Java IOs

RES, Lecture 1

Olivier Liechti

heig-vd

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD

www.heig-vd.ch

Agenda

heig-vd
HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

- **Week 1**
  - Universal API
  - Sources, Sinks and Streams
  - Performance and Buffering
- **Week 2**
  - The Decorator Pattern and The Mighty Filter Classes
  - Binary vs. Character-Oriented IOs
  - Shit Happens... Dealing with IO Exceptions

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

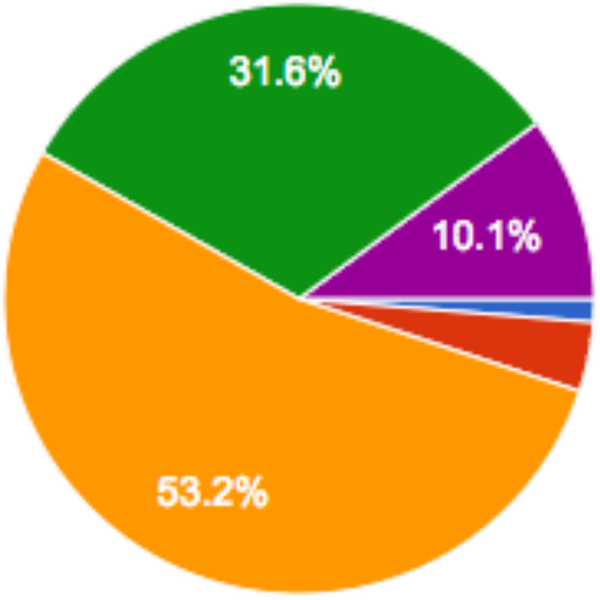
Survey

“Ca à l'air chaud quand même”
- anonymous student

“apprendre des trucs cools”
- anonymous student

Est-ce que vous avez déjà utilisé les classes d'entrées/sorties en Java (InputStream, Reader, OutputStream, Writer, etc.)?

79 responses

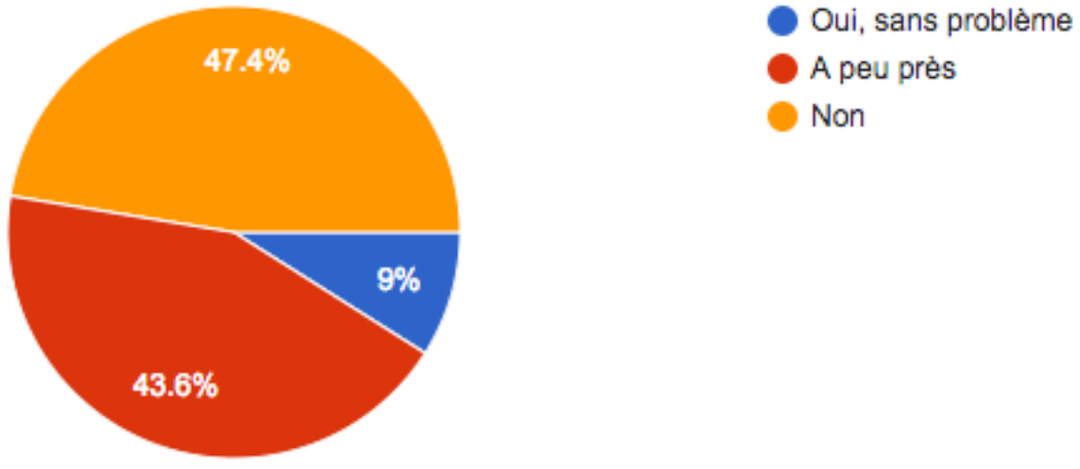


Seriez-vous capable d'expliquer la notion de tampon (buffer) dans le cadre des entrées-sorties?

79 responses

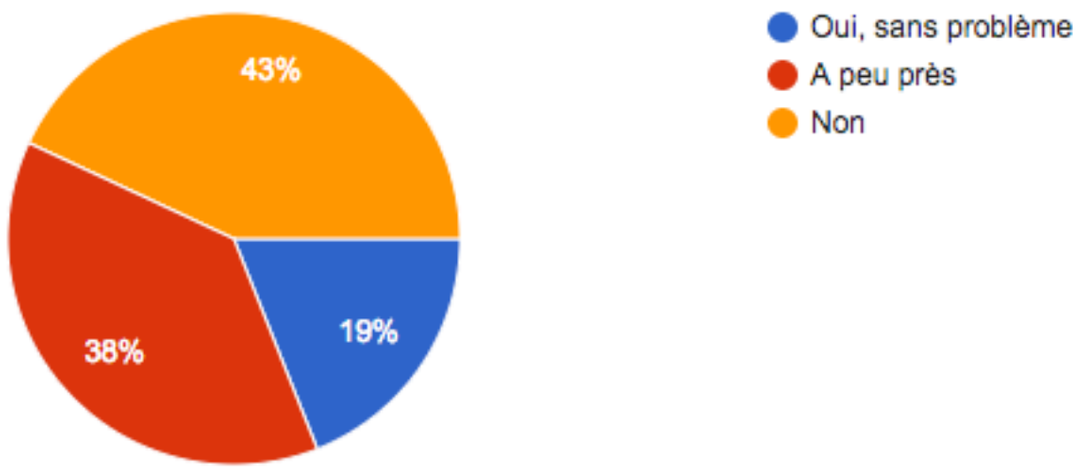
Seriez-vous capable d'expliquer la différence entre des entrées-sorties blocantes et des entrées-sorties non blocantes?

78 responses



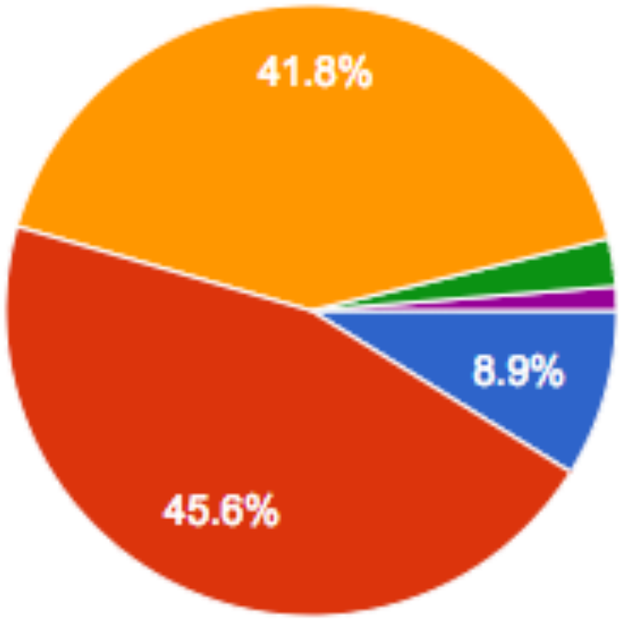
Seriez-vous capable d'expliquer ce qu'est une fonction de callback?

79 responses



Si on vous demandait d'écrire un programme Java qui compte le nombre de voyelles dans un fichier texte...

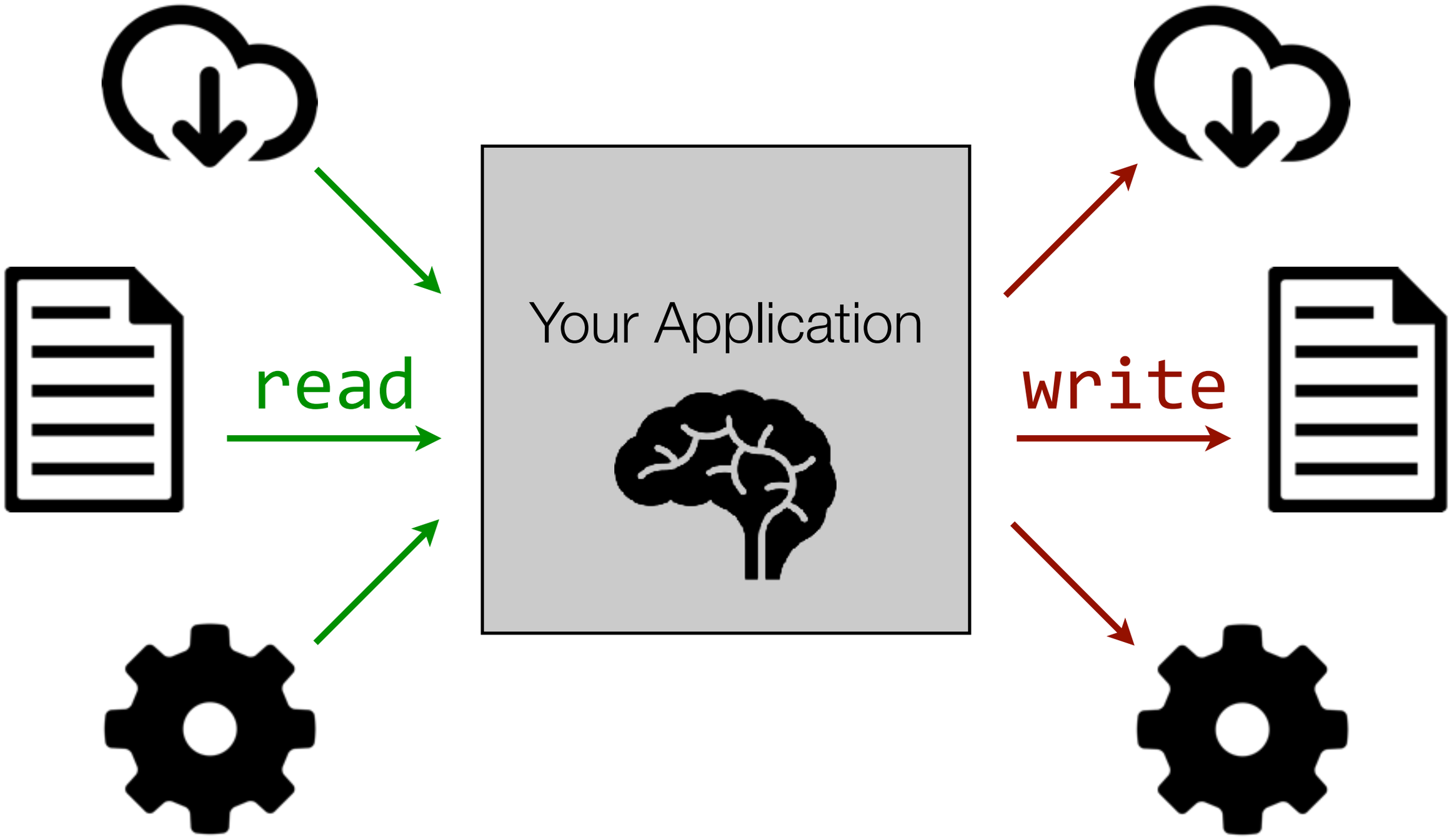
79 responses



- C'est trivial, je sais exactement comment faire ça.
- C'est simple, en regardant la doc du package java.io, je m'en sortirais rapidement.
- Je ne sais pas trop comment faire, mais ça devrait aller.
- Je ne saurais absolument pas commencer.
- Option 5

Universal API

What do we mean by IO?



A Universal API

*At the end of the day, whether you are "talking" to a **file**, to a **network endpoint** or to a **process** does not matter.*

*You are always doing the same thing: **reading** and/or **writing** bytes or characters.*

*The Java IO API is the **toolbox** that you need for that purpose.*

Sources, Sinks & Streams

System.out.println("I like IO");

out

```
public static final PrintStream out
```

The "standard" output stream. This stream is already open and ready to accept output data. Typically this stream corresponds to display output or another output destination specified by the host environment or user.

For simple stand-alone Java applications, a typical way to write a line of output data is:

```
System.out.println(data)
```

See the `println` methods in class `PrintStream`.

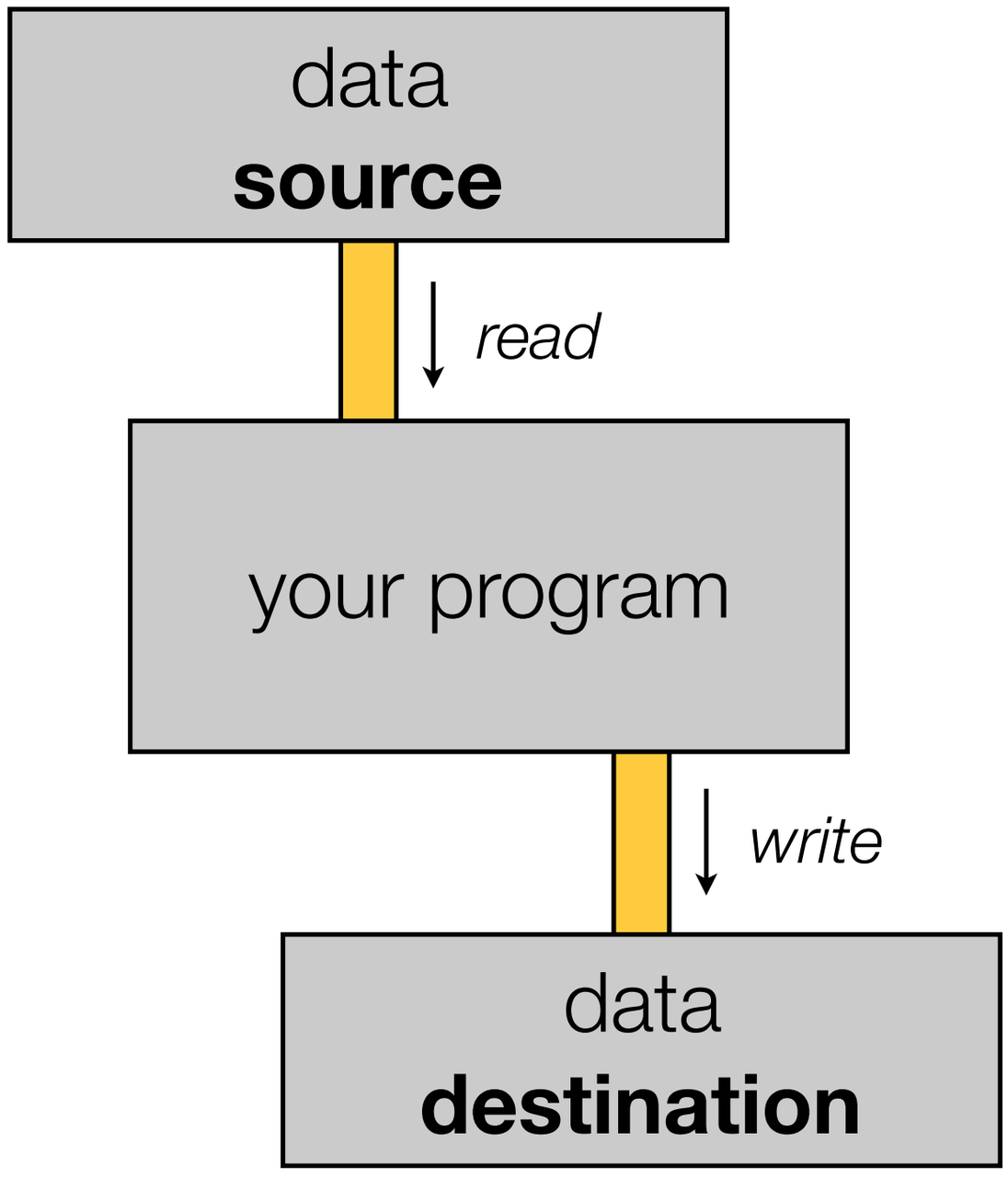
See Also:

`PrintStream.println()`, `PrintStream.println(boolean)`, `PrintStream.println(char)`, `PrintStream.println(char[])`, `PrintStream.println(double)`, `PrintStream.println(float)`, `PrintStream.println(int)`, `PrintStream.println(long)`, `PrintStream.println(java.lang.Object)`, `PrintStream.println(java.lang.String)`

Streams

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

- When we talk about "input/output", or IOs, we think about **producing** and **consuming streams of data**.
- There are different **sources** that contain or produce data.
- There are also different **destinations** that receive or consume data.
- Think about **files**, **network endpoints**, **memory**, **processes**, etc.
- Your **program** can **read data** from a stream. Your program can **write data** to a stream.



What can you do with IO streams?

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

- **Read data from a file**. Think about a program that reads an XML configuration file at start-up. Think about your favorite text editor.
- **Write data to a file**. Think about a java compiler that produces .class files.
- **Receive data over the network**. Think about a web server receiving requests from clients.
- **Send data over the network**. Think about the same server sending responses to the clients.
- **Exchange data with other programs** (processes) running on the same machine.
- **Etc.**

In the end, it's **always the same thing**... it's about **reading** and **writing** data!

If it's the same thing, then we want to have a **single API** for dealing with all sorts of IOs!

Classes in the `java.io` package (1)

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

InputStream

OutputStream

FileInputStream

FileOutputStream

ByteArrayInputStream

ByteArrayOutputStream

PipedInputStream

PipedOutputStream

FilterInputStream

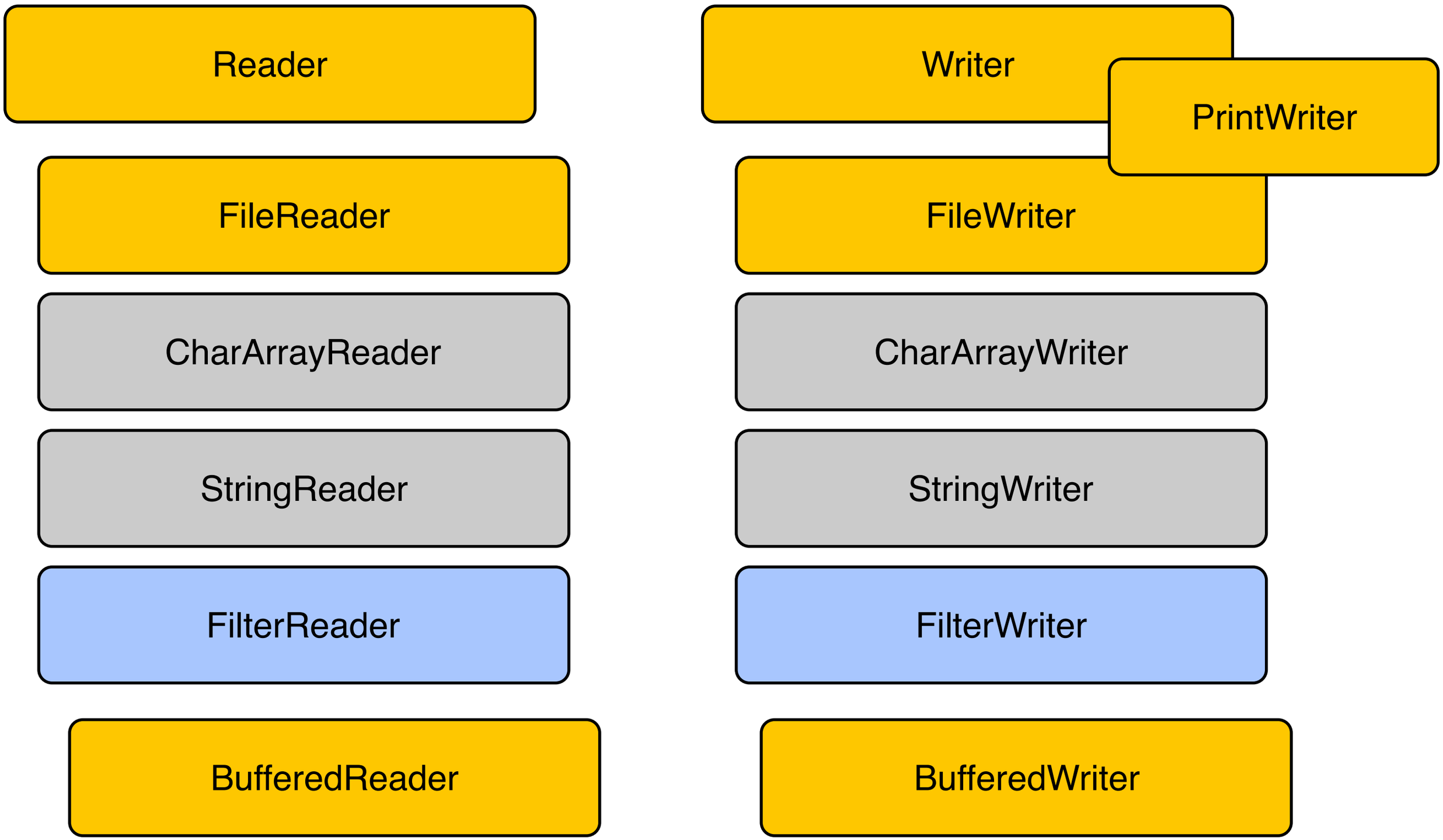
FilterOutputStream

BufferedInputStream

BufferedOutputStream

Classes in the `java.io` package (2)

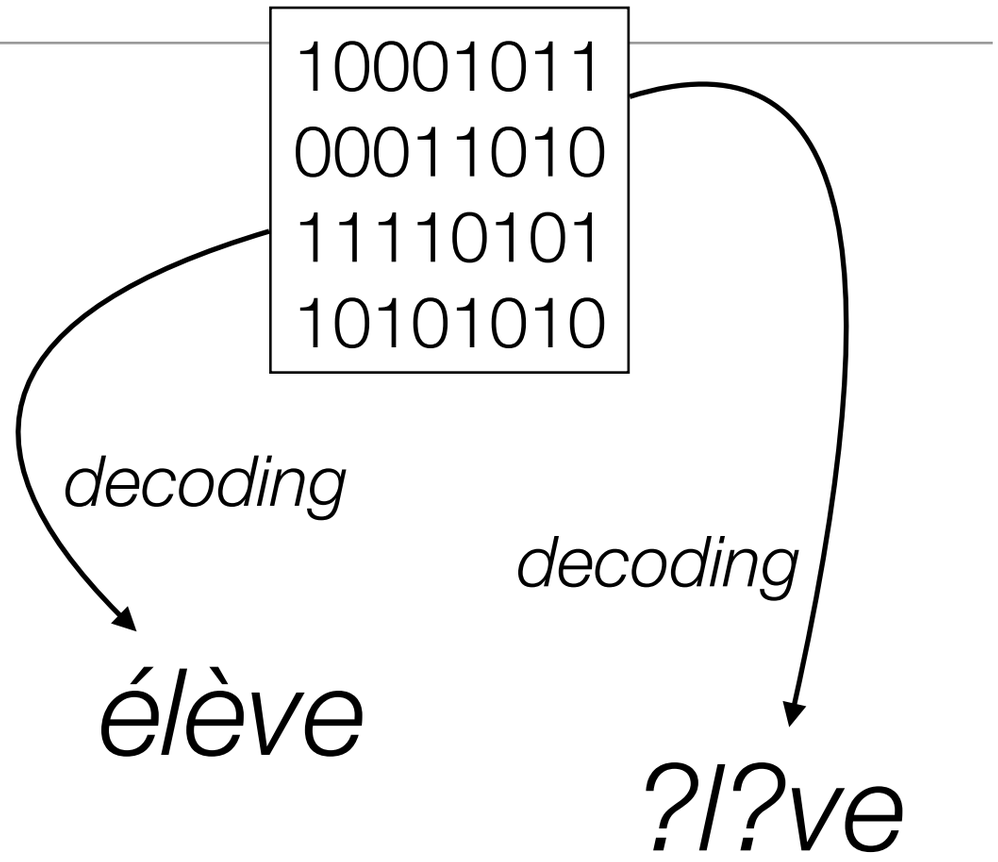
HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch



Bytes vs. Characters

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

- There are different types of data sources. Some sources produce **binary data**. Other sources produce **textual data**.
- It's always a **series of 0's and 1's**. The real question is : "how do you interpret these bits"?
- When you deal with **textual data**, the interpretation is not always the same. It depends on the "**character encoding system**"?
- When you deal with IOs, you have to **use different classes** for processing binary, respectively textual data. **Otherwise, you will corrupt data**!



For binary data, use **inputStreams** and **outputStreams**.
For text data, use **readers** and **writers**.

Reading Bytes, One at The Time

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

```
int b = fis.read();
while ( b != -1 ) {

  // b has an int value between -1 and 255
  // -1 indicates that we are at the end of the stream
  // b can be casted to a byte, remember that in Java,
  // bytes are signed and have values in the range [-128..127].

  b = fis.read();
}
```

This is a blocking call

Or if you absolutely want to save 1 line...

```
int b;
while ( (b = fis.read()) != -1 ) {

}
```

Reading Bytes, in Blocks

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

```
final int BUFFERSIZE = 255;
byte[] buffer = new byte[BUFFERSIZE];

int numberOfNewBytes = fis.read(buffer);

  // we know that numberOfNewBytes bytes have been read
  // we know that these bytes are available in the buffer
  // WARNING: there might be left-over junk after these bytes!

while ( numberOfNewBytes != -1 ) {

  // do something with the bytes in buffer[0..numberOfNewBytes-1]
  // ignore what is left in buffer[numberOfNewBytes.. BUFFERSIZE]

  // read the next chunk of bytes
  numberOfNewBytes = fis.read(buffer);

}
```

Writing Bytes

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

Write 1 byte

```
OutputStream os = ...;
int b;
// The byte to be written is the eight low-order b.
// The 24 high-order bits of b are ignored. So, b can have an
// int value in the range [0..255]
os.write(b); ←——— This is a blocking call
```

Write a block of bytes

```
OutputStream os = ...;
byte[] data = new byte[BLOCK_SIZE];

data[3] = 22; data[4] = 5; data[5] = 9; data[6] = 7;
// data[0..2] contains junk, data[7..BLOCK_SIZE-1] too

int offset = 3; // because we have started to fill at slot 3
int length = 4; // because we have filled 4 slots
os.write(data, offset, length);
```

Design **Your Code** to Be Universal

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

```
/**
 * This interface will work only for data sources on the file
 * system. In the method implementation, I would need to create
 * a FileInputStream from f and read bytes from it.
 */
 public interface IPoorlyDesignedService {
  public void readAndProcessBinaryDataFromFile(File f);
}
```

data source

VS

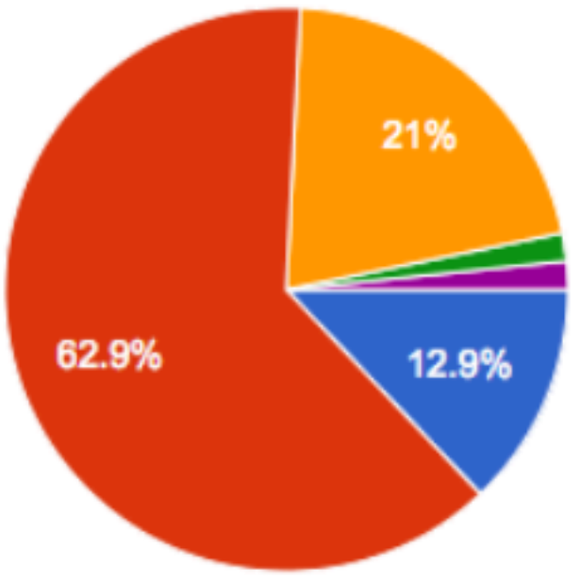
```
/**
 * This interface is much better. The client using the service
 * has a bit more responsibility (and work). It is up to the
 * client to select a data source (which can still be a file,
 * but can be something else). The method implementation
 * will ignore where it is reading bytes from. Nice for reuse,
 * nice for testing.
 */
public interface INicelyDesignedService {
  public void readAndProcessBinaryData(InputStream is);
}
```

stream

Back to Lab-00...

Connaissez-vous git?

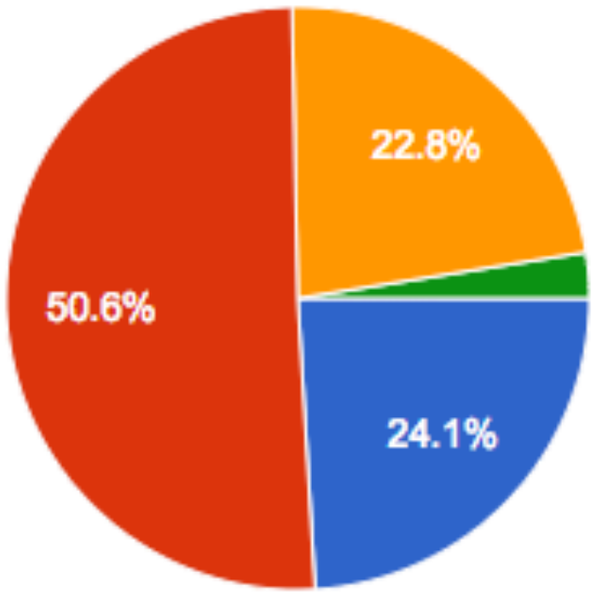
62 responses



- Oui, je connais très bien les commandes et je suis à l'aise avec les branches
- Oui, je l'utilise assez régulièrement mais je ne suis pas 100% à l'aise avec les commandes
- Oui, mais je ne l'ai pas encore utilisé
- Non, je sais vaguement ce que c'est
- Non, aucune idée

Connaissez-vous GitHub?

79 responses



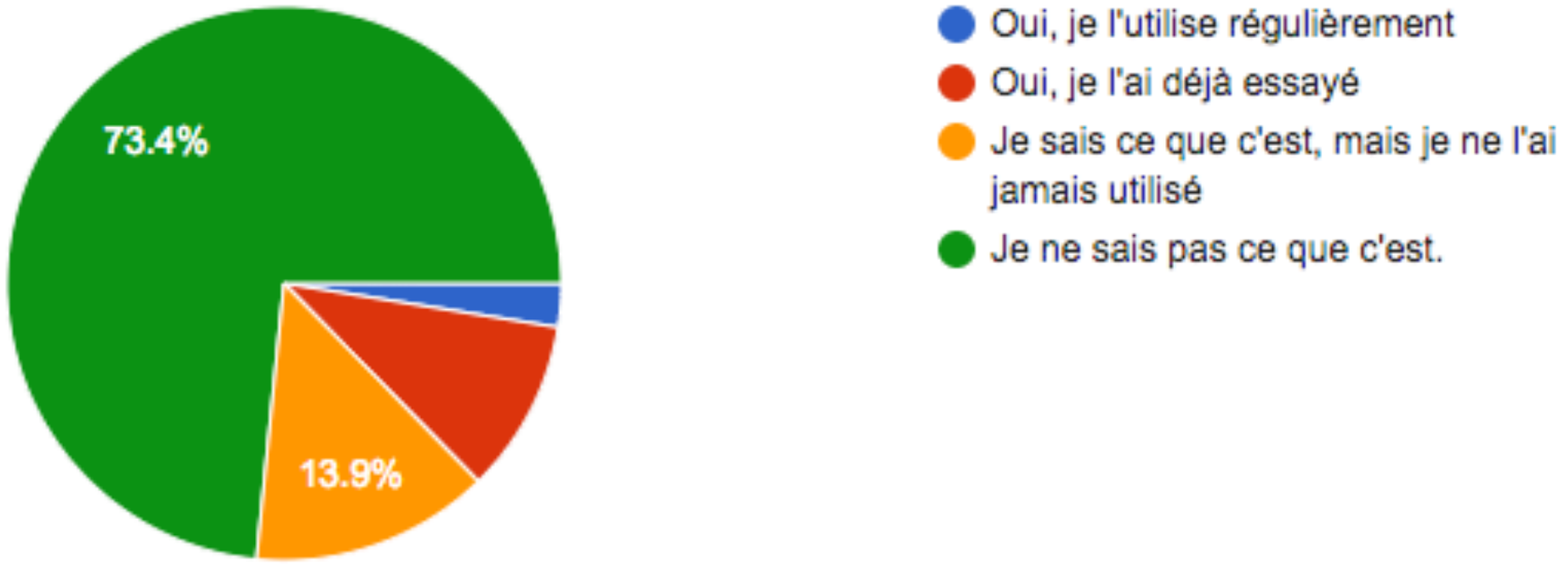
- Oui, je l'utilise régulièrement et je connais la différence entre un fork et un clone.
- Oui, je l'utilise mais j'en fait une utilisation de base.
- Oui, mais je ne l'ai jamais utilisé.
- Non.

Seriez-vous capable d'expliquer ce qu'est une "pull request"?

79 responses

Est-ce que vous connaissez l'outil apache maven?

79 responses



Connaissez vous JUnit et les tests unitaires?

79 responses

Using Git locally

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

```
$ mkdir my-project
$ cd my-project
$ git init
$ ls -al
```

- **You do not *have to* use a server**: Git is already useful to manage versions of your files on your local machine.
- The **`git init` command** creates a **local repository**. If you look carefully, you will see **a hidden `.git` directory**, where Git keeps all of his data.
- **Important**: your **`my-project`** directory is your **working directory**. If you simply create files in it, they will not immediately be part of your repository!

Using Git locally

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

```
$ echo "text a" > a.txt
$ git status
$ git add a.txt
$ git commit -m “First version of a.txt”
$ echo “my mod on text a” > a.txt
$ git status
```

- A **commit** is a **snapshot** of your repository. Git maintains a **graph of commits** and you can always **recover the state** of a particular commit.
- When **you have modified files in your working directory**, you need to specify which ones should be **part of the next commit**.
- You use the **`git add`** command to add a file to the so-called **staging area**. It will be part of the next commit.
- You use the **`git status`** command to **check the content** of your working directory and of your staging area.

More info: http://git-scm.com/book/en/Git-Basics-Recording-Changes-to-the-Repository

The Git Object Model

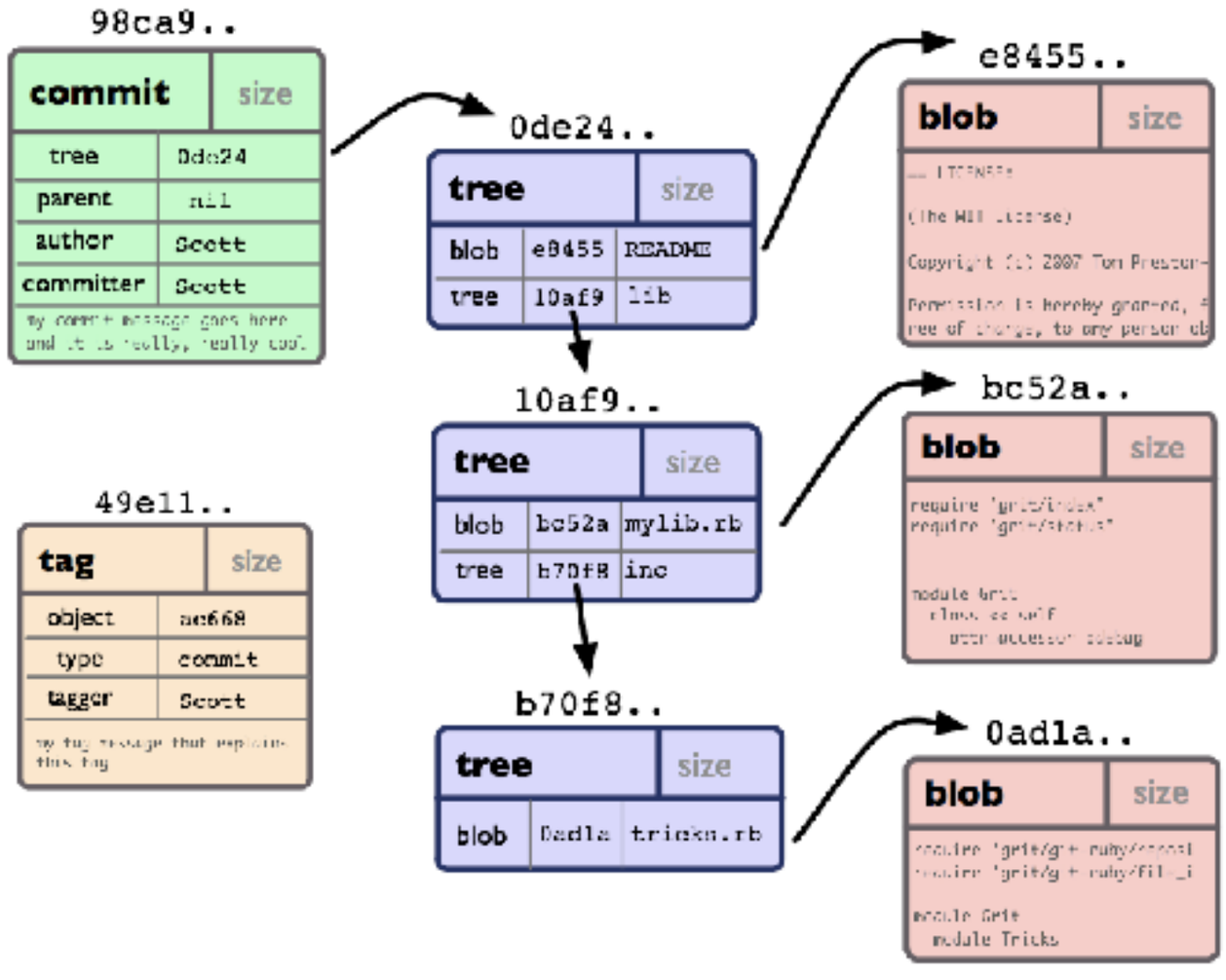
heig-vd
HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

- **Git is a "content-addressable file system"**
- **Git uses a key-value store:**
  - When you store a file in a repo, git computes a SHA-1 hash of its content.
  - The hash is used as a key to index the file in the store.
  - For this reason, two files with exactly the same content are stored only once in the git repository.
  - **Go in the .git hidden directory, have a look at the ./objects directory and you will find this key-value store.**

The Git Object Model

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

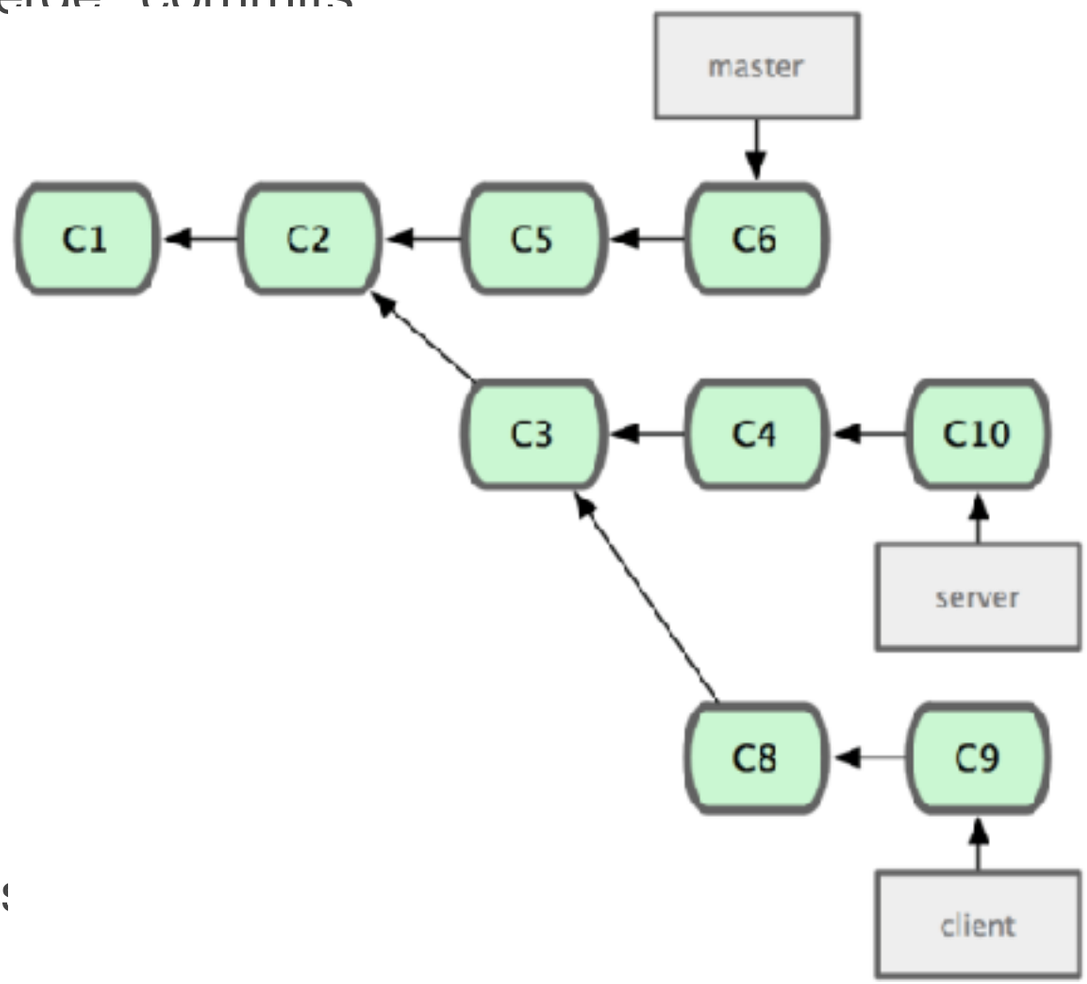
- **BLOBs** store **files**.
- **Trees** store **directory structures** in the file system. A tree has pointers to subtrees and BLOBs.
- **Commits** identify snapshots in the history. **Every commit points to a version of the top-level tree**. A commit also stores metadata: author, parent, message, etc.
- **Annotated Tags** (created with git tag -a) are used to mark releases.
- Recommendation: do not use lightweight tags (created with git tag, without -a)

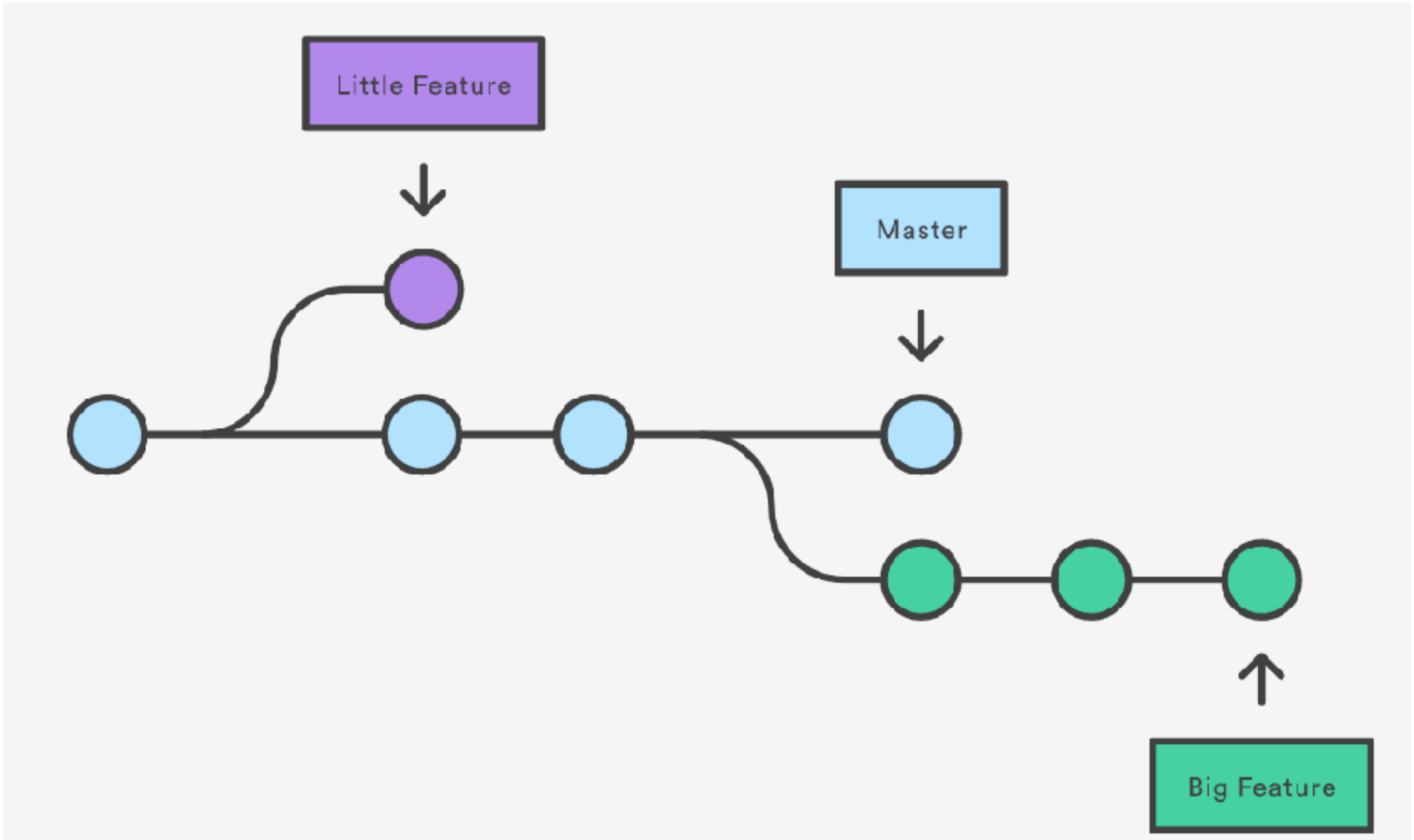


The Git history is a DAG structure

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

- **A DAG is a Directed Acyclic Graph**
  - Except for the initial commit, every commit has a least one parent.
  - Commits with more than one parent are "merge" commits
- **In this example:**
  - C1 is the initial commit
  - There are no merge commits
  - master, server and client are branches
  - The branches have never been merged
  - The last commit on the client branch is C9
- All branches have C1 and C2 as common ance



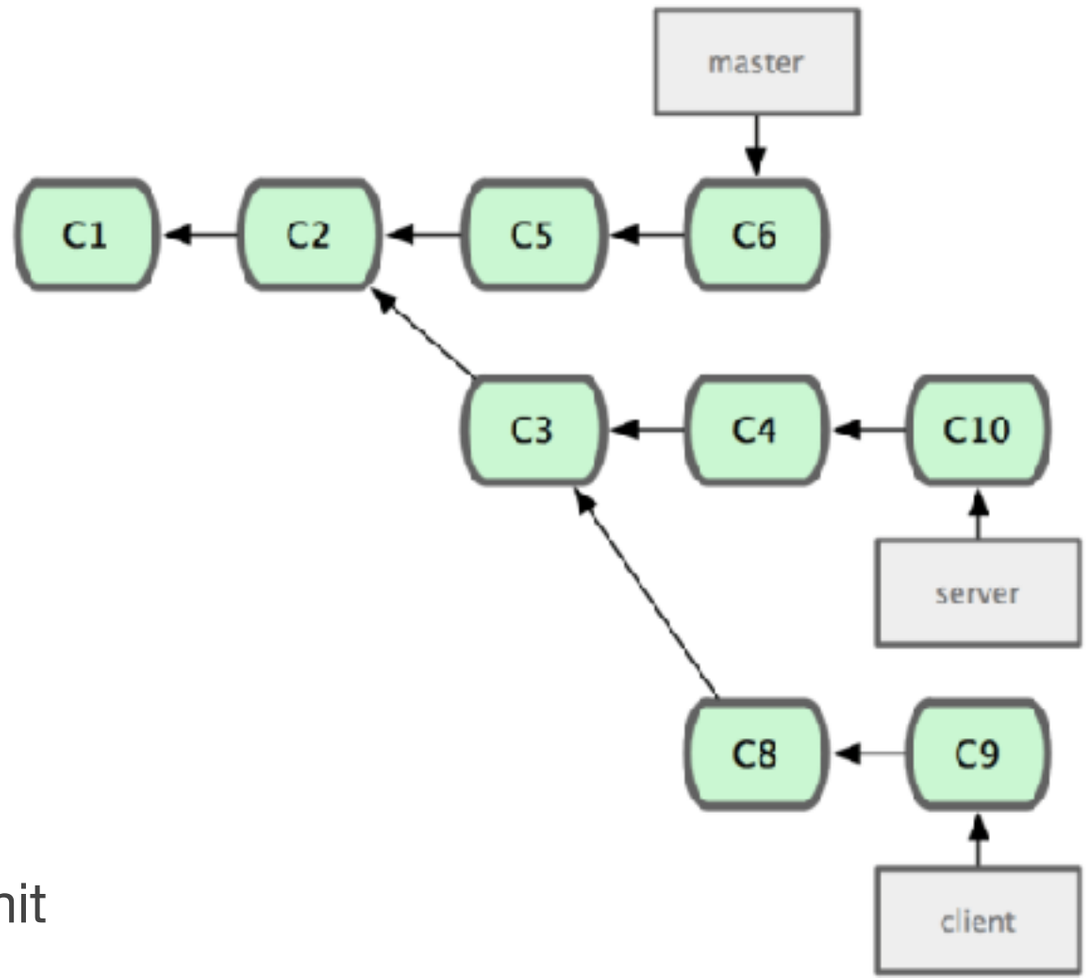


https://www.atlassian.com/git/tutorials/using-branches

Git branches are only pointers

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

- Branches are stored in files located in the **.git/refs/heads** directory.
- **Each branch is stored in a file, which only contains the hash of a commit.**
- **What is HEAD?**
  - HEAD means "the tip of the current branch" (last commit of the current branch)
  - When you checkout a branch, HEAD is the name of this branch.
- **What is HEAD^n?**
  - It is a reference to the nth parent of HEAD
  - HEAD^1 means the parent of the last commit
  - HEAD^2 means its grand-parent



Performance and Buffering

Buffered IOs

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

Do you know what happens when you do a

```
int c = read();
```

?

Buffered IOs

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

It's a bit like when you are thirsty and feel like **drinking** something...

!

Buffered IOs

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

It takes you 56' to sip a beer

20 min

5 min

10 min

20 min

1 min

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

Thirsty again...

?

Buffered IOs

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

It takes you 56' again to sip the next beer

20 min

5 min

10 min

20 min

1 min

Buffered IOs

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

Can we do better...

?

Buffered IOs

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

It still takes you 56' to bring back a pack of beers...

20 min

5 min

10 min

20 min

1 min

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

Thirsty again...

?

Buffered IOs

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

It now only takes 2 minutes to sip the next one!

1 min

1 min

Buffered IOs

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

Coming back to

```
int c = read();
```

- If you don't use buffered IOs, calling read() will issue **one system call** to retrieve **one single byte**... which is not efficient.

- With buffered IOs, calling read() will **pre-fetch "several" bytes** and store it in a **temporary memory space** (i.e. in a **buffer**). "several" defines the **buffer size**.

- Subsequent calls to read() will be able to **fetch bytes directly from the buffer**, which is very fast.

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

What about

```
write(c);
```

?

Buffered IOs

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

It's the same thing! There is on gotcha:

Sometimes, you want to immediately send the content of the buffer to the output stream.

```
os.flush();
```

Buffered IOs in Java

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

- Later on, we will introduce the **Decorator** Design Pattern
- Using buffered IOs is **as simple as decorating any of your byte or character streams** (don't forget about flushing buffered output streams when required!).

```
InputStream slow;
BufferedInputStream fast = new BufferedInputStream(slow);

OutputStream slow;
BufferedOutputStream fast = new BufferedOutputStream(slow);

Reader slow;
BufferedReader fast = new BufferedReader(slow);

Writer slow;
BufferedWriter fast = new BufferedWriter(slow);
```

Example: **01-BufferedIOBenchmark**

*What is the **real** impact of buffered IOs on performance?*

BufferedIOBenchmark

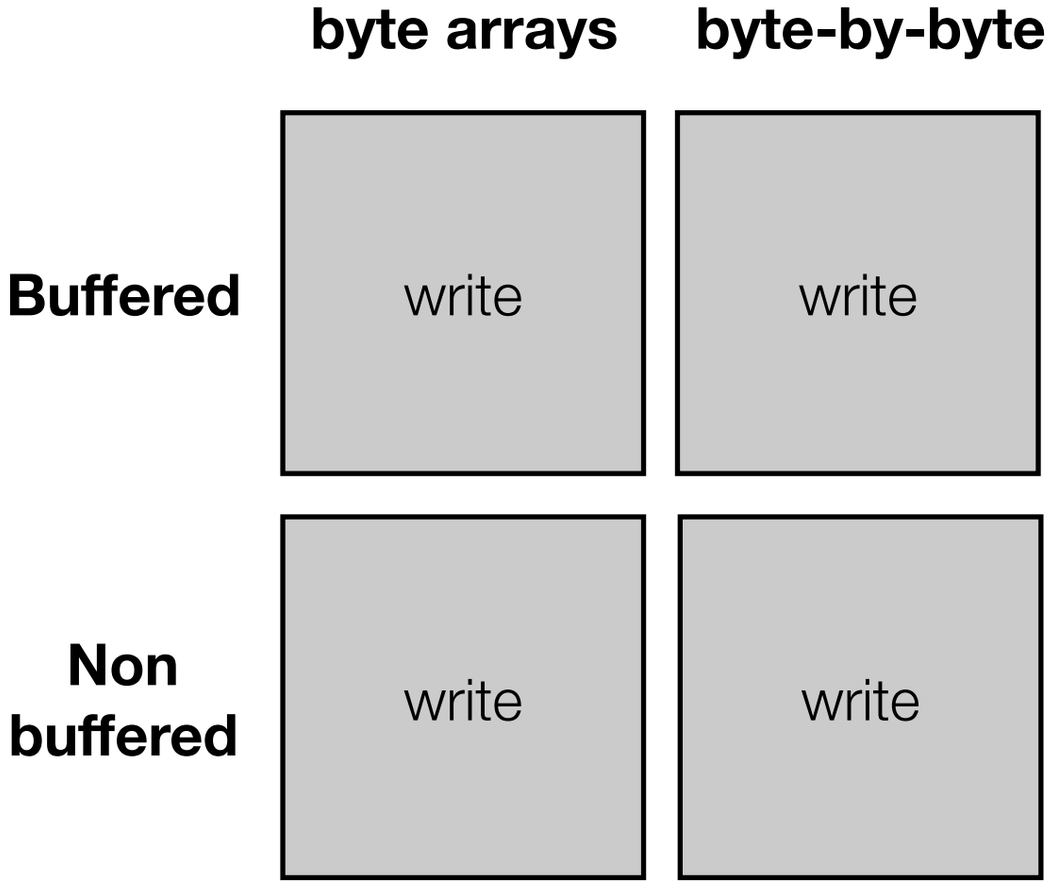
HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

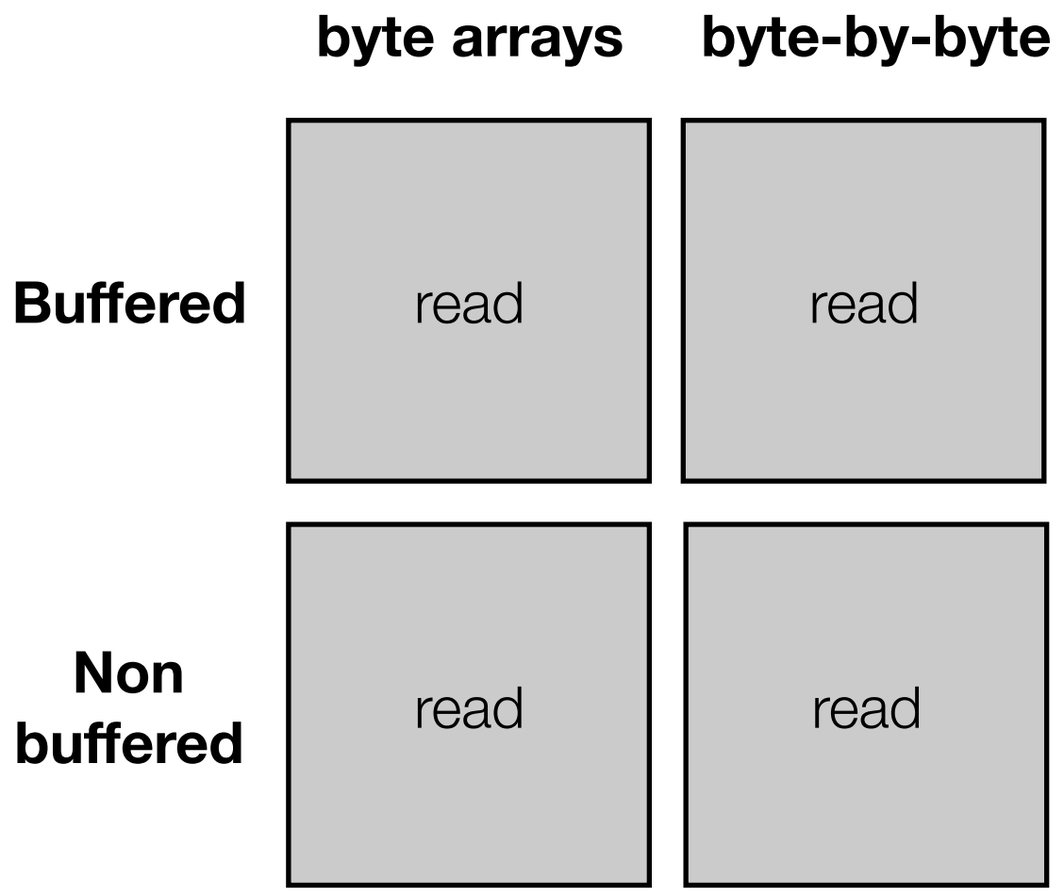
- **Step 1**: code walkthrough and live demo
- **Step 2**: how can we improve the code to be able to analyze results?

Code walkthrough

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

- **Write** and **read** random data to/from **disk**.
- 4 **strategies** (with/without BufferedIOs, operations on byte arrays/single bytes)





Code walkthrough

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

```
/**
 * This enum is used to describe the 4 different strategies for doing the IOs
 */
public enum IOStrategy {
     ByteByByteWithoutBufferedStream,
     ByteByByteWithBufferedStream,
     BlockByBlockWithoutBufferedStream,
     BlockByBlockWithBufferedStream
 };

 // we will write and read test files at this location
 final static String FILENAME_PREFIX = "test-data";

// we will write and read 10 MB files
final static long NUMBER_OF_BYTES_TO_WRITE = 1024 * 1024 * 10;
```

Code walkthrough

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

```
// this takes care of opening CONCRETE output streams (connected to files)

private void produceTestData(IOStrategy ioStrategy, long numberOfBytesToWrite, int blockSize);


// this takes care of producing a random stream of bytes, based on a strategy

private void produceDataToStream(OutputStream os, IOStrategy ioStrategy, long
numberOfBytesToWrite, int blockSize) throws IOException;


// this takes care of opening CONCRETE input streams

private void consumeTestData(IOStrategy ioStrategy, int blockSize);


// this takes care of consuming the stream of bytes, based on a strategy

private void consumeDataFromStream(InputStream is, IOStrategy ioStrategy, int blockSize)
throws IOException
```

Timer

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

```
public class Timer {

     private static long timestamp = 0;

     public static void start() {
          timestamp = System.currentTimeMillis();
     }

     public static long takeTime() {
          if (timestamp == 0) {
               timestamp = System.currentTimeMillis();
          }
          long ellapsedTime = System.currentTimeMillis() - timestamp;
          timestamp = System.currentTimeMillis();
       return ellapsedTime;
     }

}
```

heig-vd
HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

Printing the measures on the console is a first step, but...

- *What is we have 10, 20 strategies?*
- *What if we have 5 parameters for each strategy?*
- *How can we analyze the data more efficiently?*
- *How can we archive the results?*
- *What is we want to analyze/chart in Excel?*

What would you do to improve the code?

The Mighty Filter Classes

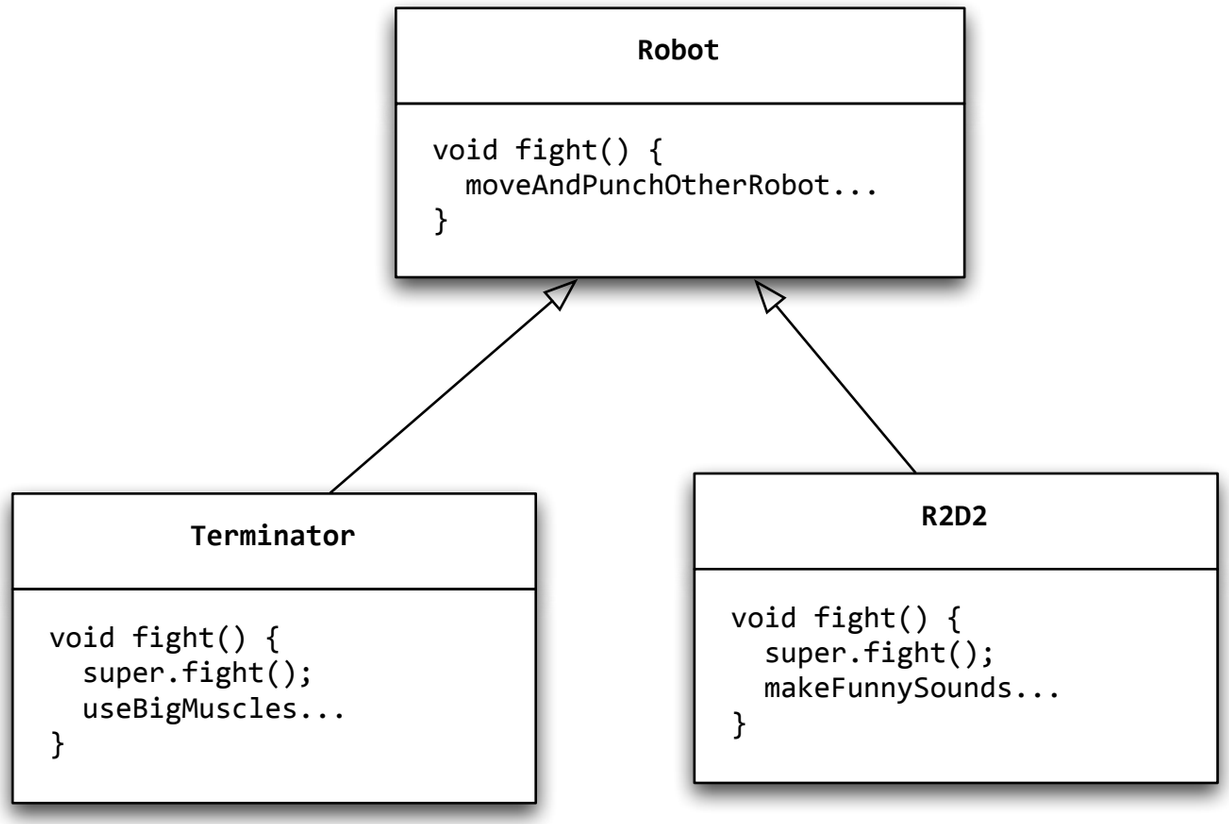
The Decorator Design Pattern

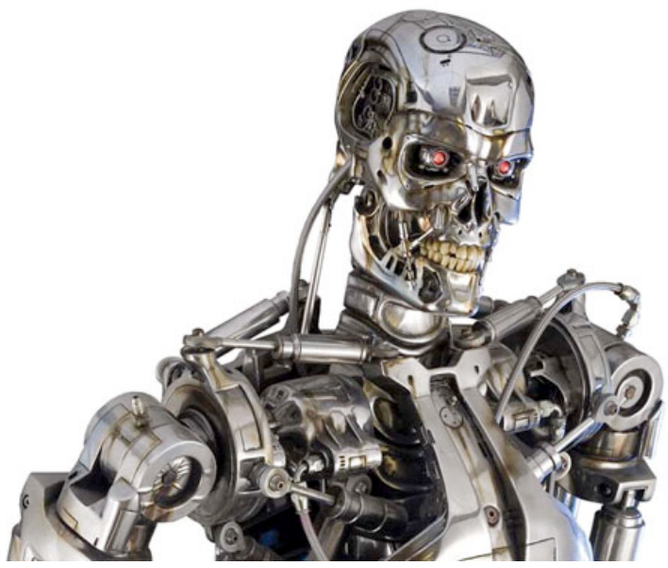
heig-vd
HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

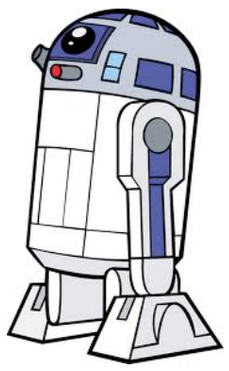
- The **decorator design pattern** is a solution that is often used when creating object-oriented models.
- It makes it possible to **add behavior to an existing class**, without modifying the code of this class. In other words, it makes it possible to **decorate** an existing class with additional behaviors.
- The design pattern also makes it possible to **define a collection of decorators**, and to **combine** them in arbitrary ways at runtime. In other words, it is possible to decorate a class with several behaviors.

Example : Fighting Robots

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch



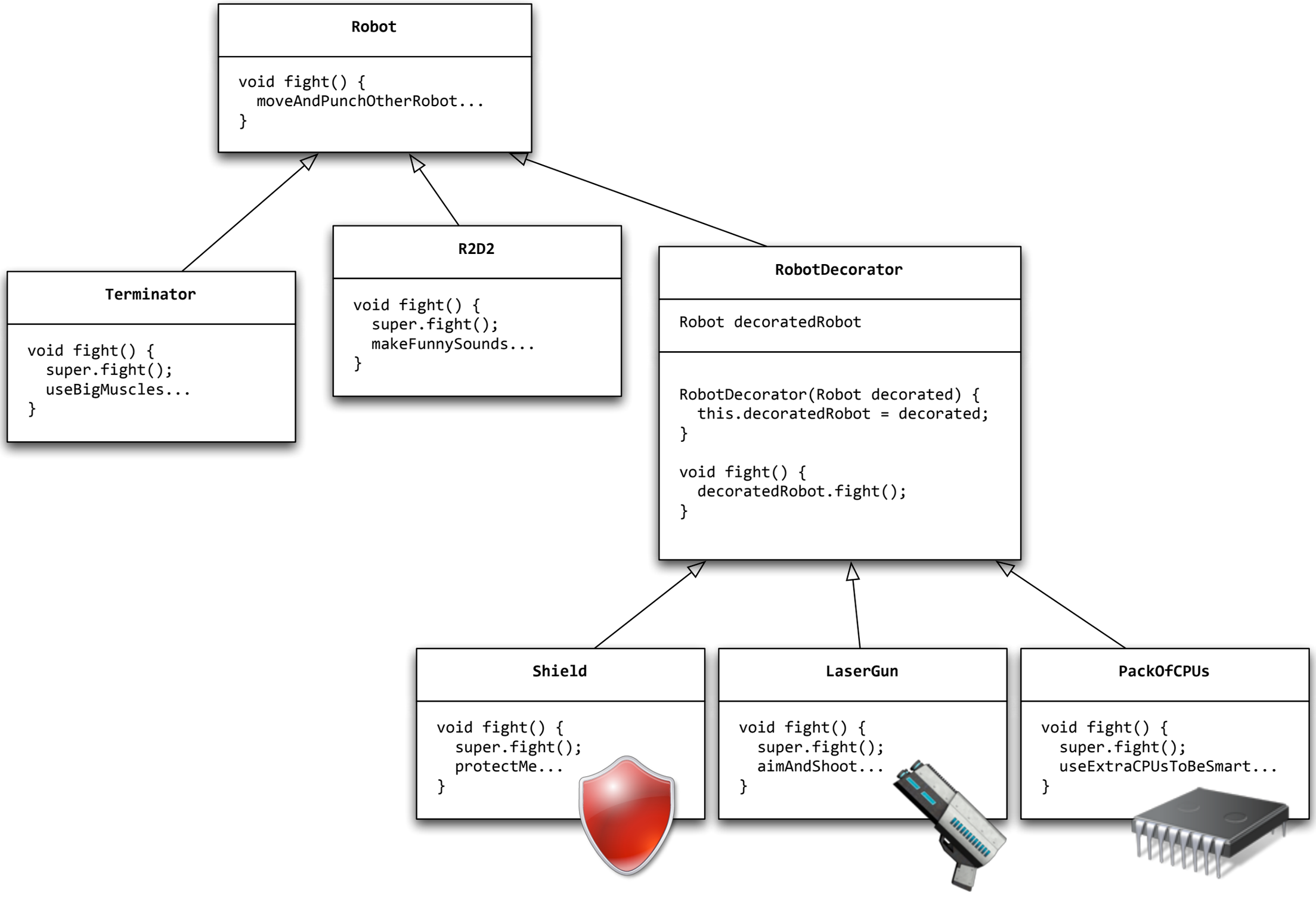




```
Robot bigOne = new Terminator();
Robot smallOne = new R2D2();

bigOne.fight();
smallOne.fight();

--
bigOne   > useBigMuscles
smallOne > makeFunnySounds

bigOne wins.
```

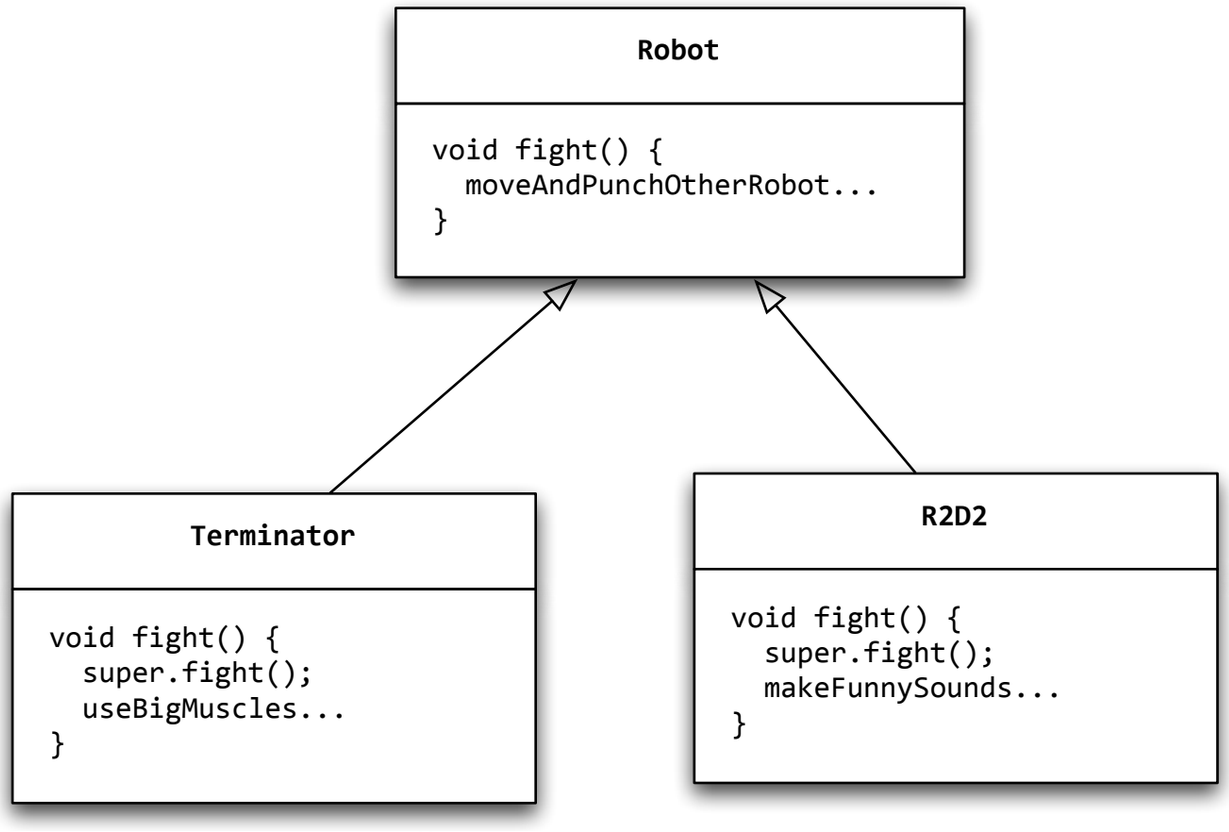
Example : Decorators for Robots

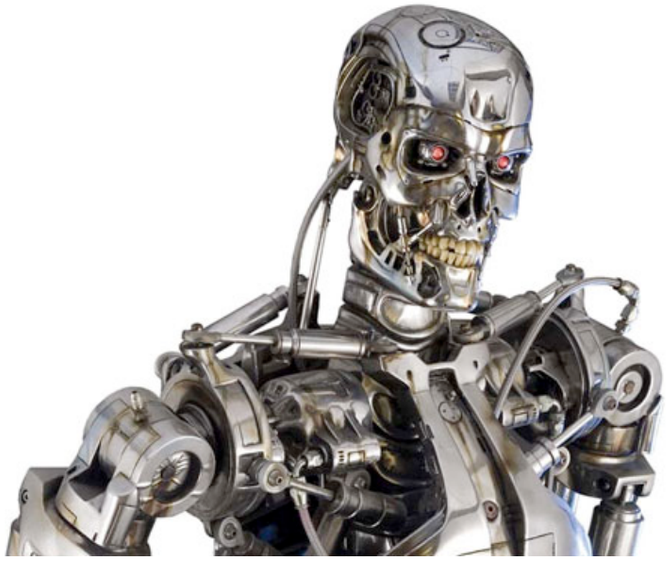
HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch



Example : Decorated Robots

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

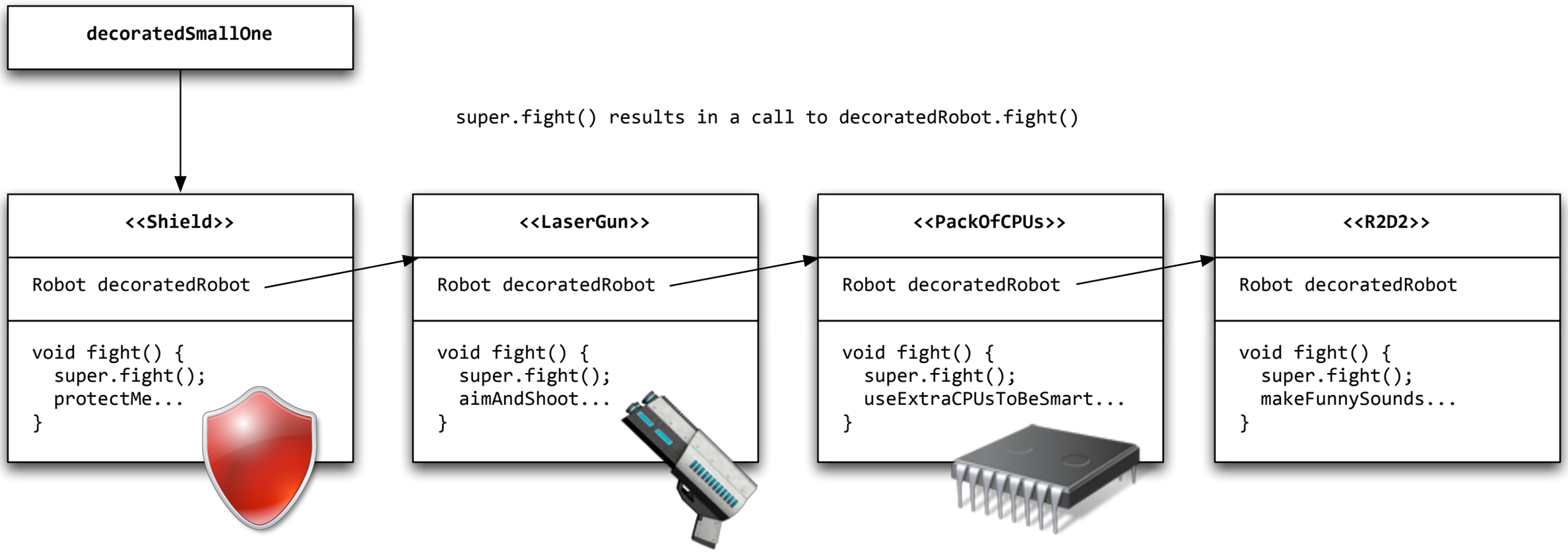




```
Robot bigOne = new Terminator();
Robot decoratedSmallOne =
  new Shield(
    new LaserGun(
      new PackOfCPU(
        new R2D2()
      )
    )
  );

bigOne.fight();
decoratedSmallOne.fight();

--
bigOne    > useBigMuscles
smallOne > makeFunnySounds,
useExtraCPUsToBeSmart, aimAndShoot,
protectMe

decoratedSmallOne wins.
```

Invocation Chain

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch




```
decoratedSmallOne.fight()
-> shield.fight();
-> laserGun.fight() + protectMe;
-> packOfCPUs.fight() + aimAndShoot + protectMe
-> R2D2.fight() + useExtraCPUsToBeSmart + aimAndShoot + protectMe
-> makeFunnySounds + useExtraCPUsToBeSmart + aimAndShoot +
protectMe.
```

Example: **02-FileIOExample** (2)

Binary vs Character-Oriented IOs

Classes in the `java.io` package (1)

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

InputStream

OutputStream

FileInputStream

FileOutputStream

ByteArrayInputStream

ByteArrayOutputStream

PipedInputStream

PipedOutputStream

FilterInputStream

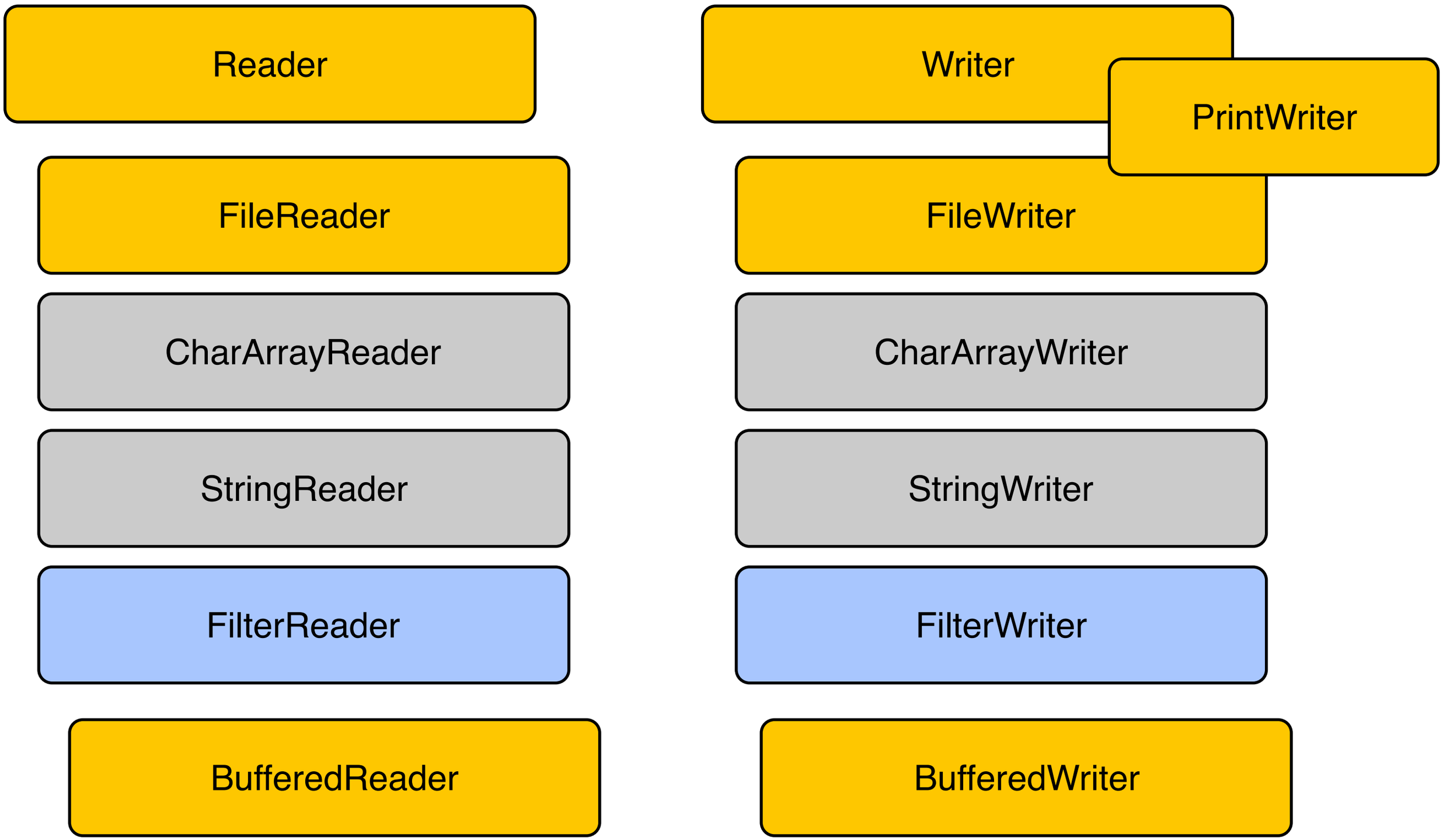
FilterOutputStream

BufferedInputStream

BufferedOutputStream

Classes in the `java.io` package (2)

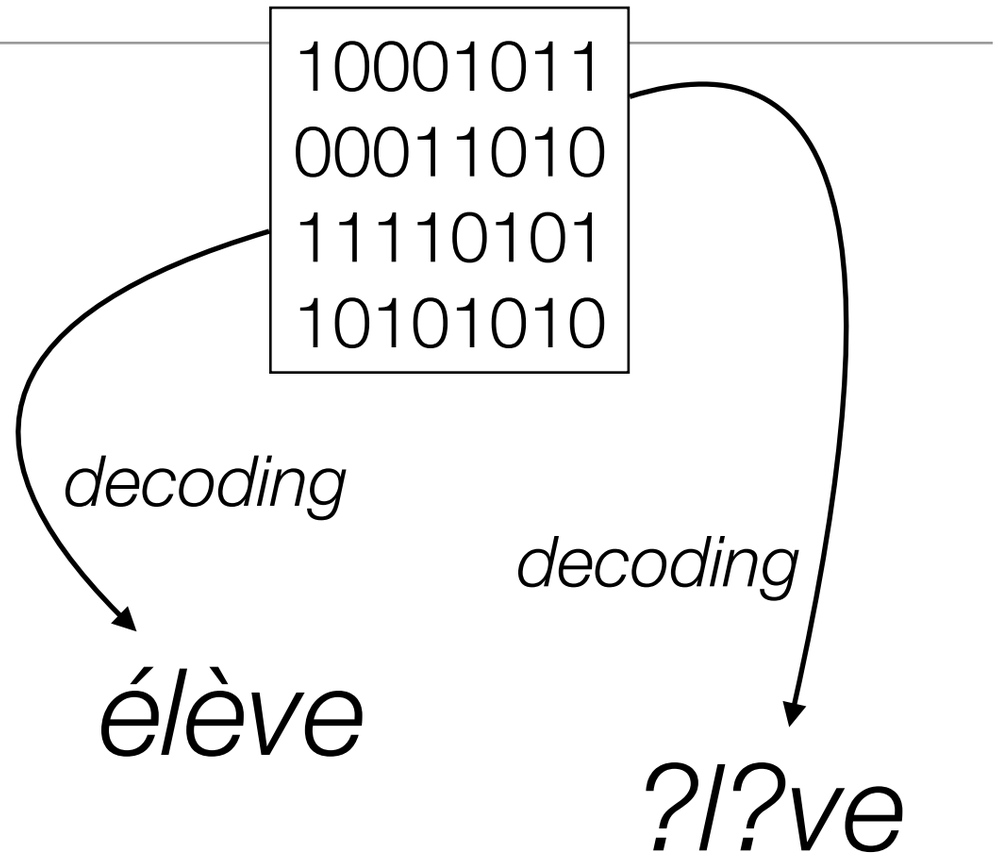
HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch



Bytes vs. Characters

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

- There are different types of data sources. Some sources produce **binary data**. Other sources produce **textual data**.
- It's always a **series of 0's and 1's**. The real question is : "how do you interpret these bits"?
- When you deal with **textual data**, the interpretation is not always the same. It depends on the "**character encoding system**"?
- When you deal with IOs, you have to **use different classes** for processing binary, respectively textual data. **Otherwise, you will corrupt data**!



For binary data, use **inputStreams** and **outputStreams**.
For text data, use **readers** and **writers**.

Bytes vs. Characters

“A byte is a sequence of 8 bits. Period.”

1000001 → 1000001 → 1000001

*“A character is the **interpretation** of a sequence of n bits. Producer and consumer need to **agree** on how to do this interpretation.”*

✓ ‘A’ ⇝ 1000001 ⇝ ‘A’

✗ ‘A’ ⇝ 1000001 ⇝ ‘▯’

Example : the ASCII encoding

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

'A' **1000001** **65**

64-32-16-08-04-02-01

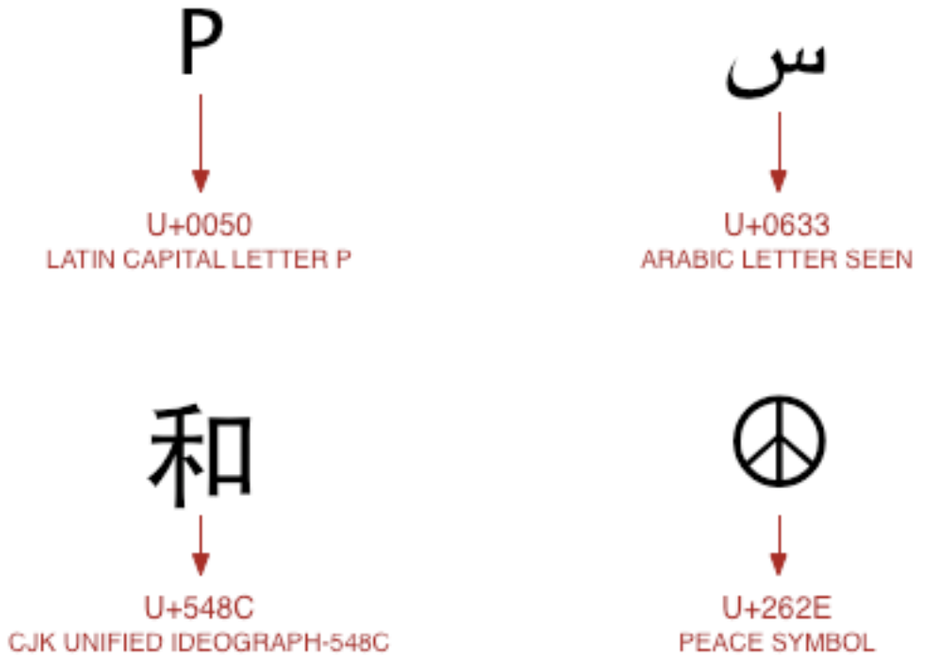
USASCII code chart

b_7 b_6 b_5 → / b_4 b_3 b_2 b_1 ↓ Bits / Row ↓ Column →	0 0 0	0 0 1	0 1 0	0 1 1	1 0 0	1 0 1	1 1 0	1 1 1
0 0 0 0	0 NUL	1 DLE	2 SP	3 0	4 @	5 P	6 `	7 p
0 0 0 1	1 SOH	DC1	!	1	A	Q	a	q
0 0 1 0	2 STX	DC2	"	2	B	R	b	r
0 0 1 1	3 ETX	DC3	#	3	C	S	c	s
0 1 0 0	4 EOT	DC4	$	4	D	T	d	t
0 1 0 1	5 ENQ	NAK	%	5	E	U	e	u
0 1 1 0	6 ACK	SYN	&	6	F	V	f	v
0 1 1 1	7 BEL	ETB	'	7	G	W	g	w
1 0 0 0	8 BS	CAN	(	8	H	X	h	x
1 0 0 1	9 HT	EM	)	9	I	Y	i	y
1 0 1 0	10 LF	SUB	*	:	J	Z	j	z
1 0 1 1	11 VT	ESC	+	;	K	[	k	{
1 1 0 0	12 FF	FS	,	<	L	\	l	|
1 1 0 1	13 CR	GS	-	=	M	]	m	}
1 1 1 0	14 SO	RS	.	>	N	^	n	~
1 1 1 1	15 SI	US	/	?	O	_	o	DEL

Example : Unicode & UTF-8 encoding

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

- **Unicode** is an industry standard for representing text in almost all world languages (e.g. japanese, hebrew, arabic, etc.).
- With Unicode, **every character is associated with a "code point"**, which is nothing else than a number. It is expressed as 'U+' followed by an **hexadecimal** value.
- For instance, 'A' has the code point **U+41** (41 is 65 in hexadecimal).
- The code points for **latin characters** have the same value as in the ASCII encoding.
- **UTF-8** is **one of the encoding systems** used to represent characters of the Unicode character set.
- UTF-8 is a **variable-length encoding system**. Some characters are encoded with 1 byte, others with 2 bytes, etc.



http://wiki.secondlife.com/wiki/Unicode_In_5_Minutes

Java & Unicode

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

- Java uses the **unicode character encoding system**. This means that your program can manipulate characters in different languages and alphabets.
- Every **char** variable is defined by **2 bytes**, i.e. 16 bits. The two bytes
- Think about **what happens when you read data from a source**. You will see a series of 1's and 0's. You know that these bits represent characters, but how do you know how to interpret them? The answer will depend on the source! Or more precisely, **it will depend on the encoding system used by the source**.
- **Same problem when you produce data**. You have text data in memory (char and String variables). You want to understand and control how this data is transformed in a series of bits. This is important if you want that other parties are able to read what you have produced!

Exploring the Unicode Charset

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

http://www.fileformat.info/info/unicode/char/41/index.htm

Unicode Character 'LATIN CAPITAL LETTER A' (U+0041)

A

Browser Test Page
Outline (as SVG file)
Fonts that support U+0041

Unicode Data	
Name	LATIN CAPITAL LETTER A
Block	Basic Latin
Category	Letter, Uppercase [Lu]
Combine	0
BIDI	Left-to-Right [L]
Mirror	N
Index entries	Latin Uppercase Alphabet Uppercase Alphabet, Latin Capital Letters, Latin
Lower case	U+0061
Version	Unicode 1.1.0 (June, 1993)

Encodings	
HTML Entity (decimal)	A
HTML Entity (hex)	A
How to type in Microsoft Windows	Alt +41 Alt 065 Alt 65
UTF-8 (hex)	0x41 (41)
UTF-8 (binary)	01000001
UTF-16 (hex)	0x0041 (0041)
UTF-16 (decimal)	65
UTF-32 (hex)	0x00000041 (41)
UTF-32 (decimal)	65
C/C++/Java source code	"\u0041"
Python source code	u"\u0041"
More...	

Example: **03-CharacterIODemo**

*How do I avoid these damn **?l?ves** when I want to see **élèves**?*

CharacterIODemo

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

- **Step 1**: code walkthrough and live demo (10')
- **Step 2**: self-study of the code and individual experiments

Demo

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

```
When I encode 'ABC élève 広島' (12 chars) with encoding US-ASCII, I generate 12 bytes.
01000001     65
01000010     66
01000011     67
00100000     32
00111111     63
01101100    108
00111111     63
01110110    118
01100101    101
00111111     63
00111111     63
00111111     63
If I decode the result with the default encoding for this JVM (UTF-8), I get: ABC ?l?ve???
If I decode the result with the same encoding (US-ASCII) for this JVM (UTF-8), I get: ABC ?l?ve???
```

Demo

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

```
When I encode 'ABC élève 広島' (12 chars) with encoding ISO-8859-15, I generate 12 bytes.
01000001     65
01000010     66
01000011     67
00100000     32
11101001    -23
01101100    108
11101000    -24
01110110    118
01100101    101
00111111     63
00111111     63
00111111     63
If I decode the result with the default encoding for this JVM (UTF-8), I get: ABC lve???
If I decode the result with the same encoding (ISO-8859-15) for this JVM (UTF-8), I get: ABC élève???
```

Demo

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

```
When I encode 'ABC élève  広島' (12 chars) with encoding UTF-8, I generate 20 bytes.
01000001     65
01000010     66
01000011     67
00100000     32
11000011    -61
10101001    -87
01101100    108
11000011    -61
10101000    -88
01110110    118
01100101    101
11100011    -29
10000000   -128
10000000   -128
11100101    -27
10111010    -70
10000011   -125
11100101    -27
10110011    -77
10110110    -74
If I decode the result with the default encoding for this JVM (UTF-8), I get: ABC élève  広島
If I decode the result with the same encoding (UTF-8) for this JVM (UTF-8), I get: ABC élève  広島
```

Demo

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

```
When I encode 'ABC élève 広島' (12 chars) with encoding UTF-16, I generate 26 bytes.
11111110    -2
11111111    -1
00000000     0
01000001    65
00000000     0
01000010    66
00000000     0
01000011    67
00000000     0
00100000    32
00000000     0
11101001   -23
00000000     0
01101100   108
00000000     0
11101000   -24
00000000     0
01110110   118
00000000     0
01100101   101
00110000    48
00000000     0
01011110    94
10000011  -125
01011100    92
11110110   -10
If I decode the result with the default encoding for this JVM (UTF-8), I get: ABC lve0^\
If I decode the result with the same encoding (UTF-16) for this JVM (UTF-8), I get: ABC élève 広島
```

Code walkthrough

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

```
// Lets create an output stream, which will send written bytes to a memory zone
ByteArrayOutputStream os = new ByteArrayOutputStream();
...
byte[] encodedMessage = os.toByteArray();
...
decodedMessage = new String(encodedMessage, encoding);
```

```
OutputStreamWriter writer = new OutputStreamWriter(os, encoding);
writer.write(message);
```

Beware of the FileReader and FileWriter!

heig-vd
HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

no way to specify an encoding!!!

Constructor Summary

Constructors

Constructor and Description
`FileReader(File file)` Creates a new `FileReader`, given the `File` to read from.
`FileReader(FileDescriptor fd)` Creates a new `FileReader`, given the `FileDescriptor` to read from.
`FileReader(String fileName)` Creates a new `FileReader`, given the name of the file to read from.

InputStreamReader to the rescue

Constructor Summary

Constructors

Constructor and Description
`InputStreamReader(InputStream in)` Creates an InputStreamReader that uses the default charset.
`InputStreamReader(InputStream in, Charset cs)` Creates an InputStreamReader that uses the given charset.
`InputStreamReader(InputStream in, CharsetDecoder dec)` Creates an InputStreamReader that uses the given charset decoder.
`InputStreamReader(InputStream in, String charsetName)` Creates an InputStreamReader that uses the named charset.

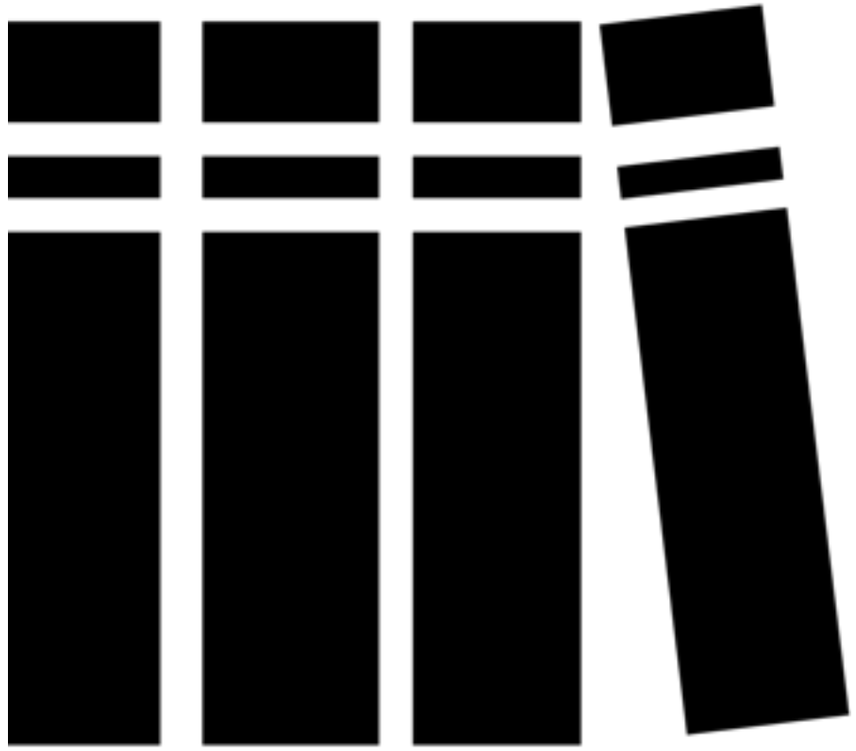
```
Reader reader = new InputStreamReader(new FileInputStream(file), “UTF-8”);
is safer than
Reader reader = new FileReader(file); // what is the “default” encoding???
```

Shit Happens...
Dealing with IO Exceptions

Exception Summary

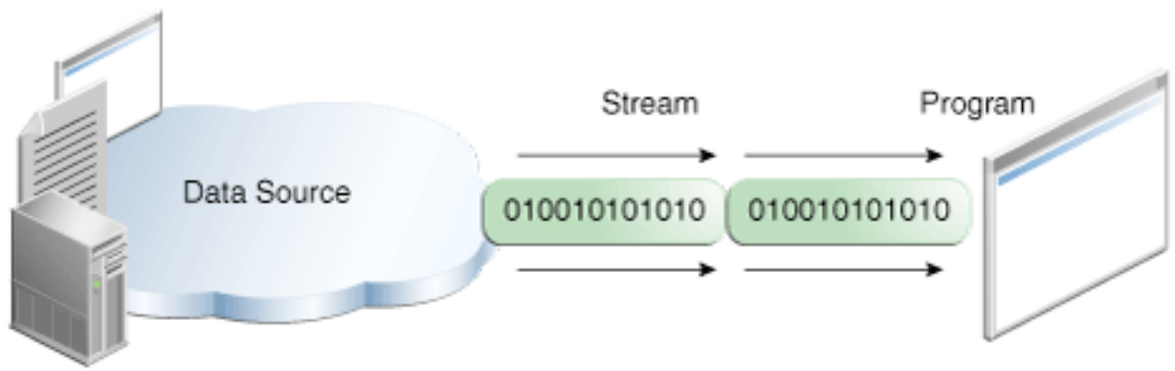
Exception	Description
CharConversionException	Base class for character conversion exceptions.
EOFException	Signals that an end of file or end of stream has been reached unexpectedly during input.
FileNotFoundException	Signals that an attempt to open the file denoted by a specified pathname has failed.
InterruptedIOException	Signals that an I/O operation has been interrupted.
InvalidClassException	Thrown when the Serialization runtime detects one of the following problems with a Class.
InvalidObjectException	Indicates that one or more deserialized objects failed validation tests.
IOException	Signals that an I/O exception of some sort has occurred.
NotActiveException	Thrown when serialization or deserialization is not active.
NotSerializableException	Thrown when an instance is required to have a Serializable interface.
ObjectStreamException	Superclass of all exceptions specific to Object Stream classes.
OptionalDataException	Exception indicating the failure of an object read operation due to unread primitive data, or the end of data belonging to a serialized object in the stream.
StreamCorruptedException	Thrown when control information that was read from an object stream violates internal consistency checks.
SyncFailedException	Signals that a sync operation has failed.
UnsupportedEncodingException	The Character Encoding is not supported.
UTFDataFormatException	Signals that a malformed string in modified UTF-8 format has been read in a data input stream or by any class that implements the data input interface.
WriteAbortedException	Signals that one of the ObjectStreamExceptions was thrown during a write operation.

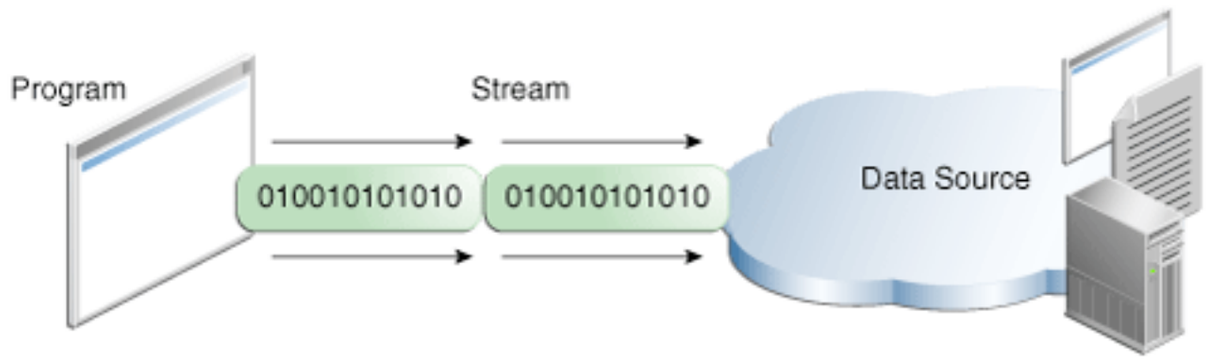
References



The IO Trail in the Java Tutorial

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

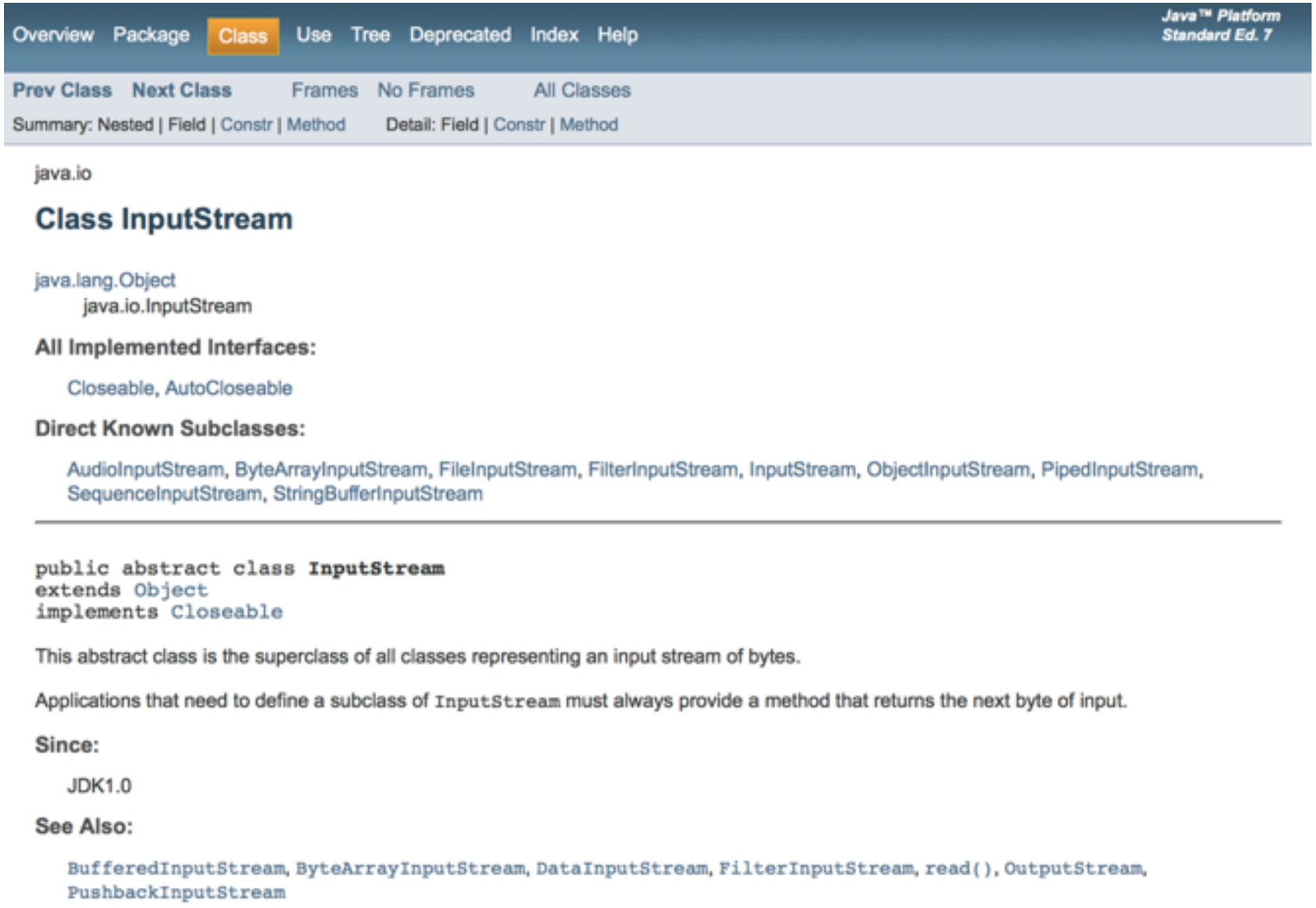




http://docs.oracle.com/javase/tutorial/essential/io/index.html

The `java.io` API documentation

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch



Overview Package Class Use Tree Deprecated Index Help — Java™ Platform Standard Ed. 7

Prev Class Next Class Frames No Frames All Classes

Summary: Nested | Field | Constr | Method Detail: Field | Constr | Method

java.io

Class InputStream

java.lang.Object
    java.io.InputStream

All Implemented Interfaces:

Closeable, AutoCloseable

Direct Known Subclasses:

AudioInputStream, ByteArrayInputStream, FileInputStream, FilterInputStream, InputStream, ObjectInputStream, PipedInputStream, SequenceInputStream, StringBufferInputStream

---

```
public abstract class InputStream
extends Object
implements Closeable
```

This abstract class is the superclass of all classes representing an input stream of bytes.

Applications that need to define a subclass of `InputStream` must always provide a method that returns the next byte of input.

Since:

JDK1.0

See Also:

`BufferedInputStream`, `ByteArrayInputStream`, `DataInputStream`, `FilterInputStream`, `read()`, `OutputStream`, `PushbackInputStream`

http://docs.oracle.com/javase/7/docs/api/java/io/package-summary.html

Books

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch