47th CREST Open Workshop - CREST 10th Anniversary

APP STORE ANALYSIS

Yue Jia, CREST, UCL

Anthony Finkelstein

Mark Harman

Yue Jia

Yuanyuan Zhang

Federica Sarro

William Martin

Afnan A. AlSubaihin

http://www0.cs.ucl.ac.uk/staff/F.Sarro/projects/UCLappA

CURRENT WORK AT CREST

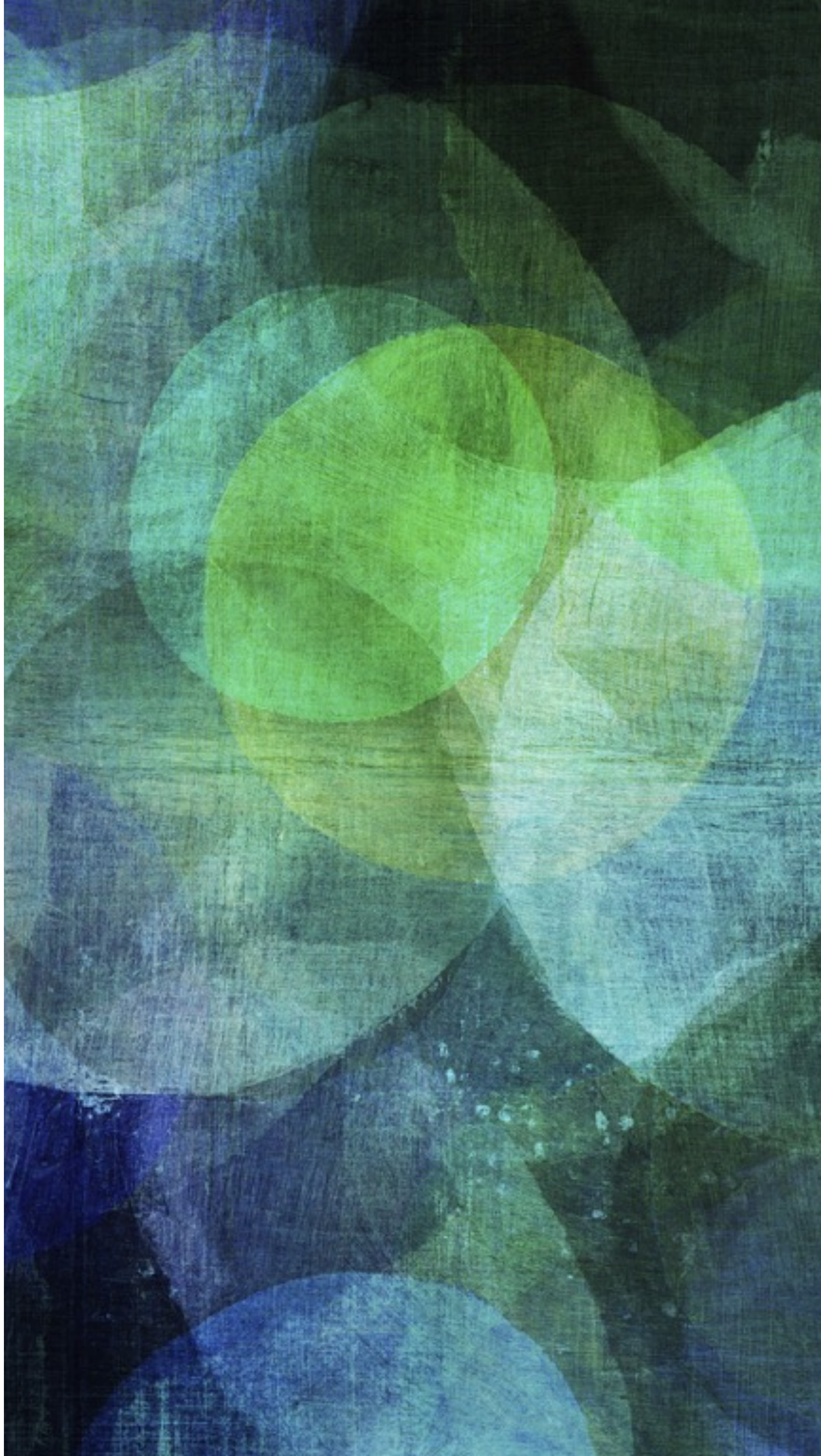
- Feature Analysis
- Clustering Mobile Apps
- Predicting Price and Rating
- Feature Migration
- Causal Impact Analysis
- Sampling Bias Issues
- App Developer Interviews and Survey
- Android Test Data Generation
- Mobile Energy Optimisation

CURRENT WORK AT CREST

- **Feature Analysis**
- **Clustering Mobile Apps**
- **Predicating Price and Rating**
- **Feature Migration**
- Causal Impact Analysis
- Sampling Bias Issues
- App Developer Interviews and Survey
- Android Test Data Generation
- Mobile Energy Optimisation

FEATURE ANALYSIS

App Store Mining and Analysis: MSR for App Stores (MSR'12)



APP STORE: THE TREMENDOUS SUCCESS

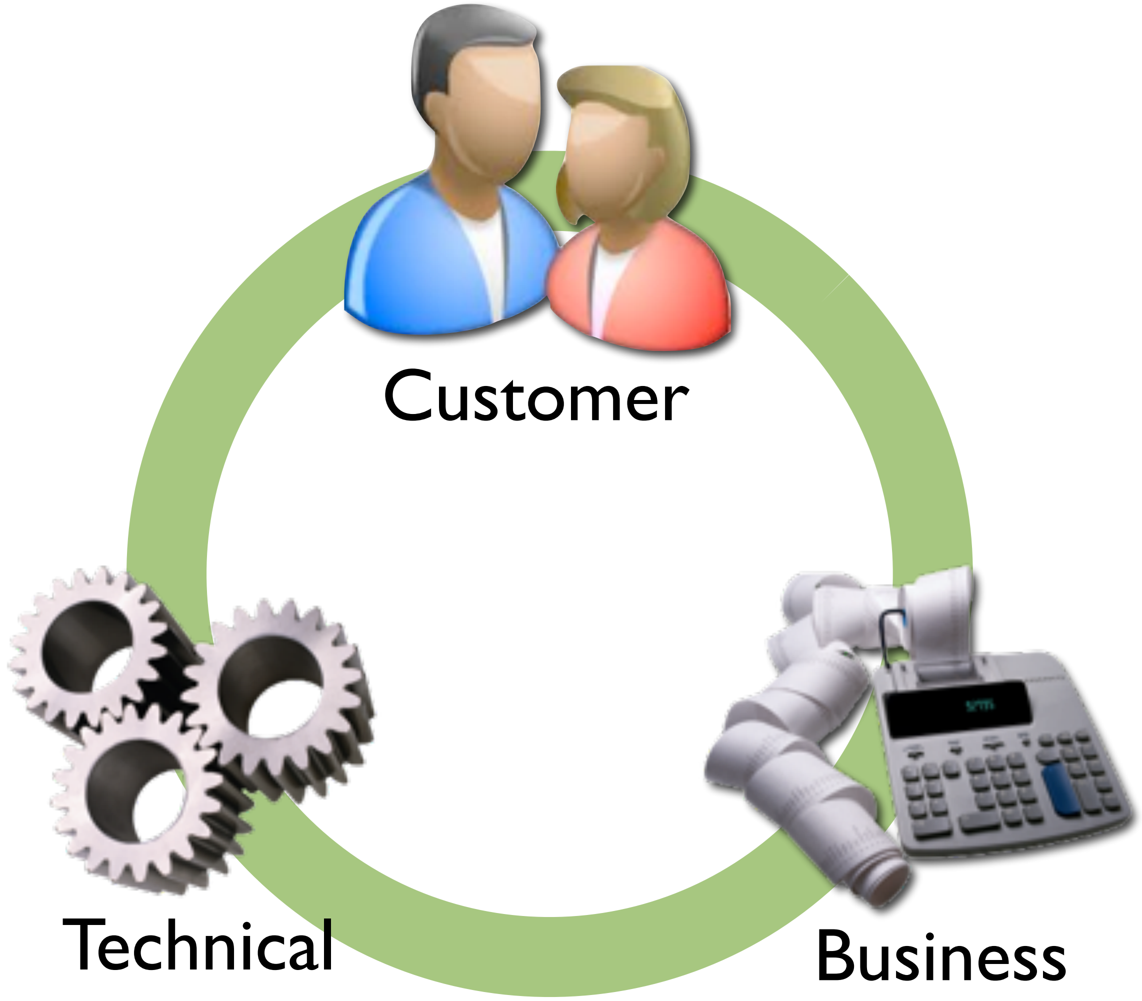
130 BILLIONS IOS DOWNLOADS

1.4 BILLIONS ANDROID DEVICES

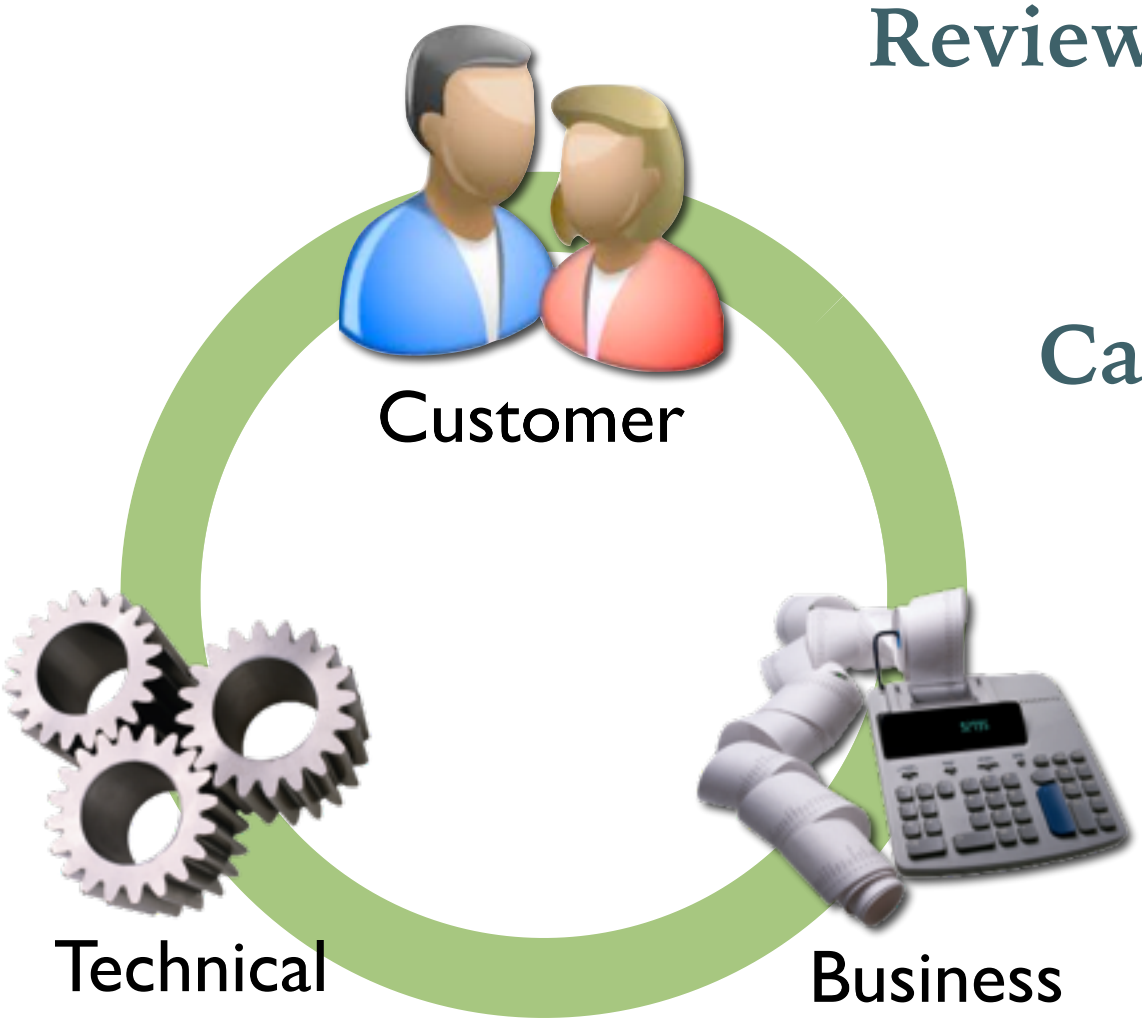
25 BILLIONS $ REVENUE

APP STORE: A NEW FORM OF SOFTWARE REPOSITORY

APP STORE: A NEW FORM OF SOFTWARE REPOSITORY



APP STORE: A NEW FORM OF SOFTWARE REPOSITORY



Description

Ratings

Size

Review

Authors

Discussions

Category

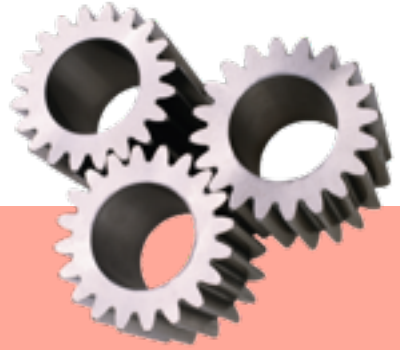
Releases

Issues

Price

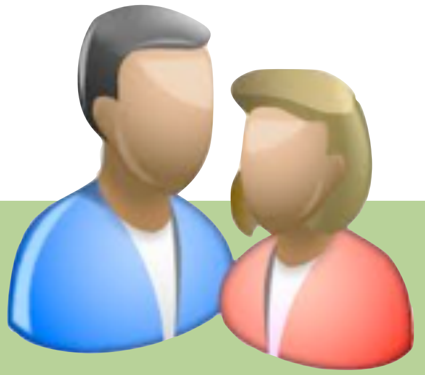
In-app purchases

Versions



Technical

Features



Customer

Ratings

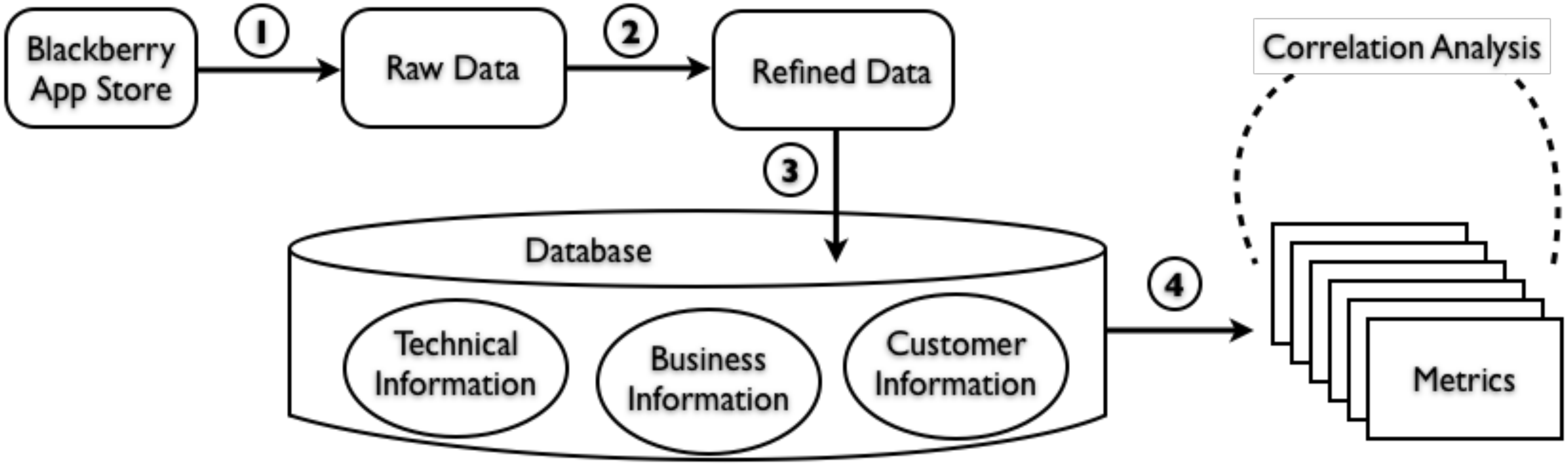
Popularity

Business

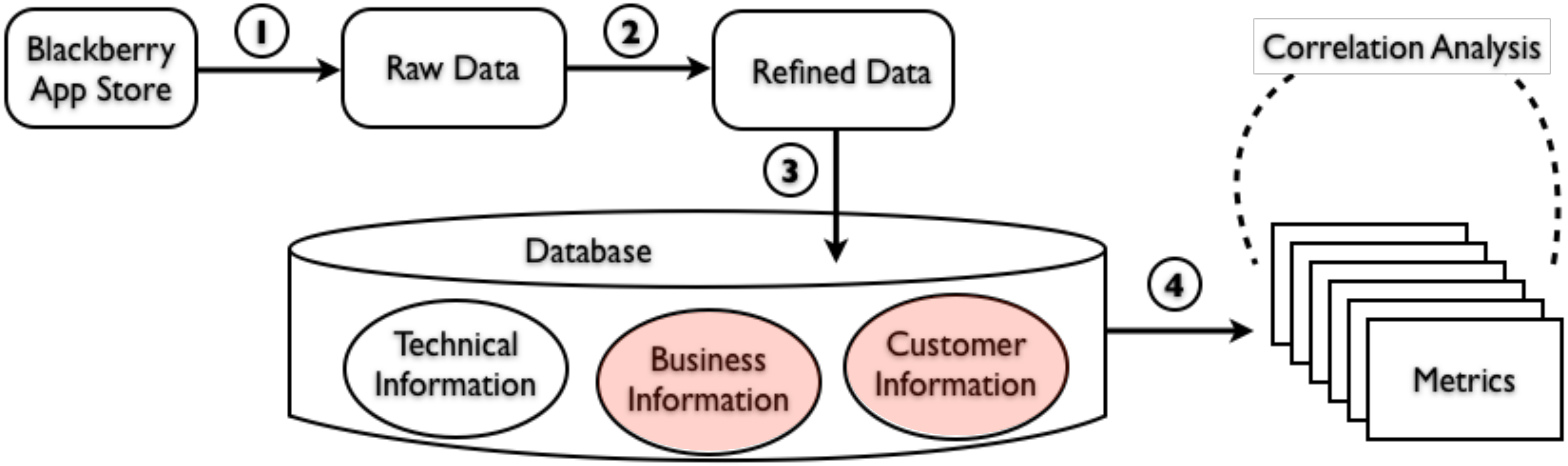
Price

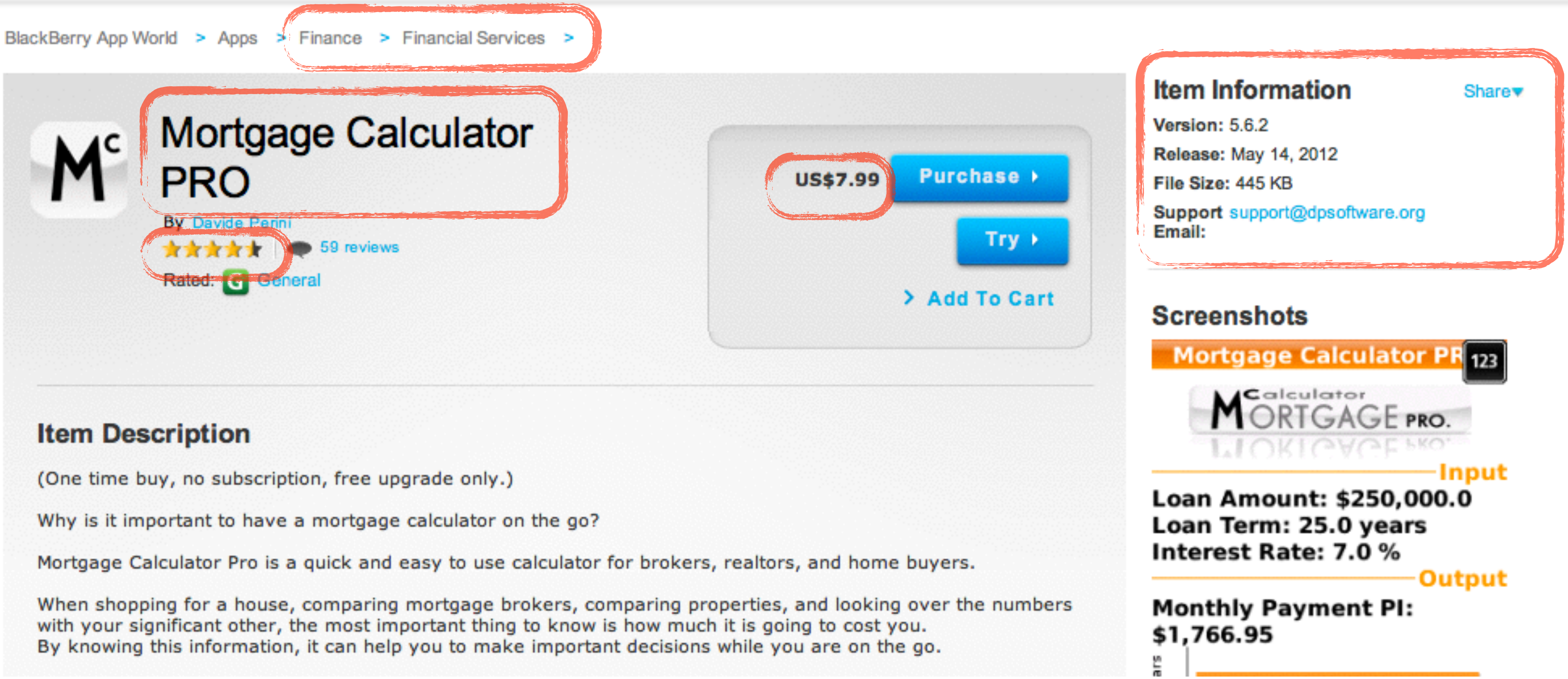
App Store Repository

Extracting features from description of apps



Mark Harman, Yue Jia, Yuanyuan Zhang: App store mining and analysis: MSR for app stores. MSR 2012: 108-111





BlackBerry App World > Apps > Finance > Financial Services >

Mortgage Calculator PRO

By Davide Perini

★★★★★ 59 reviews

Rated: G General

US$7.99 Purchase ›

Try ›

› Add To Cart

Item Information Share▾

Version: 5.6.2

Release: May 14, 2012

File Size: 445 KB

Support Email: support@dpsoftware.org

Screenshots

Mortgage Calculator PR 123

Calculator MORTGAGE PRO.

Input

Loan Amount: $250,000.0

Loan Term: 25.0 years

Interest Rate: 7.0 %

Output

Monthly Payment PI: $1,766.95

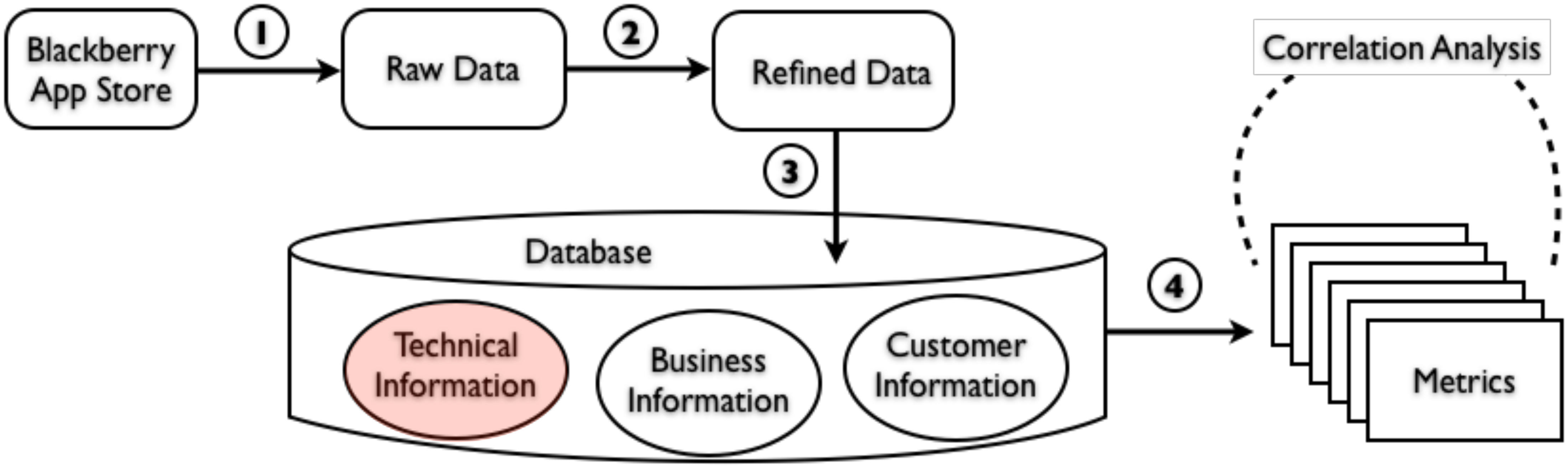
Item Description

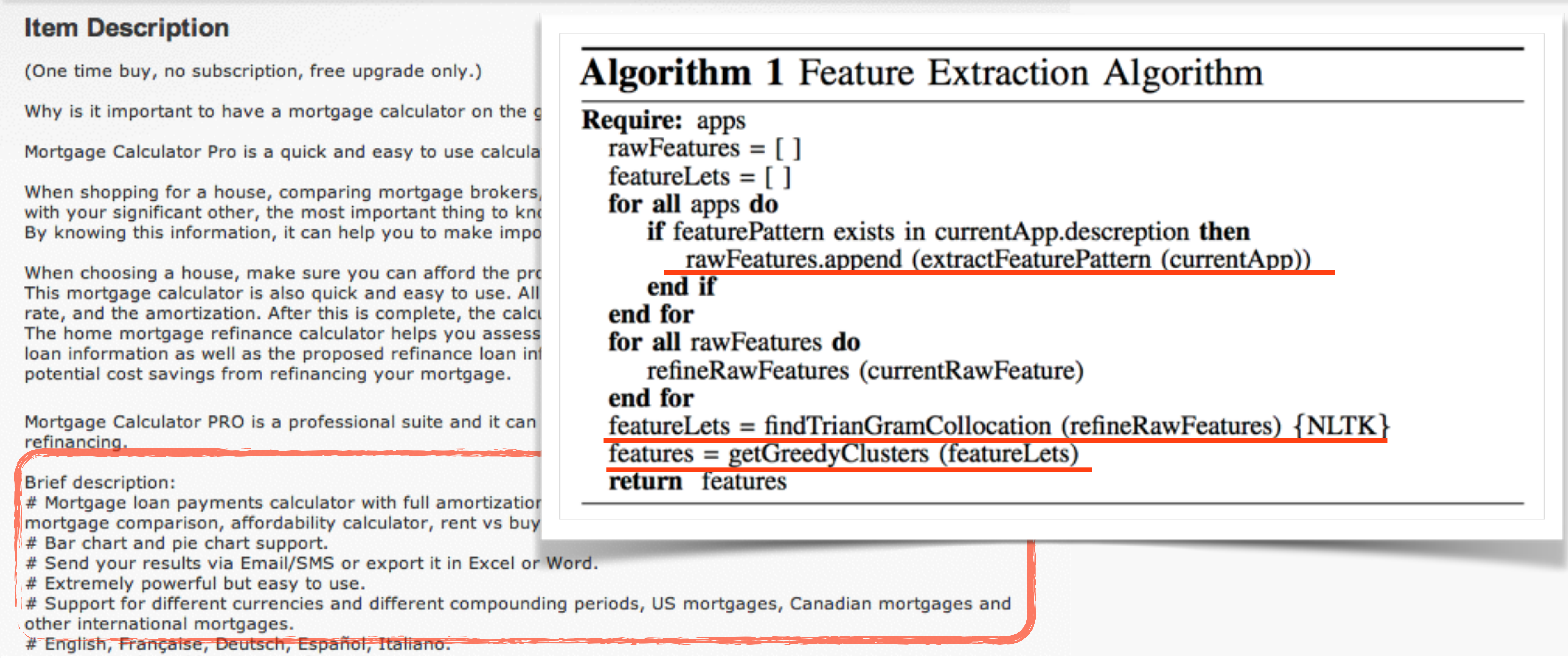
(One time buy, no subscription, free upgrade only.)

Why is it important to have a mortgage calculator on the go?

Mortgage Calculator Pro is a quick and easy to use calculator for brokers, realtors, and home buyers.

When shopping for a house, comparing mortgage brokers, comparing properties, and looking over the numbers with your significant other, the most important thing to know is how much it is going to cost you.
By knowing this information, it can help you to make important decisions while you are on the go.





Item Description

(One time buy, no subscription, free upgrade only.)

Why is it important to have a mortgage calculator on the g

Mortgage Calculator Pro is a quick and easy to use calcula

When shopping for a house, comparing mortgage brokers,
with your significant other, the most important thing to kno
By knowing this information, it can help you to make impo

When choosing a house, make sure you can afford the pro
This mortgage calculator is also quick and easy to use. All
rate, and the amortization. After this is complete, the calcu
The home mortgage refinance calculator helps you assess
loan information as well as the proposed refinance loan inf
potential cost savings from refinancing your mortgage.

Mortgage Calculator PRO is a professional suite and it can
refinancing.

Brief description:
Mortgage loan payments calculator with full amortizatio
mortgage comparison, affordability calculator, rent vs buy
Bar chart and pie chart support.
Send your results via Email/SMS or export it in Excel or Word.
Extremely powerful but easy to use.
Support for different currencies and different compounding periods, US mortgages, Canadian mortgages and
other international mortgages.
English, Française, Deutsch, Español, Italiano.

Algorithm 1 Feature Extraction Algorithm

```
Require: apps
  rawFeatures = [ ]
  featureLets = [ ]
  for all apps do
      if featurePattern exists in currentApp.descreption then
          rawFeatures.append (extractFeaturePattern (currentApp))
      end if
  end for
  for all rawFeatures do
      refineRawFeatures (currentRawFeature)
  end for
  featureLets = findTrianGramCollocation (refineRawFeatures) {NLTK}
  features = getGreedyClusters (featureLets)
  return  features
```

Extracting features from description of apps

A **feature** to be a property, captured by a set of words in the app description and shared by a set of apps.

e.g. Finance

- setup, bank, accounts
- calculate, monthly, expenses
- e-mail, alerts, stock
- create, watch, lists
- financial, business, news

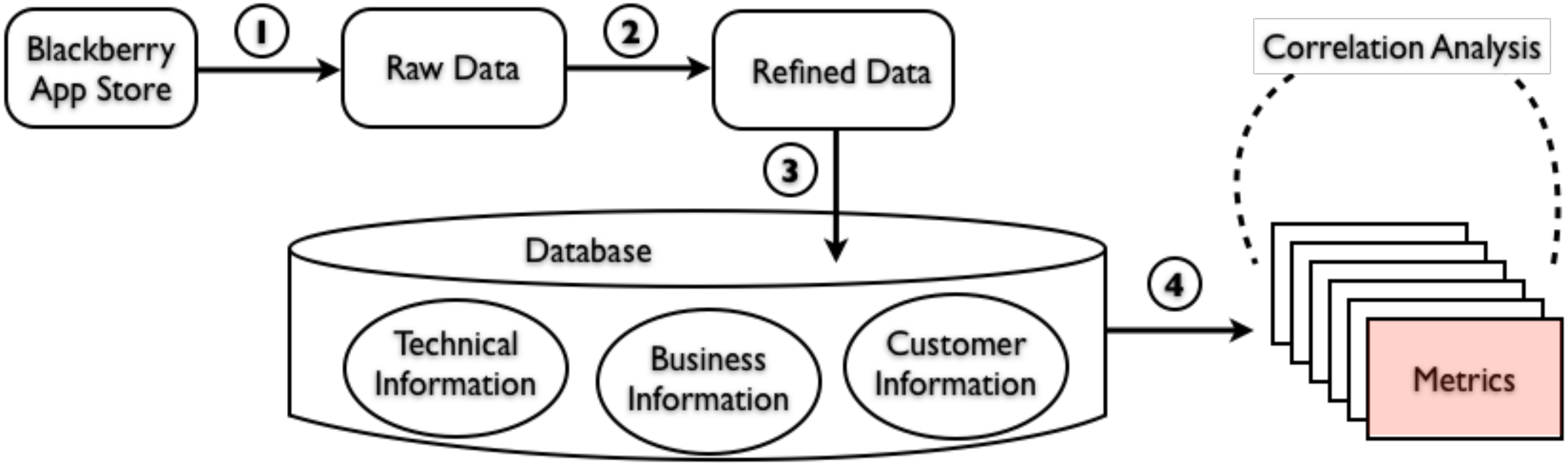
e.g. Travel

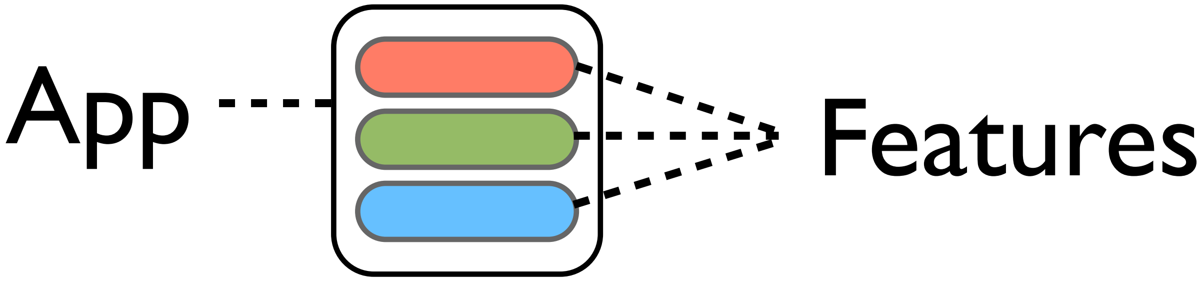
- free, wifi
- wifi, hotspot, near
- download, offline, use
- restaurants, plotted, map
- bus, service

Feature Attributes

Features have price, rating and popularity

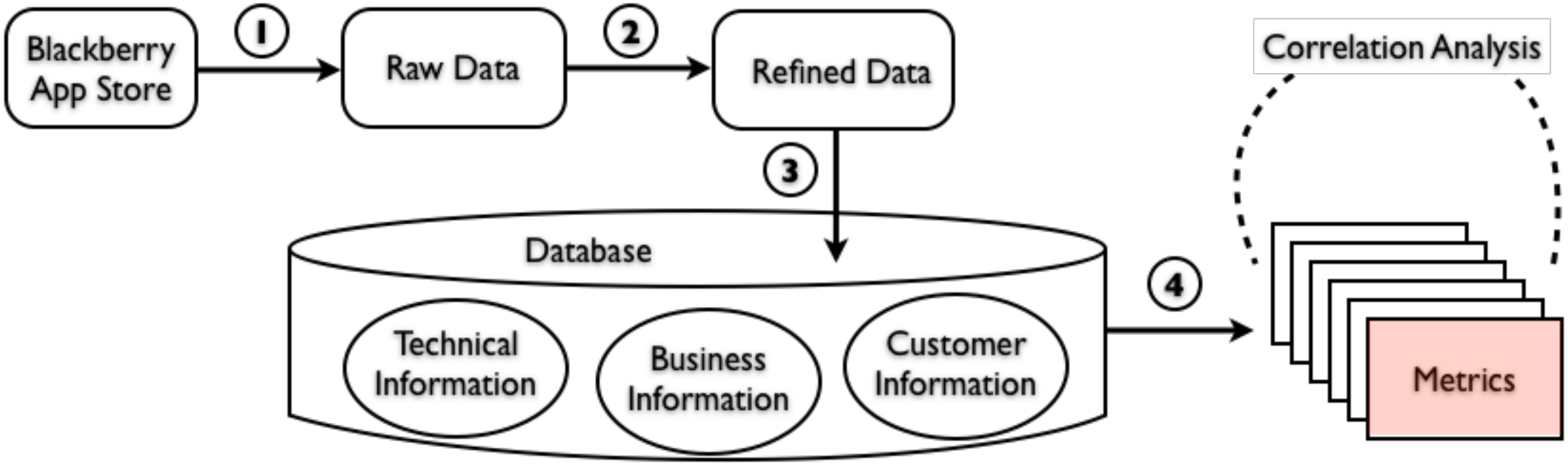
- by extension (aggregated over apps)





App

Features



Blackberry App Store

1

Raw Data

2

Refined Data

3

Database

Technical Information

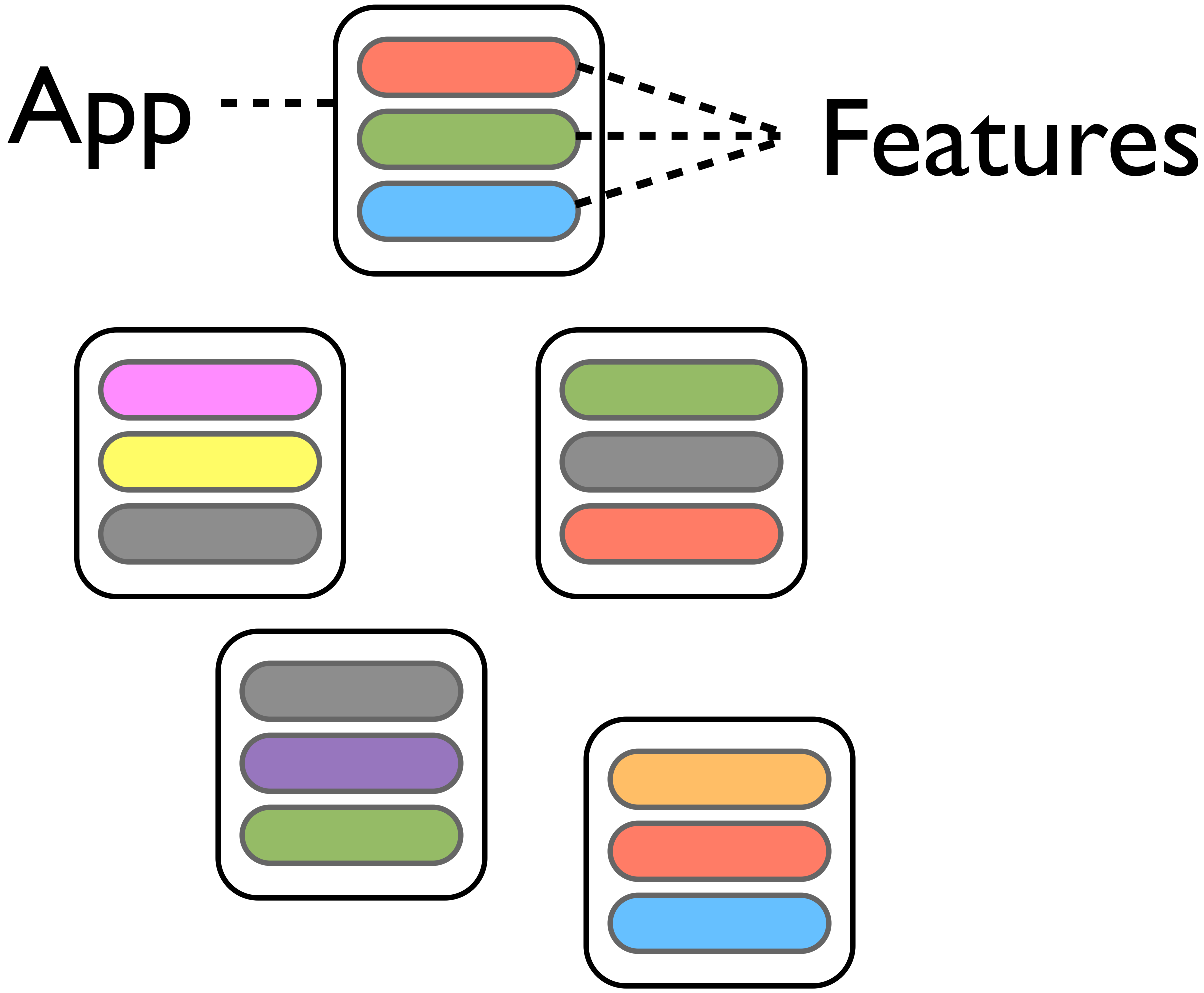
Business Information

Customer Information

4

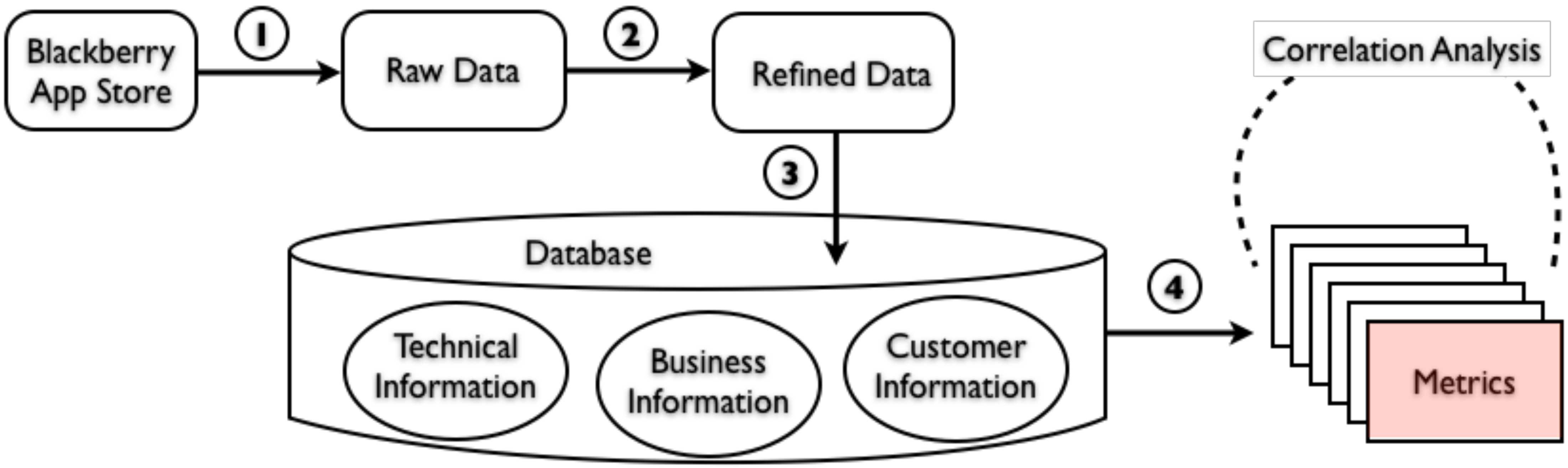
Correlation Analysis

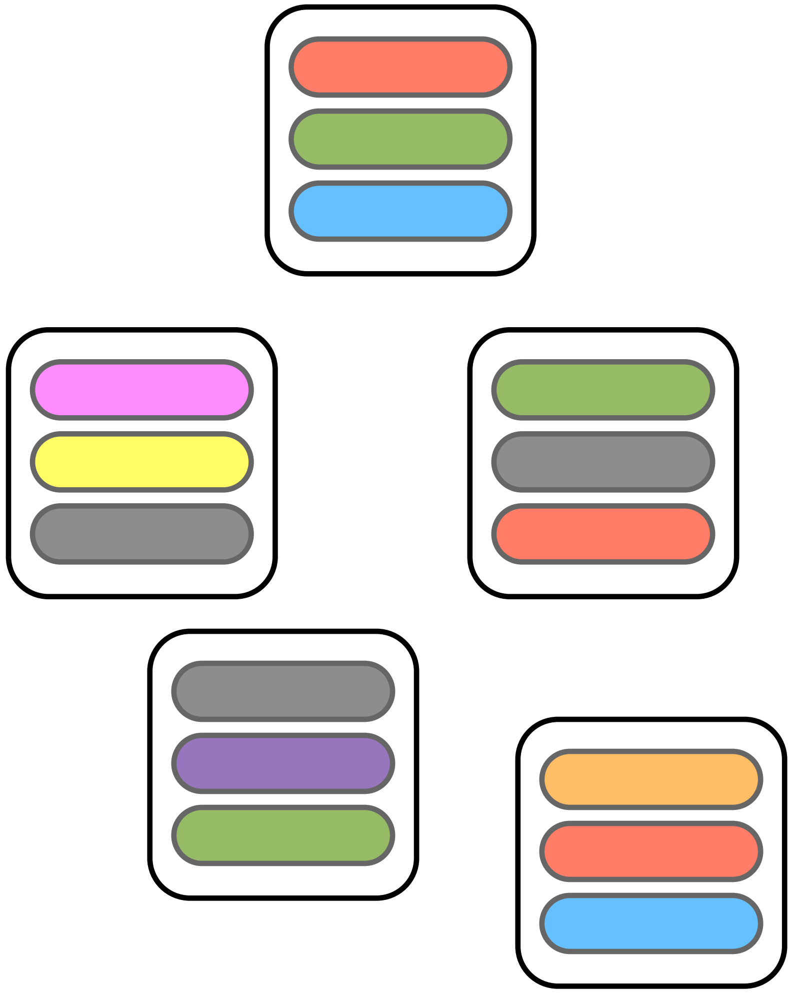
Metrics



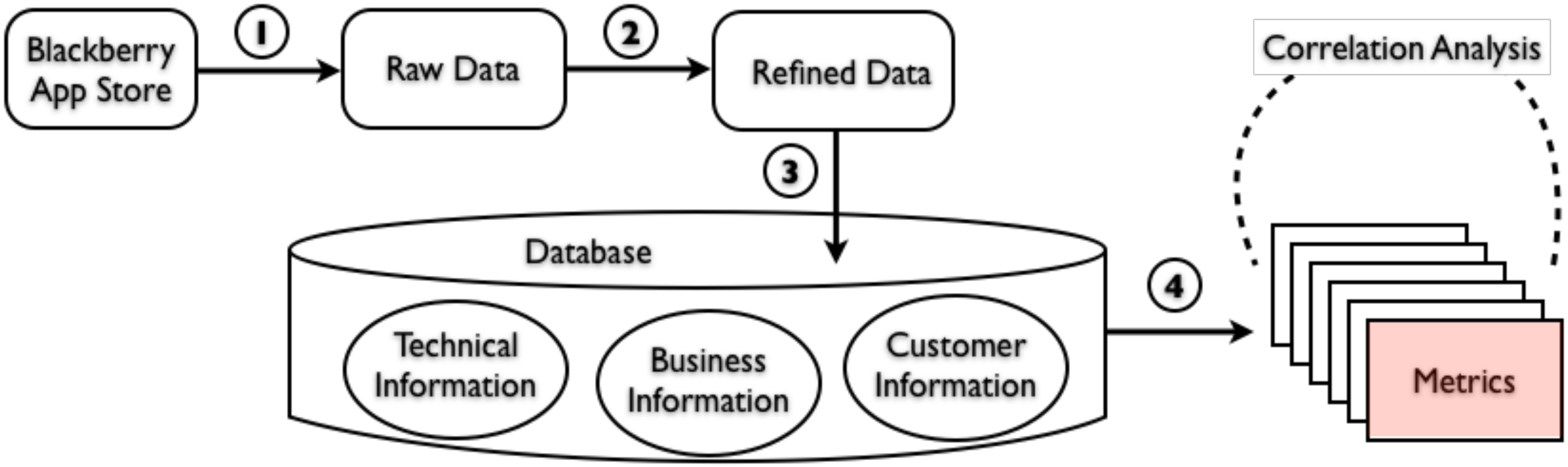
App

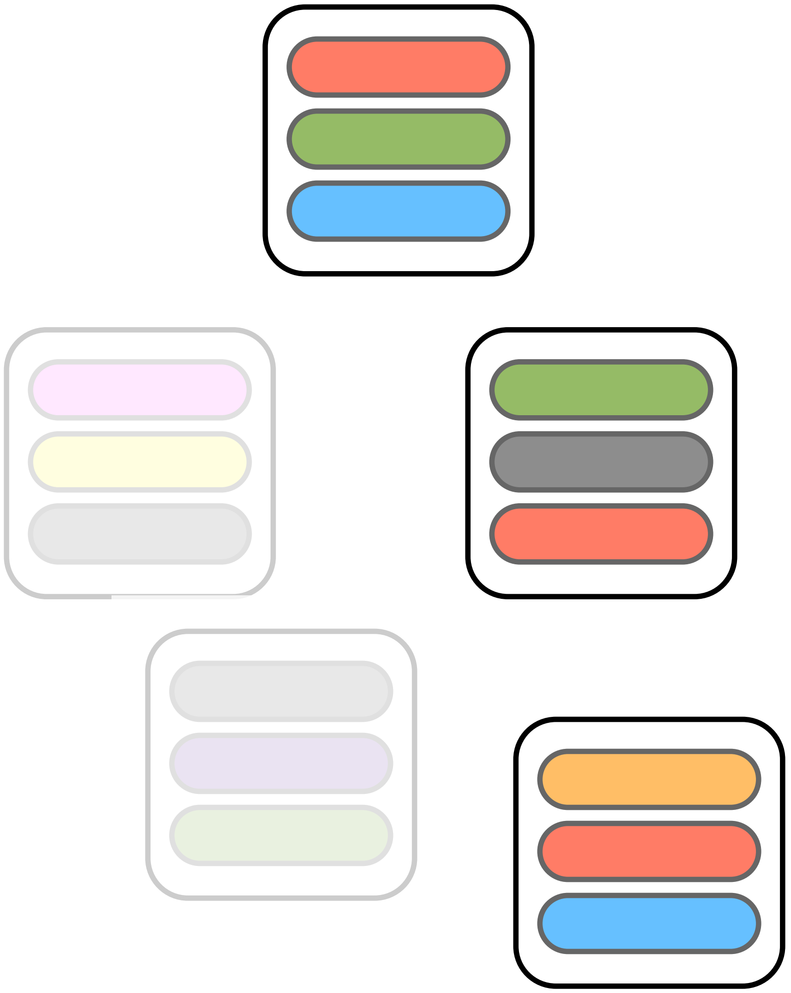
Features





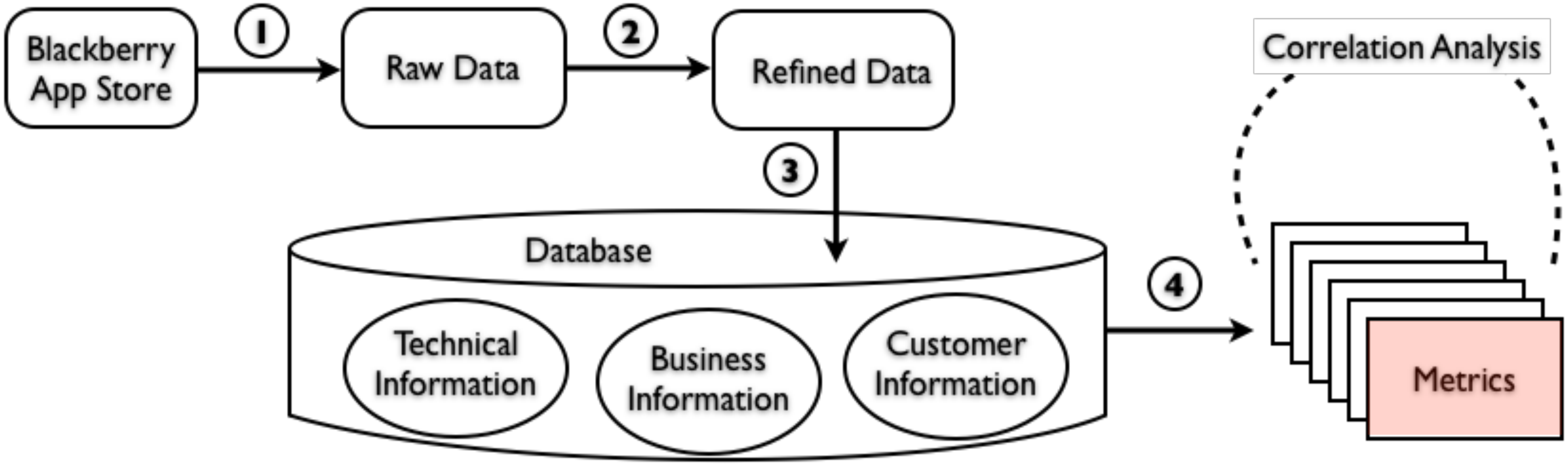
E.g cost for features 🟥





E.g cost for features 🟥

C (🟥🟩🟦)+C(🟩⬛🟥)+C(🟧🟥🟦)
―――――――――――――――
3



Blackberry App Store

1

Raw Data

2

Refined Data

3

Database

Technical Information

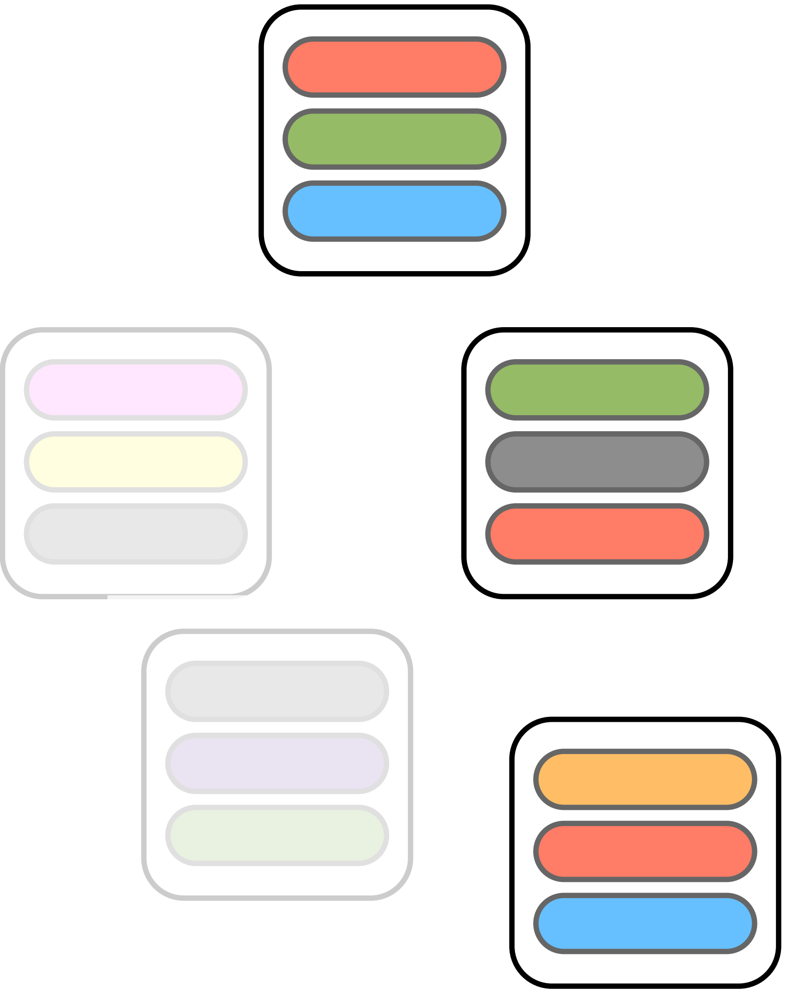
Business Information

Customer Information

4

Metrics

Correlation Analysis



E.g cost for features 🟥

$$\frac{C\,(\text{🟥🟩🟦})+C(\text{🟩⬛🟥})+C(\text{🟧🟥🟦})}{3}$$

$$F(f,d) = \frac{\sum_{a_i \in S(f,d)} A(a_i,d)}{\sharp(S(f,d))}$$

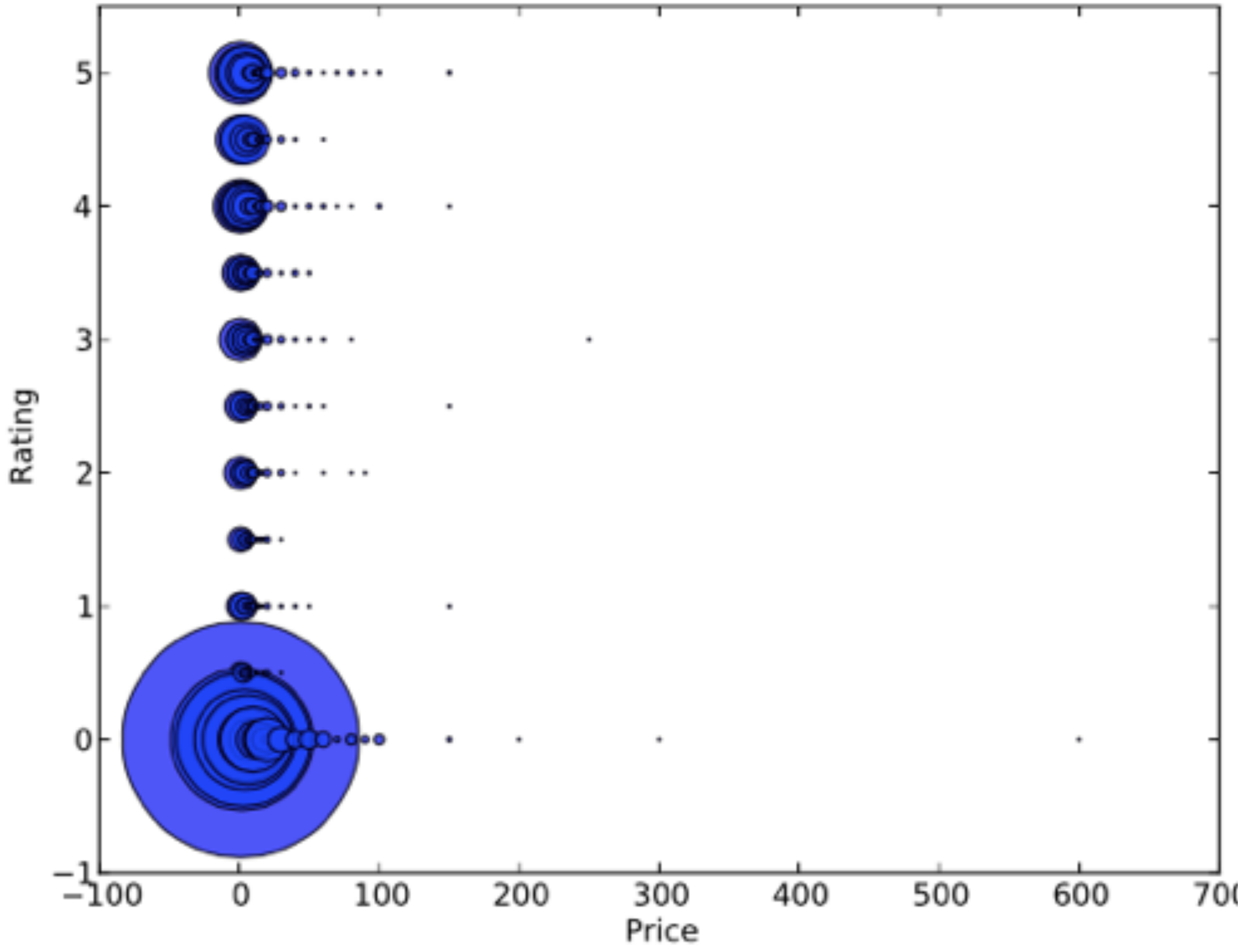
DATA SET

SNAPSHOT ON THE
1ST OF SEPTEMBER 2011

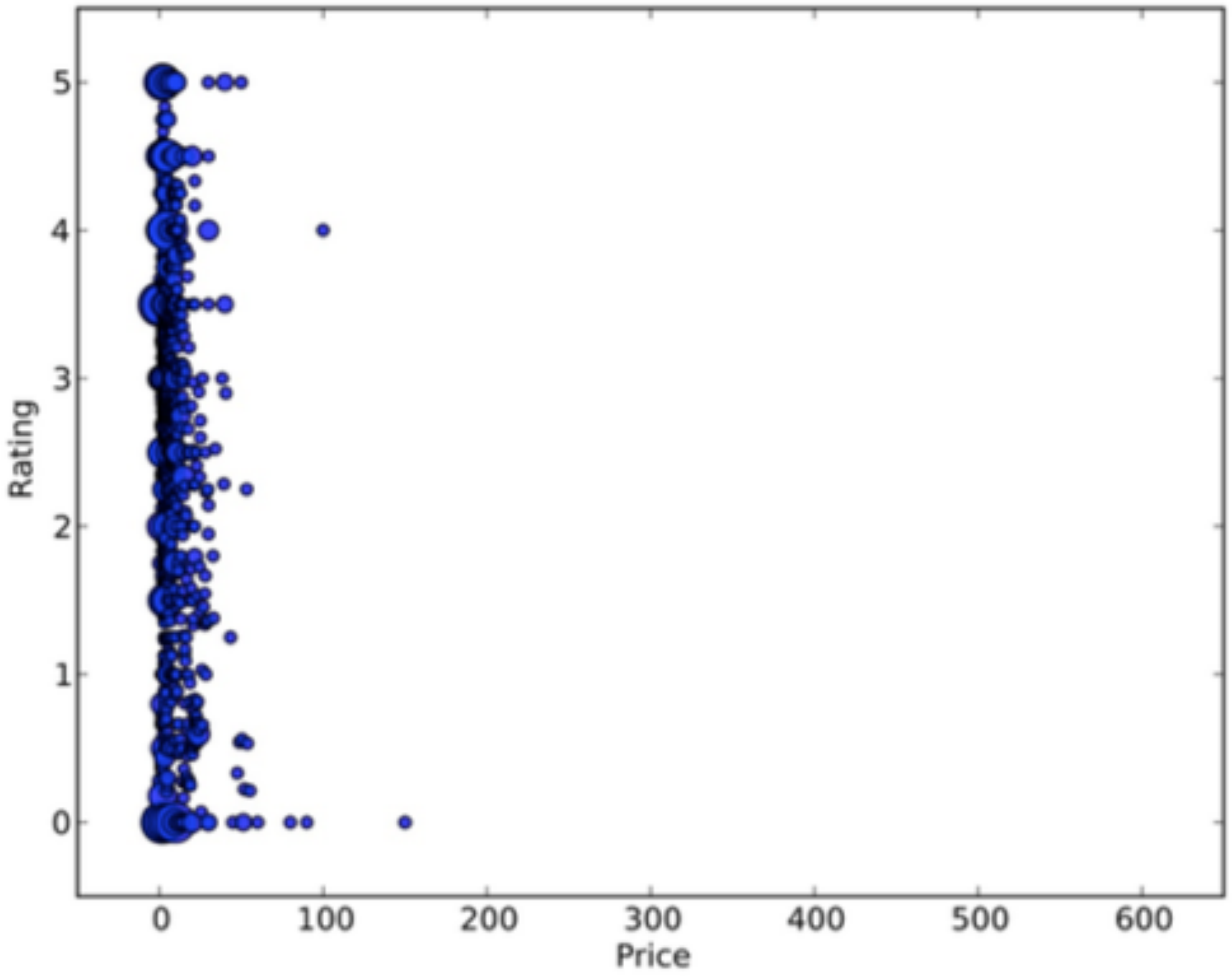
19 CATEGORIES FOR
32108 NON-FREE AND 9984 FREE APPS

EXTRACTED 1008 FEATURES

PRICE VS RATING CORRELATION

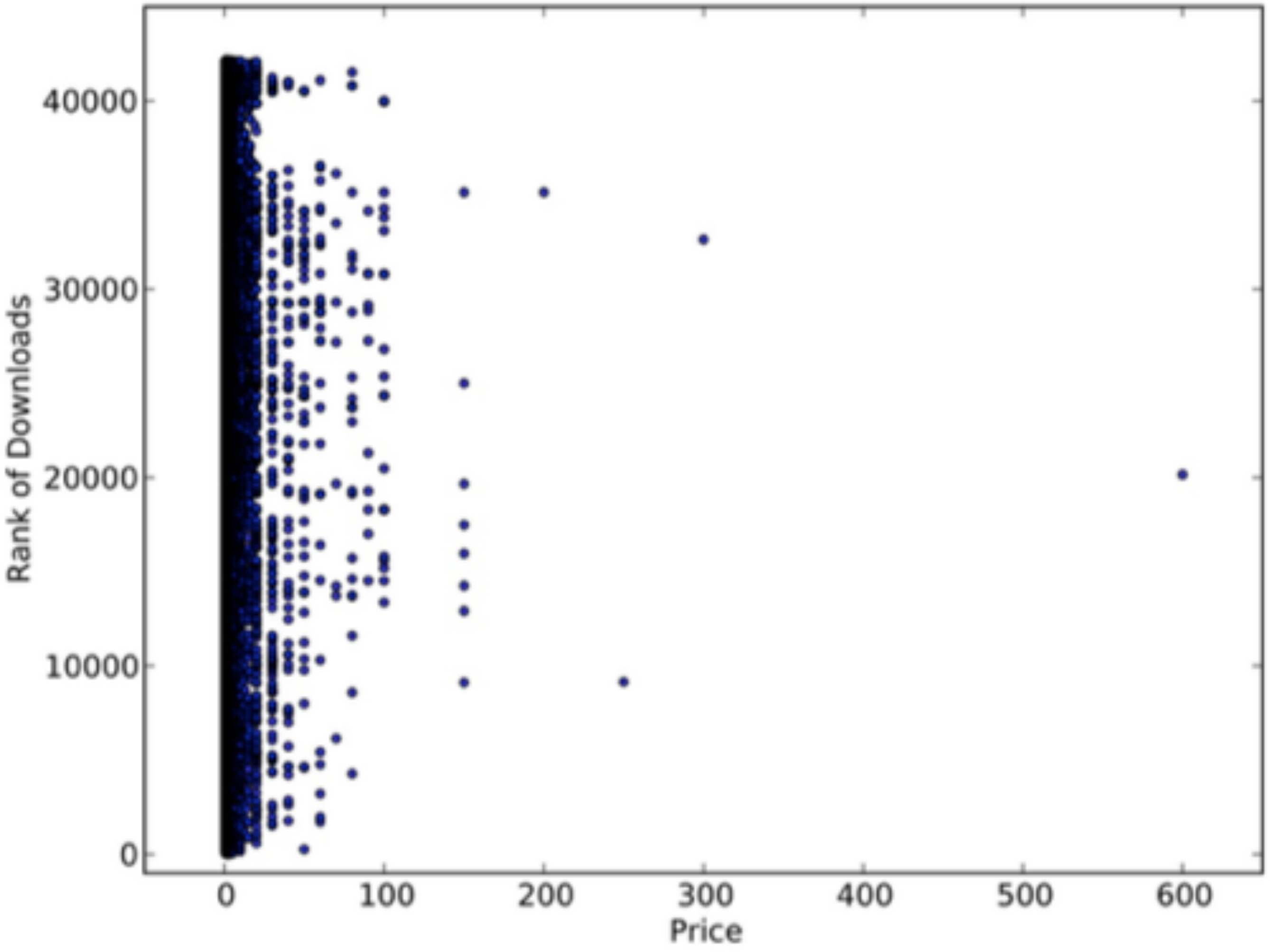


(a) PR non-free apps

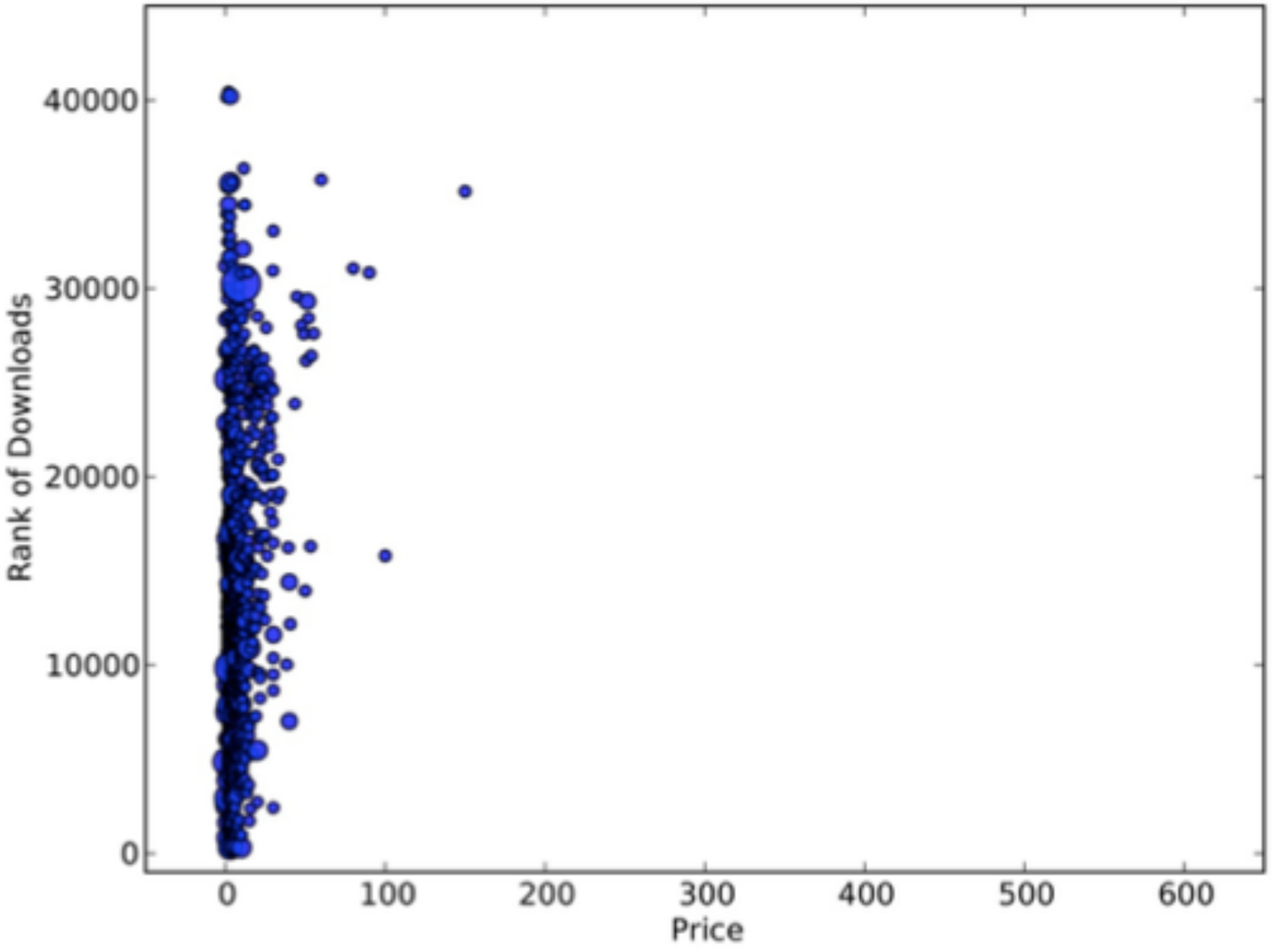


(i) PR non-free features

PRICE VS POPULARITY CORRELATION

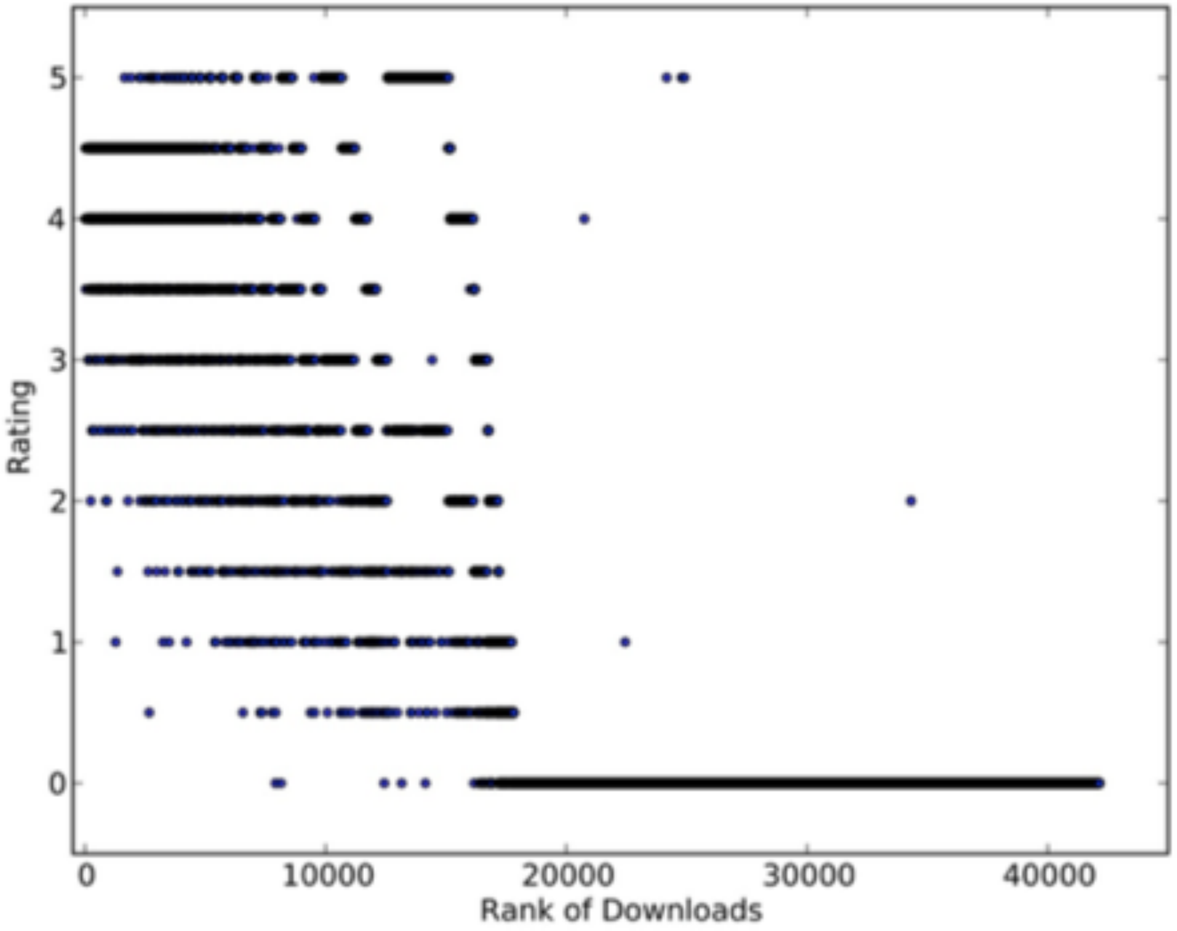


(b) PD non-free apps

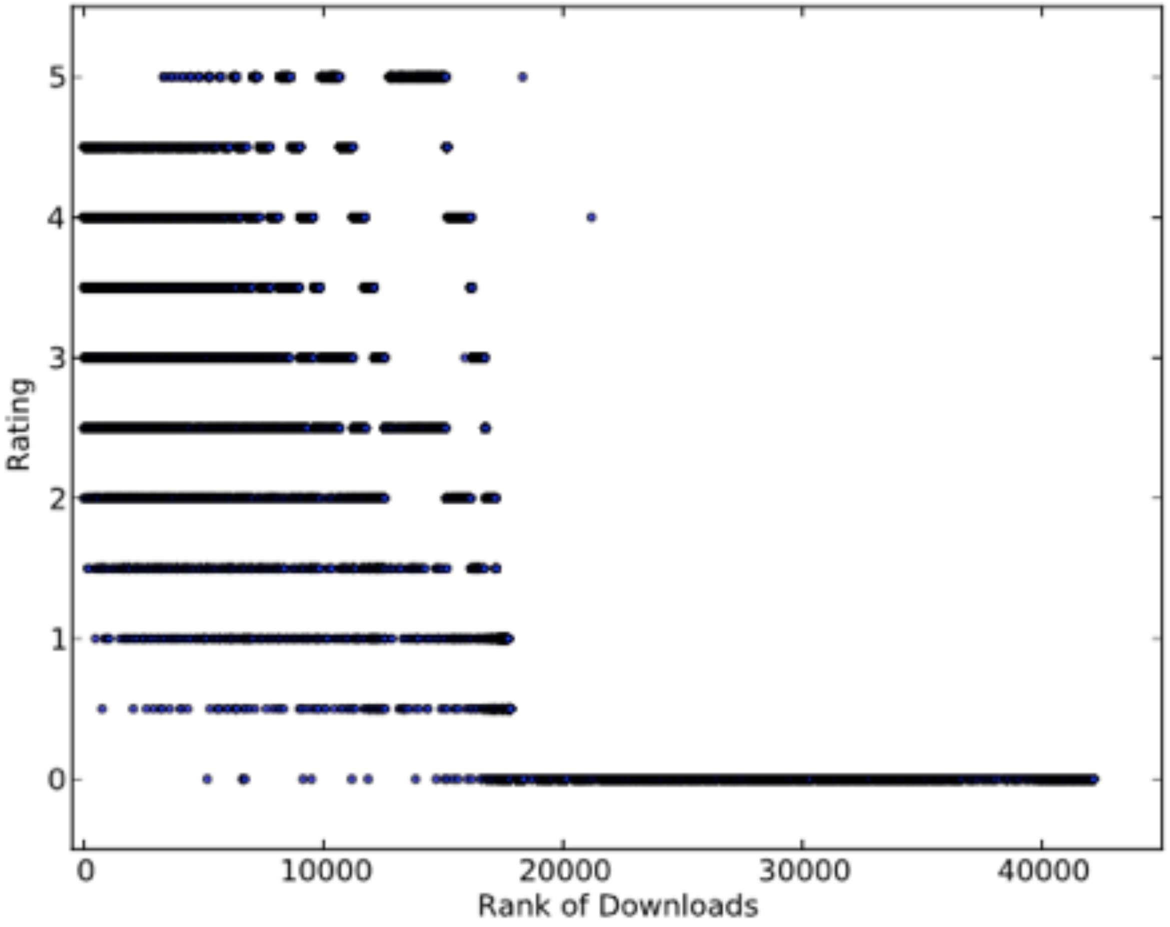


(j) PD non-free features

RATING VS POPULARITY CORRELATION

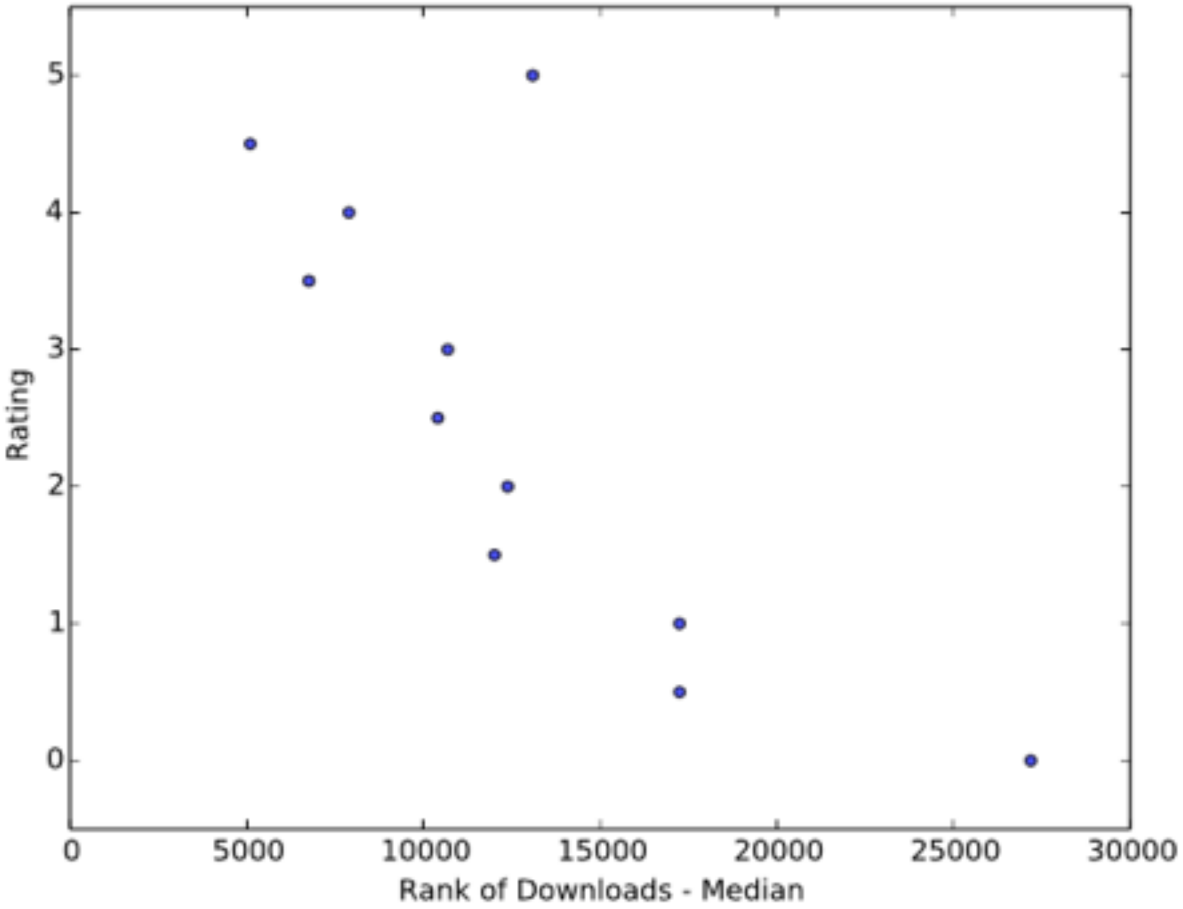


(c) RD non-free apps

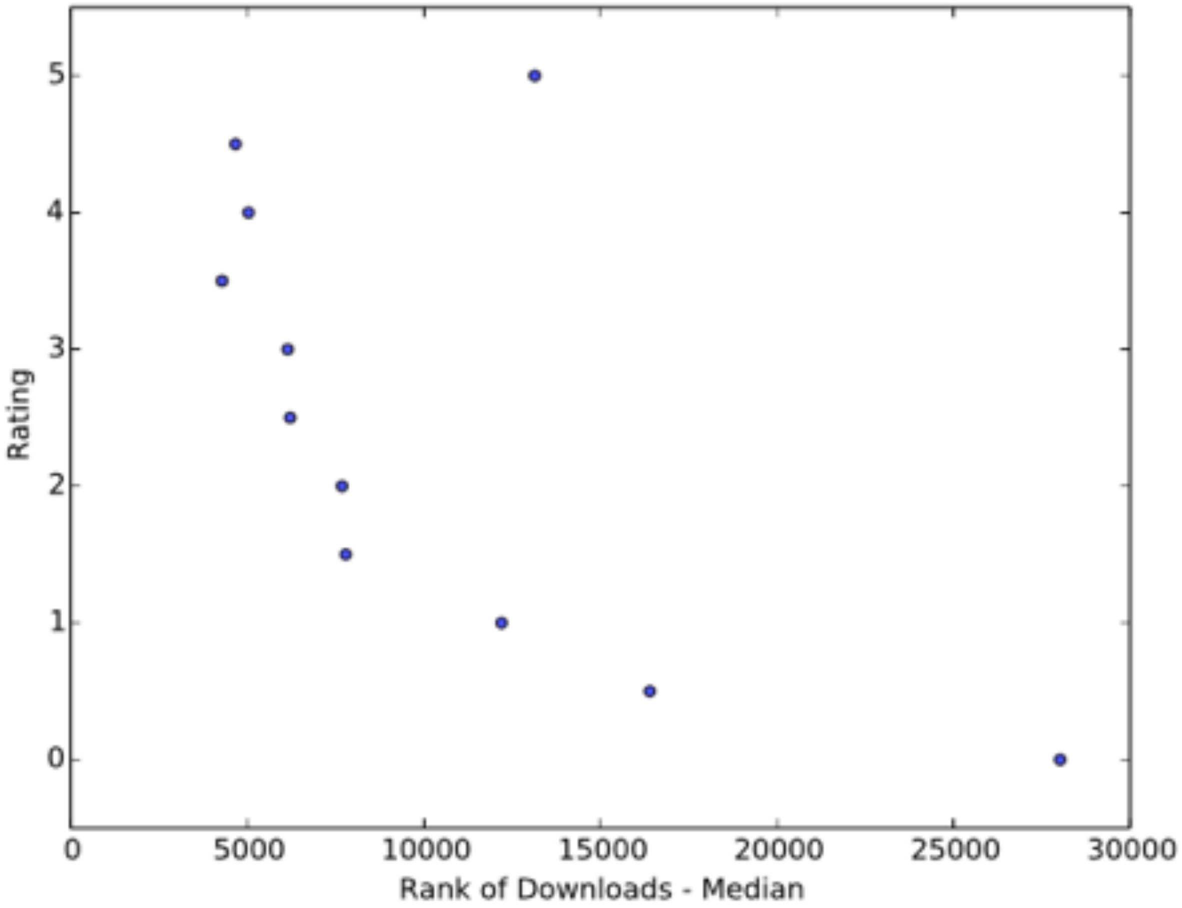


(d) RD free apps

RATING VS POPULARITY CORRELATION

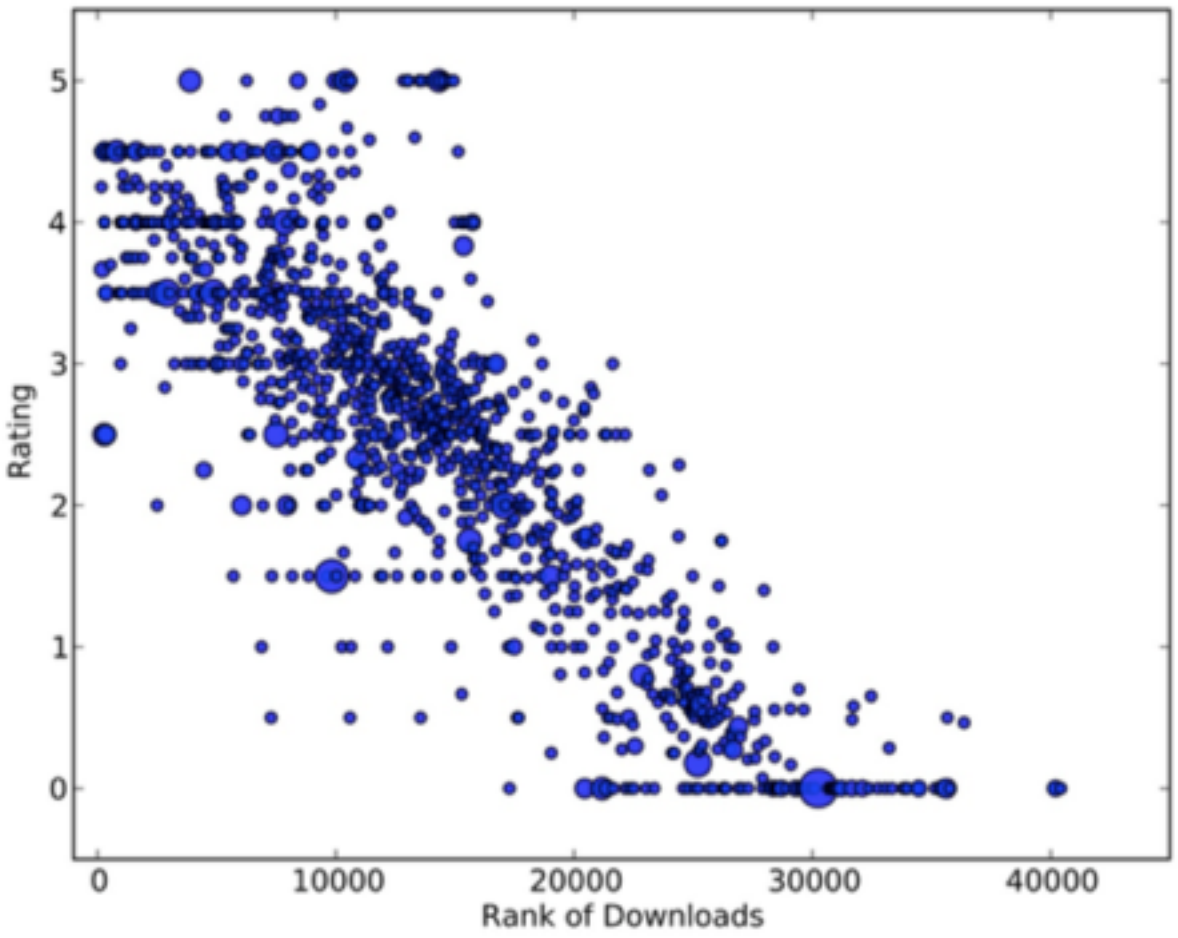


(e) MedianRD non-free apps

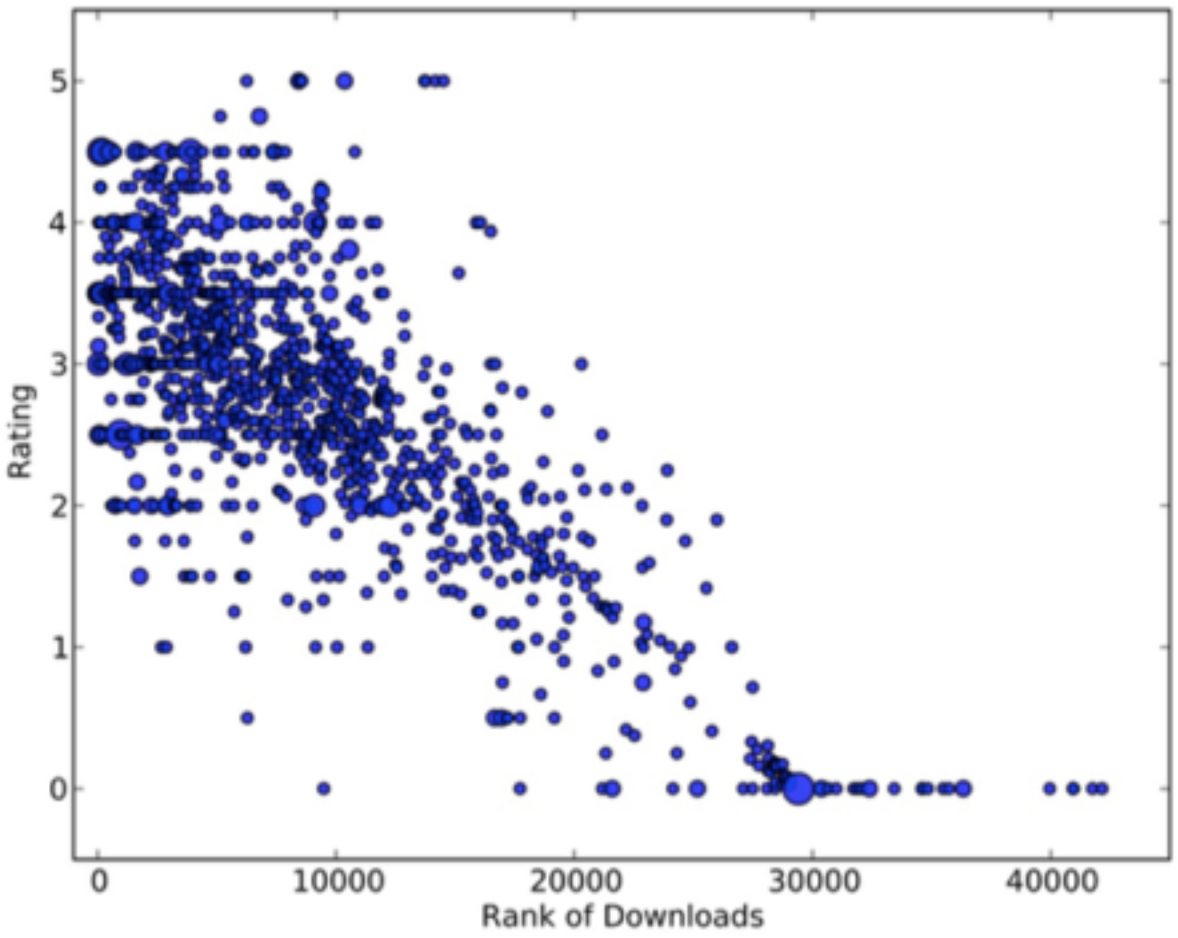


(f) MedianRD free apps

RATING VS POPULARITY CORRELATION



(k) RD non-free features

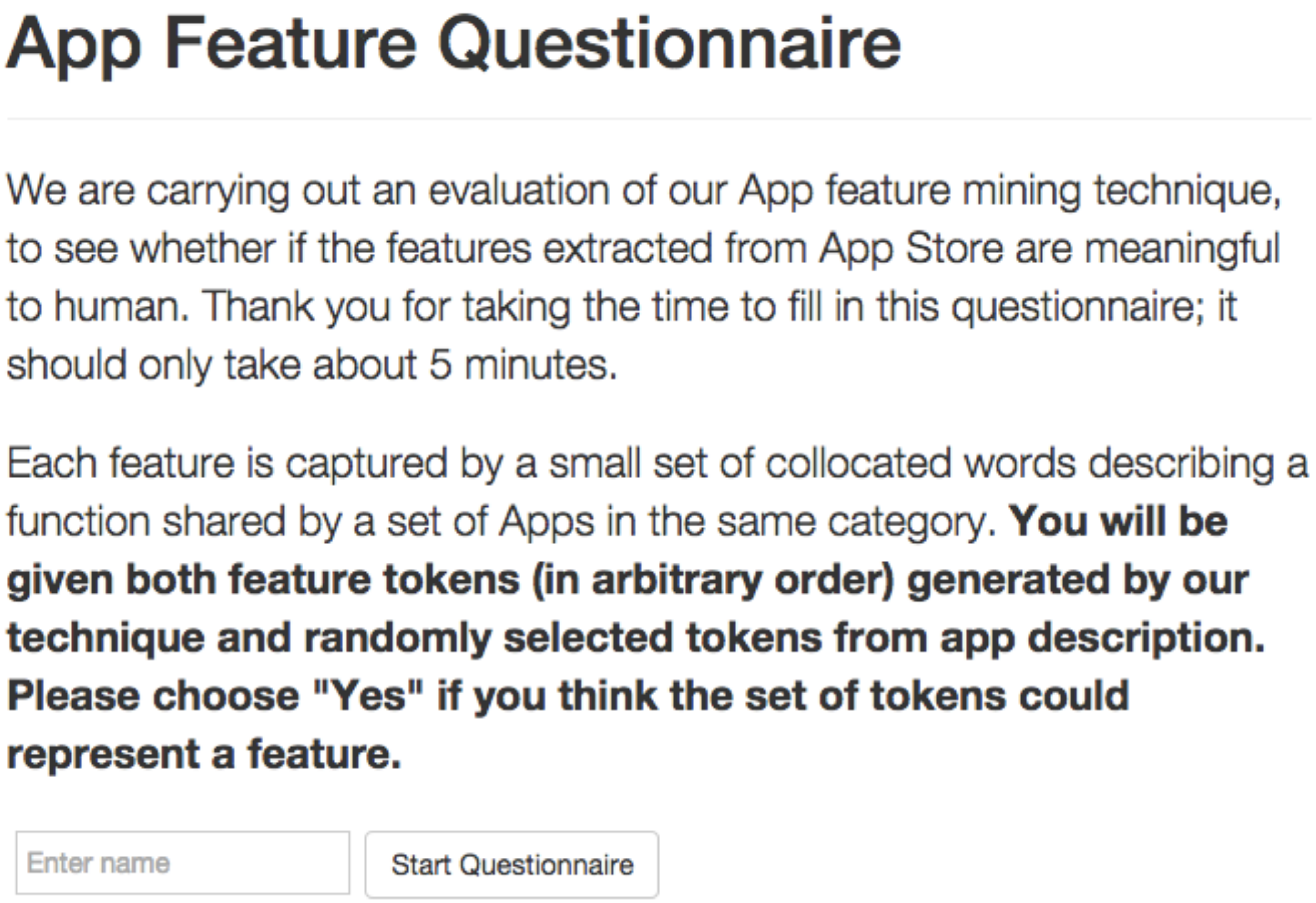


(l) RD free features

“ RATING MATTERS

Our results show that there is **a correlation** between **customer rating and the rank of app downloads** for apps and the features extracted from them and for **both free and non-free apps and features**. However, there is **very little evidence for any correlation** between **price and either rating or popularity.**

MEANINGFUL FEATURES?

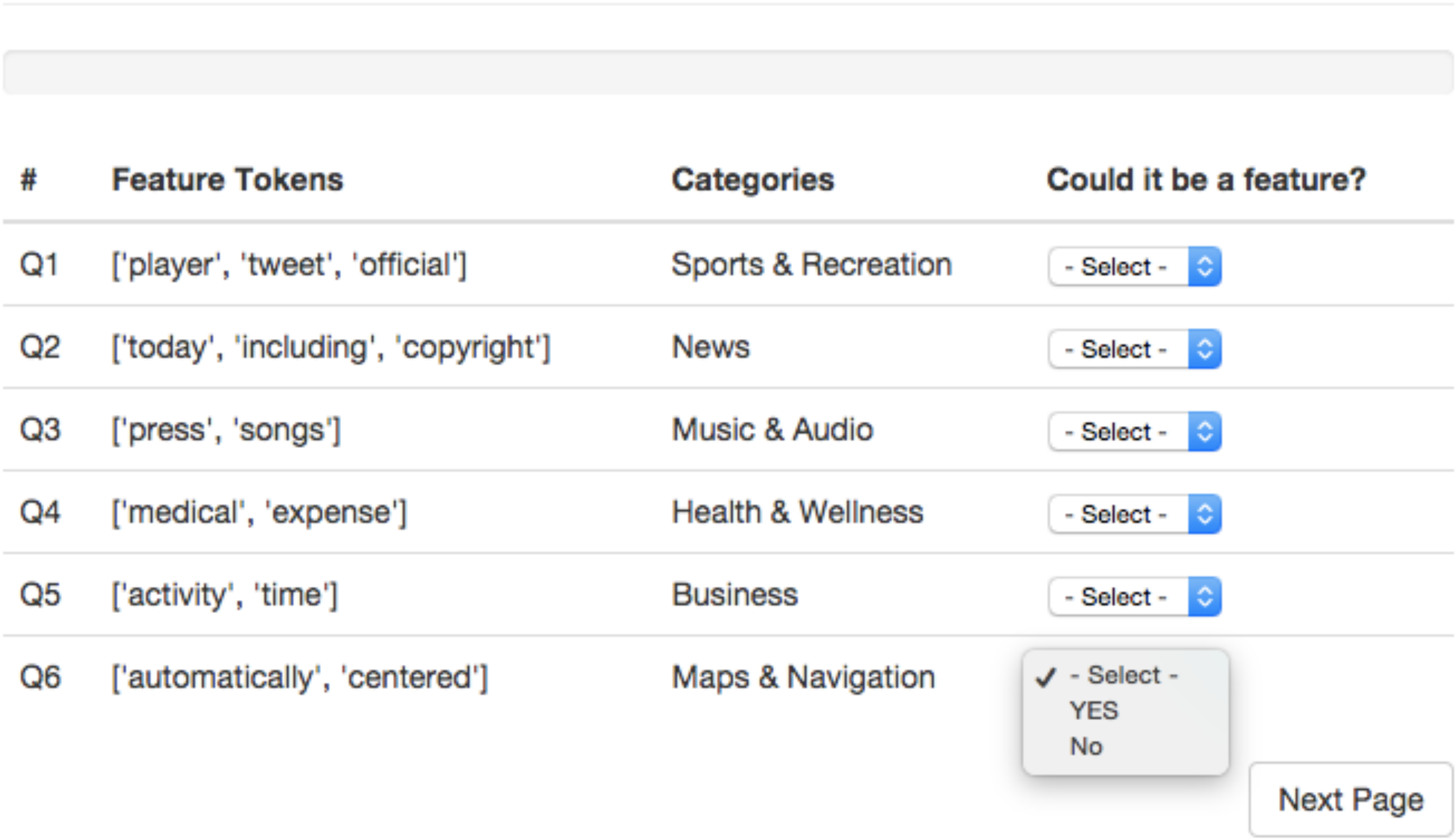


App Feature Questionnaire

We are carrying out an evaluation of our App feature mining technique, to see whether if the features extracted from App Store are meaningful to human. Thank you for taking the time to fill in this questionnaire; it should only take about 5 minutes.

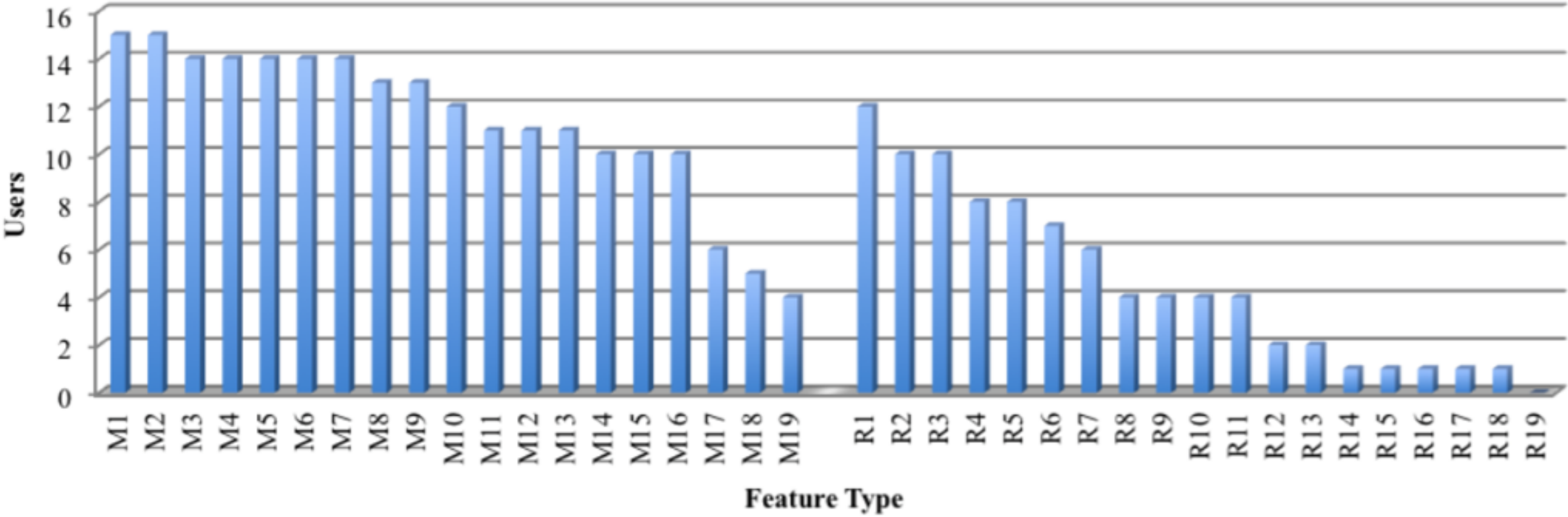
Each feature is captured by a small set of collocated words describing a function shared by a set of Apps in the same category. **You will be given both feature tokens (in arbitrary order) generated by our technique and randomly selected tokens from app description. Please choose "Yes" if you think the set of tokens could represent a feature.**

MEANINGFUL FEATURES?



#	Feature Tokens	Categories	Could it be a feature?
Q1	['player', 'tweet', 'official']	Sports & Recreation	- Select -
Q2	['today', 'including', 'copyright']	News	- Select -
Q3	['press', 'songs']	Music & Audio	- Select -
Q4	['medical', 'expense']	Health & Wellness	- Select -
Q5	['activity', 'time']	Business	- Select -
Q6	['automatically', 'centered']	Maps & Navigation	

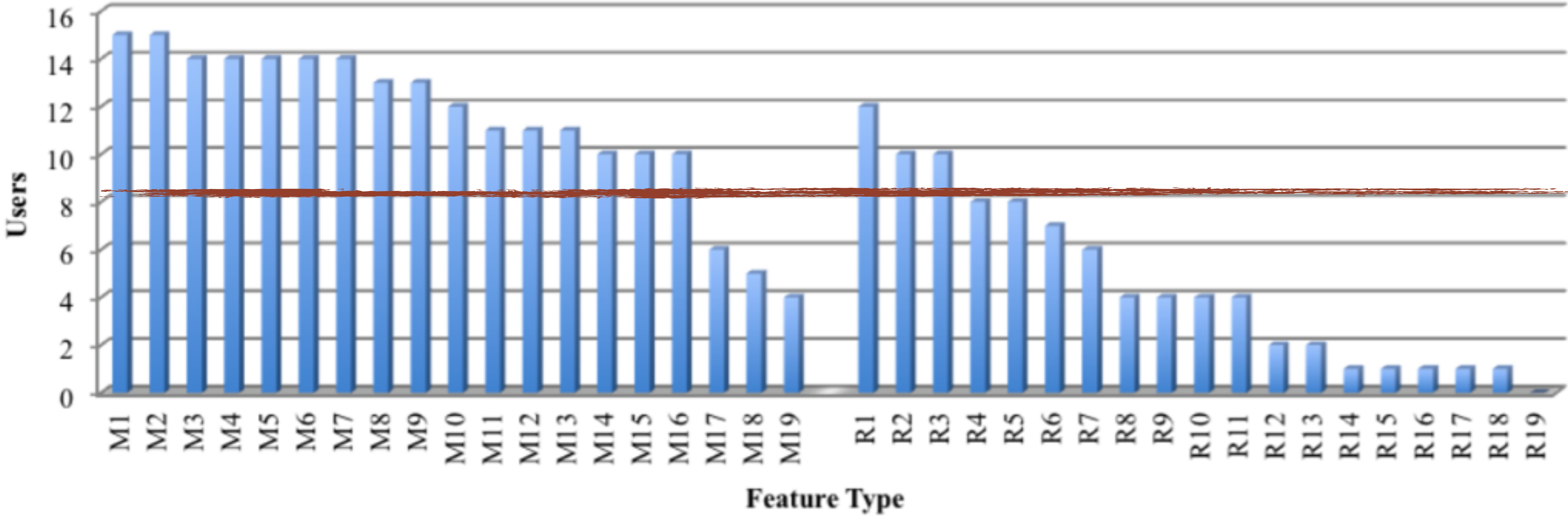
MEANINGFUL FEATURES?



Algorithm Extracted

Random Generated

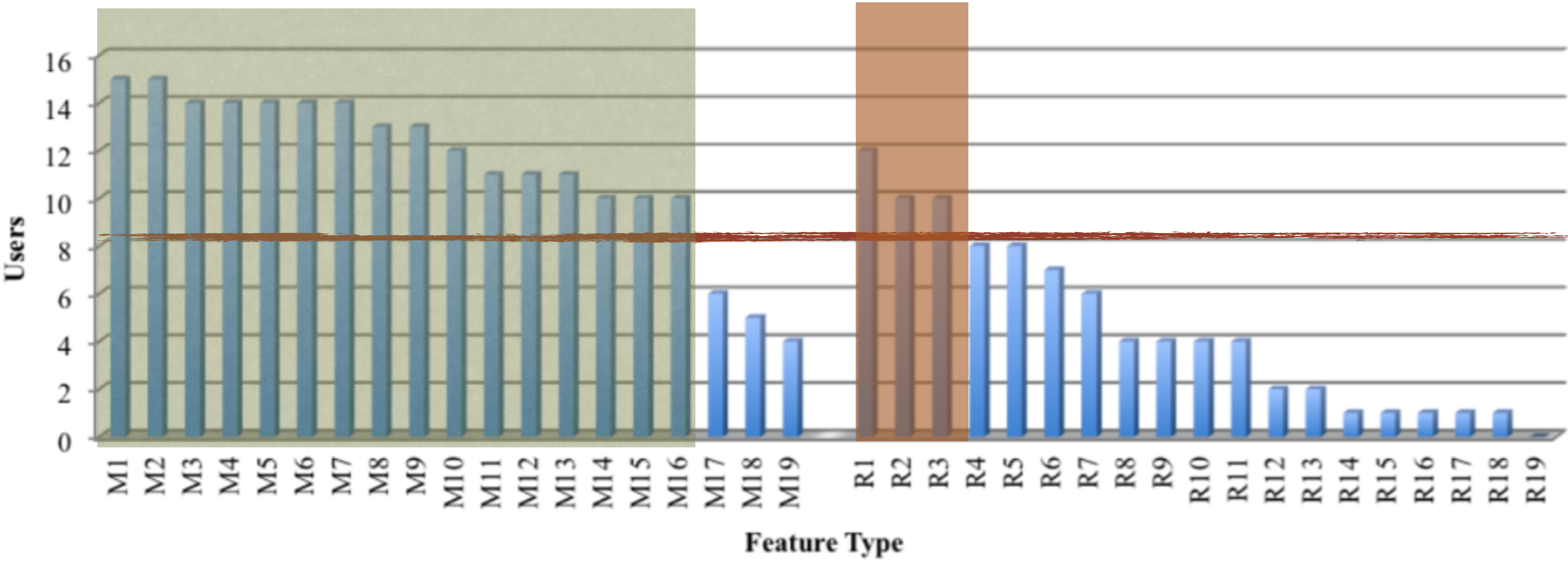
MEANINGFUL FEATURES?



Algorithm Extracted

Random Generated

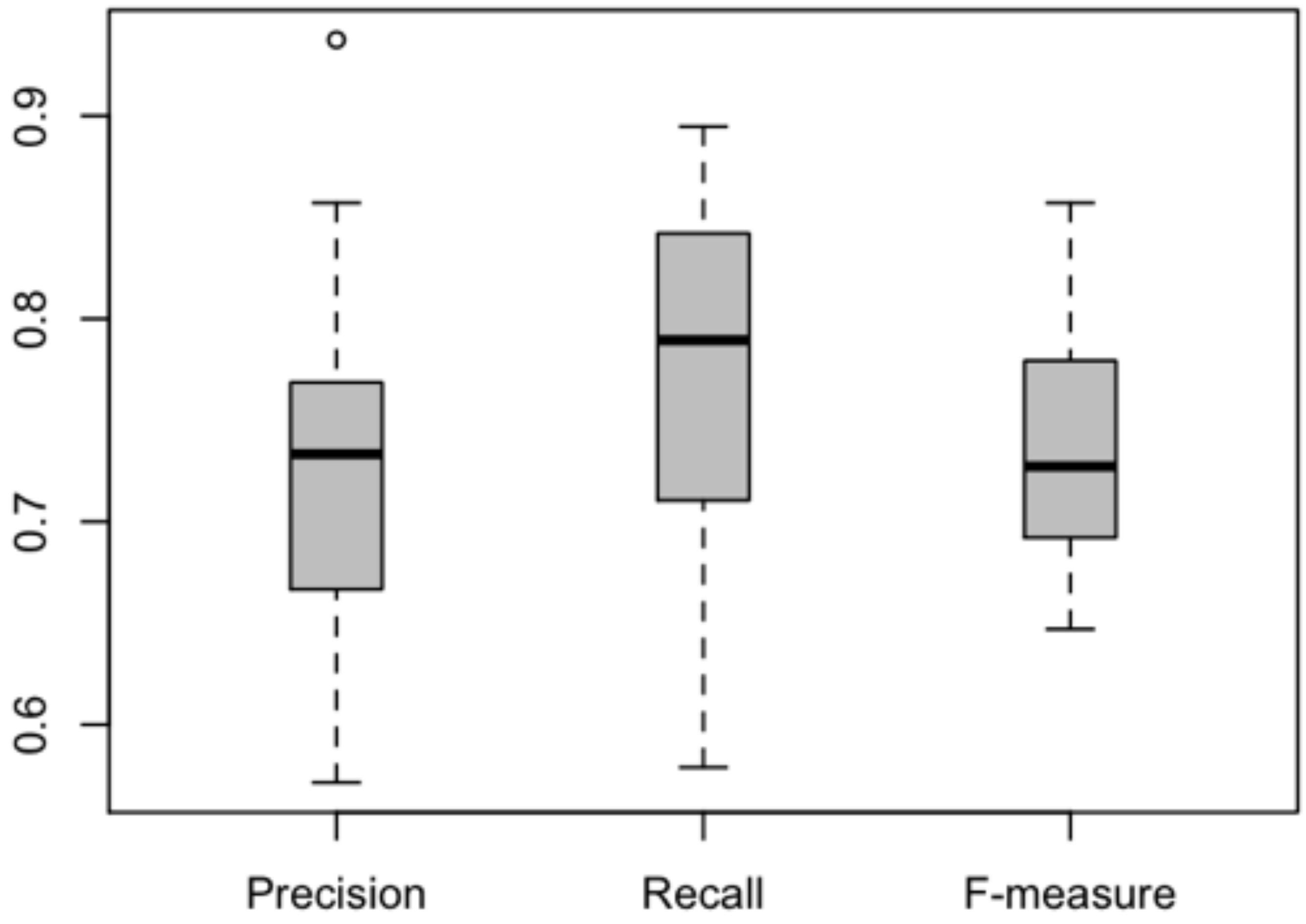
MEANINGFUL FEATURES?



Algorithm Extracted

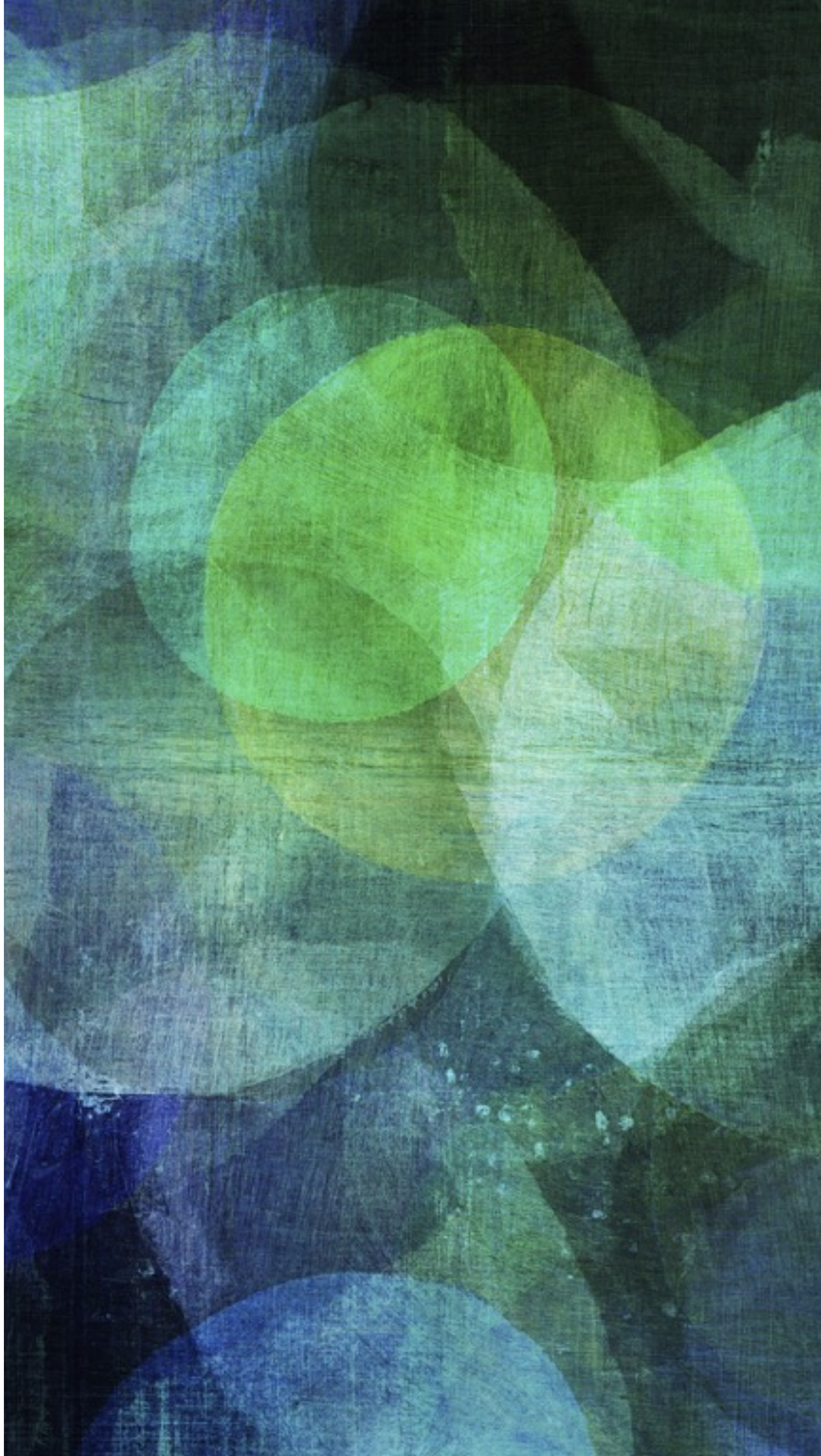
Random Generated

“ There is evidence that the bitri-grams of features extracted are meaningful to humans.

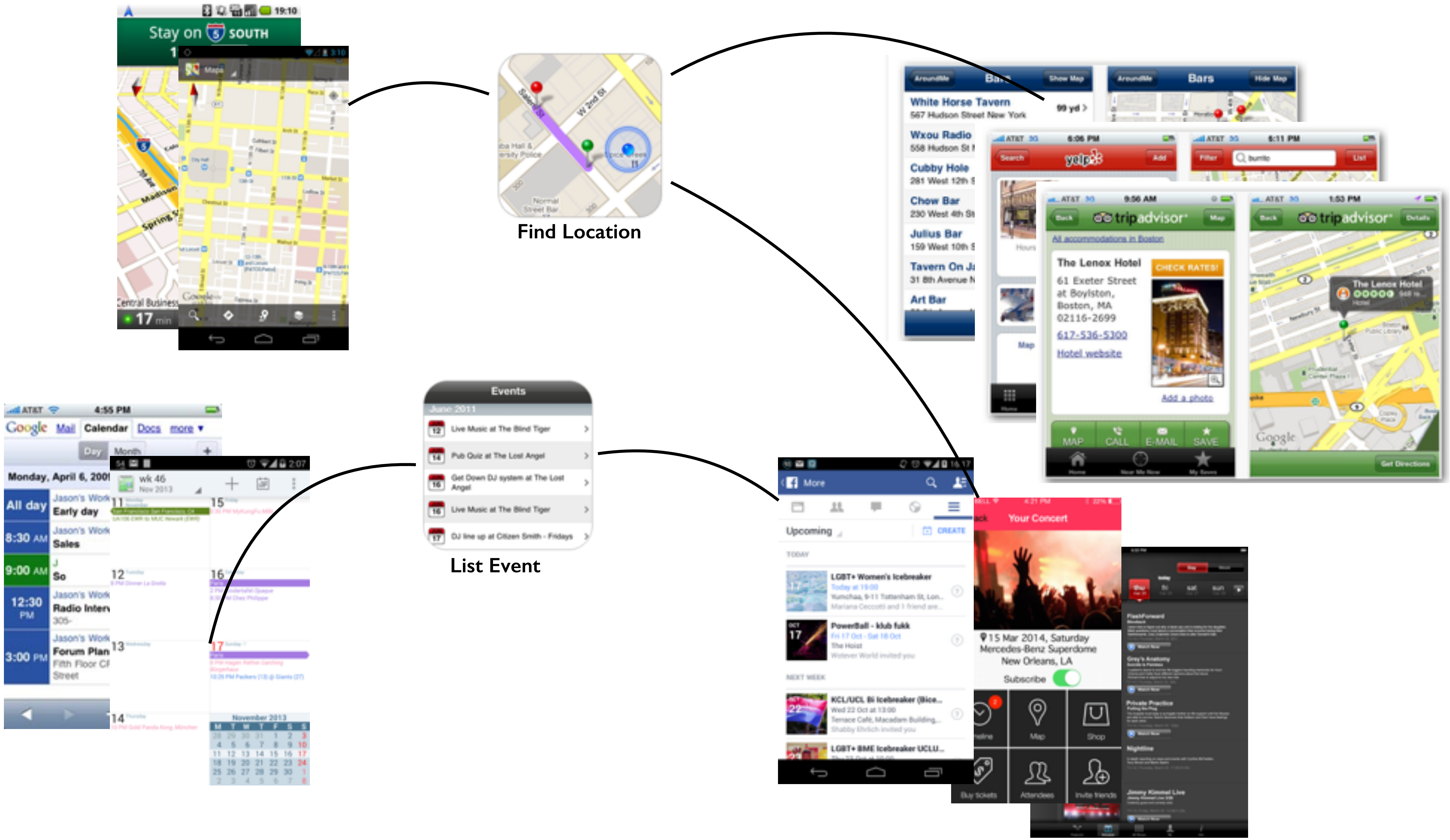


FEATURE MIGRATION

Feature lifecycles as they spread, migrate, remain, and die in app stores (RE'16)



Feature Migration



Find Location

List Event

We can ask

Does Migration follow the money?

We can ask

Popularity implies migration ?

Points Of Interest

June 2011

12 Live Music at The Blind Tiger

14 Pub Quiz at The Lost Angel

16 Get Down DJ system at The Lost Angel

16 Live Music at The Blind Tiger

17 DJ line up at Citizen Smith - Fridays

List Events

Chico State Wrec
150 Cherry St.
Chico, Ca, 95928

530-898-4444

http://www.aschico.com/wrec

wrecinfo@csuchico.edu

Show Contact Detail

Send us a photo

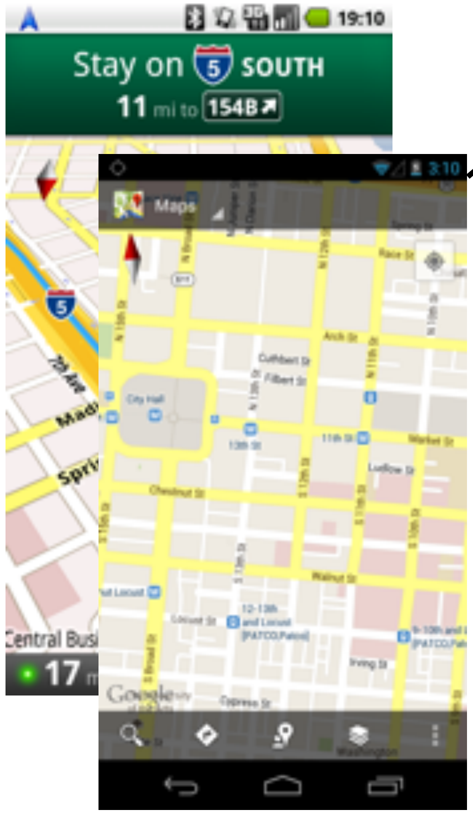
Got in an accident? Send a picture now to start your case.

Email Picture

What Developers may Ask

Which categories are more likely to migrate features to one other?

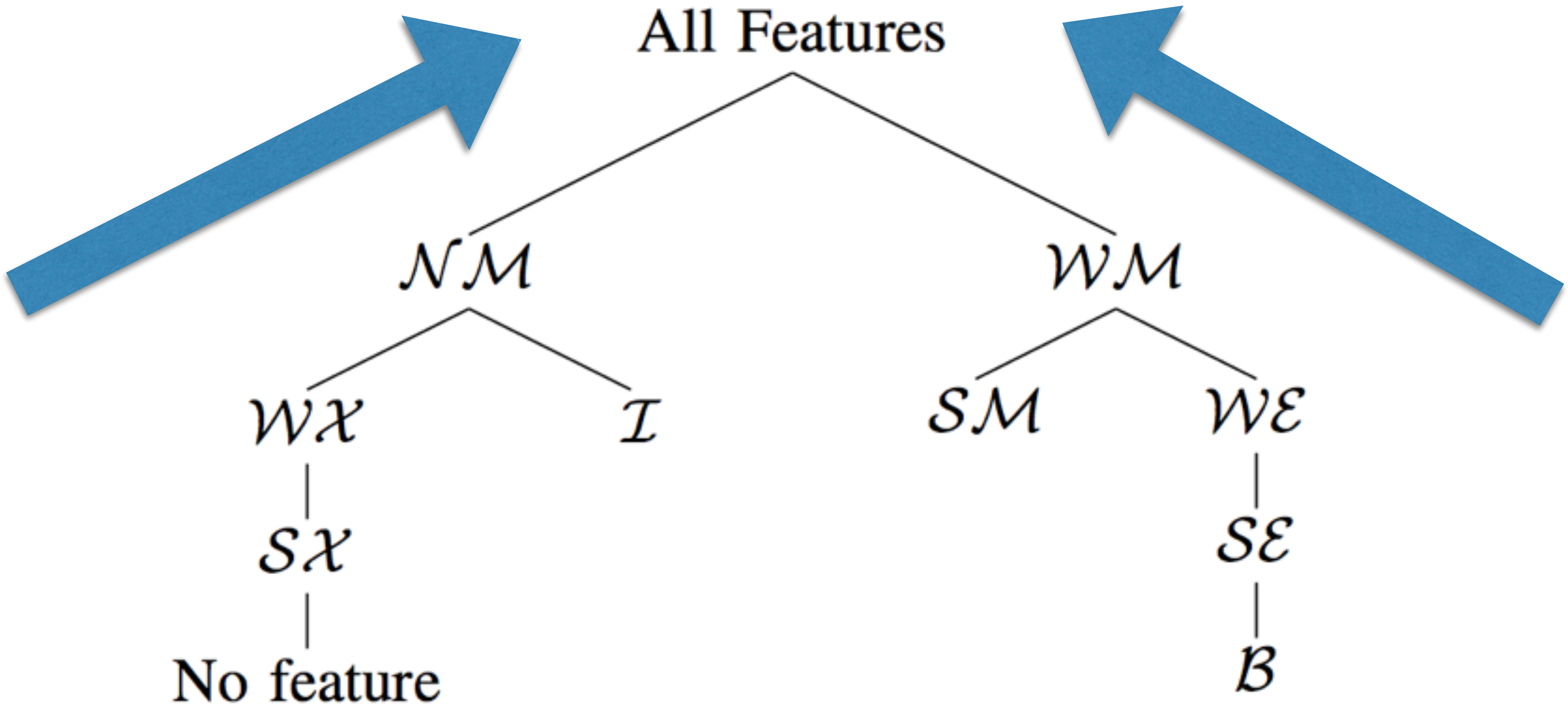
Maps & Navigation



Find Location

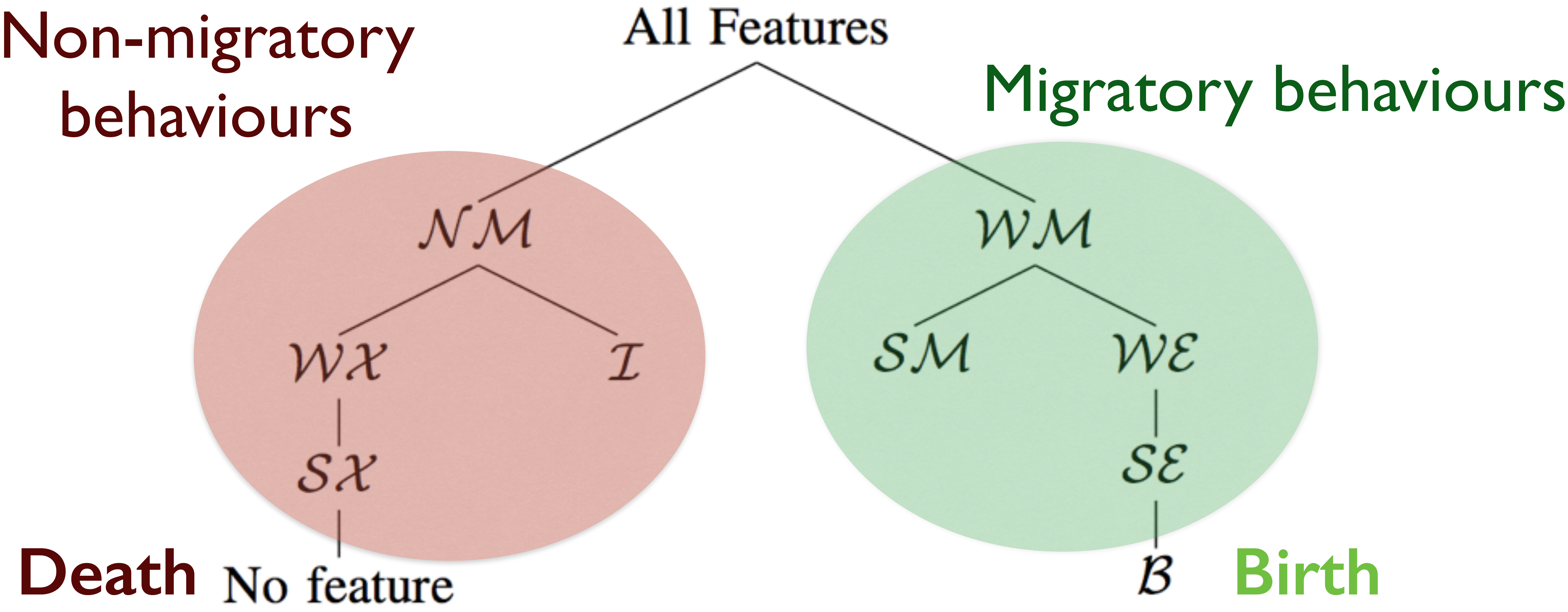
Travel Apps

Set Theoretic Characterisation of App Store Feature Migration



The Theoretical Feature Migration Subsumption Hierarchy

Set Theoretic Characterisation of App Store Feature Migration



Non-migratory behaviours

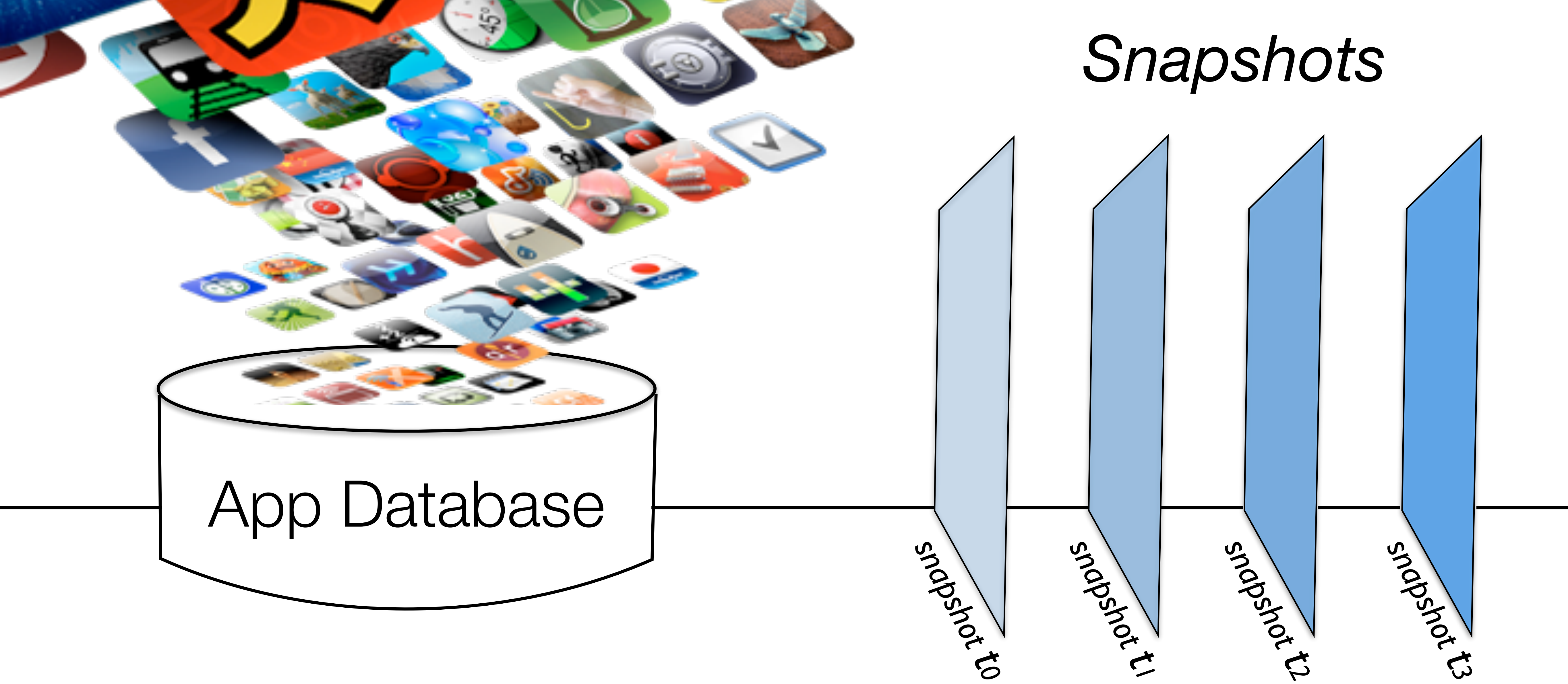
Migratory behaviours

Death

Birth

The Theoretical Feature Migration Subsumption Hierarchy

Snapshots



App Database

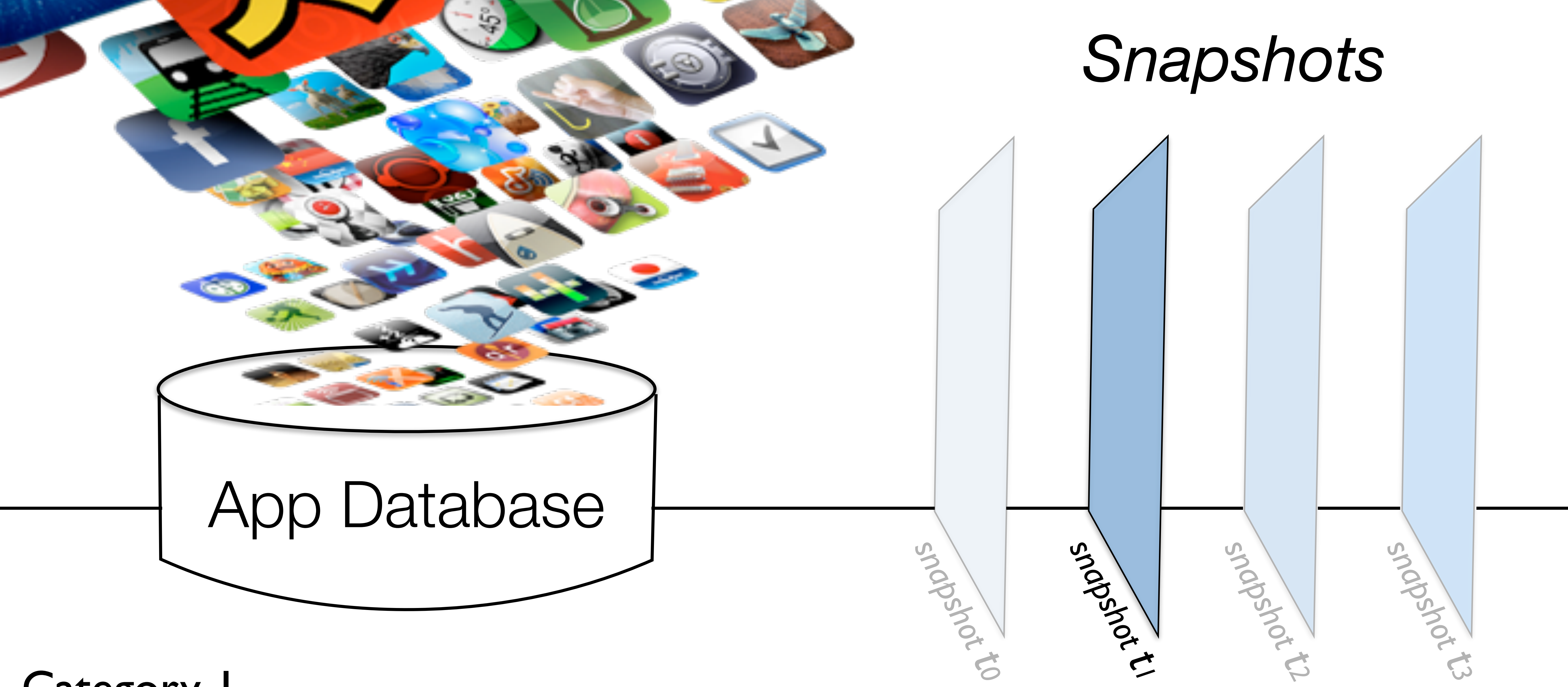
snapshot t0

snapshot t1

snapshot t2

snapshot t3

Snapshots



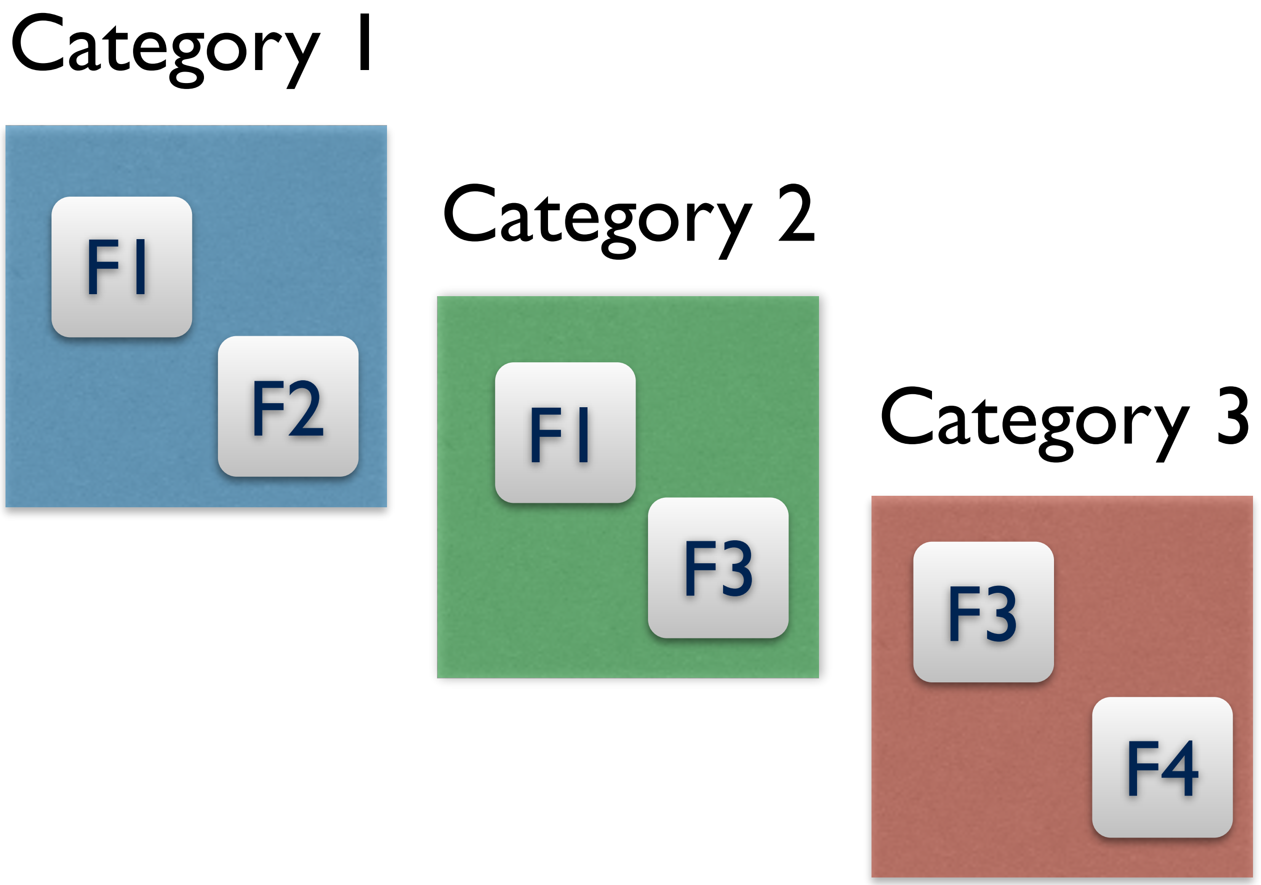
App Database

snapshot t_0

snapshot t_1

snapshot t_2

snapshot t_3



Category 1

F1

F2

Category 2

F1

F3

Category 3

F3

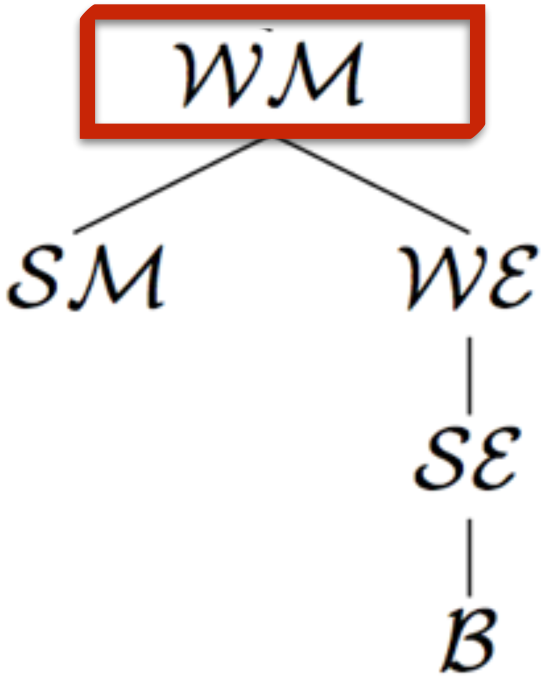
F4

Category Membership $c^{f}_{D\{t\}}$

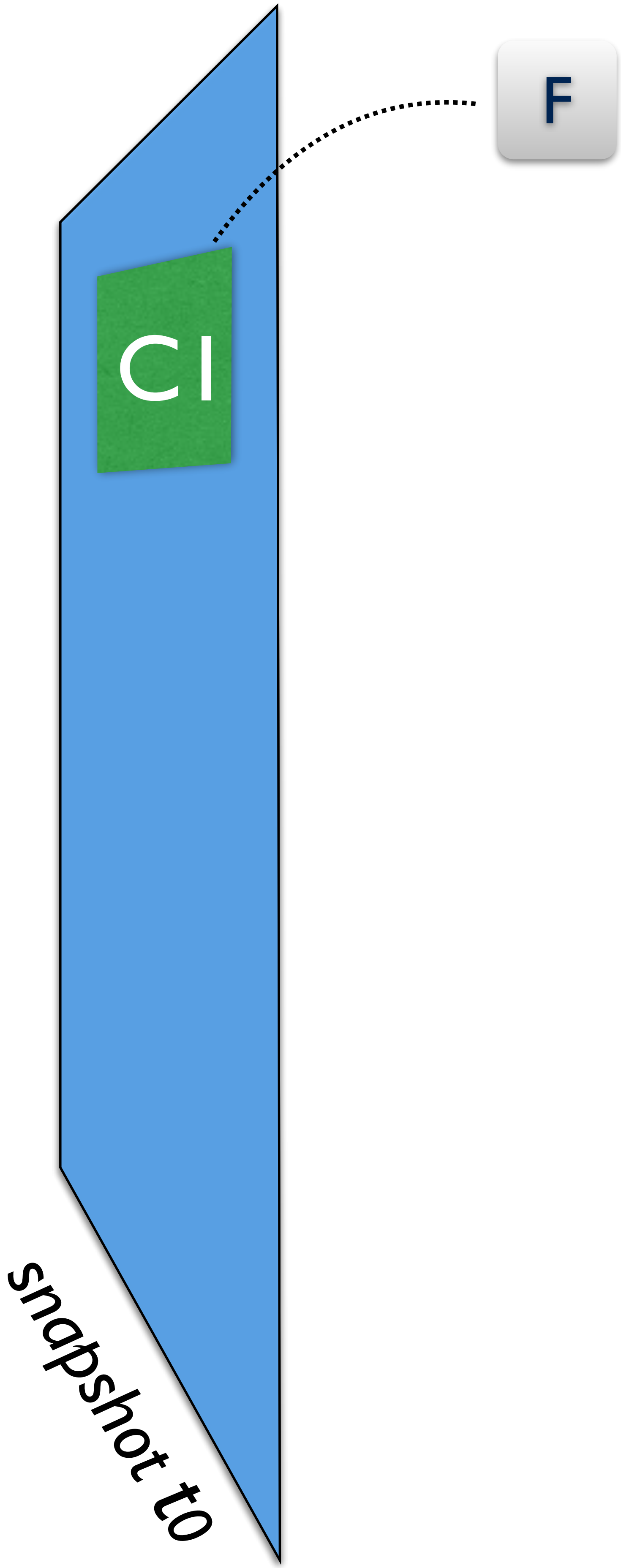
F1 is member of { 🟦 🟩 }

F3 is member of { 🟩 🟥 }

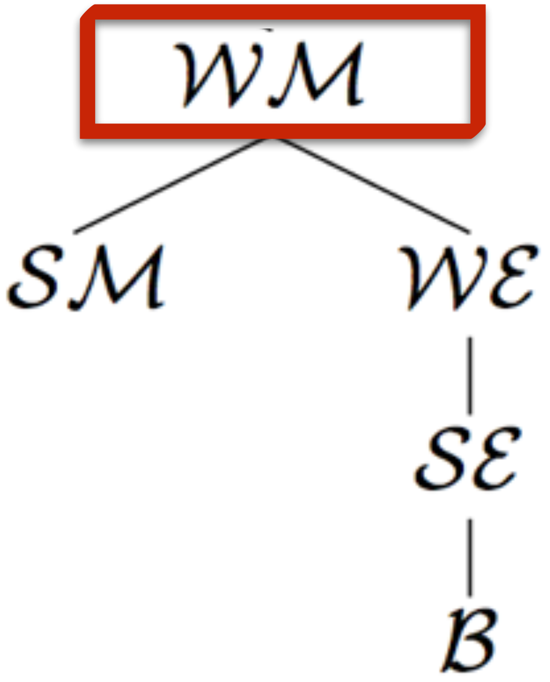
Weak Migration



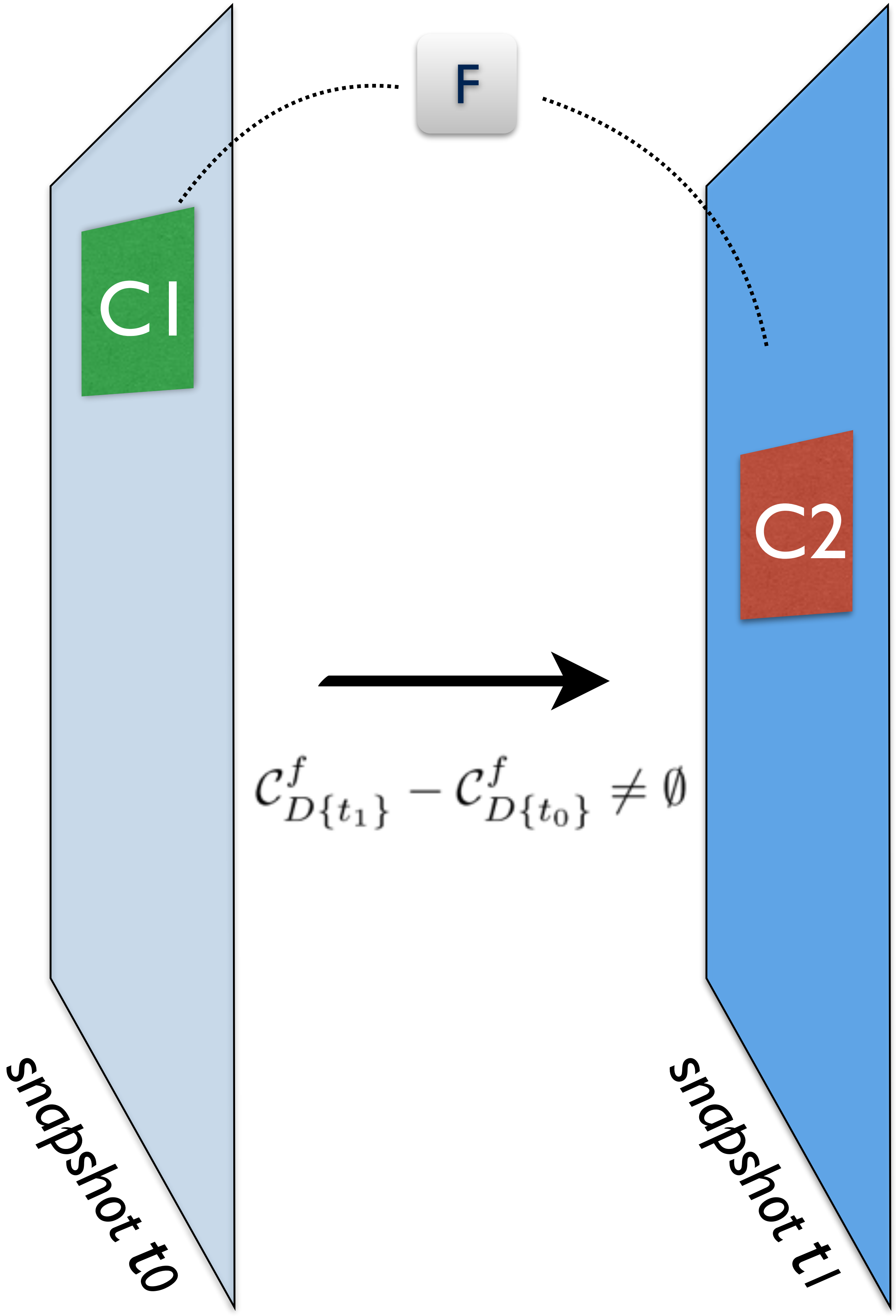
A feature *migrates* if it resides in at least one *new* category at the end of the time period considered (*WM*)



Weak Migration

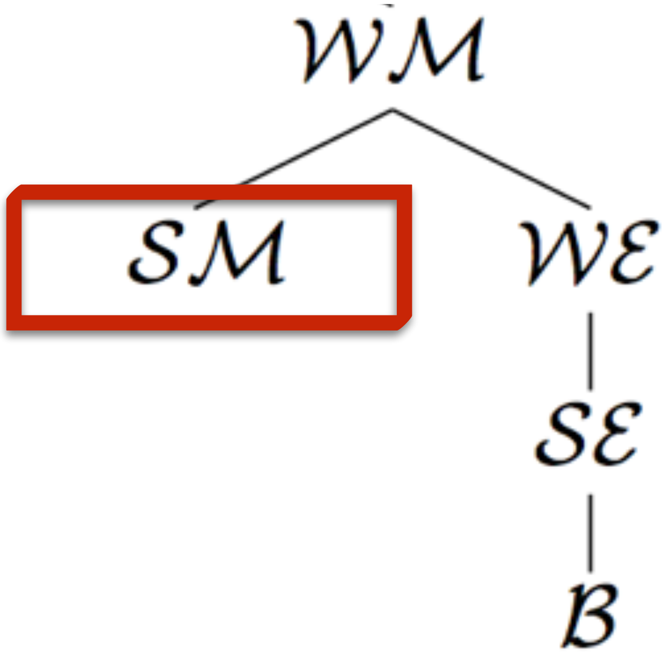


A feature *migrates* if it resides in at least one *new* category at the end of the time period considered (*WM*)

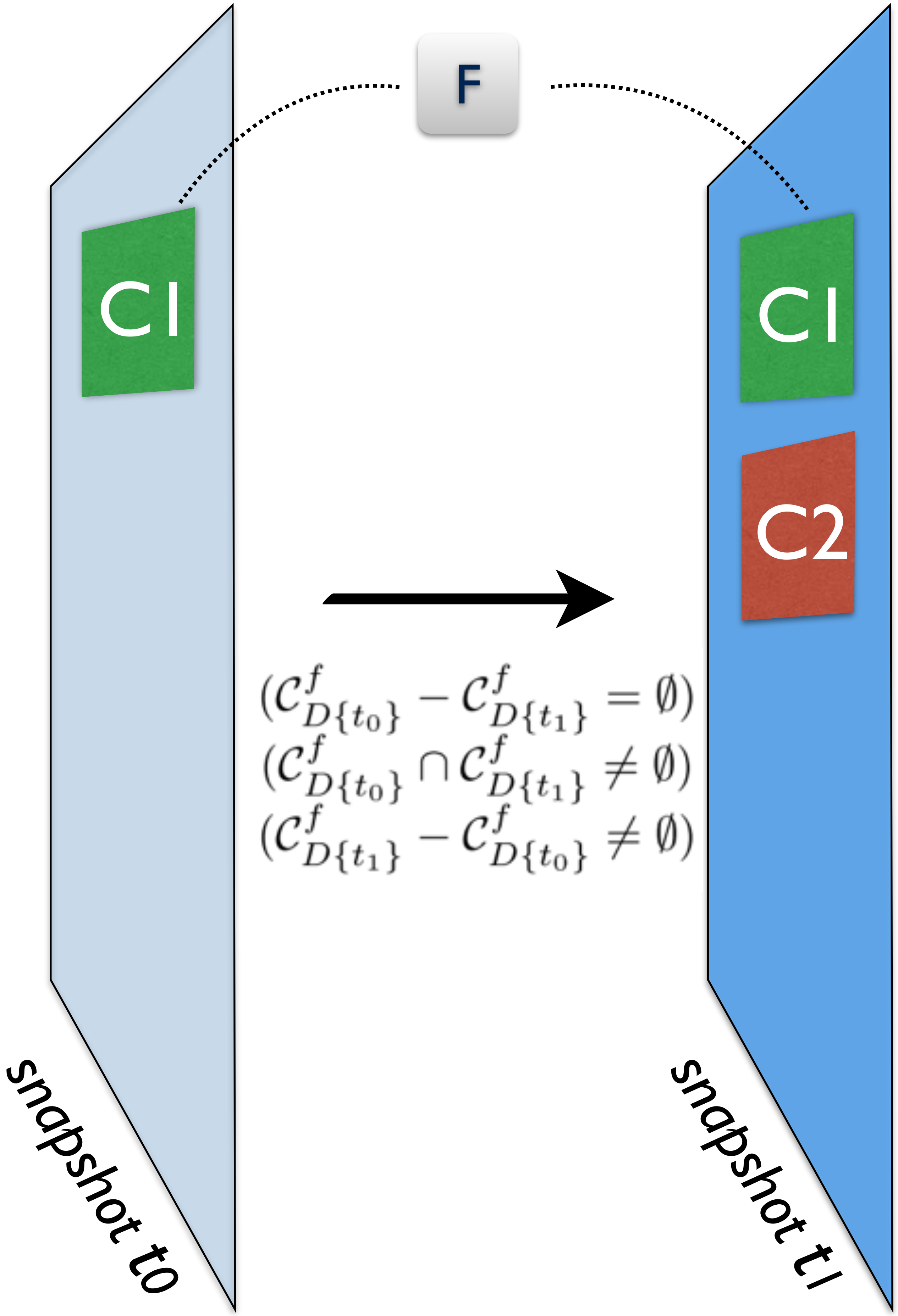


$$\mathcal{C}^f_{D\{t_1\}} - \mathcal{C}^f_{D\{t_0\}} \neq \emptyset$$

Strong Migration

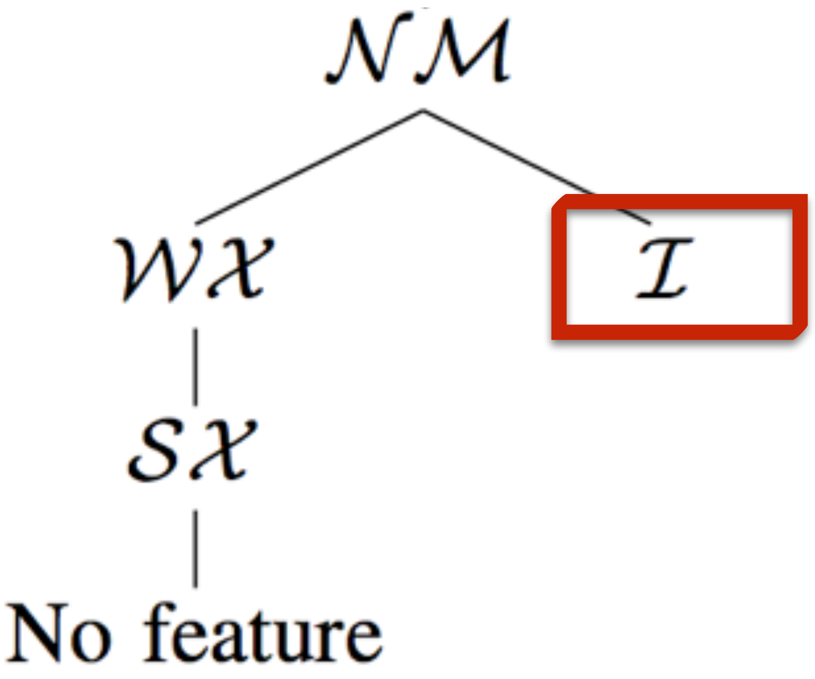


A feature spreads from at least one category to at least one new category and remains in all categories in which it originated (*SM*).

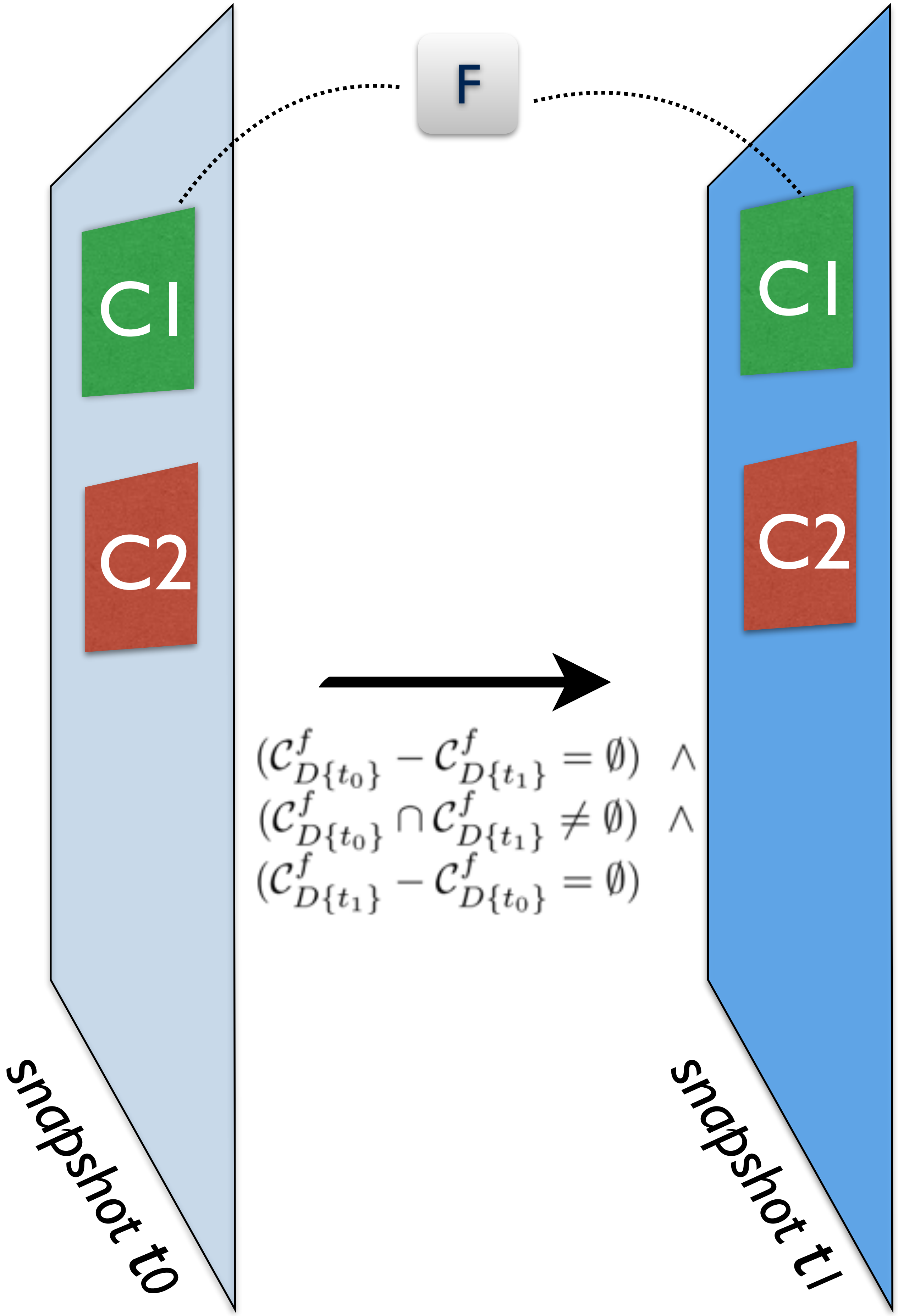


$$(\mathcal{C}^f_{D\{t_0\}} - \mathcal{C}^f_{D\{t_1\}} = \emptyset)$$
$$(\mathcal{C}^f_{D\{t_0\}} \cap \mathcal{C}^f_{D\{t_1\}} \neq \emptyset)$$
$$(\mathcal{C}^f_{D\{t_1\}} - \mathcal{C}^f_{D\{t_0\}} \neq \emptyset)$$

Intransitive

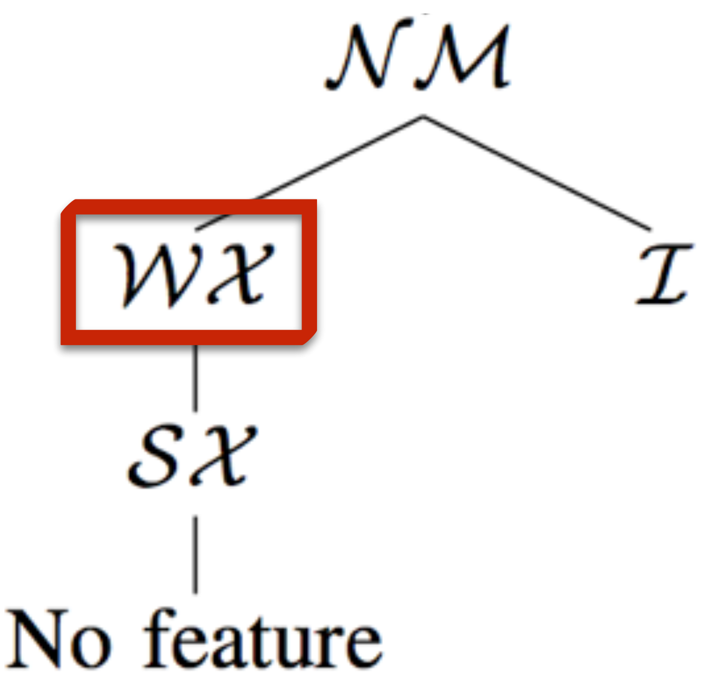


An intransitive feature neither appears in any new categories nor does it disappear from any between the start and the end of the time period considered (I).

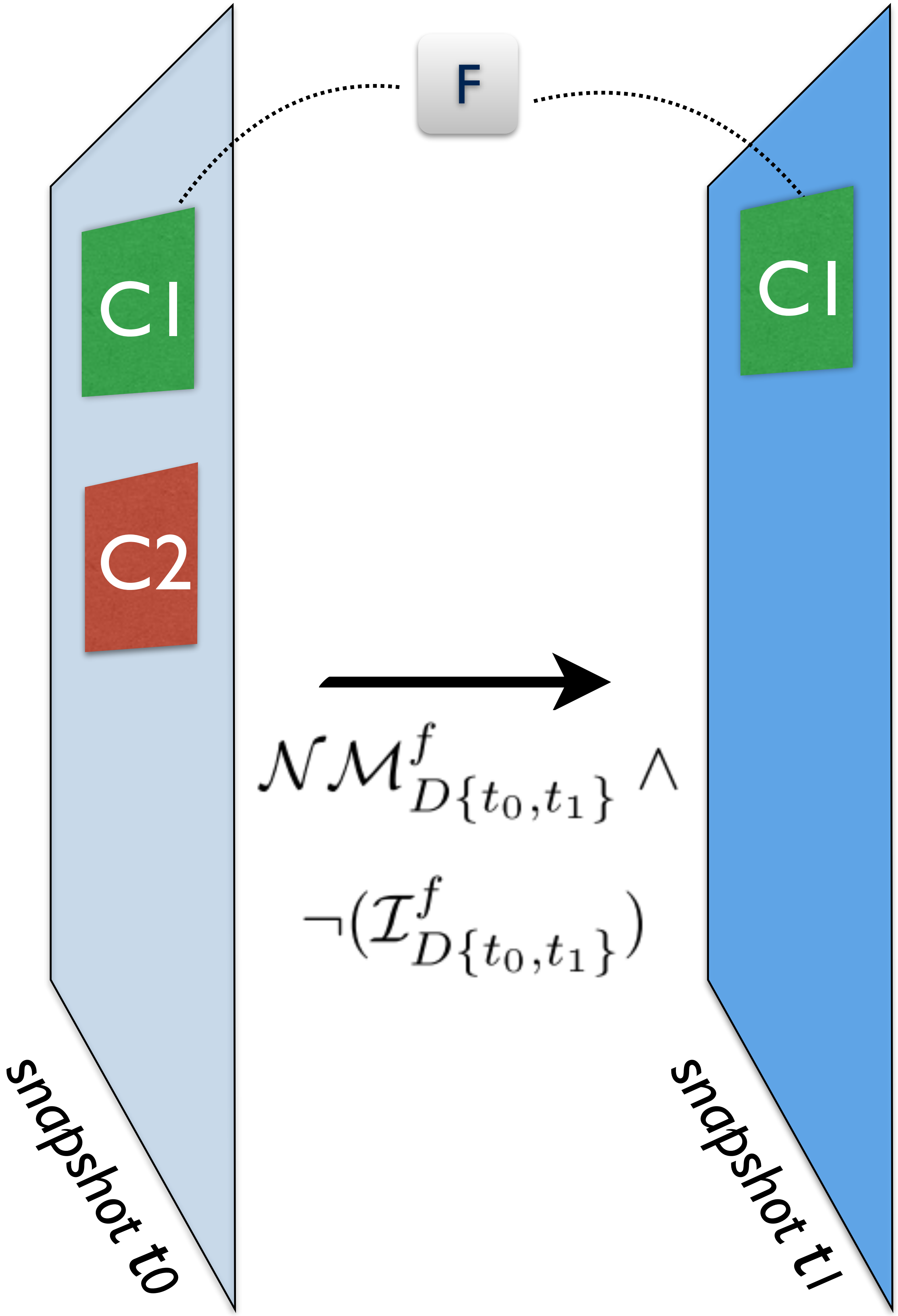


$$
\begin{aligned}
&(\mathcal{C}^{f}_{D\{t_0\}} - \mathcal{C}^{f}_{D\{t_1\}} = \emptyset) \quad \wedge\\
&(\mathcal{C}^{f}_{D\{t_0\}} \cap \mathcal{C}^{f}_{D\{t_1\}} \neq \emptyset) \quad \wedge\\
&(\mathcal{C}^{f}_{D\{t_1\}} - \mathcal{C}^{f}_{D\{t_0\}} = \emptyset)
\end{aligned}
$$

Weak Extinction



A feature disappears from *at least one* category in which it resided and does not migrate to any new ones (*WX*).

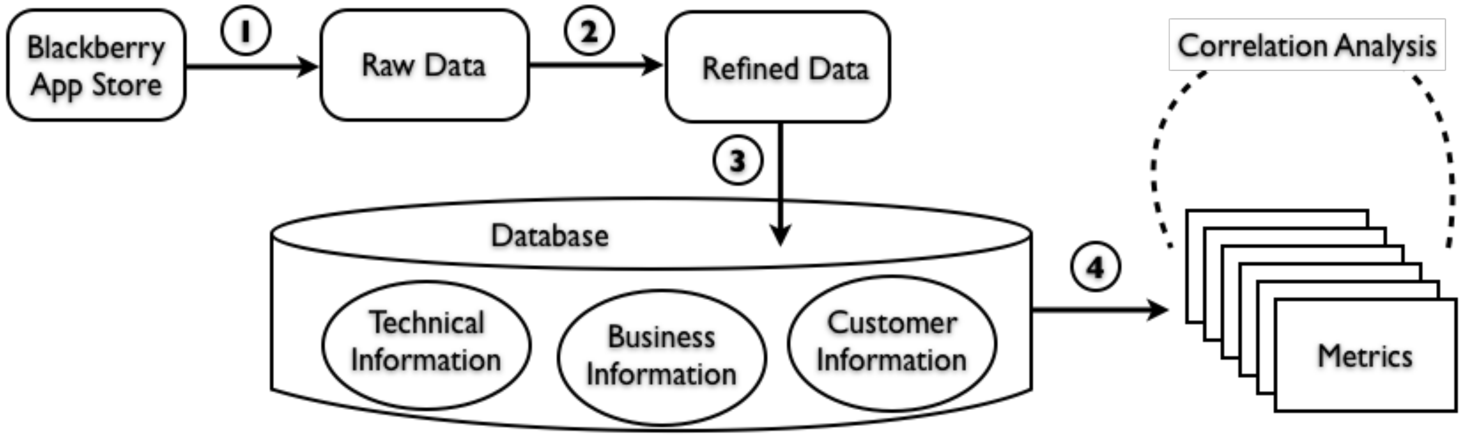


$$\mathcal{NM}^f_{D\{t_0,t_1\}} \wedge \neg(\mathcal{I}^f_{D\{t_0,t_1\}})$$

DATA SET

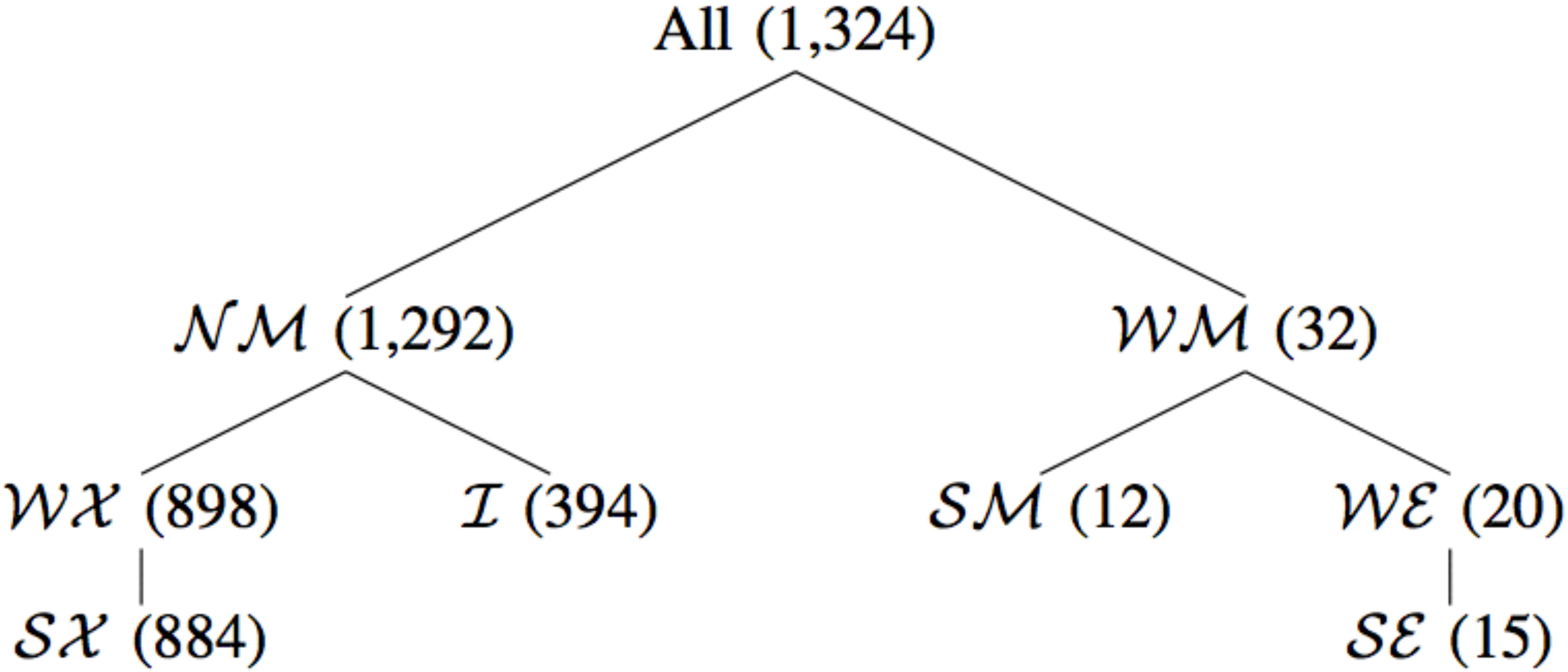
BlackBerry App World™

Week 3 and Week 36 in 2011

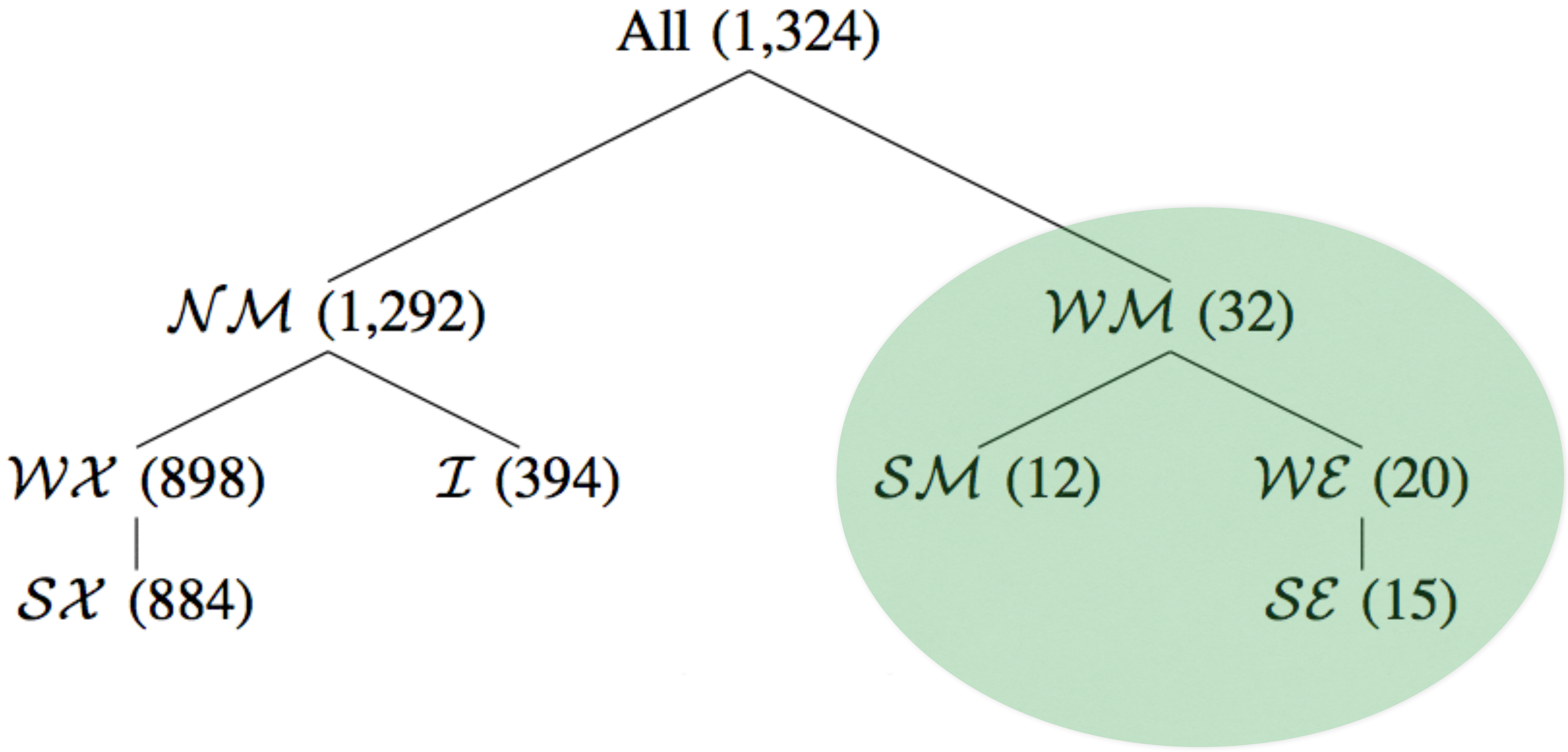


1,324 features

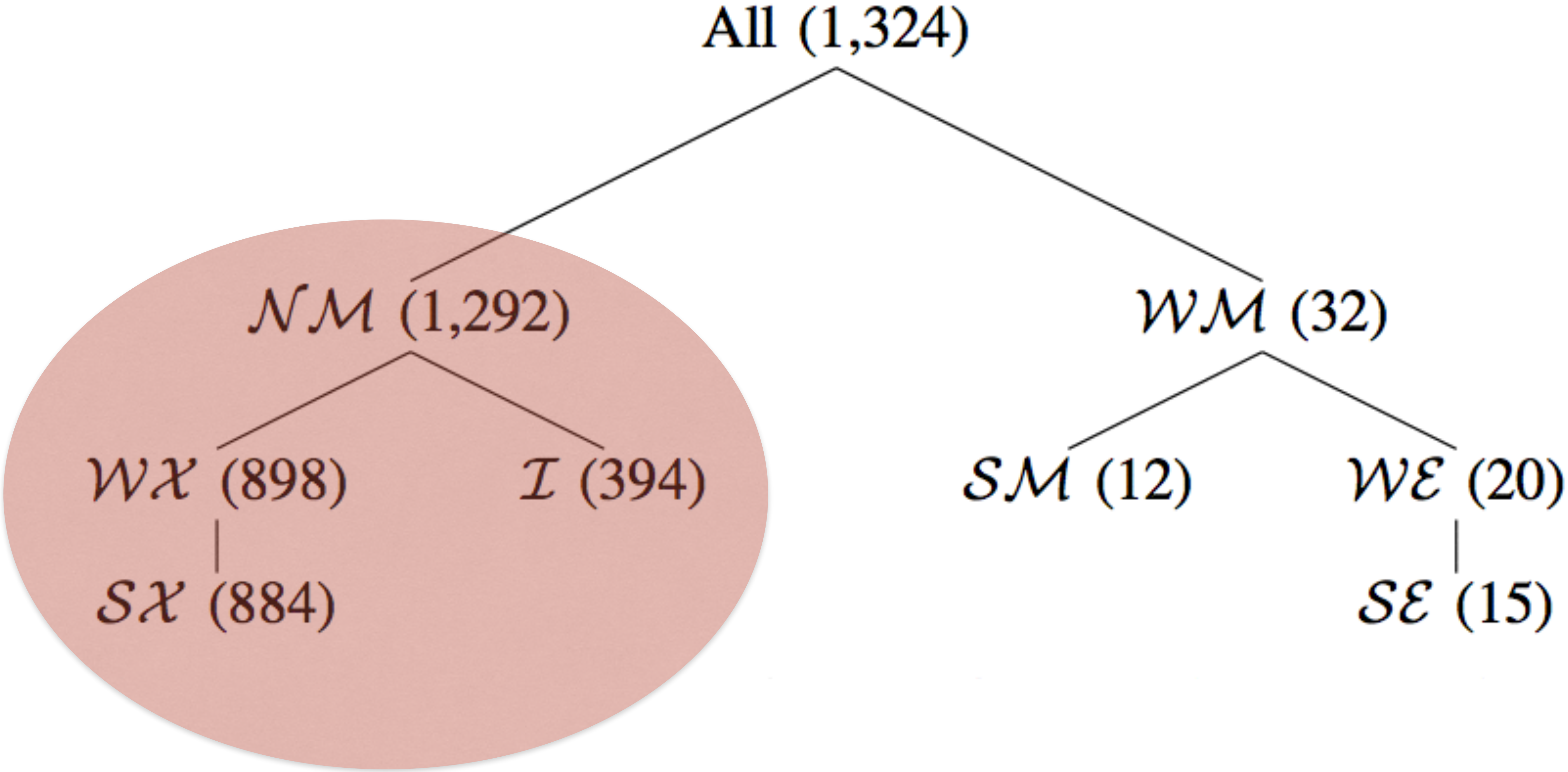
OBSERVED NUMBER OF FEATURES FOR EACH MIGRATORY BEHAVIOUR



OBSERVED NUMBER OF FEATURES FOR EACH MIGRATORY BEHAVIOUR



OBSERVED NUMBER OF FEATURES FOR EACH MIGRATORY BEHAVIOUR

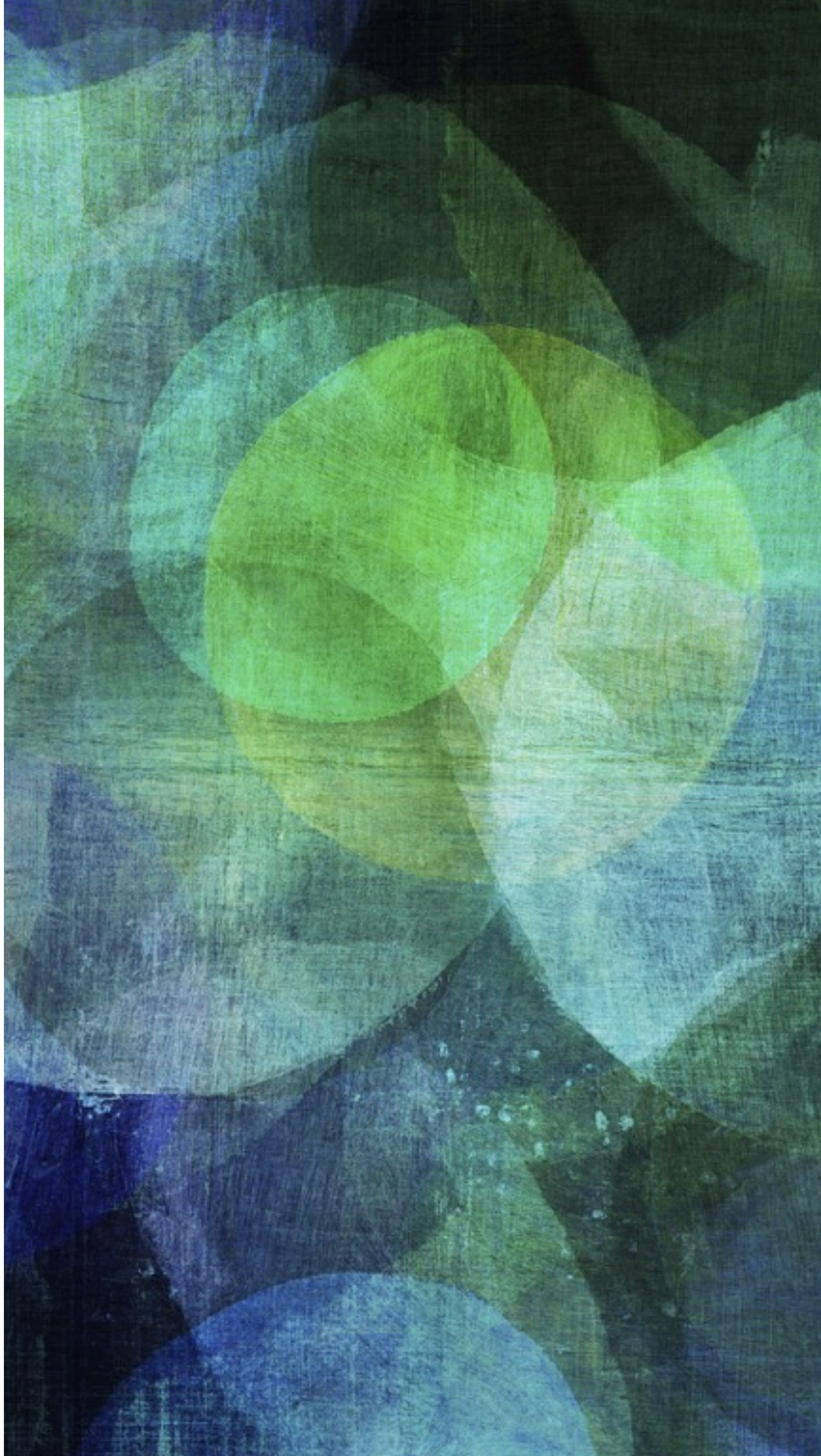


Strongly migratory features are cheaper and less popular

Intransitive features carry the highest monetary value; notably higher than either those features that migrate or those that die out.

APP CLUSTERING

Clustering Mobile Apps Based on Mined Textual Features (ESEM’16)



GOOD APP CATEGORISATION

User

More exposure to newly emerging apps

Developer

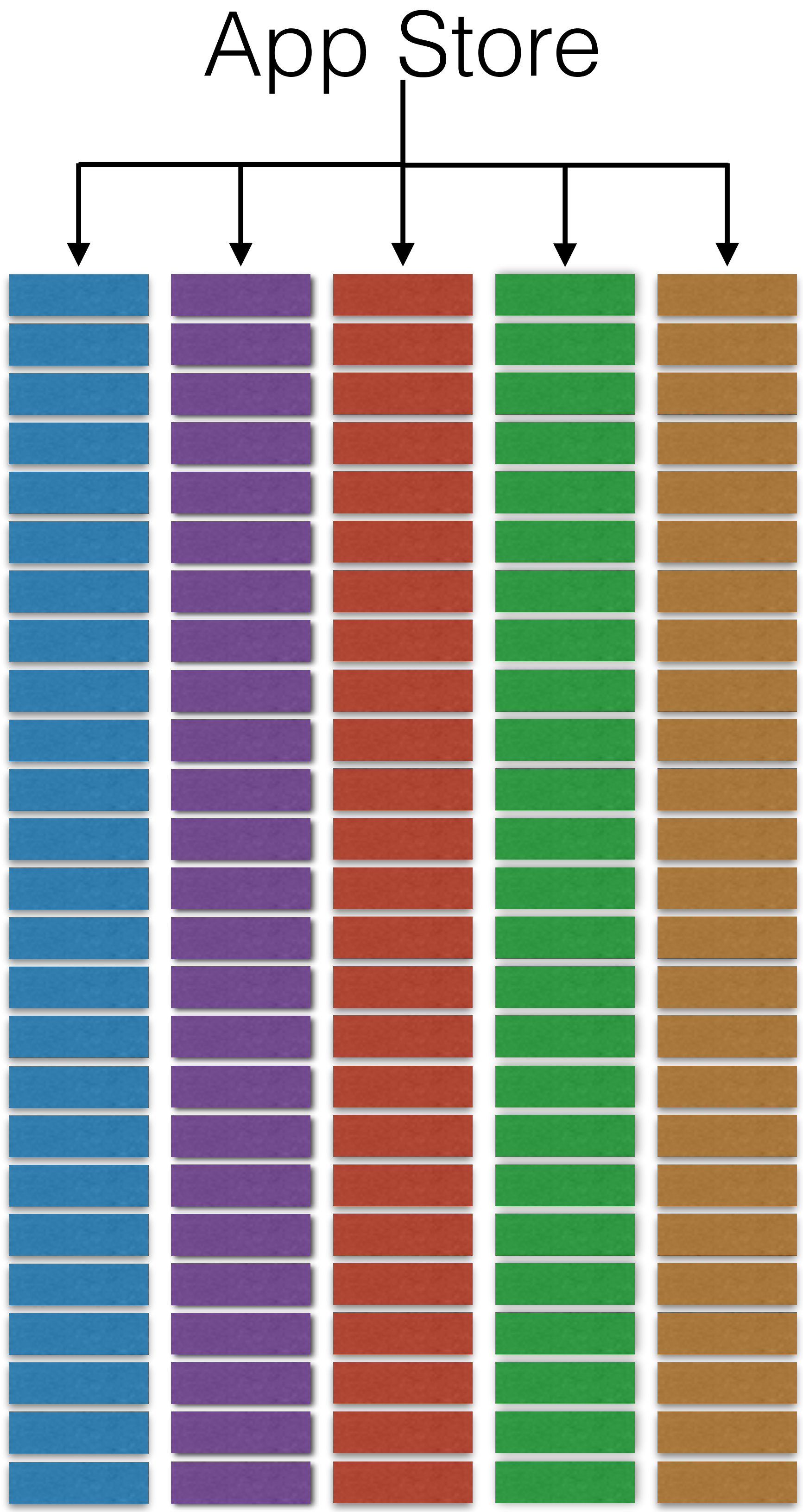
Locating desirable features and technical trends

App store owners

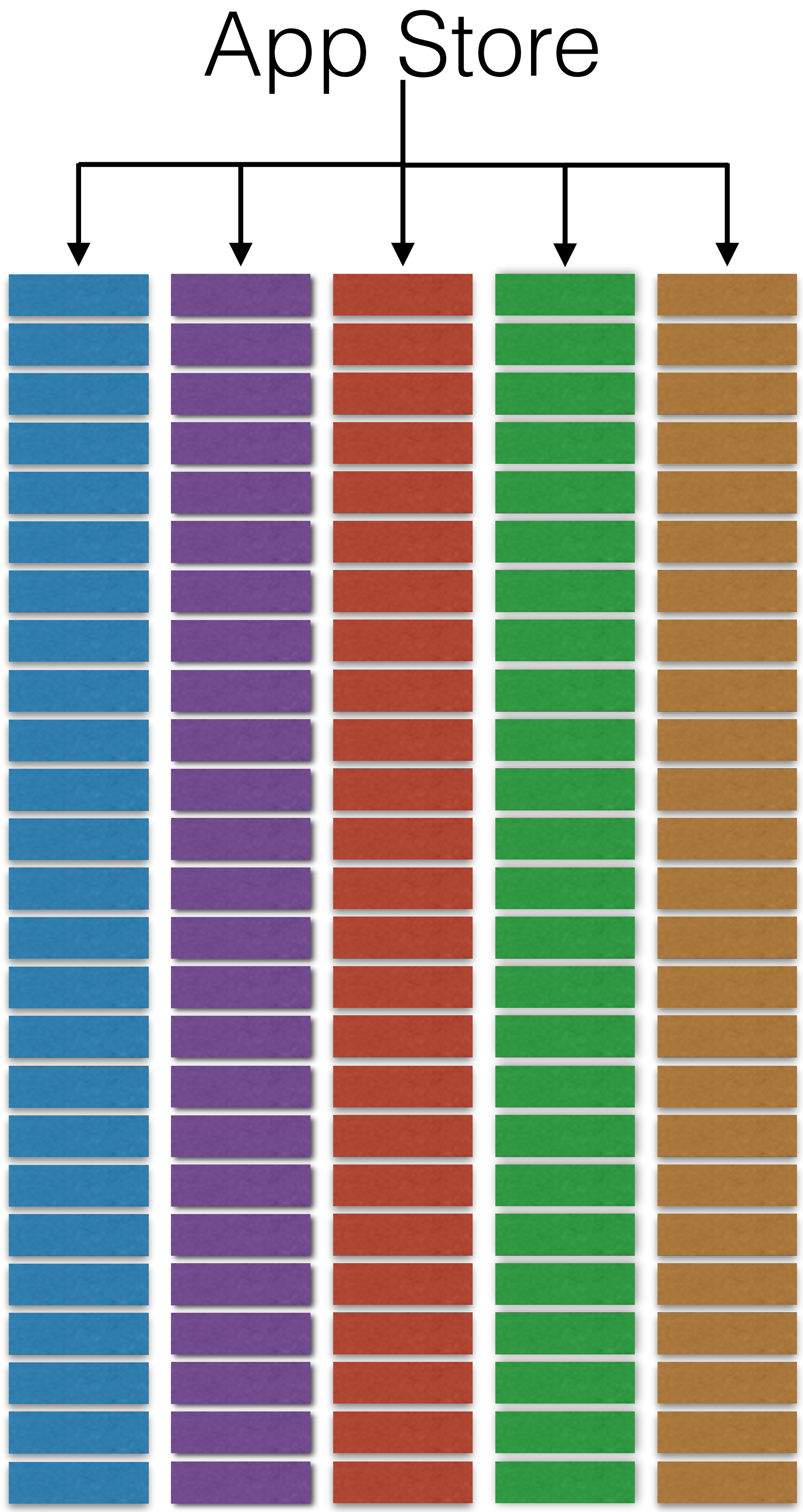
Detecting malicious apps and clones

APPS: HUGE PILES OF UNSORTED PRODUCTS

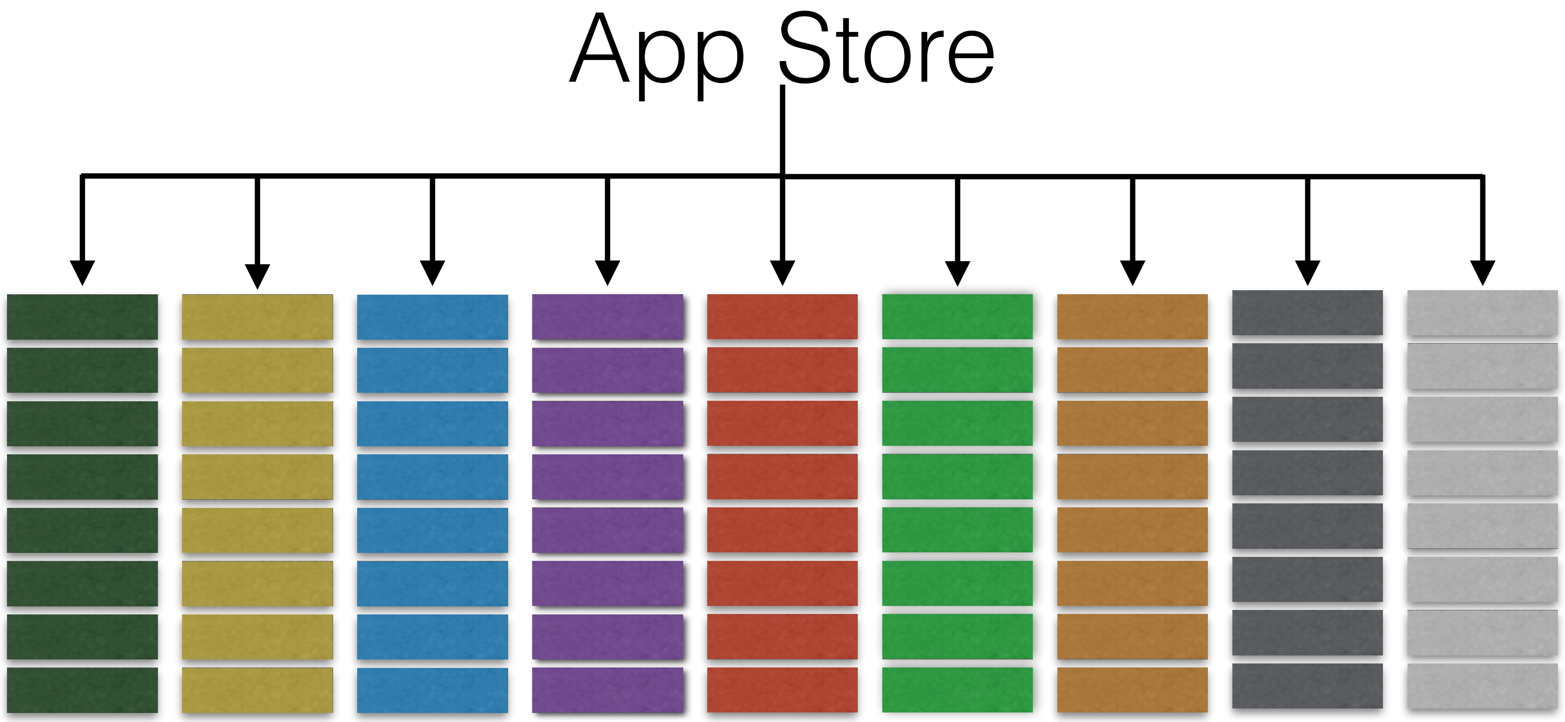
APPS: HUGE PILES OF UNSORTED PRODUCTS



APPS: HUGE PILES OF UNSORTED PRODUCTS



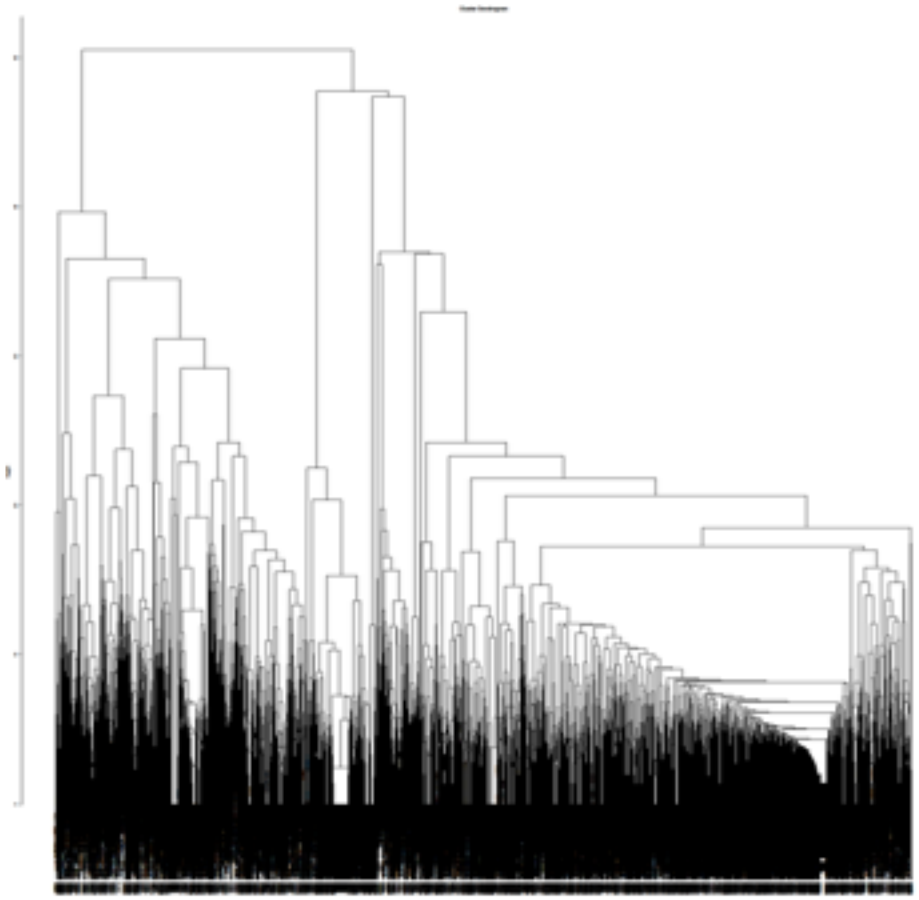
App Store



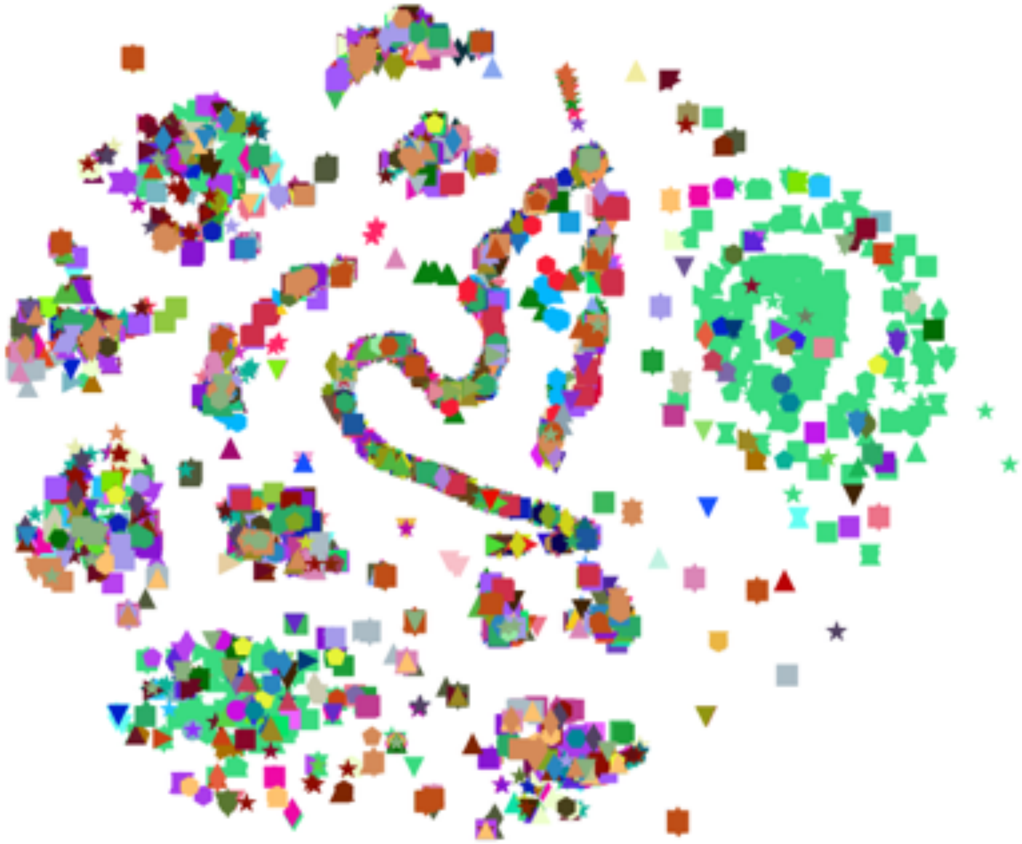
App Store

Feature Based

HIERARCHICAL CLUSTERING APPS



Agglomerative Hierarchical Clustering
Using Cosine Similarity

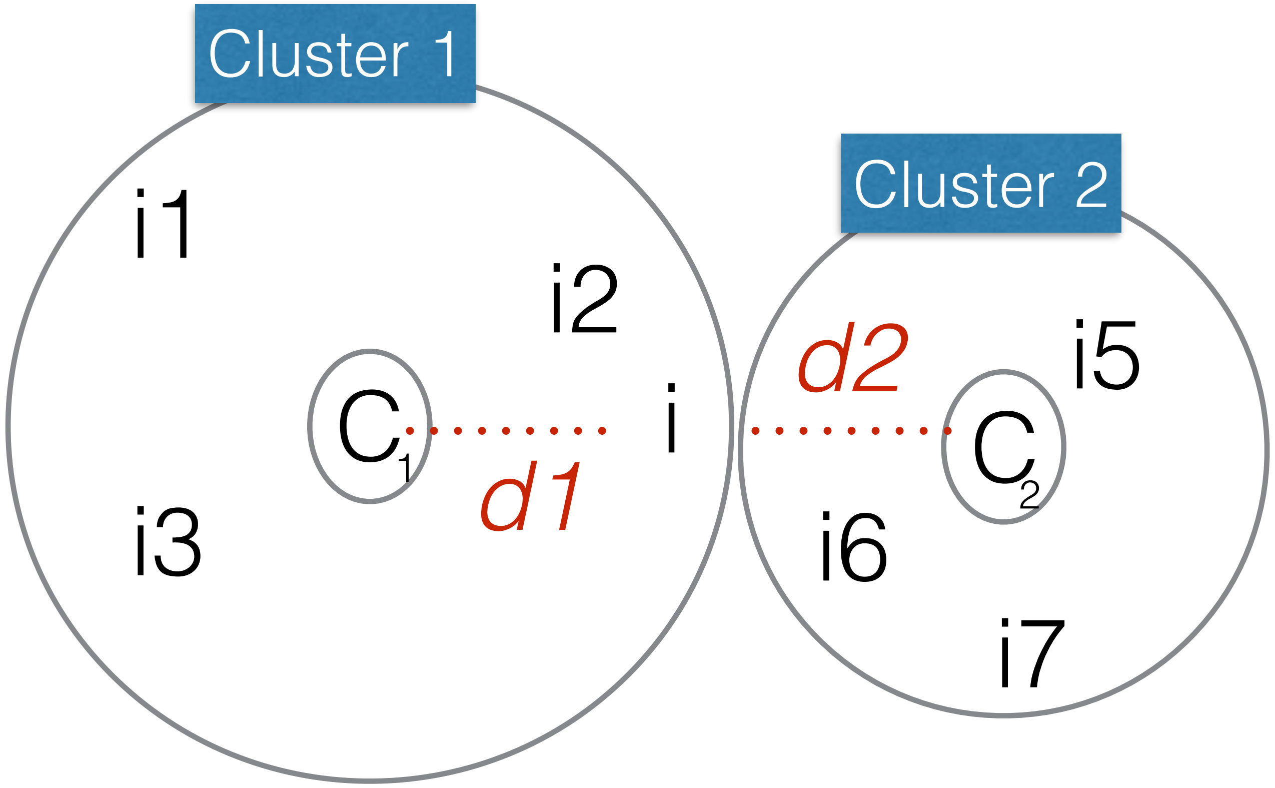


Plotted using t-SNE. Shape is original category
colour is assigned cluster
k = 368

THE SILHOUETTE SCORE

The silhouette of point *i* indicates how well it was classified

d1 = how far *i* is from its cluster
d2 = How far it is from closest cluster



$$\mathrm{sil}(i) = \frac{d2 - d1}{\max\{d1, d2\}}$$

ONLY TWO DEFAULT CATEGORY BOTH FARE BETTER IN TERMS OF SILHOUETTE SCORE

Category	Size	Avg. Sil.
Books	142	0
Business	813	-0.02
Education and Reference	1260	-0.04
Entertainment	1595	-0.03
Finance	588	0.02
Health and Fitness	506	-0.04
Music and Audio	1025	0.08
Navigation and Travel	953	0
News and Magazines	1474	0.21
Photo and Video	753	0.03
Productivity	974	-0.01
Shopping	144	-0.01
Social	668	-0.02
Sports	439	0.05
Utilities	2832	-0.02
Weather	92	0.15

BlackBerry App World

Category	Size	Avg. Sil.
Books and Reference	34	0.002
Business	23	0.031
Communication	65	0.017
Education	90	-0.005
Entertainment	164	-0.041
Family	79	0.012
Finance	20	0.218
Games	2002	-0.016
Health and Fitness	84	0.046
Lifestyle	59	-0.052
Media and Video	40	0.019
Music and Audio	98	0.051
News and Magazines	18	0.108
Personalization	121	0.008
Photography	89	0.083
Productivity	99	-0.012
Shopping	42	0.009
Sports	213	-0.015
Social	56	0.047
Tools	144	-0.018
Transport	33	0.048
Travel and Local	69	0.002
Weather	31	0.223

Google Play

HIERARCHICAL CLUSTERING IMPROVED SILHOUETTE SCORE

Category	Granularity	Silhouette
Books	76	0.58
Business	397	0.33
Education and Reference	706	0.46
Entertainment	816	0.54
Finance	325	0.32
Health and Fitness	248	0.37
Music and Audio	473	0.57
Navigation and Travel	480	0.34
News and Magazines	662	0.62
Photo and Video	401	0.36
Productivity	460	0.26
Shopping	83	0.34
Social	379	0.31
Sports	179	0.49
Utilities	1974	0.34
Weather	67	0.32

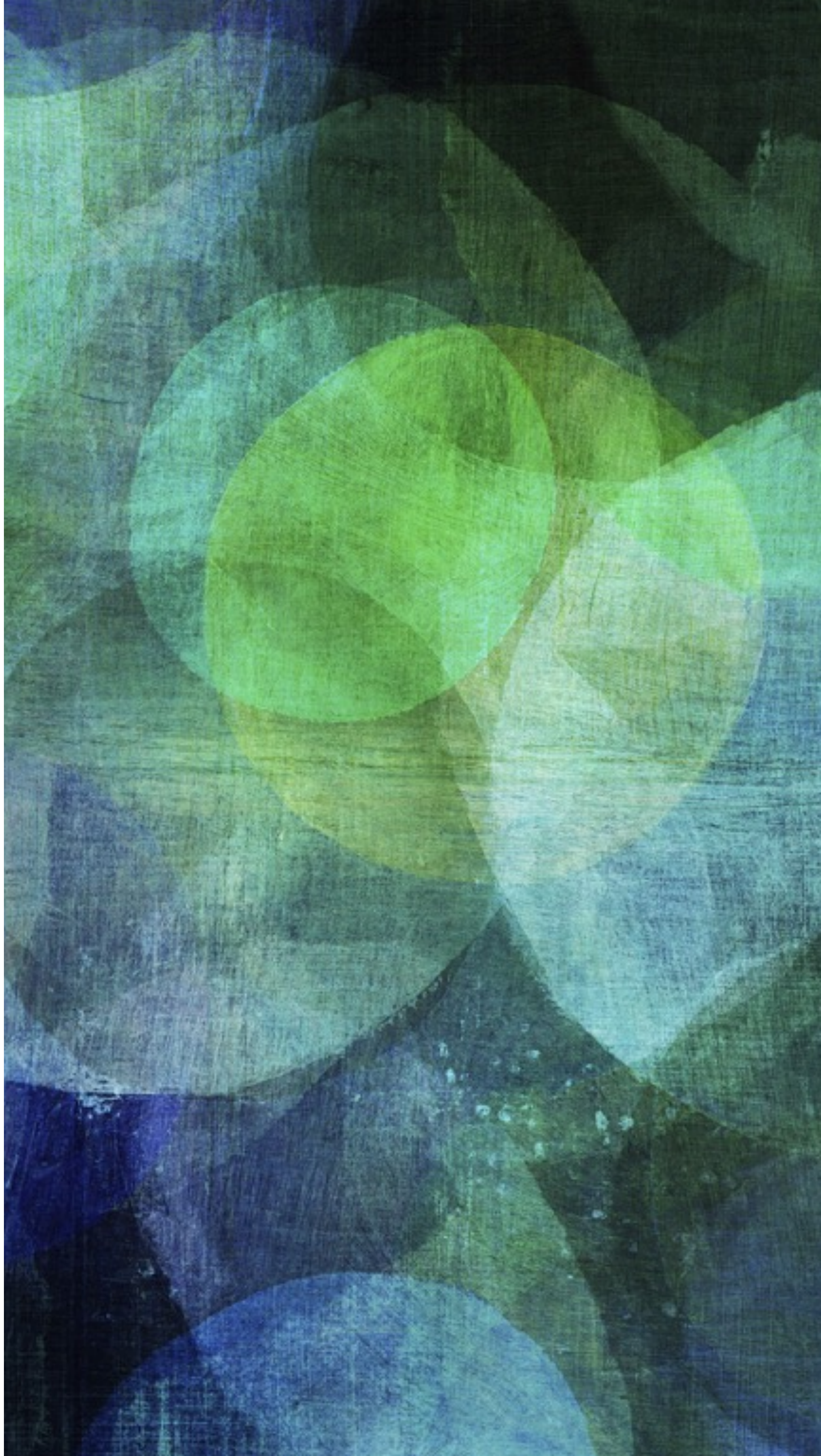
BlackBerry
App World

Category	Granularity	Silhouette
Books and Reference	20	0.2
Business	17	0.35
Communication	26	0.17
Education	58	0.27
Entertainment	70	0.22
Family	46	0.19
Finance	11	0.2
Games	964	0.21
Health and Fitness	46	0.23
Lifestyle	32	0.2
Media and Video	22	0.24
Music and Audio	57	0.2
News & Magazines	4	0.23
Personalization	53	0.32
Photography	53	0.19
Productivity	58	0.19
Shopping	14	0.17
Sports	120	0.19
Social	28	0.15
Tools	66	0.23
Transport	26	0.37
Travel and Local	37	0.2
Weather	24	0.24

Google Play

PREDICTIVE MODELLING

Mining App Stores: Extracting Technical, Business and Customer Rating Information for Analysis and Prediction

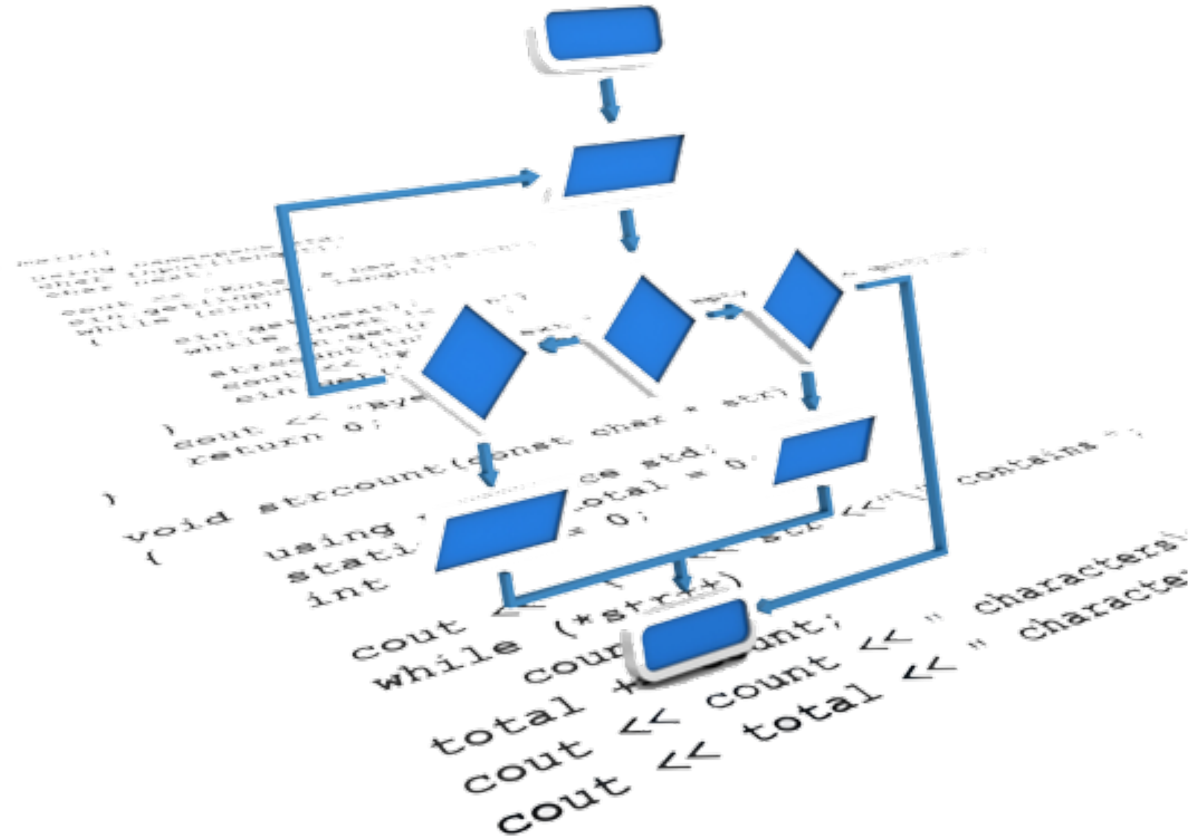


MOTIVATION

In 2012 more than 60% of the apps in the App Store **have never been downloaded,** even once

Source: analytical firm Adeven, 2012

MOTIVATION



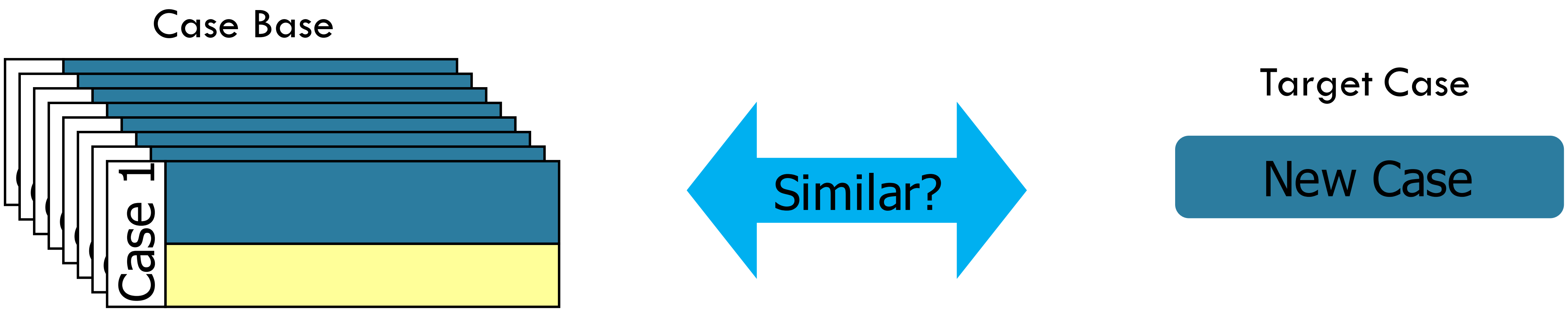
developers

Target Market
Product
Place
Price
Promotion

marketing experts

CASE BASED PREDICTION

- □ AI approach where knowledge of similar past cases is used to solve new cases
  - Compare new problem to each case
  - Select most similar



Case Base

Case 1

Similar?

Target Case

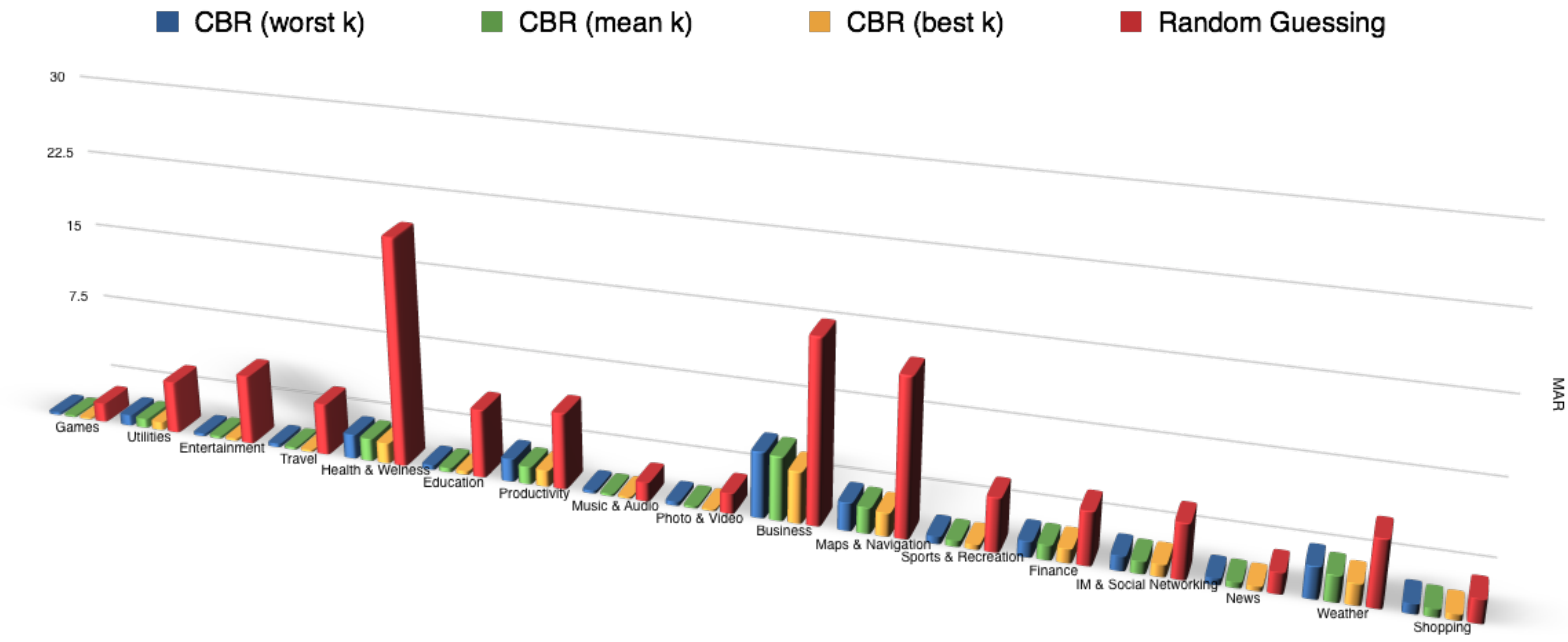
New Case

apps available in the store characterized in terms of a set of features

similarity of the features' vector

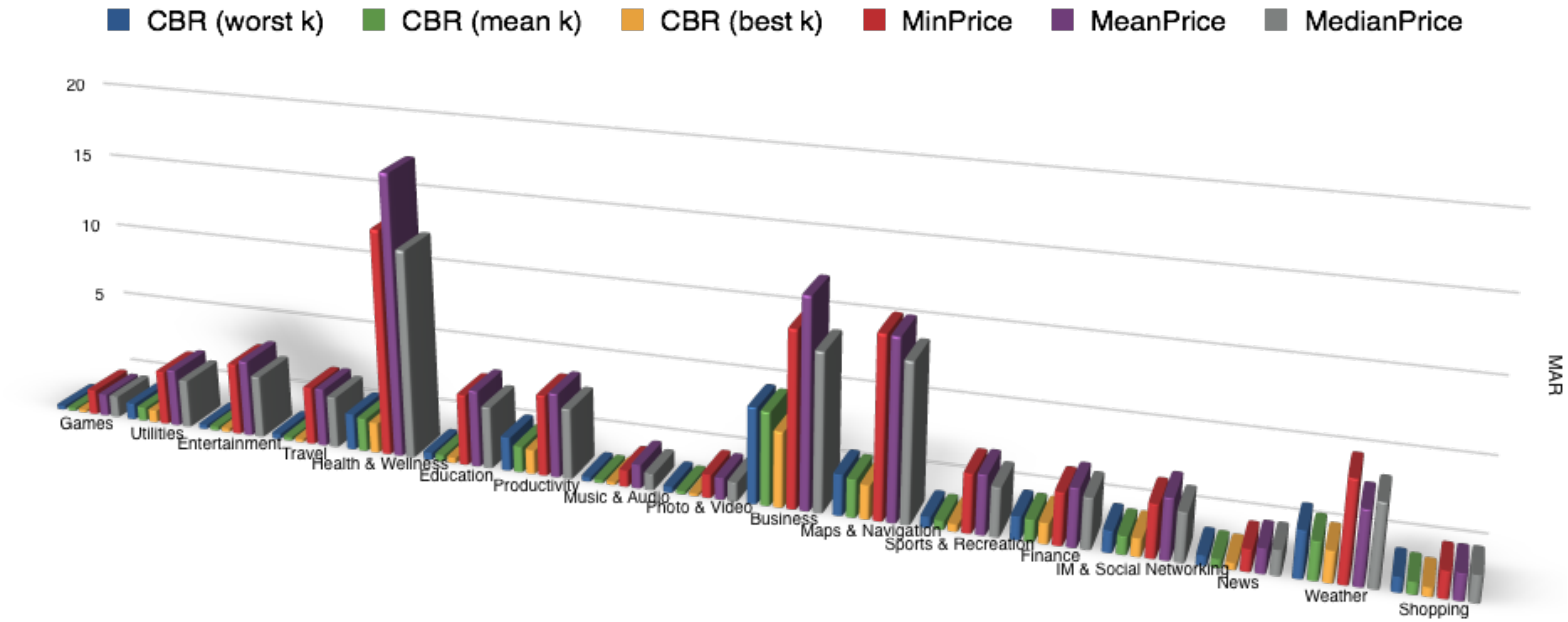
new app characterized by *n* features

PREDICT PRICE (VS RANDOM GUESSING)



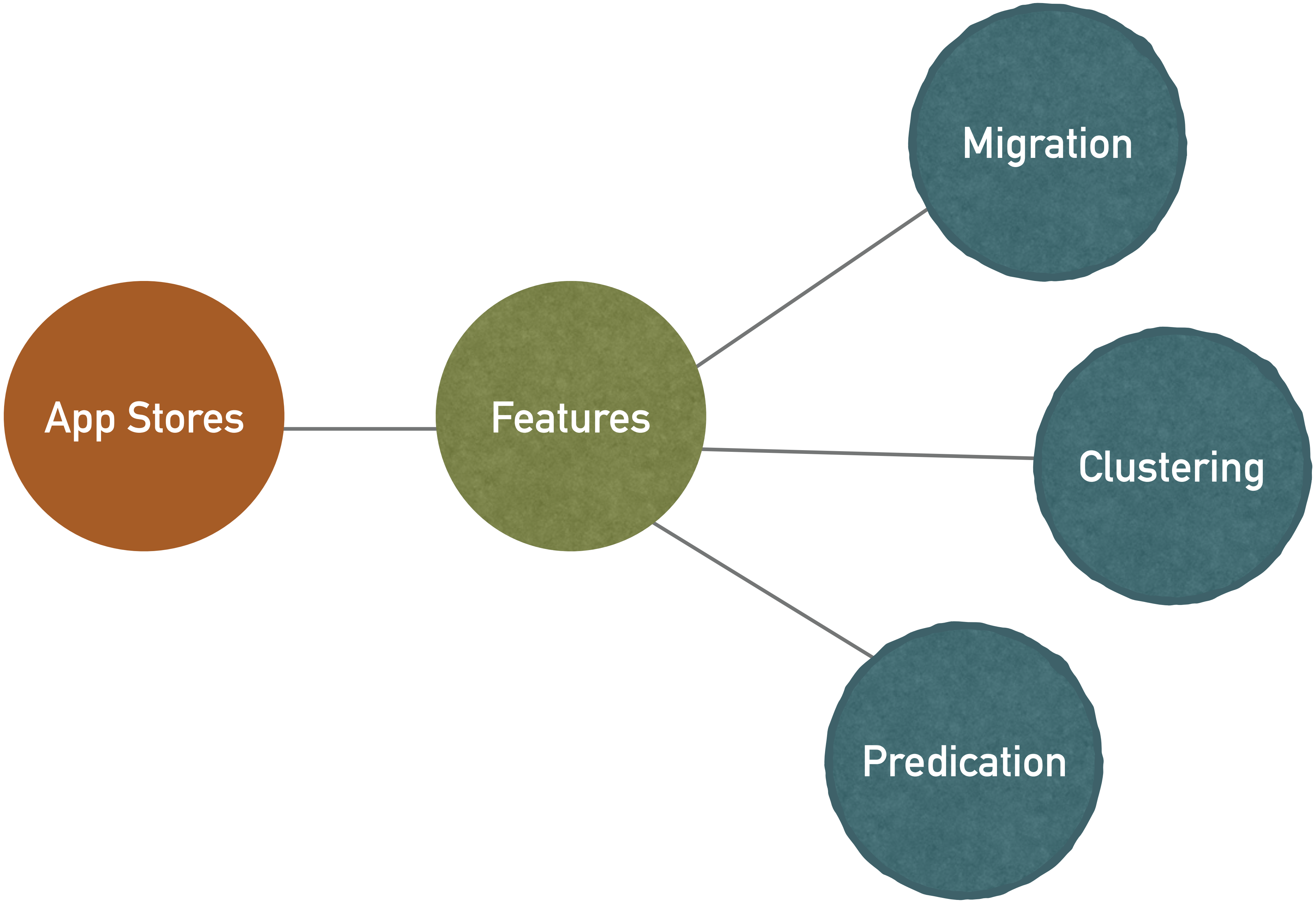
CBR significantly better than RG with high effect size

PREDICT PRICE (VS MEDIAN PRICE)



CBR (worst, mean and best k) achieved the lowest MAR values on all the categories

FEATURE ANALYSIS



“

This is not software engineering …

Apps, … just GUI interface…

- The third reviewer

A Survey of App Store Analysis for Software Engineering

William Martin, Federica Sarro, Yue Jia, Yuanyuan Zhang and Mark Harman

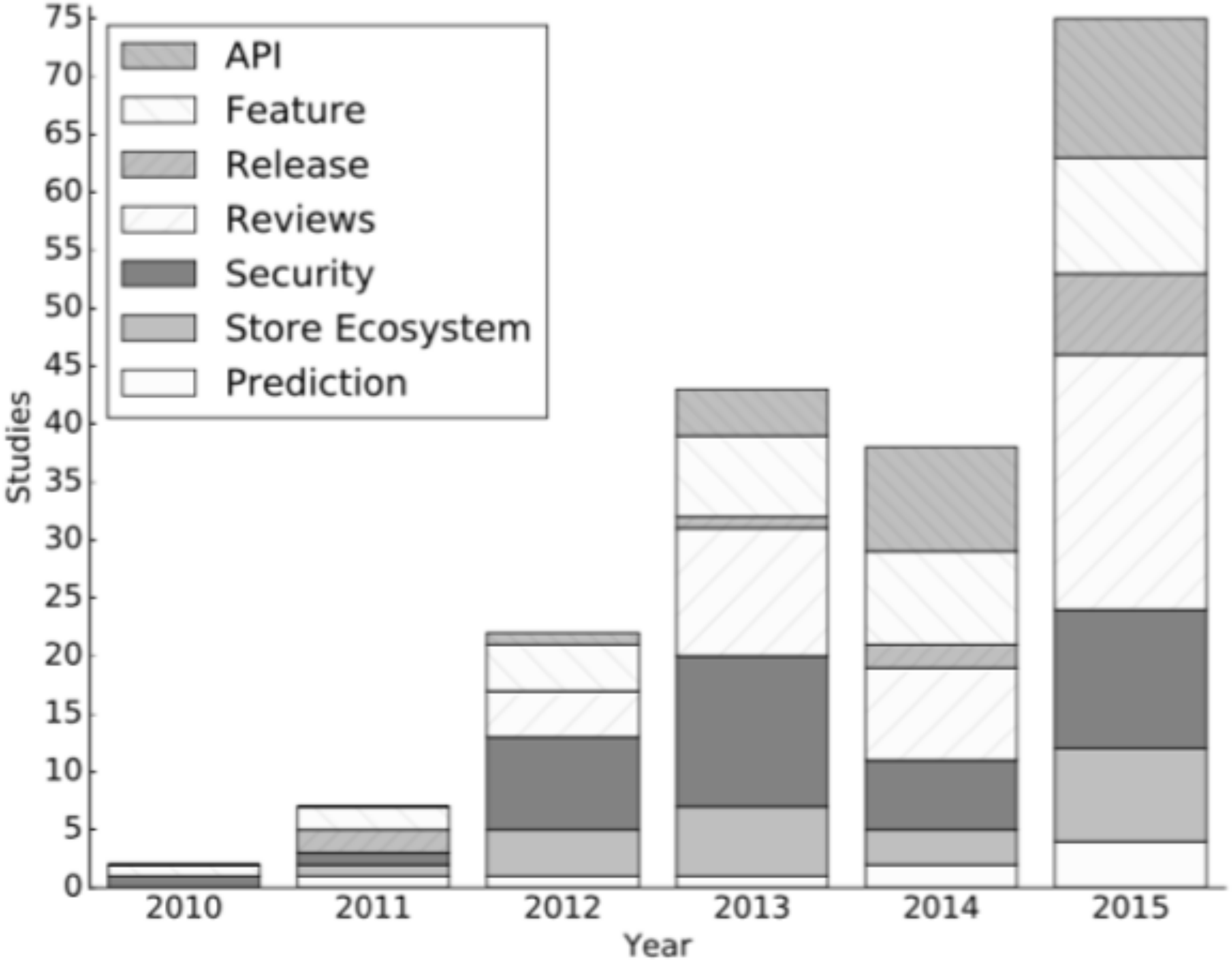


Fig. 3. **Histogram of sub-field trends** showing the period from 2010 to November 27, 2015.

APP STORE ANALYSIS

- ➤ Feature Analysis
- ➤ Clustering Mobile Apps
- ➤ Predicting Price and Rating
- ➤ Feature Migration
- ➤ Causal Impact Analysis
- ➤ Sampling Bias Issues
- ➤ App Developer Interviews
- ➤ Android Test Data Generation
- ➤ Mobile Energy Optimisation