

# A gentle introduction to machine learning

Aidan Finn  
@aidanf



# What is machine learning?

- Programs that learn by example
- Data-mining - applying machine learning to large databases (Big Data)

# Some applications

- Shopping. Predict what the customer will buy
- Finance. Credit scoring, fraud detection
- Medical diagnosis
- Search engines

# Example: spam classification

[Walk In Showers](#) - [bathingsolutions.co.uk/Showers](#) - Discover Our Walk In Shower Range. Request Your Free

[Delete all spam messages now](#) (messages that have been in Spam more than 30 days will

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<b>CasinoClub</b>	<b>30 free spins</b> - Casino No Deposit Promotions. is an America
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<b>Directory Network@authse.</b>	<b>A Free Cross-Platform Anti-Virus For All Your Systems</b> - E
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<b>FreeGames</b>	<b>Daily update - no deposit of the day</b> - Casino No Deposit F
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<b>B2B Search Engine</b>	<b>Dig Deeper Into Your Twitter Analytics, Find Out What Yo</b>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<b>iEntry</b>	<b>Determine Changes Made To File Versions Then Merge T</b>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<b>Change.org</b>	<b>Transphobia on Facebook</b> - Change.org Aidan - There's a
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<b>Uk Lottery Orgaization</b>	<b>UK Lotto US\$4.6MILLION free-ticket e-mail address winni</b>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<b>FIT Newsletter</b>	<b>Making the Most of the Final 3 Gameweeks – by James D</b>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<b>St. Baldrick's Foundation</b>	<b>Chase's Best Shot: One boy's cancer story</b> - Chase's mon
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<b>Expedited Shipping</b>	<b>Delivery Status Notification</b> - USPS .COM Notification Our c
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<b>B2B Search Engine</b>	<b>Assess Your Strengths and Weaknesses For Starting a N</b>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<b>Fantasy iTeam</b>	<b>FIT Summary Gameweek 35</b> - Find us on Facebook Follow u
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<b>SmallWebBusiness</b>	<b>Get the Number 1 CRM Solution for Small Business - Free</b>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<b>David Kash</b>	<b>In This Email: \$10-\$30 Coupon</b> - Casino No Deposit Promol

# Example: item recommendation

2.



[Why Grizzly Bears Should Wear Underpants \(The Oatmeal\)](#)  
by Matthew Inman



Don't use for recommendations

# Anatomy of a classification system

- Problem - what do I want to classify?
- Instances - examples where we know the classification
- Features - how do I represent it?
- ML algorithm - can be treated as a black box
- Training - learn from initial data
- Evaluation - does it work?
- Tuning - improve over time

# Example: Classifying news articles

http://www.reuters.com/politics/elections-2012

1,388 captures  
3 Jul 11 - 30 Apr 14

MAY JUN JUL  
2011 30 2012 2013

**REUTERS** EDITION: U.S. Register Sign In Search News & Quotes

Home Business Markets World **Politics** Tech Opinion Breakingviews Money Life Pictures Video

## Election 2012

Related Topics: POLITICAL PUNCHLINES TALES FROM THE TRAIL POLITICS CAMPAIGN POLLING ADD TO MY WIRE



### A tax or a penalty? Romney, Obama camps debate healthcare ruling

4:37pm EDT  
WASHINGTON - Republican presidential candidate Mitt Romney took his fight against President Barack Obama's newly upheld healthcare law out on the campaign trail on Friday, attempting to use it to galvanize support for his bid to oust Obama on November 6.

#### LATEST ELECTION NEWS

- Congress poised to wrap up transport, loans, flood bill**  
1:44pm EDT
- Romney fundraising keeps spiking, small donors giving too** VIDEO  
1:30pm EDT
- Young, worried, and unsure - about both candidates**  
11:28am EDT
- Supreme Court upholds Obama's healthcare law** VIDEO  
10:18am EDT
- Analysis: Why Roberts saved Obama's healthcare law**  
8:39am EDT
- "Now Teddy can rest," Pelosi tells Kennedy's widow**  
7:54am EDT
- Analysis: After healthcare victory in court,**

#### FOLLOW ELECTION 2012 NEWS

Follow RSS Email Error

## AMERICAN MOSAIC

- Part IV: Recent college graduates**  
A Reuters/Ipsos poll of recent graduates reveals a drop in support for President Barack Obama compared with 2008, but it shows no movement toward the GOP. [Full Article](#)
- Part III: Florida bingo set**
- Part II: The Rust Belt**
- Part I: Military families**

COUNTDOWN TO NOVEMBER 6 7 AM ET  
188 15 42

Latest from

# What do I want to classify?

- buddy-roemer
- ron-paul
- gary-johnson
- newt-gingrich
- rick-santorum
- michele-bachmann
- jon-huntsman
- thaddeus-mccotter
- tim-pawlenty



# How do I represent it?

- Title and body keywords

title\_romney  
title\_bachman  
title\_mccotter  
body\_washington  
body\_mccotter  
...  
body\_healthcare

# ML Algorithm

- Information Gain
- Figure out which combination of keywords are good at distinguishing between categories.
- Build rules

# Training

- Initial training data is manually annotated
- Generate rules based on this data
- `title_washington + body_romney + body_huntsman`  
`=> jon_huntsman`
- Repeat for each class

# Evaluation

- Does it work?
- Check on training data
- Check on new data

# Tuning

- Manually check for errors repeatedly
- Add mistakes to training data and re-generate rules
- Eventually good enough (or as good as it will get)

# DIY

- Libraries
- APIs
- Simple prediction services (Upload a spreadsheet)

Aidan Finn  
@aidanf

[www.aidanf.net](http://www.aidanf.net)

[www.windmill.ie](http://www.windmill.ie)

