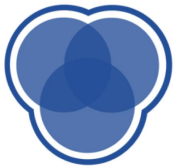
INFO 1998: Introduction to Machine Learning



CDS Education

We explore, learn, and educate big minds.

Lecture 10: Real-World Applications of Data Science

INFO 1998: Introduction to Machine Learning

CDS Education

We explore, learn, and educate big minds.

Agenda

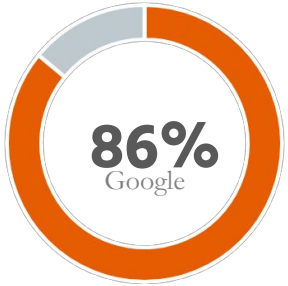
- Advertising
- Healthcare
- Media
- Social Impact
- Ethics
- Cornell Data Science projects :)

Advertising

Machine Learners: The Modern Mad Men

Context

Big Tech companies earn their the bulk of their revenue through ads
One usually earns money when the ad is 'clicked' by the user
Users are most likely to click on ads when the ads are relevant to them
Ads could be tailored to users only when there is data on the users



c_id	ip	loc	city	state	link	time	timestamp
3d5wf31	128.83.126	(68.3, 98.5)	Hoboken	NJ	../falltrends19	143s	07:56:31
6d1wd34	128.45.313	(62.3, 89.5)	SYR	NY	…/shoestobuy	9s	07:56:35
3d5wf31	341.34.345	(68.5, 98.6)	NYC	NY	../excelhelp	552s	14:42:23

Sample Data (Extremely small slice): What can you interpret?

Advertising

c_id	ip	loc	city	state	link	time	timestamp
3d5wf31	128.83.126	(68.3, 98.5)	Hoboken	NJ	../falltrends19	143s	07:56:31
6d1wd34	128.45.313	(62.3, 89.5)	SYR	NY	…/shoestobuy	9s	07:56:35
3d5wf31	341.34.345	(68.5, 98.6)	NYC	NY	../excelhelp	552s	14:42:23

c_id	ip	loc	city	state	link	time	timestamp
3d5wf31	128.83.126	(68.3, 98.5)	Hoboken	NJ	../falltrends19	143s	07:56:31
	341.34.345	(68.5, 98.6)	NYC	NY	../excelhelp	552s	14:42:23
6d1wd34	128.45.313	(62.3, 89.5)	SYR	NY	…/shoestobuy	9s	07:56:35

Objective: Get data on the users

Advertising

c_id	ip	loc	city	state	link	time	timestamp
3d5wf31	128.83.126	(68.3, 98.5)	Hoboken	NJ	../cutefallskirts	143s	07:56:31
	341.34.345	(68.5, 98.6)	NYC	NY	../excelhelp	552s	14:42:23

Hypotheses:

- Lives in NJ and works in NYC
- Lives in area with average rent: $r
- Lives in area with average income: $i
- Works in area with average salary: $s
- Falls in k income bracket (Estimated)
- Takes NJTransit to work
- Takes the 67 Train at 8:05am
- Works at XYZ Company
- Works in Business/Data Analytics
- Is a Female
- Is interested in topics A, B, C

With **enough data** and **testing**, the hypotheses could be affirmed or rejected.

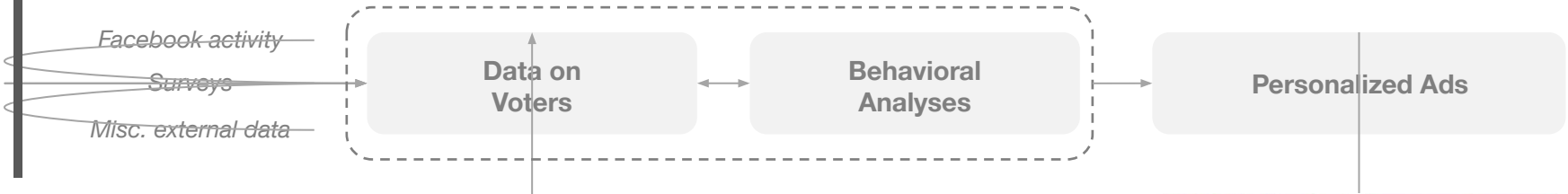
Cambridge Analytica: Data Science in Political Campaigning

Case Study

Overview

Cambridge Analytica combined *data analytics, behavioral sciences,* and *innovative ad tech* to influence voters
Widely regarded as instrumental in the result of the 2016 Elections, and many more across the globe

Methodology

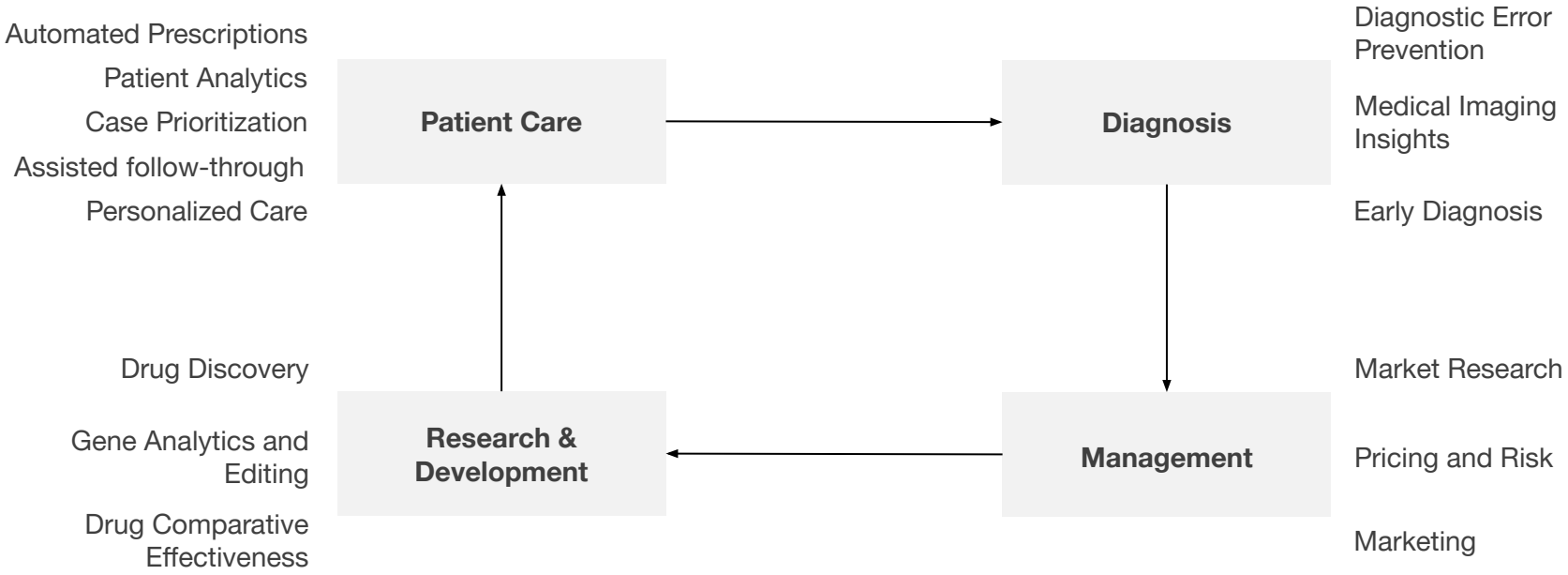


Example

Source: towardsdatascience.com/effect-of-cambridge-analyticas-facebook-ads-on-the-2016-us-presidential-election-dacb5462155d

Healthcare

All-round betterment in the healthcare industry



Source: https://blog.appliedai.com/healthcare-ai/

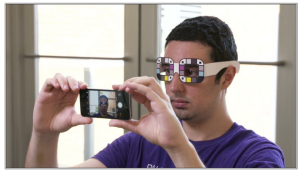
BiliScreen: A Selfie to Diagnose Pancreatic Cancer

Case Study

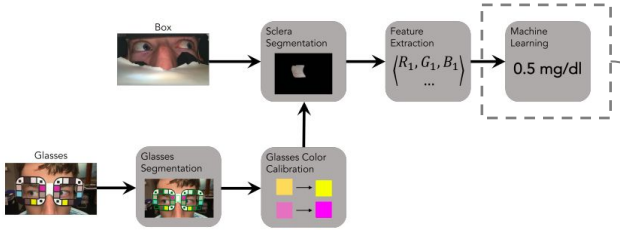
Overview

A smartphone app that captures pictures of the eye and produces an estimate of a person's bilirubin level

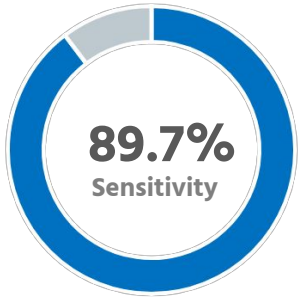
Uses: (1) A 3D-printed box that controls the eyes' exposure to light
(2) Paper glasses with colored squares for calibration



Methodology



Machine Learning Algorithms Used?



Source: ubicomplab.cs.washington.edu/pdfs/biliscreen.pdf

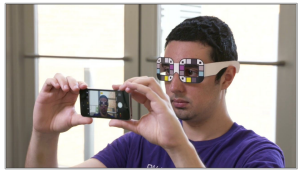
BiliScreen: A Selfie to Diagnose Pancreatic Cancer

Case Study

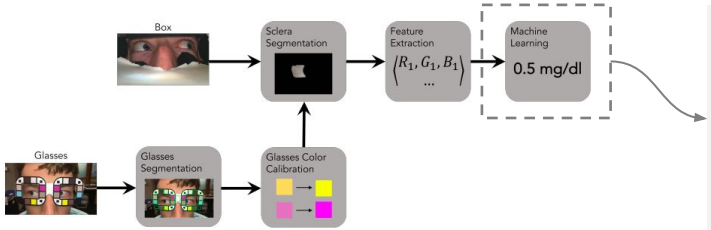
Overview

A smartphone app that captures pictures of the eye and produces an estimate of a person's bilirubin level

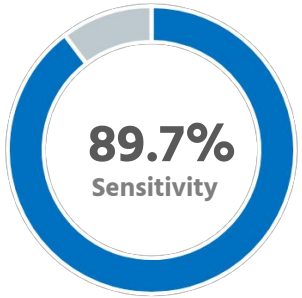
Uses: (1) A 3D-printed box that controls the eyes' exposure to light
(2) Paper glasses with colored squares for calibration



Methodology



Random Forest
with 10-fold Cross Validation



Source: ubicomplab.cs.washington.edu/pdfs/biliscreen.pdf, medium.com/sciforce/top-ai-algorithms-for-healthcare-aa5007ffa330

Media: Recommender Systems

How Netflix keeps you hooked

Overview

Most of Netflix's views (~80%) come through recommendations
The famous Netflix Challenge offered $1m to the participant that could do better than Netflix's recommender system
These algorithms are relatively simple and intuitive, but extremely effective

c_id	movie	tags	time	duration	rating
A	Avengers	Action, Superhero	07:56:31	112m	5/5
	Mr. Bean	Comedy	07:36:35	3s	2/5
B	Batman	Superhero	14:42:23	59m	4/5
	Black Mirror	Sci-Fi	07:56:34	142m	5/5

Sample: What would you recommend A next?

Usually, many other features and tags for the movies/shows exist in the database as well

Media: Recommender Systems

How Netflix keeps you hooked

c_id	movie	tags	time	duration	rating
A	Avengers	Action, Superhero	07:56:31	112m	5/5
	Mr. Bean	Comedy	07:36:35	3s	2/5
B	Batman	Superhero	14:42:23	59m	4/5
	Black Mirror	Sci-Fi	07:56:34	142m	5/5

Sample: What would you recommend A next?

Read More: towardsdatascience.com/introduction-to-recommender-systems-6c66cf15ada

Where else are recommender systems applicable?

Social Impact

Data Science for Social Good

Overview

Advanced analytics for social impact is becoming increasingly popular due to innumerable low-cost and high-impact applications

Education

Adaptive-learning technology that could **recommend** material based on student's success and engagement

Public Sector

Identifying tax-fraud using alternate data such as browsing history, retail data, or payments history.

Crisis

Predicting the progression of wildfires to optimize the response of firefighters.

Read More: https://www.mckinsey.com/featured-insights/artificial-intelligence/applying-artificial-intelligence-for-social-good

Predicting End Location: Tackling Human Trafficking

Case Study

Overview

Human trafficking is a great cause of concern, especially in developing countries
ML could be leveraged to aid ground rescue operations for trafficking victims

Predicting End Location: Tackling Human Trafficking

Case Study

Overview

Human trafficking is a great cause of concern, especially in developing countries
ML could be leveraged to aid ground rescue operations for trafficking victims

An Important Note on Ethics

It’s easy to get caught up in the technical challenge, but it is important to know that your work may affect other people directly or indirectly, now or in the future. Ask yourself the following questions often:

- Does your data or analysis impede on anyone’s privacy?
- Did the people give consent for their data to be used?
- Were the people given the option to opt out?
- Who has the right of access to your data?
- Who owns the data?
- Was the data anonymized sufficiently?
- Was there any bias in your dataset against certain sections of the society?
- Are you introducing any bias?
- Should you include any features that may be discriminatory?
- Is your analysis transparent?
- Are the end users aware of shortcomings?

Extreme Example: Black Mirror

- In “Be Right Back” (S2, E1), a widow discovers a chat-bot that can mimic the responses of her recently deceased husband
- Went as far as creating a robot that looked like her deceased husband, and responded to actions using this chatbot…
- How difficult would this be to create?

Microsoft Did it 😳

Microsoft patented a chatbot that would let you talk to dead people. It was too disturbing for production

By Clare Duffy, CNN Business

Updated 7:04 AM EST, Wed January 27, 2021

Source: https://www.cnn.com/2021/01/27/tech/microsoft-chat-bot-patent/index.html

Takeaway

- Data science has amazing potential for improvements in fields like advertising, healthcare, and media
- Can also have great social impacts
- However, with great power comes great responsibility
    - The ethics of applications of data science must be considered!

CDS Project Highlight: MathSearch!

- **Goal**: Create a web application that can take as input, a PDF and a LaTeX equation query. Find matches of that equation inside the PDF and then return pages with bounding boxes around most similar equations
- Combines Data Science, and Machine Learning Engineering
  - A lot of the work has to do with learning how to use cloud computing and Amazon Web Services to create an ML pipeline which is able to handle concurrent requests and run large models quickly!!
- Come see us present at CDS' Final Showcase
  - May 5th, check our instagram @cornelldatascience for more details!

That’s all folks!

- **Final Project Due**: May 1st. Come to OH if you have project-related questions!
- End-of-semester feedback form (extra credit!):

Thank you all for taking this class, and for an incredible semester.

CDS Education

We explore, learn, and educate big minds.