# INTRODUCTION TO PATTERN MINING

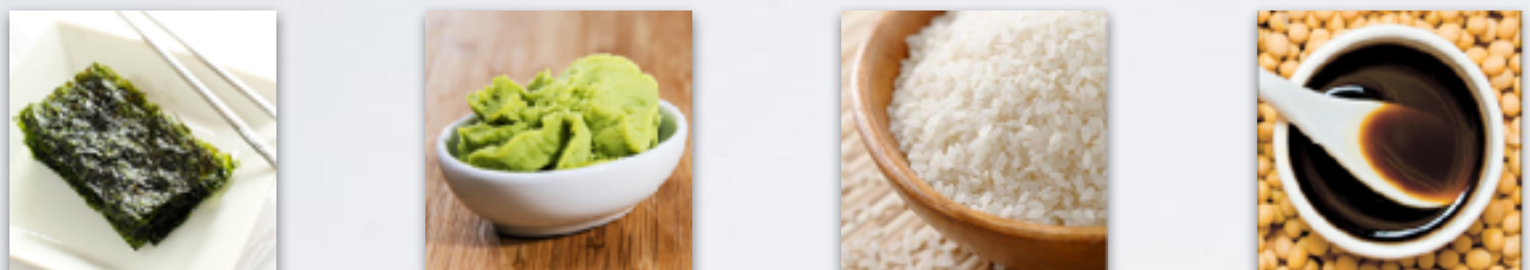
Vincent Leroy

# REFERENCES

- Christian Borgelt
  http://www.borgelt.net/teach/fpm/

# CONTEXT

## ItemSet



Support = 133

## Association Rule

Recall = 27%
Confidence = 81%

# DEFINITIONS (1)

- Set of items *I*
  (e.g. *I*={potatoes, milk, sugar …}, all products sold in the store)

- Transaction *T* is a subset of *I*
  (in our case a customer receipt)

- Database *D* is a collection of transactions
  (e.g. all receipts in a store for a month)

# DEFINITIONS (2)

- An Itemset *P* is a subset of *I*, and *P* occurs in a transaction *T* if *P* is a subset of *T*
(e.g. the itemset {milk, eggs} occurs in the transaction {milk, eggs, chocolate})

- The support of an itemset *P* in a database *D* is the number of transactions in *D* in which *P* occurs

# ITEMSET MINING EXAMPLE

- ItemSets with support at least 3

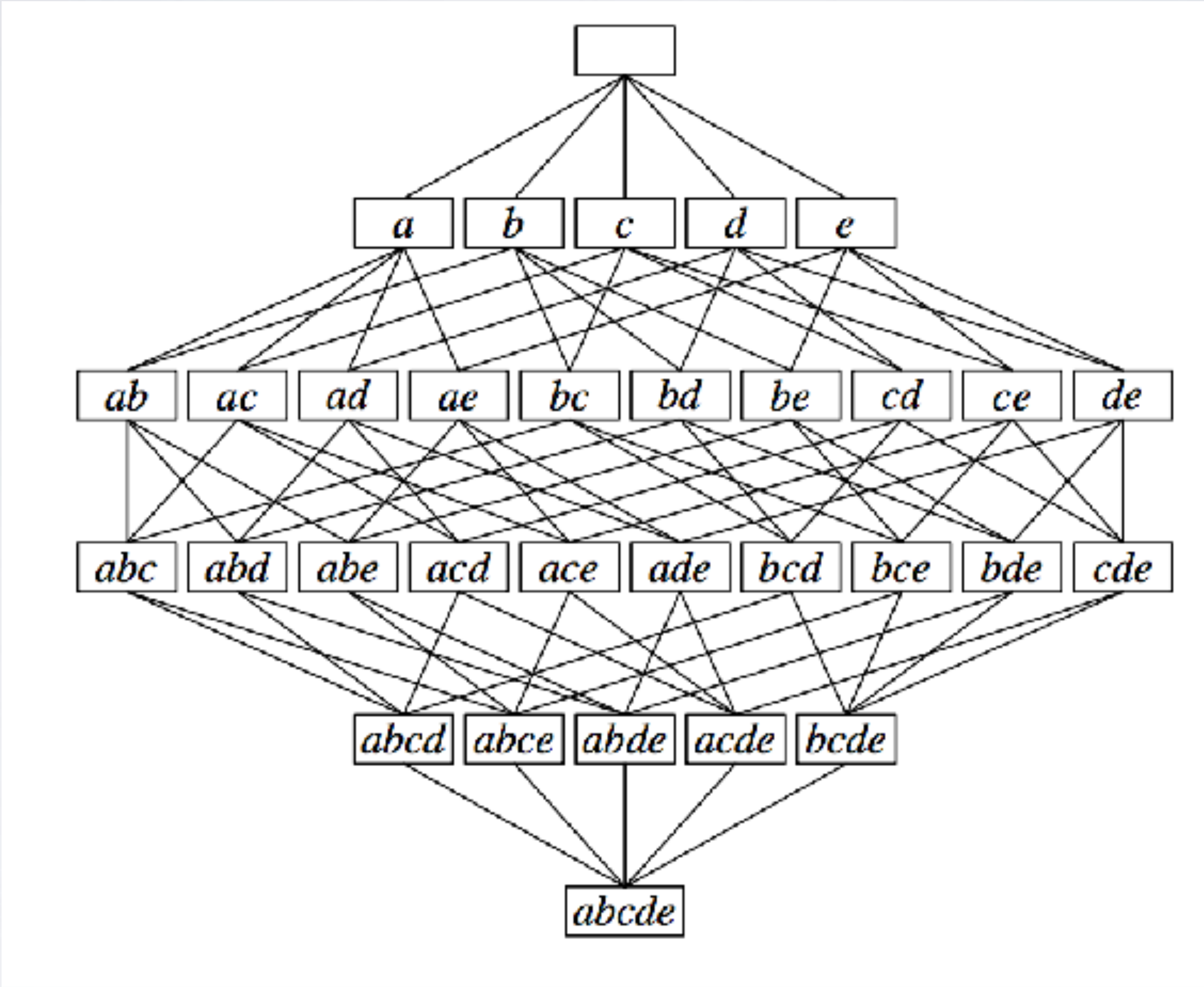
transaction database

1: $\{a, d, e\}$
2: $\{b, c, d\}$
3: $\{a, c, e\}$
4: $\{a, c, d, e\}$
5: $\{a, e\}$
6: $\{a, c, d\}$
7: $\{b, c\}$
8: $\{a, c, d, e\}$
9: $\{b, c, e\}$
10: $\{a, d, e\}$

frequent item sets

| 0 items | 1 item | 2 items | 3 items |
|---|---|---|---|
| $\emptyset$: 10 | $\{a\}$: 7 | $\{a, c\}$: 4 | $\{a, c, d\}$: 3 |
| | $\{b\}$: 3 | $\{a, d\}$: 5 | $\{a, c, e\}$: 3 |
| | $\{c\}$: 7 | $\{a, e\}$: 6 | $\{a, d, e\}$: 4 |
| | $\{d\}$: 6 | $\{b, c\}$: 3 | |
| | $\{e\}$: 7 | $\{c, d\}$: 4 | |
| | | $\{c, e\}$: 4 | |
| | | $\{d, e\}$: 4 | |

# FINDING FREQUENT ITEMSETS




- Hasse diagram of itemsets: $2^l$ possibilities

  A retail store sells 10k products, brute-force too expensive!

# APRIORI PROPERTY

- Given an ItemSet *P* and *P'*, a superset of *P*
  *support(P',D)* ≤ *support(P,D)*
  e.g. there are less customers that buy {rice, seaweed, soy sauce} than customers that buy {rice, soy sauce}

- If we know that *P* is not frequent, there is no need to check if supersets of *P* are frequent

# APRIORI ALGORITM (1)

- Recursive algorithm
  Find frequent ItemSets of size *k*, then generate candidates of size *k+1*

- Given a candidate *P* of size *k+1*, if any of its subsets of size *k* is not frequent, then *P* cannot be frequent
  No need to compute its support! (expensive)

# APRIORI ALGORITM (2)

```
function apriori (B, T, s_min)
begin                                   (* — Apriori algorithm *)
  k   := 1;                             (* initialize the item set size *)
  E_k := ∪_{i∈B} {{i}};                 (* start with single element sets *)
  F_k := prune(E_k, T, s_min);          (* and determine the frequent ones *)
  while F_k ≠ ∅ do begin                (* while there are frequent item sets *)
    E_{k+1} := candidates(F_k);         (* create candidates with one item more *)
    F_{k+1} := prune(E_{k+1}, T, s_min); (* and determine the frequent item sets *)
    k       := k + 1;                   (* increment the item counter *)
  end;
  return ∪_{j=1}^{k} F_j;               (* return the frequent item sets *)
end (* apriori *)
```

$E_j$: candidate item sets of size $j$, $F_j$: frequent item sets of size $j$.

# ASSOCIATION RULES (1)

- ItemSet {rice, seaweed, soy sauce}

| | | |
|---|---|---|
| {rice} | → | {seaweed, soy sauce} |
| {seaweed} | → | {rice, soy sauce} |
| {soy sauce} | → | {rice, seaweed} |
| {rice, seaweed} | → | {soy sauce} |
| {rice, soy sauce} | → | {seaweed} |
| {soy sauce, seaweed} | → | {rice} |

# ASSOCIATION RULES (II)

- Association rule *A* → *B*
- Support of the rule: *support(A)*
- Confidence (precision): *support(A∪B)/support(A)* equivalent to conditional probability of *B* given *A*

# RETAIL DATASETS

*Apriori* algorithm by Agrawal & Srikant (1994)


x47k

Intermarché (2014)


x300M

x200k

?


x9M

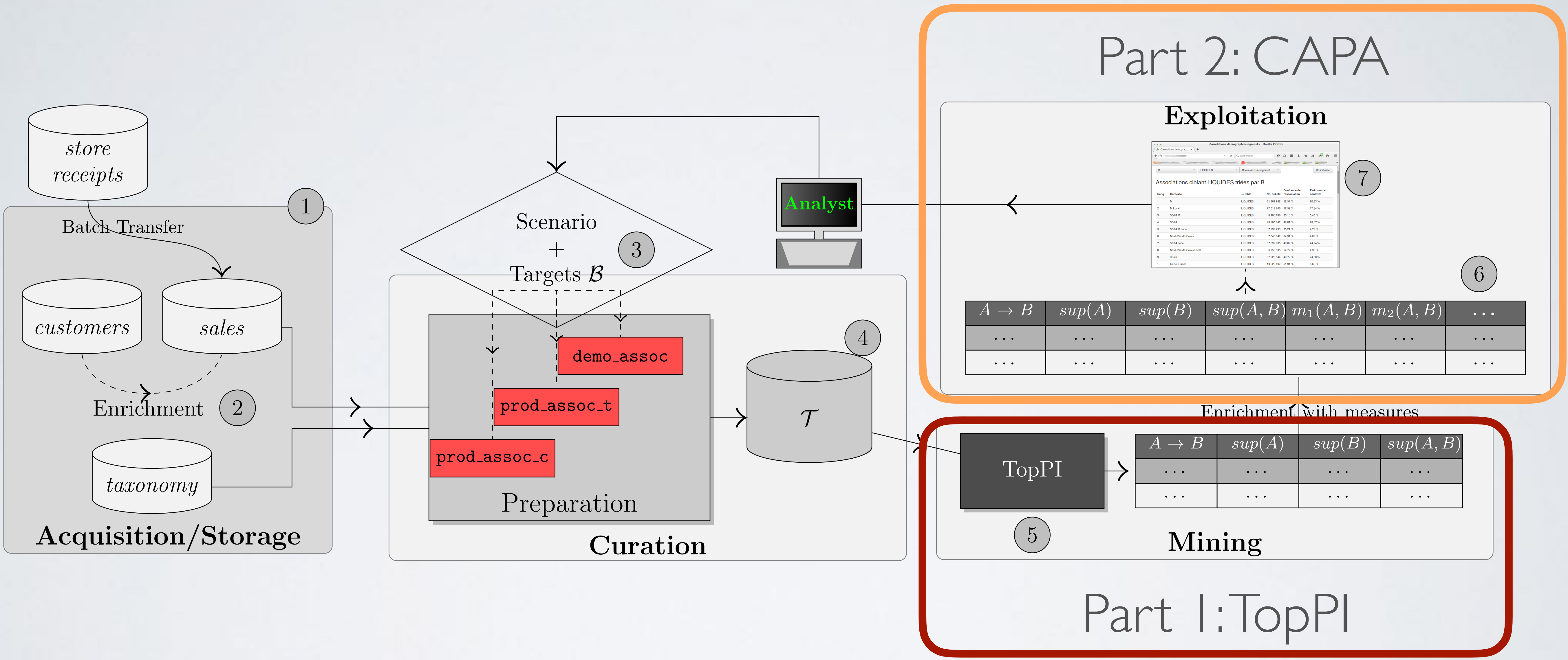
# MINING LARGE-SCALE RETAIL DATA

Vincent Leroy

# GOALS

- Coverage: generate rules about any item
- Scalability: process millions of receipts
- Quality: identify the most remarkable rules

# ARCHITECTURE OVERVIEW

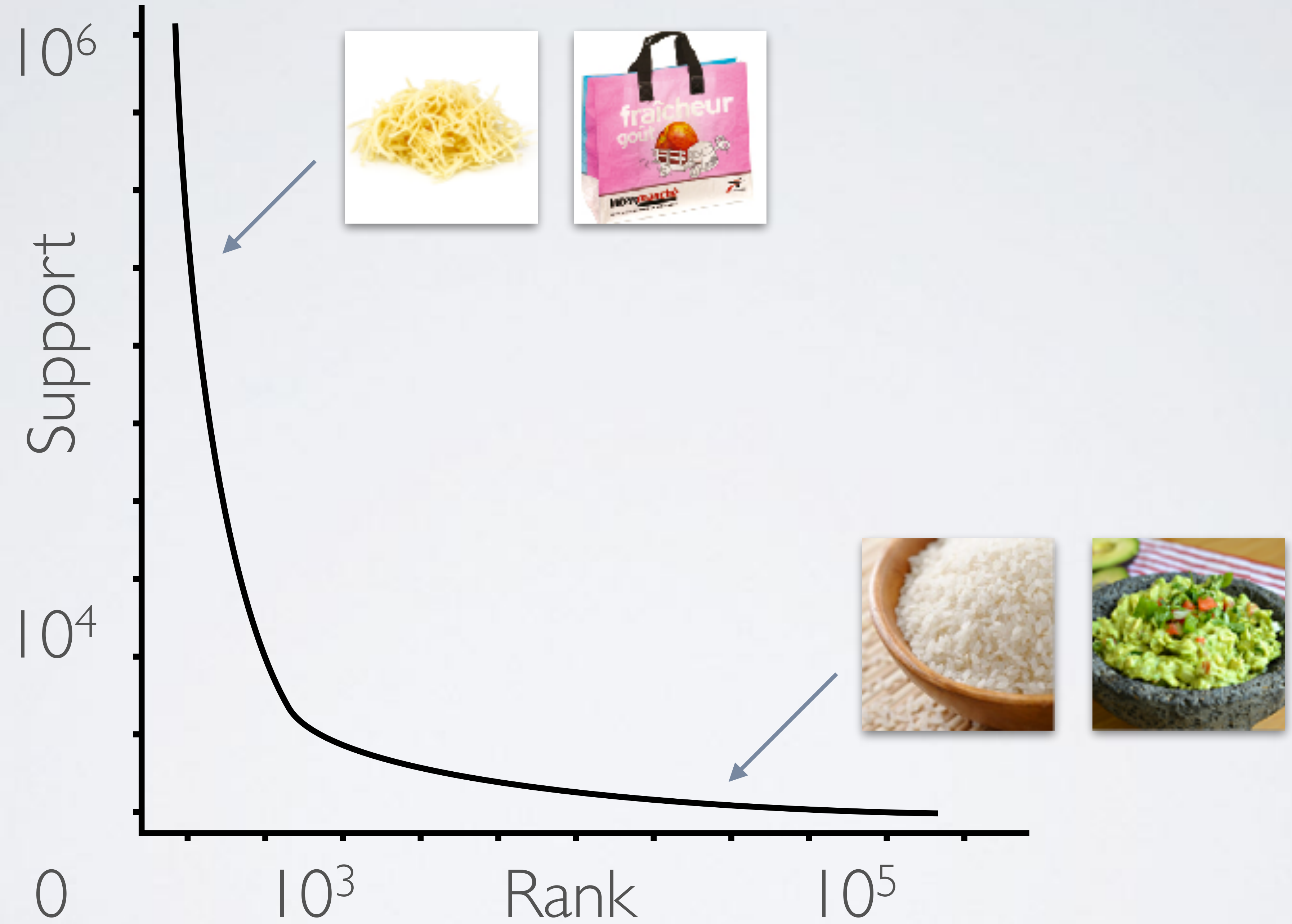


# TopPI:
# MINING THE LONG TAIL

# ITEMSET MINING: SOTA

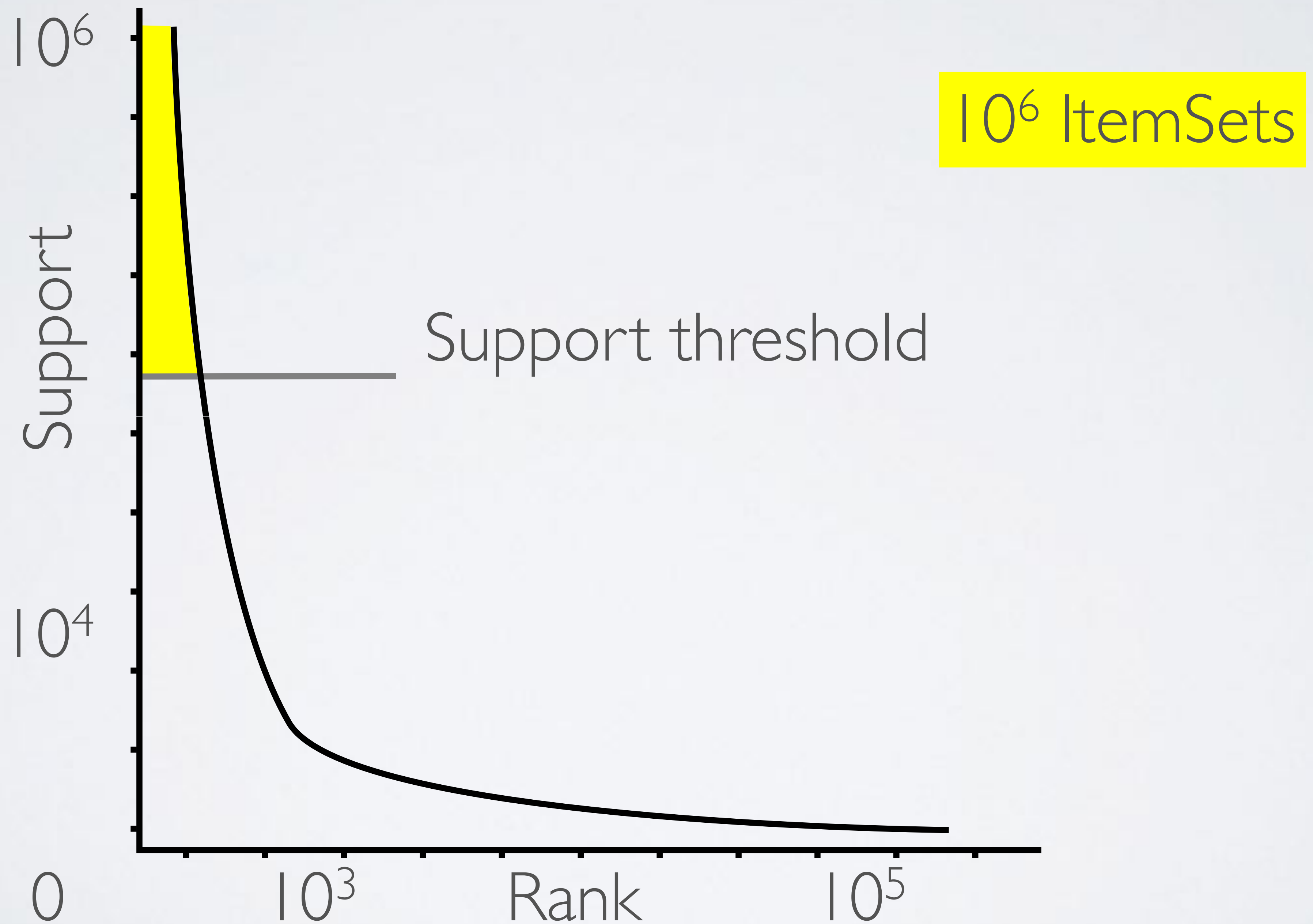
- 20 years of research to accelerate mining
  - Standardized benchmark datasets
  - LCM algorithm as a building block
- Focus on high support ItemSets
  - Find all ItemSets s.t. support > threshold
  - Find the *k* most frequent ItemSets

# LONG-TAILED DISTRIBUTION

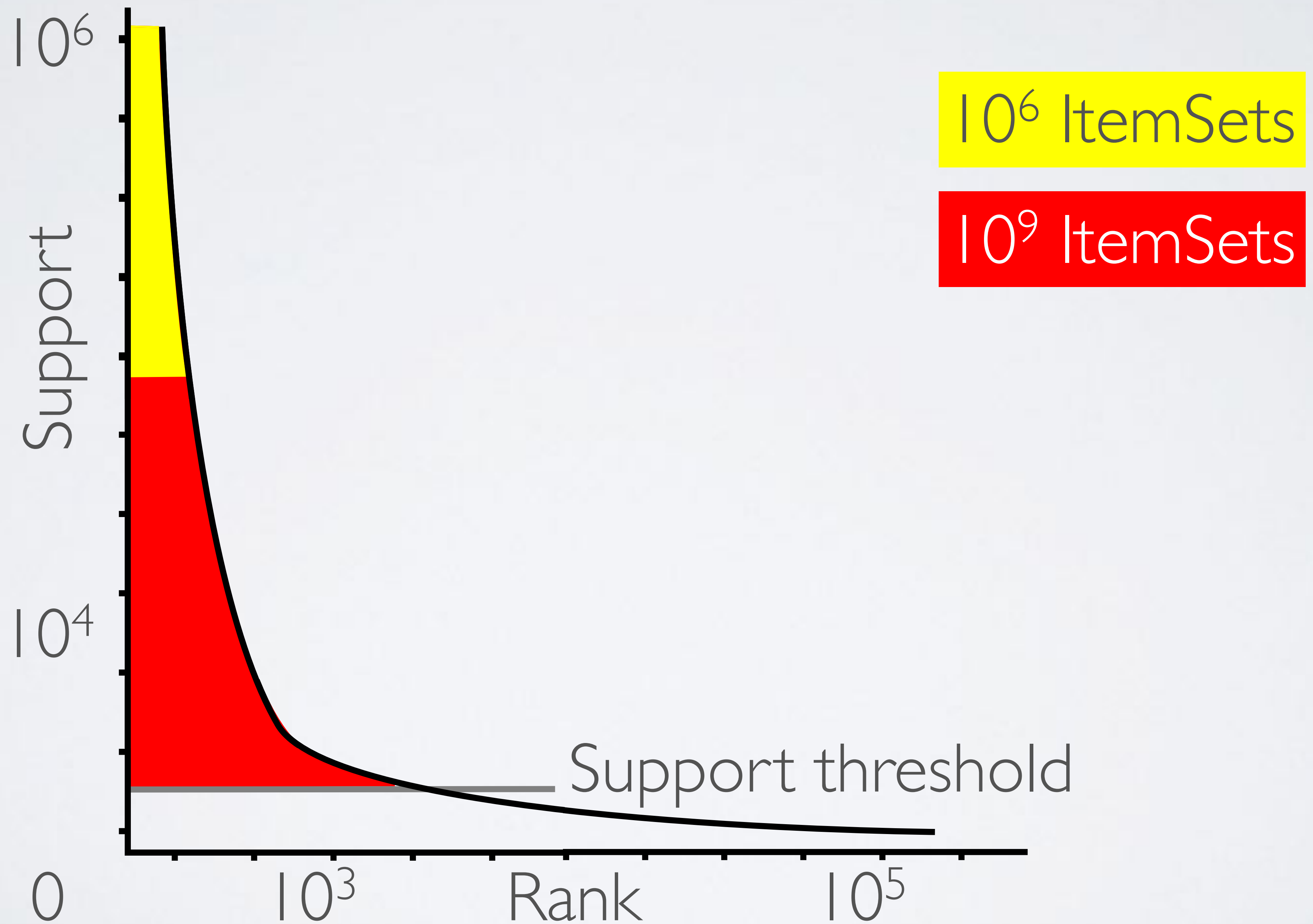



Anatomy of the long tail: ordinary people with extraordinary tastes
Goel, Broder, Gabrilovich, Pang @ WSDM'10

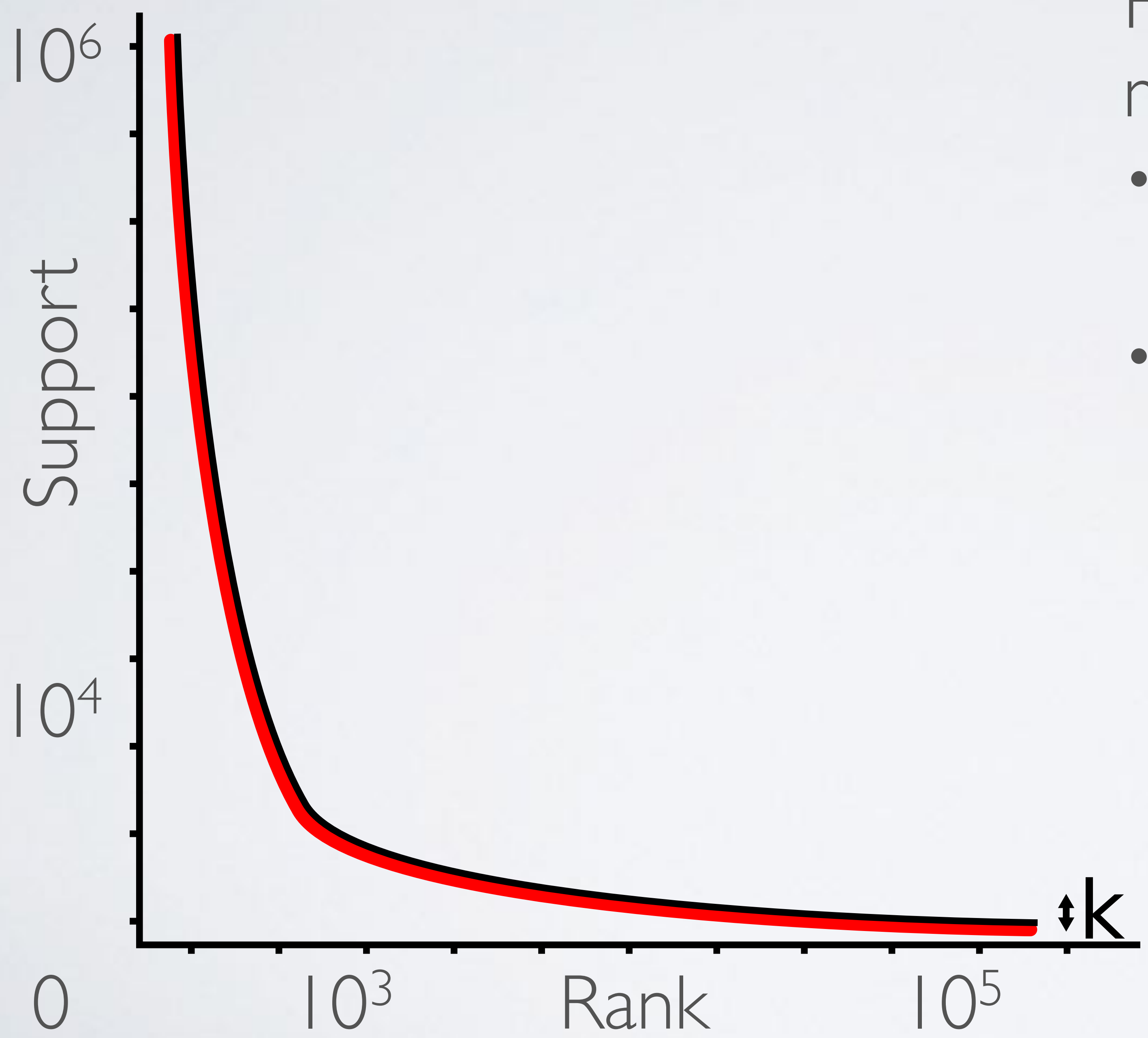
# FREQUENT ITEMSETS AND THE LONG TAIL (1)



# FREQUENT ITEMSETS AND THE LONG TAIL (2)



# ITEM-CENTRIC MINING




For each item, find the *k* most frequent ItemSets

- *k* results for each item (coverage)
- At most *k*nbItems* ItemSets (scalability)

**Top**-*k* **P**er **I**tem

# TopPI OUTPUT

top(Grated Cheese)

| Support | ItemSet |
|---|---|
| 9,395,643 | Grated Cheese |
| 861,304 | Grated Cheese, Cream |
| 793,310 | Grated Cheese, 10 eggs |
| 652,493 | Grated Cheese, Butter |
| 597,144 | Grated Cheese, Bacon |

top(Sushi Rice)

| Support | ItemSet |
|---|---|
| 14,887 | Sushi Rice |
| 5,935 | Sushi Rice, Seaweed |
| 3,669 | Sushi Rice, Rice Vinegar |
| 1,843 | Sushi Rice, Seaweed, Rice Vinegar |
| 1,762 | Sushi Rice, Wasabi |

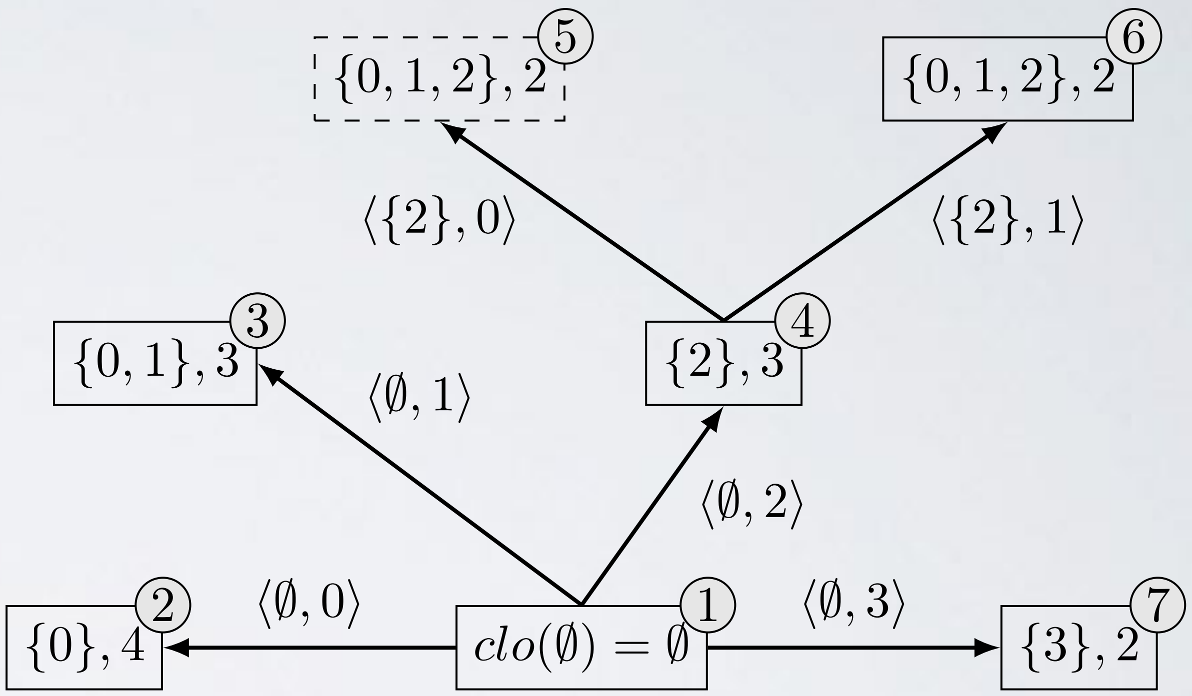
# TREE-SHAPED MINING

| TID | Transaction |
|---|---|
| $t_0$ | $\{0,1,2\}$ |
| $t_1$ | $\{0,1,2\}$ |
| $t_2$ | $\{0,1\}$ |
| $t_3$ | $\{2,3\}$ |
| $t_4$ | $\{0,3\}$ |

(a) Input $\mathcal{D}$

| item $i$ | $top(i)$: $P$, $support(P)$ $1^{st}$ | $2^{nd}$ |
|---|---|---|
| 0 | $\{0\},4$ | $\{0,1\},3$ |
| 1 | $\{0,1\},3$ | $\{0,1,2\},2$ |
| 2 | $\{2\},3$ | $\{0,1,2\},2$ |
| 3 | $\{3\},2$ | |

(b) TopPI results for $k=2$




Recursion using an *expand* function to add an item to an ItemSet (LCM)

*Anti-monotony* of ItemSet support for pruning

# TOP-*k* PROCESSING

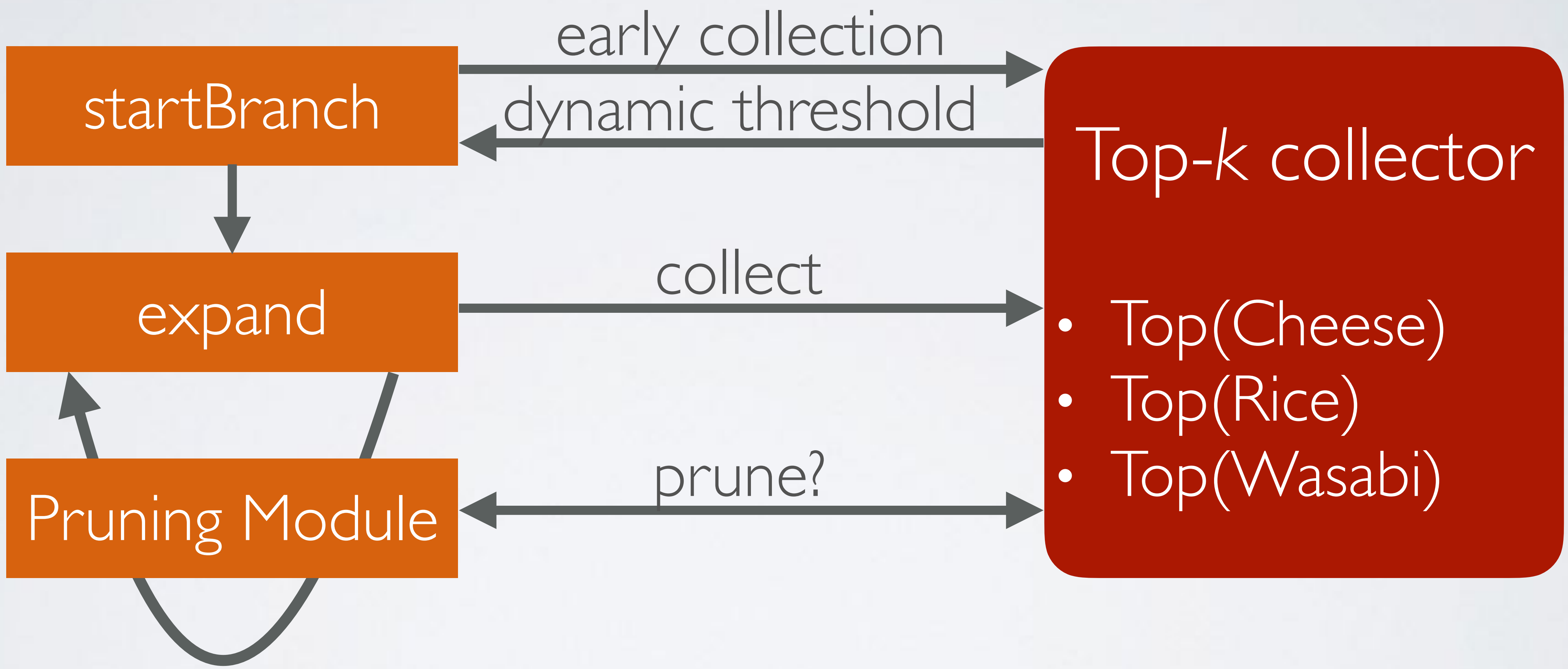
- General principle
  - Maintain Lower Bound on current top-*k*
  - Maintain Upper Bound on unseen results
  - Prune when UB < LB
- Applicability in TopPI
  - One top-*k* per item
  - Bounds on support of ItemSets (anti-monotony)

# OVERVIEW OF TopPI (1)

## Tree-shaped exploration guided by top-*k*(s) pruning

- Support-based pruning
  - Carefully determine which ItemSets can be generated in a branch
  - Eliminate branches that cannot improve any top-*k* (result-space is **not** monotonic!)
  - Insert (partial) ItemSets early to raise LBs
- Dataset reduction
  - Use LBs to eliminate items that will not be in relevant ItemSets
  - Instantiate in memory when gain is significant (RAM, L3, L2, L1)
- Optimize, optimize, optimize

# OVERVIEW OF TopPI (2)



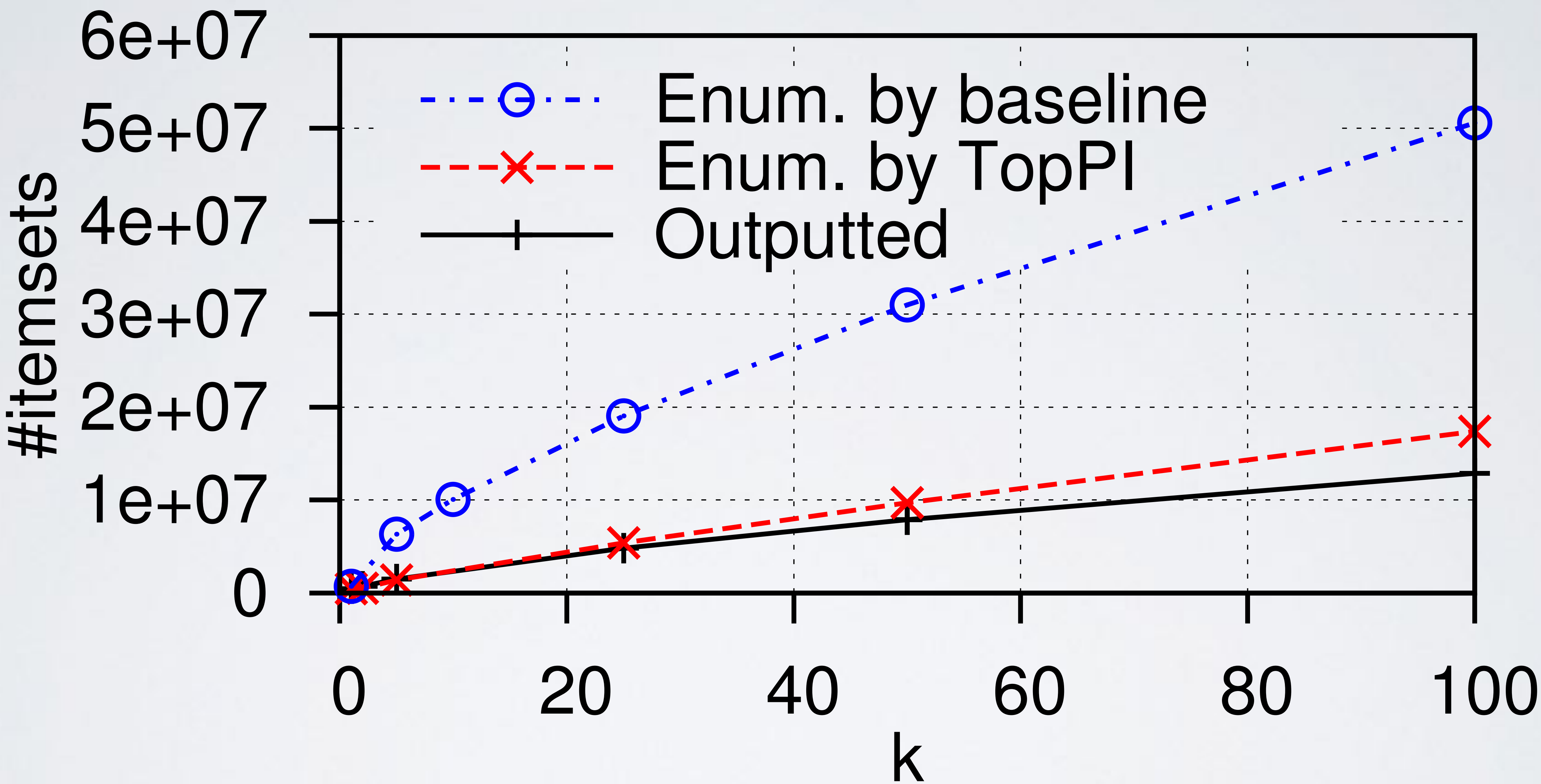
# SINGLE SERVER PERFORMANCE

| Dataset | #items | #transactions | Exec. time |
|---|---|---|---|
| Receipts | 222k | 290M | 4 min |
| Receipts, by client | 222k | 9M | 11 min |

32 threads, 128 GB RAM, *k*=50

1 year of sales in minutes on a single server

# MINING OVERHEAD




Mining focused on useful results, pruning almost optimal

# DISTRIBUTED ALGORITHM

*Fortunately*, some (Web) datasets are more challenging

ex: WebDocs 8h on a single server

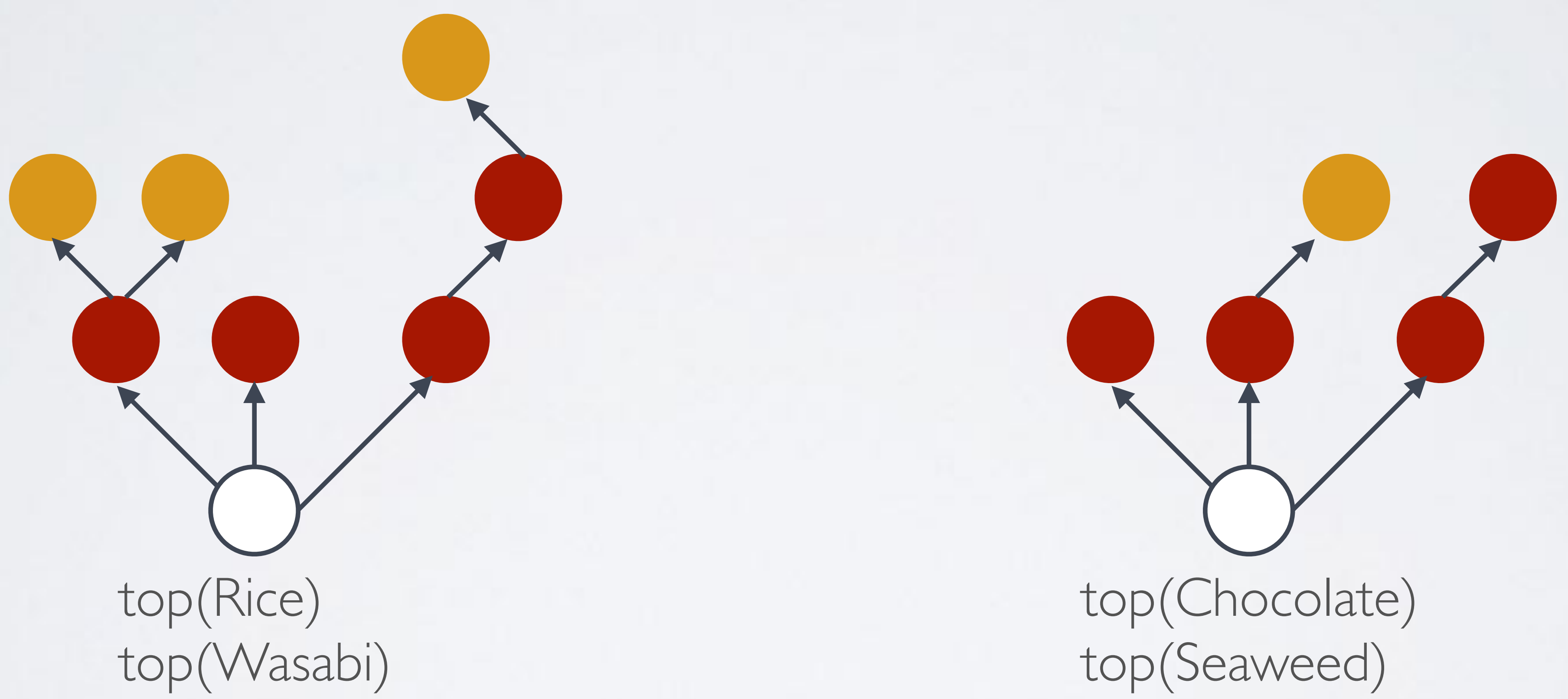
Distribute on Hadoop

Challenges:

- Partition into many mining task
- No communication between tasks

# PARTITIONING ITEMS

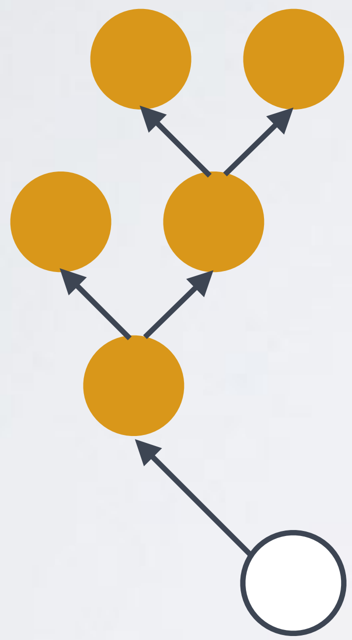
Each task is responsible for a set of items and produces all necessary patterns



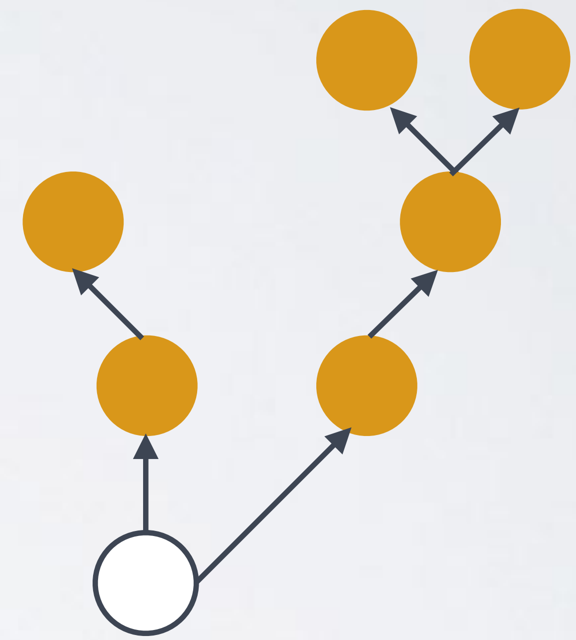

Drawback: some ItemSets are enumerated multiple times

# PARTITIONING THE TREE

Each task is responsible for some branches of the enumeration tree and computes a top-*k* for all items



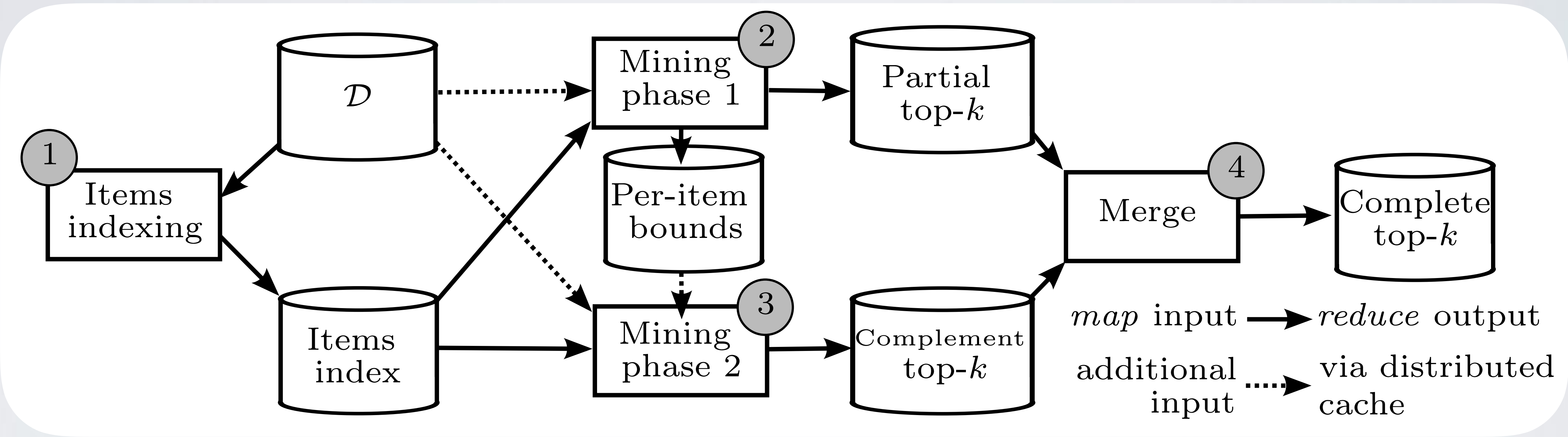
top(Rice)
top(Wasabi)
top(Chocolate)
top(Seaweed)



top(Rice)
top(Wasabi)
top(Chocolate)
top(Seaweed)

Drawback: multiple top-*k* per item, less pruning

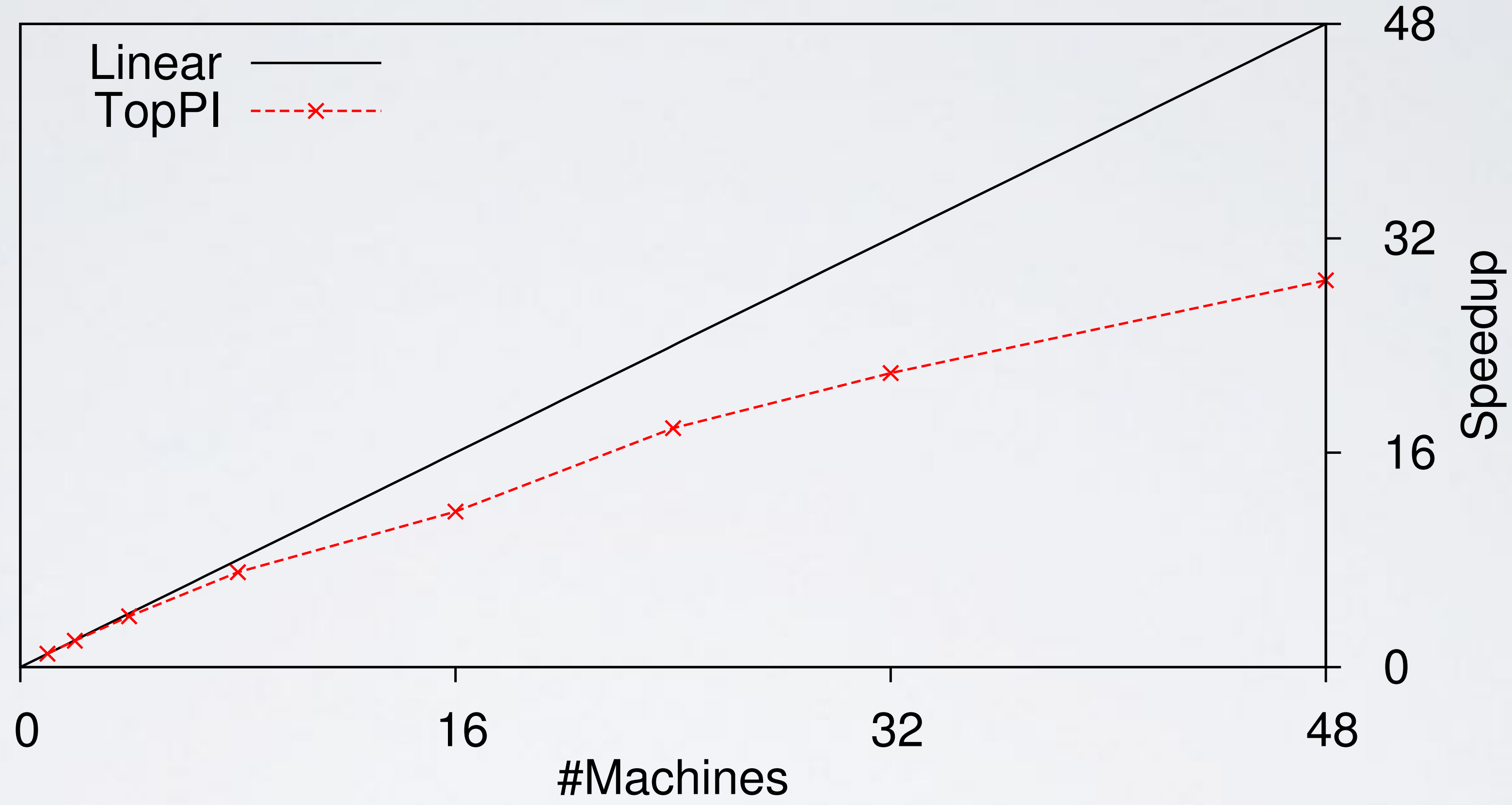
# HYBRID STRATEGY




Phase 1: Partition items and tree
Phase 2: Partition tree

Close to no mining overhead

# HADOOP SPEEDUP




Supermarket dataset, $k$=1000

- Excellent CPU time scalability
- I/O overhead

# TopPI CONCLUSION

- Item-centric mining
  - Redefines mining target for long-tailed data
  - Provides coverage and scalability → Intermarché
- Top-*k* based on support
  - Enables pruning (anti-monotony)
  - Frequency is important, but is all that is frequent interesting?

# QUALITY?

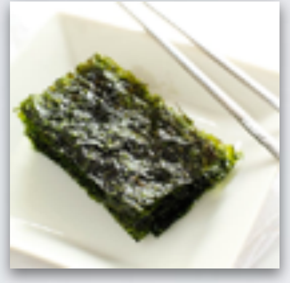
top(Chocolate cream)

| Support | ItemSet |
|---|---|
| 581,042 | Chocolate cream |
| 58,569 | Chocolate cream, Vanilla cream |
| 32,701 | Chocolate cream, Grated cheese |
| 30,451 | Chocolate cream, Cola |
| 29,671 | Chocolate cream, Butter |

# CAPA:
# IDENTIFYING INTERESTING RULES

# ASSOCIATION RULES

## Product rule




## Demographic rule


# RANKING RULES

- Basic measures
  - Recall
  - Confidence
- Advanced measures
  - Over 34!
  - No guidelines

| Measure | Formula |
|---|---|
| One-Way Support | $P(B\|A) \times log_2 \frac{P(AB)}{P(A)P(B)}$ |
| Relative Risk | $P(B\|A)/P(B\|\neg A)$ |
| Odd Multiplier | $\frac{P(AB)P(\neg B)}{P(B)P(A\neg B)}$ |
| Zhang | $\frac{P(AB)-P(A)P(B)}{max(P(AB)P(\neg B),P(B)P(A\neg B))}$ |
| Yule's Q $\diamond$ | $\frac{P(AB)P(\neg A\neg B)-P(A\neg B)P(B\neg A)}{P(AB)P(\neg A\neg B)+P(A\neg B)P(B\neg A)}$ |
| Yule's Y $\diamond$ | $\frac{\sqrt{P(AB)P(\neg A\neg B)}-\sqrt{P(A\neg B)P(B\neg A)}}{\sqrt{P(AB)P(\neg A\neg B)}+\sqrt{P(A\neg B)P(B\neg A)}}$ |
| Odds Ratio $\diamond$ | $\frac{P(AB)P(\neg A\neg B)}{P(A\neg B)P(B\neg A)}$ |
| Information Gain $*\ominus$ | $log(P(AB)/(P(A)P(B)))$ |
| Lift $*\ominus$ | $P(AB)/(P(A)P(B))$ |
| Added Value $*$ | $P(B\|A) - P(B)$ |
| Certainty Factor $*$ | $(P(B\|A) - P(B))/(1 - P(B))$ |
| Confidence / Precision $*\otimes$ | $P(B\|A)$ |
| Laplace Correction $*\otimes$ | $\frac{support(AB)+1}{support(A)+2}$ |
| Loevinger $\dagger$ | $1 - \frac{P(A)P(\neg B)}{P(A\neg B)}$ |
| Conviction $\dagger$ | $\frac{P(A)P(\neg B)}{P(A\neg B)}$ |
| Example and Counter-example Rate | $1 - \frac{P(A\neg B)}{P(AB)}$ |
| Sebag-Schoenauer | $\frac{P(AB)}{P(A\neg B)}$ |
| Leverage | $P(B\|A) - P(A)P(B)$ |
| Least Contradiction | $\frac{P(AB)-P(A\neg B)}{P(B)}$ |
| Accuracy | $P(AB) + P(\neg A\neg B)$ |
| Pearson's $\chi^2$ $\triangleright$ | $\|\mathcal{T}\| \times \left(\frac{(P(AB)-P(A)P(B))^2}{P(A)P(B)} + \frac{(P(\neg AB)-P(\neg A)P(B))^2}{P(\neg A)P(B)}\right)$ $+\|\mathcal{T}\| \times \left(\frac{(P(A\neg B)-P(A)P(\neg B))^2}{P(A)P(B)} + \frac{(P(\neg A\neg B)-P(\neg A)P(\neg B))^2}{P(\neg A)P(\neg B)}\right)$ |
| Gini Index $\triangleright$ | $P(A) \times (P(B\|A)^2 + P(\neg B\|A)^2) + P(\neg A) \times (P(B\|\neg A)^2 +$ $P(\neg B\|\neg A)^2) - P(B)^2 - P(\neg B)^2$ |
| J-measure | $P(AB)log(\frac{P(B\|A)}{P(B)}) + P(A\neg B)log(\frac{P(\neg B\|A)}{P(\neg B)})$ |
| $\Phi$ Linear Correlation Coefficient | $\frac{P(AB)-P(A)P(B)}{\sqrt{P(A)P(B)P(\neg A)P(\neg B)}}$ |
| Two-Way Support Variation | $P(AB) \times log_2 \frac{P(AB)}{P(A)P(B)} + P(A\neg B) \times log_2 \frac{P(A\neg B)}{P(A)P(\neg B)} +$ $P(\neg AB) \times log_2 \frac{P(\neg AB)}{P(\neg A)P(B)} + P(\neg A\neg B) \times log_2 \frac{P(\neg A\neg B)}{P(\neg A)P(\neg B)}$ |
| Fisher's exact test | $\frac{\binom{\|\mathcal{T}\| \times P(B)}{\|\mathcal{T}\| \times P(AB)}\binom{\|\mathcal{T}\| \times P(\neg B)}{\|\mathcal{T}\| \times P(A\neg B)}}{\binom{\|\mathcal{T}\|}{\|\mathcal{T}\| \times P(A)}}$ |
| Jaccard | $P(AB)/(P(A) + P(B) - P(AB))$ |
| Cosine | $\frac{P(AB)}{\sqrt{P(A)P(B)}}$ |
| Two-Way Support | $P(AB) \times log_2 \frac{P(AB)}{P(A)P(B)}$ |
| Piatetsky-Shapiro | $P(AB) - P(A)P(B)$ |
| Klosgen | $\sqrt{P(AB)}max(P(B\|A) - P(B), P(A\|B) - P(A))$ |
| Specificity | $P(\neg B\|\neg A)$ |
| Recall | $P(A\|B)$ |
| Collective Strength | $\frac{P(AB)+P(\neg B\|\neg A)}{P(A)P(B)+P(\neg A)P(\neg B)} \times \frac{1-P(A)P(B)-P(\neg A)P(\neg B)}{1-P(AB)-P(\neg B\|\neg A)}$ |

# EXAMPLE OF RANKINGS (1)

| **by confidence** | |
|---|---|
| $\{> 65, F, \text{Aube}\}$ | $\rightarrow$ *Dairy* |
| $\{> 65, F, \text{Aveyron}\}$ | $\rightarrow$ *Dairy* |
| $\{> 65, F, \text{Val de Marne}\}$ | $\rightarrow$ *Dairy* |
| $\{> 65, F, \text{Seine S}^{t}\text{ Denis}\}$ | $\rightarrow$ *Dairy* |
| $\{> 65, F, \text{Haute Saone}\}$ | $\rightarrow$ *Dairy* |

$$P(B|A)$$

# EXAMPLE OF RANKINGS (2)

| **by Piatetsky-Shapiro [5]** | |
|---|---|
| $\{*, *, Nord\}$ | $\rightarrow$ *Liquids* |
| $\{*, *, Nord\}$ | $\rightarrow$ *Soft drinks* |
| $\{*, *, Nord\}$ | $\rightarrow$ *Beers* |
| $\{*, *, Nord\}$ | $\rightarrow$ *Spreads* |
| $\{*, F, Nord\}$ | $\rightarrow$ *Soft drinks* |

$$P(AB) - P(A)P(B)$$

# EXAMPLE OF RANKINGS (3)

| **by Pearson's $\chi^2$** | |
|---|---|
| $\{*, *, Somme\}$ | $\rightarrow$ *Cut cheese* |
| $\{*, F, Somme\}$ | $\rightarrow$ *Cut cheese* |
| $\{> 65, *, Morbihan\}$ | $\rightarrow$ *Fresh milk* |
| $\{> 65, *, Somme\}$ | $\rightarrow$ *Cut cheese* |
| $\{*, *, Finistere\}$ | $\rightarrow$ *Canned pork* |

$$
\begin{aligned}
&|\mathcal{T}| \times \left(\frac{(P(AB)-P(A)P(B))^2}{P(A)P(B)} + \frac{(P(\neg AB)-P(\neg A)P(B))^2}{P(\neg A)P(B)}\right)\\
&+|\mathcal{T}| \times \left(\frac{(P(A\neg B)-P(A)P(\neg B))^2}{P(A)P(B)} + \frac{(P(\neg A\neg B)-P(\neg A)P(\neg B))^2}{P(\neg A)P(\neg B)}\right)
\end{aligned}
$$

# COMPARATIVE ANALYSIS OF PATTERNS

- Framework for ranking association rules
  - Ranking matters, not absolute scores
- Study in 2 phases
  - Empirical evaluation to reduce options (34 initially)
  - User study with marketing experts

# RANKING SIMILARITY (1)

## Uniform: Spearman and Kendall's τ

$$Spearman(L_{\mathcal{R}}^{m}, L_{\mathcal{R}}^{m'}) = 1 - \frac{6 \sum_{r \in \mathcal{R}} (r^{m} - r^{m'})^2}{|\mathcal{R}|(|\mathcal{R}|^2 - 1)}$$

## Top Biased: Overlap@*k* and NDCC

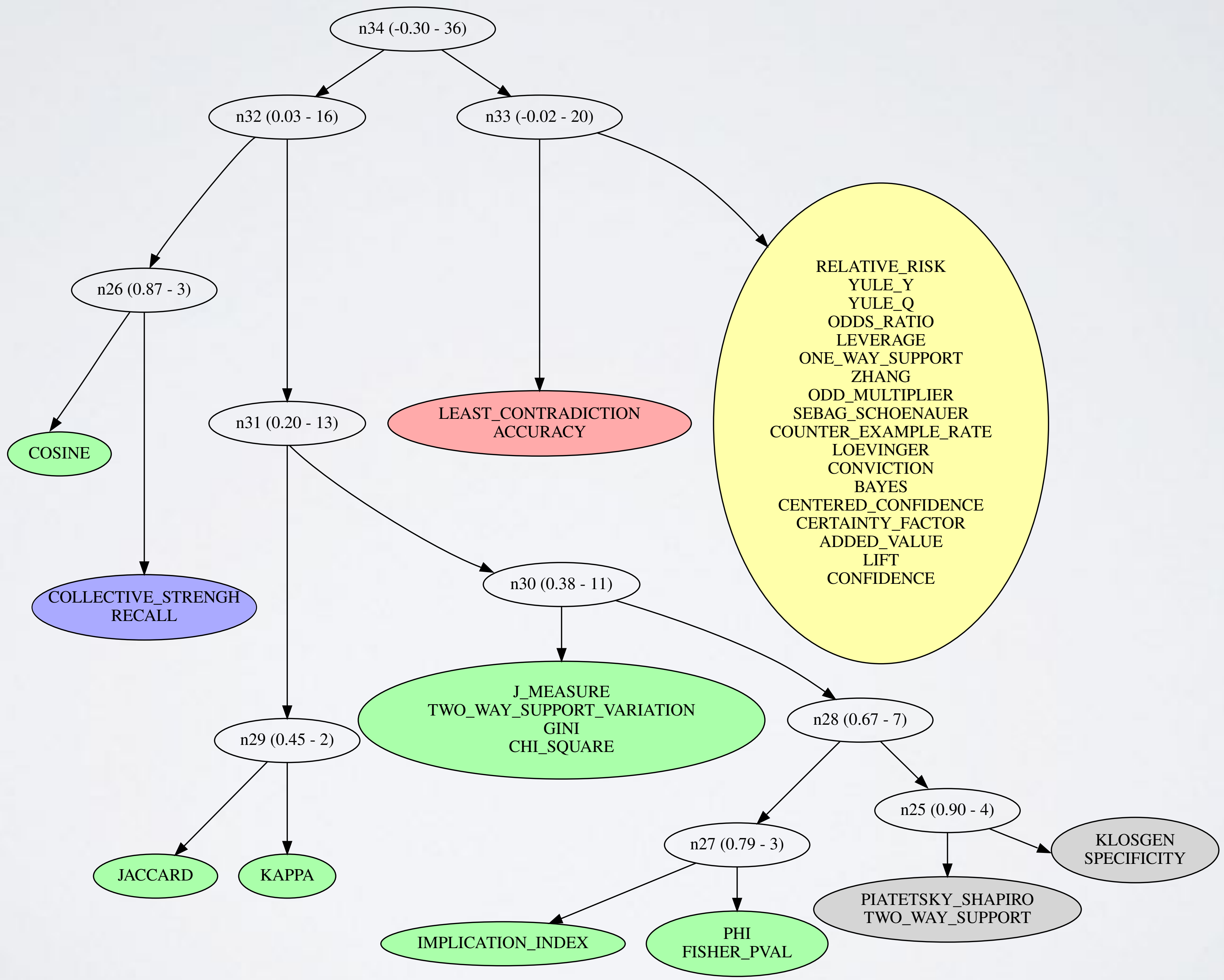
$$DCC(L_{\mathcal{R}}^{m}, L_{\mathcal{R}}^{m'}) = \sum_{r \in \mathcal{R}} \frac{1}{\log(1 + r^{m'}) \log(1 + r^{m})}$$

# RANKING SIMILARITY (2)

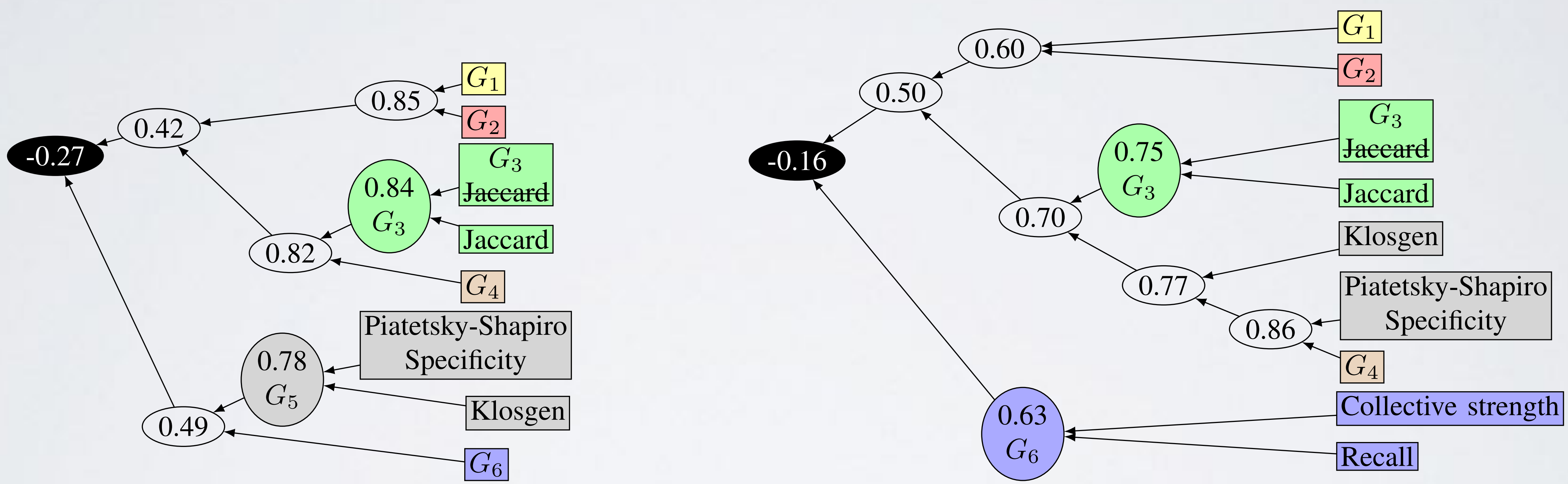

| | $Spearman$ | $\tau$ | $Overlap@2$ | $NDCC$ |
|---|---|---|---|---|
| $L^2$ | $0.80$ | $0.67$ | $1$ | $0.20$ |
| $L^3$ | $0.80$ | $0.67$ | $1$ | $0.97$ |
| $L^4$ | $0.40$ | $0.33$ | $0.5$ | $-0.18$ |

# COMPARING RANKINGS

# HIERARCHICAL CLUSTERING



# IDENTIFYING GROUPS OF MEASURES

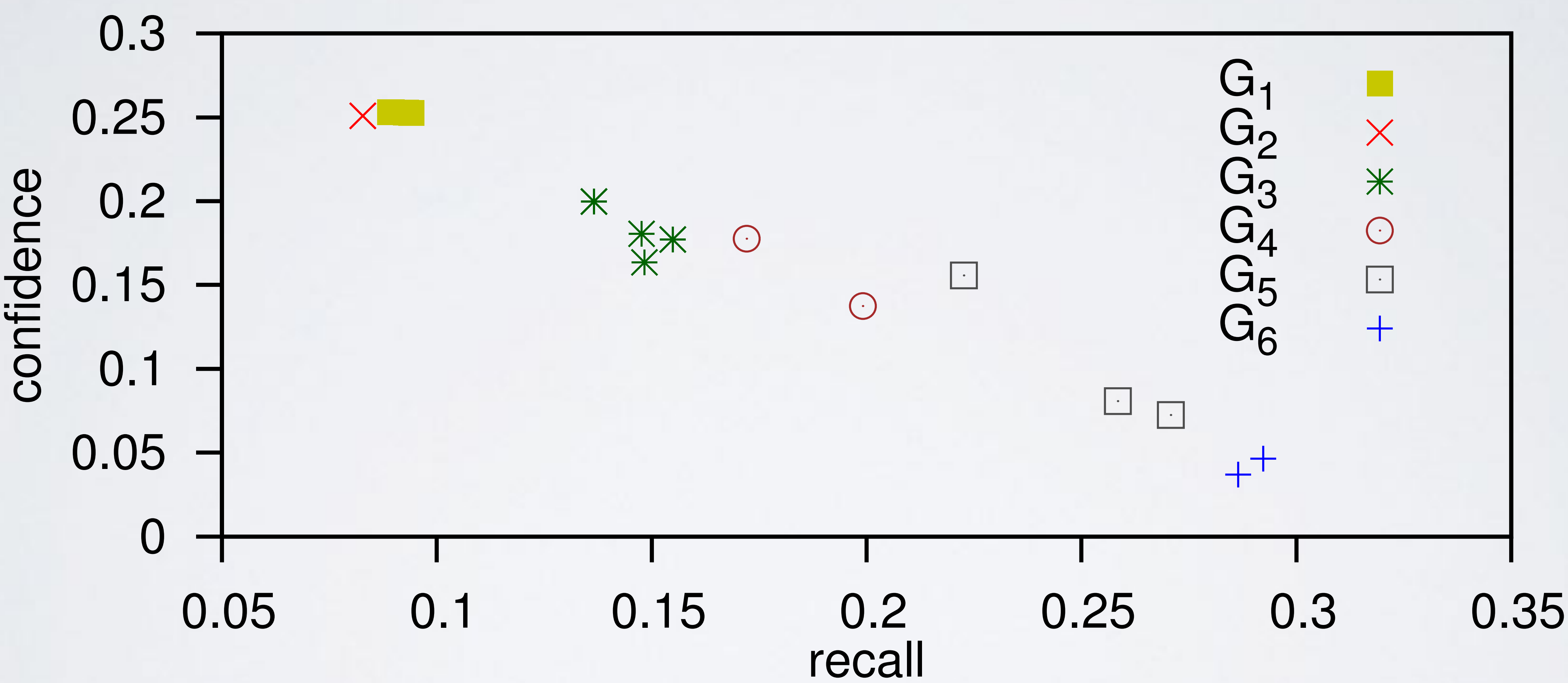



(a) $NDCC$ (b) $\tau$

# GROUPS OF MEASURES

| **Measure** | **Formula** | **Group** |
|---|---|---|
| One-Way Support | $P(B \mid A) \times log_2 \frac{P(AB)}{P(A)P(B)}$ | $G_1^a$ |
| Relative Risk | $P(B \mid A)/P(B \mid \neg A)$ | |
| Odd Multiplier | $\frac{P(AB)P(\neg B)}{P(B)P(A\neg B)}$ | |
| Zhang | $\frac{P(AB)-P(A)P(B)}{max(P(AB)P(\neg B),P(B)P(A\neg B))}$ | |
| Yule's Q $\diamond$ | $\frac{P(AB)P(\neg A\neg B)-P(A\neg B)P(B\neg A)}{P(AB)P(\neg A\neg B)+P(A\neg B)P(B\neg A)}$ | |
| Yule's Y $\diamond$ | $\frac{\sqrt{P(AB)P(\neg A\neg B)}-\sqrt{P(A\neg B)P(B\neg A)}}{\sqrt{P(AB)P(\neg A\neg B)}+\sqrt{P(A\neg B)P(B\neg A)}}$ | |
| Odds Ratio $\diamond$ | $\frac{P(AB)P(\neg A\neg B)}{P(A\neg B)P(B\neg A)}$ | |
| Information Gain $*\ominus$ | $log(P(AB)/(P(A)P(B)))$ | |
| Lift $*\ominus$ | $P(AB)/(P(A)P(B))$ | |
| Added Value $*$ | $P(B \mid A) - P(B)$ | $G_1^b$ |
| Certainty Factor $*$ | $(P(B \mid A) - P(B))/(1 - P(B))$ | |
| Confidence / Precision $*\otimes$ | $P(B \mid A)$ | |
| Laplace Correction $*\otimes$ | $\frac{support(AB)+1}{support(A)+2}$ | |
| Loevinger $\dagger$ | $1 - \frac{P(A)P(\neg B)}{P(A\neg B)}$ | |
| Conviction $\dagger$ | $\frac{P(A)P(\neg B)}{P(A\neg B)}$ | |
| Example and Counter-example Rate | $1 - \frac{P(A\neg B)}{P(AB)}$ | |
| Sebag-Schoenauer | $\frac{P(AB)}{P(A\neg B)}$ | |
| Leverage | $P(B \mid A) - P(A)P(B)$ | |
| Least Contradiction | $\frac{P(AB)-P(A\neg B)}{P(B)}$ | $G_2$ |
| Accuracy | $P(AB) + P(\neg A\neg B)$ | |
| Pearson's $\chi^2$ $\triangleright$ | $\lvert\mathcal{T}\rvert \times \left(\frac{(P(AB)-P(A)P(B))^2}{P(A)P(B)} + \frac{(P(\neg AB)-P(\neg A)P(B))^2}{P(\neg A)P(B)}\right)$ $+\lvert\mathcal{T}\rvert \times \left(\frac{(P(A\neg B)-P(A)P(\neg B))^2}{P(A)P(B)} + \frac{(P(\neg A\neg B)-P(\neg A)P(\neg B))^2}{P(\neg A)P(\neg B)}\right)$ | $G_3$ |
| Gini Index $\triangleright$ | $P(A) \times (P(B \mid A)^2 + P(\neg B \mid A)^2) + P(\neg A) \times (P(B \mid \neg A)^2 +$ $P(\neg B \mid \neg A)^2) - P(B)^2 - P(\neg B)^2$ | |
| J-measure | $P(AB)log(\frac{P(B \mid A)}{P(B)}) + P(A\neg B)log(\frac{P(\neg B \mid A)}{P(\neg B)})$ | |
| $\Phi$ Linear Correlation Coefficient | $\frac{P(AB)-P(A)P(B)}{\sqrt{P(A)P(B)P(\neg A)P(\neg B)}}$ | |
| Two-Way Support Variation | $P(AB) \times log_2 \frac{P(AB)}{P(A)P(B)} + P(A\neg B) \times log_2 \frac{P(A\neg B)}{P(A)P(\neg B)} +$ $P(\neg AB) \times log_2 \frac{P(\neg AB)}{P(\neg A)P(B)} + P(\neg A\neg B) \times log_2 \frac{P(\neg A\neg B)}{P(\neg A)P(\neg B)}$ | |
| Fisher's exact test | $\frac{\binom{\lvert\mathcal{T}\rvert \times P(B)}{\lvert\mathcal{T}\rvert \times P(AB)}\binom{\lvert\mathcal{T}\rvert \times P(\neg B)}{\lvert\mathcal{T}\rvert \times P(A\neg B)}}{\binom{\lvert\mathcal{T}\rvert}{\lvert\mathcal{T}\rvert \times P(A)}}$ | |
| Jaccard | $P(AB)/(P(A) + P(B) - P(AB))$ | |
| Cosine | $\frac{P(AB)}{\sqrt{P(A)P(B)}}$ | $G_4$ |
| Two-Way Support | $P(AB) \times log_2 \frac{P(AB)}{P(A)P(B)}$ | |
| Piatetsky-Shapiro | $P(AB) - P(A)P(B)$ | $G_5$ |
| Klosgen | $\sqrt{P(AB)}max(P(B \mid A) - P(B), P(A \mid B) - P(A))$ | |
| Specificity | $P(\neg B \mid \neg A)$ | |
| Recall | $P(A \mid B)$ | $G_6$ |
| Collective Strength | $\frac{P(AB)+P(\neg B \mid \neg A)}{P(A)P(B)+P(\neg A)P(\neg B)} \times \frac{1-P(A)P(B)-P(\neg A)P(\neg B)}{1-P(AB)-P(\neg B \mid \neg A)}$ | |

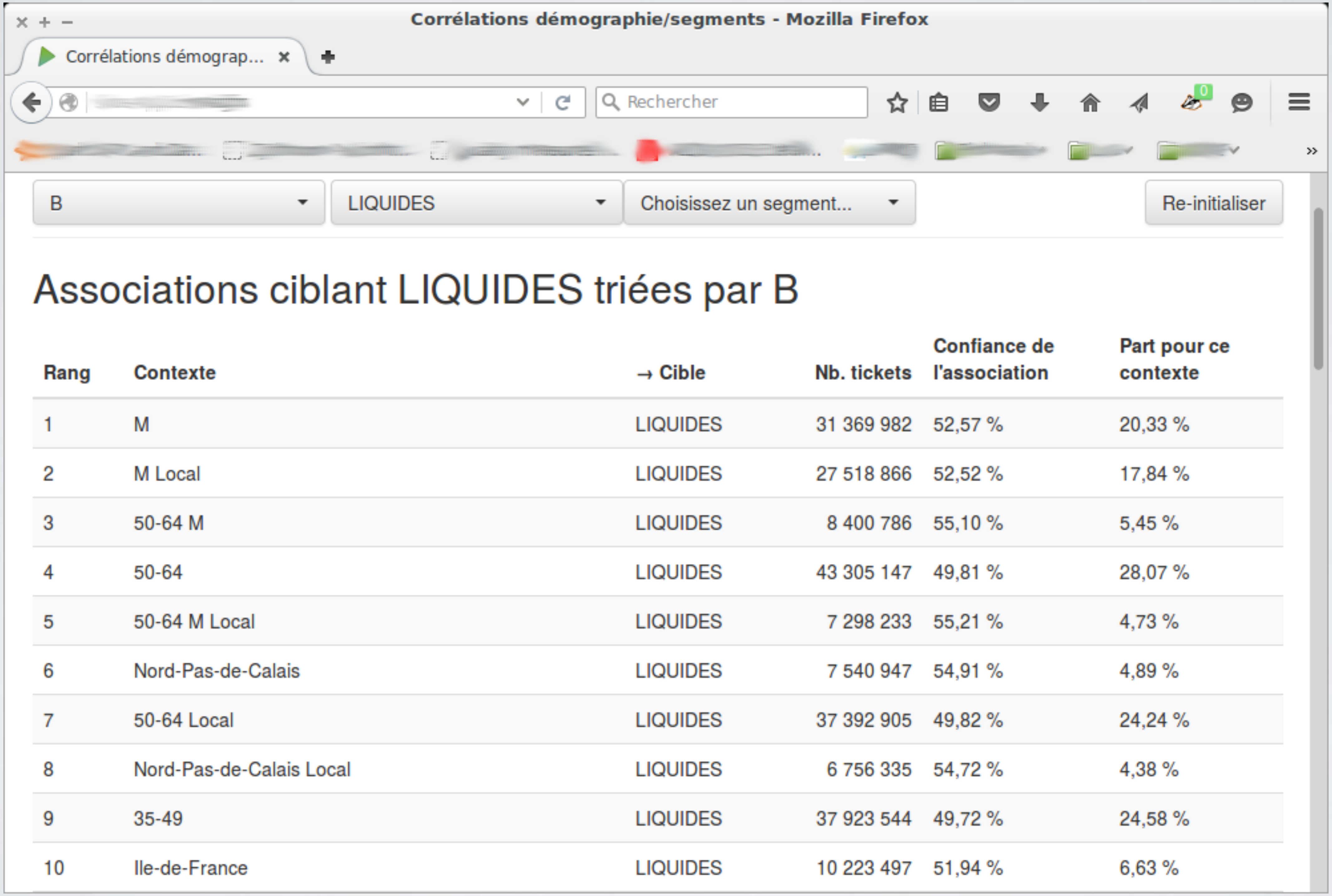
# CHARACTERISTICS OF GROUPS



# USER STUDY

- 2 experienced analysts from Intermarché
- 6 groups of measures
  - 1 representative per group
- For a given target, find which measure highlights the most interesting/usable rules

# USER STUDY INTERFACE




## Associations ciblant LIQUIDES triées par B

| Rang | Contexte | → Cible | Nb. tickets | Confiance de l'association | Part pour ce contexte |
|---|---|---|---|---|---|
| 1 | M | LIQUIDES | 31 369 982 | 52,57 % | 20,33 % |
| 2 | M Local | LIQUIDES | 27 518 866 | 52,52 % | 17,84 % |
| 3 | 50-64 M | LIQUIDES | 8 400 786 | 55,10 % | 5,45 % |
| 4 | 50-64 | LIQUIDES | 43 305 147 | 49,81 % | 28,07 % |
| 5 | 50-64 M Local | LIQUIDES | 7 298 233 | 55,21 % | 4,73 % |
| 6 | Nord-Pas-de-Calais | LIQUIDES | 7 540 947 | 54,91 % | 4,89 % |
| 7 | 50-64 Local | LIQUIDES | 37 392 905 | 49,82 % | 24,24 % |
| 8 | Nord-Pas-de-Calais Local | LIQUIDES | 6 756 335 | 54,72 % | 4,38 % |
| 9 | 35-49 | LIQUIDES | 37 923 544 | 49,72 % | 24,58 % |
| 10 | Ile-de-France | LIQUIDES | 10 223 497 | 51,94 % | 6,63 % |

# USER STUDY RESULTS

- Precision is key
  - Rules feel reliable and usable
- Unexpected results are unsettling

- Filtering eliminates surprises
  - Precision/Recall trade-off favored
- Scrolling for anti-associations

# CAPA CONCLUSION

- Generic framework for comparing rankings
  - In retail, Lift is a good choice
  - How do we promote anti-associations?
- Two-phase ranking to combine efficiency and quality
  - TopPI mines top-1000 using support, then CAPA re-ranks for top-50 using Lift