

Introduction to Data Science

Lecture 6; May 2nd, 2016

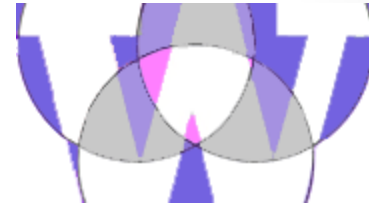
Ernst Henle

ErnstHe@UW.edu

Skype: ernst-henle

(1)

Agenda



- Announcements
 - Encourage Group Homework and ask questions on LinkedIn
 - Guest Lectures in May
 - Business side of Data Science by Marius Marcu on May 9th 2016
 - Data Visualization by Tatyana Yakushev on May 16th 2016
 - Building a Data Science Group by Sarmila Basu May 23rd 2016
- Midpoint Retrospective
- Review Homework (Accuracy Measures)
- Quiz 06a (Accuracy Measures)
- Predictive Anecdotes
- Break
- Relational Algebra I
- Quiz 06b (Relational Algebra)
- Relational Algebra II
- Quiz 06c (Product, Join, Division)
- Break
- Data as Sparse Matrices
- Assignment (Complete all assignments items from all assignment slides)

Midpoint Retrospective

Midpoint Retrospective

- The class has done excellent work. I haven't commented enough on your excellent work and obvious talent.
- Philosophy of instruction: The point of a class like ours, as opposed to a MOOC, is the personal feedback and dedicated community.
 - The feedback from the instructor and students are the most valuable learning tool.
 - The community is special, because you know everyone and everyone has committed to participate in the group.

Midpoint Retrospective

- Some topics we covered (Focus on Analysis)
 - Data Preparation for Data Science
 - Introduction to R
 - Predictive Analytics:
 - Unsupervised learning: K-Means
 - Supervised learning: Classification
 - Classification Statistics
- Some topics we will cover (Focus on Persistence)
 - NoSQL (Scale out and CAP)
 - Relational Algebra and RDBMS
 - Graph Data
 - SPARQL
 - EAV and Sparse Matrices
 - Hadoop (HDFS and MapReduce)

Midpoint Retrospective

Accuracy Measures Exercise

Homework Review

1. Question: Training vs Test Data

- a) In general, for any modeling data, why are performance metrics better on training data than on test data? Answer: Because model is optimized for (trained on) training data
- b) Given modeling data, how do you determine which of this data will become training data and which data will become test data? Answer: Random partition of mutually exclusive datasets.
- c) Given two datasets, one that was the training data and the other that is the test data, how can you determine which is which? Answer: You can identify the training data in that the model is optimized for those data and will produce “better” results.

Homework Review

- The Confusion Matrix
 - Calculate the accuracy measures including the F-measure for the Homework. Positive and negative are just points-of-view:
 - Illness is positive (as in a test to determine if one is ill)
 - Health is positive (as in: its positive to be healthy)

Homework Review

- A model was trained on 300 individuals where 149 had the cold and 151 were healthy.
 - These numbers are irrelevant.
 - The accuracy measures are assessed by predictions and the test data.
 - Accuracy is not assessed with the training data.
- The model was tested on 100 individuals where 10 were ill.
 - Total population: 100
 - Support for ill: 10
 - Therefore, support for healthy: 90
- The model correctly predicted that 85 of the healthy individuals were indeed healthy
 - Correct predictions of healthy: 85
 - Therefore, incorrect prediction of ill (they were actually healthy): 5
 - (90 healthy - 85 correct predictions of healthy -> 5 healthy that were not predicted as healthy)
- and correctly predicted that 7 of the ill individuals were indeed ill.
 - Correct predictions of ill: 7
 - Therefore, incorrect prediction of healthy (they were actually ill): 3
 - (10 ill - 7 correct predictions of ill -> 3 ill that were not predicted as ill)

Homework Review

85 predicted healthy and were healthy
3 predicted healthy but were ill
5 predicted ill but were healthy
7 predicted ill and were ill

- A model was trained on 300 individuals where 149 had the cold and 151 were healthy.
 - These numbers are irrelevant.
 - The accuracy measures are assessed by predictions and the test data.
 - Accuracy is not assessed with the training data.
- The model was tested on 100 individuals where 10 were ill.
 - Total population: 100
 - Support for ill: 10
 - Therefore, support for healthy: 90
- The model correctly predicted that 85 of the healthy individuals were indeed healthy
 - Correct predictions of healthy: 85
 - Therefore, incorrect prediction of ill (they were actually healthy): 5
 - (90 healthy - 85 correct predictions of healthy -> 5 healthy that were not predicted as healthy)
- and correctly predicted that 7 of the ill individuals were indeed ill.
 - Correct predictions of ill: 7
 - Therefore, incorrect prediction of healthy (they were actually ill): 3
 - (10 ill - 7 correct predictions of ill -> 3 ill that were not predicted as ill)

Homework Review

85 predicted healthy and were healthy
3 predicted healthy but were ill
5 predicted ill but were healthy
7 predicted ill and were ill

	Actual	Predicted
Healthy	90	88
Ill	10	12

This is not a
confusion matrix!

- A model was trained on 300 individuals where 149 had the cold and 151 were healthy.
 - These numbers are irrelevant.
 - The accuracy measures are assessed by predictions and the test data.
 - Accuracy is not assessed with the training data.
- The model was tested on 100 individuals where 10 were ill.
 - Total population: 100
 - Support for ill: 10
 - Therefore, support for healthy: 90
- The model correctly predicted that 85 of the healthy individuals were indeed healthy
 - Correct predictions of healthy: 85
 - Therefore, incorrect prediction of ill (they were actually healthy): 5
 - (90 healthy - 85 correct predictions of healthy -> 5 healthy that were not predicted as healthy)
- and correctly predicted that 7 of the ill individuals were indeed ill.
 - Correct predictions of ill: 7
 - Therefore, incorrect prediction of healthy (they were actually ill): 3
 - (10 ill - 7 correct predictions of ill -> 3 ill that were not predicted as ill)

Homework: Confusion Matrix

85 predicted healthy and were healthy
3 predicted healthy but were ill
5 predicted ill but were healthy
7 predicted ill and were ill

	Actual	Predicted
Healthy	90	88
Ill	10	12

This is not a
confusion matrix!

Homework: Confusion Matrix

85 predicted healthy and were healthy
3 predicted healthy but were ill
5 predicted ill but were healthy
7 predicted ill and were ill

Homework: Confusion Matrix

85 predicted healthy and were healthy
3 predicted healthy but were ill
5 predicted ill but were healthy
7 predicted ill and were ill

Positive and negative are just points-of-view:

- Illness could be positive (as in a test to determine if one is ill)
- Health could be positive (as in: it's a positive thing to be healthy)

Homework: Confusion Matrix

85 predicted healthy and were healthy
3 predicted healthy but were ill
5 predicted ill but were healthy
7 predicted ill and were ill

Health is Positive

		Actual	
		P	N
Predicted	P'	TP	FP
	N'	FN	TN

Homework: Confusion Matrix

85 predicted healthy and were healthy
3 predicted healthy but were ill
5 predicted ill but were healthy
7 predicted ill and were ill

		Actual	
		P	N
Predicted	P'	85	3
	N'	5	7
		Health is Positive	

		Actual	
		P	N
Predicted	P'	TP	FP
	N'	FN	TN

Homework: Confusion Matrix

85 predicted healthy and were healthy
 3 predicted healthy but were ill
 5 predicted ill but were healthy
 7 predicted ill and were ill

		Actual	
		P	N
Predicted	P'	85	3
	N'	5	7
		Health is Positive	

Illness is Positive

		Actual	
		P	N
Predicted	P'	TP	FP
	N'	FN	TN

		P	N
Predicted	P'	TP	FP
	N'	FN	TN

Homework: Confusion Matrix

85 predicted healthy and were healthy
 3 predicted healthy but were ill
 5 predicted ill but were healthy
 7 predicted ill and were ill

		Actual	
		P	N
Predicted	P'	85	3
	N'	5	7

Health is Positive

		Actual	
		P	N
Predicted	P'	7	5
	N'	3	85

Illness is Positive

		Actual	
		P	N
Predicted	P'	TP	FP
	N'	FN	TN

		Actual	
		P	N
Predicted	P'	TP	FP
	N'	FN	TN

Homework: Confusion Matrix

85 predicted healthy and were healthy
3 predicted healthy but were ill
5 predicted ill but were healthy
7 predicted ill and were ill

		Actual	
		P	N
Predicted	P'	85	3
	N'	5	7

Health is Positive

		Actual	
		P	N
Predicted	P'	7	5
	N'	3	85

Illness is Positive

Homework: Confusion Matrix

85 predicted healthy and were healthy
3 predicted healthy but were ill
5 predicted ill but were healthy
7 predicted ill and were ill

		Actual	
		P	N
Predicted	P'	85	3
	N'	5	7

Health is Positive

- True Positive: 85
- True Negative: 7
- False Positive: 3
- False Negative: 5

		Actual	
		P	N
Predicted	P'	7	5
	N'	3	85

Homework: Confusion Matrix

85 predicted healthy and were healthy
 3 predicted healthy but were ill
 5 predicted ill but were healthy
 7 predicted ill and were ill

		Actual	
		P	N
Predicted	P'	85	3
	N'	5	7

Health is Positive

- True Positive: 85
- True Negative: 7
- False Positive: 3
- False Negative: 5

		Actual	
		P	N
Predicted	P'	7	5
	N'	3	85

Illness is Positive

- True Positive: 7
- True Negative: 85
- False Positive: 5
- False Negative: 3

Homework: Confusion Matrix

85 predicted healthy and were healthy
 3 predicted healthy but were ill
 5 predicted ill but were healthy
 7 predicted ill and were ill

		Actual	
		P	N
Predicted	P'	85	3
	N'	5	7

Health is Positive

- True Positive: 85
- True Negative: 7
- False Positive: 3
- False Negative: 5

		Actual	
		P	N
Predicted	P'	7	5
	N'	3	85

Illness is Positive

- True Positive: 7
- True Negative: 85
- False Positive: 5
- False Negative: 3

- Sensitivity*: $tp / (tp + fn)$
- Specificity: $tn / (tn + fp)$
- Accuracy: $(tp + tn) / (tp + fp + tn + fn)$
- Precision : $tp / (tp + fp)$
- Recall*: $tp / (tp + fn)$
- F-measure: $2tp / (2tp + fn + fp)$

Homework: Confusion Matrix

85 predicted healthy and were healthy
 3 predicted healthy but were ill
 5 predicted ill but were healthy
 7 predicted ill and were ill

		Actual	
		P	N
Predicted	P'	85	3
	N'	5	7

Health is Positive

- True Positive: 85
- True Negative: 7
- False Positive: 3
- False Negative: 5

		Actual	
		P	N
Predicted	P'	7	5
	N'	3	85

Illness is Positive

- True Positive: 7
- True Negative: 85
- False Positive: 5
- False Negative: 3

- Sensitivity*: $tp / (tp + fn)$
- Specificity: $tn / (tn + fp)$
- Accuracy: $(tp + tn) / (tp + fp + tn + fn)$
- Precision : $tp / (tp + fp)$
- Recall*: $tp / (tp + fn)$
- F-measure: $2tp / (2tp + fn + fp)$

- Sensitivity*: 0.94
- Specificity: 0.7
- Accuracy: 0.92
- Precision: 0.97
- Recall*: 0.94
- F-measure: 0.95

Homework: Confusion Matrix

85 predicted healthy and were healthy
 3 predicted healthy but were ill
 5 predicted ill but were healthy
 7 predicted ill and were ill

		Actual	
		P	N
Predicted	P'	85	3
	N'	5	7

Health is Positive

- True Positive: 85
- True Negative: 7
- False Positive: 3
- False Negative: 5

- Sensitivity*: $tp / (tp + fn)$
- Specificity: $tn / (tn + fp)$
- Accuracy: $(tp + tn) / (tp + fp + tn + fn)$
- Precision : $tp / (tp + fp)$
- Recall*: $tp / (tp + fn)$
- F-measure: $2tp / (2tp + fn + fp)$

- Sensitivity*: 0.94
- Specificity: 0.7
- Accuracy: 0.92
- Precision: 0.97
- Recall*: 0.94
- F-measure: 0.95

		Actual	
		P	N
Predicted	P'	7	5
	N'	3	85

Illness is Positive

- True Positive: 7
- True Negative: 85
- False Positive: 5
- False Negative: 3

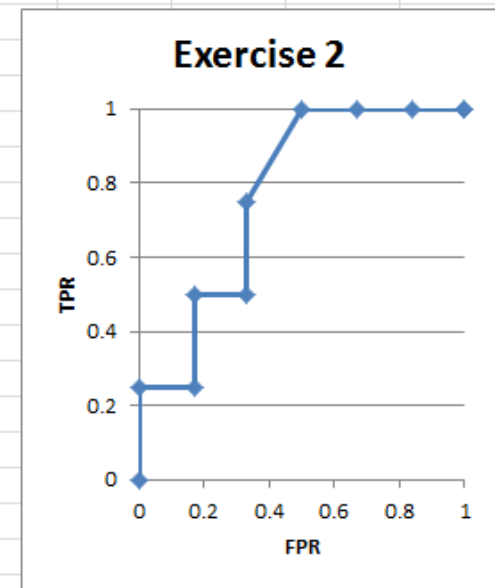
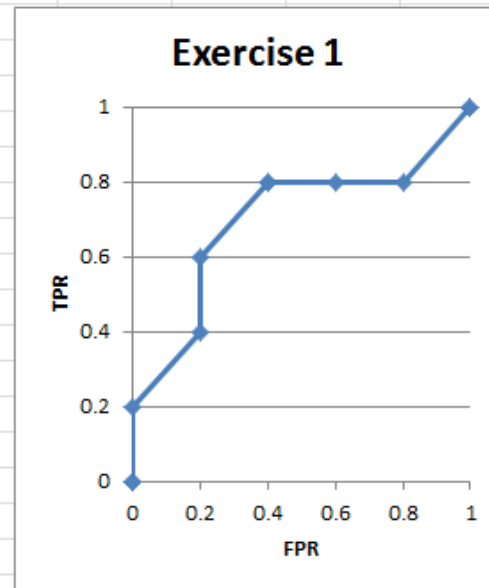
- Sensitivity*: 0.7
- Specificity: 0.94
- Accuracy: 0.92
- Precision: 0.58
- Recall*: 0.7
- F-measure: 0.63

Homework: Make an ROC

- [HowToMakeAnROC_Results.xls](#)

Results: Exercise 1	
FPR	TPR
1	1
1	1
0.8	0.8
0.6	0.8
0.4	0.8
0.4	0.8
0.2	0.6
0.2	0.4
0	0.2
0	0
0	0

Results: Exercise 2	
FPR	TPR
1	1
1	1
0.84	1
0.67	1
0.5	1
0.33	0.75
0.33	0.5
0.17	0.5
0.17	0.25
0	0.25
0	0

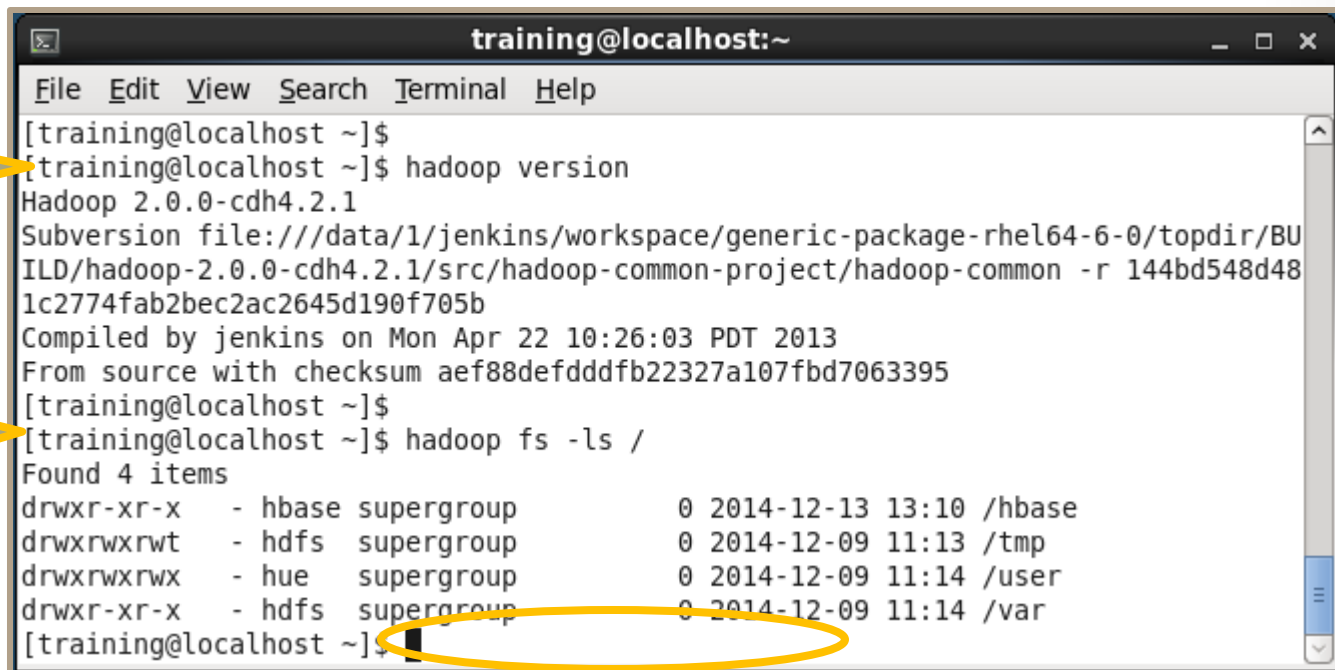


Homework: Prepare VM

- Enter into Console: `hadoop version`
- Enter into Console: `hadoop fs -ls /`

Check that Hadoop is installed

List directories in HDFS



A terminal window titled 'training@localhost:~' showing the execution of Hadoop commands. The window has a menu bar with 'File', 'Edit', 'View', 'Search', 'Terminal', and 'Help'. The terminal output shows the Hadoop version and the contents of the root directory in HDFS. A yellow oval highlights the prompt '[training@localhost ~]' at the end of the terminal output.

```
training@localhost:~  
File Edit View Search Terminal Help  
[training@localhost ~]$  
[training@localhost ~]$ hadoop version  
Hadoop 2.0.0-cdh4.2.1  
Subversion file:///data/1/jenkins/workspace/generic-package-rhel64-6-0/topdir/BU  
ILD/hadoop-2.0.0-cdh4.2.1/src/hadoop-common-project/hadoop-common -r 144bd548d48  
1c2774fab2bec2ac2645d190f705b  
Compiled by jenkins on Mon Apr 22 10:26:03 PDT 2013  
From source with checksum aef88defdddfb22327a107fbd7063395  
[training@localhost ~]$  
[training@localhost ~]$ hadoop fs -ls /  
Found 4 items  
drwxr-xr-x - hbase supergroup 0 2014-12-13 13:10 /hbase  
drwxrwxrwt - hdfs supergroup 0 2014-12-09 11:13 /tmp  
drwxrwxrwx - hue supergroup 0 2014-12-09 11:14 /user  
drwxr-xr-x - hdfs supergroup 0 2014-12-09 11:14 /var  
[training@localhost ~]$
```

Type your name into the console and take a screen shot

Homework:

Classification Accuracy (0)

- ClassificationAccuracy.R

Homework:

Classification Accuracy (1)

- # Problem statement
 - # I A Classification is tested on 1000 cases.
 - # II The false positive rate is 0.4
 - # III The true positive rate is 0.8.
 - # IV The accuracy is 0.7.
-
- # Problem statement expressed using TP, FP, FN, TN
 - # I $N = TP + FP + FN + TN = 1000$
 - # II $FPR = FP / (FP + TN) = 0.4$
 - # III $TPR = TP / (TP + FN) = 0.8$
 - # IV $(TP + TN) / (TP + FP + FN + TN) = 0.7$
-
- # Problem statement expressed as linear equations
 - # I $1 \cdot TP + 1 \cdot FP + 1 \cdot FN + 1 \cdot TN = 1000$
 - # II $0 + 3 \cdot FP + 0 - 2 \cdot TN = 0$
 - # III $1 \cdot TP + 0 - 4 \cdot FN + 0 = 0$
 - # IV $-3 \cdot TP + 7 \cdot FP + 7 \cdot FN - 3 \cdot TN = 0$

Homework:

Classification Accuracy (2)

- # Problem statement expressed as linear equations
- # I $1*TP + 1*FP + 1*FN + 1*TN = 1000$
- # II $0 + 3*FP + 0 - 2*TN = 0$
- # III $1*TP + 0 - 4*FN + 0 = 0$
- # IV $-3*TP + 7*FP + 7*FN - 3*TN = 0$

- # Problem statement expressed in terms of linear algebra:
- # We want to solve the linear equation: $Ax = b$
- # Where:
- # A is the matrix
- # x is a vector of TP, FP, FN, TN
- # b is the right-hand side of the linear equation

```

# -----
#   matrix A           vector b
# -----
# TP   FP   FN   TN   | b
# -----
#  1    1    1    1   | 1000
#  0    3    0   -2   |  0
#  1    0   -4    0   |  0
# -3    7    7   -3   |  0
# -----

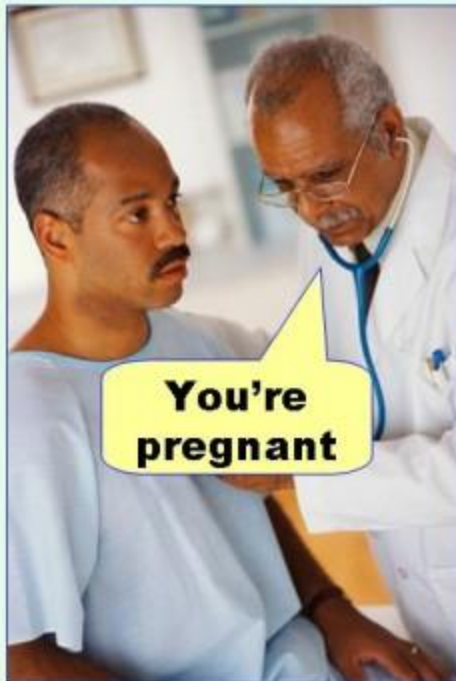
```

Accuracy

- Links
 - http://en.wikipedia.org/wiki/Accuracy_and_precision
 - http://en.wikipedia.org/wiki/F1_score
 - http://en.wikipedia.org/wiki/Precision_and_recall
- Exercise
 - Question 1: Why is the following statement both correct and useless? “My pregnancy test has a 95% accuracy”.
 - Question 2: What is the precision of the pregnancy test with the following measures?
 - A pregnancy test correctly predicted pregnancy 80% of the time among pregnant women.
 - 10% of all the women were predicted pregnant but were actually not pregnant.
 - The accuracy of the test was 89%.

Pregnancy Test Exercise

Type I error
(false positive)



Type II error
(false negative)



Pregnancy Test Exercise

- Question 1: Accuracy does not address Recall or Precision. For instance, 95% Accuracy could mean 95% TN and 0% TP. Both Recall and Precision would be 0%
- Question 2: Use ClassificationAccuracy.R (homework) as a template to complete PregnancyExercise.R
- **PregnancyExercise.R**
- Problem Statement
 - I $TP + FN + FP + TN = 1$
 - II $TP / (TP + FN) = \text{Recall} = 0.80$
 - III $(TP + TN) / (TP + FP + TN + FN) = 0.89$
 - IV $FP = 0.1$
- Algebra on statements II and III
 - II $FN = TP * 0.20 / 0.80$
 - III*I $TN = 0.89 - TP$
- Substitute FN, TN, and FP:
 - I,II,III,
 - $TP * 0.$
 - $TP = 0$
- Results:
 - $TP = 0$
 - Precision

Accuracy Measures Exercise

Quiz 06a

- Confusion Matrix and Accuracy Measures
- Last question is similar to the last question of the homework review. Complete PregnancyExercise.R by using ClassificationAccuracy.R as an example

An Expert is a person who tells you a simple thing in a confused way in such a fashion as to make you think the confusion is your own fault.

William Castle

Predictive Anecdotes

Review: Facts and Theories

- “It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.”

Sir Arthur Conan Doyle as the character of Sherlock Holmes



Predictive Anecdotes (1)

- We were using predictive analytics to look for causes of dropouts in a nursing school.
- At one point we looked for professors who were associated with high dropouts or high retention.
- We found one professor whose students had a 100% retention rate. We thought that this result was significant.
- It turned out that this professor had the final class in this two-year program. In other words, drop-outs occurred prior to this professor's class. In fact her class was a pro-seminar and all the students for this class had essentially already graduated.

Predictive Anecdotes (2)

- In the same nursing school we found that if the students race was “Missing” then the students were more likely to dropout.
- At first we thought that this missing race information indicated that their was an ethnicity that pre-disposed these students to drop out.
- But, we could not find any ethnicity that had a significantly higher retention or dropout rate.
- In fact, further investigation revealed that the proportion of ethnicities was the same for the students. It did not matter whether race was categorized as “Missing” or if the students race had been entered into the database.
- Later, we determined that most of the students who filled out the forms themselves did not enter information on their ethnicity. Only those students who were personally assisted by a (diligent) registrar entered a value for race. Further analysis indicated that personal assistance by a registrar, regardless of race, correlated with high retention rate.

Predictive Anecdotes (3)

- Many Years ago, a convict in Italy, wrote to 80 stockbrokers from prison. He claimed to have insider information from a fellow convict on a large local manufacturing firm.
- To 40 stockbrokers he wrote that the stock price would rise in the next two days. To the other 40 stockbrokers he wrote that the stock price would fall.
- After two days he followed up letters to the 40 stockbrokers who received the correct prediction. To half of those he wrote that the stock price would rise and to the other half he wrote that the stock price would fall.
- The prisoner repeated this pattern three more times and then requested a fee from the stockbrokers for additional predictions.

Predictive Anecdotes(4)



- “If you torture the data long enough, it will confess,”
 - Ronald Coase, Professor of Economics,
 - University of Chicago
- Real Story
 - After a failed, very large, epidemiological study, the researchers wanted to justify their grant. They looked for any pattern in their data.
 - When they found a pattern they retrospectively formulated a hypothesis and then they determined if that hypothesis could be verified by their data to a 95% certainty, as is common in such studies. A 95% certainty means that there is a 5% probability that the hypothesis does not account for the patterns. Actually, it means that there is a 5% chance that the pattern is fortuitous.
 - The researchers announced many (50) “verified” hypotheses. Soon colleagues educated them: Constructing a post-facto hypothesis, is similar to re-using training data as testing data.
 - Then the researchers randomly partitioned their data into a pattern search dataset and a pattern corroboration dataset. Although, they corroborated 1 of the patterns, this search was still statistically insignificant because we expect that 5% of these patterns are fortuitous.

Predictive Anecdotes (5)

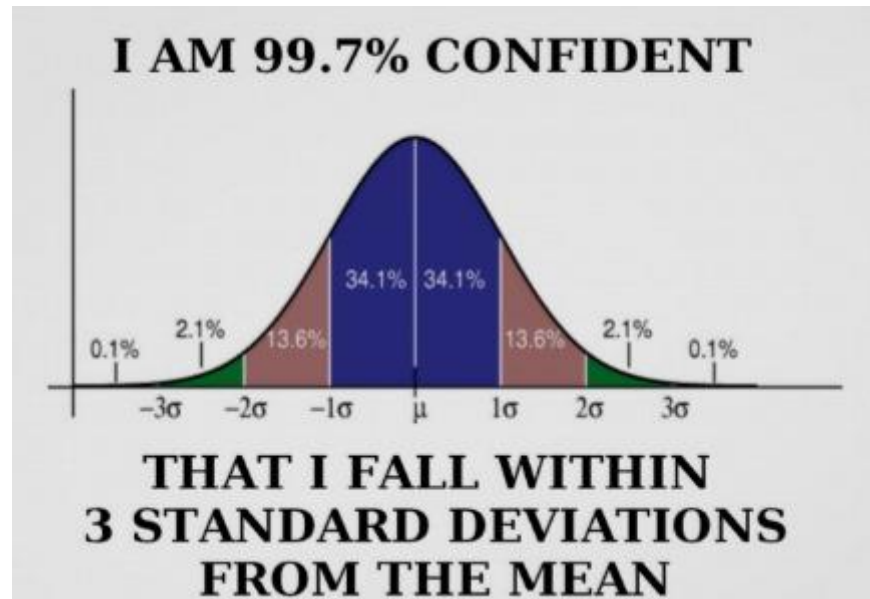
- Do Jelly Beans Cause Acne with $p < 0.05$?
- Green Jelly Beans Cause Acne $p < 0.05$!
- <http://xkcd.com/882/>
- The null hypothesis states that the observed variations are by chance (aka random). If you choose enough null hypotheses then there is an increasing chance that you will find a null hypothesis that is below the p-value.
- In biology we typically use a p-value of 0.05. That means that there is “only” a 5% chance that the null hypothesis is true.
- If the observed p-value < 0.05 , then we know that there is only a 5% chance that it is random. In other words there is a 95% chance that it is not random.
- How many hypotheses (n) should we test if we expect ($> 50\%$ chance) to find by chance 1 or more p-values (p) at less than 5%?
 - $0.5 < 1 - (1 - p)^n$; for $p = 0.05$ we find: $n \geq 14$

$p < 0.05$



Predictive Anecdotes (6)

- Tautology



Prediction Anecdotes(7)

- Redskins Rule:
 - http://en.wikipedia.org/wiki/Redskins_Rule
 - <http://abbottanalytics.blogspot.com/2012/11/why-predictive-modelers-should-be.html>



“Our algorithms have linked funny cat videos, UFO reports and searches for tofu pizza. We’re now on alert about a suspicious group of cat aliens who infiltrated our pizza industry.”

Predictive Anecdotes (8)

- http://www.finanzaonline.com/forum/attachments/econometria-e-modelli-di-trading-operativo/903701d1213616349-variazione-della-vix-e-rendimento-dello-s-p500-dataminejune_2000.pdf (See: dataminejune_2000.pdf)
- **S&P500 ~ Butter Production in Bangladesh + Butter Production in United States + United States Cheese Production + Sheep Population in Bangladesh + Sheep Population in United States**
- **S&P500 ~ Butter Production in Bangladesh**
-
- See Also:
<http://www.forbes.com/sites/davidleinweber/2012/07/24/stupid-data-miner-tricks-quants-fooling-themselves-the-economic-indicator-in-your-pants/>
- **Exact prediction of S&P 500 returns** by Ivan O. Kitov, Oleg I. Kitov (See: SSRN-id1045281.pdf)

Predictive Anecdotes (9)

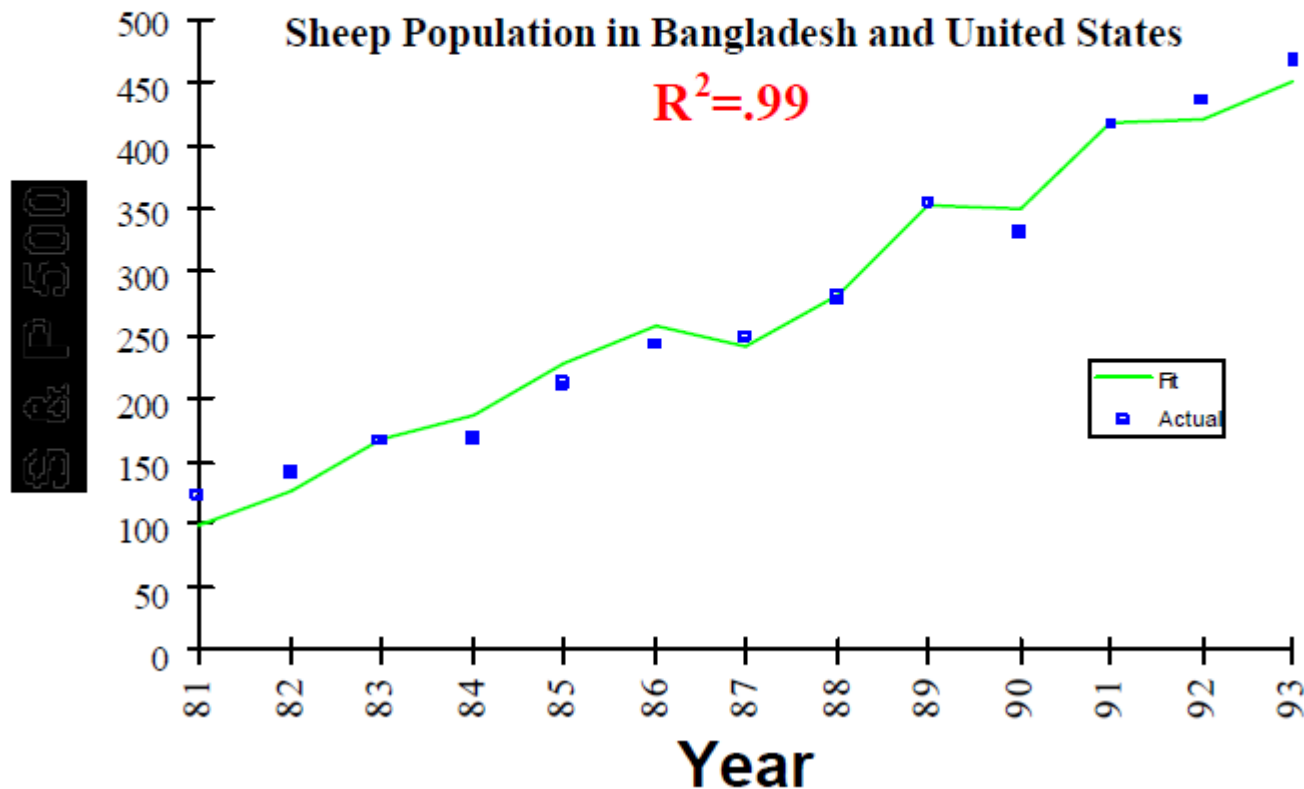
Overfitting the S & P 500

Butter Production in Bangladesh and United States

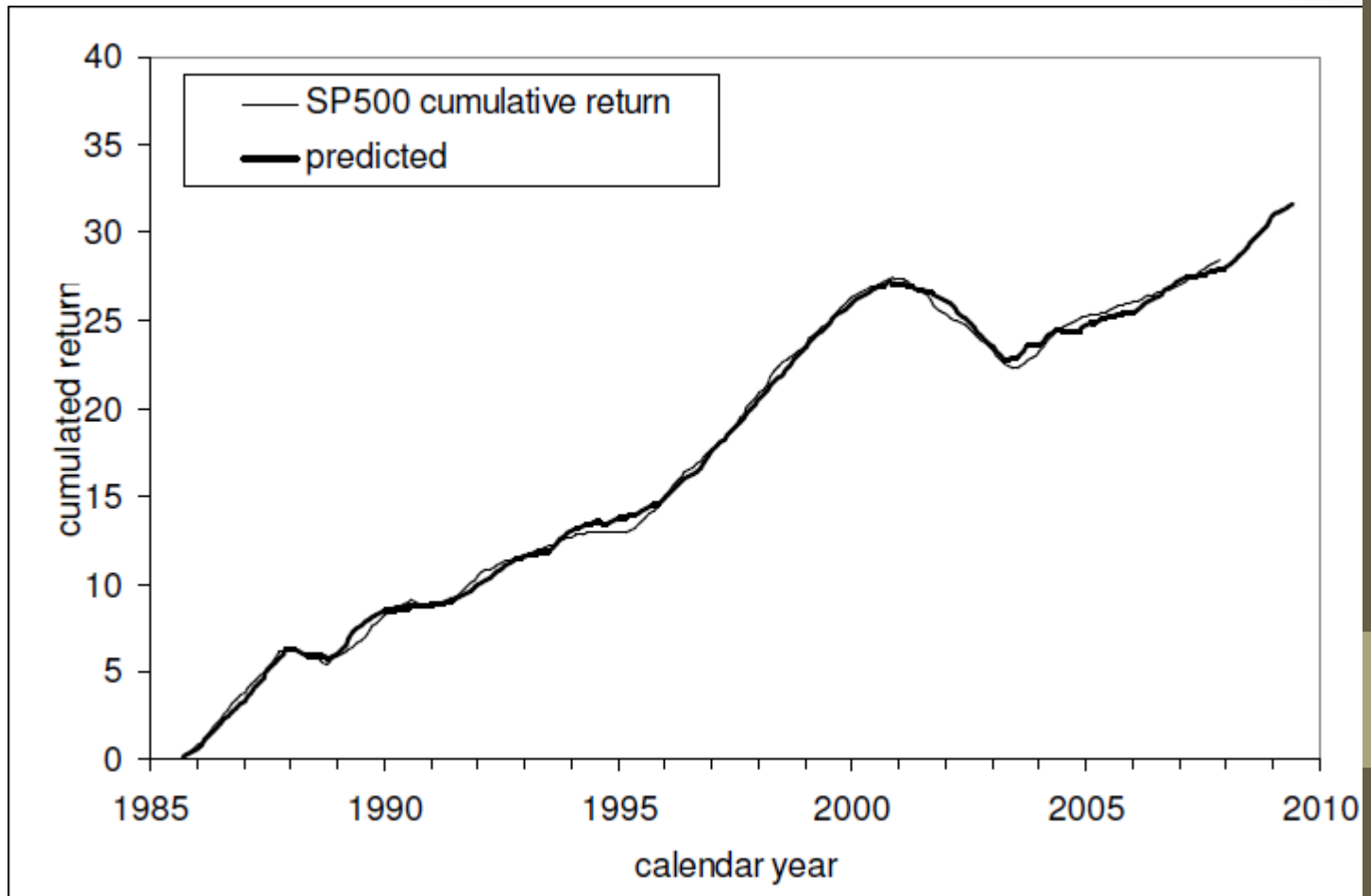
United States Cheese Production

Sheep Population in Bangladesh and United States

$R^2 = .99$



Predictive Anecdotes (10)



Predictive Anecdotes (11)

- Orange Cars Predicted to be better!
- "An orange used car is least likely to be a lemon" (http://seattletimes.com/html/business/technology/2017983961_apusscienceassport.html): *Of the 72,983 used cars, 8,976 were bad buys (12.3%). Yet, of the 415 orange cars in the dataset, only 34 were bad (8.2%)*
-
- *Debunked in this paper: "Are Orange Cars Really not Lemons?" by Ben Bullard and John Elder. See: orange cars.pdf.*

Predictive Anecdotes (12)

- Miscellaneous
 - Confusing Correlation with Causation:
 - http://en.wikipedia.org/wiki/Spurious_relationship
 - Proxy Columns and Audience Gullibility:
 - Scam artists use proxy attributes in their “predictions”
 - A true story from about 20 years ago. A fortune teller went on a radio talk show on KGO in the Bay Area. He demonstrated how he could mimic psychic abilities and get people to divulge information without their knowledge. After the show, this confessed scam artist was flooded with requests for psychic readings. The audience preferred to believe in his psychic powers and not his confessions.



“Twitter and Facebook can’t predict the election, but they did predict what you’re going to have for lunch: a tuna salad sandwich. You’re having the wrong sandwich.”

Predictive Anecdotes

Break

Relational Algebra

The Theory behind Relational Databases

Relational Algebra: What and Why

- Ted Codd introduced relational algebra to databases and created the relational model.
- Relational algebra provides a theoretical foundation for relational databases, and particularly for query languages like SQL.
- Why do you want a theoretical foundation?
 - If you want to optimize a query or a database
 - If you are thinking about using NOSQL, then you should be aware of the limitations and advantages of NOSQL data management. In other words, relational algebra assists in comparing SQL with NOSQL (NOT-SQL, Not-Oonly-SQL, KNOW-SQL, http://www.youtube.com/watch?v=sh1YACOK_bo)
- Use these files as examples for this lecture:
 - RelationalAlgebraAndSQL.pdf
 - RelationalAlgebraAndSQL.sql

New Terminology (1)

Term	Comments
<u>Table</u>	Part of a database
<u>Relation</u>	A table where rows are unique. Operand in Relational Algebra/Calculus
<u>Tuple</u>	<u>single</u> , <u>double</u> , <u>triple</u> , <u>quadruple</u> , <u>quintuple</u> , <u>sextuple</u> ; Like a row in a table
<u>Arity</u>	<u>unary</u> , <u>binary</u> , <u>ternary</u> , <u>quaternary</u>
<u>Closure</u>	Operation on a type produces a value of that same type. Natural Numbers have closure under + and * ($3 * 5 = 15$) Natural Numbers do not have closure under – or /; $5 - 3 = -2$

New Terminology (2)

Term	Comments
<u>Procedural</u>	Step-by-step solution to solving problem or achieving goal. I will drive to Bellevue, enter the class room and listen to the lecture. (Relational Algebra is <u>procedural</u> or <u>imperative</u>)
<u>Declarative</u>	Stating what one wants in non-ambiguous terms without describing how one is to achieve ones goal. Example: I want to know what was said in class last week. I don't care if you use the slide deck, your memory, or the recording to get me that information. (SQL is <u>declarative</u>)
<u>Relational Algebra</u>	The algebra that describes relations as operands and results
<u>Relational Calculus</u>	The calculus that uses relations as operands and results (SQL)

New Terminology (3)

Operation	Symbols	Comments
<u>Selection</u>	σ (sigma); $\sigma_{\varphi}(R)$;	SELECT * FROM <table name> <u>WHERE</u> <u>Column1 = 1</u>
<u>Projection</u>	π (pi); $\pi_{c_1, c_2, \dots, c_n}(R)$	SELECT <u>Column1, Column 2</u> FROM <table name>
<u>Rename</u>	ρ (rho)	
<u>Union</u>	\cup	$A \cup B$; $A = \{1, 2, 3, 5\}$; $B = \{0, 2\}$; $\{1, 2, 3, 5\} \cup \{0, 2\} = \{0, 1, 2, 3, 5\}$
<u>Intersection</u>	\cap	$A \cap B$; $A = \{1, 2, 3, 5\}$; $B = \{0, 2\}$; $\{1, 2, 3, 5\} \cap \{0, 2\} = \{2\}$
<u>Difference</u>	$\setminus, -$	$B \setminus A = B - A$; $\{0, 2\} - \{1, 2, 3, 5\} = \{0\}$

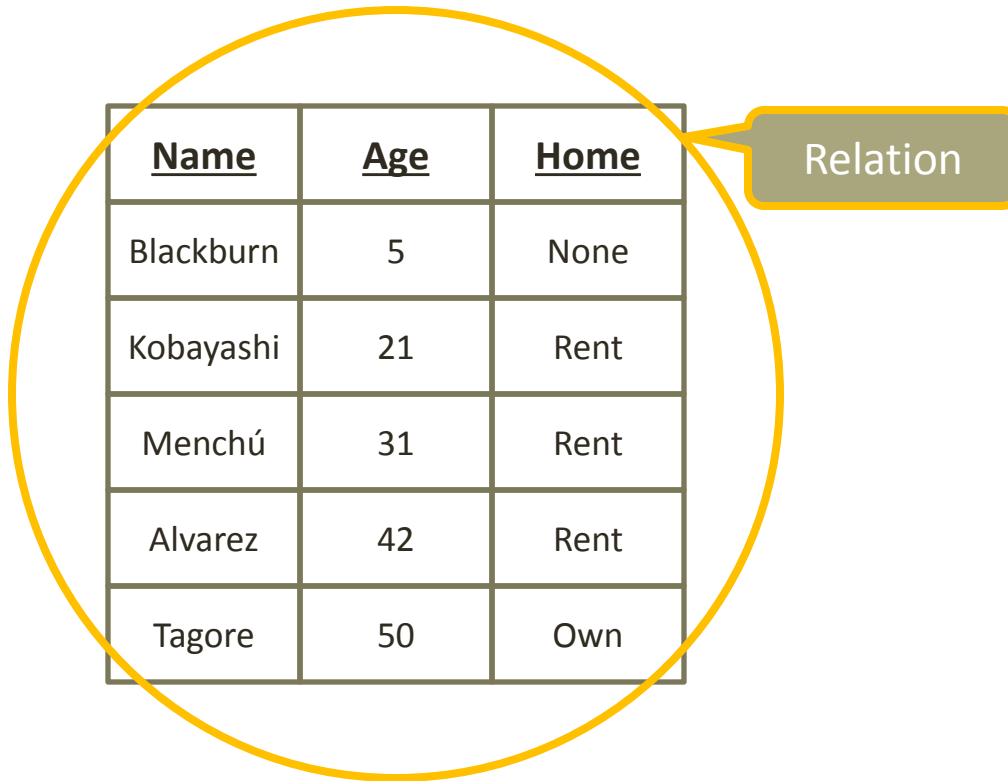
New Terminology (4)

Operation	Symbols	Comments
<u>Product</u>	\times	$A \times B$ $A=\{1,2,3,5\}$; $B=\{0,2\}$; $\{1,2,3, 5\} \times \{0,2\} = \{\{1,0\}, \{2,0\}, \{3,0\}, \{5,0\}, \{1,2\}, \{2,2\}, \{3,2\}, \{5,2\}\}$
<u>Join</u>	\bowtie_{φ}	$B \bowtie_{\varphi} A$; $\varphi: A > B$; $A=\{1,2,3,5\}$; $B=\{0,2\}$; $\{1,2,3,5\} \bowtie_{\varphi} \{0,2\} = \{\{1,0\}, \{2,0\}, \{3,0\}, \{3,2\}, \{5,0\}, \{5,2\}\}$
<u>Division</u>	\div	$A \div B = C$; Project to show me the columns in A that are not in B; Select to show me the tuples in A that are a superset of the a tuple in B.

Relational Algebra

<u>Name</u>	<u>Age</u>	<u>Home</u>
Blackburn	5	None
Kobayashi	21	Rent
Menchú	31	Rent
Alvarez	42	Rent
Tagore	50	Own

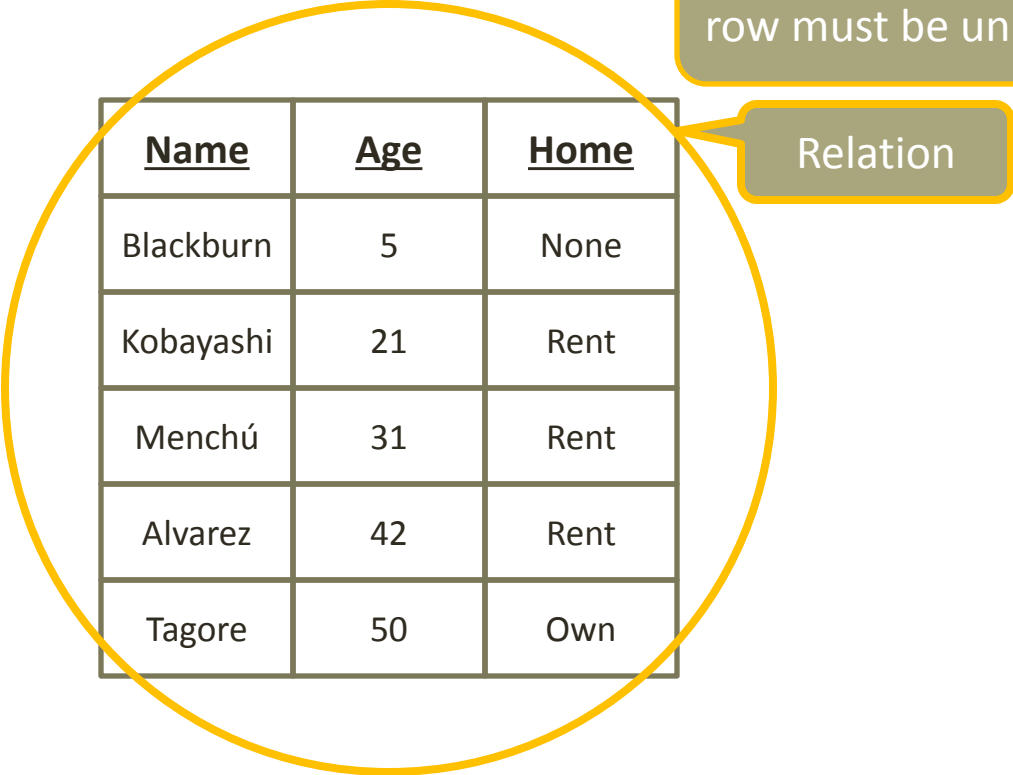
Relational Algebra: Relation



<u>Name</u>	<u>Age</u>	<u>Home</u>
Blackburn	5	None
Kobayashi	21	Rent
Menchú	31	Rent
Alvarez	42	Rent
Tagore	50	Own

Relational Algebra: Relation

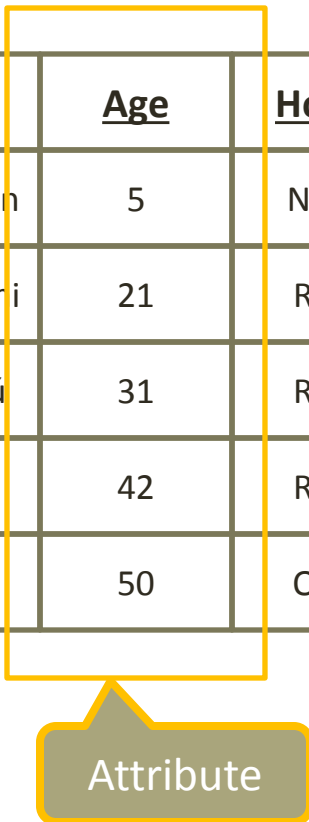
Relation is like a table except that each row must be unique like in a set



<u>Name</u>	<u>Age</u>	<u>Home</u>
Blackburn	5	None
Kobayashi	21	Rent
Menchú	31	Rent
Alvarez	42	Rent
Tagore	50	Own

Relation

Relational Algebra: Attribute



<u>Name</u>	<u>Age</u>	<u>Home</u>
Blackburn	5	None
Kobayashi	21	Rent
Menchú	31	Rent
Alvarez	42	Rent
Tagore	50	Own

Attribute

Relational Algebra: Attribute

Attribute:


Must be of the same data type.
Have a name

<u>Name</u>	<u>Age</u>	<u>Home</u>
Blackburn	5	None
Kobayashi	21	Rent
Menchú	31	Rent
Alvarez	42	Rent
Tagore	50	Own

Attribute

Relational Algebra: Tuple

<u>Name</u>	<u>Age</u>	<u>Home</u>
Blackburn	5	None
Kobayashi	21	Rent
Menchú	31	Rent
Alvarez	42	Rent
Tagore	50	Own



A diagram illustrating a tuple in a relational database. A yellow rectangular box highlights the row containing 'Kobayashi', '21', and 'Rent'. A yellow arrow points from this box to a separate yellow rounded rectangle labeled 'tuple'.

Relational Algebra: Tuple

tuple from: singlele, doublele, triplele,
quadruple, quintuple
arity from: unary, binary, ternary

<u>Name</u>	<u>Age</u>	<u>Home</u>
Blackburn	5	None
Kobayashi	21	Rent
Menchú	31	Rent
Alvarez	42	Rent
Tagore	50	Own

tuple with arity of 3

Relational Algebra: Operands and Simple Operations

- Operand
 - Relation (Table)
- Operations
 - UNION
 - INTERSECT
 - PROJECT
 - SELECT
 - PRODUCT
 - DIVISION

Relational Algebra: Union

Combine Relations

<u>Name</u>	<u>Age</u>	<u>Home</u>
Blackburn	5	None
Kobayashi	21	Rent

<u>Name</u>	<u>Age</u>	<u>Home</u>
Menchú	31	Rent
Alvarez	42	Rent
Tagore	50	Own

Relational Algebra: Union

Combine Relations

<u>Name</u>	<u>Age</u>	<u>Home</u>
Blackburn	5	None
Kobayashi	21	Rent

<u>Name</u>	<u>Age</u>	<u>Home</u>
Menchú	31	Rent
Alvarez	42	Rent
Tagore	50	Own

Relational Algebra Union:
 $R \cup S$

Relational Algebra: Union

Combine Relations

<u>Name</u>	<u>Age</u>	<u>Home</u>
Blackburn	5	None
Kobayashi	21	Rent

<u>Name</u>	<u>Age</u>	<u>Home</u>
Menchú	31	Rent
Alvarez	42	Rent
Tagore	50	Own

SQL Statement:

```
SELECT * FROM MyTableR UNION  
SELECT * FROM MyTableS
```

Relational Algebra Union:
 $R \cup S$

Relational Algebra: Union

Combine Relations

<u>Name</u>	<u>Age</u>	<u>Home</u>
Blackburn	5	None
Kobayashi	21	Rent

<u>Name</u>	<u>Age</u>	<u>Home</u>
Menchú	31	Rent
Alvarez	42	Rent
Tagore	50	Own



<u>Name</u>	<u>Age</u>	<u>Home</u>
Blackburn	5	None
Kobayashi	21	Rent
Menchú	31	Rent
Alvarez	42	Rent
Tagore	50	Own

Relational Algebra Union:
 $R \cup S$

Relational Algebra: Intersect

<u>Name</u>	<u>Age</u>	<u>Home</u>
Blackburn	5	None
Kobayashi	21	Rent
Menchú	31	Rent
Tagore	50	Own

Same Rows

<u>Name</u>	<u>Age</u>	<u>Home</u>
Menchú	31	Rent
Alvarez	42	Rent
Tagore	50	Own

Relational Algebra: Intersect

<u>Name</u>	<u>Age</u>	<u>Home</u>
Blackburn	5	None
Kobayashi	21	Rent
Menchú	31	Rent
Tagore	50	Own

Same Rows

<u>Name</u>	<u>Age</u>	<u>Home</u>
Menchú	31	Rent
Alvarez	42	Rent
Tagore	50	Own

Relational Algebra: Intersect

<u>Name</u>	<u>Age</u>	<u>Home</u>
Blackburn	5	None
Kobayashi	21	Rent
Menchú	31	Rent
Tagore	50	Own

Same Rows

<u>Name</u>	<u>Age</u>	<u>Home</u>
Menchú	31	Rent
Alvarez	42	Rent
Tagore	50	Own

Relational Algebra Intersection:
 $R \cap S$

Relational Algebra: Intersect

<u>Name</u>	<u>Age</u>	<u>Home</u>
Blackburn	5	None
Kobayashi	21	Rent
Menchú	31	Rent
Tagore	50	Own

Same Rows

<u>Name</u>	<u>Age</u>	<u>Home</u>
Menchú	31	Rent
Alvarez	42	Rent
Tagore	50	Own

SQL Statement:

```
SELECT * FROM MyTableR  
INTERSECT  
SELECT * FROM MyTableS
```

Relational Algebra Intersection:
 $R \cap S$

Relational Algebra: Intersect

<u>Name</u>	<u>Age</u>	<u>Home</u>
Blackburn	5	None
Kobayashi	21	Rent
Menchú	31	Rent
Tagore	50	Own

Same Rows

<u>Name</u>	<u>Age</u>	<u>Home</u>
Menchú	31	Rent
Alvarez	42	Rent
Tagore	50	Own



<u>Name</u>	<u>Age</u>	<u>Home</u>
Menchú	31	Rent
Tagore	50	Own

Relational Algebra Intersection:
 $R \cap S$

Relational Algebra: Examples

- $R \cup S$
 - `SELECT * FROM MyTableR UNION SELECT * FROM MyTableS`
- `SELECT * FROM MyTableR UNION SELECT * FROM MyTableS`
 - $R \cup S$ or $S \cup R$
- $R \cap S$
 - `SELECT * FROM MyTableR INTERSECT SELECT * FROM MyTableS`
- `SELECT * FROM MyTableR INTERSECT SELECT * FROM MyTableS`
 - $R \cap S$ or $S \cap R$
- In General:
 - An operation with \cup or \cap produces a relation
 - $R \cup S = S \cup R$
 - $R \cap S = S \cap R$
 - $(R \cup S) \cap T = (R \cap T) \cup (S \cap T)$
 - $(R \cap S) \cup T = (R \cup T) \cap (S \cup T)$

Relational Algebra: Project

<u>Name</u>	<u>Age</u>	<u>Home</u>
Blackburn	5	None
Kobayashi	21	Rent
Menchú	31	Rent
Alvarez	42	Rent
Tagore	50	Own

Vertical partition

Relational Algebra: Project

<u>Name</u>	<u>Age</u>	<u>Home</u>
Blackburn	5	None
Kobayashi	21	Rent
Menchú	31	Rent
Alvarez	42	Rent
Tagore	50	Own

Vertical partition

Relational Algebra Project:

$\pi_{c1, c2, \dots, cn}(R)$

where

$c1, c2, \dots, cn$: Age, Home

R: MyTable

Relational Algebra: Project

<u>Name</u>	<u>Age</u>	<u>Home</u>
Blackburn	5	None
Kobayashi	21	Rent
Menchú	31	Rent
Alvarez	42	Rent
Tagore	50	Own

SQL Statement:

```
SELECT Age, Home FROM  
MyTable
```

Vertical partition

Relational Algebra Project:

$$\pi_{c1, c2, \dots, cn}(R)$$


where

$c1, c2, \dots, cn$: Age, Home

R: MyTable

Relational Algebra: Project

<u>Name</u>	<u>Age</u>	<u>Home</u>
Blackburn	5	None
Kobayashi	21	Rent
Menchú	31	Rent
Alvarez	42	Rent
Tagore	50	Own



<u>Age</u>	<u>Home</u>
5	None
21	Rent
31	Rent
42	Rent
50	Own

Relational Algebra Project:

$\pi_{c1, c2, \dots, cn}(R)$


where

$c1, c2, \dots, cn$: Age, Home

R: MyTable

Relational Algebra: Project

<u>Name</u>	<u>Age</u>	<u>Home</u>
Blackburn	5	None
Kobayashi	21	Rent
Menchú	31	Rent
Alvarez	42	Rent
Tagore	50	Own



<u>Age</u>	<u>Home</u>
5	None
21	Rent
31	Rent
42	Rent
50	Own

The result of a projection is a relation with 0 to n attributes where n is the number of attributes in the operand

Relational Algebra Project:

$\pi_{c1, c2, \dots, cn}(R)$

where

c1, c2, ..., cn: Age, Home

R: MyTable

Relational Algebra: Select

<u>Name</u>	<u>Age</u>	<u>Home</u>
Blackburn	5	None
Kobayashi	21	Rent
Menchú	31	Rent
Alvarez	42	Rent
Tagore	50	Own

Horizontal partition

Relational Algebra: Select

<u>Name</u>	<u>Age</u>	<u>Home</u>
Blackburn	5	None
Kobayashi	21	Rent
Menchú	31	Rent
Alvarez	42	Rent
Tagore	50	Own

Horizontal partition

Relational Algebra Select:

$\sigma_{\varphi}(R)$
where

φ : Home = "Rent"

R: MyTable

Relational Algebra: Select

SQL Statement:

```
SELECT * FROM MyTable WHERE  
Home = "Rent"
```

<u>Name</u>	<u>Age</u>	<u>Home</u>
Blackburn	5	None
Kobayashi	21	Rent
Menchú	31	Rent
Alvarez	42	Rent
Tagore	50	Own

Horizontal partition

Relational Algebra Select:


$\sigma_{\varphi}(R)$
where

φ : Home = "Rent"

R: MyTable

Relational Algebra: Select

<u>Name</u>	<u>Age</u>	<u>Home</u>
Blackburn	5	None
Kobayashi	21	Rent
Menchú	31	Rent
Alvarez	42	Rent
Tagore	50	Own



<u>Name</u>	<u>Age</u>	<u>Home</u>
Kobayashi	21	Rent
Menchú	31	Rent
Alvarez	42	Rent

The result of a selection is a relation with 0 to n tuples where n is the number of tuples in the operand

Relational Algebra Select:

$\sigma_{\varphi}(R)$
where

φ : Home = "Rent"

R: MyTable

Relational Algebra: Examples

- $\pi_{\text{Age, Home}}(R)$
 - SELECT Age, Home FROM MyTable
- $\sigma_{\text{Home}=\text{"Rent"}}(R)$
 - SELECT * FROM MyTable WHERE Home = "Rent"
- SELECT Age, Home FROM MyTable WHERE Home = "Rent"
 - $\pi_{\text{Age, Home}}(\sigma_{\text{Home}=\text{"Rent"}}(R))$ or $\sigma_{\text{Home}=\text{"Rent"}}(\pi_{\text{Age, Home}}(R))$
- In General:
 - An operation with σ produces a relation
 - An operation with π produces a relation
 - $\sigma_{\varphi_1}(\sigma_{\varphi_2}(R)) = \sigma_{\varphi_2}(\sigma_{\varphi_1}(R))$
 - $\pi_{[c1]}(\pi_{[c2]}(R)) = \pi_{[c2]}(\pi_{[c1]}(R))$
 - $\pi_{[c]}(\sigma_{\varphi}(R)) = \sigma_{\varphi}(\pi_{[c]}(R))$ (**only if** φ is not dependent on $[c]$)

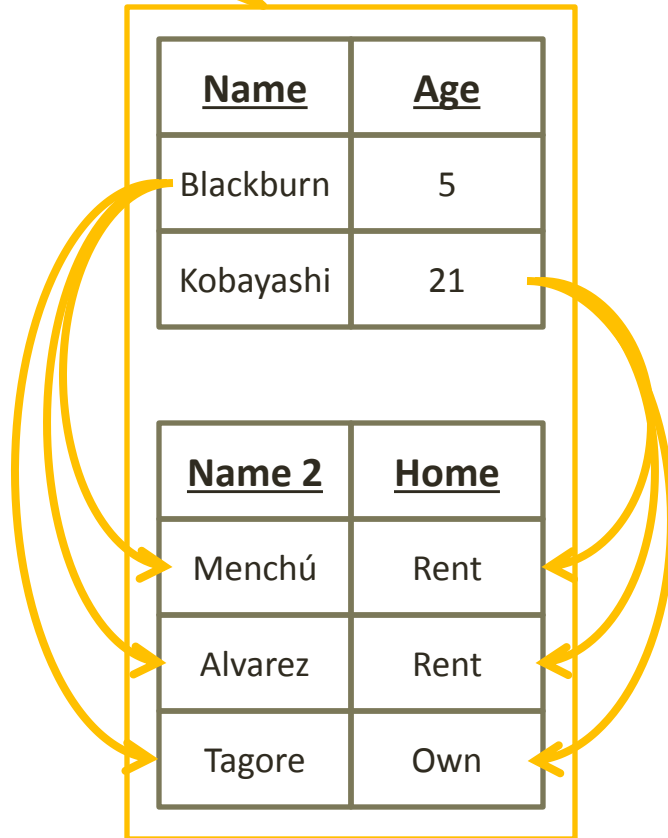
Quiz 06b

- Quiz 06b (Relational Algebra)



Relational Algebra: Product

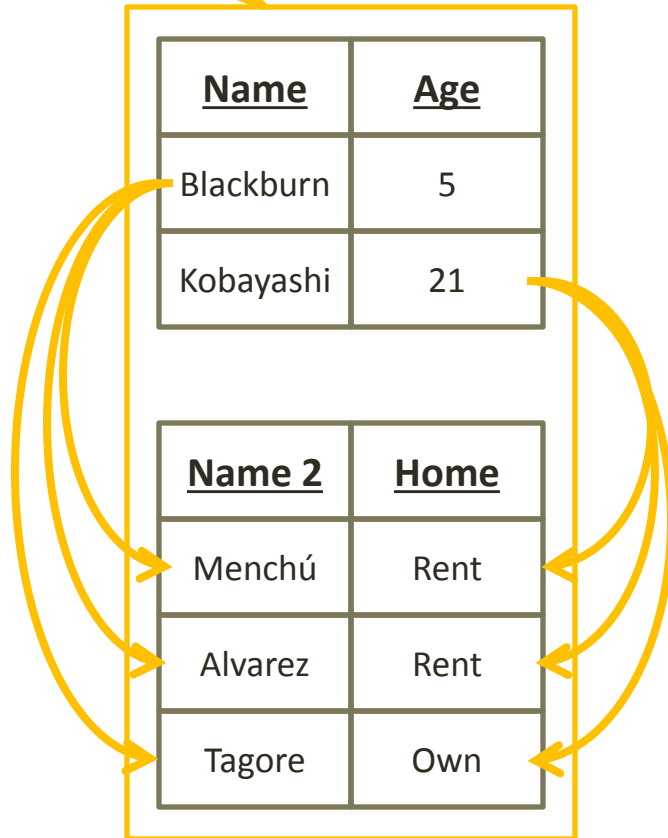
Combine Rows



Relational Algebra Product:
 $R \times S$

Relational Algebra: Product

Combine Rows



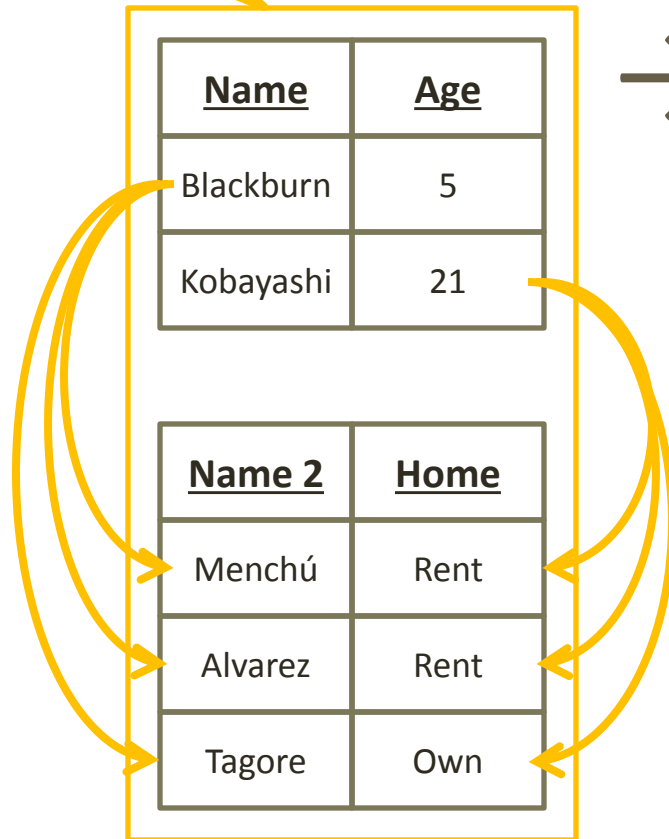
SQL Statement:

```
SELECT * FROM TableR, TableS
```

Relational Algebra Product:
 $R \times S$

Relational Algebra: Product

Combine Rows



<u>Name</u>	<u>Age</u>	<u>Name 2</u>	<u>Home</u>
-------------	------------	---------------	-------------

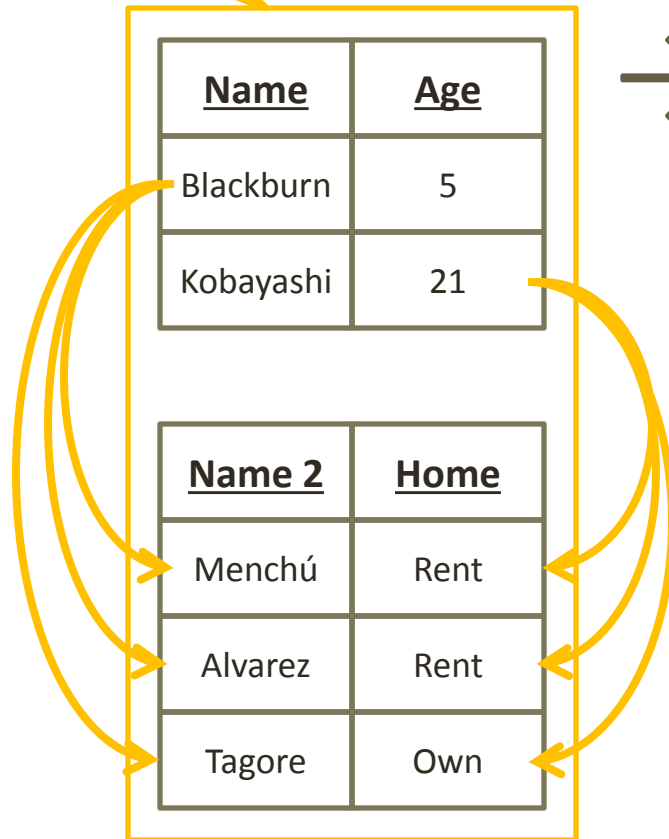
Blackburn	5	Menchú	Rent
		Alvarez	Rent
		Tagore	Own

Kobayashi	21	Menchú	Rent
		Alvarez	Rent
		Tagore	Own

Relational Algebra Product:
 $R \times S$

Relational Algebra: Product

Combine Rows



<u>Name</u>	<u>Age</u>	<u>Name 2</u>	<u>Home</u>
-------------	------------	---------------	-------------

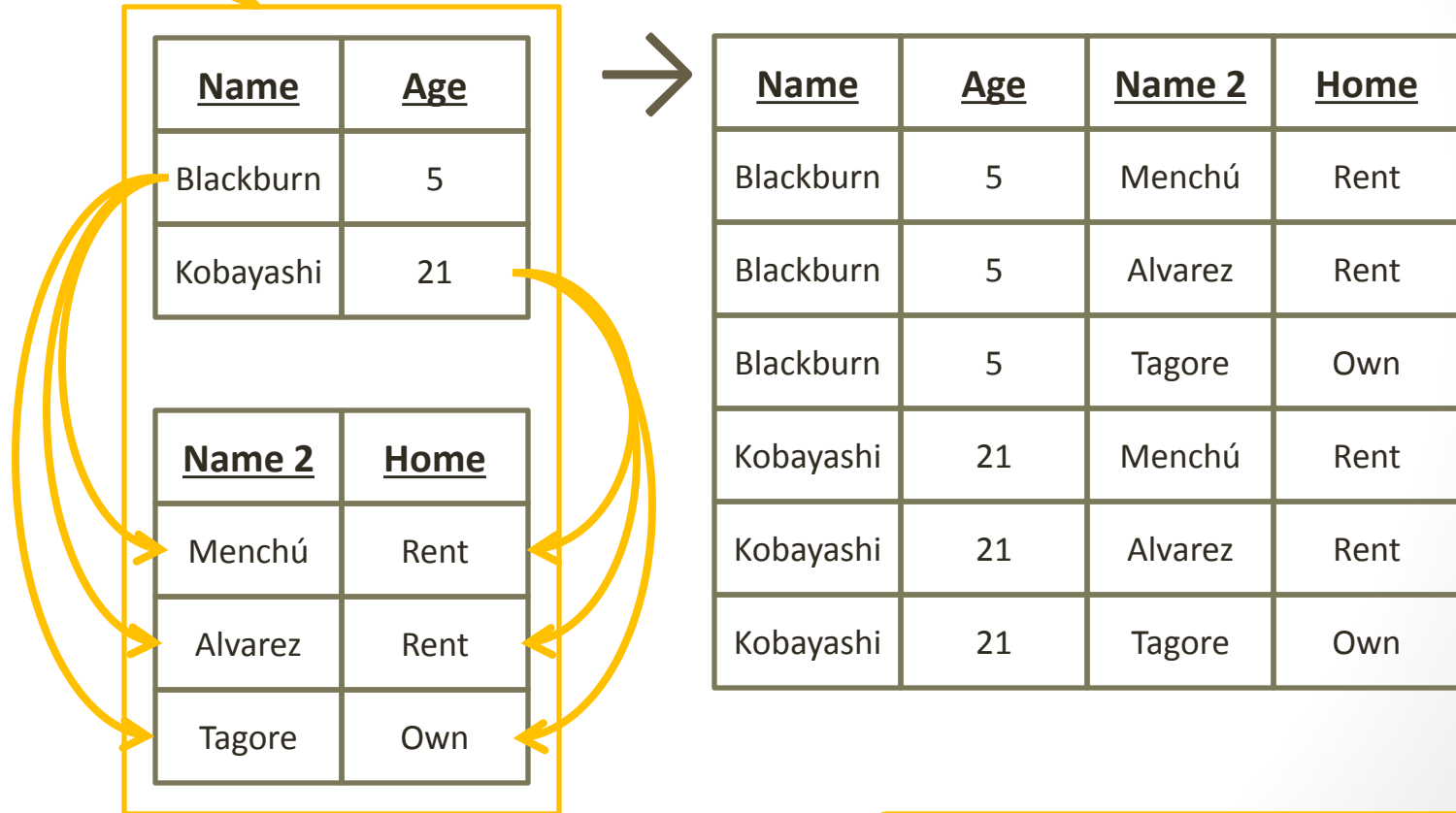
Blackburn	5	Menchú	Rent
Blackburn	5	Alvarez	Rent
Blackburn	5	Tagore	Own

Kobayashi	21	Menchú	Rent
Kobayashi	21	Alvarez	Rent
Kobayashi	21	Tagore	Own

Relational Algebra Product:
 $R \times S$

Relational Algebra: Product

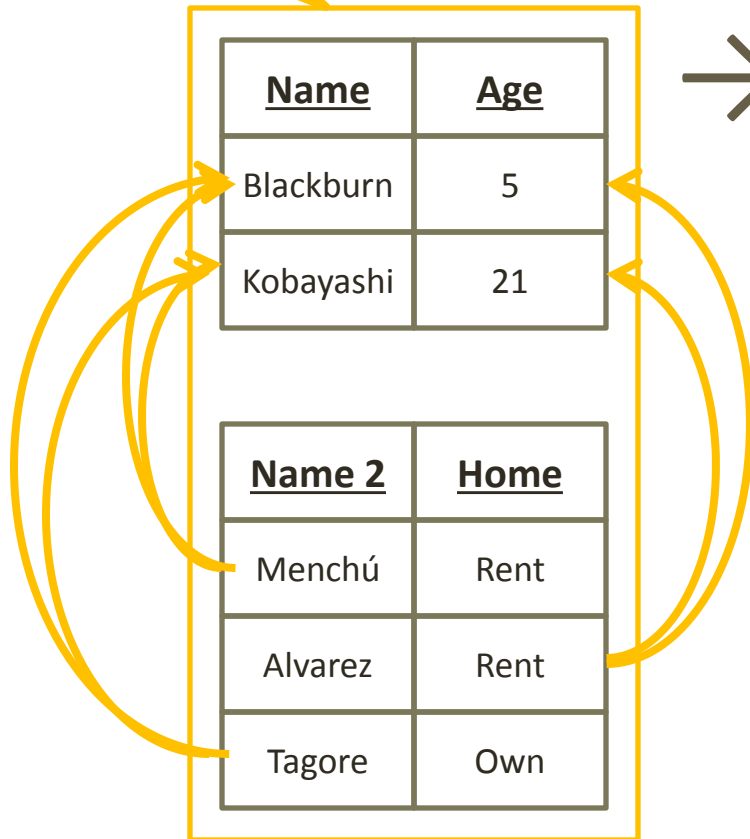
Combine Rows



Relational Algebra Product:
 $R \times S$

Relational Algebra: Product

Combine Rows



<u>Name</u>	<u>Age</u>	<u>Name 2</u>	<u>Home</u>
Blackburn	5	Menchú	Rent
Kobayashi	21		
Blackburn	5	Alvarez	Rent
Kobayashi	21		
Blackburn	5	Tagore	Own
Kobayashi	21		

Relational Algebra Product:
 $R \times S$

Relational Algebra: Product

Combine Rows

<u>Name</u>	<u>Age</u>
Blackburn	5
Kobayashi	21

<u>Name 2</u>	<u>Home</u>
Menchú	Rent
Alvarez	Rent
Tagore	Own



<u>Name</u>	<u>Age</u>	<u>Name 2</u>	<u>Home</u>
-------------	------------	---------------	-------------

Blackburn	5	Menchú	Rent
Kobayashi	21	Menchú	Rent

Blackburn	5	Alvarez	Rent
Kobayashi	21	Alvarez	Rent

Blackburn	5	Tagore	Own
Kobayashi	21	Tagore	Own

Relational Algebra Product:
 $R \times S$

Relational Algebra: Product

Combine Rows

<u>Name</u>	<u>Age</u>
Blackburn	5
Kobayashi	21

<u>Name 2</u>	<u>Home</u>
Menchú	Rent
Alvarez	Rent
Tagore	Own



<u>Name</u>	<u>Age</u>	<u>Name 2</u>	<u>Home</u>
Blackburn	5	Menchú	Rent
Kobayashi	21	Menchú	Rent
Blackburn	5	Alvarez	Rent
Kobayashi	21	Alvarez	Rent
Blackburn	5	Tagore	Own
Kobayashi	21	Tagore	Own

Relational Algebra Product:
 $R \times S$

Relational Algebra: Product

Combine Rows

The result of a product is a relation with $n*m$ tuples where n and m are the number of tuples in the operands. The arity of the result is $i + j$ where i and j are the arities of the operands

<u>Name</u>	<u>Age</u>
Blackburn	5
Kobayashi	21

<u>Name 2</u>	<u>Home</u>
Menchú	Rent
Alvarez	Rent
Tagore	Own



<u>Name</u>	<u>Age</u>	<u>Name 2</u>	<u>Home</u>
Blackburn	5	Menchú	Rent
Kobayashi	21	Menchú	Rent
Blackburn	5	Alvarez	Rent
Kobayashi	21	Alvarez	Rent
Blackburn	5	Tagore	Own
Kobayashi	21	Tagore	Own

Relational Algebra Product:
 $R \times S$

Relational Algebra: Product

Combine Rows

The result of a product is a relation with $n \times m$ tuples where n and m are the number of tuples in the operands. The arity of the result is $i + j$ where i and j are the arities of the operands

<u>Name</u>	<u>Age</u>
Blackburn	5
Kobayashi	21

<u>Name 2</u>	<u>Home</u>
Menchú	Rent
Alvarez	Rent
Tagore	Own



<u>Name</u>	<u>Age</u>	<u>Name 2</u>	<u>Home</u>
Blackburn	5	Menchú	Rent
Blackburn	5	Alvarez	Rent
Blackburn	5	Tagore	Own
Kobayashi	21	Menchú	Rent
Kobayashi	21	Alvarez	Rent
Kobayashi	21	Tagore	Own

Relational Algebra Product:
 $R \times S$

Relational Algebra: Join

Combine Rows

<u>Name</u>	<u>Age</u>
Blackburn	5
Kobayashi	21

<u>Name 2</u>	<u>Home</u>
Menchú	Rent
Alvarez	Rent
Tagore	Own



<u>Name</u>	<u>Age</u>	<u>Name 2</u>	<u>Home</u>
Blackburn	5	Menchú	Rent
Kobayashi	21	Menchú	Rent
Blackburn	5	Alvarez	Rent
Kobayashi	21	Alvarez	Rent
Blackburn	5	Tagore	Own
Kobayashi	21	Tagore	Own

Relational Algebra Product with Select:
 $\sigma_{\varphi}(R \times S)$ where $\varphi: \text{Home} = \text{"Rent"}$
Relational Algebra Join:
 $R \bowtie_{\varphi} S$ where $\varphi: \text{Home} = \text{"Rent"}$

Relational Algebra: Join

Combine Rows

<u>Name</u>	<u>Age</u>
Blackburn	5
Kobayashi	21

<u>Name 2</u>	<u>Home</u>
Menchú	Rent
Alvarez	Rent
Tagore	Own



<u>Name</u>	<u>Age</u>	<u>Name 2</u>	<u>Home</u>
Blackburn	5	Menchú	Rent
Kobayashi	21	Menchú	Rent
Blackburn	5	Alvarez	Rent
Kobayashi	21	Alvarez	Rent
Blackburn	5	Tagore	Own
Kobayashi	21	Tagore	Own

Relational Algebra Product with Select:
 $\sigma_{\varphi}(R \times S)$ where $\varphi: \text{Home} = \text{"Rent"}$
Relational Algebra Join:
 $R \bowtie_{\varphi} S$ where $\varphi: \text{Home} = \text{"Rent"}$

Relational Algebra: Join

Combine Rows

<u>Name</u>	<u>Age</u>
Blackburn	5
Kobayashi	21

<u>Name 2</u>	<u>Home</u>
Menchú	Rent
Alvarez	Rent
Tagore	Own



<u>Name</u>	<u>Age</u>	<u>Name 2</u>	<u>Home</u>
Blackburn	5	Menchú	Rent
Kobayashi	21	Menchú	Rent
Blackburn	5	Alvarez	Rent
Kobayashi	21	Alvarez	Rent

Relational Algebra Product with Select:
 $\sigma_{\varphi}(R \times S)$ where $\varphi: \text{Home} = \text{"Rent"}$
Relational Algebra Join:
 $R \bowtie_{\varphi} S$ where $\varphi: \text{Home} = \text{"Rent"}$

Relational Algebra: Join

- A Join is a Product with a select statement
- Product followed by Select
 - `SELECT * FROM TableR, TableS WHERE Home = "Rent"`
 - $\sigma_{\varphi}(R \times S)$ where $\varphi: \text{Home} = \text{"Rent"}$
- JOIN
 - `SELECT * FROM TableR JOIN TableS ON Home = "Rent"`
 - $R \bowtie_{\varphi} S$ where $\varphi: \text{Home} = \text{"Rent"}$

Relational Algebra: Division

This was a Product
Operand

<u>Name</u>	<u>Age</u>
Blackburn	5
Kobayashi	21

<u>Name 2</u>	<u>Home</u>
Menchú	Rent
Alvarez	Rent
Tagore	Own

This was a Product Operand

This was the result
of a Product

<u>Name</u>	<u>Age</u>	<u>Name 2</u>	<u>Home</u>
Blackburn	5	Menchú	Rent
Blackburn	5	Alvarez	Rent
Blackburn	5	Tagore	Own
Kobayashi	21	Menchú	Rent
Kobayashi	21	Alvarez	Rent
Kobayashi	21	Tagore	Own

Relational Algebra Division:
 $R \div S$

Relational Algebra: Division

A Division is sort of like the reverse of a Product

This was a Product
Operand

<u>Name</u>	<u>Age</u>
Blackburn	5
Kobayashi	21

This was the result
of a Product

<u>Name</u>	<u>Age</u>	<u>Name 2</u>	<u>Home</u>
Blackburn	5	Menchú	Rent
Blackburn	5	Alvarez	Rent
Blackburn	5	Tagore	Own
Kobayashi	21	Menchú	Rent
Kobayashi	21	Alvarez	Rent
Kobayashi	21	Tagore	Own

<u>Name 2</u>	<u>Home</u>
Menchú	Rent
Alvarez	Rent
Tagore	Own

This was a Product Operand

Relational Algebra Division:
 $R \div S$

Relational Algebra: Division

A Division is sort of like the reverse of a Product

This was a Product
Operand

<u>Name</u>	<u>Age</u>
Blackburn	5
Kobayashi	21



<u>Name 2</u>	<u>Home</u>
Menchú	Rent
Alvarez	Rent
Tagore	Own

This was a Product Operand

This was the result
of a Product

<u>Name</u>	<u>Age</u>	<u>Name 2</u>	<u>Home</u>
Blackburn	5	Menchú	Rent
Blackburn	5	Alvarez	Rent
Blackburn	5	Tagore	Own
Kobayashi	21	Menchú	Rent
Kobayashi	21	Alvarez	Rent
Kobayashi	21	Tagore	Own



Relational Algebra Division:
 $R \div S$

Relational Algebra: Division

<u>Name</u>	<u>Age</u>
Blackburn	5
Kobayashi	21



<u>Name 2</u>	<u>Home</u>
Menchú	Rent
Alvarez	Rent
Tagore	Own



<u>Name</u>	<u>Age</u>	<u>Name 2</u>	<u>Home</u>
Blackburn	5	Menchú	Rent
Blackburn	5	Alvarez	Rent
Blackburn	5	Tagore	Own
Kobayashi	21	Menchú	Rent
Kobayashi	21	Alvarez	Rent
Kobayashi	21	Tagore	Own
Sancar	54	Tagore	Own

Add another row to this table that did not result from the product.

Relational Algebra Division:
 $R \div S$

Relational Algebra: Division

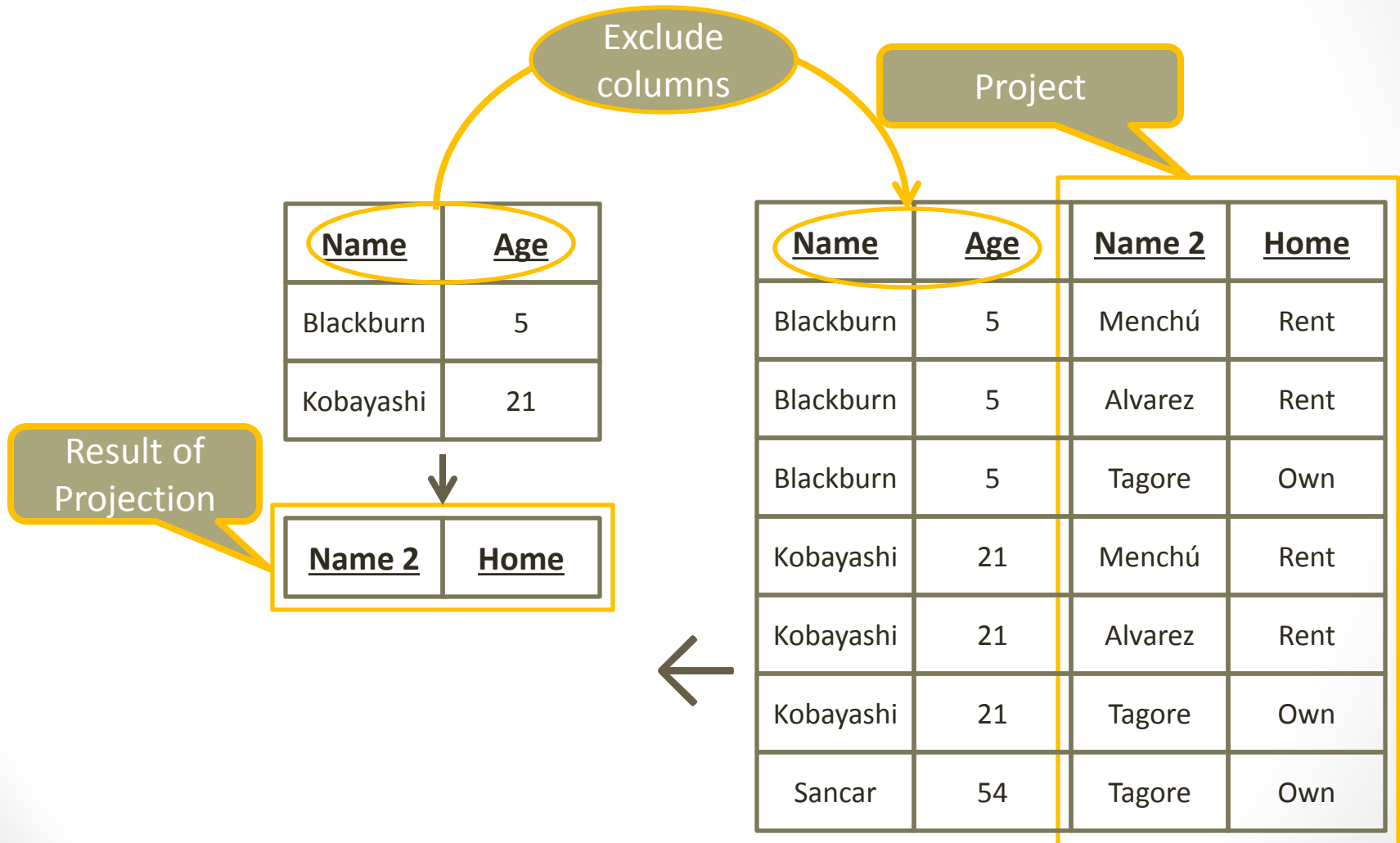
<u>Name</u>	<u>Age</u>
Blackburn	5
Kobayashi	21



<u>Name</u>	<u>Age</u>	<u>Name 2</u>	<u>Home</u>
Blackburn	5	Menchú	Rent
Blackburn	5	Alvarez	Rent
Blackburn	5	Tagore	Own
Kobayashi	21	Menchú	Rent
Kobayashi	21	Alvarez	Rent
Kobayashi	21	Tagore	Own
Sancar	54	Tagore	Own

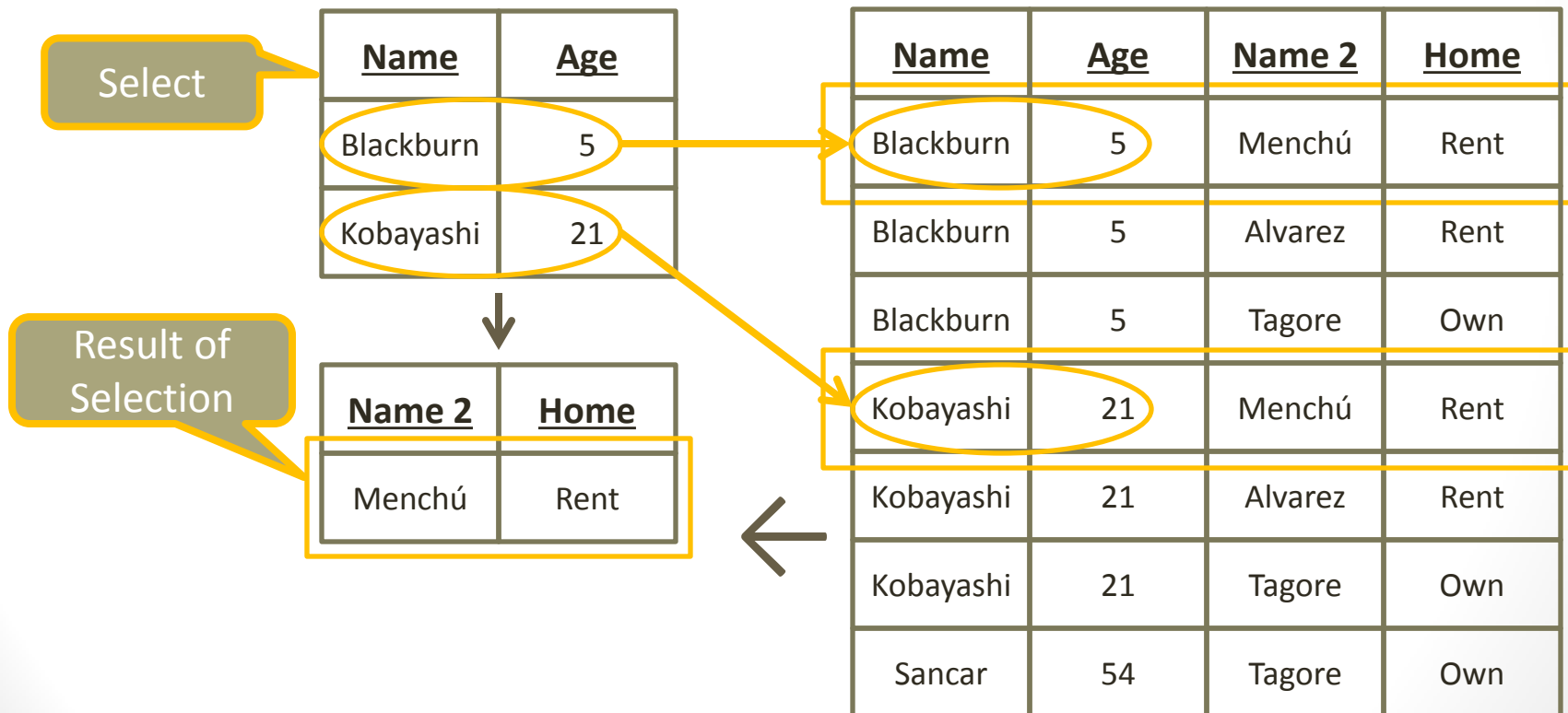
Relational Algebra Division:
 $R \div S$

Relational Algebra: Division



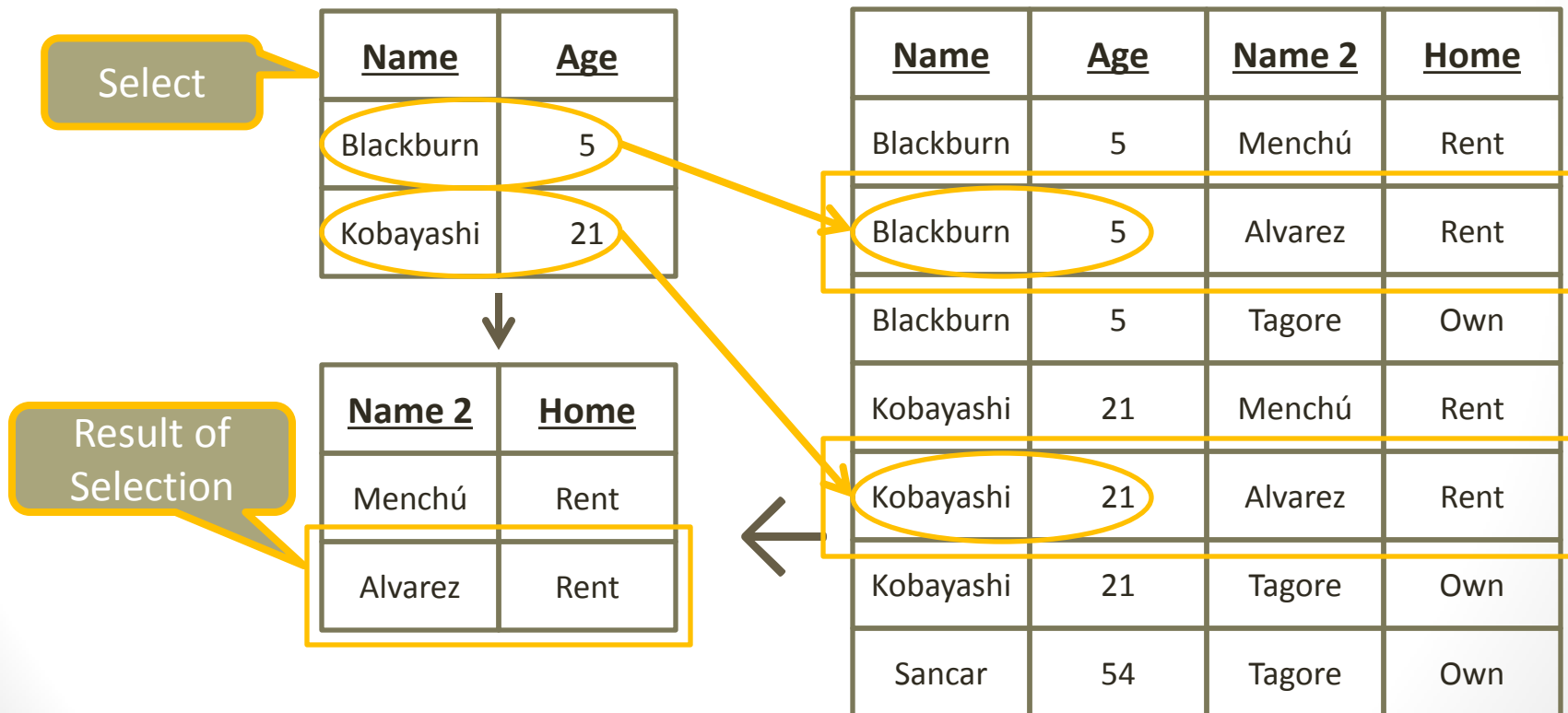
Relational Algebra Division:
 $R \div S$

Relational Algebra: Division



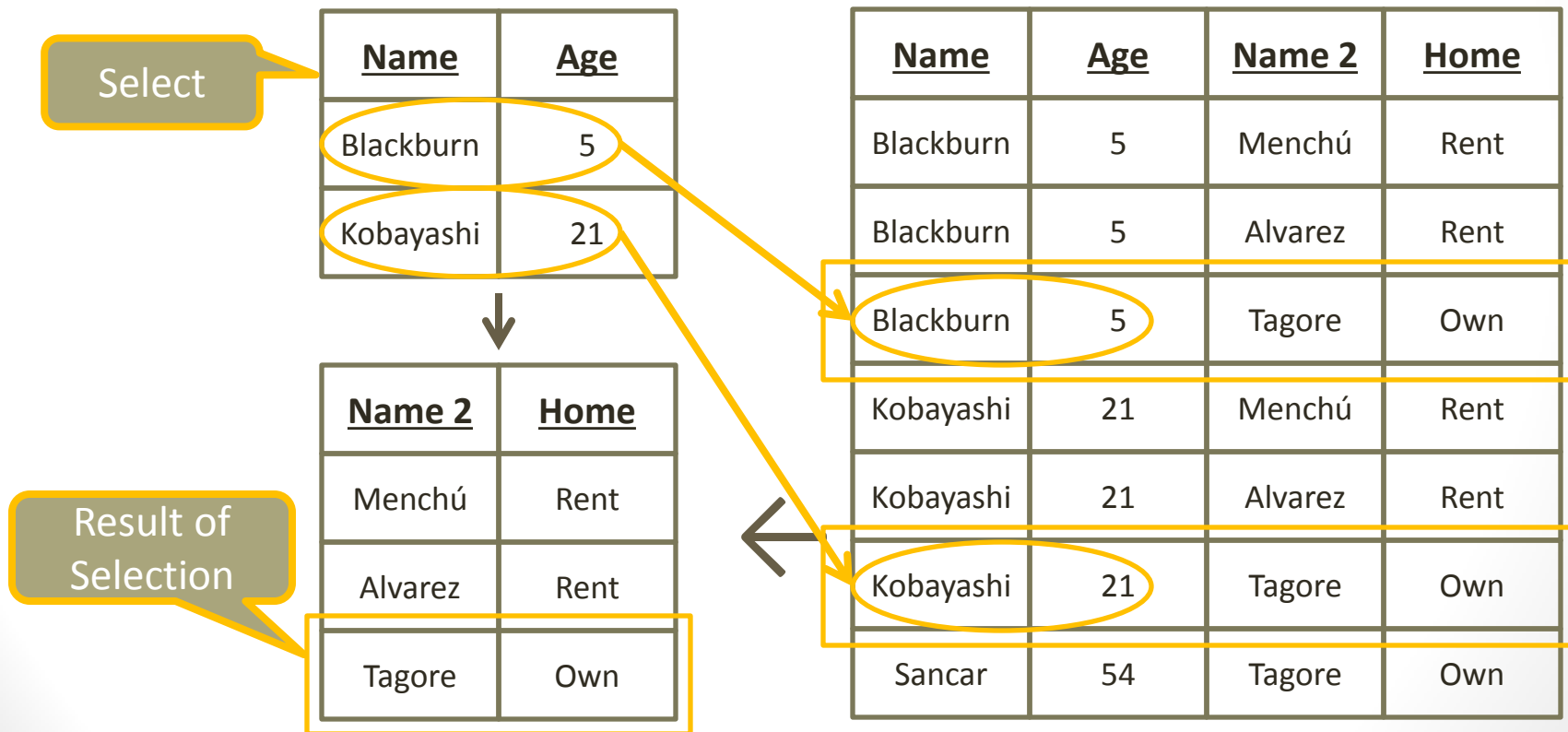
Relational Algebra Division:
 $R \div S$

Relational Algebra: Division



Relational Algebra Division:
 $R \div S$

Relational Algebra: Division



Relational Algebra Division:
 $R \div S$

Relational Algebra: Division

[Menchú, Rent] is in the same tuple as
[Blackburn, 5] and [Kobayashi, 21]

<u>Name</u>	<u>Age</u>
Blackburn	5
Kobayashi	21



<u>Name 2</u>	<u>Home</u>
Menchú	Rent
Alvarez	Rent
Tagore	Own



<u>Name</u>	<u>Age</u>	<u>Name 2</u>	<u>Home</u>
Blackburn	5	Menchú	Rent
Blackburn	5	Alvarez	Rent
Blackburn	5	Tagore	Own
Kobayashi	21	Menchú	Rent
Kobayashi	21	Alvarez	Rent
Kobayashi	21	Tagore	Own
Sancar	54	Tagore	Own

Relational Algebra Division:
 $R \div S$

Relational Algebra: Division

[Alvarez, Rent] is in the same tuple as
[Blackburn, 5] and [Kobayashi, 21]

<u>Name</u>	<u>Age</u>
Blackburn	5
Kobayashi	21



<u>Name 2</u>	<u>Home</u>
Menchú	Rent
Alvarez	Rent
Tagore	Own



<u>Name</u>	<u>Age</u>	<u>Name 2</u>	<u>Home</u>
Blackburn	5	Menchú	Rent
Blackburn	5	Alvarez	Rent
Blackburn	5	Tagore	Own
Kobayashi	21	Menchú	Rent
Kobayashi	21	Alvarez	Rent
Kobayashi	21	Tagore	Own
Sancar	54	Tagore	Own

Relational Algebra Division:
 $R \div S$

Relational Algebra: Division

[Tagore, Own] is in the same tuple as
[Blackburn, 5] and [Kobayashi, 21]

<u>Name</u>	<u>Age</u>
Blackburn	5
Kobayashi	21



<u>Name 2</u>	<u>Home</u>
Menchú	Rent
Alvarez	Rent
Tagore	Own

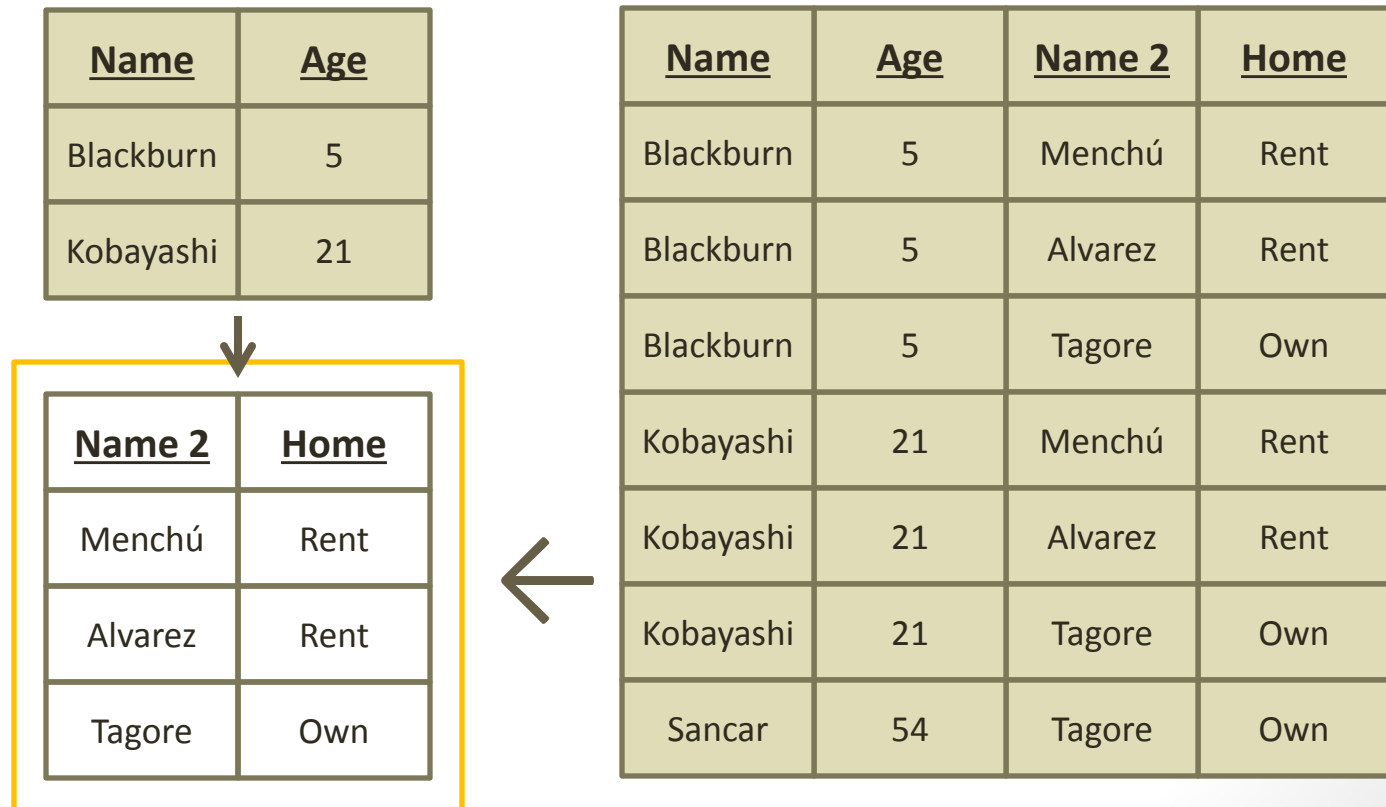


<u>Name</u>	<u>Age</u>	<u>Name 2</u>	<u>Home</u>
Blackburn	5	Menchú	Rent
Blackburn	5	Alvarez	Rent
Blackburn	5	Tagore	Own
Kobayashi	21	Menchú	Rent
Kobayashi	21	Alvarez	Rent
Kobayashi	21	Tagore	Own
Sancar	54	Tagore	Own

Relational Algebra Division:
 $R \div S$

Relational Algebra: Division

The result of a division is a relation with n tuples of arity l where the dividend operand has at least $n \cdot m$ tuples of arity $i + j$ and the divisor operand has exactly m tuples of arity j that are a subset of the of the dividend tuples.



Relational Algebra Division:
 $R \div S$

Relational Algebra: Resources

- Examples for Relational Algebra and SQL
 - RelationalAlgebraAndSQL.pdf
 - RelationalAlgebraAndSQL.sql
- Links for definitions and concepts:
 - http://en.wikipedia.org/wiki/Cartesian_product
 - http://en.wikipedia.org/wiki/Commutative_property
 - http://en.wikipedia.org/wiki/Associative_property
 - [http://en.wikipedia.org/wiki/Closure_\(mathematics\)](http://en.wikipedia.org/wiki/Closure_(mathematics))
 - http://en.wikipedia.org/wiki/Relational_calculus
 - http://en.wikipedia.org/wiki/Relational_algebra
 - http://en.wikipedia.org/wiki/Edgar_F._Codd
 - http://en.wikipedia.org/wiki/Relational_model
 - http://en.wikipedia.org/wiki/Relational_database
 - http://en.wikipedia.org/wiki/Query_language
 - <http://en.wikipedia.org/wiki/SQL>
 - <http://en.wikipedia.org/wiki/NoSQL>

Quiz 06c

- Quiz 06c (Product, Join, Division)



Break



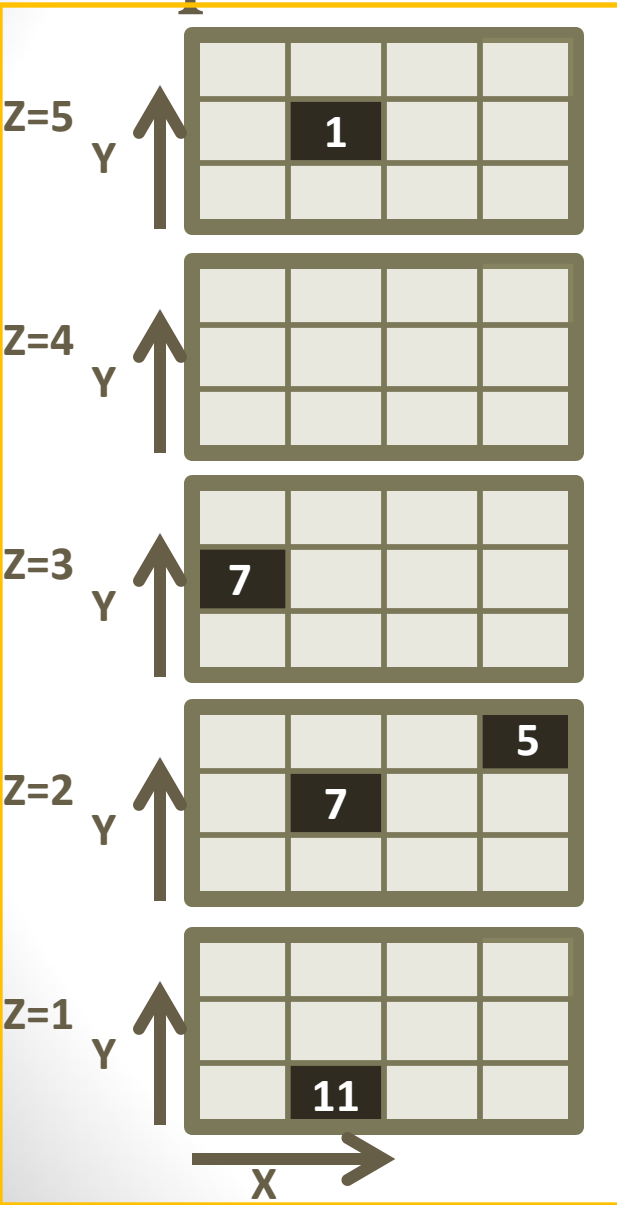
Relational Algebra

Data as Sparse Matrices

Cartesian Product

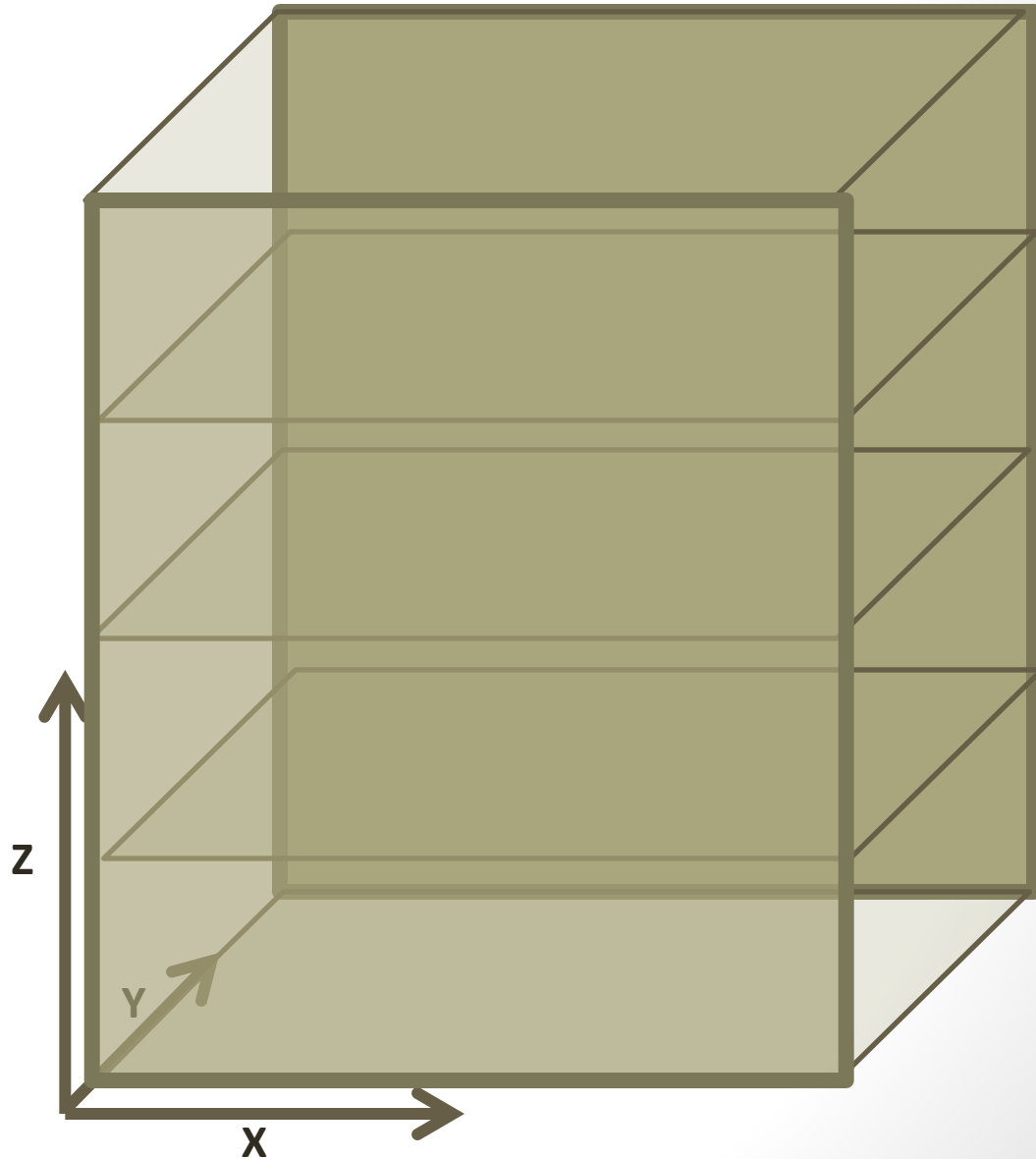
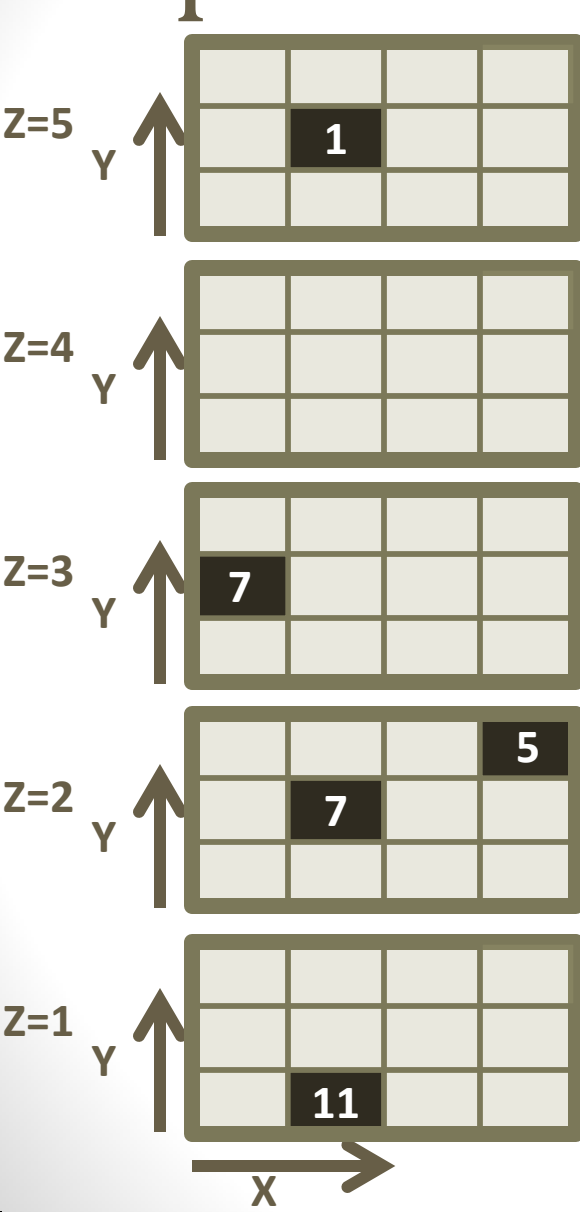
- Cartesian product
- http://en.wikipedia.org/wiki/Cartesian_product
- The Cartesian product of two sets A and B is the set of all ordered pairs ab , where a is element of A and b is element of B.
- Relational Algebra
- http://en.wikipedia.org/wiki/Relational_algebra
- In Relational Algebra we need the Cartesian product to combine tuples into a single tuple. The Cartesian product creates a new schema (relation) from other relations.
- Hyperrectangle (Sparse Multi-Dimensional Matrix)
- <http://en.wikipedia.org/wiki/Hyperrectangle>
- Hyperrectangle is the generalization of a rectangle for higher dimensions and is defined as the Cartesian product of intervals

Sparse Matrices

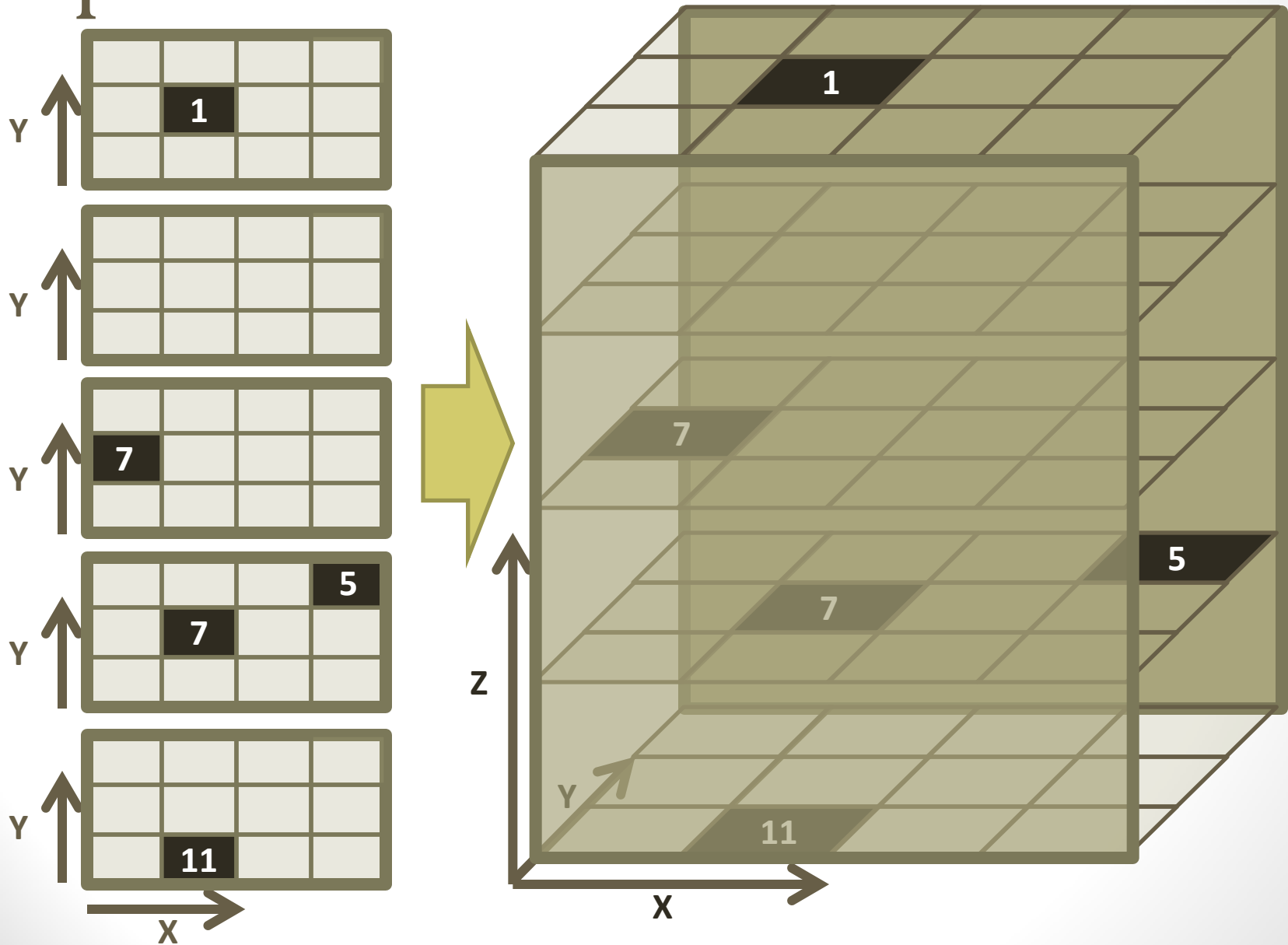


A series of equal-sized 2-dimensional matrices is a 3-dimensional matrix

Sparse Matrices



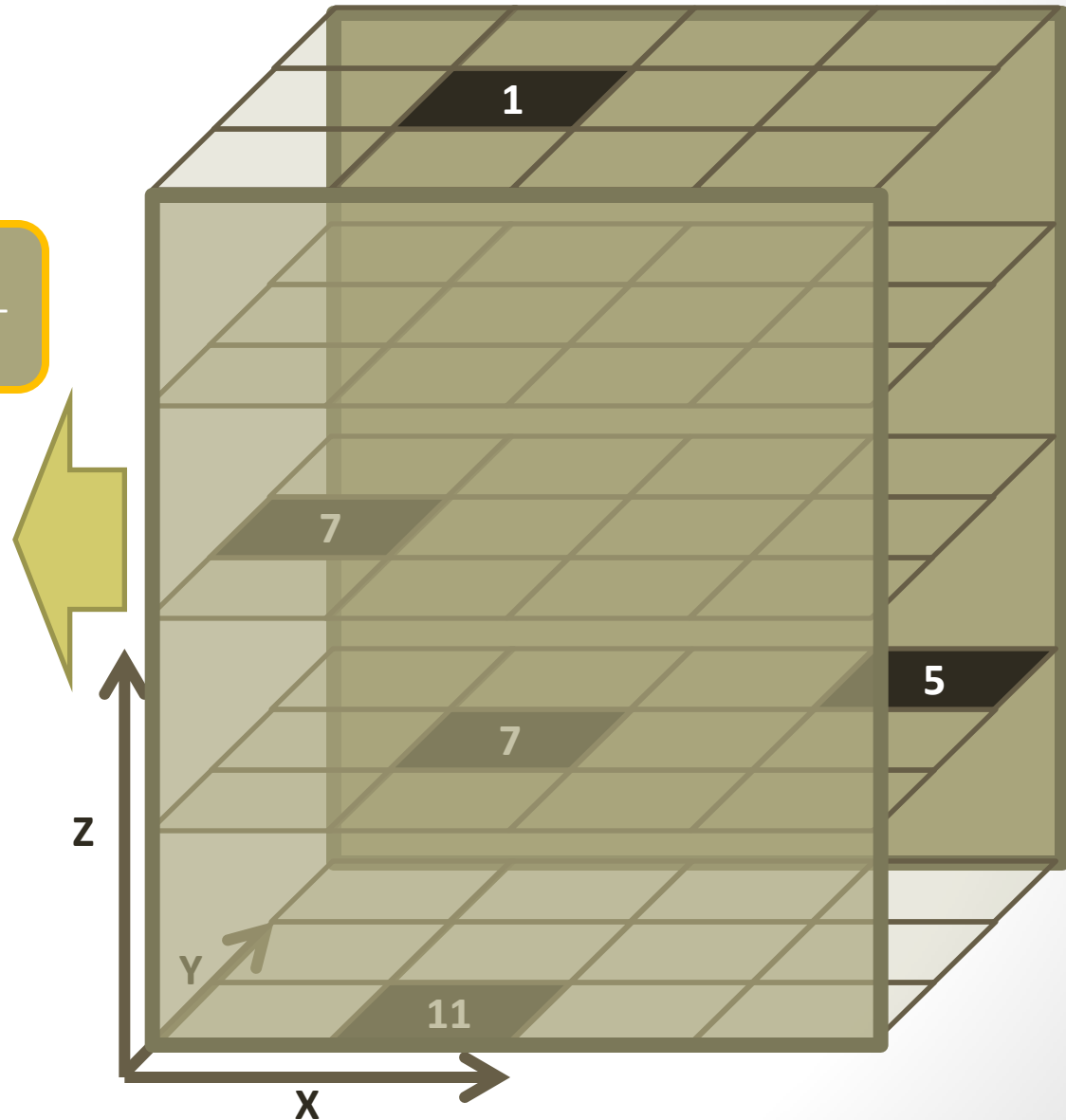
Sparse Matrices



Sparse Matrices

A table with n columns represents values in an $n-1$ dimensional matrix

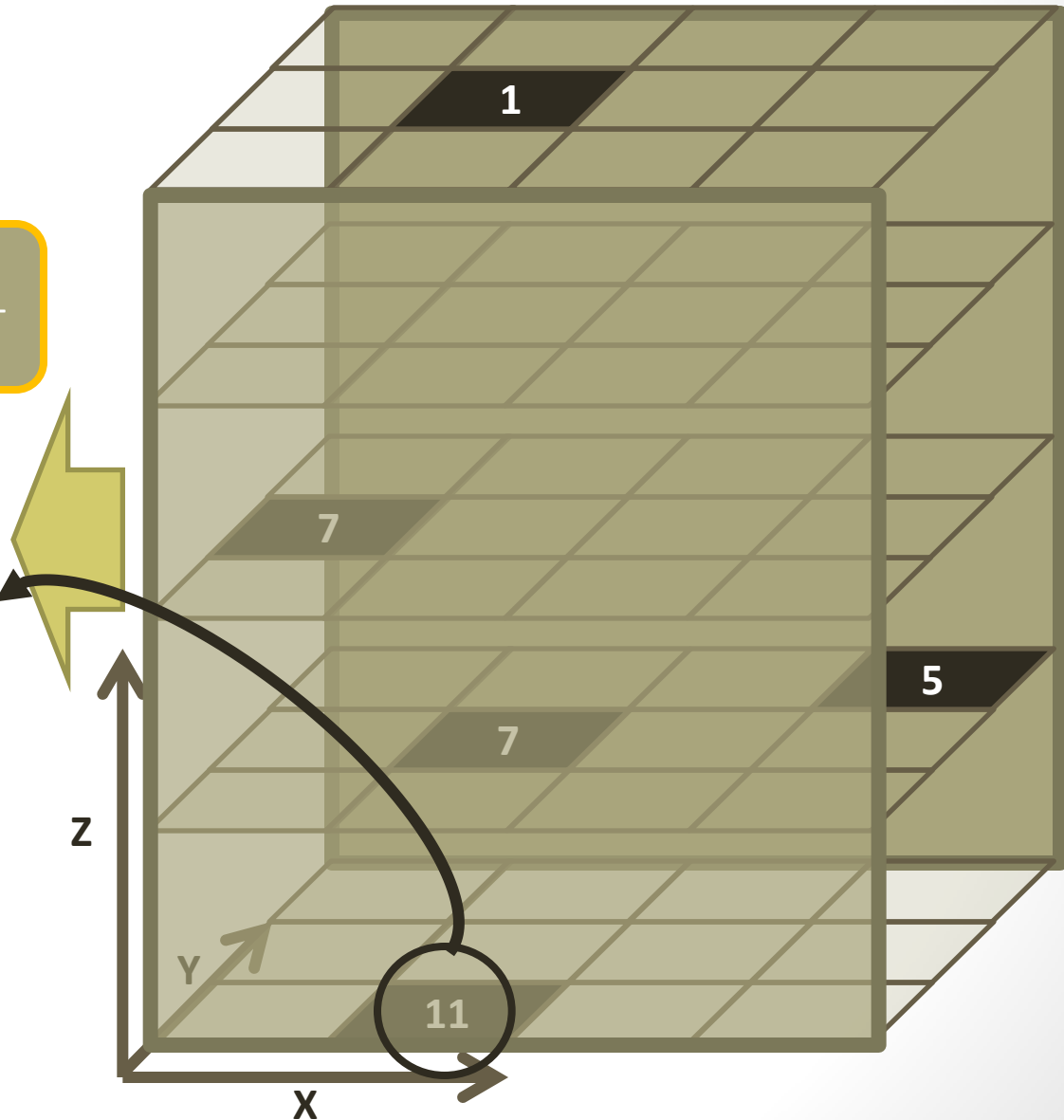
<u>X</u>	<u>Y</u>	<u>Z</u>	<u>V</u>



Sparse Matrices

A table with n columns represents values in an $n-1$ dimensional matrix

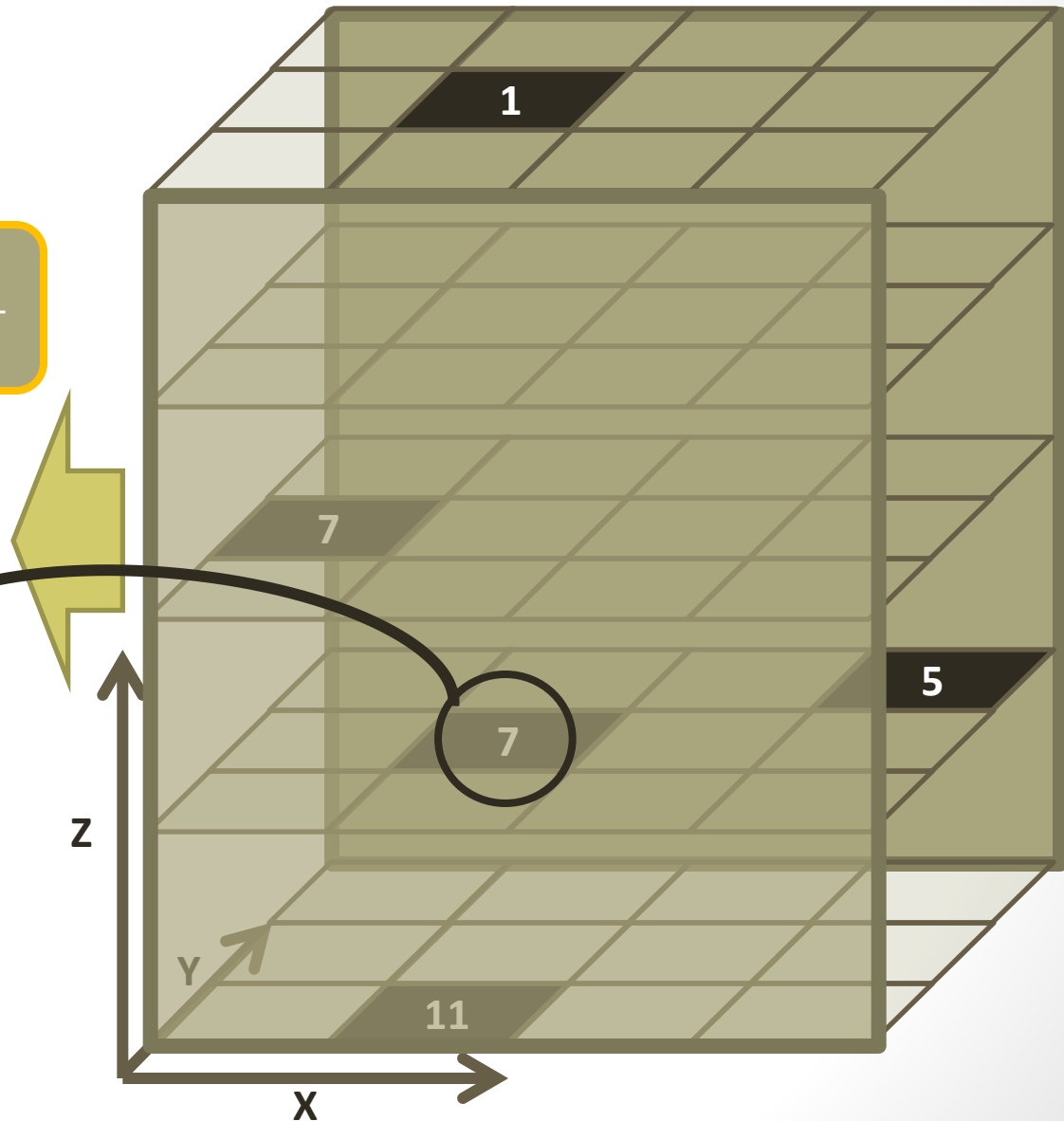
<u>X</u>	<u>Y</u>	<u>Z</u>	<u>V</u>
2	1	1	11



Sparse Matrices

A table with n columns represents values in an $n-1$ dimensional matrix

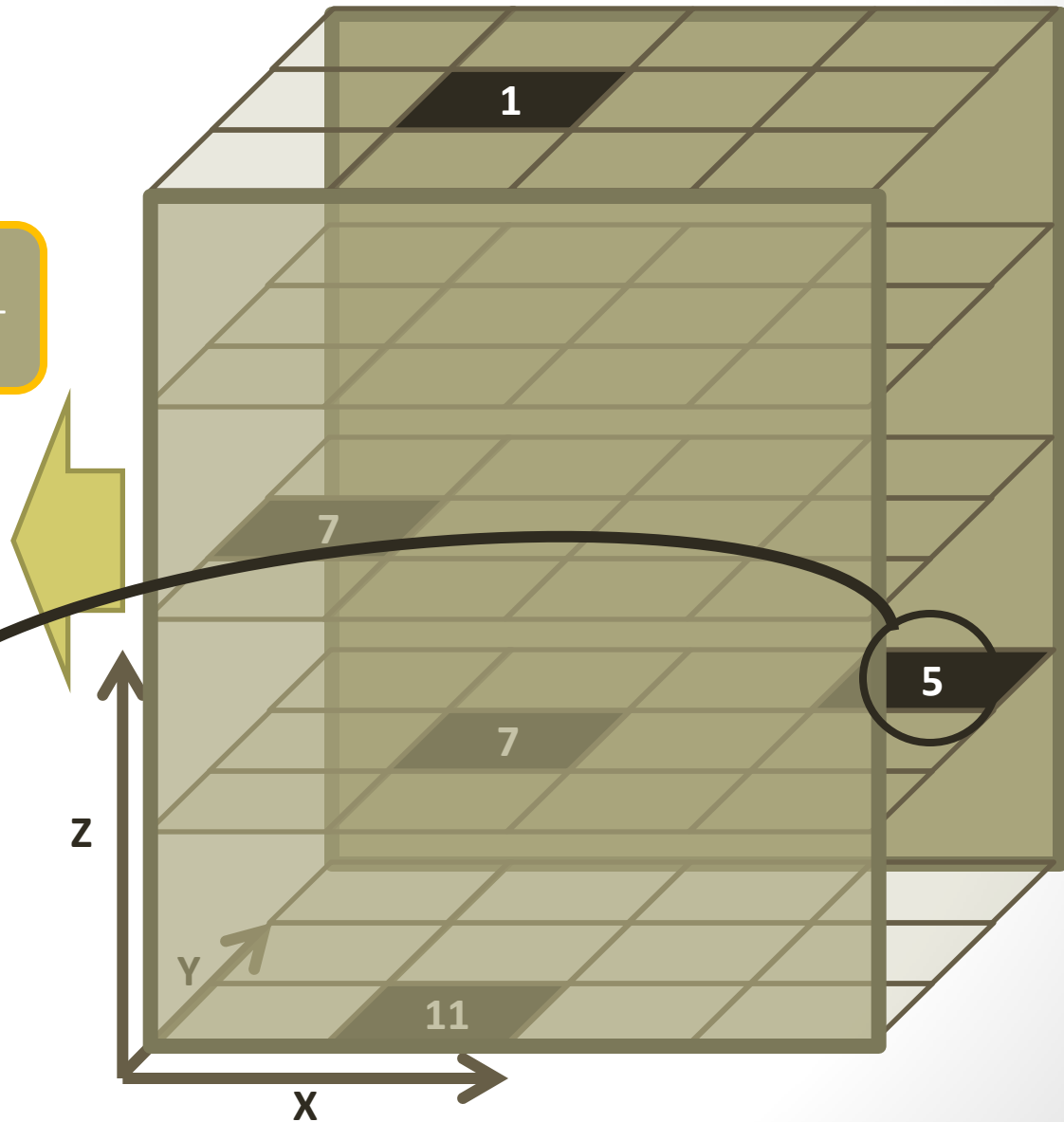
<u>X</u>	<u>Y</u>	<u>Z</u>	<u>V</u>
2	1	1	11
2	2	2	7



Sparse Matrices

A table with n columns represents values in an $n-1$ dimensional matrix

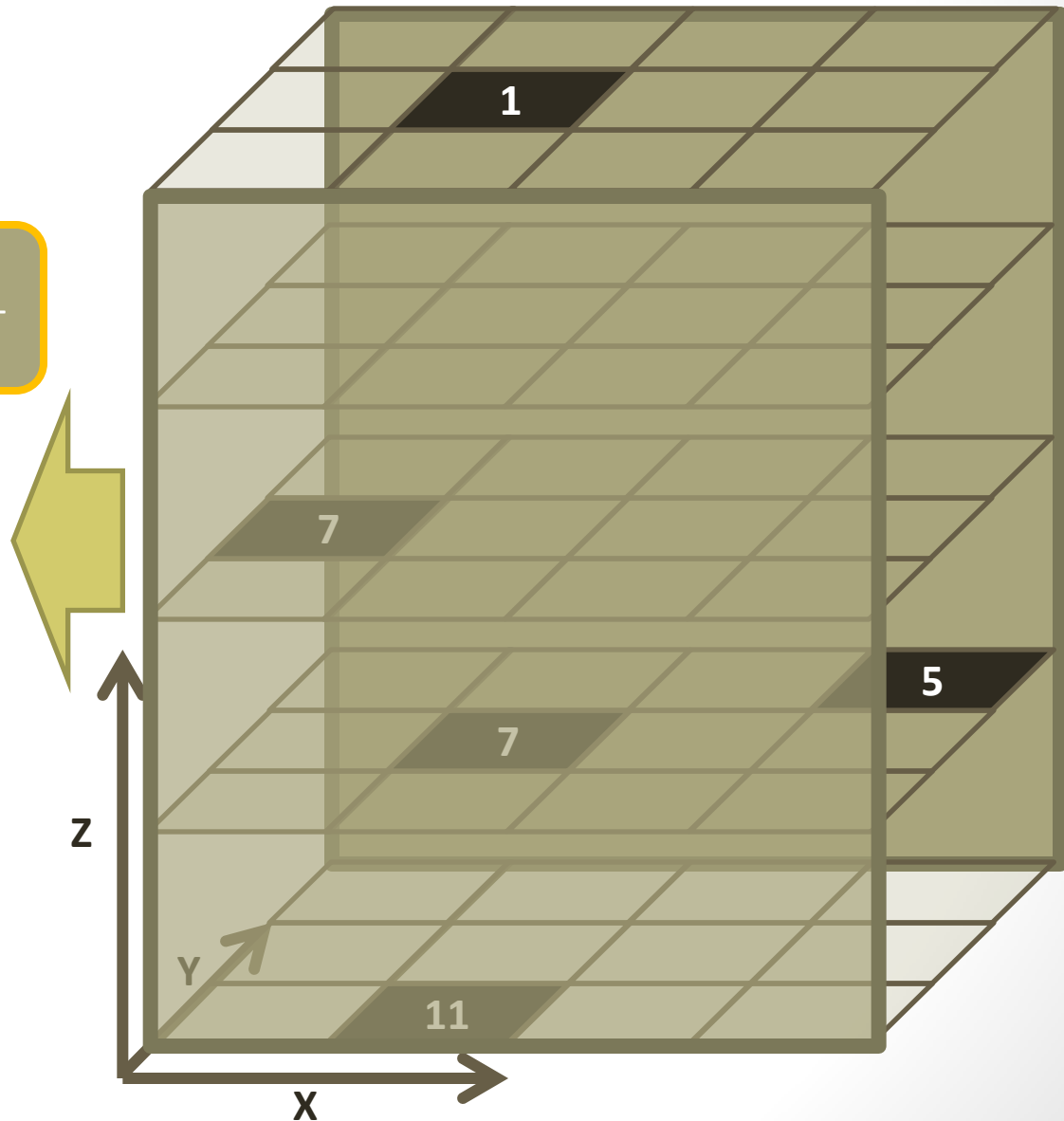
<u>X</u>	<u>Y</u>	<u>Z</u>	<u>V</u>
2	1	1	11
2	2	2	7
4	3	2	5



Sparse Matrices

A table with n columns represents values in an $n-1$ dimensional matrix

<u>X</u>	<u>Y</u>	<u>Z</u>	<u>V</u>
2	1	1	11
2	2	2	7
4	3	2	5
1	2	3	7
2	2	5	1



Sparse Matrices

<u>X</u>	<u>Y</u>	<u>Z</u>	<u>V</u>
2	1	1	11
2	2	2	7
4	3	2	5
1	2	3	7
2	2	5	1

Sparse Matrices

<u>X</u>	<u>Y</u>	<u>Z</u>	<u>V</u>
2	1	1	11
2	2	2	7
4	3	2	5
1	2	3	7
2	2	5	1

Think of V as just
another dimension

Sparse Matrices

A table with n columns represents points in an n -dimensional matrix

<u>X</u>	<u>Y</u>	<u>Z</u>	<u>V</u>
2	1	1	11
2	2	2	7
4	3	2	5
1	2	3	7
2	2	5	1

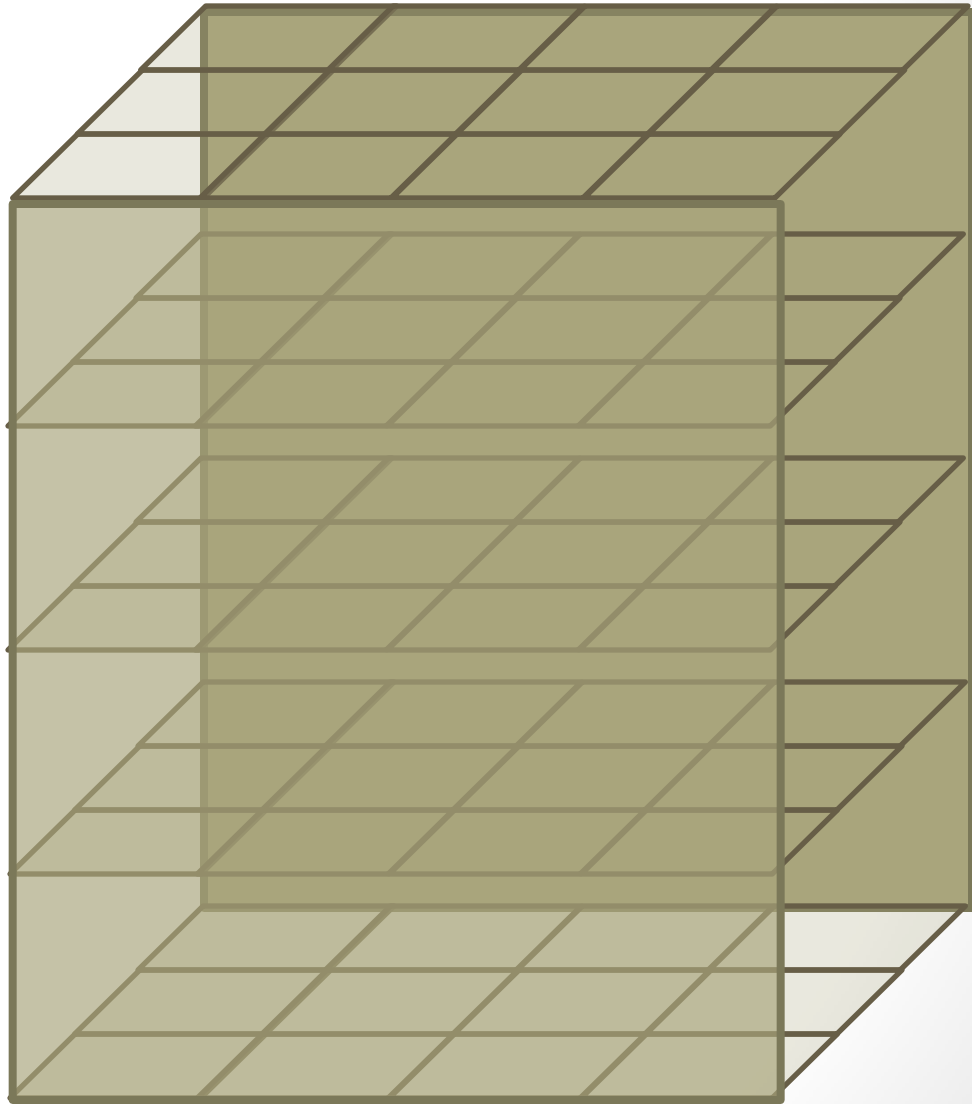
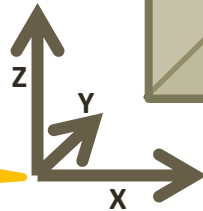
Think of V as just another dimension

Sparse Matrices

This table represents
points in 4-Dimensional
Space.

<u>X</u>	<u>Y</u>	<u>Z</u>	<u>V</u>
2	1	1	11
2	2	2	7
4	3	2	5
1	2	3	7
2	2	5	1

3 Dimensions

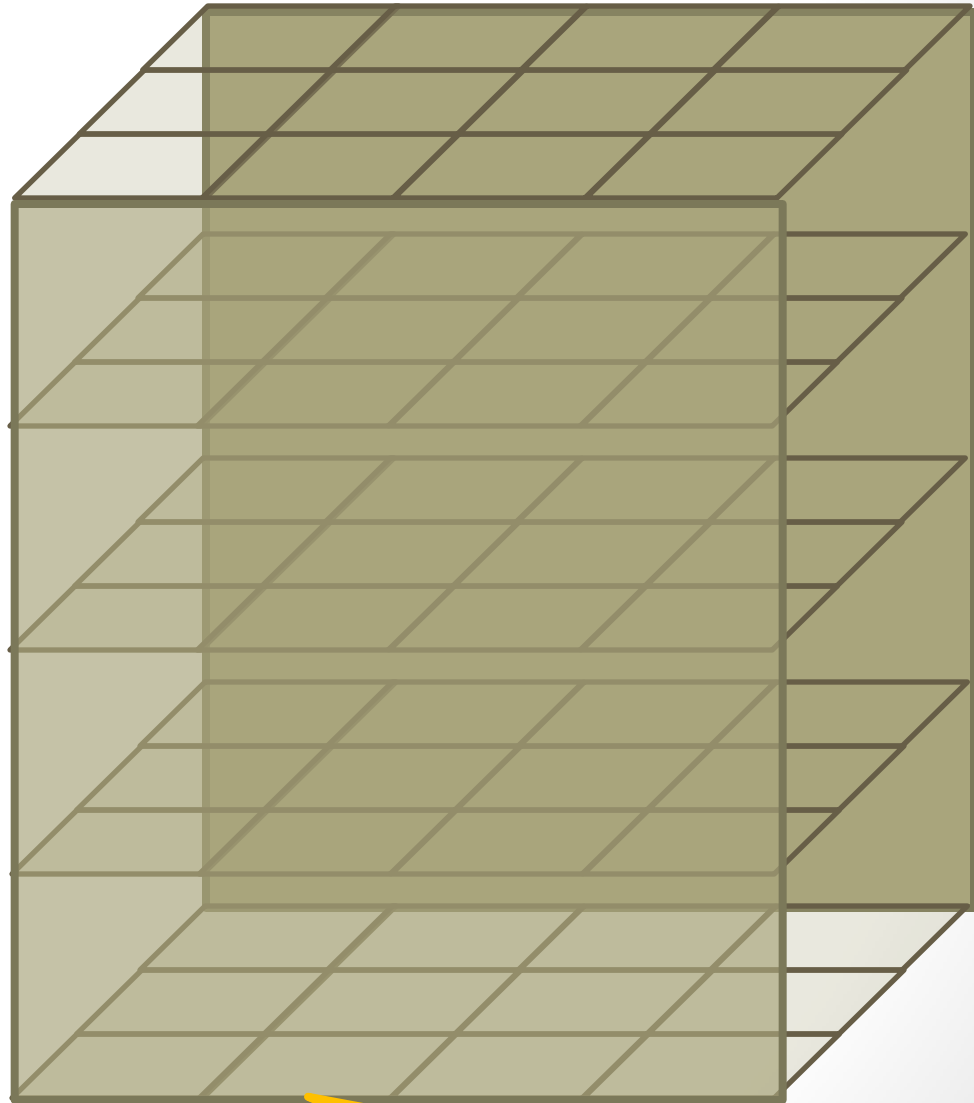
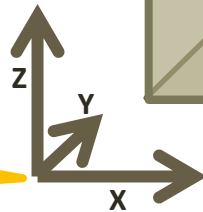


Sparse Matrices

This table represents
points in 4-Dimensional
Space.

<u>X</u>	<u>Y</u>	<u>Z</u>	<u>V</u>
2	1	1	11
2	2	2	7
4	3	2	5
1	2	3	7
2	2	5	1

3 Dimensions



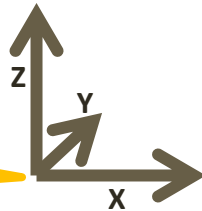
3-Dimensional Space

Sparse Matrices

This table represents
points in 4-Dimensional
Space.

<u>X</u>	<u>Y</u>	<u>Z</u>	<u>V</u>
2	1	1	11
2	2	2	7
4	3	2	5
1	2	3	7
2	2	5	1

3 Dimensions



3-Dimensional Space

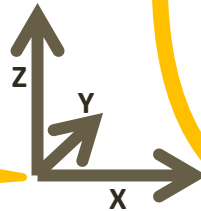
Sparse Matrices

This table represents
points in 4-Dimensional
Space.

<u>X</u>	<u>Y</u>	<u>Z</u>	<u>V</u>
2	1	1	11
2	2	2	7
4	3	2	5
1	2	3	7
2	2	5	1



3 Dimensions



?

4-Dimensional
Space

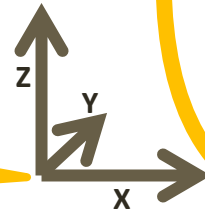


Sparse Matrices

This table represents points in 4-Dimensional Space.

<u>X</u>	<u>Y</u>	<u>Z</u>	<u>V</u>
2	1	1	11
2	2	2	7
4	3	2	5
1	2	3	7
2	2	5	1

3 Dimensions



?

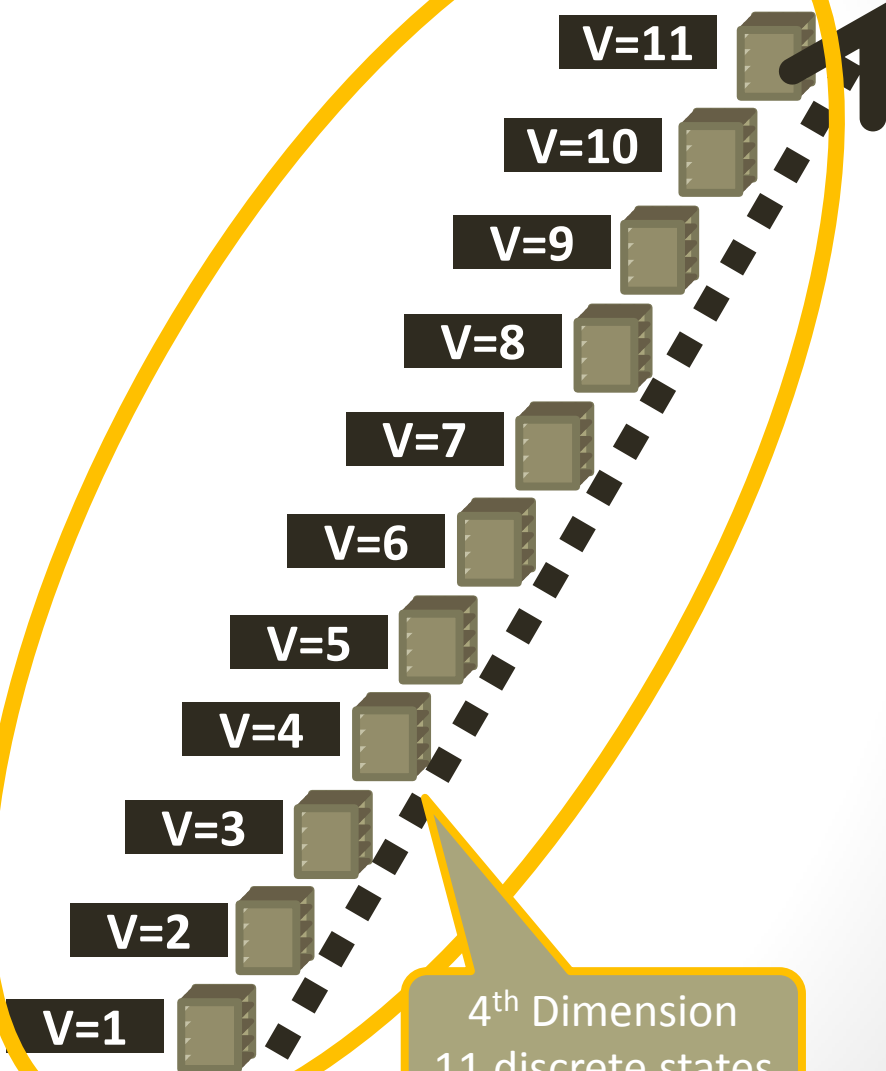
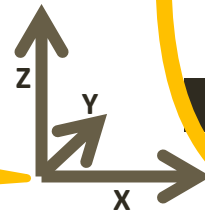
4th Dimension

Sparse Matrices

This table represents points in 4-Dimensional Space.

<u>X</u>	<u>Y</u>	<u>Z</u>	<u>V</u>
2	1	1	11
2	2	2	7
4	3	2	5
1	2	3	7
2	2	5	1

3 Dimensions

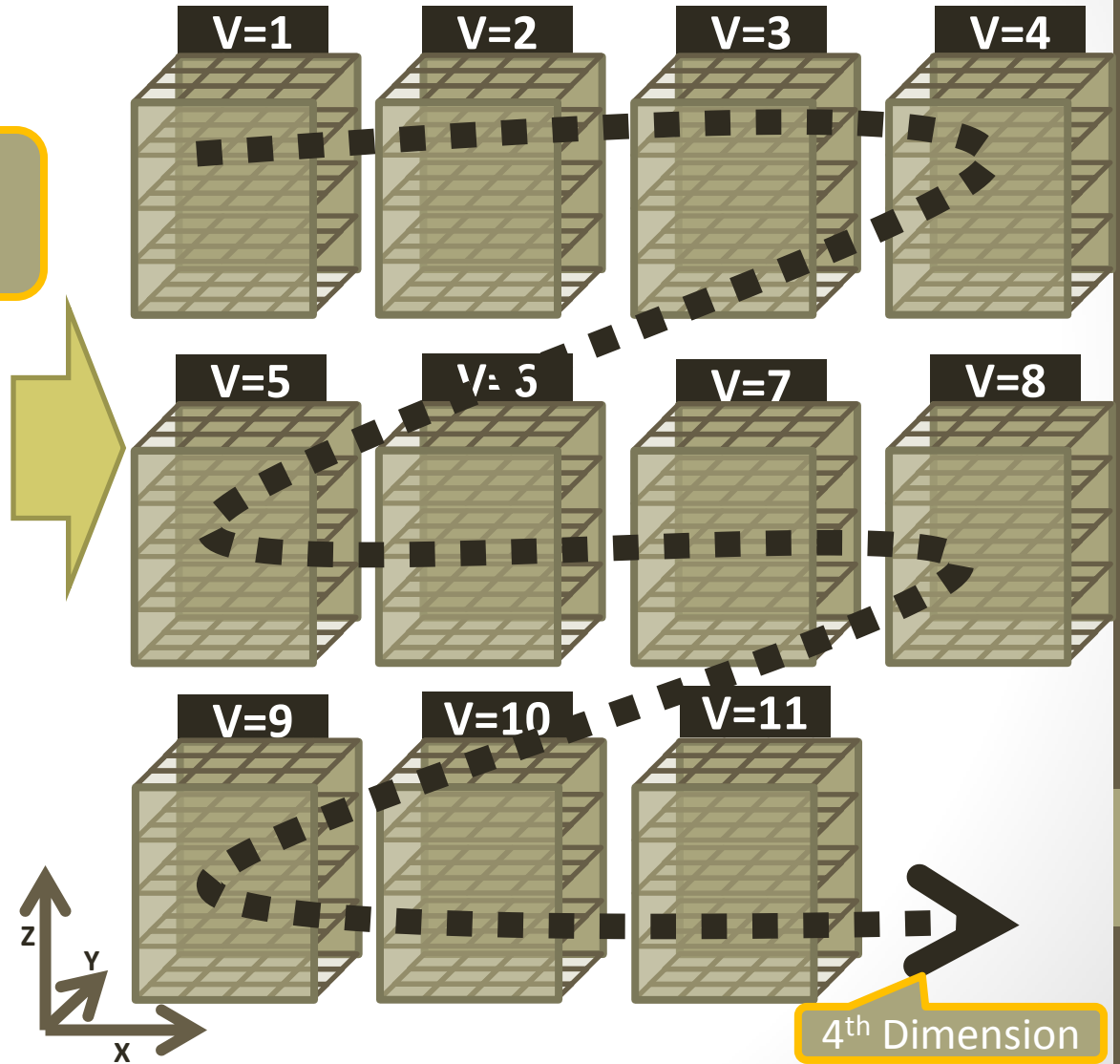


4th Dimension
11 discrete states

Sparse Matrices

This table represents points in 4-Dimensional Space.

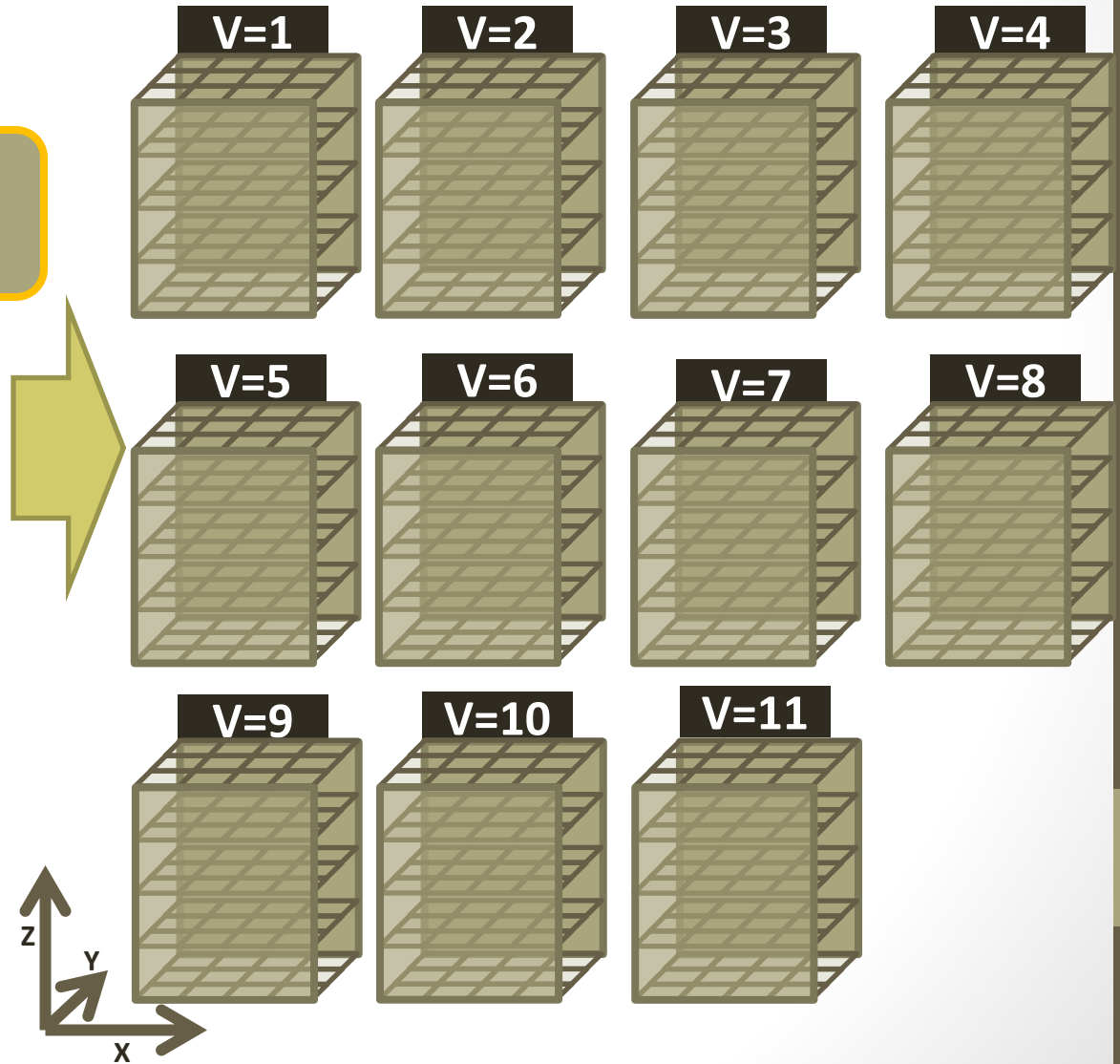
<u>X</u>	<u>Y</u>	<u>Z</u>	<u>V</u>
2	1	1	11
2	2	2	7
4	3	2	5
1	2	3	7
2	2	5	1



Sparse Matrices

This table represents
points in 4-Dimensional
Space.

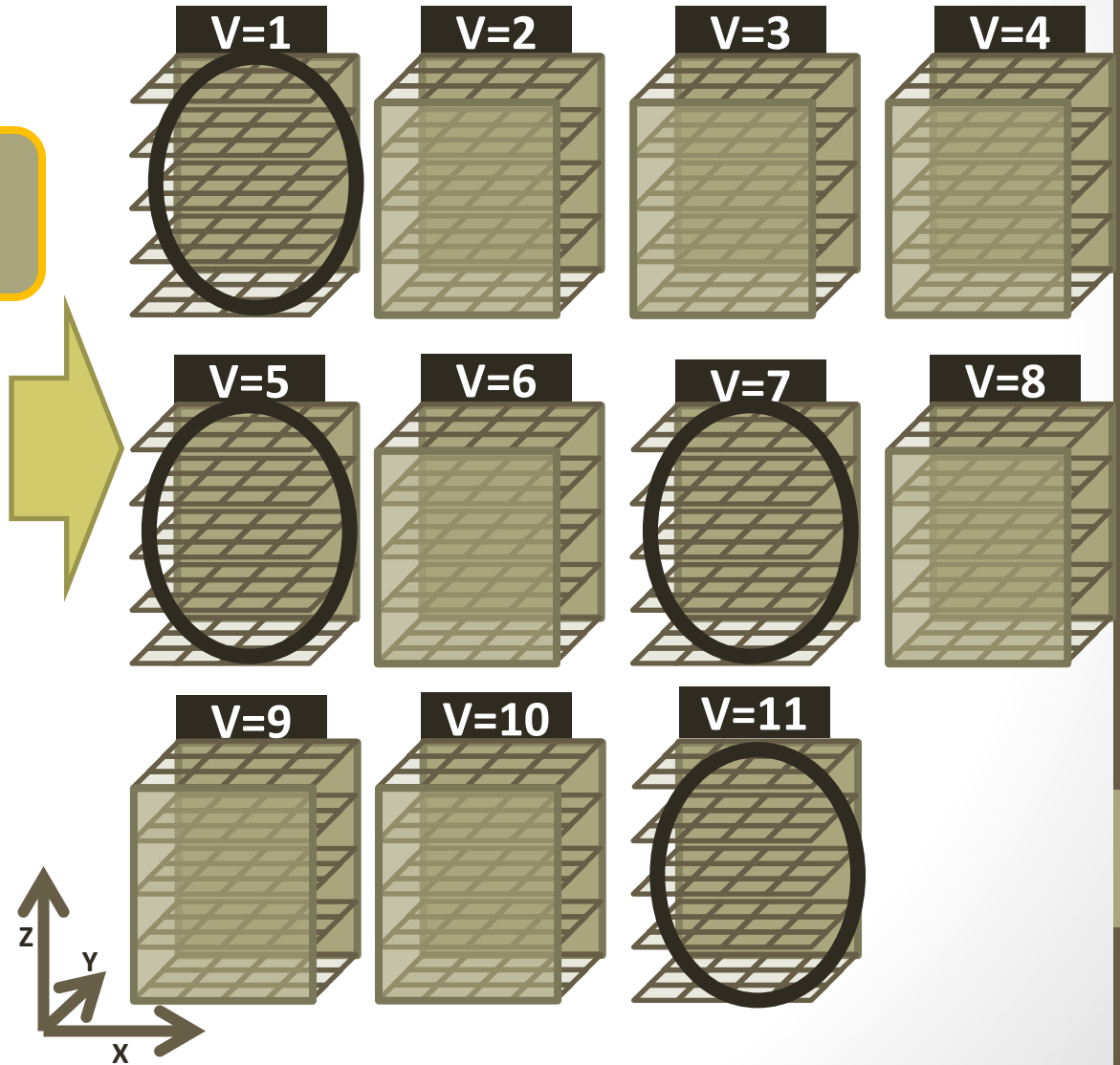
<u>X</u>	<u>Y</u>	<u>Z</u>	<u>V</u>
2	1	1	11
2	2	2	7
4	3	2	5
1	2	3	7
2	2	5	1



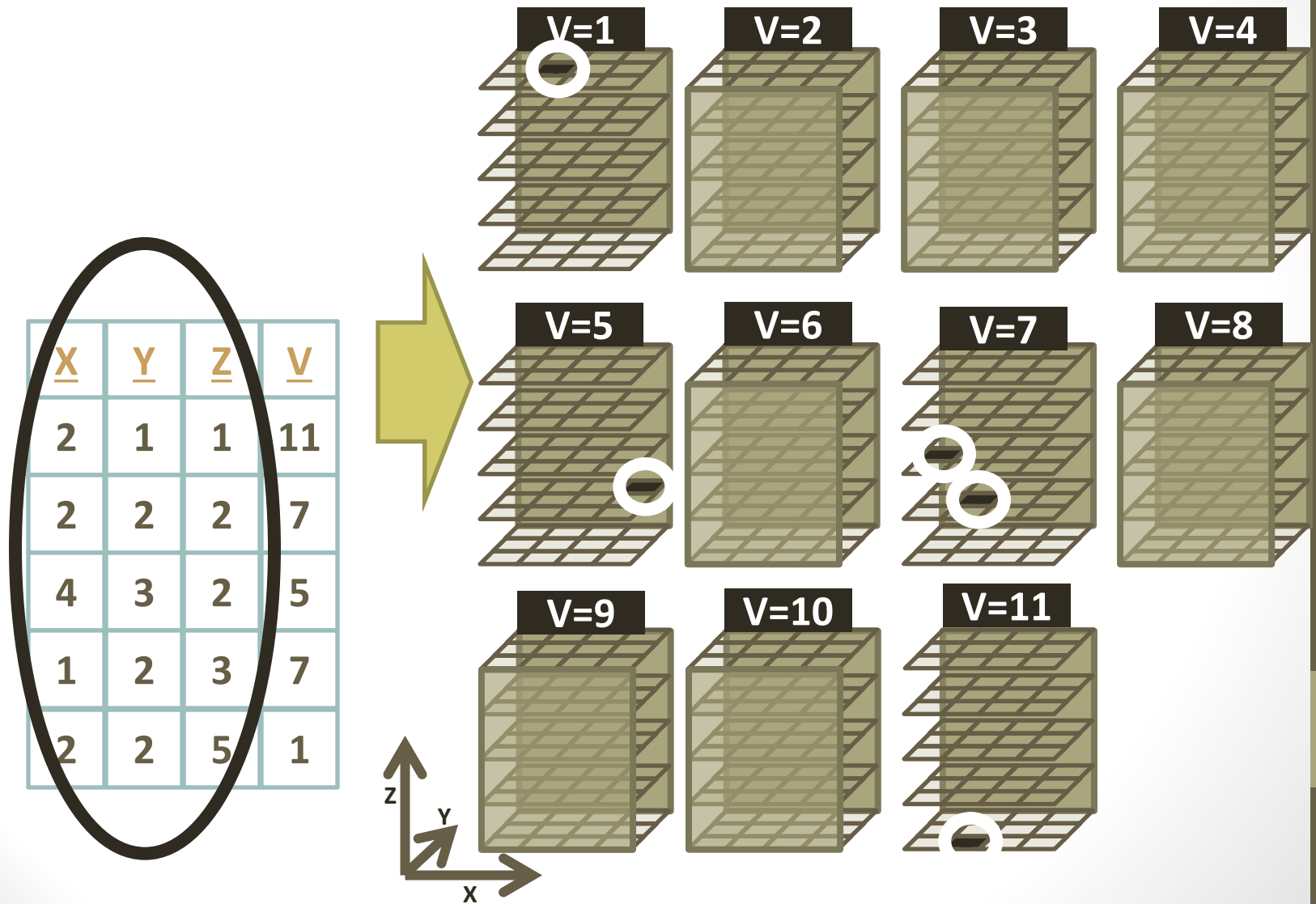
Sparse Matrices

This table represents points in 4-Dimensional Space.

<u>X</u>	<u>Y</u>	<u>Z</u>	<u>V</u>
2	1	1	11
2	2	2	7
4	3	2	5
1	2	3	7
2	2	5	1



Sparse Matrices



Sparse Matrices: EAV

<u>X</u>	<u>Y</u>	<u>Z</u>	<u>V</u>
2	1	1	11
2	2	2	7
4	3	2	5
1	2	3	7
2	2	5	1

Sparse Matrices: EAV

A table represents points
in n-Dimensional Space.

<u>X</u>	<u>Y</u>	<u>Z</u>	<u>V</u>
2	1	1	11
2	2	2	7
4	3	2	5
1	2	3	7
2	2	5	1

Sparse Matrices: EAV

A table represents points in n-Dimensional Space.

<u>X</u>	<u>Y</u>	<u>Z</u>	<u>V</u>
2	1	1	11
2	2	2	7
4	3	2	5
1	2	3	7
2	2	5	1

- Can we represent all tables in a single schema?
- Any table or matrix cell can be described by row, column and value.
- Represent each cell of a table in its own row.
- Entity-attribute-value model

Sparse Matrices: EAV

Row ID. Needs to be unique for a given row in the original table. Does not need to be a number or sequential

Column Name

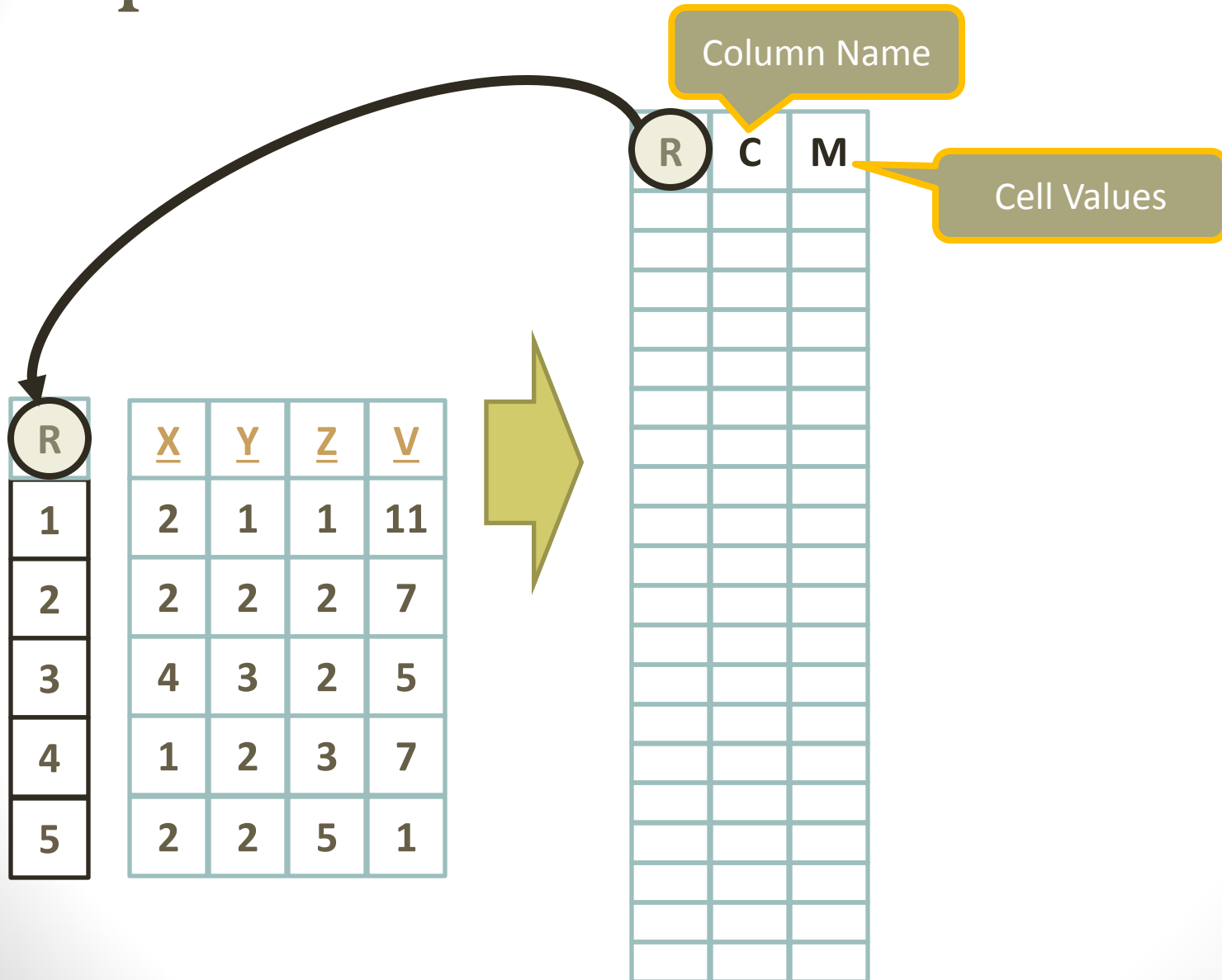
Cell Values

<u>X</u>	<u>Y</u>	<u>Z</u>	<u>V</u>
2	1	1	11
2	2	2	7
4	3	2	5
1	2	3	7
2	2	5	1

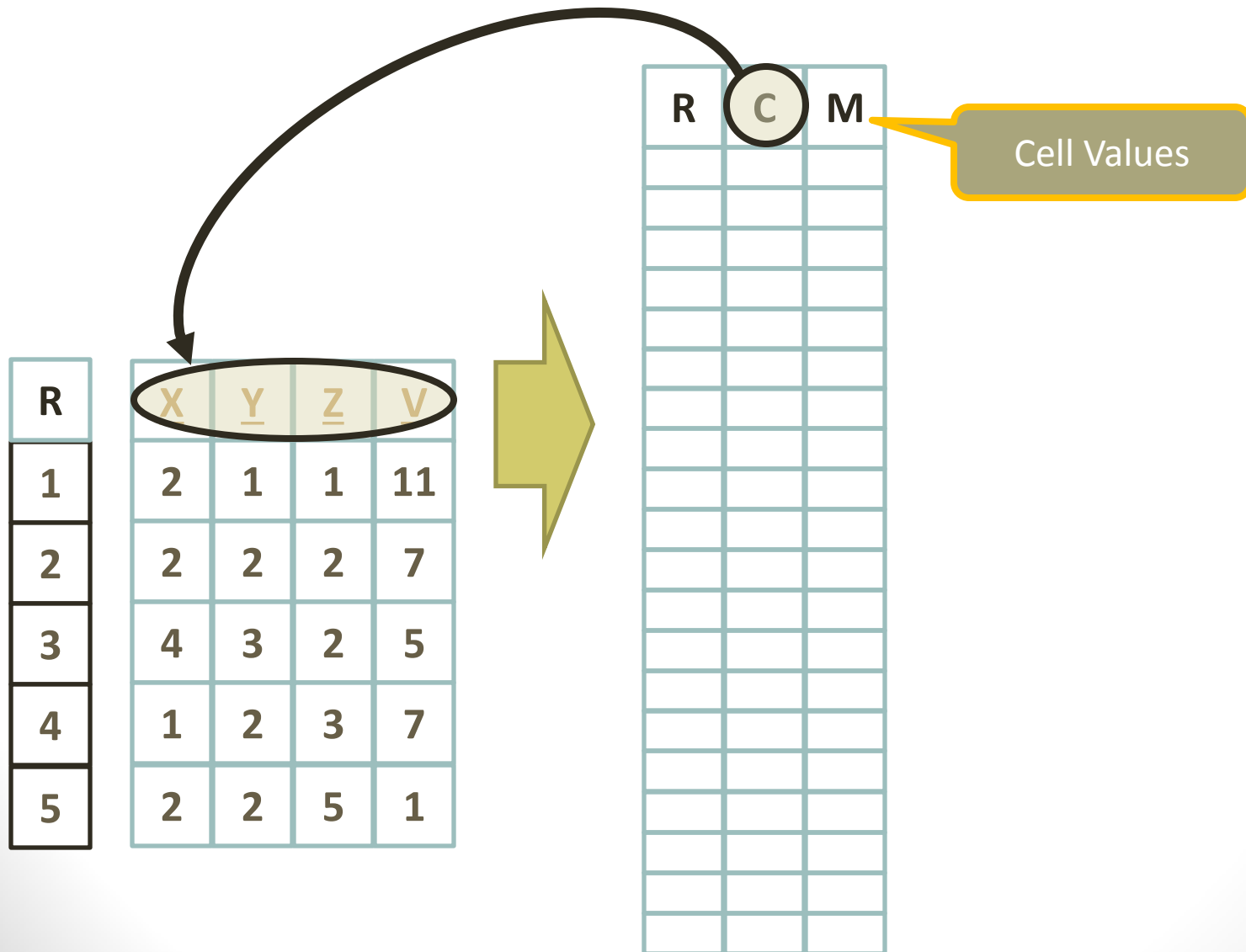
[illegible]

- Can we represent all tables in a single schema?
- Any table or matrix cell can be described by row, column and value.
- Represent each cell of a table in its own row.
- Entity-attribute-value model

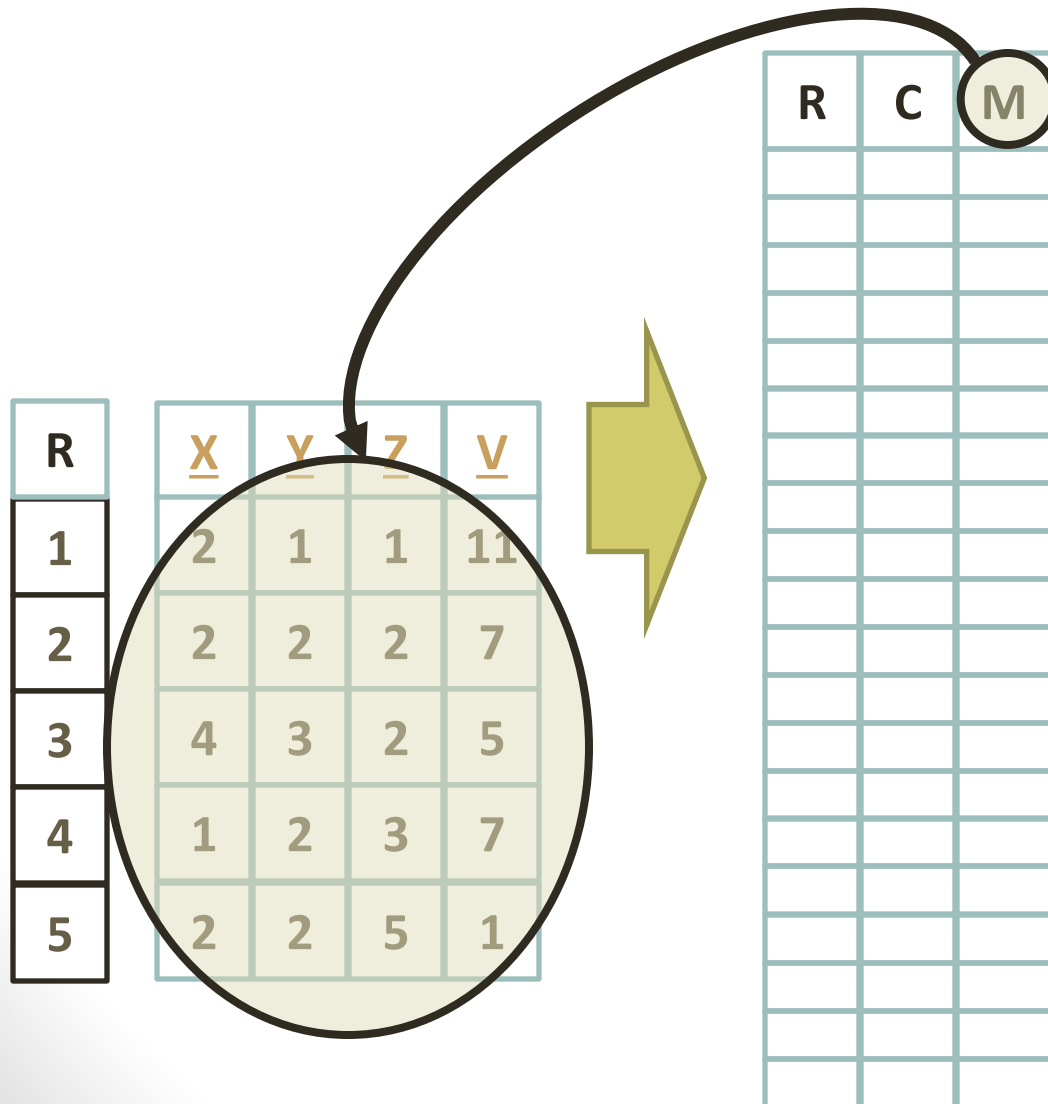
Sparse Matrices: EAV



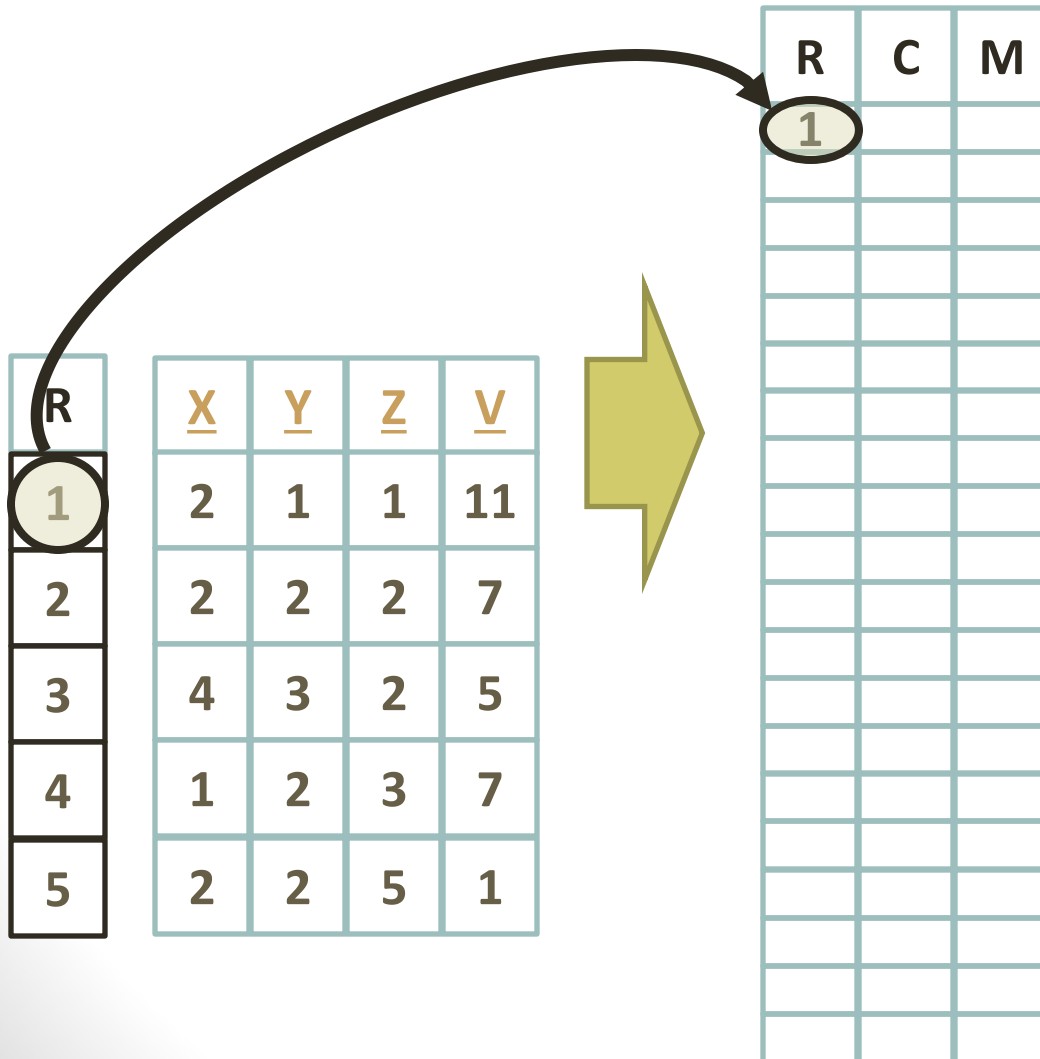
Sparse Matrices: EAV



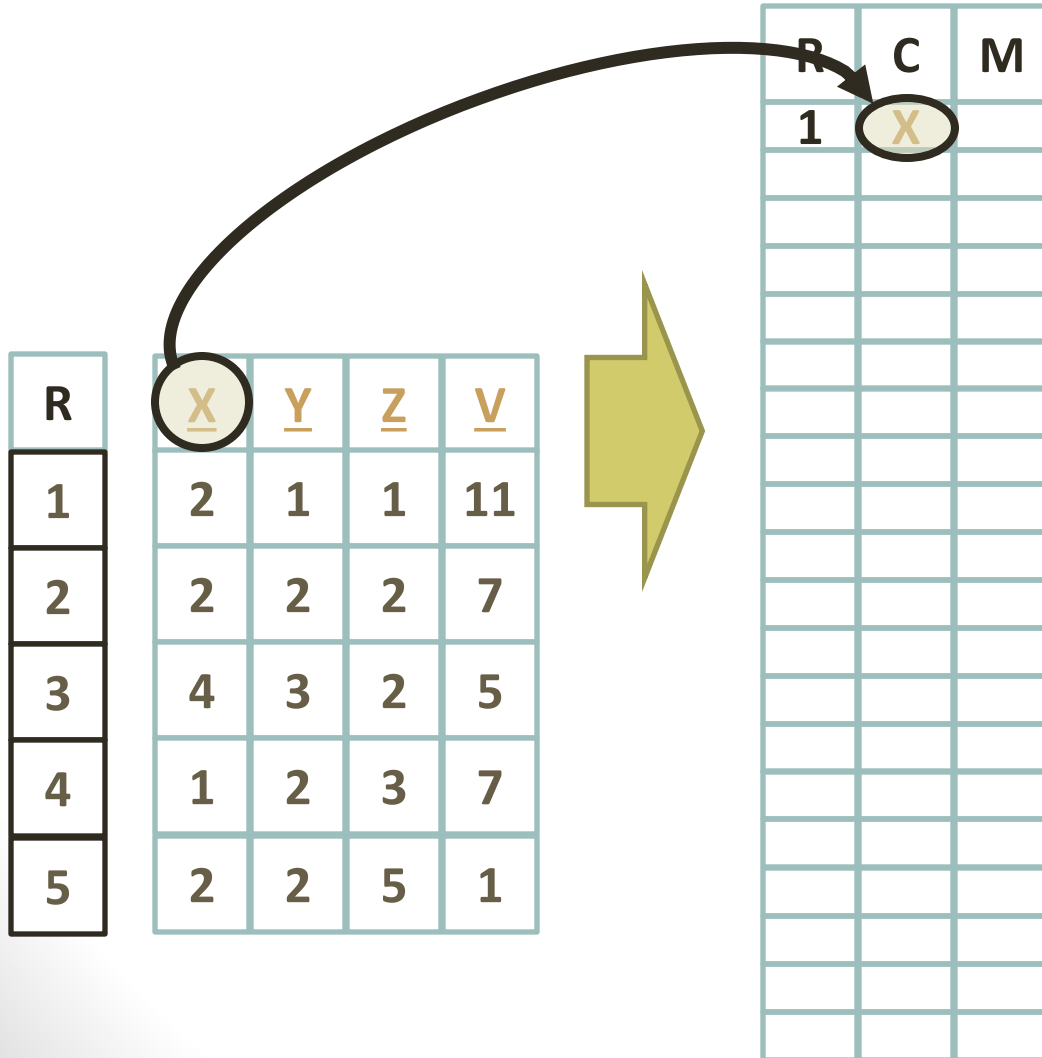
Sparse Matrices: EAV



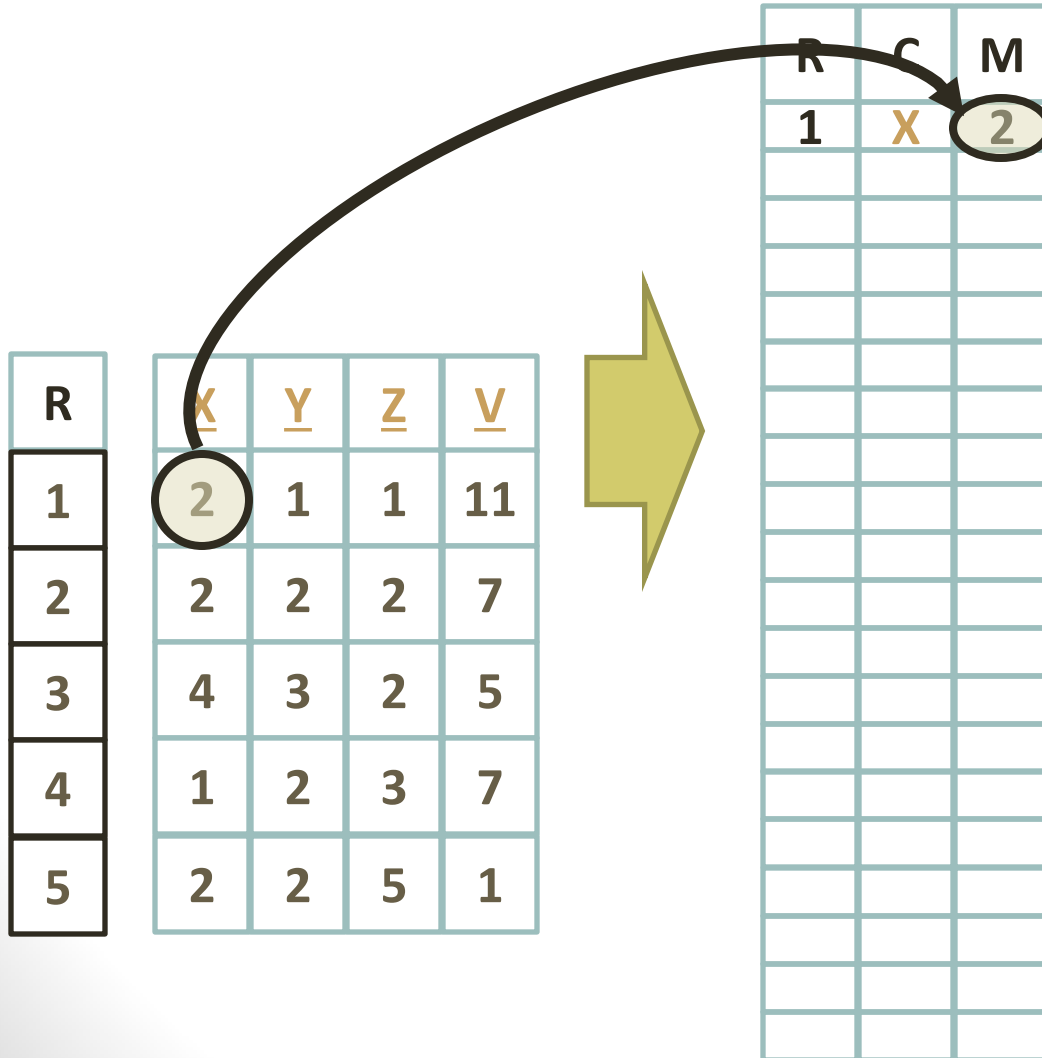
Sparse Matrices: EAV



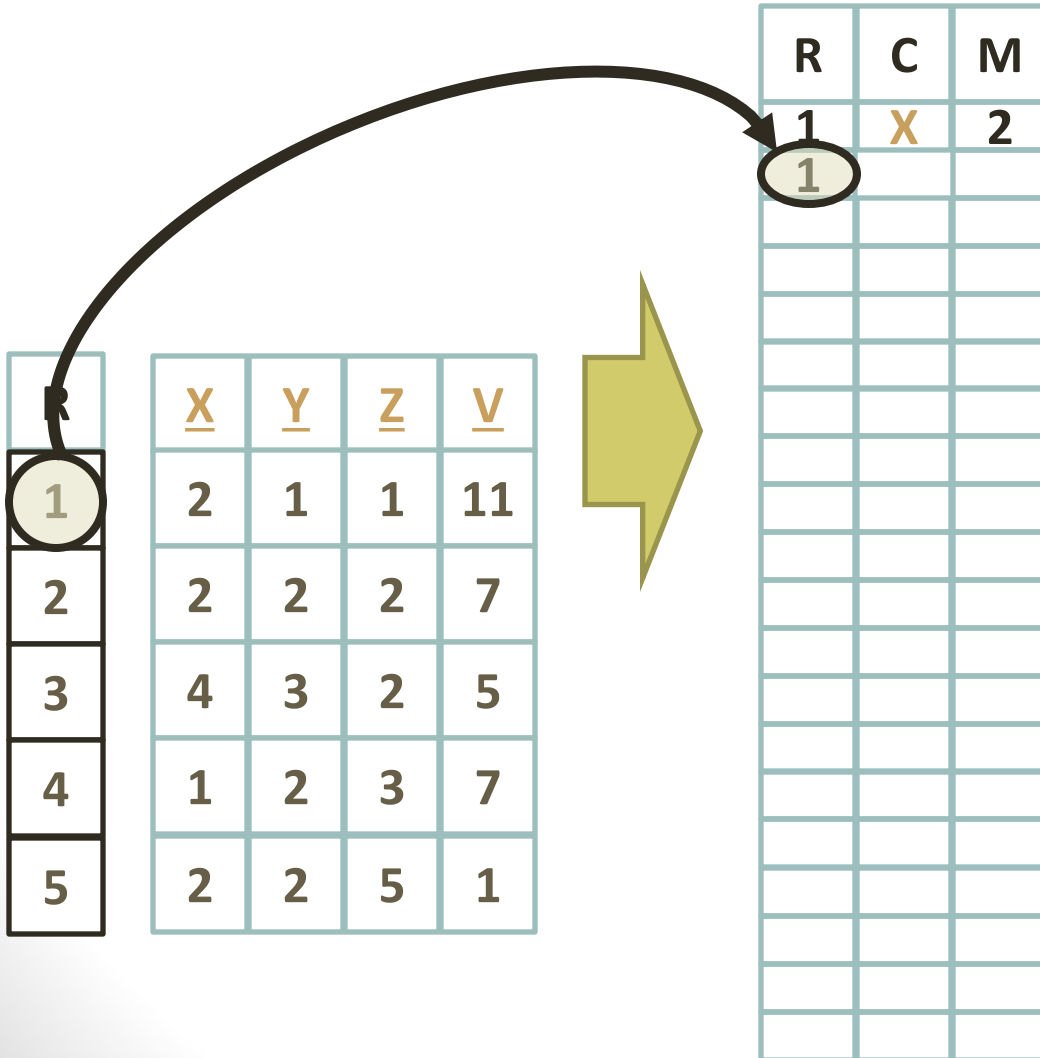
Sparse Matrices: EAV



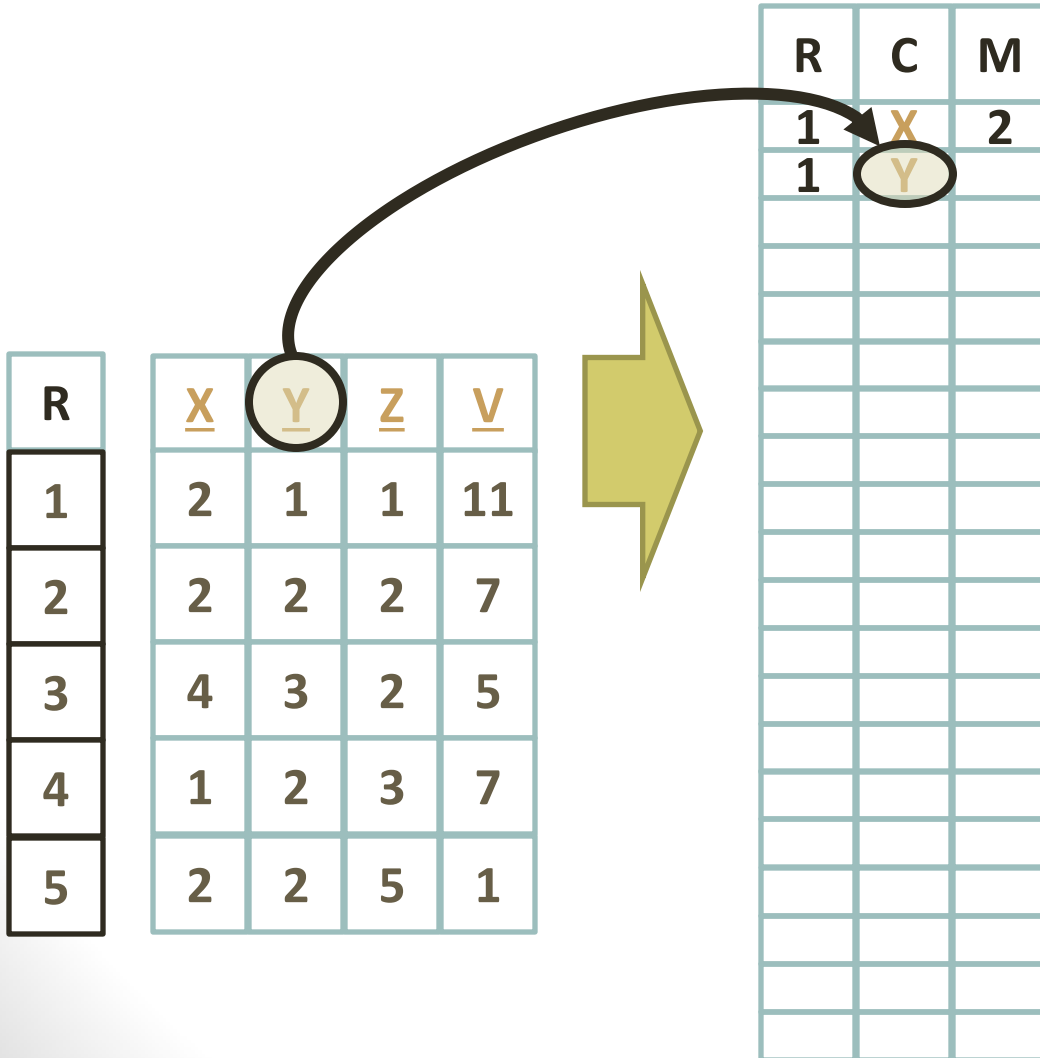
Sparse Matrices: EAV



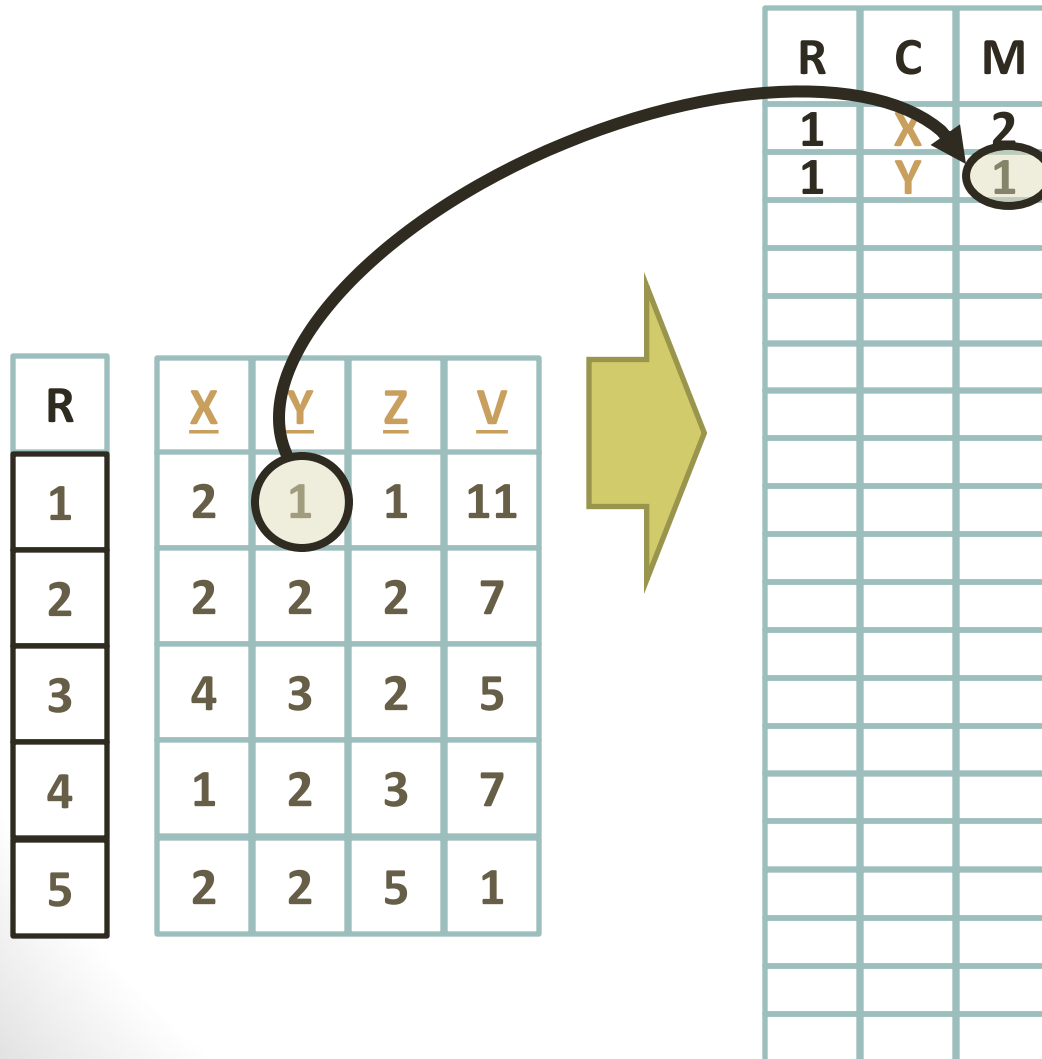
Sparse Matrices: EAV



Sparse Matrices: EAV



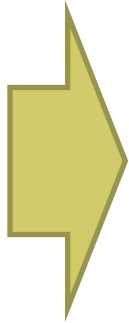
Sparse Matrices: EAV



Sparse Matrices: EAV

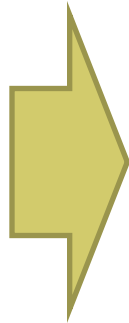
R
1
2
3
4
5

<u>X</u>	<u>Y</u>	<u>Z</u>	<u>V</u>
2	1	1	11
2	2	2	7
4	3	2	5
1	2	3	7
2	2	5	1

[illegible]

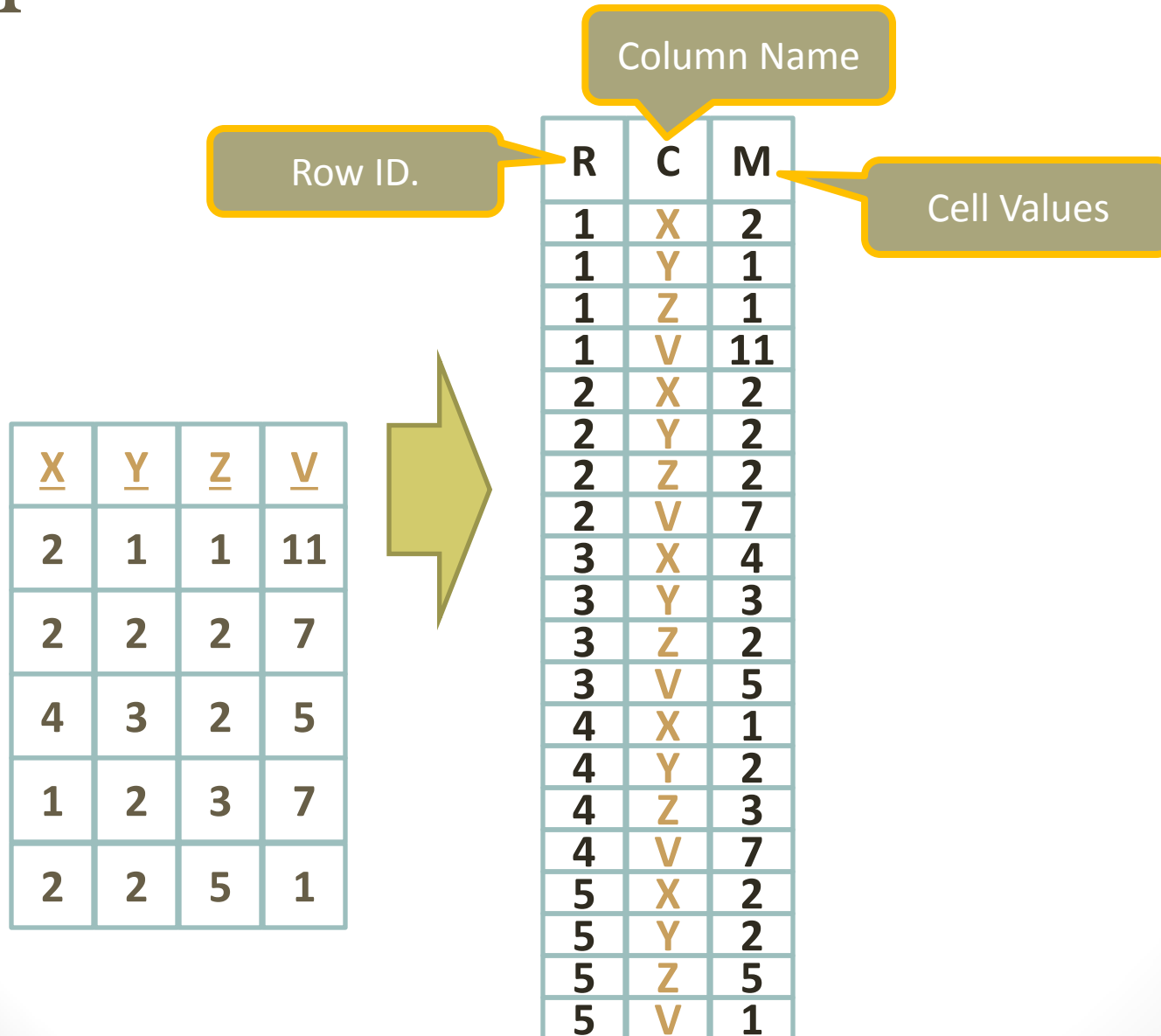
Sparse Matrices: EAV

<u>X</u>	<u>Y</u>	<u>Z</u>	<u>V</u>
2	1	1	11
2	2	2	7
4	3	2	5
1	2	3	7
2	2	5	1



R	C	M
1	X	2
1	Y	1
1	Z	1
1	V	11
2	X	2
2	Y	2
2	Z	2
2	V	7
3	X	4
3	Y	3
3	Z	2
3	V	5
4	X	1
4	Y	2
4	Z	3
4	V	7
5	X	2
5	Y	2
5	Z	5
5	V	1

Sparse Matrices: EAV



Data as Sparse Matrices

Assignment (1)

1. $\{a, b, c\}$ is a relation that contains the tuples a , b , and c . In the following cases the tuples have arity of 1. Calculate the following:
 - a. $(\{1, 2, 3\} \cup \{5, 7, 11\}) \cap \{2, 4, 6, 8, 10\}$
 - b. $(\{1, 2, 3\} \cap \{2, 4, 6, 8, 10\}) \cup (\{5, 7, 11\} \cap \{2, 4, 6, 8, 10\})$
2. A relation exists with 4 columns, named Column1, Column2, Column3, and Column4. Column1 is of type text. Column2, Column3, and Column4 are of type int:
 - a. Use relational algebra to fulfill the intent of the following SQL.
 - **SELECT Column1, Column3 FROM MyTable WHERE Column2 = Column3**
 - b. Reverse the order of projection and selection in your algebraic formulation from item 2a. What is the result of the new algebraic expression?
3. $\pi_{c1, c2}(\sigma_{\varphi1}(\sigma_{\varphi2}(\pi_{c1, c2, c3, c5}(R))))$
Where
 - $\varphi1: C1 = C5;$
 - $\varphi2: C5 = \text{"Test"};$
 - $R: \text{MyTable};$
 - a. Write a SQL statement that declares the intent of the algebraic notation
 - b. Simplify the algebraic statement. Simplification means minimize the number of parentheses and terms.

Assignment (2)

4. `SELECT * FROM T1 JOIN T2 ON T1.C1 = T2.C1`
 - a. Write out an equivalent in relational algebra using the join operator
 - b. Write out an equivalent in relational algebra without using the join operator
5. $\pi_{S.C1, R.C2}(\sigma_{\varphi1}(R) \bowtie_{\varphi2} S)$
where
 - $\varphi1 = (R.C2 = 'A')$
 - $\varphi2 = (R.C1 = S.C2)$
 - Write out equivalent SQL and test this SQL using relations R and S that you create for this example. The relations R and S in `RelationalAlgebraAndSQL.pdf` and `RelationalAlgebraAndSQL.sql` don't quite work because their column types do not match for this assignment.
6. Install and setup the Hadoop VM according to `SetupVirtualMachine.pdf`. Make sure that you get results similar to those shown in `SetupVirtualMachine.pdf`. If you completed this item last week, then you do not need to do it again. Otherwise, submit the requested screenshot to Canvas.

Assignment (3)

7. Submit answers to items 1 through 5 in a txt, doc/docx, or sql file. I will need to copy and paste the SQL statements. Submit the screen shot from item 6. Submission due date is Saturday 11:57 PM.
8. Discuss a topic in the class LinkedIn group. The topic should be related to something here:
 - The preview section in the class overview.
 - The new terminology at the end of this slide.
 - Read: Google file system:
<http://static.googleusercontent.com/media/research.google.com/en/us/archive/gfs-sosp2003.pdf>
 - Read MapReduce:
<http://static.googleusercontent.com/media/research.google.com/en/us/archive/mapreduce-osdi04.pdf>

New Terminology

- Hadoop
- Master Node
- Data Node
- Cluster
- Hive
- Impala
- MapReduce
- HDFS
- Doug Cutting
- Scalability
- AWS
- Elastic Cloud
- NoSQL
- CAP Theorem
- Consistency (CAP)
- Availability (CAP)
- Partition Tolerance (CAP)
- Eric Brewer
- RDBMS
- ACID
- Atomic (ACID)
- Consistent (ACID)
- Isolation (ACID)
- Durability (ACID)
- BASE
- Eventual Consistency
- Paxos
- Sqoop
- CouchDB
- Shared Data
- Stale Data
- Scale-out
- Scale-up
- Grace Hopper
- Data Replication
- Horizontal Partitioning
- Vertical Partitioning
- Heartbeats
- Multi-Version Concurrency Control
- EAV
- Relational Algebra
- Relational Calculus
- Relational Model
- Ted Codd
- Codd's Theorem
- Transaction Shell
- Column-oriented DBMS
- Row-oriented
- SPARQL

Introduction to Data Science