

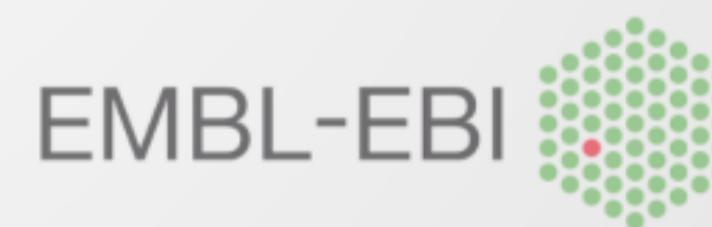
# Using Elasticsearch to store, integrate and mine diverse biological data

Andrea Pierleoni

CTTV Core Team - EBI



[goo.gl/hmTI6j](http://goo.gl/hmTI6j)



4 Feb 2016

# Centre for Therapeutic Target Validation

## CTTV

# CTTV

- Private-Public Initiative
  - EBI - Sanger - GSK
- Aims to provide evidence on the biological validity of therapeutic targets using genome-scale experiments and analysis
- CTTV Core
  - The Core Team is in charge of organise and integrate the available data and serve it through a Web App and a REST API
  - Our software stack should be easy to replicate in internal mirrors by partner organisations

# Overview

- Data
- Data Integration
- Data Mining
- Architecture

# Data

# Core Data

# Core Data

## Data Sources

- ChEMBL
- COSMIC
- Europe PMC
- EVA
- Expression Atlas
- GWAs Catalog
- Phenodigm
- Uniprot
- ... and eventually more  
in private deployments

# Core Data

## Data Sources

- ChEMBL
- COSMIC
- Europe PMC
- EVA
- Expression Atlas
- GWAs Catalog
- Phenodigm
- Uniprot
- ... and eventually more  
in private deployments

## Data Types

- Genetic association
- Somatic mutation
- RNA expression
- Known drugs
- Animal models
- Text mining

# Public Database Data

# Public Database Data

- Target related generic data
- Disease related data (EFO Ontology)
- Evidence Ontology
- Reactome pathways
- Protein and RNA expression data
- more to come...

# CTTV Building Blocks

TARGET

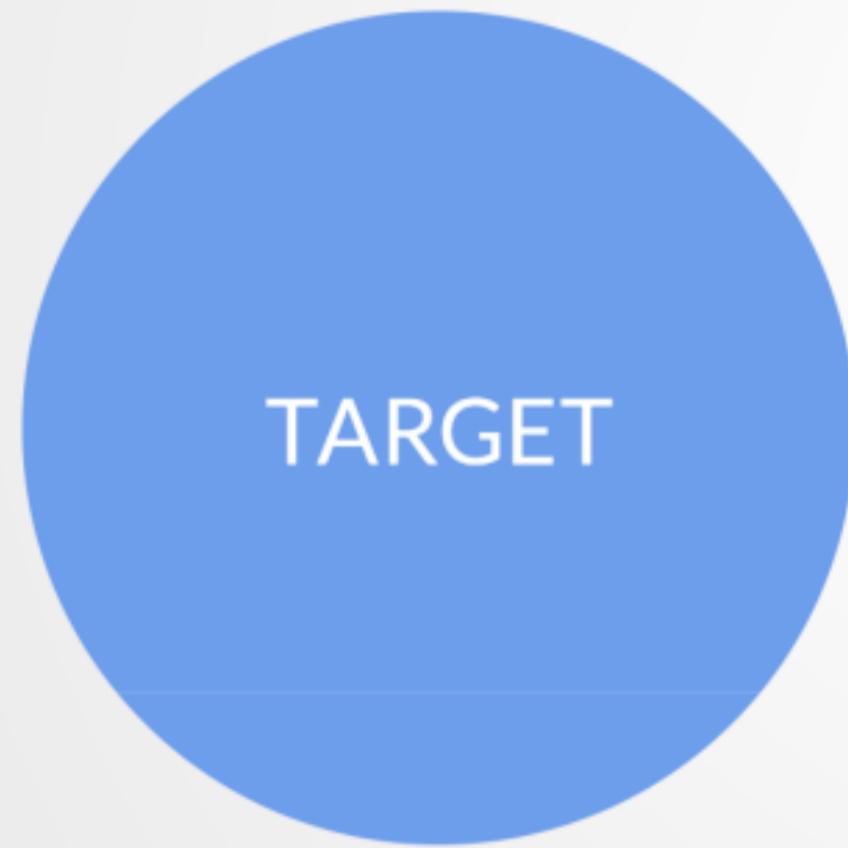
# CTTV Building Blocks

## Relate a Target and a Disease through an Evidence

TARGET

# CTTV Building Blocks

Relate a Target and a Disease through  
an Evidence



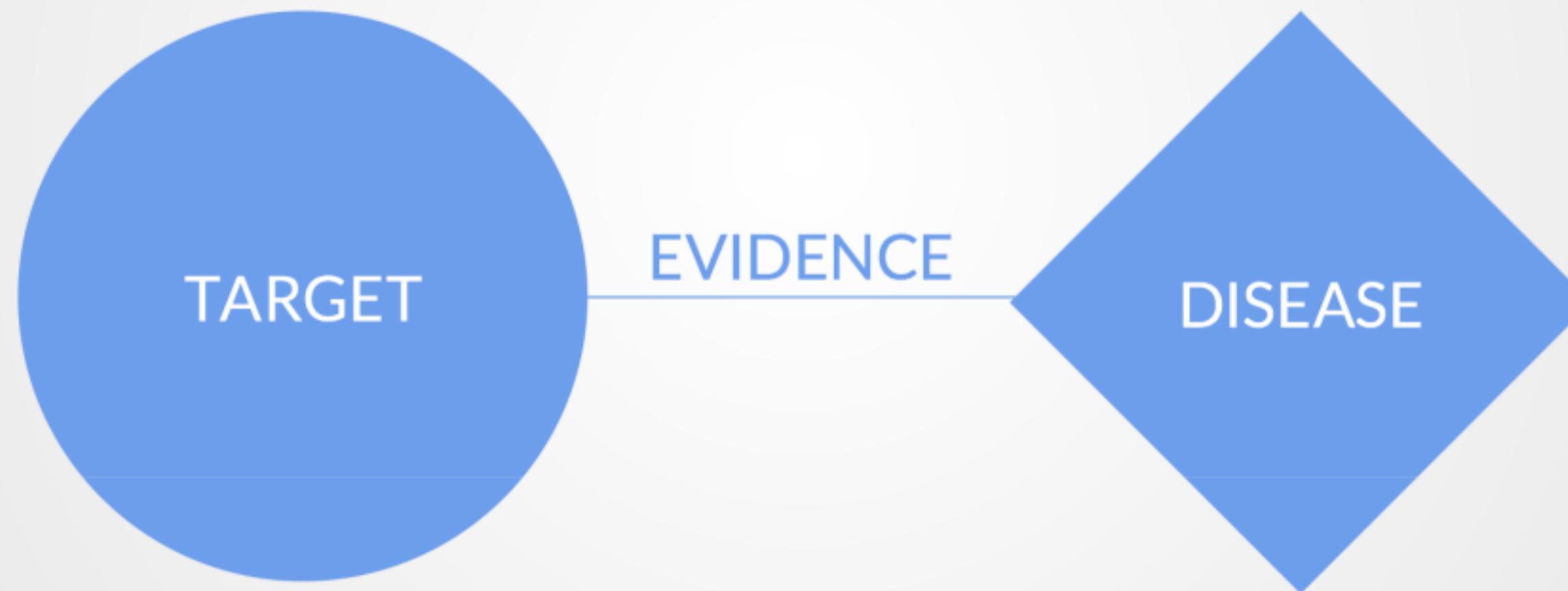
# CTTV Building Blocks

Relate a Target and a Disease through  
an Evidence



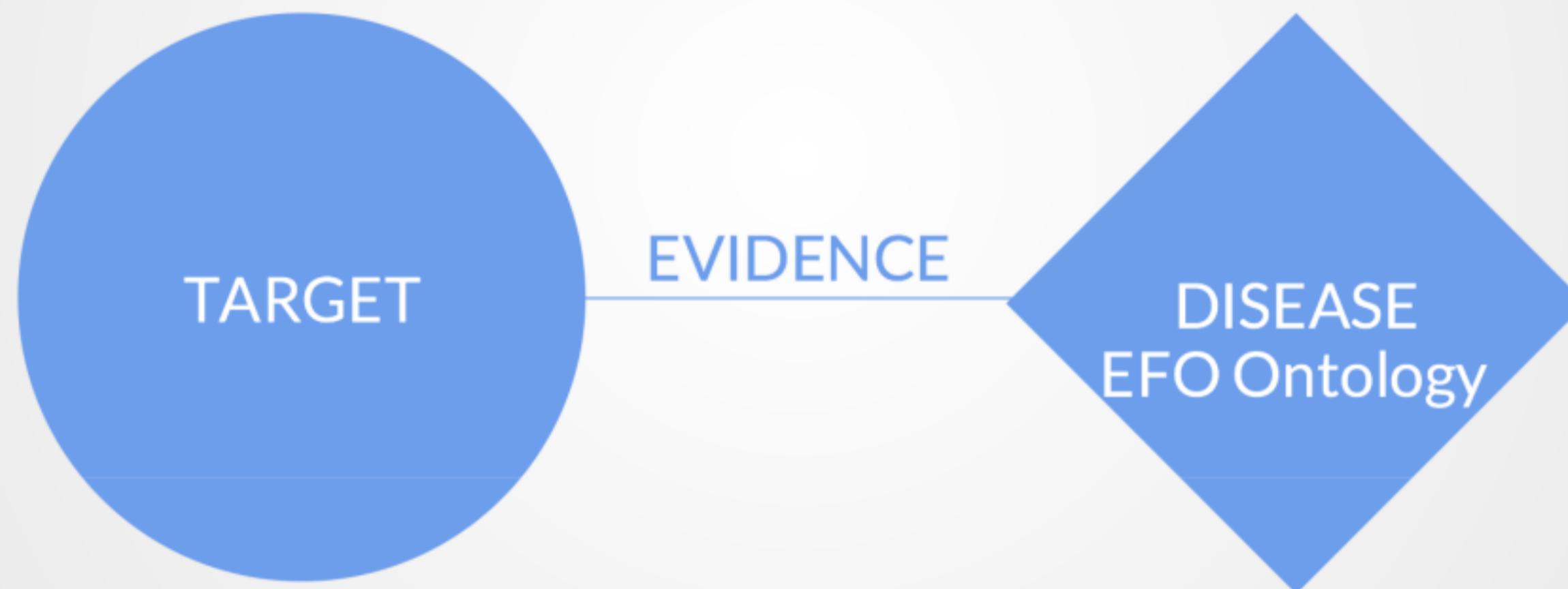
# CTTV Building Blocks

Relate a Target and a Disease through  
an Evidence



# CTTV Building Blocks

Relate a Target and a Disease through  
an Evidence



# Building blocks in real life

# Building blocks in real life

## JSON objects with a flexible format: Evidence String

# Building blocks in real life

## JSON objects with a flexible format: Evidence String

```
{  
    "sourceID": "expression_atlas",  
    "type": "rna_expression",  
    "target": {  
        "id": [  
            "http://identifiers.org/ensembl/ENSG00000112984"  
        ],  
        ...  
    },  
    "disease": {  
        "id": [  
            "http://www.ebi.ac.uk/efo/EFO_0002917"  
        ],  
        ...  
    },  
    "evidence": {  
        "unique_experiment_reference": "STUDYID_E-GEOID-14001",  
        "resource_score": {  
            "value": 0.0168,  
            "type": "pvalue"  
        },  
        "evidence_codes": [  
            "http://purl.obolibrary.org/obo/ECO_0000356"  
        ],  
        "log2_fold_change": {  
            ...  
        }  
    }  
}
```

# 4.2 Millions Evidence Strings

# Storing Evidence Strings in Elasticsearch

# Storing Evidence Strings in Elasticsearch

PUT JSON into ES

EASY. Just be sure to use a document type per datasource  
to avoid automatic mapping clashes

# Storing Evidence Strings in Elasticsearch

PUT JSON into ES

EASY. Just be sure to use a document type per datasource  
to avoid automatic mapping clashes

Filter for Target, Disease and other metadata  
Needs ad-hoc mapping

# Storing Evidence Strings in Elasticsearch

PUT JSON into ES

EASY. Just be sure to use a document type per datasource  
to avoid automatic mapping clashes

Filter for Target, Disease and other metadata

Needs ad-hoc mapping

How many Indexes?

Document count per source range from 20K to 3.3M

Best configuration with a separate index for text mining

# Storing Evidence Strings in Elasticsearch

PUT JSON into ES

EASY. Just be sure to use a document type per datasource  
to avoid automatic mapping clashes

Filter for Target, Disease and other metadata

Needs ad-hoc mapping

How many Indexes?

Document count per source range from 20K to 3.3M

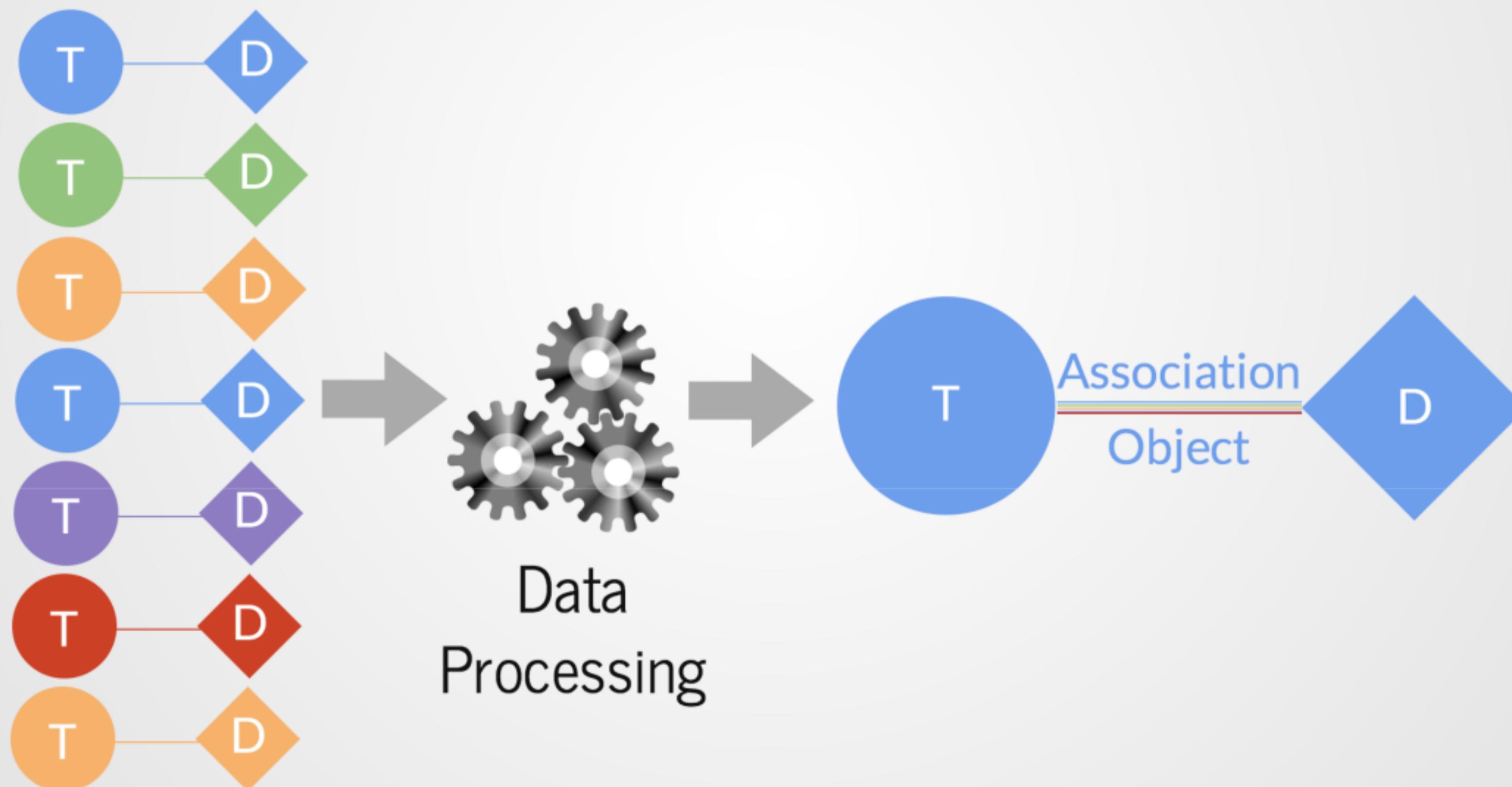
Best configuration with a separate index for text mining

Routing

Routing by target ID distribute nicely the data and  
improves filter by target query performances

# Data Integration

# From single evidence to target-disease association



# Data Processing

Strategy #1: compute in Python-based REST API

- Retrieve all relevant evidence strings from ES
- Process them in the api
- Return computed object as JSON

# Data Processing

Strategy #1: compute in Python-based REST API

- Retrieve all relevant evidence strings from ES
- Process them in the api
- Return computed object as JSON

PROS

- Easy to code, and experiment with

# Data Processing

Strategy #1: compute in Python-based REST API

- Retrieve all relevant evidence strings from ES
- Process them in the api
- Return computed object as JSON

## PROS

- Easy to code, and experiment with

## CONS

- Very slow to retrieve a big number of evidence strings through the network
- Python is not the fastest language out there
- Heaviest query took 20 min to run!!

# Data Processing

Strategy #2: compute with ES scripted metric aggregation

- Compute heavy processing steps in ES
- Process results the api
- Return computed object as JSON

# Data Processing

Strategy #2: compute with ES scripted metric aggregation

- Compute heavy processing steps in ES
- Process results the api
- Return computed object as JSON

## PROS

- Much faster, more than 10X over first attempt

# Data Processing

Strategy #2: compute with ES scripted metric aggregation

- Compute heavy processing steps in ES
- Process results the api
- Return computed object as JSON

## PROS

- Much faster, more than 10X over first attempt

## CONS

- Tricky to code and debug
- Heavy load in ES cluster as it requires lots of CPU
- Cannot compute aggregations

# Data Processing

Strategy #3: precompute association JSON objects

- Precompute JSON Object in a Python-based pipeline
- Filter and return computed object from the Python REST API

# Data Processing

Strategy #3: precompute association JSON objects

- Precompute JSON Object in a Python-based pipeline
- Filter and return computed object from the Python REST API

## PROS

- Faster, more than 100X over first attempt.
- All queries have a similar weight on the backend
- Can compute aggregations

# Data Processing

Strategy #3: precompute association JSON objects

- Precompute JSON Object in a Python-based pipeline
- Filter and return computed object from the Python REST API

## PROS

- Faster, more than 100X over first attempt.
- All queries have a similar weight on the backend
- Can compute aggregations

## CONS

- Not able to tune computation parameters in the live system

# 2.1 Millions Association Objects

# Enabling Full Text Search

- Map Association objects to full text search entities
  - Targets
  - Diseases
  - more to come (SNPs, Publications, Drugs, etc...)
- 70K search entities so far (Ensembl Gene Ids + EFO)

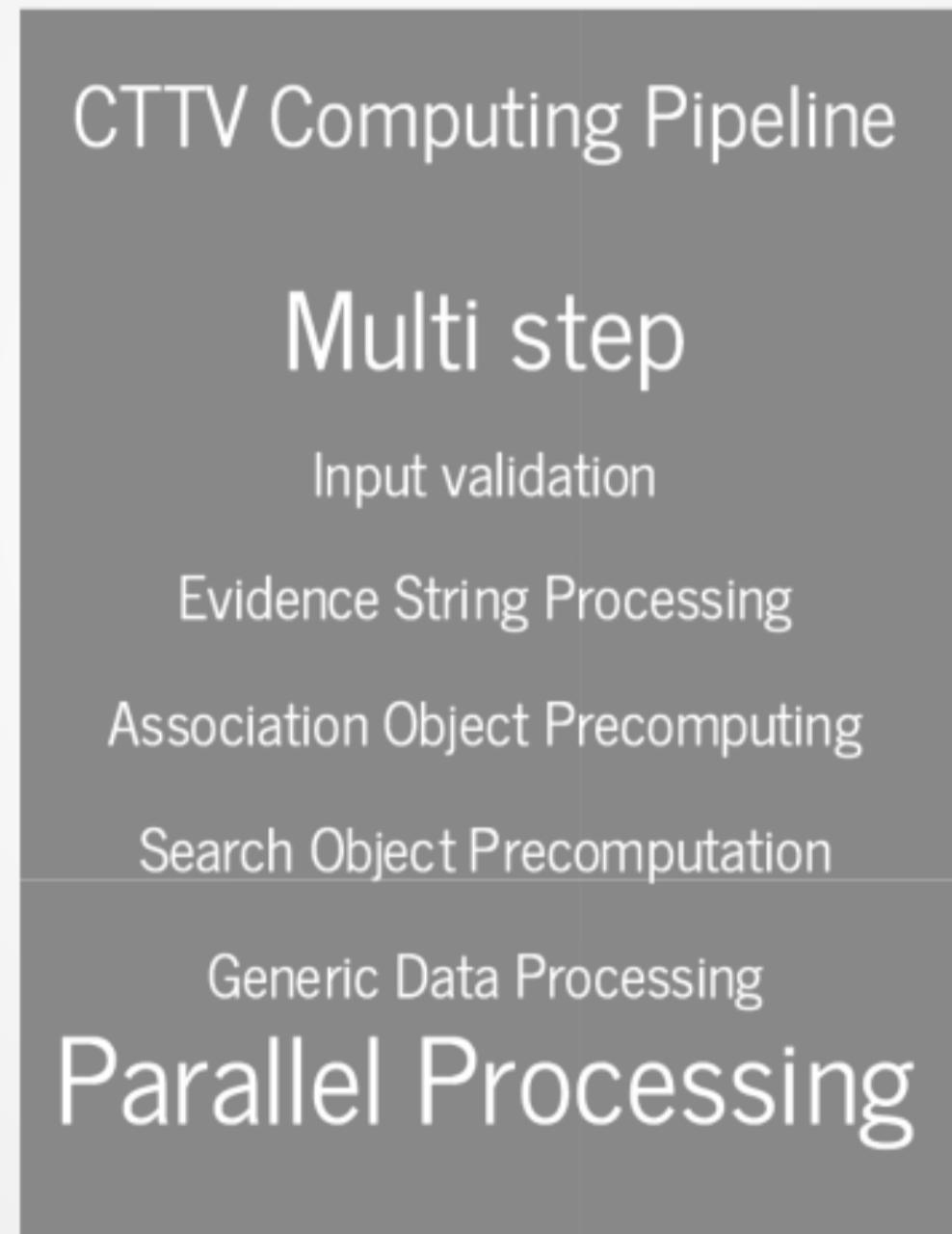
# Data Processing

## Computing Pipeline

# Data Processing Computing Pipeline



Postgres 9.4



Elasticsearch  
1.7.2

# Data Processing Computing Pipeline



Postgres 9.4



CTTV Computing Pipeline

Multi step

Input validation

Evidence String Processing

Association Object Precomputing

Search Object Precomputation

Generic Data Processing

Parallel Processing



Elasticsearch  
1.7.2

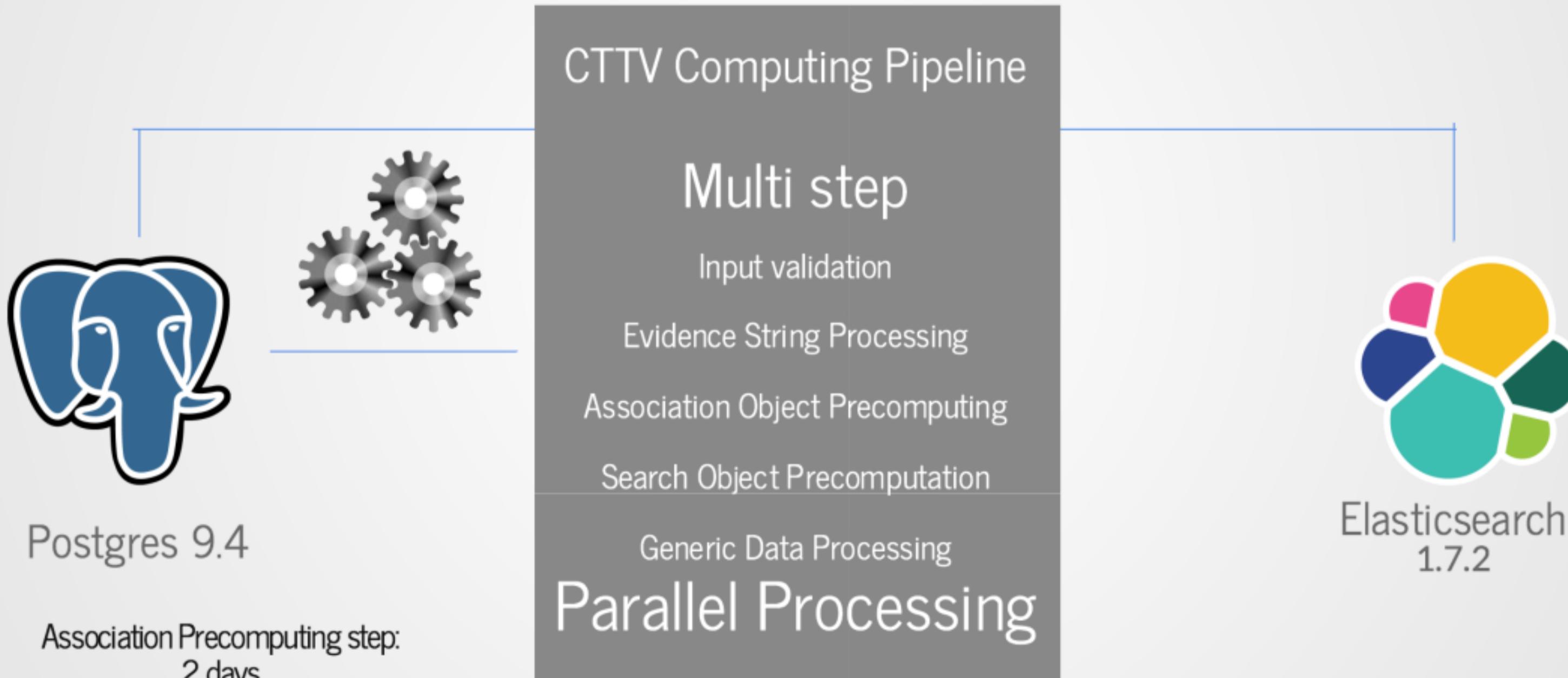


python™

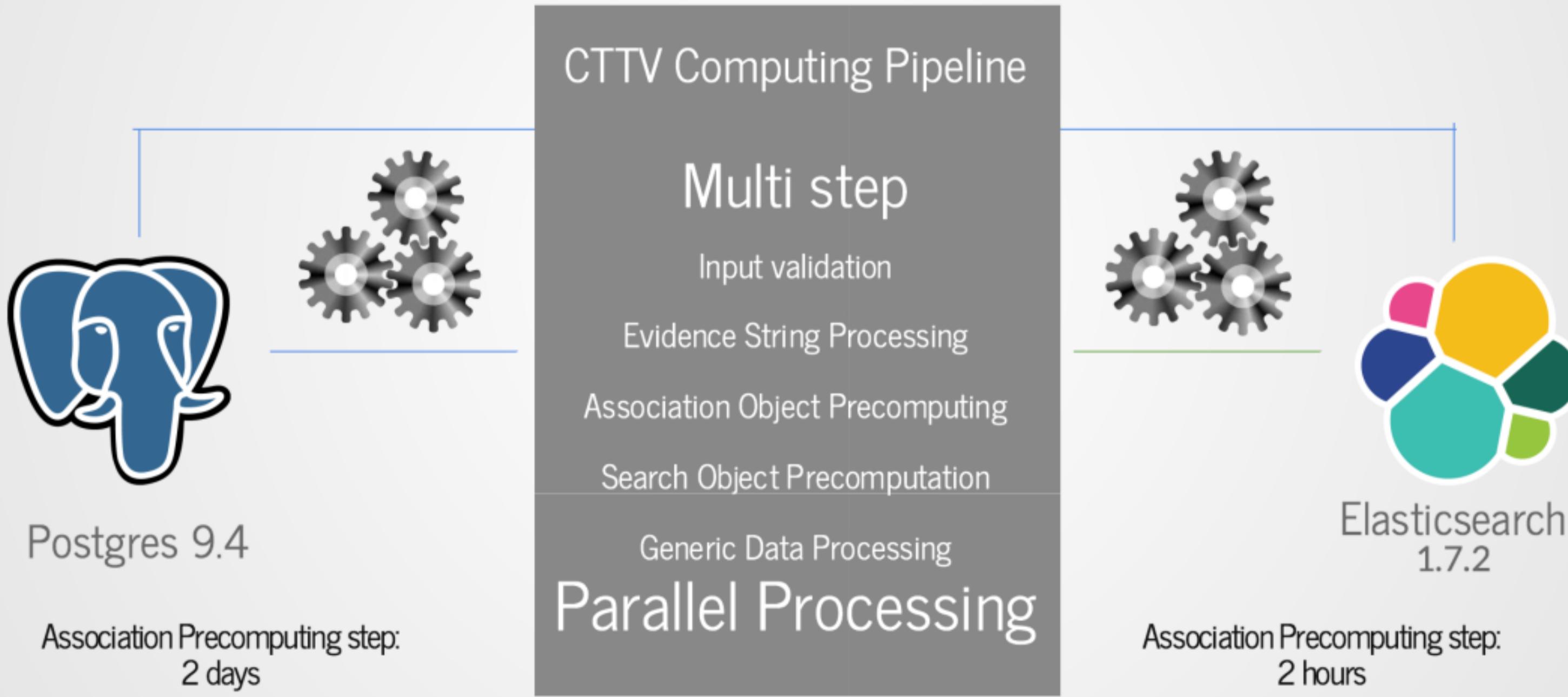


redis

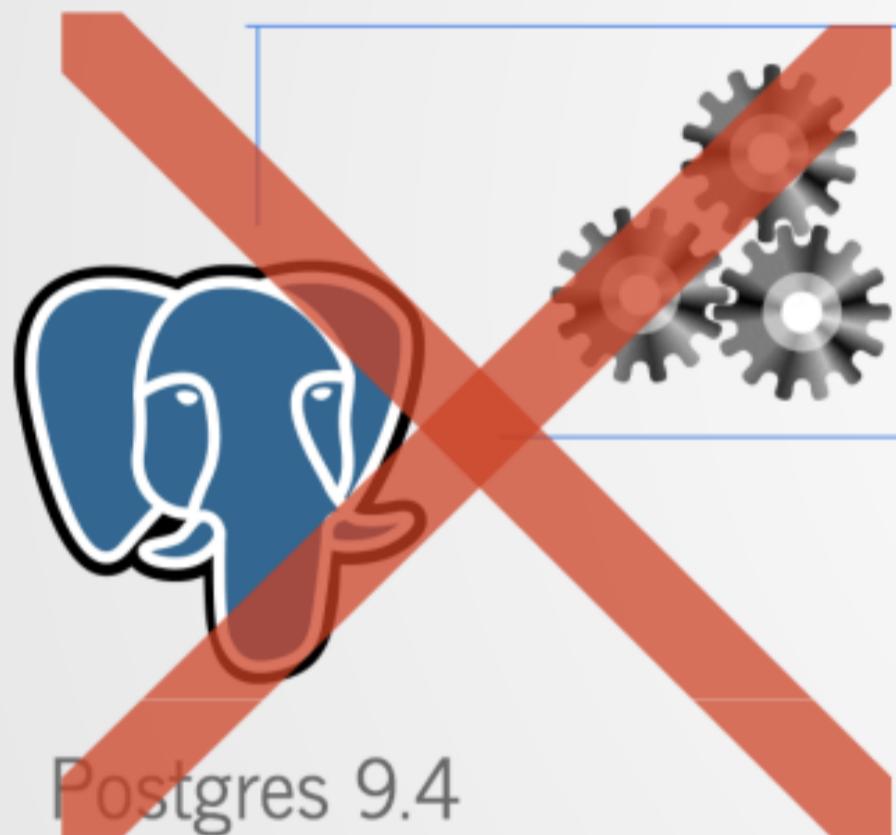
# Data Processing Computing Pipeline



# Data Processing Computing Pipeline

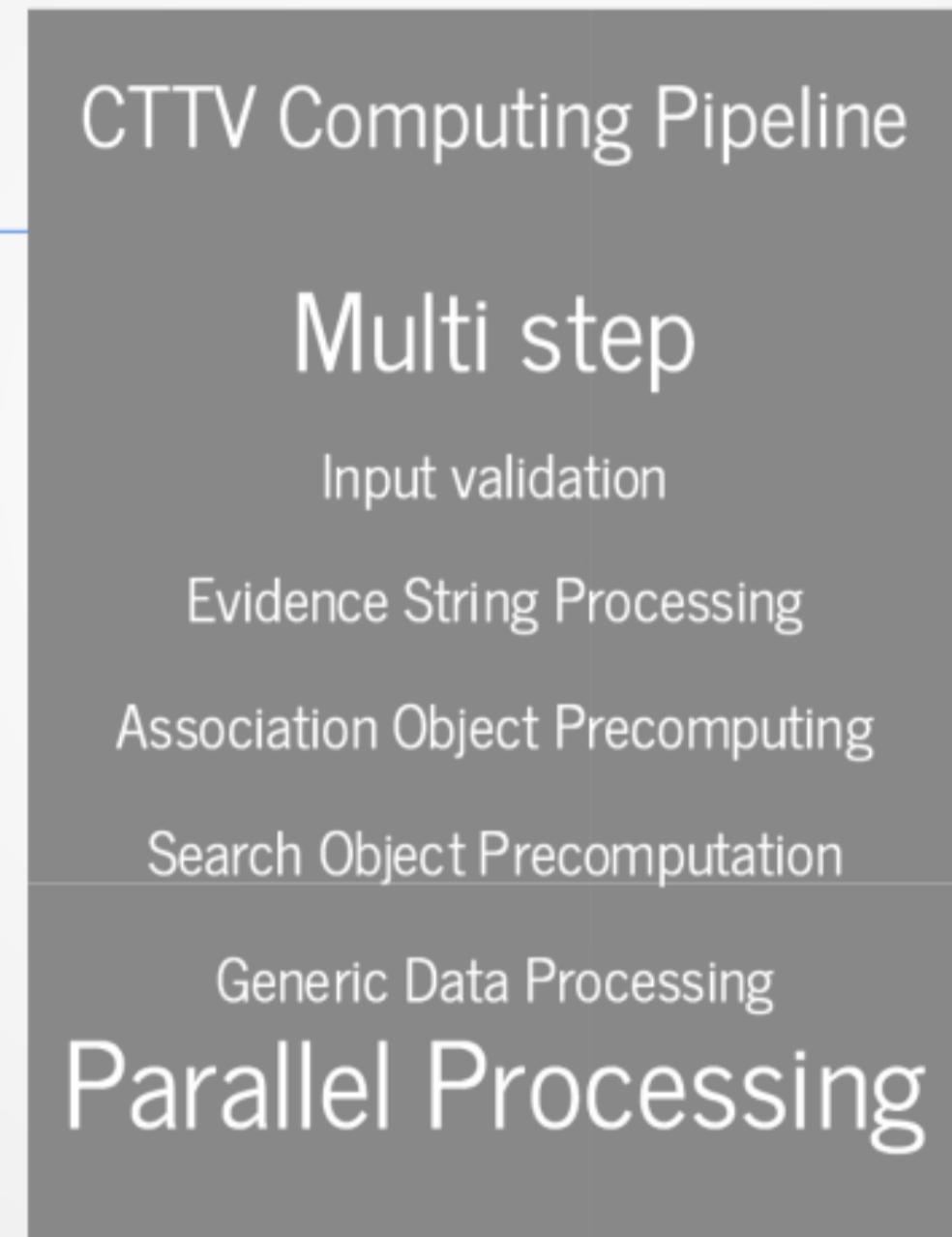


# Data Processing Computing Pipeline



Postgres 9.4

Association Precomputing step:  
2 days



Elasticsearch  
1.7.2

Association Precomputing step:  
2 hours



# Mining the Data

# CTTV REST API

CTTV REST API

Show/Hide | List Operations | Expand Operations | Raw

**default : CTTV REST API**

GET	/getbyid	Get an evidence from its id
POST	/getbyid	
GET	/filterby	Get a list of evidences filtered by gene, efo and/or eco codes
POST	/filterby	Get a list of evidences filtered by gene, efo and/or eco codes
GET	/association	Get association objects
POST	/association	get a list of association objects by target and/or disease .
GET	/disease/{code}	get EFO information from a code
GET	/eco/{code}	get ECO information from a code
GET	/target/{target_id}	Get gene information
GET	/expression	Get expression data for a gene
POST	/expression	Get expression data for a gene
GET	/search	Search for gene and disease
GET	/quicksearch	Suggest best terms for gene and disease
GET	/autocomplete	Suggest best terms for gene and disease
GET	/auth/request_token	

[ BASE URL: [http://localhost:8088/api/latest/cttv/\\_resource\\_list.json](http://localhost:8088/api/latest/cttv/_resource_list.json) , API VERSION: 1.0 ]

Not public yet (coming soon...)

# REST API Implementation

Enable to search for target and disease by full text search  
(more options coming)

Most methods are thin wrappers around elasticsearch filters

Returns aggregations for associations to enable faceted search

Auth built in

# Full Text Search

⚠ WARNING ⚠

Demo Time

# CTTV Web APP

[www.targetvalidation.org](http://www.targetvalidation.org)

# Full Text Search

# Full Text Search

I am interested in...  
Which diseases are modulated by D?

CTTV  
Centre for Therapeutic Target Validation

braf

**BRAF target**  
B-Raf proto-oncogene, serine/threonine kinase  
Protein kinase involved in the transduction of mitogenic signals from the cell membrane to the nucleus. May play a role in the postsynaptic responses of hippocampal neuron. Phosphorylates MAP2K1, and thereby contributes to the MAP kinase signal transduction pathway.

Targets

BRAFP1 BRAF pseudogene 1  
HMG20B high mobility group 20B  
ARAF A-Raf proto-oncogene, serine/threonine kinase

disease D:  
modulated to D?

# Full Text Search

# Full Text Search

- Support autocomplete

# Full Text Search

- Support autocompletion
- Disambiguate between different document types
  - interleukin
    - TARGET: interleukin-1, interleukin-2, ...
    - DISEASE: Immunodeficiency due to interleukin-1 receptor-associated kinase-4 deficiency

# Full Text Search

- Support autocompletion
- Disambiguate between different document types
  - interleukin
    - TARGET: interleukin-1, interleukin-2, ...
    - DISEASE: Immunodeficiency due to interleukin-1 receptor-associated kinase-4 deficiency
- Support for acronyms and many, many target synonyms
  - query: "SLE"
    - SLE<sup>ep?</sup> (EFO term)
    - SLEED1? (Target)
    - **systemic lupus erythematosus** (what they were actually looking for)

# Full Text Search

- Support autocompletion
- Disambiguate between different document types
  - interleukin
    - TARGET: interleukin-1, interleukin-2, ...
    - DISEASE: Immunodeficiency due to interleukin-1 receptor-associated kinase-4 deficiency
- Support for acronyms and many, many target synonyms
  - query: "SLE"
    - SLE<sup>ep?</sup> (EFO term)
    - SLEED1? (Target)
    - **systemic lupus erythematosus** (what they were actually looking for)
- Data driven search score
  - Increase with the number of associations linked
  - Decrease if down in the ontology tree

# Full Text Search

- Support autocompletion
- Disambiguate between different document types
  - interleukin
    - TARGET: interleukin-1, interleukin-2, ...
    - DISEASE: Immunodeficiency due to interleukin-1 receptor-associated kinase-4 deficiency
- Support for acronyms and many, many target synonyms
  - query: "SLE"
    - SLE<sup>ep?</sup> (EFO term)
    - SLEED1? (Target)
    - **systemic lupus erythematosus** (what they were actually looking for)
- Data driven search score
  - Increase with the number of associations linked
  - Decrease if down in the ontology tree
- Track user behaviour?

REST API powers a search driven web app with rich visualisation

# REST API powers a search driven web app with rich visualisation

The screenshot shows a search interface for the Centre for Therapeutic Target Validation (CTTV) website. A search bar at the top contains the query "braf". Below the search bar, the results are displayed:

- BRAF target**:  
B-Raf proto-oncogene, serine/threonine kinase  
Protein kinase involved in the transduction of mitogenic signals from the cell membrane to the nucleus. May play a role in the postsynaptic responses of hippocampal neurons. Phosphorylates MAP2K1, and thereby contributes to the MAP kinase signal transduction pathway.
- Targets**:  
BRAFP1 BRAF pseudogene 1  
HMG20B high mobility group 20B  
ARAF A-Raf proto-oncogene, serine/threonine kinase

On the left side of the results, there are two dropdown menus:

- "I am interested in" dropdown, currently showing "Which disease modulates BRAF?"
- "Which disease modulates BRAF?" dropdown, currently showing "modulated to disease D?"

# REST API powers a search driven web app with rich visualisation



# REST API powers a search driven web app with rich visualisation

**CTTV**  
Centre for Therapeutic Target Validation

braf

BRAF target

B-Raf proto-oncogene, serine/threonine kinase

Protein kinase involved in the transduction of mitogenic signals from the cell membrane to the nucleus. May play a role in the postsynaptic responses of hippocampal neurons. Phosphorylates MAP2K1, and thereby contributes to the MAP kinase signal transduction pathway.

Targets

BRAFP1 BRAF pseudogene 1

HMG20B high mobility group 20B

ARAF A-Raf proto-oncogene, serine/threonine kinase

disease D:  
modulated to  
D?

Filter by

Data types

- Clear all  Select all
- Genetic associations (119)
- Somatic mutations (120)
  - Cancer Gene Census (118)
  - European Variation Archive ... (21)
- Drugs (48)
- Affected pathways (2)
- RNA expression (27)
- Expression Atlas (27)
- Text mining (260)
- Animal models (48)

Bubbles Table Tree

All

Showing 1 to 50 of 340 entries

Search:

Therapeutic areas

Neoplasm (212)

Genetic disorder (60)

Hematological system disease (37)

Skin disease (36)

Endocrine system disease (33)

Digestive system disease (22)

Other (25)

Phenotype (23)

Nervous system disease (22)

Cardiovascular disease (22)

Chromosomes

Autosomal genes

Chromosomal alterations

Chromosomal mutations

Drugs

Affected pathways

RNA expression

Text mining

Animal models

Therapeutic area

Chromosomes	Autosomal genes	Chromosomal alterations	Chromosomal mutations	Drugs	Affected pathways	RNA expression	Text mining	Animal models	Therapeutic area
neoplasm									neoplasm
cancer									neoplasm
lung disease									respiratory system disease
lung carcinoma									neoplasm, respiratory system di...
skin disease									hematological system disease, ...
lymphoid neoplasm									neoplasm
adenocarcinoma									neoplasm, respiratory system di...
non-small cell lung carcinoma									hematological system disease, ...
genetic disorder									neoplasm
lymphoma									hematological system disease, ...
melanoma									skin disease, neoplasm
sarcoma									neoplasm
colonic neoplasm									neoplasm, digestive system di...
colorectal adenocarcinoma									digestive system disease, neopl...
kidney neoplasm									neoplasm
neoplasm of mature B-cells									hematological system disease, ...
Cardiofaciocutaneous syndrome									skin disease, genetic disorder, c...
neoplasms of mature B-cells									neoplasm
squamous cell carcinoma									neoplasm
thyroid disease									endocrine system disease
thyroid carcinoma									endocrine system disease, neop...
non-Hodgkin lymphoma									hematological system disease, ...
renal cell carcinoma									neoplasm
myeloid neoplasm									hematological system disease, ...
breast carcinoma									neoplasm
Noonan syndrome									reproductive system disease, sk...
brain neoplasm									nervous system disease, neopl...
plasma cell neoplasm									hematological system disease, ...

# REST API powers a search driven web app with rich visualisation

**CTTV**  
Centre for Therapeutic Target Validation

braf

BRAF target

B-Raf proto-oncogene, serine/threonine kinase

Protein kinase involved in the transduction of mitogenic signals from the cell membrane to the nucleus. May play a role in the postsynaptic responses of hippocampal neurons. Phosphorylates MAP2K1, and thereby contributes to the MAP kinase signal transduction pathway.

Targets

BRAFP1 BRAF pseudogene 1

HMG20B high mobility group 20B

ARAF A-Raf proto-oncogene, serine/threonine kinase

Filter by

Data types

- Clear all
- Select all
- Genetic associations (119)
- Somatic mutations (120)
  - Cancer Gene Census (118)
  - European Variation Archive ... (21)
- Drugs (48)
- Affected pathways (2)
- RNA expression (27)
- Expression Atlas (27)
- Text mining (260)
- Animal models (48)

Therapeutic areas

Neoplasm (212)

Genetic disorder (60)

Hematological system disease (37)

Skin disease (36)

Endocrine system disease (33)

Digestive system disease (22)

Other (25)

Phenotype (23)

Nervous system disease (22)

Cardiovascular disease (22)

Bubbles Table Tree

All

Showing 1 to 50 of 340 entries

Search:

Chromosomal Alterations score Chromosomal instability Drugs Affected pathways RNA expression Text mining Animal models Therapeutic area

Chromosomal Alterations score	Chromosomal Instability	Drugs	Affected Pathways	RNA Expression	Text Mining	Animal Models	Therapeutic Area
neoplasm							neoplasm
cancer							neoplasm
lung disease							respiratory system disease
lung carcinoma							neoplasm, respiratory system di...
skin disease							hematological system disease, ...
lymphoid neoplasm							neoplasm
adenocarcinoma							neoplasm, respiratory system di...
non-small cell lung carcinoma							neoplasm, respiratory system di...
genetic disorder							hematological system disease, ...
lymphoma							skin disease, neoplasm
melanoma							neoplasm
sarcoma							neoplasm, digestive system di...
colonic neoplasm							digestive system disease, neopl...
colorectal adenocarcinoma							neoplasm
kidney neoplasm							hematological system disease, ...
neoplasm of mature B-cells							skin disease, genetic disorder, c...
Cardiofaciocutaneous syndrome							neoplasm
neoplasms of mature B-cells							neoplasm
squamous cell carcinoma							endocrine system disease
thyroid disease							endocrine system disease, neop...
thyroid carcinoma							hematological system disease, ...
non-Hodgkin lymphoma							neoplasm
renal cell carcinoma							hematological system disease, ...
myeloid neoplasm							neoplasm
breast carcinoma							hematological system disease, ...
Noonan syndrome							neoplasm
brain neoplasm							reproductive system disease, sk...
plasma cell neoplasm							nervous system disease, neopl...
							hematological system disease, ...

Diagram illustrating the relationship between cancer types and their subtypes:

```
graph TD; cancer[cancer] --> colonColon[colon cancer]; cancer --> sarcoma[sarcoma]; cancer --> msvg[motheaten sweat gland cancer]; cancer --> lhd[lymphoid neoplasm]; cancer --> tc[tumour of cranial and spinal nerves]; cancer --> ocv[oral cavity cancer]; cancer --> meso[mesothelioma]; cancer --> pharynx[pharyngeal cancer]; cancer --> gum[gum cancer]; lhd --> ocv; lhd --> meso; lhd --> pharynx; lhd --> tc;
```

# REST API powers a search driven web app with rich visualisation

**CTTV**  
Centre for Therapeutic Target Validation

**braf**

**BRAF target**  
B-Raf proto-oncogene, serine/threonine kinase  
Protein kinase involved in the transduction of mitogenic signals from the cell membrane to the nucleus. May play a role in the postsynaptic responses of hippocampal neurons. Phosphorylates MAP2K1, and thereby contributes to the MAP kinase signal transduction pathway.

**Targets**  
BRAFP1 BRAF pseudogene 1  
HMG20B high mobility group 20B  
ARAF A-Raf proto-oncogene, serine/threonine kinase

**Filter by**

**Data types**

- Clear all  Select all
- Genetic associations (119)
- Somatic mutations (120)
  - Cancer Gene Census (118)
  - European Variation Archive ... (21)
- Drugs (48)
- Affected pathways (2)
- RNA expression (27)
- Expression Atlas (27)
- Text mining (260)
- Animal models (48)

**Therapeutic areas**

Neoplasm (212)  
Genetic disorder (60)  
Hematological system disease (37)  
Skin disease (36)  
Endocrine system disease (33)  
Digestive system disease (22)  
Other (25)  
Phenotype (23)  
Nervous system disease (22)  
Cardiovascular disease (22)

**Bubbles** **Table** **Tree**

All

Showing 1 to 50 of 340 entries

Search:

**Evidence for BRAF in cancer**

**BRAF**  
B-Raf proto-oncogene, serine/threonine kinase  
Synonyms: BRAF1, RAF1  
Protein kinase involved in the transduction of mitogenic signals from the cell membrane to the nucleus. May play a role in the postsynaptic responses of hippocampal neurons. Phosphorylates MAP2K1, and thereby contributes to the MAP kinase signal transduction pathway.

**cancer**  
Synonyms: Malignant neoplasm, malignant neoplasia, malignant tumor, malignant tumour  
A malignant neoplasm in which new abnormal tissue grows by uncontrolled division and proliferates more rapidly than normal and continues to grow after the stimuli that initiated the new growth ce...

**Genetic associations**

Table Browser

Human Chr: 7

human:7:140678239-140965851

Variants in common diseases

Variants in rare diseases

sequence

Genes / Transcripts

# Architecture

# ES Cluster



- Running in the EBI Embassy cloud (VMWare)
- CoreOS + Docker
- 3 Nodes
  - Production: 4 CPU / 12 Gb RAM / 40 Gb SSD
  - Staging: 4 CPU / 8 Gb RAM / 100 Gb spinning disk

# Architecture evolution

- Migrate from 1.7.2 to 2.1
- Drop production/staging paradigm and move to one big cluster
  - hot and cold nodes
  - tagged version of the data
  - support multiple version of data / api / webapp within the same infrastructure

# Summary

## Elasticsearch at CTTV

# Elasticsearch at CTV

Flexible, can easily handle many types of data

Main document datastore

Backend for computational pipeline

Powering REST API and WEB APP

Potential to run on demand map reduce computation



IT WORKS!

# Acknowledgements

## Core platform development team



The image displays eight headshots of individuals arranged in two rows of four. The top row contains Andrea Pierleoni, Gautier Koscielny, Ian Dunham, and Jessica Vamathevan. The bottom row contains Luca Fumis, Mick Maguire, Miguel Pignatelli, and Niki Karamanis. Samiul Hasan's headshot is present but not visible in the image.

Andrea Pierleoni	Gautier Koscielny	Ian Dunham	Jessica Vamathevan
Luca Fumis	Mick Maguire	Miguel Pignatelli	Niki Karamanis
Samiul Hasan			

?