Automatic Text Evaluation through the Lens of Wasserstein Barycenters

Oral Presentation at EMNLP 2021

Pierre Colombo, Guillaume Staerman, Chloé Clavel, Pablo Piantanida

TELECOM Paris

IP PARIS

L2S Laboratoire Signaux & Systèmes

IBM

Importance of Evaluation of NLG

Importance of Evaluation of NLG

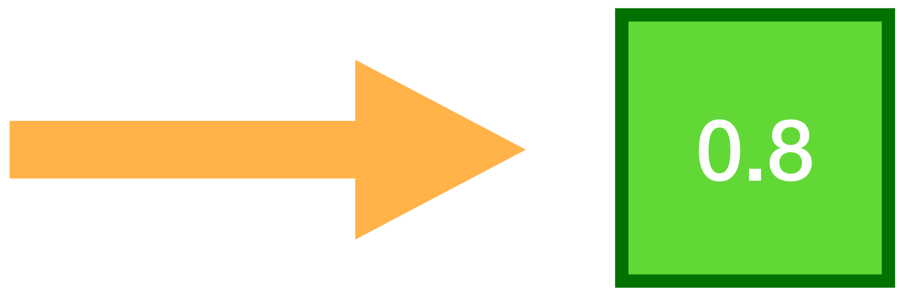
What is automatic evaluation?

Importance of Evaluation of NLG

What is automatic evaluation?

R: The weather is cold today.

C: It is freezing today



Importance of Evaluation of NLG

What is automatic evaluation?

R: The weather is cold today.

C: It is freezing today

→ 0.8

R: I like those cats.

C: It is freezing today

→ 0.1

Importance of Evaluation of NLG

What is automatic evaluation?

R: The weather is cold today.

C: It is freezing today

0.8

R: I like those cats.

C: It is freezing today

0.1

Why is automatic evaluation popular ?

Importance of Evaluation of NLG

What is automatic evaluation?

R: The weather is cold today.
C: It is freezing today → 0.8

R: I like those cats.
C: It is freezing today → 0.1

Why is automatic evaluation popular ?

1. Cheap: compared to human evaluation

Importance of Evaluation of NLG

What is automatic evaluation?

R: The weather is cold today.

C: It is freezing today

0.8

R: I like those cats.

C: It is freezing today

0.1

Why is automatic evaluation popular ?

1. Cheap: compared to human evaluation
2. Fast: you can label “instantaneously”

Importance of Evaluation of NLG

What is automatic evaluation?

R: The weather is cold today.

C: It is freezing today

→ 0.8

R: I like those cats.

C: It is freezing today

→ 0.1

Why is automatic evaluation popular ?

1. Cheap: compared to human evaluation
2. Fast: you can label “instantaneously”
3. Reproductible: two sentence always get the same score

Importance of Evaluation of NLG

What is automatic evaluation?

R: The weather is cold today.

C: It is freezing today

0.8

R: I like those cats.

C: It is freezing today

0.1

Why is automatic evaluation popular ?

1. Cheap: compared to human evaluation
2. Fast: you can label “instantaneously”
3. Reproductible: two sentence always get the same score
4. Easy to use: don’t need to train anotators, ask the right questions

Importance of Evaluation of NLG

What is automatic evaluation?

R: The weather is cold today.
C: It is freezing today
→ 0.8

R: I like those cats.
C: It is freezing today
→ 0.1

Why is automatic evaluation popular ?

1. Cheap: compared to human evaluation
2. Fast: you can label "instantaneously"
3. Reproductible: two sentence always get the same score
4. Easy to use: don't need to train anotators, ask the right questions

Why do we need evaluation of NLG?

Importance of Evaluation of NLG

What is automatic evaluation?

R: The weather is cold today.
C: It is freezing today
→ 0.8

R: I like those cats.
C: It is freezing today
→ 0.1

Why is automatic evaluation popular ?

1. Cheap: compared to human evaluation
2. Fast: you can label “instantaneously”
3. Reproductible: two sentence always get the same score
4. Easy to use: don’t need to train anotators, ask the right questions

Why do we need evaluation of NLG?

1. Debug NLG systems without annotators

Importance of Evaluation of NLG

What is automatic evaluation?

R: The weather is cold today.

C: It is freezing today

0.8

R: I like those cats.

C: It is freezing today

0.1

Why is automatic evaluation popular ?

1. Cheap: compared to human evaluation
2. Fast: you can label "instantaneously"
3. Reproductible: two sentence always get the same score
4. Easy to use: don't need to train anotators, ask the right questions

Why do we need evaluation of NLG?

1. Debug NLG systems without annotators
2. Improve learning of systems by deriving new loss

Importance of Evaluation of NLG

What is automatic evaluation?

R: The weather is cold today.

C: It is freezing today

0.8

R: I like those cats.

C: It is freezing today

0.1

Why is automatic evaluation popular ?

1. Cheap: compared to human evaluation
2. Fast: you can label “instantaneously”
3. Reproductible: two sentence always get the same score
4. Easy to use: don’t need to train anotators, ask the right questions

Why do we need evaluation of NLG?

1. Debug NLG systems without annotators
2. Improve learning of systems by deriving new loss
3. Compare different systems

Existing Methods

Existing Methods

Edit Based

Existing Methods

Edit Based

Operations

- Insertion (I)
- Deletion (D)
- Substitution (S).

tailor -> sailor (S)
sailor -> sailir (S)
sailir -> sailin (S)
sailin -> sailing (I)

Existing Methods

Edit Based

Operations

- Insertion (I)
- Deletion (D)
- Substitution (S).

tailor -> sailor (S)
sailor -> sailir (S)
sailir -> sailin (S)
sailin_ -> sailing (I)

Distance is 4 !

Existing Methods

Edit Based

Operations

- Insertion (I)
- Deletion (D)
- Substitution (S).

tailor -> sailor (S)
sailor -> sailir (S)
sailir -> sailin (S)
sailin -> sailing (I)

Distance is 4 !

N-gram Based

Existing Methods

Edit Based

Operations

- Insertion (I)
- Deletion (D)
- Substitution (S).

tailor -> sailor (S)
sailor -> sailir (S)
sailir -> sailin (S)
sailin -> sailing (I)

Distance is 4 !

N-gram Based

C : I like these very nice pies !

R : I like those cakes !

Unigrams

C : I like these very nice pies !

R : I like those cakes !

Existing Methods

Edit Based

Operations

- Insertion (I)
- Deletion (D)
- Substitution (S).

tailor -> sailor (S)
sailor -> sailir (S)
sailir -> sailin (S)
sailin_ -> sailing (I)

Distance is 4 !

N-gram Based

C : I like these very nice pies !

R : I like those cakes !

Unigrams

C : $\boxed{\text{I}}$ $\boxed{\text{like}}$ these very nice pies $\boxed{\text{!}}$

R : $\boxed{\text{I}}$ $\boxed{\text{like}}$ those cakes $\boxed{\text{!}}$

Bigrams

C : $\boxed{\text{I like}}$ these very nice pies !

R : $\boxed{\text{I like}}$ those cakes !

Existing Methods

Edit Based

Operations

- Insertion (I)
- Deletion (D)
- Substitution (S).

tailor -> sailor (S)
sailor -> sailir (S)
sailir -> sailin (S)
sailin_ -> sailing (I)

Distance is 4 !

N-gram Based

C : I like these very nice pies !

R : I like those cakes !

Unigrams

C : $\boxed{\text{I}}\boxed{\text{like}}$ these very nice pies $\boxed{\text{!}}$

R : $\boxed{\text{I}}\boxed{\text{like}}$ those cakes $\boxed{\text{!}}$

Bigrams

C : $\boxed{\text{I like}}$ these very nice pies !

R : $\boxed{\text{I like}}$ those cakes !

Embedding Based

Existing Methods

Edit Based

Operations

- Insertion (I)
- Deletion (D)
- Substitution (S).

tailor -> sailor (S)
sailor -> sailir (S)
sailir -> sailin (S)
sailin_ -> sailing (I)

Distance is 4 !

N-gram Based

C : I like these very nice pies !

R : I like those cakes !

Unigrams

C : $\boxed{\text{I}}$ $\boxed{\text{like}}$ these very nice pies $\boxed{\text{!}}$

R : $\boxed{\text{I}}$ $\boxed{\text{like}}$ those cakes $\boxed{\text{!}}$

Bigrams

C : $\boxed{\text{I like}}$ these very nice pies !

R : $\boxed{\text{I like}}$ those cakes !

Embedding Based

Word Mover distance

BertScore

MoverScore

Sentence Mover

Existing Methods

Edit Based

Operations

- Insertion (I)
- Deletion (D)
- Substitution (S).

tailor -> sailor (S)
sailor -> sailir (S)
sailir -> sailin (S)
sailin -> sailing (I)

Distance is 4 !

N-gram Based

C : I like these very nice pies !

R : I like those cakes !

Unigrams

C : I like these very nice pies !

R : I like those cakes !

Bigrams

C : I like these very nice pies !

R : I like those cakes !

Embedding Based

Word Mover distance

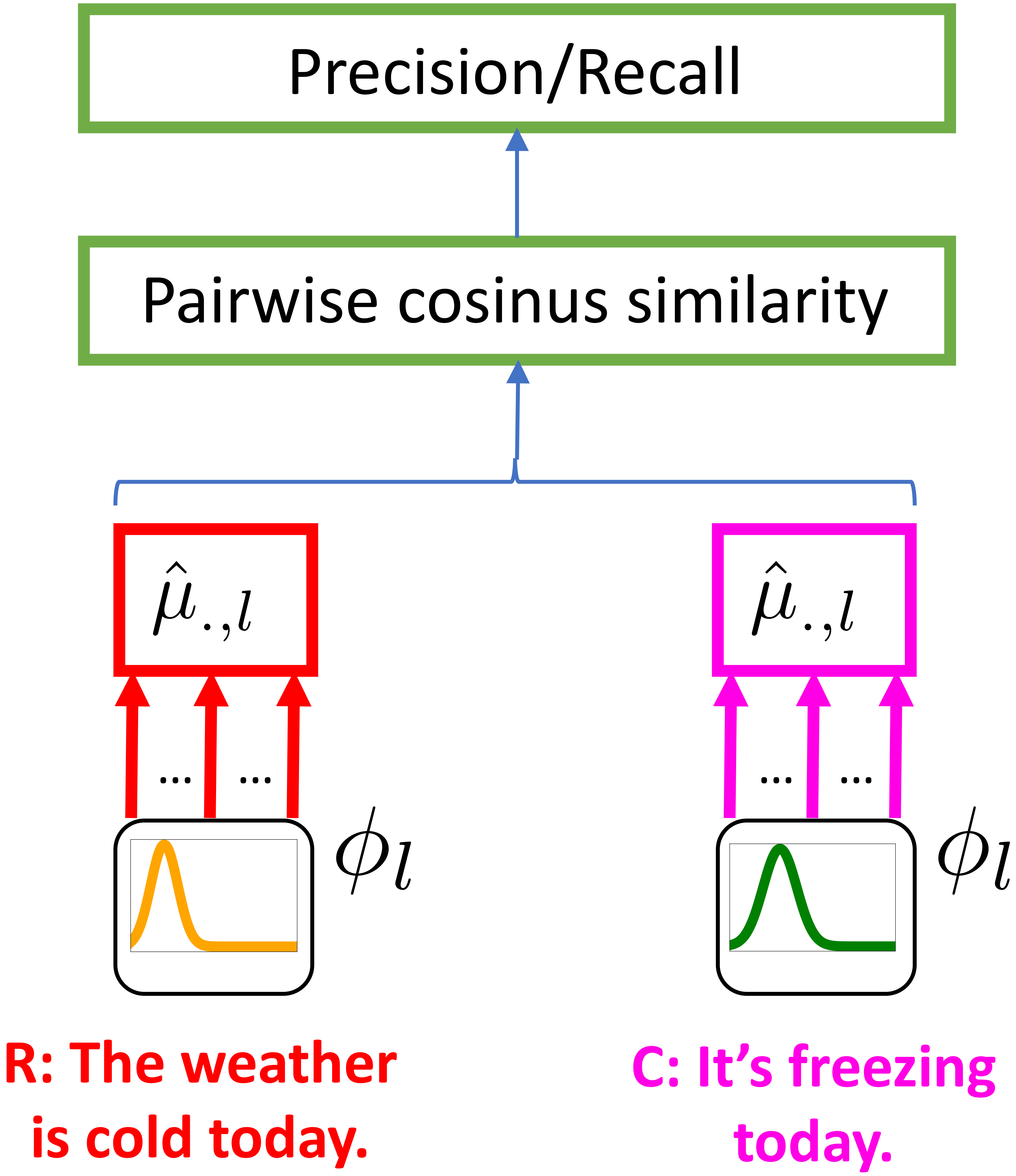
BertScore

MoverScore

Sentence Mover

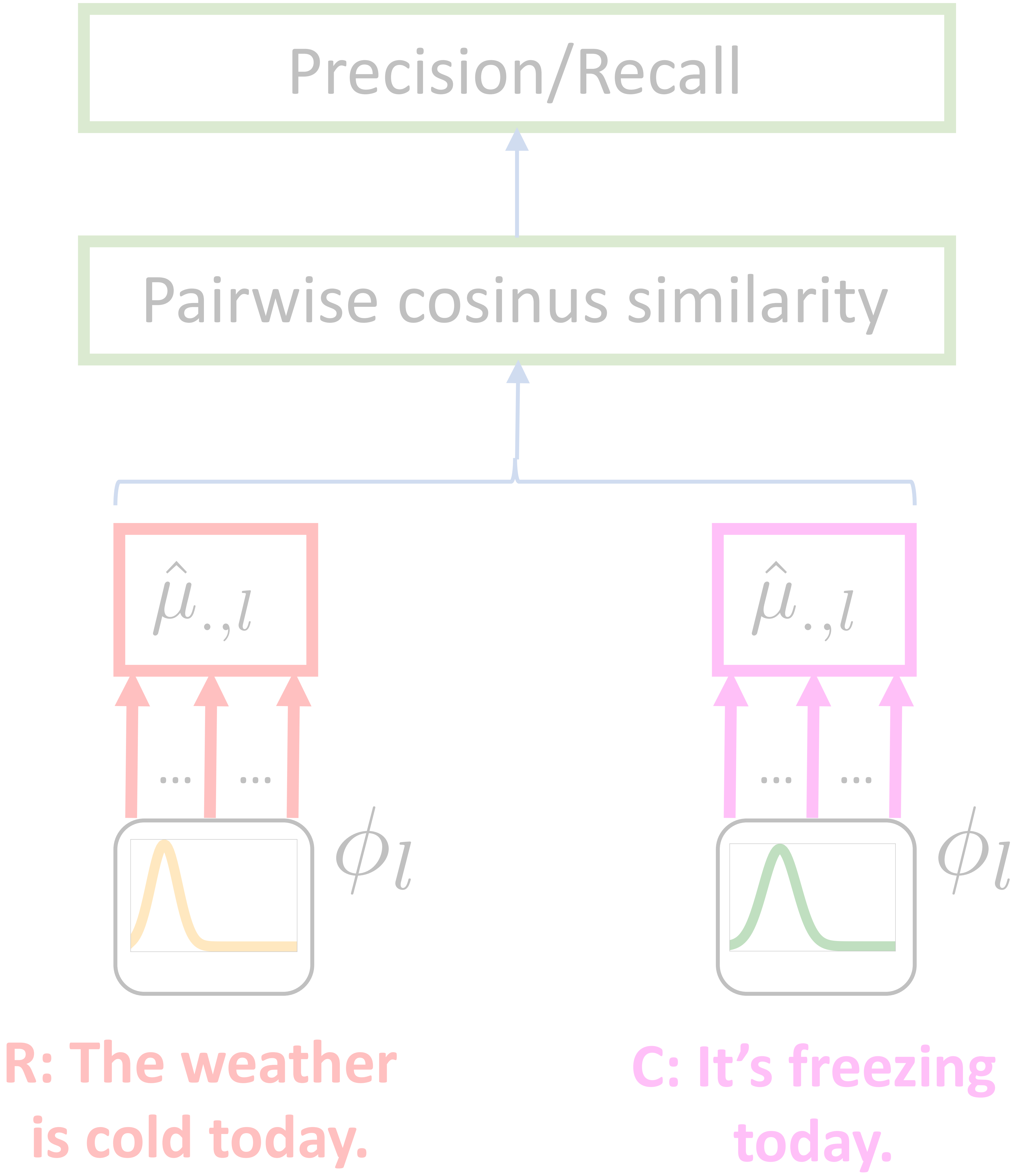
BertScore

BertScore



Precision/Recall
Pairwise cosinus similarity
$\hat{\mu}_{.,l}$
$\hat{\mu}_{.,l}$
...
...
...
...
ϕ_l
ϕ_l
R: The weather is cold today.
C: It's freezing today.

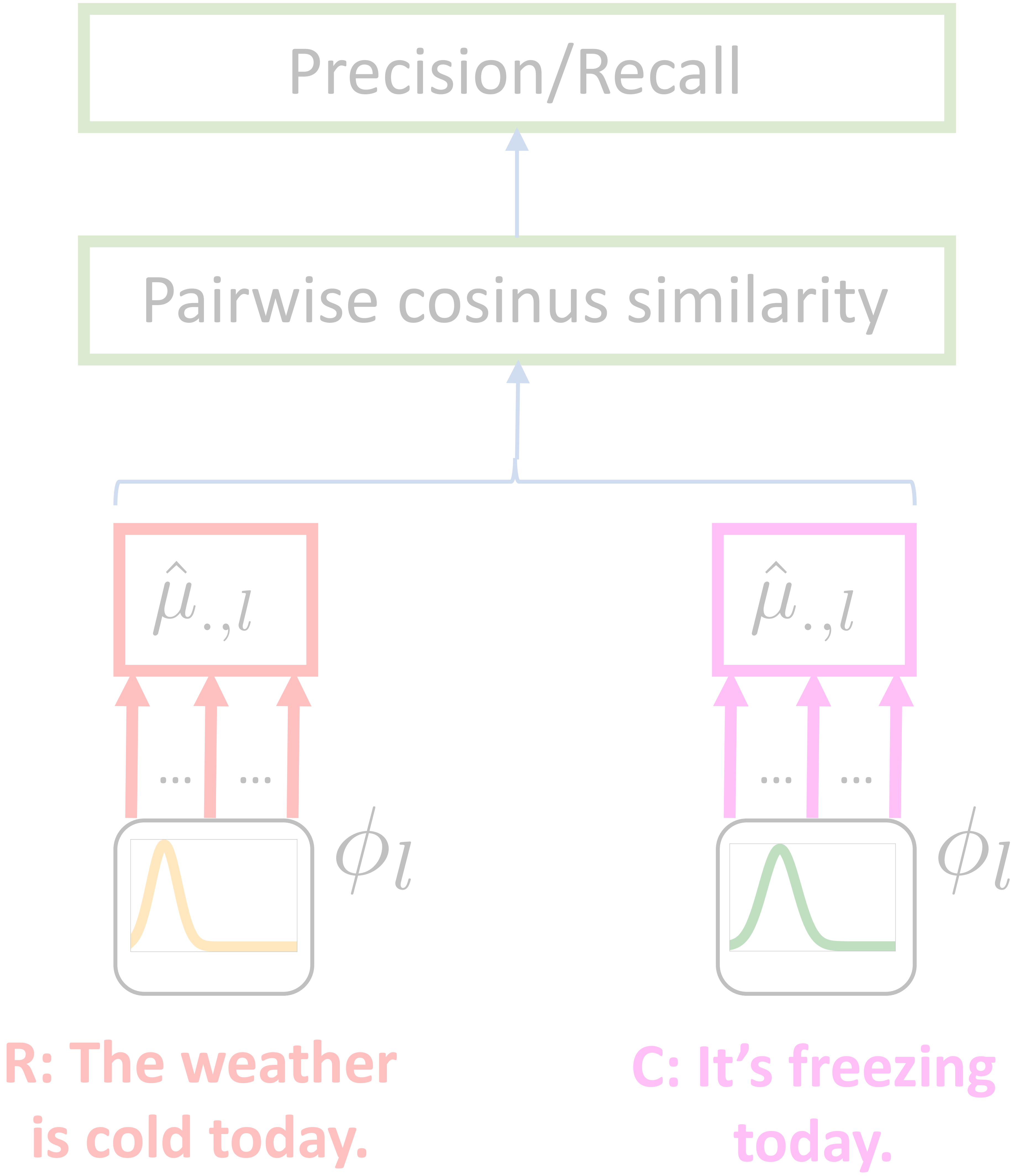
BertScore



Advantage

1. Deal with paraphrases
2. Include "semantic"

BertScore



Advantage

1. Deal with paraphrases
2. Include "semantic"

Limitations

1. Use only one layer
2. Use arbitrary sequence of operation

Optimal Transport

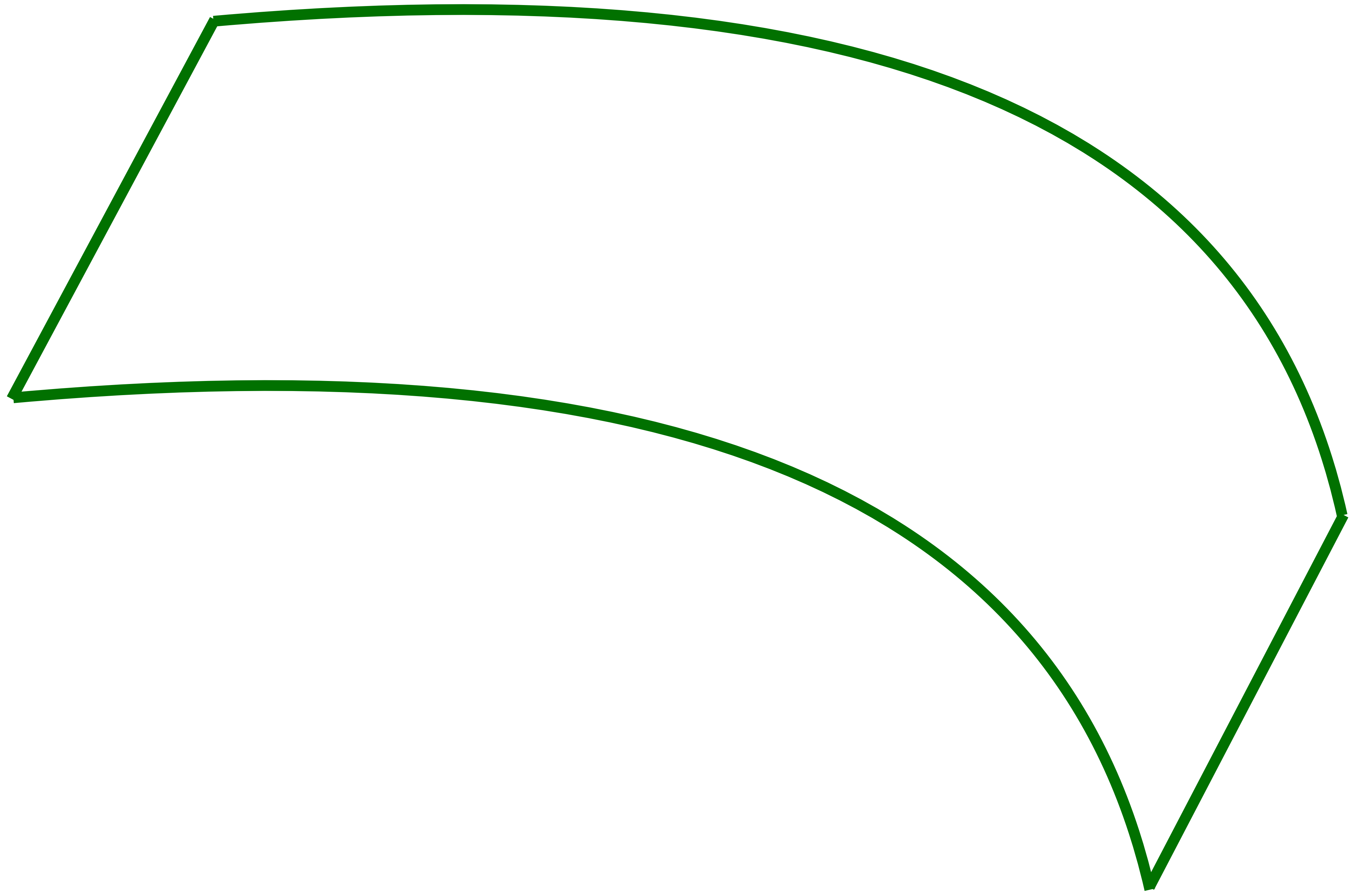
Optimal Transport

Goal Compute distance between probability measures (μ,ν)

Optimal Transport

Goal Compute distance between probability measures (μ,ν)

Input Discret Measures

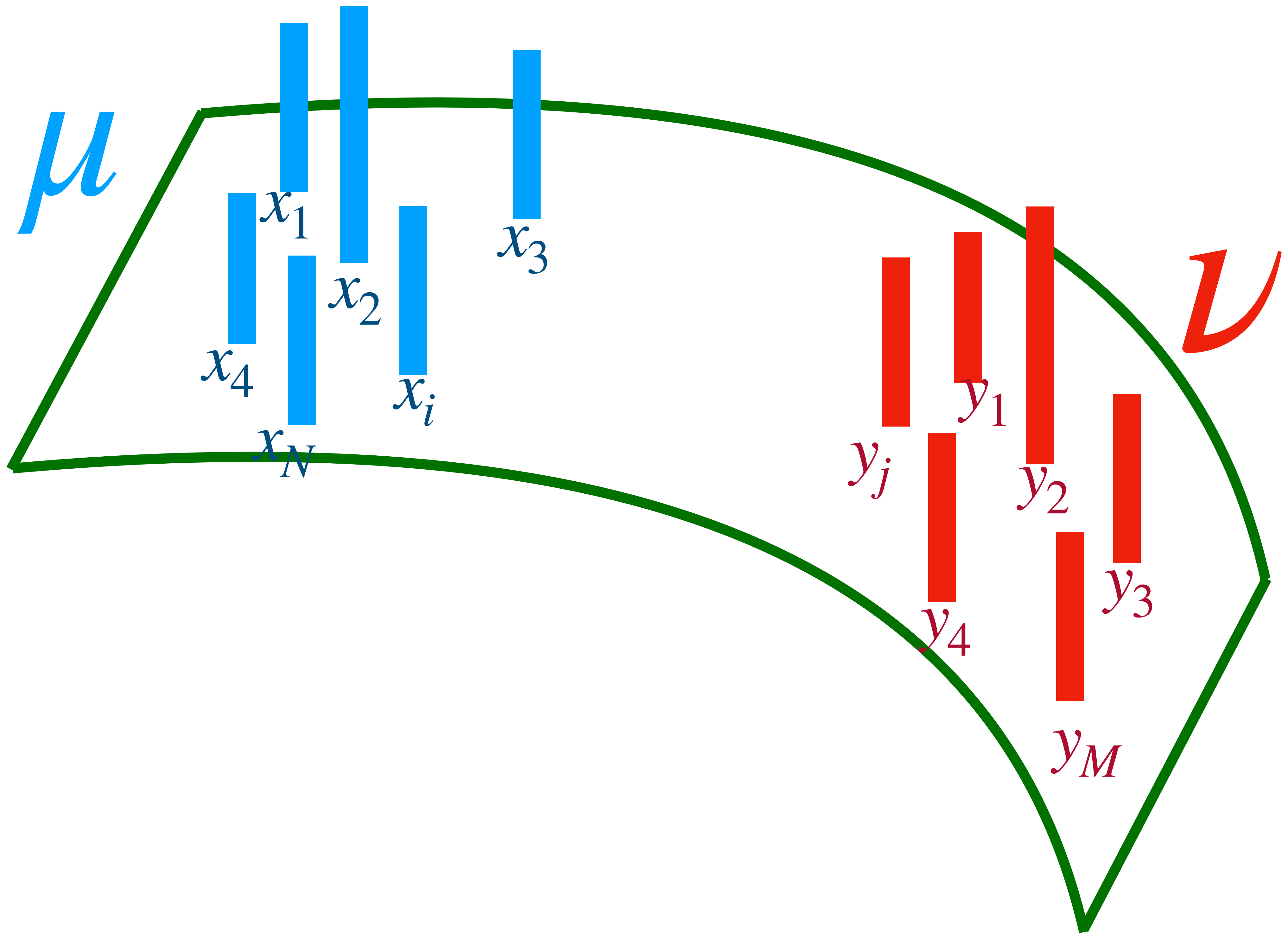


Optimal Transport

Goal Compute distance between probability measures (μ,ν)

Input Discret Measures

$$\nu = \sum_{j=1}^{M} \beta_j \delta_{x_j} \qquad \mu = \sum_{i=1}^{N} \alpha_i \delta_{x_i}$$

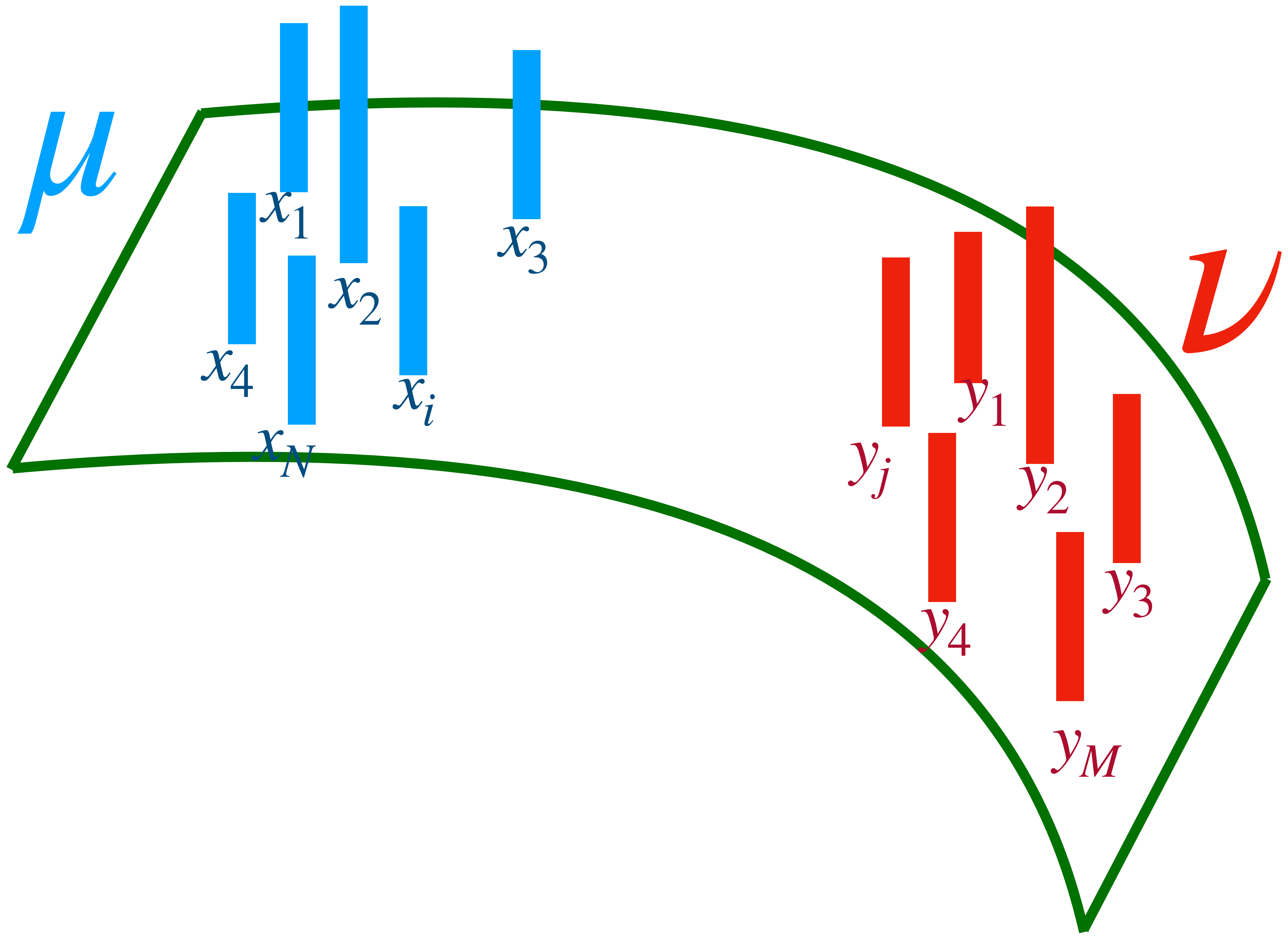


Optimal Transport

Goal Compute distance between probability measures (μ,ν)

Input Discret Measures

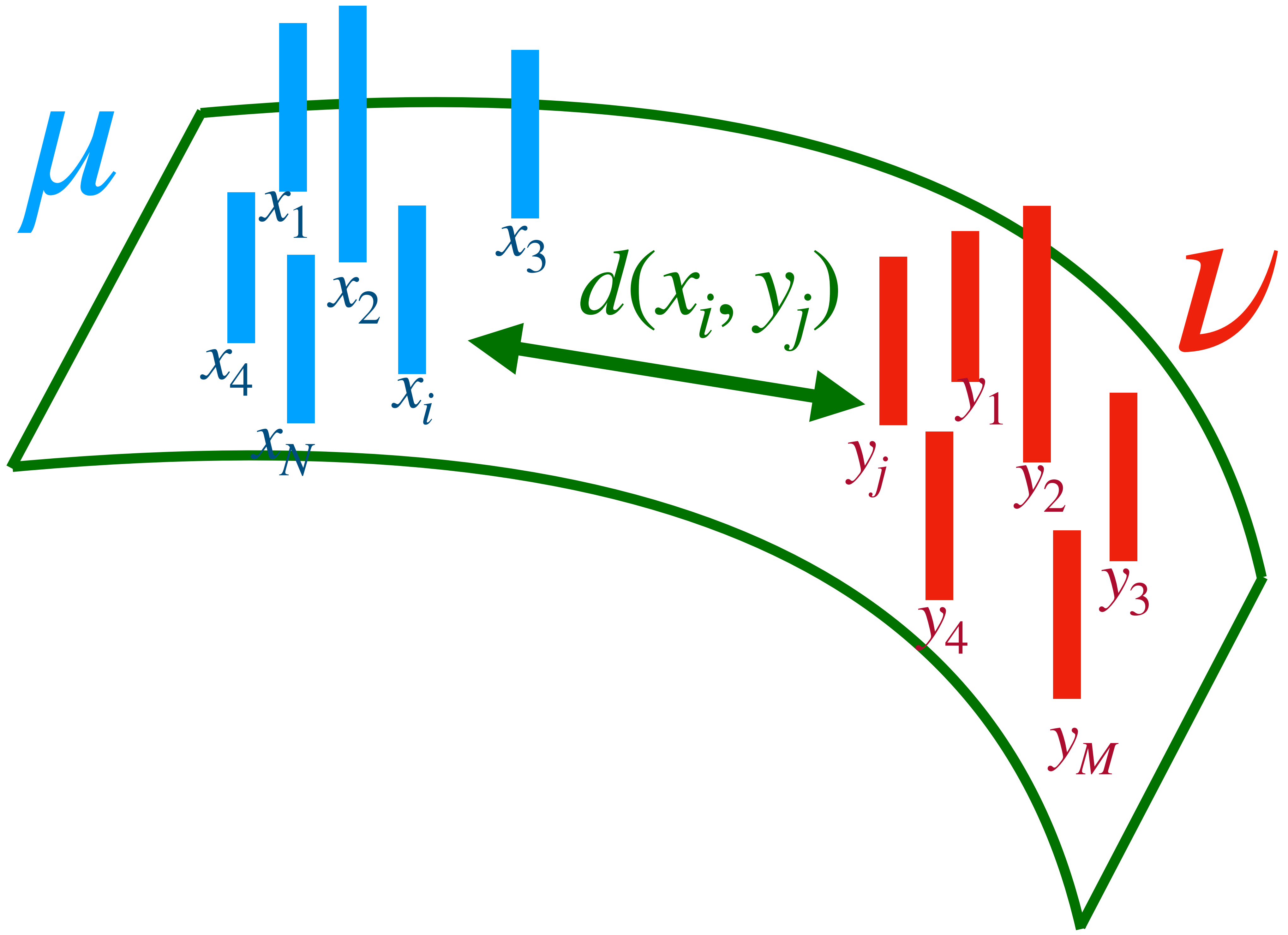
$$\nu = \sum_{j=1}^{M} \beta_j \delta_{x_j} \qquad \mu = \sum_{i=1}^{N} \alpha_i \delta_{x_i}$$



Cost Matrix

Optimal Transport

Goal Compute distance between probability measures (μ,ν)



Input Discret Measures

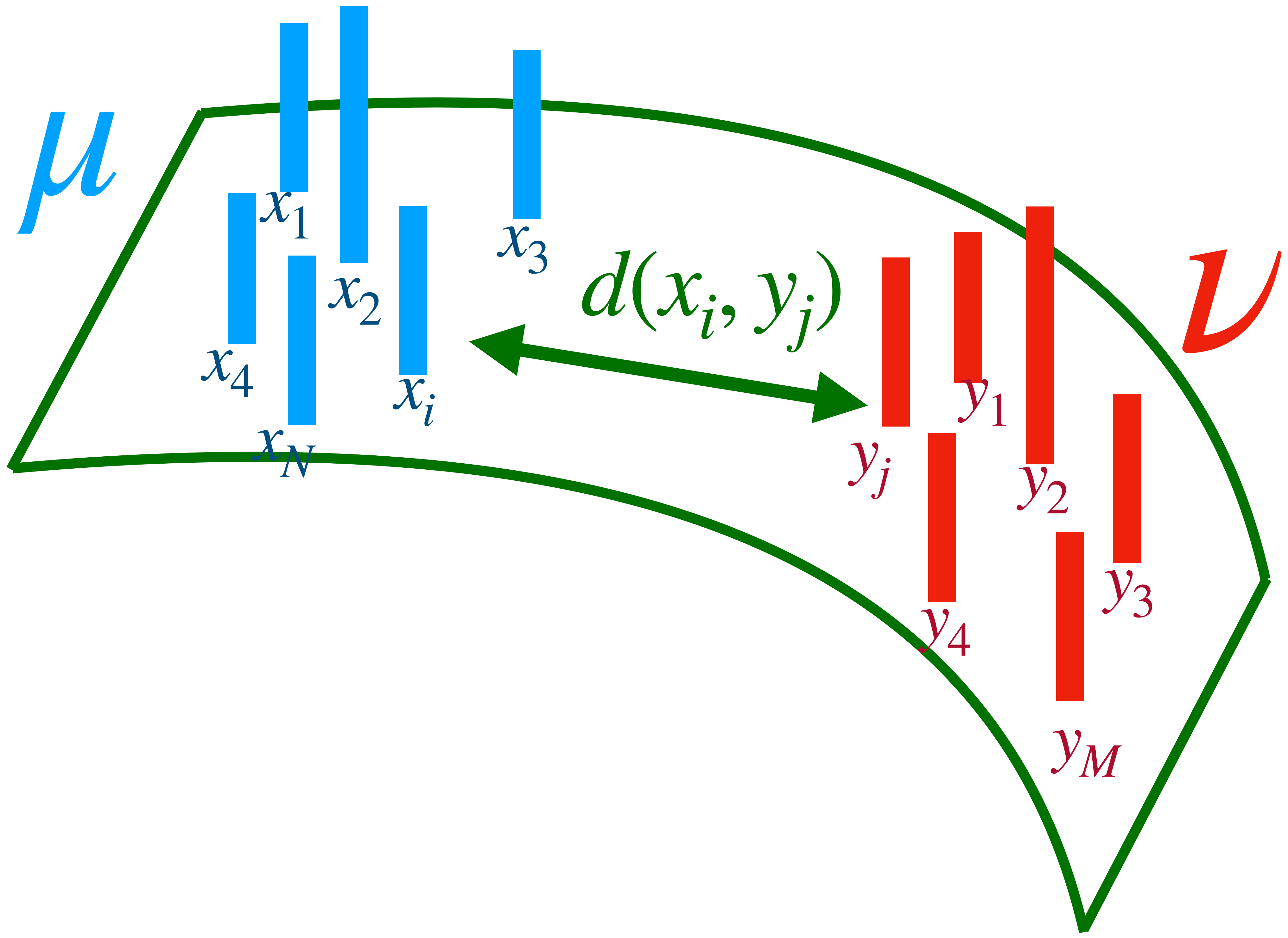
$$\nu = \sum_{j=1}^{M} \beta_j \delta_{x_j} \qquad \mu = \sum_{i=1}^{N} \alpha_i \delta_{x_i}$$

Cost Matrix

$$C = \begin{pmatrix} d(x_1, y_1) & \cdots & d(x_1, y_M) \\ \cdots & \cdots & \cdots \\ d(x_N, y_1) & \cdots & d(x_N, y_M) \end{pmatrix}$$

Optimal Transport

Goal **Compute distance between probability measures (μ,ν)**



Input Discret Measures

$$\nu = \sum_{j=1}^{M} \beta_j \delta_{x_j} \qquad \mu = \sum_{i=1}^{N} \alpha_i \delta_{x_i}$$

Cost Matrix

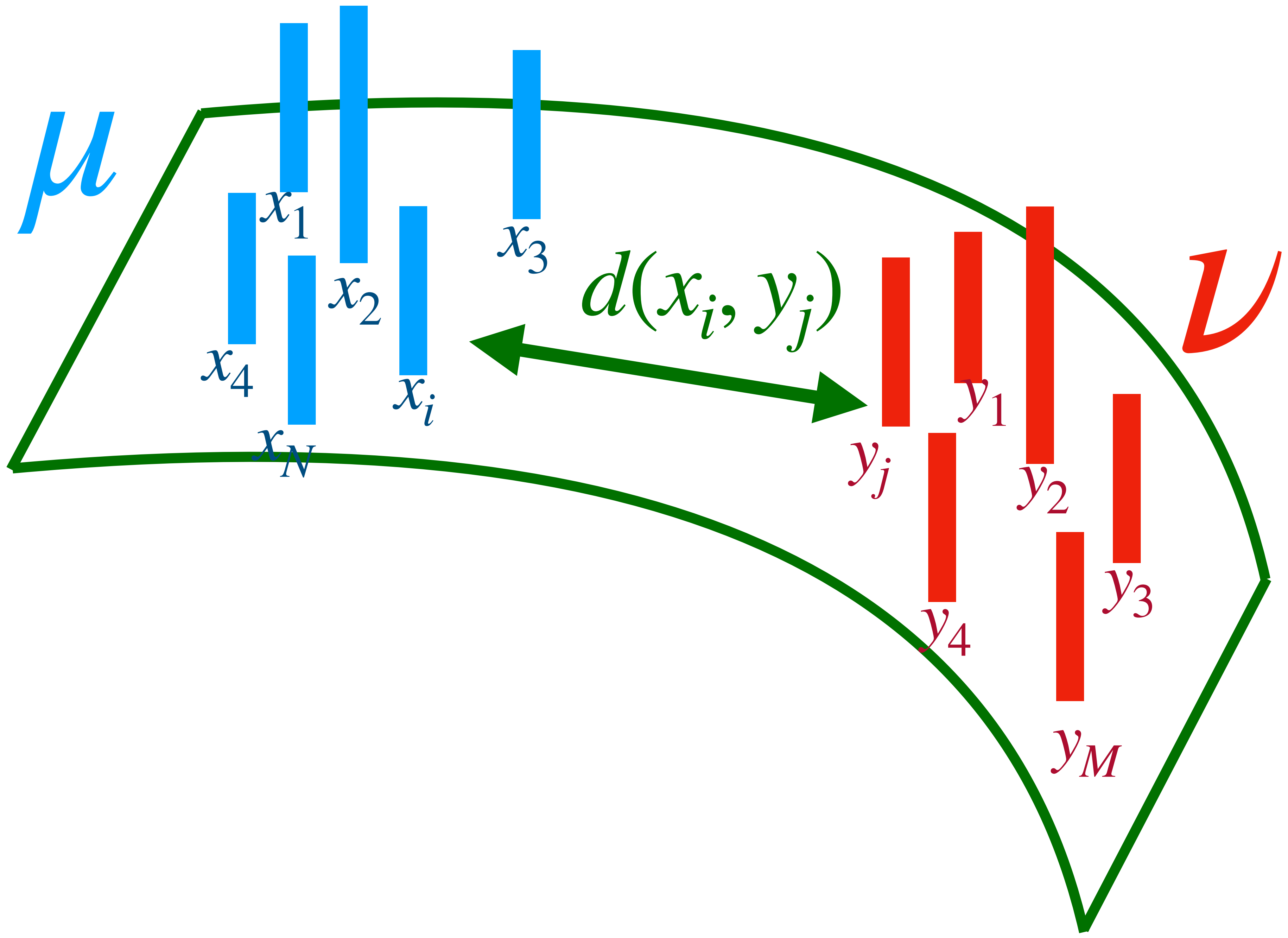
$$C = \begin{pmatrix} d(x_1, y_1) & \cdots & d(x_1, y_M) \\ \cdots & \cdots & \cdots \\ d(x_N, y_1) & \cdots & d(x_N, y_M) \end{pmatrix}$$

Transport Plan

$$\begin{array}{cc} \begin{array}{ccc} \alpha_1 & & \alpha_M \\ \uparrow & & \uparrow \end{array} & \\ \Pi = \begin{pmatrix} \pi_{11} & \cdots & \pi_{1M} \\ \cdots & \cdots & \cdots \\ \pi_{N1} & \cdots & \pi_{NM} \end{pmatrix} & \begin{array}{c} \longrightarrow \beta_1 \\ \\ \longrightarrow \beta_N \end{array} \end{array}$$

Optimal Transport

Goal Compute distance between probability measures (μ,ν)



Input Discret Measures

$$\nu = \sum_{j=1}^{M} \beta_j \delta_{x_j} \qquad \mu = \sum_{i=1}^{N} \alpha_i \delta_{x_i}$$

Cost Matrix

$$C = \begin{pmatrix} d(x_1, y_1) & \cdots & d(x_1, y_M) \\ \cdots & \cdots & \cdots \\ d(x_N, y_1) & \cdots & d(x_N, y_M) \end{pmatrix}$$

Transport Plan

$$\begin{array}{cccccc} & \alpha_1 & & \alpha_M & & \\ & \uparrow & & \uparrow & & \\ \Pi = & \left(\pi_{11}\right. & \cdots & \left.\pi_{1M}\right) & \longrightarrow & \beta_1 \\ & \cdots & \cdots & \cdots & & \\ & \left(\pi_{N1}\right. & \cdots & \left.\pi_{NM}\right) & \longrightarrow & \beta_N \end{array}$$

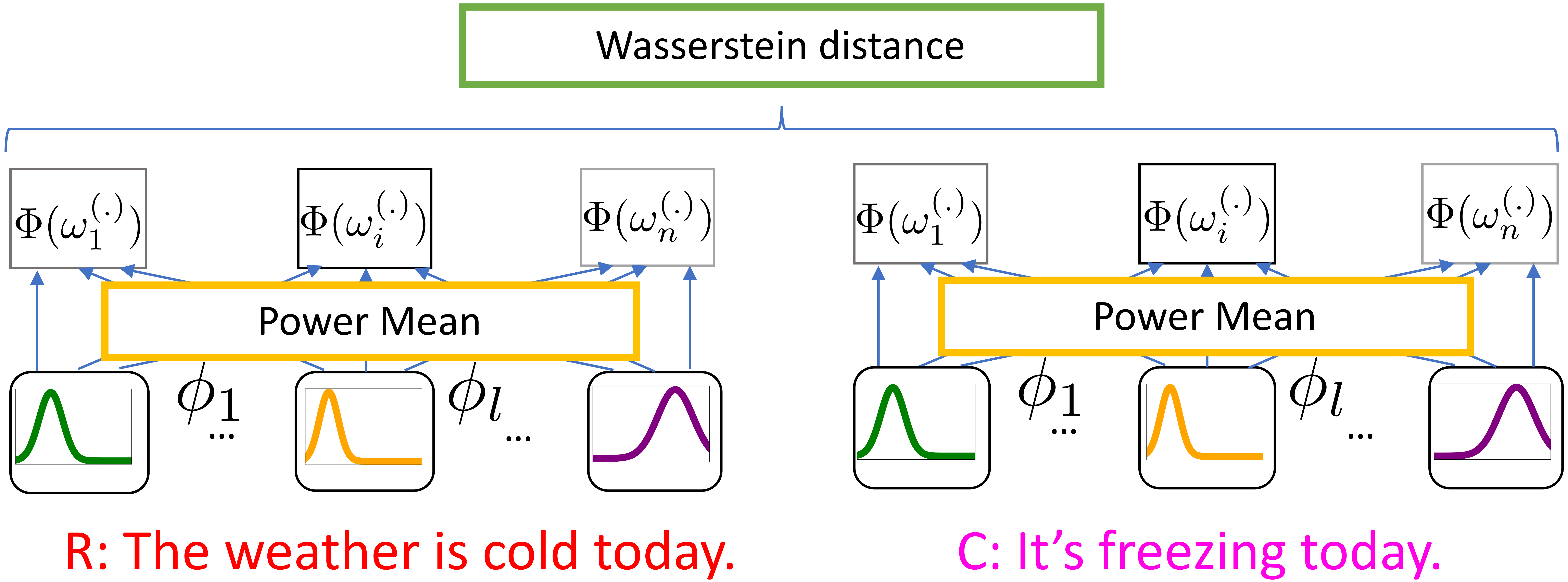
Wasserstein Distance

$$OT(\nu, \mu) = \min_{\Pi} \sum_{ij} C_{i,j} \times \Pi_{i,j}$$

$$\Pi 1 = \alpha, \Pi^T 1 = \beta$$

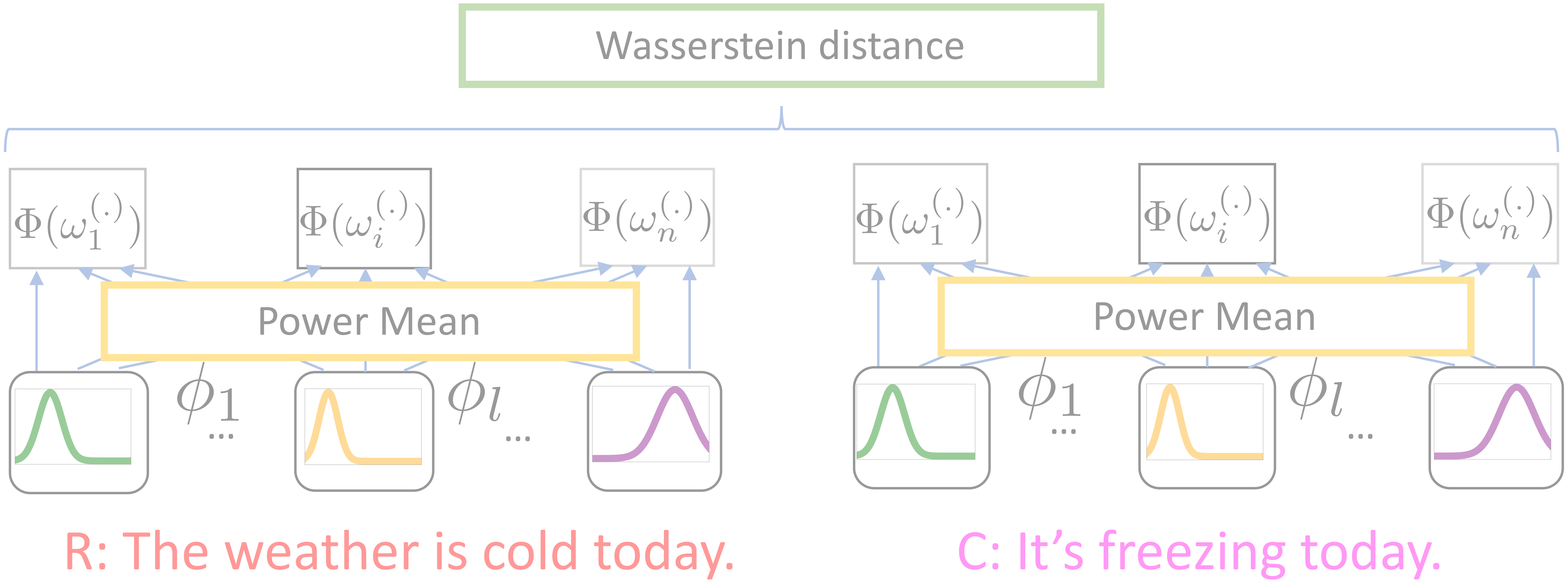
MoverScore

MoverScore



Wasserstein distance
$\Phi(\omega_1^{(.)})$
$\Phi(\omega_i^{(.)})$
$\Phi(\omega_n^{(.)})$
$\Phi(\omega_1^{(.)})$
$\Phi(\omega_i^{(.)})$
$\Phi(\omega_n^{(.)})$
Power Mean
Power Mean
ϕ_1 ...
ϕ_l ...
ϕ_1 ...
ϕ_l ...
R: The weather is cold today.
C: It's freezing today.

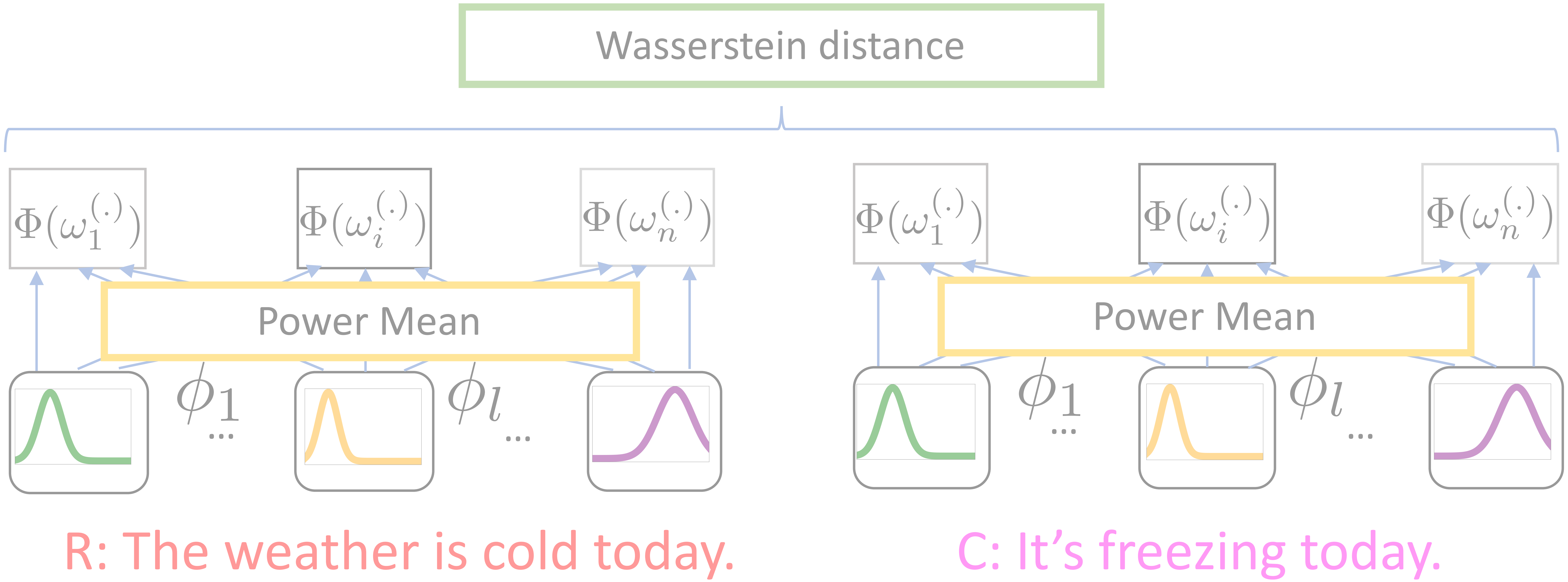
MoverScore



Advantage

1. Deal with paraphrases
2. Include "semantic"
3. Use several layers

MoverScore



Advantage

1. Deal with paraphrases
2. Include "semantic"
3. Use several layers

Limitations

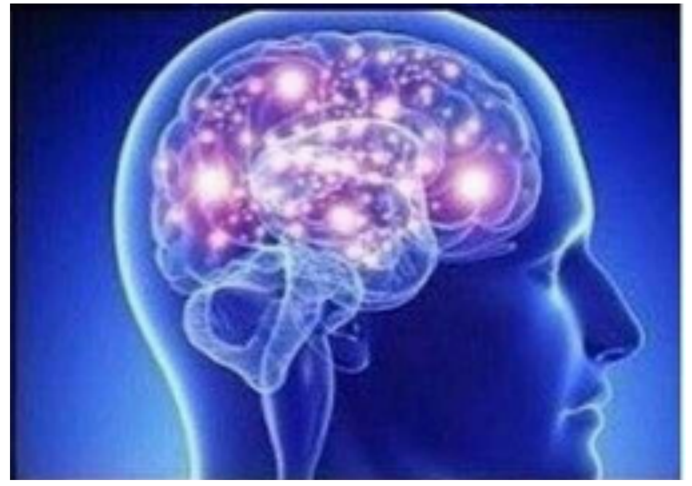
1. Use arbitrary sequence of operation

(euclidean aggregation function

Wasserstein distance)

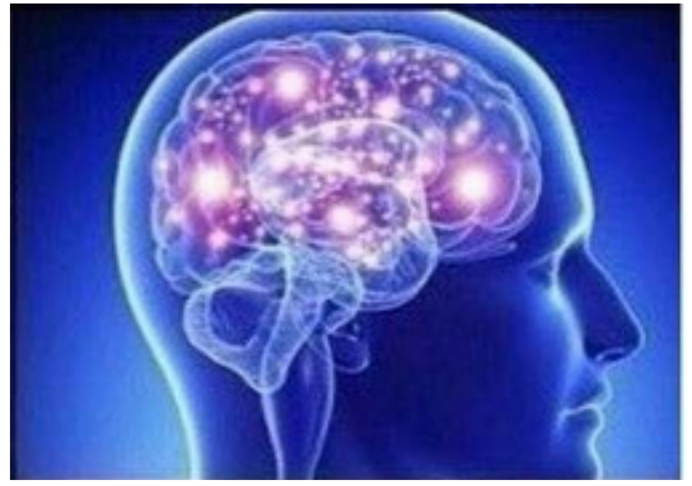
Our Contributions

Our Contributions



A novel metric called BaryScore

Our Contributions



A novel metric called BaryScore

Previously

1 Take one layer

2 Do a series of operations
(Wasserstein)

BertScore

Our Contributions

A novel metric called BaryScore

Previously

1 Take one layer	1 Take several layers
2 Do a series of operations (Wasserstein)	2 Aggregate using euclidean distance
	3 Do a series of Operations (Wasserstein)
BertScore	**MoverScore**

Our Contributions

A novel metric called BaryScore

Previously

Best of all worlds

1 Take one layer

1 Take several layers

2 Do a series of operations (Wasserstein)

2 Aggregate using euclidean distance

3 Do a series of Operations (Wasserstein)

BertScore

MoverScore

Our Contributions

A novel metric called BaryScore

Previously		Best of all worlds
1 Take one layer	1 Take several layers	1 Take several layers
2 Do a series of operations (Wasserstein)	2 Aggregate using euclidean distance	2 Aggregate using Wasserstein distance
	3 Do a series of Operations (Wasserstein)	3 Do a series of operation (Wasserstein)
BertScore	MoverScore	BaryScore

Wasserstein Barycenters

Wasserstein Barycenters

Euclidean Interpolation

$$\nu = \sum_{i=1}^{N} \alpha_i l_2(\mu_i, \mu)$$

Wasserstein Barycenters

Euclidean Interpolation

$$\nu = \sum_{i=1}^{N} \alpha_i l_2(\mu_i, \mu)$$

Wasserstein Interpolation

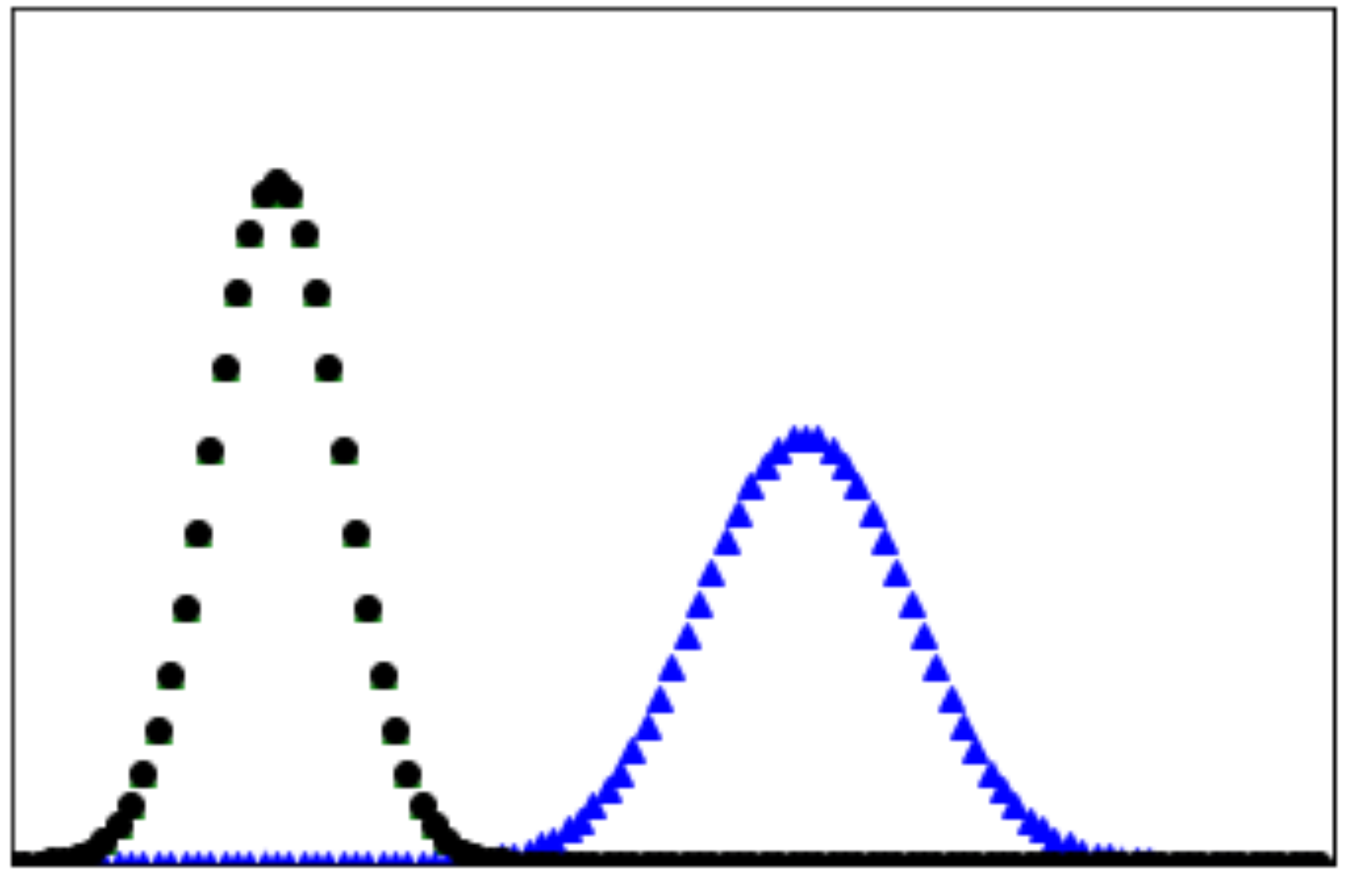
$$\nu = argmin_{\mu} \sum_{i=1}^{N} \alpha_i W(\mu_i, \mu)$$

Wasserstein Barycenters

Euclidean Interpolation

$$\nu = \sum_{i=1}^{N} \alpha_i l_2(\mu_i, \mu)$$

$$\nu = \alpha_i l_2(\mu_i, \mu) + (1 - \alpha_i) l_2(\mu_i, \mu)$$



α_i **Varies**

Wasserstein Interpolation

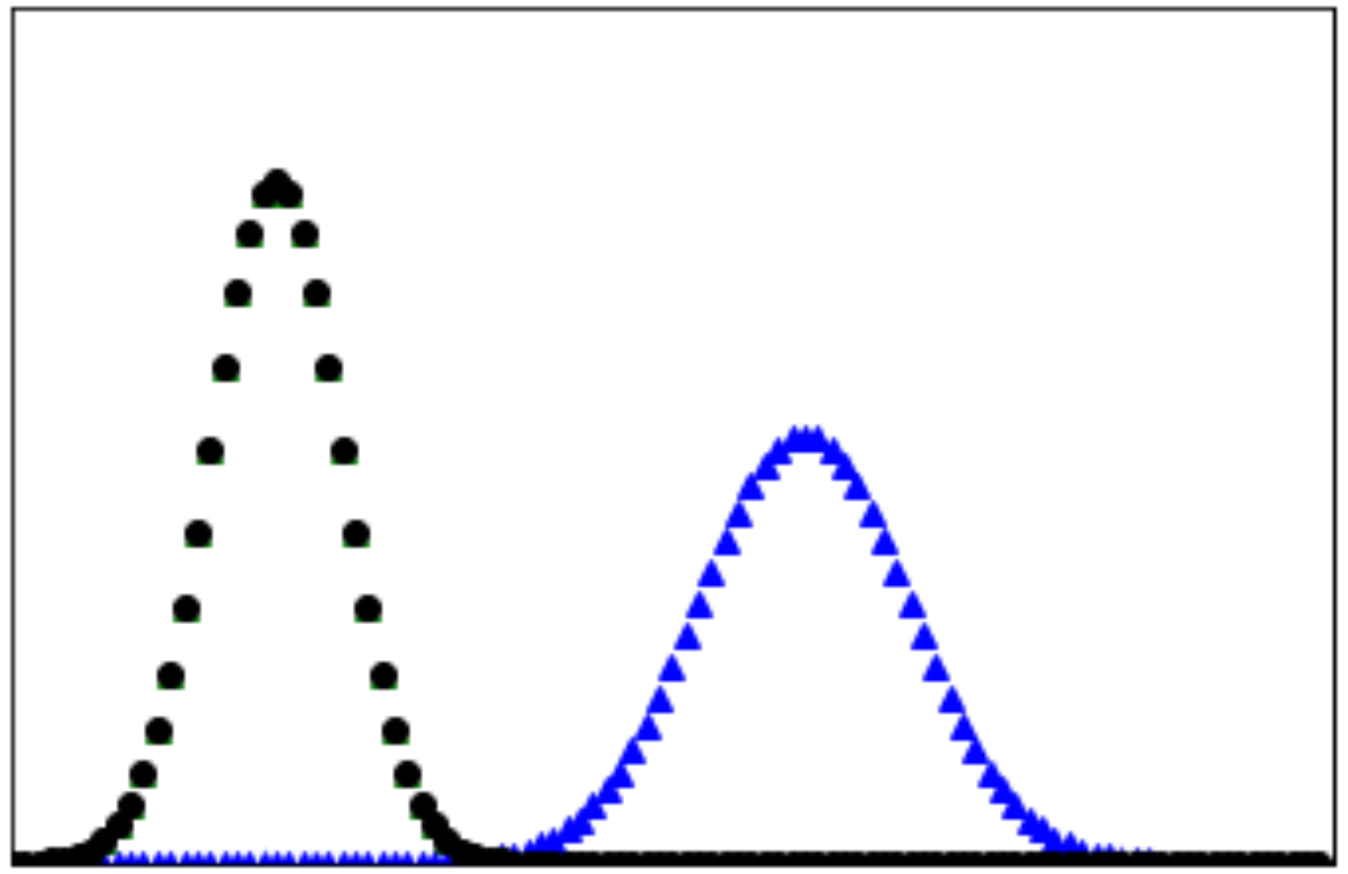
$$\nu = argmin_{\mu} \sum_{i=1}^{N} \alpha_i W(\mu_i, \mu)$$

Wasserstein Barycenters

Euclidean Interpolation

$$\nu = \sum_{i=1}^{N} \alpha_i l_2(\mu_i, \mu)$$

$$\nu = \alpha_i l_2(\mu_i, \mu) + (1 - \alpha_i) l_2(\mu_i, \mu)$$



α_i **Varies**

Do not look like a gaussian !

Wasserstein Interpolation

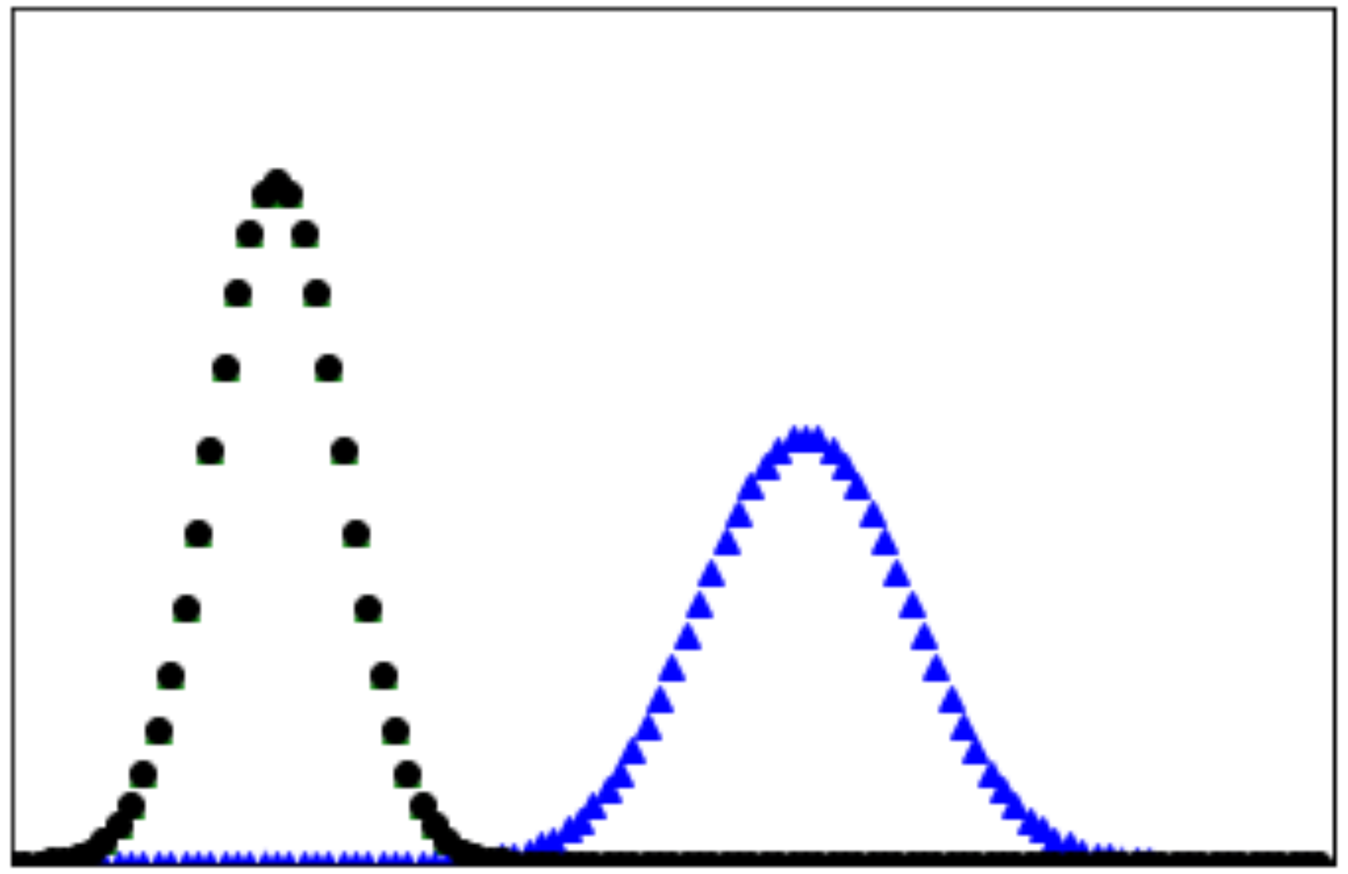
$$\nu = argmin_\mu \sum_{i=1}^{N} \alpha_i W(\mu_i, \mu)$$

Wasserstein Barycenters

Euclidean Interpolation

$$\nu = \sum_{i=1}^{N} \alpha_i l_2(\mu_i, \mu)$$

$$\nu = \alpha_i l_2(\mu_i, \mu) + (1 - \alpha_i) l_2(\mu_i, \mu)$$



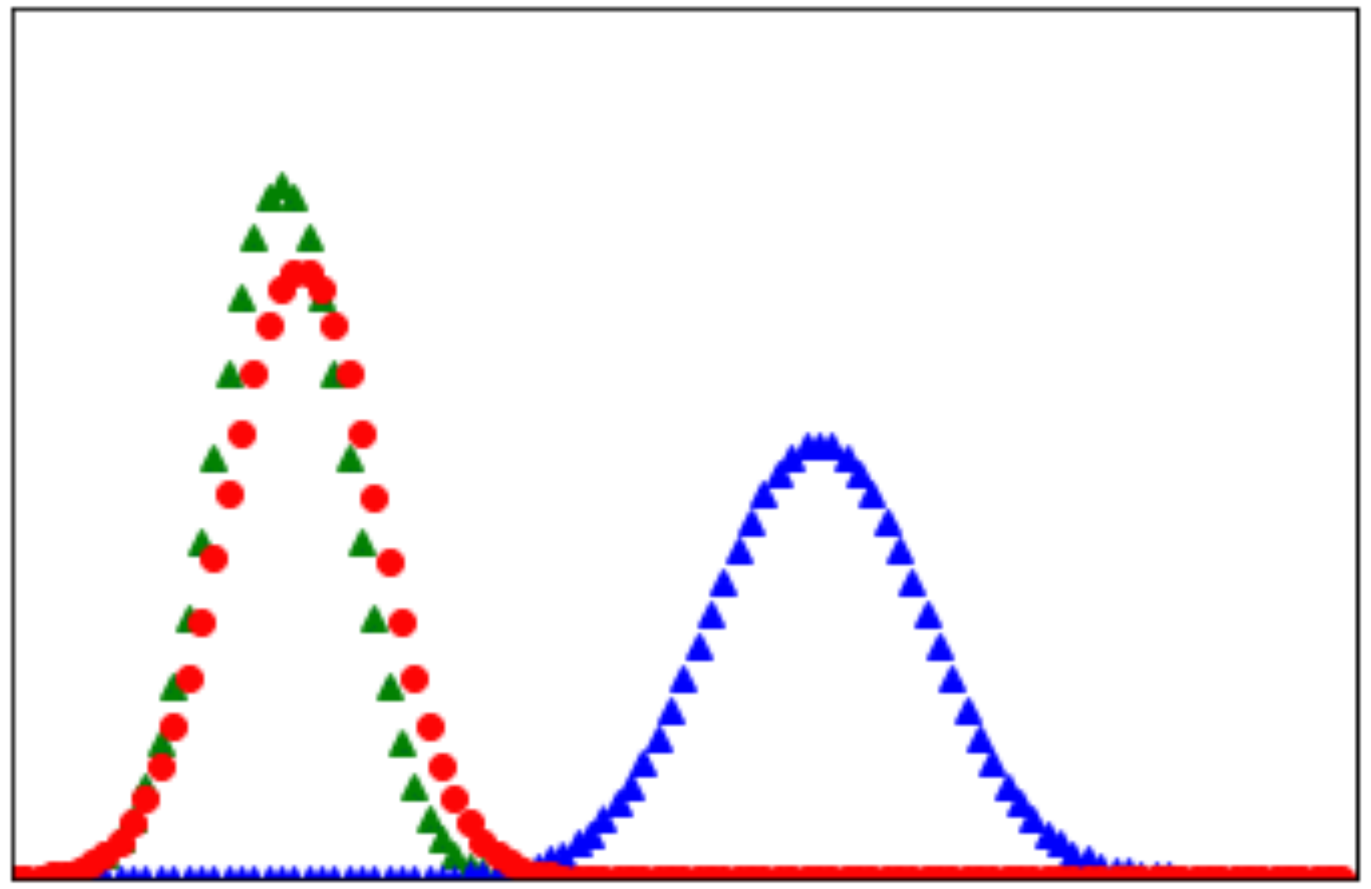
Do not look like a gaussian !

α_i **Varies**

Wasserstein Interpolation

$$\nu = argmin_\mu \sum_{i=1}^{N} \alpha_i W(\mu_i, \mu)$$

$$\nu = argmin_\mu \alpha_i W(\mu_i, \mu) + (1 - \alpha_i) W(\mu_i, \mu)$$



Wasserstein Barycenters

Euclidean Interpolation

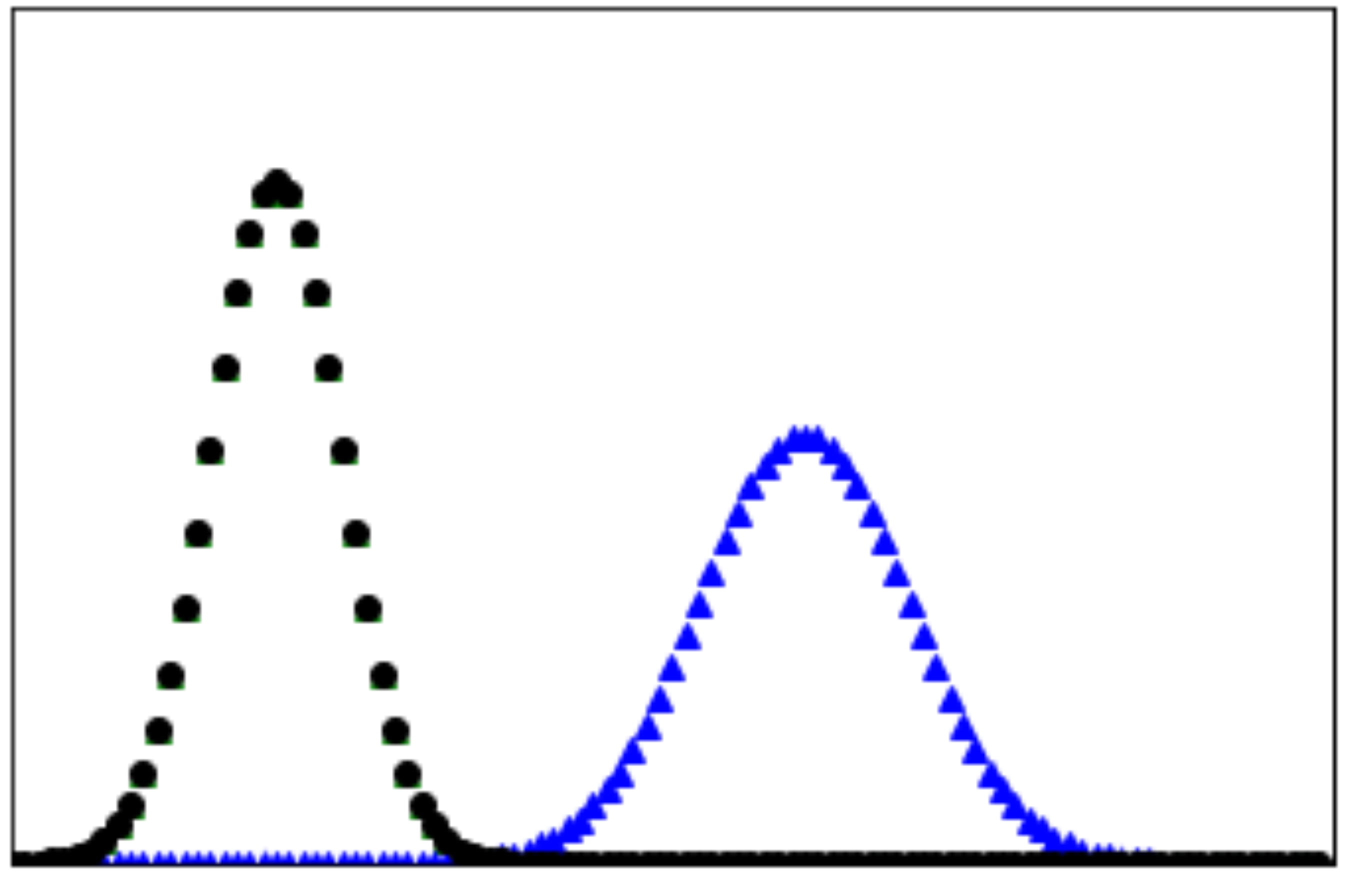
$$\nu = \sum_{i=1}^{N} \alpha_i l_2(\mu_i, \mu)$$

$$\nu = \alpha_i l_2(\mu_i, \mu) + (1 - \alpha_i) l_2(\mu_i, \mu)$$

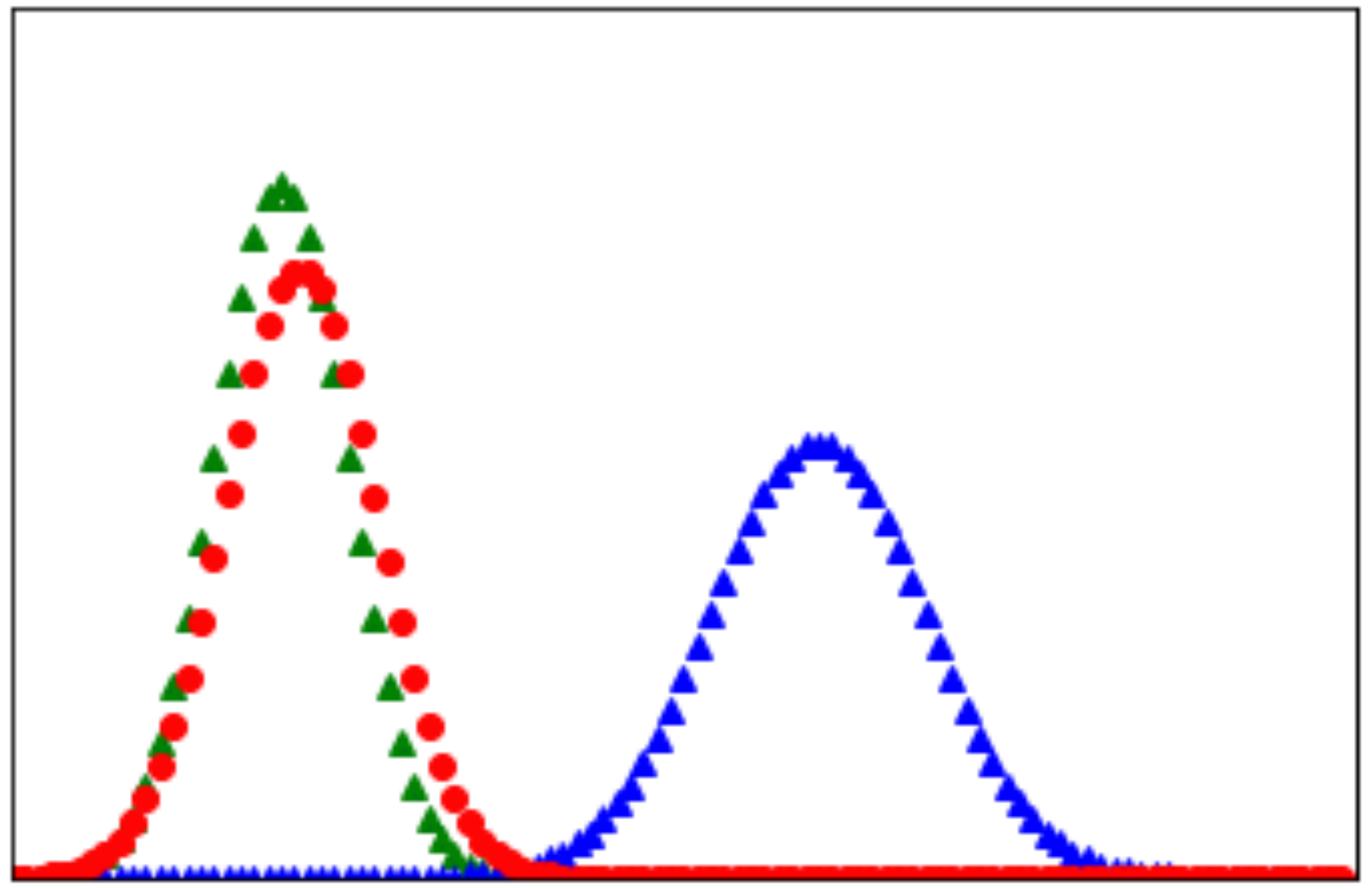
Wasserstein Interpolation

$$\nu = argmin_\mu \sum_{i=1}^{N} \alpha_i W(\mu_i, \mu)$$

$$\nu = argmin_\mu \alpha_i W(\mu_i, \mu) + (1 - \alpha_i) W(\mu_i, \mu)$$



α_i **Varies**



Do not look like a gaussian !

Preserve the gaussian!

BaryScore

BarySсore

Reference: R

BaryScore

Reference: R **Candidate:** C

BaryScore

Reference: R **Candidate:** C

Goal : metric $m : (\boxed{C}, \boxed{R}) \mapsto m(C, R) \in \mathbb{R}_+$

BaryScore

Reference: R **Candidate:** C

Goal : metric $m : (\boxed{C}, \boxed{R}) \mapsto m(C, R) \in \mathbb{R}_+$

Algorithm

Algorithm 1 `BaryScore`

INPUT: $C = \{\omega_1^c, \ldots, \omega_{n_c}^c\}$, $R = \{\omega_1^r, \ldots, \omega_{n_r}^r\}$, $(\phi_1, \ldots, \phi_L)$ pre-trained layers from BERT or ELMo.

Compute layers embeddings:
$\phi_\ell(C)$ and $\phi_\ell(R)$ for every $1 \le \ell \le L$.

Compute measures: $\{\hat{\mu}_{C,\ell}, \hat{\mu}_{R,\ell}\}_{\ell=1}^L$.

Compute Wasserstein barycenters:

$$\hat{\mu}_C = \underset{\hat{\mu}}{\operatorname{argmin}} \sum_{\ell=1}^{L} \mathcal{W}(\hat{\mu}_{C,\ell}, \hat{\mu}),$$

$$\hat{\mu}_R = \underset{\hat{\mu}}{\operatorname{argmin}} \sum_{\ell=1}^{L} \mathcal{W}(\hat{\mu}_{R,\ell}.\hat{\mu}),$$

OUTPUT: $\mathcal{W}(\hat{\mu}_R, \hat{\mu}_C)$.

BaryScore

Reference: R Candidate: C

Goal : metric $m : (\boxed{C}, \boxed{R}) \mapsto m(C, R) \in \mathbb{R}_+$

Algorithm

1. Find the Wasserstein barycentric distributions of BERT layers for C and R

Algorithm 1 `BaryScore`

INPUT: $C = \{\omega_1^c, \ldots, \omega_{n_c}^c\}$, $R = \{\omega_1^r, \ldots, \omega_{n_r}^r\}$, $(\phi_1, \ldots, \phi_L)$ pre-trained layers from BERT or ELMo.

Compute layers embeddings:
$\phi_\ell(C)$ and $\phi_\ell(R)$ for every $1 \leq \ell \leq L$.

Compute measures: $\{\hat{\mu}_{C,\ell}, \hat{\mu}_{R,\ell}\}_{\ell=1}^L$.

Compute Wasserstein barycenters:

$$\hat{\mu}_C = \underset{\hat{\mu}}{\text{argmin}} \sum_{\ell=1}^{L} \mathcal{W}(\hat{\mu}_{C,\ell}, \hat{\mu}),$$

$$\hat{\mu}_R = \underset{\hat{\mu}}{\text{argmin}} \sum_{\ell=1}^{L} \mathcal{W}(\hat{\mu}_{R,\ell}.\hat{\mu}),$$

OUTPUT: $\mathcal{W}(\hat{\mu}_R, \hat{\mu}_C)$.

BaryScore

Reference: R Candidate: C

Goal : metric $m : (\boxed{C}, \boxed{R}) \mapsto m(C, R) \in \mathbb{R}_+$

Algorithm

1. Find the Wasserstein barycentric distributions of BERT layers for C and R

Algorithm 1 `BaryScore`

INPUT: $C = \{\omega_1^c, \ldots, \omega_{n_c}^c\}$, $R = \{\omega_1^r, \ldots, \omega_{n_r}^r\}$,
$(\phi_1, \ldots, \phi_L)$ pre-trained layers from BERT or ELMo.

Compute layers embeddings:
$\phi_\ell(C)$ and $\phi_\ell(R)$ for every $1 \leq \ell \leq L$.

Compute measures: $\{\hat{\mu}_{C,\ell}, \hat{\mu}_{R,\ell}\}_{\ell=1}^{L}$.

Compute Wasserstein barycenters:

$$\hat{\mu}_C = \underset{\hat{\mu}}{\text{argmin}} \sum_{\ell=1}^{L} \mathcal{W}(\hat{\mu}_{C,\ell}, \hat{\mu}),$$

$$\hat{\mu}_R = \underset{\hat{\mu}}{\text{argmin}} \sum_{\ell=1}^{L} \mathcal{W}(\hat{\mu}_{R,\ell}.\hat{\mu}),$$

OUTPUT: $\mathcal{W}(\hat{\mu}_R, \hat{\mu}_C)$.

BaryScore

Reference: R Candidate: C

Goal : metric $m : (\boxed{C}, \boxed{R}) \mapsto m(C, R) \in \mathbb{R}_+$

Algorithm

1. Find the Wasserstein barycentric distributions of BERT layers for C and R

2. Evaluate these barycentric distributions using the Wasserstein distance.

Algorithm 1 `BaryScore`

INPUT: $C = \{\omega_1^c, \ldots, \omega_{n_c}^c\}$, $R = \{\omega_1^r, \ldots, \omega_{n_r}^r\}$, $(\phi_1, \ldots, \phi_L)$ pre-trained layers from BERT or ELMo.

Compute layers embeddings:
$\phi_\ell(C)$ and $\phi_\ell(R)$ for every $1 \le \ell \le L$.

Compute measures: $\{\hat{\mu}_{C,\ell}, \hat{\mu}_{R,\ell}\}_{\ell=1}^L$.

Compute Wasserstein barycenters:

$$\boxed{\hat{\mu}_C = \operatorname*{argmin}_{\hat{\mu}} \sum_{\ell=1}^{L} \mathcal{W}(\hat{\mu}_{C,\ell}, \hat{\mu}),}$$

$$\boxed{\hat{\mu}_R = \operatorname*{argmin}_{\hat{\mu}} \sum_{\ell=1}^{L} \mathcal{W}(\hat{\mu}_{R,\ell}.\hat{\mu}),}$$

OUTPUT: $\mathcal{W}(\hat{\mu}_R, \hat{\mu}_C)$.

BaryScore

Reference: R Candidate: C

Goal : metric $m : (\boxed{C}, \boxed{R}) \mapsto m(C, R) \in \mathbb{R}_+$

Algorithm

1. Find the Wasserstein barycentric distributions of BERT layers for C and R

2. Evaluate these barycentric distributions using the Wasserstein distance.

Algorithm 1 `BaryScore`

INPUT: $C = \{\omega_1^c, \ldots, \omega_{n_c}^c\}$, $R = \{\omega_1^r, \ldots, \omega_{n_r}^r\}$, $(\phi_1, \ldots, \phi_L)$ pre-trained layers from BERT or ELMo.

Compute layers embeddings:

$\phi_\ell(C)$ and $\phi_\ell(R)$ for every $1 \leq \ell \leq L$.

Compute measures: $\{\hat{\mu}_{C,\ell}, \hat{\mu}_{R,\ell}\}_{\ell=1}^L$.

Compute Wasserstein barycenters:

$$\hat{\mu}_C = \underset{\hat{\mu}}{\text{argmin}} \sum_{\ell=1}^{L} \mathcal{W}(\hat{\mu}_{C,\ell}, \hat{\mu}),$$

$$\hat{\mu}_R = \underset{\hat{\mu}}{\text{argmin}} \sum_{\ell=1}^{L} \mathcal{W}(\hat{\mu}_{R,\ell}.\hat{\mu}),$$

OUTPUT: $\boxed{\mathcal{W}(\hat{\mu}_R, \hat{\mu}_C).}$

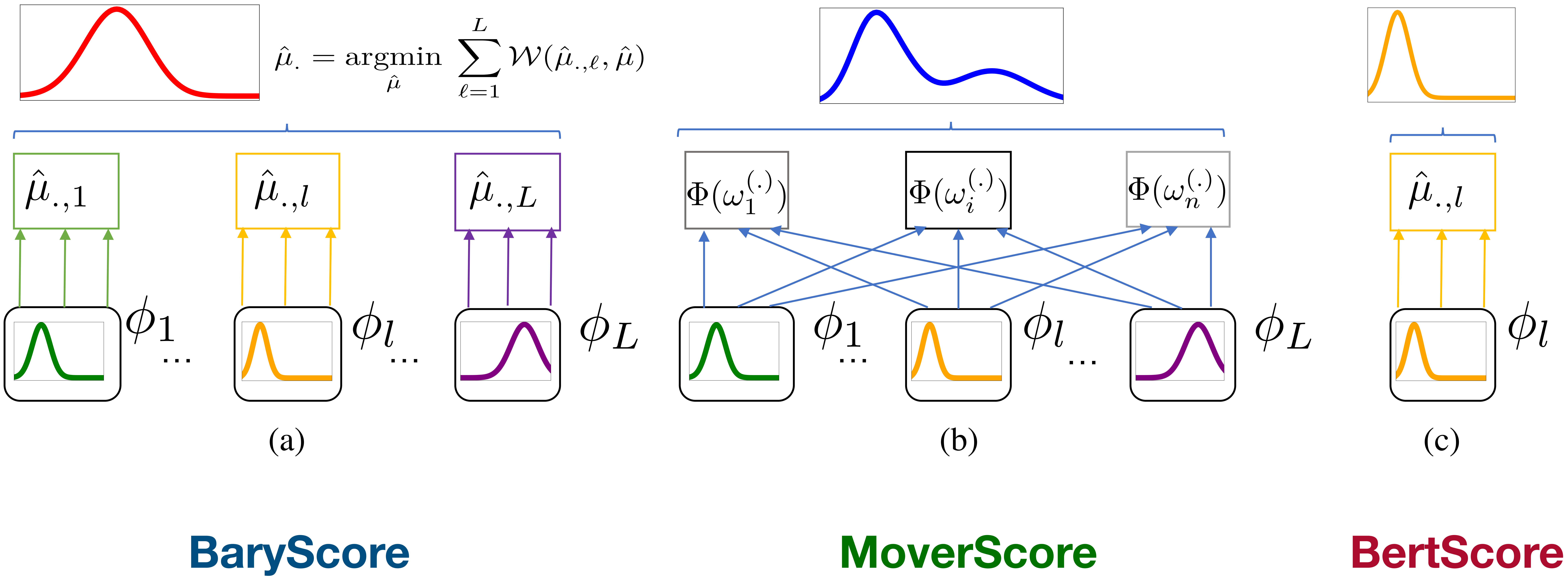
BaryScore vs BertScore vs MoverScore

BaryScore vs BertScore vs MoverScore

Comparison between aggregation functions

BaryScore vs BertScore vs MoverScore

Comparison between aggregation functions



Evaluating Metric Scores

Evaluating Metric Scores

Notations

S systems

N texts

R_i i-th reference

C_i^j i-th text candidate generated by j-th system

$h(C_i^j)$ human score

Evaluating Metric Scores

Notations

S systems

N texts

R_i i-th reference

C_i^j i-th text candidate generated by j-th system

$h(C_i^j)$ human score

Can the metric be used to compare the performance of two systems?

$$K_{sys} = K(M^{sy}, H^{sy})$$

$$M^{sy} = \left[\frac{1}{N}\sum_{i=1}^{N} m(R_i, C_i^1), \cdots, \frac{1}{N}\sum_{i=1}^{n} m(R_i, C_i^s)\right]$$

$$H^{sy} = \left[\frac{1}{N}\sum_{i=1}^{N} h(C_i^1), \cdots, \frac{1}{N}\sum_{i=1}^{N} h(C_i^S)\right]$$

Evaluating Metric Scores

Notations

S systems

N texts

R_i i-th reference

C_i^j i-th text candidate generated by j-th system

$h(C_i^j)$ human score

Can the metric be used to compare the performance of two systems?

$$K_{sys} = K(M^{sy}, H^{sy})$$

$$M^{sy} = \left[\frac{1}{N}\sum_{i=1}^{N} m(R_i, C_i^1), \cdots, \frac{1}{N}\sum_{i=1}^{n} m(R_i, C_i^s)\right]$$

$$H^{sy} = \left[\frac{1}{N}\sum_{i=1}^{N} h(C_i^1), \cdots, \frac{1}{N}\sum_{i=1}^{N} h(C_i^S)\right]$$

System Aggregation !
Compare vector of length S

Evaluating Metric Scores

Notations

S systems

N texts

R_i i-th reference

C_i^j i-th text candidate generated by j-th system

$h(C_i^j)$ human score

Can the metric be used to compare the performance of two systems?

$$K_{sys} = K(M^{sy}, H^{sy})$$

$$M^{sy} = \left[\frac{1}{N}\sum_{i=1}^{N} m(R_i, C_i^1), \cdots, \frac{1}{N}\sum_{i=1}^{n} m(R_i, C_i^S)\right]$$

$$H^{sy} = \left[\frac{1}{N}\sum_{i=1}^{N} h(C_i^1), \cdots, \frac{1}{N}\sum_{i=1}^{N} h(C_i^S)\right]$$

System Aggregation !
Compare vector of length S

Can the metric be used as a loss or reward of a system?

$$K_{text} = \frac{1}{N}\sum_{i=1}^{N} K(M_i^{text}, H_i^{text})$$

$$H_i^{text} = \left[h(C_i^1), \cdots, h(C_i^S)\right]$$

$$M_i^{text} = \left[m(R_i, C_i^1), \cdots, m(R_i, C_i^S)\right]$$

Evaluating Metric Scores

R_i i-th reference

C_i^j i-th text candidate generated by j-th system

$h(C_i^j)$ human score

Notations

S systems

N texts

Can the metric be used to compare the performance of two systems?

$$K_{sys} = K(M^{sy}, H^{sy})$$

$$M^{sy} = \left[\frac{1}{N}\sum_{i=1}^{N} m(R_i, C_i^1), \cdots, \frac{1}{N}\sum_{i=1}^{n} m(R_i, C_i^S)\right]$$

$$H^{sy} = \left[\frac{1}{N}\sum_{i=1}^{N} h(C_i^1), \cdots, \frac{1}{N}\sum_{i=1}^{N} h(C_i^S)\right]$$

System Aggregation !
Compare vector of length S

Can the metric be used as a loss or reward of a system?

$$K_{text} = \frac{1}{N}\sum_{i=1}^{N} K(M_i^{text}, H_i^{text})$$

$$H_i^{text} = \left[h(C_i^1), \cdots, h(C_i^S)\right]$$

$$M_i^{text} = \left[m(R_i, C_i^1), \cdots, m(R_i, C_i^S)\right]$$

Text Aggregation !
Averaged correlation

Experimental Setting

Experimental Setting

Machine Translation

- Results on WMT17/WMT18
- All metrics are measures on en only
- Pairs includes cs-en de-en ru-en fi-en ro-en tr-en

Experimental Setting

Machine Translation

- Results on WMT17/WMT18
- All metrics are measures on en only
- Pairs includes cs-en de-en ru-en fi-en ro-en tr-en

Data2text Generation

- Results on WebNLG 2020
- Correctness / Data Coverage / Relevance
- Results on English only

Experimental Setting

Machine Translation

- Results on WMT17/WMT18
- All metrics are measures on en only
- Pairs includes cs-en de-en ru-en fi-en ro-en tr-en

Data2text Generation

- Results on WebNLG 2020
- Correctness / Data Coverage / Relevance
- Results on English only

Image Captioning

- Results on MSCOCO
- Results on English only

Experimental Setting

Machine Translation

- Results on WMT17/WMT18
- All metrics are measures on en only
- Pairs includes cs-en de-en ru-en fi-en ro-en tr-en

Data2text Generation

- Results on WebNLG 2020
- Correctness / Data Coverage / Relevance
- Results on English only

Image Captioning

- Results on MSCOCO
- Results on English only

Summary Generation

- Results on SummEval
- Correlation with pyramid score
- Results on English only

Experimental Setting

Machine Translation

- Results on WMT17/WMT18
- All metrics are measures on en only
- Pairs includes cs-en de-en ru-en fi-en ro-en tr-en

Data2text Generation

- Results on WebNLG 2020
- Correctness / Data Coverage / Relevance
- Results on English only

Image Captioning

- Results on MSCOCO
- Results on English only

Summary Generation

- Results on SummEval
- Correlation with pyramid score
- Results on English only

Results

Results

Task

Results

Task

(John_Blaha birthDate 1942_08_26)
(John_Blaha birthPlace San_Antonio)
(John_E_Blaha job Pilot)

John Blaha, born in San Antonio on 1942-08-26, worked as a pilot

Results

Task

(John_Blaha birthDate 1942_08_26)
(John_Blaha birthPlace San_Antonio)
(John_E_Blaha job Pilot)

John Blaha, born in San Antonio on 1942-08-26, worked as a pilot

Criterion

Correctness

Data coverage

Relevance

Results

Task

(John_Blaha birthDate 1942_08_26)
(John_Blaha birthPlace San_Antonio)
(John_E_Blaha job Pilot)

John Blaha, born in San Antonio on 1942-08-26, worked as a pilot

Criterion

Correctness

Data coverage

Relevance

	Correctness			Data Coverage			Relevance		
Metric	r	ρ	τ	r	ρ	τ	r	ρ	τ
Correct	100.0	100.0	100.0	97.6	85.2	73.3	99.1	89.7	75.0
DataC	85.2	97.6	73.3	100.0	100.0	100.0	96.0	93.8	81.6
Relev	89.7	99.1	75.0	96.0	93.8	81.6	100.0	100.0	100.0
BaryS	**91.7**	**90.0**	**78.3**	**87.8**	78.2	61.6	**89.4**	**82.6**	70.0
BaryS$^+$	90.5	89.5	76.6	87.7	**85.0**	**70.0**	89.2	86.4	71.6
BertS	85.5	83.4	73.3	74.7	68.2	53.3	83.3	79.4	65.0
MoverS	84.1	84.1	73.3	78.7	66.2	53.3	82.1	77.4	65.0
BLEU	77.6	66.3	60.0	55.7	50.2	36.6	63.0	65.2	51.6
R-1	80.6	65.0	65.0	76.5	76.3	60.3	64.3	69.2	56.7
R-2	73.6	63.3	58.3	54.7	43.1	35.0	62.0	60.8	46.7
R-WE	60.9	73.4	60.0	40.2	58.2	40.1	49.9	64.1	48.3
METEOR	86.5	66.3	70.0	77.3	50.2	46.6	82.1	65.2	58.6
TER	79.6	78.3	58.0	69.7	58.2	38.0	75.0	70.2	**77.6**

Correlation score for different coefficient Pearson r, Spearman ρ and Kendall τ

Results

Task

(John_Blaha birthDate 1942_08_26)
(John_Blaha birthPlace San_Antonio)
(John_E_Blaha job Pilot)

John Blaha, born in San Antonio on 1942-08-26, worked as a pilot

Criterion

Correctness

Data coverage

Relevance

	Correctness			Data Coverage			Relevance		
Metric	r	ρ	τ	r	ρ	τ	r	ρ	τ
Correct	100.0	100.0	100.0	97.6	85.2	73.3	99.1	89.7	75.0
DataC	85.2	97.6	73.3	100.0	100.0	100.0	96.0	93.8	81.6
Relev	89.7	99.1	75.0	96.0	93.8	81.6	100.0	100.0	100.0
`BaryS`	**91.7**	**90.0**	**78.3**	**87.8**	78.2	61.6	**89.4**	**82.6**	70.0
`BaryS`$^+$	90.5	89.5	76.6	87.7	**85.0**	**70.0**	89.2	86.4	71.6
BertS	85.5	83.4	73.3	74.7	68.2	53.3	83.3	79.4	65.0
MoverS	84.1	84.1	73.3	78.7	66.2	53.3	82.1	77.4	65.0
BLEU	77.6	66.3	60.0	55.7	50.2	36.6	63.0	65.2	51.6
R-1	80.6	65.0	65.0	76.5	76.3	60.3	64.3	69.2	56.7
R-2	73.6	63.3	58.3	54.7	43.1	35.0	62.0	60.8	46.7
R-WE	60.9	73.4	60.0	40.2	58.2	40.1	49.9	64.1	48.3
METEOR	86.5	66.3	70.0	77.3	50.2	46.6	82.1	65.2	58.6
TER	79.6	78.3	58.0	69.7	58.2	38.0	75.0	70.2	**77.6**

Correlation score for different coefficient Pearson r, Spearman ρ and Kendall τ

Thanks for listening

Link to Paper

Title: **Automatic Text Evaluation through the Lens of Wasserstein Barycenters**

Corresponding Authors:

Pierre Colombo