Recommended for You: An Algorithmic View into Product Recommendations

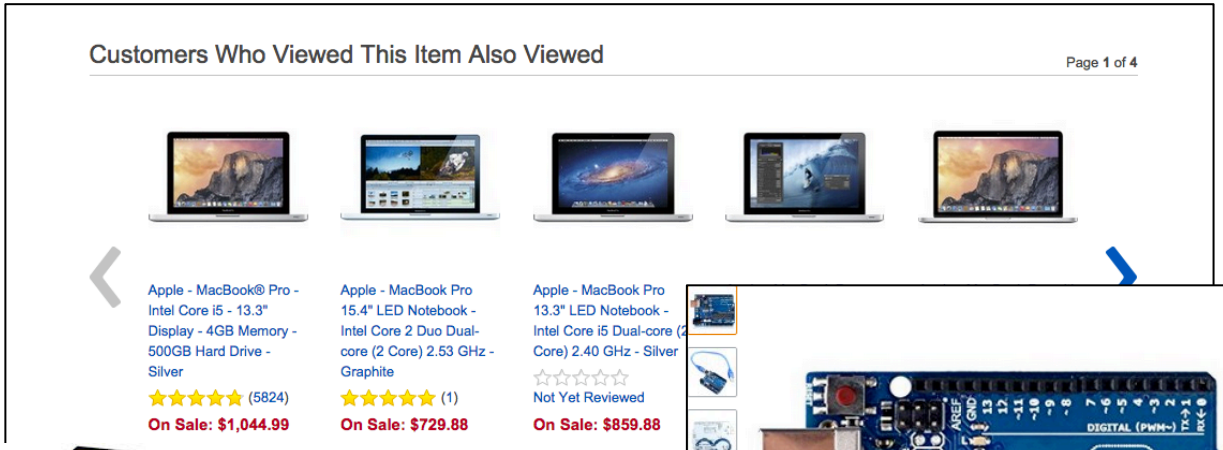
by Josh Janzen

BEST BUY DATA SCIENCE

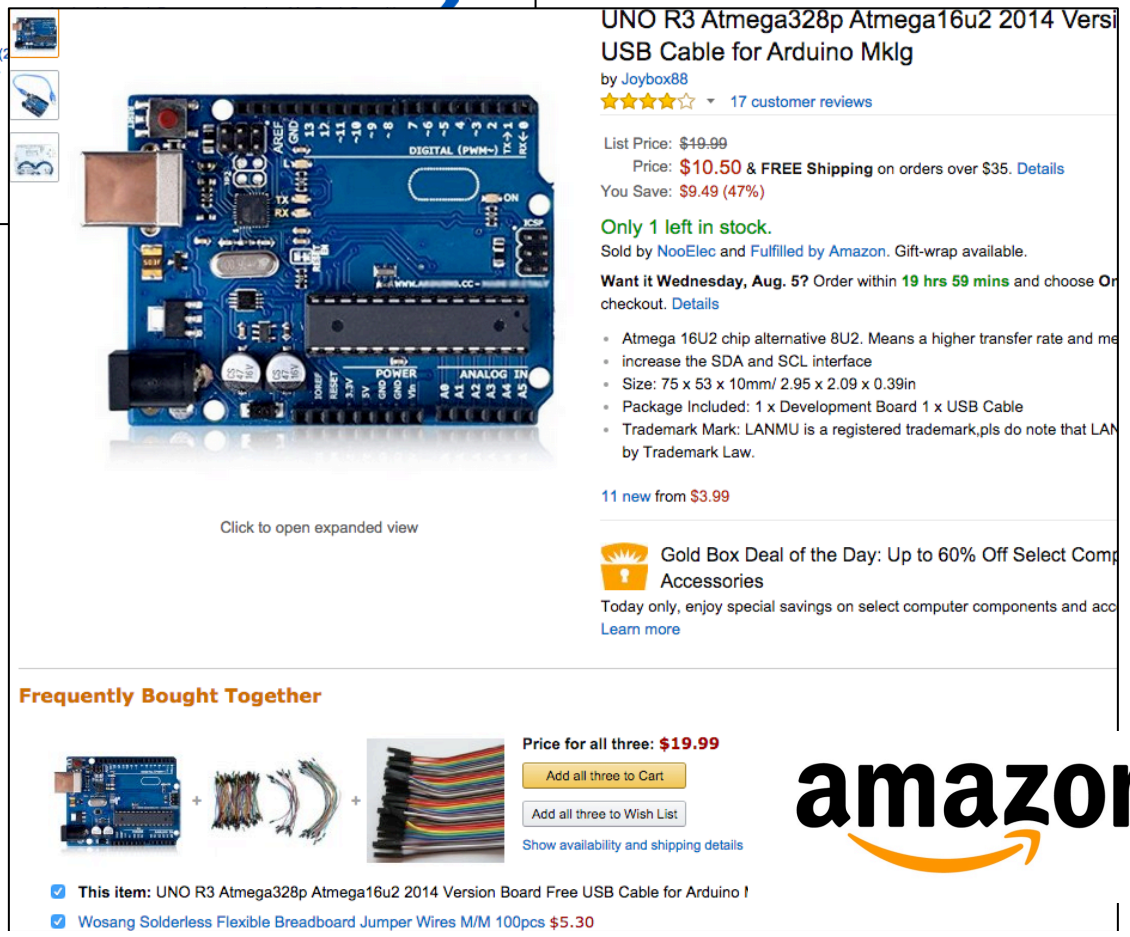
Outline

- Intro to Recommendations
- Demo (begin)
- Algorithm Deep-dive
  - Collaborative Filtering
  - Bayes' Theorem for Conditional Probability
- Demo (results)
- Questions?

Recommendations

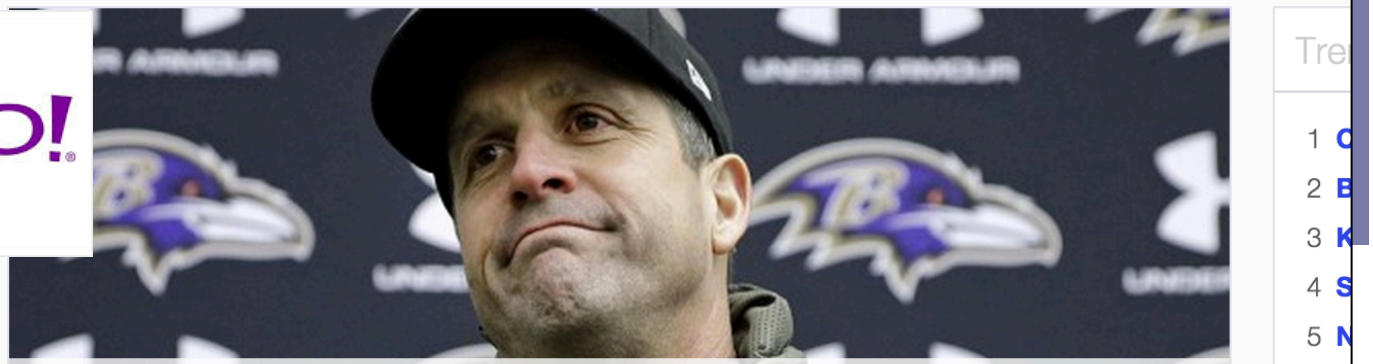


BEST BUY



amazon

known-user



Ravens deny role in deflate-gate allegations

Baltimore says it never tipped off the Colts to the Patriots' underinflated footballs in last season's playoffs. **'We knew nothing' »**

1 – 5 of 85


Tenn. theater shooting


Obama on Iran deal


MH370 debris confirmed


Ravens deny allegations


NFL star suffers serious injury

All Stories | News | Local | Entertainment | Sports | More

Tom Brady's reaction to being told he has to buy an $8,500 pool cover is priceless

A massive pile of Tom Brady's personal emails were made public as part of the NFLPA's...

Business Insider

Five Vikings facing a make-or-break season in 2015

The Vikings and Packers tied 26-26. The Minnesota Vikings are a transforming team under second-year head coach Mike Zimmer. Whether it's the result of an upcoming salary-cap hit, impending free agency or waning

Fox News

Browns running backs dropping with injuries

YAHOO!

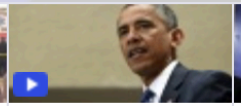
unknown-user

Reunion Island debris confirmed to be from MH370

Malaysia says the fragment discovered more than a year after the jet's disappearance is from the doomed flight. **'Physical evidence' »**

1 – 5 of 85


Tenn. theater shooting


Obama on Iran deal


MH370 debris confirmed

Obamacare subsidy issue

How J. Crew lost its way

All Stories | News | Local | Entertainment | Sports | More

6 Things to Know About Christine Ouzounian, the Nanny at the Center of the Ben Affleck-Jennifer Garner Divorce

Though Ouzounian maintains no public social media presence, there are a few things we do know for sure about the woman who's alleged to have been

Yahoo Celebrity 32 mins ago

Credit Cards Are Now Offering 0% APR Through 2017

Credit card companies are now offering 0% APR for 21 months for all purchases and balance transfers or 40k bonus miles ($400 bonus) and double miles.

Demo

Will use participants in this room to create a "crowd sourced" Recommender

Step 1: go to www.joshjanzen.com/demo

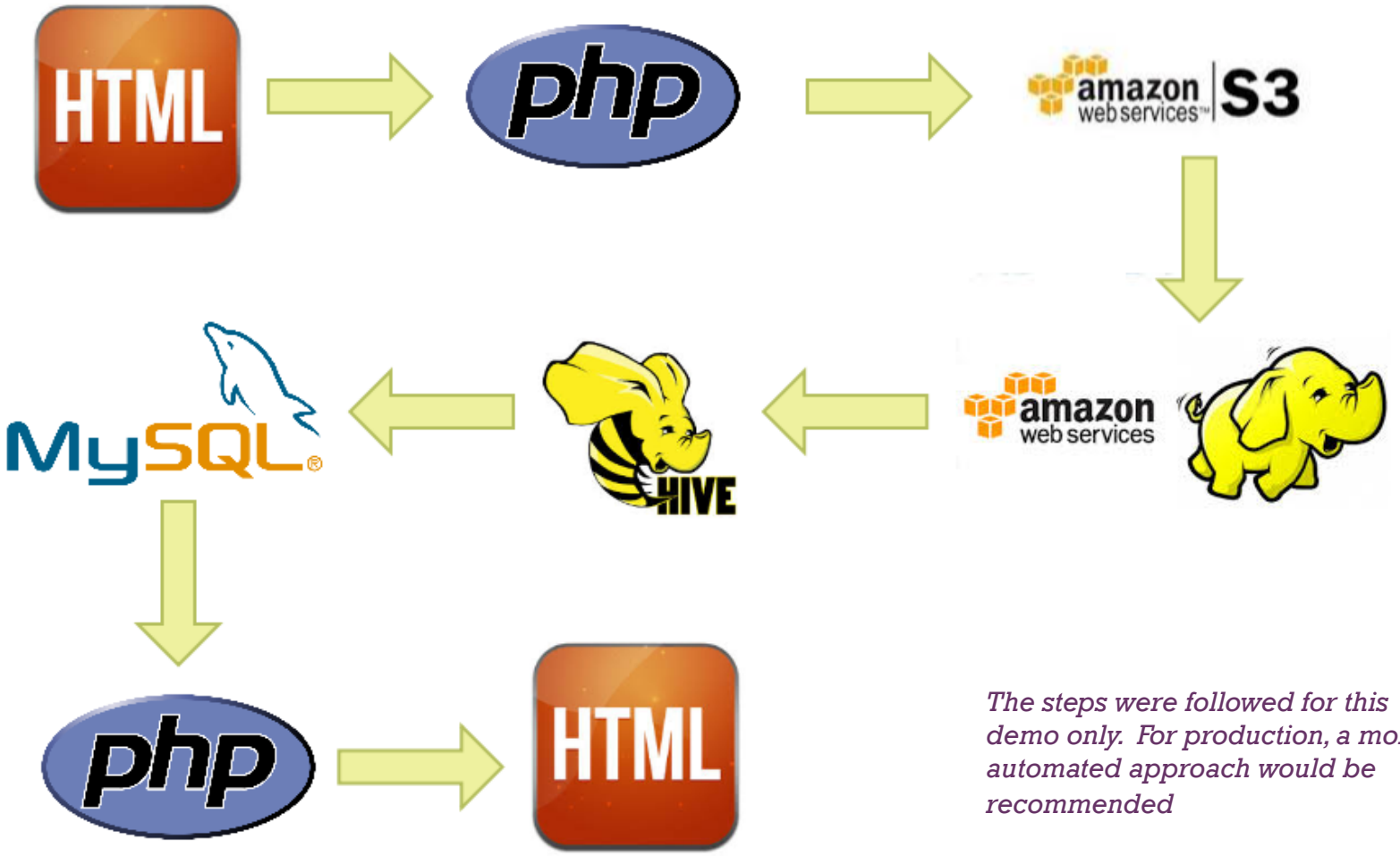
Step 2: choose activities in MN you ENJOY doing

Step 3: enter your email address

Step 4: click 'Submit'

Let's process the data and make some recommendations!

Demo (how data is processed)



The steps were followed for this demo only. For production, a more automated approach would be recommended

Collaborative Filtering (aka crowd-sourcing)

- Crowd based recommender algorithm which aggregates users preferences
- "Customers Who Viewed This Item Also Viewed"
- **Pros**
  - simple to write in SQL
  - deep data-sets
  - favors frequent items, such as "top-selling" or "most-viewed"
- **Cons**
  - difficulty in finding highly unique item-sets
  - significant processing power
  - "cold-start" problem

How Collaborative Filtering Works

All User Activity History

Name	Art Fair	Fishing	Shovel Snow	Wedding
Jon	No	Yes	Yes	No
Jane	Yes	No	No	Yes
Jill	Yes	No	Yes	Yes

How Collaborative Filtering Works

All User Activity History

Name	Art Fair	Fishing	Shovel Snow	Wedding
Jon	No	Yes	Yes	No
Jane	Yes	No	No	Yes
Jill	Yes	No	Yes	Yes

Scenario: a new user, Homer likes **Weddings**, and we need to recommend him other activities he may enjoy.

- Find Wedding in User history above and identify users who also enjoyed Weddings: U{Jane, Jill}
- Identify other activities U{Jane, Jill} enjoyed, and rank by count

Recommendation Output

Activity	Rank	Count of User (co-occurrence)
Art Fair	1	2
Shovel Snow	2	1
Fishing	3	0

Bayes' Theorem for Conditional Probability

- Uses probabilities to classify and rank recommendations
- "Frequently Bought Together"
- **Pros**
  - probability based, easy to understand and tweak
  - depth of data-set can be easily adjusted
- **Cons**
  - more complex SQL statements and processing power

How Bayes' Theorem Works

All User Activity History

Name	Art Fair	Fishing	Shovel Snow	Wedding
Jon	No	Yes	Yes	No
Jane	Yes	No	No	Yes
Jill	Yes	No	Yes	Yes

How Bayes' Theorem Works

All User Activity History

Name	Art Fair	Fishing	Shovel Snow	Wedding
Jon	No	Yes	Yes	No
Jane	Yes	No	No	Yes
Jill	Yes	No	Yes	Yes

Scenario: a new user, Homer likes **Weddings**, and we need to recommend him other activities he may enjoy. We need to calculate probability of an activity occurring with Weddings.

- **Art Fair (AF) & Wedding (W)**
- **Fishing (F) & Wedding (W)**
- **Shovel Snow (SS) & Wedding (W)**

Bayes' Theorem

Name	Art Fair	Fishing	Shovel Snow	Wedding
Jon	No	Yes	Yes	No
Jane	Yes	No	No	Yes
Jill	Yes	No	Yes	Yes

P(Art Fair | Wedding)

- P(AF) = 2/3 = 67%
- P(W) = 2/3 = 67%
- P(W|AF) = 2/2 = 100%
- Posterior Probability = (67% * 100%) / 67% = **100%**

P(Fishing | Wedding)

- P(F) = 1/3 = 33%
- P(W) = 2/3 = 67%
- P(W|F) = 0/1 = 0%
- Posterior Probability = (33% * 0%) / 67% = **0%**

P(Shovel Snow | Wedding)

- P(SS) = 2/3 = 67%
- P(W) = 2/3 = 67%
- P(W|SS) = 1/2 = 50%
- Posterior Probability = (67% * 50%) / 67% = **50%**

Wedding Recommendation (CF vs. BT)

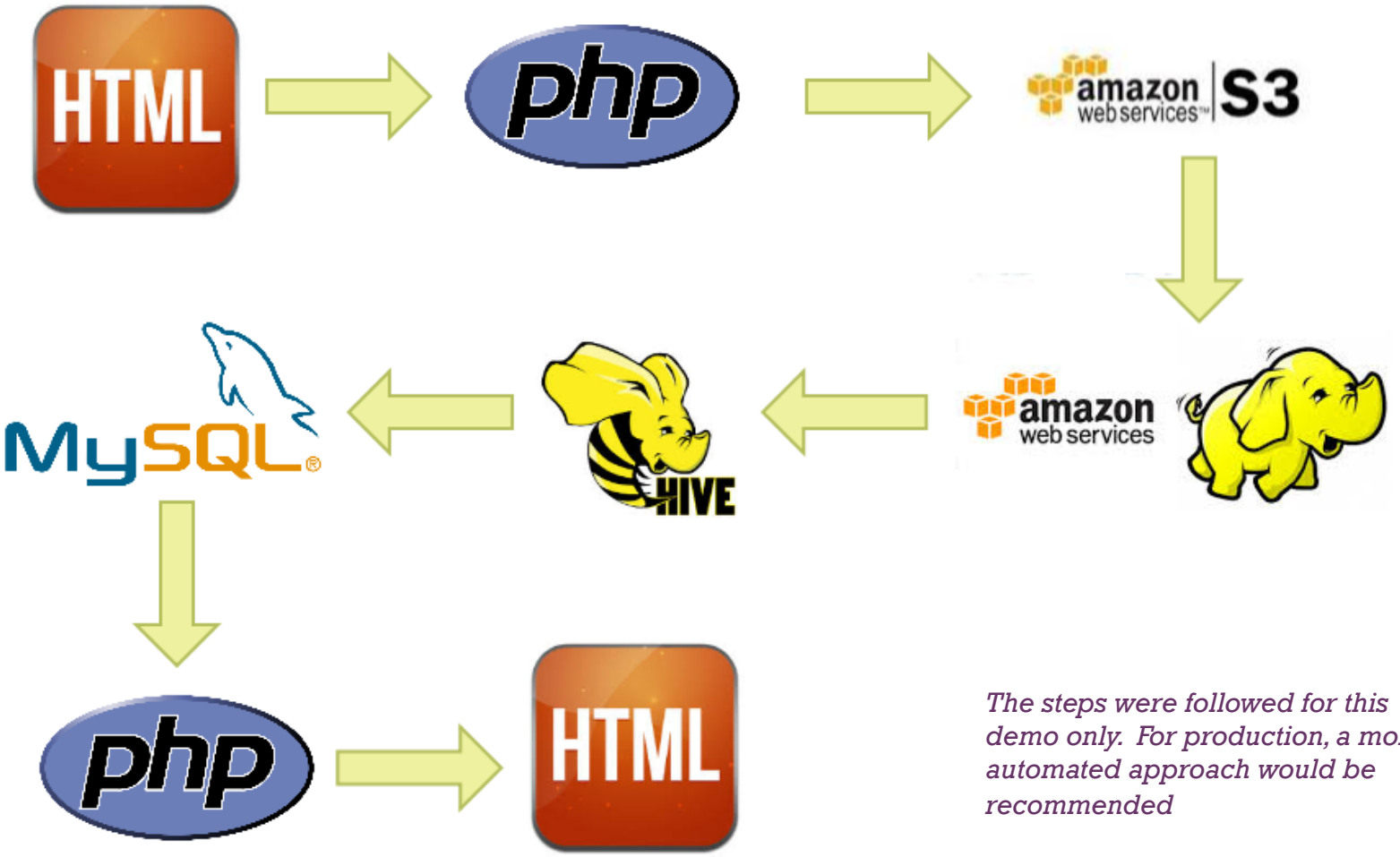
Recommendation Output – Collaborative Filtering

Activity	Rank	Count of User (co-occurrence)
Art Fair	1	2
Shovel Snow	2	1
Fishing	3	0

Recommendation Output – Bayes' Theorem

Activity	Rank	Conditional Probability
Art Fair	1	100%
Shovel Snow	2	50%
Fishing	3	0%

Demo (finish up processing)



The steps were followed for this demo only. For production, a more automated approach would be recommended

Demo Results

All of your (the "crowd") submissions have been calculated!

Step 1: go to http://www.joshjanzen.com/demo-results

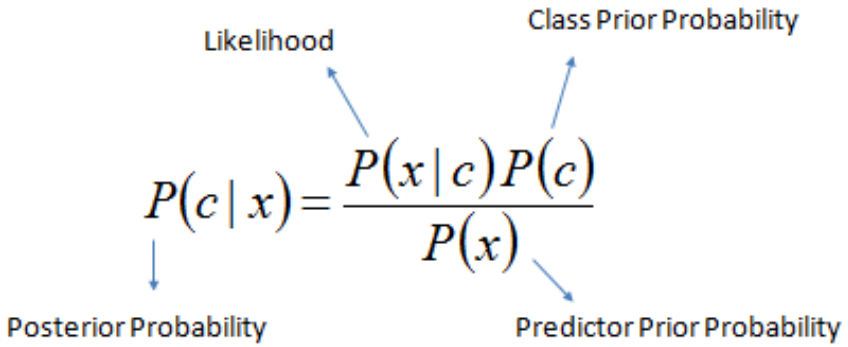
Step 2: select any activity you enjoy

Step 3: click 'Submit'

Step 4: the top 10 recommended activities are listed

Questions?

Appendix: Bayes' Theorem



$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

- $P(c \mid x)$ is the posterior probability of *class* (*target*) given *predictor* (*attribute*).
- $P(c)$ is the prior probability of *class*.
- $P(x \mid c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*.

Source: http://www.saedsayad.com/naive_bayesian.htm

Appendix: Demo (how data is processed)

- **Step 1:** User submits data through a website form. Using PHP, this is saved as a CSV in AWS S3
- **Step 2:** CSVs on S3 are processed with Hive on AWS EMR Hadoop cluster (any database will work). This is where a series of HQL (SQL) statements in Hive are ran to prepare the data for output. An output CSV is put into S3
- **Step 3:** CSV output is moved from S3 to MySQL
- **Step 4:** the website connects to MySQL via PHP to display recommendations on website

The steps were followed for this demo only. For production, a more automated approach would be recommended