CEPT Word-SDRs

or why “Jaguar” minus “Porsche” equals “Tiger”

Numenta Hackathon 2013

The Problem

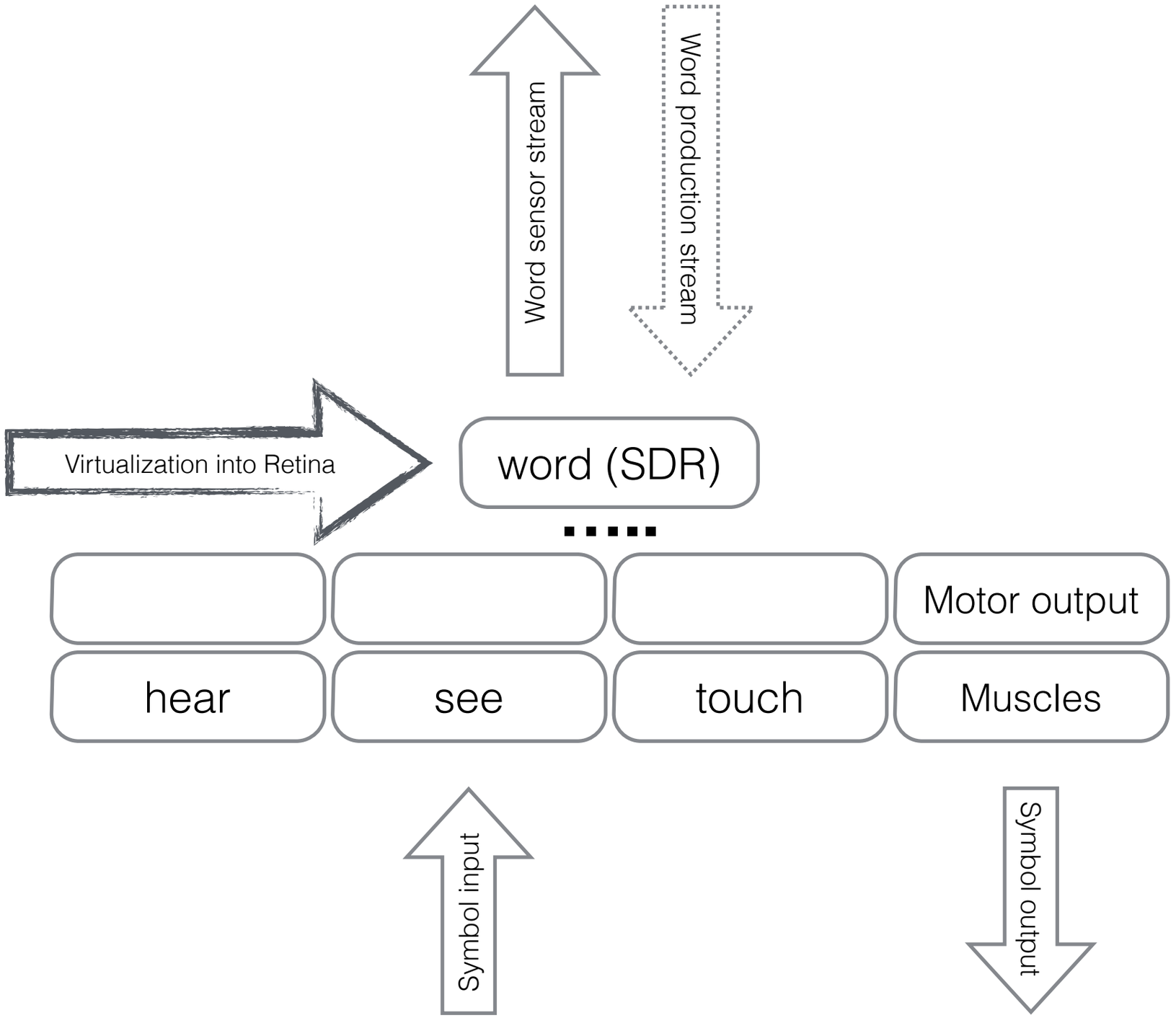
- Language —>Text—>Symbols
- Knowledge —>Meaning —>Representation

- Language is a creation of the neocortex
- Used to send (sensorial) experiences from brain to brain
- Brain and language complexity seem to go hand in hand
- Better understanding the brain helps better understand language

Brain Assumptions

- Modular composition of the NC —> Bigger is better
- Modular homology —> one CLA fits all
- CLA inputs and outputs are SDRs
- All input data originates in sensors —> All sensors output SDRs
- Organization in regions, layers, hierarchies with SDR I/O

The Word Layer



Qualifying SDR

- Binary vector —> Every bit is a semantic feature
- Sparsity —> Large combinatorial space
- Stability —> Resistance to noise or bit-failure
- Compositionality —> Semantic features add up
- Similarity —> Similar SDRs mean similar things
- *Topology —> Semantic features represent a map*

Word Vectors

- Describing words using collections of features
- Metric space —> Similarity by distance metric
- Problem: feature generation
- Examples: manual features - random indexing - Collocation (word2vec)

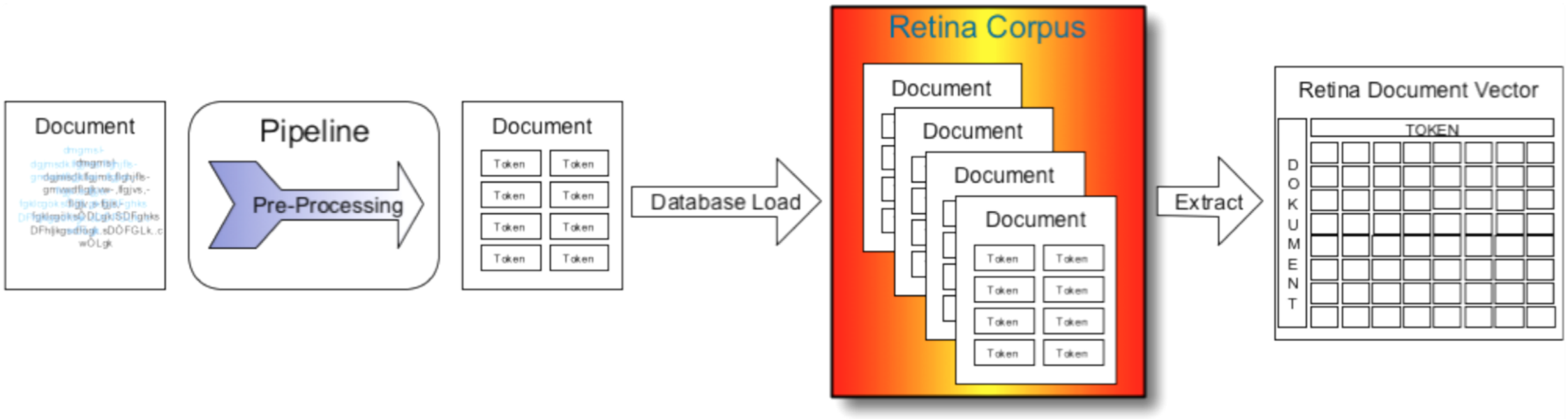
Contexts

- We teach words by building lists of (known) words
- We form our individual context repository by aggregating many special case contexts (lists)
- Every special case context is consistent
- Words can have many (independent) contexts
- The more context we have the better we understand

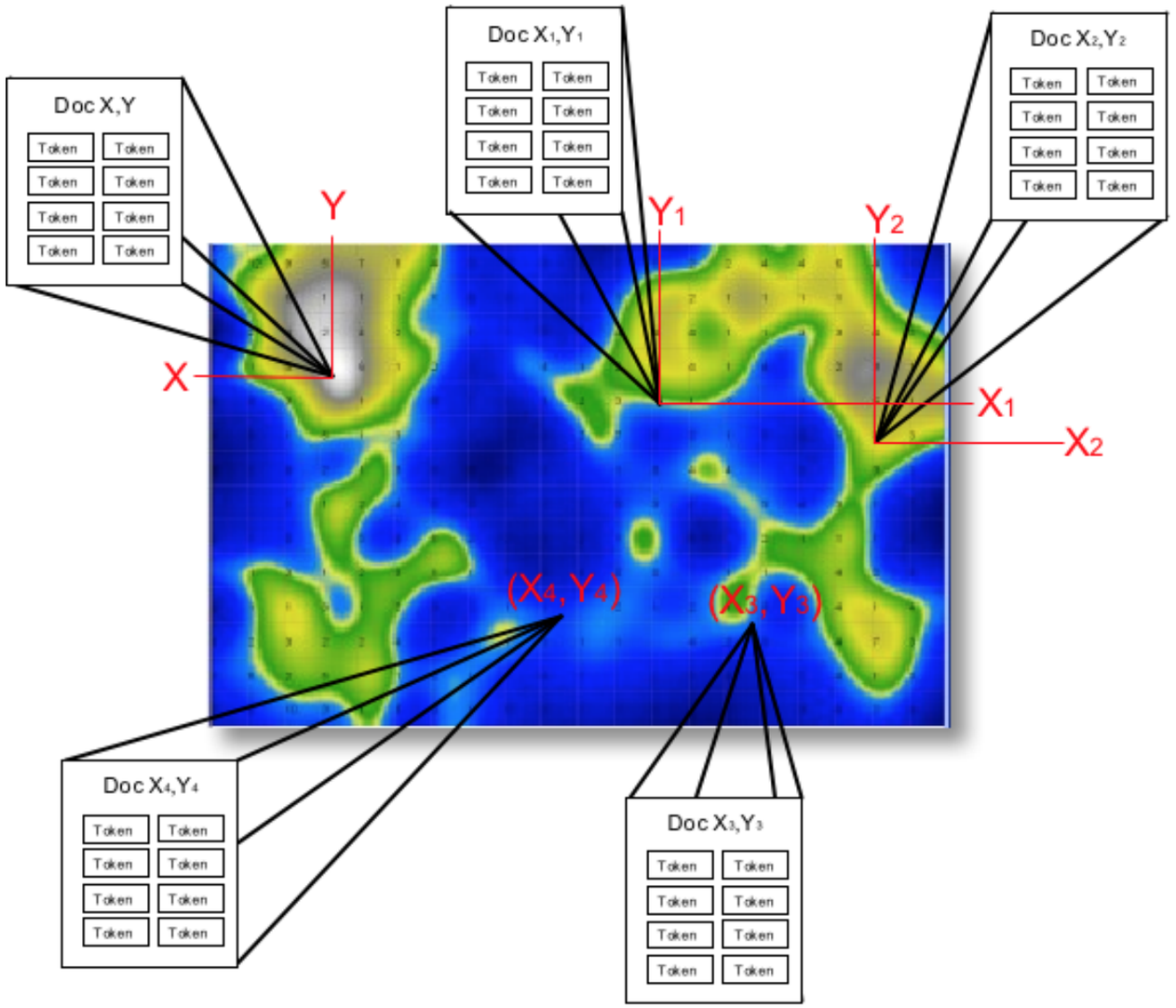
Retina Corpus

- Simulating the individual context repository
- Setting up a corpus of training documents
- Each document corresponds to a consistent single context (for example Wikipedia)
- We employ traditional NLP “tricks” to prepare the corpus
- Creation of the corpus term list —> words known by the retina

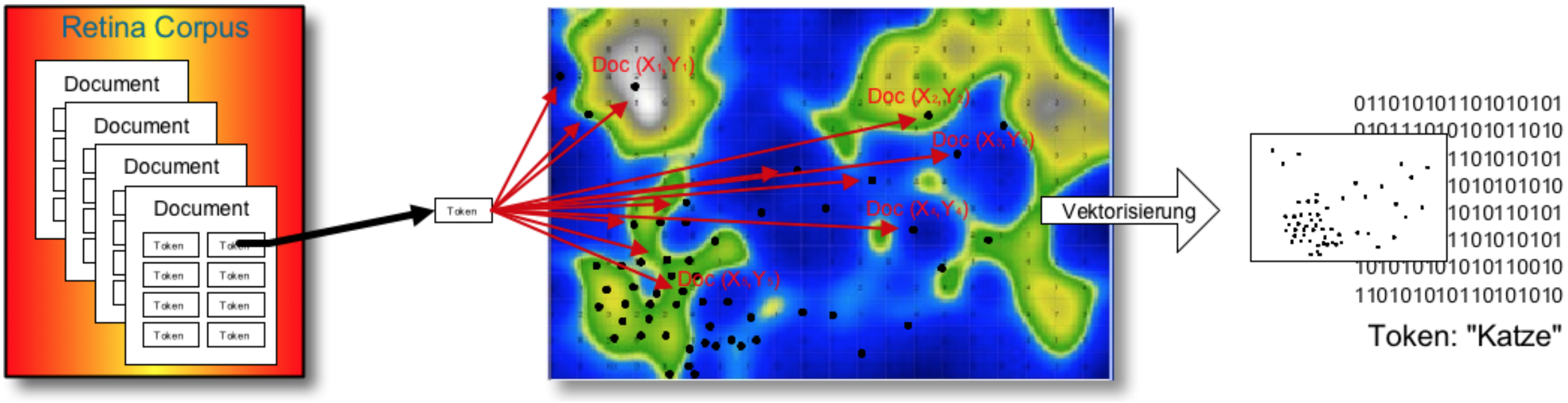
Mapping out Semantic Space



Mapping out 2D Space



Definition of Words in 2D Contexts

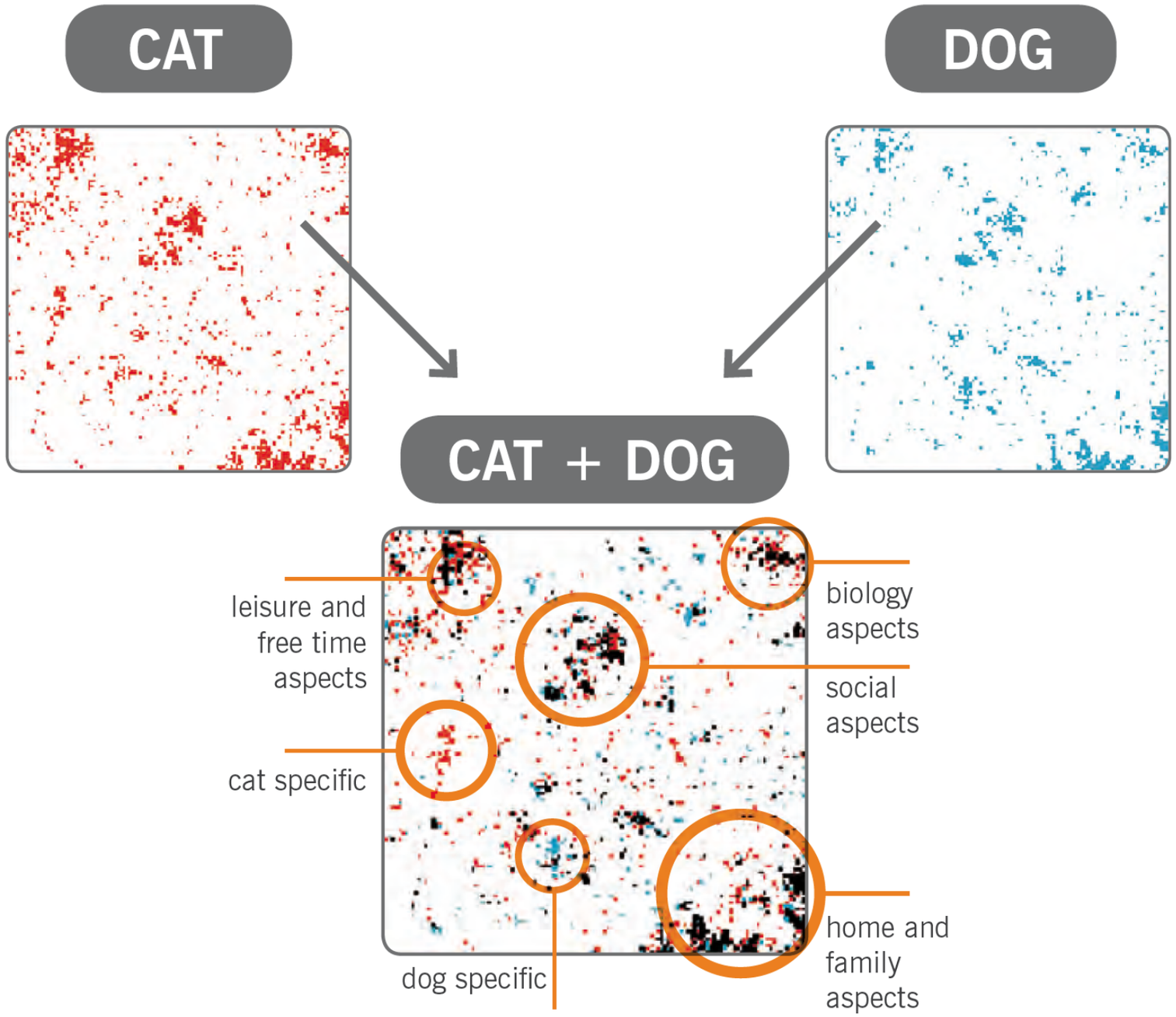


Two levels of semantics

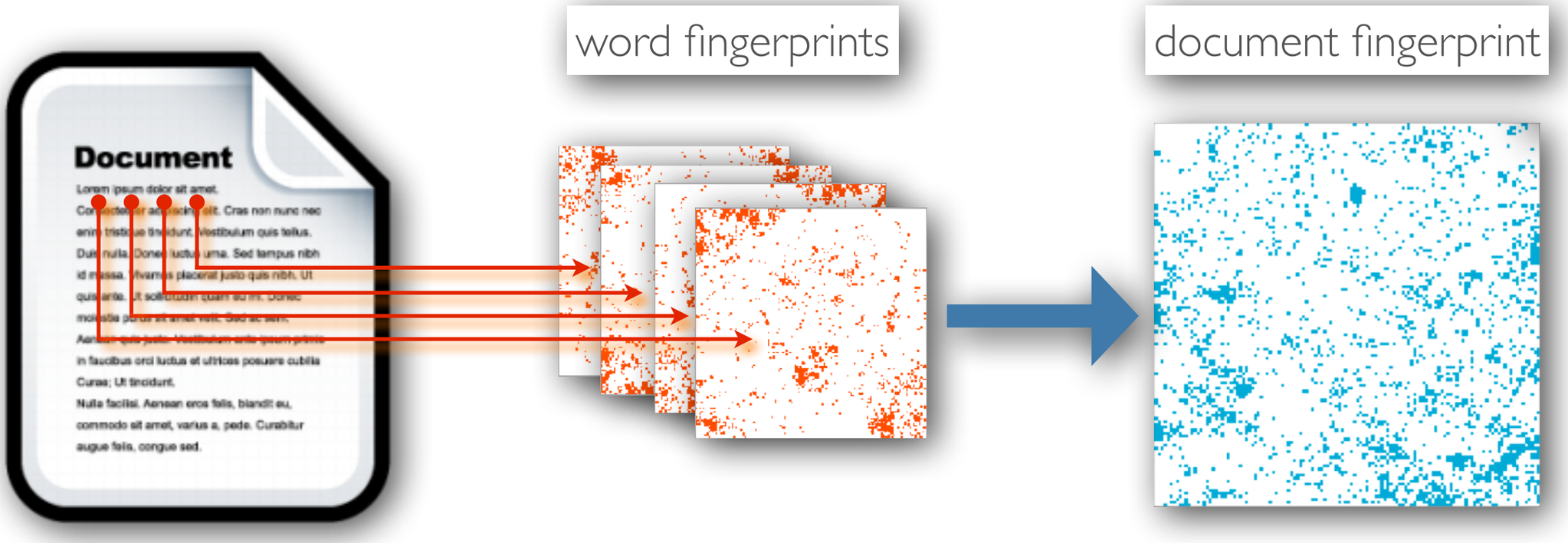
- **Lexical Semantics**
  Semantic Fingerprinting (word-SDRs)

- **Grammatical Semantics**
  Learning by capturing streams of word-SDRs in a CLA

Similarity

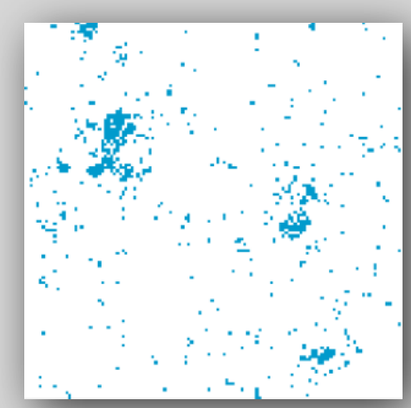

CAT
DOG
CAT + DOG
leisure and free time aspects
cat specific
dog specific
biology aspects
social aspects
home and family aspects

Compositionality

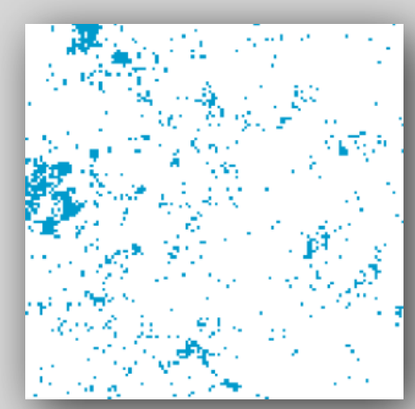


Set-Theoretic Operations

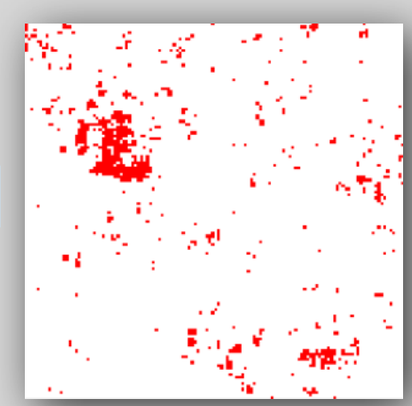
Apple − Fruit = Computer



SUB



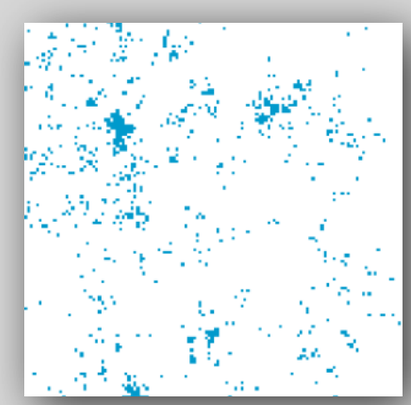
RETURN



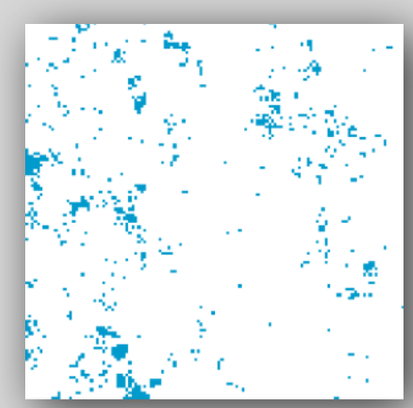
Expression Engine

Set-Theoretic Operations

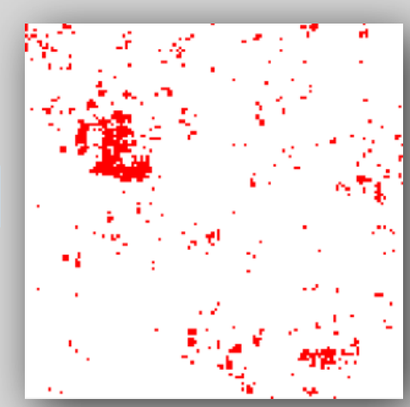
Source − Water = Computer



SUB



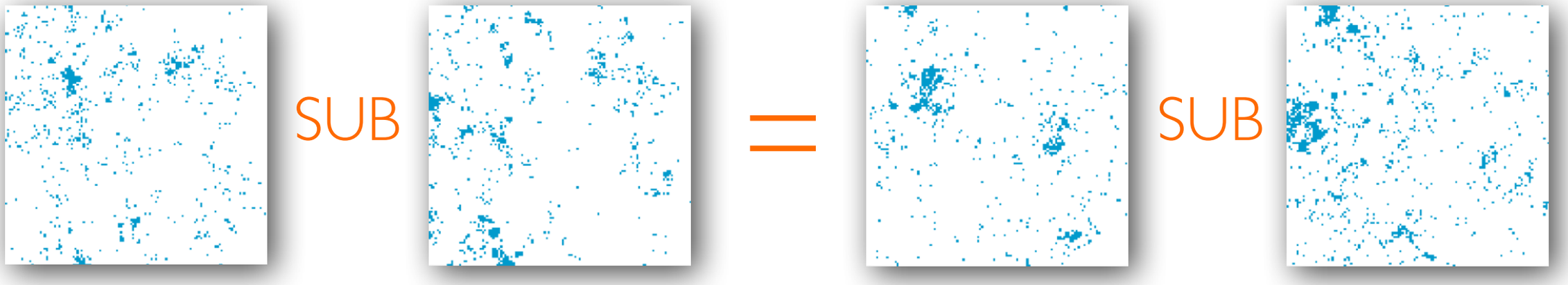
RETURN



Expression Engine

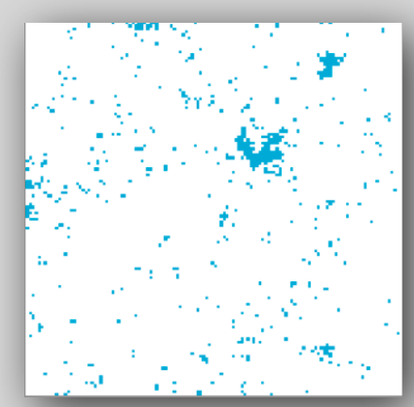
ERGO

Source − Water = Apple − Fruit

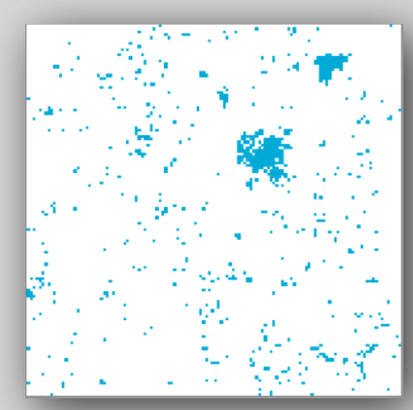


Set-Theoretic Operations

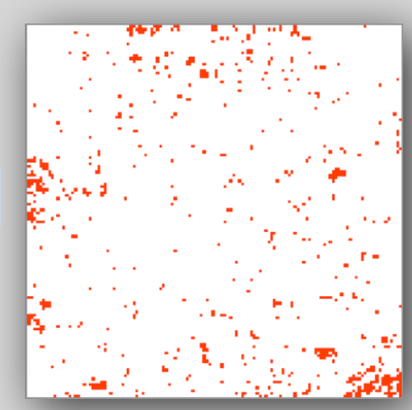
Jaguar − Porsche = Tiger



SUB

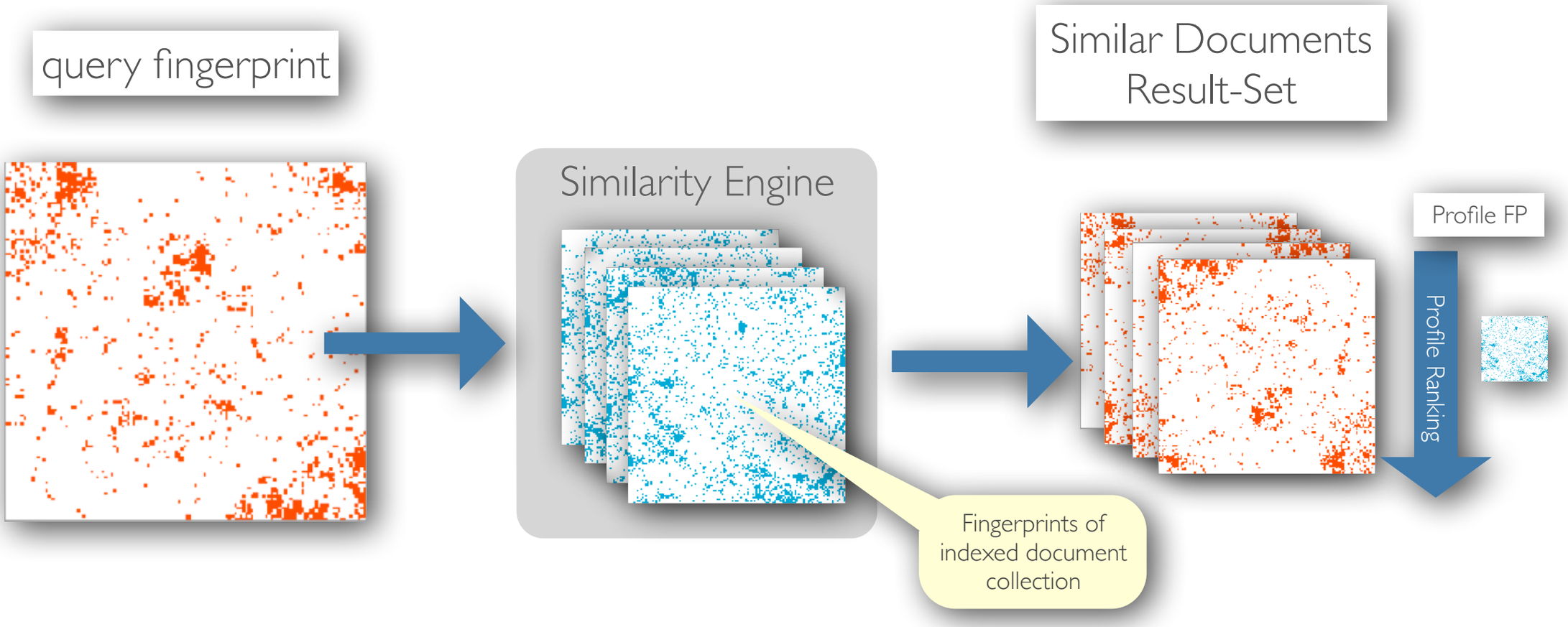


RETURN



Expression Engine

Practical Application: Search Engine



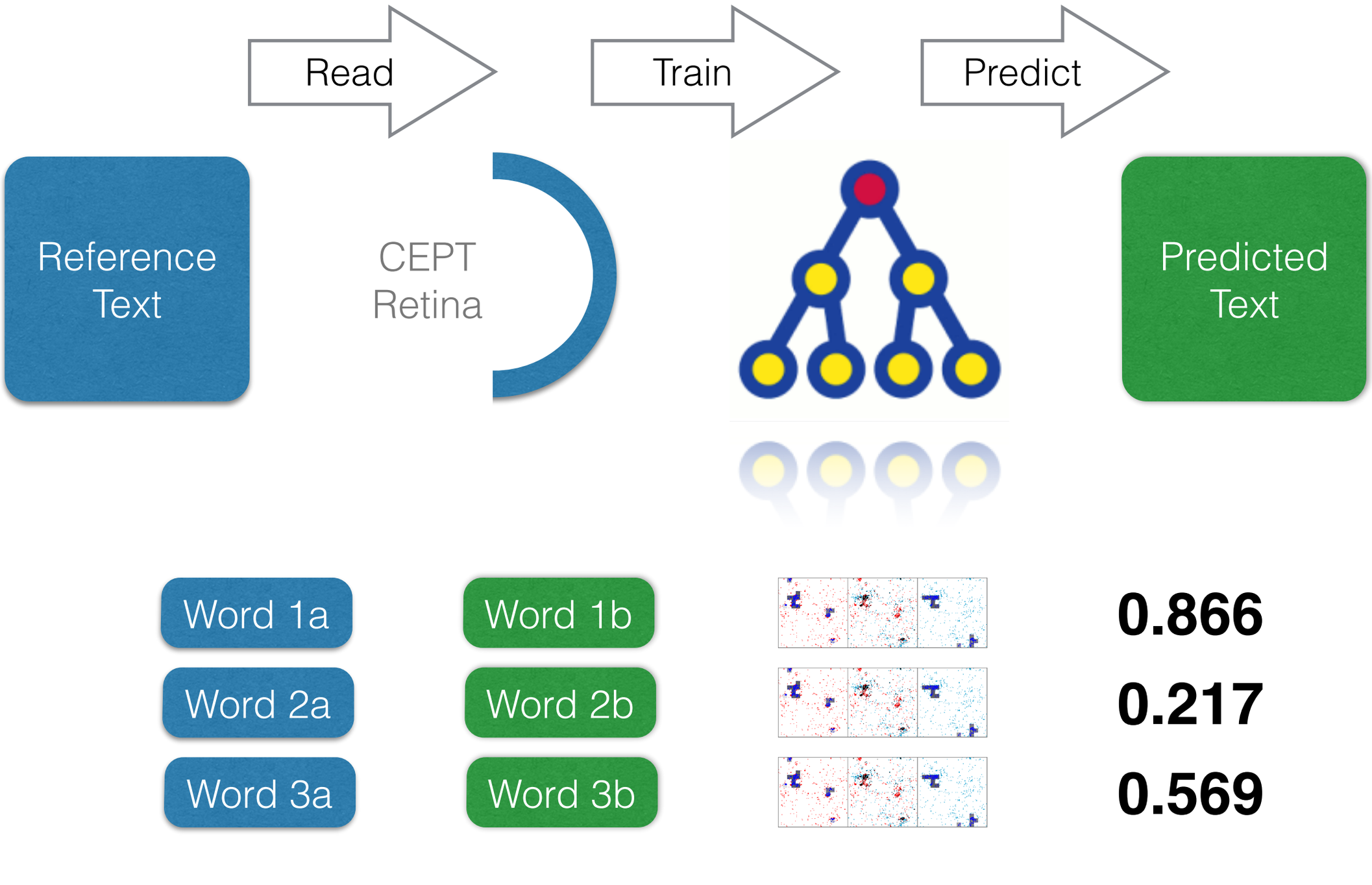
CEPT API

- Subscribe
- Upgrade to Beta Program
- Documentation
- Convert word to SDR
- Map SDR to closest word

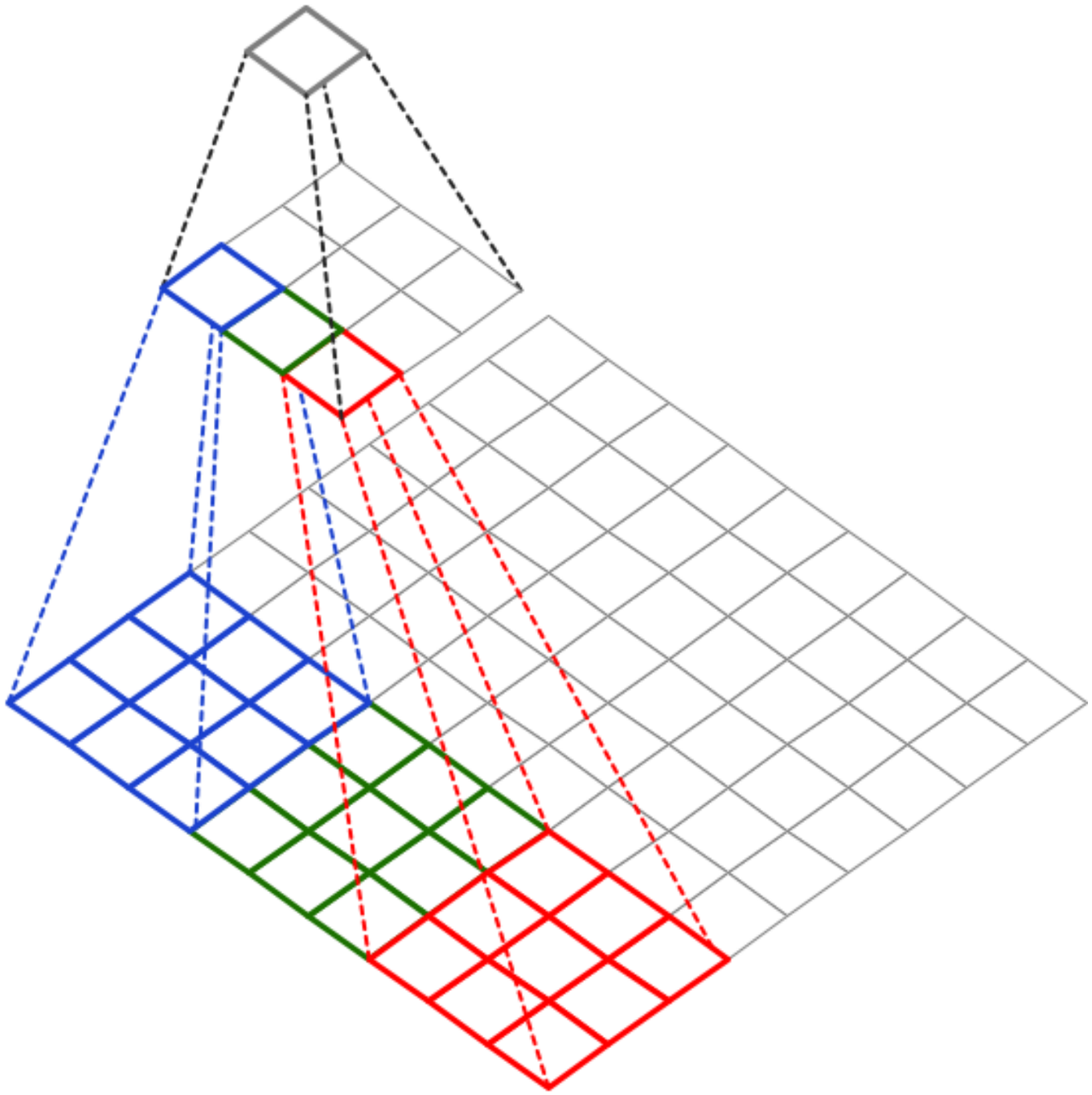
CEPT-SDRs & CLA

- Capture the meaning of sequences of word(-SDRs)
- Generation of text by:
  - prediction output (perceptive “talk”)
  - layer output (descriptive “talk”)
  - *future: motor output (reactive “talk”)*
- Statements, Arguments, Metaphors, Dialogs etc.

CLA Evaluation Framework



CLA - CEPT Retina Link-up



Examples

- Sentiment Analysis
- Automatic Abstracting
- Speech Recognition Improvement by Text-Feedback
- Dialogue System
- Automatic Reporting System
- Machine Translation