Preference-Based CBR

General Ideas and Basic Principles

Eyke Hüllermeier
University of Marburg
Germany

Weiwei Cheng
Amazon
Germany

GOALS AND CONTRIBUTIONS

We make a first step toward a

preference-based

methodological framework of CBR.

- in-between high-level models (like CBR cycle) and concrete implementations
- sufficiently general and abstract, so as to allow for the development of generic algorithms, for analyzing formal properties, etc.
- sufficiently concrete, so as to support the development of specific applications

PREFERENCES IN AI

"**Early work in AI focused on the notion of a goal—an explicit target that must be achieved**—and this paradigm is still dominant in AI problem solving. But as application domains become more complex and realistic, it is apparent that **the dichotomic notion of a goal**, while adequate for certain puzzles, **is too crude in general**. The problem is that in many contemporary application domains ... **the user has little knowledge about the set of possible solutions or feasible items,** and what she typically seeks is the best that's out there. But since the user does not know what is the best achievable plan or the best available document or product, she typically cannot characterize it or its properties specifically. **As a result, she will end up either asking for an unachievable goal, getting no solution in response, or asking for too little, obtaining a solution that can be substantially improved.**" [Brafman & Domshlak, 2009]

PREFERENCES IN AI

"**Early work in AI focused on the notion of a goal—an explicit target that must be achieved**—and this paradigm is still dominant in AI problem solving. But as application domains become more complex and realistic, it is apparent that **the dichotomic notion of a goal**, while adequate for certain puzzles, **is too crude in general**. The problem is that in many contemporary application domains ... **the user has little knowledge about the set of possible solutions or feasible items,** and what she typically seeks is the best that's out there. But since the user does not know what is the best achievable plan or the best available document or product, she typically cannot characterize it or its properties specifically. **As a result, she will end up either asking for an unachievable goal, getting no solution in response, or asking for too little, obtaining a solution that can be substantially improved.**" [Brafman & Domshlak, 2009]

PREFERENCES IN AI

"**Early work in AI focused on the notion of a goal—an explicit target that must be achieved**—and this paradigm is still dominant in AI problem solving. But as application domains become more complex and realistic, it is apparent that **the dichotomic notion of a goal**, while adequate for certain puzzles, **is too crude in general**. The problem is that in many contemporary application domains ... **the user has little knowledge about the set of possible solutions or feasible items,** and what she typically seeks is the best that's out there. But since the user does not know what is the best achievable plan or the best available document or product, she typically cannot characterize it or its properties specifically. **As a result, she will end up either asking for an unachievable goal, getting no solution in response, or asking for too little, obtaining a solution that can be substantially improved.**" [Brafman & Domshlak, 2009]

PREFERENCES IN AI

"**Early work in AI focused on the notion of a goal—an explicit target that must be achieved**—and this paradigm is still dominant in AI problem solving. But as application domains become more complex and realistic, it is apparent that **the dichotomic notion of a goal**, while adequate for certain puzzles, **is too crude in general**. The problem is that in many contemporary application domains ... **the user has little knowledge about the set of possible solutions or feasible items,** and what she typically seeks is the best that's out there. But since the user does not know what is the best achievable plan or the best available document or product, she typically cannot characterize it or its properties specifically. **As a result, she will end up either asking for an unachievable goal, getting no solution in response, or asking for too little, obtaining a solution that can be substantially improved.**"

[Brafman & Domshlak, 2009]

→ applies to AI in general and to CBR in particular!

→ modeling case-based experience in terms of preferences!

AGENDA

1. Basic ideas of Preference-Based CBR
2. Case-Based Inference (CBI)
3. CBR as Preference-Guided Search
4. Case Study

CASE-BASED REPRESENTATION OF EXPERIENCE

The standard representation of experience in terms of **problem/solution** pairs

$$(\boldsymbol{x}, \boldsymbol{y}) \in \mathbf{X} \times \mathbf{Y}$$

may cause disadvantages:

- assumes **existence of "correct"** (and perhaps even unique) **solution**
- assumes that a certain level of **optimality can be proved**
- a single solution does not necessarily reflect the whole experience gathered during a problem solving episode (**loss of information**)
- provides **limited guidance** if a retrieved solution fails

CASE-BASED REPRESENTATION OF EXPERIENCE

Our basic idea is to replace experiences of the form (x, y), meaning

"solution y (optimally) solves problem x",

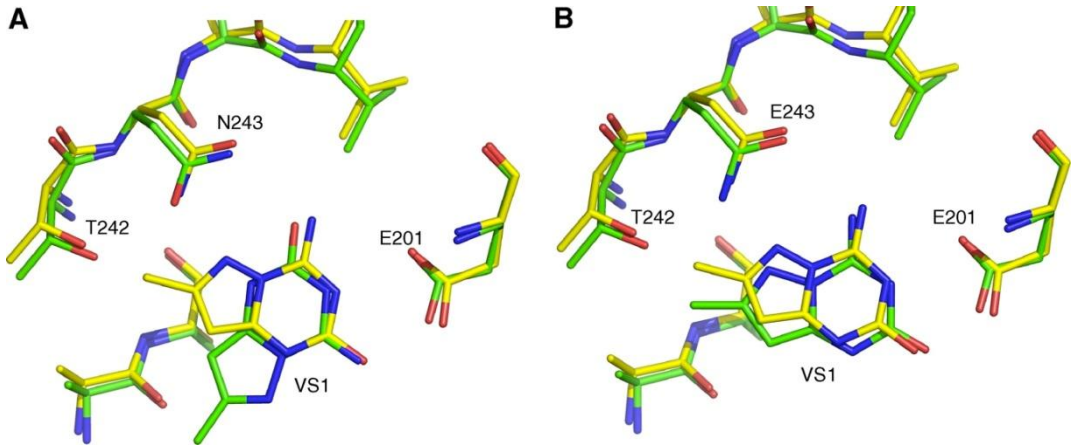
by "contextualized preferences" of the form $y \succ_x y'$, meaning

"y is better (more preferred) than y' as a solution for x".

- This is relatively **weak, qualitative knowledge**, which is easy to acquire.
- Thus, the above problems (existence of correct solutions, proof of optimality, loss of information, limited guidance) can be alleviated.
- Suggests recommendation for a new problem in the form of a **ranking**:

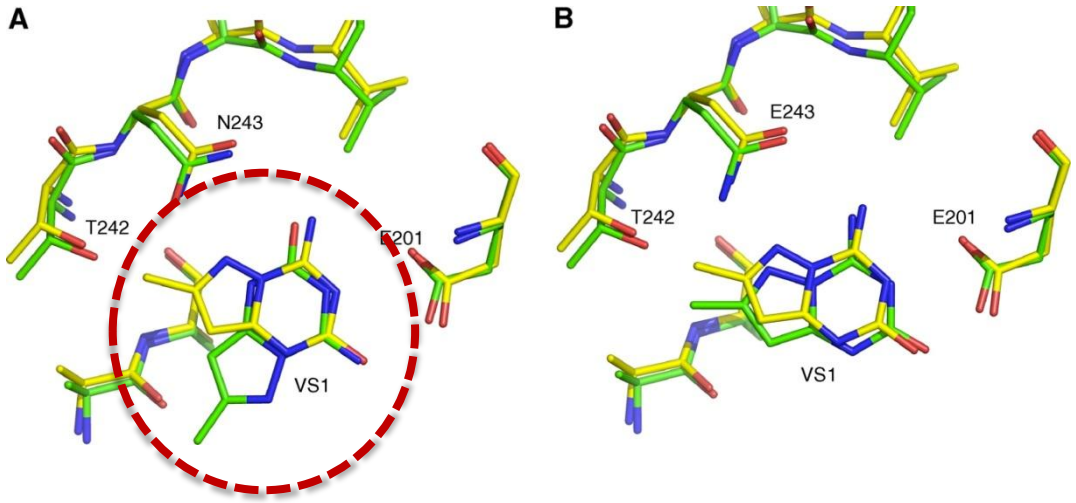
$$y_1 \succ_x y_2 \succ_x y_3 \succ_x \cdots \succ_x y_n$$

CASE-BASED REPRESENTATION OF EXPERIENCE



- **Drug discovery:** Finding ligands (small molecules) with high binding affinity to a target protein.
- **CBR perspective**: protein = problem, ligand = solution

CASE-BASED REPRESENTATION OF EXPERIENCE

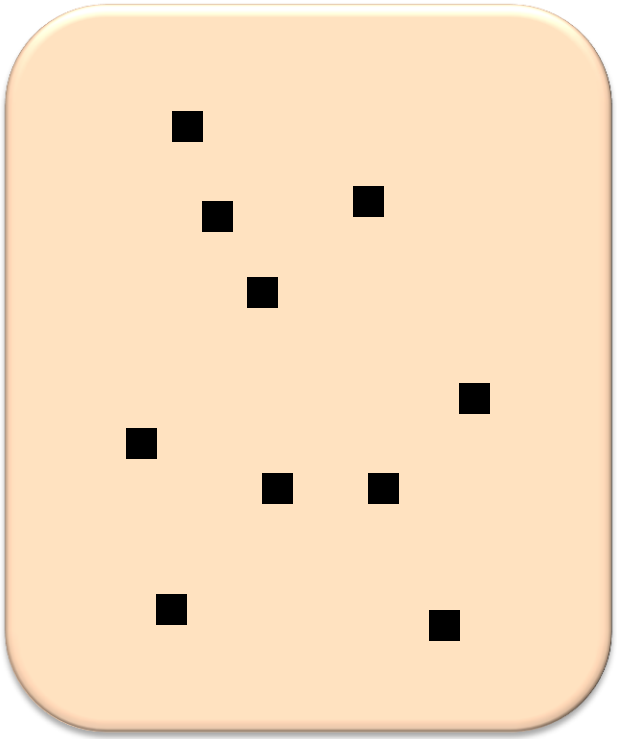


Given a protein as a "problem", molecule B is preferred as a "solution" to molecule A.

- Showing two docking poses to a domain expert (chemist, pharmacist), she can easily decide which of the molecules fits better.
- In contrast to this, she will find it difficult to assign a numerical score to an individual molecule.
- Moreover, the notion of "optimality" is not well defined (the space of molecules is huge and only partly known).

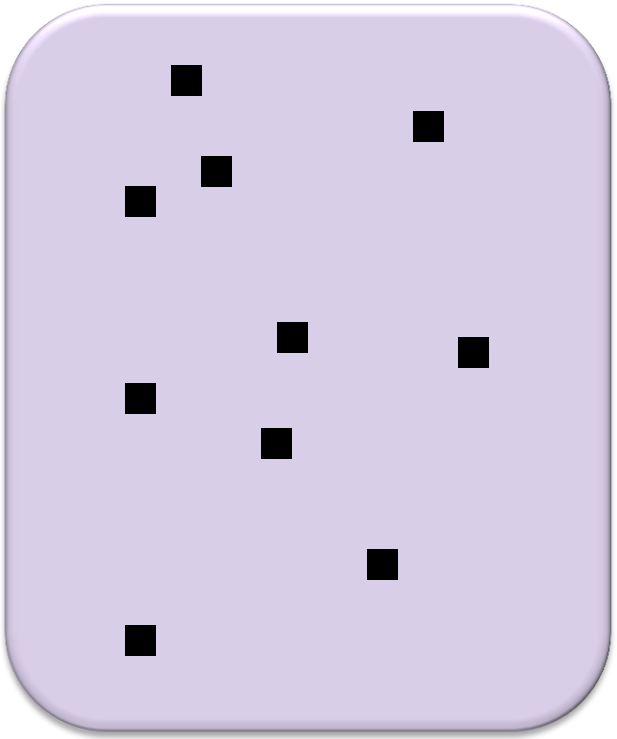
A FORMAL FRAMEWORK

problem space $\mathbf{X}$



retrieval

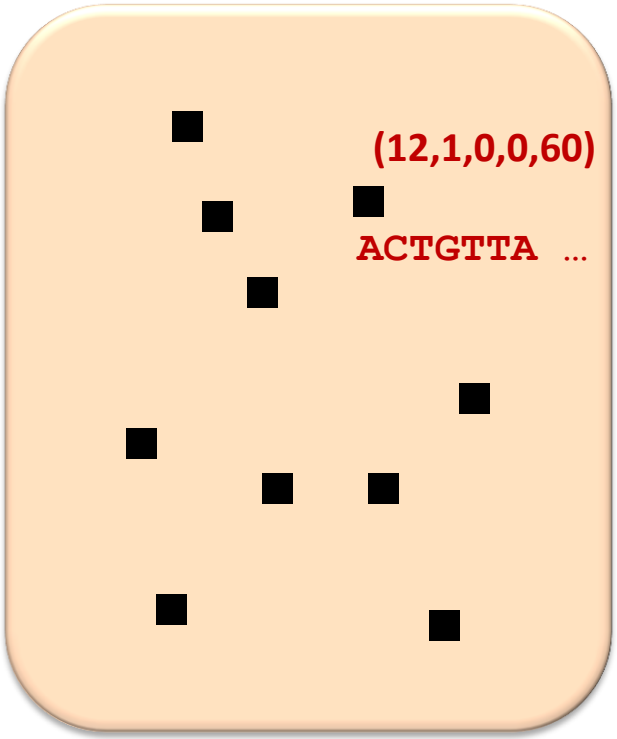
solution space $\mathbf{Y}$



inference

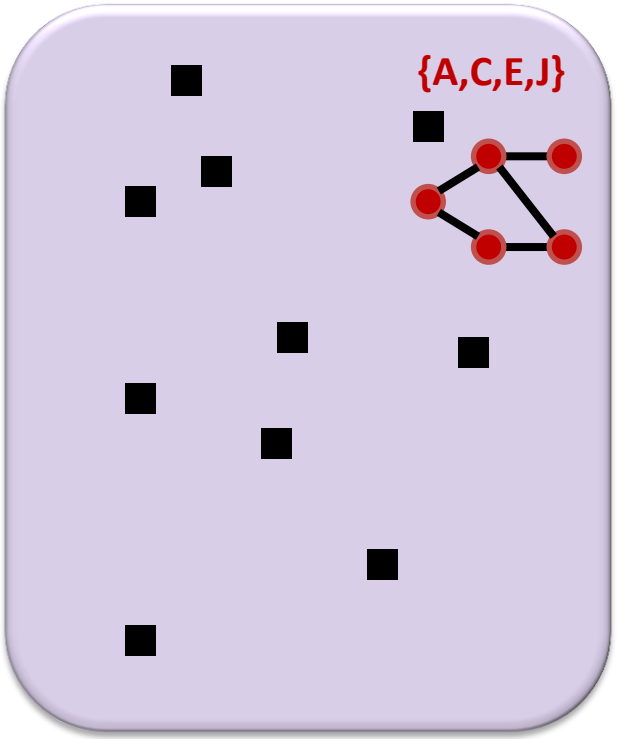
A FORMAL FRAMEWORK

problem space $\mathbf{X}$



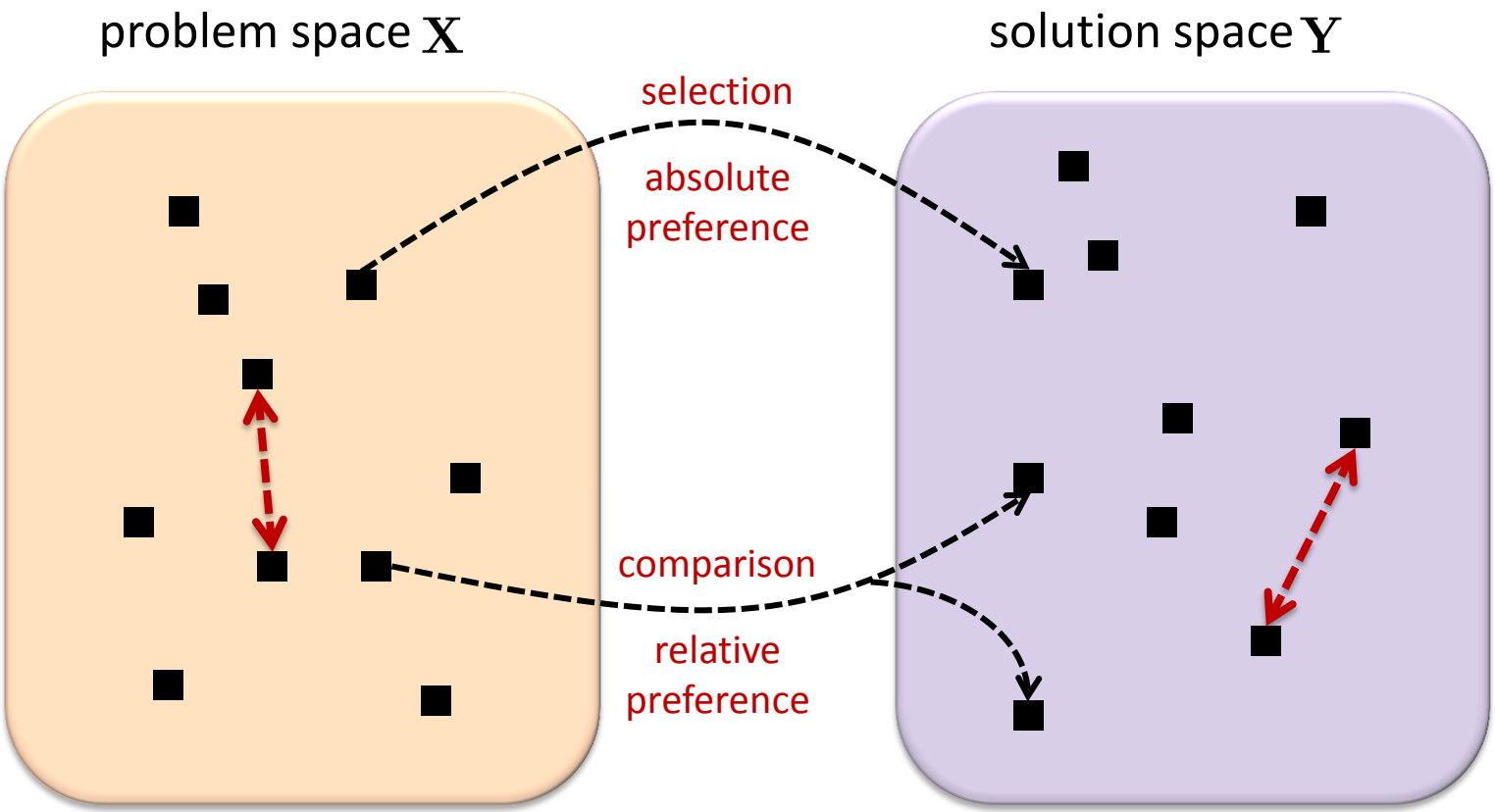
retrieval

solution space $\mathbf{Y}$



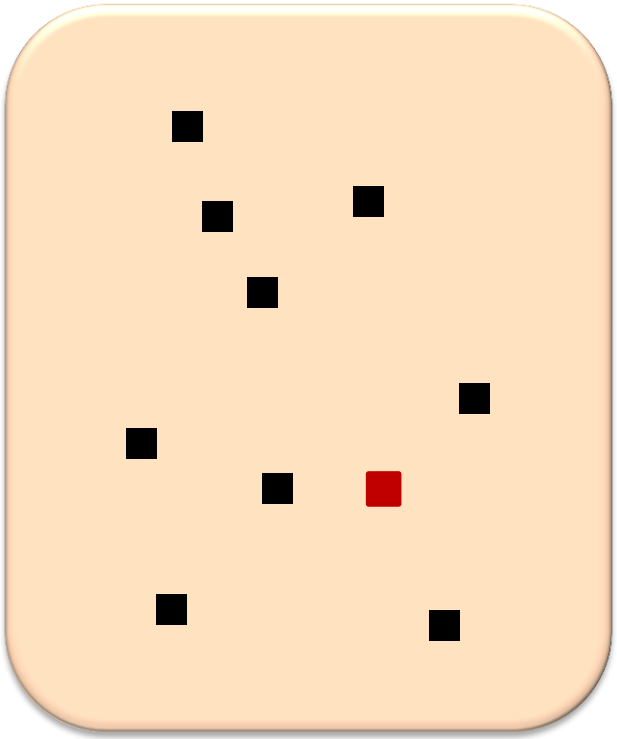
inference

A FORMAL FRAMEWORK

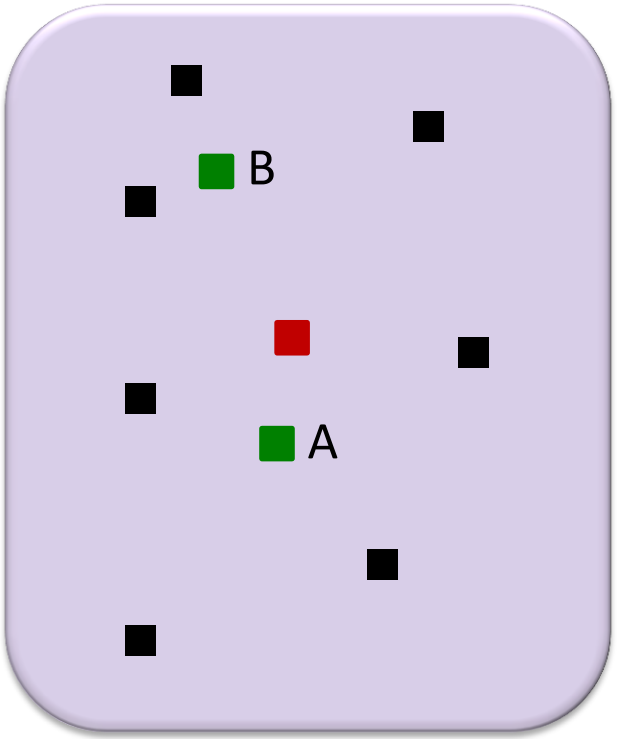


SIMILARITY/DISTANCE AND PREFERENCE

problem space $\mathbf{X}$



solution space $\mathbf{Y}$



We assume that preference is related to similarity: *the closer a candidate solution to the "ideal" solution (though perhaps fictitious) for the given problem, the more it is preferred*!

AGENDA

1. Basic ideas of Preference-Based CBR
2. **Case-Based Inference (CBI)**
3. CBR as Preference-Guided Search
4. Case Study

PREFERENCES ON SOLUTIONS: A DISCRETE CHOICE MODEL

Assuming an "ideal" solution $\boldsymbol{y}^*$ (given query $\boldsymbol{x}$), Δ induces a ranking on $\mathbf{Y}$:

$$\boldsymbol{y} \succeq_{\boldsymbol{x}} \boldsymbol{y}' \quad \Leftrightarrow \quad \Delta(\boldsymbol{y}, \boldsymbol{y}^*) \leq \Delta(\boldsymbol{y}', \boldsymbol{y}^*)$$

- In our approach, **preferences are "noisy"**, since they may come from imperfect information sources (Internet sources, humans, computer simulations, ...).
- Therefore, we adopt a **probabilistic approach** for modeling and inference!

PREFERENCES ON SOLUTIONS: A DISCRETE CHOICE MODEL

Assuming an "ideal" solution $\boldsymbol{y}^*$ (given query $\boldsymbol{x}$), Δ induces a ranking on $\mathbf{Y}$:

$$\boldsymbol{y} \succeq_{\boldsymbol{x}} \boldsymbol{y}' \quad \Leftrightarrow \quad \Delta(\boldsymbol{y}, \boldsymbol{y}^*) \leq \Delta(\boldsymbol{y}', \boldsymbol{y}^*)$$

Comparisons are made according to the **latent utility** of $\boldsymbol{y}$, namely

$$U(\boldsymbol{y}) = -\beta\, \Delta\big(\boldsymbol{y}, \boldsymbol{y}^*\big) + \epsilon \ ,$$

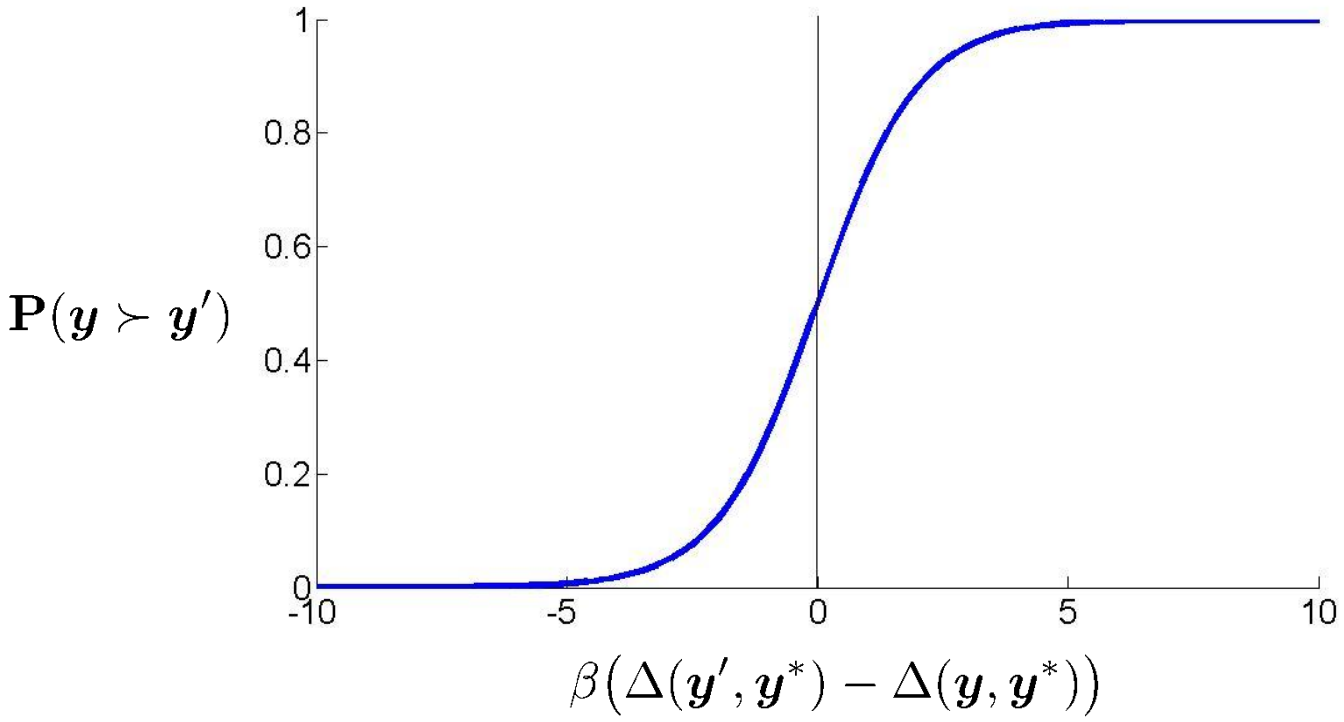
which leads to the **logit model of discrete choice**:

$$\mathbf{P}(\boldsymbol{y} \succ \boldsymbol{y}') = \frac{1}{1 + \exp\Big(-\beta\big(\Delta(\boldsymbol{y}', \boldsymbol{y}^*) - \Delta(\boldsymbol{y}, \boldsymbol{y}^*)\big)\Big)}$$

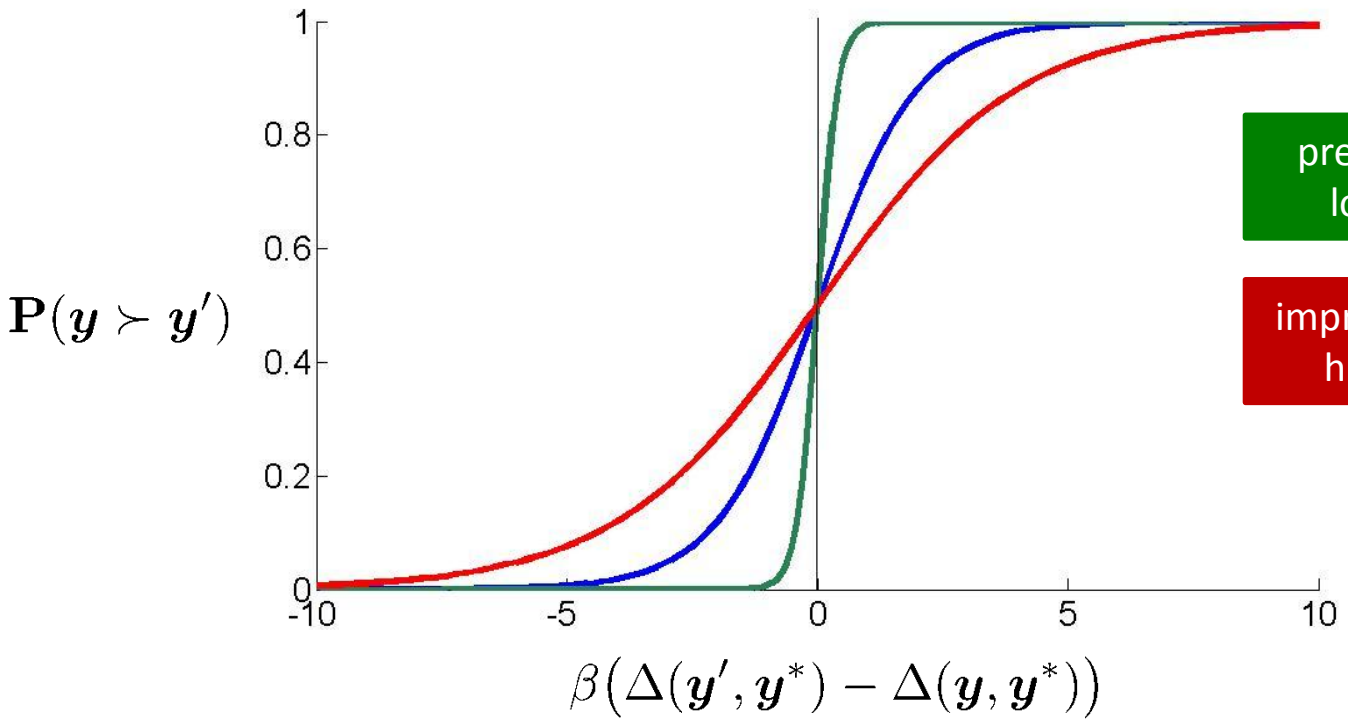
precision parameter

ideal solution

PREFERENCES ON SOLUTIONS: A DISCRETE CHOICE MODEL



PREFERENCES ON SOLUTIONS: A DISCRETE CHOICE MODEL



precise observations, low level of noise

imprecise observations, high level of noise

CASE-BASED INFERENCE: A PROBABILISTIC APPROACH

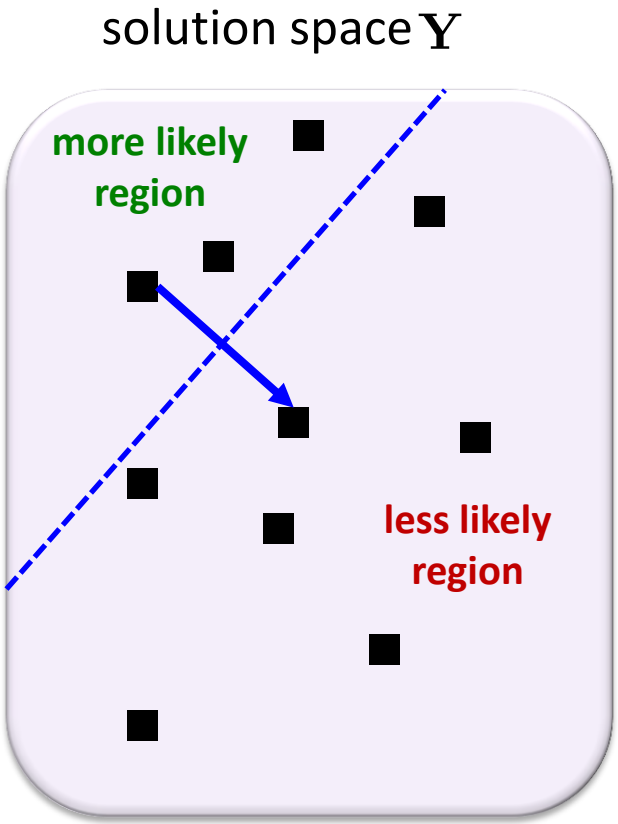
CASE-BASED INFERENCE:

Given a set of observed pairwise preferences

$$\mathcal{D} = \left\{ \boldsymbol{y}^{(i)} \succ \boldsymbol{z}^{(i)} \right\}_{i=1}^{N} ,$$

assumed to be representative for the current problem $\boldsymbol{x}$, what is the most plausible "ideal" solution for $\boldsymbol{x}$ among a given set of candidates $\mathbf{Y}_0 \subseteq \mathbf{Y}$?

CASE-BASED INFERENCE: A PROBABILISTIC APPROACH



Each pairwise preference provides a hint at the ideal solution!

CASE-BASED INFERENCE: A PROBABILISTIC APPROACH

Formal approach via maximum-likelihood estimation:

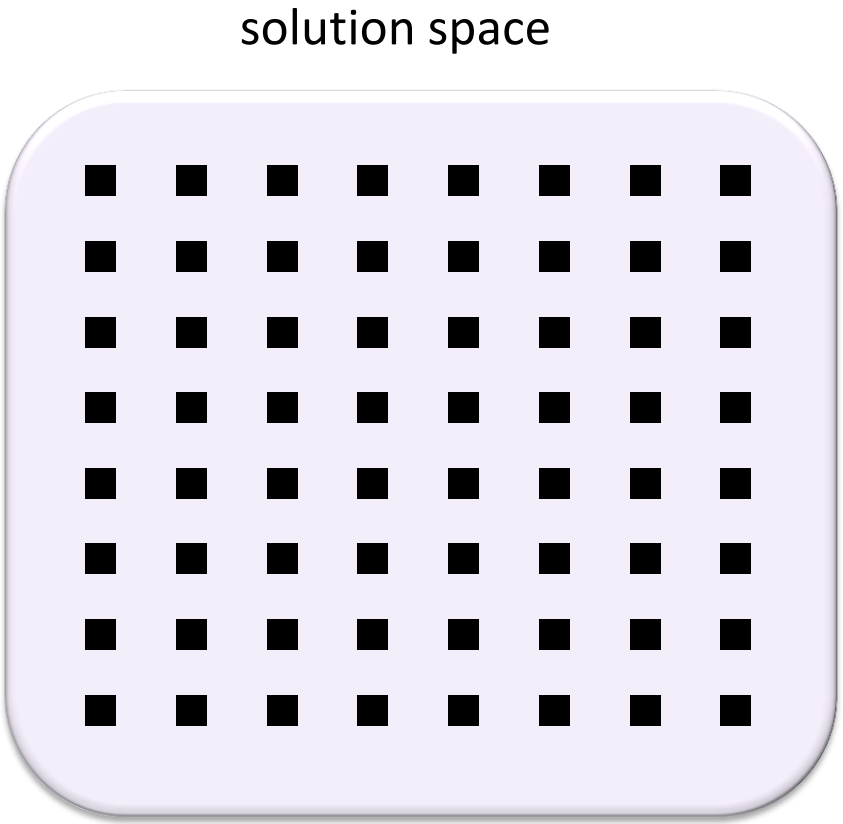
$$(\boldsymbol{y}^*, \beta^*) = \arg \max_{\boldsymbol{y} \in \mathbf{Y}_0,\, \beta \in \mathbb{R}_+} \ell(\boldsymbol{y}, \beta \,|\, \mathcal{D})$$

$$= \arg \min_{\boldsymbol{y} \in \mathbf{Y}_0,\, \beta \in \mathbb{R}_+} \sum_{i=1}^{N} \log \left(1 + \exp \left(-\beta(\Delta(\boldsymbol{z}^{(i)}, \boldsymbol{y}) - \Delta(\boldsymbol{y}^{(i)}, \boldsymbol{y})\right)\right)$$

AGENDA

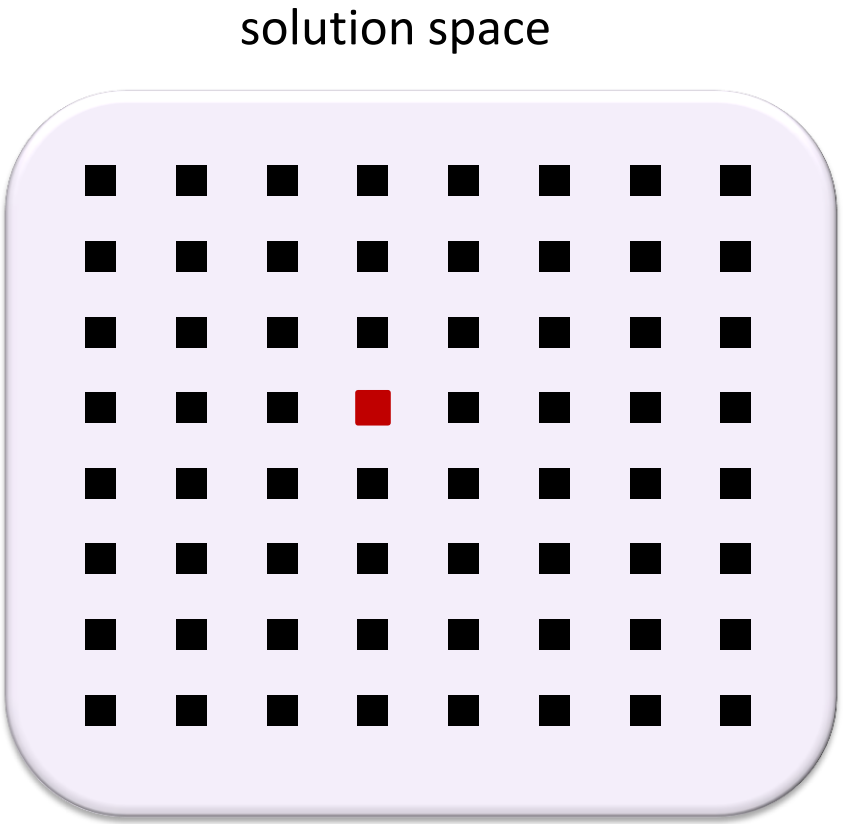
1. Basic ideas of Preference-Based CBR
2. Case-Based Inference (CBI)
3. **CBR as Preference-Guided Search**
4. Case Study

CBR AS PREFERENCE-GUIDED SEARCH



Case-based problem solving as (heuristic) search in the solution space.

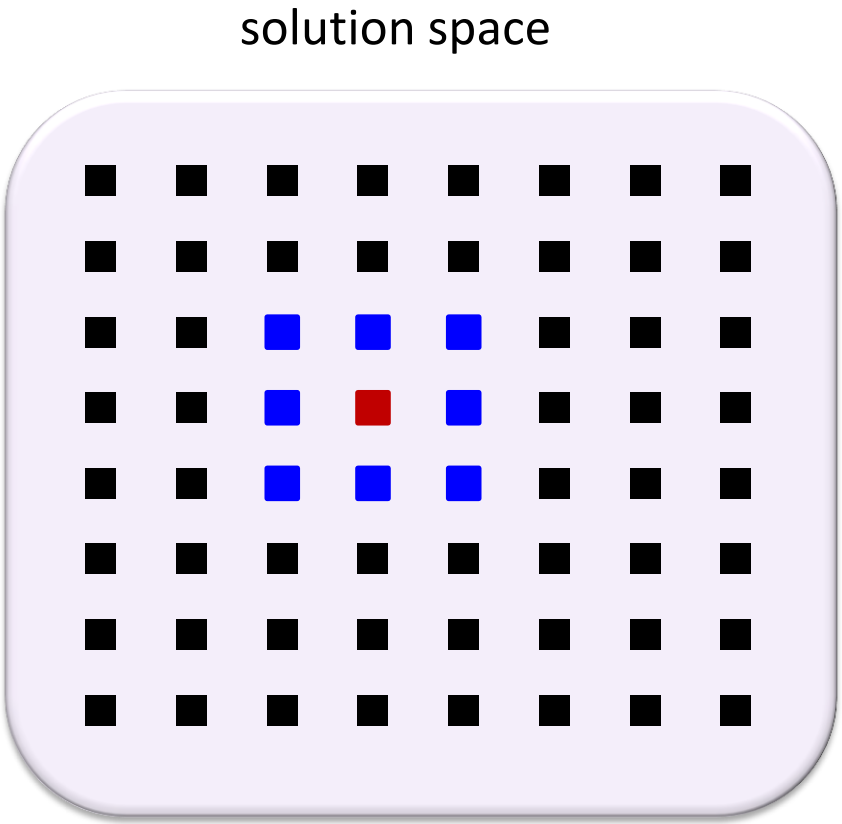
CBR AS PREFERENCE-GUIDED SEARCH



- Start with an **initial solution**

Case-based problem solving as (heuristic) search in the solution space.

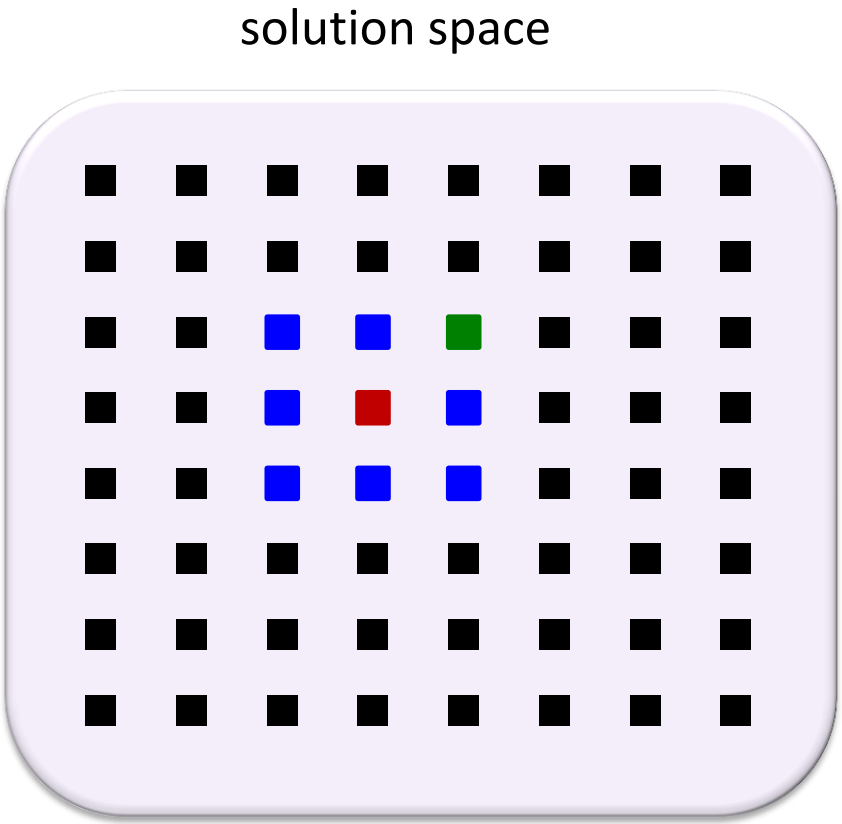
CBR AS PREFERENCE-GUIDED SEARCH



- Start with an initial solution
- Consider the **neighbors** of the current solutions as new candidates.

Case-based problem solving as (heuristic) search in the solution space.

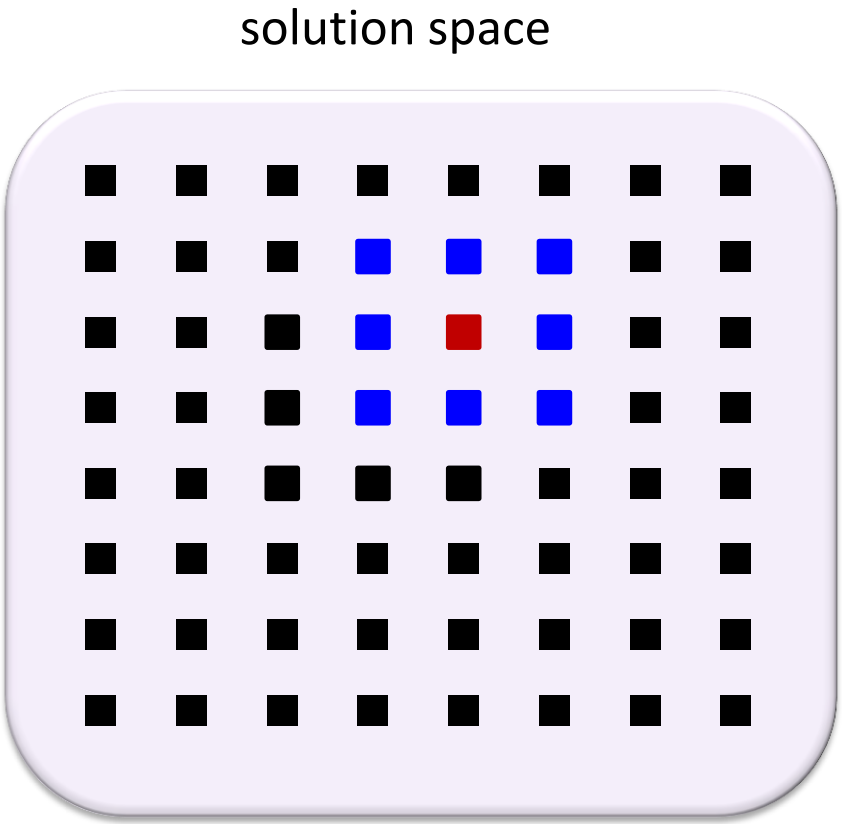
CBR AS PREFERENCE-GUIDED SEARCH



- Start with an initial solution
- Consider the neighbors of the current solutions as new candidates.
- Select a **promising neighbor**, compare with **current solution** and adopt the better one.

Case-based problem solving as (heuristic) search in the solution space.

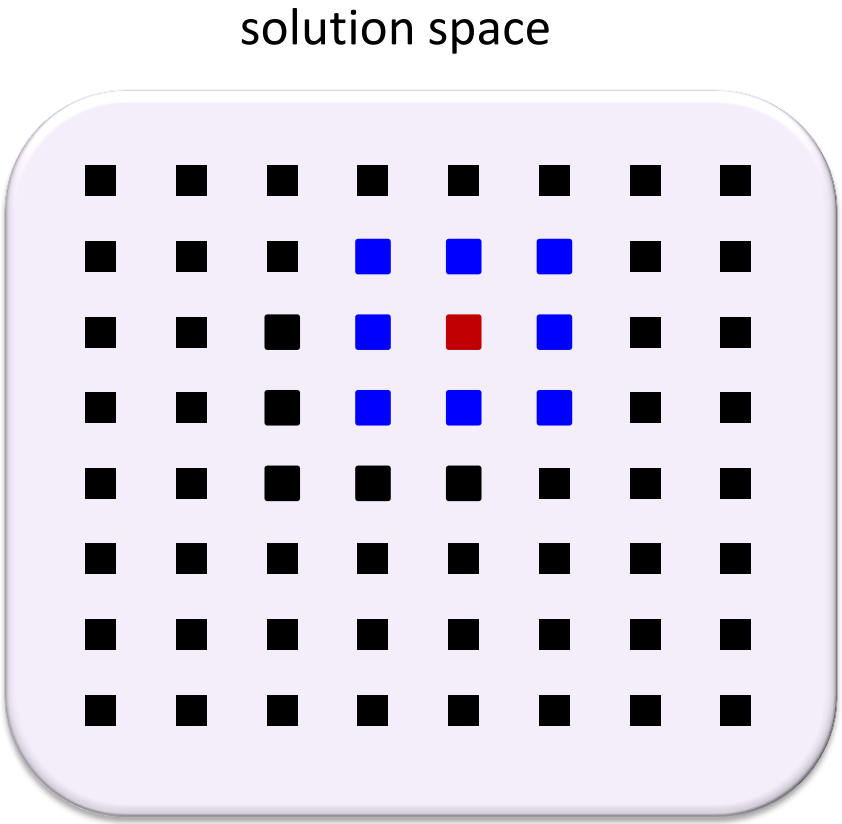
CBR AS PREFERENCE-GUIDED SEARCH



- Start with an initial solution
- Consider the neighbors of the current solutions as new candidates.
- Select a promising neighbor, compare with current solution and adopt the better one.
- Repeat till no further improvement or maximum number of iterations reached.

Case-based problem solving as (heuristic) search in the solution space.

CBR AS PREFERENCE-GUIDED SEARCH



- **Start with an initial solution**
- Consider the neighbors of the current solutions as new candidates.
- **Select a promising neighbor**, compare with current solution and adopt the better one.
- Repeat till no further improvement or maximum number of iterations reached.

We seek to support these steps through case-based experience!

PREF-CBR CASE BASE

x1	y12 ≻ y72	y42 ≻ y41	y76 ≻ y21	y42 ≻ y72
x2	y05 ≻ y53	y92 ≻ y43	y32 ≻ y56	y65 ≻ y84
x3	y39 ≻ y37	y33 ≻ y67	y65 ≻ y76	y76 ≻ y37
x4	y72 ≻ y98	y47 ≻ y27	y34 ≻ y34	y76 ≻ y65
x5	y39 ≻ y49	y29 ≻ y81	y32 ≻ y26	y76 ≻ y11
x6	y46 ≻ y11	y46 ≻ y28	y68 ≻ y28	y22 ≻ y42

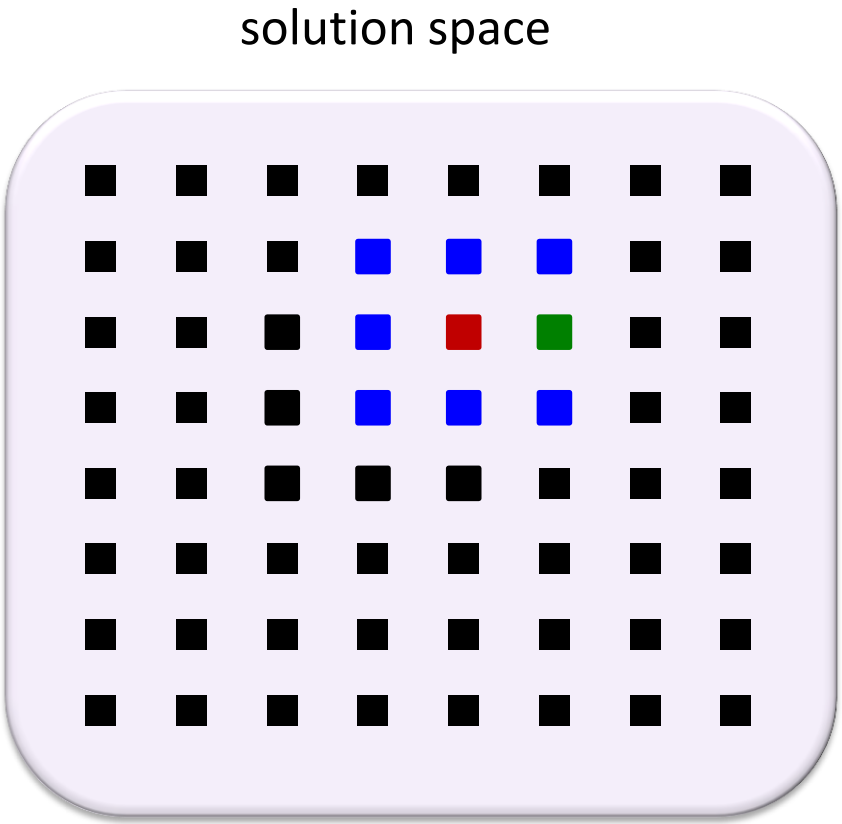
Problems are stored together with observed pairwise preferences.

FINDING AN INITIAL SOLUTION

x1	y12 ≻ y72	y42 ≻ y41	y76 ≻ y21	y42 ≻ y72
x2	y05 ≻ y53	y92 ≻ y43	y32 ≻ y56	y65 ≻ y84
x3	y39 ≻ y37	y33 ≻ y67	y65 ≻ y76	y76 ≻ y37
x4	y72 ≻ y98	y47 ≻ y27	y34 ≻ y34	y76 ≻ y65
x5	y39 ≻ y49	y29 ≻ y81	y32 ≻ y26	y76 ≻ y11
x6	y46 ≻ y11	y46 ≻ y28	y68 ≻ y28	y22 ≻ y42

- Given an new problem, find the **nearest neighbors** in the case base and collect the associated preferences into an initial preference set.
- The **initial solution** is then found by applying CBI to this set of preferences (with the complete solution space as candidates).

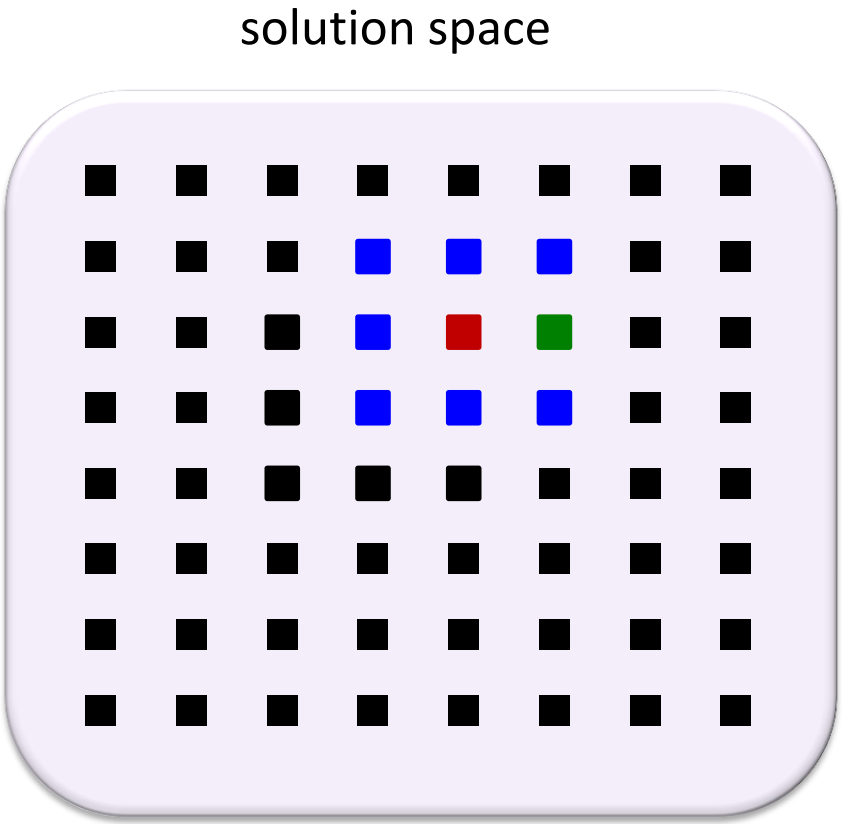
CBR AS PREFERENCE-GUIDED SEARCH



- In each iteration, CBI is applied to the **neighbors of the current solution** to find the **most promising candidate**.

We seek to support these steps through case-based experience!

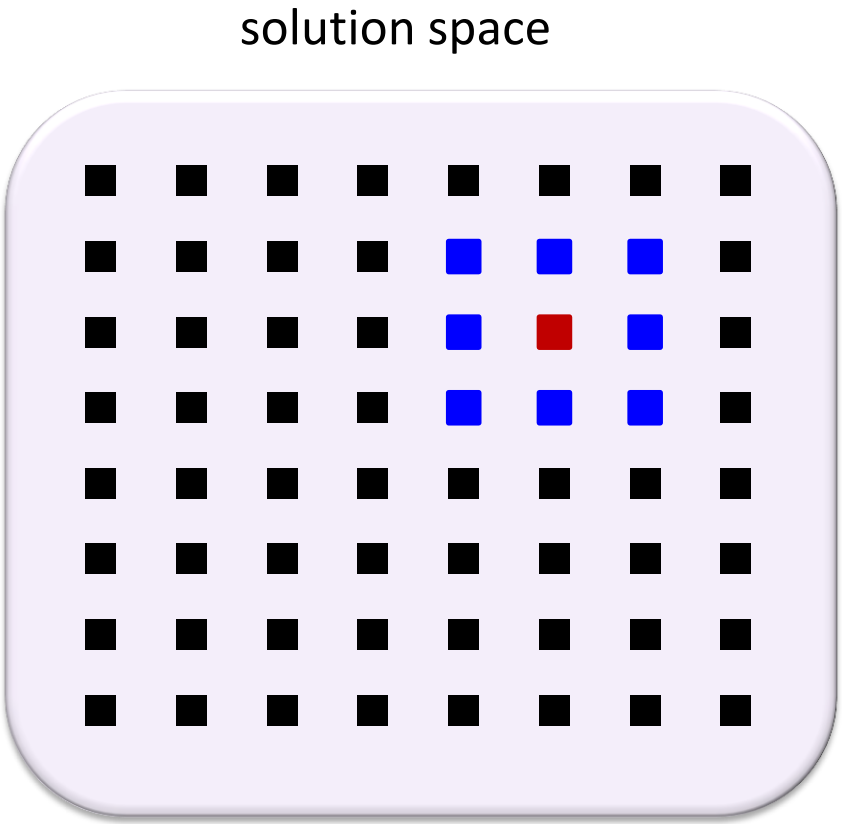
CBR AS PREFERENCE-GUIDED SEARCH



- In each iteration, CBI is applied to the **neighbors of the current solution** to find the **most promising candidate**.
- The two solutions are compared, the better one is adopted, and the **new preference is added to preference set.**

We seek to support these steps through case-based experience!

CBR AS PREFERENCE-GUIDED SEARCH



- In each iteration, CBI is applied to the **neighbors of the current solution** to find the **most promising candidate**.
- The two solutions are compared, the better one is adopted, and the **new preference is added to preference set.**
- The process stops after a predefined number of iterations, and the current best solution is returned.

We seek to support these steps through case-based experience!

PREF-CBR CASE BASE

x1	y12 ≻ y72	y42 ≻ y41	y76 ≻ y21	y42 ≻ y72
x2	y05 ≻ y53	y92 ≻ y43	y32 ≻ y56	y65 ≻ y84
x3	y39 ≻ y37	y33 ≻ y67	y65 ≻ y76	y76 ≻ y37
x4	y72 ≻ y98	y47 ≻ y27	y34 ≻ y34	y76 ≻ y65
x5	y39 ≻ y49	y29 ≻ y81	y32 ≻ y26	y76 ≻ y11
x6	y46 ≻ y11	y46 ≻ y28	y68 ≻ y28	y22 ≻ y42
x7	Y62 ≻ y22	Y62 ≻ y81	Y71 ≻ y62	Y77 ≻ y71

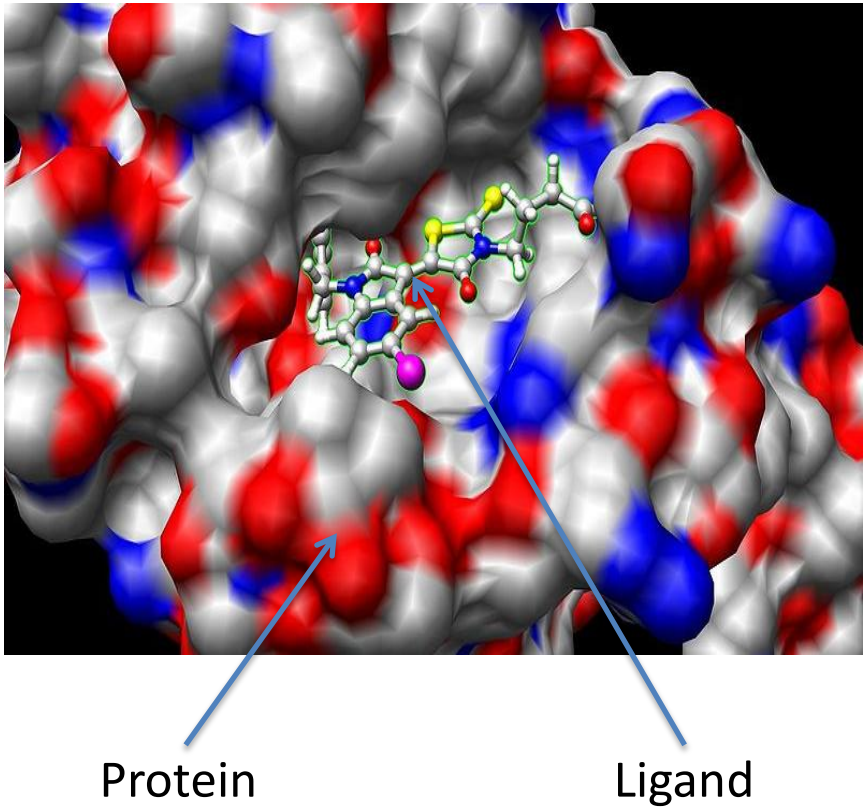
The new problem is stored together with the pairwise preferences collected during the problem solving process.

AGENDA

1. Basic ideas of Preference-Based CBR
2. Case-Based Inference (CBI)
3. CBR as Preference-Guided Search
4. **Case Study**

APPLICATION TO DRUG DISCOVERY

- Ligands (small molecules) bind to protein surface, thereby blocking or enhancing its biochemical activity.
- Identification and selection of ligands targeting a specific protein is of high interest for drug development and de-novo design.



APPLICATION TO DRUG DISCOVERY

Search for a ligand that well interacts with a target protein!

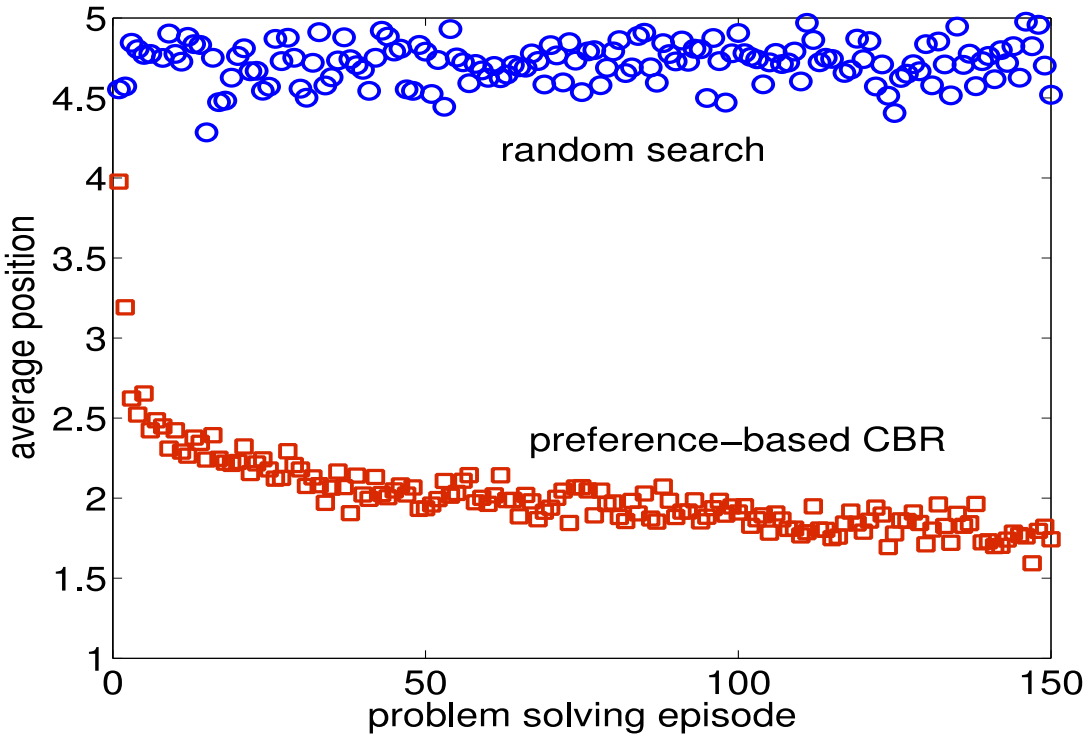
CBR perspective

- problems = proteins, solutions = ligands (molecules)
- **Problem similarity**: A measure that compares proteins in terms of spatial and physicochemical properties of their respective binding sites
- **Solution similarity**: Tanimoto similarity of molecules (SMILES representation) to determine similarity between ligands

APPLICATION TO DRUG DISCOVERY

- Experiments done with a data set consisting of 588 proteins and 38 molecules.
- For each protein/molecule pair, the data contains an affinity score (pairwise binding energy) computed by a docking tool (these scores are very noisy).
- We make use of these scores in order to mimic a human expert: Given a protein and two candidate molecules, the expert‘s preference is determined by the corresponding affinity scores.

APPLICATION TO DRUG DISCOVERY



Solution quality measured in terms of the position of the solution returned by Pref-CBR in the ground-truth ranking (ligand ranked according to binding affinity to the target protein).

SUMMARY & CONCLUSION

- Our goal is a **methodological framework of preference-based CBR** disposing of a sound theoretical basis while accommodating a wide spectrum of potential applications.
- Two main contributions so far:
  - A basic (probabilistic) method for **case-based inference**.
  - A generic framework of **CBR as preference-guided search**, formalizing CBR as heuristic search in the solution space, guided through case-based experience.
- Pref-CBR is especially suitable if candidate solutions can be **compared only qualitatively** and **comparisons are expensive** (e.g., involving human interaction).

OUTLOOK

- Ideally, a user can easily “parameterize” the framework by choosing the **type of solution space** and the **distance measure** on this space, while the methods themselves are completely generic.
- Our approach still needs to be instantiated for **different types of solution spaces**.
- Besides, **other CBR issues** need to be addressed (case base maintenance, efficient retrieval, etc.)