

# Transforming CodeNet to dataset for code translation

*Proposals*

Vladimir Zolotov

*January 5, 2020*

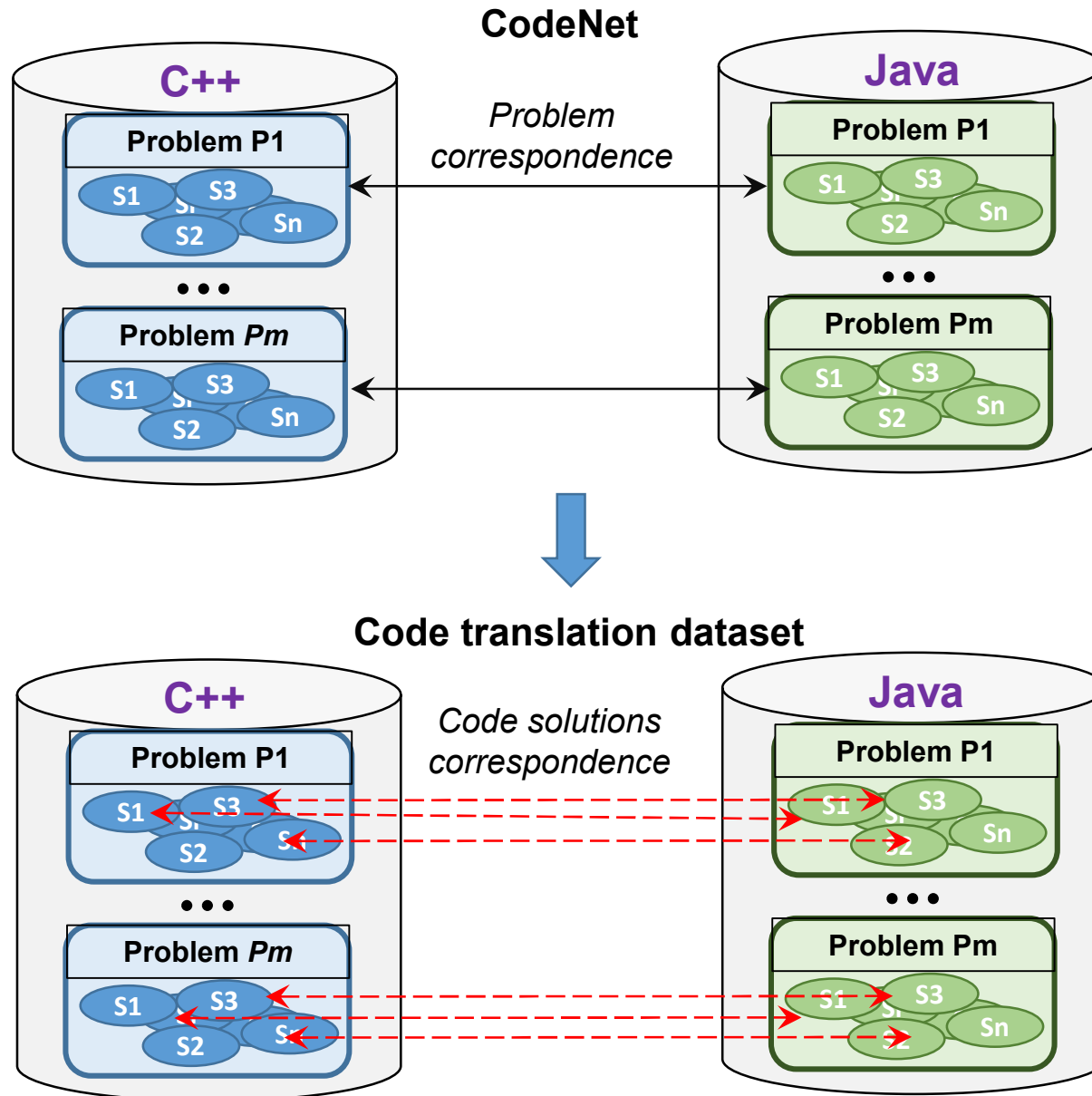
# Problem formulation

- ML code translation requires large labeled dataset:
  - Pairs of similar code samples in source and target languages
- CodeNet has code samples labeled with problems solved:
  - Problems are solved in different languages: C, C++, Java, Python, etc.
  - Only ~ 2000 problems in C++ (most popular language in CodeNet)
  - However more than 4,000,000 solutions
  - Many solution samples of each problem
    - From several to almost 20,000

## **Problem:**

- Build code translation dataset from CodeNet
  - Compute relation between problem solutions in different languages
    - Ideally the relation is 1-to-1
  - Solution code in language 1 corresponds to similar code in language 2
    - Similarity is very important to for correct training of ML code translation engine

# Problem formulation illustrated

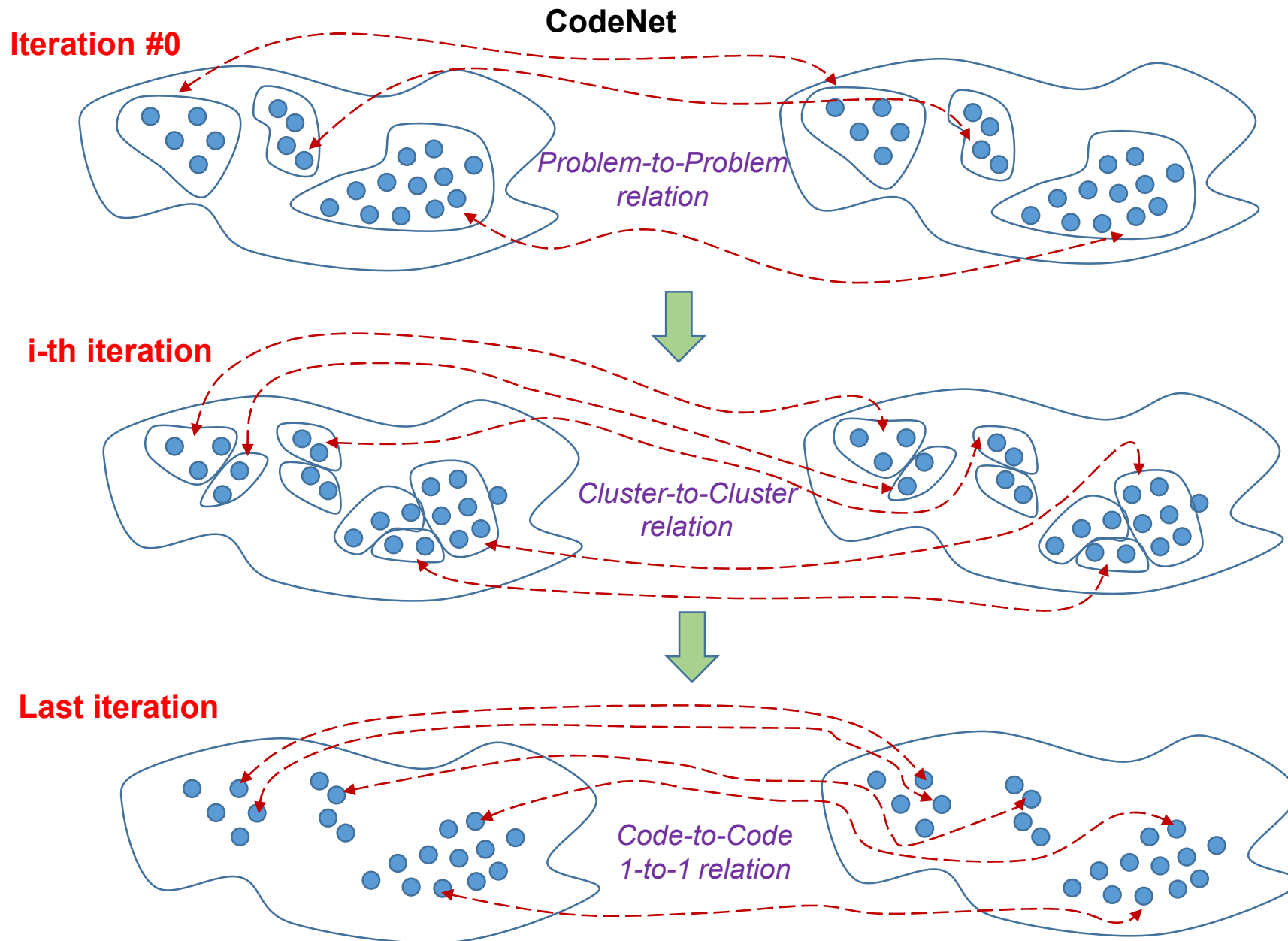


# Plan of solution



1. Develop and train Cross-Language similarity DNN-analyzer:
  - Predict if 2 source codes written in 2 languages (C++ and Java) solve the same or different problems
  - DNN output is “probability” of similarity of 2 codes
    - Metric of similarity of 2 codes written in different languages
2. Construct relation between codes written in different languages:
  - Two source codes are related if:
    1. They solve the same problem
    2. Similarity between them is higher than a threshold
  - The relation is Many-to-Many, as the similarity metric is not ideal
3. Retrain Cross-Language similarity DNN-analyzer using newly constructed similarity relation
  - Make similarity prediction more correct
4. Improve relation between codes written in different languages
  - Use newly computed metric of similarity
  - Improved relation is more selective: closer to 1-to-1 relation
5. Repeat improving relation till it is as close to 1-to-1 as possible
  - The resulting relation can be used for training code translation ML engine

# Evolution of relation between code solutions



# Difficulties and Risks



- Data related difficulties:

- Possibly not many problems have solution code in both languages
- Same problems have different number of solutions in different languages
  - 1-to-1 relation between solution source codes is impossible
- Even similar solutions can be too different for ML translation engine
  - Though is not clear what are good samples for training ML translation engine

- Development items:

- Accurate cross-Language similarity DNN-analyzer
  - So far only inter-language similarity DNN-analyzer with ~95% accuracy
- Current similarity DNN-analyzer considers only sequence operators and key words
  - To extend it to consider variables, functions, and program structure through syntax tree
- Fast and efficient clustering engine for particularizing code relation
  - Non-trivial problem as similarity metric is not Euclidian and has many peculiarities
  - Very large size of the problem:
    - Millions of code solutions, quadratic number of potential candidates for relations

# Conclusions



- Problem of constructing dataset for training ML code translation is *interesting, novel and hard*
- Its solution is useful not only for CodeNet but as a general technique for constructing datasets for training ML code translation
- No guarantee for success due to both algorithmic and dataset issues
- Cross-language similarity DNN-analyzer is planned irrespective to decision to work on ML code translation
- Improvement of code similarity DNN-analyzer by using syntax trees is planned irrespective to decision to work ML code translation
- Experiments with cross-language similarity analysis will clarify feasibility of transforming CodeNet into dataset for training ML engine