Designing an ML-Minded Product and a Product-Minded ML System

ACM Webinar
January 23, 2019

Grace Huang
Personalized Homefeed
Personalized Homefeed

Personalization: Scoring and ranking
Picking the best of the best among candidates
• **Supervised learning** with labels:
  - $1 =$ some positive engagements
  - $0 =$ no engagement or negative actions

• Learns to predict a positive engagement (ranking) score

• Pins are then sorted by engagement score $= f(pin, user, ...)$
Data collection → Feature engineering → Model training → Prediction (Store and serve)
Components of A production ML system

Data
- Data pipeline (training/test)
- Data (to make predictions)
- Data (for training/offline testing)

Training
- Model
- Predictions

Serving
- Predictions

Evaluation
- On-line experiments
- Offline evaluations

Launch
- Launch (production model)
We will focus on data, evaluation and shipping.
Considerations for a data pipeline

Data pipeline (training/test)

Model

Predictions

On-line experiments

Offline evaluations

Launch (production model)

Data (to make predictions)

Data (for training/offline testing)

Data (for training/offline testing)

Predictions

Serving

Evaluation

Shipping
Engagement score = \( f(pin, user...) \)

Pin

The perfect path to cold brew

Caffeinated Inc.

Omar Seyal
Cravings

User profile

User’s past actions: engagement signals

Derived user profiles from past actions

Derived pin information
Considerations for a data pipeline

- Logging (and changes)
- Aggregations (ETLs)
- ETL management libraries
- Data validation
- Monitoring and alerts for the pipeline
Training data should be carefully managed

- Sampling scheme
- Version control
- Monitoring feature distribution changes
- Feature extraction and transformations
- Feature value validation
- Shared feature store or individual pipelines
Training and serving data discrepancy (skew)?

- Training data sampled differently from serving data?
- There is a lag to certain features being populated? (e.g. takes a long time to compute)
- Logging change?
- ETL breaks?
- Seasonality
- Market differences
How to evaluate a candidate model

- Your favorite offline performance measures
- Human evaluation
- Custom tools (e.g. side by side, simulated debuggers for sanity check, funnels..etc)
How to evaluate a candidate model

- Goal metrics
- Leading indicators
- Debug metrics
- Guardrail metrics
- Custom tools
- Metrics vs. loss function
Shipping criteria should include:

- Metrics
- Infrastructure cost
- Maintenance overhead (regularization!)
- Product vision
- Cannibalization
- Speed vs. iteration
Once shipped, continue to monitor

- Continuous monitoring:
  - Goal metrics on dashboards
  - Alerts for data and prediction distribution drifts
  - Runbook, tools and delegation for investigations

Data pipeline (training/test)

Data (to make predictions)

Data (for training/offline testing)

Training

Predictions

Serving

Evaluation

On-line experiments

Launch

Launch (production model)
Automation is key
Lessons learned

#1 Beware of Data and System Bias
#2 Testing & Monitoring .....(Do it!)
#3 Good Infrastructure Speeds Up Iteration
#4 Measurement and Understanding are Crucial
#5 Build a Sustainable Ecosystem
#6 Design a ML Minded Product, and a Product Minded ML System
#1 Beware of Data and System Bias
Engagement data complements pin information
Engagement data is a double-sided sword!
Remove bias and effects of the existing system as much as possible (so rich doesn’t get richer)

**Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market**

Matthew J. Salganik, Peter Sheridan Dodds, Duncan J. Watts

Hitting the market with new songs and movies is often a daunting task, where success is largely determined by numerous factors. While some songs and movies become hits, others struggle to find an audience. This variability can be explained by the existing system, which has biases and effects that propagate and influence the outcomes.

### How can success in cultural markets be at once strikingly distinct from average performance (1–4), and yet so hard to anticipate for profit-minded experts armed with extensive market research (4–5)?

One explanation (9) for the observed inequality of outcomes is that the mapping from “quality” to success is convex (i.e., differences in quality correspond to larger differences in success, leading to what has been called the “superstar” effect (9), or “winner-take-all” markets (10). However, models of this type become incommensurately assigned to conditions—indeed distinguished only by formation on the pre-existing independent conditions about which only the names of it. While listening to a song, the opportunity (but not the stars) often come.

### Abstract

In this paper, we seek to understand the relationship between online popularity of a song and its intrinsic quality. Prior experimental work suggests that the relationship between popularity and quality can be very distorted due to factors like social influence bias and inequality in visibility. We conduct a study of popularity on two different social news aggregators, Reddit and Hacker News. We define quality as the relative number of votes an article would have received if each article was shown in a bias-free way, to an equal number of users. The results of this study show that the existing system, which is biased and unpredictable, can lead to unequal distributions of popularity.

### Social Influence Bias: A Randomized Experiment

**Lev Muchnik, Sten Aron, Sean T. Taylor**

Our society is increasingly relying on the digitized, aggregated opinions of others to make decisions. We therefore designed and analyzed a large-scale randomized experiment on a social news aggregation Web site to investigate whether knowledge of such aggregate interests distorts decision-making. Prior ratings created significant bias in individual rating behavior, and positive and negative social influences created asymmetric herding effects. Whereas negative social influence inspired users to correct manipulated ratings, positive social influence increased the likelihood of positive ratings by 32% and created accumulating positive herding that increased final ratings by 25% on average. This positive herding was topic and affected by whether individuals were viewing the opinions of friends or enemies. A mixture of changing opinion and greater turnover under both manipulations together with a natural tendency for up-votes on the site combined to create the herding effects. Such findings help inform collective judgment accuracy and avoid social media on true quality is rarely observable in isolation. To achieve this goal, we developed a large-scale randomized experiment to quantify the effect of social influence on users’ ratings and discursions on a social news aggregation Web site, where users contribute news articles and discuss them. Users of the site that were not aware of the experiment were exposed to a random selection of news articles, and their ratings were recorded. We found that the existing system, which is biased and unpredictable, can lead to unequal distributions of popularity.
#2 Testing & Monitoring
.....(Do it!)
Some important metric

Not good!!!
Some important metric

Weeks…….

GBDT Migration to Neural Network

Not good!!!
Offline data distribution != Online data distribution
Data coverage drop or corruption -> Silent failures
#3 Good Infrastructure Speeds Up Iteration
Can multiple engineers work on the system simultaneously?

• Are there automated training/deploy pipelines? Can they ship multiple experiments at once?

• Are there effective offline analysis tools to help reduce amount of live experiments needed?
#4 Measurement and Understanding are Crucial
**Offline performance != Online performance**

- Final bar is running on live traffic
- Run experiments to learn

<table>
<thead>
<tr>
<th>Homemedia Events</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homemedia Users</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>21</td>
<td>22</td>
</tr>
<tr>
<td>Homemedia Repinners</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>24</td>
<td>25</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>Homemedia Closeupers</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>24</td>
<td>25</td>
<td>26</td>
<td>27</td>
<td>28</td>
<td>29</td>
<td>30</td>
<td>31</td>
<td>32</td>
<td>33</td>
<td>34</td>
<td>35</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>Homemedia Clickthroughers</td>
<td>36</td>
<td>37</td>
<td>38</td>
<td>39</td>
<td>40</td>
<td>41</td>
<td>42</td>
<td>43</td>
<td>44</td>
<td>45</td>
<td>46</td>
<td>47</td>
<td>48</td>
<td>49</td>
<td>50</td>
<td>51</td>
<td>52</td>
<td>53</td>
<td>54</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>Homemedia R/Cers</td>
<td>46</td>
<td>47</td>
<td>48</td>
<td>49</td>
<td>50</td>
<td>51</td>
<td>52</td>
<td>53</td>
<td>54</td>
<td>55</td>
<td>56</td>
<td>57</td>
<td>58</td>
<td>59</td>
<td>60</td>
<td>61</td>
<td>62</td>
<td>63</td>
<td>64</td>
<td>65</td>
<td></td>
</tr>
</tbody>
</table>

Users with no daily flow

Users with less than 25 new pins

Users with pins in flow

Following Feed Users

Following Feed Repinners

Following Feed Closeupers

Following Feed Clickthroughers

Following Feed R/Cers

Users with following feed pins in flow
Invest in toolings and experiments to understand the blackbox

- Ablation experiments
- Are sub-populations of users disproportionally impacted
- Analyses and tools to help us understand long term, ecosystem effect
It’s easy to get what you wish for, but not what you want....... 

"You have just one more wish. Are you sure you want another can of tuna?"

(Goodharts Law)
#5 Build a Sustainable Ecosystem
Do we handle cold starts elegantly?

Are we taking care of fresh, less impressed content?

<table>
<thead>
<tr>
<th>New Score</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
</tbody>
</table>

Lower Ranking Score

Higher Ranking Score

Fresher

Older
Do we handle cold starts elegantly?

Are we taking care of content with missing features (or features whose generation is delayed)?

Streaky, offensive content!
Build a system with tight negative feedback, and make use of (explicit) negative signals as much as possible

- **Model / Objective Function** - Change label / prediction target / model architecture so that negative events are tied to the objective function we optimize

- **Features** - Add more features that help in predicting negative events

But separate spam/racy filtering from negative signal incorporation in ML models
#6 Design a ML Minded Product, and a Product Minded ML System
Do you really need ML?

- 2013: Time order
- 2014-15: Linear Model
- 2016: Gradient Boosted Decision Trees
- 2017: Neural Net
For complex problems like diversity and freshness, ML components need to work in concert.

Beware of bottleneck!!
Important to have a way to build policy and product vision into the ML system
Independent surfaces for exploitation vs. exploration

Exploration →

Exploitation →
Build a system for users tomorrow (or users you really care about)

<table>
<thead>
<tr>
<th>Global engagement</th>
<th>Local engagement</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Homefeed Users</strong></td>
<td><strong>Homefeed Local Users</strong></td>
</tr>
<tr>
<td><img src="image" alt="1" /> Aggregate</td>
<td><img src="image" alt="1" /> Aggregate</td>
</tr>
<tr>
<td><strong>Homefeed Repinners</strong></td>
<td><strong>Homefeed Local Repinners</strong></td>
</tr>
<tr>
<td><img src="image" alt="7" /> Aggregate</td>
<td><img src="image" alt="12" /> Aggregate</td>
</tr>
<tr>
<td><strong>Picked For You Users</strong></td>
<td><strong>Picked For You Local Users</strong></td>
</tr>
<tr>
<td><img src="image" alt="1" /> Aggregate</td>
<td><img src="image" alt="9" /> Aggregate</td>
</tr>
<tr>
<td><strong>Picked For You Repinners</strong></td>
<td><strong>Picked For You Local Repinners</strong></td>
</tr>
<tr>
<td><img src="image" alt="12" /> Aggregate</td>
<td><img src="image" alt="21" /> Aggregate</td>
</tr>
</tbody>
</table>
Thank you