CPSBayes
naïve bayesian classifier for CPS

Tarek Ziadé, Nuxeo
tz@nuxeo.com

Who am i

- I am engineer at Nuxeo
- I work on CPS, the famous ECM Platform ;)

What is a naïve Bayesian classifier ?

it's quite simple :

$$p(C|F_1,\ldots,F_n)$$

$$p(C|F_1,\ldots,F_n) = \frac{p(C)\,p(F_1,\ldots,F_n|C)}{p(F_1,\ldots,F_n)}.$$

$$\text{classify}(f_1,\ldots,f_n) = \text{argmax}_c\; p(C=c)\prod_{i=1}^{n} p(F_i=f_i|C=c)$$

$$= p(C)\,p(F_1|C)\,p(F_2,\ldots,F_n|C,F_1)$$

take a pencil, resolve the equation, you have 5 minutes

CPS

nuxeo

What is a naïve Bayesian classifier ?

- A simple probabilistic classifier
- "Bayesian" stands for the algorithm in use to evaluate data
- "Naïve" stands for strong independence assumptions on data
- Adapts itself with supervised learning

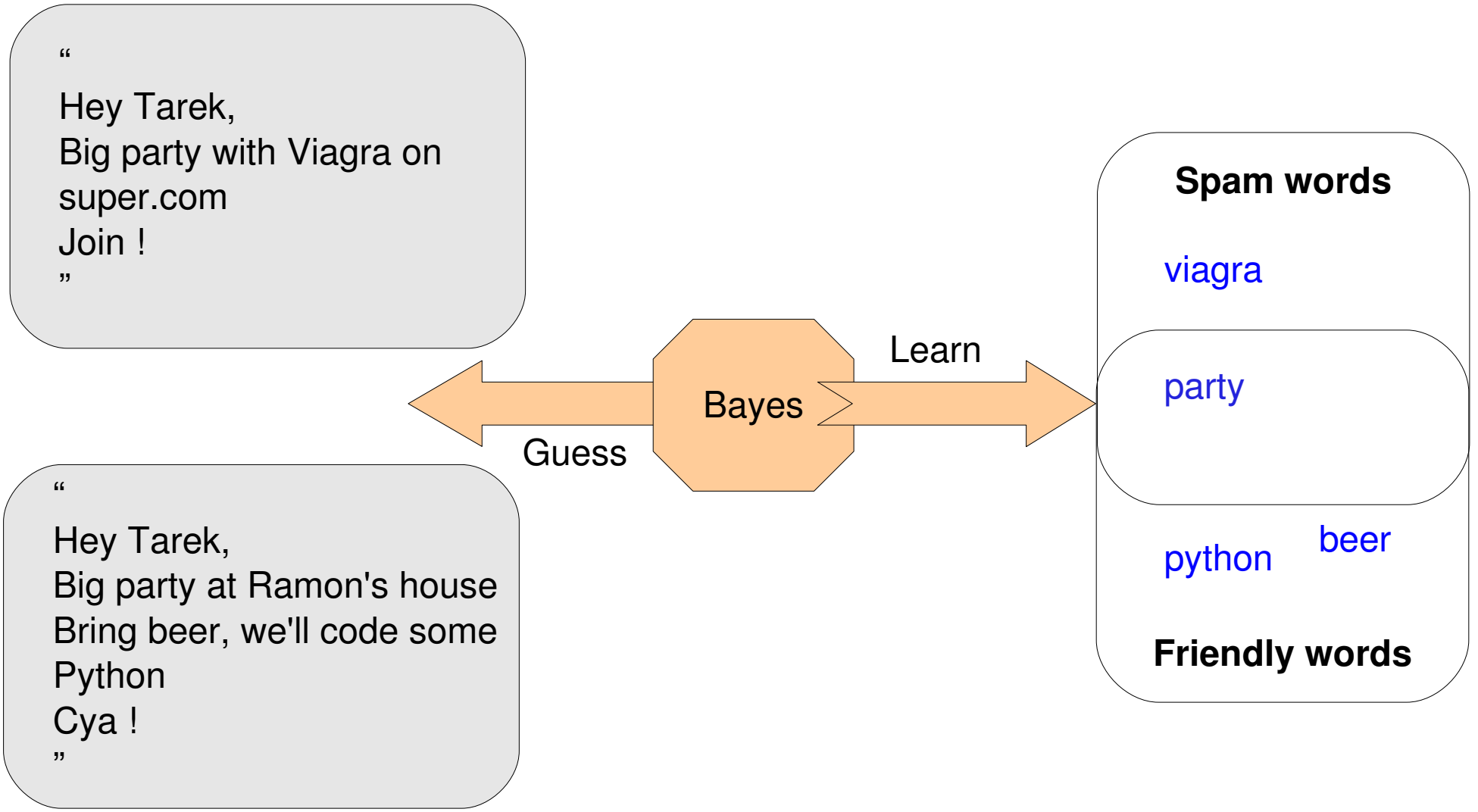
CPS

nuxeo

What is used for ?

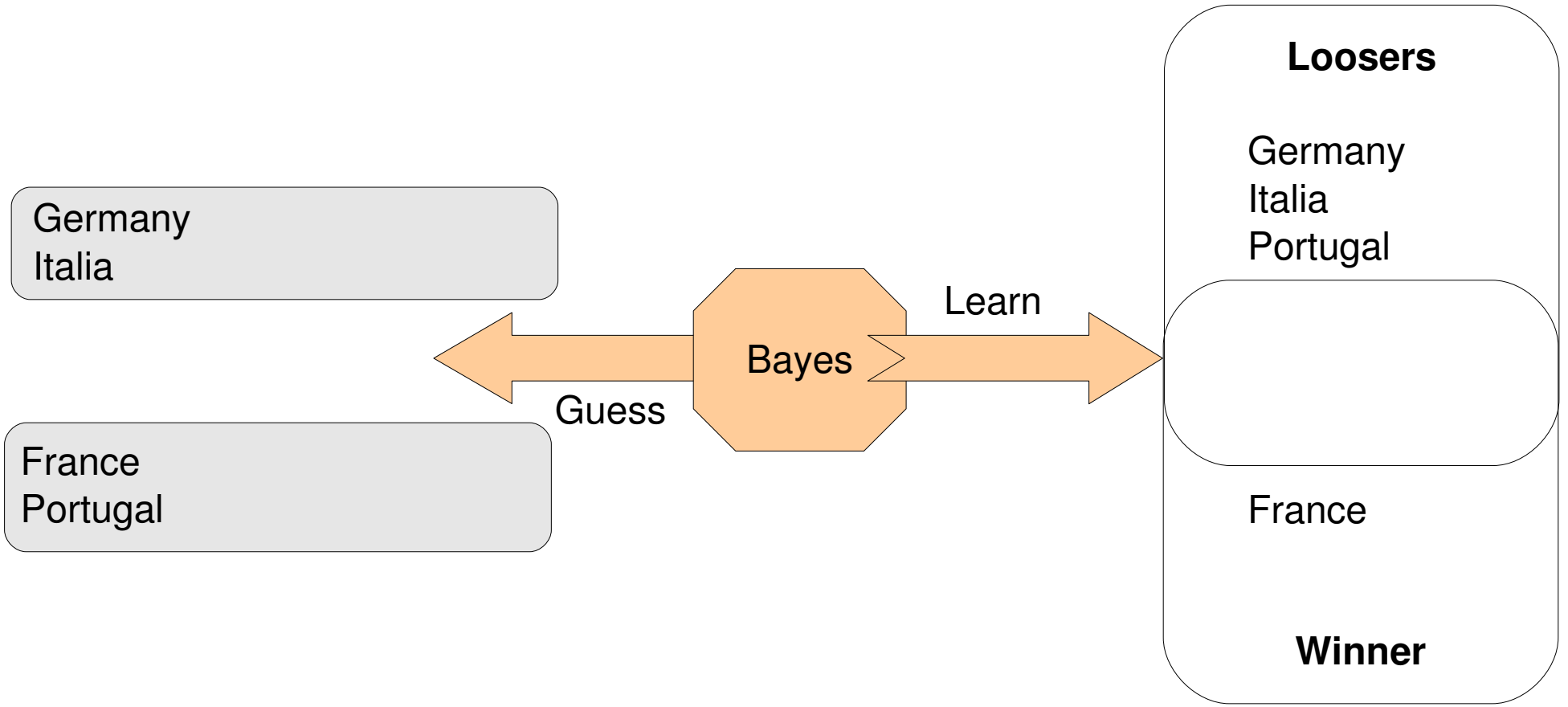
- Document classification
- Anti-spam software
- TextMining
- DataMining
- etc.

The big picture of inference



Anti-spam example

The big picture of inference



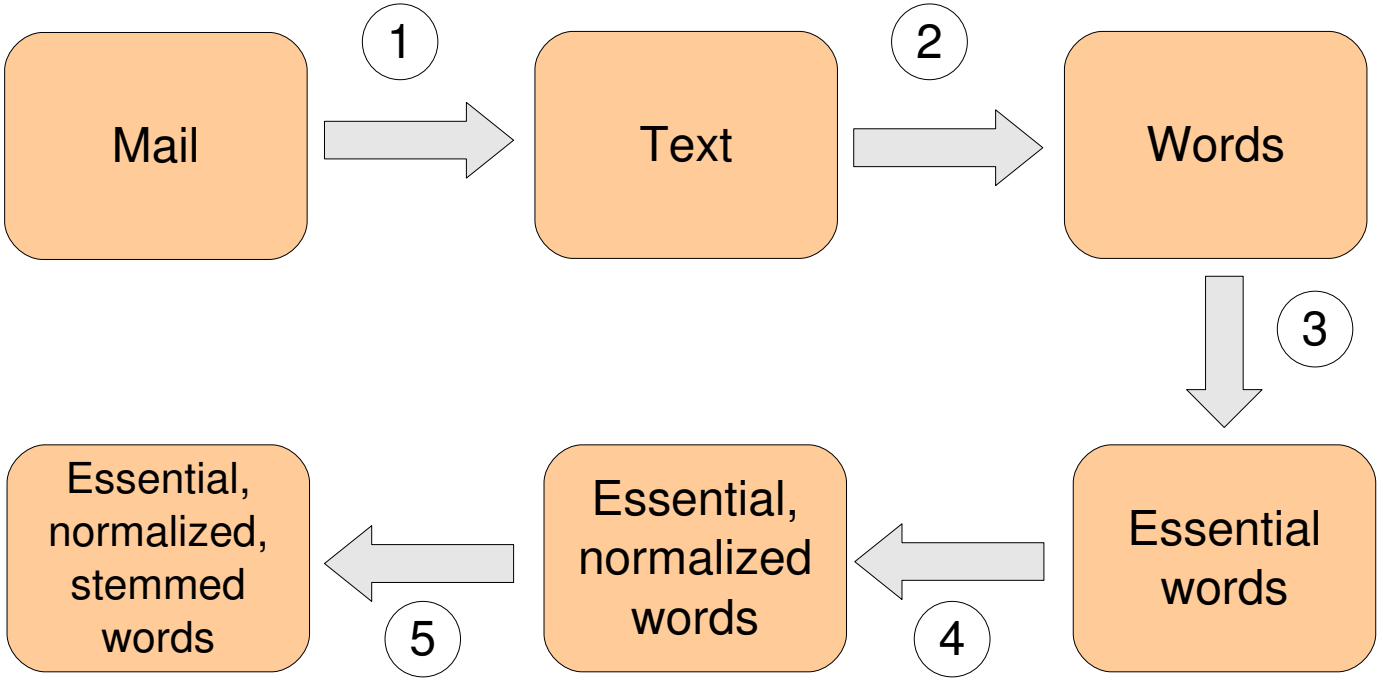
Another quite self-explanatory example

Anti-spamming complete process

How antispam is done ?

- Words are extracted form Mail
- Words are processed into the Bayesian methods
- Weight of categorized words are kept, up to date in a database

Mail preprocessing



The zopyx.txng3 package

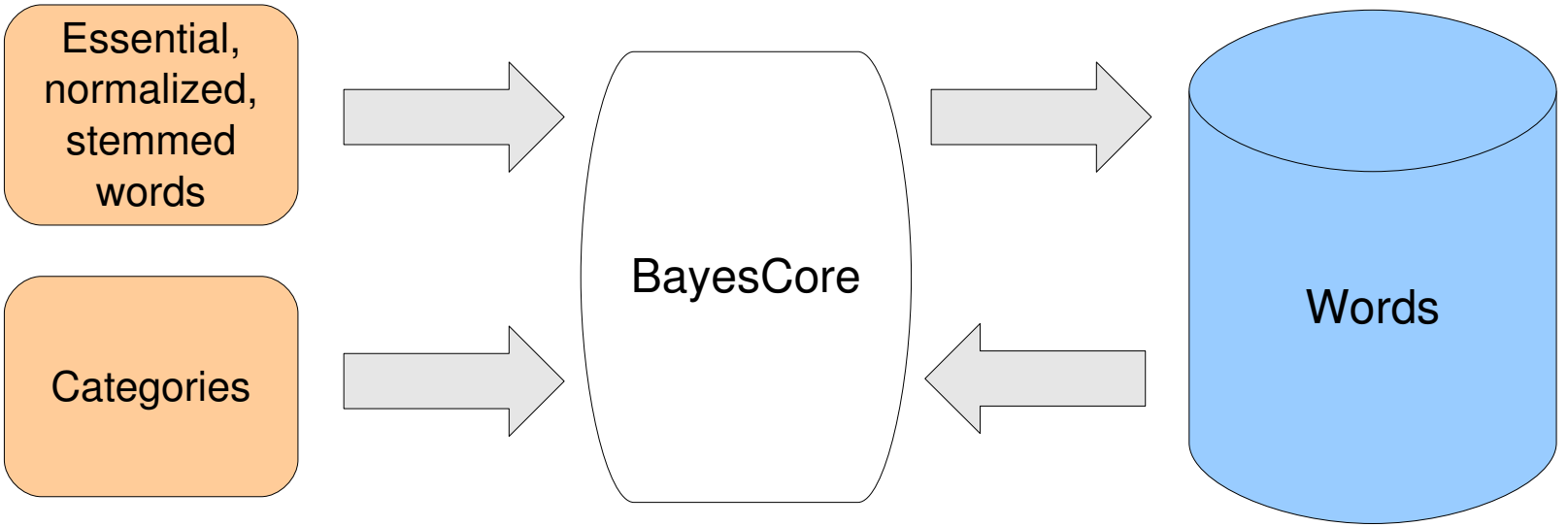
- Coded and maintained by Andreas Jung
- Derivates from TextIndexNG a Zope 2 product
- Provides:
  - Splitting
  - Normalizing
  - Stemming (Snowball)
  - etc.

nuxeo

Mail preprocessing

- Text body extracted by Python email module
- Text splitting
- Stop words applied
- Normalizing
- Stemming

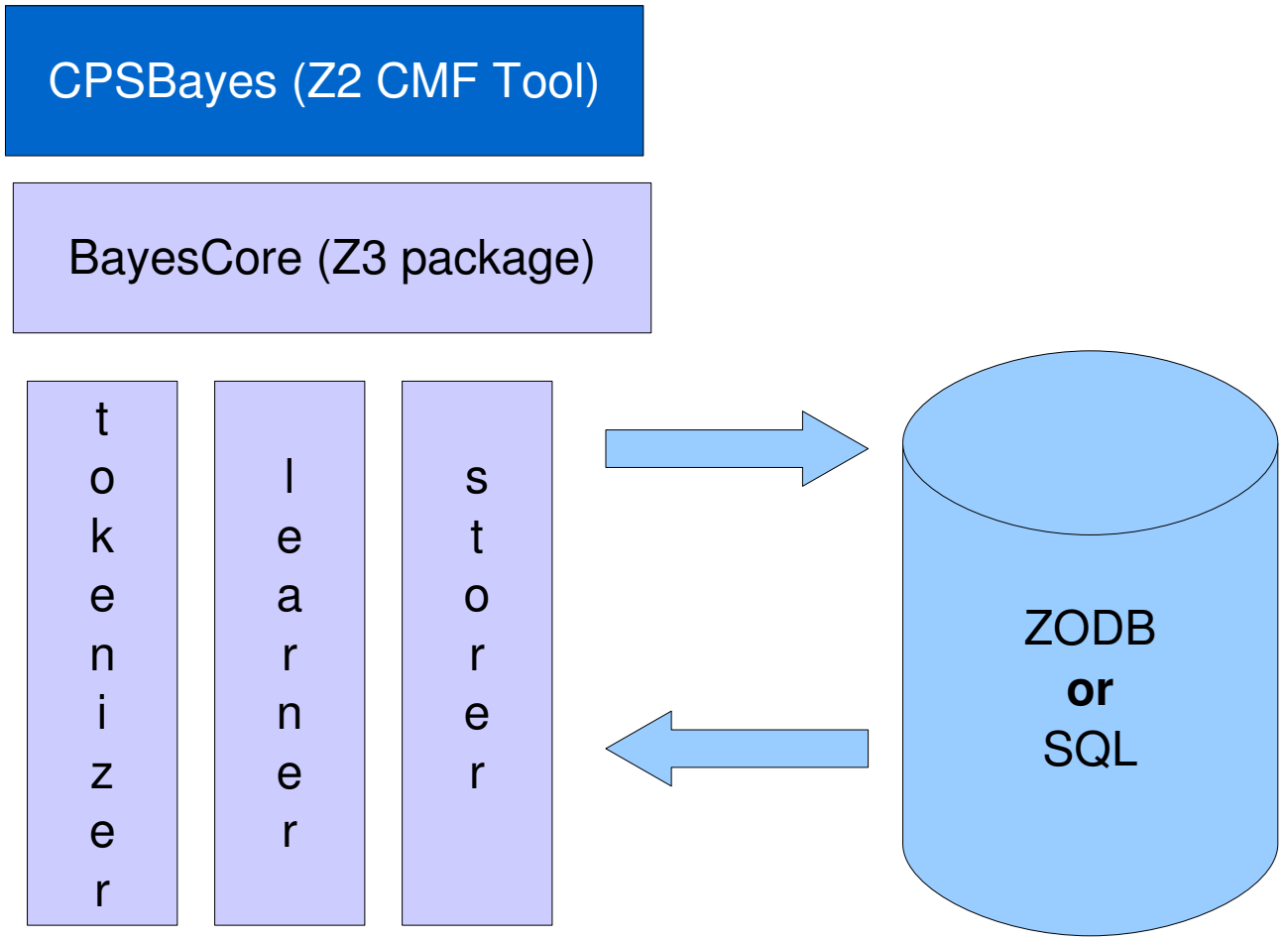
Mail processing



CPSCourier presentation

nuxeo

CPSBayes


CPSBayes (Z2 CMF Tool)
BayesCore (Z3 package)
tokenizer
learner
storer
ZODB
or
SQL

Other use cases

- Metadata automatic filling
- Automatic linking between documents
- ...

Demo: CPSBayesExample

Resources

- Nuxeo blogs: http://blogs.nuxeo.org
- CPS Project: http://cps-project.org
- Zope 3: http://zope.org
- CPSCourrier:

  http://www.nuxeo.com/solutions/cps-c

My pronostic:

ALLEMAGNE	2	mar 04/07 21h
ITALIE	1	
mar 04/07 - Dortmund (Westfalenstadion)		

Finale

ALLEMAGNE	0	dim 09/07 20h
FRANCE	1	
dim 09/07 - Berlin (Olympia Stadion)		

PORTUGAL	0	mer 05/07 21h
FRANCE	2	
mer 05/07 - Munich (Allianz-Arena)		

nuxeo