science-miner

Charles Napier Hemy,
The Fisherman 1888 (Wikimedia)

Patrice Lopez

October 29, 2017

WIKI DATA CON

entity-fishing

- Repo: https://github.com/kermit2/nerd
- Demo: http://entity-fishing.science-miner.com
- Doc: http://nerd.readthedocs.io
- Open source Apache 2 (including dependencies)
- Resources/models CC0

First World War excerpt

AUSTRIA invaded and fought the SERBIAN ARMY at the BATTLE OF CER and BATTLE OF KOLUBARA beginning on 12 AUGUST.

The army, led by general PAUL VON HINDENBURG defeated RUSSIA in a series of battles collectively known as the First BATTLE OF TANNENBERG (17 AUGUST – 2 SEPTEMBER). But the failed RUSSIAN invasion, causing the fresh GERMAN TROOPS to move to the east, allowed the tactical ALLIED victory at the FIRST BATTLE OF THE MARNE.

Unfortunately for the ALLIES, the pro-GERMAN KING CONSTANTINE I dismissed the pro-ALLIED GOVERNMENT of E. VENIZELOS before the ALLIED expeditionary force could arrive. Beginning in 1915, the ITALIANS under CADORNA mounted eleven offensives on the ISONZO front along the ISONZO RIVER, northeast of TRIESTE.

At the SIEGE OF MAUBEUGE about 40000 FRENCH soldiers surrendered, at the BATTLE OF GALICIA Russians took about 100-120000 AUSTRIAN captives, at the BRUSILOV OFFENSIVE about 325 000 to 417 000 GERMANS and AUSTRIANS surrendered to RUSSIANS, at the BATTLE OF TANNENBERG 92,000 RUSSIANS surrendered.

After marching through BELGIUM, LUXEMBOURG and the ARDENNES, the GERMAN ARMY advanced, in the latter half of AUGUST, into northern FRANCE where they met both the FRENCH ARMY, under JOSEPH JOFFRE, and the initial six divisions of the BRITISH EXPEDITIONARY FORCE, under SIR JOHN FRENCH. A series of engagements known as the BATTLE OF THE FRONTIERS ensued. Key battles included the BATTLE OF CHARLEROI and the BATTLE OF MONS. In the former battle the FRENCH 5TH ARMY was almost destroyed by the German 2nd and 3RD ARMIES and the latter delayed the GERMAN advance by a day. A general ALLIED retreat followed, resulting in more clashes such as the BATTLE OF LE CATEAU, the SIEGE OF MAUBEUGE and the BATTLE OF ST. QUENTIN (Guise).

The GERMAN ARMY came within 70 km (43 mi) of PARIS, but at the FIRST BATTLE OF THE MARNE (6–12 SEPTEMBER), FRENCH and BRITISH troops were able to force a GERMAN retreat by exploiting a gap which appeared between the 1ST and 2ND ARMIES, ending the GERMAN advance into FRANCE. The GERMAN ARMY retreated north of the AISNE RIVER and dug in there, establishing the beginnings of a static western front that was to last for the next three years. Following this GERMAN setback, the opposing forces tried to OUTFLANK each other in the RACE FOR THE SEA, and quickly extended their TRENCH SYSTEMS from the NORTH SEA to the SWISS frontier. The resulting GERMAN-OCCUPIED TERRITORY held 64% of FRANCE's pig-IRON PRODUCTION, 24% of its steel manufacturing, dealing a serious, but not crippling setback to FRENCH industry.

ALLIES

Normalized: **Allies of World War I**

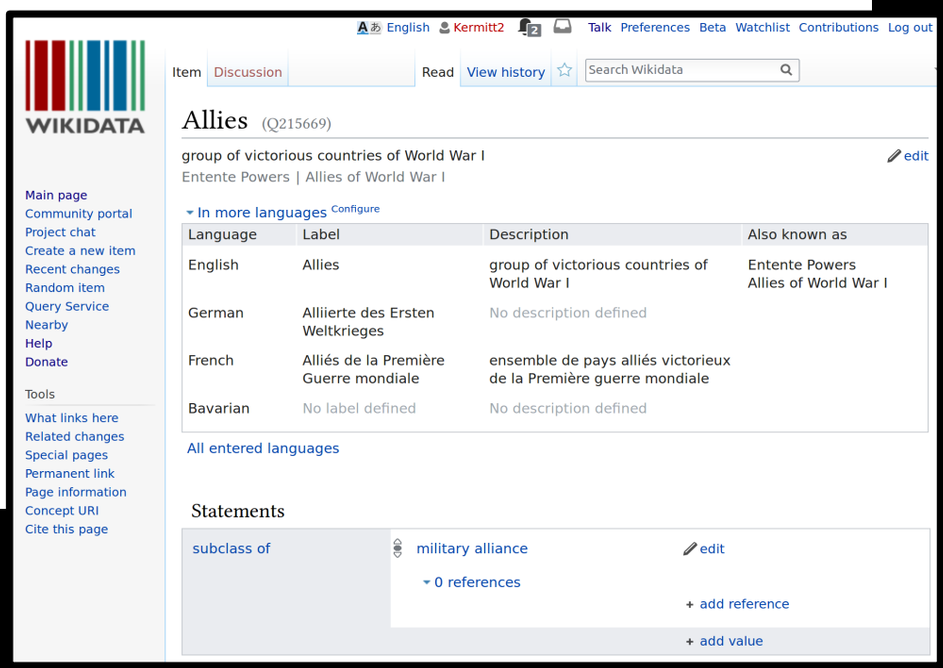
Domains: **Administration, Military**

conf: *0.7626*

The **Allies of World War I,** or **Entente Powers,** were the countries that opposed the Central Powers in the First World War.

subclass of	Military alliance
Quora topic ID	Allies-1

References:



science-miner

The challenge is to disambiguate mentions in context. For instance, "*allies*" refers most likely in the English Wikimedia to the Second World War allies entity, WW1 allies being only fourth with only ~6% prob.

Service to call: Term look-up

allies | en

Submit

Entities | Response

Number of ambiguous concepts: 32

Allies of World War II

Cond. prob.: 0.3558282208588957

The **Allies of World War II**, called the **United Nations** from the 1 January 1942 declaration, were the countries that together opposed the Axis powers during the Second World War (1939–1945). The Allies promoted the alliance as seeking to stop German, Japanese and Italian aggression.

Alliance

Cond. prob.: 0.20245398773006135

An **alliance** is a relationship among people, groups, or states that have joined together for mutual benefit or to achieve some common purpose, whether or not explicit agreement has been worked out among them. Members of an alliance are called **allies**. Alliances form in many settings, including political alliance, military alliance, and business alliance. When the term is used in the context of war or armed struggle, such associations may also be called **allied powers**, especially when discussing World War I or World War II.A formal military alliance is not required for being perceived as an ally—co-belligerence, fighting alongside someone, is enough. According to this usage, allies become so not when concluding an alliance treaty but when struck by war.

Straight ally

Cond. prob.: 0.09202453987730061

A **straight ally** or **heterosexual ally** is a heterosexual and cisgender person who supports equal civil rights, gender equality, LGBT social movements, and challenges homophobia, biphobia and transphobia. Despite this, some people who meet this definition do not identify themselves as straight allies. A straight ally believes that LGBT people face discrimination and thus are socially and economically disadvantaged. They aim to use their position as heterosexual or cisgender individuals in a society focused on heteronormativity to fight homophobia, biphobia and transphobia.Most LGBT organizations have straight members involved; others actively encourage straight participation. A gay–straight alliance is a student-run club that brings together LGBT and straight students to create a platform for activism to fight homophobia and transphobia. There are also some groups that unite the LGBT community to work together with straight allies. Founded in 1973, Parents, Families and Friends of Lesbians and Gays (PFLAG) is the original straight ally organization, started by Jeanne Manford, mother of the Straight Ally movement. Based in the United States, PFLAG unites parents, families, friends, and straight allies with the LGBT community to move equality forward for LGBT people. In 2007, the organization launched a new project, Straight for Equality to help more straight allies become engaged in the LGBT movement in the workplace, healthcare, and now in faith communities. Gay & Lesbian Advocates & Defenders (GLAD) is another organization specifically formed to group allies of this cause.

Allies of World War I

Cond. prob.: 0.06134969325153374

The **Allies of World War I,** or **Entente Powers,** were the countries that opposed the Central Powers in the First World War.

science-miner

Catching Wikidata entities in scholar PDF article

Morphological variation in hybrids between *Salmo marmoratus* and alien *Salmo* species in the Volarja stream, Soca River basin, Slovenia

B. DELLING*||, A. J. CRIVELLI†, J-F. RUBIN‡ AND P. BERREBI¶

*Department of Vertebrate Zoology, Swedish Museum of Natural History, P.O. Box 50007, SE-104 05 Stockholm and Department of Zoology, Stockholm University, SE-106 91 Stockholm, Sweden; †Station Biologique de la Tour du Valat, Le Sambuc, F-13200 Arles, France; ‡Institut d'Ecologie, Lausanne University, CH-1000 Dorigny, Switzerland; ¶Laboratoire Génome et Populations, Université Montpellier II, CC063, Place Bataillon, F-34095 Montpellier Cédex 05, France

(*Received 11 February 2000, Accepted 16 June 2000*)

There were significant correlations between colour pattern, *LDH-5** genotype and certain meristic characters in 59 hybrid trout *Salmo* sp. from the Volarja stream, Soca River basin, Slovenia. It is concluded that panmixia between native *Salmo marmoratus* and introduced *S. trutta* of Atlantic, Danubian and Mediterranean origin had not been reached in this zone, despite the long period of introgression. The result is in agreement with other studies dealing with introgression in *Salmo*, and for management purposes certain morphological characters, especially colour pattern, can be a valuable tool in restoring the marble trout population in the Soca River.

© 2000 The Fisheries Society of the British Isles

Key words: introgression; hybridization; allozymes; morphometrics; *Salmo marmoratus*.

INTRODUCTION

Hybridization and introgression among fish species is well documented (Verspoor & Hammar, 1991; Leary *et al.*, 1995). Among the Eurasian *Salmo* species, hybrids between Atlantic salmon *Salmo salar* L. and brown trout *S. trutta* L. are reported frequently from the entire range of sympatry of the two

SALMO MARMORATUS

Normalized: **Salmo marmoratus**

Domains: **Animals**

conf: *0.7247*

Salmo marmoratus (**marble trout**) is a species of freshwater fish in the Salmonidae family. It is characterized by a distinctive marbled color pattern and high growth capacity. The marble trout is found in only three basins and two rivers of the Adriatic basin, namely the Po with only northern/left tributaries and the Adige, Brenta, Piave, Tagliamento and Livenza basins in Italy, the Soča basin in Slovenia and Italy, the Neretva river in Bosnia and Herzegovina and Croatia, and the Morača river in Montenegro. While once present in the Drin river basin in Albania fish is almost certainly extirpated there.

taxon rank	Subspecies
taxon name	Salmo trutta marmoratus
Commons category	Salmo marmoratus
parent taxon	Brown trout
Freebase ID	/m/02wygvg
instance of	Taxon
image	Salmo marmoratus.jpg
Encyclopedia of Life ID	1157691
taxon synonym	Q22231158

References: W

Search query disambiguation for “*concrete pump sensor*” (response time 5-10ms)

concrete

Conf: 0.36

Concrete is a composite material composed of coarse aggregate bonded together with a fluid cement that hardens over time. Most concretes used are lime-based concretes such as Portland cement concrete or concretes made with other hydraulic cements, such as ciment fondu. However, asphalt concrete, which is frequently used for road surfaces, is also a type of concrete, where the cement material is bitumen, and polymer concretes are sometimes used where the cementing material is a polymer.

concrete pump

Conf: 0.72

A **concrete pump** is a machine used for transferring liquid concrete by pumping. There are two types of concrete pumps. The first type of concrete pump is attached to a truck or longer units are on semi-trailers. It is known as a boom concrete pump because it uses a remote-controlled articulating robotic arm (called a *boom*) to place concrete accurately. Boom pumps are used on most of the larger construction projects as they are capable of pumping at very high volumes and because of the labour saving nature of the placing boom. They are a revolutionary alternative to line-concrete pumps.

pump

Conf: 0.36

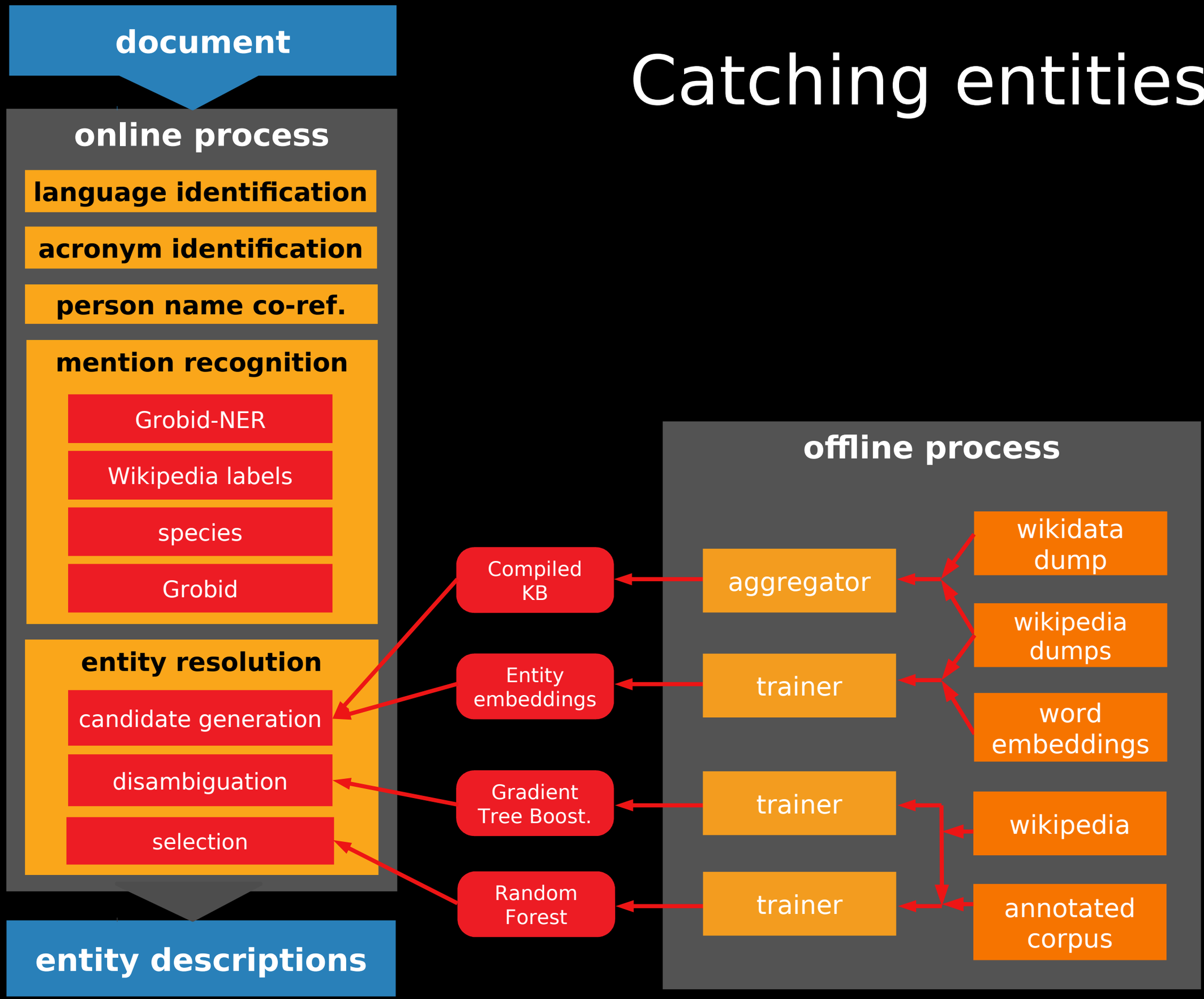
A **pump** is a device that moves fluids (liquids or gases), or sometimes slurries, by mechanical action. Pumps can be classified into three major groups according to the method they use to move the fluid: *direct lift*, *displacement*, and *gravity* pumps.

sensor

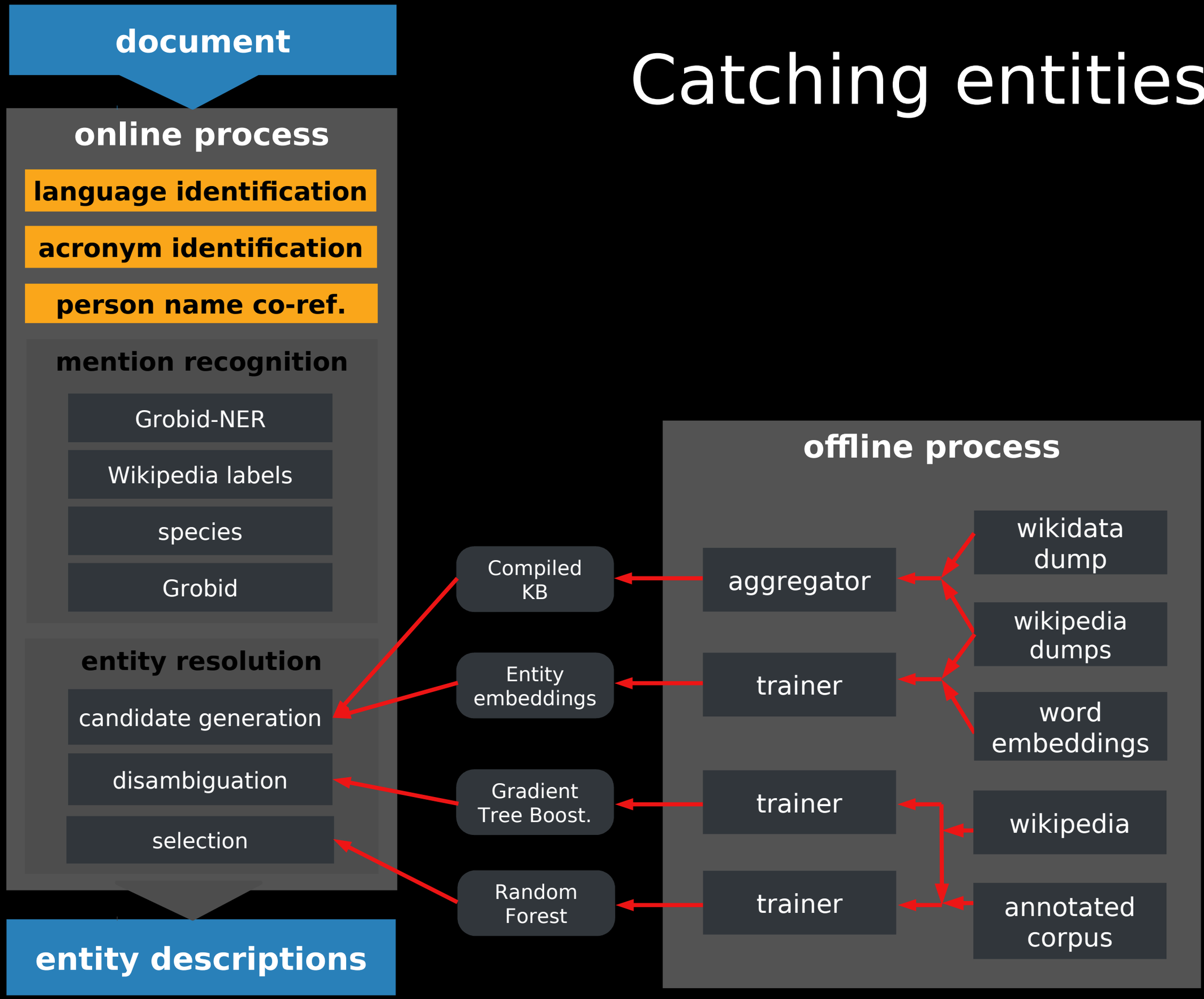
Conf: 0.72

In the broadest definition, a **sensor** is an electronic component, module, or subsystem whose purpose is to detect events or changes in its environment and send the information to other electronics, frequently a computer processor. A sensor is always used with other electronics, whether as simple as a light or as complex as a computer.

Catching entities

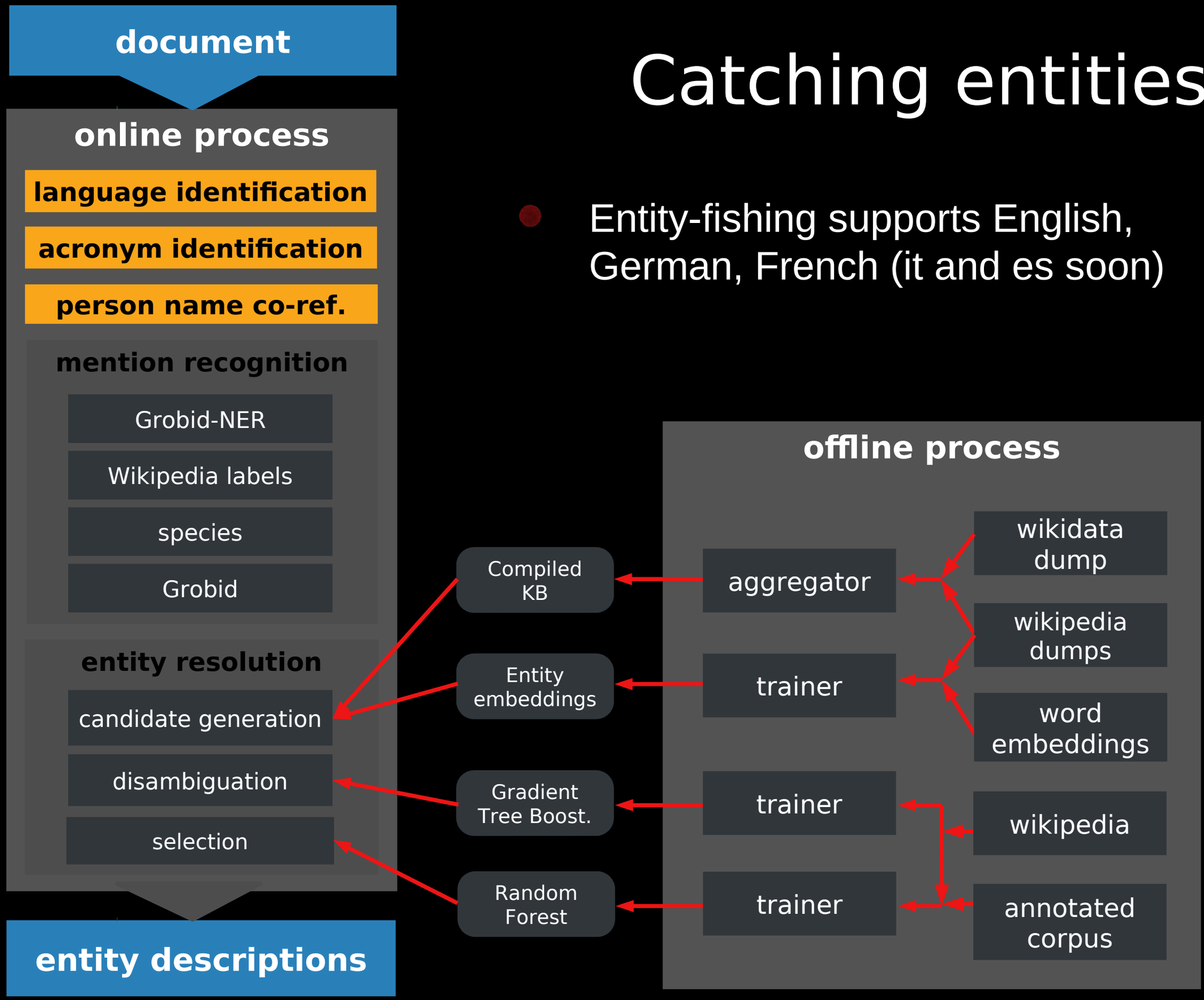


Catching entities

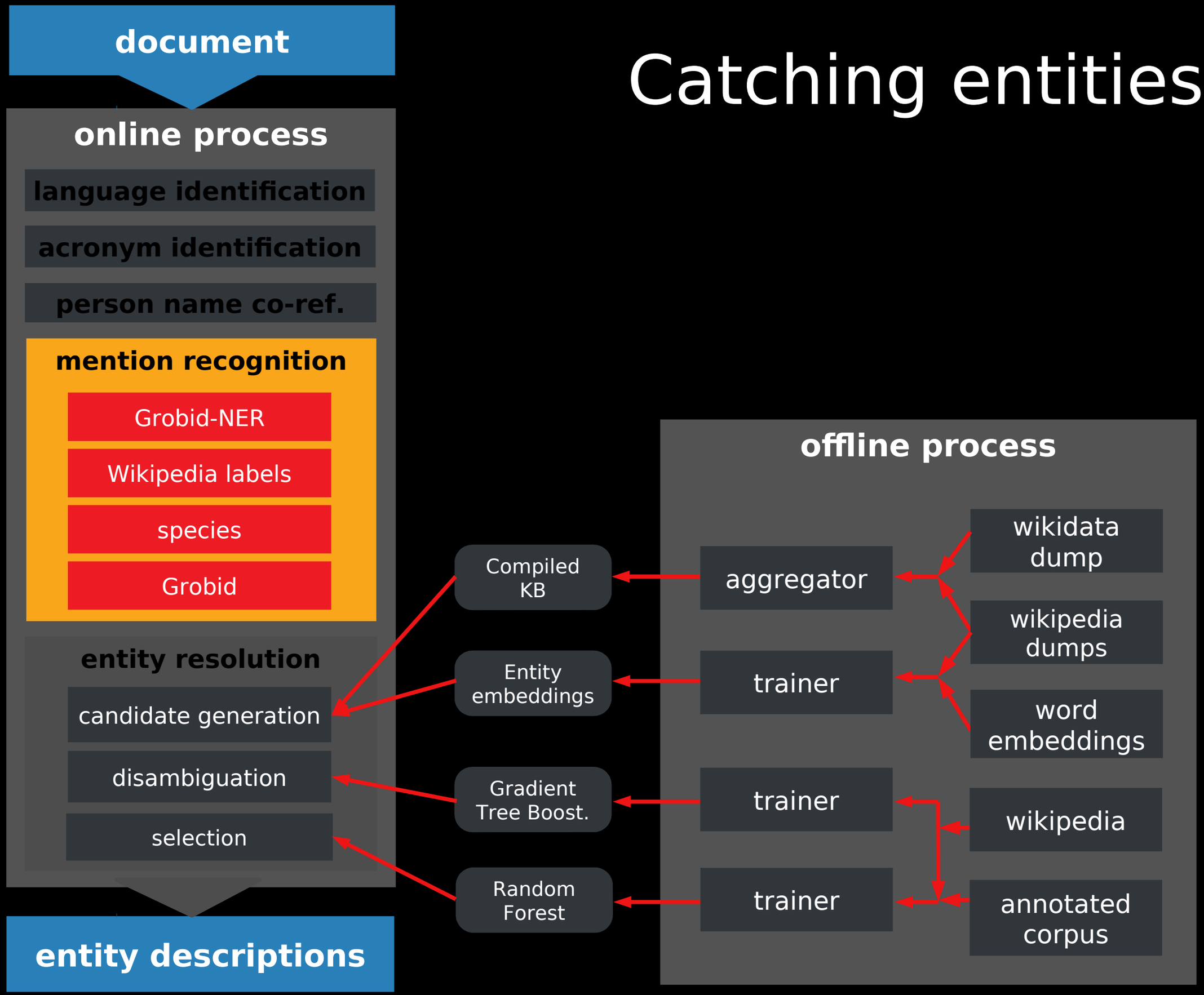


Catching entities

- Entity-fishing supports English, German, French (it and es soon)



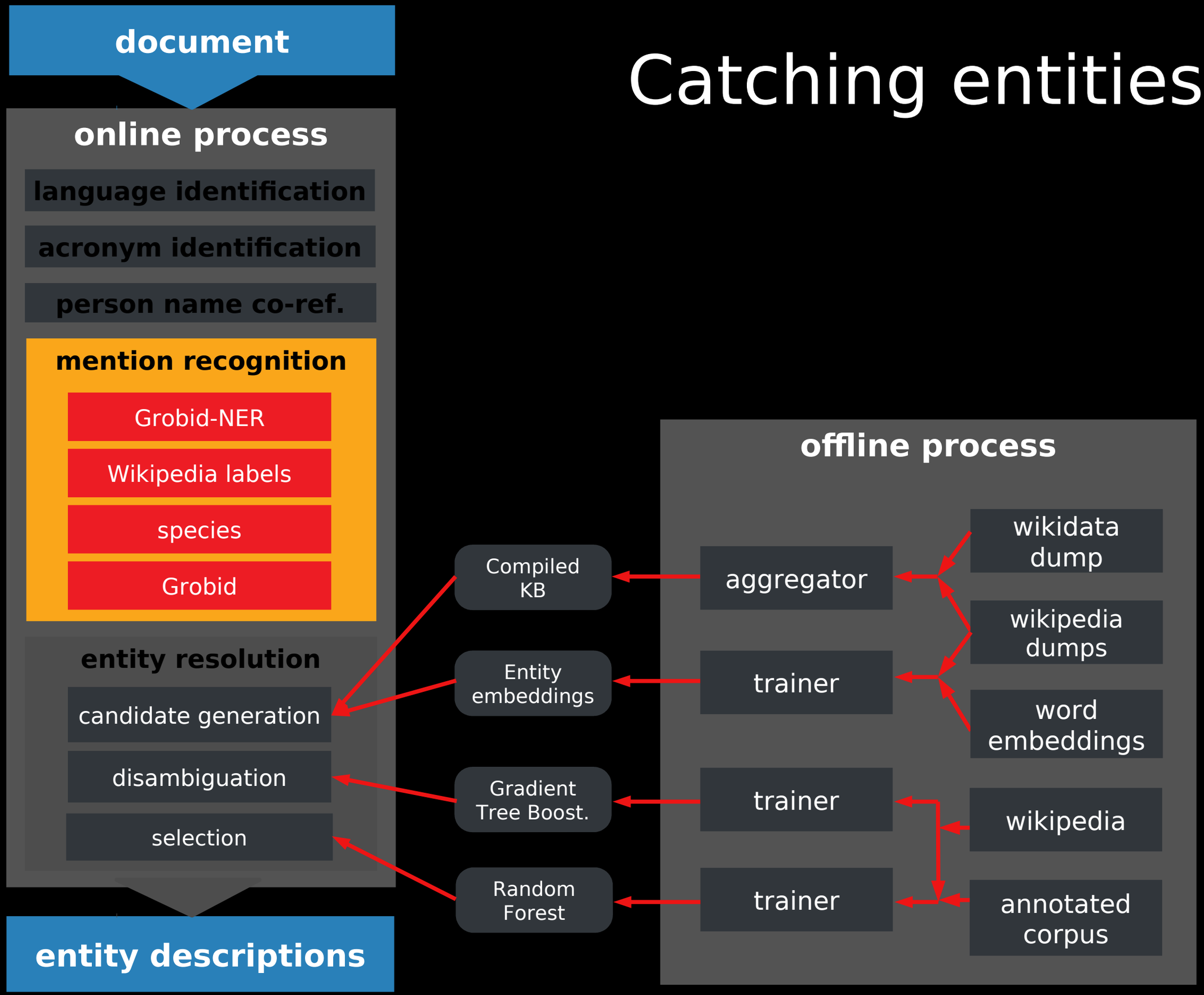
Catching entities



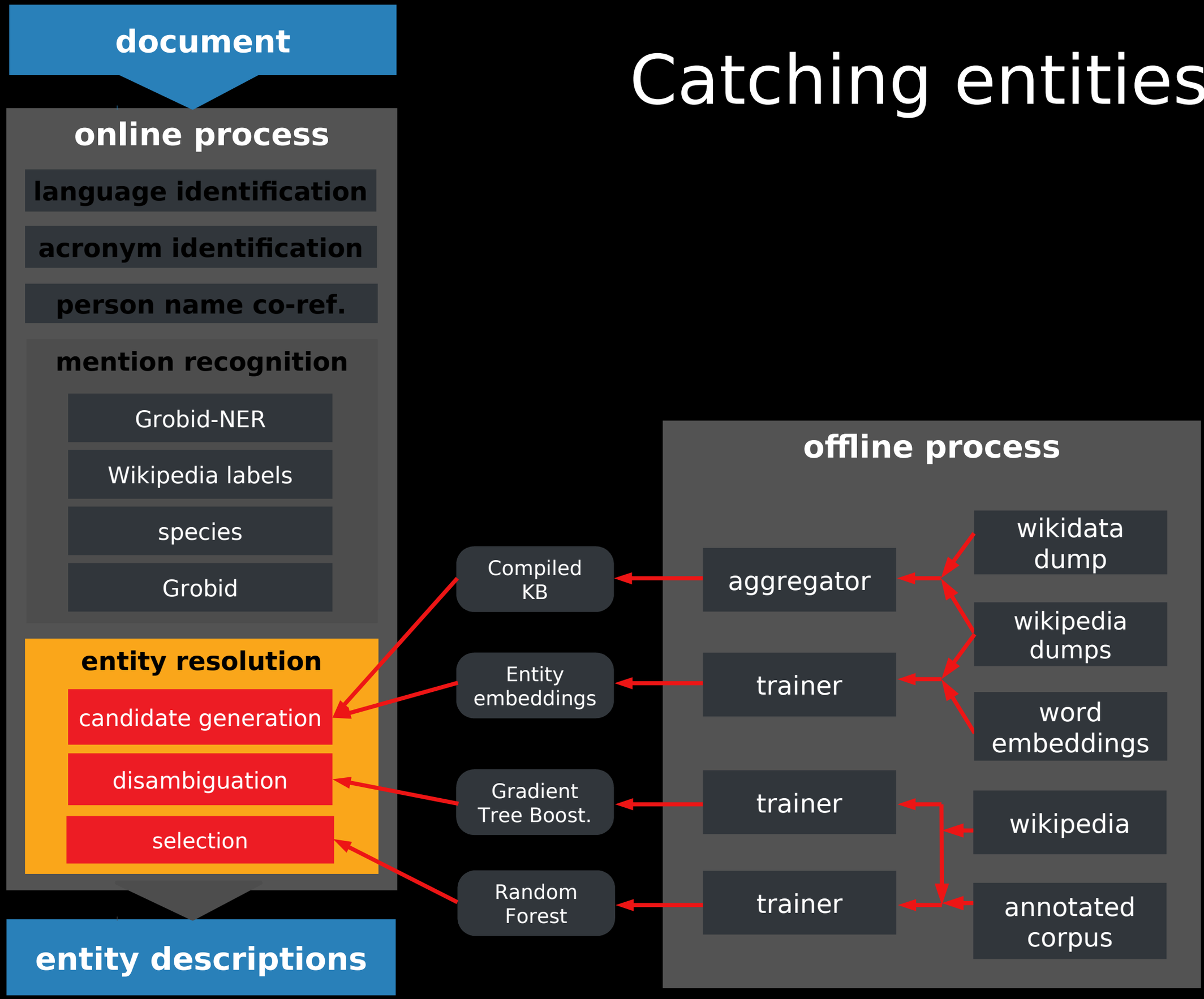
Mention detection

- A **mention** is a text string that can refer to an entity
- Traditional mentions are identified by
  - ➡ A Named-Entity Recognizer, for names, locations, organisations, etc.
  - ➡ Wikipedia titles and anchors
- But **Wikidata** entities are much more heterogeneous than in usual NERD, for example:
  - ➡ Many scientific entities, e.g. chemical formula, name of species, astronomical objects, etc.
  - ➡ Bibliographical objects

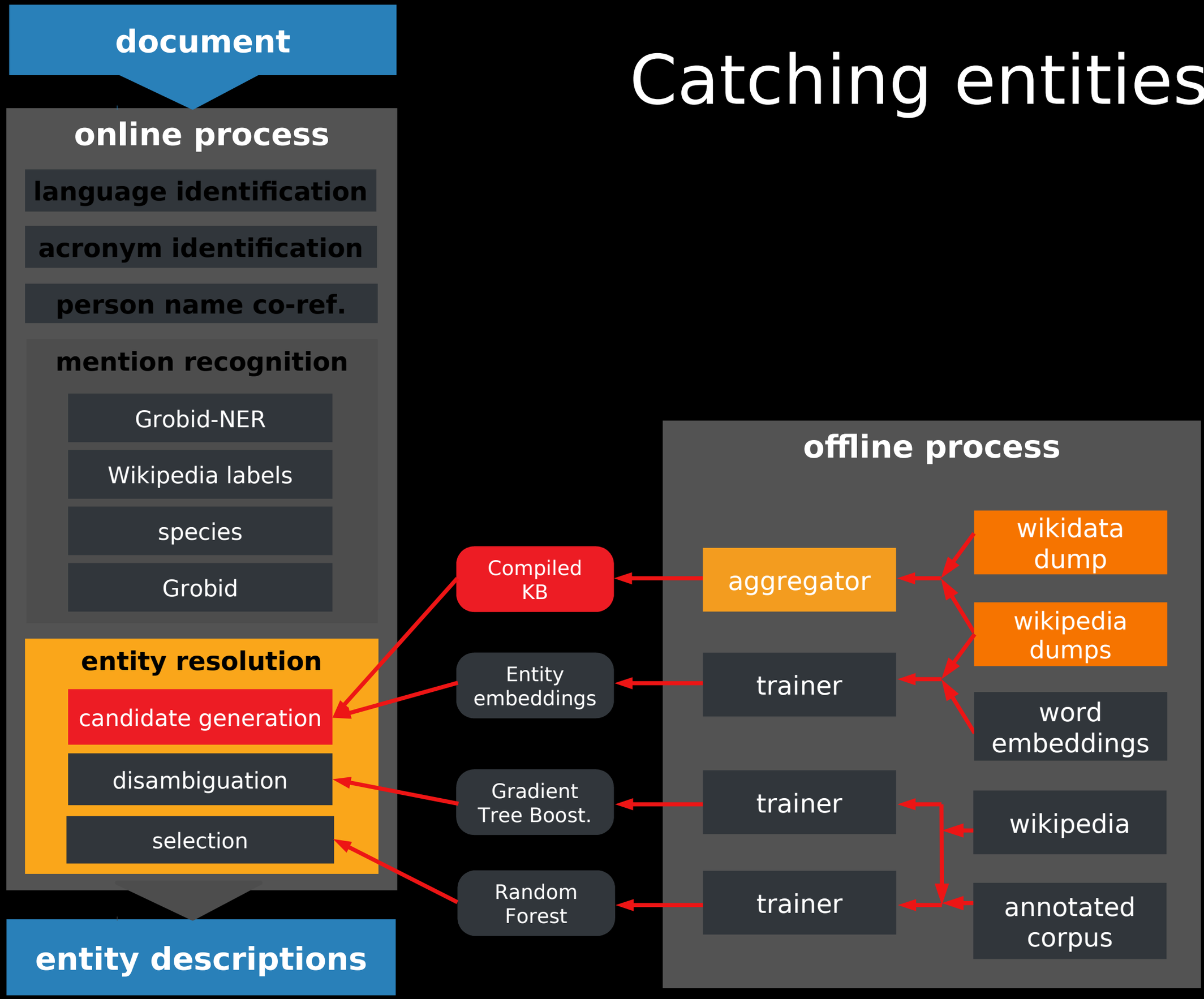
Catching entities



Catching entities

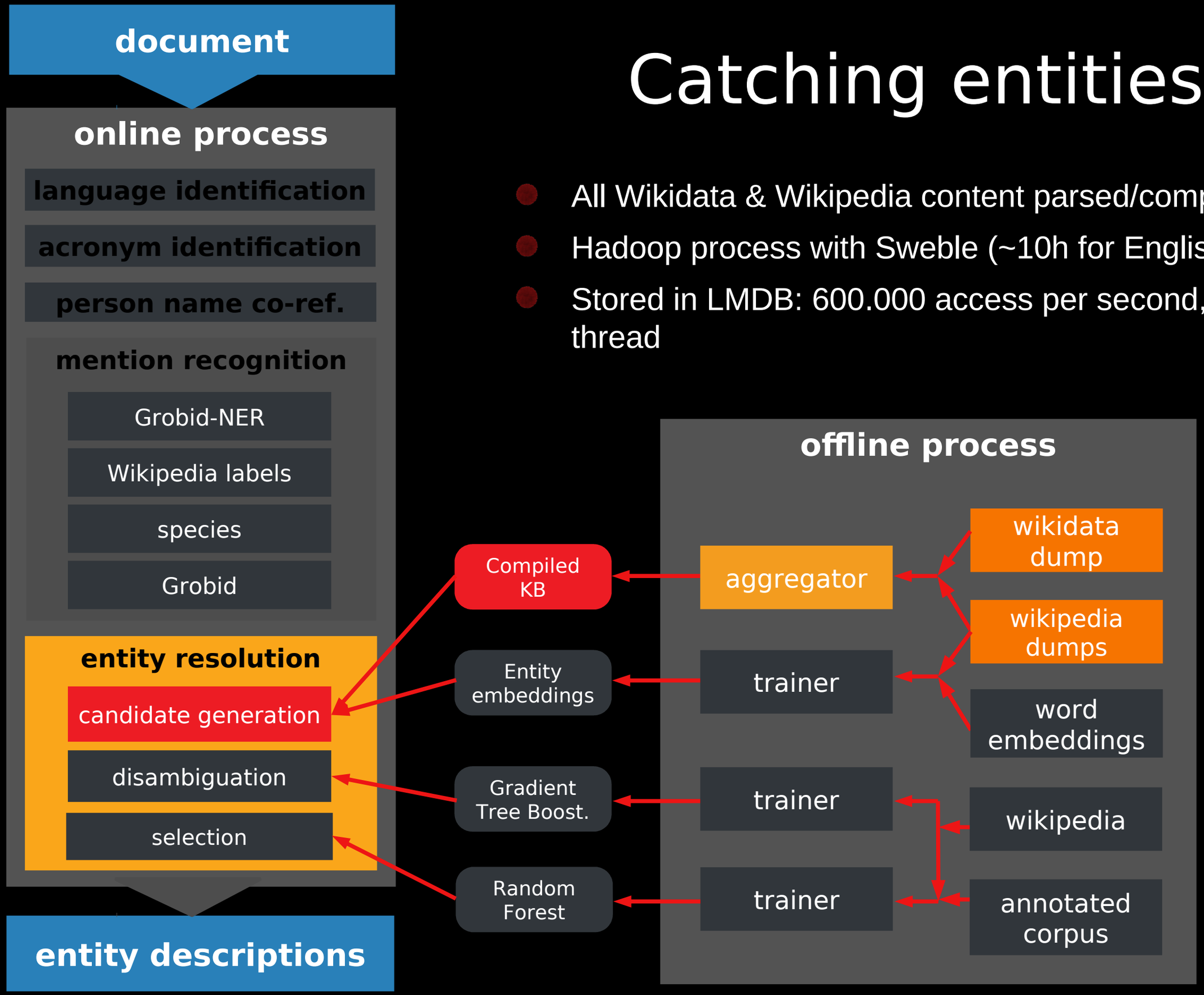


Catching entities

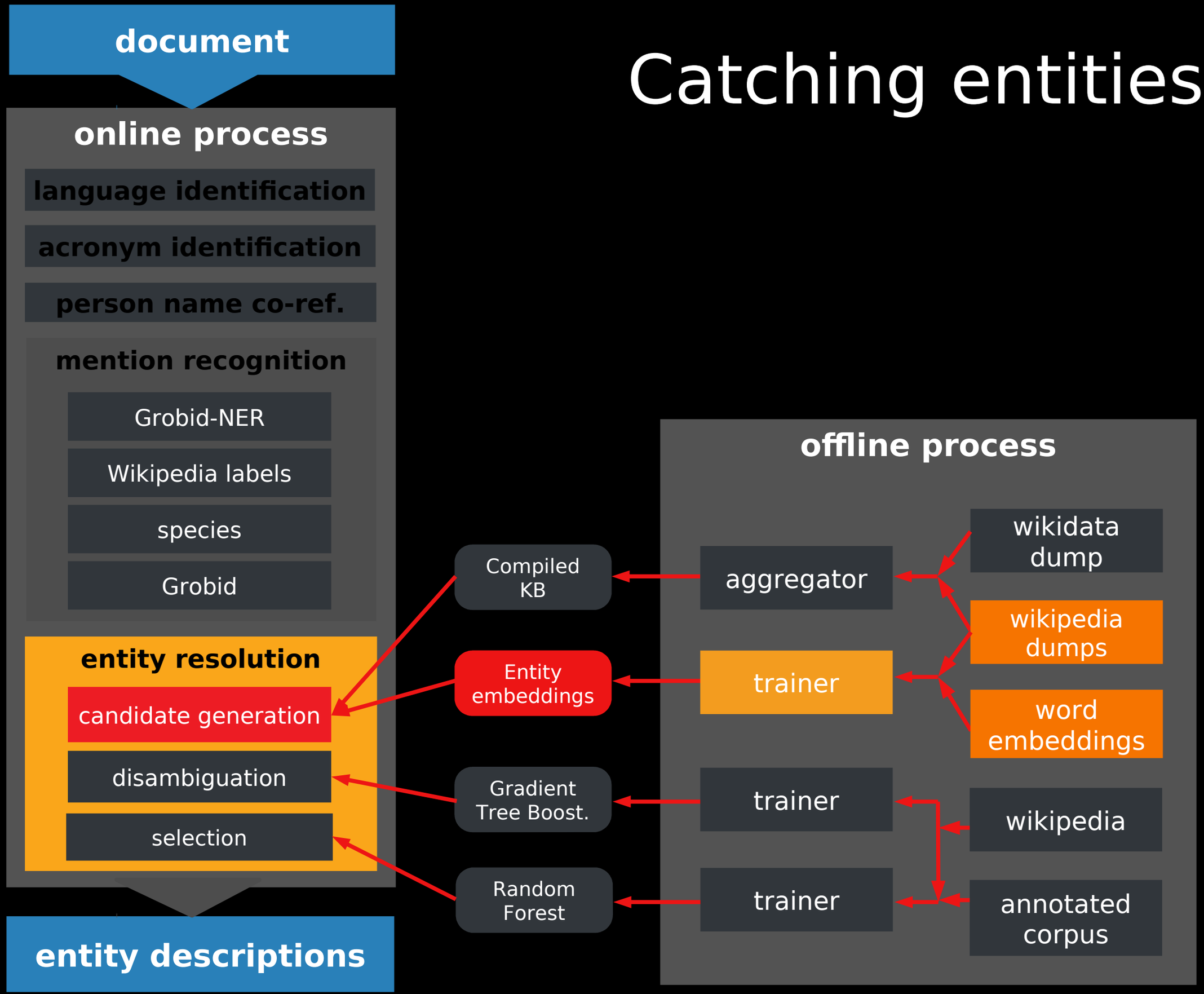


Catching entities

- All Wikidata & Wikipedia content parsed/compiled
- Hadoop process with Sweble (~10h for English)
- Stored in LMDB: 600.000 access per second, per thread

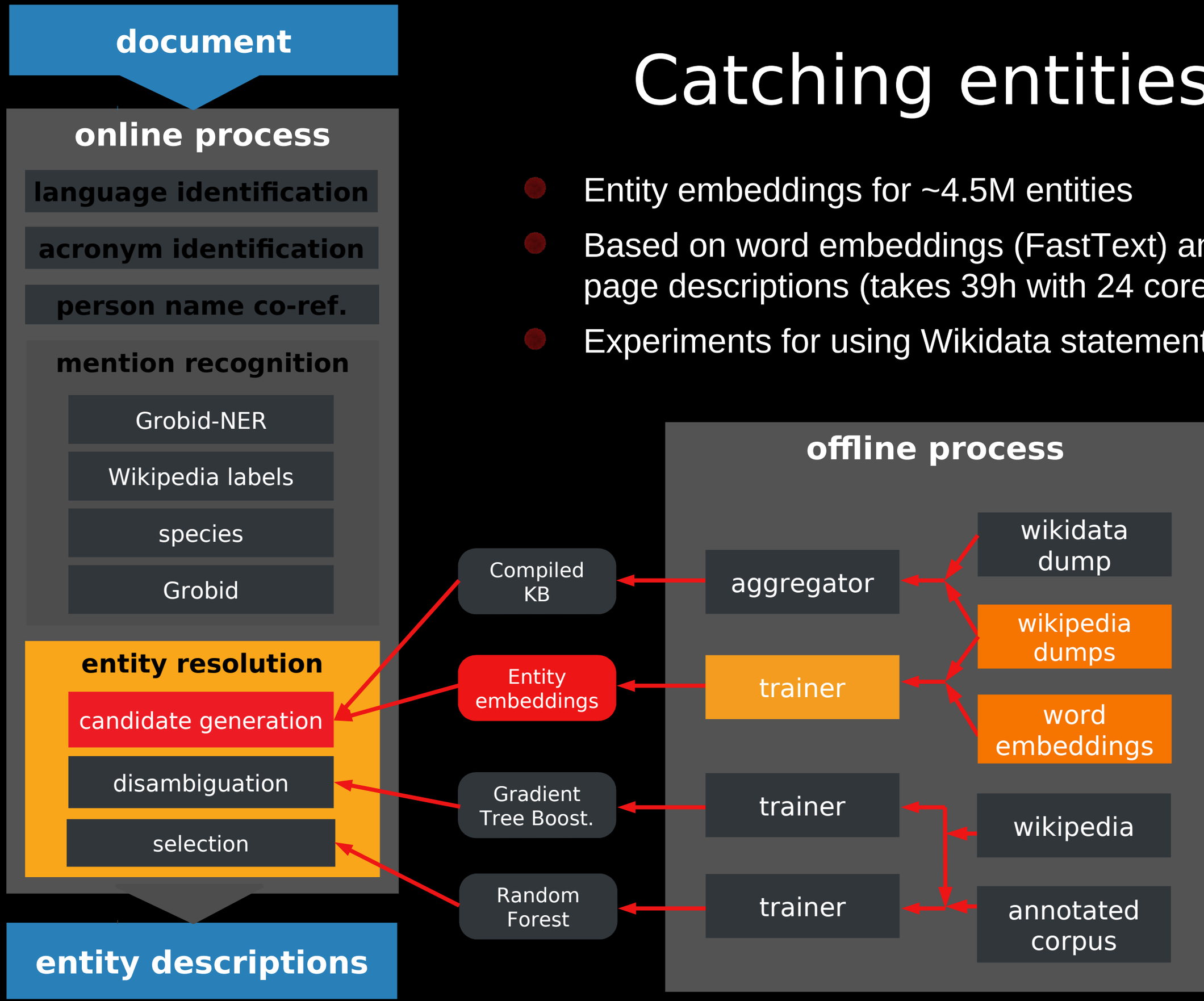


Catching entities

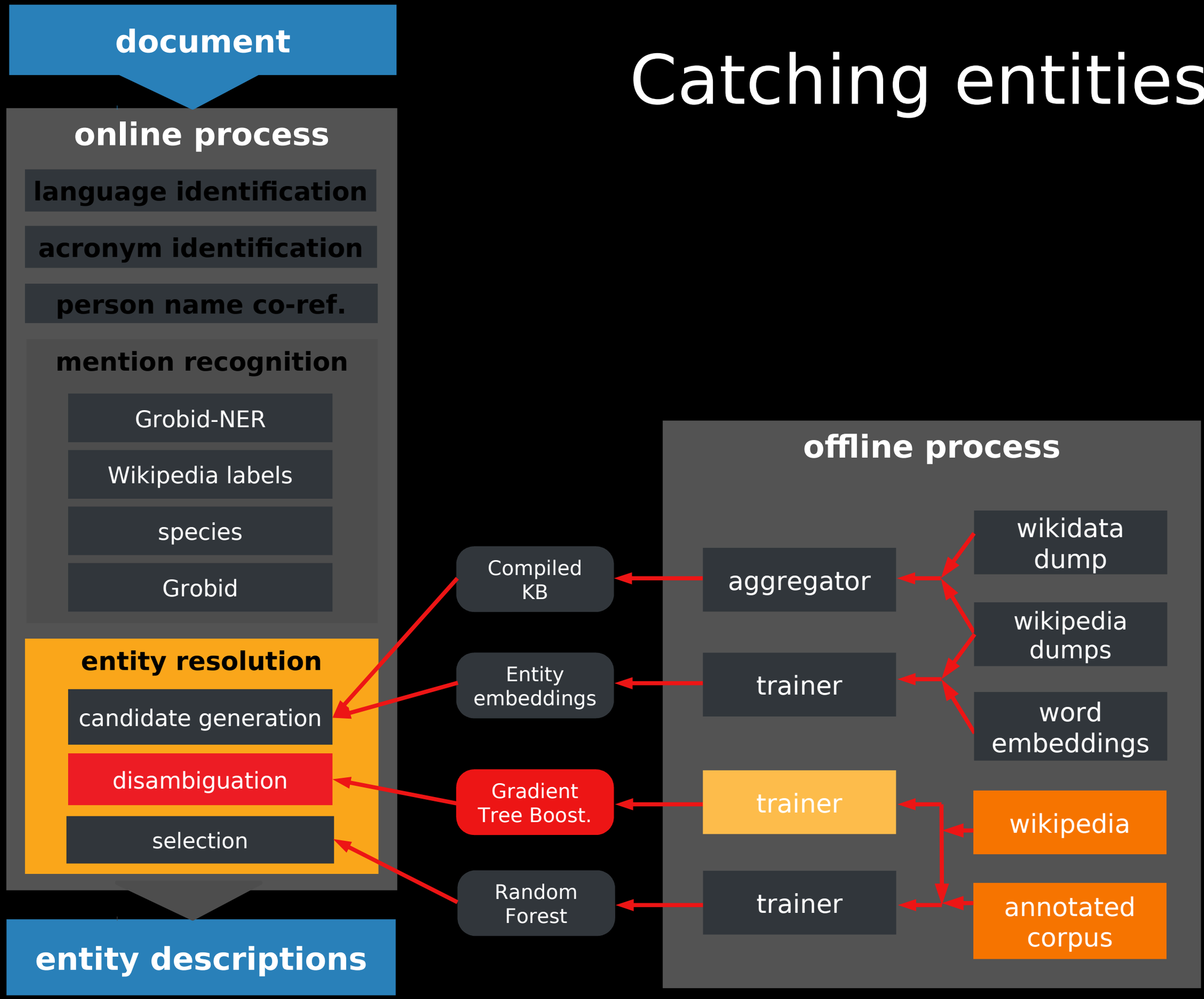


Catching entities

- Entity embeddings for ~4.5M entities
- Based on word embeddings (FastText) and page descriptions (takes 39h with 24 cores)
- Experiments for using Wikidata statements



Catching entities



Entity disambiguation

- One model per language
- Ranking entity candidates with Gradient Tree Boosting and features:
  - ➡ Milne & Witten relatedness
  - ➡ Embeddings cosine entity and word context
  - ➡ Prior probability *text* → *entity*, based on anchors in Wikipedia
  - ➡ Context quality

Milne & Witten (2009) relatedness

Number of links
with target c

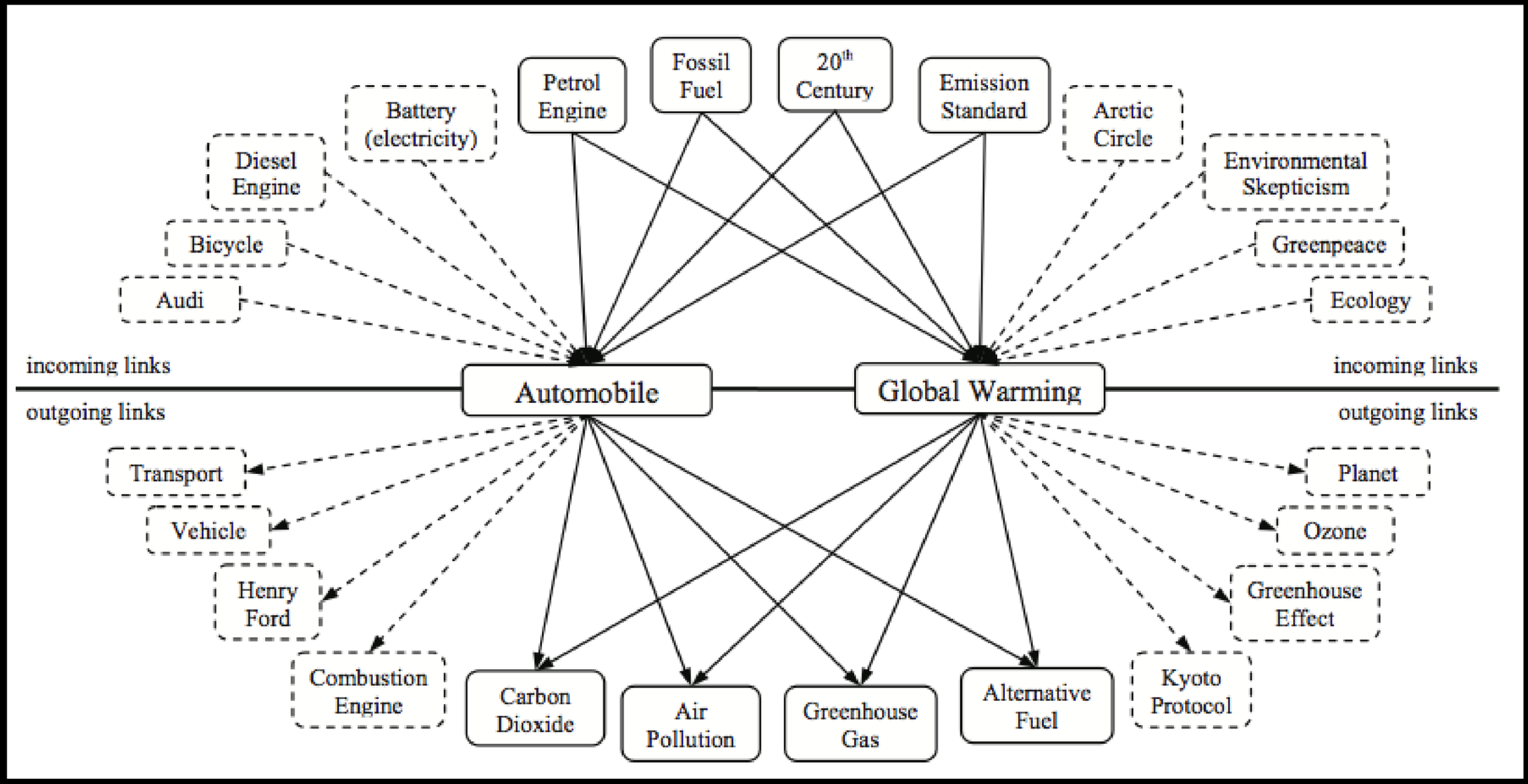
Intersection of inlinks
with target c and c'

$$\text{relateness}(c, c') = \frac{\log(\max(|L_c|, |L_{c'}|)) - \log(|L_c \cap L_{c'}|)}{\log(|WP|) - \log(\min(|L_c|, |L_{c'}|))}$$

Total number of Wikipedia articles

science-miner

Milne & Witten (2009) relatedness



➡ Extend well to Wikidata relations

Entity disambiguation

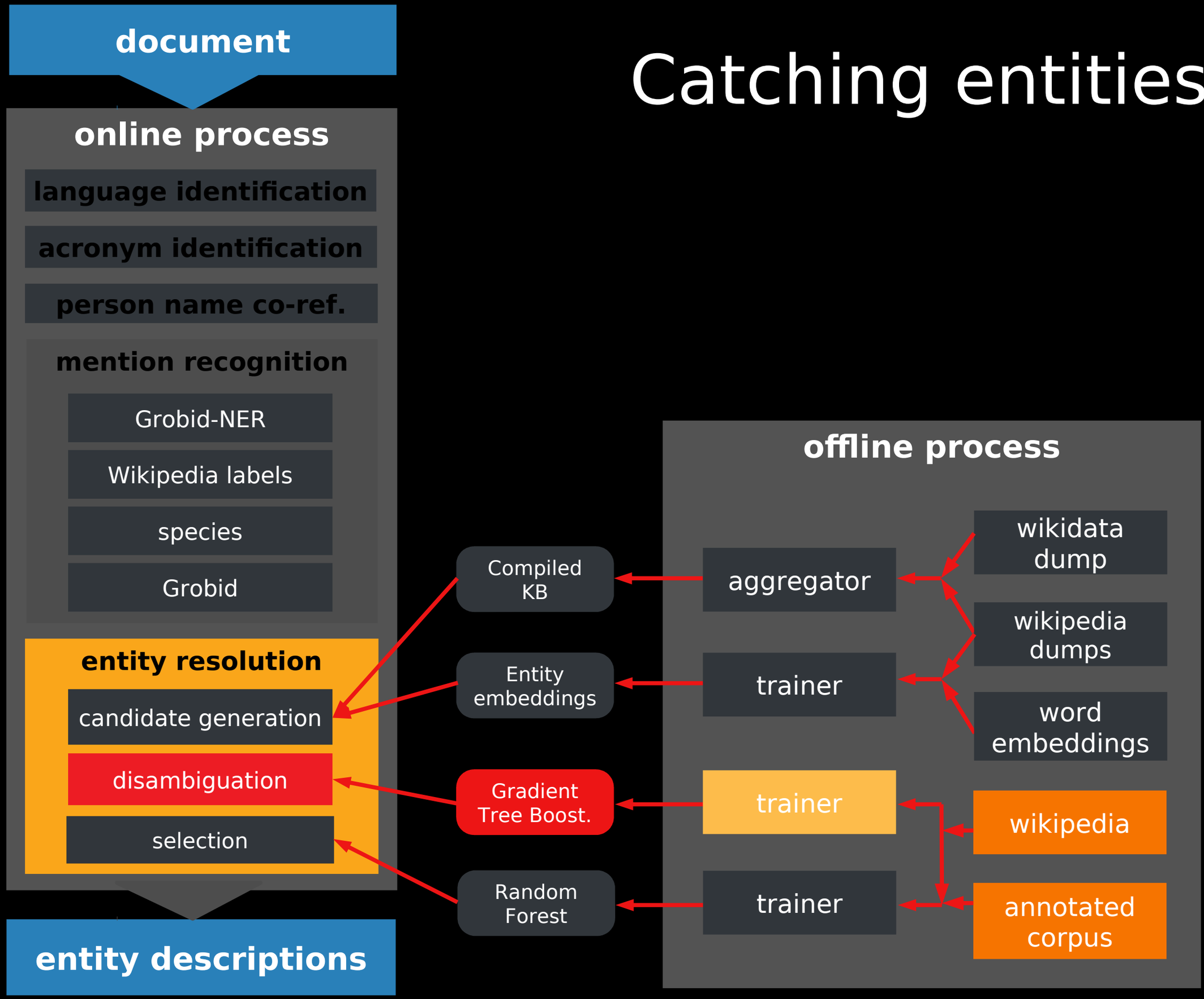
- One model per language
- Ranking entity candidates with Gradient Tree Boosting and features:
  - ➡ Milne & Witten relatedness
  - ➡ Embeddings cosine entity and word context
  - ➡ Prior probability text → entity, based on anchors in Wikipedia
  - ➡ Context quality

Entity disambiguation accuracy

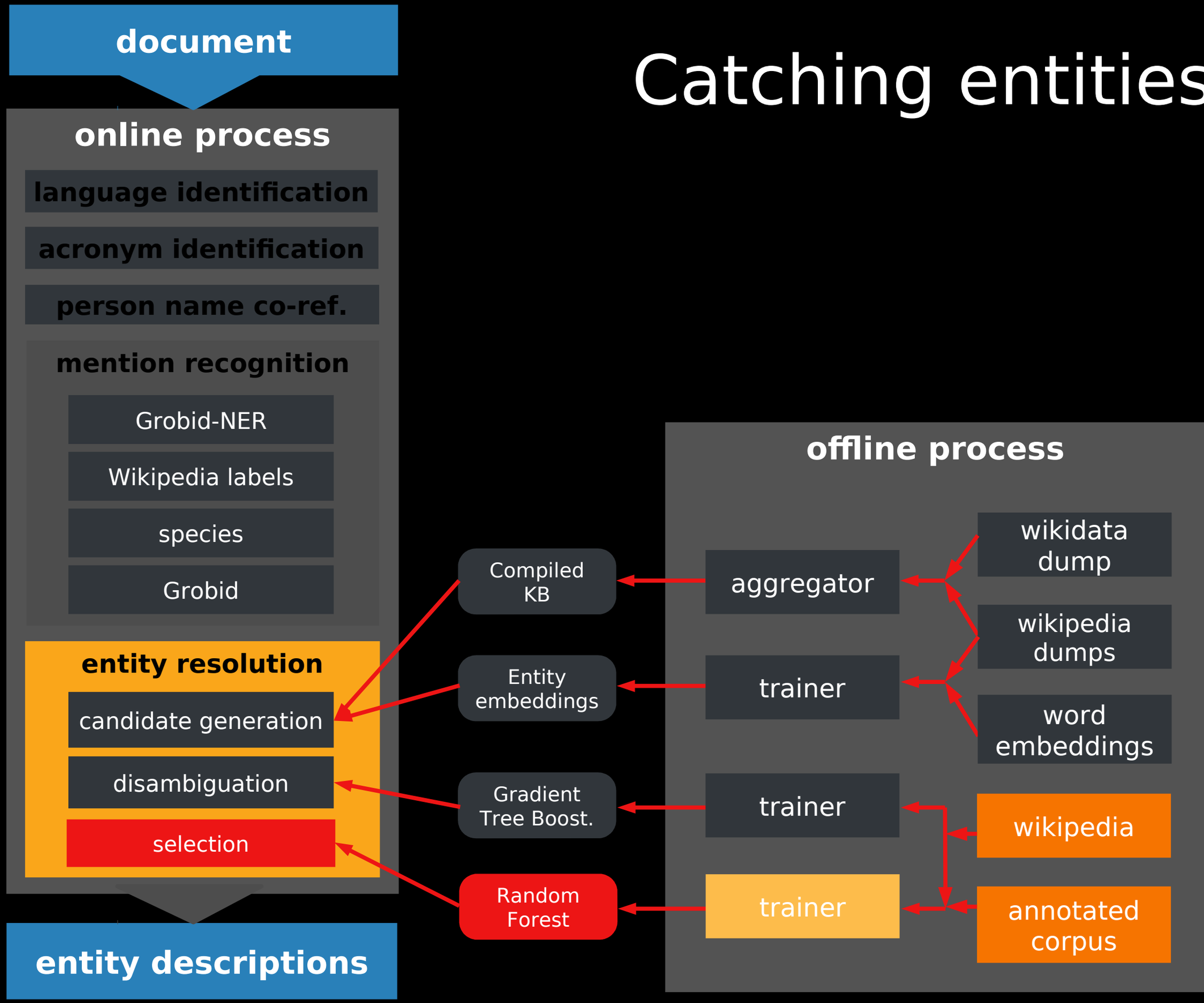
	Priors	entity-fishing	Wikifier	DoSeR	AIDA	Spotlight	Babelfy	WAT	(Ganea & Hofmann, 2017)
ACE2004	83.1	83.5	83.4	**90.7**	81.5	71.3	56.1	80.0	88.5
AIDA-CONLL-testb	66.1	76.5	77.7	78.4	77.4	59.3	59.2	84.3	**92.2**
AQUAINT	80.3	**89.1**	86.2	84.2	53.2	71.3	65.2	76.8	88.5
MSNBC	71.1	86.7	85.1	91.1	78.2	51.1	60.7	77.7	**93.7**

- only disambiguation of entities (mentions are given)
- only named entities (person, location, organisation, misc.)
- results from (Zwicklbauer & al., 2016), (Ganea & Hoffman, 2017) and GERBIL
- entity-fishing is work-in-progress and this will be improved

Catching entities



Catching entities



Scaling

# concurrent clients	1	5	6	10
text tokens/s	1371	3796	4800	3756
PDF pages/s	2.6	8.92	9.86	8.17
PDF tokens/s	1108.2	3796	4077	3376.7

- entity-fishing runs with 2GB RAM (4GB ideally)
- For comparison: AIDA 40GB, Wikifier 8-16GB (named-entity only), DoSeR 25GB (disambiguation only), ...

Some usages

- Scientific entity recognition and disambiguation from PDF (and structure-aware annotation via GROBID)
- Search engine – query disambiguation
- Key-phrase and concept extraction from scientific extraction

- And also
  - ➡ Taxonomy mapping to Wikidata (astro-thesaurus)
  - ➡ Natural language command processing
  - ➡ Bibliographical citation matching in Wikidata

Semantic enrichment for scholar search engine

Amedeo Napoli (200)
Rémi Gribonval (180)
Jérôme Euzenat (177)
Olivier Festor (174)
François Chaumette (168)
Cordelia Schmid (163)
Laurent Romary (162)
Olivier OD Devillers (158)

concepts

Algorithm (781)
Scientific modelling (569)
Data (338)
Computer simulation (322)
Equation (303)
Computer software (290)
Mathematical optimization (259)
Design (251)
Conceptual model (243)
Homogeneity and heterogeneity (235)

keyterms

Machine learning (34,169)
Educational software (32,548)
Mathematical optimization (1,103)
Learning (938)
Data mining (787)
Par (score) (772)
Machine (733)
Wireless sensor network (687)
Simulation (640)
Robot (617)

hal-01242157 Journal articles

Abstract: NEGATIVE FEEDBACK circuits are a RECURRENT MOTIF in regulatory BIOLOGICAL networks, strongly linkedto the EMERGENCE of oscillatory BEHAVIOR. The theoretical analysis of the existence of oscillationsis a difficult problem and typically involves many constraints on the monotonicity of the activityfunctions. Here, we study the occurrence of periodic solutions in an N-DIMENSIONAL class of nega-tive FEEDBACK systems defined by SMOOTH VECTOR FIELDS with a window of not necessarily monotonicactivity. Our method consists in CIRCUMSCRIBING the SMOOTH system by two PIECEWISE LINEAR ones,each admitting a PERIODIC SOLUTION. It can then be shown that the SMOOTH NEGATIVE FEEDBACK systemalso has a PERIODIC ORBIT, inscribed in the TOPOLOGICAL SOLID TORUS constructed from the two piecewiselinear orbits. The interest of our approach lies in first, adopting a general class of FUNCTIONS, with anonmonotonicity window, which permits a better fitting between theoretical models and experimen-tal DATA, and second, establishing a more accurate location for the PERIODIC SOLUTION, which is usefulfor computational purposes in HIGH DIMENSIONS. As an ILLUSTRATION, a model for the “repressilator”synthetic system is analyzed and compared to REAL DATA, and shown to admit a PERIODIC ORBIT, for arange of activity FUNCTIONS.

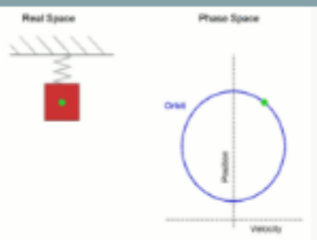
Keywords: Piecewise linear systems, negative feedback circuits, periodic oscillations, Poincaré maps AMS subject classifications 34, 92

PERIODIC ORBIT

Normalized: **Orbit (dynamics)**

Domains: **Physics**

conf: *0.37*



In mathematics, in the study of dynamical systems, an **orbit** is a collection of points related by the evolution function of the dynamical system. The orbit is a subset of the phase space and the set of all orbits is a partition of the phase space, that is different orbits do not intersect in the phase space. Understanding the properties of orbits by using topological method is one of the objectives of the modern theory of dynamical systems.

Reference: W

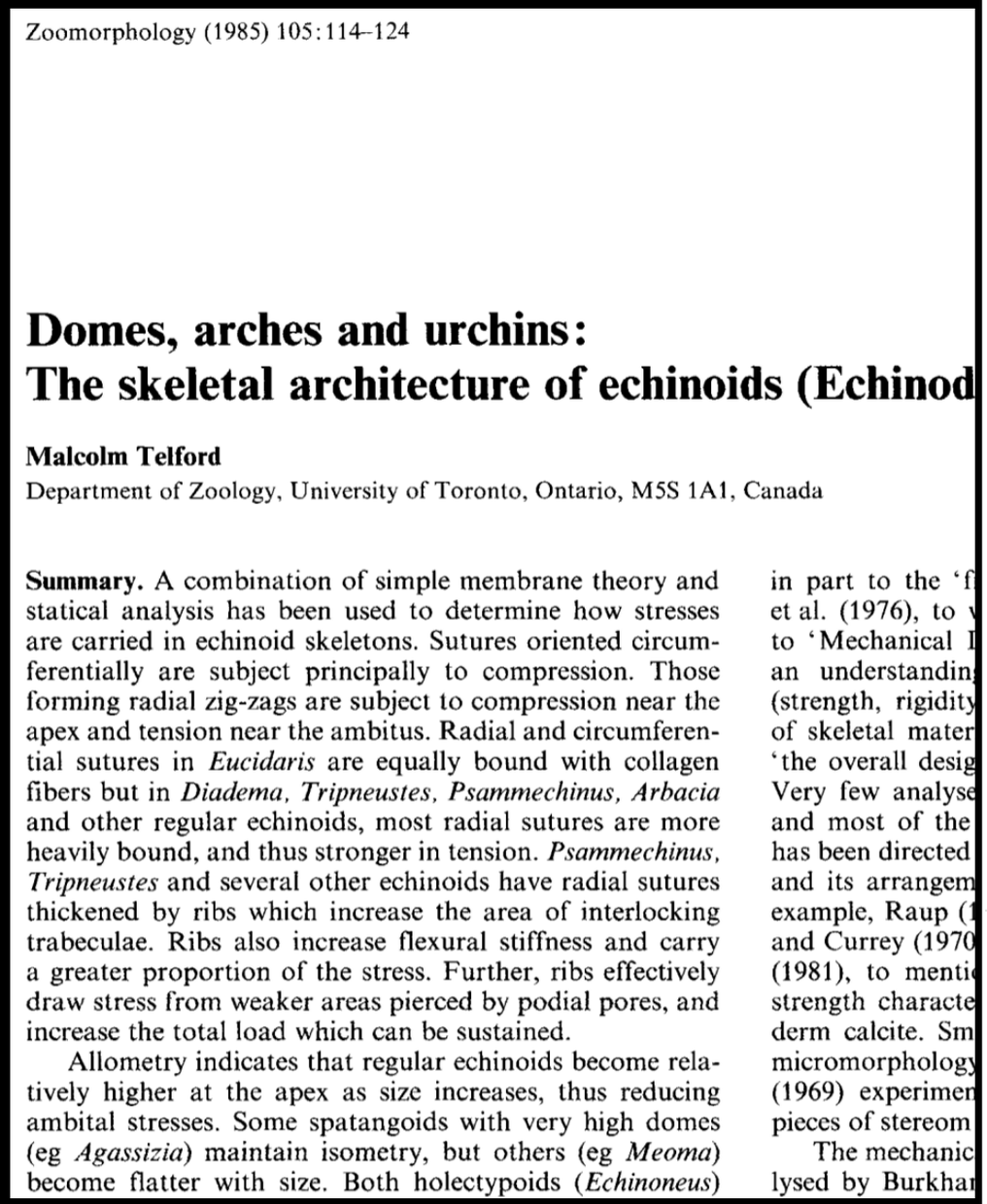
Lagrange–Schwarz Waveform Relaxation domain decomposition methods for linear and nonlinear quantum wave problems

Xavier Antoine, Emmanuel Lorin - *Applied Mathematics Letters - 2016*

Abstract/Keywords

science-miner

Key-concept extraction from scholar articles



Zoomorphology (1985) 105:114–124

Domes, arches and urchins:
The skeletal architecture of echinoids (Echinod

Malcolm Telford
Department of Zoology, University of Toronto, Ontario, M5S 1A1, Canada

Summary. A combination of simple membrane theory and statical analysis has been used to determine how stresses are carried in echinoid skeletons. Sutures oriented circumferentially are subject principally to compression. Those forming radial zig-zags are subject to compression near the apex and tension near the ambitus. Radial and circumferential sutures in *Eucidaris* are equally bound with collagen fibers but in *Diadema, Tripneustes, Psammechinus, Arbacia* and other regular echinoids, most radial sutures are more heavily bound, and thus stronger in tension. *Psammechinus, Tripneustes* and several other echinoids have radial sutures thickened by ribs which increase the area of interlocking trabeculae. Ribs also increase flexural stiffness and carry a greater proportion of the stress. Further, ribs effectively draw stress from weaker areas pierced by podial pores, and increase the total load which can be sustained.

Allometry indicates that regular echinoids become relatively higher at the apex as size increases, thus reducing ambital stresses. Some spatangoids with very high domes (eg *Agassizia*) maintain isometry, but others (eg *Meoma*) become flatter with size. Both holectypoids (*Echinoneus*)

in part to the 'f
et al. (1976), to
to 'Mechanical
an understandin
(strength, rigidity
of skeletal mater
'the overall desig
Very few analyse
and most of the
has been directed
and its arrangem
example, Raup (
and Currey (1970
(1981), to menti
strength characte
derm calcite. Sm
micromorphology
(1969) experimen
pieces of stereom

The mechanic
lysed by Burkha

membrane theory 0.0166

membrane theory

conf: 0.1680

Preferred term: History of cell membrane theory

Cell theory has its origins in seventeenth century microscopy observations, but it would be nearly two hundred years before a complete cell membrane theory be developed to explain what separates cells from the outside world. By the 19th century it was accepted that some form of semi-permeable barrier must exist around a cell. Studies of the action of anesthetic molecules led to the theory that this barrier might be made of some sort of fat (lipid), but the structure was still unknown.

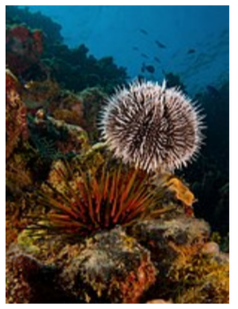
echinoid 0.0083

echinoid

conf: 0.1492

Preferred term: Sea urchin

Sea urchins or **urchins**, archaically called **sea hedgehogs**, are small, spiny, globular animals that, with their close kin, such as sand dollar, constitute the class **Echinoidea** of the echinoderm phylum. About 950 species of echinoids inhabit all oceans from the intertidal to deep. The shell, or "test", of sea urchins is round and spiny, typically from across. Common colors include black and dull shades of green, olive, brown, purple, blue, and red. Sea urchins move slowly, feeding primarily on algae. Sea otter, starfish, wolf eel, triggerfish, and other predators hunt and feed on sea urchins. The name "urchin" is an old word for hedgehog, which sea urchins resemble.



sutures 0.0062

sutures

conf: 0.1026

Preferred term: Trilobite

Trilobites (; meaning "three lobes") are a fossil group of extinct marine arachnomorph arthropod that form the class **Trilobita**. Trilobites form one of the earliest known groups of arthropods. The first appearance of trilobites in the fossil record defines the base of the Atdabanian stage of the Early Cambrian period, and they flourished throughout the lower Paleozoic era before beginning a drawn-out decline to extinction when, during the Devonian, all trilobite orders except the Proetids died out.

science-miner