Clarin:el meets Social Sciences: New Frontiers of collaboration

Haris Papageorgiou

ATHENA RC

Computational Social Science

➢ **Computational social science** means that computers are used to model, simulate, and analyze social phenomena. It refers to the academic sub-disciplines concerned with computational approaches to the social sciences.

➢ Computational social science, revolutionizes both fundamental legs of the scientific method: **empirical research**, especially through big data, by analyzing the digital footprint left behind through social online activities; and **scientific theory**, especially through computer simulation model building through social simulation.

Text Analytics

➢ The term text analytics describes a set of **linguistic statistical, and machine learning techniques** that model and structure the information content of textual sources. (Wikipedia)

➢ In other words, enhancing the value of text content by extracting entities, features, context, relationships and emotion.
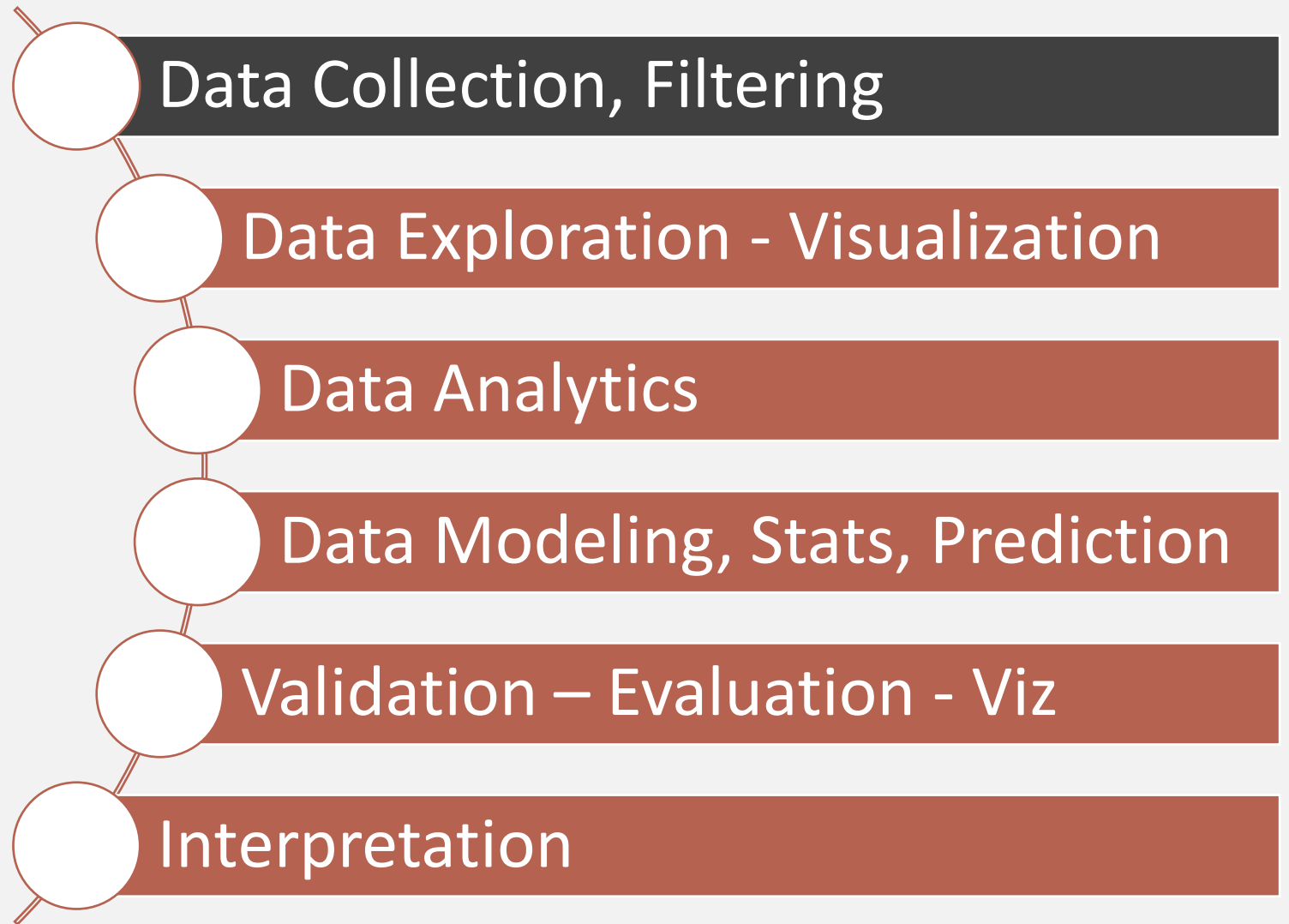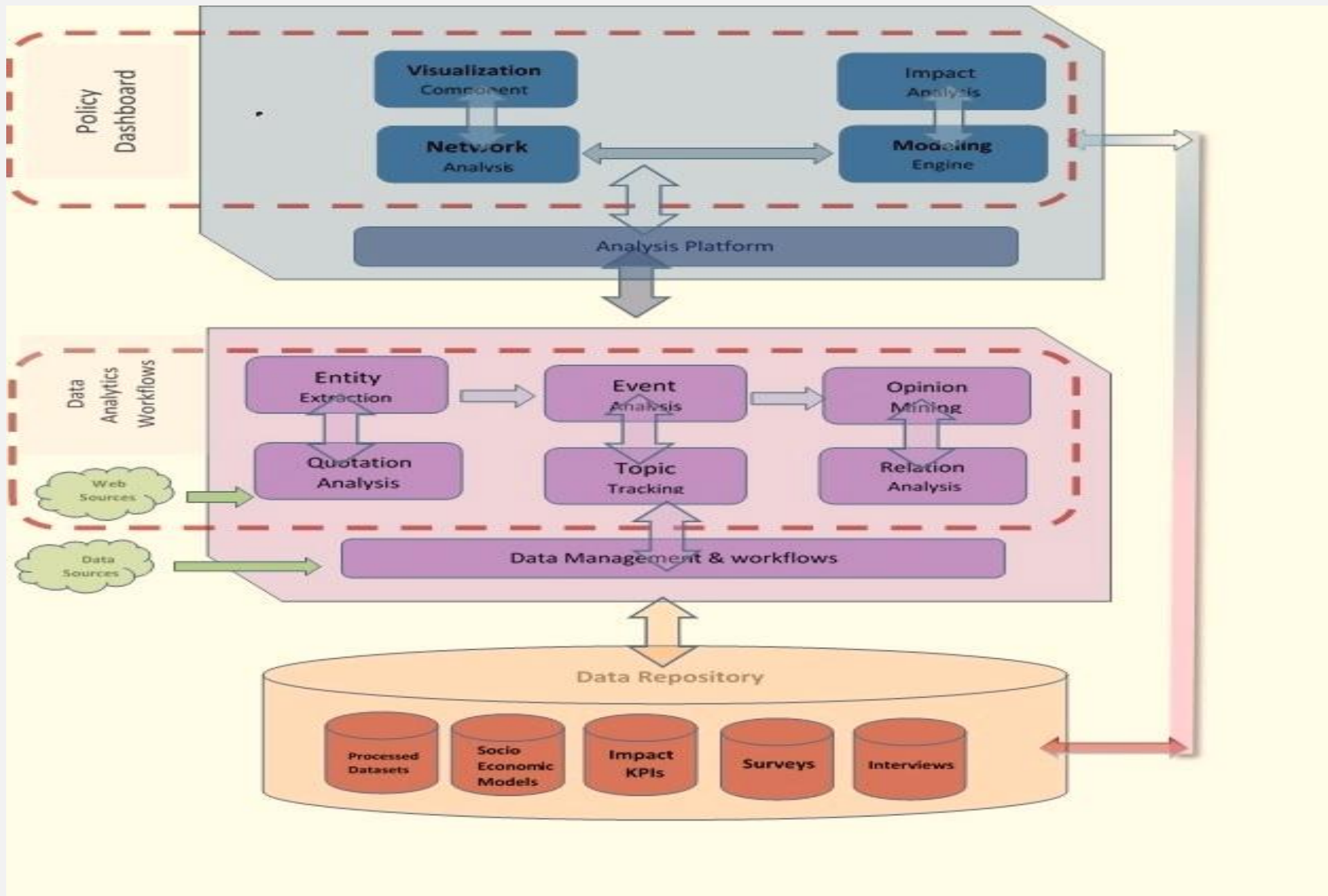
Goals

➢ Predict the outcome of elections, referendum (BREXIT) etc.

➢ **Examine xenophobia during the economic and migration crisis**

➢ Correlate the political points of view and perspectives with the financial crisis

➢ Predict conflict regions

➢ **Describe & Analyze the evolution and dynamics of social movements**

➢ Causal relationships between disasters and job offers

➢ Correlate radicalization with verbal aggressiveness on religious-related events

➢ Explain collective behaviour

➢ Test and Validate hypotheses etc. ………….

The Big three Ingredients
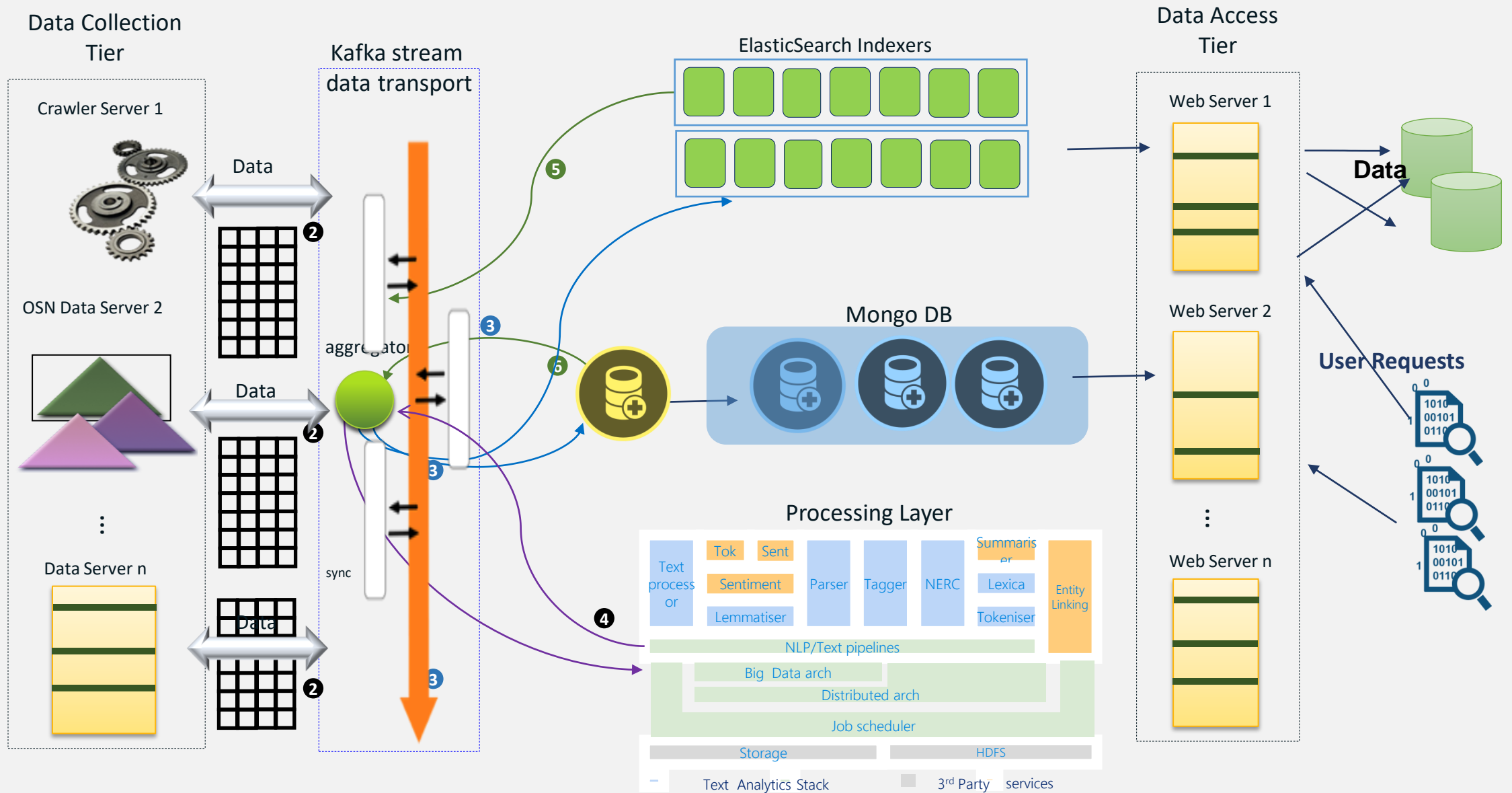
- M e t h o d o l o g y
- Architecture
- Infrastructure

PALOMAR Architecture

Analytics operation workflow

clarin:el

browse resources   community   statistics   help   about

clarin:el inventory
of language resources and services

na ne ni no na
ba be bi bo ba
sha she sho

athena repository   go!

203 language resources at your disposal

Type in your keywords, please...   search

clarin:el is the Greek national network of language resources, a nation-wide Research Infrastructure devoted to the sustainable storage, sharing, dissemination and preservation of language resources.

The Central Aggregator is the central Repository of the clarin:el Infrastructure, which is responsible for the harvesting of metadata from the local Repositories, the organisation and the presentation of the metadata descriptions in a uniform catalogue and the provision of access to the Language Resources to the network members and to the public.

Ερευνητικό Κέντρο Αθηνά   grnet

SoDaMap[1],
Xenophobia[2]

- Data Collection, Filtering
- Data Exploration - Visualization
- Data Analytics
- Data Modeling, Stats, Prediction
- Validation – Evaluation - Viz
- Interpretation
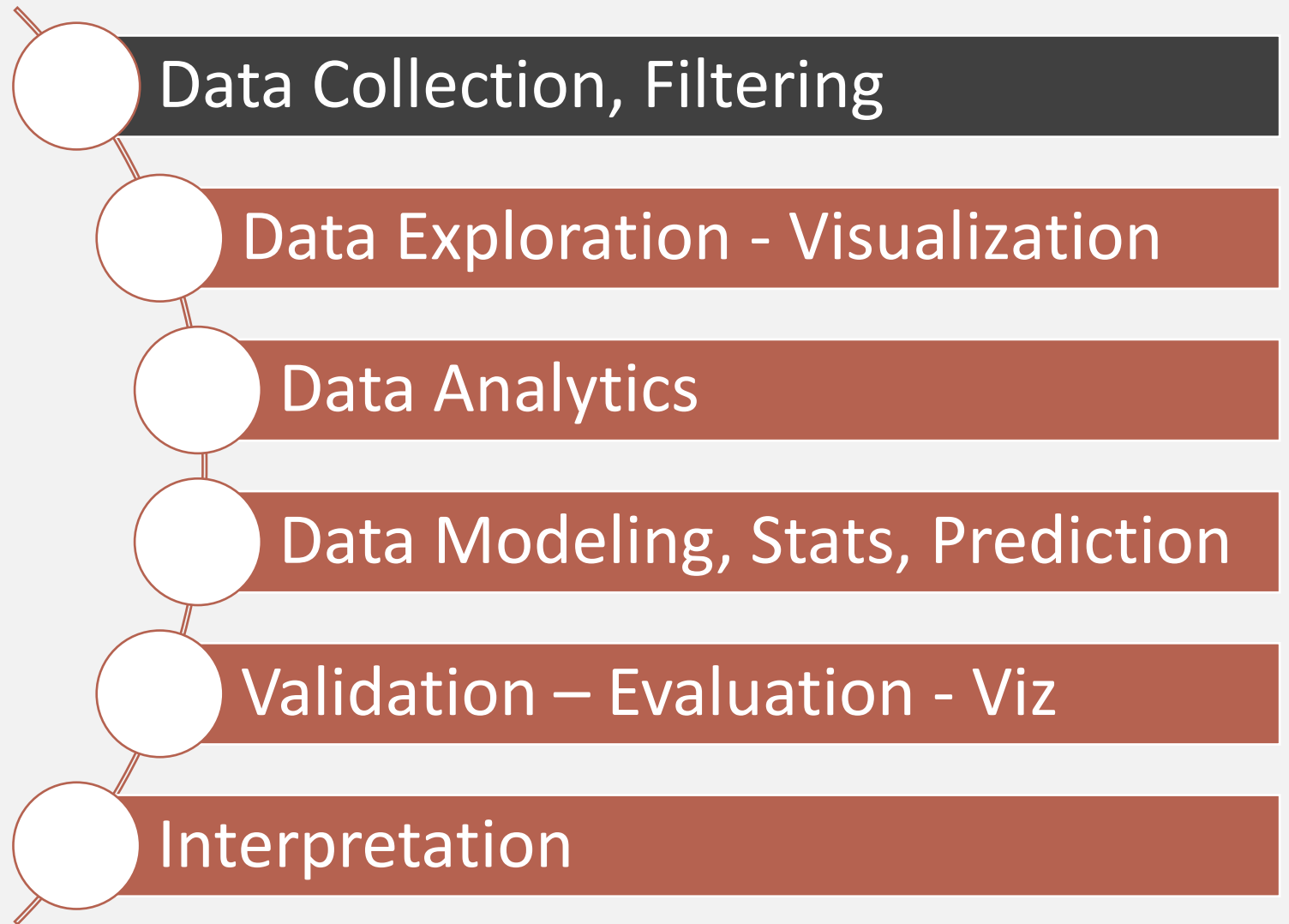
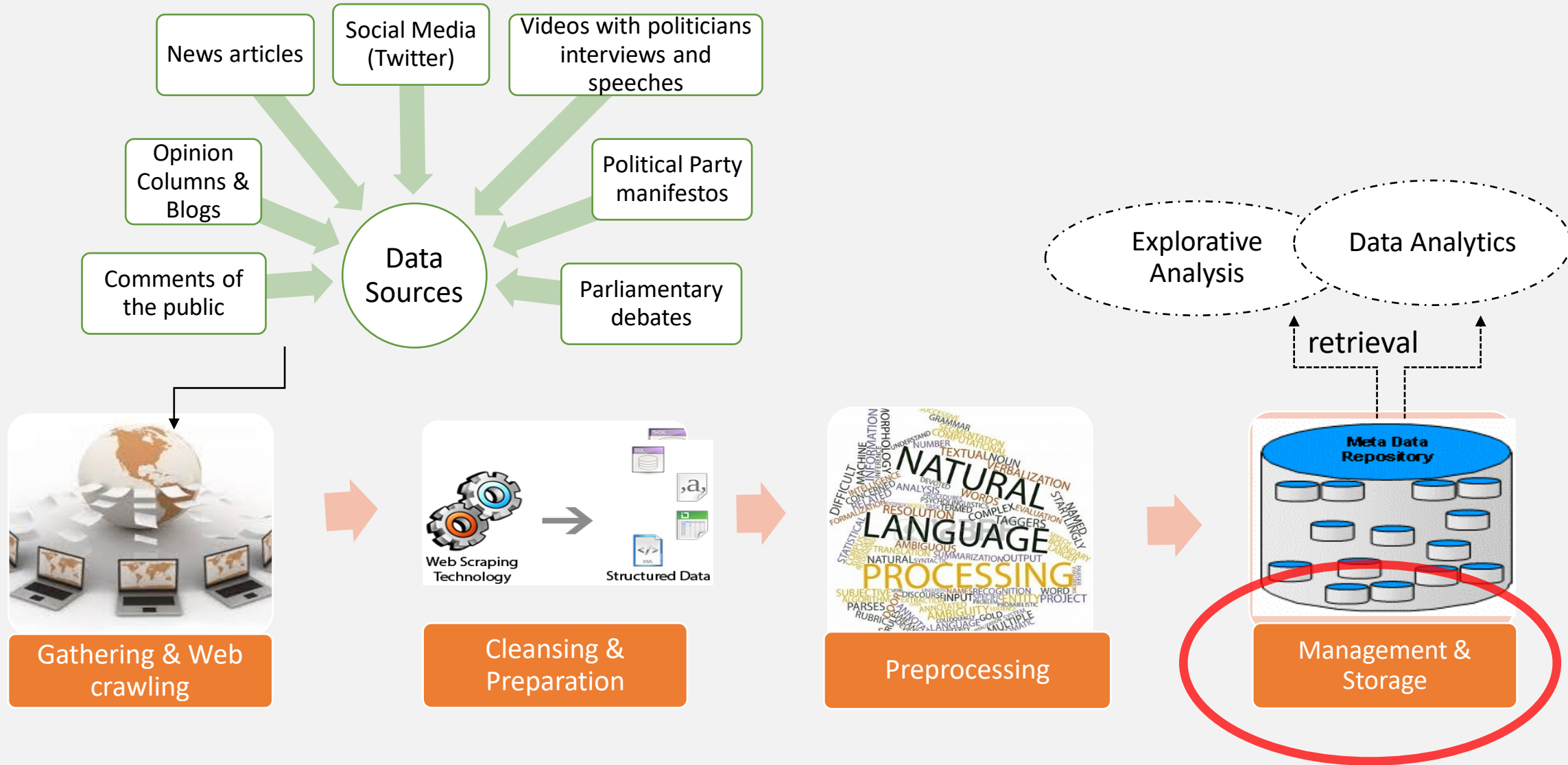1. The Sodamap Project: a longitudinal study on social movements, in collaboration with Research Director Theoni Stathopoulou at National Center for Social Research-NCSR.
2. The Xenophobia Project: examinining xenophobia during the economic and migration crisis in Greece, in collaboration with Professor of Political Science, Vassiliki Georgiadou, at the Department of Political Science and History of the Panteion University.

Research Questions

- Who are the **agents** of this particular social phenomenon (social movement, xenophobia, etc)?

- Who are the **targets** ….?

- What kind of **events** are related to the phenomenon?

- Which **sentiments** could be indicators of collective behavior manifestations (stereotypes and prejudices correlated with Xenophobia)?

Data Collection: Overview

News articles

Social Media (Twitter)

Videos with politicians interviews and speeches

Opinion Columns & Blogs

Comments of the public

Data Sources

Political Party manifestos

Parliamentary debates

Explorative Analysis

Data Analytics

retrieval

Meta Data Repository

Gathering & Web crawling

Web Scraping Technology → Structured Data

Cleansing & Preparation

NATURAL LANGUAGE PROCESSING

Preprocessing

Management & Storage

Data Collections - News

|	Total	Physical Attacks	Protests	Time Period
Avgi	792.715	152.244	60.745	1996 – 2015
Kathimerini	282.621	66.405	18.296	2002 – 2012
Eleftherotypia	429.364	81.961	23.921	2002 – 2014
Rizospastis	725.108	168.850	78.496	1995 – 2016
Ta Nea	330.190	72.713	17.578	1997 – 2007
In.gr	428.880	79.833	27.745	1999 – 2016
Naftemporiki	649.259	75.663	24.637	2000 – 2016

Data Collection

- Newspapers, ≈ 2.541.000 articles, time period 1996-2014

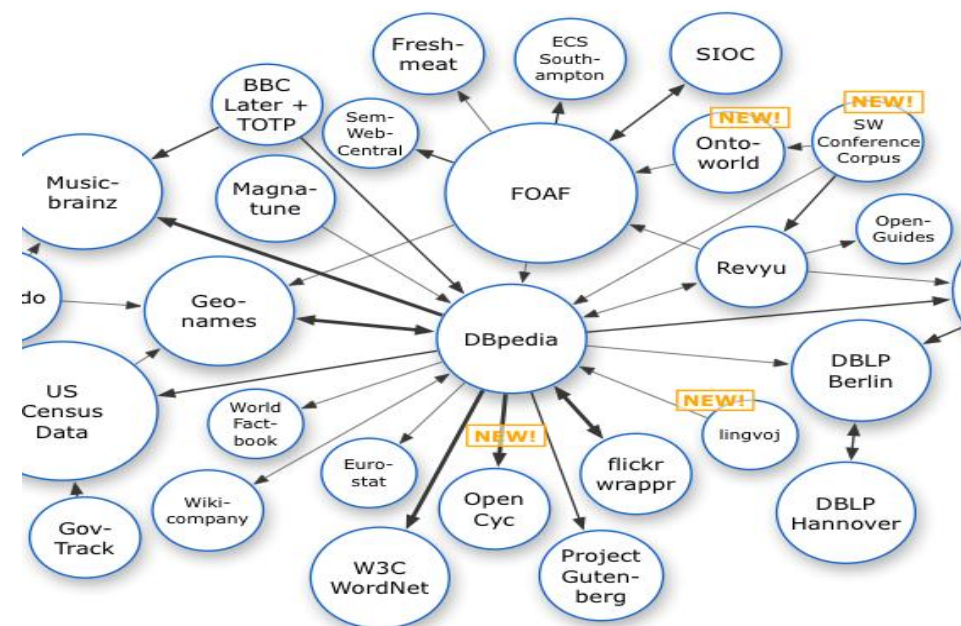- ≈ 564.000.000 tweets, time period 2012-2016

Data Exploration

QUERIES

SEARCH ENGINE

Data Collections

PALOMAR
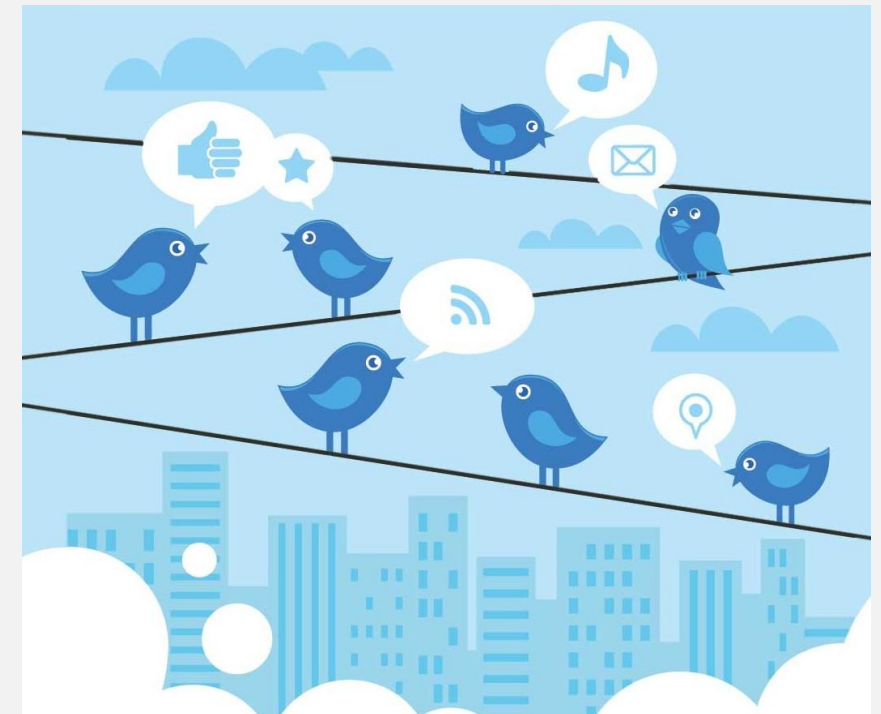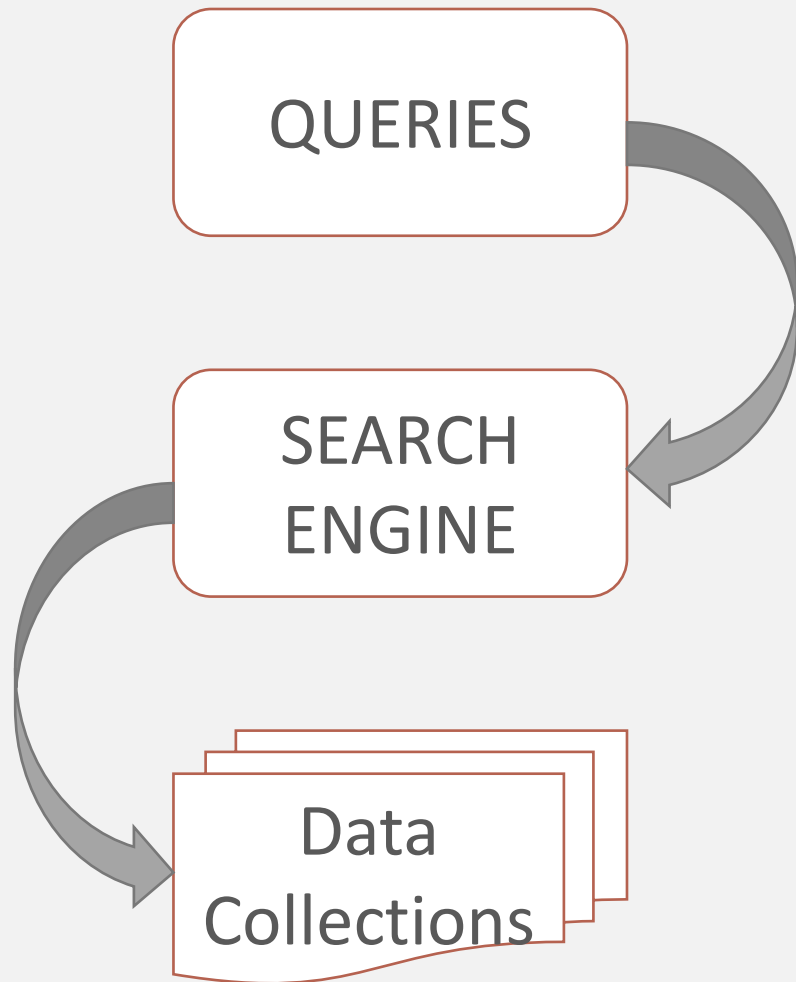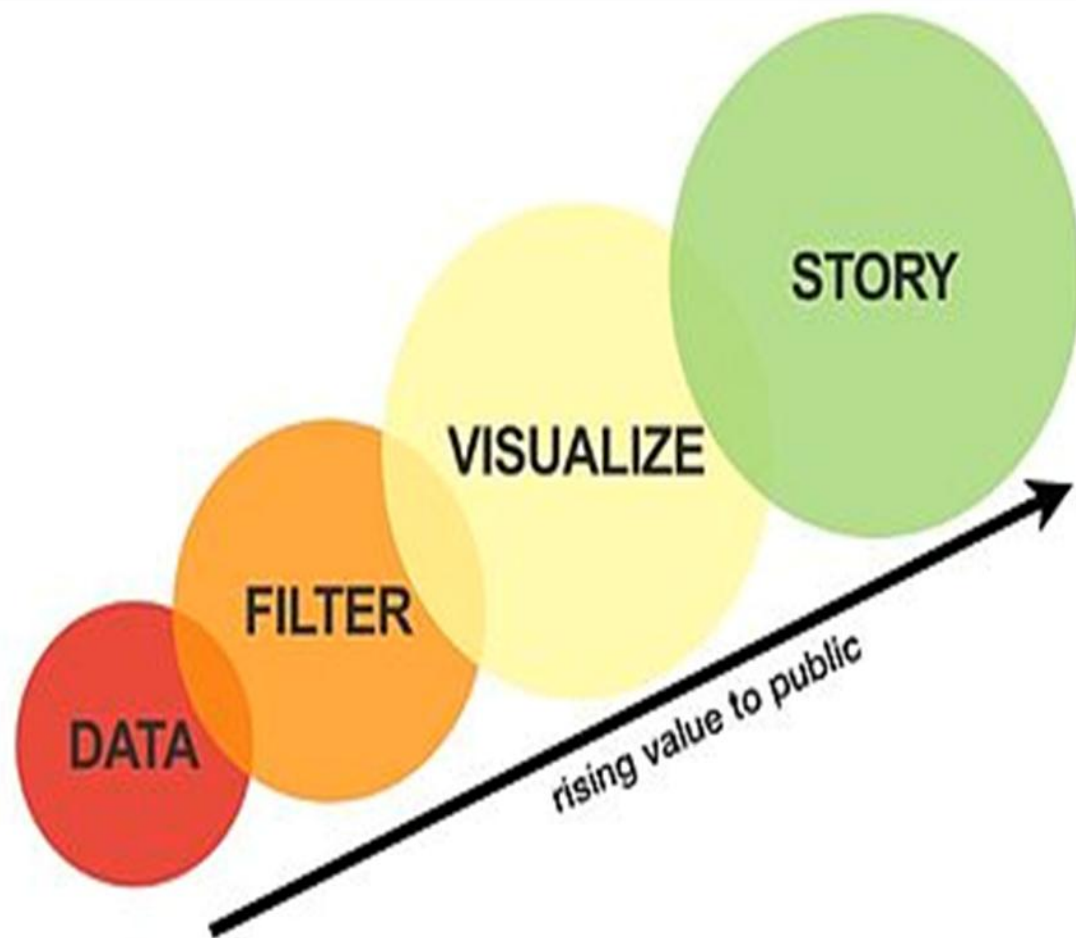
Quick

Relative

Absolute

Today	Yesterday	Last 15 minutes	Last 30 days
This week	Day before yesterday	Last 30 minutes	Last 60 days
This month	This day last week	Last 1 hour	Last 90 days
This year	Previous week	Last 4 hours	Last 6 months
The day so far	Previous month	Last 12 hours	Last 1 year
Week to date	Previous year	Last 24 hours	Last 2 years
Month to date		Last 7 days	Last 5 years
Year to date			

⌃

text: ισλάμ   🔍

xeno-data-twitter-1   65,660 hits

Selected Fields

September 5th 2011, 13:09:38.053 - September 5th 2016, 13:09:38.054 — by month

t   text

Available Fields   ⚙

Popular

⌖ @timestamp

⌖ created_at

t   in_reply_to_screen_name

in_reply_to_user_id

created_at per month

t   lang

t   source

user.id

t   user.lang

t   user.location

t   user.name

t   @version

t   _id

t   _index

_score

t   _type

? contributors

? coordinates
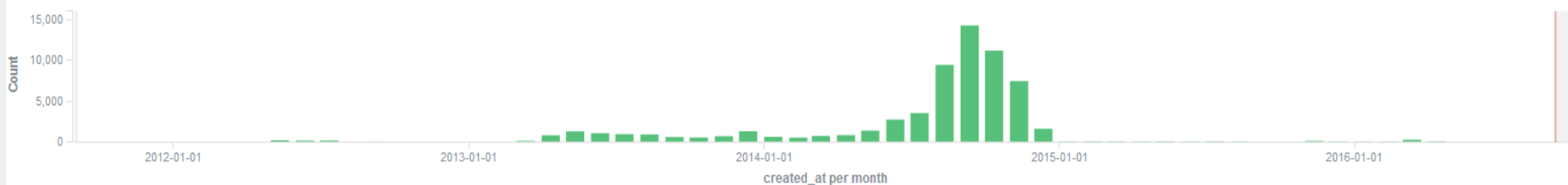
Time	text
July 15th 2016, 21:19:56.000	Η προφητεία του Καντάφι που τον δικαίωσε: Εάν πέσω, η Ευρώπη θα πέσει στα νύχια του Ισλάμ - https://t.co/P6aSmXgUO8
June 14th 2016, 14:32:37.000	RT @amna_news: ΕΚΤΑΚΤΟ: Συριακοί τηλεοπτικοί σταθμοί μεταδίδουν ότι σκοτώθηκε ο ηγέτης του Ισλαμικού Κράτους Αμπού Μπακρ αλ-Μπαγκντάντι #Συ...
June 13th 2016, 17:06:58.000	RT @Papavannesa: Δεν υπάρχει μετριοπαθές Ισλάμ δεν υπάρχει μετριοπαθής θρησκεία Η πιστεύεις τα πάντα η όχι
June 13th 2016, 02:14:41.000	Το Ισλαμικό Κράτος ανέλαβε την ευθύνη για το μακελειό στο Ορλάντο - Αναγνωρίστηκε ο δράστης (photos) https://t.co/M0jOEE4oIq
June 12th 2016, 19:08:09.000	RT @adiasistos: Ταξιδιωτική οδηγία του Ισλαμικού Κράτους προς τα μέλη του να αποφεύγουν τις πόλεις που γίνεται το γιούρο λόγω επικινδυνότητ...
June 10th 2016, 22:58:56.000	RT @tribunegr: 3 χλμ από το κέντρο της Φαλούτζα οι Ιρακινοί - Παγιδευμένο το Ισλαμικό Κράτος https://t.co/ac59wUuIqo https://t.co/8sSqQzg37Q
June 9th 2016, 18:07:32.000	RT @kanekos69: Το ισλαμικο κρατος ανελαβε την ευθυνη για την τρομοκρατικη επιθεση στον Μαυρικο, μπερδευτηκαν και χτυπησαν λαθος ακροπολη
June 8th 2016, 16:30:24.000	"Ανθέλληνας" και "μισέλληνας" γιατί γράφω στον πούτσο πρών καθολικά/ορθόδοξα/ ισλαμικά ντουβάρια. Σάλτα γαμήσου. https://t.co/x26TsdpGTM

Data Analytics

- Topics / Themes

- **Event Extraction Methodology**

- Verbal Aggressiveness /Opinion Mining

SCOPE

- Code specific event types

- Build computational tools for extracting event instances

- Record results to a database

- Visualize results

- Analyze outcome

- Correlate with impact KPIs

Claim vs Event

> A claim is defined as a **purposive unit of strategic or communicative action** in the public sphere aiming at a collective stake and having the form of physical or verbal action**.**

> **In NLP, an event** denotes a specific instance of an event type, e.g. a killing incident happening at a specific location, time and involving certain participants.

Claim

➢ A valid **claim** comprises the following components:

▪ The **Form (How**), namely the form of action through which the claim is introduced in the public sphere, or the realization of a protest event is located in a text.

▪ The **Actor (Who),** or the claimant, meaning the entity that is making the claim.

▪ The **Addressee (To Whom),** namely the entity to whom the action is addressed.

▪ The **Issue (What**) of a Claim, i.e. what the protest is about, the subject matter of a Claim.

▪ **The Location (Where),** viz where the event took place, the location of a claim in space.

▪ The **Time (When)** at which the event happened, or location of a claim in time.

Event Extraction Workflow

1 Codebook design

2 Tools design and development

3 Event Database population

Nationality Target
Age
Person Location
Candidate
Target Group
Time Actor Event

Data Exploration

Results - Evaluation

Events

Codebook

Data Analytics

Several Event Categories

1 Physical Violence Events

2 Protest Events

3 Quotations - Statements

EVENT TAXONOMY

Physical Attack

- Assault
  - Assault against life
  - Violent Assault
  - Sexual Assault
  - Verbal Attack

- Attack against Property
  - Attack against Personal Property
  - Attack against Religious Property

Protest

- Protest
  - Demonstration/March
    - Motorized March
    - Sit Down
  - Strike
  - Hunger Strike
  - Blockade
  - Occupation
  - Signature Collection
  - Boycott
  - Symbolic Violence
  - Revolt
  - Violent Demonstration

VARIOUS TYPES OF INFORMATION

- Medium
- Title
- Date (Day, Month, Year)
- Location (Category)
- Event (Event type)
- Confidence (Degree)
- Issue

ACTOR & TARGET ATTRIBUTES

- Target Group
- Nationality
- Category
- Age
- Sex
- Status

Protest Events Codebook

CLAIM=EVENT

⬇

<**Actor**, Form, Addressee, Issue, Loc, Time>

The **Law Society of Piraeus** has voted to occupy the Mortgage Registries of Piraeus and Salamis, on April 26th and 27th 2006, in protest against the serious operational problems it is facing.

Protest Events Codebook

CLAIM=EVENT

<Actor, **Form**, Addressee, Issue, Loc, Time>

The Law Society of Piraeus **has voted to occupy** the Mortgage Registries of Piraeus and Salamis, on April 26th and 27th 2006, in protest against the serious operational problems it is facing.

Protest Events Codebook

CLAIM=EVENT

<Actor, Form, **Addressee**, Issue, Loc, Time>

The Law Society of Piraeus has voted to occupy the **Mortgage Registries** of Piraeus and Salamis, on April 26th and 27th 2006, in protest against the serious operational problems it is facing.

Protest Events Codebook

CLAIM=EVENT

<Actor, Form, Addressee, Issue, **Loc**, Time>

The Law Society of Piraeus has voted to occupy the Mortgage Registries of **Piraeus** and **Salamis**, on April 26th and 27th 2006, in protest against the serious operational problems it is facing.

Protest Events Codebook

CLAIM=EVENT

<Actor, Form, Addressee, Issue, Loc, **Time**>

The Law Society of Piraeus has voted to occupy the Mortgage Registries of Piraeus and Salamis, on **April 26th and 27th 2006**, in protest against the serious operational problems it is facing.

Protest Events Codebook

CLAIM=EVENT

<Actor, Form, Addressee, **Issue**, Loc, Time>

The Law Society of Piraeus has voted to occupy the Mortgage Registries of Piraeus and Salamis, on April 26th and 27th 2006, in protest against the **serious operational problems it is facing**.

• Codebook

EVENT

⬇

<Time, **Location**, Target, Event, Actor>

On 28 March 2013, in the center of **Xanthi**, a 30-year-old Greek Muslim suffered a stub wound by members of Golden Dawn.
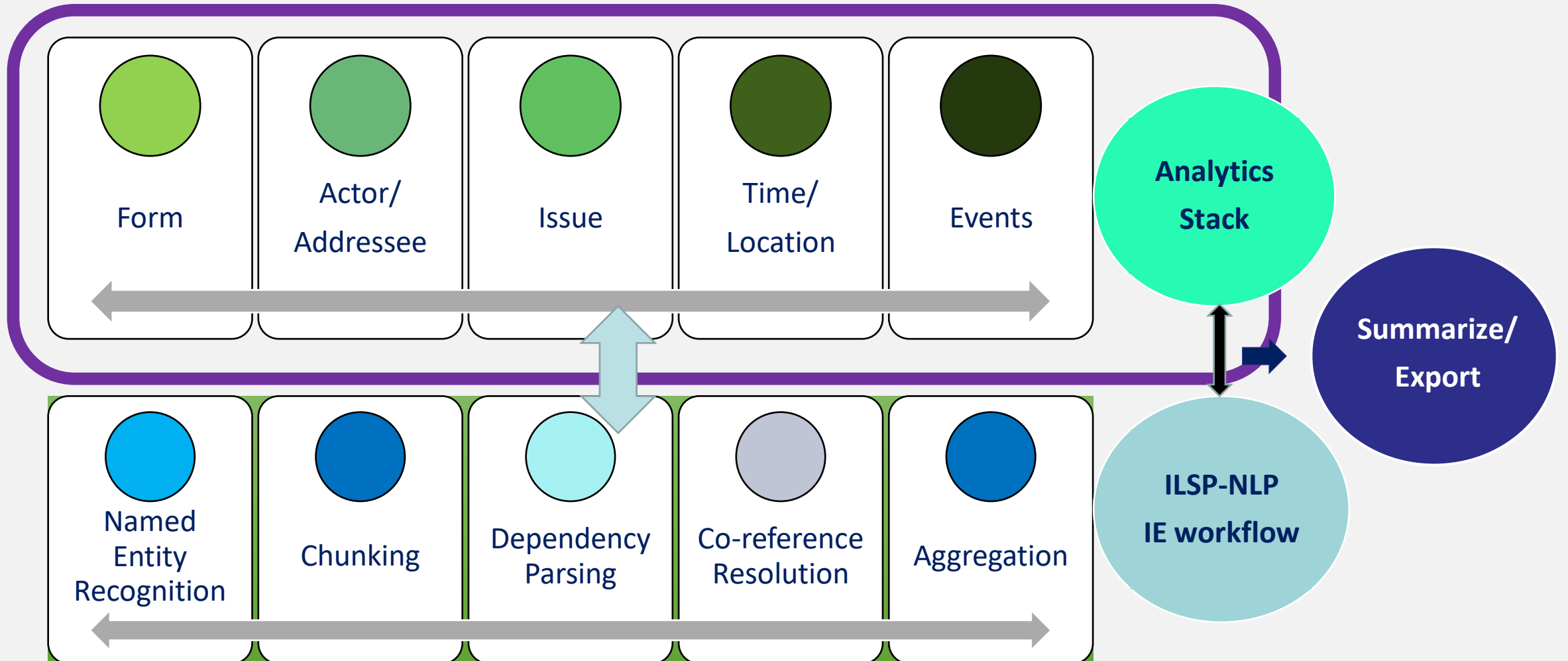
• Codebook

EVENT

⬇

<Time, Location, **Target**, Event, Actor>

On 28 March 2013, in the center of Xanthi, a **30-year-old Greek Muslim** suffered a stub wound by members of Golden Dawn.

- Codebook

EVENT

⬇

<Time, Location, Target, **Event**, Actor>

On 28 March 2013, in the center of Xanthi, a 30-year-old Greek Muslim **suffered a stub wound** by members of Golden Dawn.

- Codebook

EVENT

⬇

<Time, Location, Target, Event, **Actor**>

On 28 March 2013, in the center of Xanthi, a 30-year-old Greek Muslim suffered a stub wound by **members of Golden Dawn**.

Coding Quotations

QUOTATION=EVENT

<Actor, **Form**, Addressee, Issue, Loc, Time>

Recorded coverage of the **statement made** in the House of Commons by home secretary Theresa May on the 1989 Hillsborough disaster, from Wednesday 27 April.

Coding Quotations

QUOTATION=EVENT

<Actor, Form, **Addressee**, Issue, Loc, Time>

Recorded coverage of the statement made in the **House of Commons** by home secretary Theresa May on the 1989 Hillsborough disaster, from Wednesday 27 April.

Coding Quotations

QUOTATION=EVENT

<**Actor**, Form, Addressee, Issue, Loc, Time>

Recorded coverage of the statement made in the House of Commons by **home secretary Theresa May** on the 1989 Hillsborough disaster, from Wednesday 27 April.

Coding Quotations

QUOTATION=EVENT

<Actor, Form, Addressee, **Issue**, Loc, Time>

Recorded coverage of the statement made in the House of Commons by home secretary Theresa May on the **1989 Hillsborough disaster**, from Wednesday 27 April.

Coding Quotations

QUOTATION=EVENT

<Actor, Form, Addressee, Issue, Loc, **Time**>

Recorded coverage of the statement made in the House of Commons by home secretary Theresa May on the 1989 Hillsborough disaster, from **Wednesday 27 April**.

Coding Quotations

QUOTATION=EVENT

<**Actor**, Form, Addressee, Issue, Time, Loc>

"Of course a rise in tension and terrorist activity cannot but heighten concern. It is further proof of how fragile the situation is in Syria and demonstrates the necessity to continue active steps towards resuming talks," **Kremlin spokesman Dmitry Peskov** told reporters yesterday in Moscow.
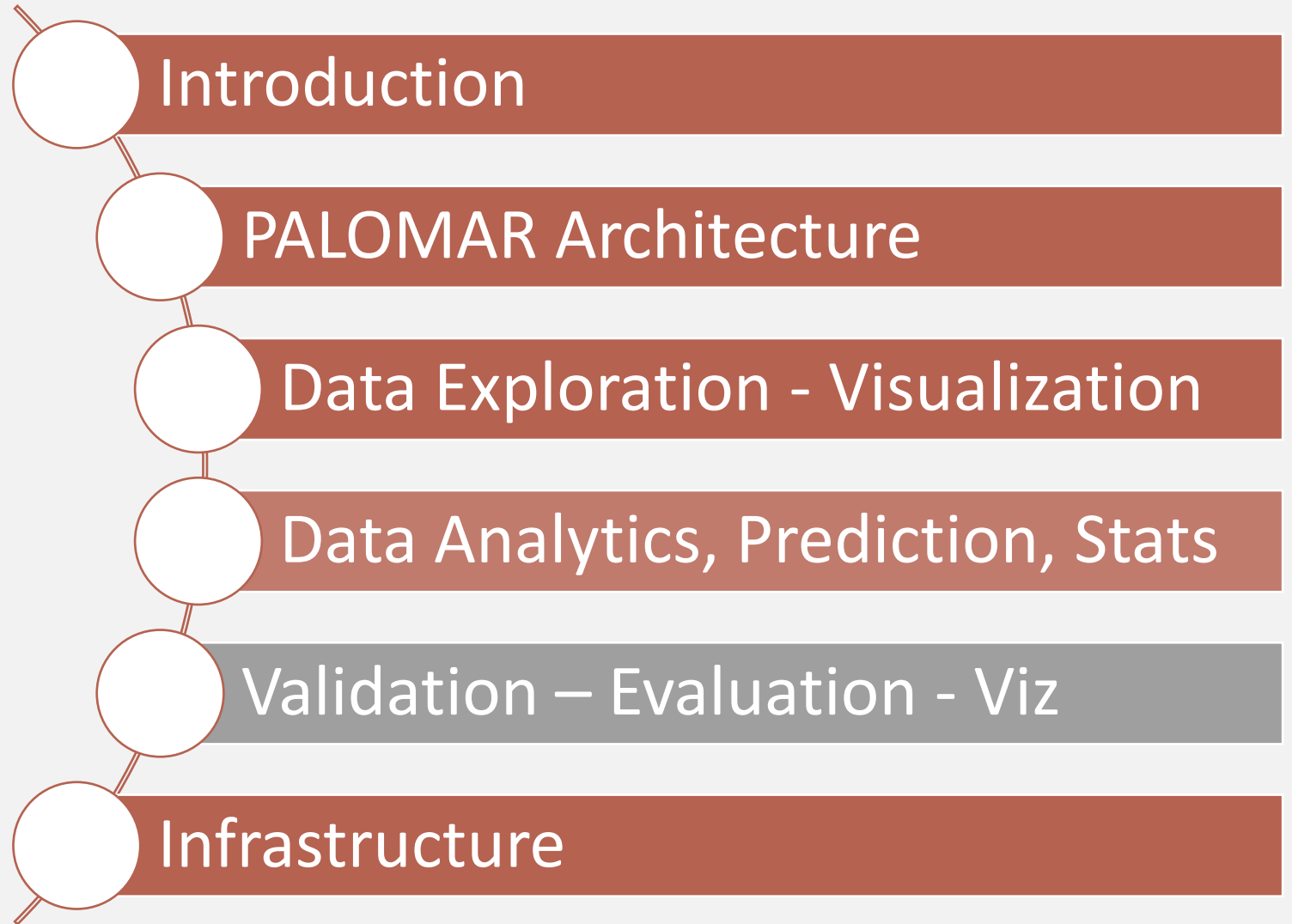
Coding Quotations

QUOTATION=EVENT

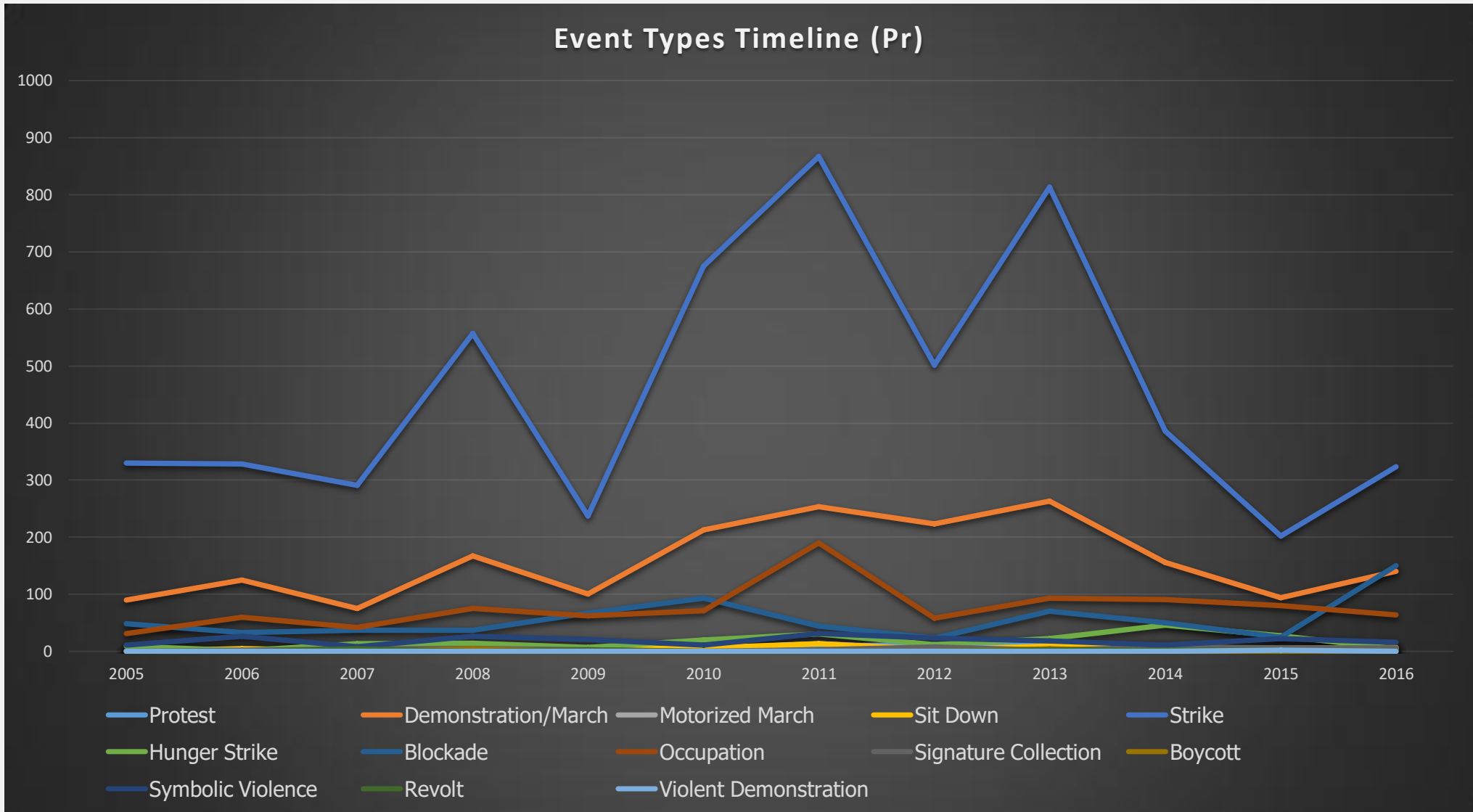<Actor, **Form**, Addressee, Issue, Time, Loc>

"Of course a rise in tension and terrorist activity cannot but heighten concern. It is further proof of how fragile the situation is in Syria and demonstrates the necessity to continue active steps towards resuming talks," Kremlin spokesman Dmitry Peskov **told** reporters yesterday in Moscow.

Coding Quotations

QUOTATION=EVENT

<Actor, Form, **Addressee**, Issue, Time, Loc>

"Of course a rise in tension and terrorist activity cannot but heighten concern. It is further proof of how fragile the situation is in Syria and demonstrates the necessity to continue active steps towards resuming talks," Kremlin spokesman Dmitry Peskov told **reporters** yesterday in Moscow.
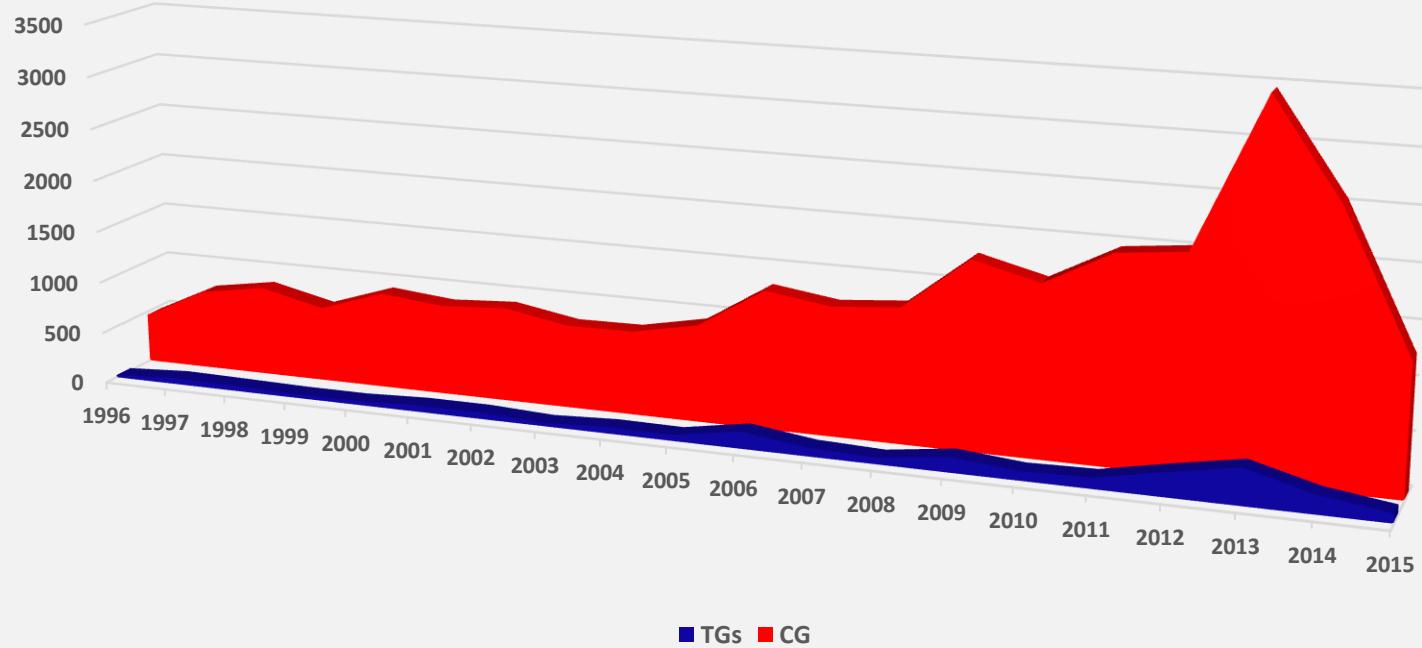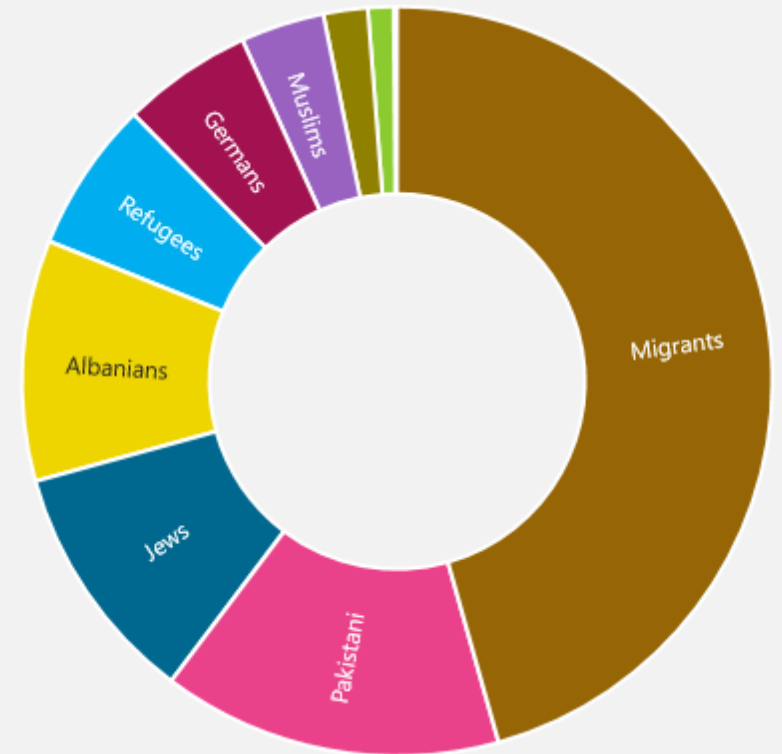
Coding Quotations

QUOTATION=EVENT

<Actor, Form, Addressee, **Issue**, Time, Loc>

"**Of course a rise in tension and terrorist activity cannot but heighten concern. It is further proof of how fragile the situation is in Syria and demonstrates the necessity to continue active steps towards resuming talks**," Kremlin spokesman Dmitry Peskov told reporters yesterday in Moscow.

Coding Quotations

QUOTATION=EVENT

<Actor, Form, Addressee, Issue, **Time**, Loc>

"Of course a rise in tension and terrorist activity cannot but heighten concern. It is further proof of how fragile the situation is in Syria and demonstrates the necessity to continue active steps towards resuming talks," Kremlin spokesman Dmitry Peskov told reporters **yesterday** in Moscow.

Coding Quotations

QUOTATION=EVENT

<Actor, Form, Addressee, Issue, Time, **Loc**>

"Of course a rise in tension and terrorist activity cannot but heighten concern. It is further proof of how fragile the situation is in Syria and demonstrates the necessity to continue active steps towards resuming talks," Kremlin spokesman Dmitry Peskov told reporters yesterday in **Moscow**.

Outline

- Introduction
- PALOMAR Architecture
- Data Exploration - Visualization
- Data Analytics, Prediction, Stats
- Validation – Evaluation - Viz
- Infrastructure

Visualizations

Event Types Timeline (Pr)

•Visualizations

PA Timeline TGs vs CG

PA against TGs

Visualizations

TG:1
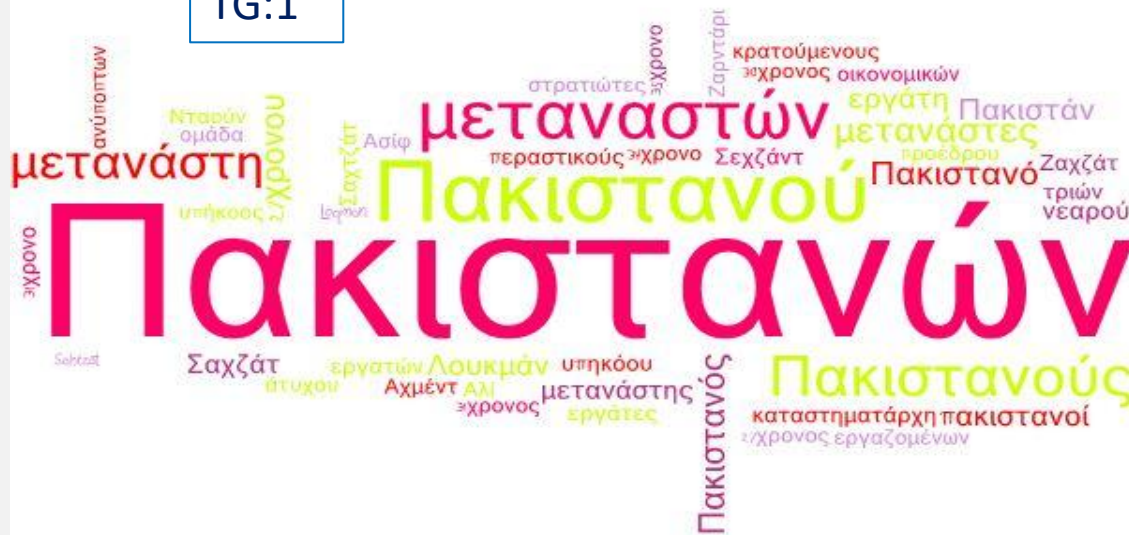
Events

Actors

Event Analysis
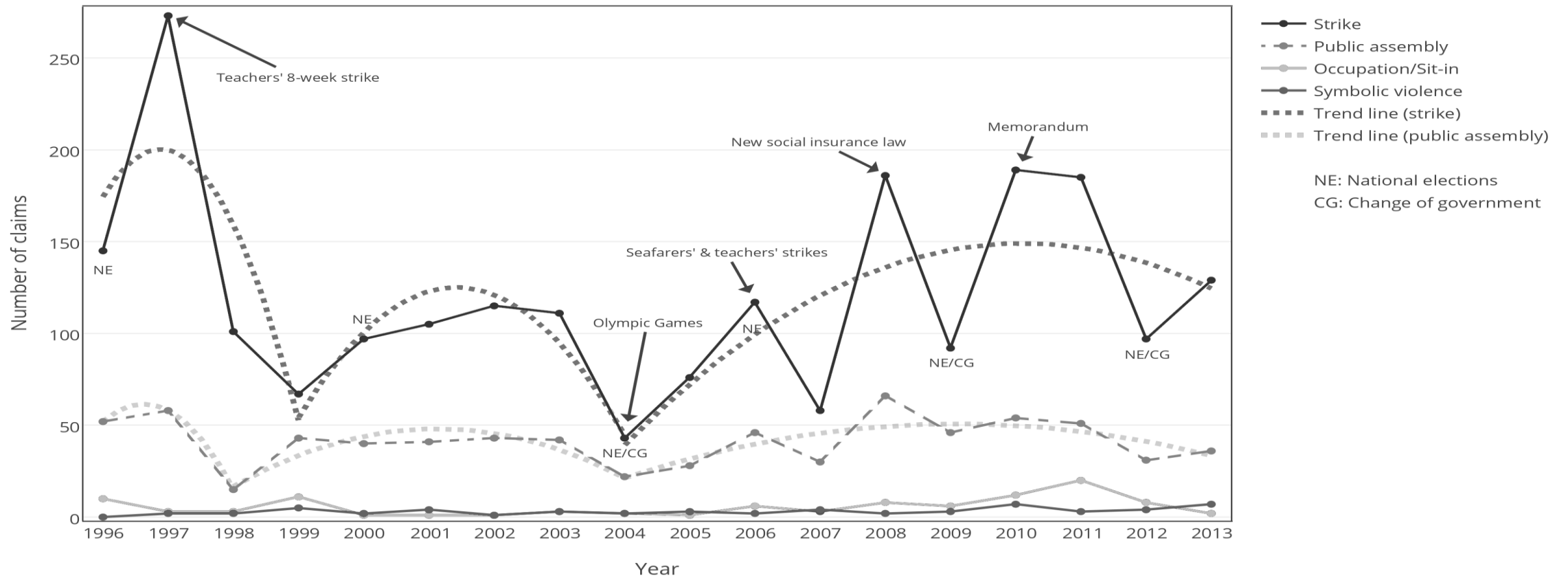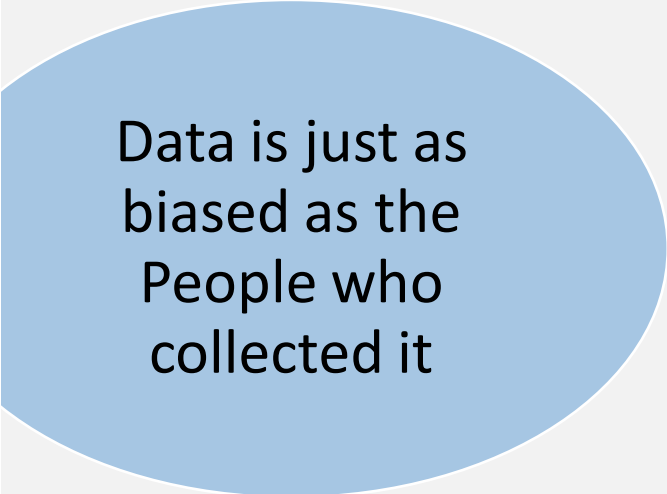
The post-election lows seem to be related to what political scientists call the electoral cycle (Miller and Mackie, 1973), referring to the fact that a <u>newly elected</u> government usually enjoys a brief honeymoon period characterised by increased popular support due to its fresh mandate.

Issues of concern

- F a i r N e s s
- Accountability & Transparency
- Interpretability

ICE hopes to build a system that would rely on government agency and law enforcement databases, as well as other public information available online, like social media activity, which will be analyzed for "continuous vetting" of travelers during their time in the U.S. Just as local law enforcement's use of predictive policing software can perpetuate existing prejudices by relying on historically biased data, so too could a national extreme-vetting system. That's because government law enforcement databases are also packed with biased information.

Data is just as biased as the People who collected it

Data Bias

> 💬 1   🔁 10   ♡ 28   ✉
>
> **Yasha Levine** ✓ @yashalevine · 10 Αυγ   ⌄
> Hitler was right is only 34% toxic, according to Google's AI.
>
> 🌐 Μετάφραση από Αγγλικά
>
> ▪▪▪ AT&T   LTE   9:58 PM   ↗ ⏱ 30% ▪
>
> 🔒 **perspectiveapi.com**   ↻
>
> ▪ 34% likely to be   SEEM
> perceived as "toxic"   WRONG?
>
> ## Hitler was right about everything
>
> 💬 21   🔁 165   ♡ 301   ✉
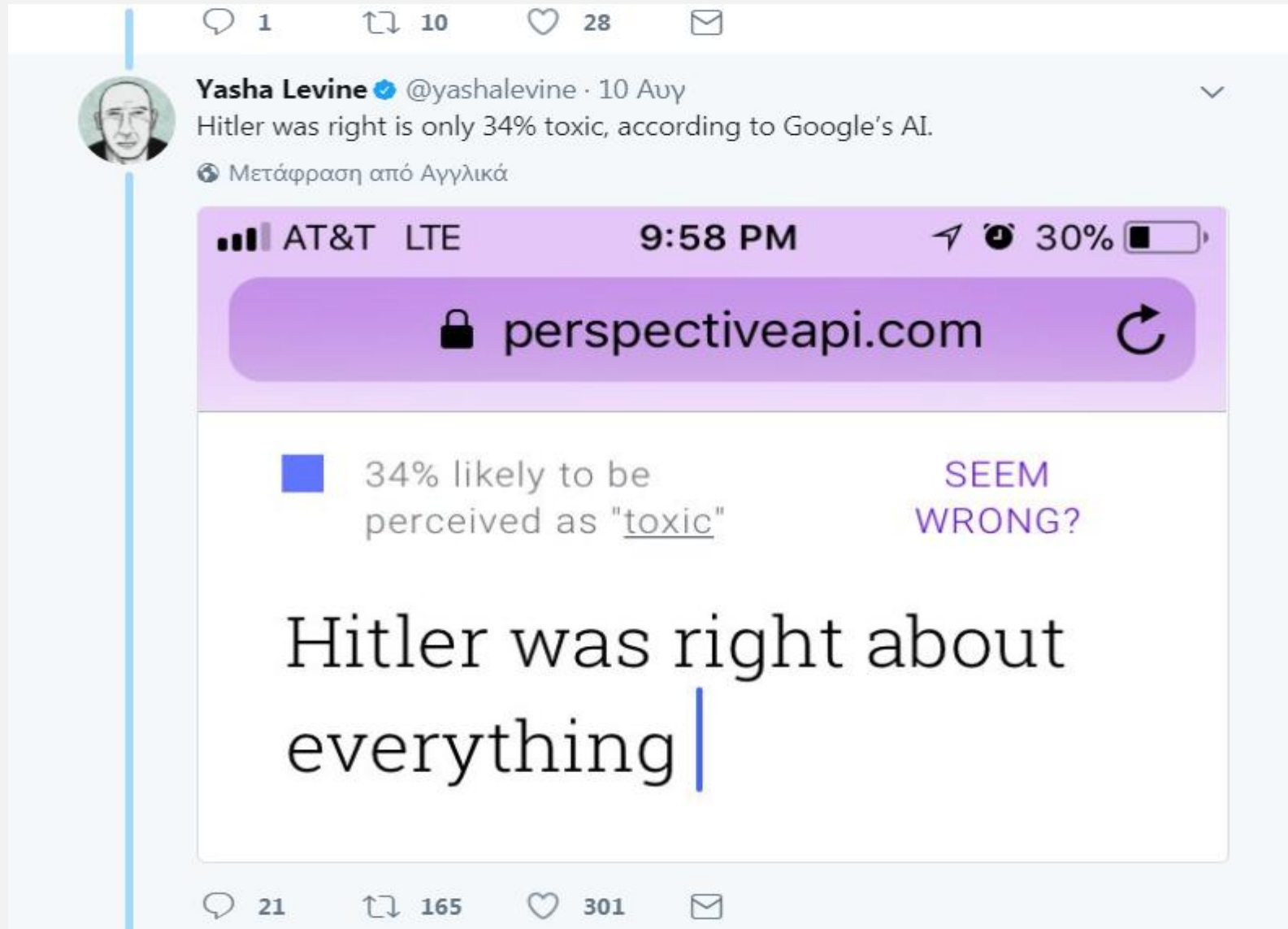
Media Bias

Al Jazeera English Follow
http://aljazeera.com
Aug 30 · 2 min read

Why we're disabling comments on aljazeera.com

Today, we disabled the ability to comment on stories on aljazeera.com. It's a decision that we've given much thought to, and one that we feel ultimately best serves our audience.

The mission of Al Jazeera is to give a voice to the voiceless, and healthy discussion is an active part of this. When we first opened up comments on our website, we hoped that it would serve as a forum for thoughtful and intelligent debate that would allow our global audience to engage with each other.

However, the comments section was hijacked by users hiding behind pseudonyms spewing vitriol, bigotry, racism and sectarianism. The possibility of having any form of debate was virtually non-existent.

Media Bias

ORIGINAL ARTICLE

A Bad Workman Blames His Tweets: The Consequences of Citizens' Uncivil Twitter Use When Interacting With Party Candidates

Yannis Theocharis[1], Pablo Barberá[2], Zoltán Fazekas[3], Sebastian Adrian Popa[1], & Olivier Parnet[4]

1 Mannheim Centre for European Social Research, University of Mannheim, Mannheim 68159, Germany
2 School of International Relations, University of California, Los Angeles, CA 90089, USA
3 Department of Political Science, University of Oslo, Oslo 0317, Norway
4 Kantar Public, Brussels 1160, Belgium

Existing studies focusing on politicians' adoption of Twitter have found that they use it primarily as a broadcasting tool. We argue that citizens' impolite and/or uncivil behavior is one possible explanation for such decisions. Social media conversations are rife with harassment and politicians are a prime target. This alters the incentive structure of engaging in dialogue on social media. We use Spanish, Greek, German, and U.K. candidates' tweets sent during the run-up to the recent European Parliament elections, and rely on automated text analysis and machine learning methods to measure their level of civility. Our contribution is an actor-oriented theory of political dialogue that incorporates Twitter's specific affordances, clarifying how and why Twitter's democratic promise may be limited.

..a better (CSS) approach

- Control: Gender
- Age Group
- Media
- Avoid misleading conclusions
- Use other, additional techniques (surveys, interviews)

Toward Algorithmic Transparency and Accountability

By Simson Garfinkel, Jeanna Matthews, Stuart S. Shapiro, Jonathan M. Smith
Comments

VIEW AS: SHARE:

Algorithms are replacing or augmenting human decision making in crucial ways. People have become accustomed to algorithms making all manner of recommendations, from products to buy, to songs to listen to, to social network connections. However, algorithms are not just recommending, they are also being used to make big decisions about people's lives, such as who gets loans, whose résumés are reviewed by humans for possible employment, and the length of prison terms. While algorithmic decision making can offer benefits in terms of speed, efficiency, and even fairness, there is a common misconception that algorithms automatically result in unbiased decisions. In reality, inscrutable algorithms can also unfairly limit opportunities, restrict services, and even improperly curtail liberty.

Information and communication technologies invariably raise these kinds of important public policy issues. How should self-driving cars be required to act? How private is information stored on a cellphone? Can electronic voting machines be trusted? How will the increasing uses of automation in the workplace impact

Interpretability

Subscribe

The Opinion Pages | OP-ED CONTRIBUTOR

How to Regulate Artificial Intelligence

By OREN ETZIONI   SEPT. 1, 2017

173

Isaac Lawrence/Agence France-Presse — Getty Images

RECENT COMMENTS

BG September 2, 2017
I assume we can have two AIRobots exchange their "personality" while humans cannot.If that happens then they will all want to be like the...

JohnV September 2, 2017
Victor Frankenstein's monster creation faced the ire and fear of the villagers but of course it was Victor who should be punished.All the...

Mark Thomason September 2, 2017
"any A.I. must have an impregnable "off switch." That must be a rest mode, not a kill mode. We can't expect actual intelligence to allow its...

SEE ALL COMMENTS

The technology entrepreneur Elon Musk recently urged the nation's governors to regulate artificial intelligence "before it's too late." Mr. Musk insists that artificial intelligence represents an "existential threat to humanity," an alarmist view that confuses A.I. science with science fiction. Nevertheless, even A.I. researchers like me recognize that there are valid concerns about its impact on weapons, jobs and privacy. It's natural to ask whether we should develop A.I. at all.

Interpretable AI: Not just for regulators

Patrick Hall (H2O.ai | George Washington University), Sri Satish (H2O.ai)
4:35pm–5:15pm Wednesday, September 27, 2017
Law, ethics, governance, Machine Learning
Location: 1A 06/07
Level: Intermediate

Description

While understanding and trusting models and their results is a hallmark of good (data) science, model interpretability is a serious legal mandate in the regulated verticals of banking, insurance, and other industries. Moreover, scientists, physicians, researchers, and humans in general have the right to understand and trust the models and modeling results that affect their work and their lives. Today, many are embracing deep learning and machine learning techniques, but what happens when people want to explain these impactful, complex technologies or when these technologies inevitably make mistakes?

Patrick Hall and Sri Satish share several approaches beyond the error measures and assessment plots typically used to interpret deep learning and machine learning models and results. Wherever possible, interpretability approaches are deconstructed into more basic components suitable for human story telling: complexity, scope, understanding, and trust.

Topics include:

- Data visualization techniques for representing high-degree interactions and nuanced data structures
- Contemporary linear model variants that incorporate machine learning and are appropriate for use in regulated industry
- Cutting-edge approaches for explaining extremely complex deep learning and machine learning models

Haris Papageorgiou

ILSP/Athena R.C.

clarin:el

@**xpapag**

xaris@ilsp.gr