Discovering Hidden Structure of House Prices with Relational Latent Manifold Model

Sumit Chopra   Trivikraman Thampy
John Leahy   Andrew Caplin   Yann LeCun

Courant Institute of Mathematical Sciences
&
Department of Economics
New York University

Relational Learning

- Traditional Learning
  - Samples are i.i.d. from an unknown distribution D
- Relational Learning
  - Samples are no longer i.i.d
  - Related to each other in complex ways: values of unknown variables depends on each other
  - Dependencies could be direct or indirect (hidden)
- Examples
  - Web page classification
  - Scientific document classification
- Need a form of collective prediction

Predicting House Price

- “Location Location Location”

Predicting House Price

- “Location Location Location”

Predicting House Price

- "Location Location Location"

Predicting House Price

- "Location Location Location"

- Price is a function of quality/desirability of neighborhood
- Which in turn is a function of quality/desirability of other houses
- This is the relational aspect of the price

Predicting House Price

- However there is more to it
  - The Non-Relational aspect of price

Predicting House Price

- However there is more to it
  - The Non-Relational aspect of price

1 bedroom, 1 bathroom

Predicting House Price

- However there is more to it
  - The Non-Relational aspect of price

1 bedroom, 1 bathroom

4 bedroom, 3 bathroom

Predicting House Price

- However there is more to it
  - The Non-Relational aspect of price

1 bedroom, 1 bathroom

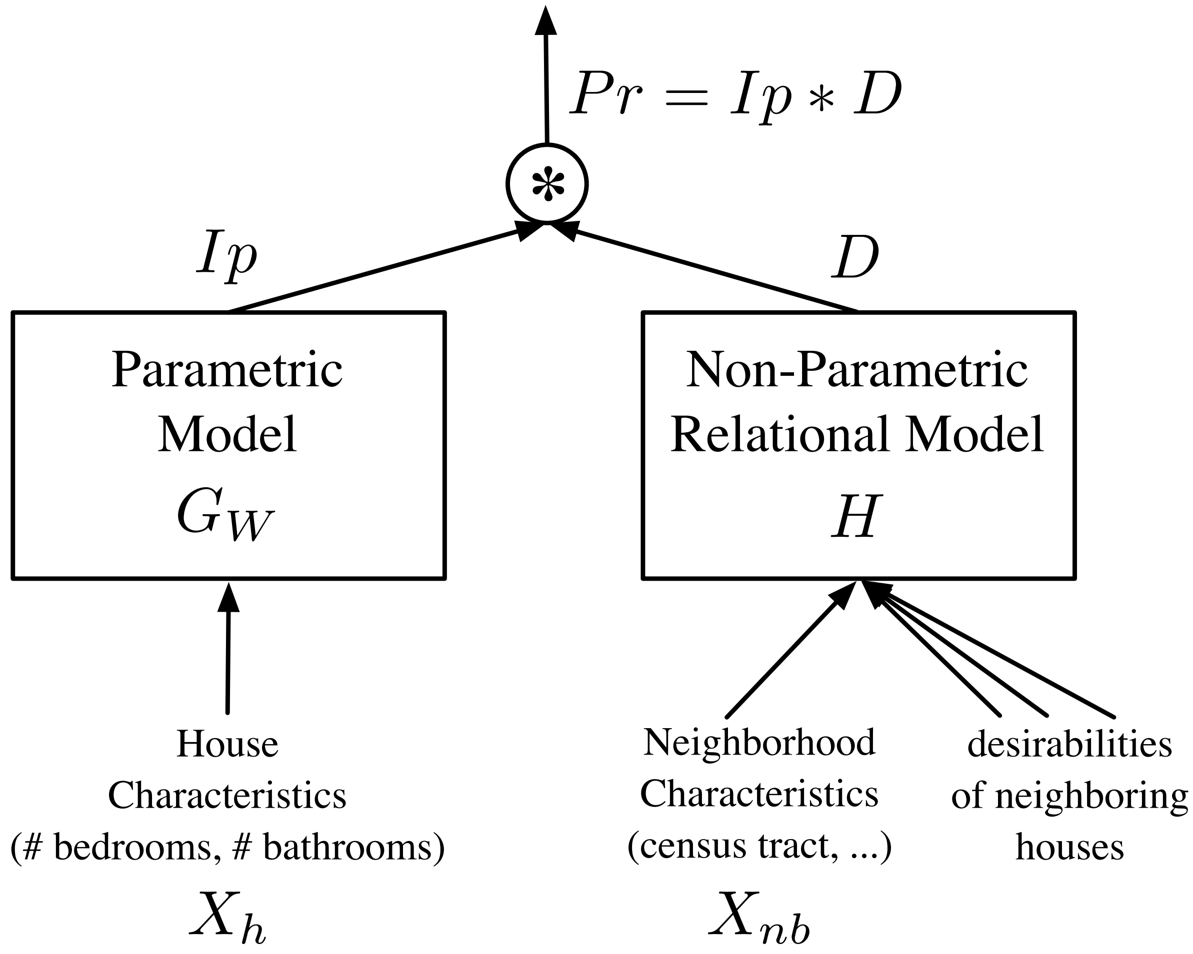
4 bedroom, 3 bathroom

- One can view this as the "intrinsic price"

The Idea

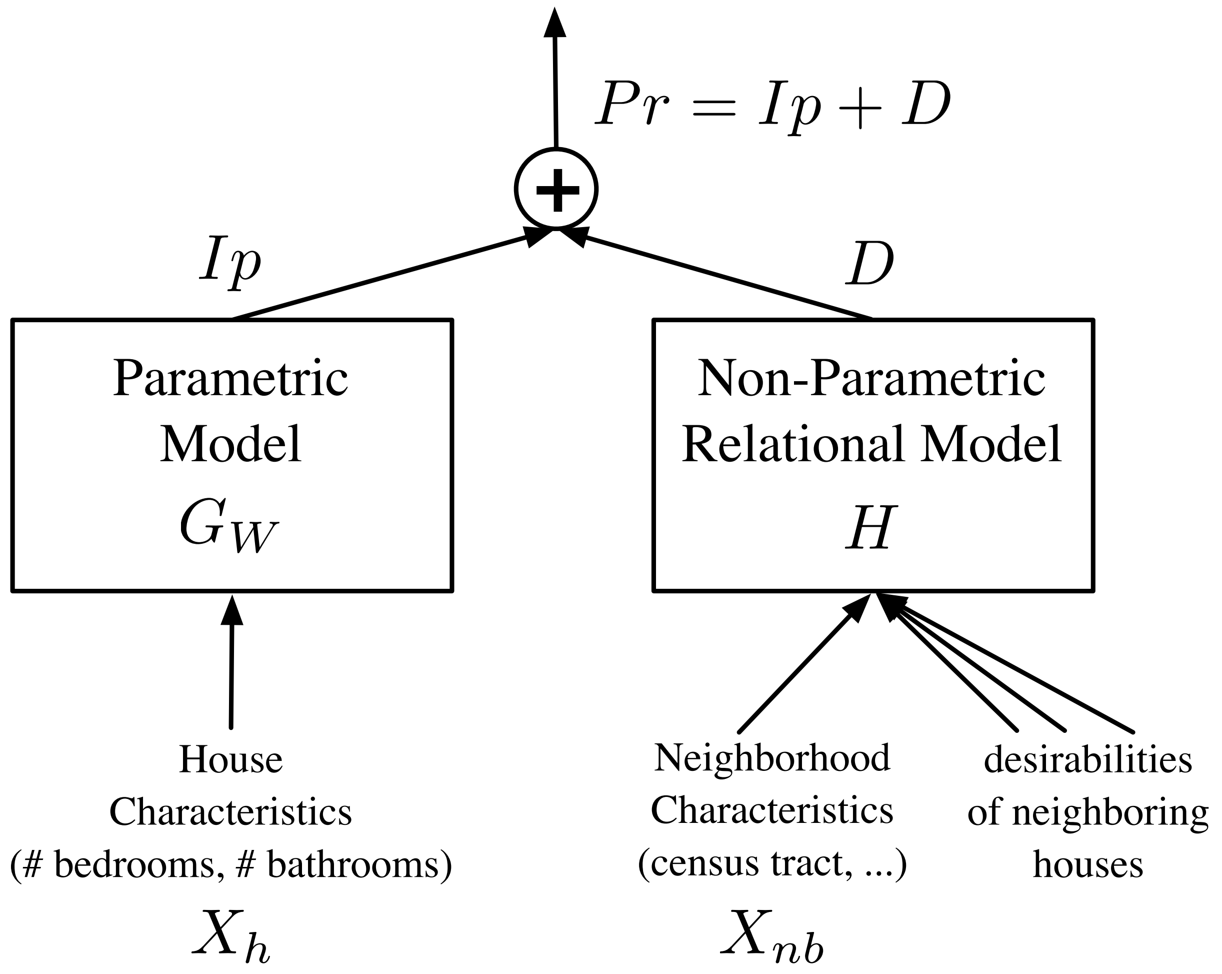
- Price of a house is modeled as

$$price = (intrinsic\, price) * (desirability\, of\, its\, location)$$



The Idea

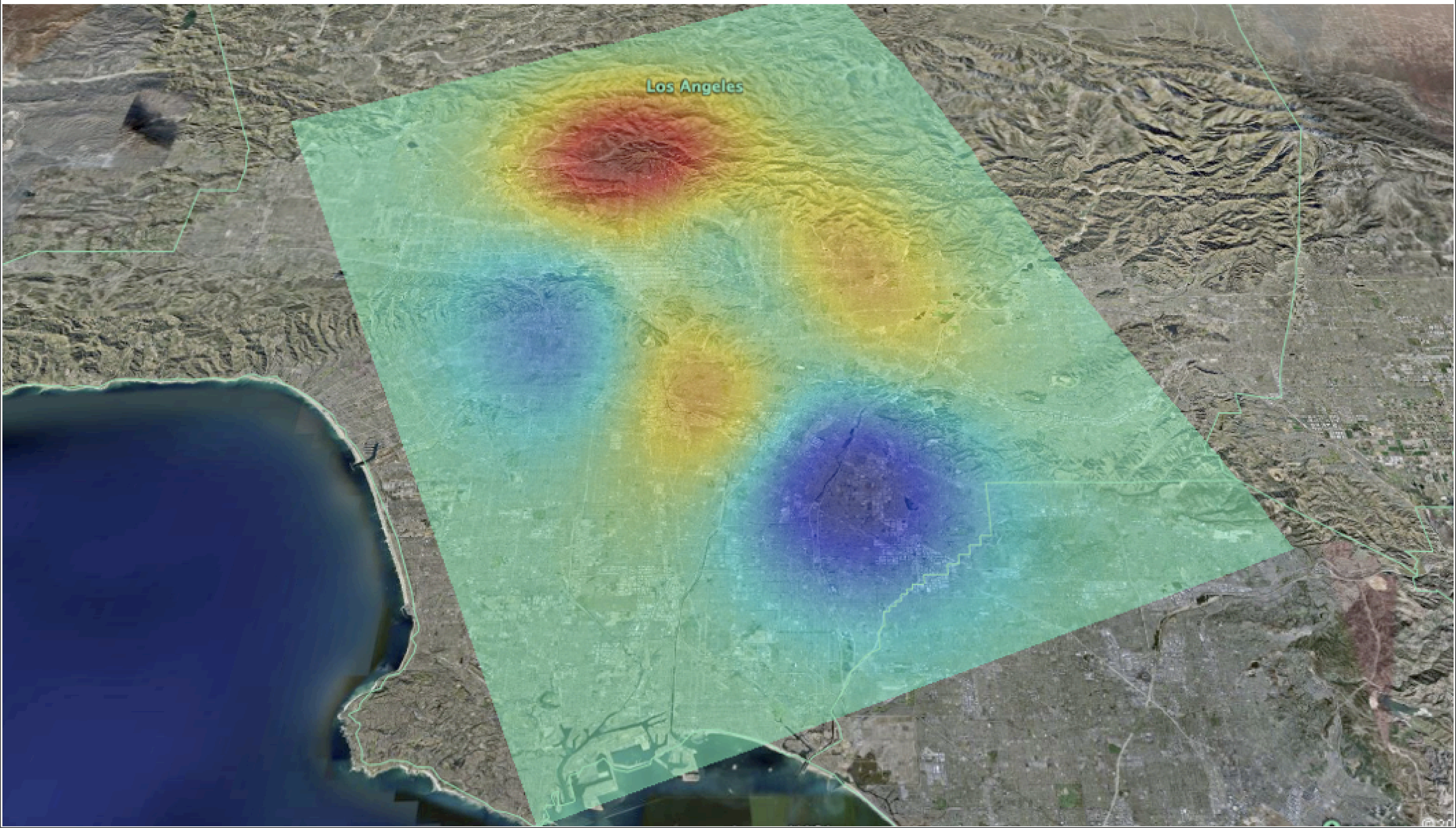
- In fact we compute the log of the price



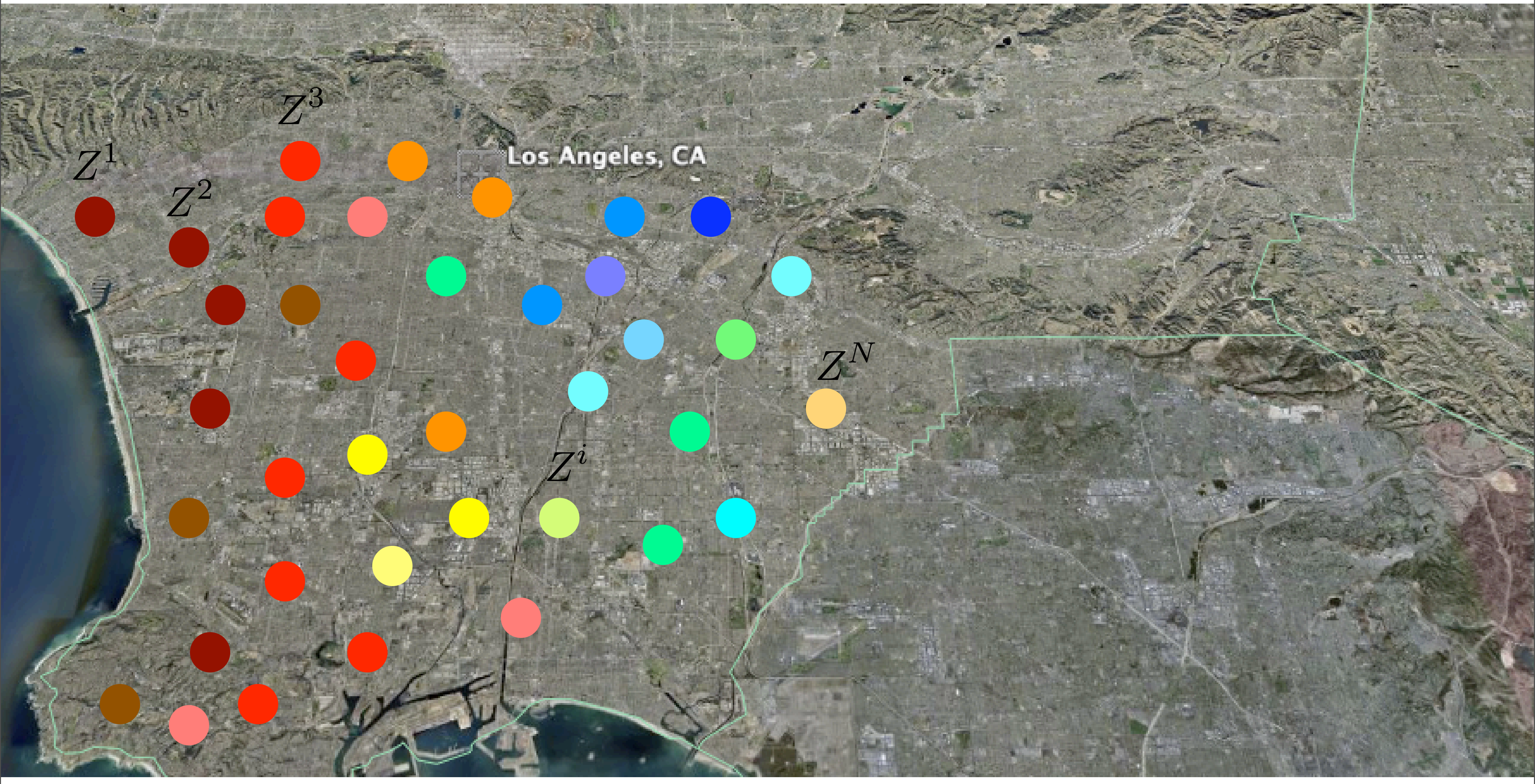
Our Contribution

- A novel technique for relational regression problems
- Allows relationships via the hidden variables
- Allows non-linear likelihood functions
- Propose efficient training and inference algorithm
- Apply it to the house price prediction problem

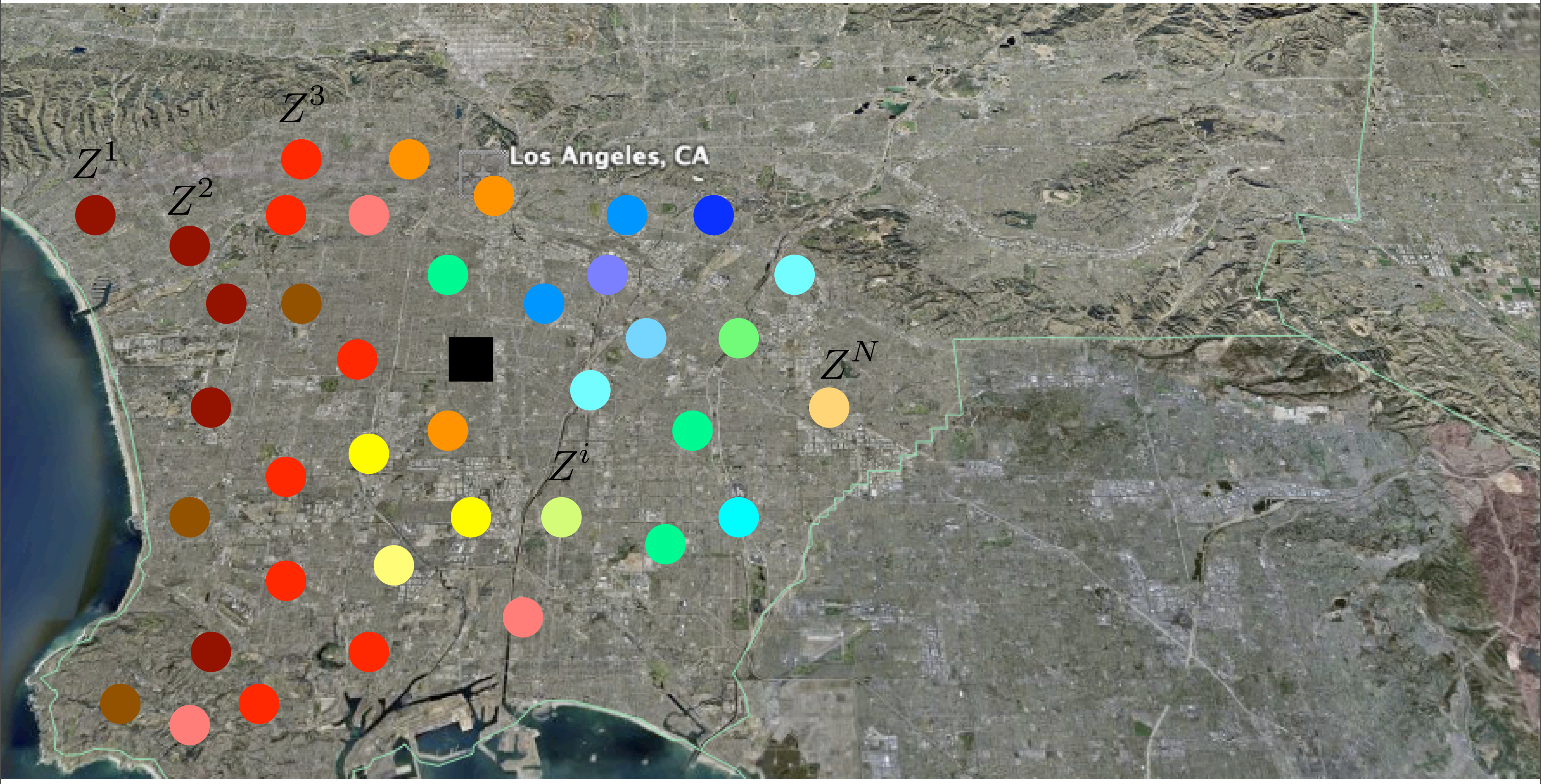
Estimating Desirabilities



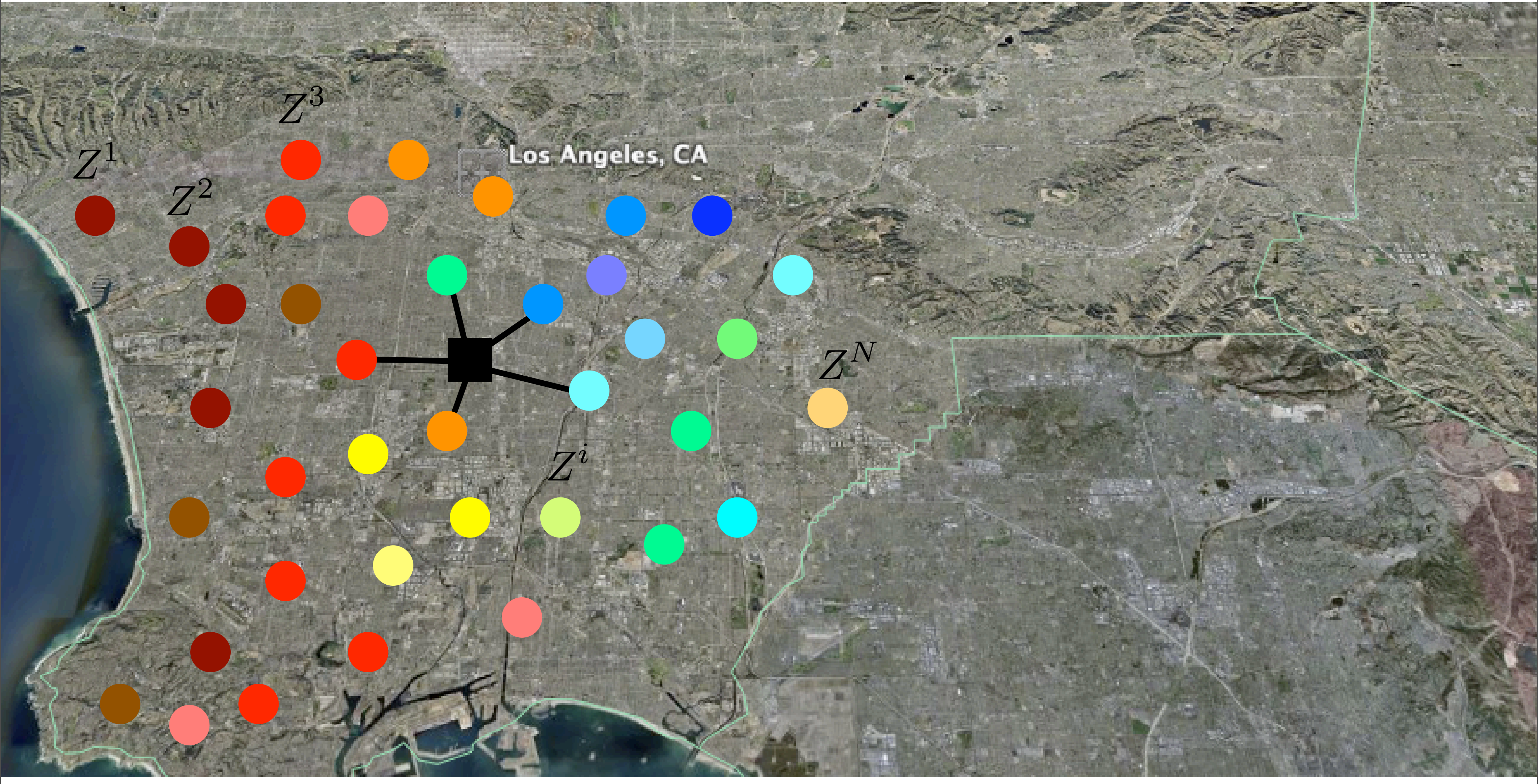
Estimating Desirabilities



Estimating Desirabilities



Estimating Desirabilities



Estimating Desirabilities

- Any form of smooth interpolation is good
- Kernel Interpolation

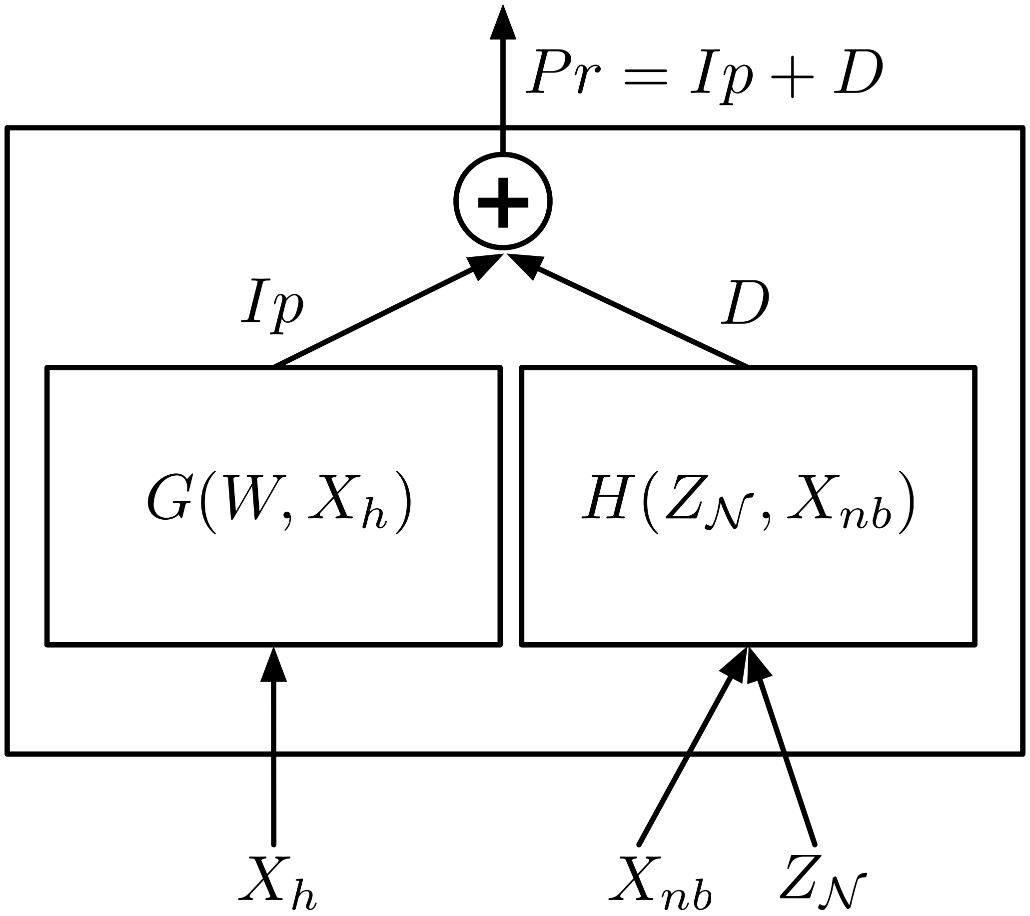
$$H(Z_{\mathcal{N}^i}, X^i_{nb}) = \sum_{j \in \mathcal{N}^i} Ker(X^i_{nb}, X^j_{nb}) Z^j$$

- Local Weighted Linear Regression

$$(\beta^*_{Z_{\mathcal{N}^i}}, \alpha^*_{Z_{\mathcal{N}^i}}) = \mathrm{argmin}_{\beta,\alpha} \sum_{j \in \mathcal{N}^i} (Z^j - (\beta + \alpha X^j))^2 Ker(X^i_{nb}, X^j_{nb})$$

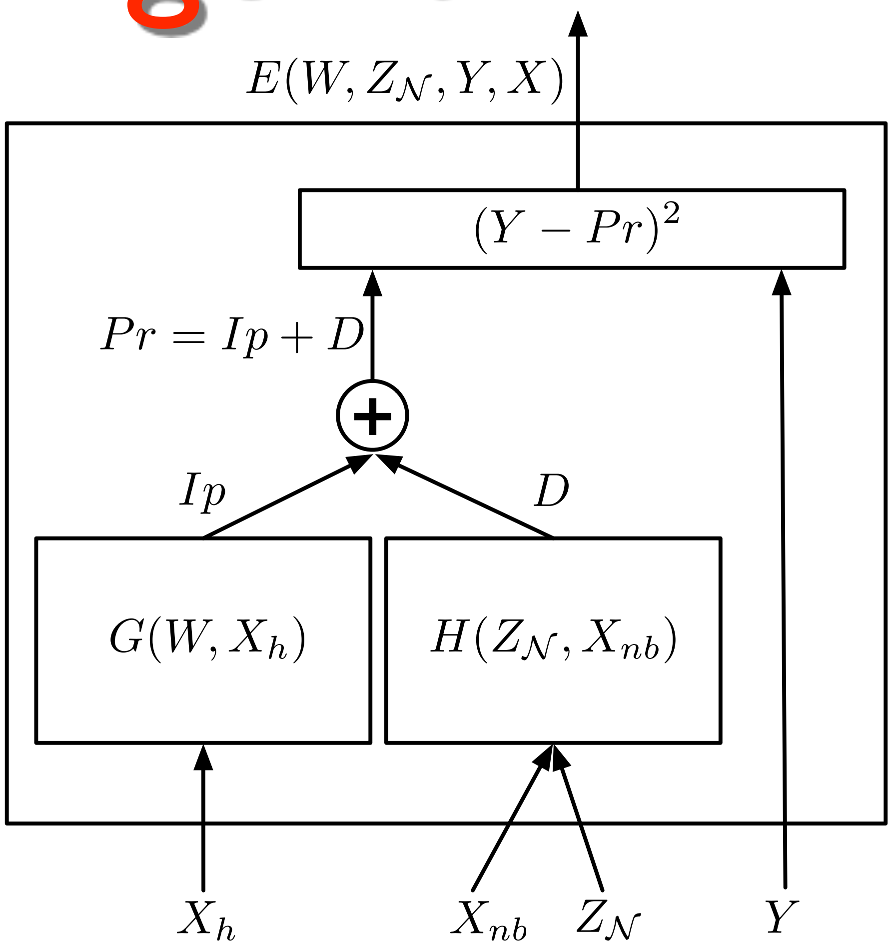
$$\begin{aligned} H(Z_{\mathcal{N}^i}, X^i_{nb}) &= \beta^*_{Z_{\mathcal{N}^i}} + \alpha^*_{Z_{\mathcal{N}_i}} X^i \\ &= \sum_k U^{ik} Z^k \end{aligned}$$

The Inference Algorithm



The Learning Algorithm

- Done by maximizing the likelihood of the data
- Achieved by minimizing the negative log-likelihood function wrt W, $\mathbf{Z}$
- Boils down to minimizing energy loss



$$\mathcal{L}(W, \mathbf{Z}) = \frac{1}{2} \sum_{i=1}^{N} (Y^i - (G(W, X_h^i) + H(Z_{\mathcal{N}^i}, X_{nb}^i))^2 + R(\mathbf{Z})$$

The Learning Algorithm

$$\mathcal{L}(W,\mathbf{Z}) = \frac{1}{2}\sum_{i=1}^{N}(Y^i - (G(W,X_h^i) + H(Z_{\mathcal{N}^i}, X_{nb}^i))^2 + \frac{r}{2}||Z||^2)$$

- Two phase: generalized EM type
- Phase I:
  - Keep W fixed and optimize with respect to $\mathbf{Z}$
  - The above loss reduces to

$$\mathcal{L}(\mathbf{Z}) = \frac{r}{2}||\mathbf{Z}||^2 + \frac{1}{2}\sum_{i=1}^{N}(Y^i - (G(W,X_h^i) + U^i \cdot \mathbf{Z}))^2$$

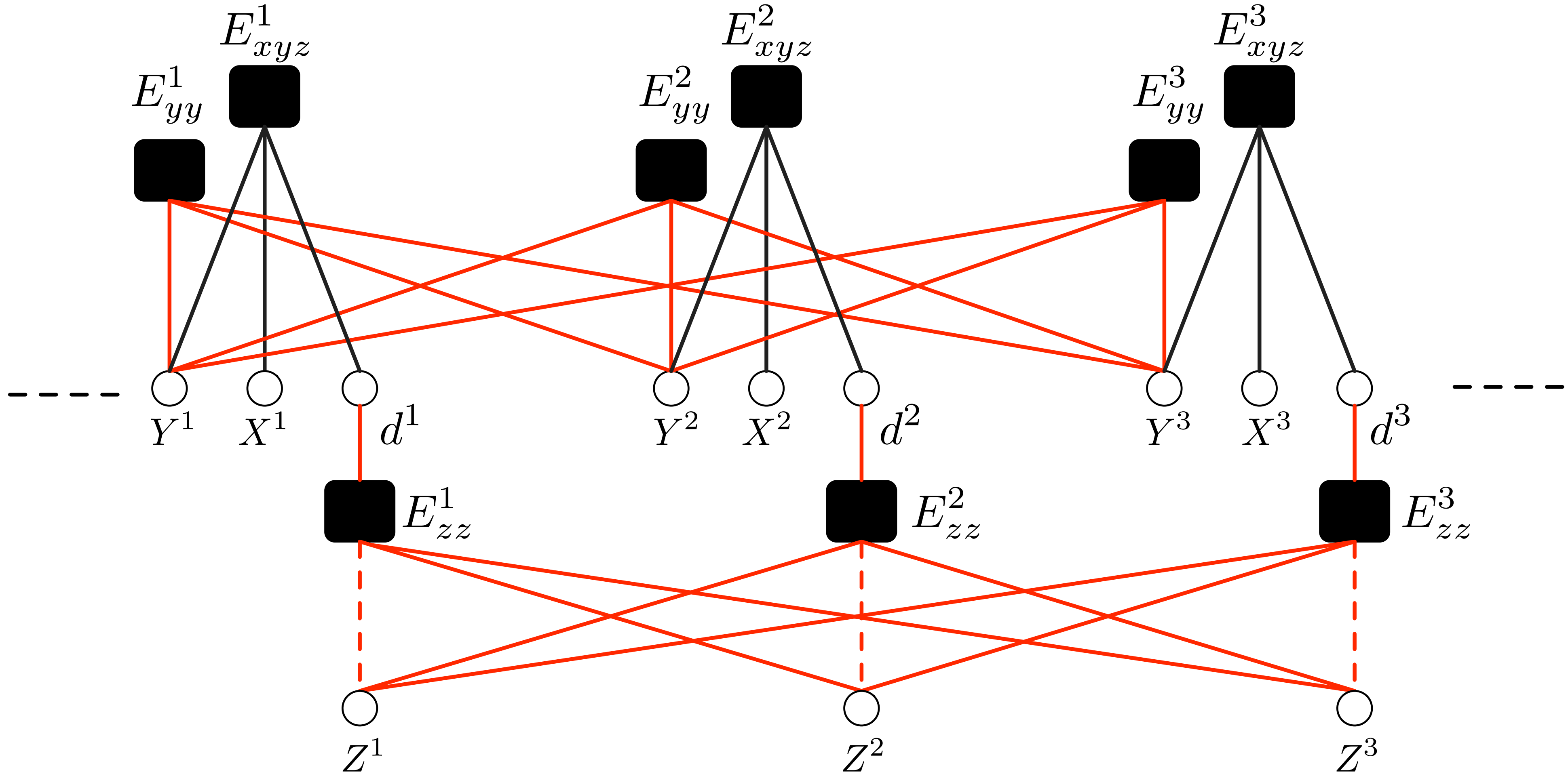
  - Sparse linear system: was solved using conjugate gradient

The Learning Algorithm

$$\mathcal{L}(W,\mathbf{Z}) = \frac{1}{2}\sum_{i=1}^{N}(Y^i - (G(W,X_h^i) + H(Z_{\mathcal{N}^i},X_{nb}^i))^2 + R(\mathbf{Z})$$

- Phase II:
  - Keep $\mathbf{Z}$ fixed and optimize with respect to W
  - The parameters are shared among samples
  - Neural network was optimized using simple gradient decent

The General Framework



$$E(W, \mathbf{Z}, \mathbf{Y}, \mathbf{X}) = \sum_{i=1}^{N} E^i_{xyz}(W_{xyz}, X^i, Y^i, D^i) + E^i_{yy}(W_{yy}, Y^i, \mathbf{Y}) + E^i_{zz}(W_{zz}, D^i, \mathbf{Z})$$

Experiments

- Dataset
  - Houses from Los Angeles Country transacted only in 2004
  - They span 1754 census tracts and 28 school district
  - A total of around 70,000 samples
  - We used a total of 19 features in X_h
    - living area, year built, # bedrooms, # bathrooms, pool, prior sale price, parking spaces, parking types, lot acerage, land value, improvement value, % improvement, new construction, foundation, roof type, heat type, site influence, and gps coordinates
  - We used 6 features as part of X_{nb}
    - median house hold income, average time of commute to work, proportion of units owner occupied, and academic performance index

Experiments

- Dataset
  - All variables containing any form of price/area/income information were mapped into log space
  - Non-numeric discrete variables were coded using a 1-of-K coding scheme
  - Only Single Family Residences were estimated
  - A total of 42025 complete labeled samples
  - Training set
    - 37822 (90%)
  - Test set
    - 4203 (10%)

Baseline Methods

- Nearest Neighbor
- Linear Regression
- Locally Weighted Linear Regression
- Fully Connected Neural Network

Results

- Absolute Relative Forecasting error is computed

$$error^i = \frac{|Pr^i - A^i|}{A^i}$$

Model Class	*Model*	*< 5%*	*< 10%*	*< 15%*
Non-Parametric	K - Nearest Neighbor	25.41	47.44	64.72
Parametric	Linear Regression	26.58	48.11	65.12
Non-Parametric	Locally Weighted Regression	32.98	58.46	75.21
Parametric	Fully Connected Neural Network	33.79	60.55	76.47
Hybrid	Relational Factor Graph	39.47	65.76	81.04

Results

Model Class	*Model*	*< 5%*	*< 10%*	*< 15%*
Non-Parametric	K - Nearest Neighbor	25.41	47.44	64.72
Parametric	Linear Regression	26.58	48.11	65.12
Non-Parametric	Locally Weighted Regression	32.98	58.46	75.21
Parametric	Fully Connected Neural Network	33.79	60.55	76.47
Hybrid	Relational Factor Graph	39.47	65.76	81.04

Results

Model Class	*Model*	*< 5%*	*< 10%*	*< 15%*
Non-Parametric	K - Nearest Neighbor	25.41	47.44	64.72
Parametric	Linear Regression	26.58	48.11	65.12
Non-Parametric	Locally Weighted Regression	32.98	58.46	75.21
Parametric	Fully Connected Neural Network	33.79	60.55	76.47
Hybrid	Relational Factor Graph	39.47	65.76	81.04

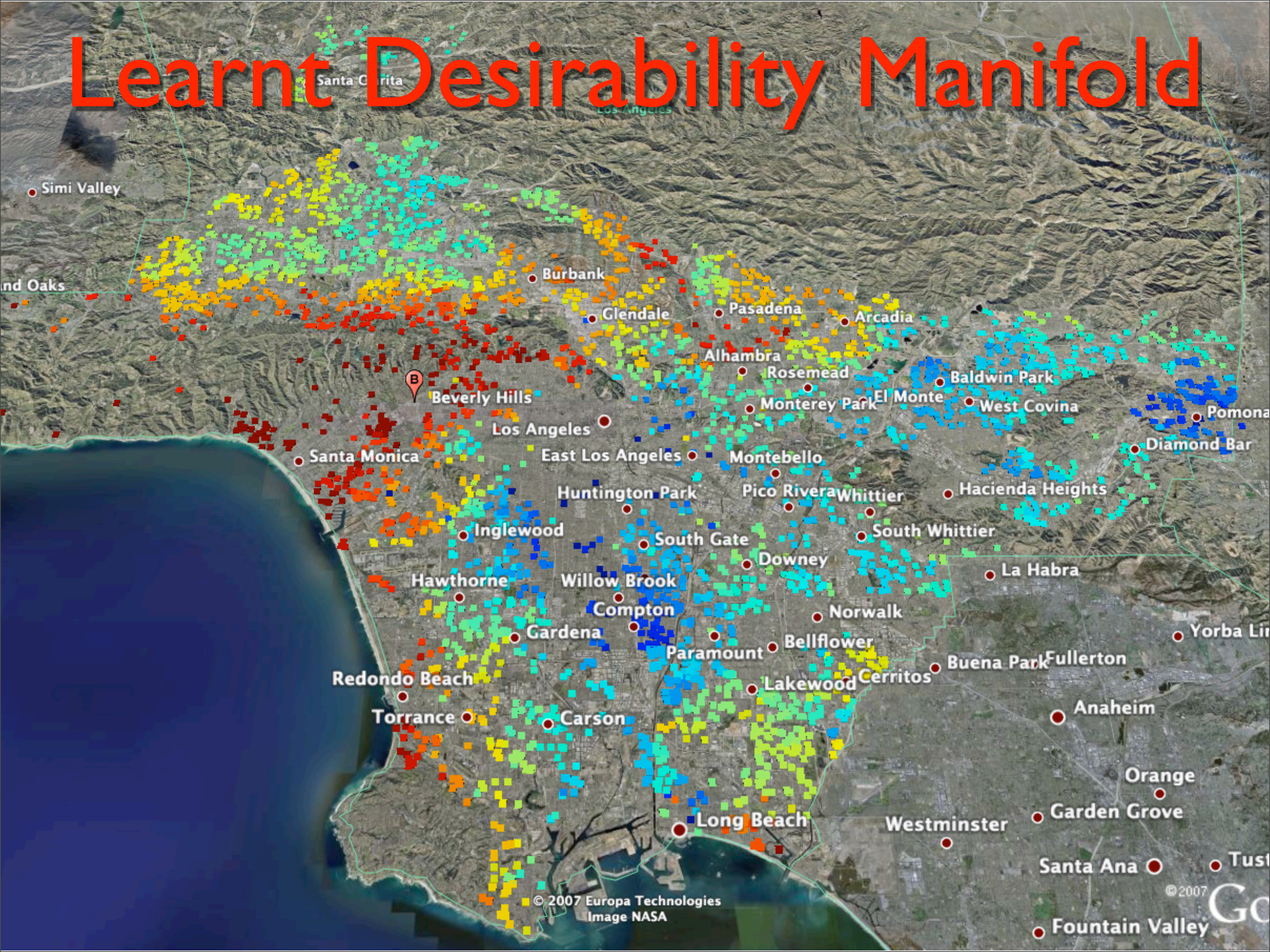
Results

Model Class	*Model*	*< 5%*	*< 10%*	*< 15%*
Non-Parametric	K - Nearest Neighbor	25.41	47.44	64.72
Parametric	Linear Regression	26.58	48.11	65.12
Non-Parametric	Locally Weighted Regression	32.98	58.46	75.21
Parametric	Fully Connected Neural Network	33.79	60.55	76.47
Hybrid	Relational Factor Graph	39.47	65.76	81.04

Results

Model Class	*Model*	*< 5%*	*< 10%*	*< 15%*
Non-Parametric	K - Nearest Neighbor	25.41	47.44	64.72
Parametric	Linear Regression	26.58	48.11	65.12
Non-Parametric	Locally Weighted Regression	32.98	58.46	75.21
Parametric	Fully Connected Neural Network	33.79	60.55	76.47
Hybrid	Relational Factor Graph	39.47	65.76	81.04

Learnt Desirability Manifold



Thank You Very Much!!!

Real Estate Price Prediction

- Direct dependencies between Y is not captured
- First factor
  - Non-relational: captures dependencies between individual variables and the price
    $$E^i_{xyz}(Y^i, X^i, D^i) = (Y^i - (G(W_{xyz}, X^i_h) + D^i))^2$$
- Second factor
  - Relational: captures the influence on the price of a house from other (related houses) via the hidden variables
    $$E^i_{zz}(D^i, H(X^i_{nb}, Z_{\mathcal{N}^i})) = (D^i - H(X^i_{nb}, Z_{\mathcal{N}^i}))^2$$
  - $H(X^i_{nb}, Z_{\mathcal{N}^i})$ non-parametric estimate of desirability of the location of the house, obtained from related houses

Real Estate Price Prediction

- $H(X_{nb}^i, Z_{\mathcal{N}^i})$ could take any smooth form
- Kernel Interpolation

$$H(X_{nb}^i, Z_{\mathcal{N}^i}) = \sum_{j \in \mathcal{N}^i} Ker(X_{nb}, X_{nb}^j) Z^j$$

- Local Weighted Linear Regression

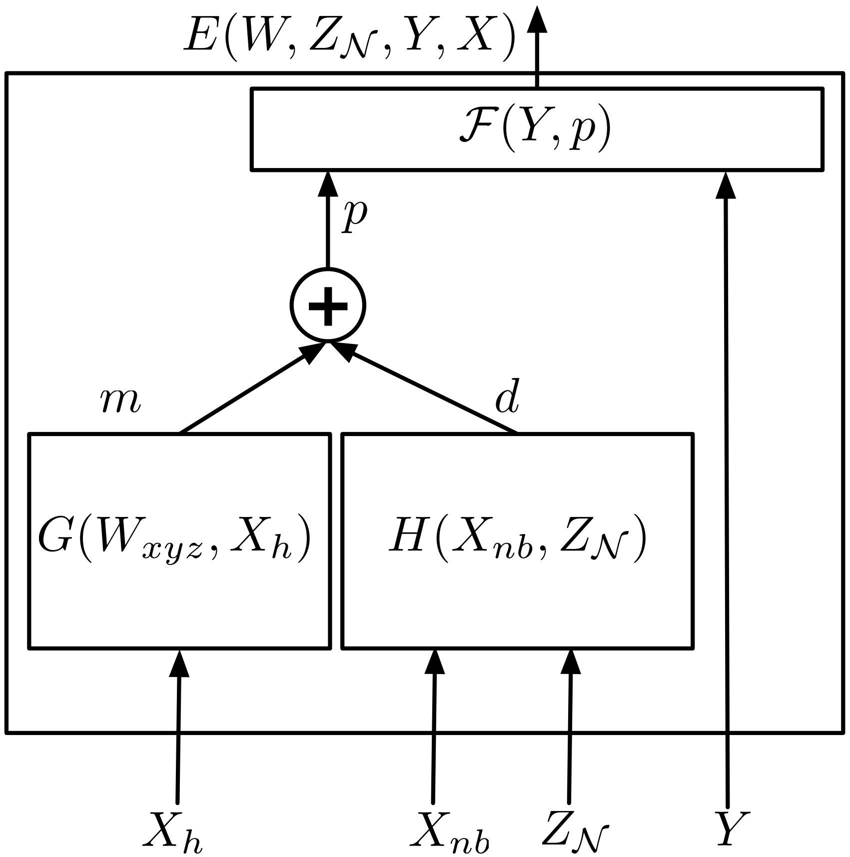
$$(\beta^*, \alpha^*) = \arg\min_{\beta,\alpha} \sum_{j \in \mathcal{N}^i} (Z^j - (\beta + \alpha X^j))^2 Ker(X^i, X^j)$$

$$\begin{aligned} H(X_{nb}^i, Z_{\mathcal{N}^i}) &= \beta^* + \alpha^* X \\ &= \sum_{j \in \mathcal{N}^i} a^j Z^j \end{aligned}$$

Real Estate Price Prediction

- The total energy associated with a single sample is

$$E^i_{xyz}(Y^i, X^i, D^i) + E^i_{zz}(D^i, H(X^i_{nb}, Z_{\mathcal{N}^i}) = (Y^i - (G(W_{xyz}, X^i_h) + D^i))^2 + (D^i - H(X^i_{nb}, Z_{\mathcal{N}^i}))^2$$



$$E^i(Y^i, X^i) = (Y^i - (G(W_{xyz}, X^i_h) + H(X^i_{nb}, Z_{\mathcal{N}^i}))^2$$

Real Estate Price Prediction

- The factor graph is