AIE Overview, Part 3

Linguistics and Text Analytics

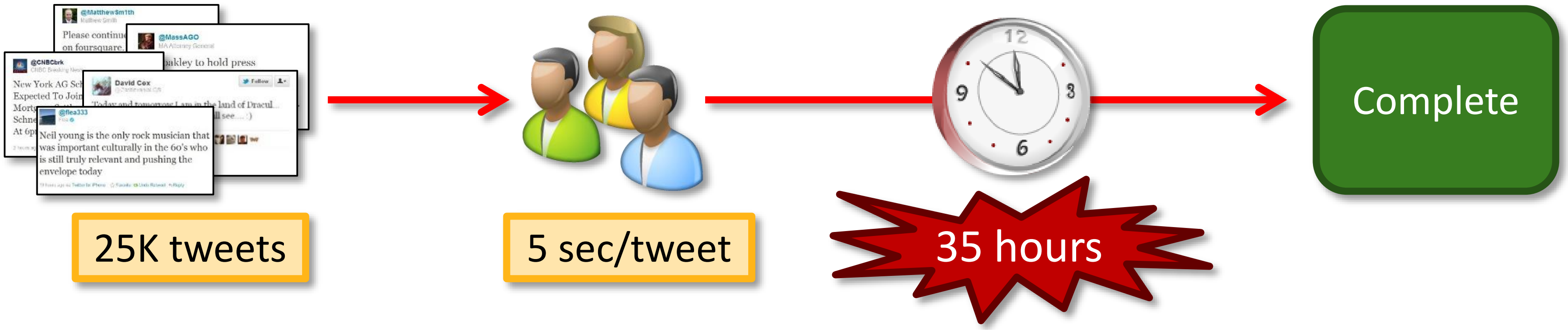
ATTIVIO

ACTIVE INTELLIGENCE

Big Data Problems

❗ Humans can't process Big Data volumes

- Example: Analyzing customer sentiment on Twitter


@MatthewSm1th
Please continu
on foursquare
@MassAGO
MA Attorney General
oakley to hold press
@CNBCbrk
New York AG Sch
Expected To Join
David Cox
Today and tomorrow I am in the land of Dracul...
@flea333
Neil young is the only rock musician that was important culturally in the 60's who is still truly relevant and pushing the envelope today
25K tweets
5 sec/tweet
35 hours
Complete

❗ Human language is too complex for a computer to understand

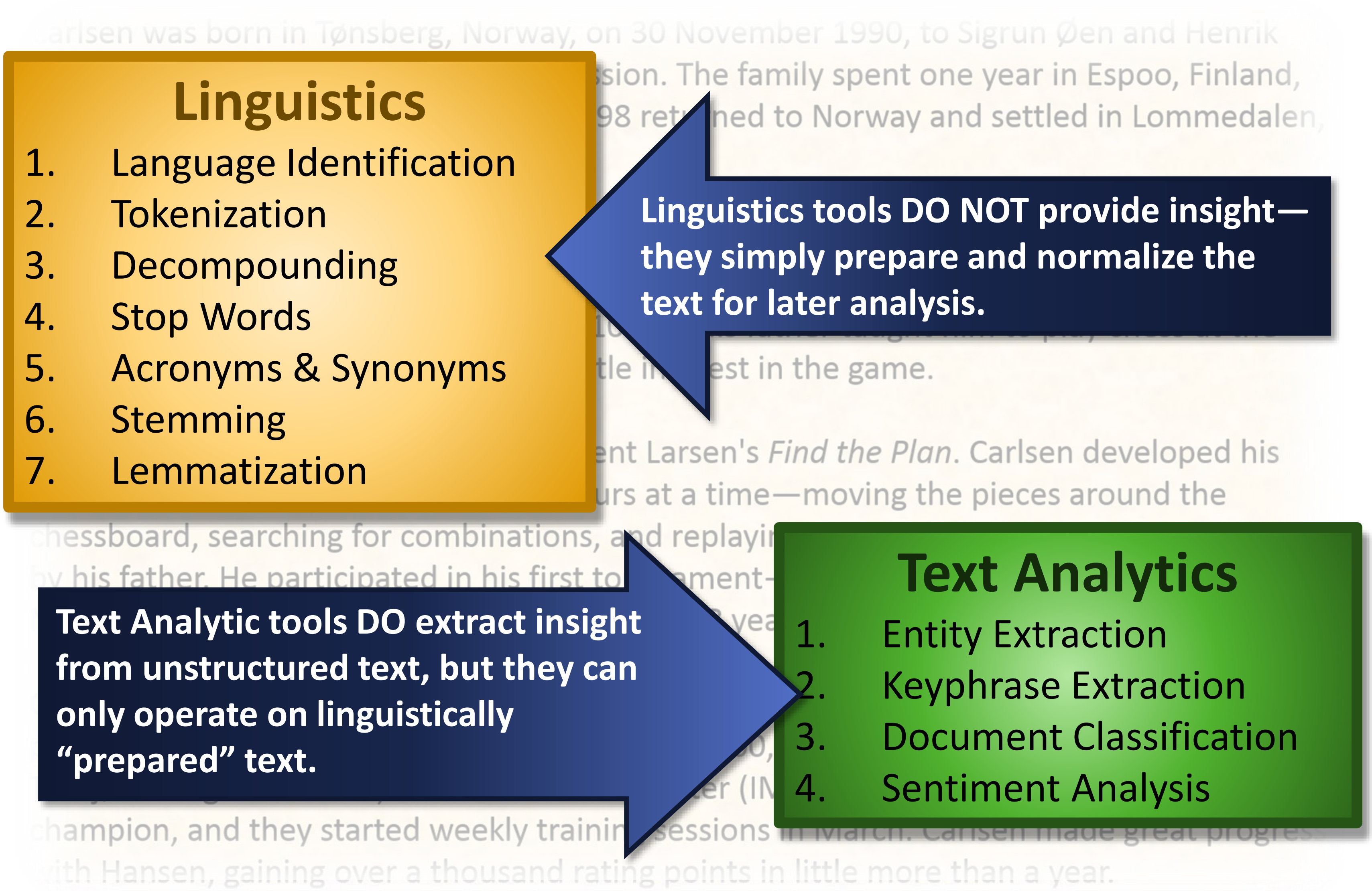
- Meaning depends on context and shared human experiences

Solution: Natural Language Processing

- **Natural Language Processing (NLP)**: the process of using a computer to extract information from human-created text
  - ✓ Allows a computer to perform useful work on unstructured data
  - ✓ Uses pattern recognition and algorithms in place of analyzing meaning
  - ✓ Scales well to Big Data volumes

- **Linguistics** and **Text Analytics** are layman terms for NLP

Linguistics vs. Text Analytics



Linguistics

1. Language Identification
2. Tokenization
3. Decompounding
4. Stop Words
5. Acronyms & Synonyms
6. Stemming
7. Lemmatization

Linguistics tools DO NOT provide insight—they simply prepare and normalize the text for later analysis.

Text Analytic tools DO extract insight from unstructured text, but they can only operate on linguistically “prepared” text.

Text Analytics

1. Entity Extraction
2. Keyphrase Extraction
3. Document Classification
4. Sentiment Analysis

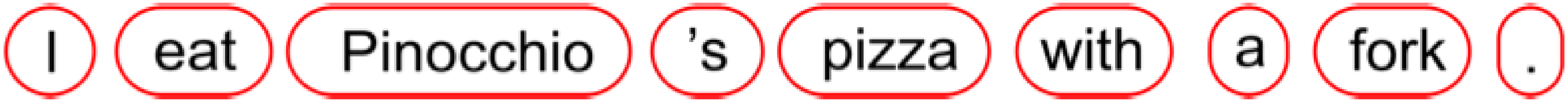
Linguistics #1: Language Identification

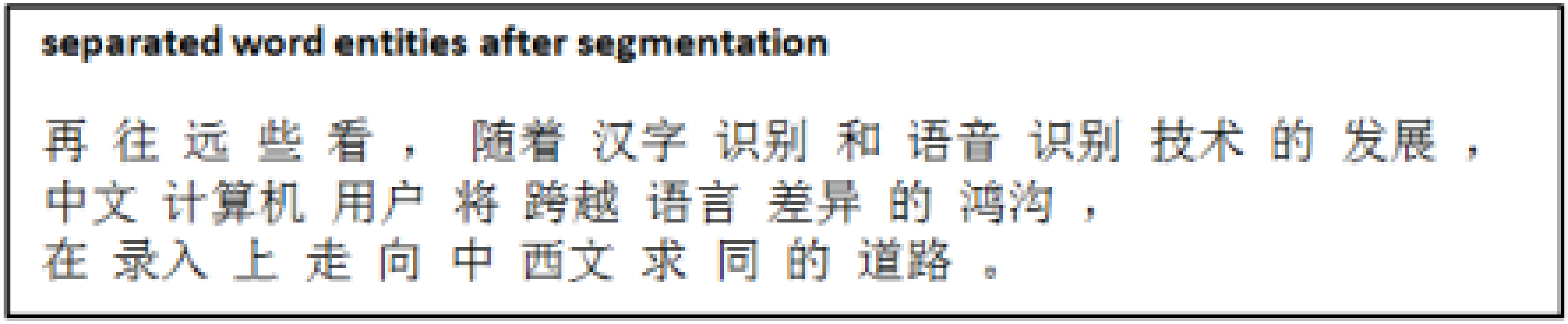
.id	doc_001
text	…
language	en
languages	en, fr

- Language ID always occurs first
  - All other text analytic tools are language specific

French proverbs are interesting. Although their literal translations may sound peculiar to Anglophones, many have an English equivalent. For example, “On n'apprend pas aux vieux singes à faire des grimaces” translates to “You cannot teach old monkeys to make faces,” which suggests the English proverb “You cannot teach old dogs new tricks.”

Linguistics #2: Tokenization





separated word entities after segmentation

再 往 远 些 看 ， 随着 汉字 识别 和 语音 识别 技术 的 发展 ，
中文 计算机 用户 将 跨越 语言 差异 的 鸿沟 ，
在 录入 上 走 向 中 西文 求 同 的 道路 。

- Token ≈ word
- Tokenization <u>always</u> occurs second
    - Most text analytic tools operate on tokens, not raw strings
- Tokenization rules are language specific

Linguistics #2.1: Decompounding

- **Decompounding**: Tokenization within a single complex word

Abwasserbehandlungsanlage

Abwasser + behandlungs + anlage

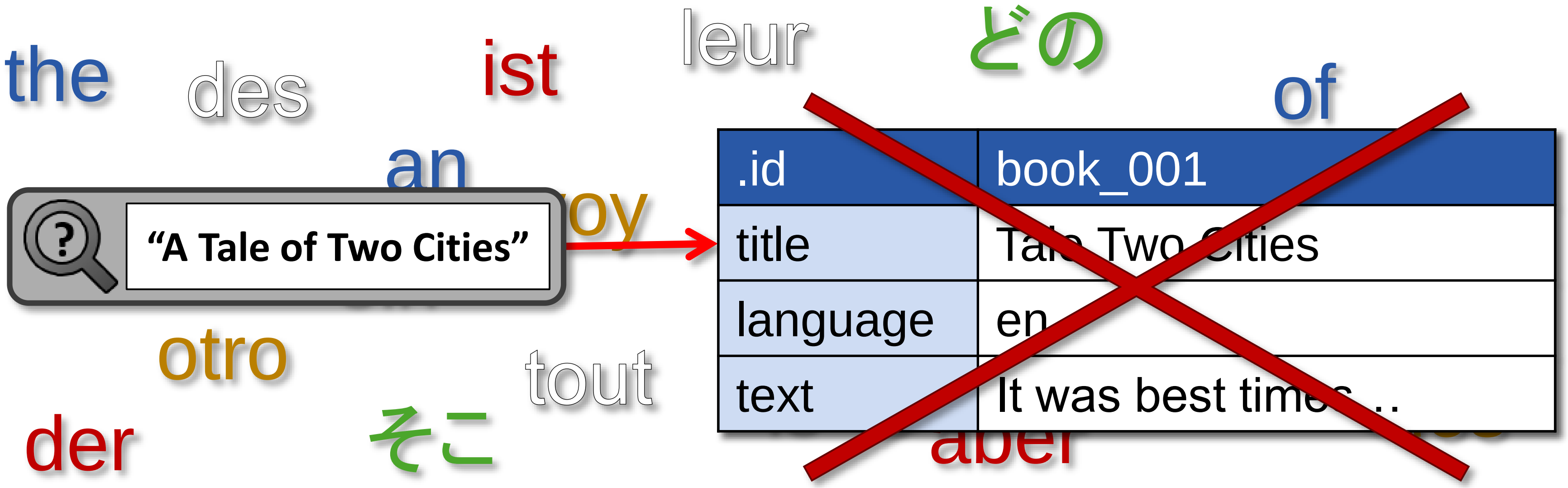
(wastewater treatment plant)

Jugendarbeitslosigkeit

Jugend + arbeits + losigkeit

(youth unemployment)

Linguistics: Stop Words



“A Tale of Two Cities”

.id	book_001
title	Tale Two Cities
language	en
text	It was best times …

- **Stop words**: common, less-meaningful words
- Dictionary-based
- Handled at ingestion and/or query time
  - ✓ Deleting at ingestion time makes the index smaller
  - ! But phrase searches become less accurate

Linguistics: Acronyms and Synonyms

ACRONYMS

BBC ≈ British Broadcasting Corporation

IBM ≈ International Business Machines

USA ≈ United States of America

SYNONYMS

student ≈ pupil

buy ≈ purchase

sick ≈ ill

quickly ≈ speedily

- When two terms have similar or identical meanings, a search for one term will find all related terms
- Dictionary-based
- Dictionaries are unidirectional

Linguistics: Stemming and Lemmatization

❗ Stemming

General
Generally
Generals
Generating
Generation
Generations
Generative

- By algorithm—on or off

✅ Lemmatization

Man

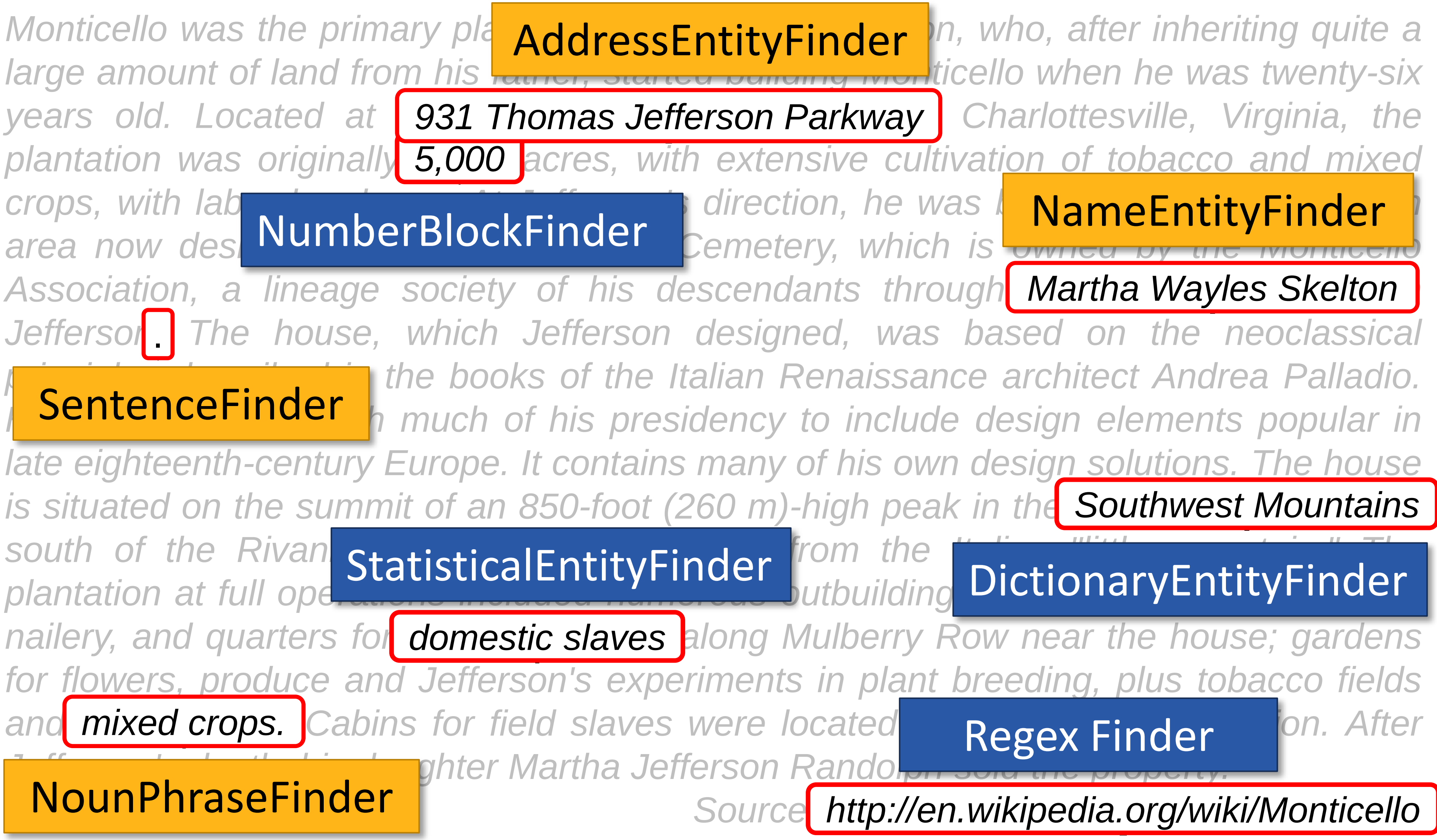
↕

Men
Manly
Manned
Manning

- Dictionary-based
- Concurrent with tokenization

Text Analytics: Entity Extraction

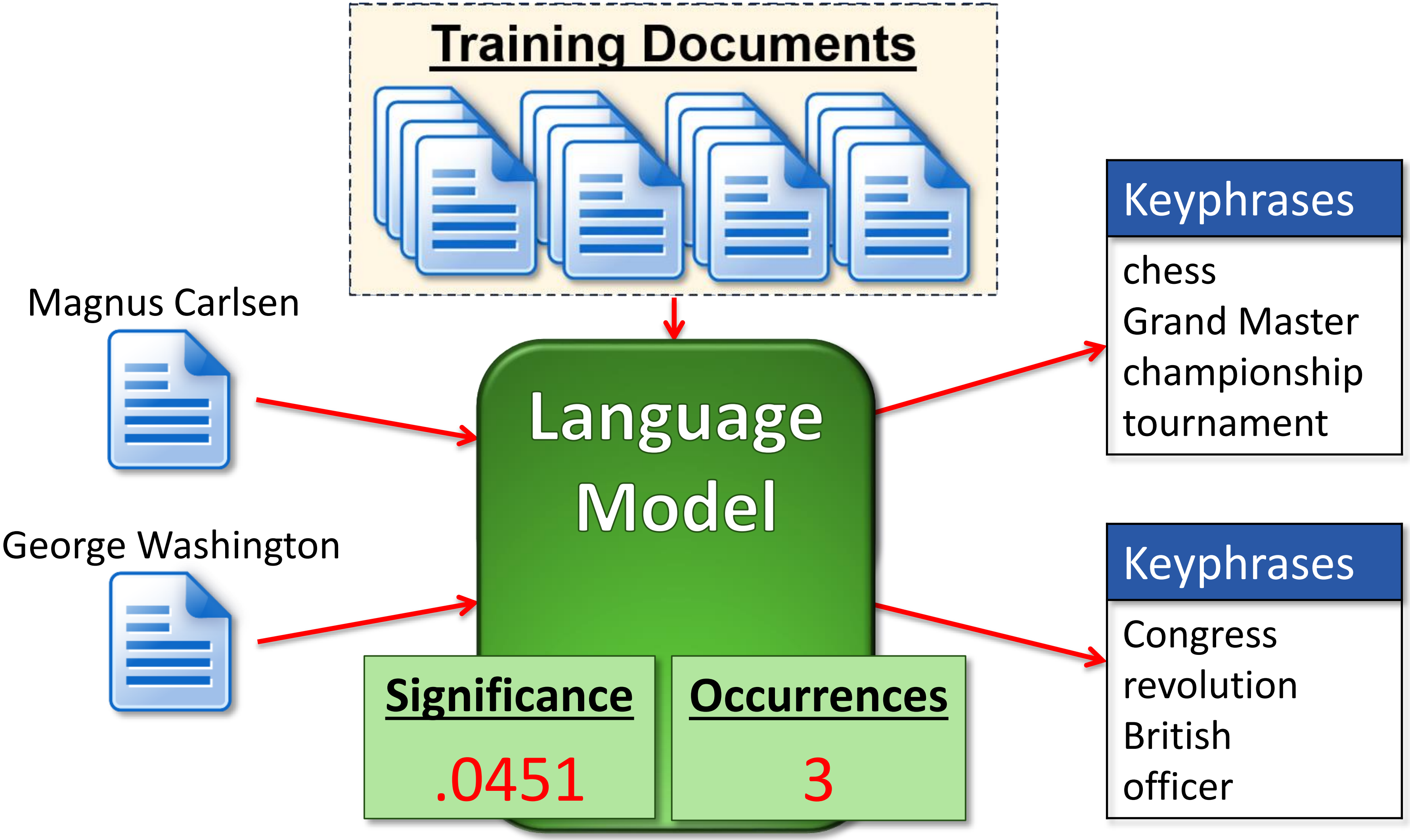
ATTIVIO
ACTIVE INTELLIGENCE



Monticello was the primary pla… n, who, after inheriting quite a large amount of land from his father, started building Monticello when he was twenty-six years old. Located at 931 Thomas Jefferson Parkway Charlottesville, Virginia, the plantation was originally 5,000 acres, with extensive cultivation of tobacco and mixed crops, with lab… At Jeff… direction, he was … area now des… Cemetery, which is owned by the Monticello Association, a lineage society of his descendants through Martha Wayles Skelton Jefferson. The house, which Jefferson designed, was based on the neoclassical p… the books of the Italian Renaissance architect Andrea Palladio. … much of his presidency to include design elements popular in late eighteenth-century Europe. It contains many of his own design solutions. The house is situated on the summit of an 850-foot (260 m)-high peak in the Southwest Mountains south of the Rivan… from the … plantation at full ope… outbuilding… nailery, and quarters for domestic slaves along Mulberry Row near the house; gardens for flowers, produce and Jefferson's experiments in plant breeding, plus tobacco fields and mixed crops. Cabins for field slaves were located … on. After … ghter Martha Jefferson Rando…

Source: http://en.wikipedia.org/wiki/Monticello

Text Analytics: Keyphrase Extraction



Training Documents

Magnus Carlsen

George Washington

Language Model

Significance

.0451

Occurrences

3

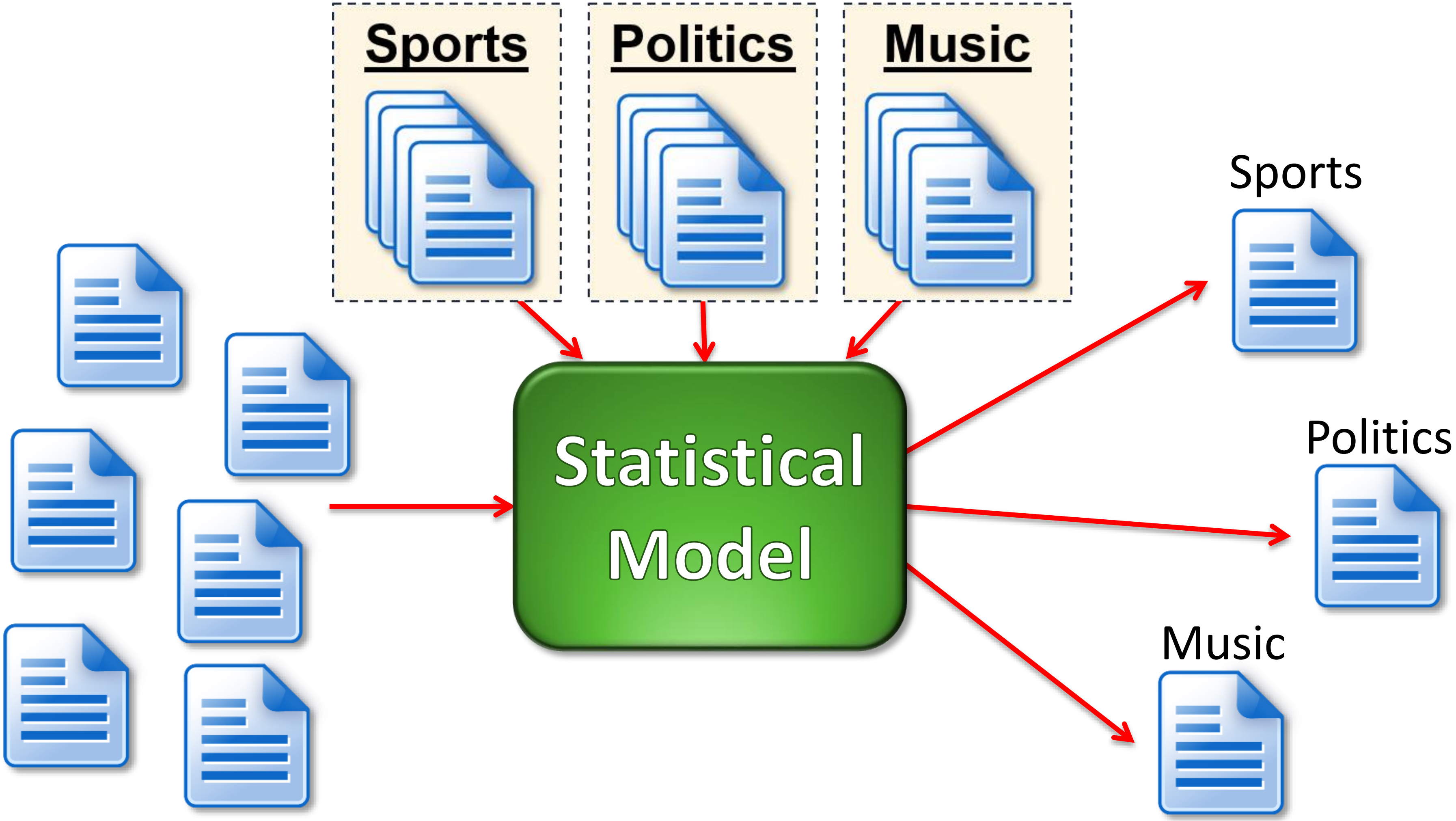
Keyphrases

chess
Grand Master
championship
tournament

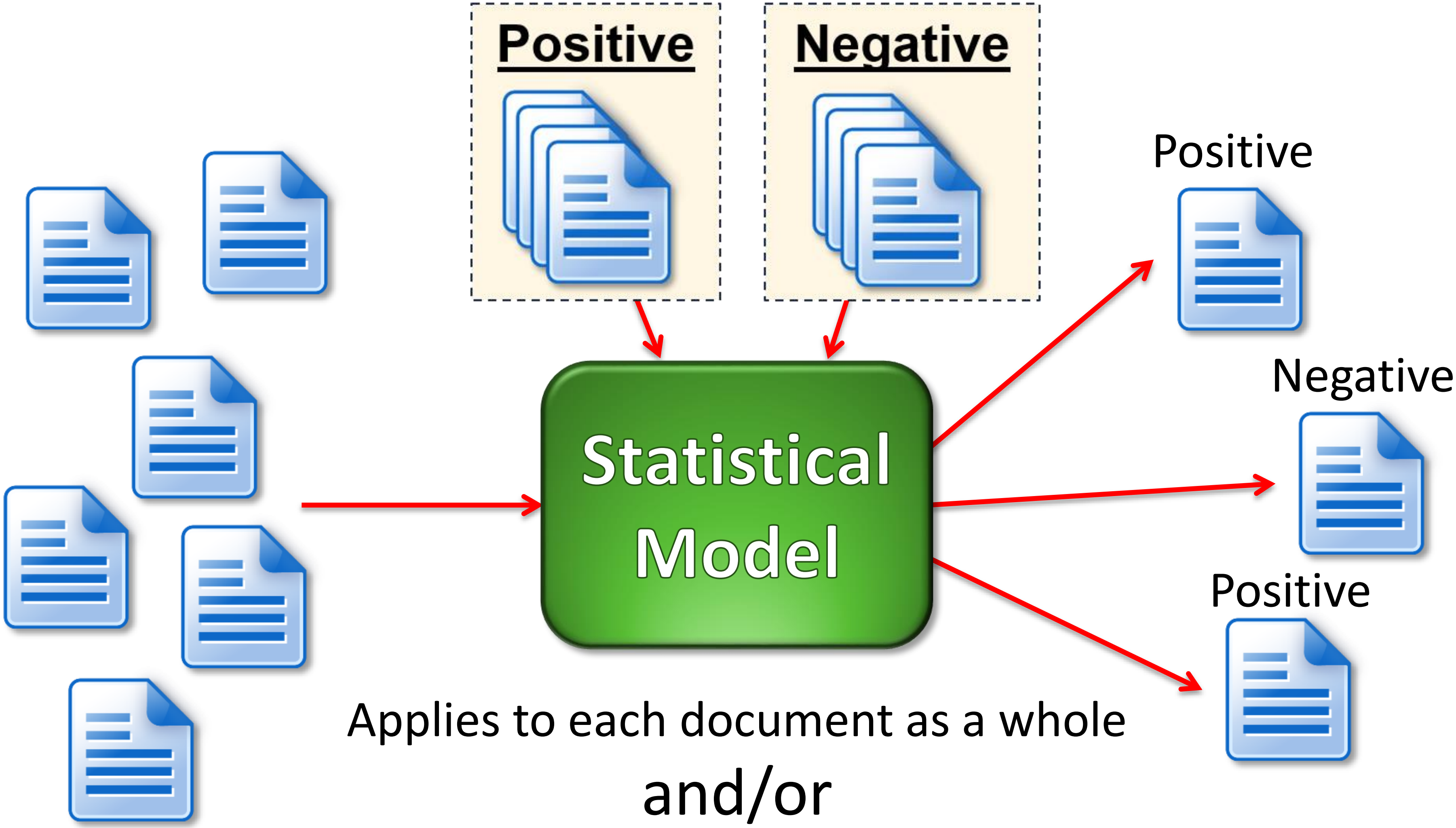
Keyphrases

Congress
revolution
British
officer

Text Analytics: Document Classification



Text Analytics: Sentiment Analysis



Applies to each document as a whole

and/or

to individual terms within each document