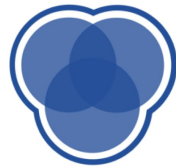# INFO 1998: Introduction to Machine Learning

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

Advertising

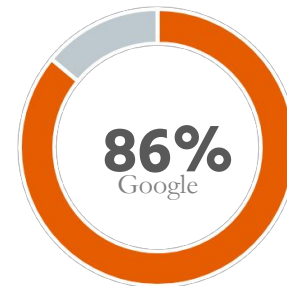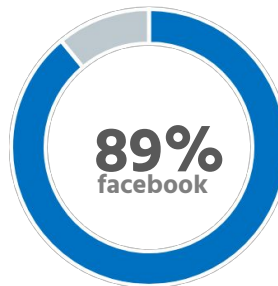*Machine Learners: The Modern Mad Men*

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

89% facebook

86% Google

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.

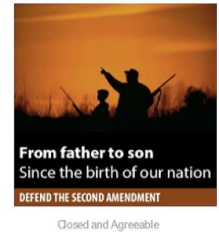# Cambridge Analytica: Data Science in Political Campaigning
Case Study

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
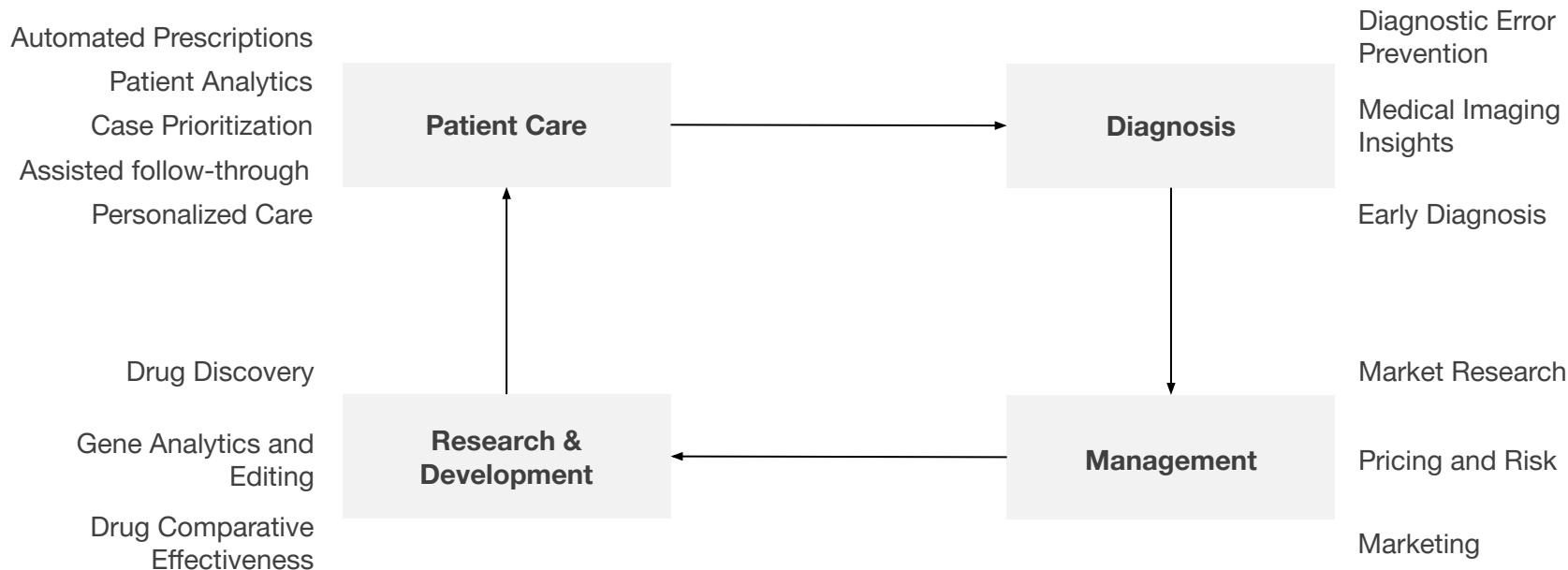
Methodology

Facebook activity
Surveys
Misc. external data

| Data on Voters | ⟷ | Behavioral Analyses | → | Personalized Ads |

Example

Likes, Comments, Surveys, etc. →

O OPENNESS
Do they enjoy
new experiences?

C CONSCIENTIOUSNESS
Do they prefer
plans and order?

E EXTRAVERSION
Do they like spending
time with others?

A AGREEABLENESS
Do they put people's
needs before theirs?

N NEUROTICISM
Do they tend
to worry a lot?

The Second Amendment isn't just
a right. **It's an insurance policy.**
DEFEND THE RIGHT TO BEAR ARMS

High Neuroticism and Conscientious

From father to son
Since the birth of our nation
DEFEND THE SECOND AMENDMENT

Closed and Agreeable

Source: Cambridge Analytica

Advertising

Healthcare

All-round betterment in the healthcare industry

Automated Prescriptions

Patient Analytics

Case Prioritization

Assisted follow-through

Personalized Care

Patient Care → **Diagnosis**

Diagnostic Error Prevention

Medical Imaging Insights

Early Diagnosis

Drug Discovery

Gene Analytics and Editing

Research & Development

Drug Comparative Effectiveness

Management

Market Research

Pricing and Risk

Marketing

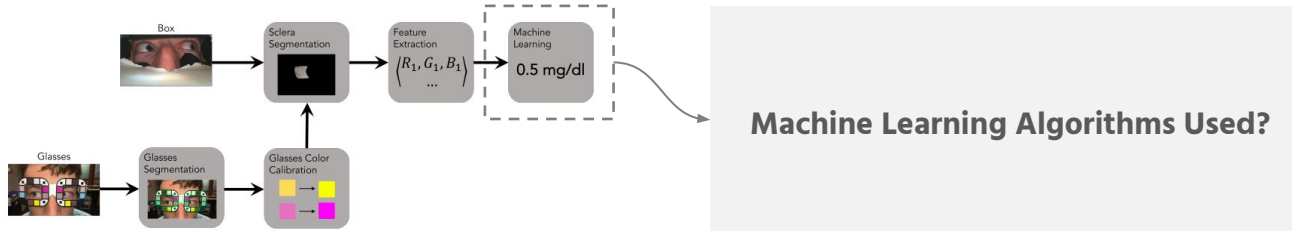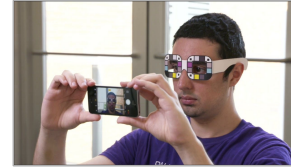# BiliScreen: A Selfie to Diagnose Pancreatic Cancer
Case Study

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

96.8%
Specificity

Methodology

Machine Learning Algorithms Used?

89.7%
Sensitivity

Healthcare

BiliScreen: A Selfie to Diagnose Pancreatic Cancer
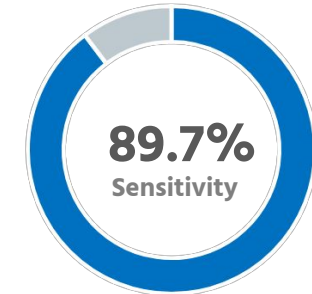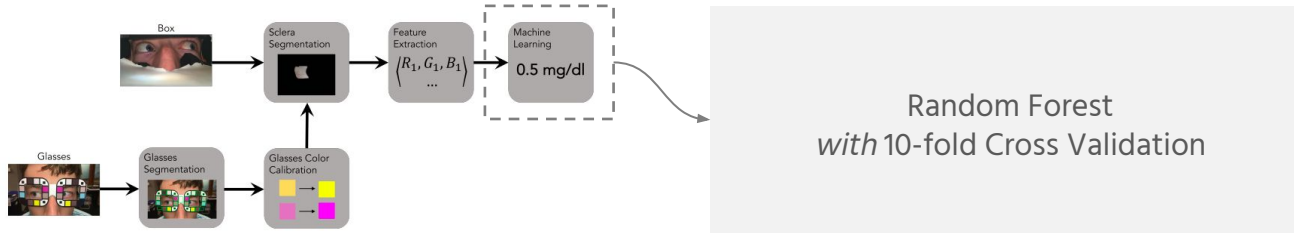*Case Study*

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

96.8%
Specificity

Methodology

Random Forest
with 10-fold Cross Validation

89.7%
Sensitivity

Healthcare

Media: Recommender Systems
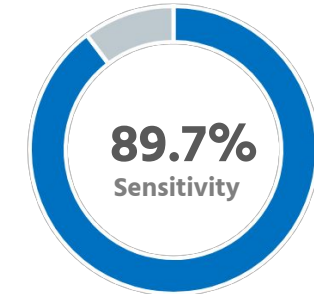
How Netflix keeps you hooked

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
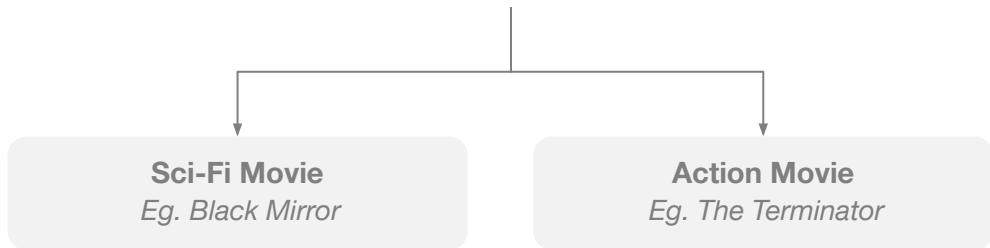
Media: Recommender Systems

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?

Sci-Fi Movie	**Action Movie**
Eg. Black Mirror	*Eg. The Terminator*

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.

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

Rescued Victims Data

Native Location, End Location, End Industry, Age, Sex, etc.

?

Probable End Locations

Probable End Industries

Predicting End Location: Tackling Human Trafficking

Case Study

Overview

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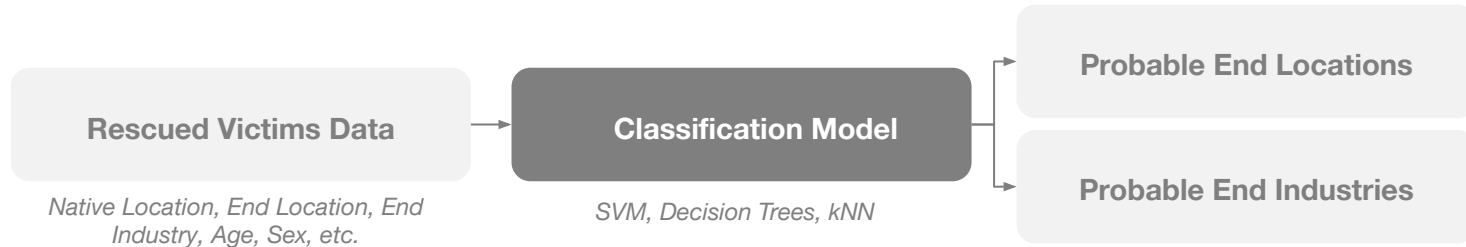ML could be leveraged to aid ground rescue operations for trafficking victims

Rescued Victims Data	→	Classification Model	→	Probable End Locations
Native Location, End Location, End Industry, Age, Sex, etc.		*SVM, Decision Trees, kNN*		Probable End Industries

Social Impact

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?

That's all folks!

- **Final Project Due**: Tonight

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

CDS Education

We explore, learn, and educate big minds.