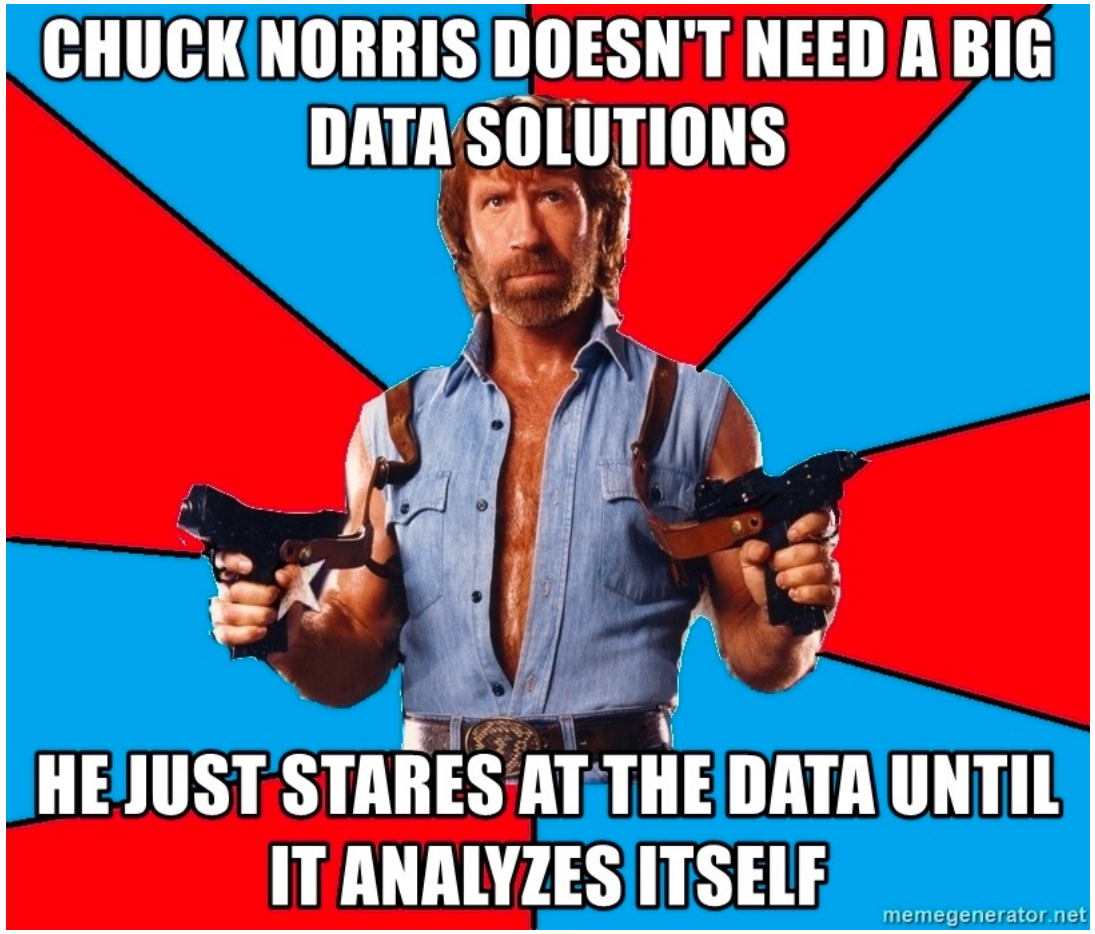
CHUCK NORRIS DOESN'T NEED A BIG DATA SOLUTIONS
HE JUST STARES AT THE DATA UNTIL IT ANALYZES ITSELF
memegenerator.net

HEALTH INC.

Health Insurers Are Vacuuming Up Details About You — And It Could Raise Your Rates

July 17, 2018 · 5:00 AM ET
Heard on Morning Edition

https://www.npr.org/sections/health-shots/2018/07/17/629441555/health-insurers-are-vacuuming-up-details-about-you-and-it-could-raise-your-rates?utm_source=facebook.com&utm_medium=social&utm_campaign=npr&utm_term=nprnews&utm_content=20180717&fbclid=IwAR1TVQOKKp00zMQu50IwEpLiPiyzFrMqz81p_OyiN-WEHkJDC7H5eb43LCc

MARSHALL ALLEN

FROM P

3-Minute Listen + PLAYLIST

Justin Volz for ProPublica

The companies are tracking your race, education level, TV habits, marital status, net worth. They're collecting what you post on social media, whether you're behind on your bills, what you order online. Then they feed this information into complicated computer algorithms that spit out predictions about how much your health care could cost them.

Are you a woman who recently changed your name? You could be newly married and have a pricey pregnancy pending. Or maybe you're stressed and anxious from a recent divorce. That, too, the computer models predict, may run up your medical bills.

Low-income and a minority? That means, the data brokers say, you are more likely to live in a dilapidated and dangerous neighborhood, increasing your health risks.

"We sit on oceans of data," said Eric McCulley, director of strategic solutions for LexisNexis Risk Solutions. And he isn't apologetic about using it. "The fact is, our data is in the public domain," he said. "We didn't put it out there."

Scalability of Association Analysis

- Market basket data can be huge
- Database (I/O) cost
- Merging work across distributed systems
- Memory size limitations
- The Tyranny of Counting Pairs
  - The most memory is required for determining frequent pairs

Apriori Rule Gen

Rule Generation

- Want to generate association rules from the list of frequent itemsets to find those that meet *minconf*
- Brute force: All non-empty subsets of every frequent itemset, f, are enumerated. And for every such subset, a, a rule of the form a → (f-a) is generated and tested against *minconf*

Ex: If {A,B,C,D} is a frequent itemset, candidate rules would be:

A →BCD	B →ACD
AB →CD	BC →AD
AC →BD	BD →AC
AD →BC	BCD →A
ABC →D	C →ABD
ABD →C	CD →AB
ACD →B	D →ABC

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

Calculating Confidence

Frequent Itemsets	Support Count
{Bread}	5
{Milk}	5
{Beer}	3
{Diaper}	5
{Bread, Milk}	4
{Bread, Diaper}	4
{Milk, Diaper}	4
{Beer, Diaper}	3
{Bread, Diapers, Milk}	3

- Computing confidence does not require any additional scans of the data!

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

$$c(\{\text{Bread, Diapers}\} \rightarrow \{\text{Milk}\}) = \frac{\sigma(\{Bread, Diapers, Milk\})}{\sigma(\{Bread, Diapers\})} = \frac{3}{4}$$

Confidence Based Pruning

- Confidence of rules generated from the **same itemset** have an anti-monotone property:
- Confidence is anti-monotone with regards to number of items on the right-hand side of the rule
- Ex: Frequent Itemset = {A,B,C,D}:

$$c(\{A,B,C\} \rightarrow \{D\}) \geq c(\{A,B\} \rightarrow \{C,D\}) \geq c(\{A\} \rightarrow \{B,C,D\})$$

- This does not apply to rules that are not generated from the same itemset!

  $c(\{A,B\} \rightarrow \{C,D\})$ can be larger or smaller than $c(\{A,B\} \rightarrow \{D\})$

Rules with ‘a’ in the consequent can be pruned because of the anti-monotone property of confidence.

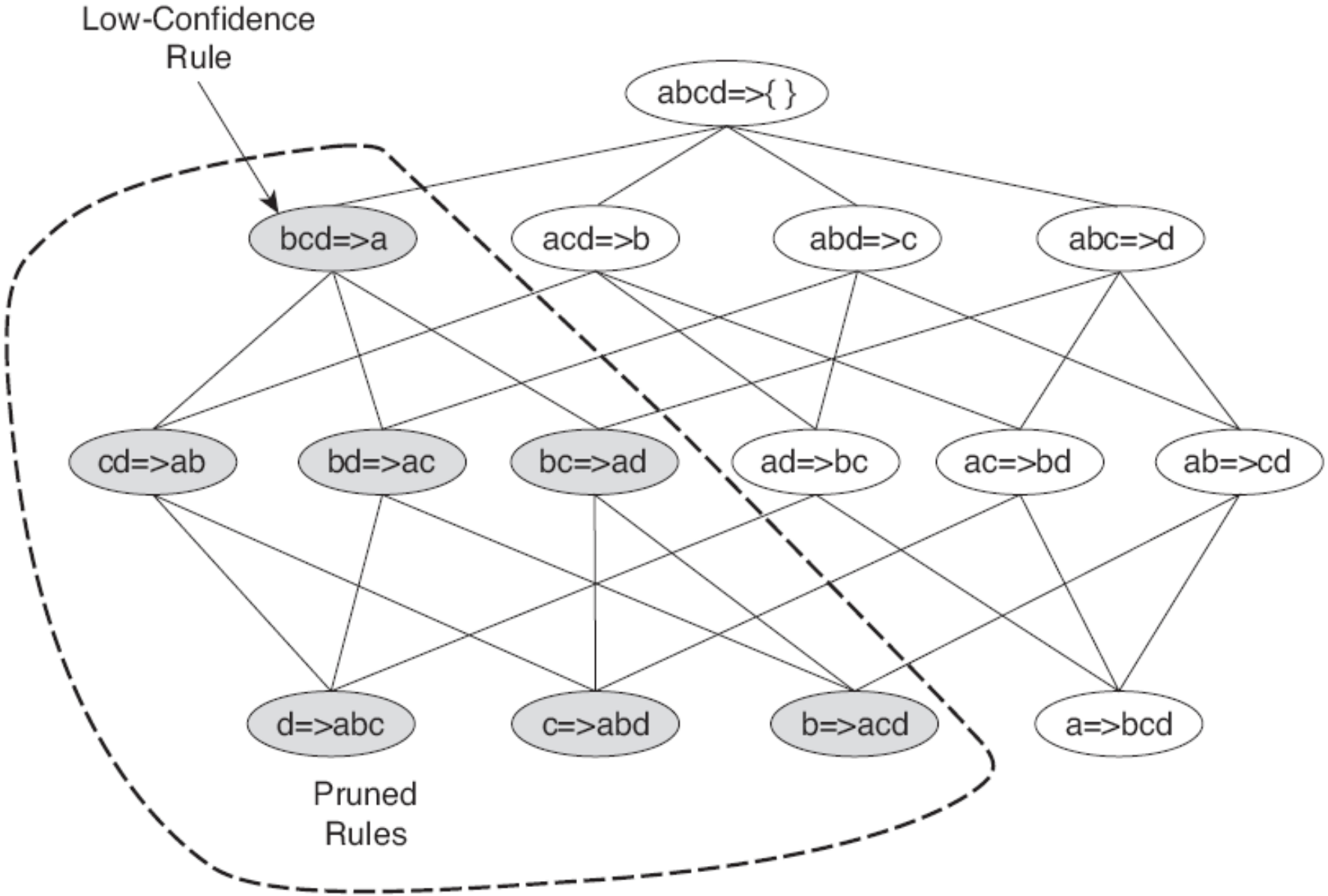

Figure 6.15. Pruning of association rules using the confidence measure.

Rule Generation Algorithm

- Apriori Rule Gen uses a level-wise approach for generating association rules
- Initially, all rules that have only one item in the consequent are generated and tested against minconf

Example: Frequent Itemset = {A,B,C,D}

Make all possible rules with 1 item in the consequent:

$ABC \rightarrow D$

$ABD \rightarrow C$

$ACD \rightarrow B$

$BCD \rightarrow A$

Check the confidence of these candidate rules.

Rule Generation Algorithm

- Apriori Rule Gen uses a level-wise approach for generating association rules
- Initially, all rules that have only one item in the consequent are generated and tested against minconf
- The high-confidence rules that are found are then used to generate the next round of candidate rules by merging consequents

Example: Frequent Itemset = {A,B,C,D}

If these three rules meet minconf:

$ABC \rightarrow D$

$ABD \rightarrow C$

$ACD \rightarrow B$

Consequents are merged to make:

$AB \rightarrow CD$

$AC \rightarrow BD$

$AD \rightarrow BC$

Now check the confidence of these candidate rules.

Rule Generation Algorithm

- Apriori Rule Gen uses a level-wise approach for generating association rules
- Initially, all rules that have only one item in the consequent are generated and tested against minconf
- The high-confidence rules that are found are then used to generate the next round of candidate rules by merging consequents

Example: Frequent Itemset = {A,B,C,D}

If these two rules meet minconf,
$AD \rightarrow BC$
$AC \rightarrow BD$

Consequents are merged to make:
$A \rightarrow BCD$

Now check the confidence of this candidate rule.

Rule Generation Algorithm

- Repeat the Rule Gen algorithm on **all** of the frequent itemsets.
- All rules that meet minconf are **strong rules**.

Rule Generation Practice

Find the strong rules from only the following frequent itemsets, using minconf=70%:

{A,B,C,D}
{A,B,C}
{A,B,D}
{A,B}
{A,D}

Frequent Itemsets	Support Count
{A}	6
{B}	7
{C}	6
{D}	4
{A,B}	4
{A,C}	2
{A,D}	3
{B,C}	2
{B,D}	3
{C,D}	3
{A,B,C}	2
{A,B,D}	3
{A,C,D}	2
{B,C,D}	2
{A,B,C,D}	2