
DATA INTEGRATION
VIDEO GAMES
TEAM 4
PLAY

LEVEL

1. INTRODUCTION

2. DATA TRANSLATION

3. IDENTITY RESOLUTION

4. DATA FUSION

Video Games' market value

Global Revenue **UP** to **USD 138 B**!

LEVEL

1. INTRODUCTION

2. DATA TRANSLATION

3. IDENTITY RESOLUTION

4. DATA FUSION

Data

- Data scraped from **4 video game rating websites**

Data

- Data scraped from **4 video game rating websites**
- Date range between 2010 – 2018

Class	Data Set	Source	Format	# of Entity	# of Entity (dupe. removed)
Game	vgchartz.csv	VGChartz.com	CSV	17,957	17,929
	igdb.json	IGDB.com	JSON	45,611	35,727
	metacritic.csv	Metacritic.com	CSV	7,370	7,370
Developer	vgchartz.csv	VGCHartz.com	CSV	3,103	3,103
	igdb.json	IGDB.com	JSON	7,827	7,827
	developer.csv	Wikipedia	CSV	569	569

Integrated Schema

- Two main classes
  - **Game** and **Developer**
- Game Class
  - **13** attributes
- Developer Class
  - **7** attributes

How the integrated schema is built

- Define the classes
- Manually identify correspondences between the attributes in our input datasets
- Assign **respective attributes** to our game and developer class
- Build the **integrated schema** by resolving conflicts by successively changing the structure of the individual schemata
- Do **schema matching** for each of our datasets and to create **unique IDs** for the games and developers class

How the integrated schema is built

- Several transformations used to transform the input data.
- Reported **userScore** and **expertScore** are transformed to a common scale ranging from **0** to **100**
- **releaseDate** attribute is normalized in all the datasets in order to have the same date format.
- For **VGChartz**, the **missing values** for **userScore** and **expertScore** are set to **-1**, since the data is of type integer.
- the **missing values** for **totalSales**, **NASales**, **PALSales**, **JapanSales** and **otherSales** are also set to **-1**
- The output of each mapping is an XML file which is used as input for the Identity Resolution.

LEVEL

1. INTRODUCTION
2. DATA TRANSLATION
3. IDENTITY RESOLUTION
4. DATA FUSION

TEAM 4

Gold Standard

- **Start with initial size of 450 records:**
  - add 100 matching games (ca. 20%)
  - add 250 non-matching games (ca. 50%)
  - add 100 ‘interesting’ corner cases (ca 30%)
- **Corner Cases:**

  Mega Man X Legacy Collection 1 <> Mega Man X Legacy Collection 2

  Need for Speed == Need for Speed (2015)
- **Subsequently update the gold standard with further corner cases!**

  → Final Size ~630 record pairs (<**Metacritic – VGChartz**>)

Data Preprocessing

- Data Preprocessing:
  - Uppercasing, removing punctuation
  - Deduplication
  - **Stop - word** removal
  - Normalization of content for **maturityRating** and **Platform**
    (**PlayStation4**, **Ps 4** → **PS4**)
  - Remove '**Read the review**' stub

Comparators and Blocking Keys

- **Comparators**
  - **Date:**
    1Year-, *2Year*, and *3YearComparator*
  - **String:**
    Equal, Jaccard, Levenshtein, Damerau, JaroWinkler, JaroWinklerTfIdf, MongeElkan, MongeElkanTfIdf (Java SecondString API)
- **Blocking Keys:**
  - **Platform** + **releaseDate**
  - **Platform** + **releaseYear**
  - **Platform** + first two letters of **nameOfGame**

➡ All blocking strategies yield a Reduction ration of > **99%** !

Machine Learning Results

< IGDB - VgChartz >

Classifier	Matching Rule	
Baseline	MachineLearning SimpleLogistic: Equal	
RandomForest (REPTree) {"-M", "3.0", "-S", "42"}	*GameDateComparator1Years, GameDateComparator2Years, GameDateComparator3Years, GameDateComparatorWeightedDate, GameNameComparatorDamerau, GameNameComparatorEqual, GameNameComparatorJaccard,*	*GameNameComparatorJaroWinkler, GameNameComparatorLevenshtein, GameNameComparatorSoftTfIdf, GameNameComparatorJaroWinklerTfIdf, GamePublisherComparatorEqual, GamePublisherComparatorJaccard, GamePublisherComparatorLevenshtein,*

Classifier	Blocker	Threshold	P	R	F1	# Corr	Run Time	Reduction Ratio
Baseline	*GameBlockingKeyByPlatformDateGenerator,*	0.5	1.00	0.39	0.56	3,862	2 sec	0.99
RandomForest (REPTree) {"-M", "3.0", "-S", "42"}	*GameBlockingKeyByPlatformNameGenerator,*	0.8	0.96	0.73	0.83	8,278	3 min 10 sec	0.99

Machine Learning Results

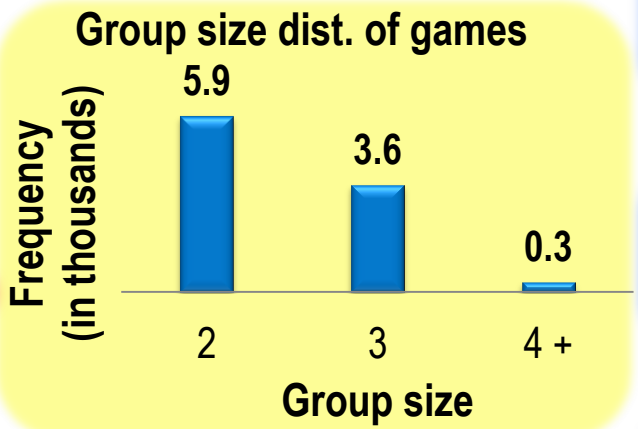
< MetaCritics - VgChartz >

Classifier	Matching Rule	
Baseline	MachineLearning SimpleLogistic: Equal	
RandomForest (REPTree) {"-I", "300", "-S", "42", "-K", "7"}	*GameDateComparator1Years, GameDateComparator2Years, GameDateComparatorWeightedDate, GameNameComparatorEqual,*	*GameNameComparatorJaccard, GameNameComparatorMongeElkan, GameNameComparatorMongeElkanTfIdf, BackwardSelection(false),*

Classifier	Blocker	Threshold	P	R	F1	# Corr	Run Time	Reduction Ratio
Baseline	*GameBlockingKeyByPlatformDateGenerator,*	0.5	1.00	0.80	0.89	3,240	1 sec	0.99
RandomForest (REPTree) {"-I", "300", "-S", "42", "-K", "7"}	*GameBlockingKeyByPlatformNameGenerator,*	0.8	0.93	0.89	0.91	4,897	20 sec	0.99

Evaluation and Error Analysis

- **Evaluation**
  - Improvement in **F1-Score**
  - Improvement in **# of correspondences**
- **Error Analysis**
  - Problems with **separate**, **versions of same game**
    (e.g. Limited Editions, Gold Editions)
  - Problems with **Add-Ons** and **Extension Packs**
  - Hatsune Miku: Project Diva Future Tone <>
    Hatsune Miku: Project Diva Future Tone - Colorful Tone



LEVEL

1. INTRODUCTION

2. DATA TRANSLATION

3. IDENTITY RESOLUTION

▶ 4. DATA FUSION

TEAM 4

Fusion Strategy

- Attribute-specific conflict resolution methods for all the attributes
- No obvious ranking or hierarchy in terms of **data provenance**
- No reliable information about their last up-date and do not allow for chronological ordering
- We implement and test a variety of different *fusers* and *evaluation rules*

Fusion STEP 1

- Attribute-specific conflict resolution methods for all the attributes

Attribute	Fuser	Evaluation rule
nameOfGame	*simple shortestString*	*Equal*
publisher	*simple shortestString*	
genre	*simple shortestString*	
ReleaseDate	*Voting*	
Platform	*favourSource*	
maturityRating	*favourSource*	
nameOfDeveloper	*simple shortestString*	

- Highest provenance: **Metacritic**

➡ This strategy leads to an overall **accuracy** of **0.53**

Fusion STEP 2

- **Finding 1**

  For **nameOfGame**: difference between **fused values** and **true values**

nameOfGame	publisher	platform	genre	releaseDate
IGDB White Knight Chronicles	Sony Computer Entertainment, Inc. (SCEI)	PS3	Role-playing (RPG)	2010-02-26
MetaCritic White Knight Chronicles International Edition	SCEA	PS3	Role-Playing, Action RPG	2010-02-05
VGChartz White Knight Chronicles: International Edition Read the review	Sony Computer Entertainment	PS3	Strategy	2010-02-02

- Changed **fusion method** from *simple shortestString* to *voting*, and **evaluation rule** from **Equal** to **tokenized Jaccard similarity,** which is case insensitive and ignores punctuation.

Fusion STEP 3

- **Finding 2**

  For **genre** and **publisher**: **Attributes** are most correctly captured by **IGDB**

	nameOfGame	publisher	platform	genre	releaseDate
IGDB	White Knight Chronicles	Sony Computer Entertainment, Inc. (SCEI)	PS3	Role-playing (RPG)	2010-02-26
MetaCritic	White Knight Chronicles International Edition	SCEA	PS3	Role-Playing, Action RPG	2010-02-05
VGChartz	White Knight Chronicles: International Edition Read the review	Sony Computer Entertainment	PS3	Strategy	2010-02-02

- Changed **highest provenance**: **IGDB**
- Changes **evaluation rule** for **genre** and **publisher** to use *FavourSource* .

➡ **Overall accuracy** increases to **0.8**

➡ **Attribute-specific accuracy** for **nameOfGame**, **publisher**, and **genre**, increases to **0.95**, **0.7** and **0.60**, respectively

Fusion STEP 4

- **Finding 3**

  Differences in **releaseDate**

	nameOfGame	publisher	platform	genre	releaseDate
IGDB	White Knight Chronicles	Sony Computer Entertainment, Inc. (SCEI)	PS3	Role-playing (RPG)	2010-02-26
MetaCritic	White Knight Chronicles International Edition	SCEA	PS3	Role-Playing, Action RPG	2010-02-05
VGChartz	White Knight Chronicles: International Edition Read the review	Sony Computer Entertainment	PS3	Strategy	2010-02-02

- Applying *evaluation rule* to allow for **a tolerance of 30 days**
  (**Start date** & **End data**)

Fusion STEPS

- **Finding 4**

  Differences in **genre** and **publisher**

	nameOfGame	publisher	platform	genre	releaseDate
IGDB	White Knight Chronicles	Sony Computer Entertainment, Inc. (SCEI)	PS3	Role-playing (RPG)	2010-02-26
MetaCritic	White Knight Chronicles International Edition	SCEA	PS3	Role-Playing, Action RPG	2010-02-05
VGChartz	White Knight Chronicles: International Edition Read the review	Sony Computer Entertainment	PS3	Strategy	2010-02-02

- Applying evaluation of **genre** and **publisher** to tokenized *Jaccard distance* with a threshold of **0.50**

➔ The overall **accuracy** increases to **0.88**

Fusion strategy (updated)

Attribute	Fuser	Evaluation rule
nameOfGame	simple shortestString ➡ Voting	Equal ➡ tokenized Jaccard similarity (threshold = 1)
publisher	simple shortestString ➡ FavourSource	Equal ➡ tokenized Jaccard similarity (threshold = 0.50)
genre	simple shortestString ➡ FavourSource	Equal ➡ tokenized Jaccard similarity (threshold = 0.50)
releaseDate	Voting	Equal ➡ + allow for a tolerance of 30 days (Start date & End data)
Platform	favourSource	Equal
maturityRating	favourSource	Equal
nameOfDeveloper	simple shortestString ➡ Voting	Equal ➡ tokenized Jaccard similarity (threshold = 0.66 by checking group records)

Accuracy and Density

Attribute	Input				Output	
	Density **IGDB**	Density **MetaCritic**	Density **VGChartz**	Overall Consistency	Overall Accuracy	Overall Density
nameOfGame	1.00	1.00	1.00	0.95	0.95	1.00
platform	1.00	1.00	1.00	1.00	1.00	1.00
maturityRating	0.31	0.84	0.00	0.97	0.95	0.76
releaseDate	1.00	1.00	1.00	0.82	0.85	1.00
genre	0.92	0.67	1.00	0.18	0.7	1.00
publisher	0.61	1.00	1.00	0.78	0.8	1.00
Avg.	0.81	0.92	0.83	0.78	0.88	0.96
nameOfDeveloper	1.00	1.00	1.00	0.84	0.90	1.00

(For nameOfDeveloper, **MetaCritic** is replaced with **developer**)

Example Result

	nameOfGame	publisher	platform	genre	releaseDate
IGDB	White Knight Chronicles	Sony Computer Entertainment, Inc. (SCEI)	PS3	Role-playing (RPG)	2010-02-26
MetaCritic	White Knight Chronicles International Edition	SCEA	PS3	Role-Playing, Action RPG	2010-02-02
VGChartz	White Knight Chronicles: International Edition Read the review	Sony Computer Entertainment	PS3	Strategy	2010-02-02

Fused

nameOfGame	publisher	platform	genre	releaseDate
White Knight Chronicles: International Edition	Sony Computer Entertainment	PS3	Role-playing (RPG)	2010-02-26

Provenance

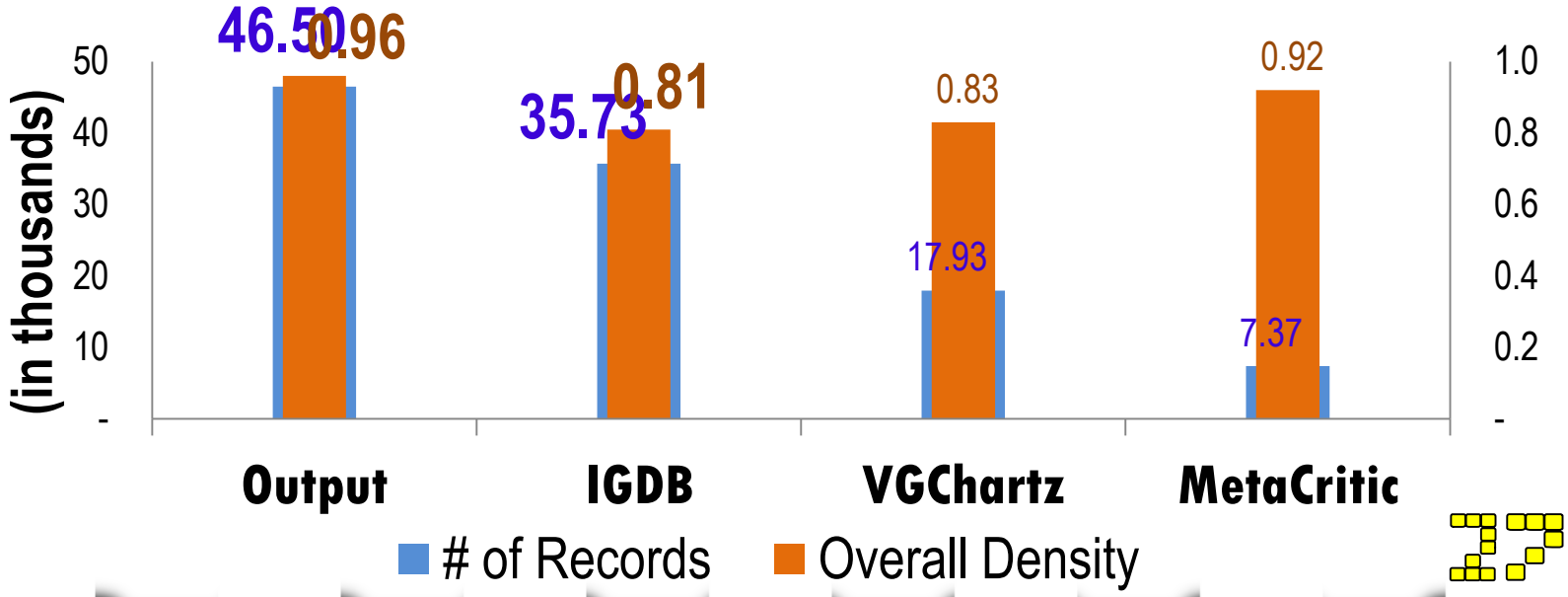
vg_117269 + igdb_game_141052	vg_117269	igdb_game_141052	igdb_game_141052	vg_117269 + mc_102683

Summary

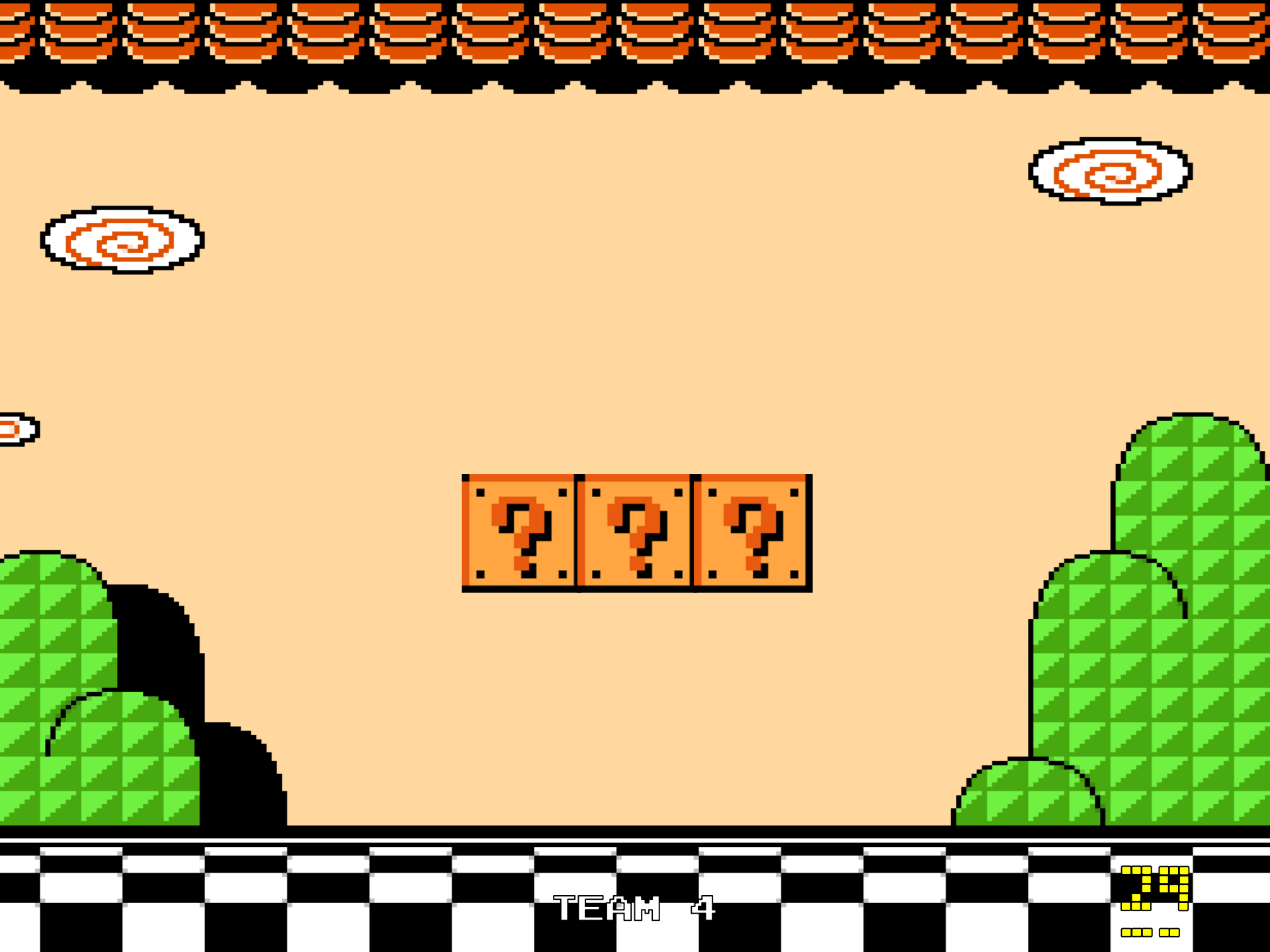
Overall Accuracy **UP** to **88**%

No. of Records **30**%

Density increases **19**%




YOU WIN!

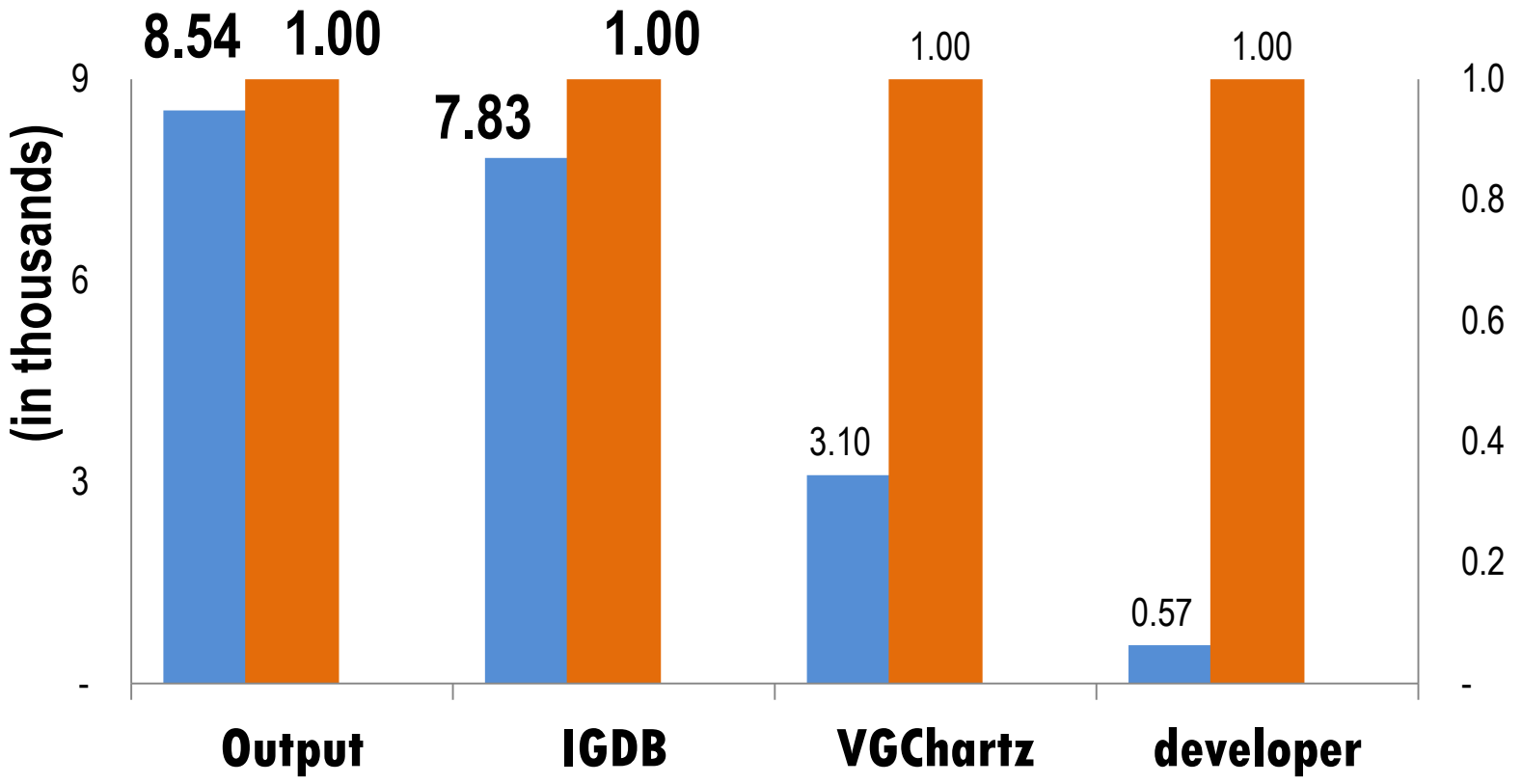

TEAM 4
74

APPENDIX

XML Data model

```
<game id="igdb_game_123231+mc_104682+vg_104115" developer="vg_dev_101063">
   <nameOfGame provenance="vg_104115+igdb_game_123231+mc_104682">DiRT 4
   </nameOfGame>
   <publisher provenance="igdb_game_123231">Codemasters</publisher>
   <platform provenance="igdb_game_123231">XOne</platform>
   <genre provenance="igdb_game_123231">Simulator, Racing, Sport</genre>
   <releaseDate provenance="vg_104115+mc_104682">2017-06-06T00:00
   </releaseDate>
   <userScore provenance="">66.0</userScore>
   <maturityRating provenance="igdb_game_123231">T</maturityRating>
   <expertScore provenance="">78.25017916502776</expertScore>
   <totalSales provenance="vg_104115">0.17</totalSales>
   <NASales provenance="vg_104115">0.06</NASales>
   <PALSales provenance="vg_104115">0.1</PALSales>
   <JapanSales provenance="" />
   <otherSales provenance="vg_104115">0.01</otherSales>
 </game>
```

Output of Developer



Data Set Information

Data Set	# of Attributes	List of Attributes
vgchartz	15	Publisher, Name, Console, Genre, Release Date, Developer User Score (MV), Critic Score (MV), VGChartz Score (MV) Total Sales (MV),NA Sales (MV), PAL Sales (MV), Japan Sales (MV), Other Sales (MV), Last Update (MV)
igdb	8	Publishers(MV), Name, Platforms, Genre(s), Release Date Developer (MV), Matruity Rating(MV), Expert Score(MV)
metacritic	8	Publisher, Name, Platform, Genre(s), ReleaseDate, UserScore, MetaScore, Rating
developers	7	Developer, City, Autonomous area (MV), Country, Est., Notable games (MV), Notes (MV)

Attribute Intersection with Integrated Schema

Class Name	Attributes Name	Datasets in which attribute is found
Game	nameOfGame	vgchartz, metacritic, igdb
Game	publisher	vgchartz, metacritic, igdb
Game	developer	vgchartz, igdb
Game	platform	vgchartz, metacritic, igdb
Game	genre	vgchartz, metacritic, igdb
Game	releaseDate	vgchartz, metacritic, igdb
Game	userScore	vgchartz, metacritic
Game	maturityRating	metacritic, igdb
Game	expertsScore	vgchartz, metacritic, igdb
Developer	nameOfDeveloper	developers, igdb

Machine Learning Results

< IGDB - MetaCritic >

Classifier	Matching Rule	Blocker	Threshold	P	R	F1	# Corr	Run Time	Reduction Ratio
Baseline	GameNameComparatorEqual()	GameBlockingKeyBy PlatformDateGenerator()	0.5	1	0.75	0.857	3,636	2.1 sec	0.9998
HoeffdingTree {"-S", "0", "-L", "2"}	GameDateComparator1Years(), GameDateComparator2Years(), GameDateComparator3Years(), GameDateComparatorWeightedDate(), GameNameComparatorDamerau(), GameNameComparatorEqual(), GameNameComparatorJaccard(), GameNameComparatorJaroWinkler(), GameNameComparatorJaroWinklerTfIdf(), GameNameComparatorLevenshtein(), GameNameComparatorMongeElkan(), GameNameComparatorMongeElkanTfIdf(), GameNameComparatorSoftTfIdf(), GamePublisherComparatorEqual(), GamePublisherComparatorJaccard(); GamePublisherComparatorJaroWinklerTfIdf(); GamePublisherComparatorLevenshtein(); GamePublisherComparatorMongeElkan(); GamePublisherComparatorMongeElkanTfIdf(); GamePublisherComparatorSoftTfIdf(), GameGenreComparatorJaccard(), BackwardSelection(true)	GameBlockingKeyBy PlatformName Generator()	0.8	0.852	0.821	0.836	5,529	1 min 20.77 sec	0.9979

Machine Learning Results (developer)

< IGDB Dev - Dev >

Classifier	Matching Rule	Blocker	Threshold	P	R	F1	# Corr	Run Time	Reduction Ratio
Baseline	MachineLearning SimpleLogistic: Equal	DeveloperBlockingKeyBy NameGenerator()	0.4	0.485	1	0.65	34,179	2.0 sec	0.9923
SimpleLogistic	DeveloperNameComparatorEqual(), DeveloperNameComparatorJaccard(), DeveloperComparatorMongeElkan(), DeveloperComparatorMongeElkanTfIdf()	DeveloperBlockingKeyByNameGenerator()	0.9	1	0.63	0.77	667	4.81 sec	0.992325

< VG Dev - Dev >

Classifier	Matching Rule	Blocker	Threshold	P	R	F1	# Corr	Run Time	Reduction Ratio
Baseline	MachineLearning SimpleLogistic: Equal	DeveloperBlockingKeyBy NameGenerator()	0.5	1	0.15	0.27	324	1.4 sec	0.9922
HoeffdingTree	DeveloperNameComparatorDamerau(), DeveloperNameComparatorEqual(), DeveloperNameComparatorJaccard(), DeveloperNameComparatorJaroWinkler(), DeveloperNameComparatorLevenshtein(), DeveloperComparatorJaroWinklerTfIdf(), DeveloperComparatorMongeElkan(), DeveloperComparatorMongeElkanTfIdf(), DeveloperComparatorSoftTfIdf()	DeveloperBlockingKeyBy NameGenerator()	0.8	0.818	0.69	0.75	514	2.43 sec	0.99225

< IGDB Dev - VG Dev >

Classifier	Matching Rule	Blocker	Threshold	P	R	F1	# Corr	Run Time	Reduction Ratio
Baseline	MachineLearning SimpleLogistic: Equal	DeveloperBlockingKeyBy NameGenerator()	0.5	0.743	1	0.85	179898	7.5 sec	0.9925
SimpleLogistic	DeveloperNameComparatorEqual(), DeveloperNameComparatorJaccard(), DeveloperComparatorMongeElkan(), DeveloperComparatorMongeElkanTfIdf()	DeveloperBlockingKeyBy NameGenerator()	0.8	0.977	0.81	0.88	2,731	17.90 sec	0.99225