Mining for Contrasting Sets (STUCCO)

Camilo Arango
Department of Computing Science
University of Alberta

What is Contrast set mining

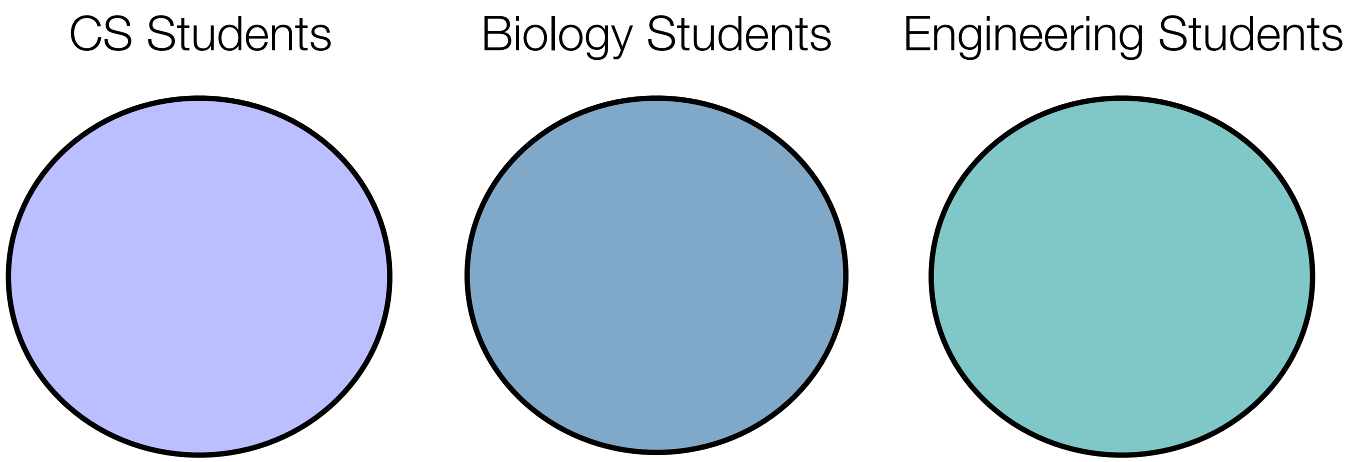
- Finding differences among groups
- Example questions:
  - Health: Which symptoms differentiate similar diseases?
  - Marketing: What are the differences between customers that spend less money and those who spend more in a particular kind of item?
  - Analysis of census data: What is the difference between people holding Ph.D. degrees and people holding Bachelor degrees?

Outline

- Definition of the problem
- STUCCO algorithm
  - Basic idea
  - Controlling error
  - Filtering of results
  - Evaluation
- Conclusions

Example

- How do prospective students for different departments differ from each other?



Data Model

- Data is a set of k-dimensional vectors where each component can take a finite number of discrete values.

Prospective Students ↓

Age	Sex	Born in US	SAT-M > 700	SAT-V > 700	Admitted
<20	M	yes	yes	yes	yes
20-25	M	yes	no	yes	no
25-30	F	no	yes	no	yes

k = 6

...

- Age = {<20, 20-25, 25-30, >30}
- Sex = {M, F}
- Born in us = {yes, no}
- SAT-M > 700 = {yes, no}
- SAT-V > 700 = {yes, no}
- Admitted = {yes, no}

Data Model

- The vectors are organized into mutually exclusive groups

	Age	Sex	Born in US	SAT-M >700	SAT-V >700	Admit
CS	<20	F	yes	yes	no	yes
	20-25	M	no	no	yes	no
	<20	F	no	yes	yes	yes
Biology	20-25	M	yes	yes	no	yes
	<20	F	yes	no	yes	no
	<20	F	no	yes	no	yes
Engineering	<20	M	yes	no	yes	yes
	20-25	M	yes	no	no	no
	25-30	F	yes	yes	no	yes
	<20	F	yes	no	yes	yes

Contrast Sets

- Differences among groups are expressed as **contrast-sets**
- A **contrast-set** is a conjunction of attribute-value pairs.

Examples

Admitted = no

Sex = F ∧ Born in US = no

Age = 20-25 ∧ Admitted = yes ∧ SAT-V > 700 = no

Support of Contrasts sets

- **Support of a contrast set in group G**: % of examples in G where the contrast set is true.

Group	Age	Sex	Born in US	SAT-M >700	SAT-V >700	Admit
CS	<20	F	yes	yes	no	yes
	20-25	M	no	no	yes	no
	<20	F	no	yes	yes	yes
Biology	20-25	M	yes	yes	no	yes
	<20	F	no	no	yes	no
	<20	F	no	yes	no	yes
Engineering	<20	M	yes	no	yes	yes
	20-25	M	yes	no	no	no
	25-30	F	yes	yes	no	yes
	<20	F	yes	no	yes	yes

sup (Sex = F $\wedge$ Born in US = no | CS) = 1 / 3 = 33%

sup (Sex = F $\wedge$ Born in US = no | Biology) = 2 / 3 = 66%

sup (Sex = F $\wedge$ Born in US = no | Biology) = 0 / 3 = 0%

Problem of finding Contrast Sets

- We want to find the contrasts sets that make one group different than another.

- In other words, we want to find the contrast-sets whose support differs *meaningfully* across groups. This contrast-sets are called **deviations**.

How can we determine this?

Defining deviations

- A **deviation** is a contrast set that is **significant** and **large**
- A contrast-set for which at least two groups *differ* in their support is called ***Significant*.**
- A contrast-set for which the *maximum difference* between supports is greater than a parameter *mindev*, is called **Large.**

Example

For the contrast set c1: *“admitted = yes ∧ age 20-25” and **mindev** = 5%*

support (admitted = yes ∧ age 20-25 | CS) = 11%

support (admitted = yes ∧ age 20-25 | Bio) = 15%

support (admitted = yes ∧ age 20-25 | Eng) = 18%

Deciding if a contrast set is large is easy:

max difference = 18% - 11% = 7%

With *mindev* = 5%, *c1* is **large**

To decide if a contrast set is significant, we use an **statistical test**

STUCCO

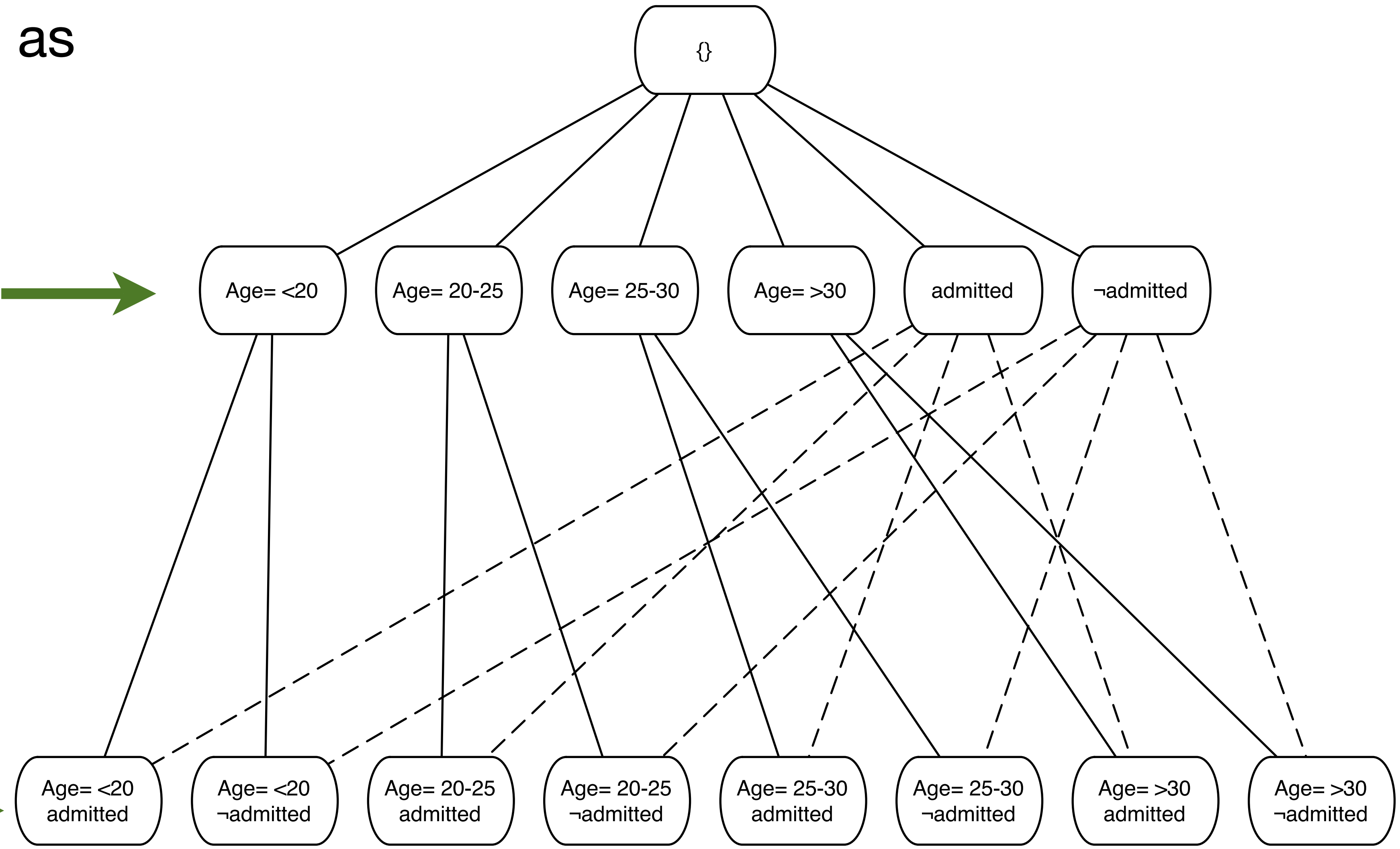
- An algorithm to find contrasts sets
- Stands for “**S**earch and **T**esting for **U**nderstandable **C**onsistent **Co**ntrast”.
- Presented by Stephen D. Bay and Michael J. Pazzani in SIGKDD 1999

STUCCO

- The problem is modeled as tree-search

All possible attribute-value pairs

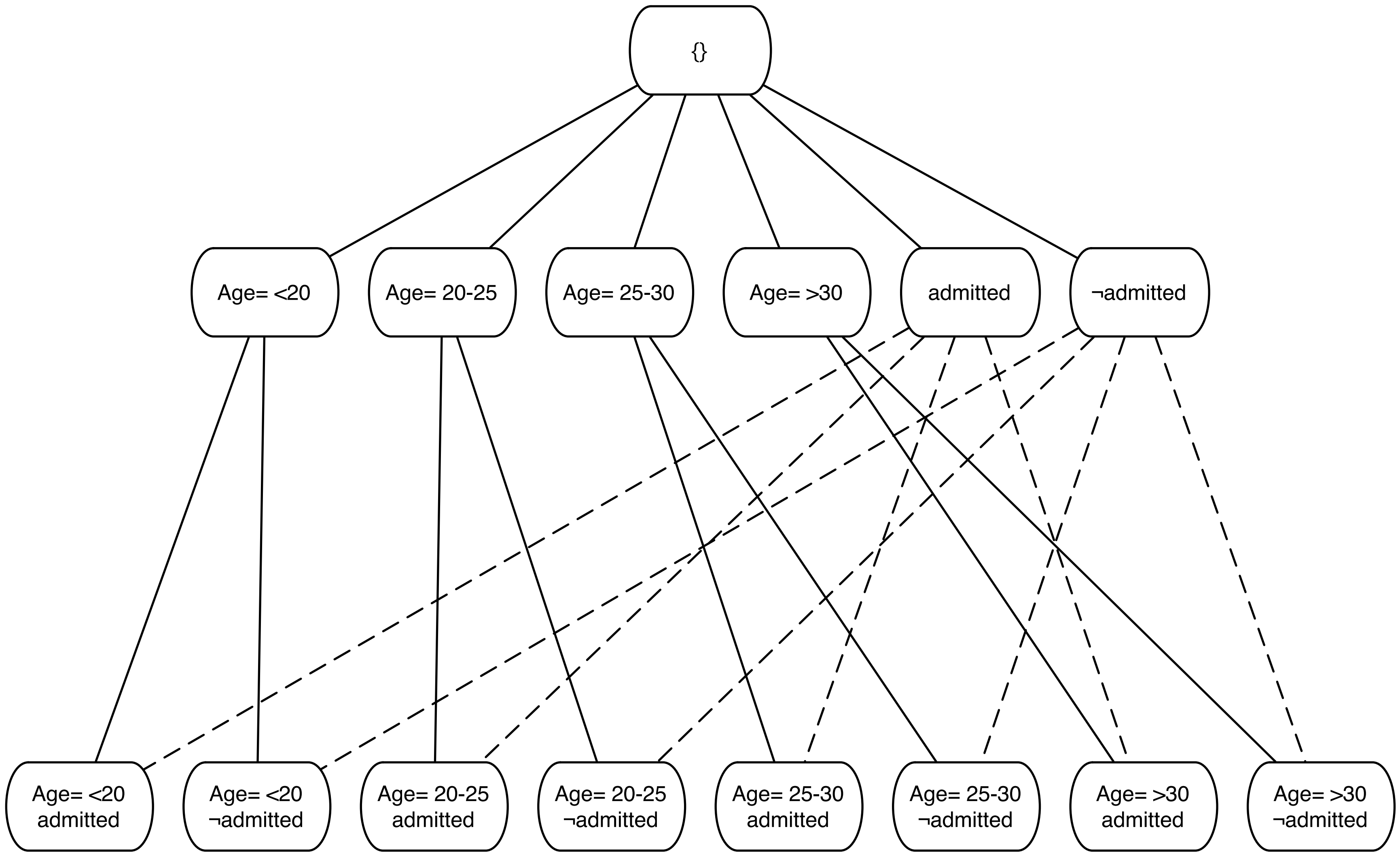
Conjunction of 2 attribute-value pairs



- Age = {<20, 20-25, 25-30, >30}
- Admitted = {yes, no}

STUCCO

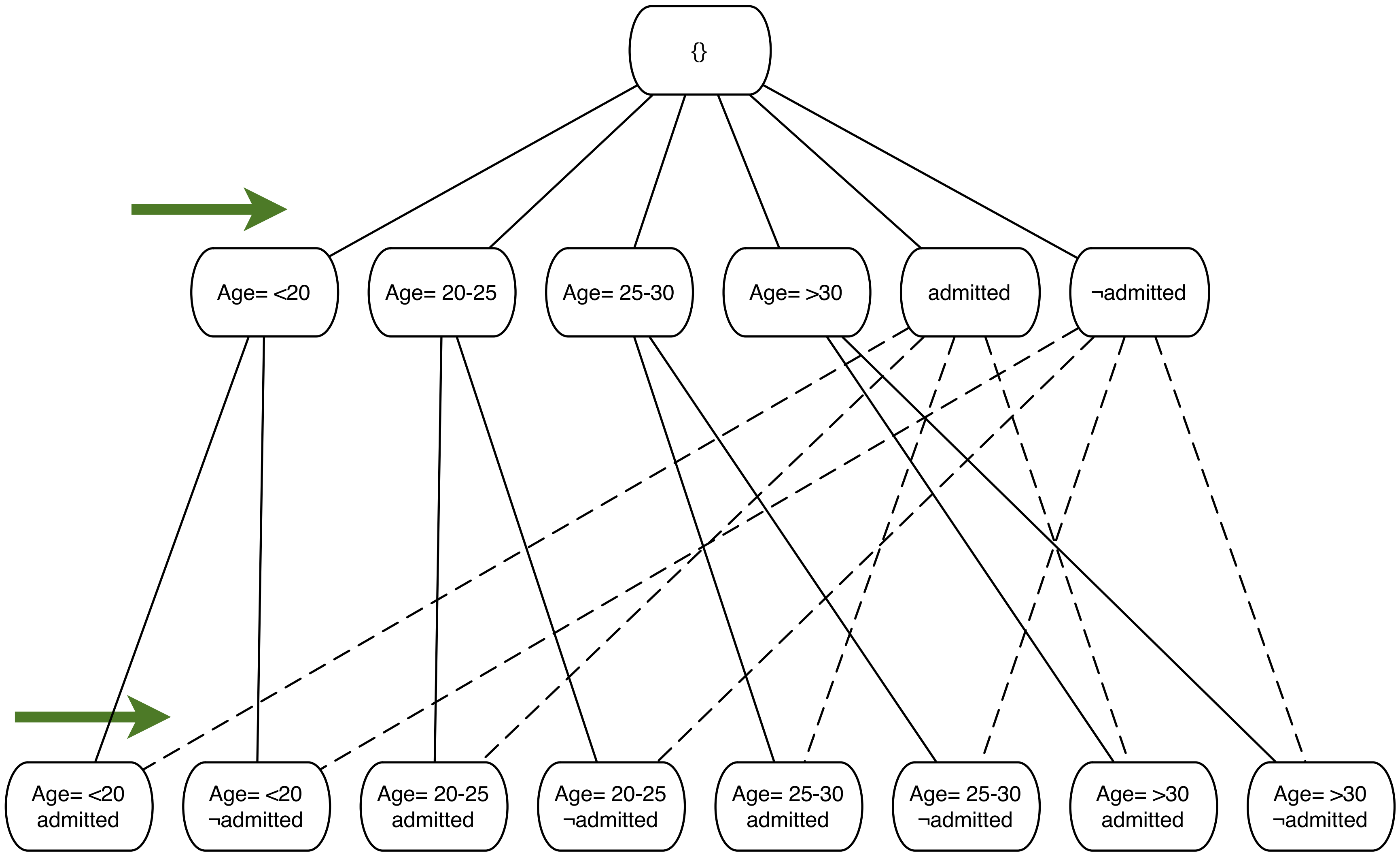
- Uses a breadth first, level by level approach.
- For each level
  - Scan database and count support for each group.
  - Determine if each node is *significant* and *large*.
  - Determine if each the node should be *pruned*.
- Display all first order deviations.
- Display other deviations only if they are *surprising*.



- Age = {<20, 20-25, 25-30, >30}
- Admitted = {yes, no}

STUCCO

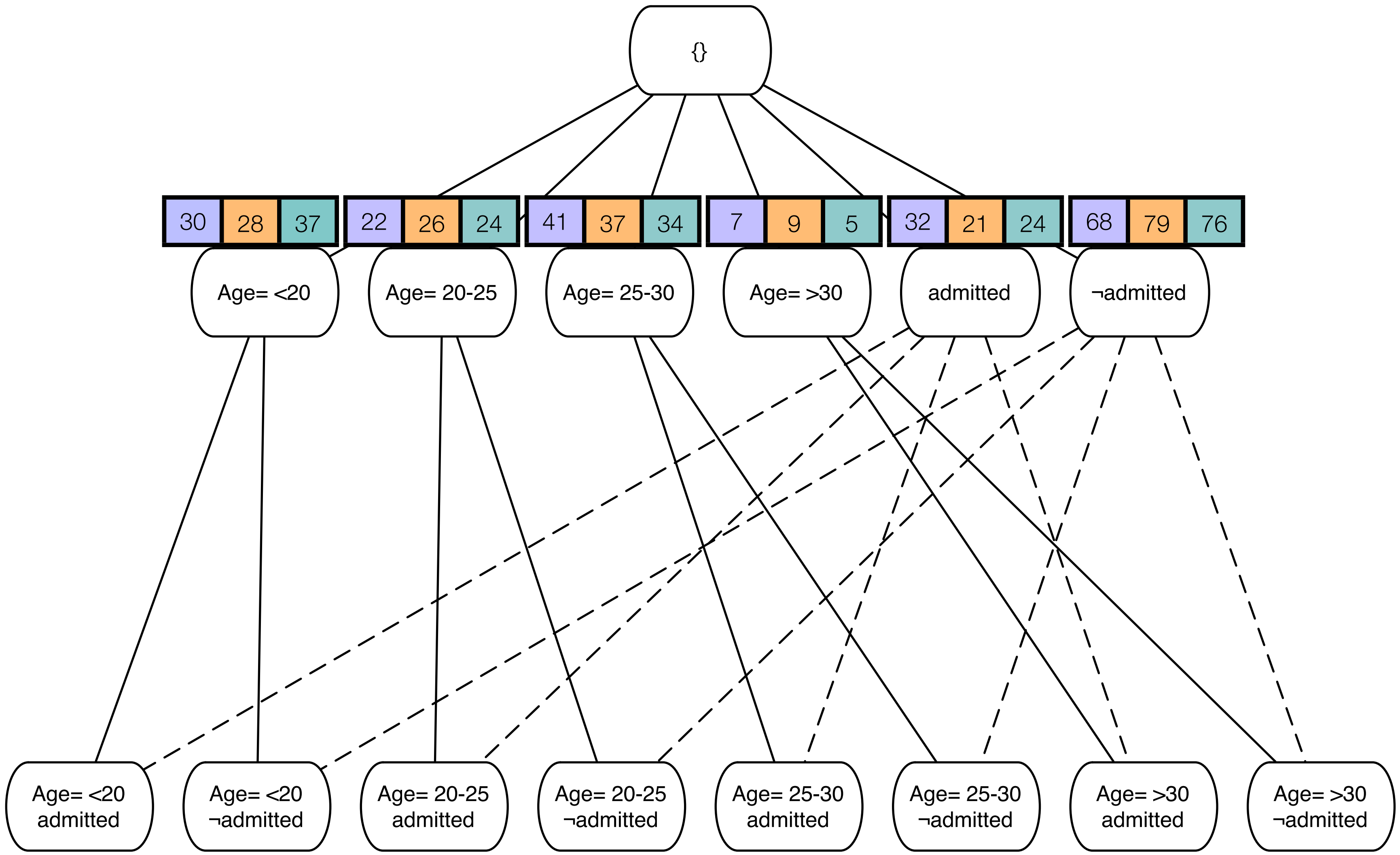
- Uses a breadth first, level by level approach.
- For each level
  - Scan database and count support for each group.
  - Determine if each node is *significant* and *large*.
  - Determine if each the node should be *pruned*.
- Display all first order deviations.
- Display other deviations only if they are *surprising*.



- Age = {<20, 20-25, 25-30, >30}
- Admitted = {yes, no}

STUCCO

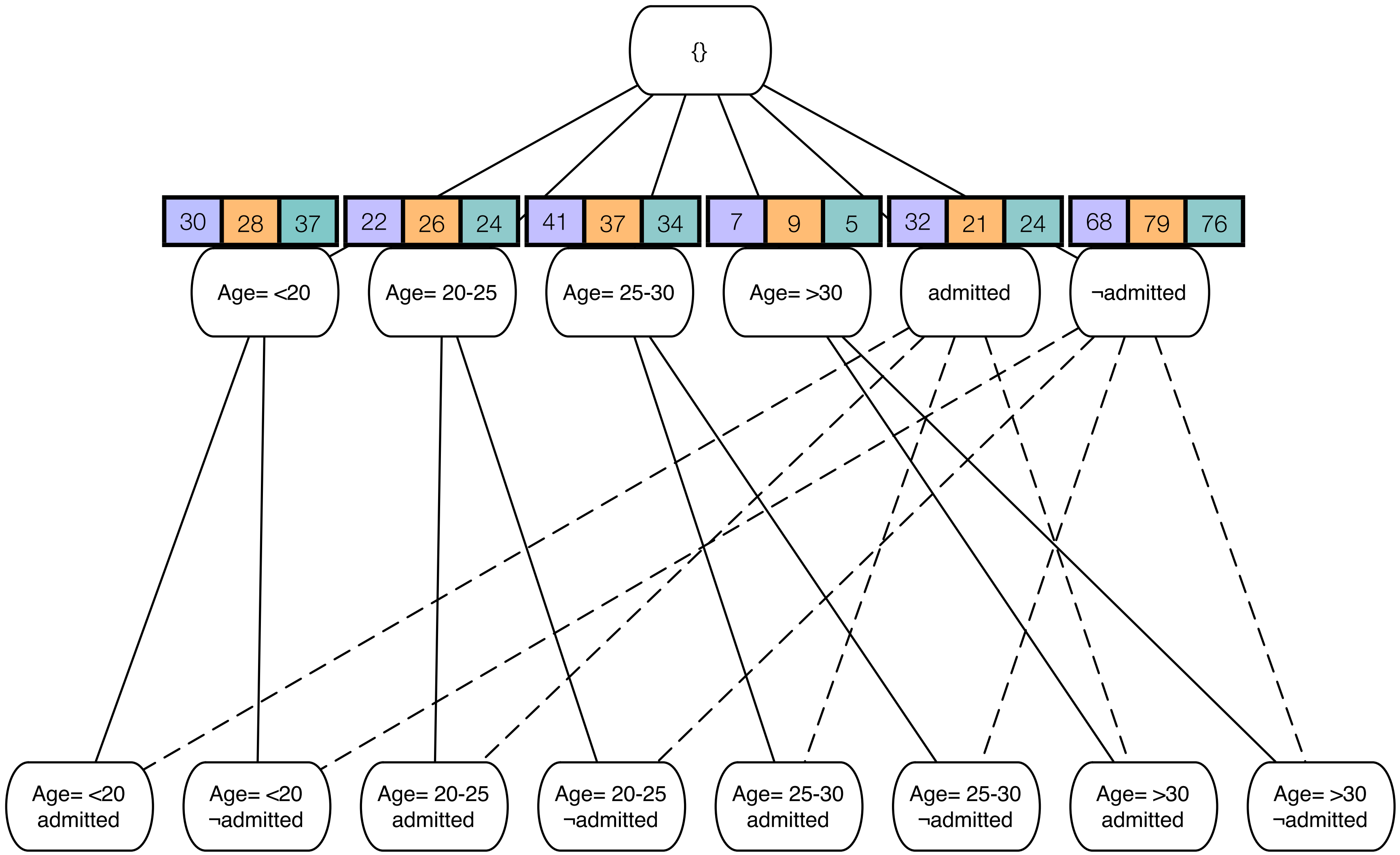
- Uses a breadth first, level by level approach.
- For each level
    - Scan database and count support for each group.
    - Determine if each node is *significant* and *large*.
    - Determine if each the node should be *pruned*.
- Display all first order deviations.
- Display other deviations only if they are *surprising*.



- Age = {<20, 20-25, 25-30, >30}
- Admitted = {yes, no}

STUCCO

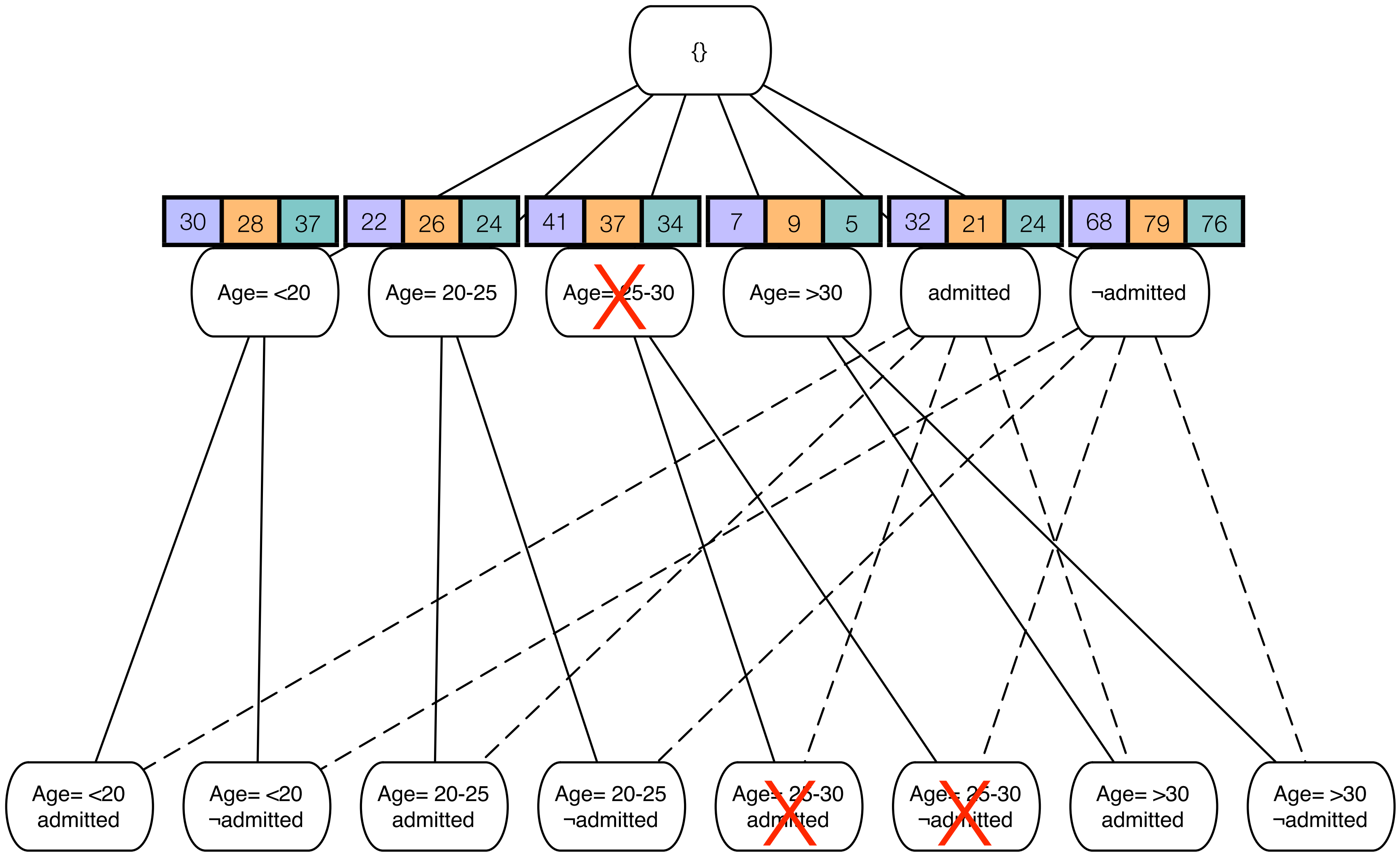
- Uses a breadth first, level by level approach.
- For each level
  - Scan database and count support for each group.
  - Determine if each node is ***significant*** and ***large***.
  - Determine if each the node should be *pruned*.
- Display all first order deviations.
- Display other deviations only if they are *surprising*.



- Age = {<20, 20-25, 25-30, >30}
- Admitted = {yes, no}

STUCCO

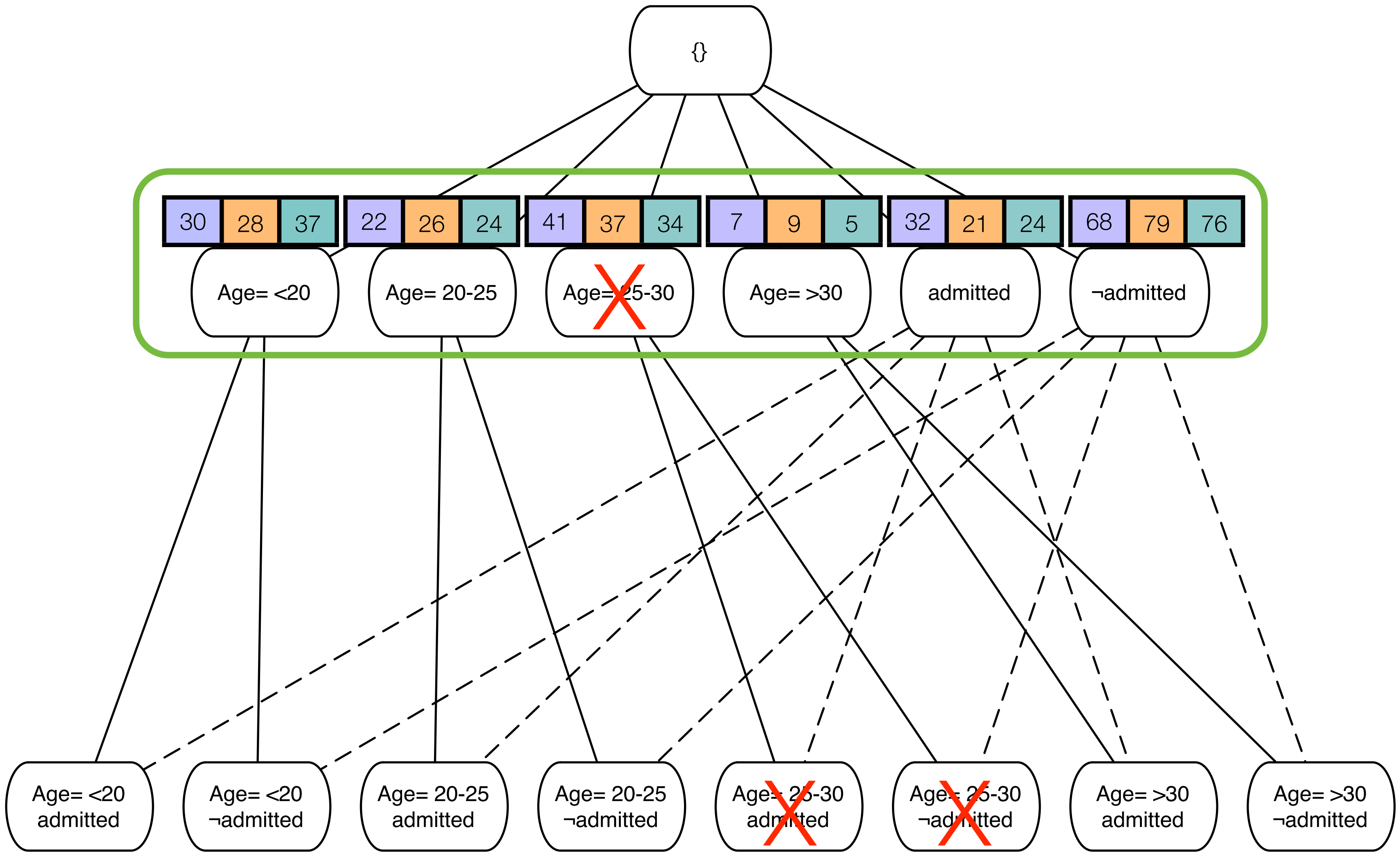
- Uses a breadth first, level by level approach.
- For each level
  - Scan database and count support for each group.
  - Determine if each node is *significant* and *large*.
  - **Determine if each the node should be *pruned*.**
- Display all first order deviations.
- Display other deviations only if they are *surprising*.



- Age = {<20, 20-25, 25-30, >30}
- Admitted = {yes, no}

STUCCO

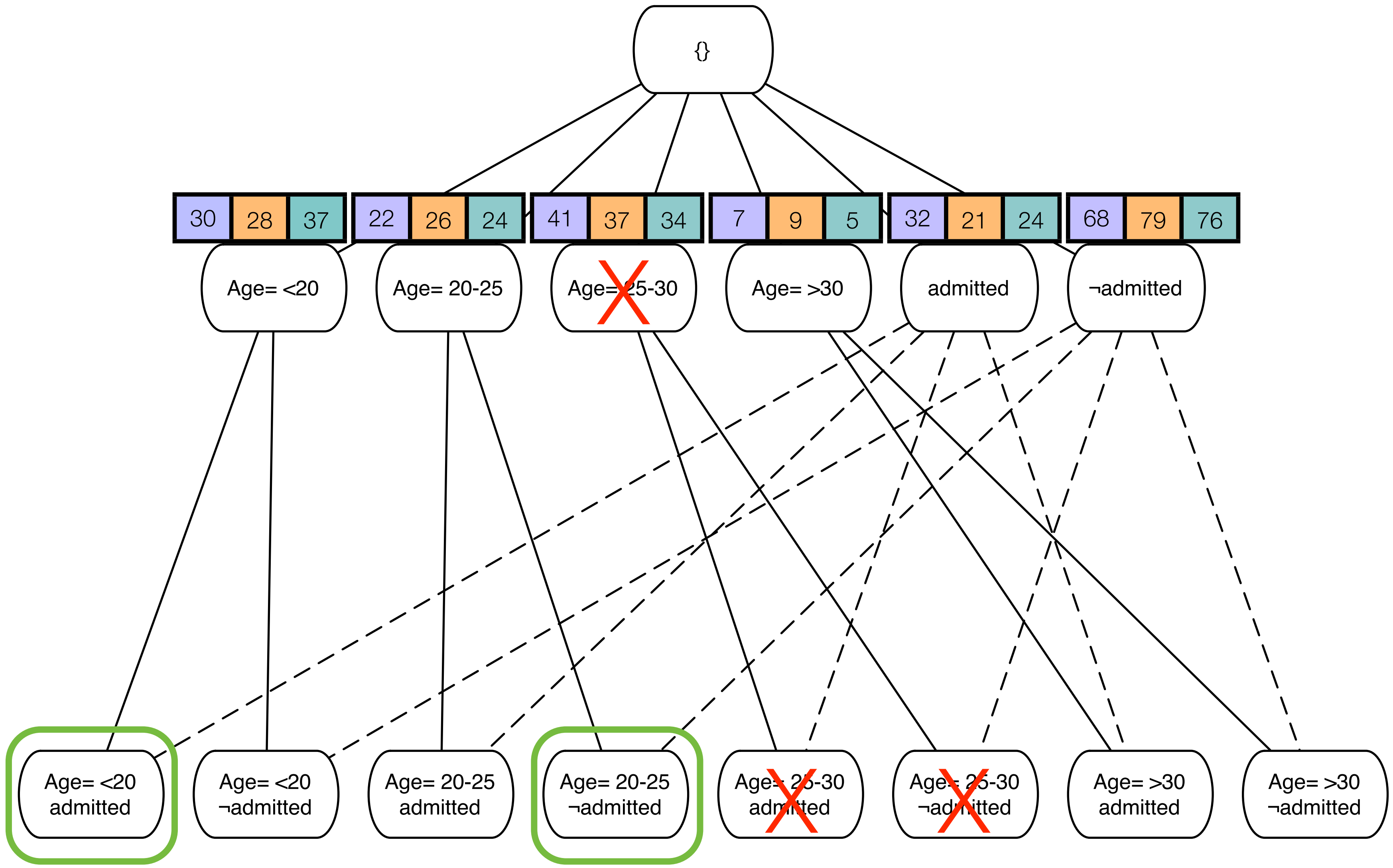
- Uses a breadth first, level by level approach.
- For each level
  - Scan database and count support for each group.
  - Determine if each node is _significant_ and _large_.
  - Determine if each the node should be _pruned_.
- Display all first order deviations.
- Display other deviations only if they are _surprising_.



- Age = {<20, 20-25, 25-30, >30}
- Admitted = {yes, no}

STUCCO

- Uses a breadth first, level by level approach.
- For each level
  - Scan database and count support for each group.
  - Determine if each node is *significant* and *large*.
  - Determine if each the node should be *pruned*.
- Display all first order deviations.
- Display more specific deviations only if they are *surprising*.



- Age = {<20, 20-25, 25-30, >30}
- Admitted = {yes, no}

STUCCO

- Uses a breadth first, level by level approach.
- For each level
  - Scan database and count support for each group.
  - Determine if each node is ***significant*** and *large*.
  - Determine if each the node should be *pruned*.
- Display all first order deviations.
- Display other deviations only if they are *surprising*.

- A contrast set for which at least two groups *differ* in their support is called ***Significant*.**
- Perform an statistical test (chi-square) for the contrast-set:
  - Null hypothesis: *"The support for the contrast-set is the same across all groups"*
  - Compute the χ^2 statistic
  - Check the value of the chi-square distribution
  - It must be less than a threshold **α**. (typically, α=0.05)

STUCCO

- Uses a breadth first, level by level approach.
- For each level
  - Scan database and count support for each group.
  - Determine if each node is ***significant*** and *large*.
  - Determine if each the node should be *pruned*.
- Display all first order deviations.
- Display other deviations only if they are *surprising*.

- To compute the χ^2 statistic we build a *2 x c* contingency table, where c is the number of groups:

c1: *"admitted = yes ∧ age 20-25"*

	CS	Bio	Eng
c1	11	15	18
¬ c1	33	11	50

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$O \rightarrow$ *Observed values*

$$E_{ij} = \frac{\sum_{i=1}^{2} O_{ij} \sum_{j=1}^{c} Oij}{N}$$

$E \rightarrow$ *Expected values*

$N \rightarrow$ *total number of observations*

STUCCO

- Uses a breadth first, level by level approach.
- For each level
  - Scan database and count support for each group.
  - Determine if each node is ***significant*** and *large*.
  - Determine if each the node should be *pruned*.
- Display all first order deviations.
- Display other deviations only if they are *surprising*.

- **How to choose α?**
  - In chi-square tests typically, α=0.05
  - α is the max probability of falsely rejecting the null hypothesis (*Type I error*).
  - That means that if we perform *1000* tests, an average of *50* Type I errors will be made!
  - **Solution**:

  Decrease the value of α progressively for each level in the tree

$$\alpha_l = min\left(\frac{\frac{\alpha}{2^l}}{|C_l|}, \alpha_{l-1} \right)$$

$|C_l| \rightarrow$ *Number of candidates in level l of the tree*

STUCCO

- Uses a breadth first, level by level approach.
- For each level
  - Scan database and count support for each group.
  - Determine if each node is *significant* and *large*.
  - Determine if each the node should be ***pruned***.
- Display all first order deviations.
- Display other deviations only if they are *surprising*.

- **Pruning strategies**:
  1. **Minimum deviation size**: If the support for a node is smaller than *mindev* for all groups.
  2. **Expected Cell Frequencies**: If the expected of a node is too small, the χ^2 test is not valid. It will also be invalid for the children.
  3. **Chi-Square Bounds**: It is possible to calculate an upper bound to the χ^2 statistic for all children of a node. If it is not high enough to pass the α test for that level, the node can be pruned.

STUCCO

- Uses a breadth first, level by level approach.
- For each level
  - Scan database and count support for each group.
  - Determine if each node is *significant* and *large*.
  - Determine if each the node should be *pruned*.
- Display all first order deviations.
- Display other deviations only if they are ***surprising***.

• Filtering Deviations

- A contrast set is considered to be ***surprising*** if their support is different from what is expected.

$P\ (Age > 30 \mid Bio) = 0.09$

$P\ (Admitted = true \mid Bio) = 0.21$

- Assuming independency, then we expect :

$P_e\ (Age > 30 \wedge Admitted = true \mid Bio)$

$= 0.09 * 0.21$

$= 0.02$

Surprise!

- If the actual value is different, the result is surprising

$P_e\ (Age = 30+ \wedge Admitted = true \mid Bio) = 0.1$

Evaluation

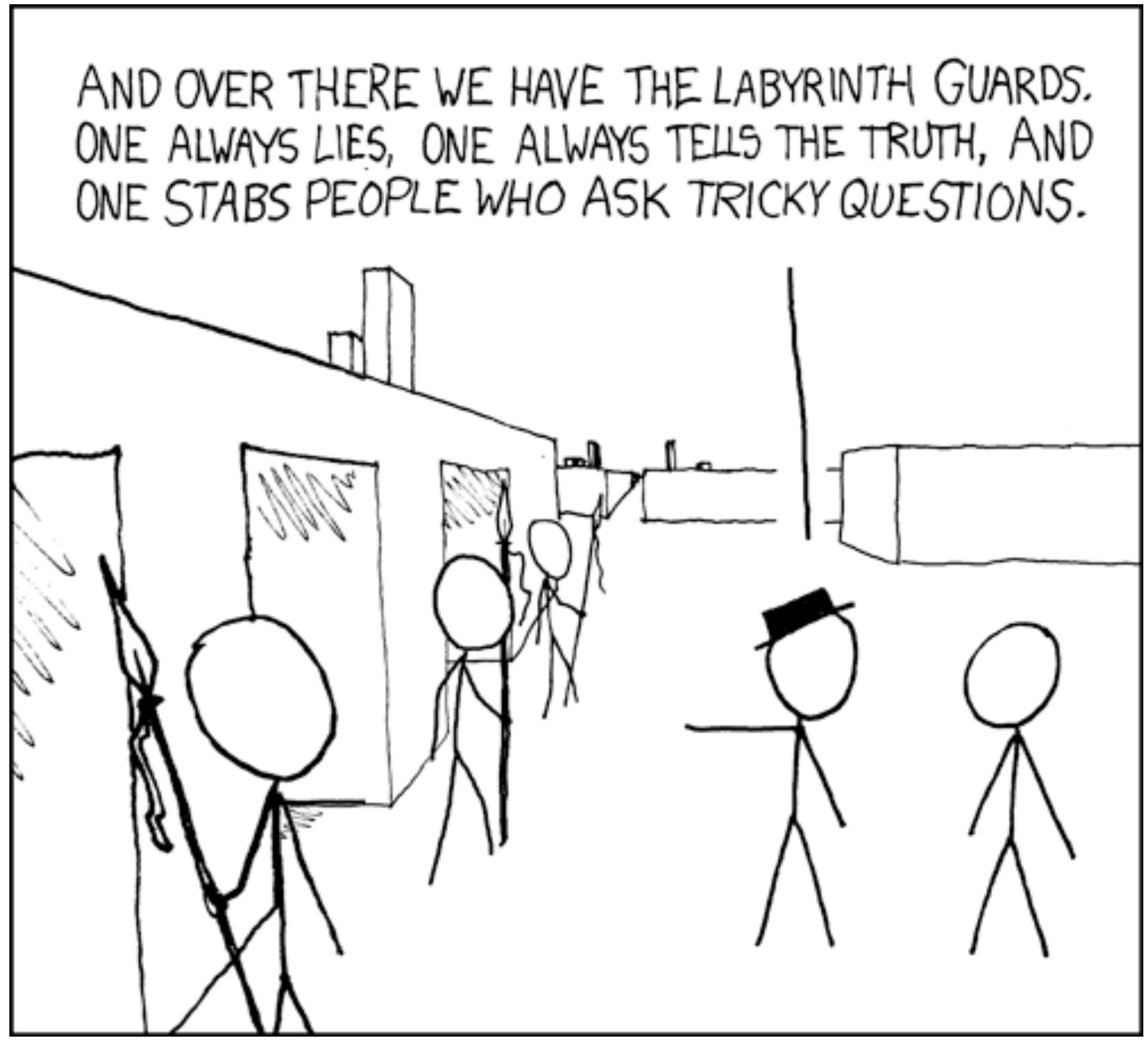
- Using the Adult census data taken from the UCI database.
  - "What are the differences between people with Ph.D. and Bachelor degrees? (mindev = 1%, α = 0.05)
    - STUCCO found 10000 deviations. Most of them not surprising, so reduced to 164.
    - Apriori found 75000 rules in the dataset.

	Observed %		Expected %			
Contrast-Set	Ph.D.	Bach.	Ph.D.	Bach.	χ^2	p
workclass = State-gov	21.0	5.4			225.1	6.9e-51
occupation = sales	2.7	15.8			74.9	4.8e-18
hour per week > 60	8.4	3.2			43.4	4.4e-11
native country = U.S.	80.5	89.5			45.9	1.3e-11
native country = Canada	1.9	0.5			18.6	1.6e-5
native country = India	1.6	0.5			15.2	9.5e-5
salary > 50K	72.6	41.3			220.2	8.3e-50
sex = male ∧ salary > 50K	61.8	34.8	58.8	28.5	173.6	1.2e-39
occupation = prof-specialty ∧ sex = female ∧ salary > 50K	7.6	2.6	10.7	3.5	48.2	3.8e-12

Conclusion

- Contrast-set mining studies techniques for finding differences across several contrasting groups.
- The STUCCO algorithm
  - Uses statistical hypothesis testing to find significant differences.
  - Provides control over false positives.
  - Implements several pruning techniques.
  - Summarization of results.

Questions?



xkcd.com

References

[1] Stephen D. Bay, Michael J. Pazzani. *Detecting Change in Categorical Data: Mining Contrast Sets*. In Proc. 1999 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining

[2] Amit Satsangi, Osmar R. Zaïane. *Contrasting the Contrast Sets: An Alternative Approach*. Database Engineering and Applications Symposium, 2007

[3] Stephen D. Bay, Michael J. Pazzani. *Detecting Group Differences: Mining Contrast Sets*. Data Mining and Knowledge Discovery. Volume 5, Number 3 / July, 2001. Pages 213-246. Springer Netherlands.