Text Mining II - Diploma en Big Data (2014-2015)

Optimización PAN-AP-2013

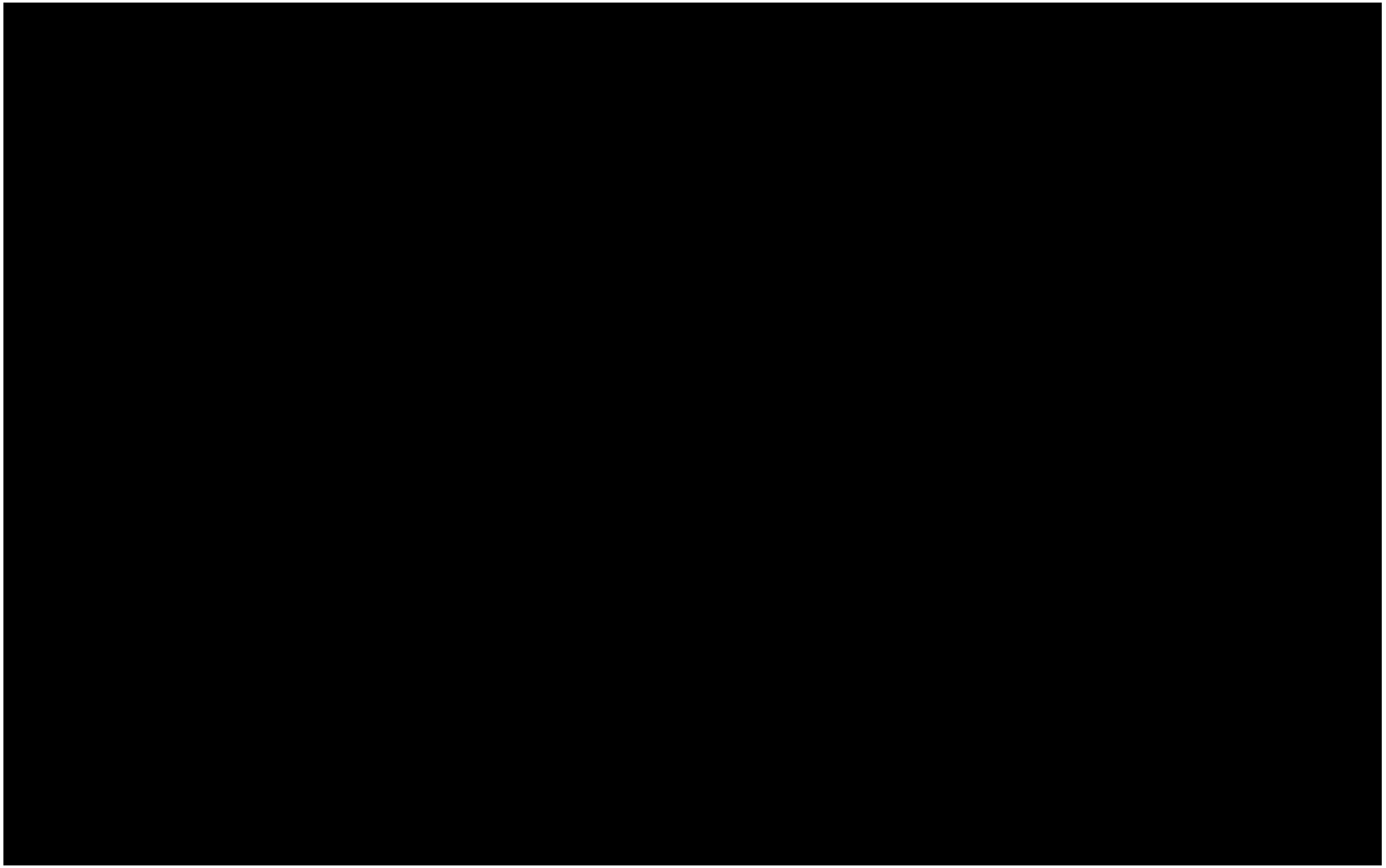
Francisco Rangel
Paolo Rosso

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

String+=

https://s3.amazonaws.com/cosmos_presentaciones/StringPlusEqual.mp4

StringBuilder.add



https://s3.amazonaws.com/cosmos_presentaciones/StringBuilder.mp4

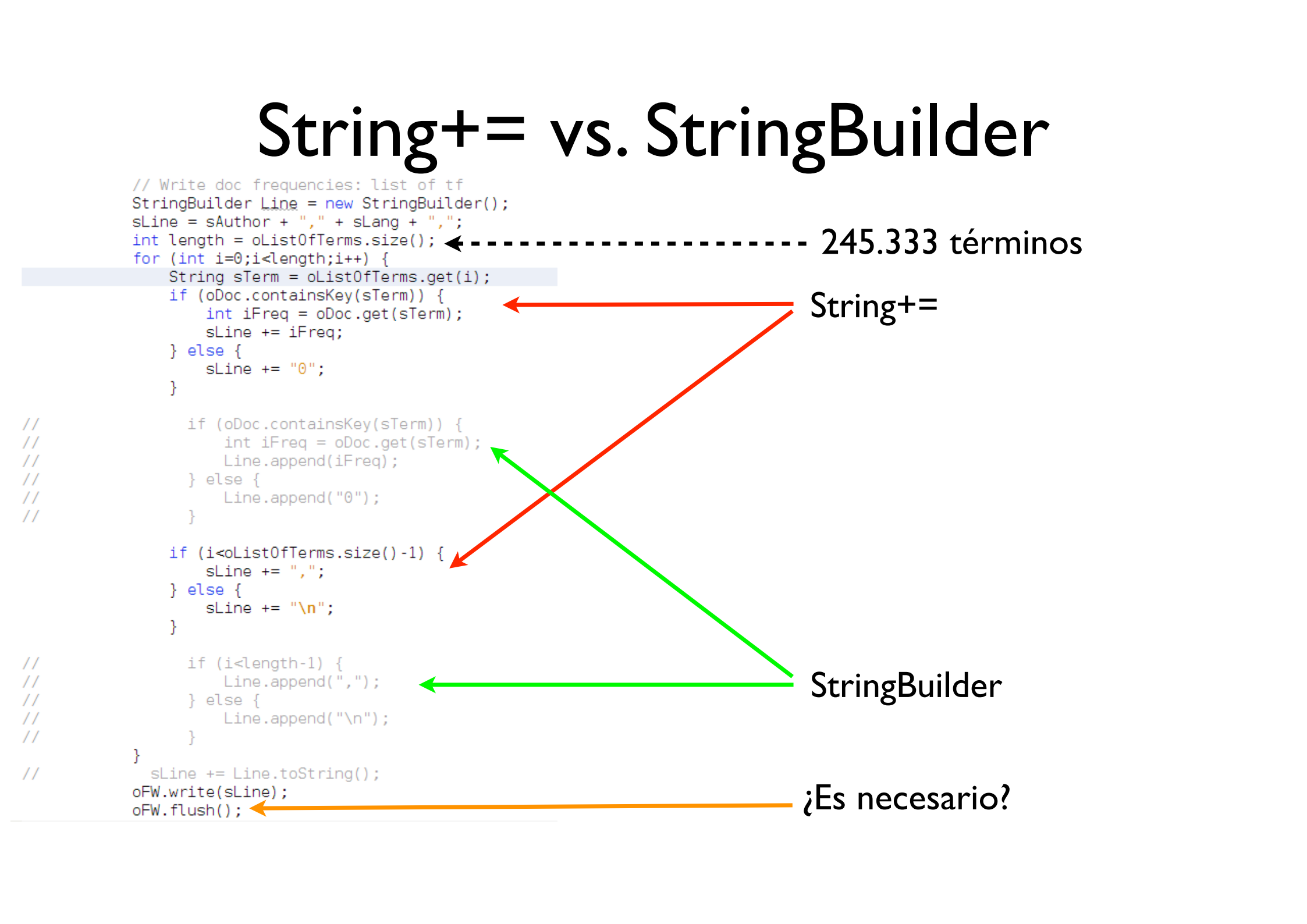
String+= vs. StringBuilder

```
            // Write doc frequencies: list of tf
            StringBuilder Line = new StringBuilder();
            sLine = sAuthor + "," + sLang + ",";
            int length = oListOfTerms.size();
            for (int i=0;i<length;i++) {
                String sTerm = oListOfTerms.get(i);
                if (oDoc.containsKey(sTerm)) {
                    int iFreq = oDoc.get(sTerm);
                    sLine += iFreq;
                } else {
                    sLine += "0";
                }

//                  if (oDoc.containsKey(sTerm)) {
//                      int iFreq = oDoc.get(sTerm);
//                      Line.append(iFreq);
//                  } else {
//                      Line.append("0");
//                  }

                if (i<oListOfTerms.size()-1) {
                    sLine += ",";
                } else {
                    sLine += "\n";
                }

//                  if (i<length-1) {
//                      Line.append(",");
//                  } else {
//                      Line.append("\n");
//                  }
            }
//              sLine += Line.toString();
            oFW.write(sLine);
            oFW.flush();
```

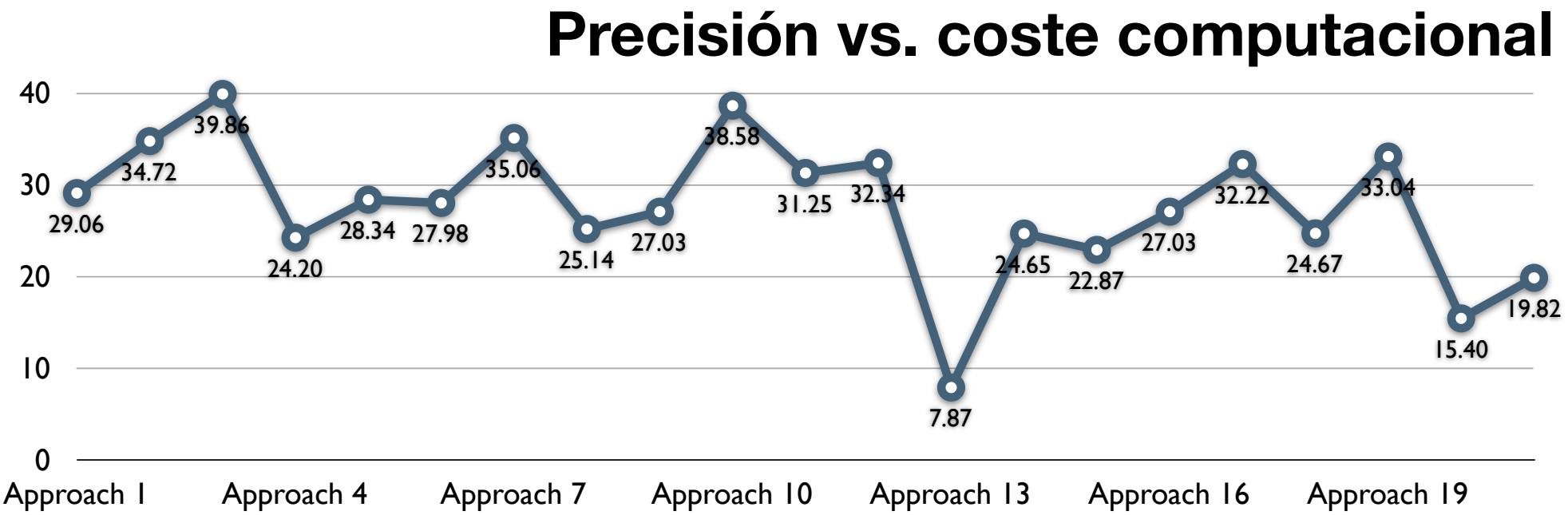


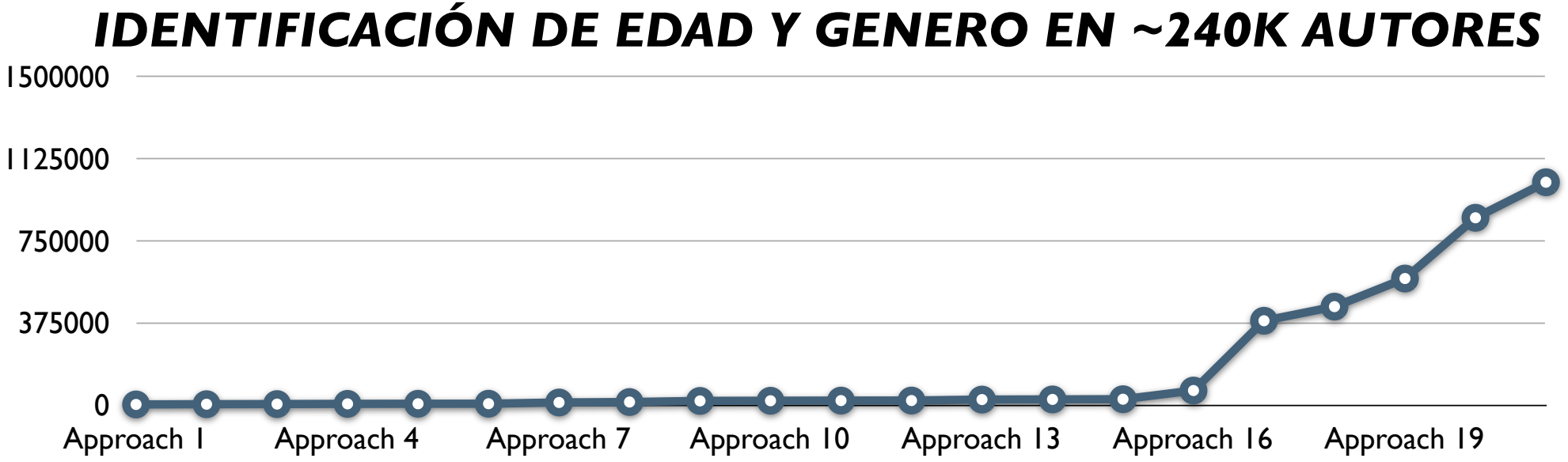
Big Data

https://s3.amazonaws.com/cosmos_presentaciones/ataque.mov

- Recuperar y almacenar
- Evolución
- Palabras y temas
- Etiquetado
- Hashtags
- Personas
- Localizaciones
- Marcas
- Tono, emoción
- Usuario, relaciones
- Influencia
- Sexo y edad
- Perfil autor
- ...

80~120 tuits/segundo =4.800~7.200 tuits/minuto =288.000~432.000 tuits/hora =6.912.000~10.368.000 tuits/día


Precisión vs. coste computacional
40
30
20
10
0
29.06
34.72
39.86
24.20
28.34
27.98
35.06
25.14
27.03
38.58
31.25
32.34
7.87
24.65
22.87
27.03
32.22
24.67
33.04
15.40
19.82
Approach 1
Approach 4
Approach 7
Approach 10
Approach 13
Approach 16
Approach 19


IDENTIFICACIÓN DE EDAD Y GENERO EN ~240K AUTORES
1500000
1125000
750000
375000
0
Approach 1
Approach 4
Approach 7
Approach 10
Approach 13
Approach 16
Approach 19