



# Web Search

MSc level Course  
João Magalhães

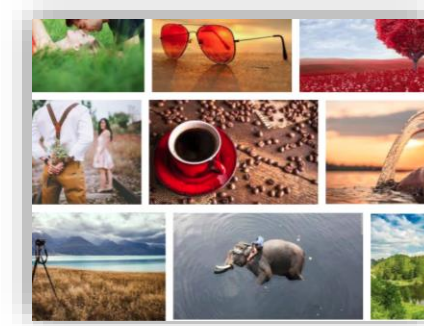
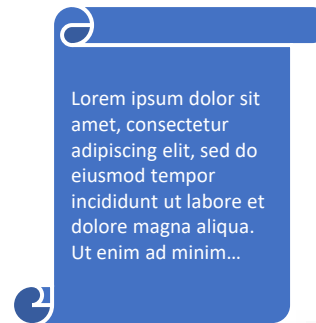
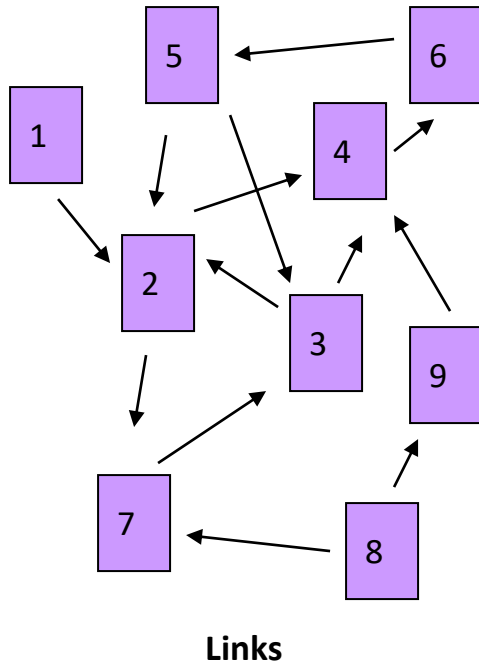
# How to search Web information?

- Textual and visual data can communicate a wide variety of information that are critical for several decision processes.
- Temporal and spatial structure adds organization and usability to information.
- Non-structured data (language and vision) puts a heavy complexity burden on standard data structures.

# Web data based search



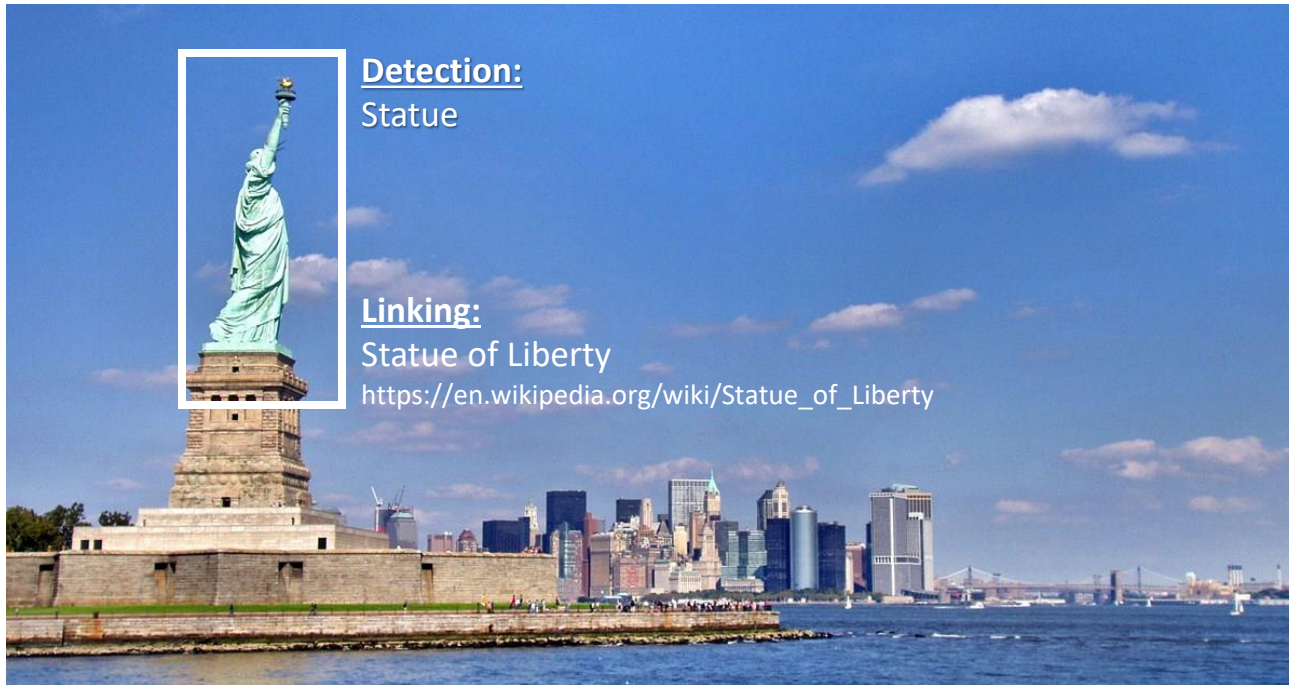
# Web graphs







# Classification, detection, linking



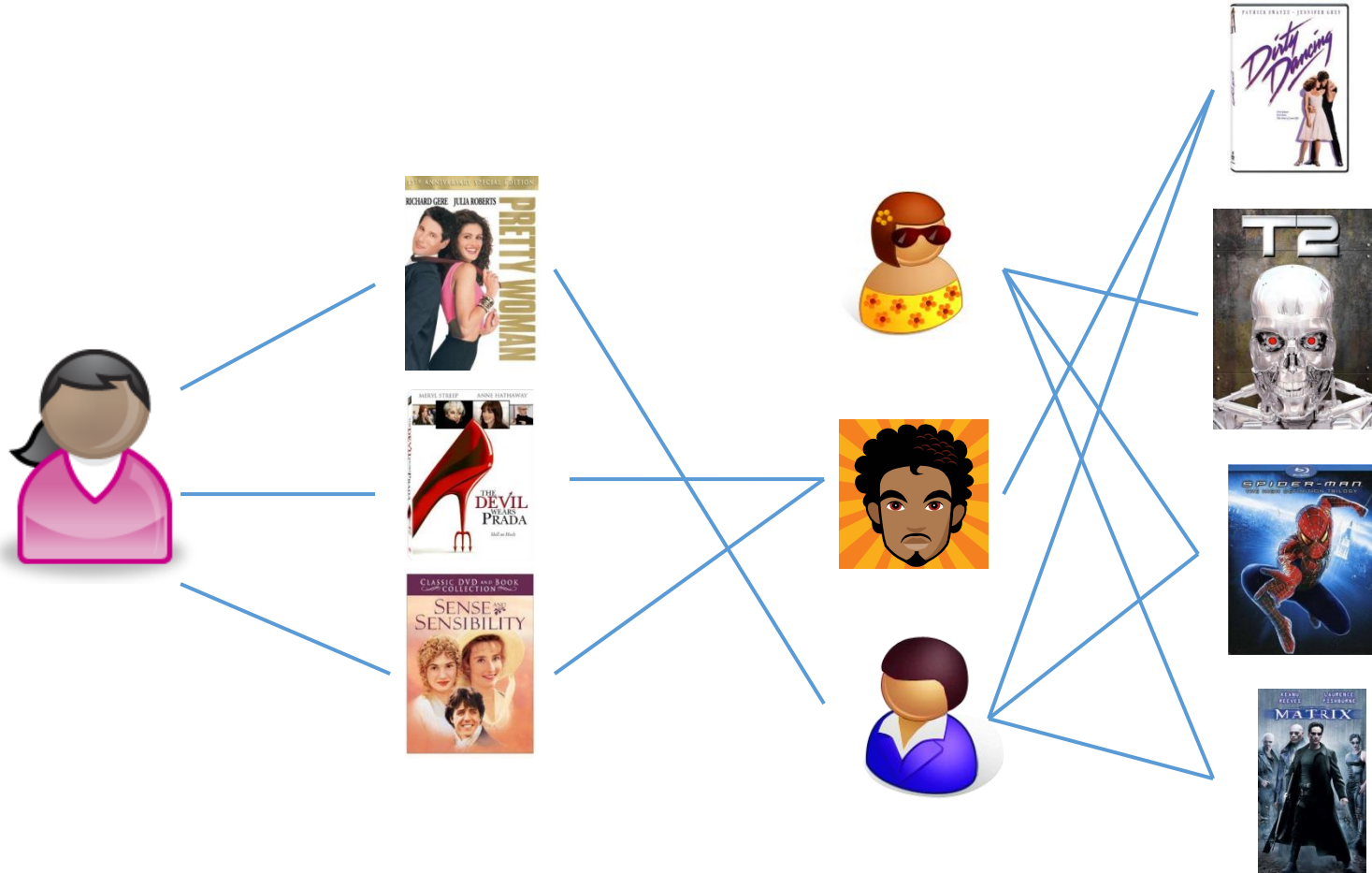
**Detection:**  
Statue

**Linking:**  
Statue of Liberty  
[https://en.wikipedia.org/wiki/Statue\\_of\\_Liberty](https://en.wikipedia.org/wiki/Statue_of_Liberty)

**Classification:**

Sea side  
Statue  
City  
Sky

# Collaborative filtering



## Funny reflex...

[Share This](#) [+](#) [ADD NOTE](#) [SEND TO GROUP](#) [+](#) [ADD TO SET](#) [BLOG THIS](#) [ALL SIZES](#) [ORDER PRINTS](#) [ROTATE](#) [EDIT PHOTO](#) [X](#) [DELETE](#) Uploaded on [August 3, 2006](#)  
by [jmc\\_mag](#)[+](#) [jmc\\_mag's photostream](#)

### Best ones (Set)

You are at  
the first  
photo.



11  
items

[browse](#) [→](#)

### Tags

- [Arizona](#) ✕
- [Reflection](#) ✕
- [Desert](#) ✕

[Add a tag](#)

### Additional Information

- [© All rights reserved](#) ([edit](#))
- [Anyone can see this photo](#) ([edit](#))

- [Add to your map](#)
- Taken with a [Canon EOS Digital Rebel XT](#).  
[More properties](#)
- Taken on [July 15, 2006](#) ([edit](#))
- [Photo stats](#)
- Viewed 26 times (Not including you)
- [Edit title, description, and tags](#)

### Comments

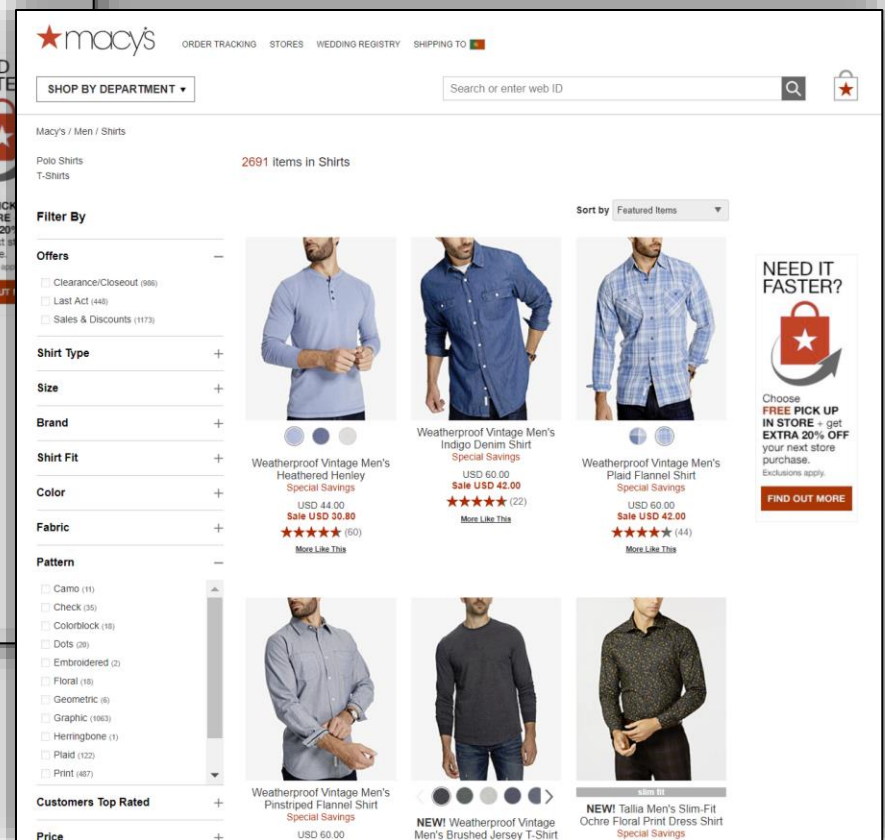
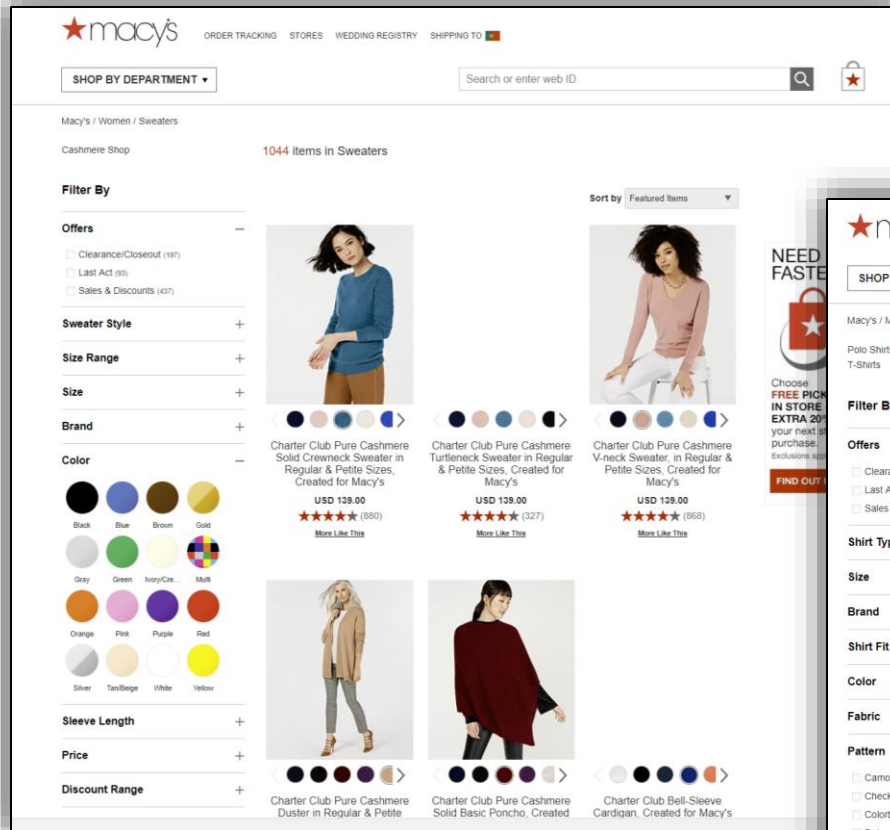
[alexei\\_322](#) [pro](#) says:

thumbs up! great shot

Posted 31 months ago. ([permalink](#) | [delete](#))

### Add your comment

# Online shopping





# Case-based search

(B) Free text query →

(A) Image drop area →

Current query →

(C) Recognized medical terms

(D) Knowledge-based assisted expansion

(E) Case-based search result

The interface displays a search bar with the following tags: **painless** × **hematuria** × **abdominal** × **tomography, spiral computed** × **renal mass** × **pelvis** × **ureter** ×. Below the search bar is a green button labeled **+ Add images (JPG only)...** and a blue button labeled **Search**. A dropdown menu titled **Also matches** is open, showing the following suggestions: **spiral volumetric ct**, **tomography, helical computed**, **spiral ct**, **helical ct**, **tomography, spiral volumetric computed**, **helical computed tomography**, and **spiral computed tomography**. The search results section is titled **Results for:** **painless hematuria abdominal tomography spiral computed renal mass left renal pelvis ureter**. It features a grid of four medical images. Below the images is the title **Massive hematuria due to a congenital renal arteriovenous malformation mimicking a renal pelvis tumor: a case report**, followed by the authors **Sountoulides, P; Zachos, I; Paschalidis, K; Asouhidou, I; Fotiadou, A; Bantis, A; Palasopoulou, M; Podimatas, T;**. The introduction text reads: **Introduction** Congenital renal arteriovenous malformations (AVMs) are very rare benign lesions. They are more common in women and rarely manifest in elderly people. In some cases they present with massive hematuria. Contemporary. To the right of the text are three small thumbnail images of medical scans.

# Course plan

- Online graph analysis: social-networks, Web graphs, etc.
- Recommender systems: movies, books, friends, etc.
- Information tagging and annotation
- Image and video tasks: automatic captioning and visual Question-Answering
- Document summarization and answer generation

# Course plan (in detail)

Web Data Mining and Search		
Week	#	Lecture
11/mar/21	1	Introduction
18/mar/21	2	Web data driven tasks
25/mar/21	3	Recommendation
01/abr/21		<b>Easter</b>
08/abr/21	4	Web Graphs
15/abr/21	5	Web document categorization
22/abr/21	6	Visual data search
29/abr/21	7	Word and Entity embeddings
06/mai/21	8	Modeling sequence data
13/mai/21	9	Transformer
20/mai/21	10	Mixed vision-and-language models
27/mai/21	11	Document summarization
08/jun/21		<b>Invited lecture: Data Science for Social Good</b>
17/jun/21		<b>Project presentation / Test</b>

# Course grading

- 40% test / exam
- 50% project
- 10% project presentation
- Groups of 3 students
- Additional rules:
  - Minimum grade on the labs or theory: 9
  - You may use one sided A4 sheet handwritten by you with your notes
    - It must be handed at the end of the test.

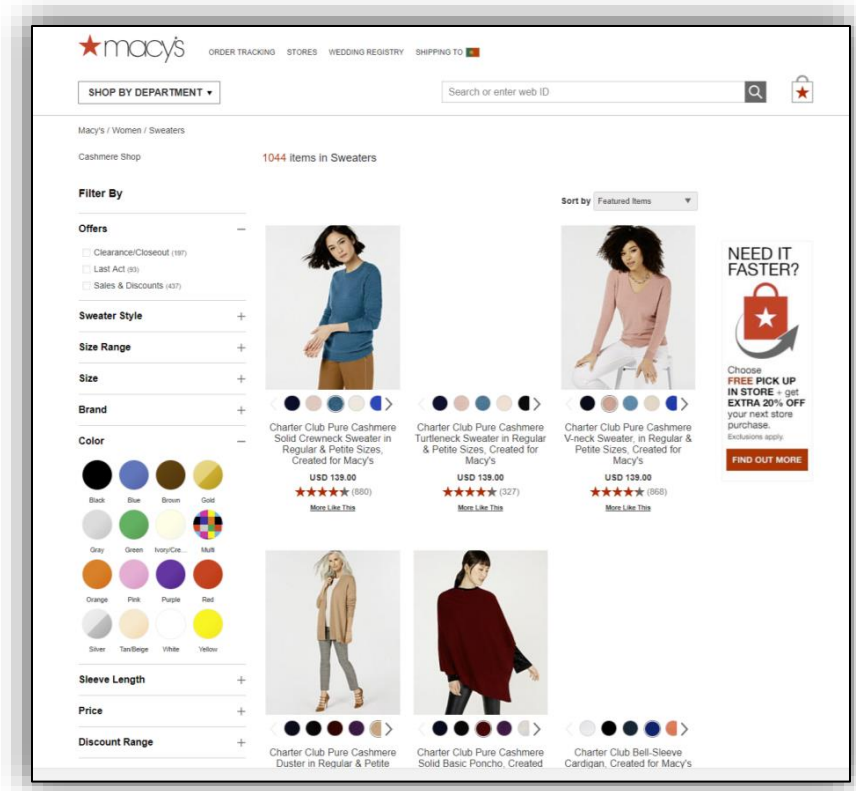


# PW is a project-based learning course!!!

- Fashion product search
- Recommendation
- Web images captioning
- Visual QA
- Timeline summarization
- Open-domain search assistant

# Fashion product search

- Fashion dataset
- ElasticSearch
- Visual Embeddings
- Faceted search and browsing



# Recommendation

- Click-through dataset
- Model past likes/don't likes
- Predict product to be purchased

**SHOPPER:** show me some espadrilles in beige/blue within \$200.

**AGENT:** Sure. let me just quickly browse through my catalogue



**AGENT:**

**AGENT:** What you can see in the fifth item, michael michael kors brings you an effortless weekend look with the wrapped espadrille heel and crisscross styling of these dana flats. And what you see in the 3rd item, steve madden, the company, was founded by steve madden, the man, in 1990 in brooklyn, where he started out making trend-advancing shoes one pair at a time.

**SHOPPER:** List more in the upper material as in the 5th image and style as the 3rd and the 5th



**AGENT:**

**AGENT:** The 2nd one, viscata espadrilles are 100% authentic and handcrafted in spain inspired by the mediterranean, viscata shoes are lightweight and made of natural jute rope, premium leathers, suedes and woven cloth.

**SHOPPER:** Can you list images of the 1st result from all orientations?



**AGENT:** Image from the front, right and back orientations respectively

**SHOPPER:** And the 3rd one?



**AGENT:** Image from the front, right, back and left orientations respectively

**SHOPPER:** I like the 4th image. Show me something like it but in style as in this image,



within my budget.



**AGENT:**

**AGENT:** Regarding the first item, platform espadrille slip on sam edelman shoes epitomize chic comfort.

# Web images captioning

- MSCOCO dataset
- PyTorch to generate captions
- External knowledge is optional

<p>A young boy is playing basketball.</p> 	<p>Two dogs play in the grass.</p> 	<p>A dog swims in the water.</p> 	<p>A little girl in a pink shirt is swinging.</p> 
<p>A group of people walking down a street.</p> 	<p>A group of women dressed in formal attire.</p> 	<p>Two children play in the water.</p> 	<p>A dog jumps over a hurdle.</p> 



# Visual QA

- MS COCO dataset
- PyTorch to generate tags
- External knowledge is optional
- Classification type of answers



## Attributes:

umbrella  
beach  
sunny  
day  
people  
sand  
laying  
blue  
green  
mountain

## Internal Textual Representation:

A group of people enjoying a sunny day at the beach with umbrellas in the sand.

## External Knowledge:

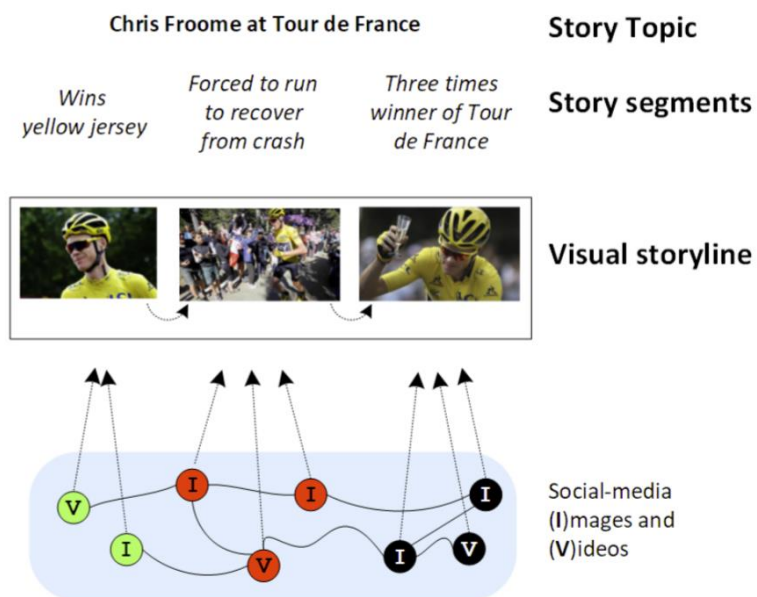
An umbrella is a canopy designed to protect against rain or sunlight. Larger umbrellas are often used as points of shade on a sunny beach. A beach is a landform along the coast of an ocean. It usually consists of loose particles, such as sand....

## Question Answering:

**Q:** Why do they have umbrellas? **A :** Shade.

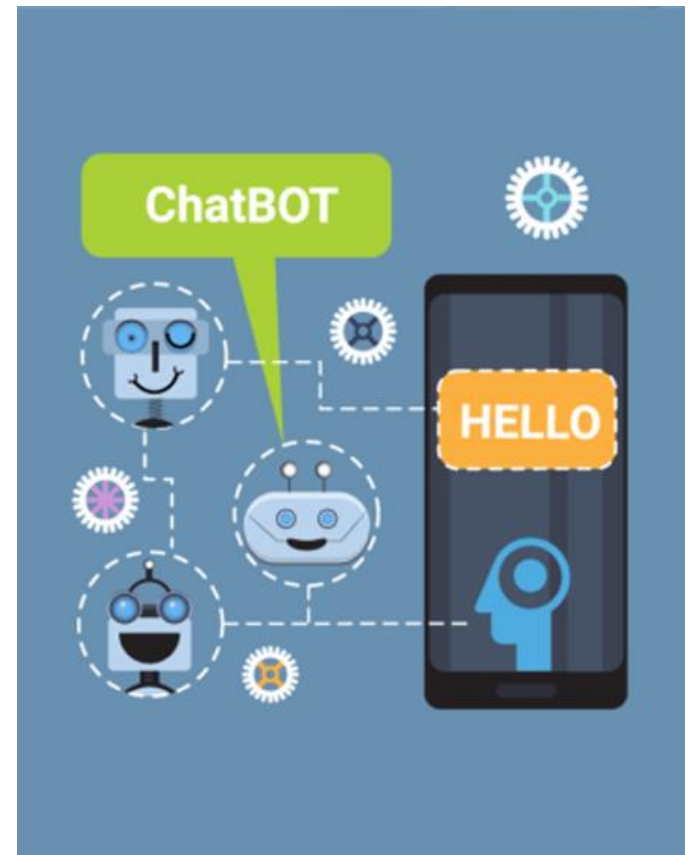
# Visual Stories - Summarization

- Create visual summaries of events
- Dataset will be provided
- Summarization of timelines



# Open-domain search assistant

- Wikipedia powered chatbot
- Open-ended conversations
- One single answer instead of document list



<https://knowledge-answer-generation.herokuapp.com/#>  
<https://www.arxiv-vanity.com/papers/2101.08197/>

# Project grading

- Report:

- MAXIMUM 8 PAGES!!!
- No cover page.
- Must include graphs, tables, etc.

- Scoring:

- Algorithms imp./int. 30%
- Critical discussion 40%
- Results analysis 30%
- OR
- UI implementation 30%

- Report organization:

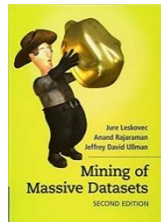
- Introduction
- Algorithms
- Implementation
- Evaluation
  - Dataset description
  - Baselines
  - Results analysis
- Critical discussion
- References



# References

- Slides and articles provided during classes.

- Books:



Jure Leskovec, Anand Rajaraman, Jeff Ullman, “Mining of Massive Datasets”, Cambridge University Press, 2011.

<http://www.mmds.org/>



Aston Zhang, Zachary Lipton, Mu Li, and Alex Smola, “Dive into Deep Learning”

<http://d2l.ai/>