A gentle introduction to machine learning

Aidan Finn
@aidanf

What is machine learning?

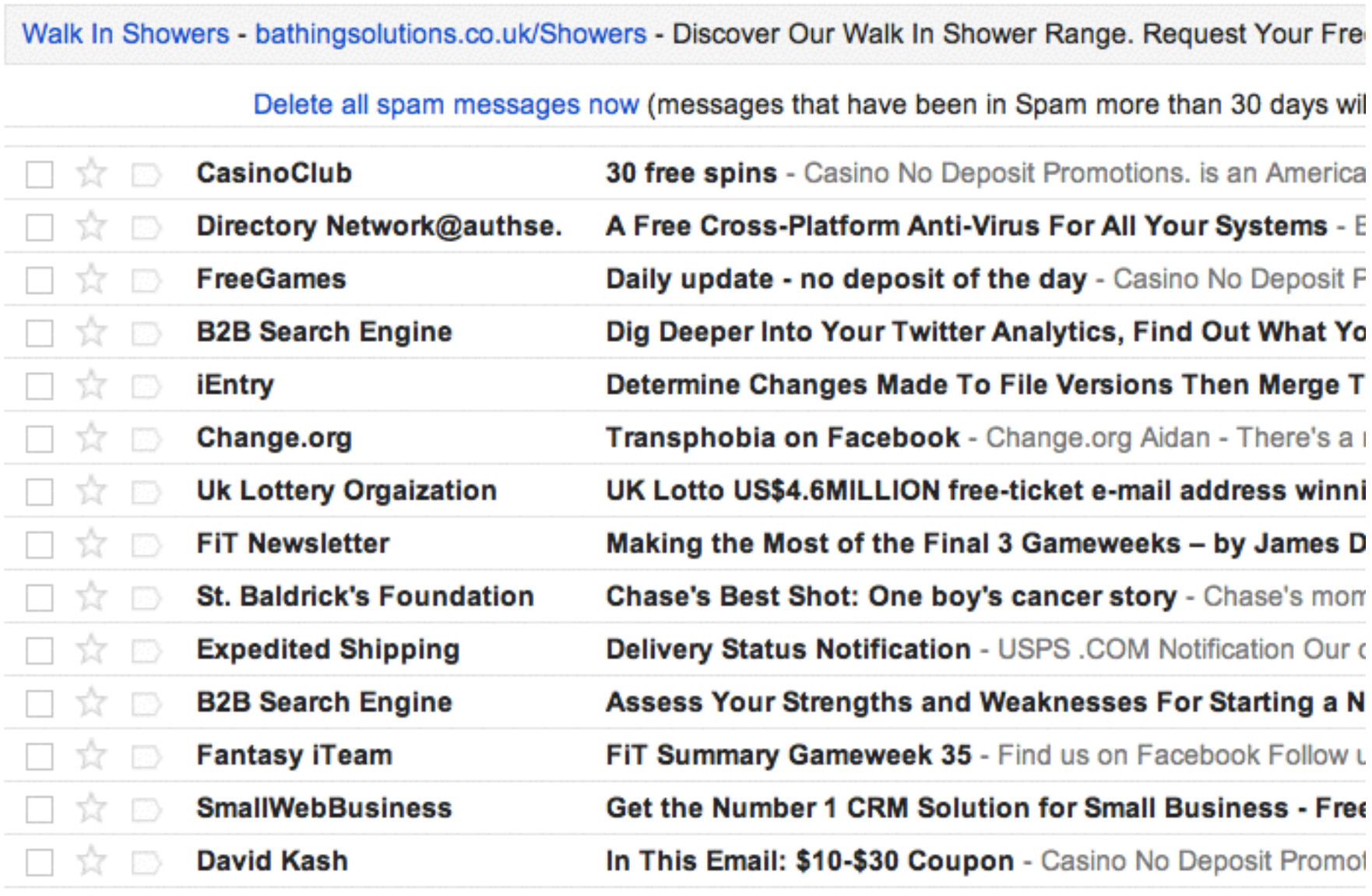
- Programs that learn by example

- 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 Fre

Delete all spam messages now (messages that have been in Spam more than 30 days wil

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

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

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

www.windmill.ie