The TwitMiner 2013 Challenge

A Machine Learning Programming Contest

Harikrishna Narasimhan and Saneem Ahmed

TwitMiner Team:

Saneem Ahmed
Anirban Laha
Adway Mitra
Harikrishna Narasimhan
Kevin Patel

Sponsored by:

amazon.com

The TwitMiner 2013 Challenge

- Part of Open Days 2013 @ CSA, IISc
- A machine learning programming contest.

The TwitMiner 2013 Challenge

- Part of Open Days 2013 @ CSA, IISc
- A machine learning programming contest.
- Spread awareness about ML among undergrads

The TwitMiner 2013 Challenge

- Part of Open Days 2013 @ CSA, IISc
- A machine learning programming contest.
- Spread awareness about ML among undergrads

Challenge

Build a system that can categorize Twitter *data*.

The TwitMiner 2013 Challenge

- Part of Open Days 2013 @ CSA, IISc
- A machine learning programming contest.
- Spread awareness about ML among undergrads

Challenge

Build a system that can categorize Twitter *data*.

In association with amazon.com

Winners to receive **internship** offers!

The TwitMiner 2013 Challenge

- Part of Open Days 2013 @ CSA, IISc
- A machine learning programming contest.
- Spread awareness about ML among undergrads

Challenge

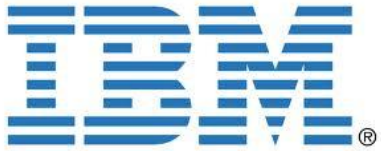
Build a system that can categorize Twitter *data*.

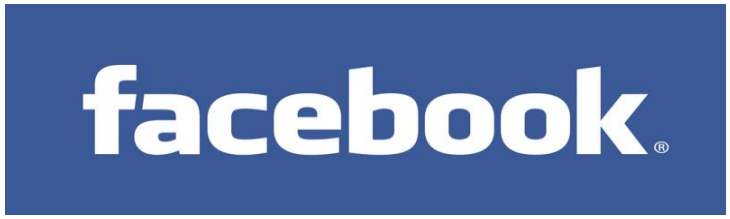
In association with amazon.com

Winners to receive **internship** offers!

But first.....

Why Machine Learning?





machine learning

Search

About 28,000,000 results (0.30 seconds)

Web

Images

Maps

Videos

News

Blogs

Books

More

Bangalore, Karnataka

Change location

The web

Pages from India

More search tools

Scholarly articles for machine learning

... in search, optimization, and **machine learning** - Goldberg - Cited by 50557
Q-**learning** - Watkins - Cited by 6525
An introduction to MCMC for **machine learning** - Andrieu - Cited by 691

Machine learning - Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/**Machine_learning**

Machine learning, a branch of artificial intelligence, is a scientific discipline concerned with the design and development of algorithms that allow computers to ...

↳ List of machine learning ... - Category:Machine learning - Weka - Transduction

Machine Learning | Coursera

www.coursera.org/course/ml

25 May 2012 – **Machine learning** is the science of getting computers to act without being explicitly programmed. In the past decade, **machine learning** has ...

Machine Learning textbook

www.cs.cmu.edu/~tom/mlbook.html

Machine Learning, Tom Mitchell, McGraw Hill, 1997. cover; **Machine Learning** is the study of computer algorithms that improve automatically through experience ...

Machine Learning

www.springer.com › ... › Computer Science › Artificial Intelligence

Machine Learning is an international forum for research on computational approaches to learning. The journal publishes articles reporting substantive results on ...

Introduction to Machine Learning

robotics.stanford.edu/~nilsson/mlbook.html

From this page you can download a draft of notes I used for a Stanford course on **Machine Learning**. Although I have tried to eliminate errors, some undoubtedly ...

amazon

Sachin's Amazon.com | Today's Deals | Gift Cards | Help

Shop by Department | Search | All

Your Amazon.com | Your Browsing History | Recommended For You | Amazon Betterizer | Improve Your Recommendations | Your Profile | Learn More

Your Amazon.com › Recommended for You

(If you're not Sachin Nagargoje, click here.)

Just For Today

Browse Recommended

Recommendations

Amazon Instant Video
Appliances
Appstore for Android
Arts, Crafts & Sewing
Automotive
Baby
Beauty
Books
Books on Kindle
Camera & Photo
Cell Phones & Accessories
Clothing & Accessories
Computers
Electronics
Grocery & Gourmet Food
Health & Personal Care
Home & Kitchen
Home Improvement
Industrial & Scientific
Jewelry
Kitchen & Dining
MP3 Downloads
Magazine Subscriptions
Movies & TV
Music
Musical Instruments
Office & School Supplies
Patio, Lawn & Garden

These recommendations are based on items you own and more.

view: **All** | New Releases | Coming Soon

1.

Nikon ML-L3 Wireless Remote Control
by Nikon (June 17, 2003)
Average Customer Review: ★★★★½ (1,215)
In Stock

List Price: ~~$20.50~~
Price: $14.95
30 used & new from $10.00

Add to Cart | Add to Wish List

I own it | Not interested | Rate this item

Recommended because you added **Nikon BM-10 LCD Cover for D90** to your Shopping Cart and more (Fix this)

2.

Peltor 97010 Ultimate-10 Hearing Protector
by Peltor (January 12, 2005)
Average Customer Review: ★★★★½ (165)
In Stock

List Price: ~~$25.99~~
Price: $20.63
45 used & new from $18.77

Add to Cart | Add to Wish List

I own it | Not interested | Rate this item

Recommended because you purchased **Peltor 97079 Combat Arms Earplugs** and more (Fix this)

3.

Nikon DK-21 Rubber Eyecup Nikon Digital SLR Cameras
by Nikon (June 17, 2003)
Average Customer Review: ★★★★½ (28)
In Stock

Price: $8.94
7 used & new from $4.25

Add to Cart | Add to Wish List

I own it | Not interested | Rate this item

Recommended because you added **Nikon BM-10 LCD Cover for D90** to your Shopping Cart and more (Fix this)

Control Panel ▸ Speech Recognition Options
Search
Advanced speech options
Text to Speech
Configure your Speech Recognition experience
Start Speech Recognition
Start using your voice to control your computer.
Set up microphone
Set up your computer to work properly with Speech Recognition.
Take Speech Tutorial
Learn to use your computer with speech. Learn basic commands and dictation.
Train your computer to better understand you
Read text to your computer to improve your computer's ability to understand your voice. Doing this isn't necessary, but can help improve dictation accuracy.
Open the Speech Reference Card
View and print a list of common commands to keep with you so you always know what to say.
If you would like to learn more about speech related tools and support from Microsoft, please visit the Microsoft Speech Online Community.
Listening
See also
Ease of Access
Audio Devices

Brief Outline

Part 1: Overview of Machine Learning

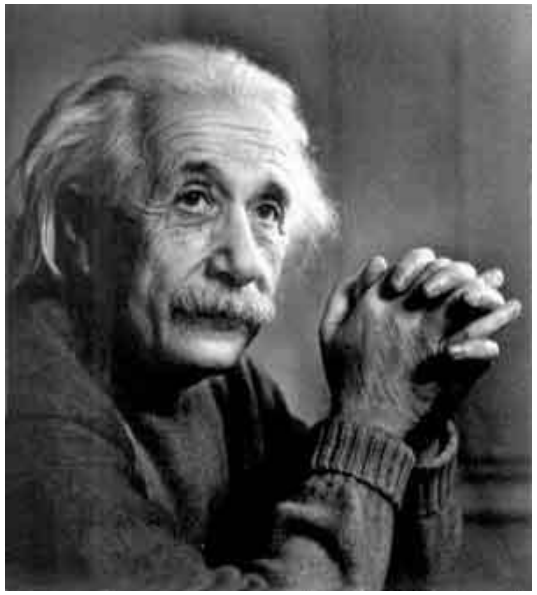
Brief Outline

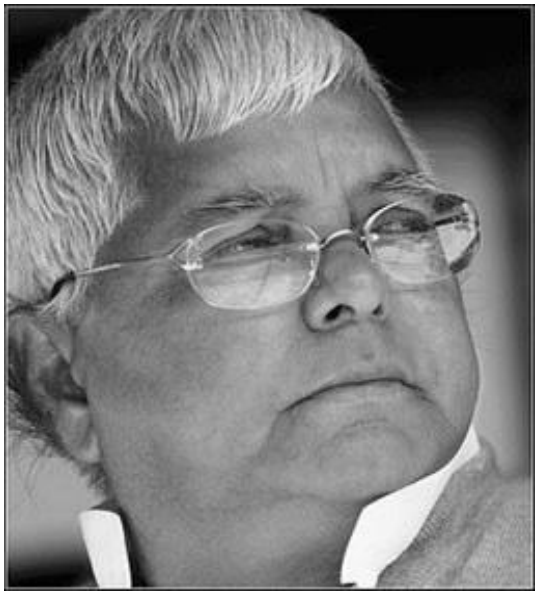
Part 1: Overview of Machine Learning

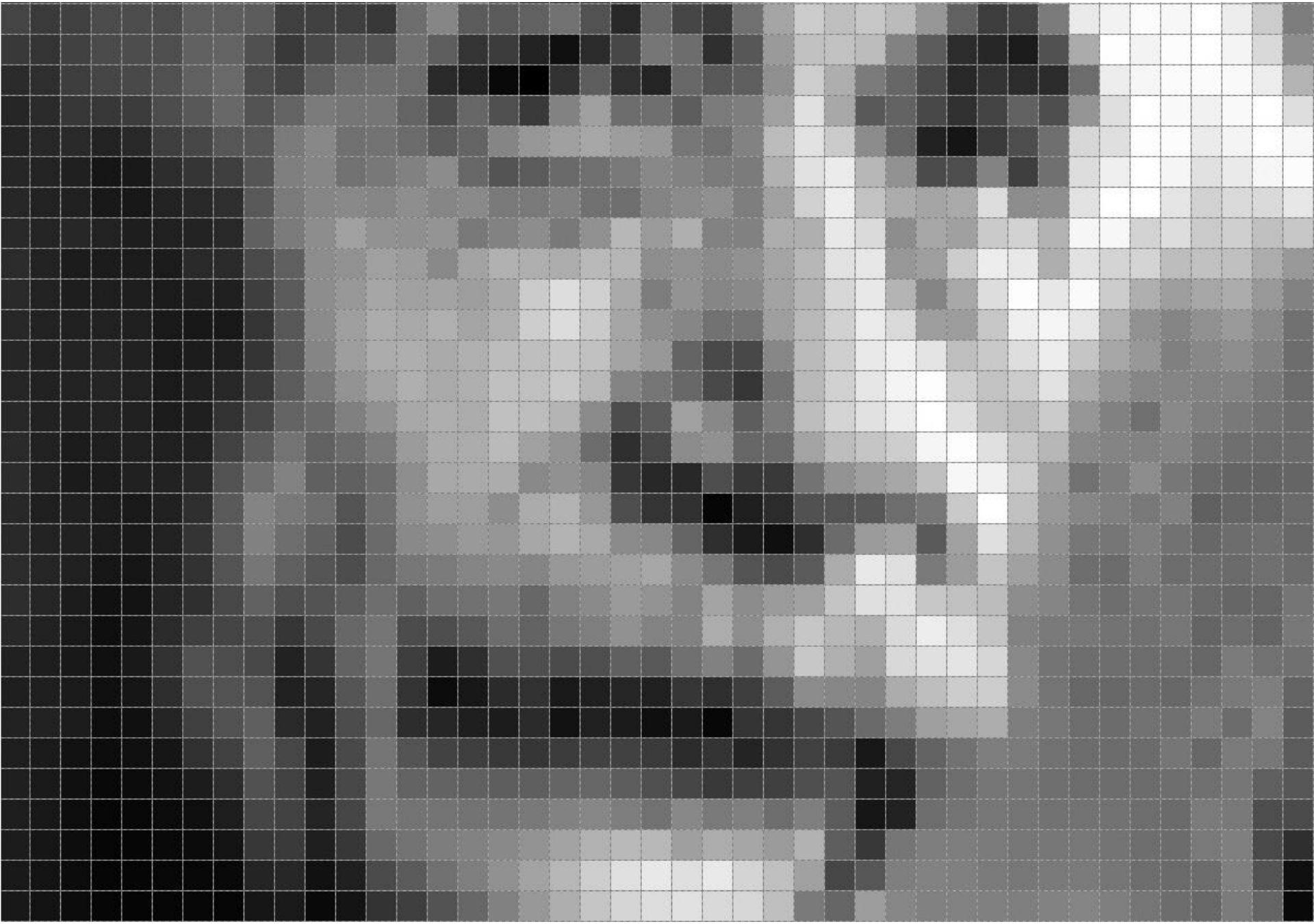
Part 2: The TwitMiner 2013 Challenge

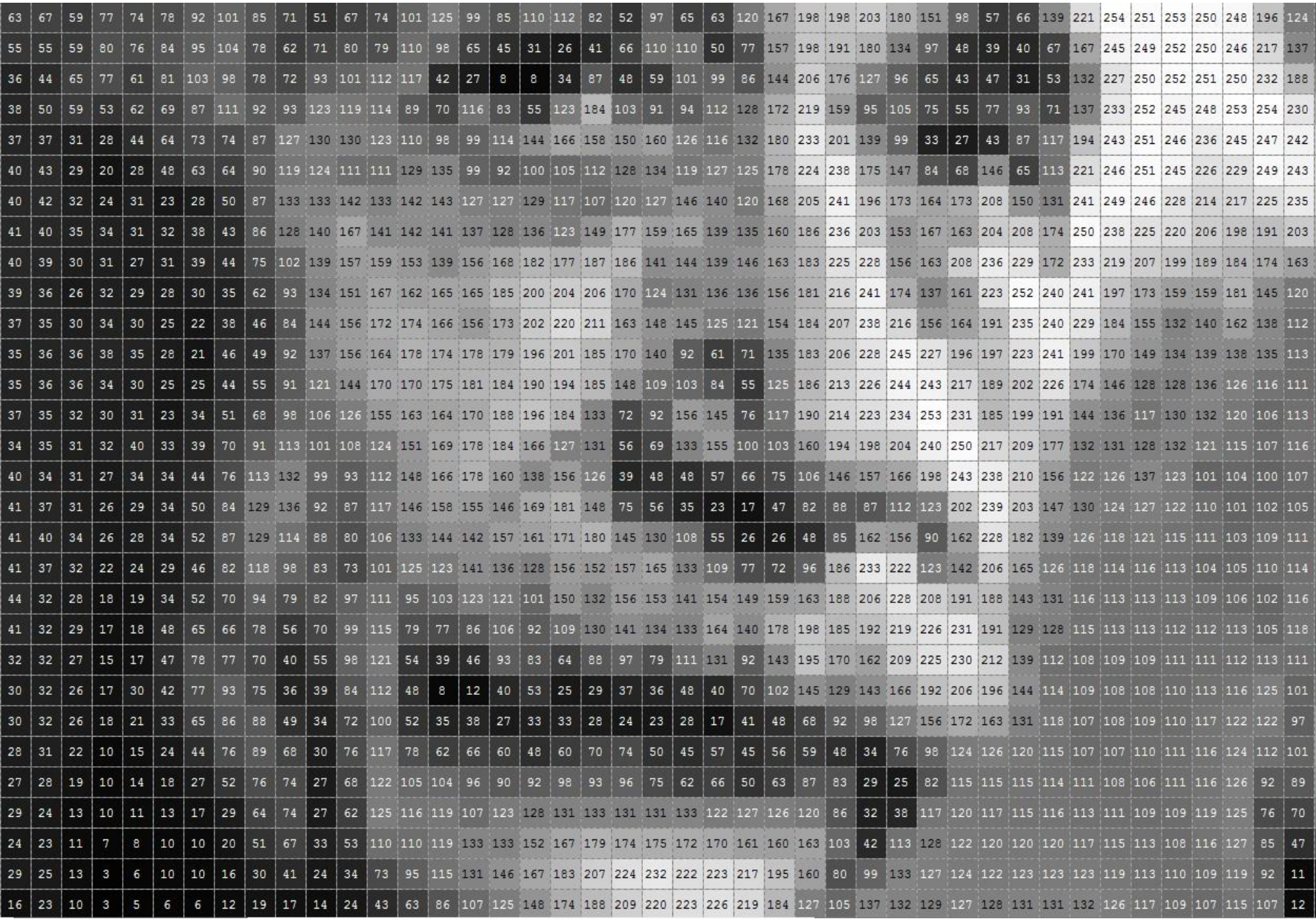
Part 1
What is Machine Learning?

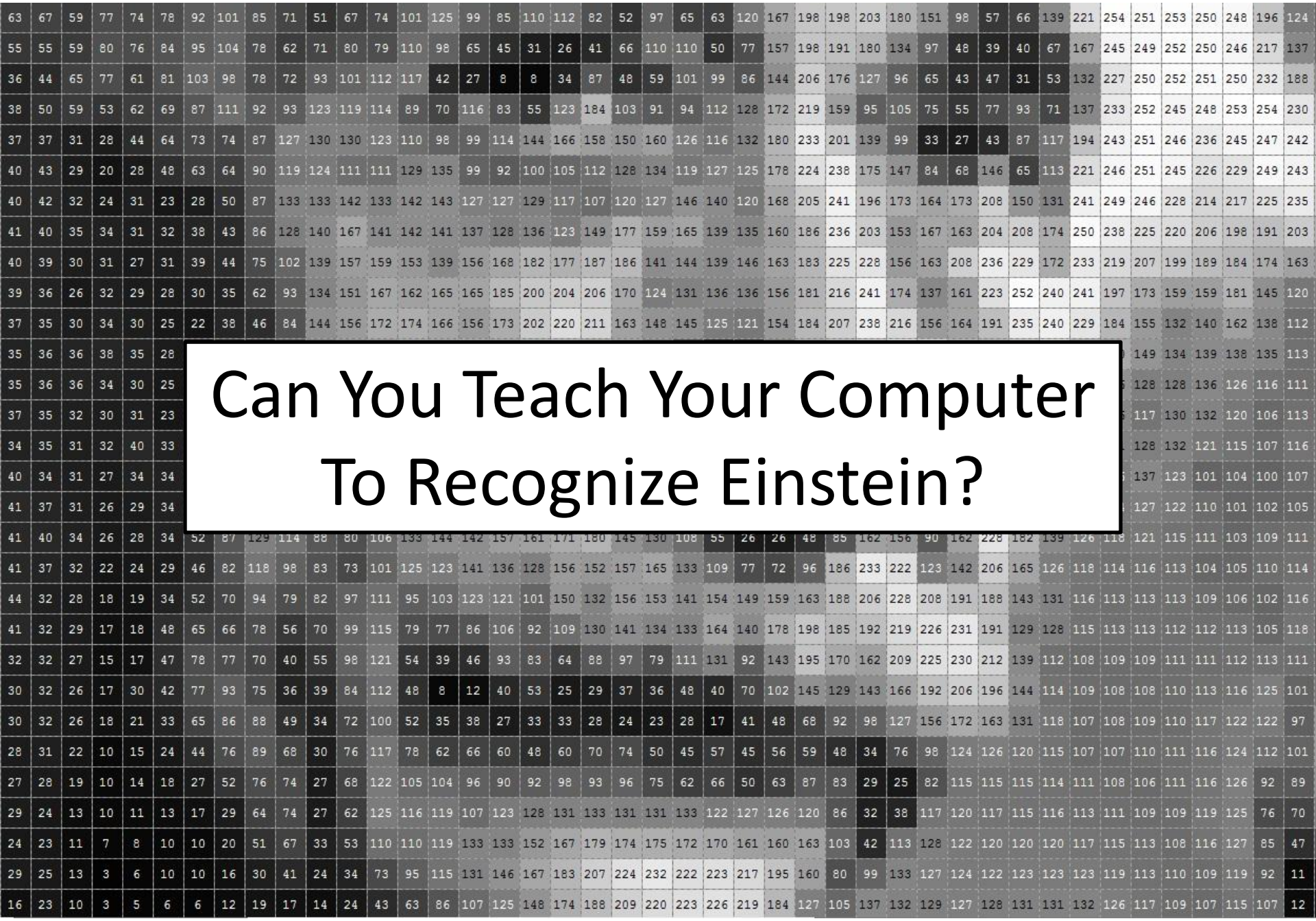
Who among these is Albert Einstein?










Can You Teach Your Computer
To Recognize Einstein?

Machine Learning

"Field of study that gives computers the ability to learn without being explicitly programmed."

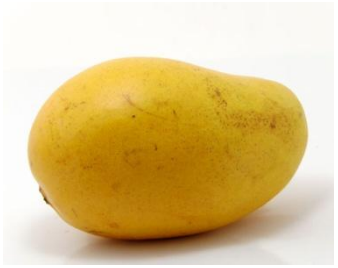
- Arthur Samuel

Raw Mango Vs. Ripen Mango

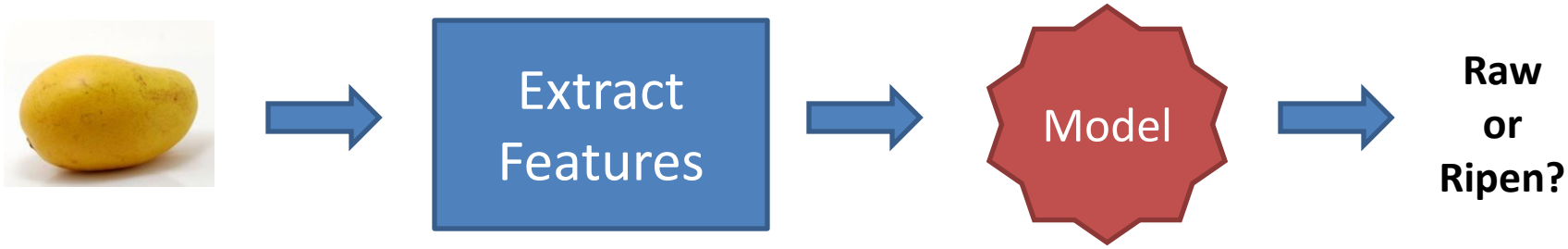
Raw Mango Vs. Ripen Mango

Raw Mango Vs. Ripen Mango

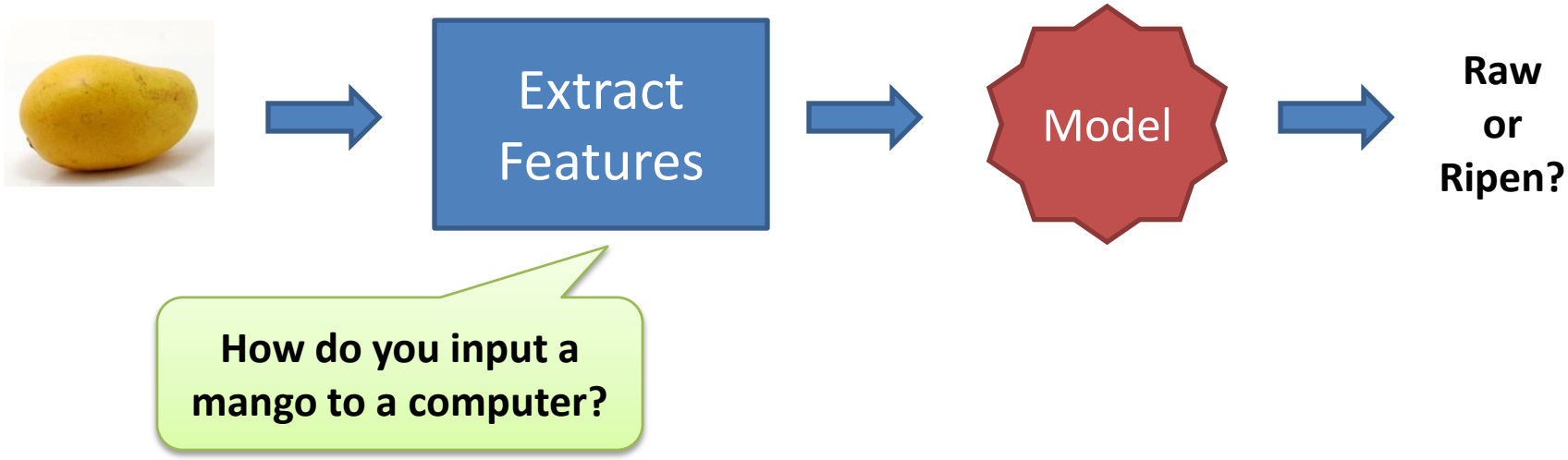
Raw Mango Vs. Ripen Mango



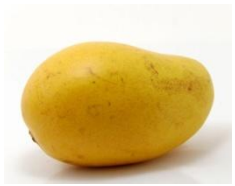
Machine Learning System



Machine Learning System



Machine Learning System



Extract Features

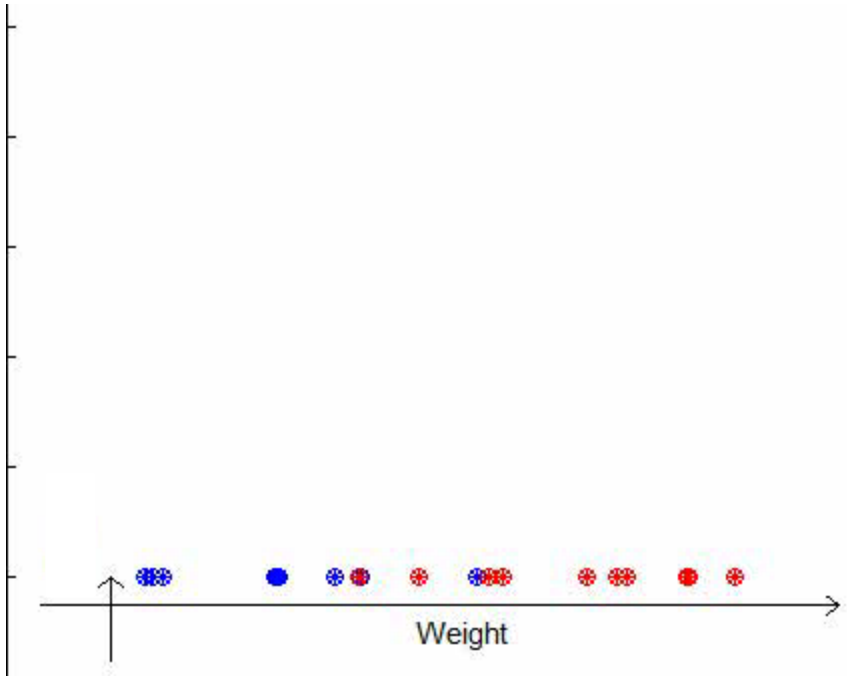
Model

Raw or Ripen?

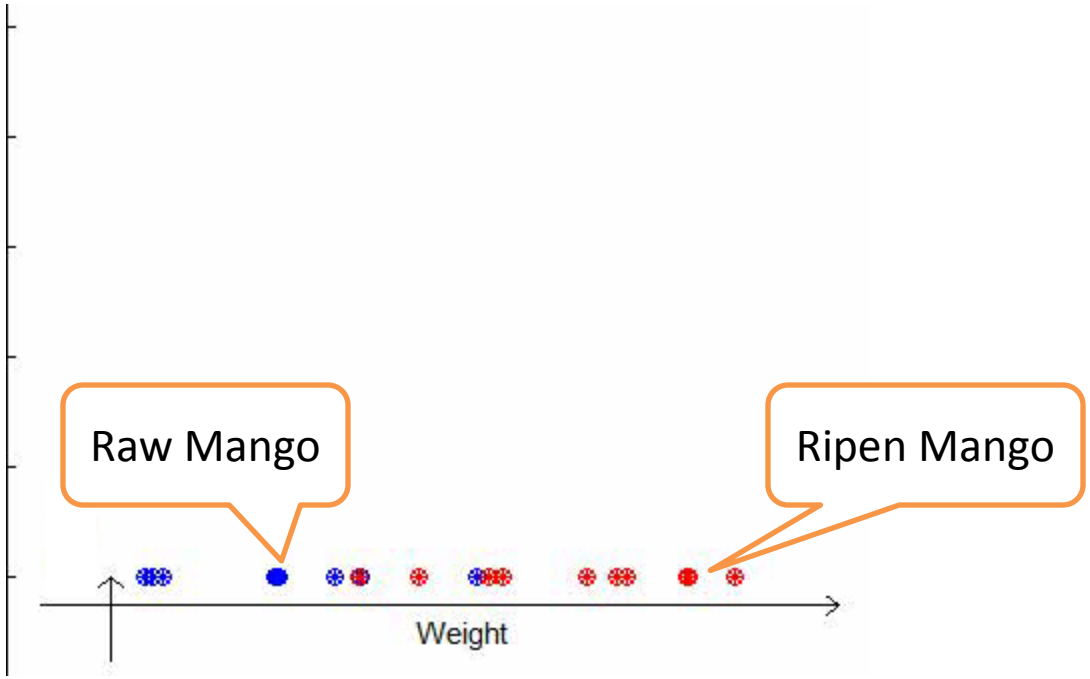
How do you input a mango to a computer?

How do you make a prediction?

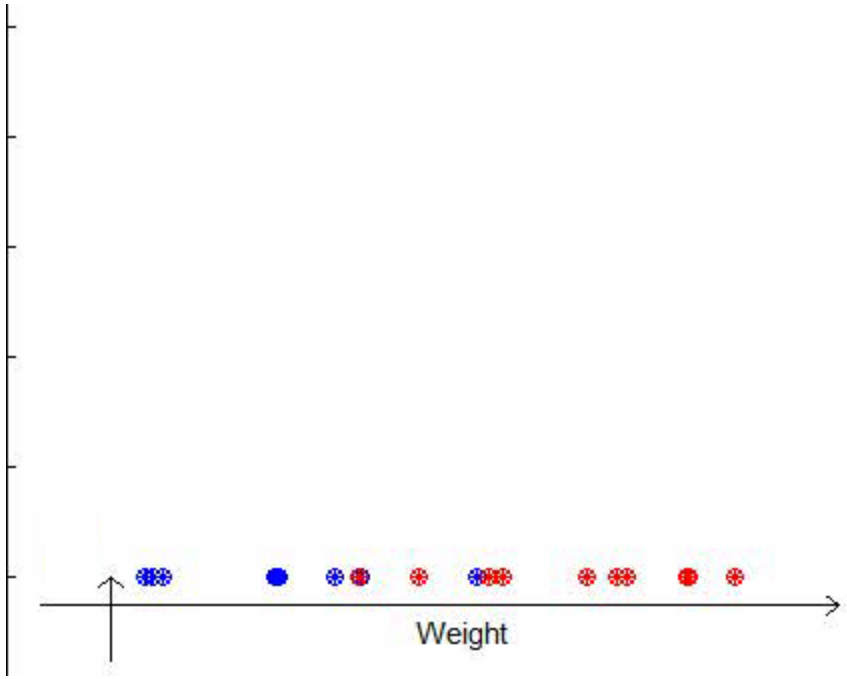
How do you represent a mango?



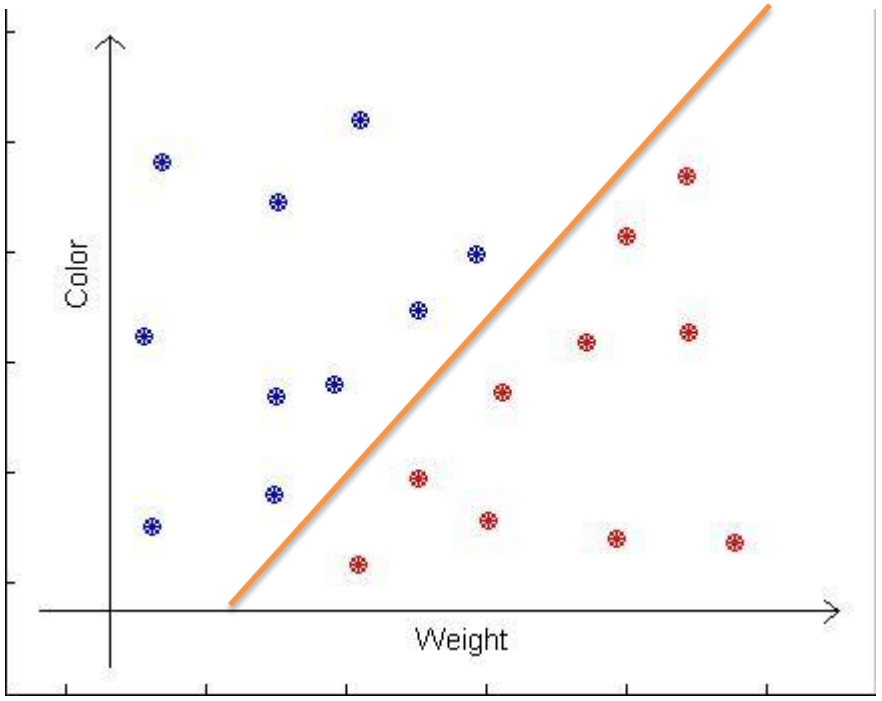
How do you represent a mango?



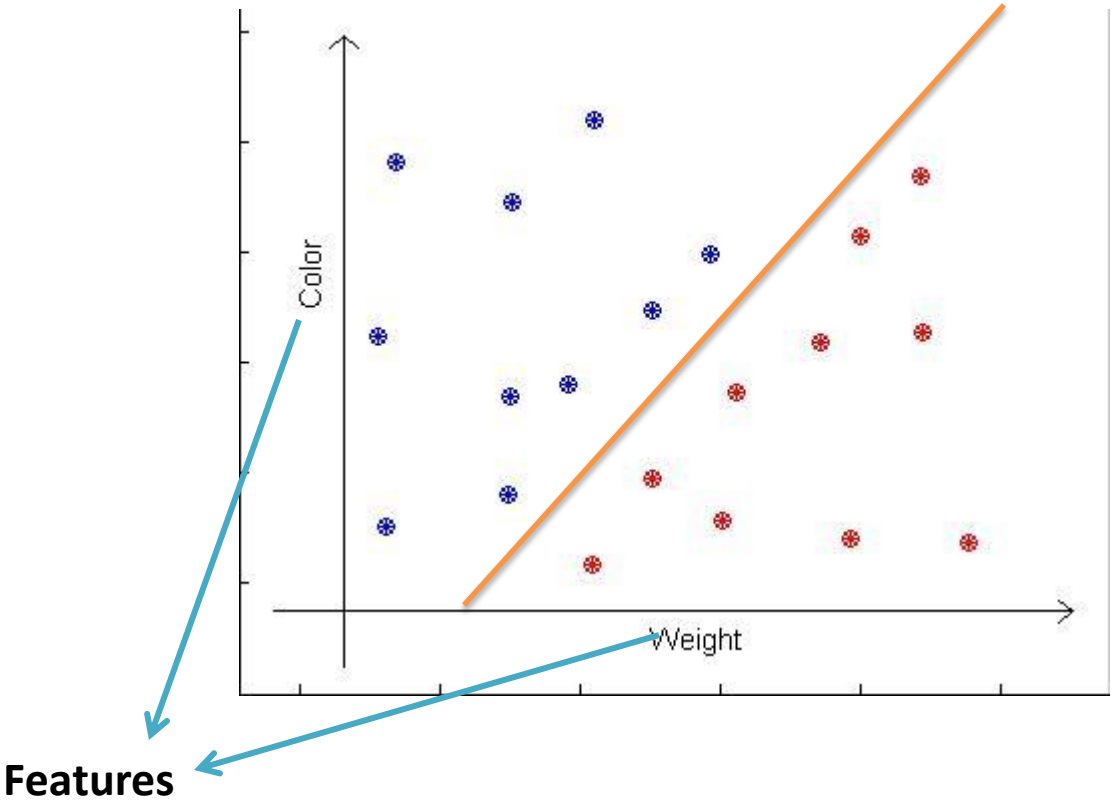
How do you represent a mango?



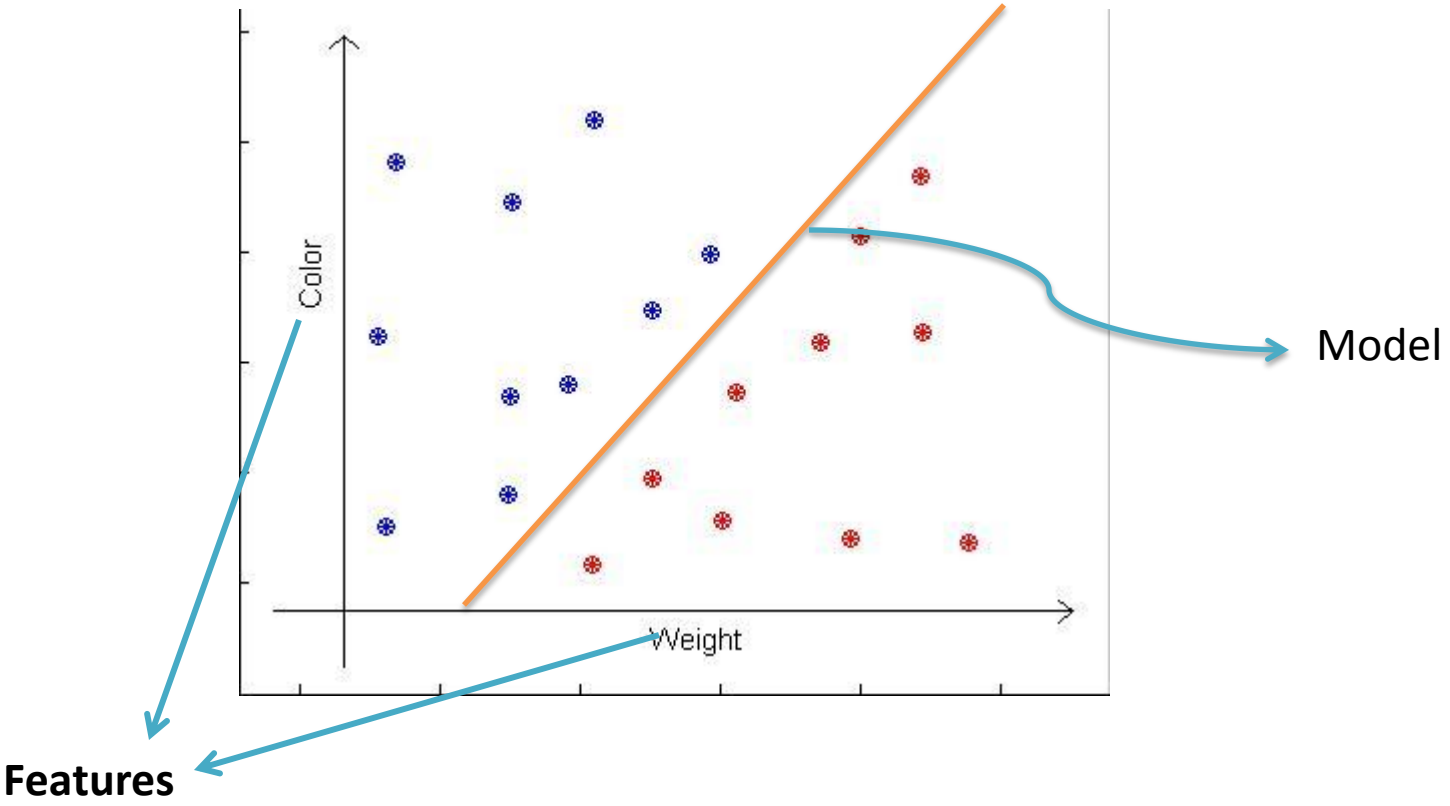
How do you represent a mango?



How do you represent a mango?



How do you represent a mango?



Spam Filter

Spam Filter

From Allen and Violet Large Spam x

saavedra saavedra@filo.uba.ar
to undisclosed recipients

May 9

Be careful with this message. Similar messages were used to steal people's personal information. Unless you trust the sender, don't click links or reply with personal information. Learn more

--

Dear Sir/Madam,

This is my fifth times of writting you this email since last year till date but no response from you. Hope you get this one, as this is a personal email directed to you.
My wife and I won a Jackpot Lottery of $11.3 million in July and have voluntarily decided to donate the sum of $500,000.00 USD to you as part of our own charity project to improve the lot of 10 lucky individuals all over the world. If you have received this email then you are one of the lucky recipients and all you have to do is get back with us so that we can send your details to the payout bank. Please note that you have to contact my private email for more informations (allenvioletlarge2@ozledim.net)

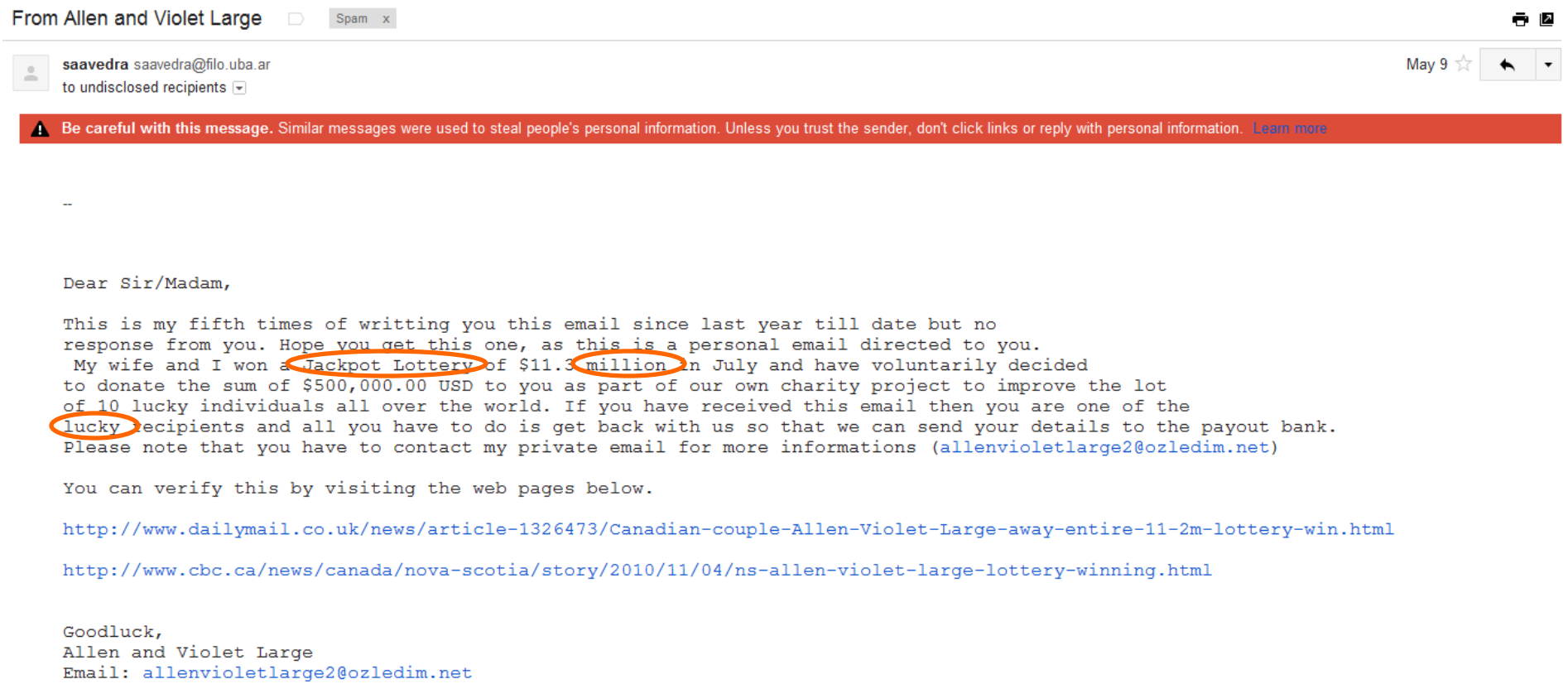
You can verify this by visiting the web pages below.

http://www.dailymail.co.uk/news/article-1326473/Canadian-couple-Allen-Violet-Large-away-entire-11-2m-lottery-win.html

http://www.cbc.ca/news/canada/nova-scotia/story/2010/11/04/ns-allen-violet-large-lottery-winning.html

Goodluck,
Allen and Violet Large
Email: allenvioletlarge2@ozledim.net

Spam Filter



From Allen and Violet Large Spam x

saavedra saavedra@filo.uba.ar
to undisclosed recipients

May 9

Be careful with this message. Similar messages were used to steal people's personal information. Unless you trust the sender, don't click links or reply with personal information. Learn more

--

Dear Sir/Madam,

This is my fifth times of writting you this email since last year till date but no response from you. Hope you get this one, as this is a personal email directed to you.
My wife and I won a Jackpot Lottery of $11.3 million in July and have voluntarily decided to donate the sum of $500,000.00 USD to you as part of our own charity project to improve the lot of 10 lucky individuals all over the world. If you have received this email then you are one of the lucky recipients and all you have to do is get back with us so that we can send your details to the payout bank.
Please note that you have to contact my private email for more informations (allenvioletlarge2@ozledim.net)

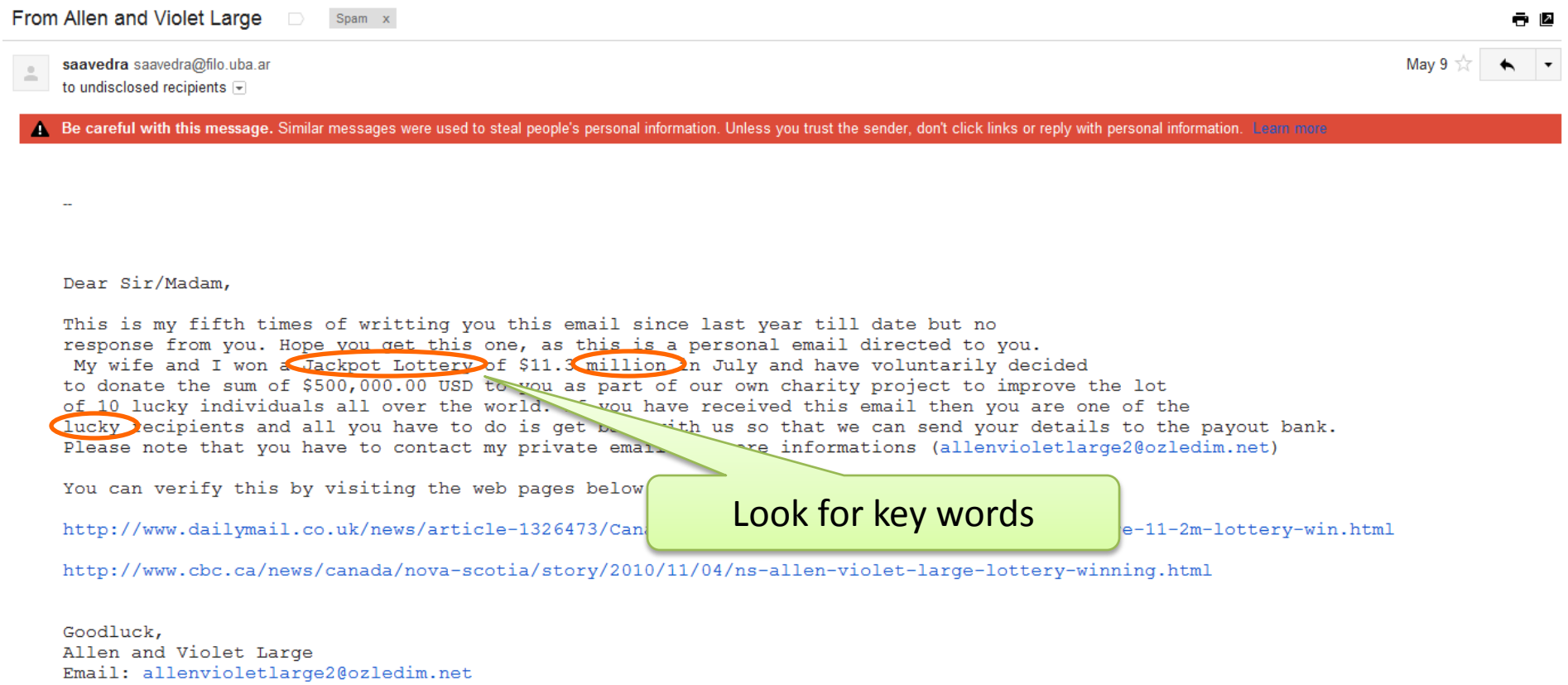
You can verify this by visiting the web pages below.

http://www.dailymail.co.uk/news/article-1326473/Canadian-couple-Allen-Violet-Large-away-entire-11-2m-lottery-win.html

http://www.cbc.ca/news/canada/nova-scotia/story/2010/11/04/ns-allen-violet-large-lottery-winning.html

Goodluck,
Allen and Violet Large
Email: allenvioletlarge2@ozledim.net

Spam Filter


From Allen and Violet Large
Spam x
saavedra saavedra@filo.uba.ar
to undisclosed recipients
May 9
Be careful with this message. Similar messages were used to steal people's personal information. Unless you trust the sender, don't click links or reply with personal information. Learn more
--
Dear Sir/Madam,
This is my fifth times of writting you this email since last year till date but no
response from you. Hope you get this one, as this is a personal email directed to you.
My wife and I won a Jackpot Lottery of $11.3 million in July and have voluntarily decided
to donate the sum of $500,000.00 USD to you as part of our own charity project to improve the lot
of 10 lucky individuals all over the world. If you have received this email then you are one of the
lucky recipients and all you have to do is get back with us so that we can send your details to the payout bank.
Please note that you have to contact my private email for more informations (allenvioletlarge2@ozledim.net)
You can verify this by visiting the web pages below
http://www.dailymail.co.uk/news/article-1326473/Can
e-11-2m-lottery-win.html
http://www.cbc.ca/news/canada/nova-scotia/story/2010/11/04/ns-allen-violet-large-lottery-winning.html
Goodluck,
Allen and Violet Large
Email: allenvioletlarge2@ozledim.net
Look for key words

Spam Filter

Spam Filter

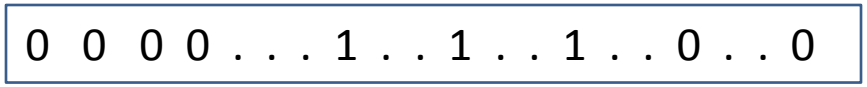
Dictionary Words

aardwolf	✗
abacus	✗
abandon	✗
abbreviate	✗
abdicate	✗
.	
.	
dollar	✓
.	
.	
jackpot	✓
.	
.	
lottery	✓
.	
.	
zygotic	✗
zymurgy	✗

Spam Filter

Dictionary Words

aardwolf	✗
abacus	✗
abandon	✗
abbreviate	✗
abdicate	✗
.	
.	
dollar	✓
.	
.	
jackpot	✓
.	
.	
lottery	✓
.	
.	
zygotic	✗
zymurgy	✗

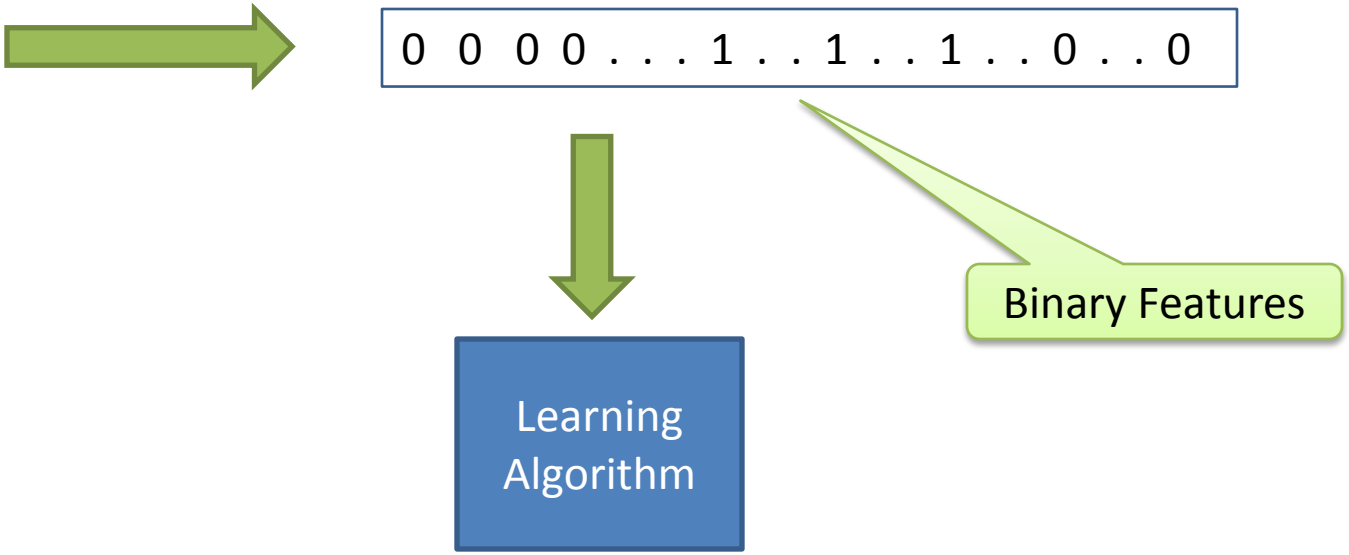


Binary Features

Spam Filter

Dictionary Words

aardwolf	✗
abacus	✗
abandon	✗
abbreviate	✗
abdicate	✗
.	
.	
dollar	✓
.	
.	
jackpot	✓
.	
.	
lottery	✓
.	
.	
zygotic	✗
zymurgy	✗



Building a Learning System

Building a Learning System

Training Data

Ripen

Raw

.
.

Raw

Ripen

Building a Learning System

Training Data

Ripen

Raw

.

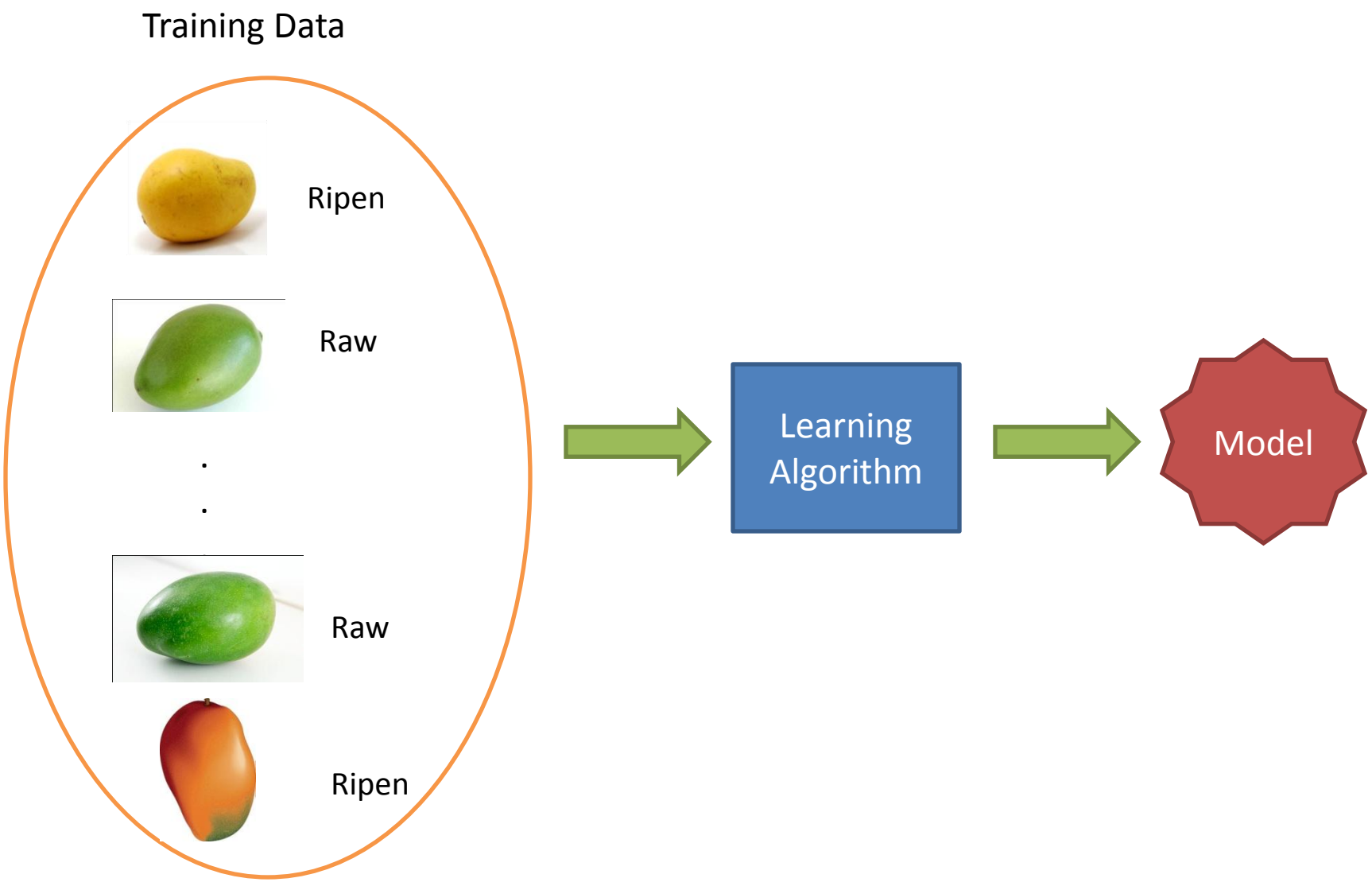
.

Raw

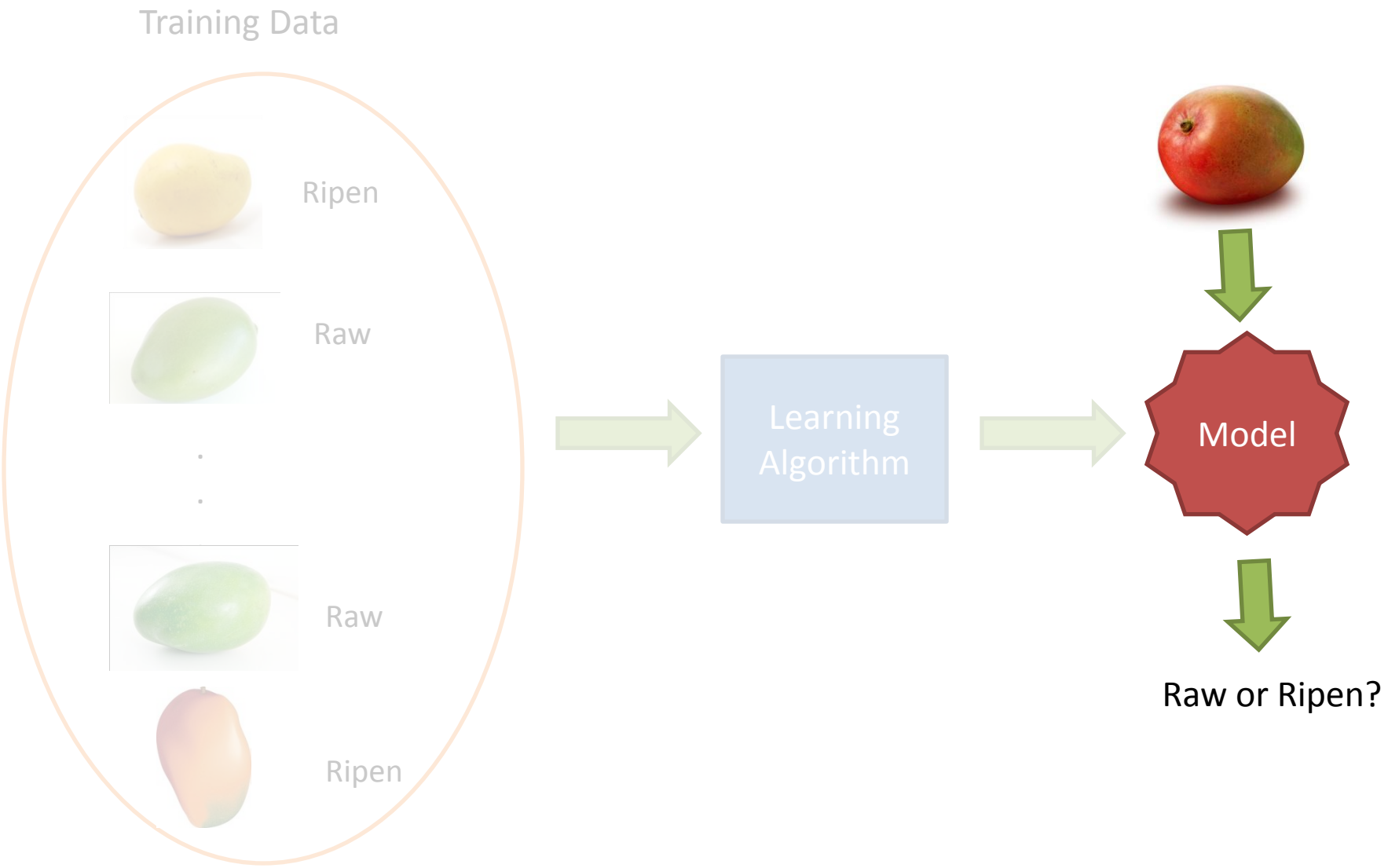
Ripen

Learning Algorithm

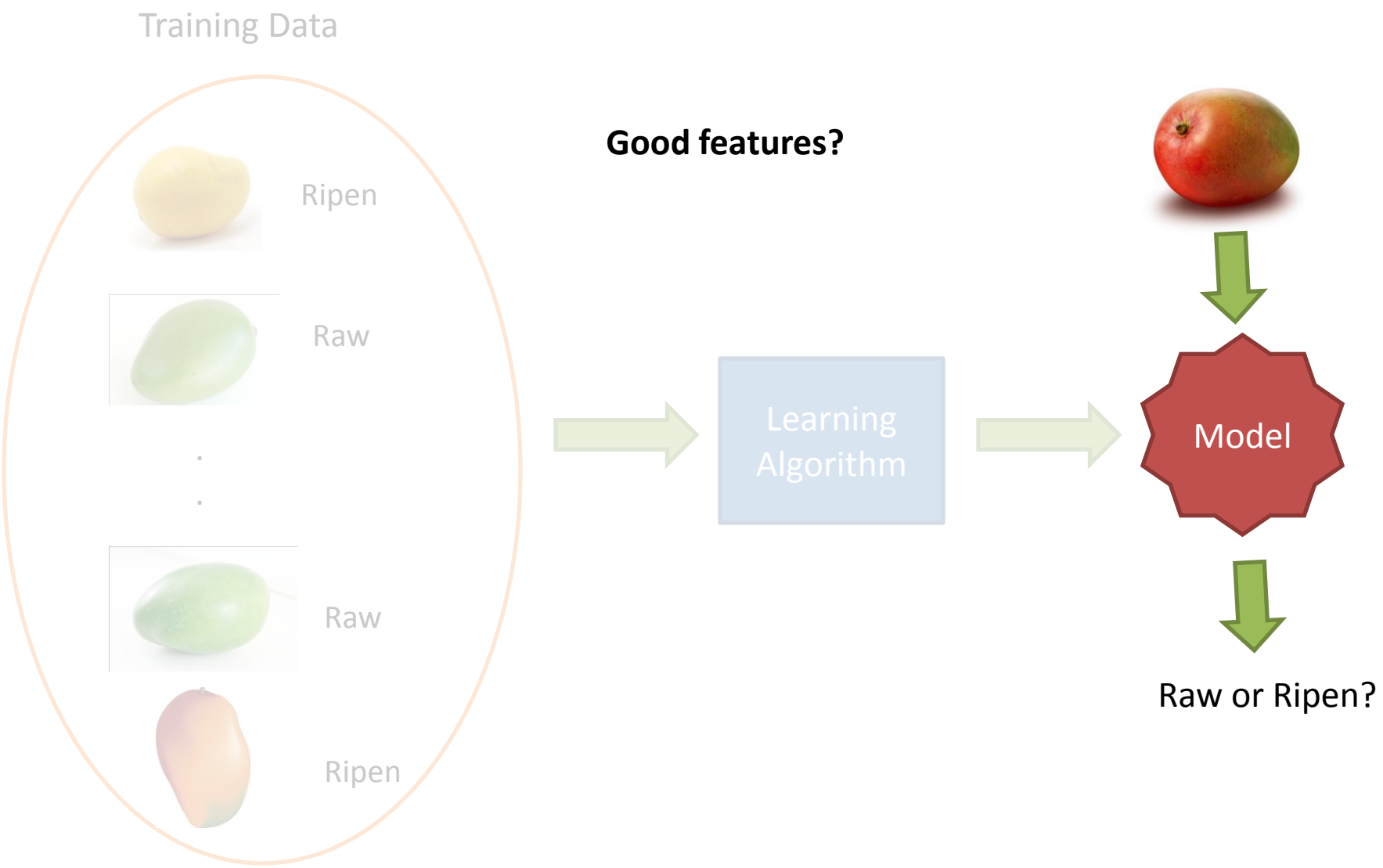
Building a Learning System



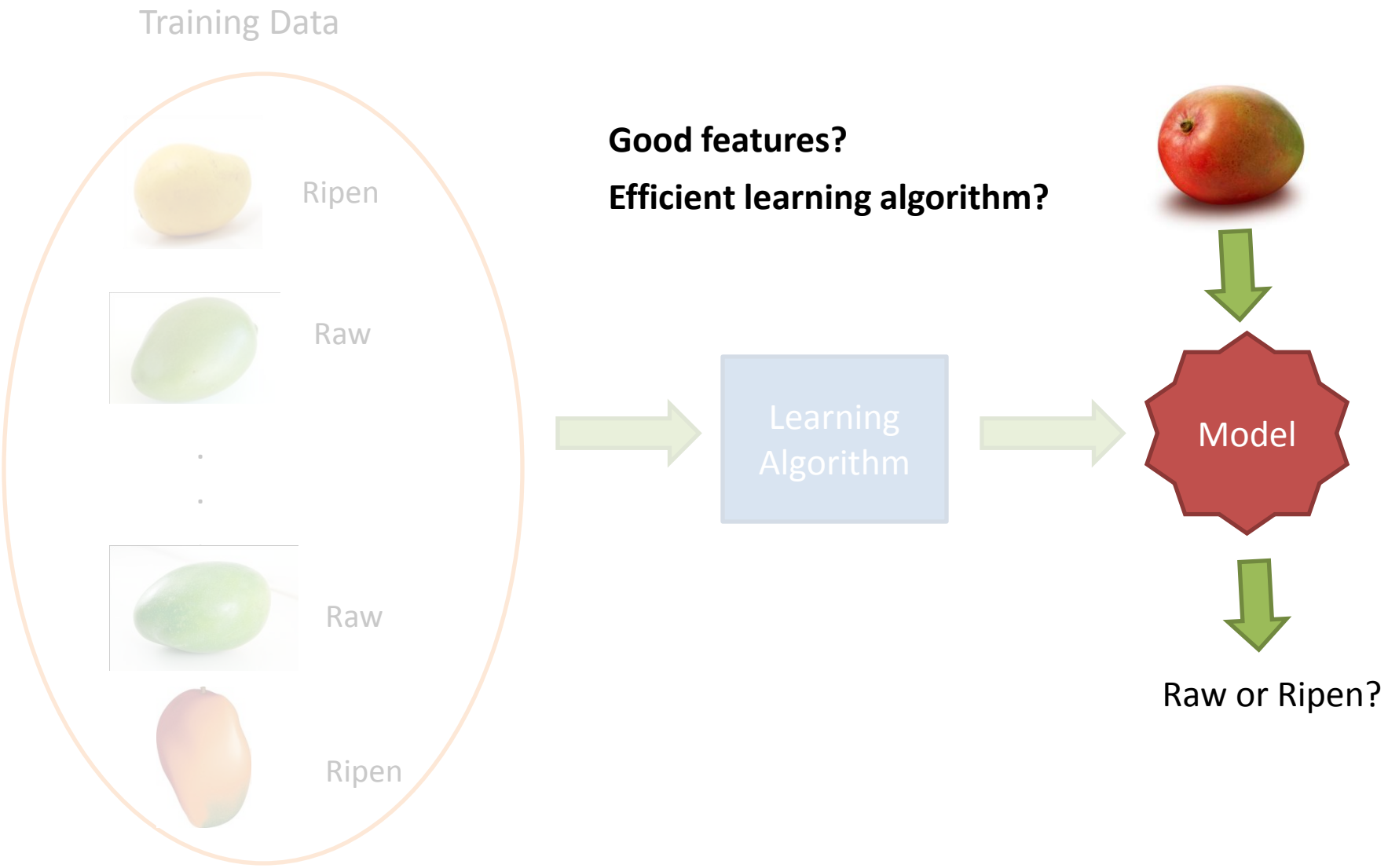
Building a Learning System



Building a Learning System



Building a Learning System



Building a Learning System

Training Data

Good features?

Efficient learning algorithm?

Good model?

Learning Algorithm

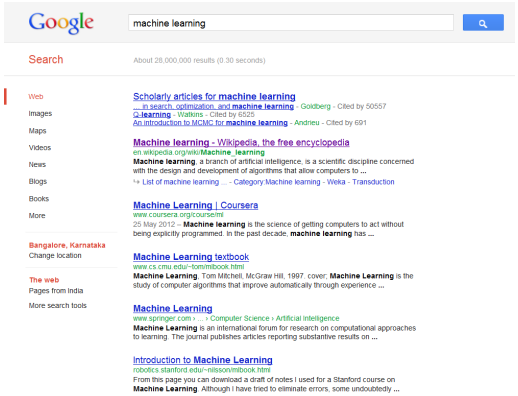
Model

Raw or Ripen?

Applications of Machine Learning

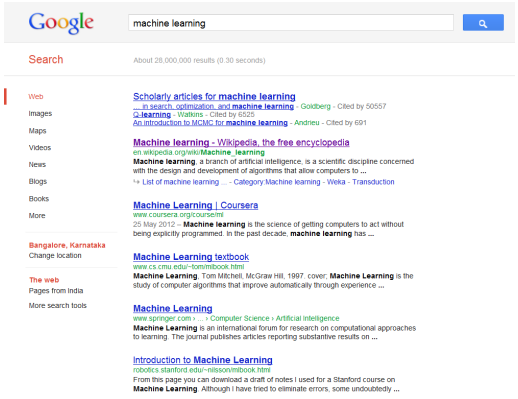
Applications of Machine Learning

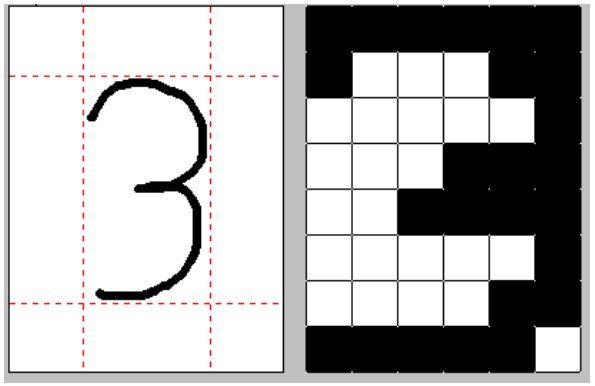
- Information Retrieval



Applications of Machine Learning

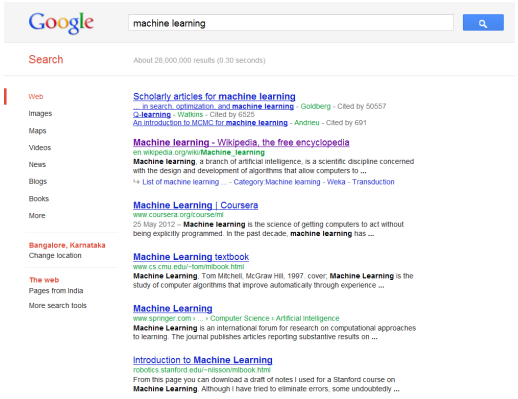
- Information Retrieval
- Computer Vision

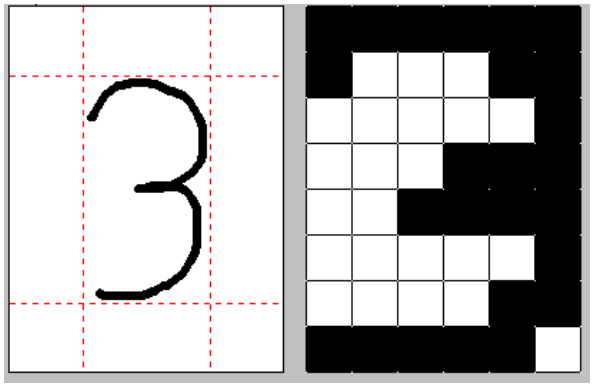


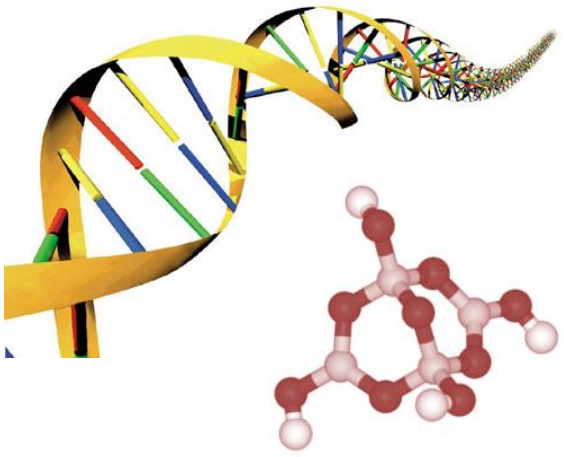


Applications of Machine Learning

- Information Retrieval
- Computer Vision
- Bioinformatics

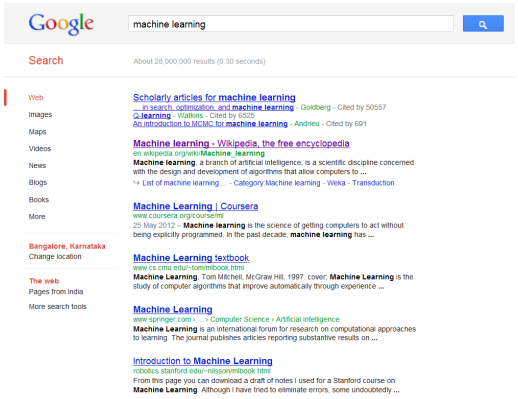


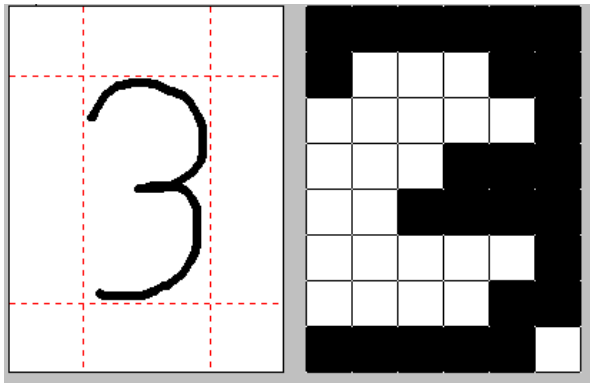


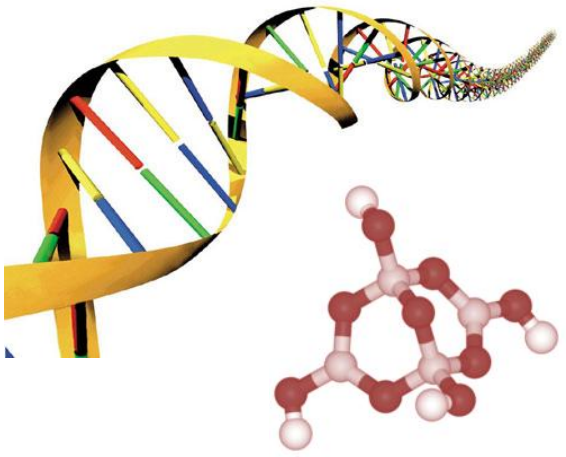


Applications of Machine Learning

- Information Retrieval
- Computer Vision
- Bioinformatics
- Natural Language Processing



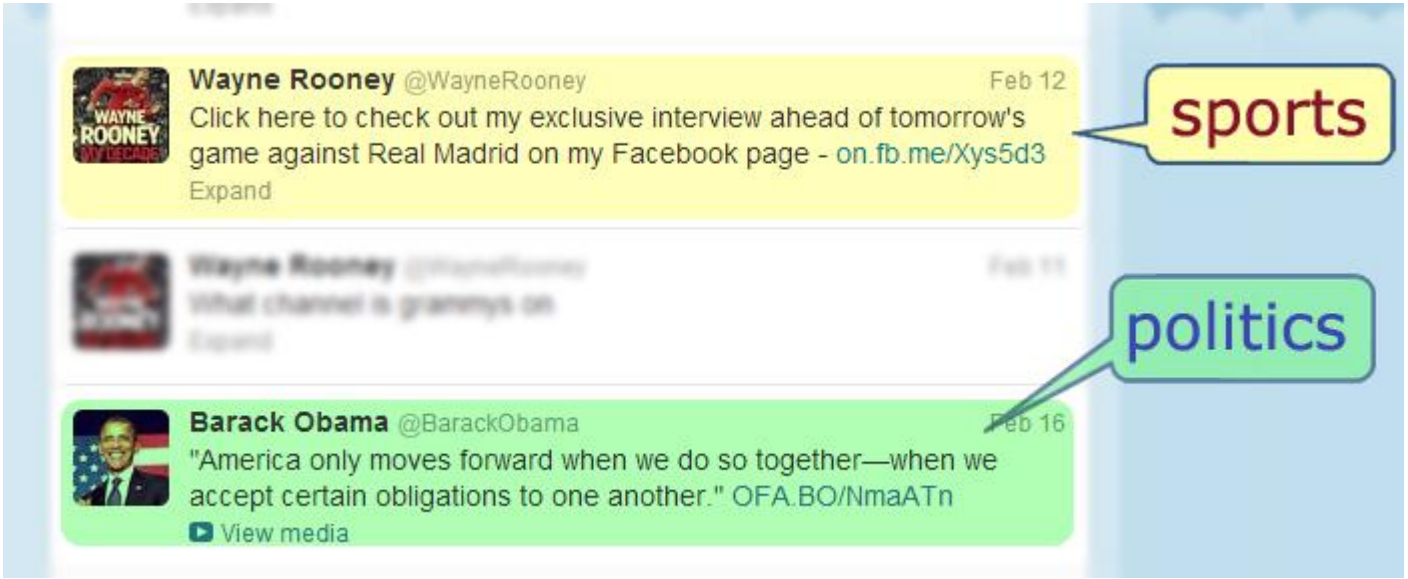




Part 2
The TwitMiner Challenge 2013

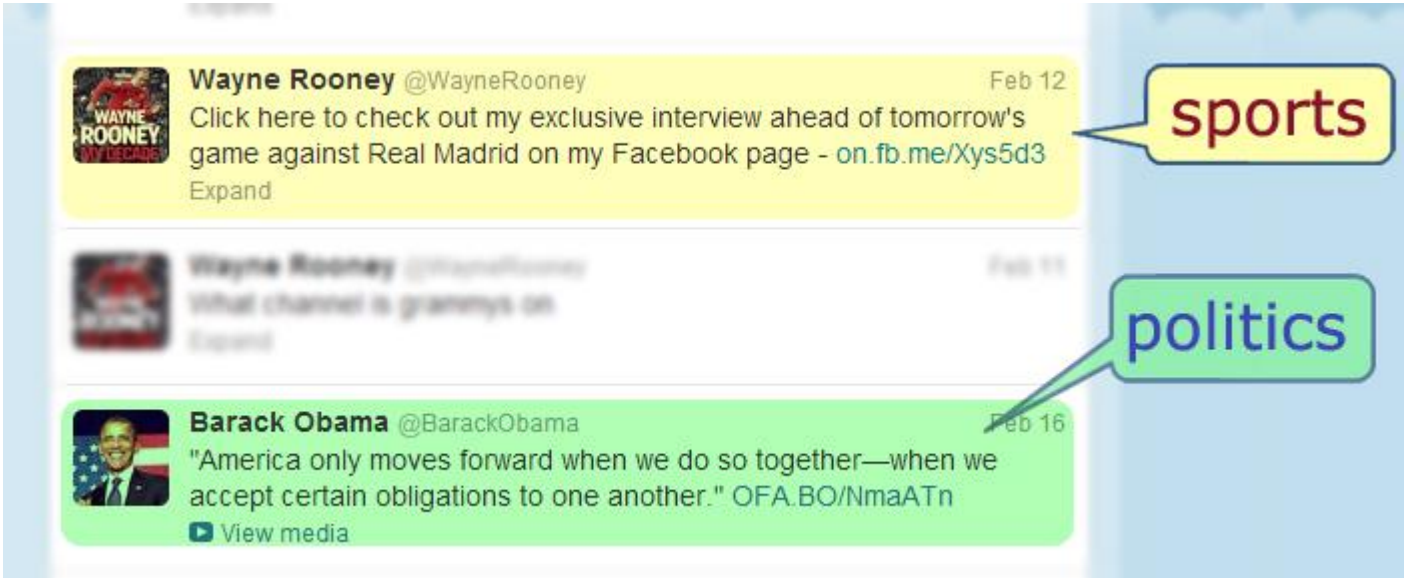
The Challenge

- Build a system that can categorize tweets into sports and politics.



The Challenge

- Build a system that can categorize tweets into sports and politics.



- Data in a single tweet is limited by number of characters.

Data

Data

- Training data set: <tweet-id, user-id, label>

Data

- Training data set: <tweet-id, user-id, label>
- Validation data set: <tweet-id, user-id >

Data

- Training data set: <tweet-id, user-id, label>
- Validation data set: <tweet-id, user-id >
- Test data set: <tweet-id, user-id >

Data

Used to build a prediction model

- Training data set: <tweet-id, user-id, label>
- Validation data set: <tweet-id, user-id >
- Test data set: <tweet-id, user-id >

Data

- Training data set: <tweet-id, user-id, label>
- Validation data set: <tweet-id, user-id >
- Test data set: <tweet-id, user-id >

Used to build a prediction model

Used to evaluate your model

Data

Used to build a prediction model

- Training data set: <tweet-id, user-id, label>
- Validation data set: <tweet-id, user-id >
- Test data set: <tweet-id, user-id >

Used to evaluate your model

Tweet text not provided!

A separate script provided to download tweet texts.

Data

Used to build a prediction model

- Training data set: <tweet-id, user-id, label>
- Validation data set: <tweet-id, user-id >
- Test data set: <tweet-id, user-id >

Used to evaluate your model

Tweet text not provided!

A separate script provided to download tweet texts.

Data to be uploaded soon.. (before March 3)

Getting Started

1. Extract features from tweet text

Getting Started

1. Extract features from tweet text
2. Build a model using a learning algorithm

Getting Started

1. Extract features from tweet text
2. Build a model using a learning algorithm
3. Tune the model using validation data

Getting Started

1. Extract features from tweet text
2. Build a model using a learning algorithm
3. Tune the model using validation data
4. Submit predicted labels on test data

Submissions

- Submission portal to be online soon..
- Predicted labels for validation data
  - Helps you tune your model
- Predicted labels for test data
  - Used to evaluate your model

Any number of times!

Only once!

Prizes

- **First Prize:** Internship/PPI offers at amazon.com
- **Second Prize:** Prizes worth INR 10000
- **Third Prize:** Prizes worth INR 6000

Rules

Rules

- Each team can have a max. of 2 members.

Rules

- Each team can have a max. of 2 members.
- Event open from March 1-15.

Rules

- Each team can have a max. of 2 members.
- Event open from March 1-15.
- Teams to upload a write up on their approach with properly documented code.

Rules

- Each team can have a max. of 2 members.
- Event open from March 1-15.
- Teams to upload a write up on their approach with properly documented code.
- Both test accuracy and submitted write up/code will be used to decide the winner.

Important Dates

Important Dates

- 1-Mar 2013: Event declared open

Important Dates

- 1-Mar 2013: Event declared open
- 3-Mar 2013: Registration begins

Important Dates

- 1-Mar 2013: Event declared open
- 3-Mar 2013: Registration begins
- 15-Mar 2013:
  - IST 12:00 AM: Test data made available
  - IST 11:59 PM: Event comes to an end

Important Dates

- 1-Mar 2013: Event declared open
- 3-Mar 2013: Registration begins
- 15-Mar 2013:
  - IST 12:00 AM: Test data made available
  - IST 11:59 PM: Event comes to an end
- 31-Mar 2013: Declaration of winners

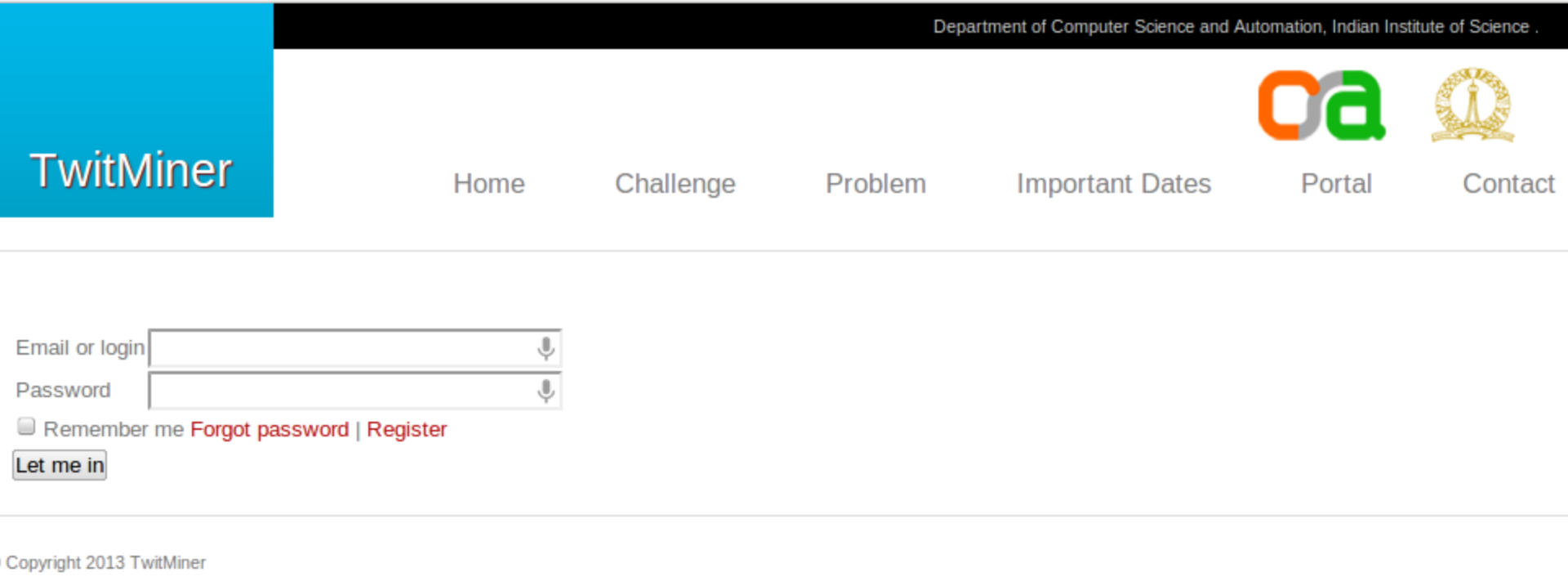
Visit us at: http://events.csa.iisc.ernet.in/opendays2013/twitminer/

Visit us at: http://events.csa.iisc.ernet.in/opendays2013/twitminer/

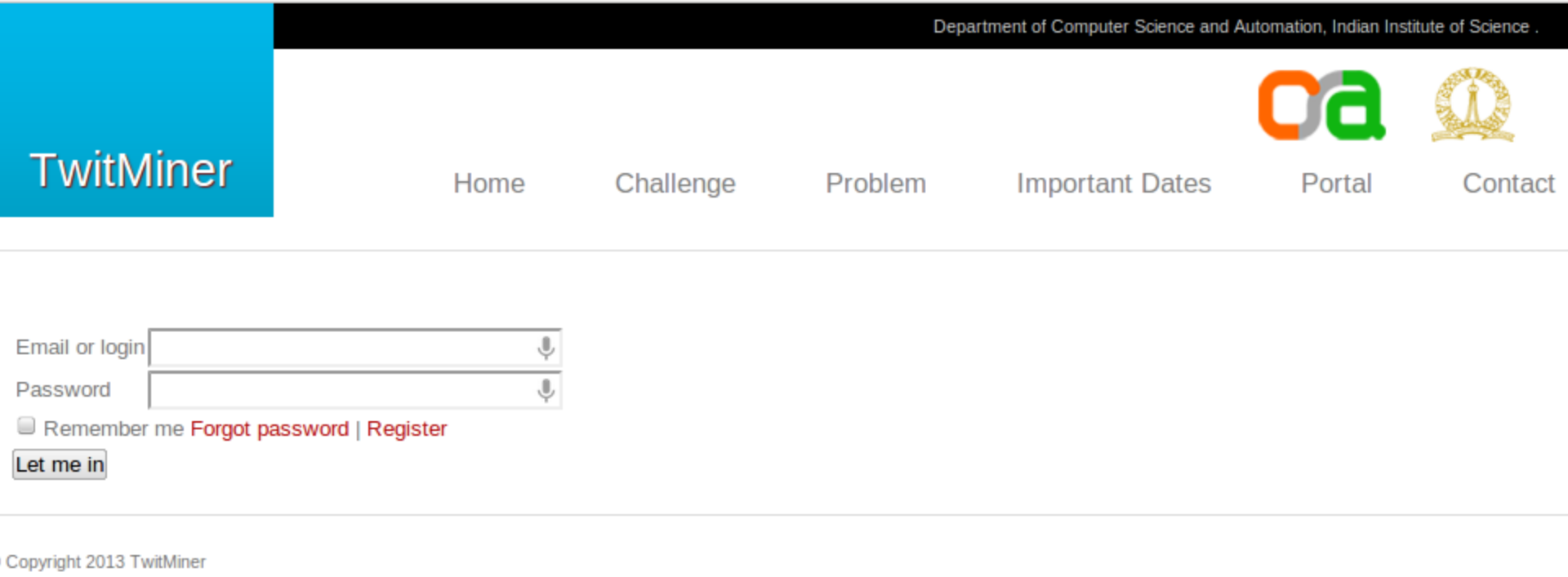
Contact Us: twitminer2013@gmail.com

Facebook: http://www.facebook.com/twitminer

Registrations open from Mar 3rd…



Registrations open from Mar 3rd…



Questions?