Information Bottleneck

Rate Distortion Functions

Agenda

- Rate Distortion Theory
  - Blahut-Arimoto algorithm
- Information Bottleneck Principle
- IB algorithms
  - iIB
  - dIB
  - aIB
- Application

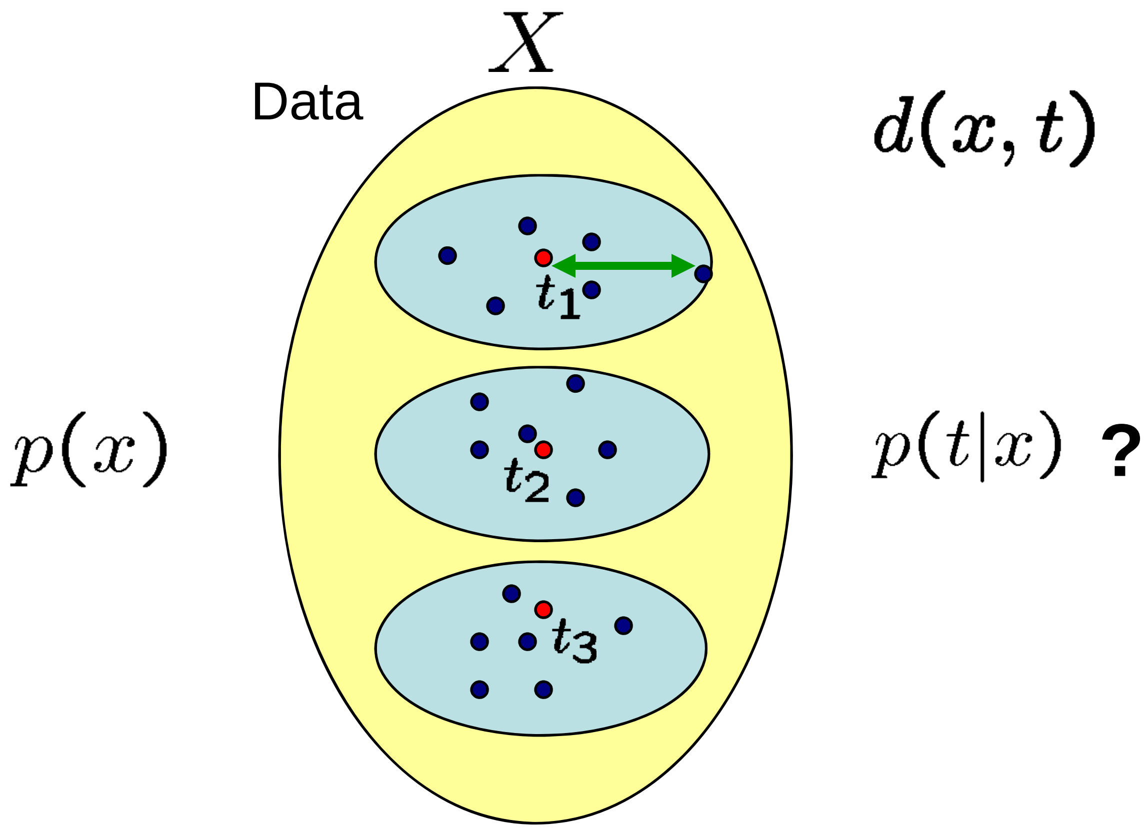
Rate Distortion Theory Introduction

- **Goal:** obtain compact clustering of the data with minimal expected distortion

- Distortion measure is a part of the problem setup

- The clustering and its quality depend on the choice of the distortion measure

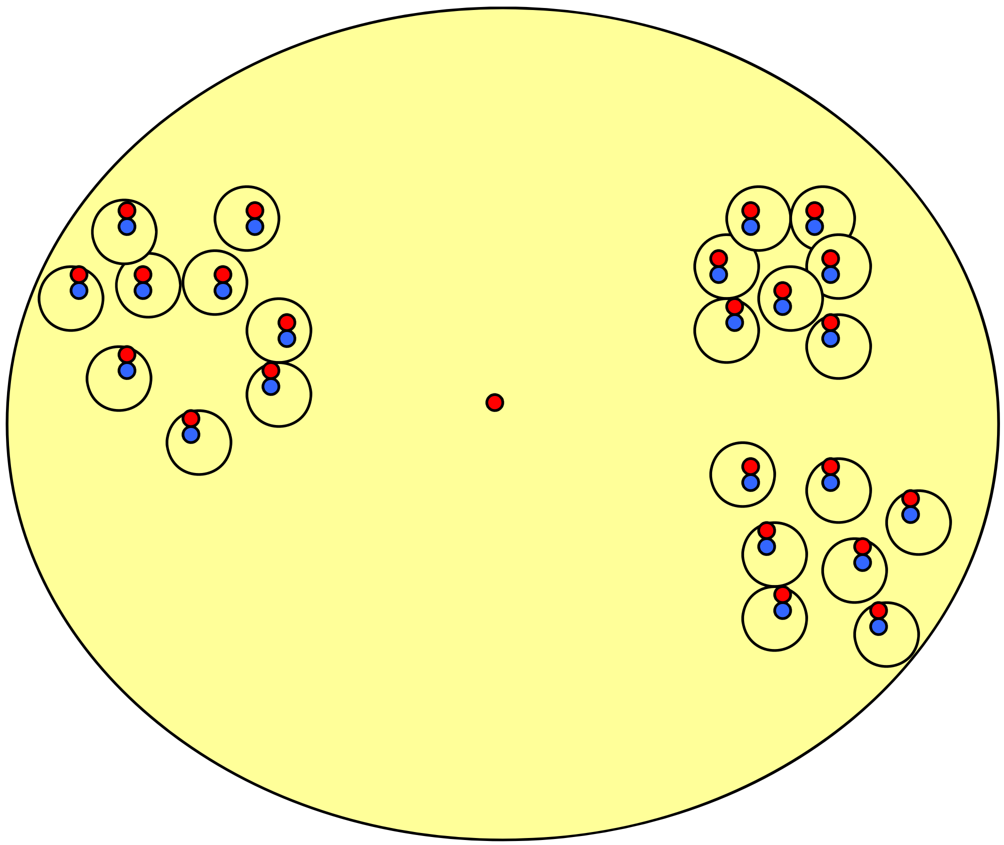
Rate Distortion Theory



- **Obtain compact clustering of the data with minimal expected distortion given fixed set of representatives** T

Cover & Thomas

Rate Distortion Theory - Intuition



- $T = X$
  - zero distortion
  - not compact

$$I(T;X) = H(X)$$

$$|T| = 1$$

  - high distortion
  - very compact

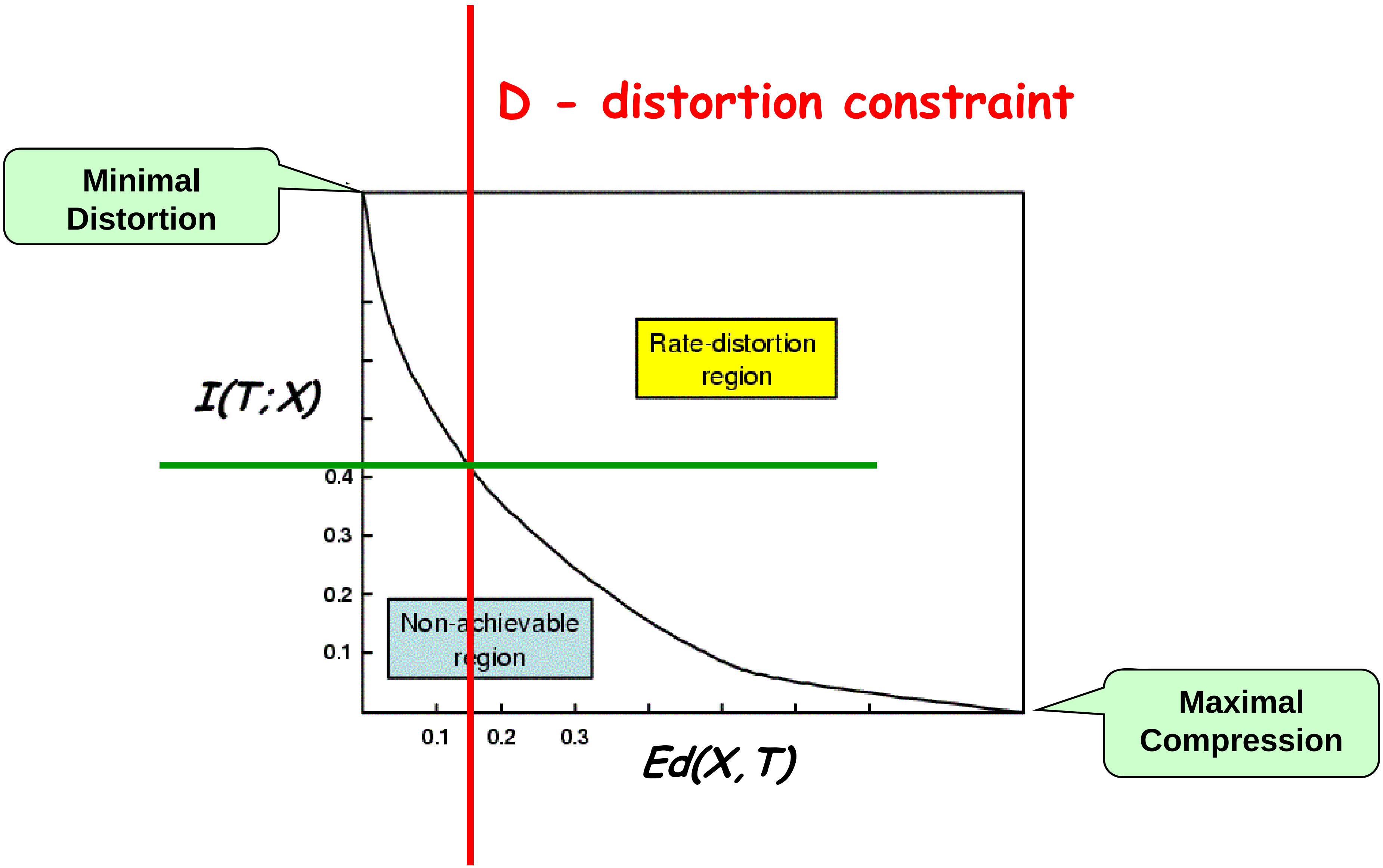
$$I(T;X) = 0$$

Rate Distortion Theory – Cont.

- **The quality of clustering is determined by**
  - **Complexity is measured by** $\boxed{I(T;X)}$
    (a.k.a. Rate)
  - **Distortion is measured by**

$$\boxed{Ed(X,T) = \sum_{i,j} p(x_i)p(t_j|x_i)d(x_i,t_j)}$$

Rate Distortion Plane



Rate Distortion Function

- **Let D be an upper bound constraint on the expected distortion**

Higher values of D mean more relaxed distortion constraint

$\Downarrow$

Stronger compression levels are attainable

- **Given the distortion constraint D find the most compact model (with smallest complexity R)**

$$R(D) \equiv \min_{\{p(t|x):Ed(X,T)\leq D\}} I(T;X)$$

Rate Distortion Function

- **Given**
  - Set of points X with prior $p(x)$
  - Set of representatives T
  - Distortion measure $d(x,t)$
- **Find**
  - The most compact soft clustering $p(t|x)$ of points of X that satisfies the distortion constraint D
- **Rate Distortion Function**

$$\boxed{R(D) \equiv \min_{\{p(t|x): Ed(X,T) \leq D\}} I(T;X)}$$

Rate Distortion Function

$$R(D) \equiv \min_{\{p(t|x):Ed(X,T)\leq D\}} I(T;X)$$

Complexity Term

Distortion Term

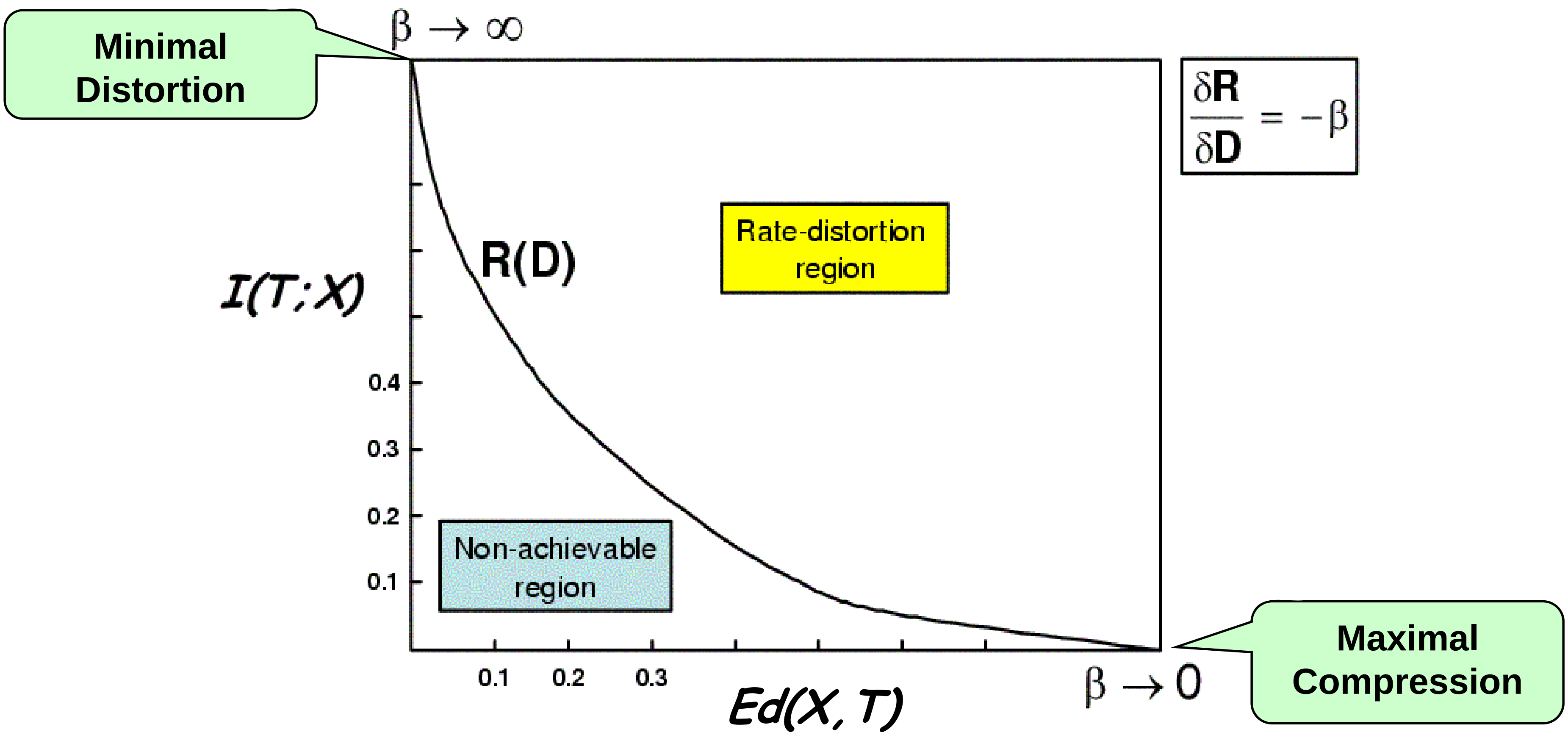
$$\mathcal{F}[p(t|x)] = I(T;X) + \beta Ed(X,T)$$

Lagrange Multiplier

Minimize $\mathcal{F}[p(t|x)]$!

Rate Distortion Curve

$$\mathcal{F}[p(t|x)] = I(T;X) + \beta Ed(X,T)$$



Rate Distortion Function

Minimize $$\boxed{\mathcal{F}[p(t|x)] = I(T;X) + \beta E d(X,T)}$$

Subject to $$\sum_t p(t|x) = 1 \;\; \forall x \in X$$

The minimum is attained when $$\frac{\partial \mathcal{F}}{\partial p(t|x)} = 0$$

$$\Rightarrow \quad \boxed{p(t|x) = \frac{p(t)}{Z(x,\beta)} \mathrm{e}^{-\beta d(x,t)}}$$

Normalization

Solution - Analysis

$$\mathcal{F}[p(t|x)] = I(T;X) + \beta E d(X,T)$$

Solution:

$$p(t|x) = \frac{p(t)}{Z(x,\beta)} \mathrm{e}^{-\beta d(x,t)}$$

The solution is implicit

Known

$$p(t) = \sum_x p(x) p(t|x)$$

Solution - Analysis

Solution:

$$p(t|x) = \frac{p(t)}{Z(x,\beta)} e^{-\beta d(x,t)}$$

For a fixed t

When x **is similar to** t

Solution - Analysis

Solution:

$$p(t|x) = \frac{p(t)}{Z(x,\beta)} e^{-\beta d(x,t)}$$

Fix t

$\beta \rightarrow 0$

Fix x

$\beta \rightarrow \infty$

Solution - Analysis

Solution:

$$p(t|x) = \frac{p(t)}{Z(x,\beta)} e^{-\beta d(x,t)}$$

Intermediate β $\Rightarrow$ soft clustering, intermediate complexity

Varying β $\Rightarrow$

Agenda

- Motivation
- Information Theory - Basic Definitions
- Rate Distortion Theory
  - Blahut-Arimoto algorithm
- Information Bottleneck Principle
- IB algorithms
  - iIB
  - dIB
  - aIB
- Application

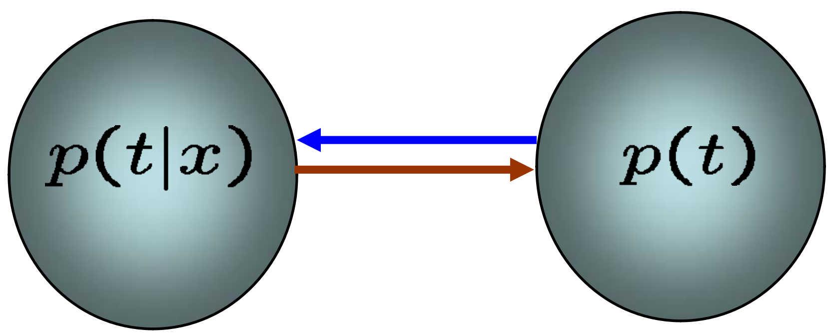
Blahut – Arimoto Algorithm

Input: $p(x),\ T,\ \beta$

Randomly init $p(t)$

$$p(t|x) = \frac{p(t)}{Z(x,\beta)} e^{-\beta d(x,t)}$$

$$p(t) = \sum_x p(x) p(t|x)$$



Optimize convex function over convex set
the minimum is global

Blahut-Arimoto Algorithm

Advantages:

- **Obtains compact clustering of the data with minimal expected distortion**

- **Optimal clustering given fixed set of representatives**

Blahut-Arimoto Algorithm

Drawbacks:

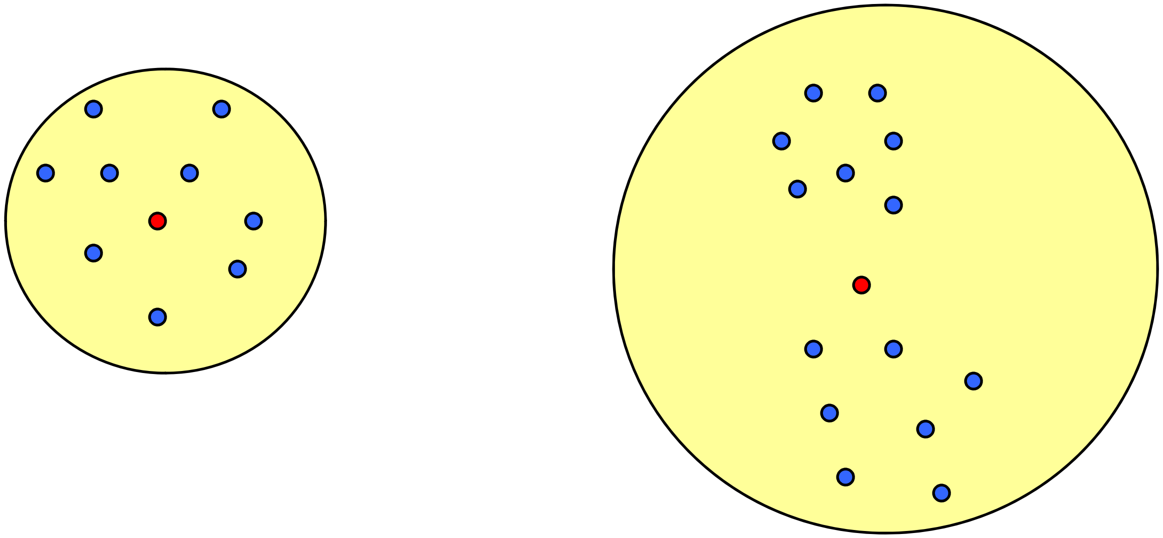
- **Distortion measure is a part of the problem setup**
  - **Hard to obtain for some problems**
  - **Equivalent to determining relevant features**

- **Fixed set of representatives**

- **Slow convergence**

Rate Distortion Theory – Additional Insights

– **Another problem would be to find optimal representatives given the clustering.**



– **Joint optimization of clustering and representatives doesn't have a unique solution. (like EM or K-means)**

Agenda

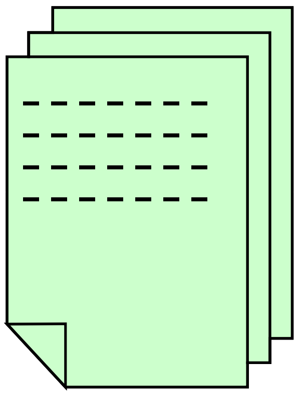
- Motivation
- Information Theory - Basic Definitions
- Rate Distortion Theory
  - Blahut-Arimoto algorithm
- Information Bottleneck Principle
- IB algorithms
  - iIB
  - dIB
  - aIB
- Application

Information Bottleneck

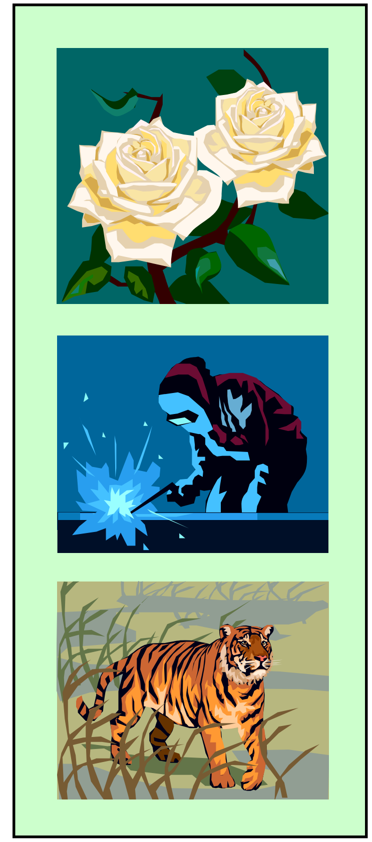
- Copes with the drawbacks of Rate Distortion approach
- Compress the data while preserving "important" (relevant) information
- It is often easier to define what information is important than to define a distortion measure.
- Replace the distortion upper bound constraint by a lower bound constraint over the relevant information

Information Bottleneck-Example

Given:

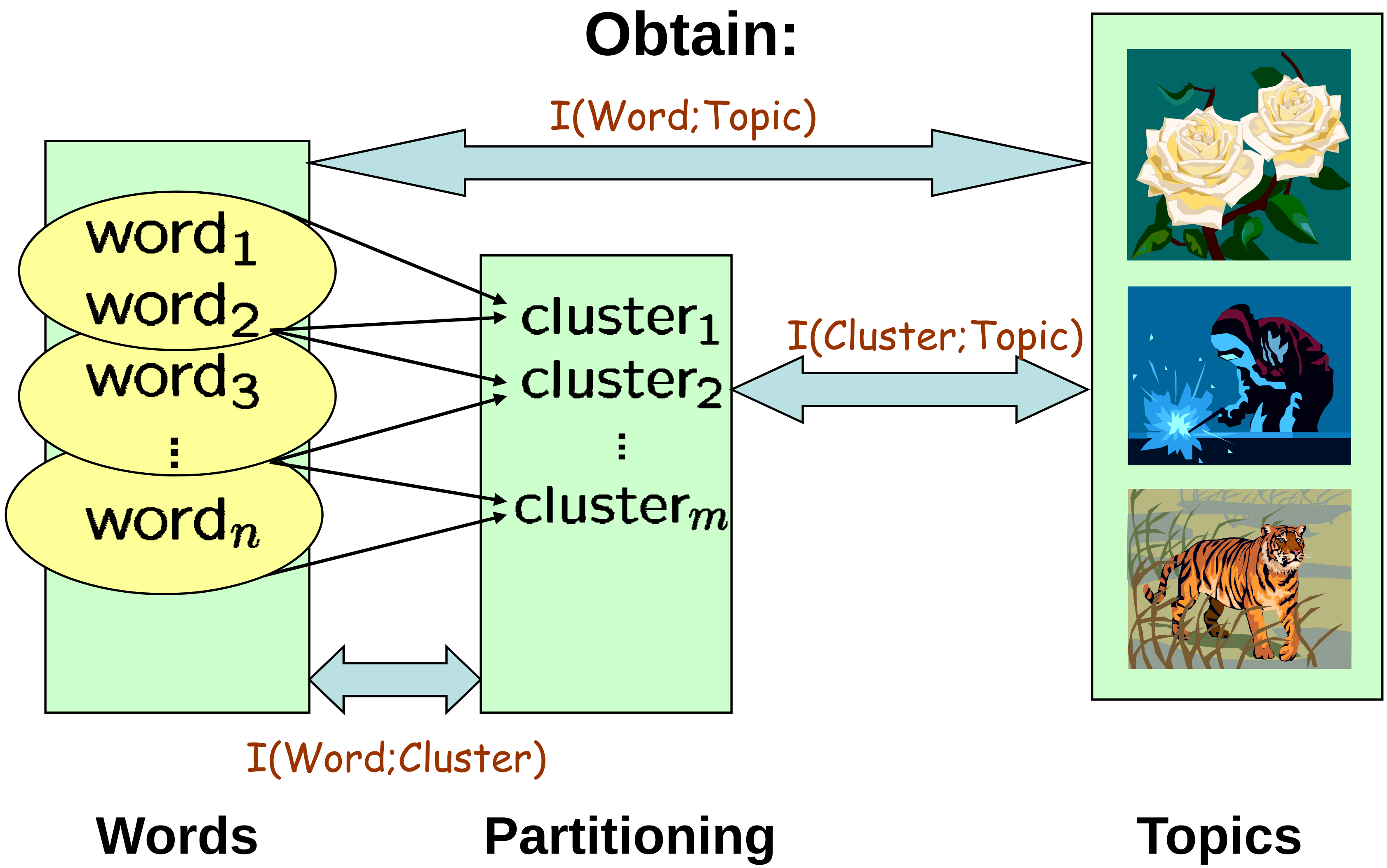


$$p(\text{word}, \text{topic})$$



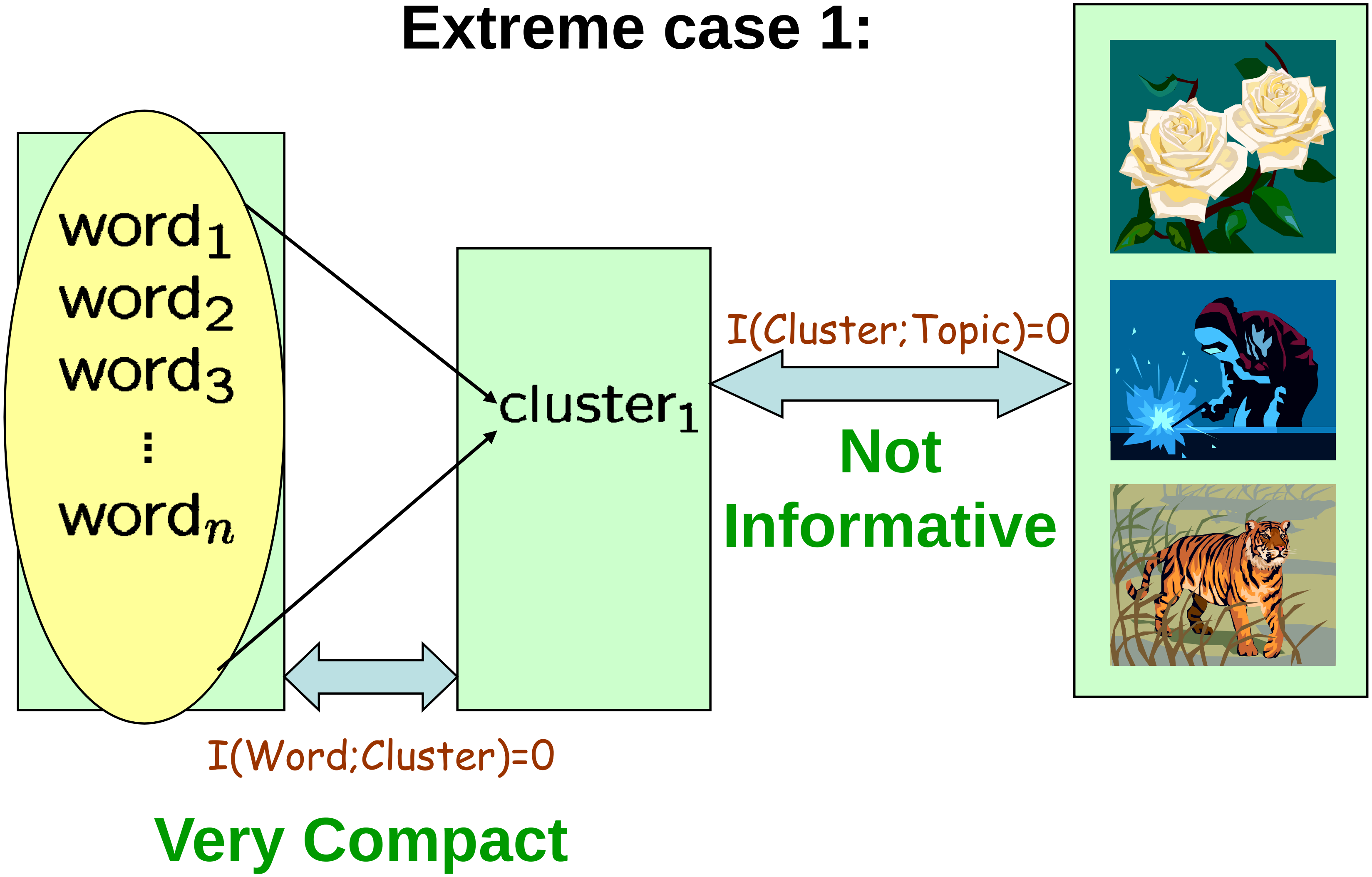
Documents **Joint prior** **Topics**

Information Bottleneck-Example



Information Bottleneck-Example

Extreme case 1:

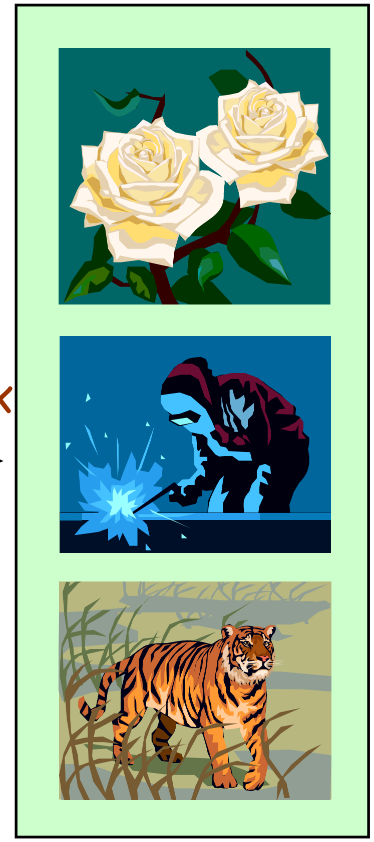


Information Bottleneck-Example

Extreme case 2:

word$_1$ → cluster$_1$
word$_2$ → cluster$_2$
word$_3$ → cluster$_3$
⋮ ⋮
word$_n$ → cluster$_n$

I(Cluster;Topic)=max



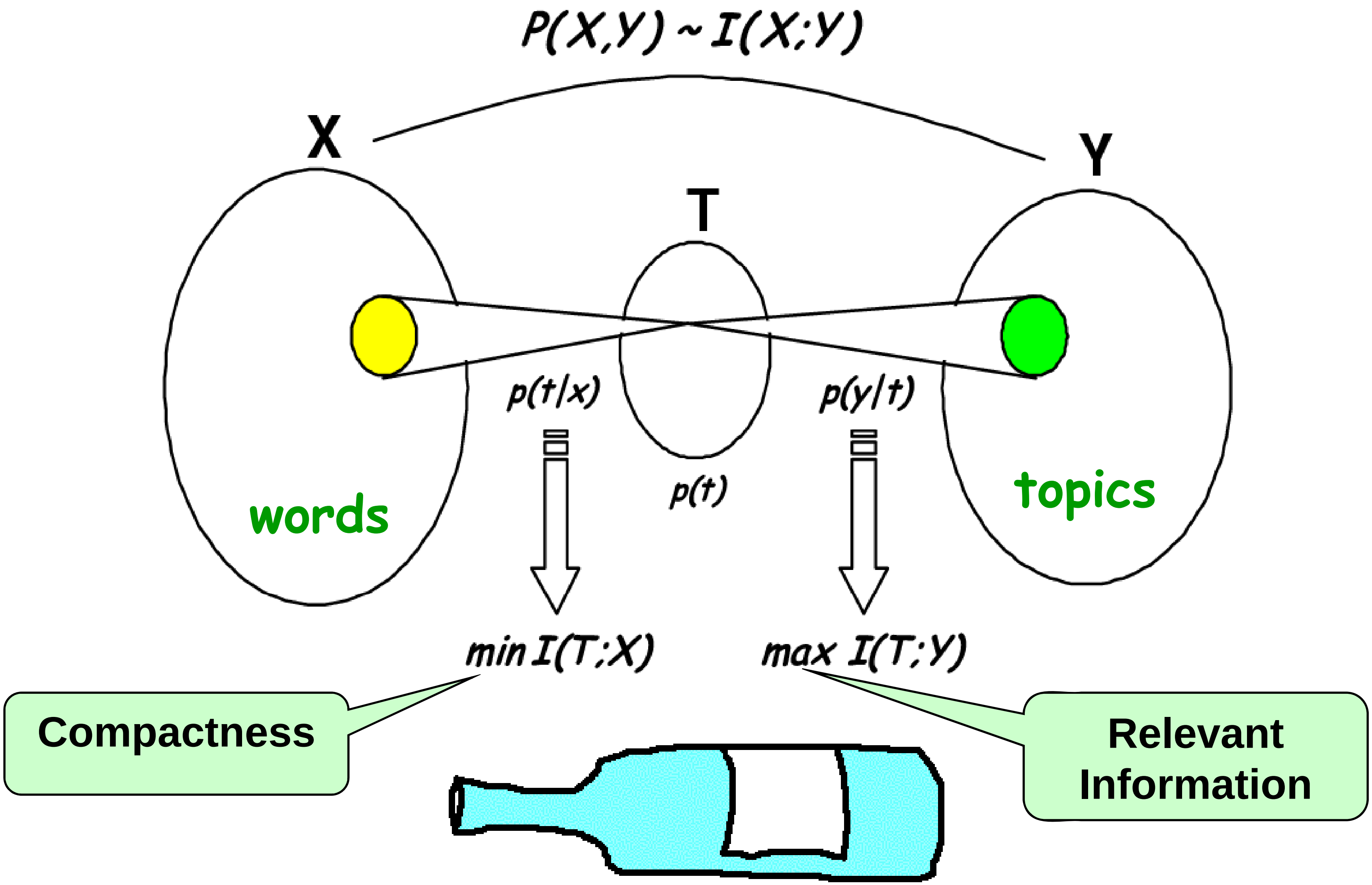
Very Informative

I(Word;Cluster)=max

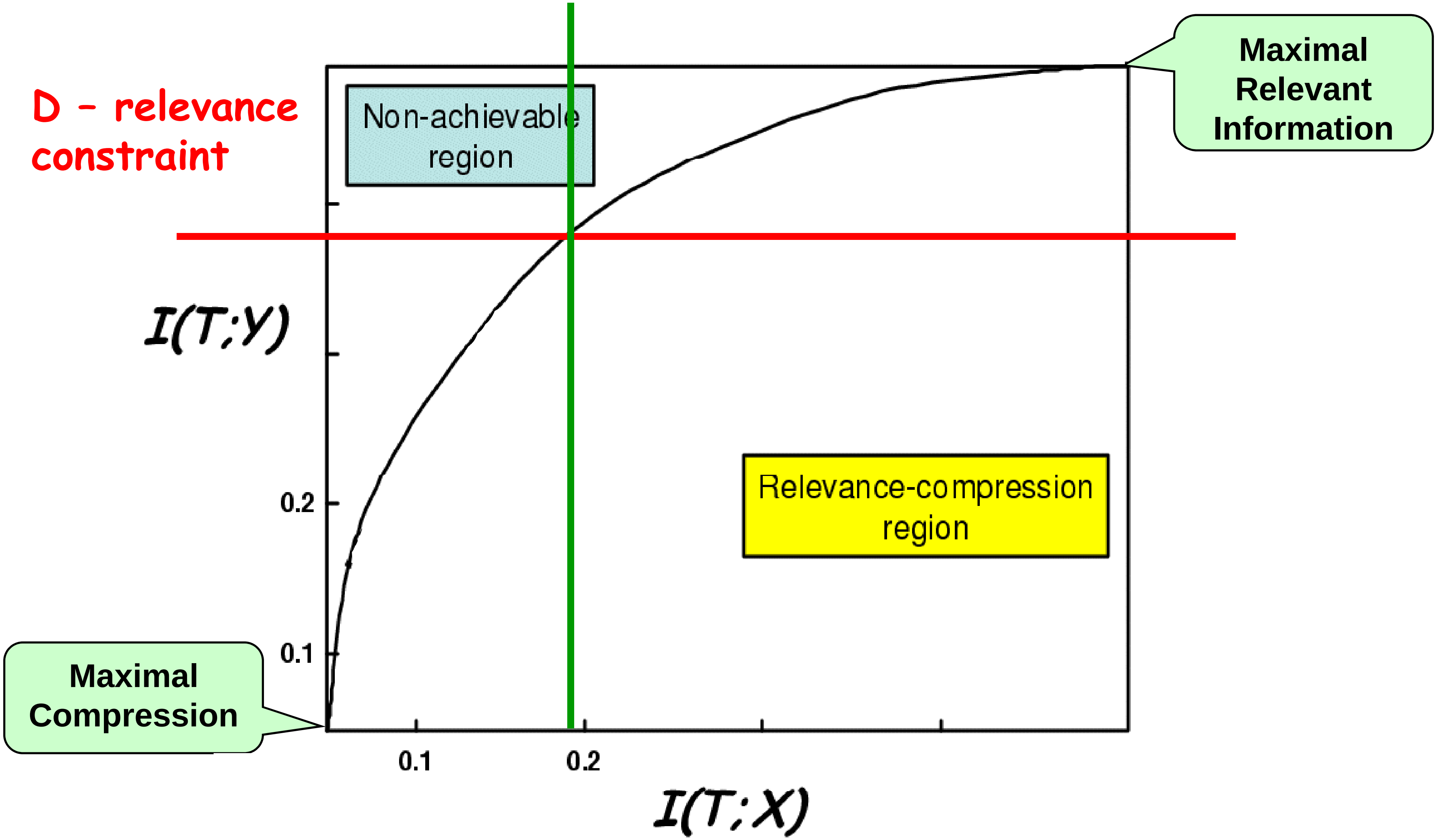
Not Compact

Minimize I(Word; Cluster) & maximize I(Cluster; Topic)

Information Bottleneck



Relevance Compression Curve



Relevance Compression Function

- **Let $\hat{D}$ be minimal allowed value of $I(T;Y)$**

Smaller $\hat{D}$ ⇒ more relaxed relevant information constraint

⇓

Stronger compression levels are attainable

- **Given relevant information constraint $\hat{D}$ Find the most compact model (with smallest $\hat{R}$)**

$$\hat{R}(\hat{D}) \equiv \min_{\{p(t|x):I(T;Y)\geq\hat{D}\}} I(T;X)$$

Relevance Compression Function

$$\hat{R}(\hat{D}) \equiv \min_{\{p(t|x): I(T;Y) \geq \hat{D}\}} I(T;X)$$

Compression Term

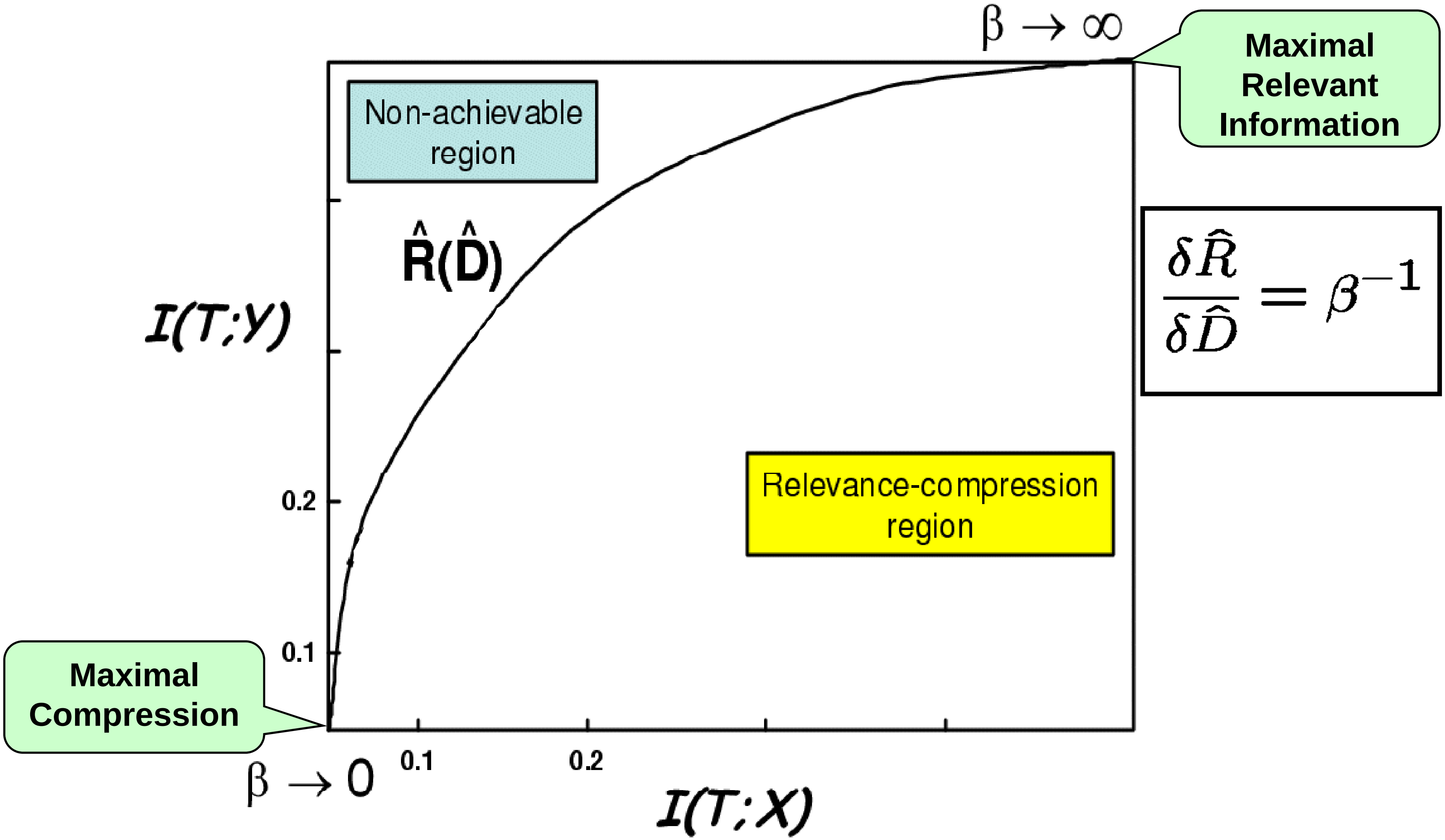
Relevance Term

$$\mathcal{L}[p(t|x)] = I(T;X) - \beta I(T;Y)$$

Lagrange Multiplier

Minimize $\mathcal{L}[p(t|x)]$ **!**

Relevance Compression Curve



Relevance Compression Function

Minimize $$\boxed{\mathcal{L}[p(t|x)] = I(T;X) - \beta I(T;Y)}$$

Subject to $$\sum_t p(t|x) = 1 \ \forall x \in X$$

The minimum is attained when $$\frac{\partial \mathcal{L}}{\partial p(t|x)} = 0$$

$$\Rightarrow \boxed{p(t|x) = \frac{p(t)}{Z(x,\beta)} e^{-\beta KL[p(y|x)||p(y|t)]}}$$

Normalization

Solution - Analysis

$$\mathcal{L}[p(t|x)] = I(T;X) - \beta I(T;Y)$$

Solution:

$$p(t|x) = \frac{p(t)}{Z(x,\beta)} e^{-\beta KL[p(y|x)||p(y|t)]}$$

Known

The solution is implicit

$$\begin{cases} p(t) = \sum_x p(x) p(t|x) \\ p(y|t) = \frac{1}{p(t)} \sum_x p(x,y) p(t|x) \end{cases}$$

Solution - Analysis

Solution: $$p(t|x) = \frac{p(t)}{Z(x,\beta)} \mathrm{e}^{-\beta KL[p(y|x)||p(y|t)]}$$

- **KL distance emerges as effective distortion measure from IB principle**

For a fixed t

When $p(y|t)$ **is similar to** $p(y|x)$

The optimization is also over cluster representatives