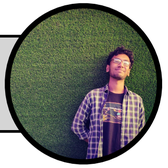
Elements of (M)ining (S)oftware (R)epositories

Dr. Antonio Mastropaolo

Instructor

Mr. Alvi Haque



Teaching Assistant

WILLIAM & MARY

CHARTERED 1693

Spring 2026

Mining Software Repositories

A **repository**, or repo, is a centralized digital storage that developers use to make and manage changes to an application's source code and more. A repo has features that allow developers to easily track code changes, simultaneously edit files, and efficiently collaborate on the same project from any location.

https://aws.amazon.com/what-is/repo/

Mining Software Repositories

REGIONS

A global community of developers

From maintainers to contributors and companies to nonprofits, people are using GitHub all over the world.

In fact, the only two places where we didn't see developer communities grow on GitHub in 2022 were Antarctica and Norfolk Island.

Everywhere else, we're seeing more developers building software on GitHub (and there are still almost 20 developers in Antarctica).

20.5M
New developers joined GitHub in 2022

India
Has the biggest developer population growth

85M+
New projects were started globally in 2022

Mining Software Repositories

100M Developers

400M Repositories

REGIONS

A global community of developers

From maintainers to contributors and companies to nonprofits, people are using GitHub all over the world.

In fact, the only two places where we didn't see developer communities grow on GitHub in 2022 were Antarctica and Norfolk Island.

Everywhere else, we're seeing more developers building software on GitHub (and there are still almost 20 developers in Antarctica).

20.5M
New developers joined GitHub in 2022

India
Has the biggest developer population growth

85M+
New projects were started globally in 2022

Mining Software Repositories

Allows the management of changes to artifacts (configuration items) for example, code

Allows to keep track of changes occurring to configuration items: What was changed, who did a change, when, and why

Allows the creation and management of branches

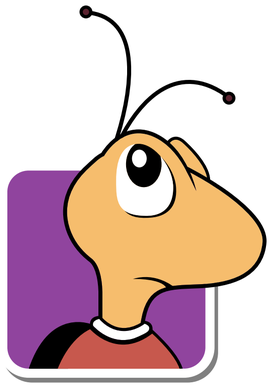
Possibility to retrieve a specific revision of a configuration item

Mining Software Repositories

Type of Repositories

Source Repositories

Bug Repositories



BugZilla

Communication Repositories

Mining Software Repositories

Type of Repositories

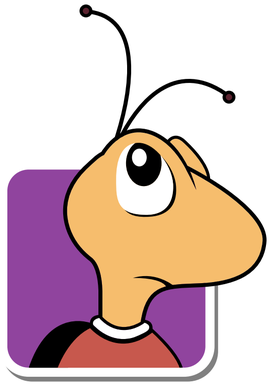
Source Repositories

git

SUBVERSION

CVS

Bug Repositories



BugZilla

Communication Repositories

Mining Software Repositories

Mining Software Repositories (MSR) leverages data available in repositories to aid development activities

Mining Software Repositories

Mining Software Repositories (MSR) leverages data available in repositories to aid development activities

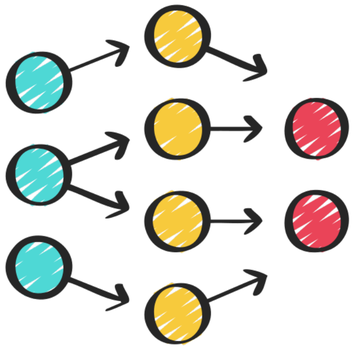
Mining Software Repositories

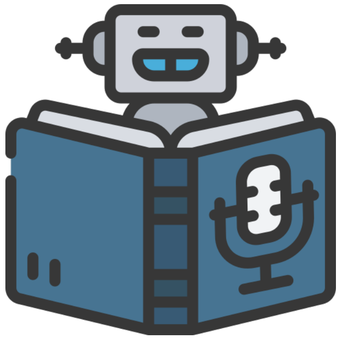
Mining Software Repositories (MSR) leverages data available in repositories to aid development activities

Our overarching goal is to build AI systems that can support developers in one or more SE-related tasks.

Mining Software Repositories

Data-driven Approaches





Rule-based Approaches

Mining Software Repositories

Data-driven Approaches

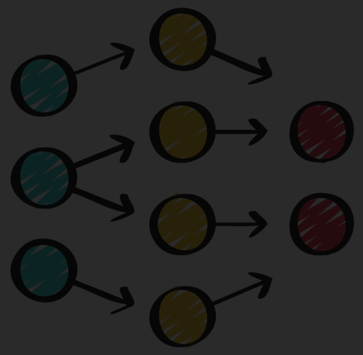
It's totally fine to have an AI-driven system that IS NOT grounded on the backbone of any data-driven methods. It's just a system that possesses some knowledge about facts, domains and methods to retrieve useful knowledge

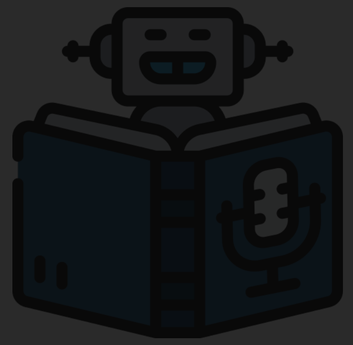
Rule-based Approaches

ing Software Repositories

EXPERT SYSTEM

Data-driven Approaches





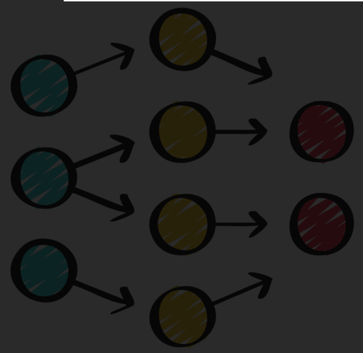
Rule-based Approaches

ng Software Repositories

EXPERT SYSTEM

Dat

IF: the screen is blue
AND: there is an error message with several alarming information THEN: it’s all good, it’s windows 😁



Rule-based Approaches

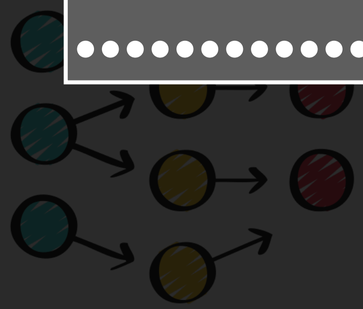
ng Software Repositories

EXPERT SYSTEM

Dat

IF: the screen is blue
AND: there is an error message with several alarming information THEN: it’s all good, it’s windows 😁

.....



Rule-based Approaches

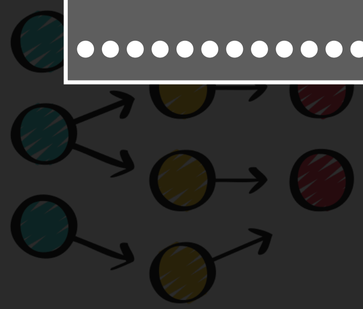
ng Software Repositories

EXPERT SYSTEM

Dat

IF: the screen is blue
AND: there is an error message with several alarming information THEN: it’s all good, it’s windows 😁

.....



Rule-based Approaches

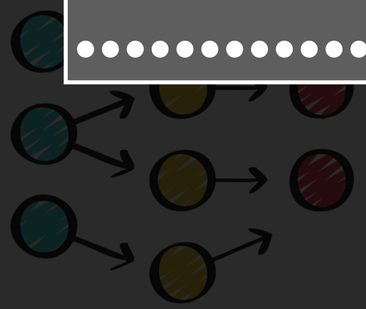
ng Software Repositories

EXPERT SYSTEM

Dat

IF: the screen is blue
AND: there is an error message with several alarming information THEN: it's all good, it's windows 😁

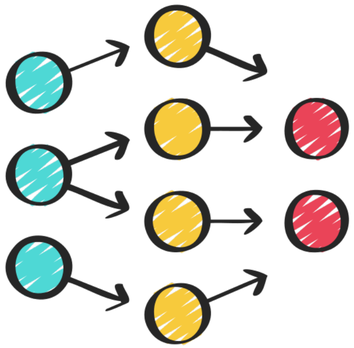
.....

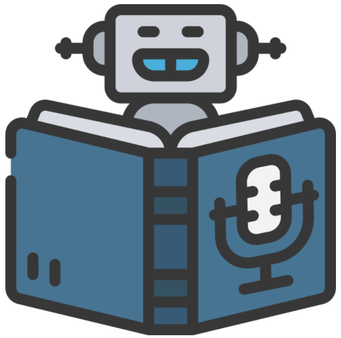


d Approaches

Mining Software Repositories

Data-driven Approaches

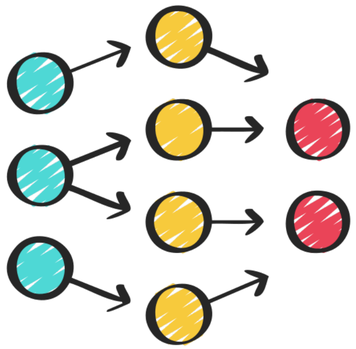


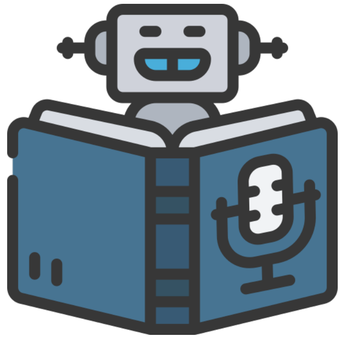


Rule-based Approaches

Mining Software Repositories

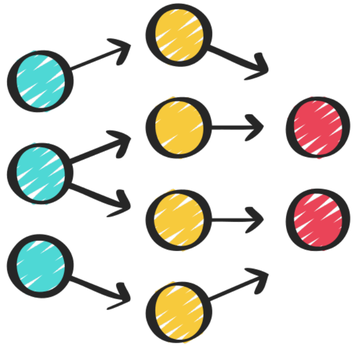
Data-driven Approaches

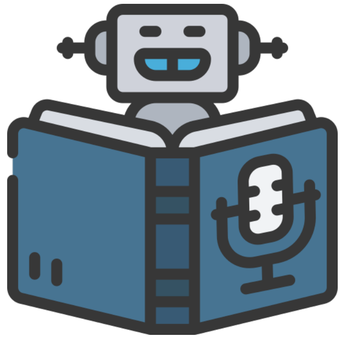




Mining Software Repositories

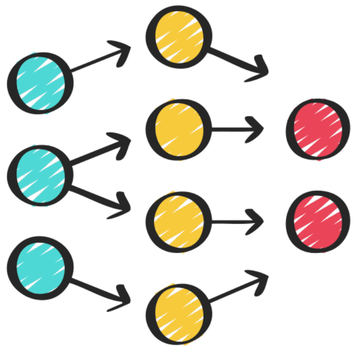
Data-driven Approaches

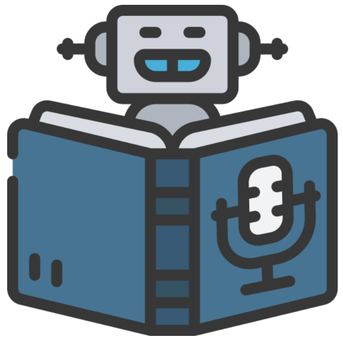




Mining Software Repositories

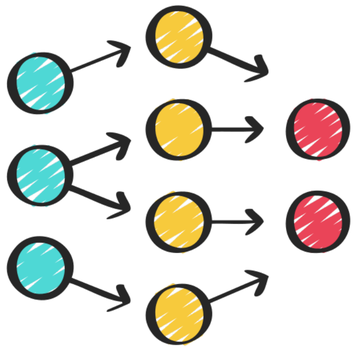
Data-driven Approaches

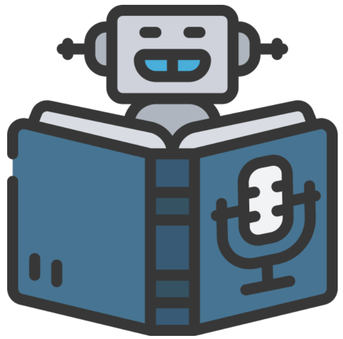




Mining Software Repositories

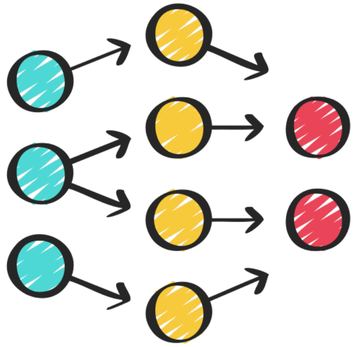
Data-driven Approaches

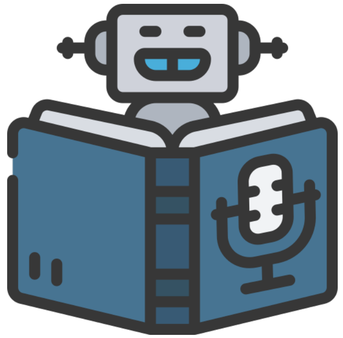




Mining Software Repositories

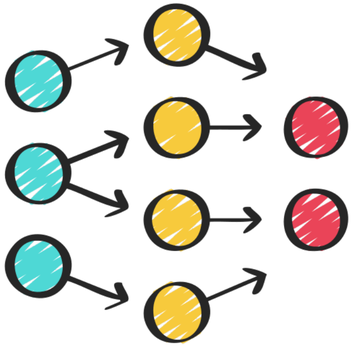
Data-driven Approaches

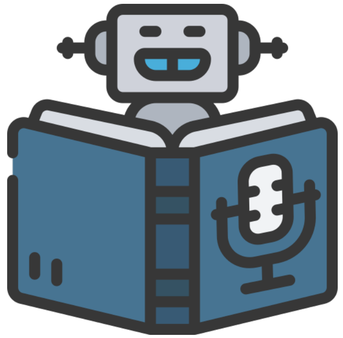




Mining Software Repositories

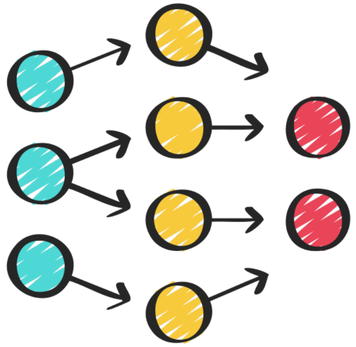
Data-driven Approaches

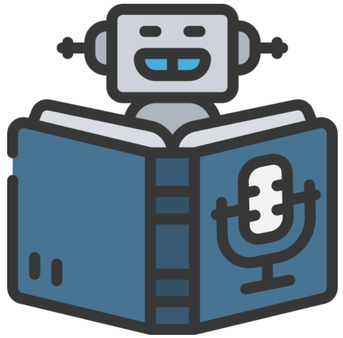


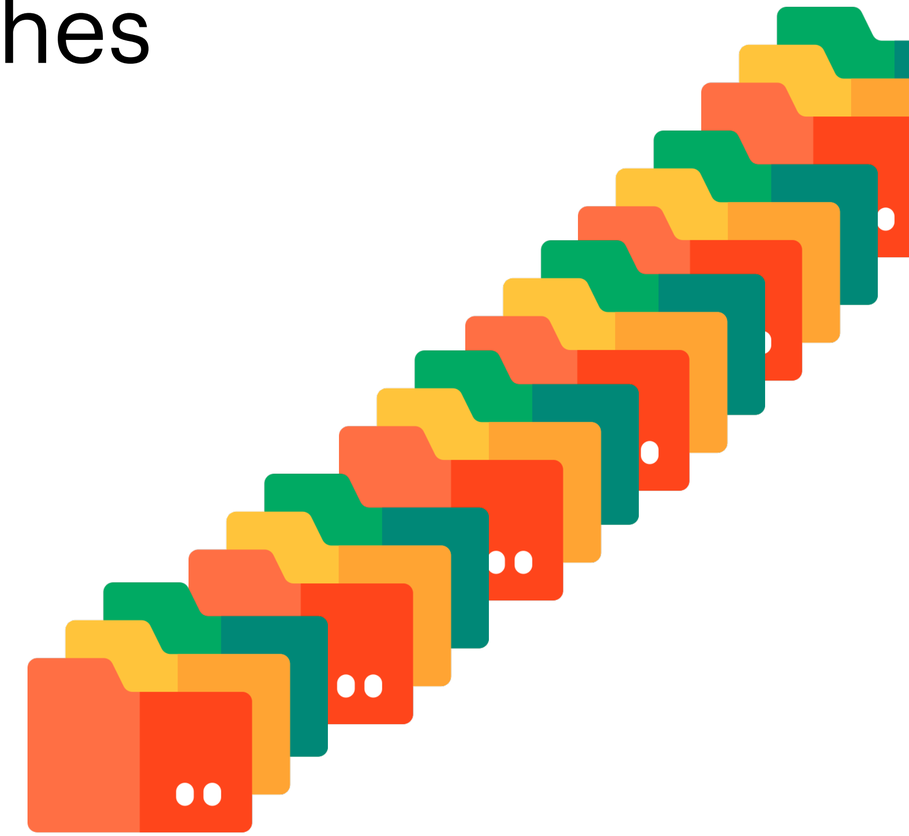


Mining Software Repositories

Data-driven Approaches







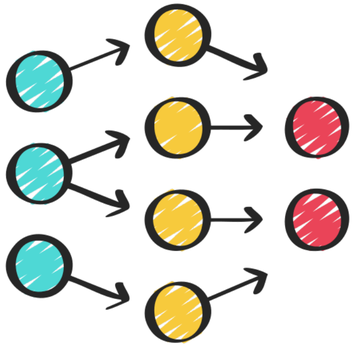
Mining Software Repositories

Why a lot and of what data Prof. are we talking about?

Mining Software Repositories

1. Why a lot of data?

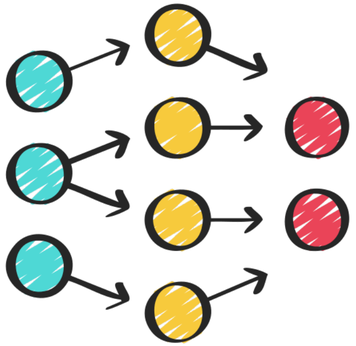
Deep Learning Models



Mining Software Repositories

1. Why a lot of data?

Deep Learning Models



Large Language Models (LLMs)

Mining Software Repositories

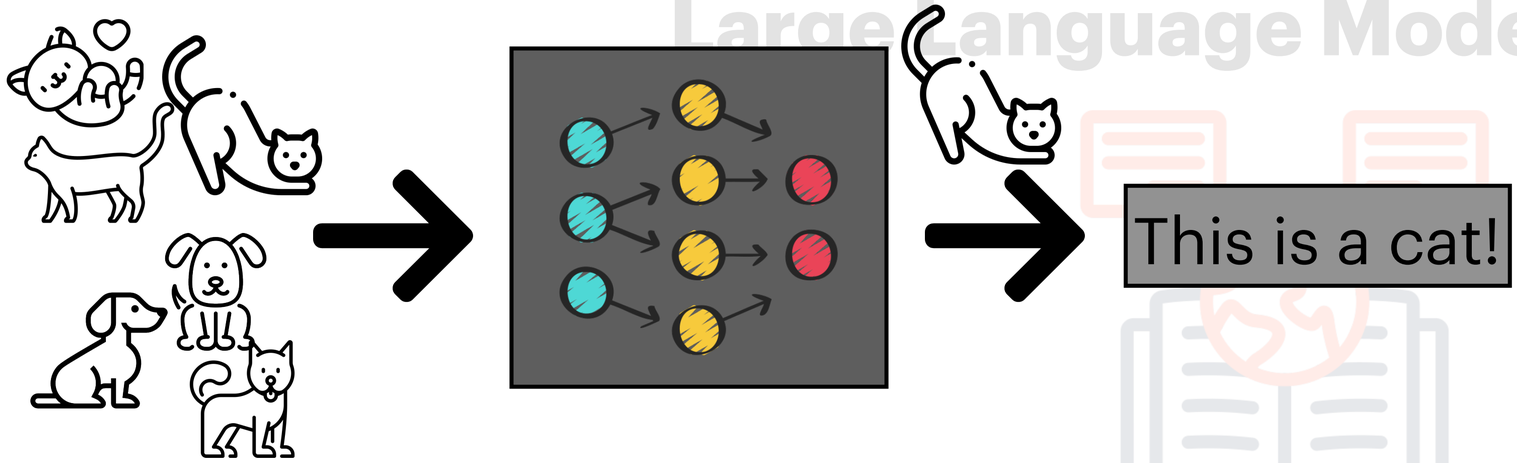
1. Why a lot of data?

Among all data-driven techniques, deep learning models, particularly large language models (LLMs), are highly dependent on data. In other words, they require vast amounts of data during training to effectively learn and generalize.

Mining Software Repositories

1. Why a lot of data?

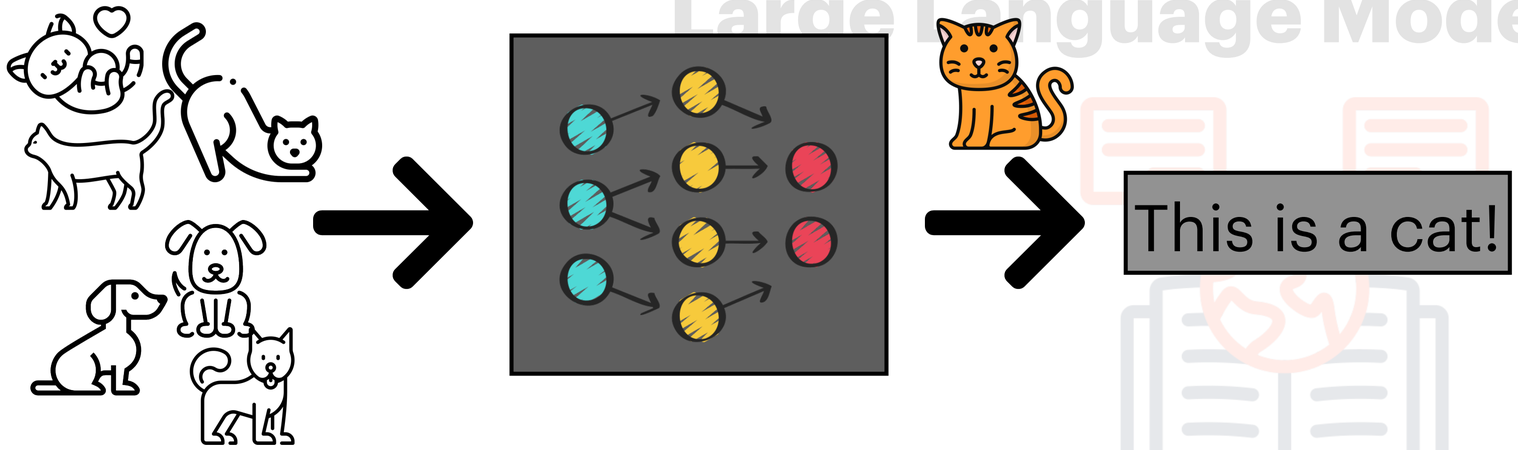
Among all data-driven techniques, deep learning models, particularly large language models (LLMs), are highly dependent on data. In other words, they require vast amounts of data during training to effectively learn and generalize.



Mining Software Repositories

1. Why a lot of data?

Among all data-driven techniques, deep learning models, particularly large language models (LLMs), are highly dependent on data. In other words, they require vast amounts of data during training to effectively learn and generalize.



Mining Software Repositories

2. What type of data?

Large Language Models (LLMs)

Mining Software Repositories

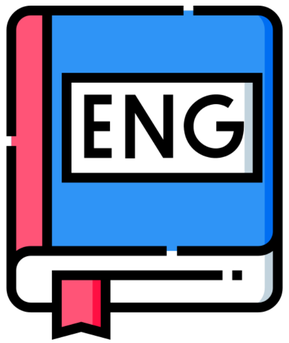
2. What type of data?

GitHub Copilot

GitHub Copilot can generate code and natural language

Mining Software Repositories

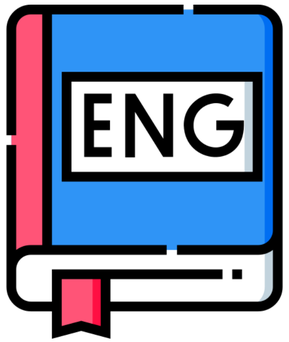
2. What type of data?



We gather the building blocks for our future approaches, specifically data from publicly available GitHub repositories, due to the vast and impressive volume of code and natural language they contain.

Mining Software Repositories

2. What type of data?



We ~~gather~~ mine the building blocks for our future approaches, specifically data from publicly available GitHub repositories, due to the vast and impressive volume of code and natural language they contain.

Mining Software Repositories

DEMO TIME

Mining Software Repositories

We need to ensure that the data we collected are of high-quality.
What does it mean?

Mining Software Repositories

We need to ensure that the data we collected are of high-quality. What does it mean?

1. Select good repositories as proxy for the quality

Mining Software Repositories

We need to ensure that the data we collected are of high-quality. What does it mean?

1. Select good repositories as proxy for the quality

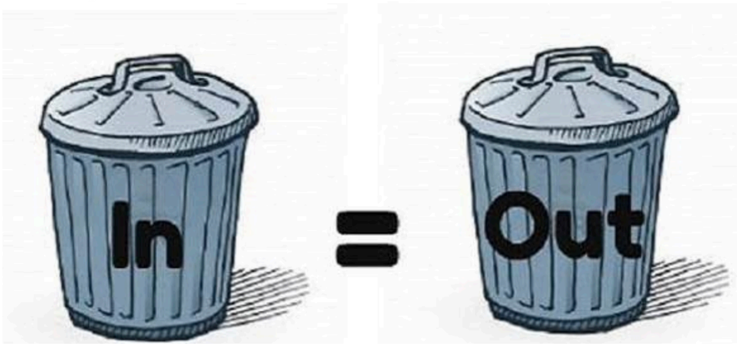
2. Once the data has been collected, we must enforce quality and sanity check

Mining Software Repositories

We need to ensure that the data we collected are of high-quality. What does it mean?

1. Select good repositories as proxy for the quality

2. Once the data has been collected, we must enforce quality and sanity check



Preprocessing Source Code:

0. No duplicates – Their presence can hinder the learning ability of the model

Preprocessing Source Code:

0. No duplicates – Their presence can hinder the learning ability of the model
1. Code that contains only ASCII characters

Preprocessing Source Code:

0. No duplicates – Their presence can hinder the learning ability of the model
1. Code that contains only ASCII characters
2. Remove outliers from the the dataset.
   We can define an outlier as a method incredibly **long** or **short**

Preprocessing Source Code:

0. No duplicates – Their presence can hinder the learning ability of the model
1. Code that contains only ASCII characters
2. Remove outliers from the the dataset.
   We can define an outlier as a method incredibly **long** or **short**
3. Remove boilerplate code (e.g., setter, getter)

Preprocessing Source Code:

0. No duplicates – Their presence can hinder the learning ability of the model
1. Code that contains only ASCII characters
2. Remove outliers from the the dataset.
   We can define an outlier as a method incredibly **long** or **short**
3. Remove boilerplate code (e.g., setter, getter)
4. Clean the method by removing all the comments within

Preprocessing Source Code:

0. No duplicates – Their presence can hinder the learning ability of the model
1. Code that contains only ASCII characters
2. Remove outliers from the the dataset.
   We can define an outlier as a method incredibly **long** or **short**
3. Remove boilerplate code (e.g., setter, getter)
4. Clean the method by removing all the comments within

N. Remove the code that does not fit the criteria we defined or someone else did for us. (e.g., I want to remove code having a **Cyclomatic Complexity** < 5, because I want to train a model on something difficult

Mining Software Repositories

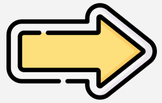
LET’S SEE SOME CODE

Mining Software Repositories

Tokenization

Tokenization is the process that breaks down text into smaller units that can be analyzed separately.

```
public int addNumbers(int a, int b){
    int sum=a+b;
    return sum;
}
```



```
public int addNumbers ( int a , int b ) {
    int sum = a + b ;
    return sum;
}
```

Mining Software Repositories

Tokenization

Lexer: is a piece of software that converts the text provided as input into a sequence of tokens. Tokens, that constitute the basic building blocks a language. For programming languages such as Java, those building blocks represent element like keywords, operators, literals...

Parser: is a piece of software that that takes the sequence of tokens generated by the lexer, and analyzes them to understand the structure and syntax of the code.

Mining Software Repositories

Tokenization

Code-tokenize *(Python) works for several programming languages*

Javalang *(Python) works for Java only*

JavaParser *(Java) works for Java only*

Pygments *(Python) works for several programming languages. This tool is technically, a Generic Syntax Highlighter, which integrates a lexer and a parser to get the job done*

Why tokenization is needed?

Reducing Complexity: Tokenization divides text into smaller units, making it easier for the model to identify patterns and relationships.

Why tokenization is needed?

Reducing Complexity: Tokenization divides text into smaller units, making it easier for the model to identify patterns and relationships.

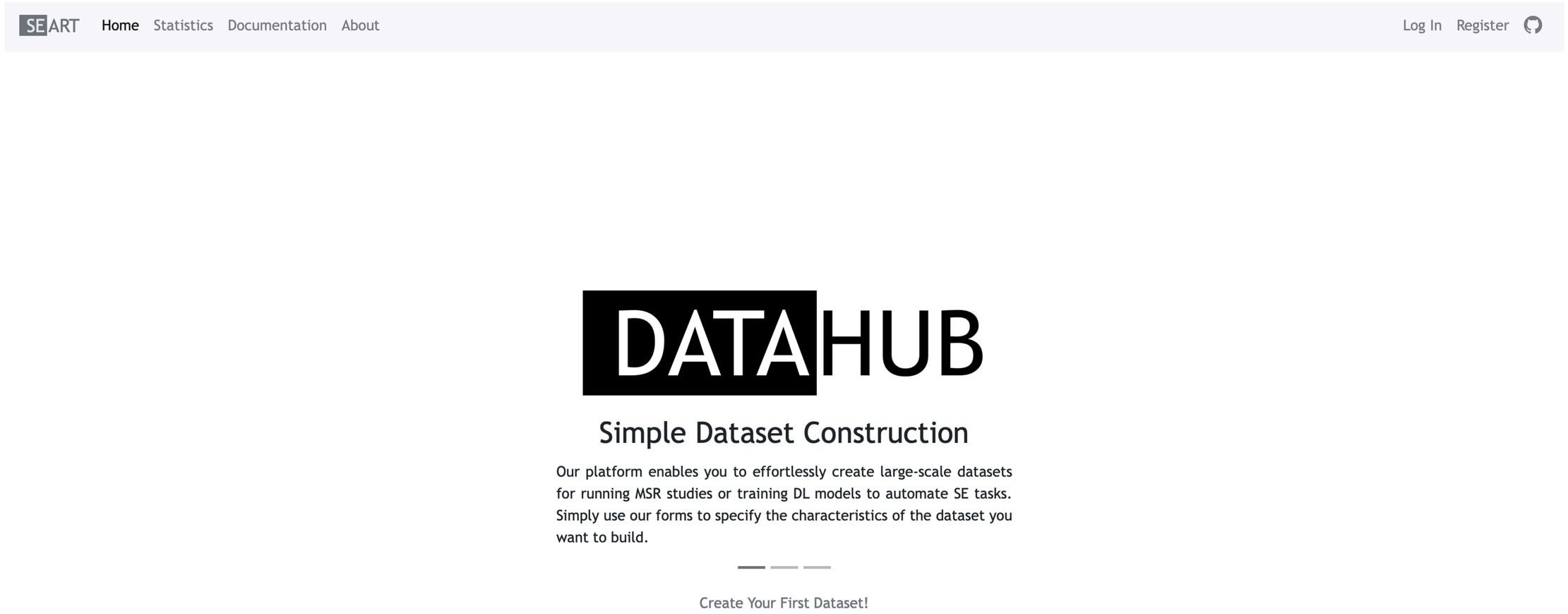
Handling the Vocabulary:

Why tokenization is needed?

Reducing Complexity: Tokenization divides text into smaller units, making it easier for the model to identify patterns and relationships.

Handling the Vocabulary: By dividing text into tokens, the model can create a numerical representation of the vocabulary, making it easier to process and understand.

Mining Software Repositories



The State-of-the-art