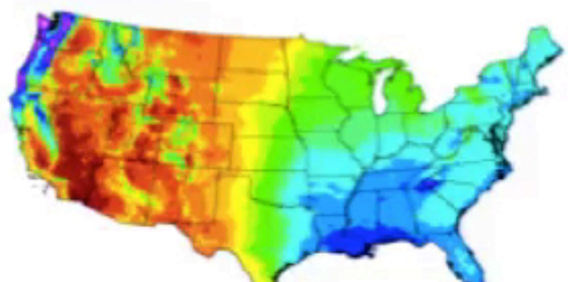


# Precipitation graph

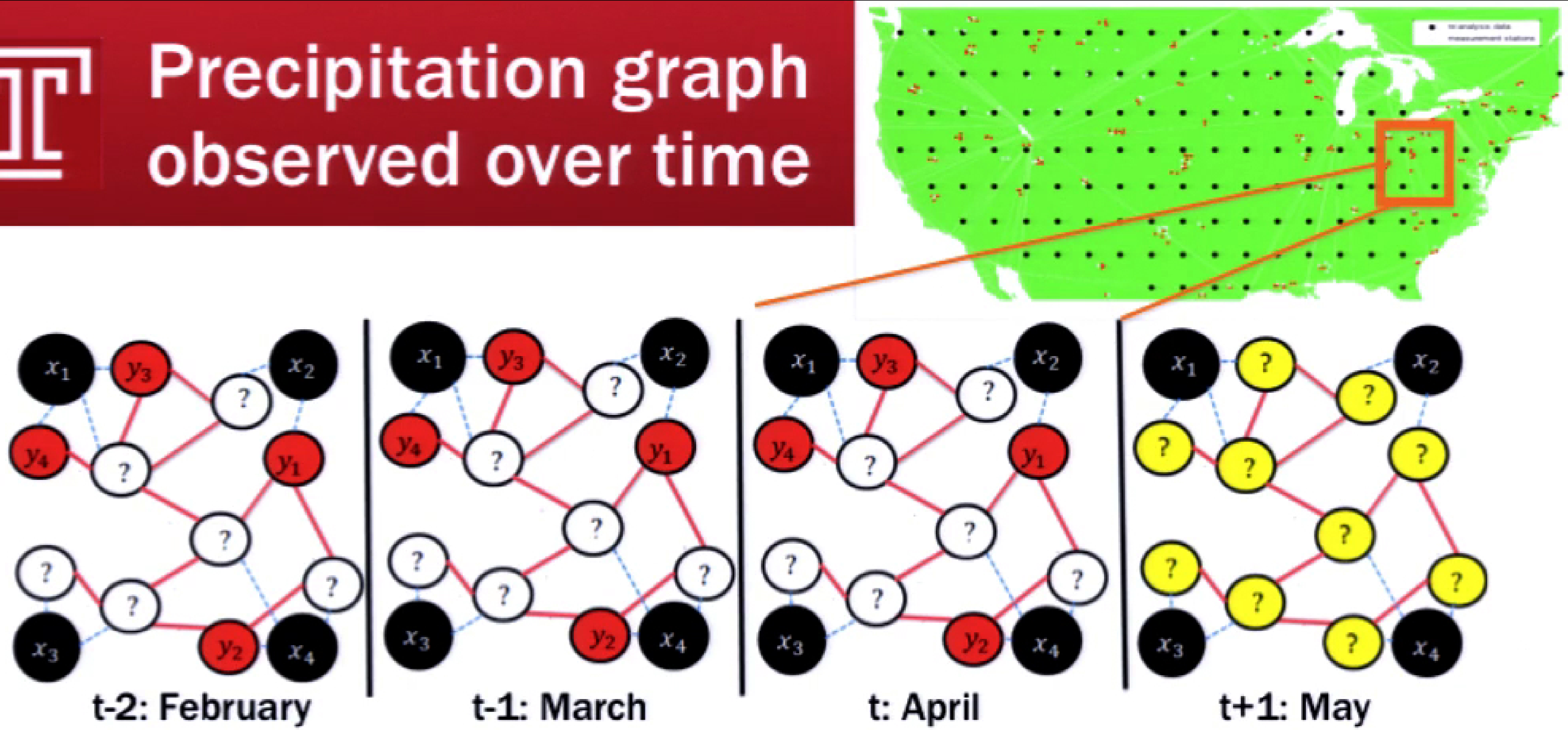


- **Read and white nodes**: 1,132 measurement stations over the whole continental US over time
- **Black nodes**: Re-analysis data- outputs of domain climate models on a coarse scale (124 locations)
- **Links**: Spatial similarities



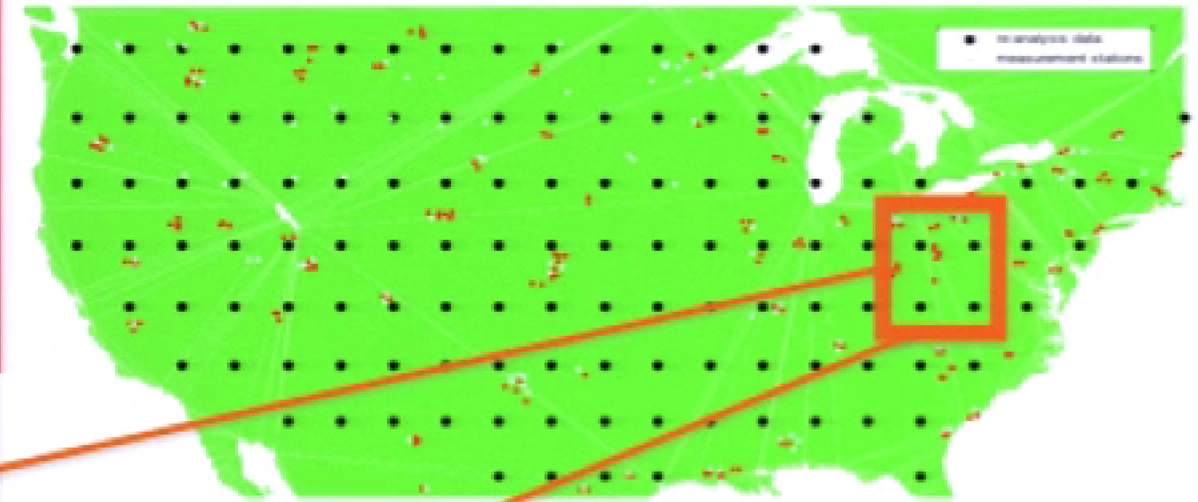



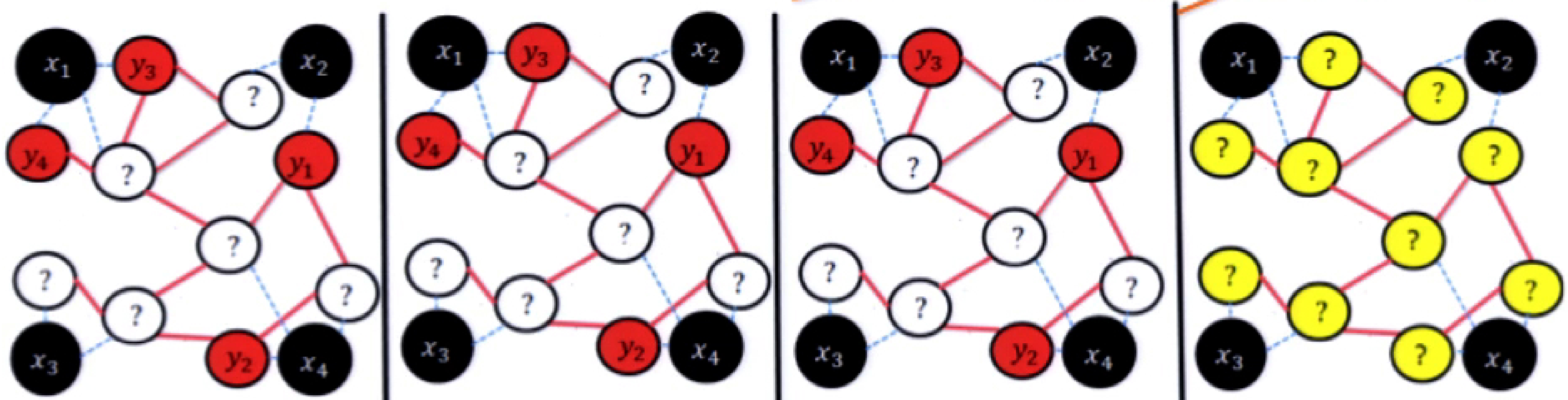
# Precipitation graph observed over time




- **Monthly** precipitation in individual stations (red and white nodes)
- **Missing** response variable (label) at some weather stations (white nodes) sometimes even through the *whole history*
- No missing values in node attributes

# Goal





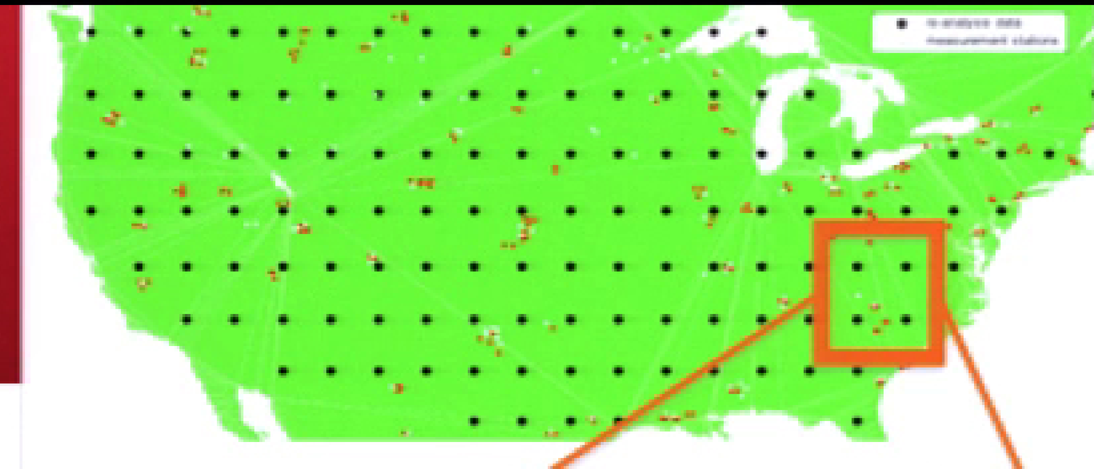

t-2: February | t-1: March | t: April | t+1: May

**Regression in evolving attributed graphs where response variables (labels) are (always) missing in large fraction of training data.**

# Possible approaches

- Conditional probabilistic graphical models - a powerful framework for structured regression in spatio-temporal datasets
  - GCRF model- not designed to cope with missing data (ignoring)
- Imputation based methods
- Learning from labeled and unlabeled nodes together, rather than expecting the missing data to be treated in a preprocessing stage

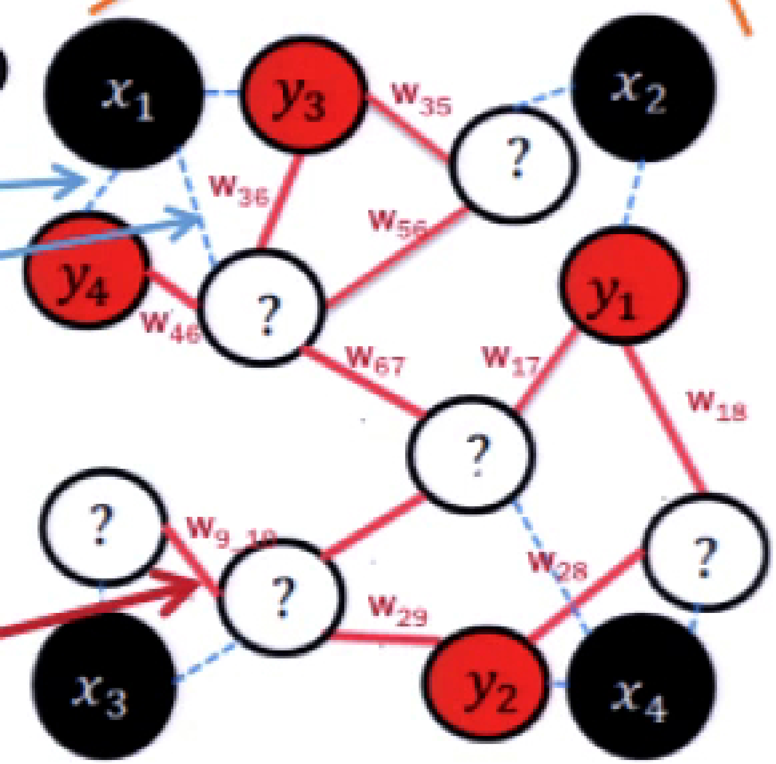
# Gaussian Conditional Random Fields



$$P(\mathbf{y}\mid\mathbf{x}) = \frac{1}{Z(\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\beta})}\exp\Big(\sum_{i=1}^{N} A(\boldsymbol{\alpha}, y_i, \mathbf{x}) + \sum_{j\sim i} I(\boldsymbol{\beta}, y_i, y_j, \mathbf{x})\Big)$$

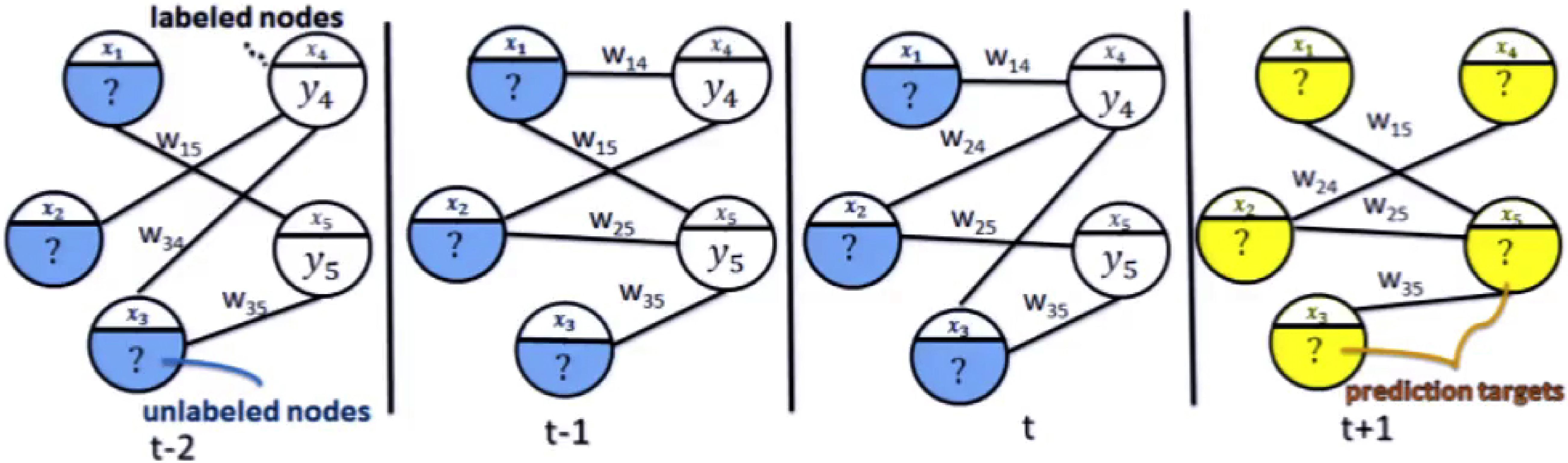
$$A(\alpha, y_i, \mathbf{x}) = -\sum_{k=1}^{K}\alpha_k\big(y_i - \underline{R_k(\mathbf{x},i)}\big)^2$$

$$I(\beta, y_i, y_j, \mathbf{x}) = -\sum_{l=1}^{L}\beta_l \underline{e_{ij}^{(l)} S_{ij}^{(l)}(\mathbf{x})}(y_i - y_j)^2$$



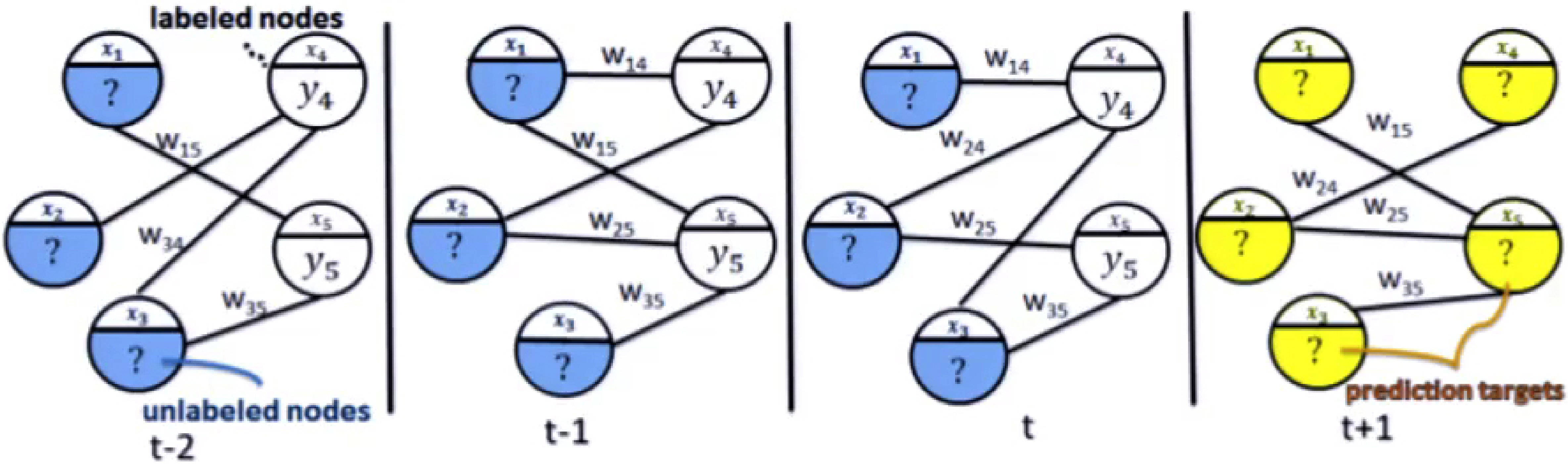

- P(y|x) is Gaussian distribution
- **Learning**: finding parameters **α** and **β** is *convex optimization*
- **Inference:** Point estimate of **y** for given **x** is **μ**, uncertainty is **Σ**, where *P*(**y**|**x**)~*N*(μ, Σ)

# i-GCRF approach




- **i-GCRF approach**: Ignoring nodes that have missing values
    - Loss of information from graph structure

# Our approach




- **Objective**: utilize entire observed structure of the graph in cases when there are missing labels in data
- Idea: Instead of ignoring nodes with missing labels, include the information that is available by marginalization over the unlabeled examples

# Marginalized Gaussian Conditional Random Field (m-GCRF) model

- **The idea: Marginalize out the effect of unlabeled data when calculating conditional probability $P(y_L \mid X)$ from joint probability $P(y_L, y_U \mid X)$ of labeled( $y_L$ ) and unlabeled data ( $y_U$ ):**

$$P\left(\begin{bmatrix} y_L \\ y_U \end{bmatrix} \middle| \begin{bmatrix} X_L \\ X_U \end{bmatrix}\right) \sim N\left(\begin{bmatrix} \mu_L \\ \mu_U \end{bmatrix}, \begin{bmatrix} Q_{LL} & Q_{LU} \\ Q_{UL} & Q_{UU} \end{bmatrix}^{-1}\right)$$

$$P(y_L \mid X) = \int_{y_U} P(y_L, y_U \mid X_L, X_U) d_{y_U}$$

- **Since the original distribution is Gaussian, marginalizing over a subset of variables yields another Gaussian distribution:**

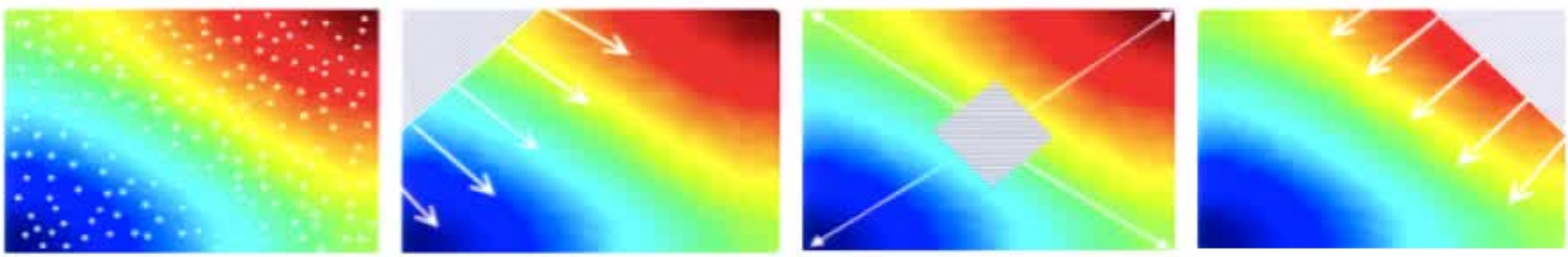
$$P(y_L \mid X) \sim N\left(\mu_L, (Q_{LL} - Q_{LU} Q_{UU}^{-1} Q_{UL})^{-1}\right)$$

# i-GCRF vs. m-GCRF

| i-GCRF | m-GCRF |
|---|---|
| | $P(y_L \mid X) = \int_{y_U} P(y_L, y_U \mid X_L, X_U) d_{y_U}$ |
| $P(y_L \mid X_L) \sim N\left(\mu_L, Q_{LL}^{-1}\right)$ | $P(y_L \mid X) \sim N\left(\mu_L, (Q_{LL} - Q_{LU} Q_{UU}^{-1} Q_{UL})^{-1}\right)$ |

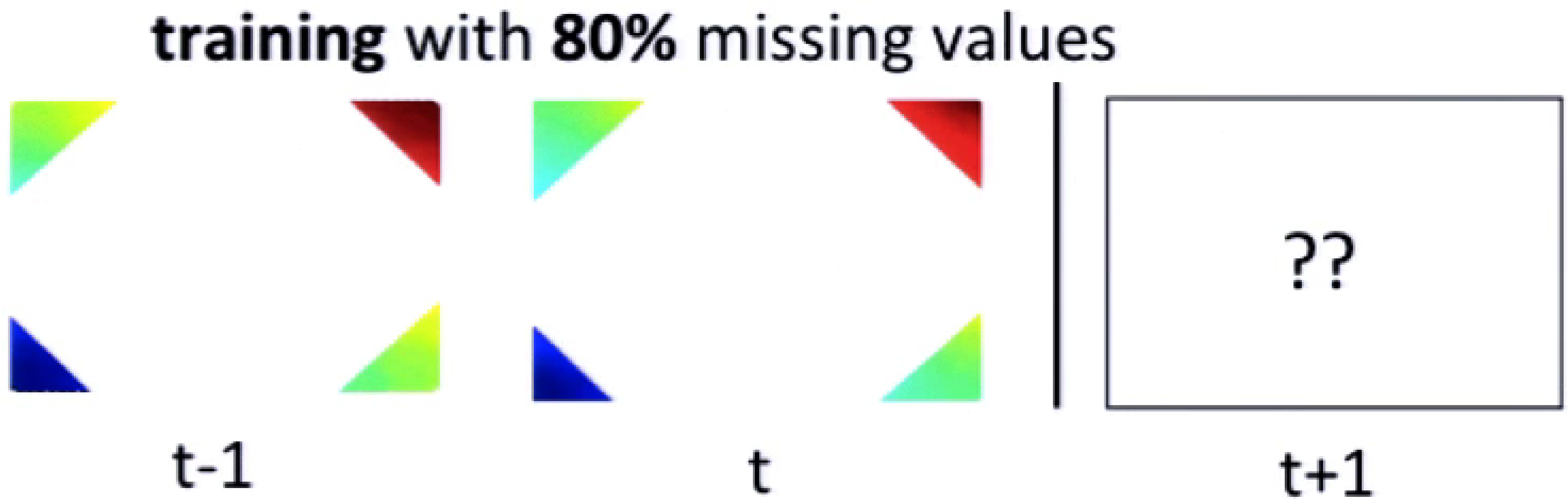
# Evaluation on Evolving Graphs with a Large Fraction of Missing Labels

Experiments on **~500** spatio-temporal **graphs** with up to **80%** of **missing** values under **7** missingness **mechanisms** (up to 15,000 nodes in 5 time steps)



Examples of missingness mechanisms

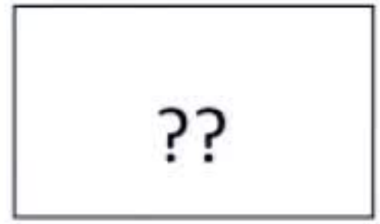
# Evaluation on Evolving Graphs with a Large Fraction of Missing Labels



**training** with **80%** missing values

t-1

t

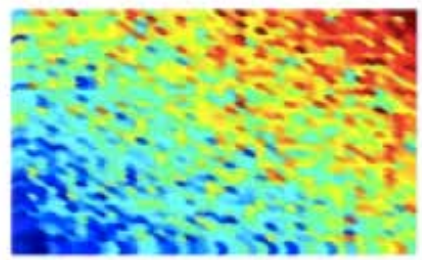



t+1

Ground truth

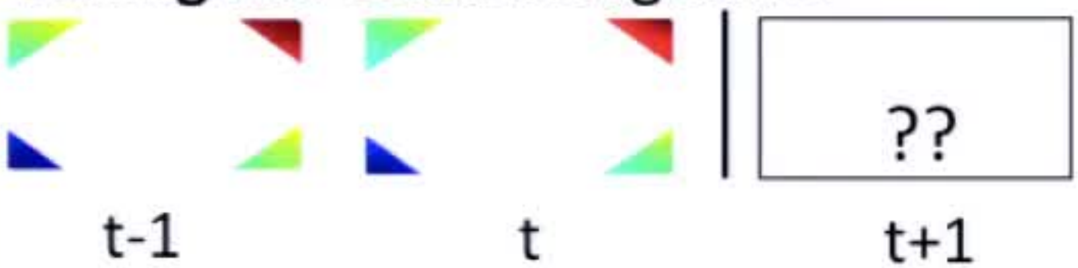
**predicted** values in time step t+1:

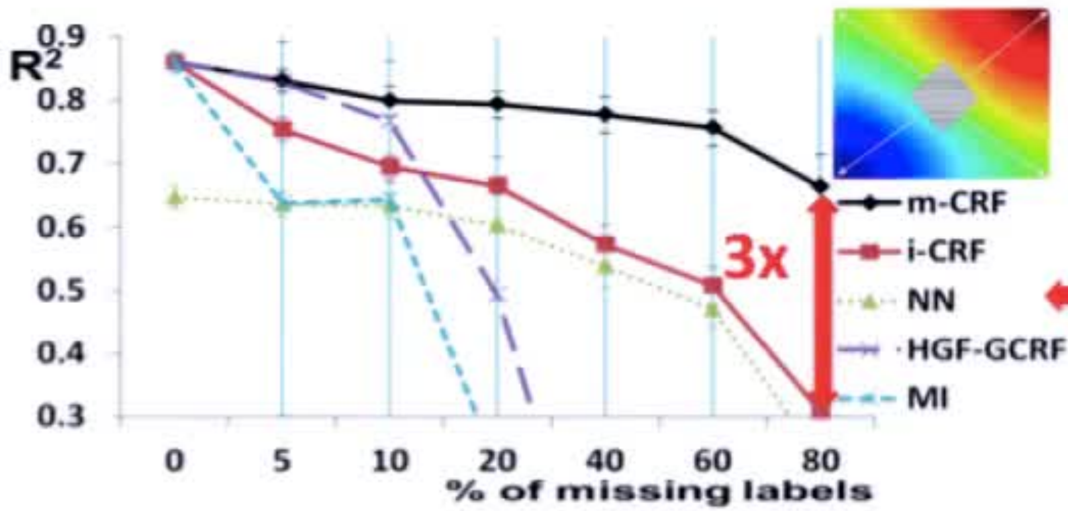
HGF-GCRF: Harmonic Gaussian Field (HGF) for imputation and GCRF for regression ($R^2$ = -1.37)



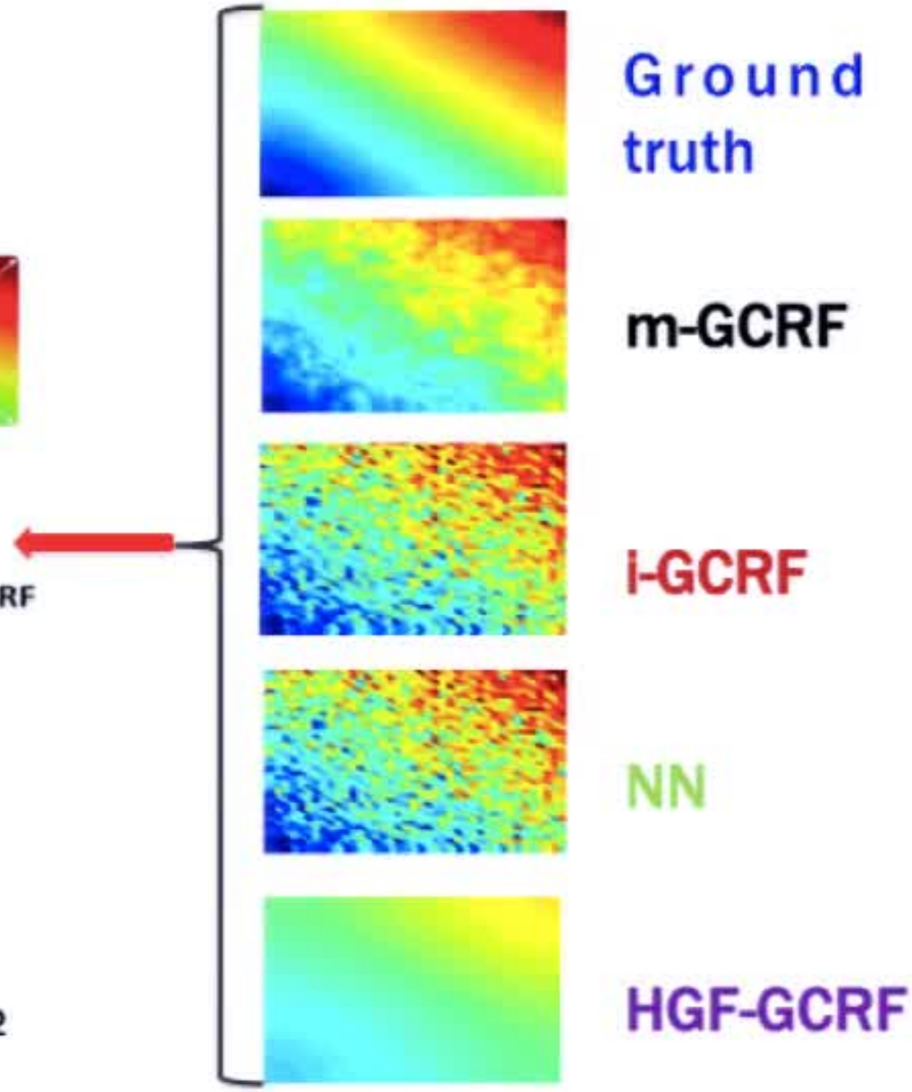
NN: nonlinear neural network ignoring nodes with missing labels ($R^2$ = 0.23)

**training** with **80%** missing values

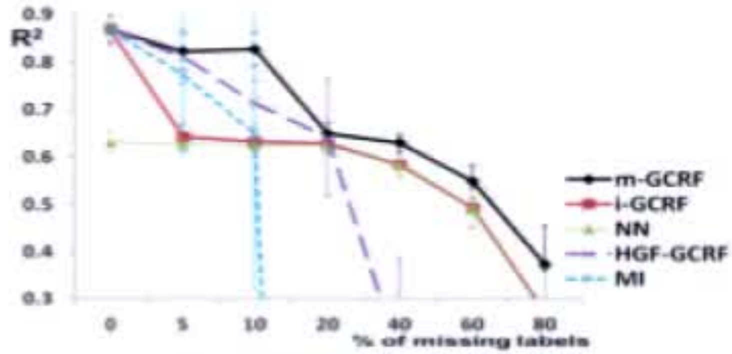


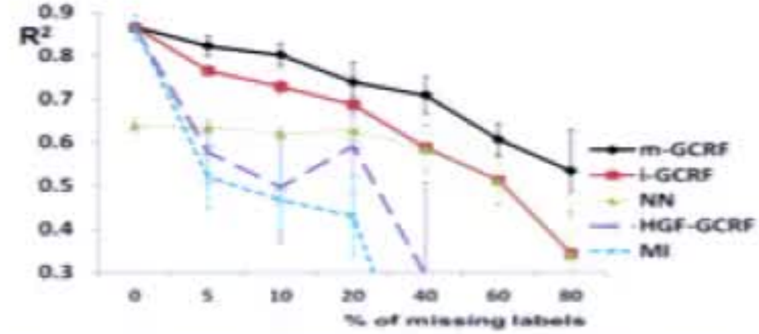

**predicted** values in time step t+1:



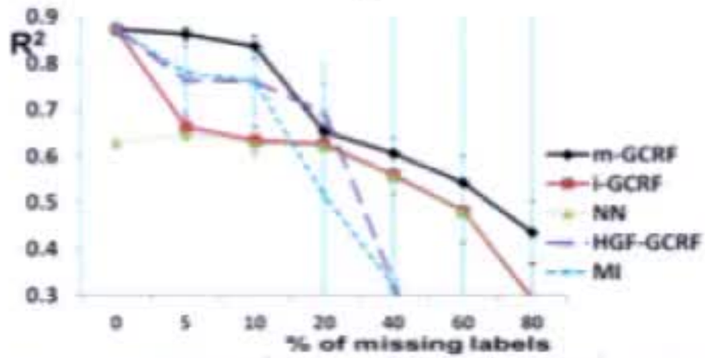

- ✓ m-GCRF is more than **twice as good as** structured i-GCRF ($R^2$ 0.67 vs. 0.3)
- ✓ m-GCRF is about **three times** better than unstructured NN, that has barely positive $R^2$
- ✓ other models are **useless** (negative $R^2$)



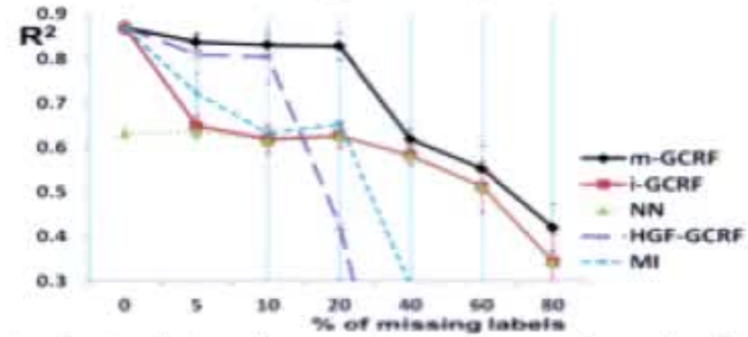

Labels **Missing at Random**



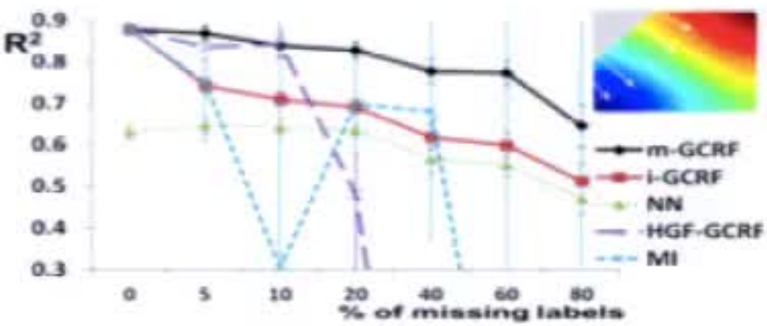

Missing Labels of **weakly connected nodes** (smaller weighted degree)



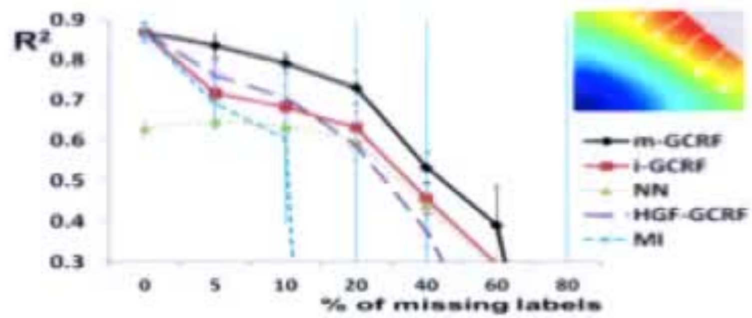

Missing Labels of **strongly connected nodes** (larger weighted degree)




Missing Labels of **strongly connected nodes** (larger weighted degree), keeping neighborhood



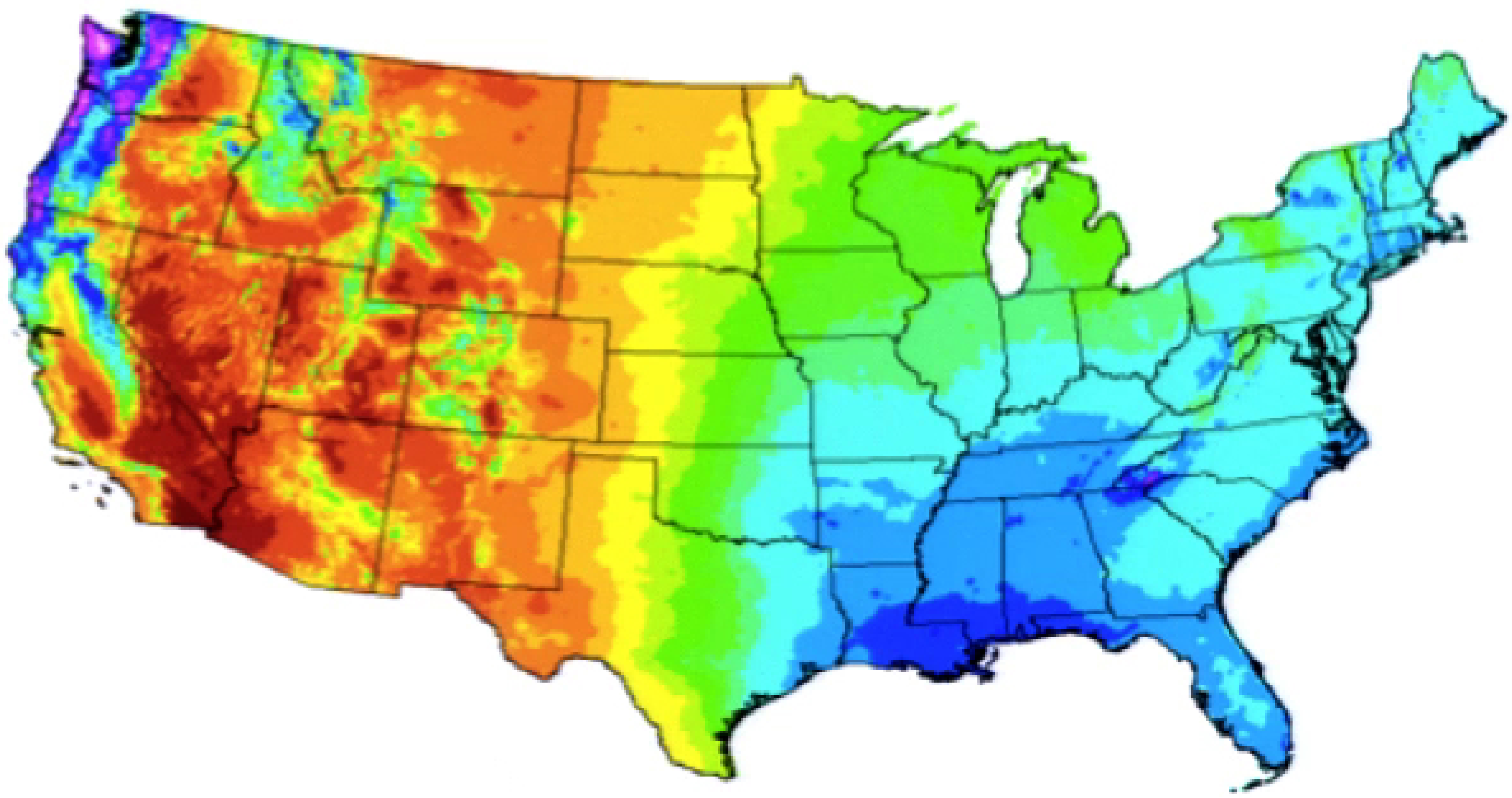

Missing labels of entire neighborhoods **(middle - range values)**




Missing labels of entire neighborhoods **(extremes)**

# Climate Application: Precipitation Prediction

1. Precipitation prediction with up to 80% missing labels
2. Data collection cost reduction



Stojanovic, J., Jovanovic, M., Gligorijevic, D., Obradovic, Z., "Semi-supervised learning for structured regression on partially observed attributed graphs", SDM2015

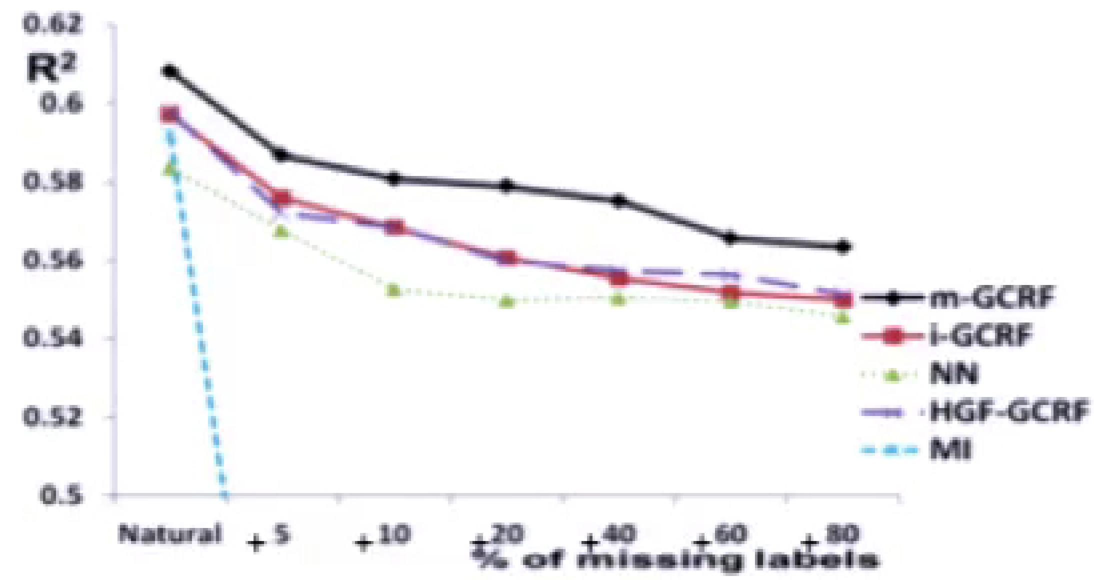
# Precipitation Prediction with up to 80% Missing Labels

✓ **Structured models** were more accurate than:

- an unstructured nonlinear model (NN)
- and statistically sound multiple imputation (MI) that cannot handle more than 10% missing labels ($R^2 < 0$)

✓ Using **m-GCRF** useful information is extracted from partially labeled graph. This was more accurate than:

- ignoring unlabeled nodes (i-GCRF)
- over-smoothing the values semi-supervised structured model HGF



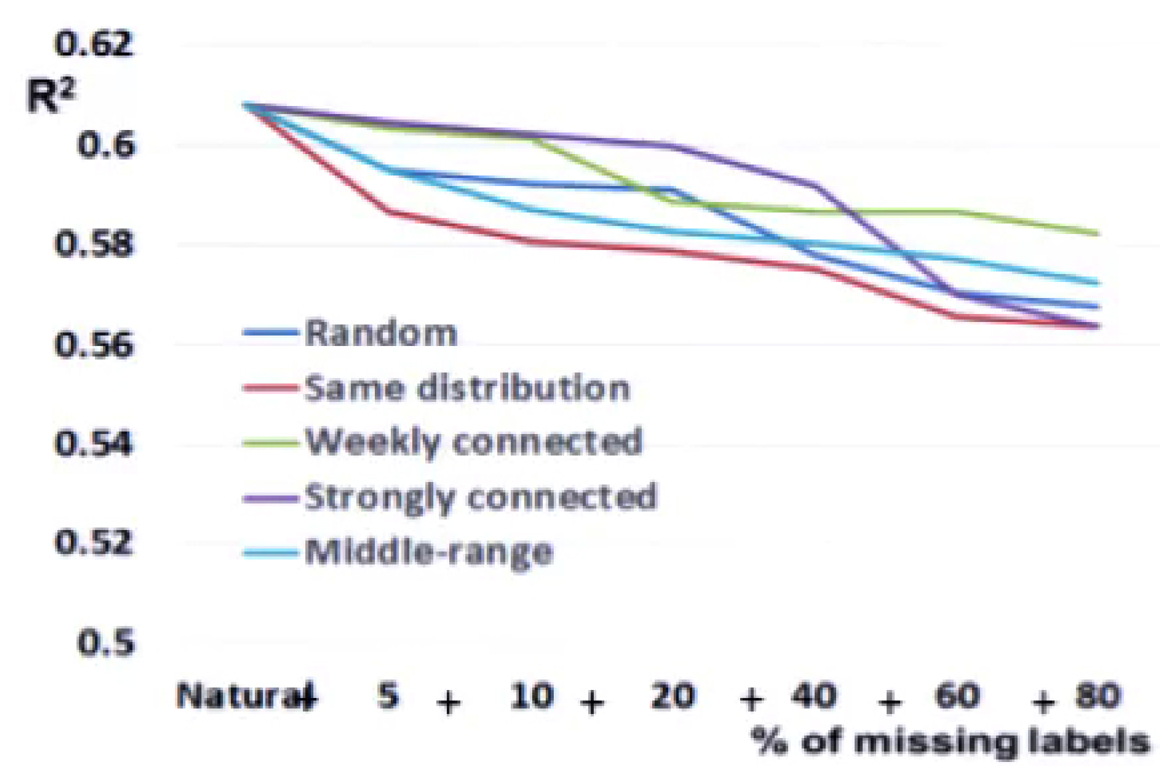

*Natural* missingness process

# Data Collection Cost Reduction

- **Objective: reduce the total number of labels** in the dataset for future data collection (e.g. in a need to reduce the cost)
- Help decision-making regarding the relevance of weather stations by examining how models behave under different missingness mechanisms

✓ Removing **most frequent missing stations** gives the worst results.

✓ Removing **strongly connected** stations preserves fairly similar accuracy when majority of stations are removed




m-GCRF under different strategies of removing labels

# Conclusion

- We proposed Marginalized GCRF method for structured regression on partially observed attributed graphs where nodes might be completely unlabeled in the history
- Experiments on ~500 spatio-temporal graphs with up to 80% of missing values provide evidence that m-GCRF under various missingness mechanisms **outperformed all of the benchmarks.**
- m-GCRF successfully applied to a challenging problem of predicting precipitation based on a temporal graph with missing observations.
- If there is a need to actively decrease the amount of labels in the data, certain data reduction strategies can be more effective

Stojanovic, J., Jovanovic, M., Gligorijevic, D., Obradovic, Z., "Semi-supervised learning for structured regression on partially observed attributed graphs", SDM2015


jelena.stojanovic@temple.edu
http://astro.temple.edu/~tue68039/
http://www.dabi.temple.edu/~zoran/code/sdm15

# Thank you for your attention!

Questions?