Summarization

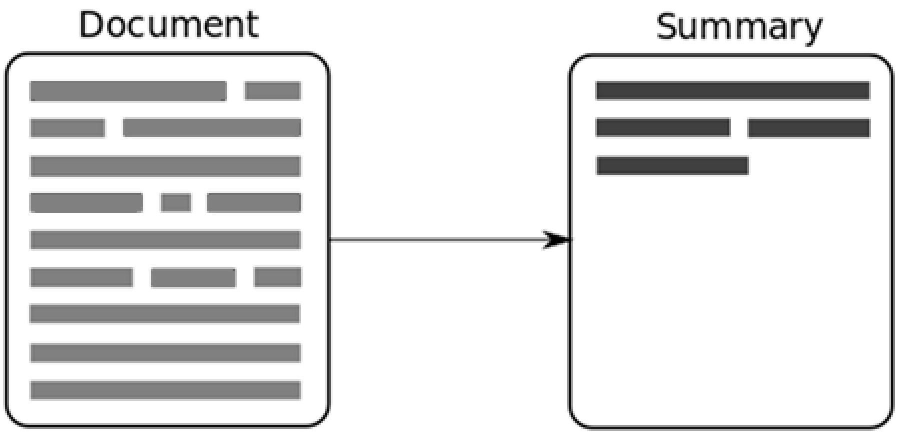
Web Data Mining and Search

Information summarization

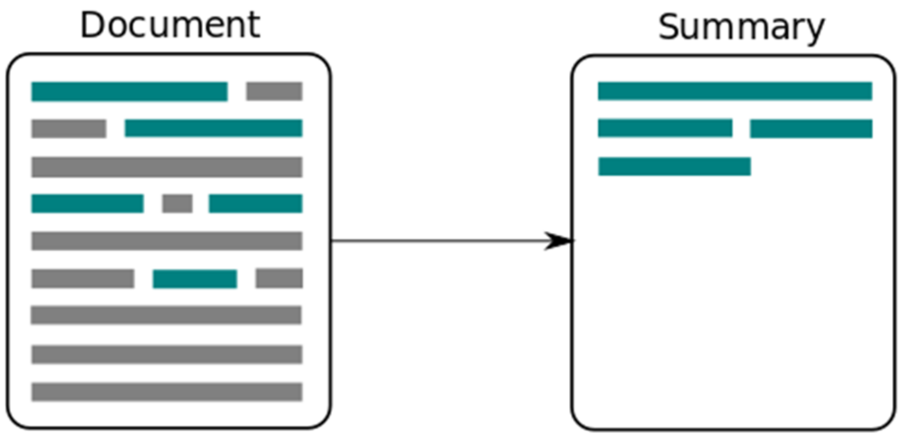
- Massive production of textual information
- Challenge to analyse all information
- Need to read a large number and diverse set of documents
  - News
  - Clinical reports
  - Economic reports
  - Technical reports
  - Standards and guidelines
  - Scienfic articles

Types of summarization

Abstractive



Extractive



Summarization

Input type	Goal	Output type
• Single document • Multiple documents	• Generic • Specific • Query based	• Extractive • Abstractive

Texto: A FCT NOVA foi fundada em 1977 e é uma das mais prestigiadas faculdades de engenharia e ciência do país. Situa-se no Monte da Caparica, a 15 minutos de Lisboa.

Resumo extrativo: A FCT NOVA é uma das mais prestigiadas faculdades do país. Situa-se 15 minutos de Lisboa.

Resumo abstrativo: A FCT NOVA é uma faculdade de prestígio que se situa a 15 minutos de Lisboa.

Abstractive summarization

Summarization example

Daily Mail

The papal bill Pope Francis insisted on paying himself... before catching the bus home after winning the election

- **Pope Francis insisted on returning to his hotel to settle the bill himself**
- **The pontiff also chose to use a bus instead of a chauffeur driven car**
- **The 76-year-old has eschewed ceremonial traditions for a more humble approach**

With the spiritual wellbeing of the world's 1.2 billion Catholics is on his shoulders he must have quite a to-do list.

But despite his new responsibilities, Francis did not forget to stop off - between engagements - to pay his hotel bill.

Staff at the central Rome priests' residence where Bergoglio was staying before the conclave, were astonished when the newly elected Pope strolled in to collect his luggage and settle the bill.

© Getty Images

Pope Francis insisted on returning to the hotel to collect his luggage and greet the staff before settling the hotel bill himself

'I need to set a good example' he joked.

He was driven to the hotel in a simple car and The Rev. Pawel Rytel-Andrianek, who teaches at the nearby Pontifical Holy Cross University and is staying at the residence, said that workers at the hotel were touched by the Pope's decision to return and bid them farewell.

'He wanted to come here because he wanted to thank the personnel, people who work in this house,' he said. 'He greeted them one by one, no rush, the whole staff, one by one.' Mr Rytel-Andrianek added that Francis apparently knew everyone by name.

A Vatican spokesman said: 'He wanted to get his luggage and the bags. He had left everything there.

'He then stopped in the office, greeted everyone and decided to pay the bill for the room... because he was concerned about giving a good example of what priests and bishops should do.'

Francis is already winning plaudits for his down-to-earth manner.

He has so far refused a motorcade and the official papal Jag for official business. And even on the night of the election he insisted on accompanying the other cardinals back to their lodgings, by mini bus, saying: 'I came on the bus, so I'll go home on the bus.'

Meeting cardinals yesterday on his second day of Papal business he eschewed protocol in favour of kissing on two cheeks, shaking hands and hugging.

He told his deputies that old people like himself are 'like good wine, getting better with age', before urging them to impart their wisdom to the young.

Francis began his reign in unorthodox fashion as he shunned public events in order to pray to the Virgin Mary.

During his first Mass since being elected as supreme pontiff, Pope Francis and his cardinals were dressed in simple yellow robes over their cassocks, rather than the formal ceremonial outfits they would normally wear on such a major occasion.

Speaking in Italian without notes, he said: 'We can walk all we want, we can build many things, but if we don't proclaim Jesus Christ, something is wrong. We would become a compassionate NGO and not a Church which is the bride of Christ.

'He who does not pray to the Lord prays to the devil. When we don't proclaim Jesus Christ, we proclaim the worldliness of the devil, the worldliness of the demon.'

'We must always walk in the presence of the Lord, in the light of the Lord, always trying to live in an irreprehensible way,' he said in a heartfelt homily of a parish priest, loaded with biblical references and simple imagery.

'When we walk without the cross, when we build without the cross and when we proclaim Christ without the cross, we are not disciples of the Lord. We are worldly,' he said

'We may be bishops, priests, cardinals, popes, all of this, but we are not disciples of the Lord,' he said.

It was a far simpler message than the dense, three-page discourse Benedict delivered in Latin during his first Mass as pope in 2005.

The difference in style was a sign of Francis' belief that the Catholic Church needs to be at one with the people it serves and not impose its message on a society that often doesn't want to hear it, Francis' authorised biographer, Sergio Rubin, said.

Francis took the helm of the 1.2 billion-member Church at a time of strife and intrigue, with the Vatican rocked by a string of sex abuse scandals, accusations of infighting within its central government and by allegations of financial wrongdoing.

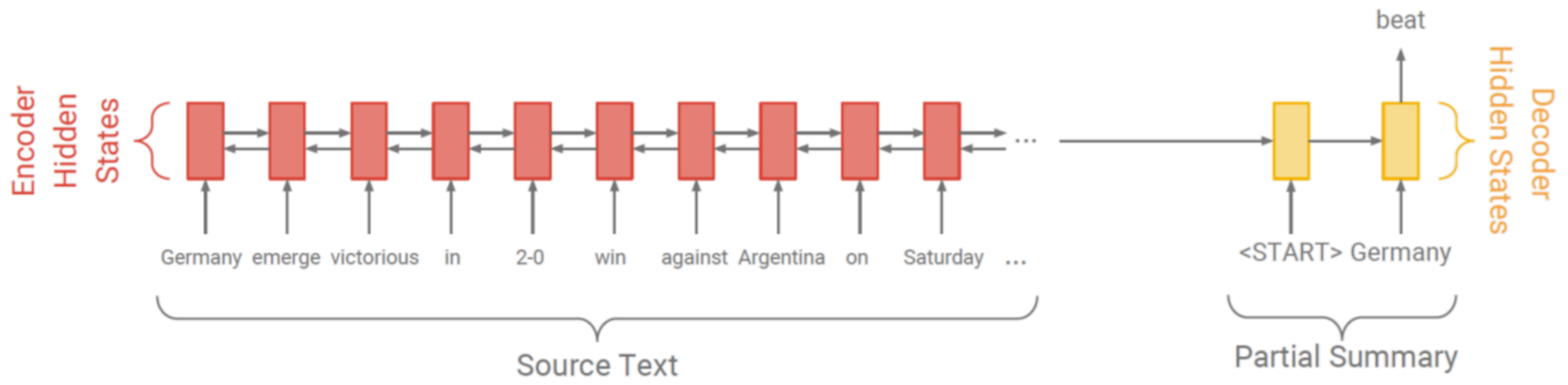
But many within the church believe he could change it for the better.

'It seems to me for now what is certain is it's a great change of style, which for us isn't a small thing,' Mr Rubin said, recalling how the former Cardinal Jorge Bergoglio would celebrate Masses with homeless people and prostitutes in Buenos Aires.

'He believes the church has to go to the streets,' he said, 'to express this closeness of the church and this accompaniment with those who are suffering.'

See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In ACL.
https://github.com/ymfa/seq2seq-summarizer

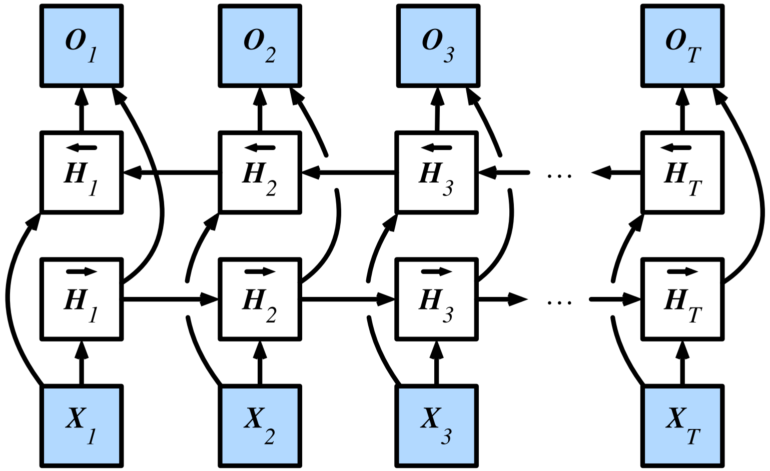
Encoder-decoder based summarization



See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In ACL.
https://github.com/ymfa/seq2seq-summarizer

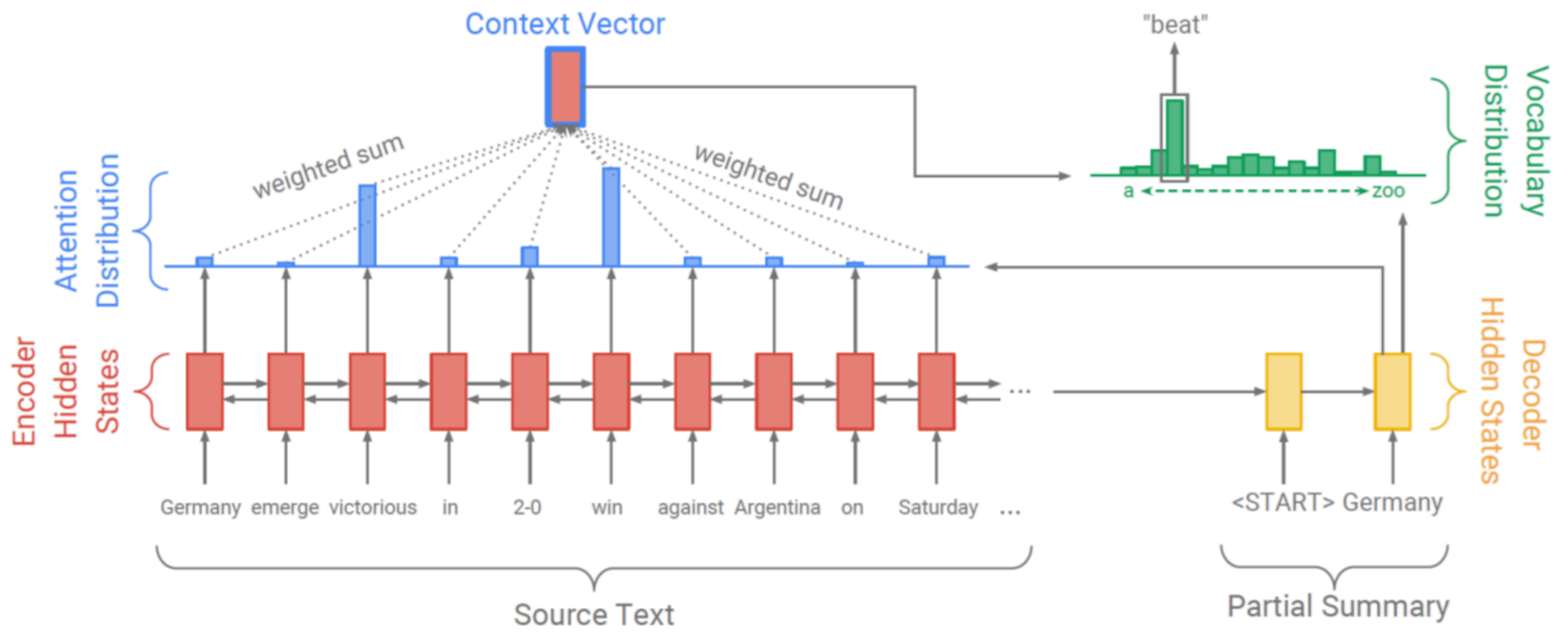
Bidirectional RNNs

- There are many tasks where the prediction is in the middle of the sequence:
  - I am _____
  - I am _____ very hungry.
  - I am _____ very hungry, I could eat half a pig



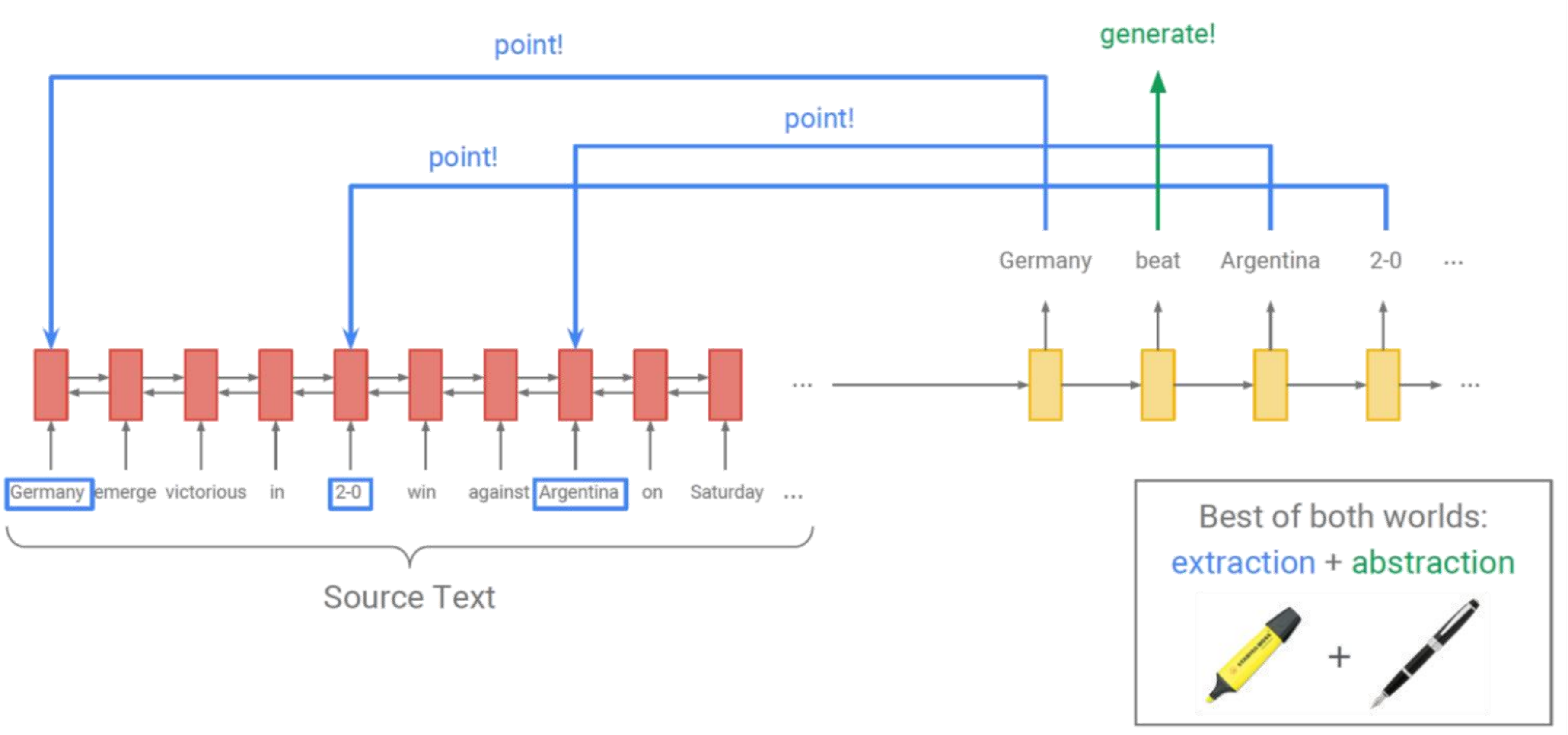
- In a bidirectional RNN information from both ends of the sequence is used to estimate the output.

Attention based summarization



See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In ACL.
https://github.com/ymfa/seq2seq-summarizer

Get to the point!



See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In ACL.
https://github.com/ymfa/seq2seq-summarizer

Extractive summarization

NFL MLB NCAA FB NASCAR Soccer NBA Boxing UFC NCAA BK ••• Live TV Highlights Shows

Other Sports | Home Videos

Tour de France cyclist walks away from terrifying mountainside crash

…

The Tour de France is a beautiful, inimitable and completely horrifying spectacle of sport. It's one of the rare athletic events where you can (…)

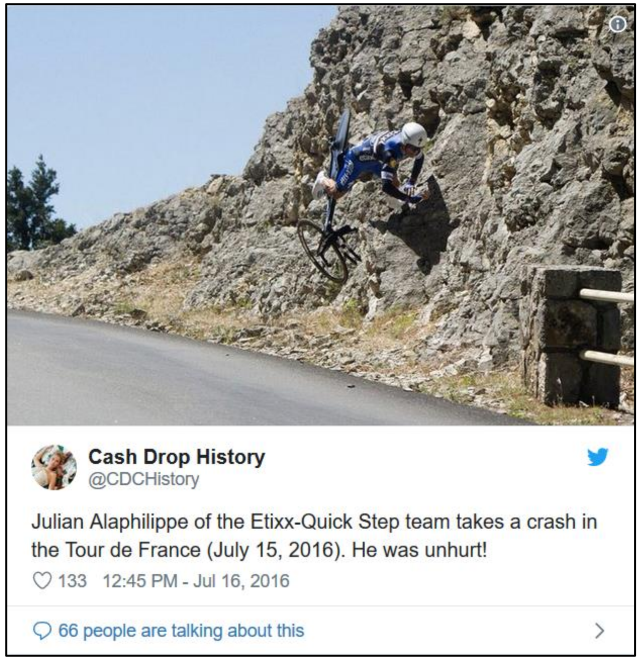
…

And unfortunately for French cyclist Julian Alaphilippe, he was the one providing the grim visuals during Friday's time trial race.

…

According to LeDauphine.com (h/t Washington Post's Marissa Payne), Alaphilippe was motoring along at 32 miles per hour when a rogue gust of wind pushed him off the road and up over his handle bars into the jagged cliffside.

…



https://www.bbc.com/news/science-environment-46547254

Social media in the newsroom

- Increasingly, social-media content is being used by news agencies.
  - Unique, high-value view point;
  - Immediacy;
  - Number of perspectives;
  - Amount available.

- This brings new challenges:
  - Variable content quality;
  - Finding relevant content;
  - Aligning content with a storyline.

Users as social media sensors

- Information published on Twitter is fresh and most of the time relevant
- Users are social sensors of live events providing live information
- News reporters can explore this proximity of users to live events to get front-line reports

Social media summarization

How to create visual storylines to illustrate news pieces using social media content?

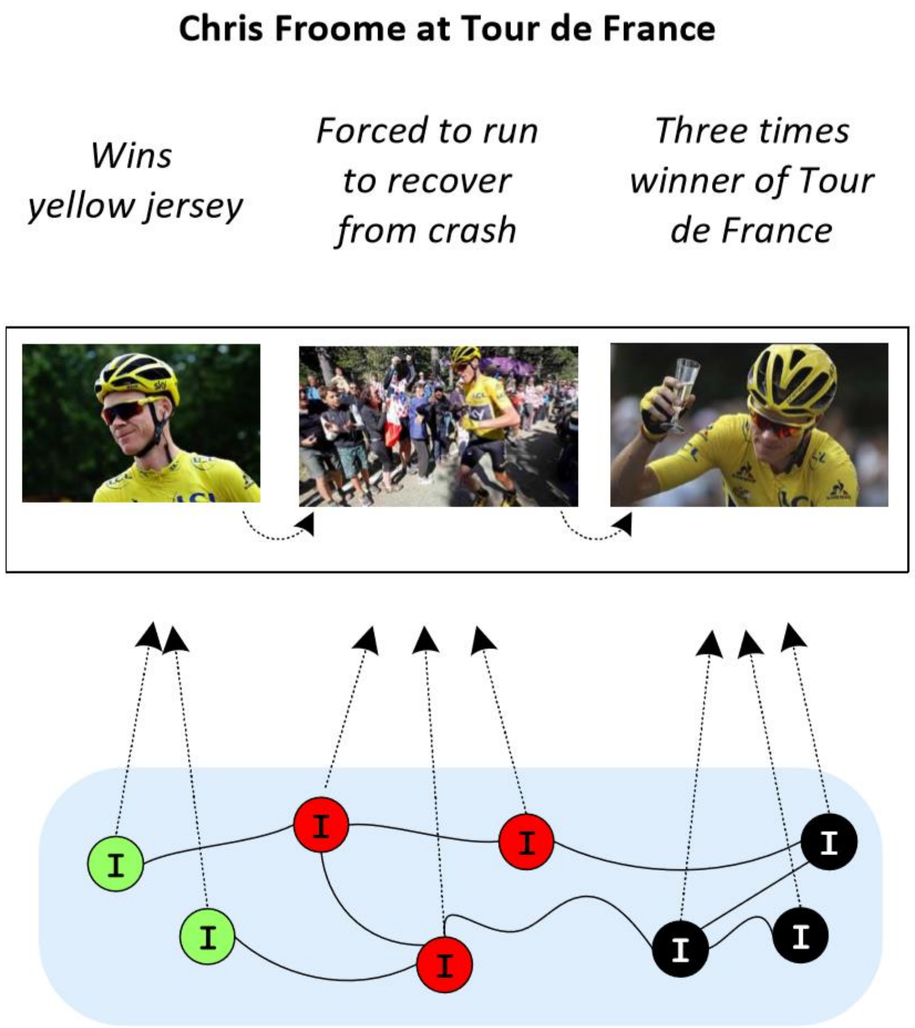
Chris Froome at Tour de France 2016 [1]

Chris Froome
Wins Yellow Jersey

Forced to run to
recover from crash

Three times winner

[1] https://www.bbc.com/sport/cycling/36879128



Story Topic

Story segments

$$Story_j = (u_1, u_2, u_3, \dots, u_N)$$

Visual storyline

$$Storyline_j = (w_1, w_2, w_3, \dots, w_N)$$

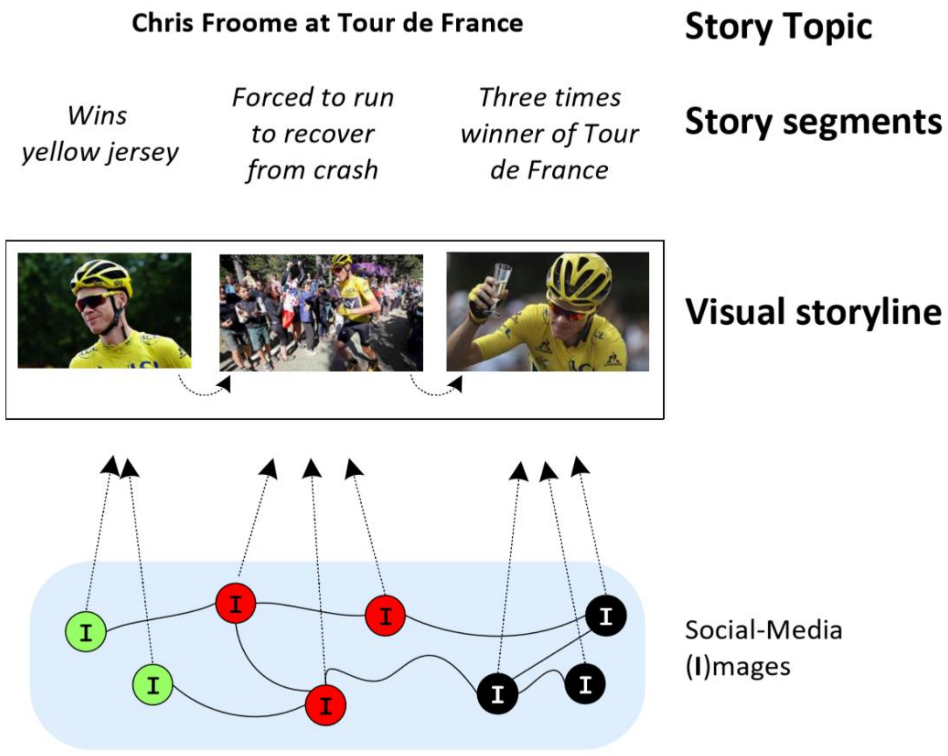
$$\forall i \in [1, N]\; w_i \in D$$

Social-Media (**I**)mages

$$Set\ of\ social\ media\ images\ D$$

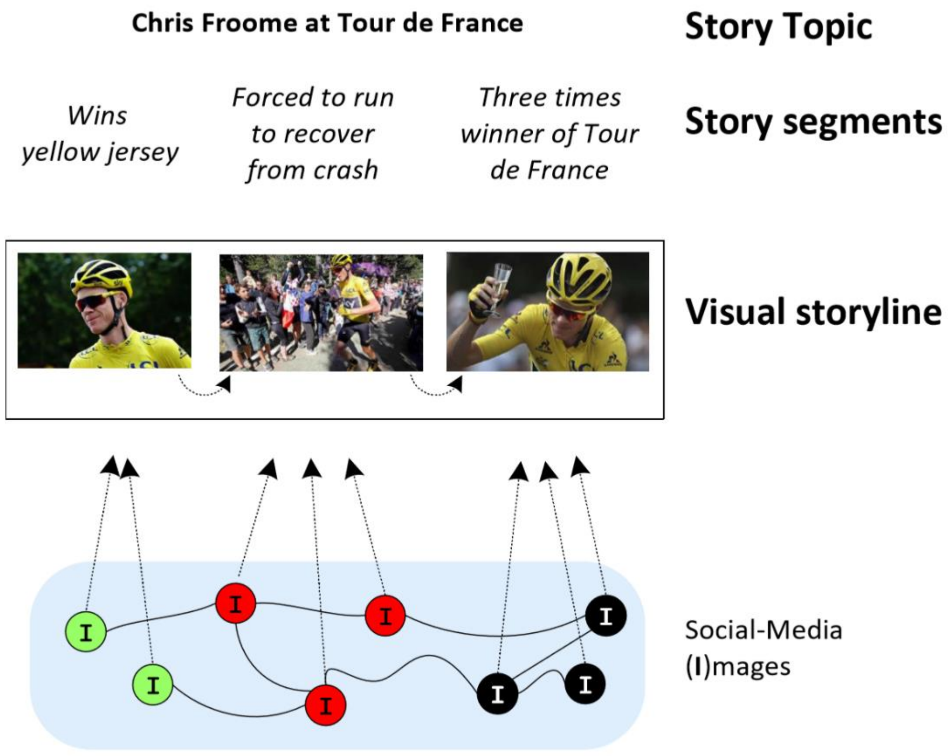
Processing steps

- How to select only **high-quality** content?
- How to **define** and **organize** the story?
- How to create a **relevant** summary?
- How to create a **coherent** and **non-redundant** summary?

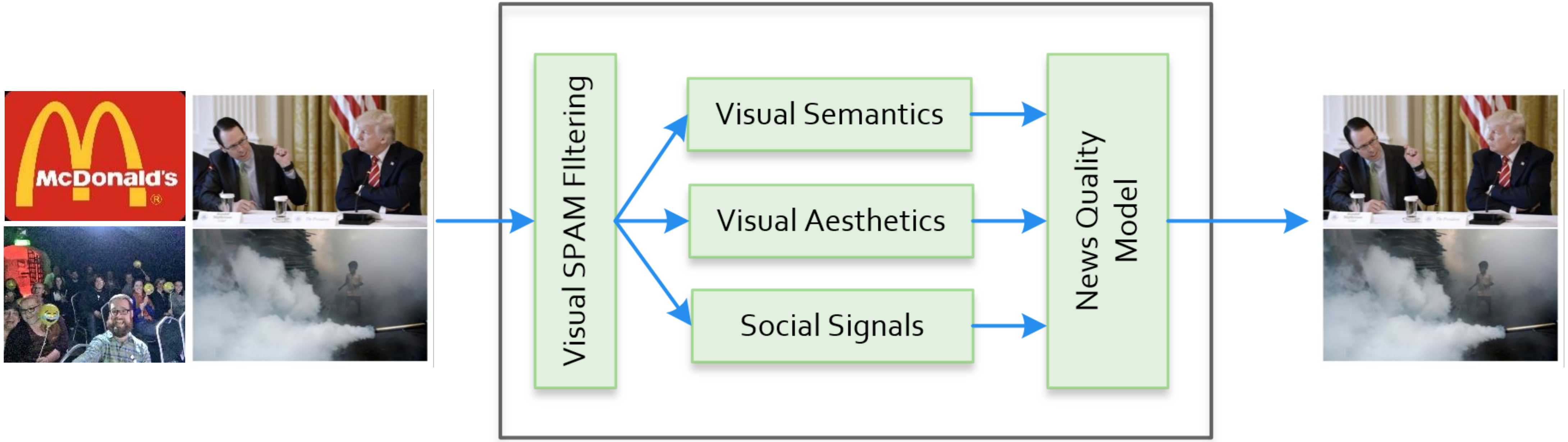


Processing steps

- **How to select only high-quality content?**
- How to **define** and **organize** the story?
- How to create a **relevant** summary?
- How to create a **coherent** and **non-redundant** summary?



Ranking by news-quality



G. Marcelino, R. Pinto, J. Magalhaes, Ranking news-quality multimedia , ACM International Conference in Multimedia Retrieval 2018. (*Best paper nomination*).

Spam detection

- Near-duplicate detection
  - pHash [1]

- Captioned images filter
  - Tesseract OCR [2].

- Synthetic image detection

[1] http://phash.org/
[2] Ray Smith. An overview of the tesseract ocr engine. In Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on , volume 2. IEEE,2007.

Spam detection

- Near-duplicate detection
  - pHash [1]
- Captioned images filter
  - Tesseract OCR [2]
- Synthetic image detection

[1] http://phash.org/
[2] Ray Smith. An overview of the tesseract ocr engine. In Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on , volume 2. IEEE,2007.

Spam detection

- Near-duplicate detection
  - pHash [1]
- Captioned images filter
  - Tesseract OCR [2]
- Synthetic image detection

[1] http://phash.org/
[2] Ray Smith. An overview of the tesseract ocr engine. In Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on , volume 2. IEEE,2007.

Synthetic image detection

- Features:
  - 3 Color fraction features.
  - Number of dominant colors.
  - Edge Histogram.
  - Number of corners.
  - Luminance.

- Model:
  - Logistic Regression
  - L1 penalty and λ=1.0
- Trained and tested:
  - NIPC dataset [1]
  - Augmented NIPC dataset (with social media images)

Best feature set	Precision	Recall	F-measure
NIPC trained	0.97	0.97	0.97
NIPC-Twitter trained	0.91	0.91	0.91

Novelty

- Previous research addressed image aesthetics [1], memorability [2] and interestingness [3].

- News media content has very specific characteristics.
- News-worthy images are informative, interestingness, memorable, and when possible have good visual aesthetics.

Quality: is this photo publishable by a news agency?

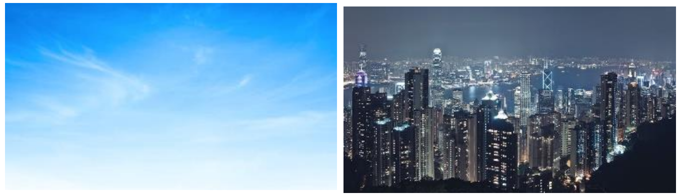
[1] Luo, Yiwen, and Xiaoou Tang. "Photo and video quality evaluation: Focusing on the subject." *European Conference on Computer Vision*. Springer, Berlin, Heidelberg, 2008.
[2] Isola, Phillip, et al. "Understanding the intrinsic memorability of images." Advances in Neural Information Processing Systems. 2011.
[3] Gygli, Michael, et al. "The interestingness of images." Computer Vision (ICCV), 2013 IEEE International Conference on. IEEE, 2013.

Social and semantic and visual features

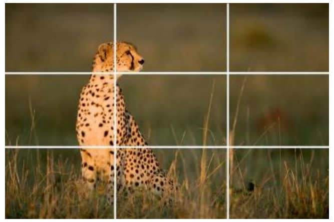
- Social (related to popularity, interestigness):
  - Number of retweets and followers.
  - Number of times a duplicate image was posted.
  - Number of times a near-duplicate image was posted.
- Semantic (related to memorability):
  - Distribution of concepts across news images.
  - Distribution of concepts across non-news images.
- Visual (related to aesthetics and interestingness)

Orientation

#Edges



Rule of 1/3



Luminance

Focus

#Faces

Area

Aspect Ratio

Entropy

Color simplicity

Colorfulness

Ranking and filtering by news quality

- **Visual Features**
  - Visual quality and aesthetics.
- **Semantic Features**
  - Probability of topic being news related.
- **Social Features**
  - Interestingness and informativeness.

- **Gradient Boosted Trees**
  - High precision.
  - Works with continuous and categorical data (e.g. *orientation and aspect ratio*).
  - Works with small and large datasets (*critical for expensive ground-truth*)
  - Is able to deal with non-linear relationships in the data.
  - Retains *some interpretability*.

Groundtruth

Train

- 500 images
  - 400 from social media (twitter)
  - 100 from news media
- Crowd sourcing with 7 annotators.

Agreem.	Images	High quality	LQ/HQ ratio
57%	124	58	1.14
71%	129	55	1.35
86%	144	39	2.69
100%	103	17	5.06
78%	500	169	1.96

Test

- Using a dataset of 1500 images from social media.
- Results pooling using 4 different models with:
  - Visual features only (GBTv)
  - Semantic features only (GBTc)
  - Social features only (GBTs)
  - Visual, Semantic and Social features (GBTf)
- Crowd sourcing with 5 annotators.

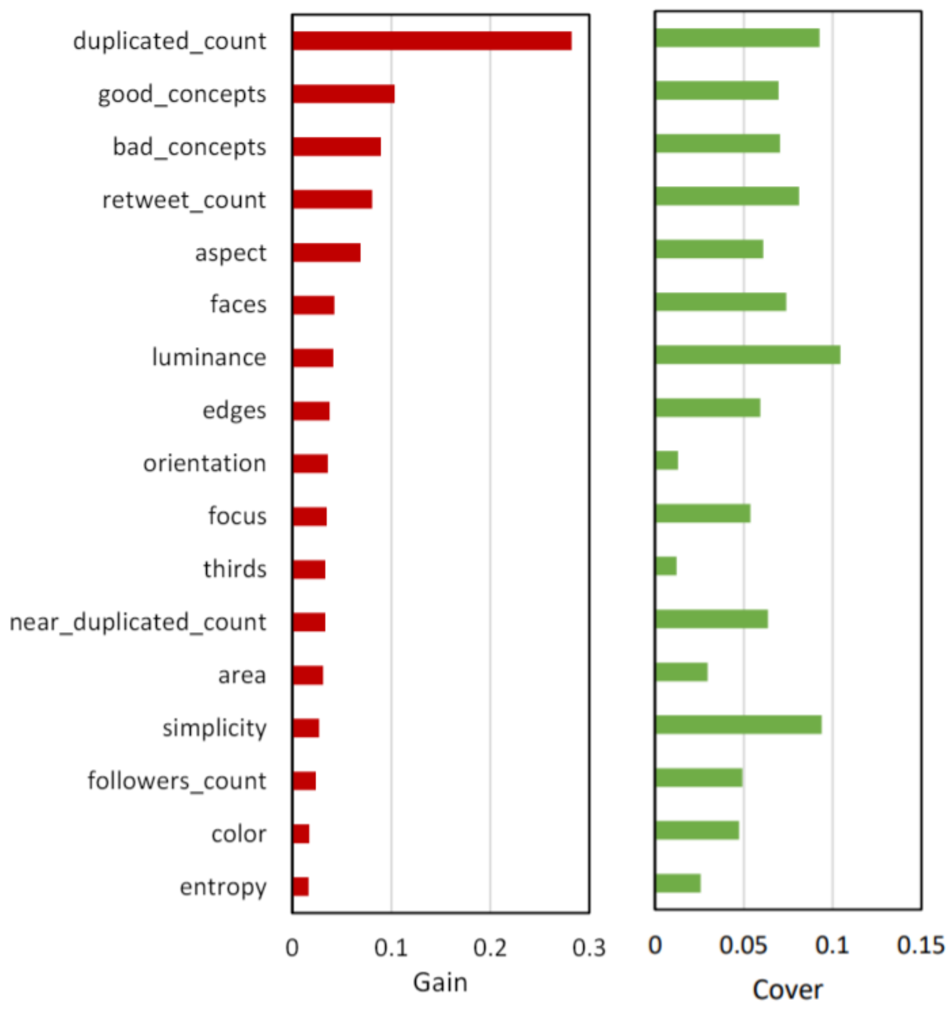
Evaluation

Features	Prec@30	nDCG@50	MAP
GBT_V	0.833	0.837	0.448
GBT_C	0.833	0.859	0.532
GBT_S	0.733	0.836	0.454
GBT_F	0.967	0.906	0.645

- GBT_V uses aesthetic features
- GBT_C uses memorability features

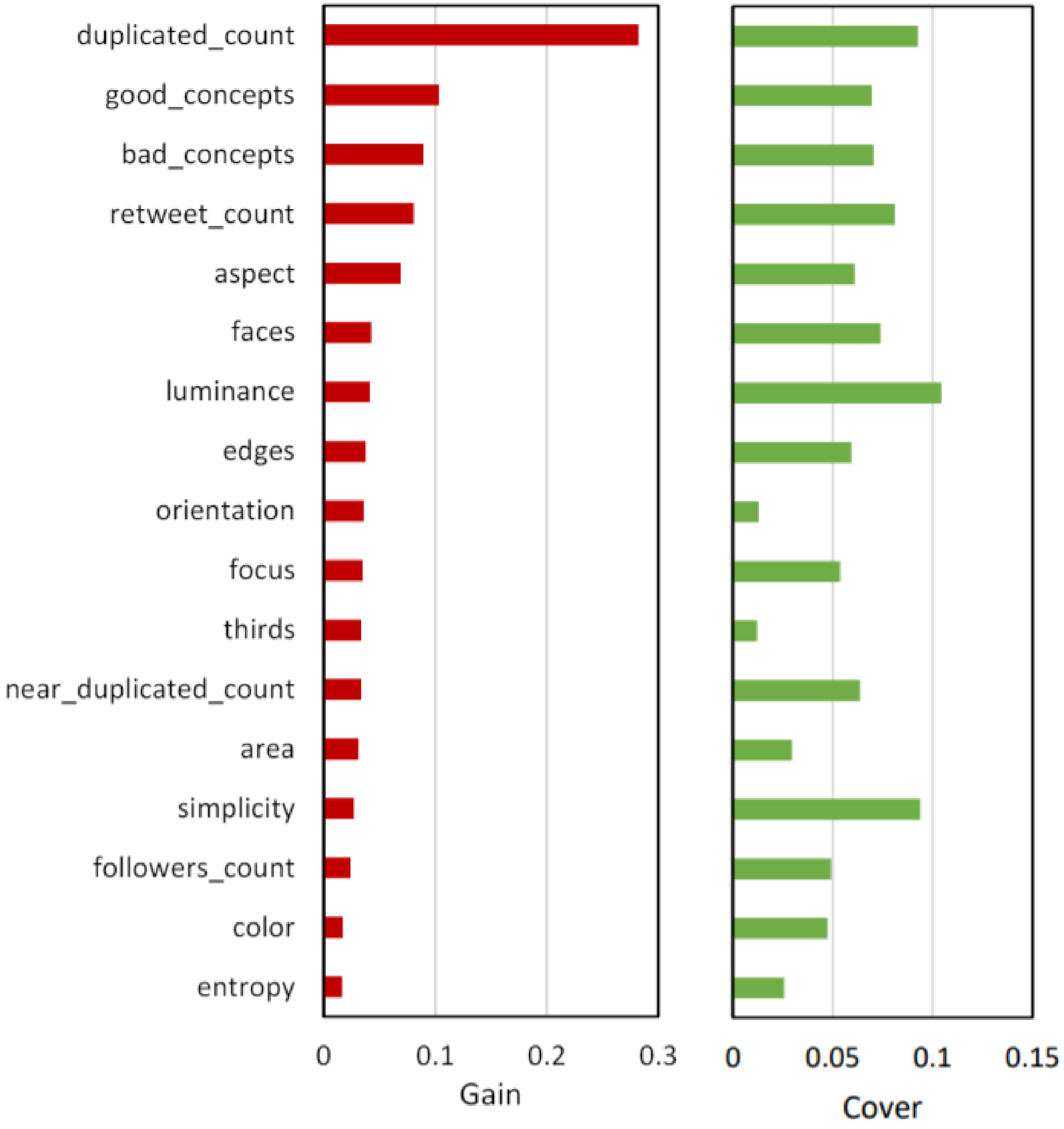
- Only combining all three feature sets in GBT_F was it possible to to attain the best results.

Interpretability



Explaining news quality

- **Gain** is the improvement in precision that was attained by splitting branches with the feature.
- **Cover** is the number of times a feature is used in the threes.
- By inspecting the **Gain** and **Cover** of each feature we confirm the importance of all three feature sets in improving the precision of the model.



Visual features

- #Edges
- Rule of 1/3
- Focus
- Aspect Ratio
- Orientation
- Colorfulness
- Faces
- Luminance
- Area
- Entropy
- Simplicity

Luminance$^{\uparrow}$, Focus$^{\uparrow}$, Color$^{\uparrow}$

Luminance$^{\uparrow}$, Focus$^{\uparrow}$,

Luminance$^{\downarrow}$, Focus$^{\downarrow}$

Aspect$^{\downarrow}$, Faces$^{\uparrow}$ Focus$^{\downarrow}$ Entropy$^{\downarrow}$

Best

Worst

Semantic features

- **Good concepts** proportional to the **number of concepts** in the image that are **commonly** found **on news media**.
- **Bad concepts** proportional to the **number of concepts** in the image that are **not commonly** found **on news media**.

Performing Arts$^{\uparrow}$, Event$^{\uparrow}$, Stage$^{\uparrow}$

Event$^{\uparrow}$, Festival$^{\uparrow}$

(No interesting concepts)

Selfie$^{\downarrow}$

Best

Worst

Social features

- Number of times an image appeared in the input set.
- Number a near-duplicate image appeared in the input set.
- Number of retweets.
- Number of followers.

#Duplicates$^{\uparrow}$, #Retweets$^{\uparrow}$

#Duplicates$^{\uparrow}$, #Retweets$^{\downarrow}$

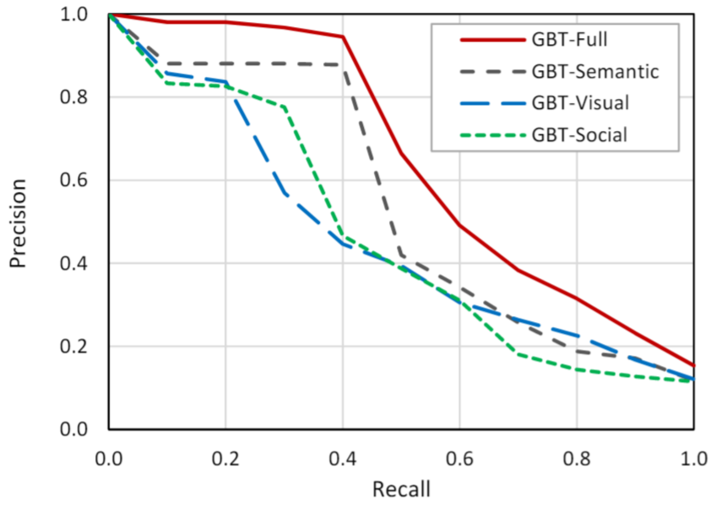
#Retweets$^{\uparrow}$ #Duplicates$^{\downarrow}$

#Duplicates$^{\downarrow}$, #Retweets$^{\downarrow}$

Best

Worst

Evaluation of the ranking model



Features	Prec@30	nDCG@50	MAP
GBT_V	0.833	0.837	0.448
GBT_C	0.833	0.859	0.532
GBT_S	0.733	0.836	0.454
GBT_F	0.967	0.906	0.645

Visual$^{\uparrow}$, Social$^{\uparrow}$, Semantic$^{\uparrow}$

Semantic$^{\uparrow}$, Visual$^{\uparrow}$

Social$^{\downarrow}$, Semantic$^{\uparrow}$

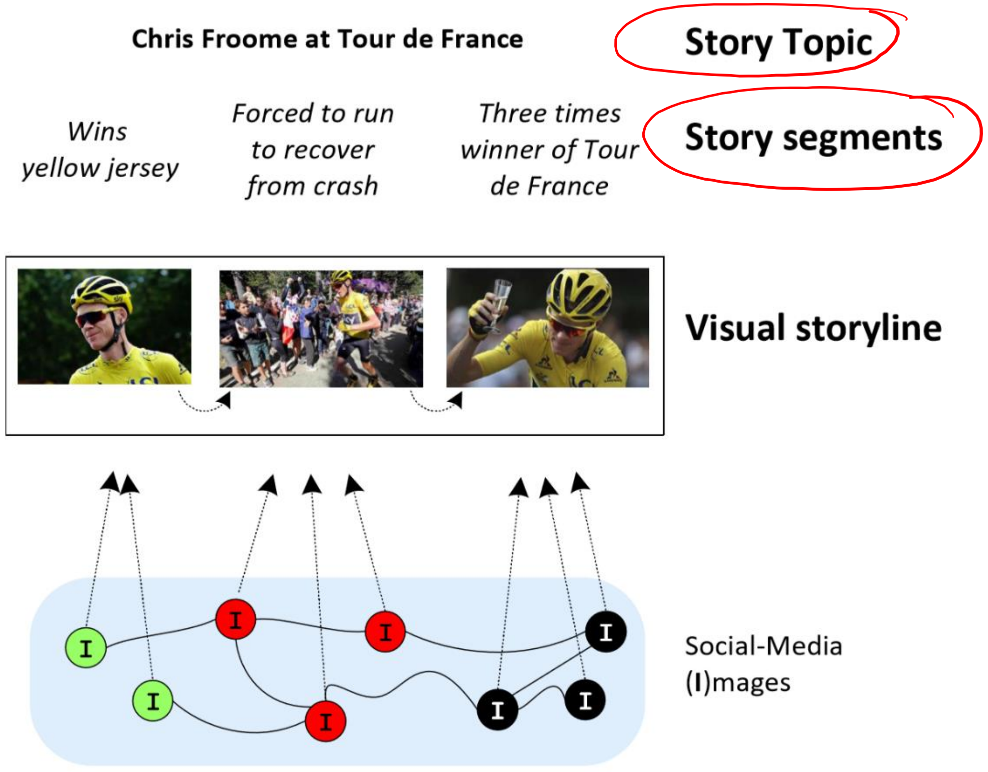
Visual$^{\downarrow}$, Social$^{\downarrow}$, Semantic$^{\downarrow}$

Best

Worst

Processing steps

- How to select only **high-quality** content?
- **How to define and organize the story?**
- How to create a **relevant** summary?
- How to create a **coherent** and **non-redundant** summary?



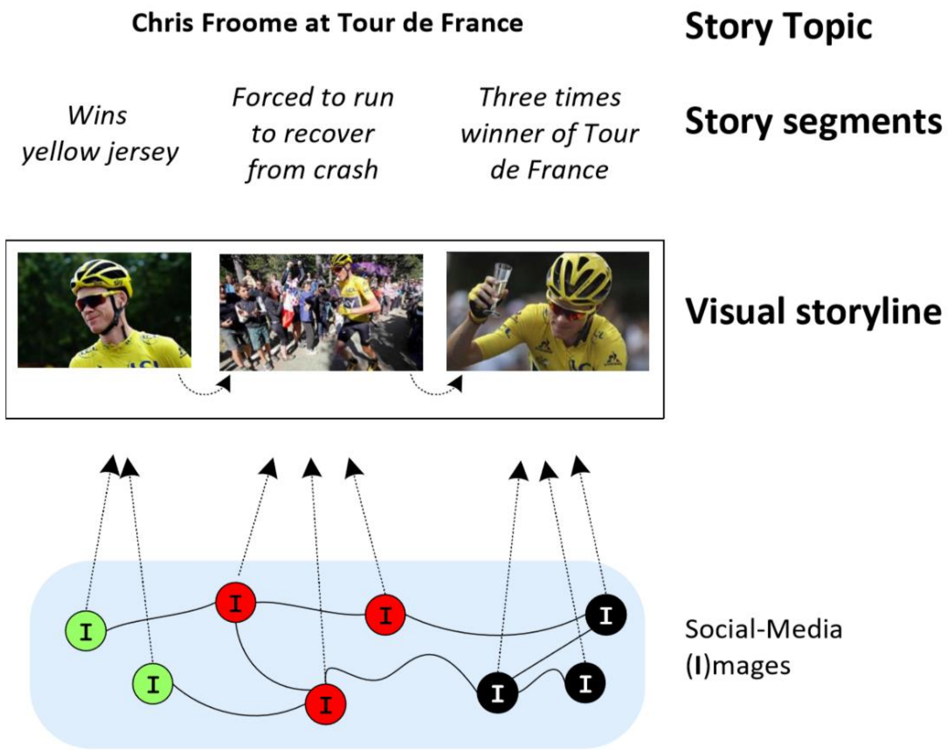
Assisted reporting of a story

- Allow media reporters to:
  - Explore social media in a structured way
  - Select the story topic
  - Organize the summary content around the story they wish to create.

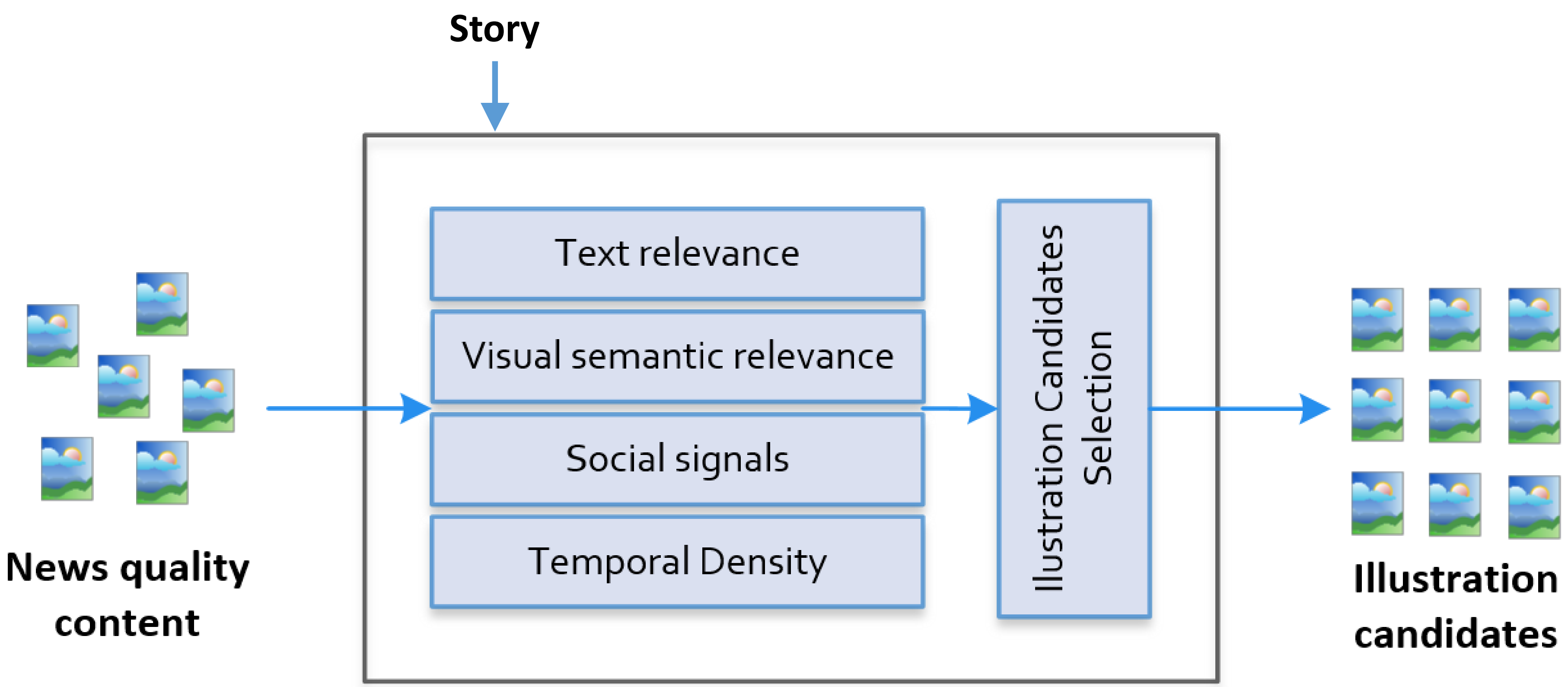
- The system can then provide reporters with the most **relevant** and **coherent** information.

Processing steps

- How to select only **high-quality** content?
- How to **define** and **organize** the story?
- **How to create a relevant summary?**
- How to create a **coherent** and **non-redundant** summary?



Selecting candidate documents



Ranking candidate images

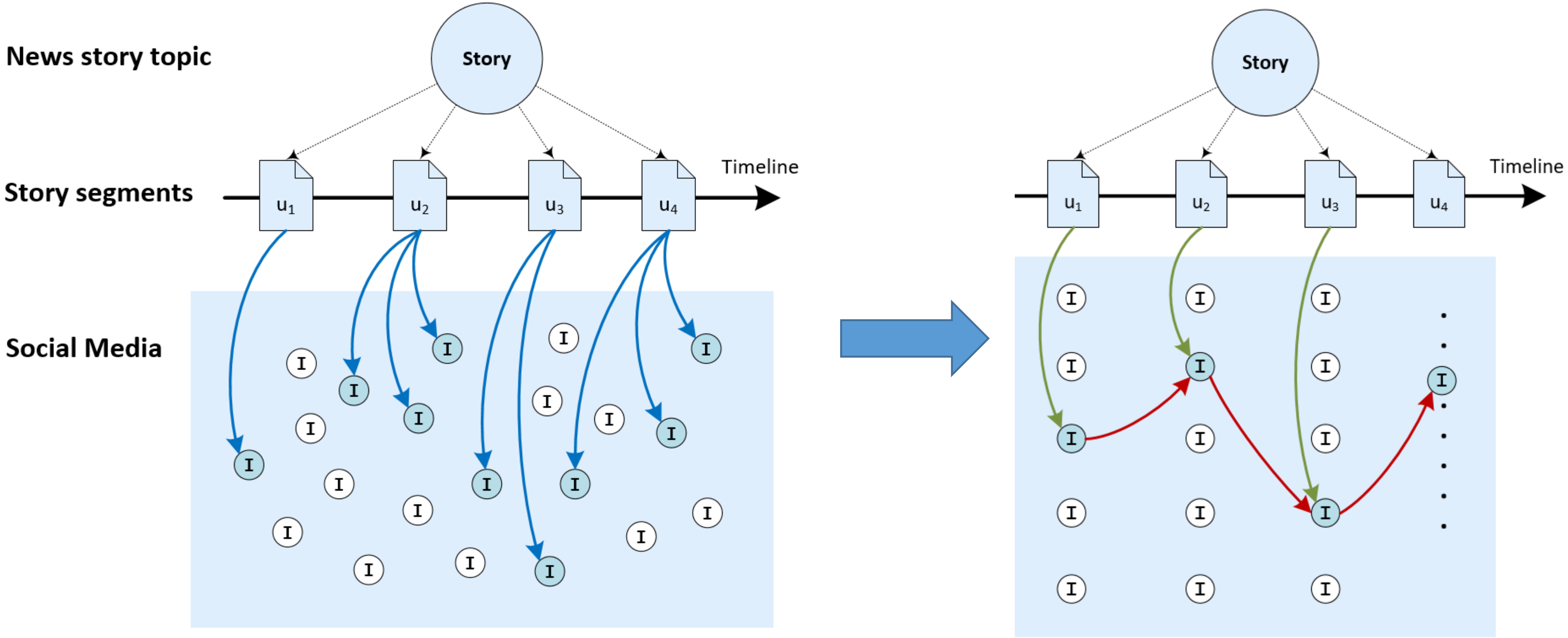
- Text retrieval techniques.
- Multimedia retrieval techniques:
  - Image concepts.
- Social media posts metadata:
  - Social traction;
  - Time of publication.

- Baselines:
  - ***BM25* (Text retrieval)**
  - *#Retweets* (Social traction)
  - *#Duplicates* (Social traction)
  - *Concept Pool* (Image concepts)
  - *Concept Query* (Image concepts)
  - *Temporal Modeling* (Time of publication)

Ranking candidate text documents

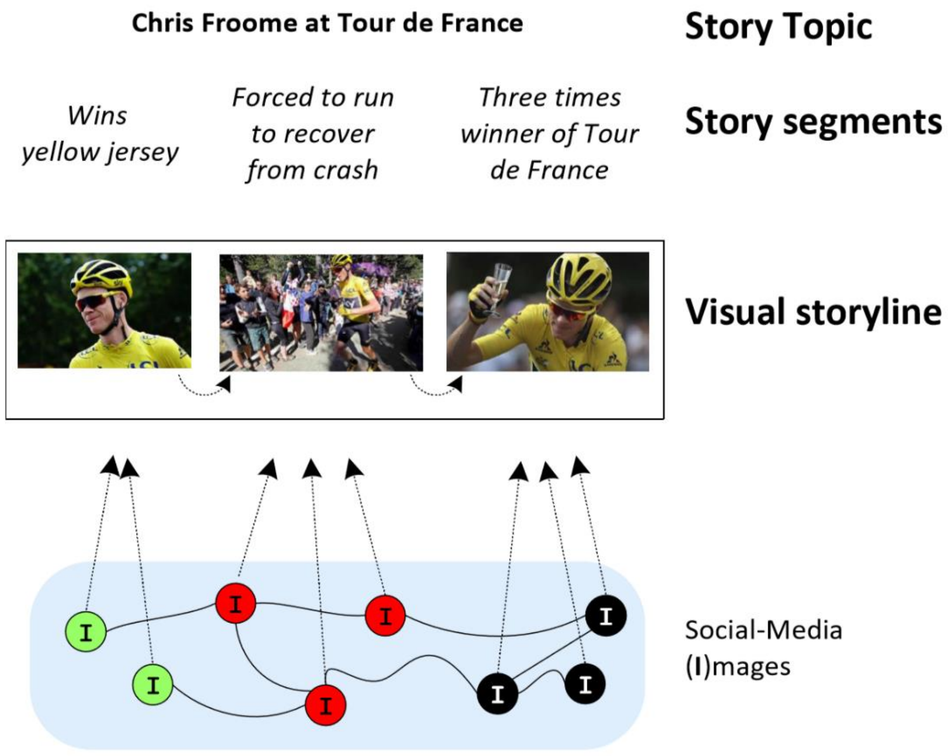
- **BM25**: using the BM25 retrieval model on publications' text.
- **#Retweets**: BM25 and re-ranking the top 20 posts by number of retweets.
- **#Duplicate**: BM25 and re-ranking the top 20 posts by number of duplicates.
- **Concept Pool**: BM25 and extracting visual concepts, using a pre-trained VGG network, from the top 10 ranked posts. The top 20 ranked posts are then re-ranked according to the number of visual concepts in the pool.
- **Concept Query**: BM25 and extracting visual concepts from top 10 ranked posts, creating a new query with those concepts. A new rank is created using the new query. We fuse the two ranks using Reciprocal Rank Fusion.
- **Temporal Modeling**: BM25 and creating a Kernel Density Estimation with the probability of a publication being posted at a given date. The publications that maximize that probability are chosen.

Relevant documents per story segment



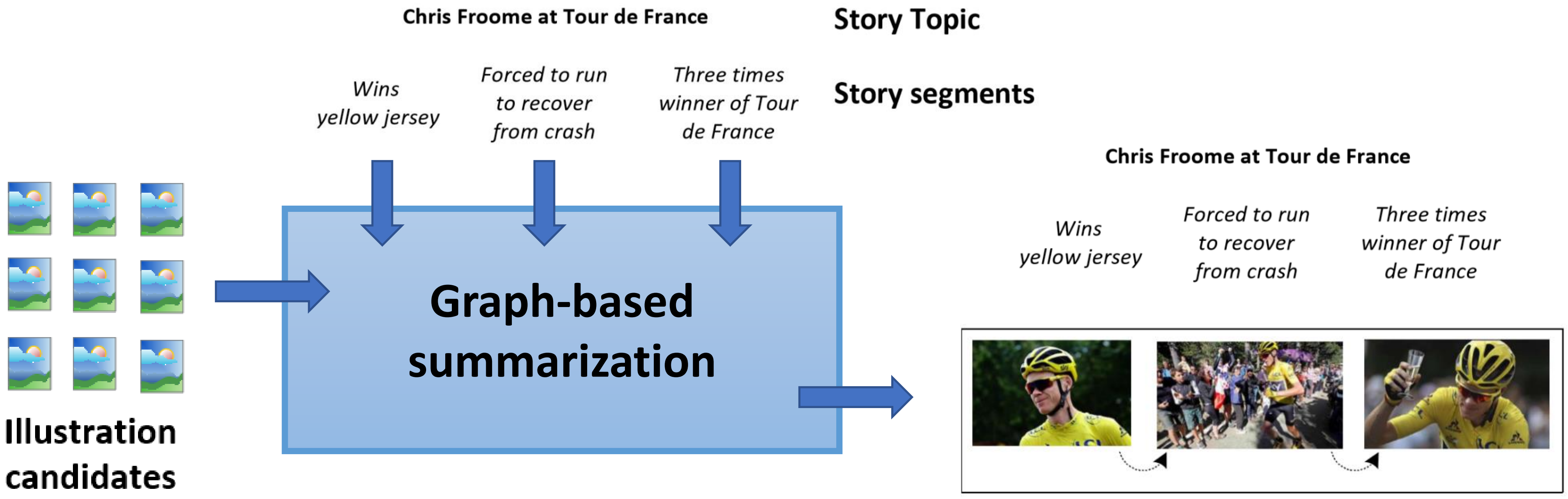
Processing steps

- How to select only **high-quality** content?
- How to **define** and **organize** the story?
- How to create a **relevant** summary?
- **How to create a coherent and non-redundant summary?**



Graph-based social media summarization

- **Graph edges** will reflect the relation between documents
- **Graph structure and path** needs to mirror the required properties of the summary



Chris Froome at Tour de France 2017

Chris Froome Wins Yellow Jersey

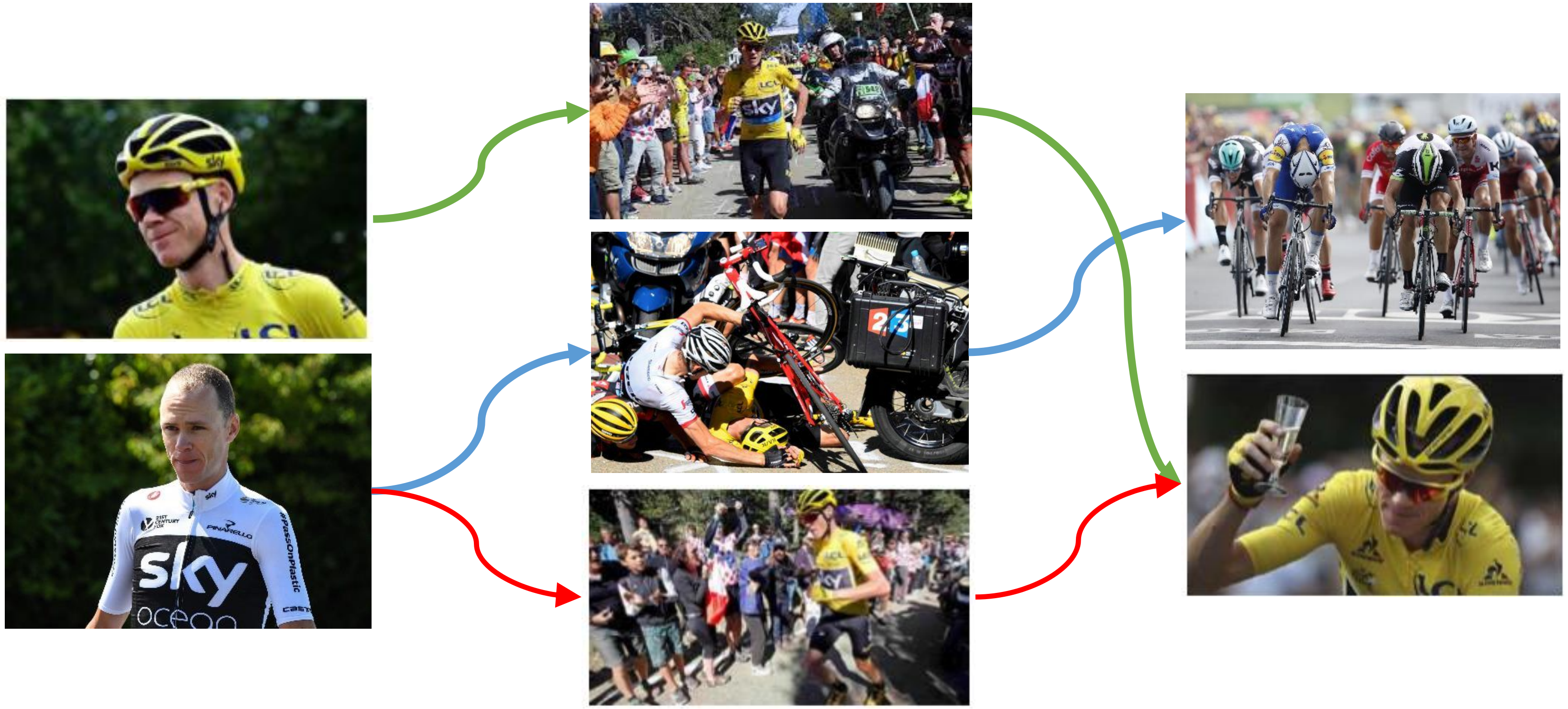
?

Forced to run to recover from crash

?

Three times winner

?



Graph edges as summary transition quality

- A *Gradient Boosted Tree* regressor was trained to predict a rating given the transition according to ground truth.

Gradient Boosted Trees

Output: Predicted transition quality

Transition similarity

- Transitions are characterized based on the relations between semantic and visual characteristics of adjacent images;

$$(\forall c \in C, distance_c(feature_c(a), feature_c(b)))$$

- Transition features are vectors of differences …

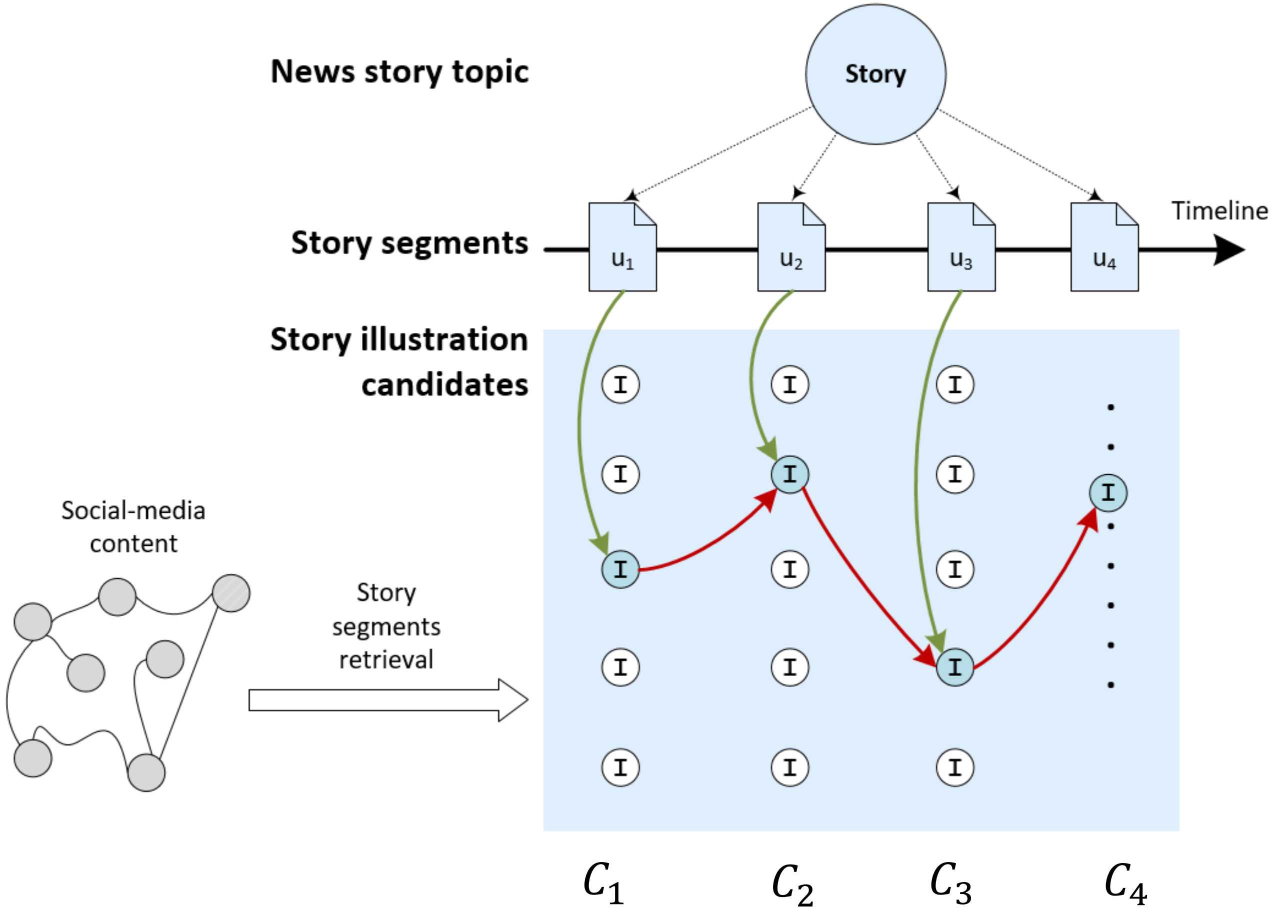
#Faces	33	1	=	32
Luminance	.5	.3		.2
…	…	…		…
Entropy	0.9	0.3		0.6

Transition features

Luminance	A positive real value representing the luminance.
Color histogram	A 3D color histogram with 16 bins per RGB channel converted to CIELAB color space.
Color moment	A vector representing the first color moment of the image in CIELAB color space.
Color correlogram	A 16 bins 3D color correlogram in CIELAB color space.
Entropy	A positve real value representing the entropy of the image.
#Edges	A vector containing the number of horizontal, vertical and diagonal edges.
pHash	A pHash vector.

Concepts	A set of image concepts extracted using VGG16.
CNN Dense	The embeddings extracted from the last layer of the ResNet CNN.
Environment	Either "outdoors"or "indoors".
Scene category	The location depicted in an image described through labels (e.g.: "bridge", "forest path", "skyscraper", etc.).
Scene attributes	The attributes of the location depicted in an image described through labels (e.g.: "man-made", "open area", "natural light", etc.).

Graph structure and summary paths



$$Story_M = (u_1, u_2, u_3, \ldots, u_N)$$

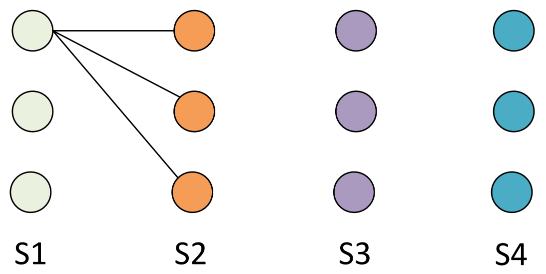
$$Storyline_M = (w_1, w_2, w_3, \ldots, w_N)$$

$$\forall i \in [1, N]\; w_i \subset C_i$$

Where C_i is the set of candidate images to illustrate segment u_1

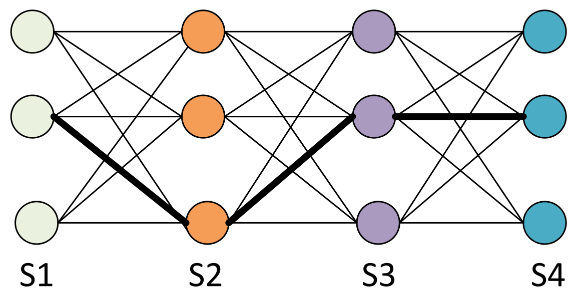
Bipartite graph - Shortest path

- A **sequence of bipartite graph** organizes story segments into groups of vertices in the graph:
  - All vertices in one group are connected to all the vertices in the neighbouring group



- The **shortest path**, selects the path with maximal similarity between vertices:

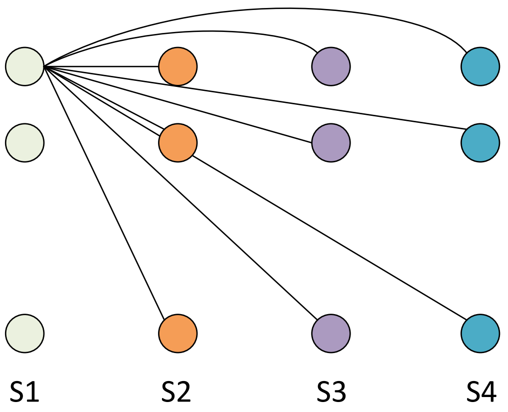
$$\min_{v_1 \in C_1, v_2 \in C_2, \ldots, v_N \in C_N} \sum_{i=1}^{N-1} pairCost(v_i, v_{i+1})$$



Gonçalo Marcelino, et al., to be submitted.

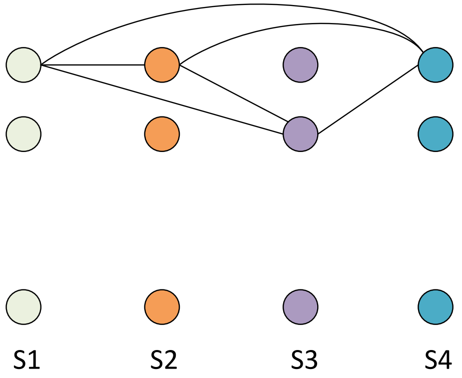
Multipartite graph - Maximal clique

- The **multipartite graph** organizes story segments into groups of vertices in the graph.
    - All vertices in one group are connected to all the other vertices but not connected to the vertices in their group.



- The **maximal clique** selects the clique with maximal intra clique similarity

$$\min_{v_1 \in C_1, v_2 \in C_2, \ldots, v_N \in C_N} \sum_{i=1}^{N-1} \sum_{k=i+1}^{N} pairCost(v_i, v_k)$$



Gonçalo Marcelino, et al., to be submitted.

Incorporating relevance

- In the first case (SeqT, FulT), the graph edges consider only the similarity between documents:

$$pairCost(v_x, v_y) = transC(v_x, v_y)$$

- In the second case (**SeqTR**, **FulTR**), the graph edges consider both the similarity and the relevance of the documents.

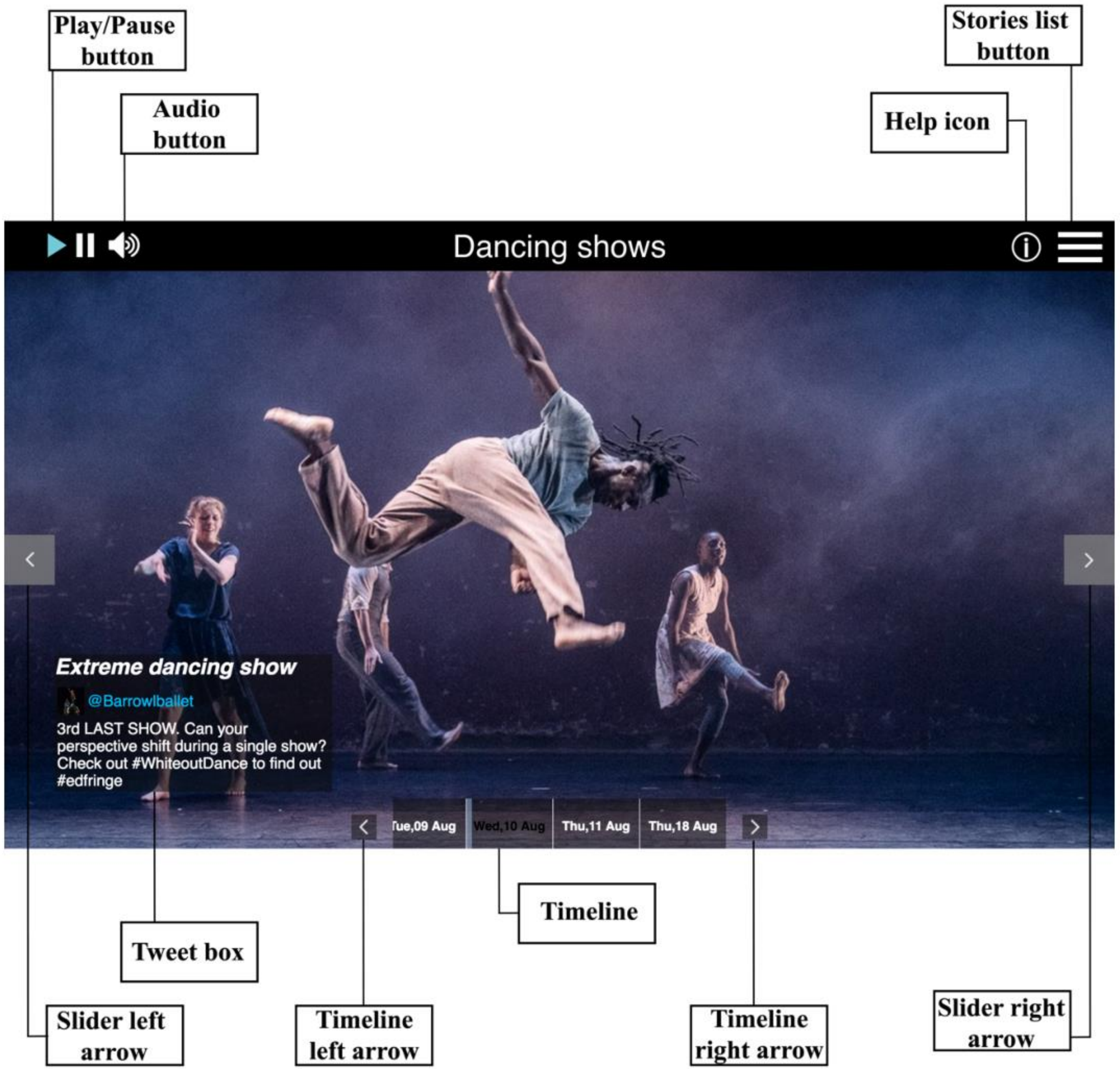
$$pairCost(v_x, v_y) = \underbrace{0.6 \cdot (relC(v_x) + relC(v_y))}_{\text{segments illustration}} + \underbrace{0.4 \cdot (relC(v_x) \cdot relC(v_y) + transC(v_x, v_y))}_{\text{transition}}$$

Evaluation framework

- To create and test this framework we resorted to 4 datasets of social media content related to 4 events.
- Training: 2016 Edinburgh Festival and Tour de France
- Test: 2017 Edinburgh Festival and Tour de France

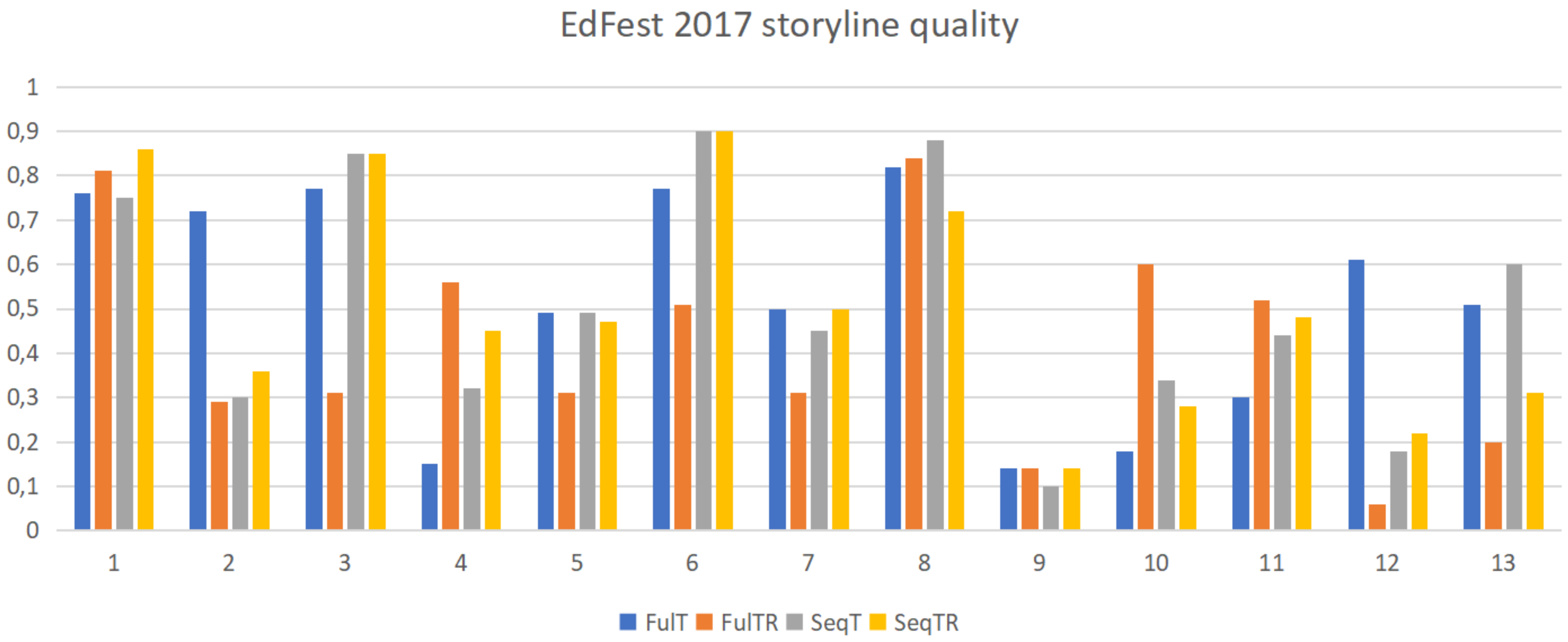
Event	Stories	Docs	Docs w/images
EdFest 2016	20	82348	15439
EdFest 2017	13	102227	34282
TDF 2016	20	325074	34865
TDF 2017	15	381529	67022

- Ground truth attained through crowd sourcing.



Play/Pause button
Audio button
Stories list button
Help icon
Dancing shows
Extreme dancing show
@Barrowlballet
3rd LAST SHOW. Can your perspective shift during a single show? Check out #WhiteoutDance to find out #edfringe
Tue,09 Aug
Wed,10 Aug
Thu,11 Aug
Thu,18 Aug
Timeline
Tweet box
Slider left arrow
Timeline left arrow
Timeline right arrow
Slider right arrow

Baseline	EdFest 2017 Relevance	EdFest 2017 Transition	EdFest 2017 Quality	TDF 2017 Relevance	TDF 2017 Transition	TDF 2017 Quality
Seq_T	0.49	0.72	0.51	0.56	0.81	0.56
Seq_{TR}	0.48	0.71	0.50	0.55	0.78	0.54
Ful_T	0.47	0.77	0.52	0.62	0.91	0.64
Ful_{TR}	0.42	0.61	0.42	0.59	0.72	0.57



What is EdFest 2017?

Music shows | *Theater and comedy* | *Circus* | *Street performances*




Ful_T




Ful_{TR}

What is EdFest 2017?

Music shows

Theater and comedy

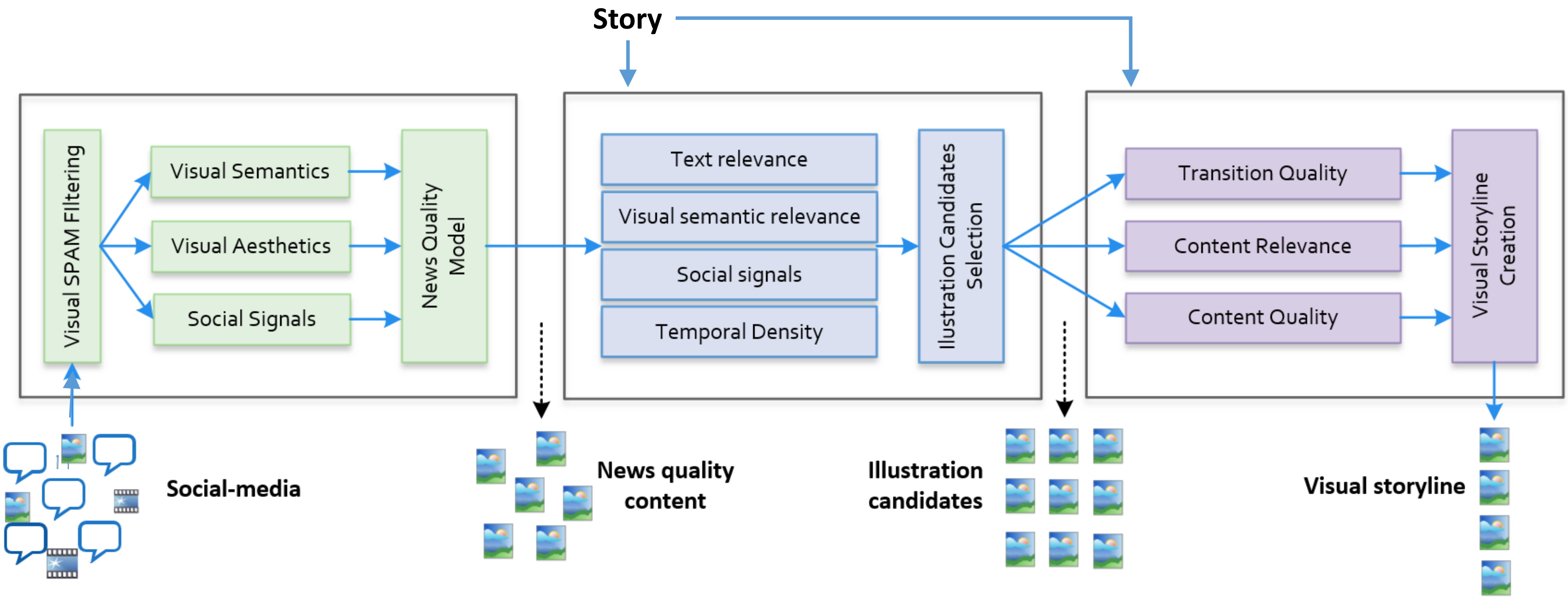
Circus

Street performances

Seq_{T}

Seq_{TR}

Data processing pipeline overview



Conclusions

- Social media data has great value but poses significant challenges in terms of noise and trustiness.
- Tools to help access social media information are critical to many domains:
  - News, Finance, Reputation monitoring
- Graph based approaches are easy to reason about and provide a meaningful way to further explore the data.

https://www.bbc.co.uk/rd/blog/2020-05-automated-news-stories-user-generated-journalism

Readings

- Paulus, R., Xiong, C., & Socher, R. (2018). A deep reinforced model for abstractive summarization. In ICLR.
- See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In ACL.
  - https://github.com/ymfa/seq2seq-summarizer