Code Clones

An Introduction to Code Duplication

Emanuel Giger

University of Zurich, Switzerland

University of Zurich
Department of Informatics

s.e.a.l.
software evolution & architecture lab

What is it?

Similar source code segments that are found in different places of a system.

- What is similar?
- How to detect?

Similar?

```
for(int = 0; i < 10; i++){
   System.out.println("Number:" + 1);
}
```

=

```
for(int = 0; i < 10; i++){
   System.out.println("Number:" + 1);
}
```

?

```
for(int = 0; i < 10; i++){
   System.out.println("Number:" + 1);
}
```

=

```
for(int = 0; i < 100; i++){
   System.out.println("Zahl:" + 1);
}
```

?

```
for(int = 0; i < 10; i++){
   System.out.println("Number:" + 1);
}
```

=

```
int = 0;
while(i < 100){
   System.out.println("Zahl:" + 1);
   i++;
}
```

?

=

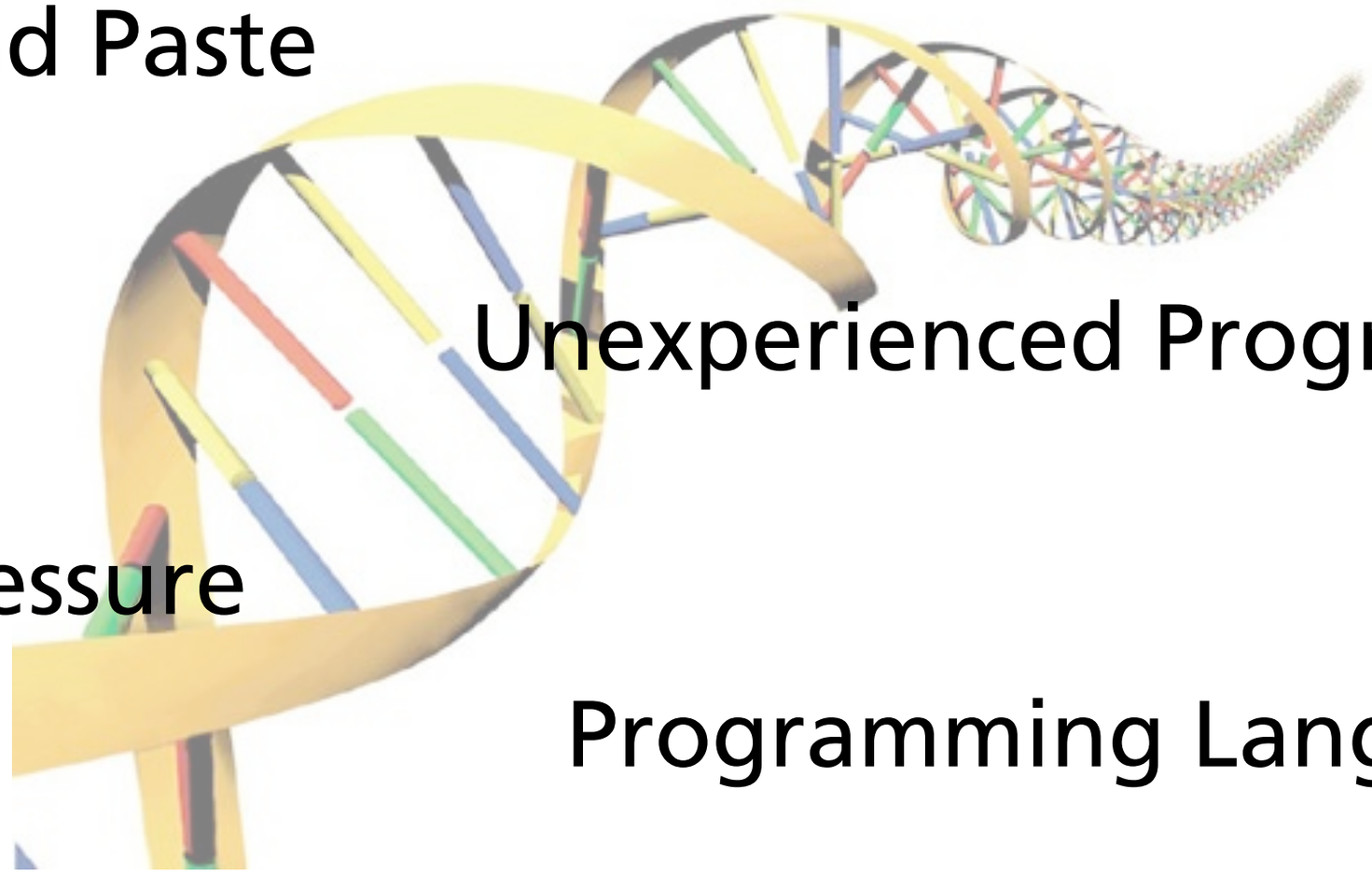
?

Similar?

Copied artifacts range from expressions, to functions, to data structures, and to entire subsystems.

Why cloning?

Copy and Paste



Unexperienced Programmer

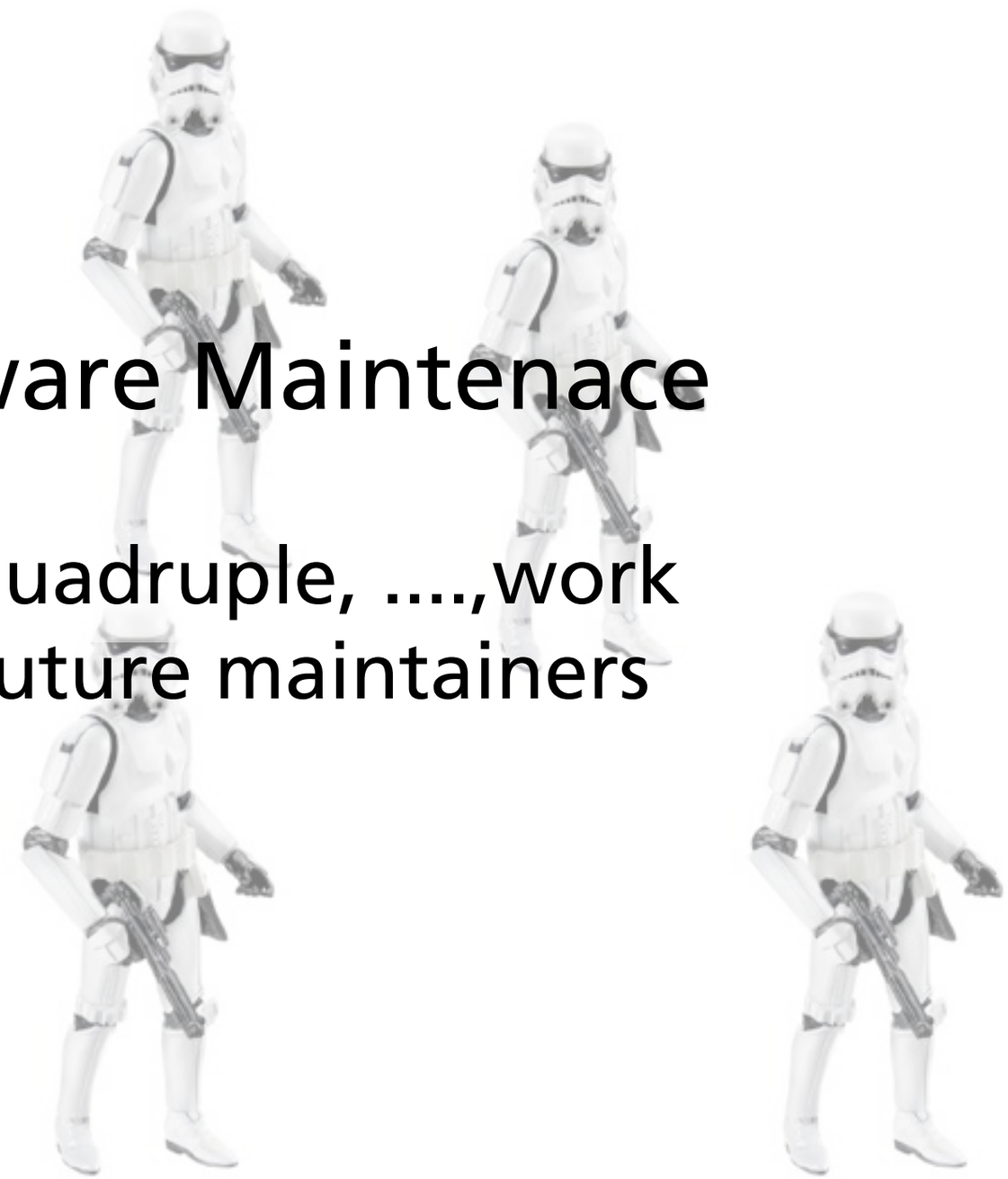
Time Pressure

Programming Language

Programming Libraries

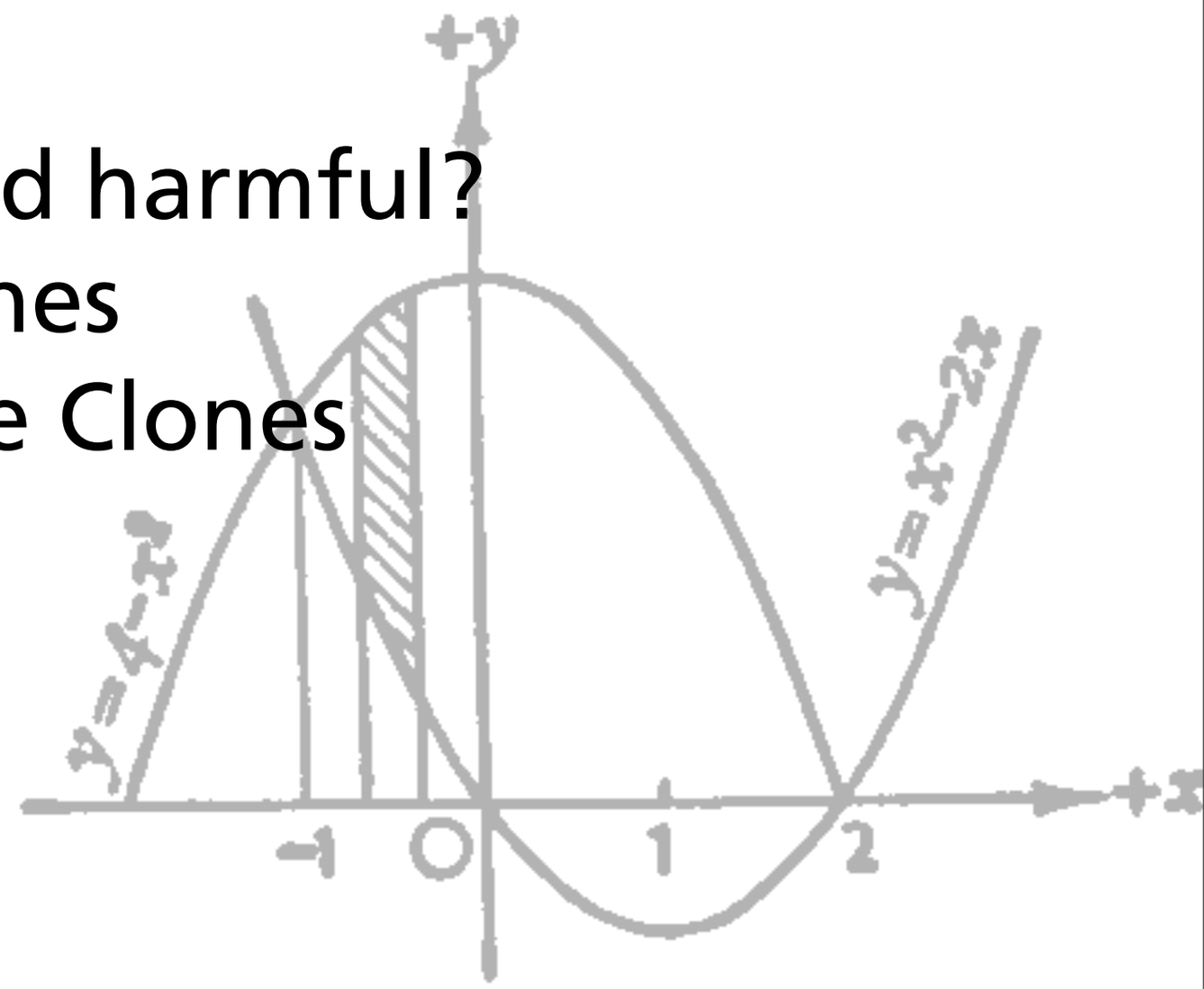
Bad Bad Clones

- General negative effect
  - Code Bloat
- Negatives effects on Software Maintenace
  - Copied Defect
  - Changes take double, triple, quadruple, ...,work
  - Add to the cognitive load of future maintainers
  - Dead Code
- Aesthetical Question



Hot Research Topic

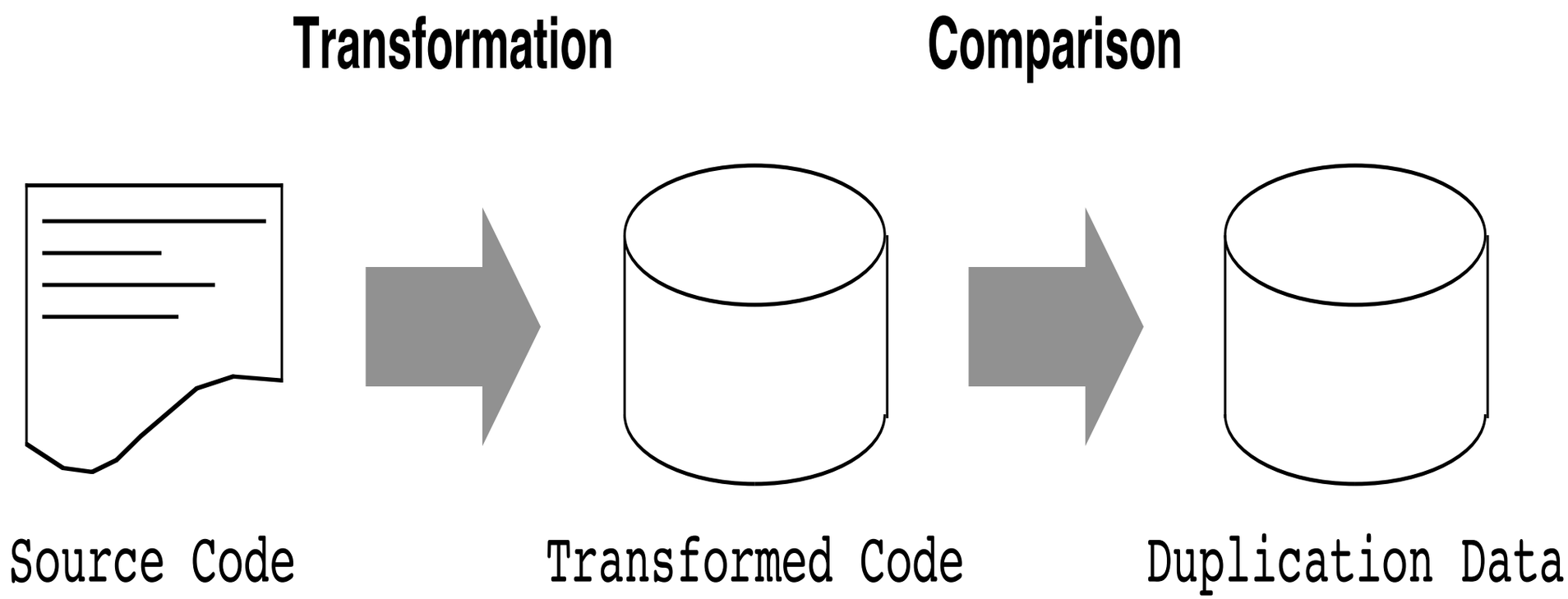
- Code Clone Detection
- Code Clones considered harmful?
- Evolution of Code Clones
- How to get rid of Code Clones
- Tools
- Visualization



Code Clone Detection

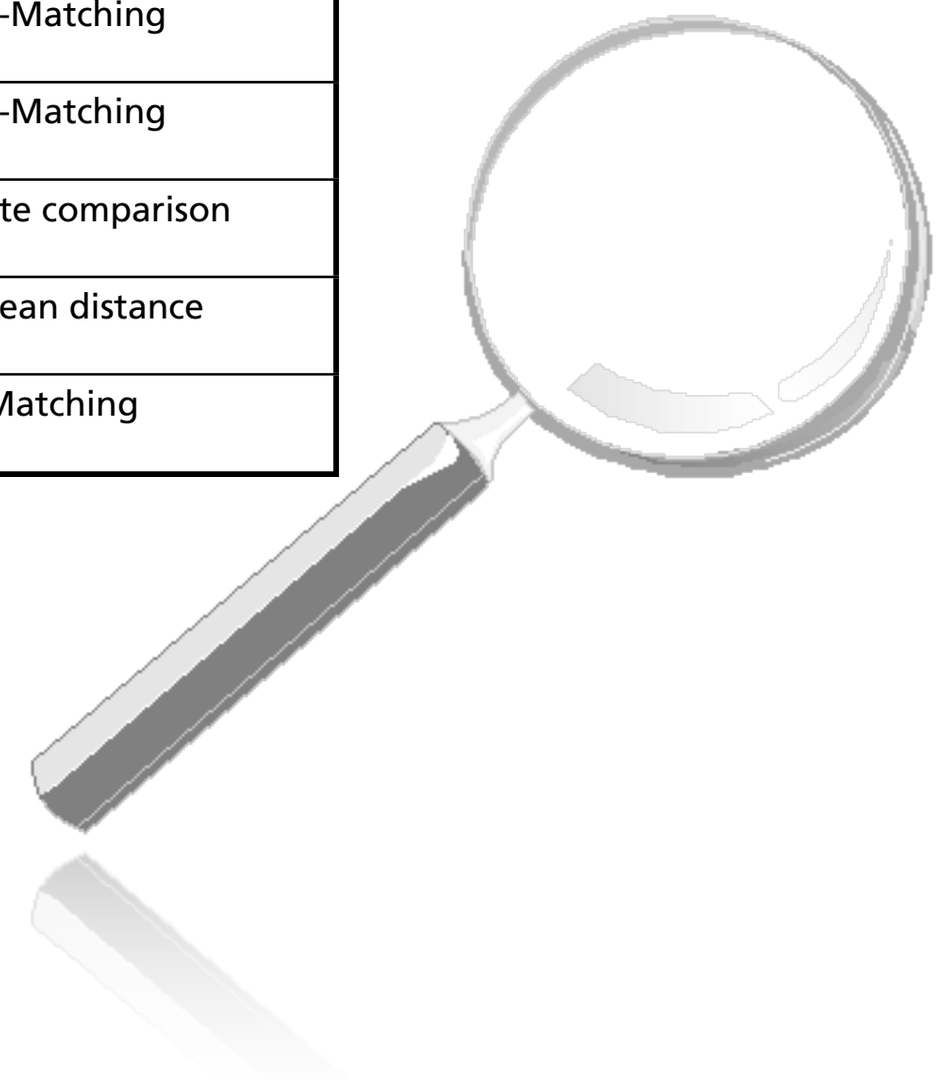
Nontrivial problem:

- No a priori knowledge about which code has been copied
- How to find all clone pairs among all possible pairs of segments?



Code Clone Detection

Author	*Level*	*Transformed Code*	*Comparison Technique*
[John94a]	Lexical	Substrings	String-Matching
[Duca99a]	Lexical	Normalized Strings	String-Matching
[Bake95a]	Syntactical	Parameterized Strings	String-Matching
[Mayr96a]	Syntactical	Metric Tuples	Discrete comparison
[Kont97a]	Syntactical	Metric Tuples	Euclidean distance
[Baxt98a]	Syntactical	AST	Tree-Matching

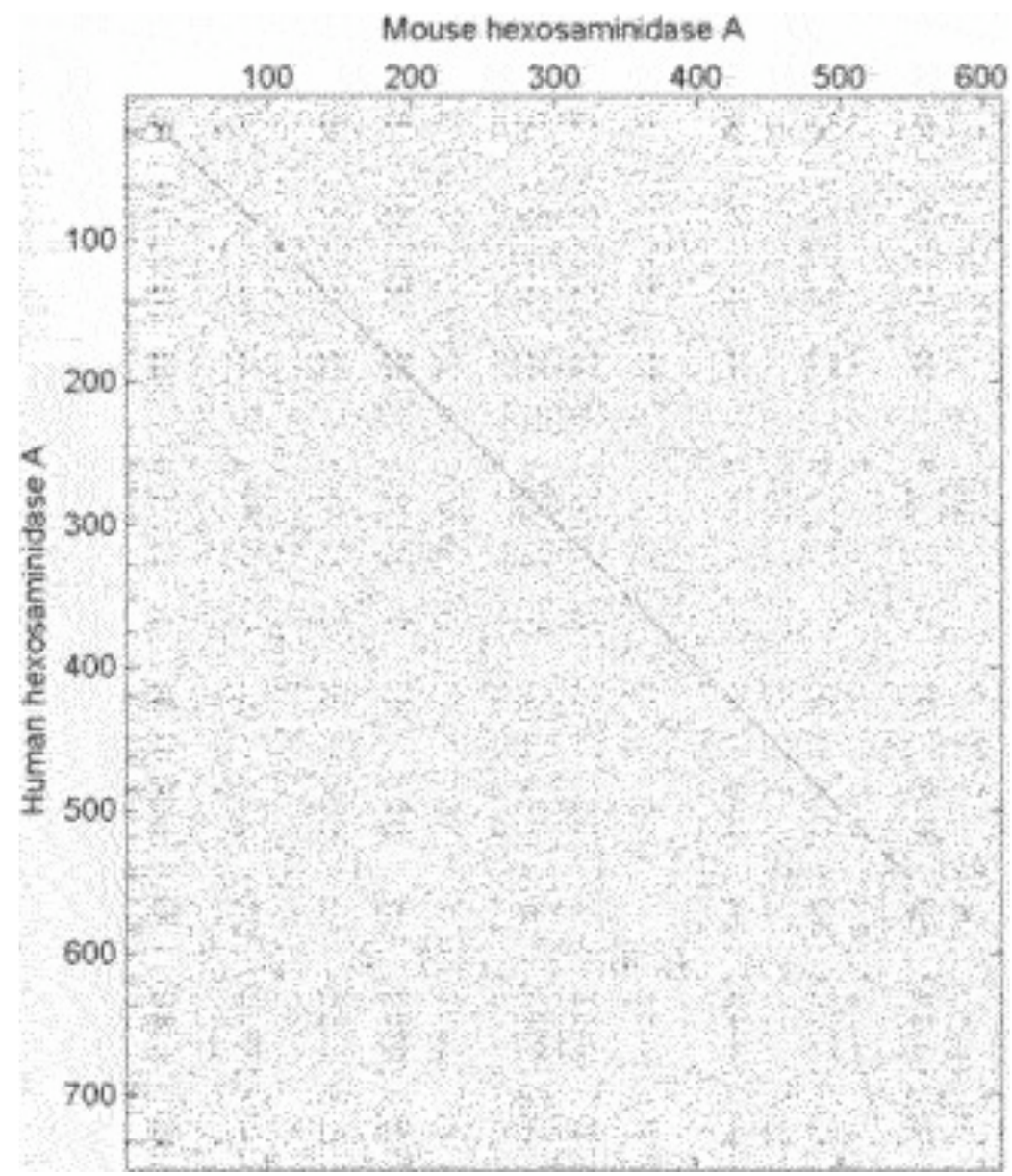


Advantages and Disadvatanges

- Performnace
- Scalability
- Configurability
- ...

Visualization

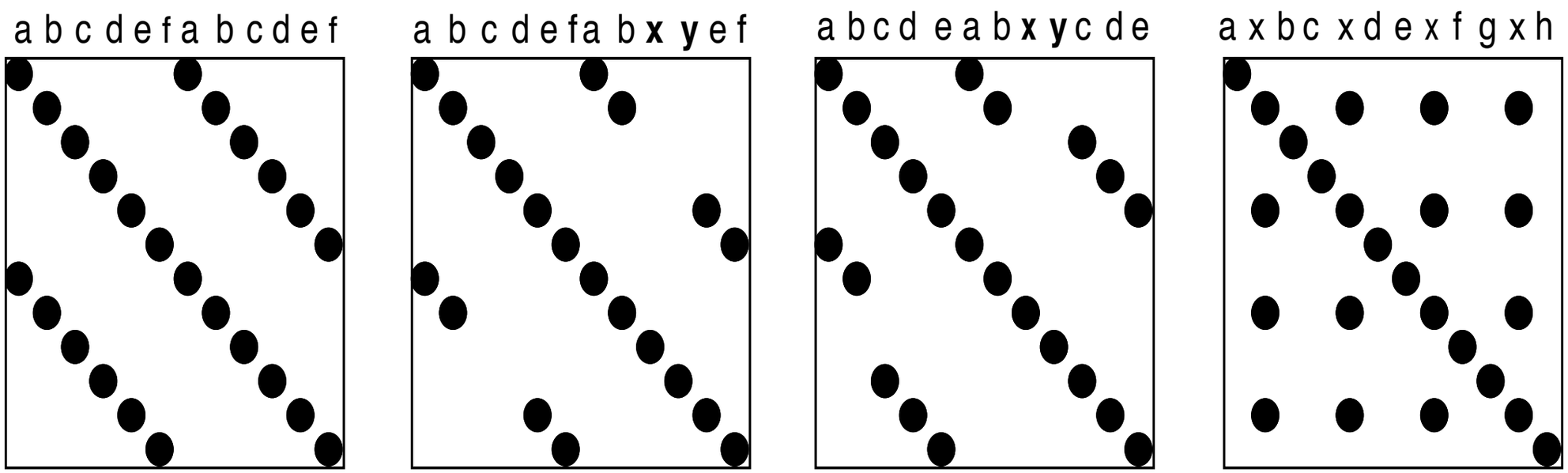
Dotplots - Technique from DNA Analysis



Visualization

Dotplot

- Code is put on vertical and horizontal axis
- A match between two elements is a dot in the matrix
- Easy visual identification of insertion, deletions, repeats, variations
- Gives a simple overall impression



Exact Copies

Copies with Variations

Inserts/Deletes

Repetitive Code Elements

CCFinderX