Surviving to social media in the misinformation era

WOMEN IN DATA SCIENCE – LA PAZ, BOLIVIA

ANTONELA TOMMASEL

ISISTAN, CONICET-UNICEN, ARGENTINA

Who am I?

- Dr Antonela Tommasel
  - PhD in Computer Sciences at UNICEN

- Assistant Researcher at ISISTAN, CONICET-UNICEN.

- Teacher Assistant at UNICEN.

- Research Interests:
  - Recommender systems
  - Text mining
  - Social media
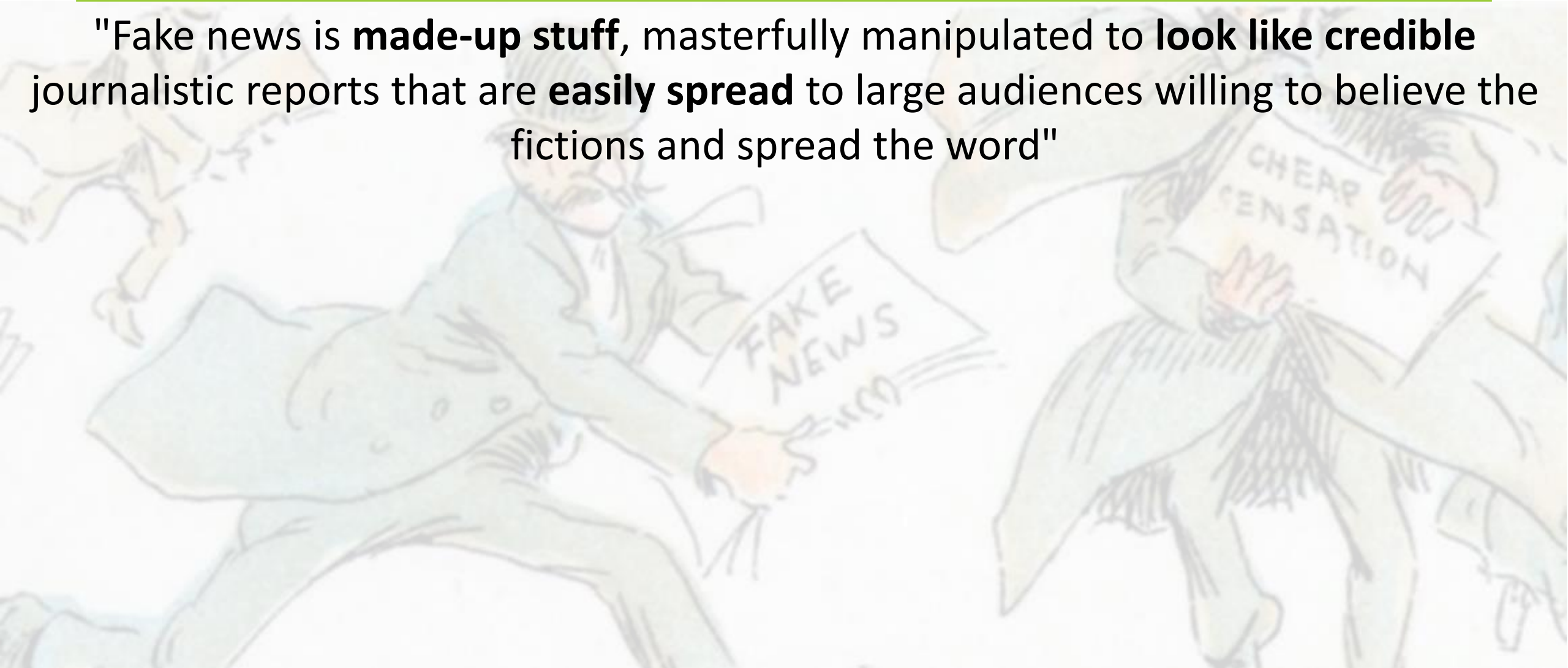 - Social computing
  - eCitizenship
  - Hate speech
  - …

Fake news are among us!

"Fake news is **made-up stuff**, masterfully manipulated to **look like credible** journalistic reports that are **easily spread** to large audiences willing to believe the fictions and spread the word"
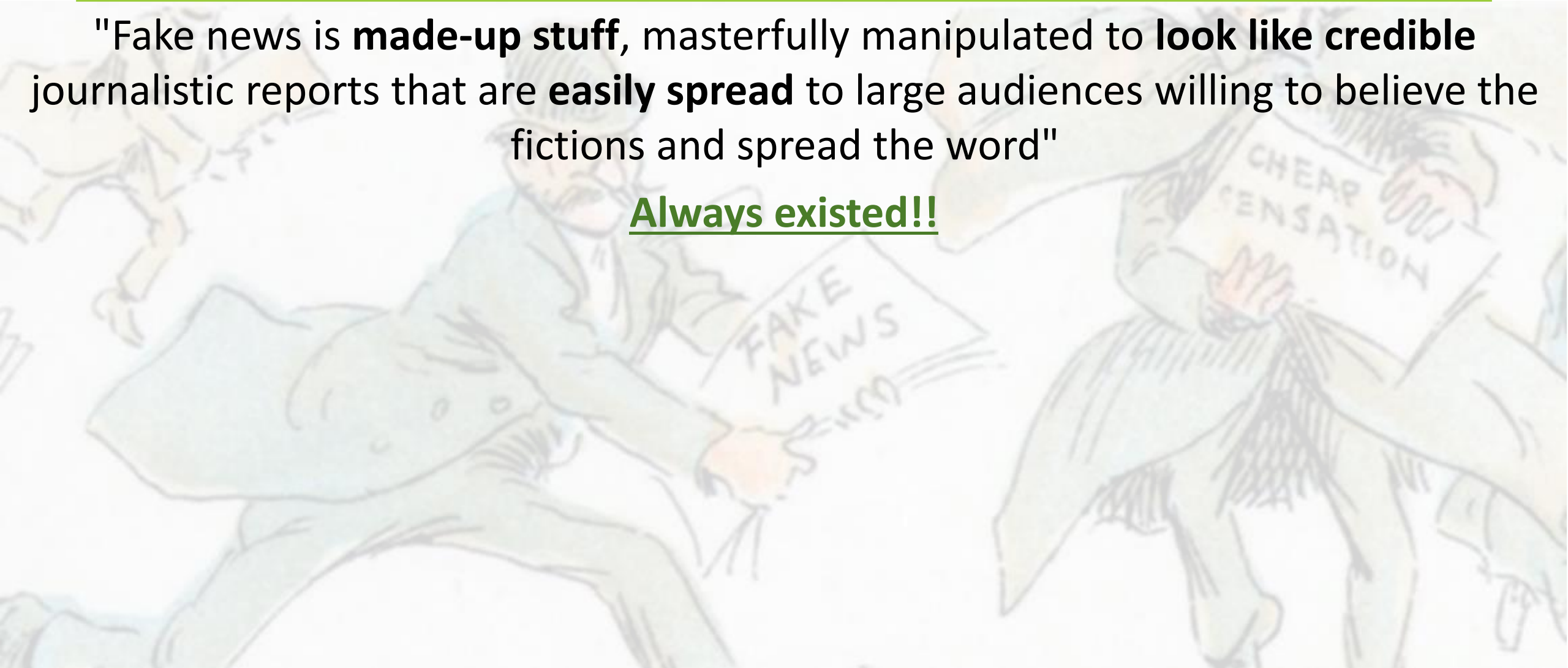
Fake news are among us!

"Fake news is **made-up stuff**, masterfully manipulated to **look like credible** journalistic reports that are **easily spread** to large audiences willing to believe the fictions and spread the word"

Always existed!!

Fake news are among us!

"Fake news is **made-up stuff**, masterfully manipulated to **look like credible** journalistic reports that are **easily spread** to large audiences willing to believe the fictions and spread the word"

Always existed!!

The King's Health is Failing
(mid 1700s – Jacobite rebellion)

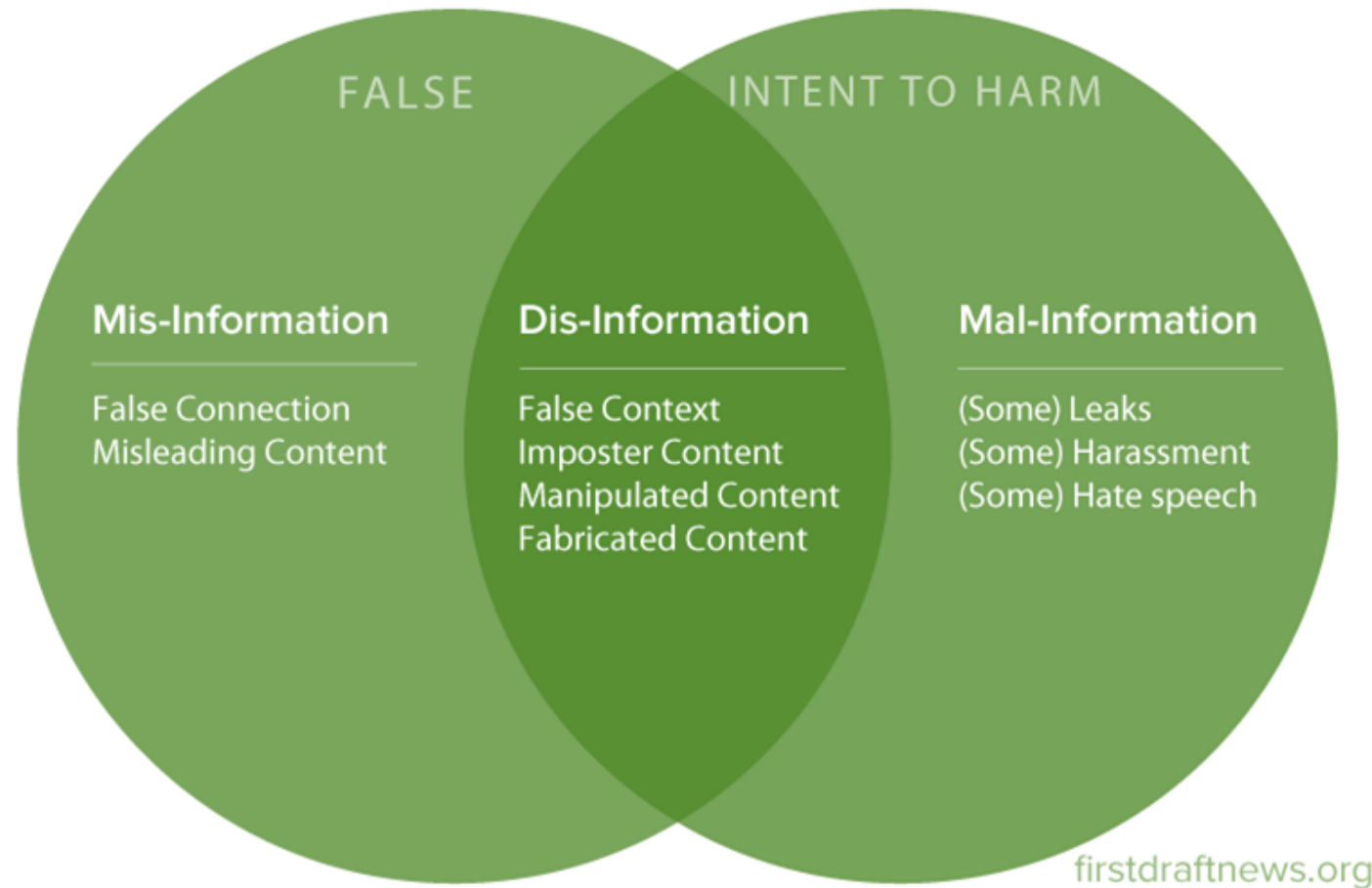Life on the Moon
(1835)

Jack the Ripper
(1888)

World War One Fake News
(1917)

FINDING THE MUTILATED BODY IN MITRE SQARE

THE GERMAN "KADAVER" FACTORIES.

LORD R. CECIL (Hitchin, U.), replying to Mr. R. M'NEILL (St. Augustine's, U.), who asked whether the Government would take steps to make it known as widely as possible in Egypt, India, and the East generally, that the Germans use the dead bodies of their own soldiers, and of their enemies when they obtain possession of them, as food for swine, and to an inquiry by Mr. DILLON (Mayo, E., Nat.) whether the Government had any solid ground for believing to be well founded the statements, widely circulated in this country, that the German Government had set up factories for extracting fat from the bodies of soldiers killed in battle, said:—The Government have no information beyond that contained in extracts from the German Press which have been published in the Press here. In view of other actions taken by the German military authorities there is nothing incredible in the present charge against them. His Majesty's Government have allowed the circulation of the facts as they appeared through the usual channels.

Fake news are among us!
...but it is not only Fake News!

FALSE

INTENT TO HARM

Mis-Information

False Connection
Misleading Content

Dis-Information

False Context
Imposter Content
Manipulated Content
Fabricated Content

Mal-Information

(Some) Leaks
(Some) Harassment
(Some) Hate speech

firstdraftnews.org

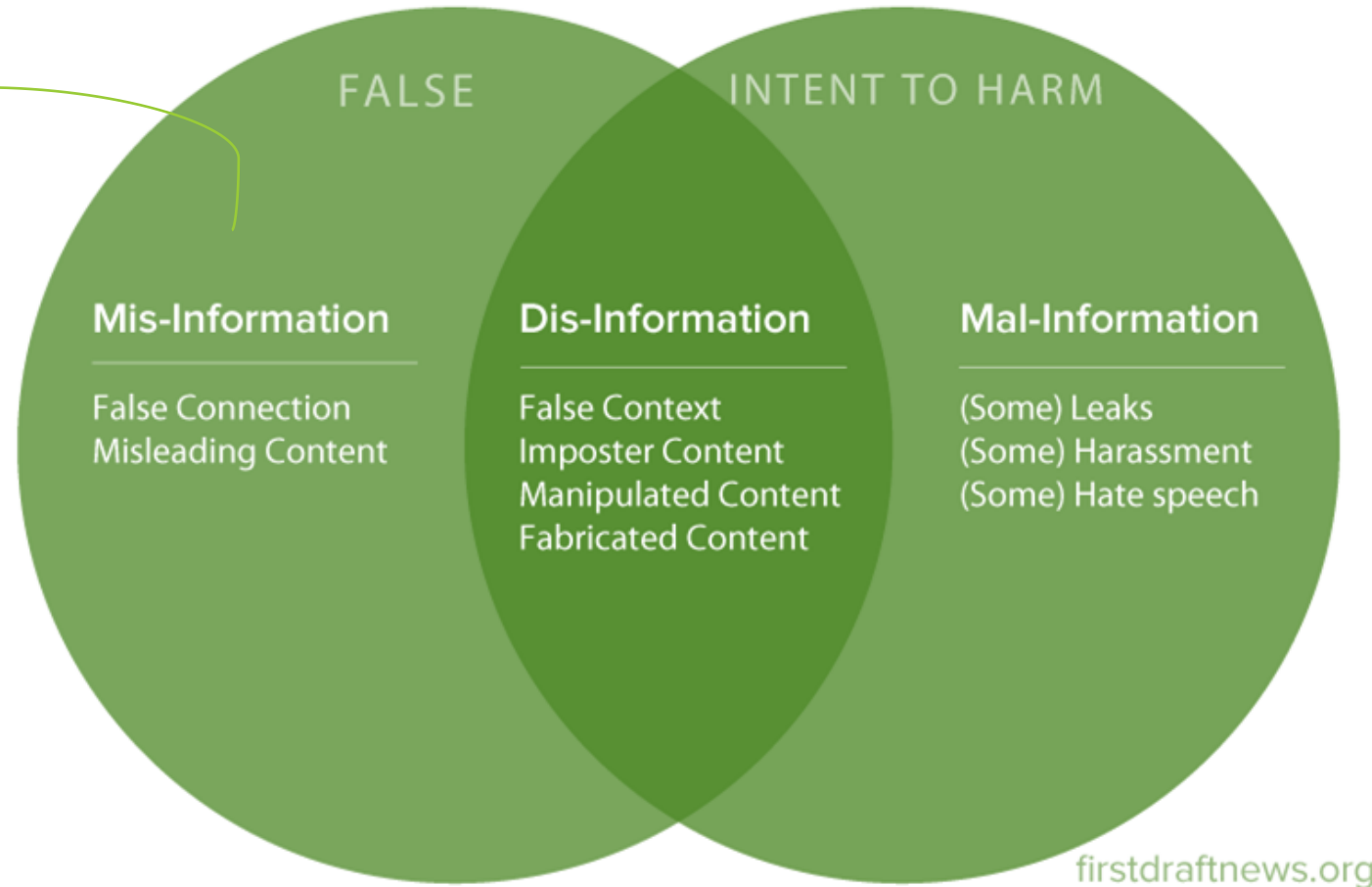https://firstdraftnews.org/latest/fake-news-complicated/

Fake news are among us!
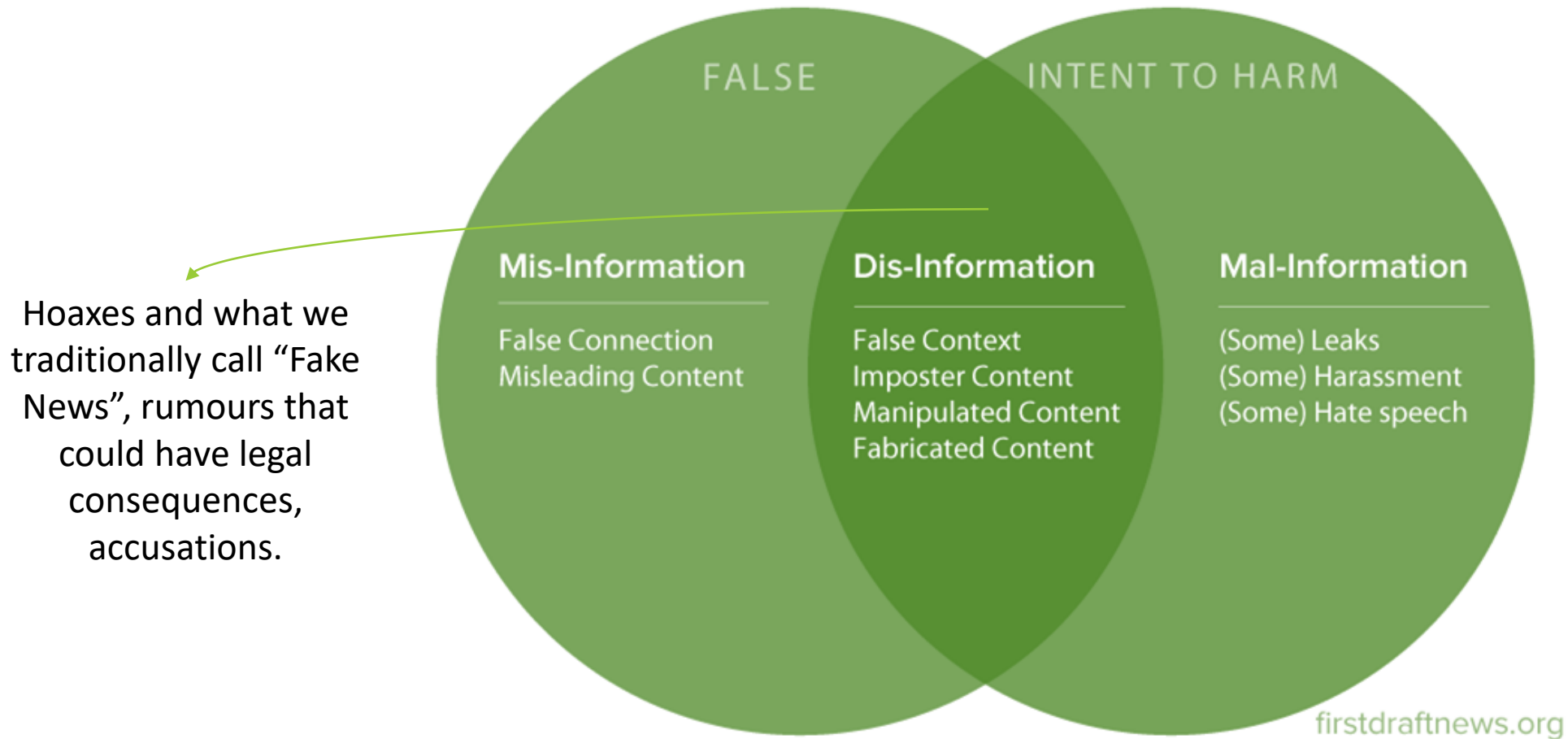## ...but it is not only Fake News!

Rumours. For example, related to the attack of Charlie Hebdo, advices for preventing COVID...

In most cases, no harm is meant, people is caught up in the moment trying to be helpful and do not inspect what they are sharing.
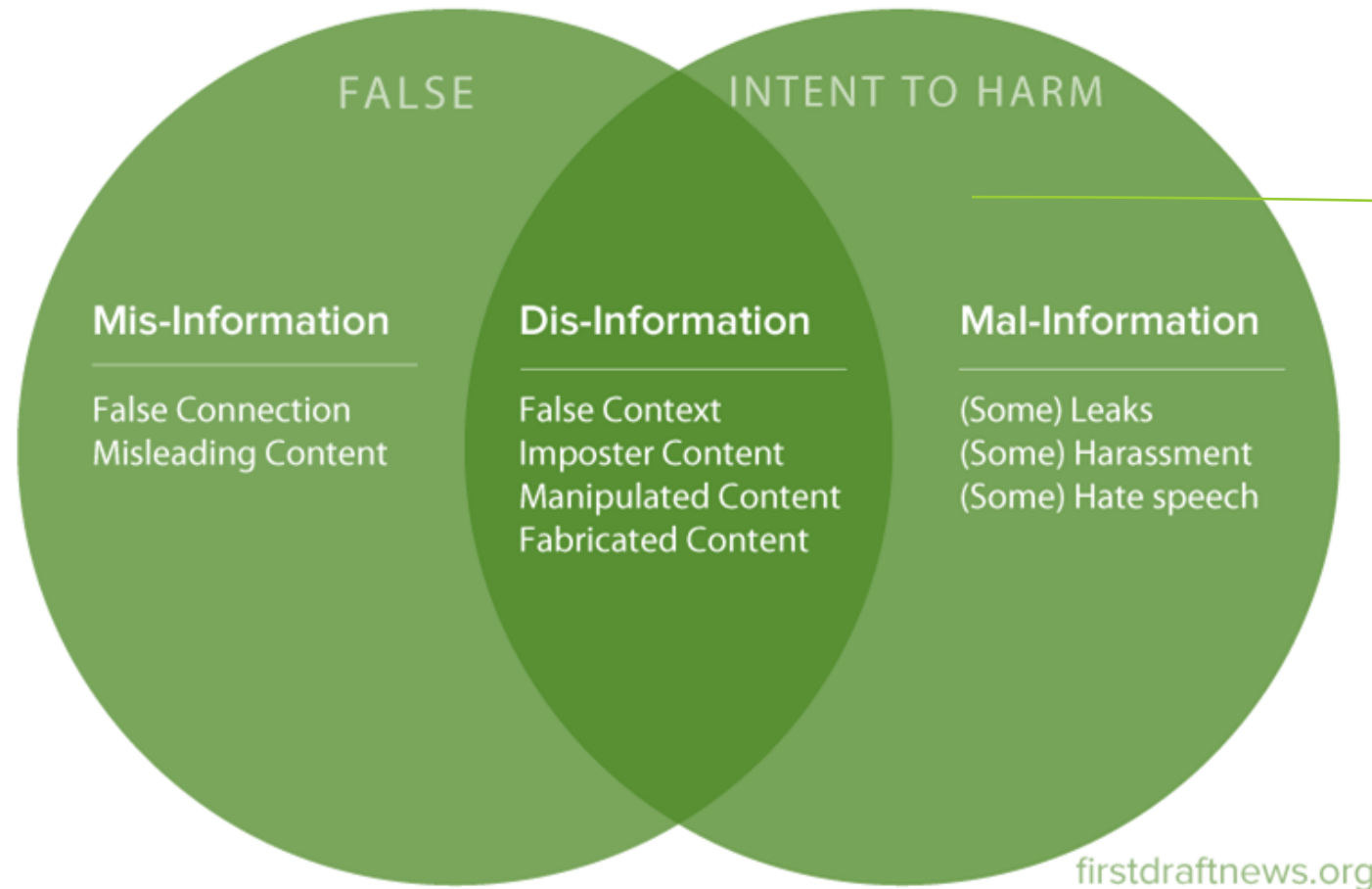
FALSE

INTENT TO HARM

Mis-Information

False Connection
Misleading Content

Dis-Information

False Context
Imposter Content
Manipulated Content
Fabricated Content

Mal-Information

(Some) Leaks
(Some) Harassment
(Some) Hate speech

firstdraftnews.org

https://firstdraftnews.org/latest/fake-news-complicated/

Fake news are among us!
...but it is not only Fake News!

Hoaxes and what we traditionally call "Fake News", rumours that could have legal consequences, accusations.

FALSE

INTENT TO HARM

Mis-Information

False Connection
Misleading Content

Dis-Information

False Context
Imposter Content
Manipulated Content
Fabricated Content

Mal-Information

(Some) Leaks
(Some) Harassment
(Some) Hate speech

firstdraftnews.org

https://firstdraftnews.org/latest/fake-news-complicated/

Fake news are among us!
...but it is not only Fake News!

FALSE

INTENT TO HARM

Mis-Information

False Connection
Misleading Content

Dis-Information

False Context
Imposter Content
Manipulated Content
Fabricated Content

Mal-Information

(Some) Leaks
(Some) Harassment
(Some) Hate speech

Leaking information with the intent to harm, but not realizing that the leaked information is false.

firstdraftnews.org

https://firstdraftnews.org/latest/fake-news-complicated/

Fake news are among us!
...Mis- and Dis-information

Satire or Parody
No intention to cause harm, but has potential to fool

False Connections
When headlines, visuals or captions don't support the content

Fabricated Content
New content is 100% false, designed to deceive and do harm

Misleading Content
Misleading use of information to frame an issue or individual

False Context
When genuine content is shared with false contextual information

Imposter Content
When genuine sources and impersonated

Manipulated Content
When genuine information or imagery is manipulated to deceive

https://firstdraftnews.org/latest/fake-news-complicated/

Fake news are among us!
...Mis- and Dis-information

Satire or Parody
No intention to cause harm, but has potential to fool

False Connections
When headlines, visuals or captions don't support the content

Fabricated Content
New content is 100% false, designed to deceive and do harm

Misleading Content
Misleading use of information to frame an issue or individual

False Context
When genuine content is shared with false contextual information

Imposter Content
When genuine sources and impersonated

Manipulated Content
When genuine information or imagery is manipulated to deceive

Why?

https://firstdraftnews.org/latest/fake-news-complicated/

Fake news are among us!
...Mis- and Dis-information

Satire or Parody
No intention to cause harm, but has potential to fool

False Connections
When headlines, visuals or captions don't support the content

Fabricated Content
New content is 100% false, designed to deceive and do harm

Misleading Content
Misleading use of information to frame an issue or individual

False Context
When genuine content is shared with false contextual information

Imposter Content
When genuine sources and impersonated

Manipulated Content
When genuine information or imagery is manipulated to deceive

Why?
- To provoke.
- To profit.
- For political influence.

- Propaganda.
- Partisanship.
- Poor journalist.

https://firstdraftnews.org/latest/fake-news-complicated/

Fake news are among us!

Social media aggravates the problem!!

A photo or comment that is posted online, and then shared by many people goes viral, spreading from one person to many as quickly as a virus does.

Fake news are among us!
Social media aggravates the problem!
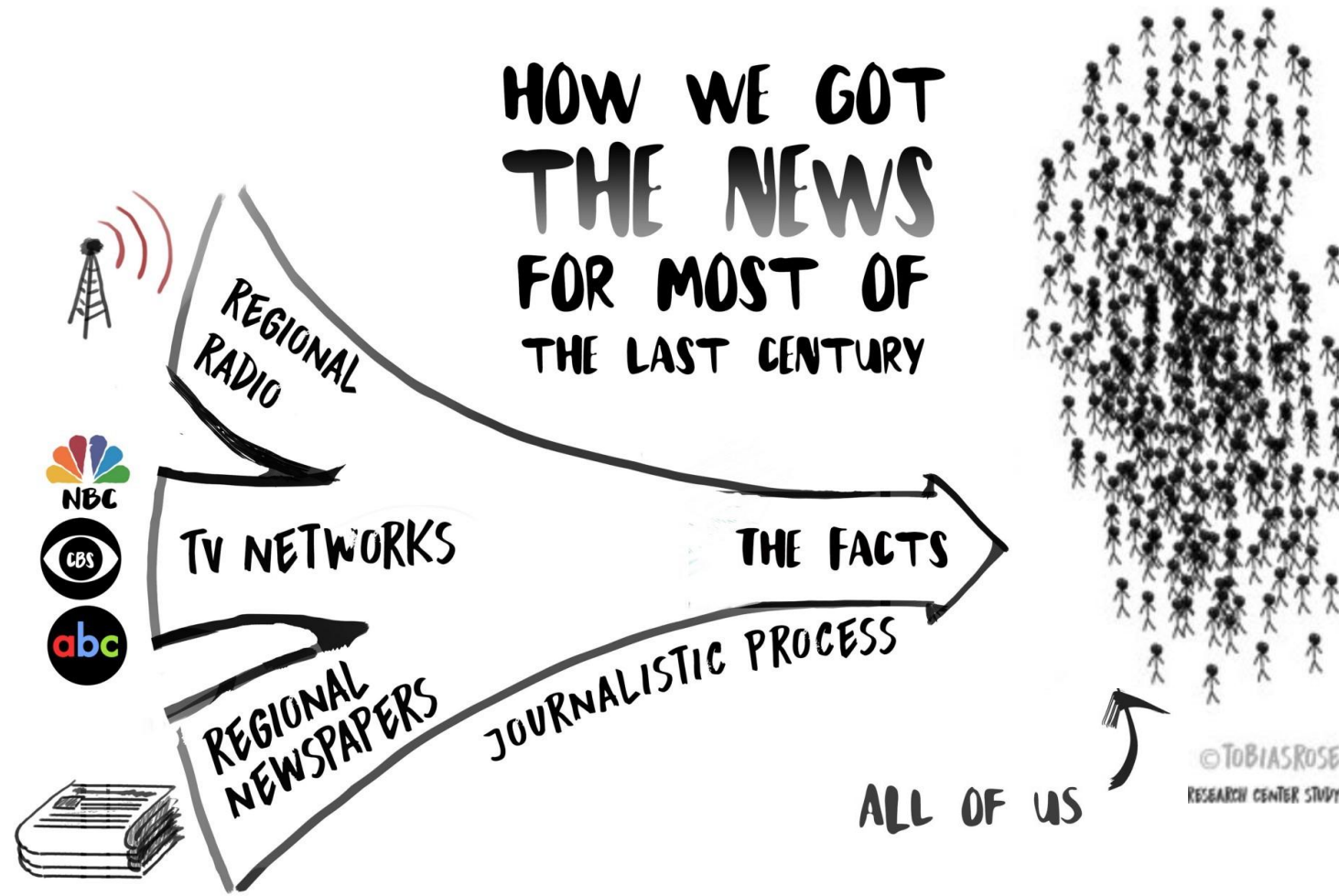
Fake news are among us!
Social media aggravates the problem!

THOUSANDS OF NEWS SOURCES (BOTH REAL AND FAKE)

MOST OF US

©TOBIASROSE

SOURCE: 2016 PEW RESEARCH CENTER STUDY
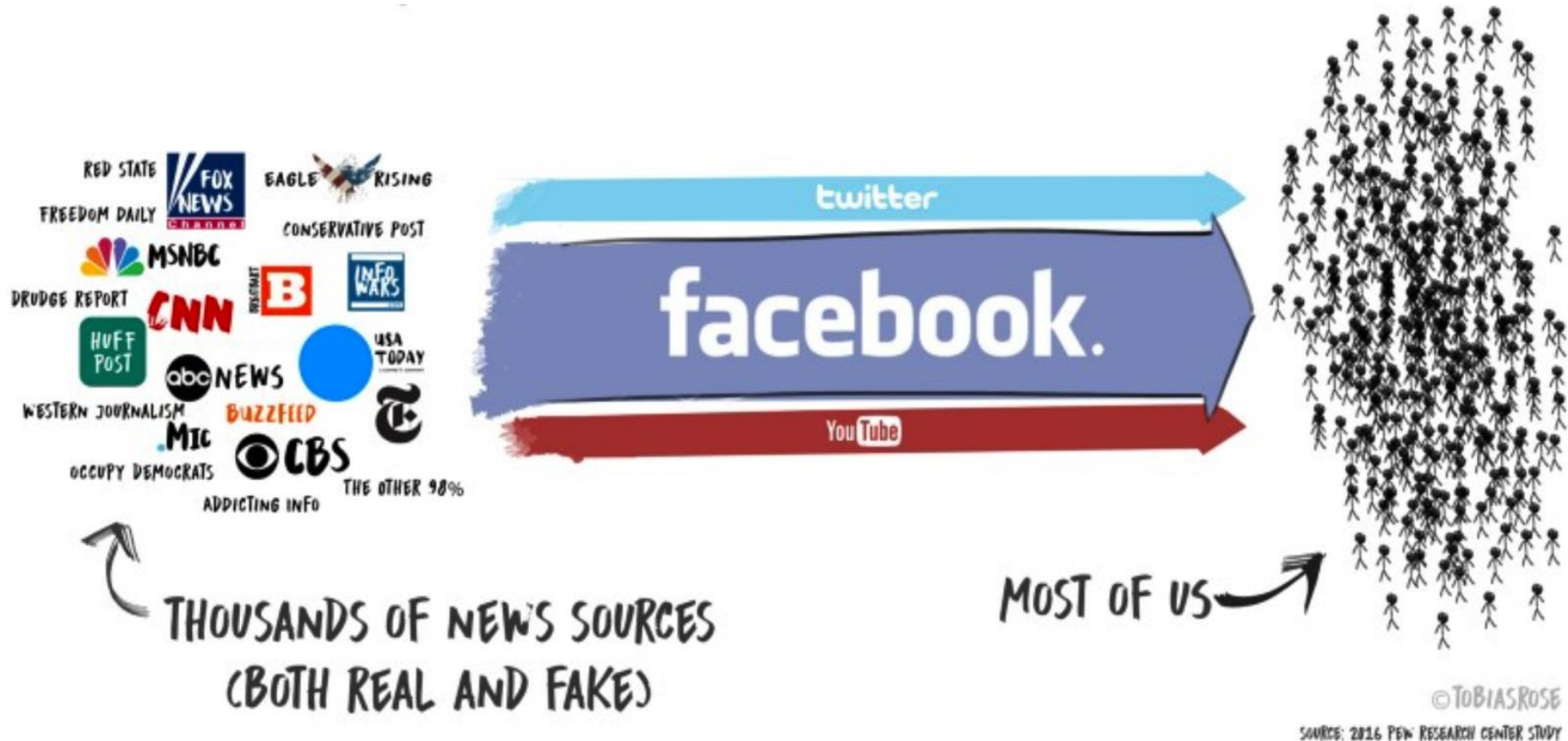
Fake news are among us!
Social media aggravates the problem!

Fake news are among us!

We are working closely together on COVID-19 response efforts. We're helping millions of people stay connected while also jointly combating fraud and misinformation about the virus, elevating authoritative content on our platforms, and sharing critical updates in coordination with government healthcare agencies around the world. We invite other companies to join us as we work to keep our communities healthy and safe.

FACEBOOK   Google   Linked**in**   Microsoft   reddit   Twitter   YouTube

Joint Industry Statement from Facebook, Google, LinkedIn, Microsoft, Reddit, Twitter and YouTube

Fake news are among us!
What is Social Media doing?

Twitter Safety ✔ @TwitterSafety · 27 May
We added a label to two @realDonaldTrump Tweets about California's vote-by-mail plans as part of our efforts to enforce our civic integrity policy. We believe those Tweets could confuse voters about what they need to do to receive a ballot and participate in the election process.

💬 7.8K   🔁 11.1K   ♡ 46.1K   ⬆️

Show this thread

Twitter Safety ✔ @TwitterSafety · 22 Apr
Since introducing our updated policies on March 18, we've removed over 2,230 Tweets containing misleading and potentially harmful content. Our automated systems have challenged more than 3.4 million accounts targeting manipulative discussions around COVID-19.

💬 173   🔁 155   ♡ 517   ⬆️

Show this thread

jack ✔ @jack
Fact check: there is someone ultimately accountable for our actions as a company, and that's me. Please leave our employees out of this. We'll continue to point out incorrect or disputed information about elections globally. And we will admit to and own any mistakes we make.

11:47 pm · 27 May 2020 from California, USA · Twitter for iPhone

40.7K Retweets   **218.4K** Likes

Twitter Safety ✔ @TwitterSafety
We have placed a public interest notice on this Tweet from @realDonaldTrump.

This Tweet violated the Twitter Rules about glorifying violence. However, Twitter has determined that it may be in the public's interest for the Tweet to remain accessible. Learn more   View

4:39 am · 29 May 2020 · Twitter Web App

2.4K Retweets   **4.8K** Likes

Twitter Support ✔ @TwitterSupport · 10 Jun
Sharing an article can spark conversation, so you may want to read it before you Tweet it.

To help promote informed discussion, we're testing a new prompt on Android — when you Retweet an article that you haven't opened on Twitter, we may ask if you'd like to open it first.

💬 2.4K   🔁 18.2K   ♡ 59.5K   ⬆️

Fake news are among us!
What is Social Media doing?

How is Facebook addressing false news through third-party fact-checkers?

→ Share article

We're committed to fighting the spread of false news on Facebook. We use both technology and human review to remove fake accounts, promote news literacy and disrupt the financial incentives of spammers. In certain countries, we also work with third-party fact-checkers who are certified through the non-partisan International Fact-Checking Network to help identify and review false news.
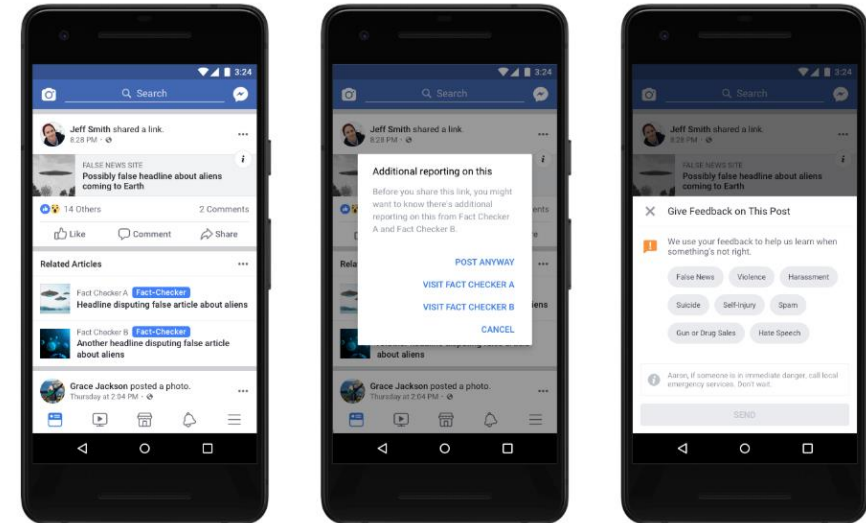
Fake news & Social Media!

- Social media represents the ideal environment for undesirable phenomena!
  - The dissemination of unwanted or unreliable content, and misinformation.

- A threat to the access to reliable and **trustworthy** information and the establishment of **reliable** social relations.

Fake news & Social Media!

- Social media represents the ideal environment for undesirable phenomena!
  - The dissemination of unwanted or unreliable content, and misinformation.

- A threat to the access to reliable and **trustworthy** information and the establishment of **reliable** social relations.

Social media provides a great
opportunity to learn about events and news.
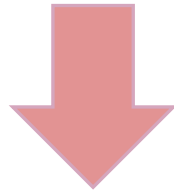
Social media produces scepticism amongst users
as relevant and accurate information coexist with
unreliable and undesired information.

Fake news & Social Media!

- Social media represents the ideal environment for undesirable phenomena!
    - The dissemination of unwanted or unreliable content, and misinformation.

- A threat to the access to reliable and **trustworthy** information and the establishment of **reliable** social relations.

Social media provides a great opportunity to learn about events and news.

Social media produces scepticism amongst users as relevant and accurate information coexist with unreliable and undesired information.

- The growing spread of undesired content motivated the assessment of the reliability of information.
- The vulnerability of individuals and society to the manipulations is still unknown.

Fake news & Social Media!

increasing availability and popularity of social media

+
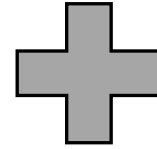
low cost of producing fraudulent sites

Fake news & Social Media!

increasing availability and popularity of social media **+** low cost of producing fraudulent sites

=

rapid creation and dissemination of misinformation

Fake news & Social Media!

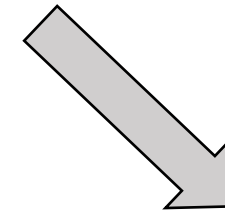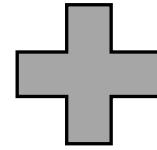increasing availability and popularity of social media

+

low cost of producing fraudulent sites

=
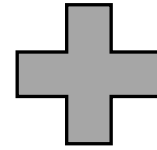
rapid creation and dissemination of misinformation
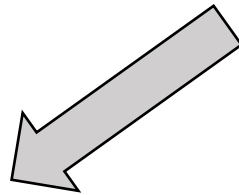
overflows legitimate users with unreliable information

influences public opinion

Fake news & Social Media!

increasing availability and popularity of social media

+

low cost of producing fraudulent sites

=

rapid creation and dissemination of misinformation
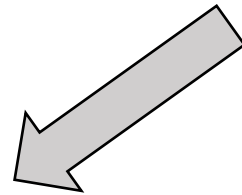
overflows legitimate users with unreliable information
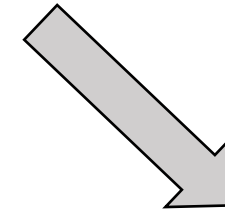
influences public opinion

The value and quality of the social Web diminishes!!

What has been done?

- The development of methods for automatically detecting undesired content is **essential**.
  - Such detection is not simple.
  - Mainly based on **one** of three aspects:

    - Textual content.

    - The responses received.

    - The identification of the content promoters.

What has been done?

- The development of methods for automatically detecting undesired content is **essential**.
  - Such detection is not simple.
  - Mainly based on **one** of three aspects:

    - Textual content.

    - The responses received.

    - The identification of the content promoters.

Each type of unwanted content may have different textual indicators.

What has been done?

- The development of methods for automatically detecting undesired content is **essential**.
    - Such detection is not simple.
    - Mainly based on **one** of three aspects:

        - Textual content.

        - The responses received.

        - The identification of the content promoters.

Each type of unwanted content may have different textual indicators.

Responses focus on content propagation requiring access to large amounts of data.

What has been done?

- The development of methods for automatically detecting undesired content is **essential**.
  - Such detection is not simple.
  - Mainly based on **one** of three aspects:

    - Textual content.

    - The responses received.

    - The identification of the content promoters.

Each type of unwanted content may have different textual indicators.

Responses focus on content propagation requiring access to large amounts of data.

- Based on the same characteristics as the detection of unreliable content.
- Mostly, techniques attempt to determine only if an account is a certain type of unwanted user (binary classification).

What has been done?
Challenges

1. Identification requires more than text analysis, hence multiple sources of information must be integrated.

What has been done?
Challenges

1. Identification requires more than text analysis, hence multiple sources of information must be integrated.

It can be challenging given the incompleteness, noise and volume of it.

What has been done?
Challenges

1. Identification requires more than text analysis, hence multiple sources of information must be integrated.

It can be challenging given the incompleteness, noise and volume of it.

2. Bots and accounts spreading misinformation modify their behaviour patterns in an attempt to go unnoticed.

What has been done?
Challenges

1. Identification requires more than text analysis, hence multiple sources of information must be integrated.

It can be challenging given the incompleteness, noise and volume of it.

2. Bots and accounts spreading misinformation modify their behaviour patterns in an attempt to go unnoticed.

Detection will be less effective if training data is not periodically updated.

What has been done?
Challenges

1. Identification requires more than text analysis, hence multiple sources of information must be integrated.

It can be challenging given the incompleteness, noise and volume of it.

2. Bots and accounts spreading misinformation modify their behaviour patterns in an attempt to go unnoticed.

Detection will be less effective if training data is not periodically updated.

3. Techniques may be over-trained for a specific type of misinformation or spam campaign, limiting its applicability in broader domains.

What has been done?
Challenges

1. Identification requires more than text analysis, hence multiple sources of information must be integrated.

It can be challenging given the incompleteness, noise and volume of it.

2. Bots and accounts spreading misinformation modify their behaviour patterns in an attempt to go unnoticed.

Detection will be less effective if training data is not periodically updated.

3. Techniques may be over-trained for a specific type of misinformation or spam campaign, limiting its applicability in broader domains.

Each type of misinformation presents particular characteristics, which must be taken into account for detection.

What has been done?
Challenges

1. Identification requires more than text analysis, hence multiple sources of information must be integrated.

It can be challenging given the incompleteness, noise and volume of it.

2. Bots and accounts spreading misinformation modify their behaviour patterns in an attempt to go unnoticed.

Detection will be less effective if training data is not periodically updated.

3. Techniques may be over-trained for a specific type of misinformation or spam campaign, limiting its applicability in broader domains.

Each type of misinformation presents particular characteristics, which must be taken into account for detection.

4. As undesired content does not appear spontaneously, it is vital to analyse who published it, its intentions and processes.

What has been done?
Challenges

1. Identification requires more than text analysis, hence multiple sources of information must be integrated.

It can be challenging given the incompleteness, noise and volume of it.

2. Bots and accounts spreading misinformation modify their behaviour patterns in an attempt to go unnoticed.

Detection will be less effective if training data is not periodically updated.

3. Techniques may be over-trained for a specific type of misinformation or spam campaign, limiting its applicability in broader domains.

Each type of misinformation presents particular characteristics, which must be taken into account for detection.

4. As undesired content does not appear spontaneously, it is vital to analyse who published it, its intentions and processes.

Affected by the lack of integration of multiple sources of information, the updating of techniques and the disregard of the interrelation between different social platforms.

And what about recommendations?

Recommending <u>similar</u> elements
is **<u>NOT</u>** sufficient to ensure reliability!

And what about recommendations?

Recommending <u>similar</u> elements
is **<u>NOT</u>** sufficient to ensure reliability!

We need **<u>Relevant</u>** **and** **<u>Trustworthy</u>** recommendations!

And what about recommendations?

Recommending <u>similar</u> elements
is **<u>NOT</u>** sufficient to ensure reliability!

We need **<u>Relevant</u>** and **<u>Trustworthy</u>** recommendations!

New challenges!

User trustworthiness

Measuring trustworthiness has been important in psychology and social sciences.

User trustworthiness

Measuring trustworthiness has been important in psychology and social sciences.

Various factors to consider:
- Personal relationships.
- Past experiences of a user with their friends.
- Actions and opinions made in the past.
- …

In social media focuses on behaviours expressed in the way information is produced and shared.

No attention to the principle of unequal participation. Largest proportion of content is created by the minority of users, whilst the other are lurkers.

User trustworthiness

Measuring trustworthiness has been important in psychology and social sciences.
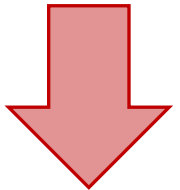
Few studies have incorporated this concept to:
- Social media
- The proliferation of unwanted content recommendations

User trustworthiness

Measuring trustworthiness has been important in psychology and social sciences.
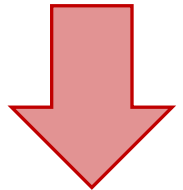
- Generally studied in the context of collaborative filtering to:
  - Determine the reliability of users' ratings.
  - Mitigate the cold start problem.

Few studies have incorporated this concept to:
- Social media
- The proliferation of unwanted content recommendations

- Aspects specifically related to unwanted content are not considered.
- Require explicit reputation indicators.
- Do not consider the dynamism of the social environments.

More challenges and questions!

How recommender systems influence the propagation of harmful content?

Is there a benchmark of data?

How to characterize users?

How trust needs to be adapted to changes in behaviour?

How to dynamically characterize users?

How to integrate trust in recommender systems?

More challenges and questions!

How recommender systems influence the propagation of harmful content?

How to characterize users?

How to dynamically characterize users?

Is there a benchmark of data?

How trust needs to be adapted to changes in behaviour?

How to integrate trust in recommender systems?

More challenges and questions!
Where is the data?

- Even though some computational solutions have been presented, the lack of common ground and public datasets has become one of the major barriers.

- Not only datasets are **rare**, but also they are **mostly limited** to only the actual shared **text**.

- **Create a publicly available dataset!**
  - Comprising multi-sourced data including:
    - Textual and multimedia content.
    - Social Context.
    - Temporal information.

- Potential uses:
  - Undesired content and user detection.
  - Evolution and engagement cycle.
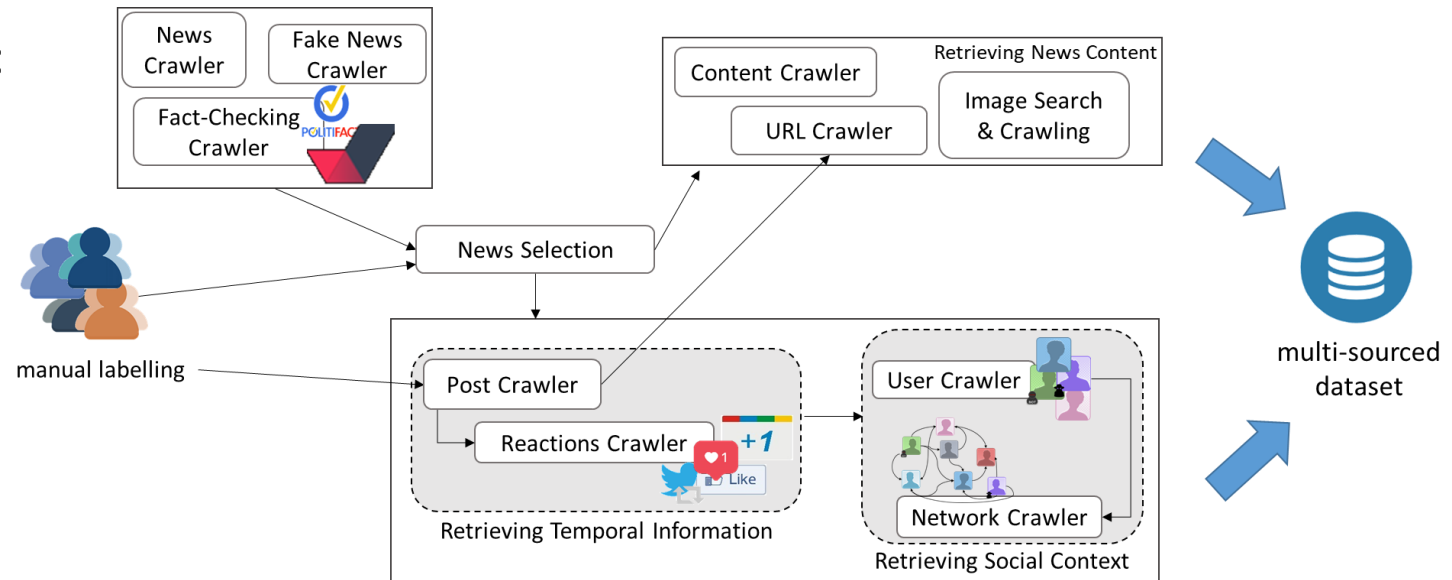 - Debunking process.

More challenges and questions!
Where is the data?

- Even though some computational solutions have been presented, the lack of common ground and public datasets has become one of the major barriers.

- Not only datasets are **rare**, but also they are **mostly limited** to only the actual shared **text**.

- **Create a publicly available dataset!**
  - Comprising multi-sourced data including:
    - Textual and multimedia content.
    - Social Context.
    - Temporal information.

- Potential uses:
  - Undesired content and user detection.
  - Evolution and engagement cycle.
  - Debunking process.

More challenges and questions!
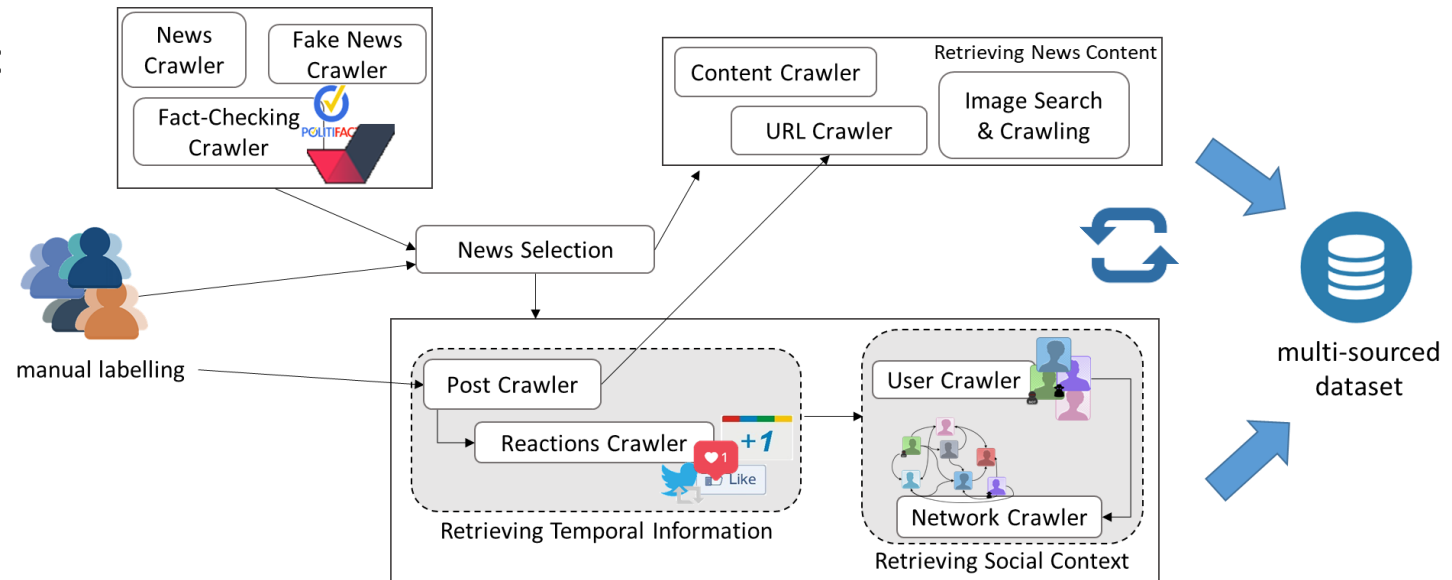## Where is the data?

- Even though some computational solutions have been presented, the lack of common ground and public datasets has become one of the major barriers.

- Not only datasets are **rare**, but also they are **mostly limited** to only the actual shared **text**.

- **Create a publicly available dataset!**
  - Comprising multi-sourced data including:
    - Textual and multimedia content.
    - Social Context.
    - Temporal information.

- Potential uses:
  - Undesired content and user detection.
  - Evolution and engagement cycle.
  - Debunking process.

More challenges and questions!
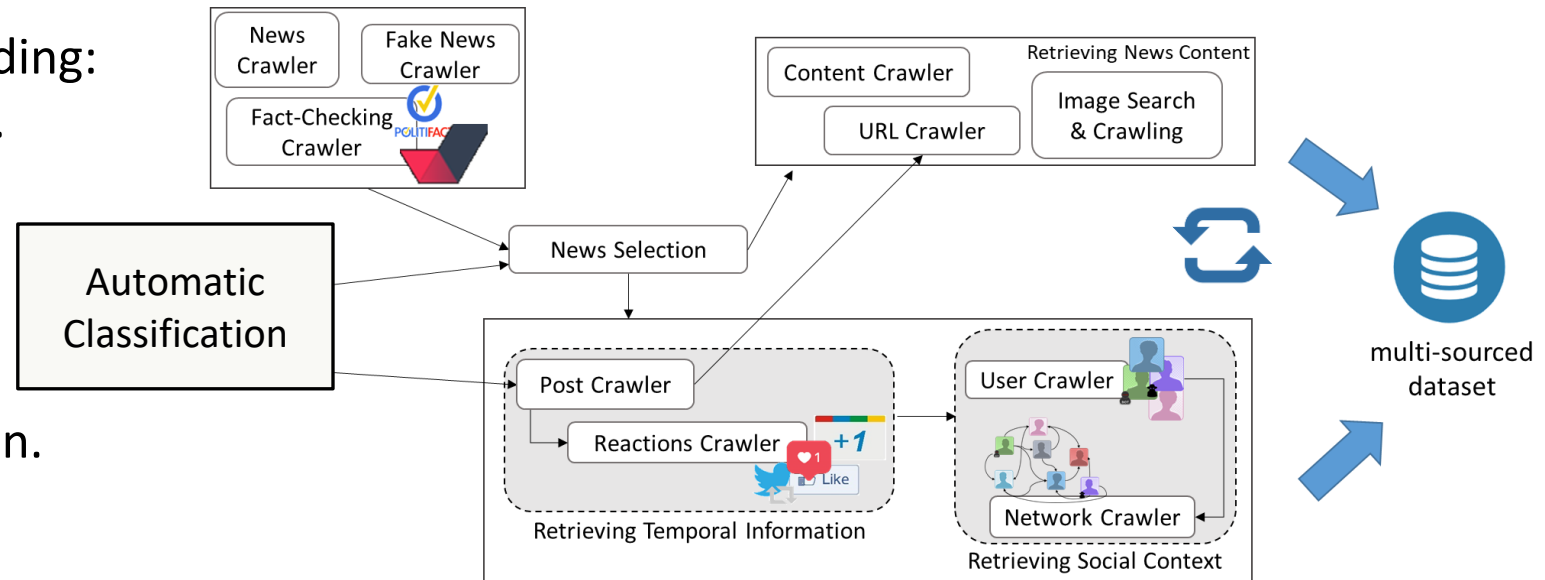## Where is the data?

- Even though some computational solutions have been presented, the lack of common ground and public datasets has become one of the major barriers.

- Not only datasets are **rare**, but also they are **mostly limited** to only the actual shared **text**.

- **Create a publicly available dataset!**
  - Comprising multi-sourced data including:
    - Textual and multimedia content.
    - Social Context.
    - Temporal information.

- Potential uses:
  - Undesired content and user detection.
  - Evolution and engagement cycle.
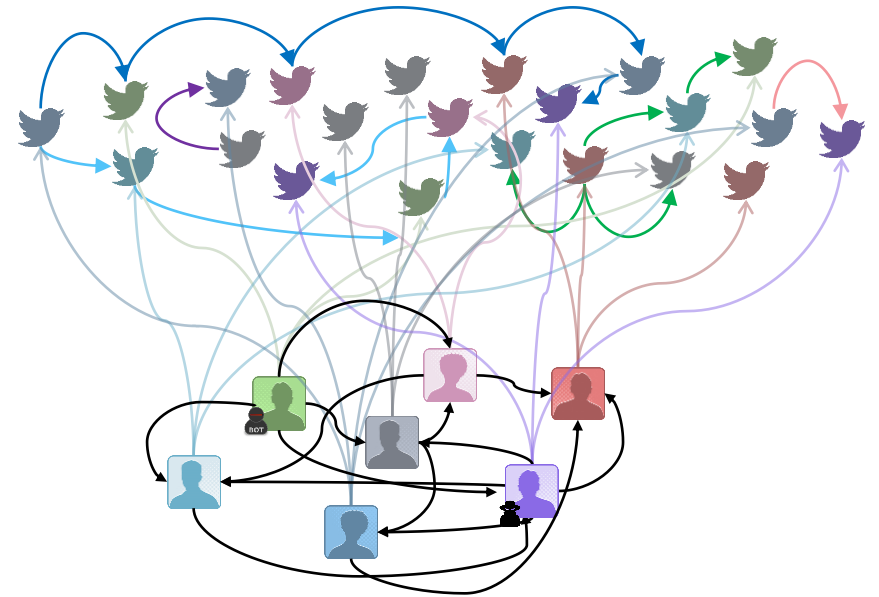 - Debunking process.

More challenges and questions!
How to characterize users?

- Most studies focus on examining users' motivations and attitudes towards adopting a particular social media network, instead of investigating the processes of information (and misinformation) diffusion.

- Define a <u>hierarchical conceptual model</u> to characterise different aspects of the <u>information diffusion process</u>, focusing not only in the <u>information being disseminated</u>, but also on the <u>role of users</u> in such process.

- Shed some light on the psychological and social motivations, and <u>attitudes towards the diffusion and consumption</u> of content in social media.

More challenges and questions!
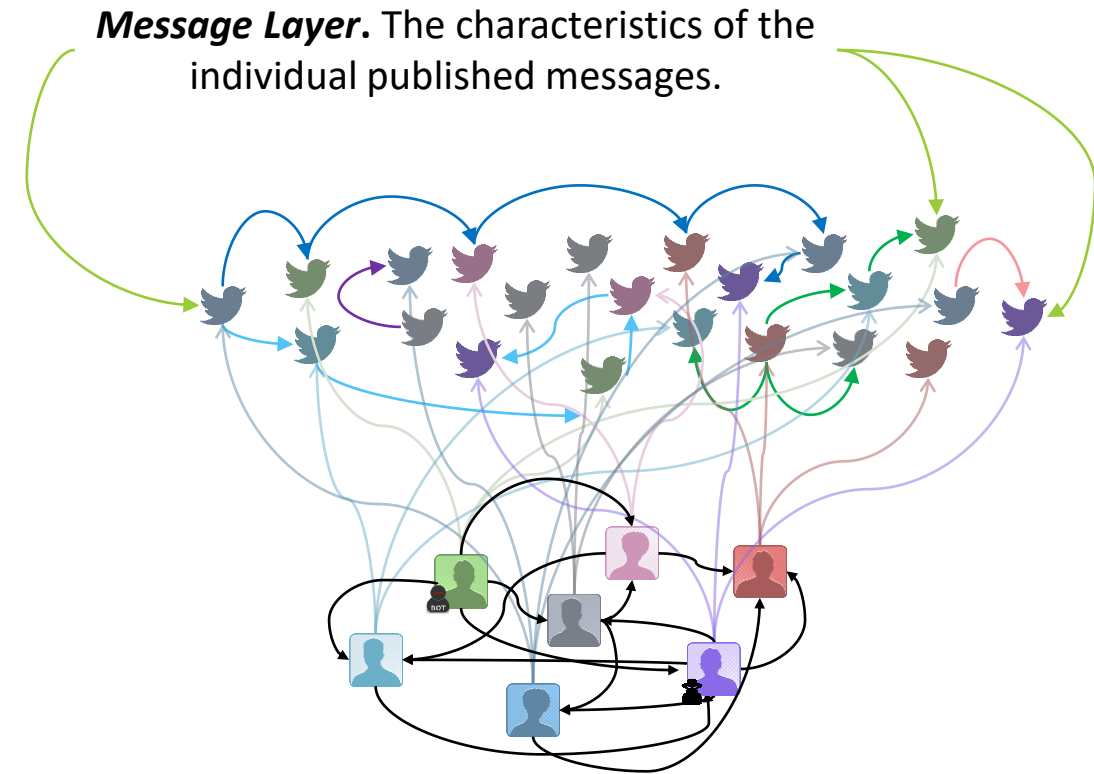## How to characterize users?

- Most studies focus on examining users' motivations and attitudes towards adopting a particular social media network, instead of investigating the processes of information (and misinformation) diffusion.

- Define a hierarchical conceptual model to characterise different aspects of the information diffusion process, focusing not only in the information being disseminated, but also on the role of users in such process.

- Shed some light on the psychological and social motivations, and attitudes towards the diffusion and consumption of content in social media.

Message Layer. The characteristics of the individual published messages.

More challenges and questions!
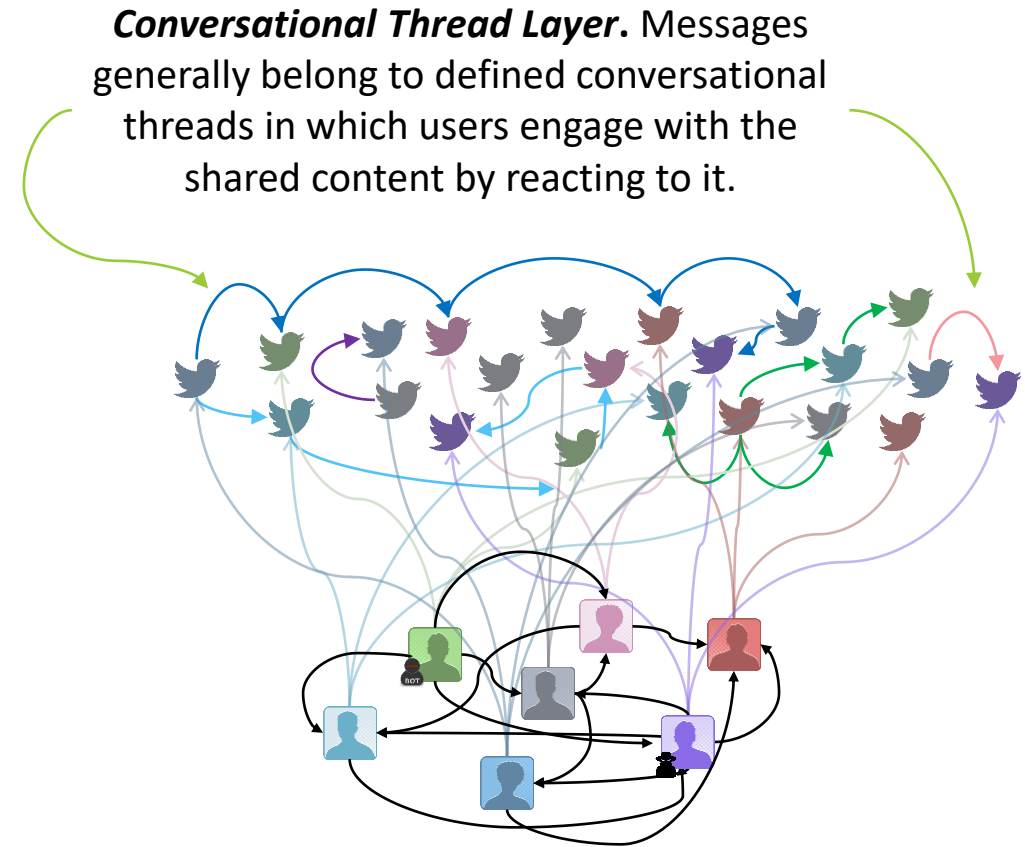## How to characterize users?

- Most studies focus on examining users' motivations and attitudes towards adopting a particular social media network, instead of investigating the processes of information (and misinformation) diffusion.

- Define a hierarchical conceptual model to characterise different aspects of the information diffusion process, focusing not only in the information being disseminated, but also on the role of users in such process.

- Shed some light on the psychological and social motivations, and attitudes towards the diffusion and consumption of content in social media.

Conversational Thread Layer. Messages generally belong to defined conversational threads in which users engage with the shared content by reacting to it.

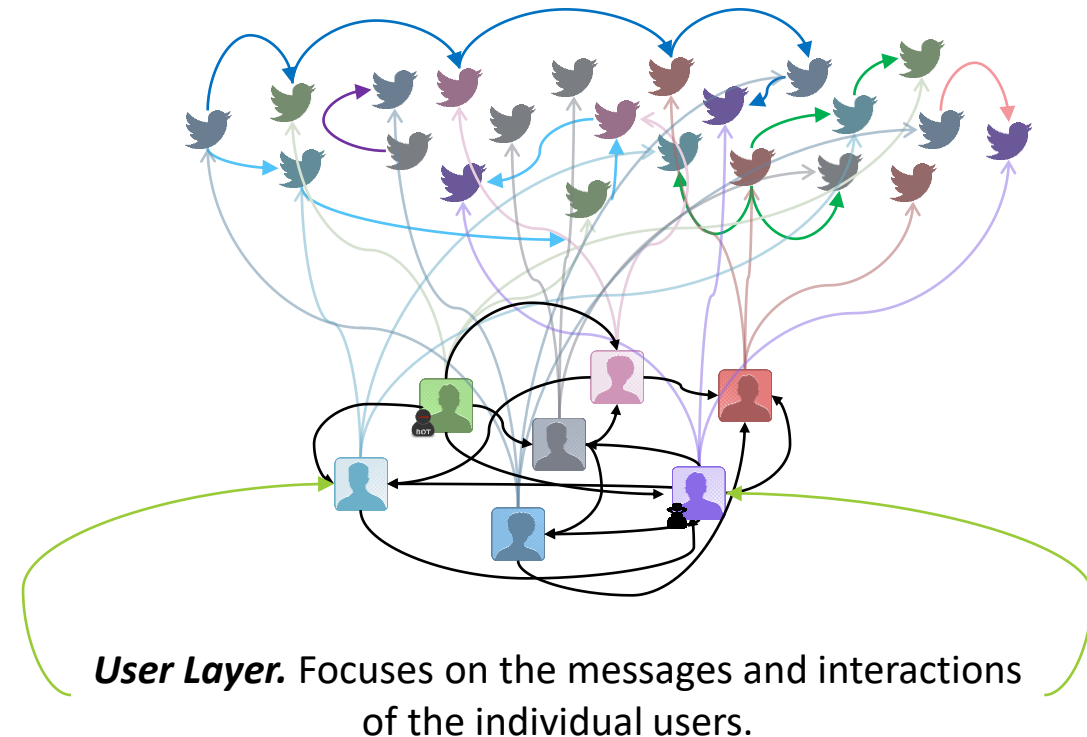# More challenges and questions!
How to characterize users?

- Most studies focus on examining users' motivations and attitudes towards adopting a particular social media network, instead of investigating the processes of information (and misinformation) diffusion.

- Define a <u>hierarchical conceptual model</u> to characterise different aspects of the <u>information diffusion process</u>, focusing not only in the <u>information being disseminated</u>, but also on the <u>role of users</u> in such process.

- Shed some light on the psychological and social motivations, and <u>attitudes towards the diffusion and consumption</u> of content in social media.

User Layer. Focuses on the messages and interactions of the individual users.

More challenges and questions!
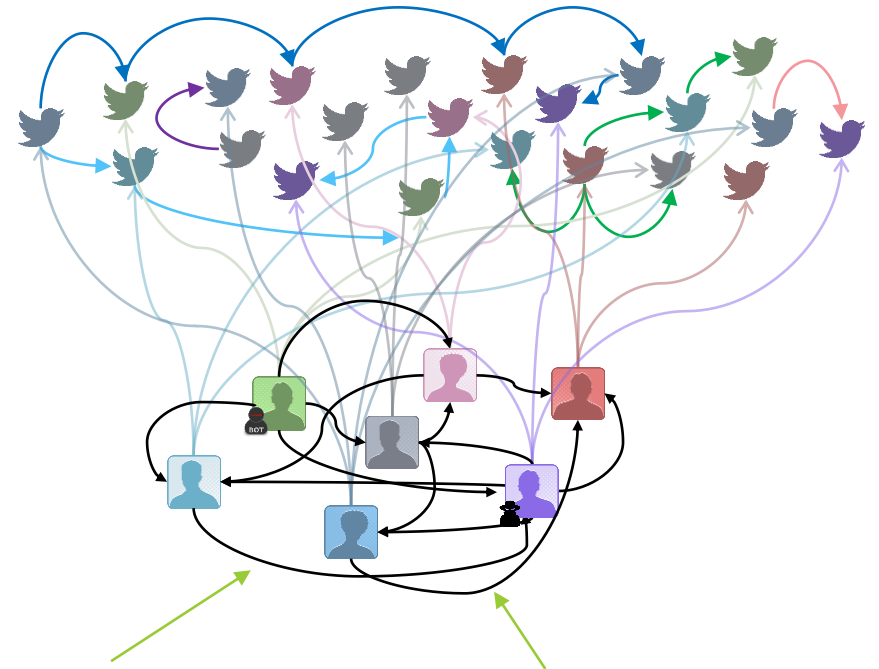## How to characterize users?

- Most studies focus on examining users' motivations and attitudes towards adopting a particular social media network, instead of investigating the processes of information (and misinformation) diffusion.

- Define a <u>hierarchical conceptual model</u> to characterise different aspects of the <u>information diffusion process</u>, focusing not only in the <u>information being disseminated</u>, but also on the <u>role of users</u> in such process.

- Shed some light on the psychological and social motivations, and <u>attitudes towards the diffusion and consumption</u> of content in social media.

Group Layer. Focuses on the behaviour of users in relation to others in terms of their activity and interaction.

More challenges and questions!
Tackling recommender systems

The same concepts in which recommenders rely make them facilitators
of such unwanted diffusion.

- Study how recommender systems affect the dissemination of unwanted content and users.
  - How patterns of exposure and viralization of the recommended elements is modified.

- Design strategies for **assessing** and **explaining** the trustworthiness of the recommended users and content, and **why the recommendations** were made.
  - Based on the current interests and behaviors of users and their exposure and interactions with their social circles.
  - Raise awareness of users towards unreliable content and users, and how to spot them!

More challenges and questions!
What have we already done?

- Explored the dynamics of social networks in terms of homophily.

- Studied the importance of personality and user behaviour in user recommendation.

- Proposed a recommendation technique that adapted over time the recommendation criteria to the characteristics of previously selected friends.

- Studied user influence.

More challenges and questions!
What have we already done?

- Exploited the linked nature of social media for community detection.

- Applied community detection for discovering groups of friends and provide recommendations tailored to the characteristics of each group.

- Studied writing styles in relation to personality and gender.

- Explored the detection of aggressive content and aggressors in the context of cyberbullying.

- Implemented Faking It!, the tool for gathering data!

More challenges and questions!
What have we already done?

- Exploited the linked nature of social media for community detection.

- Applied community detection for discovering groups of friends and provide recommendations tailored to the characteristics of each group.

- Studied writing styles in relation to personality and gender.

- Explored the detection of aggressive content and aggressors in the context of cyberbullying.

We are far from finished!!

- Implemented Faking It!, the tool for gathering data!

WHAT CAN BE DONE TO COUNTER FAKE NEWS?

What can we all do about it?

Do you recognize the source? Have you heard of it before?

Check the headline

Do links look suspicious?

Is formatting unusual?

Check the photos!

How long ago was the story published? Does the timeline make sense?

Check the story!

Understand when a message is forwarded

Check photos and media

Look for messages that look different

Check your biases

Fake news often goes viral

Verify with other sources

What can we all do about it?

What people do when they see a post...

Choose to ignore
- Too long
- Little user engagement
- It is an ad
- Not relevant

Be skeptical
- Looks edited
- Seems unbelievable
- Weird context

Choose not to investigate
- Political burnout
- Not interested
- Hard on smart phone
- No time

Investigate
- Directly investigate claims
- Evaluate information
- Evaluate source
- Compare with other sources
- Use comments

Take at face value
- Trusting the poster
- Confirmation bias
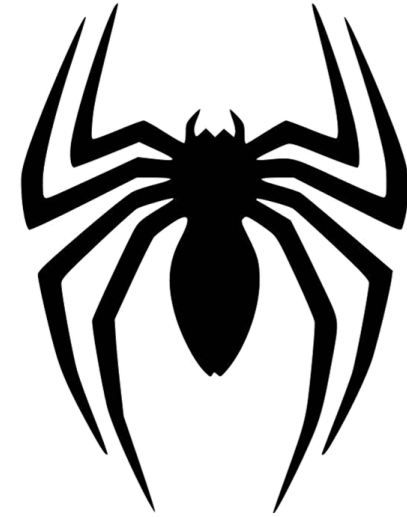- Overconfident
- Other cognitive biases

Engage
- Like or share
- Trust as is

Fake News on Twitter and Facebook: Investigating How People (Don't) Investigate
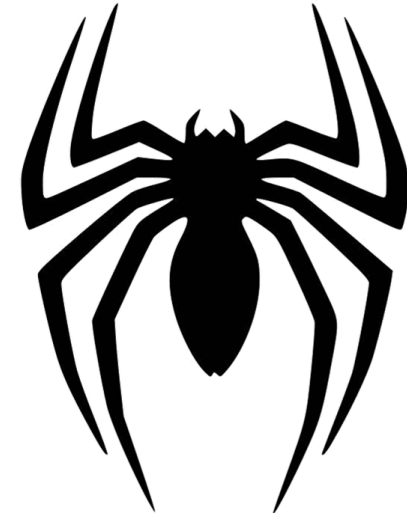https://www.washington.edu/news/2020/03/18/how-people-investigate-fake-news-on-twitter-and-facebook/

What can we all do about it?

With great **power** comes great **responsibility**.

What can we all do about it?

With great **power** comes great **responsibility**.

Question everything!!

Thanks!
Questions?