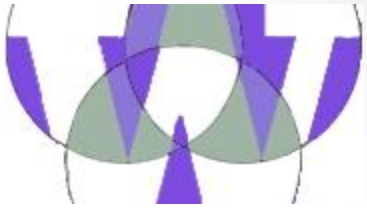
Introduction to Data Science

Lecture 7; November 16th, 2016

Ernst Henle
ErnstHe@UW.edu
Skype: ernst-henle

Agenda



- Announcements
  - The social component is a course requirement: This week's assignment includes two posts on the LinkedIn group.
  - Makeup first assignment for late admits
  - After hours with Marius on Nov 30th?
- Business Side of Data Science by Marius Marcu
- Break
- Review EAV and Sparse Matrices (Homework)
- Quiz 07a EAV and Sparse Matrices
- NoSQL Scale Out
- NoSQL vs. RDBMS
- Break
- NoSQL CAP
- Quiz 07b NoSQL Introduction
- Predictive Faux Pas
- Assignment. See assignment slides at the end of the deck. Complete all assignments items from all assignment slides. Submit as indicated on the assignment slide.

Data Science – The Business Point of View

Marius Marcu is a strategic innovator with a great passion for high-tech. His career path includes product management and product marketing roles with big companies like Intel and Microsoft, but also nimbler, high-growth startups like Smartsheet. .

While at Microsoft, Marius fell in love with big data, cloud technologies and data science, which he thinks will make a lot of people rich in the next 10 years, including himself. He is an alumnus of this certificate program and he is very passionate about leveraging data power to drive growth in the enterprise technology business. Data science professionals, like Marius, who come from a business background have unique opportunities. We are fortunate to get his analysis of the data science business landscape.

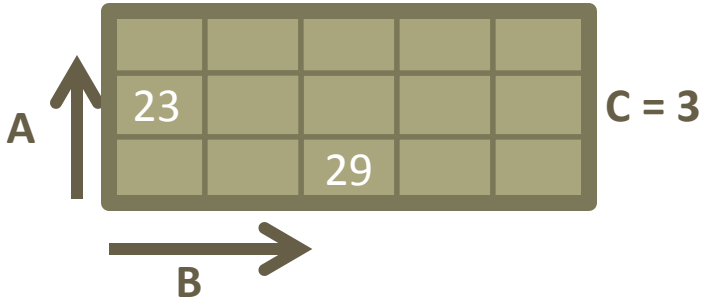
Marius' lecture slides are posted in Canvas

- https://www.linkedin.com/in/mariusmarcu
- mariusmarcu@global.t-bird.edu

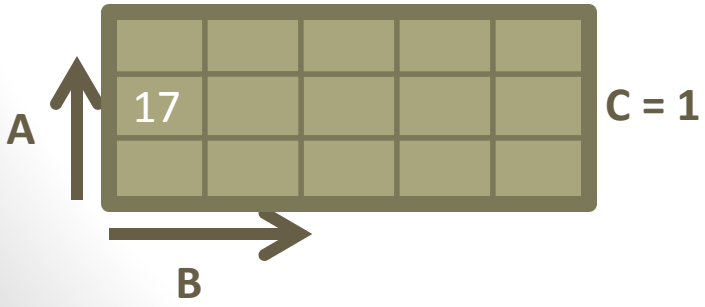
Break

Sparse Matrices and EAV Review

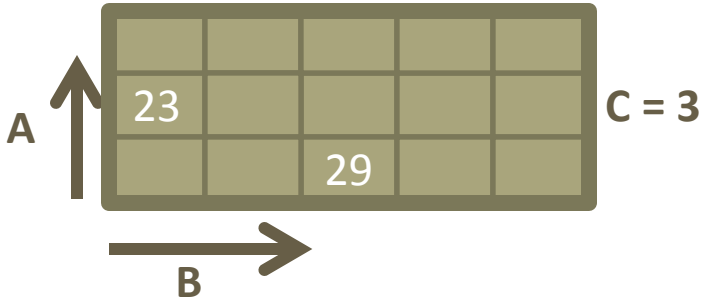
Sparse Matrices: Review (1)



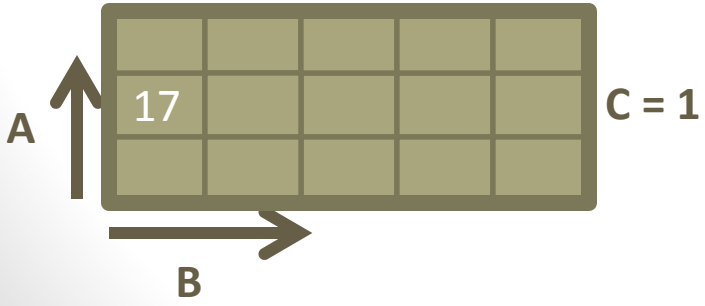
- Data: Real estate survey of single-family houses in downtown Seattle. Cell values are number (**N**) of houses found for sale.
  - **A**: Area in 1000's of square feet
  - **B**: Number of Bathrooms
  - **C**: Cost in $100,000.-
- Task: Create sparse matrices



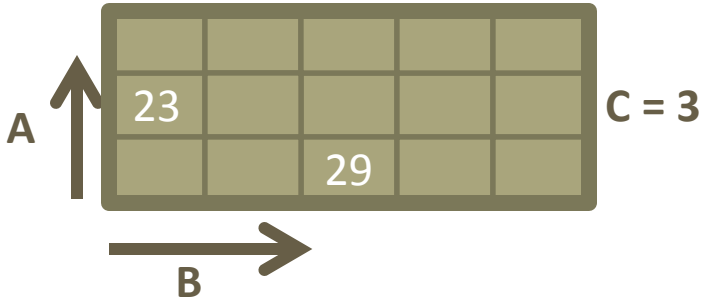
Sparse Matrices: Review (2)

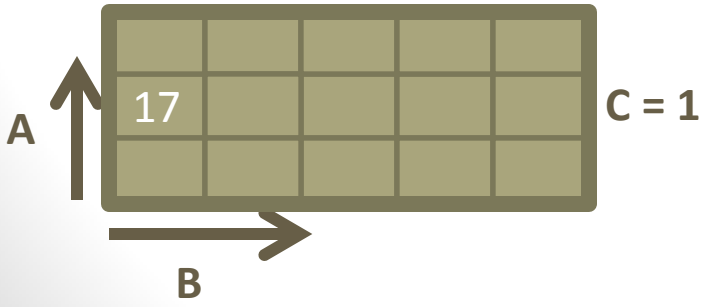


A	B	C	N
2	1	3	23
1	3	3	29
2	1	1	17



Sparse Matrices: Review (3)

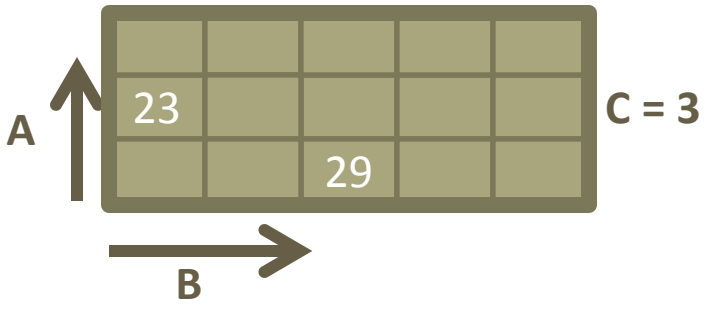


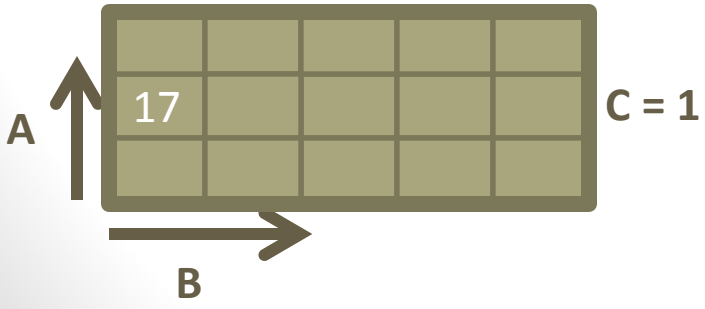


A	B	C	N
2	1	3	23
1	3	3	29
2	1	1	17

R	C	M
1	A	2
1	B	1
1	C	3
1	N	23
2	A	1
2	B	3
2	C	3
2	N	29
3	A	2
3	B	1
3	C	1
3	N	17

Sparse Matrices: Review (4)





A	B	C	N	CA
2	1	3	23	1.5
1	3	3	29	3
2	1	1	17	0.5

R	C	M
1	A	2
1	B	1
1	C	3
1	N	23
2	A	1
2	B	3
2	C	3
2	N	29
3	A	2
3	B	1
3	C	1
3	N	17
1	CA	1.5
2	CA	3
3	CA	0.5

Sparse Matrices: Review (5)

Matrix A

	1	2
1	2	-1
2		3
3		7

Multiply by 7

Matrix A

	1	2
1	14	-7
2		21
3		49

Sparse_Format _Table_A

R	C	V
1	1	2
1	2	-1
2	2	3
3	2	7

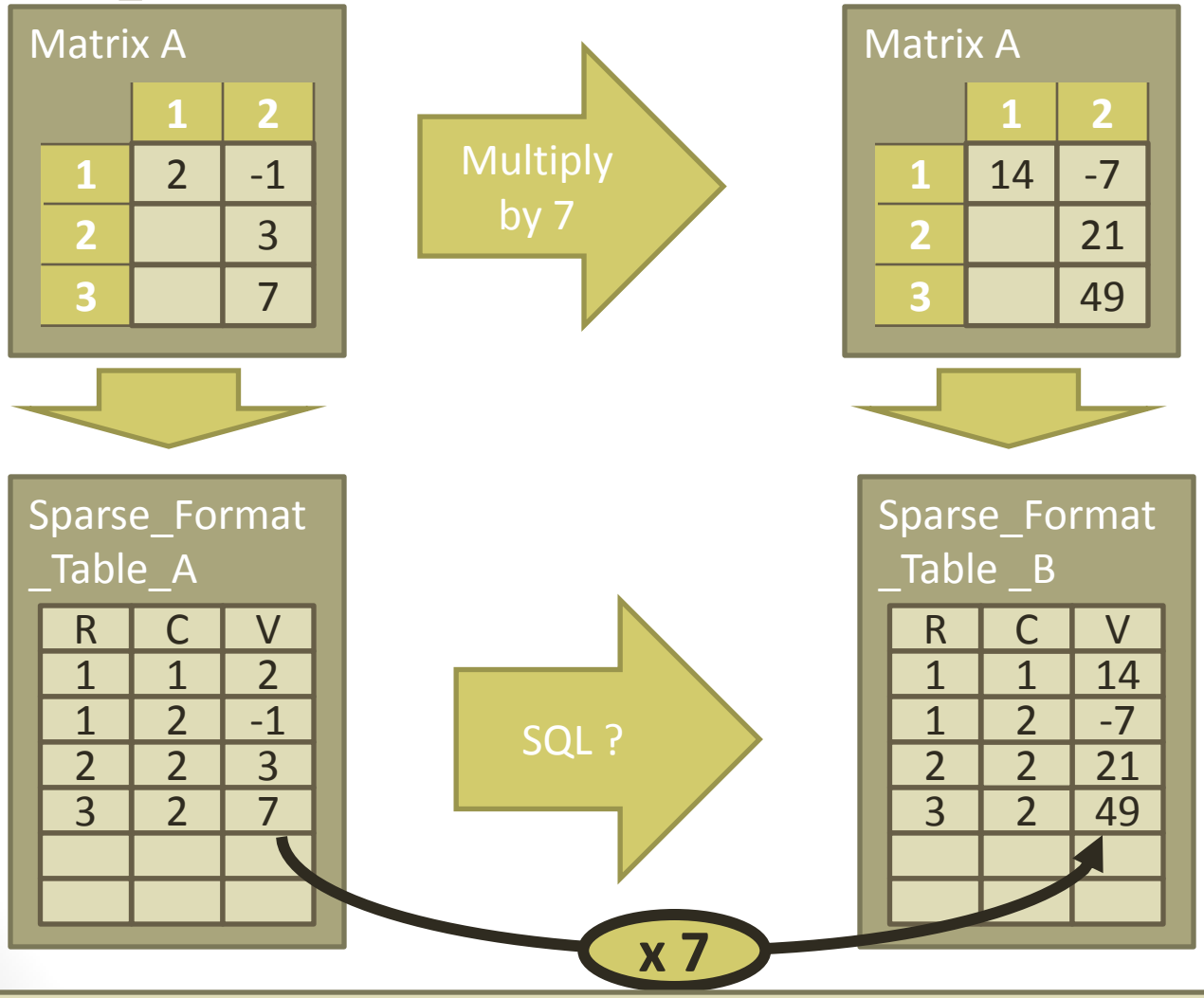
SQL ?

Sparse_Format _Table _B

R	C	V

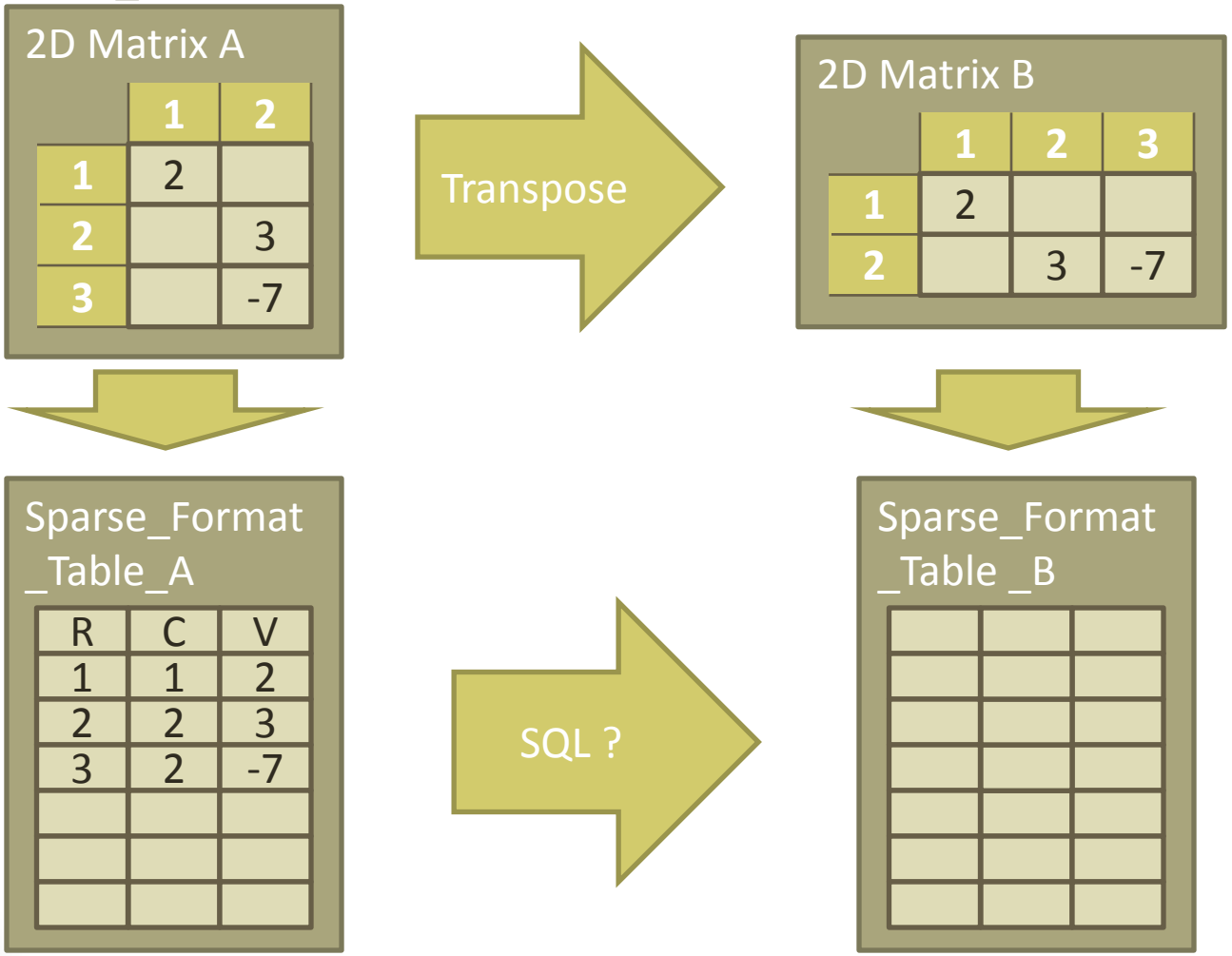
Sparse_Format_Table_A is the only table in the database. What SQL statement will present Sparse_Format_Table_B?

Sparse Matrices: Review (6)



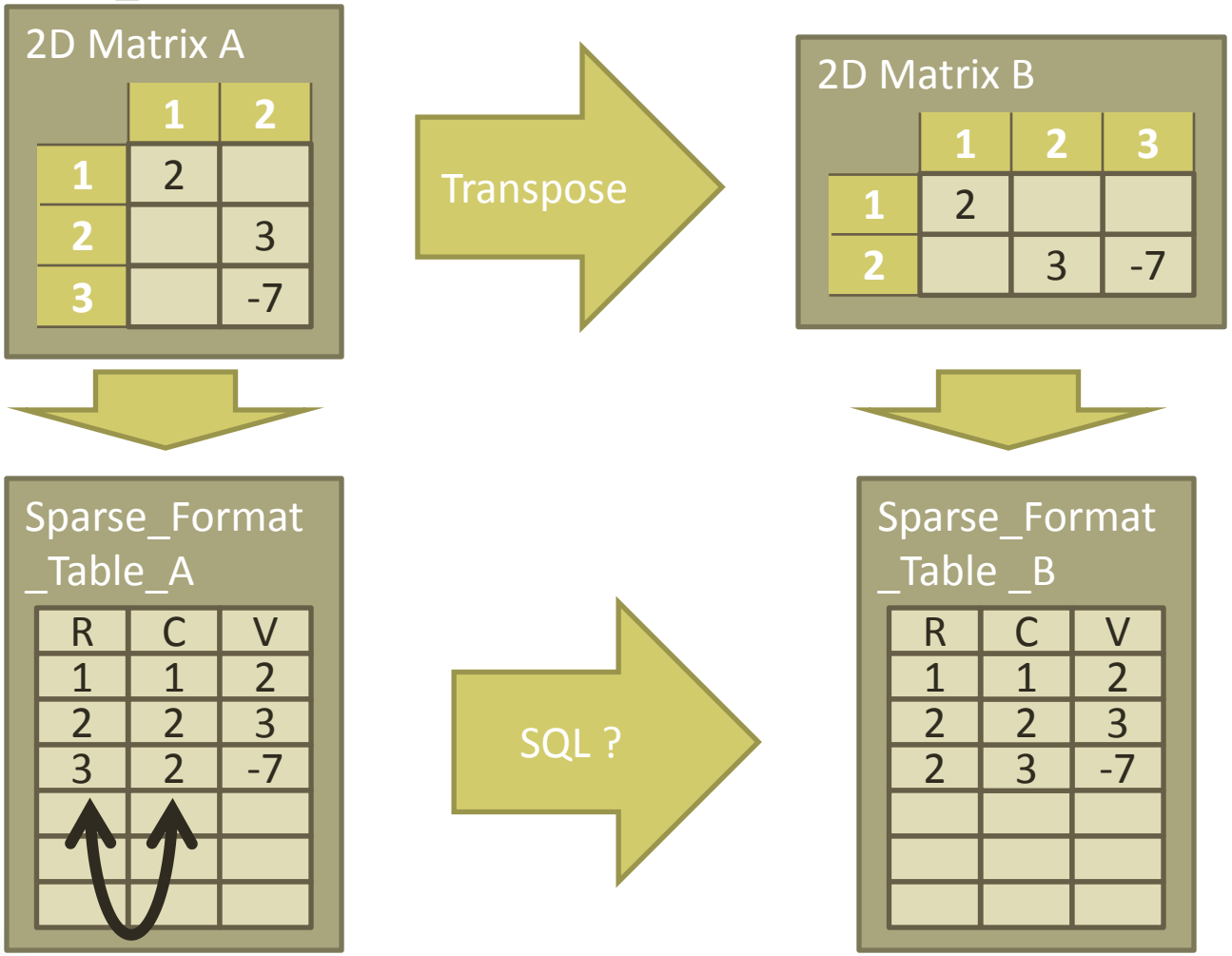
Sparse_Format_Table_A is the only table in the database. What SQL statement will present Sparse_Format_Table_B? Answer: **SELECT R, C, V*7 as V, FROM EAV_Table_A**

Sparse Matrices: Review (7)



Sparse_Format_Table_A is the only table in the database. What SQL statement will present Sparse_Format_Table_B?

Sparse Matrices: Review (8)



Sparse_Format_Table_A is the only table in the database. What SQL statement will present Sparse_Format_Table_B? Answer: **SELECT C as R, R as C, V, FROM EAV_Table_A**

Sparse Matrices: Review (9)

See: **MatrixAlgebraResults.sql**

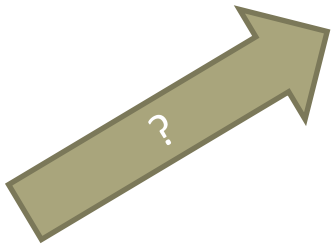
```
--Exercise 5;  Write SQL to multiply Matrix1 by 7
SELECT RowID, ColumnID, 7*Value AS Value FROM Matrix1

--Exercise 6;  Write SQL to transpose Matrix2
SELECT ColumnID AS RowID, RowID AS ColumnID, Value FROM
Matrix1

--Exercise 7;  Add two Matrices  (Add Matrix1 to Matrix3)
SELECT Matrix1.RowID AS RowID, Matrix1.ColumnID AS ColumnID,
(ISNULL(Matrix3.Value, 0) + ISNULL(Matrix1.Value, 0)) AS Value
   FROM Matrix1 FULL OUTER JOIN Matrix3 ON (Matrix1.ColumnID =
Matrix3.ColumnID) AND (Matrix1.RowID = Matrix3.RowID)
```

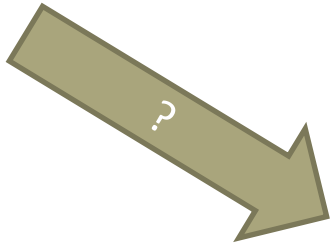
Sparse Matrices: Exercise (1)

- Exercise and Quiz Question

EAV Representation

R	C	M
1	X	1
1	Z	1
1	N	8
2	X	3
2	Z	1
2	N	6
3	X	1
3	Z	4
3	N	5





Matrix A

	Z=1	Z=4
X=1	8	5
X=3	6	

Matrix B

	Z=1	Z=3
X=1	8	5
X=4	6	

Matrix C

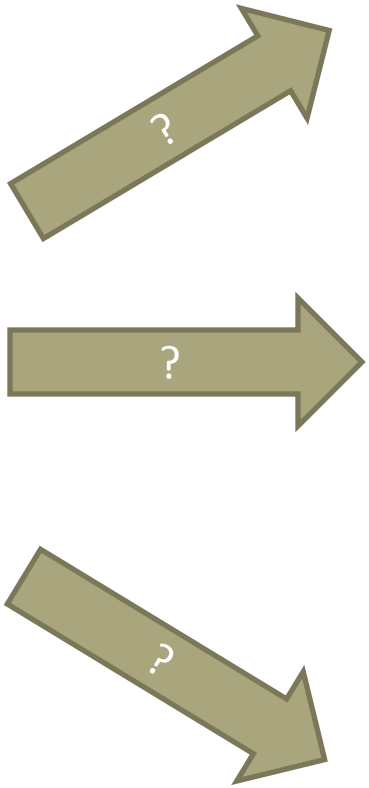
	Z=1	Z=3
X=1	8	5
X=3	6	

Sparse Matrices: Exercise (2)

- Exercise and Quiz Question

EAV Representation

R	C	M
1	X	1
1	Y	2
1	Z	1
1	N	9
2	X	3
2	Y	2
2	Z	1
2	N	5
3	X	1
3	Y	1
3	Z	4
3	N	7



Matrix A

Z=1	Y=1	Y=2
X=1	7	
X=2		
X=3		

Z=4	Y=1	Y=2
X=1		9
X=2		
X=3		5

Matrix B

Z=1	Y=1	Y=2
X=1	7	
X=2		
X=3		

Z=2	Y=1	Y=2
X=1		9
X=2		
X=3		5

Matrix C

Z=1	Y=1	Y=2
X=1		9
X=2		
X=3		5

Z=4	Y=1	Y=2
X=1	7	
X=2		
X=3		

Quiz EAV and Sparse Matrices

- You need to view the projected slide to answer questions 1 and 2.
- Tables in questions 6 and 7 are also on theses slides:
  - Sparse Matrices: Exercise (1)
  - Sparse Matrices: Exercise (2)

Sparse Matrices and EAV Review

NOSQL: Scale Out vs. Scale Up

No SQL: Scale Out

Scale-up vs. Scale-out

Before we discuss the nature of NOSQL, we should discuss the reasons for NOSQL.

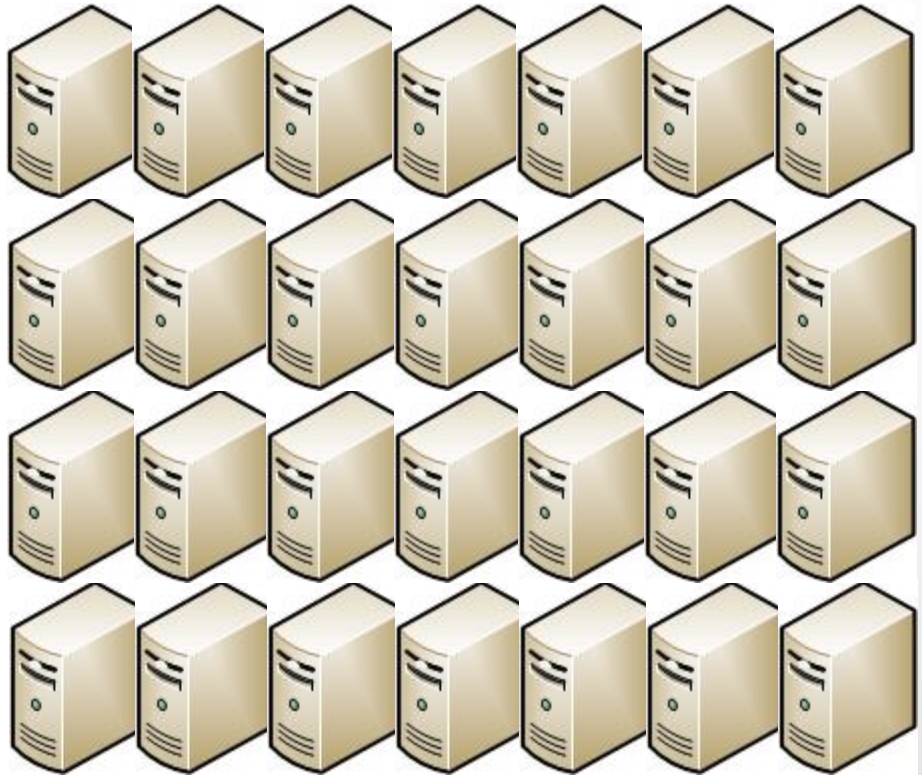
No SQL: Scale Out

Scale Up

No SQL: Scale Out

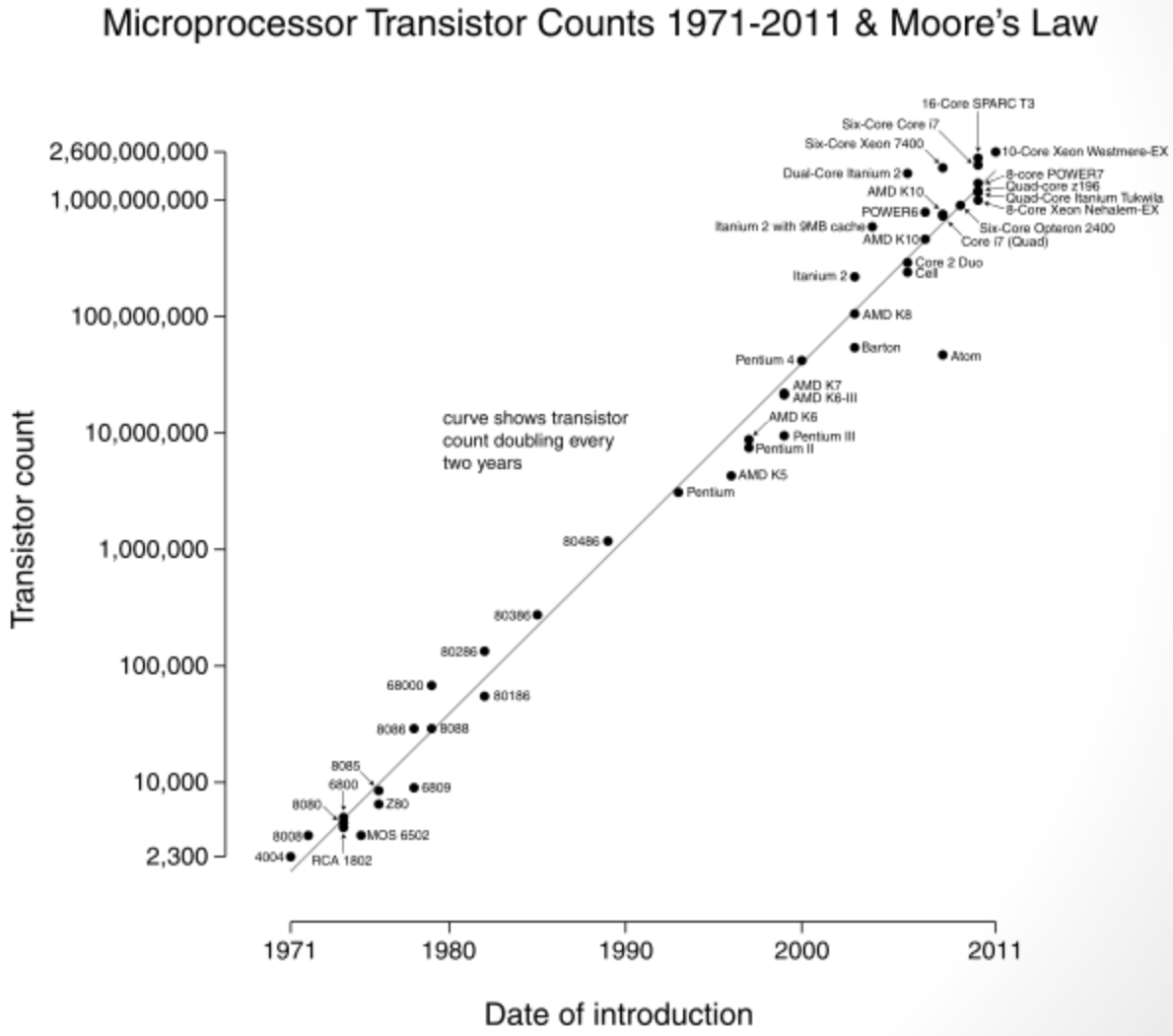
Scale Up

Scale Out



No SQL: Scale Out

- Scale-up
- Moore's Law



No SQL: Scale Out

Grace Hopper

No SQL: Scale Out

Grace Hopper

"In pioneer days they used oxen for heavy pulling, and when one ox couldn't budge a log, they didn't try to grow a larger ox. We shouldn't be trying for bigger computers, but for more systems of computers."

No SQL: Scale Out

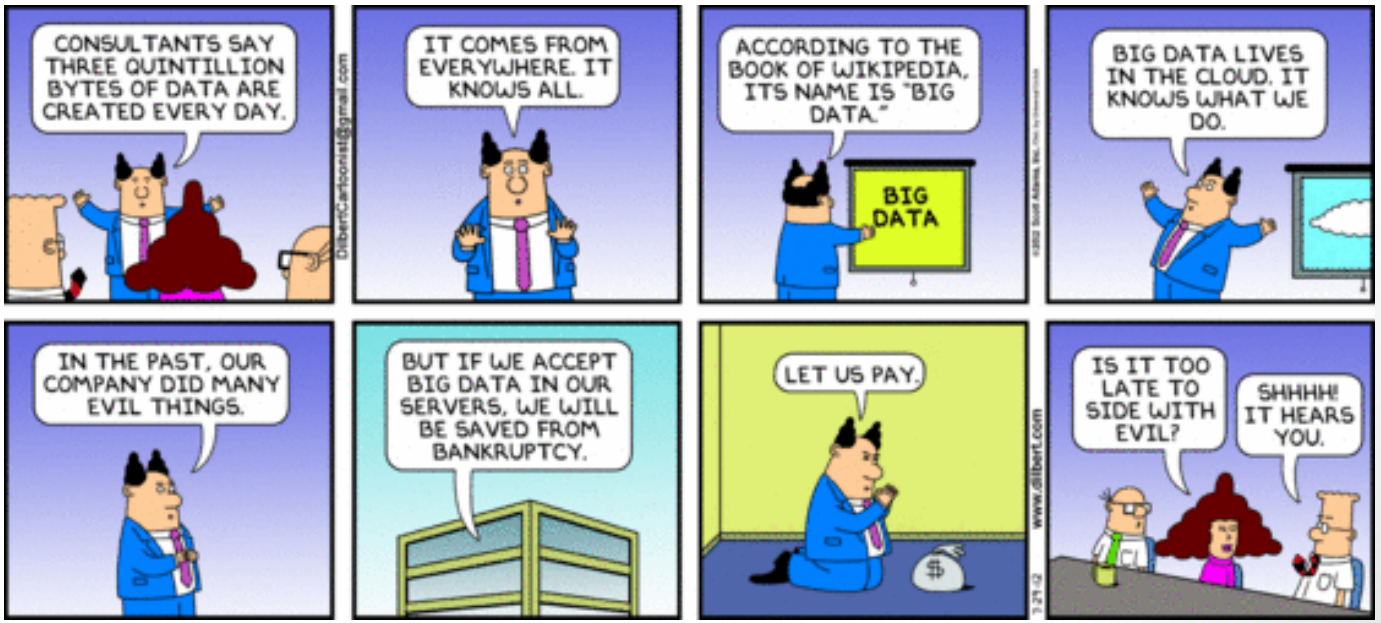
Cloud enables Scale-out

- The primary characteristic of NOSQL is scale out.
- From a practical level, scale out requires an adjustable number of commodity computers.
- Cluster Elasticity: http://en.wikipedia.org/wiki/Elasticity_%28data_store%29
- Virtual Machine
  - One computer "mimics" another computer. (A system platform supports execution of an operating system)
  - Allows hardware standardization.
  - Allows one server to "host" many computers.
  - Virtual machines in the cloud can be set up and taken down (dehydrated, reduced to an image).
- Cloud: What is the "cloud"? Remote access to a single point provides many online services like servers and storage. (http://en.wikipedia.org/wiki/Cloud_computing).

No SQL: Scale Out

Cloud: Services

Amazon Web Services, GoGrid, Google Compute Engine, Microsoft Azure, Rackspace, SoftLayer



No SQL: Scale Out

Scale-out and the "Cloud"

- **Elasticity** has made cloud computing feasible
- Clouds generally employ **virtual machines** that can be created at a moments notice, reduced to an image (dehydrated), re-started from an image, and deleted (recycled).
- How do we partition storage or usage among an unknown number of machines? Often we do not know ahead of time if new machines will become available or which machines will be recycled.
- Storage and usage are mapped to machines by a hash table. In traditional hash tables a change in the number of slots requires most keys to be remapped.
- We need a strategy to minimize remapping of storage and usage among the available computers: Consistent Hashing: http://en.wikipedia.org/wiki/Consistent_hashing

No SQL: Scale Out

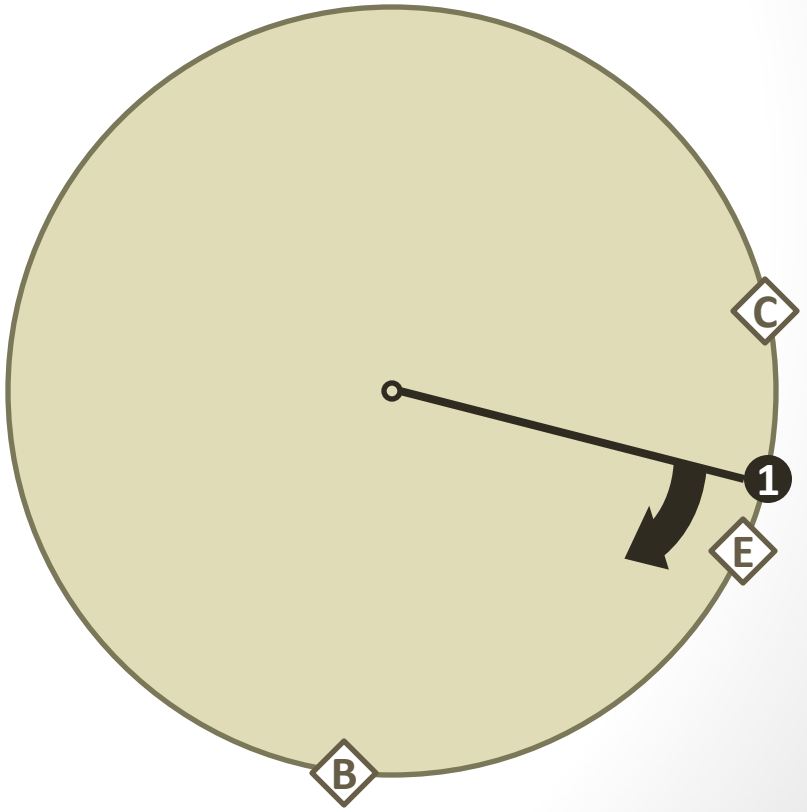
Consistent Hashing

- Consider a hash map where each object is mapped to a point on the circumference of a circle. For instance an object is mapped to the number of minutes on a clock.
- Computers, Files, Processes, etc., are mapped in this manner on the same circle.
- A computer “claims” all files and processes who have a hash that is clock wise to that computer.

No SQL: Scale Out

Consistent Hashing

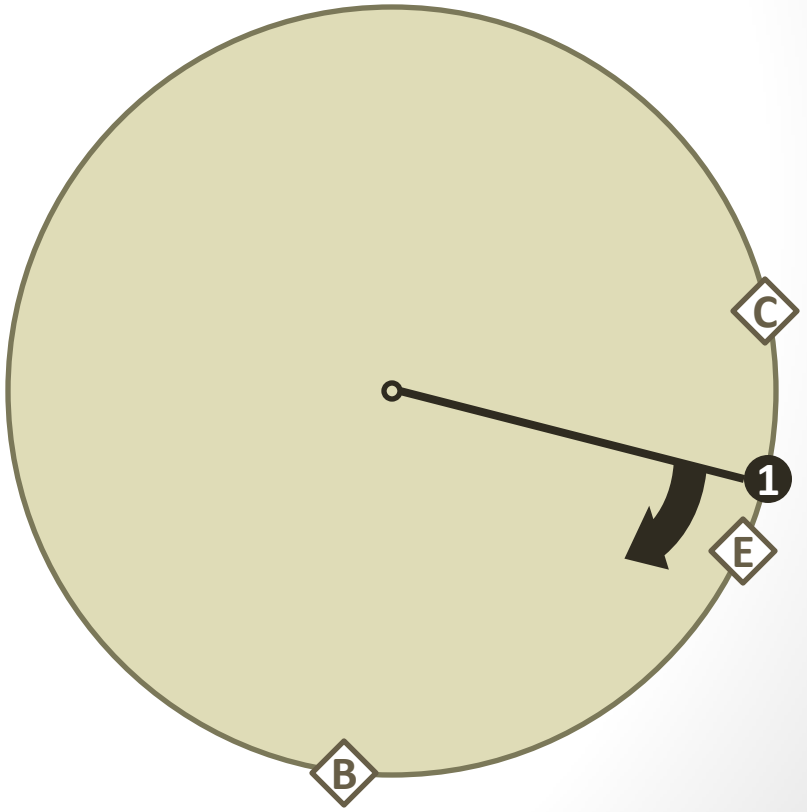
Symbol	Object Type	Hash	Relation
B	Data Object	32	1
C	Data Object	14	1
E	Data Object	18	1
1	Machine 1	17	E C B



No SQL: Scale Out

Consistent Hashing

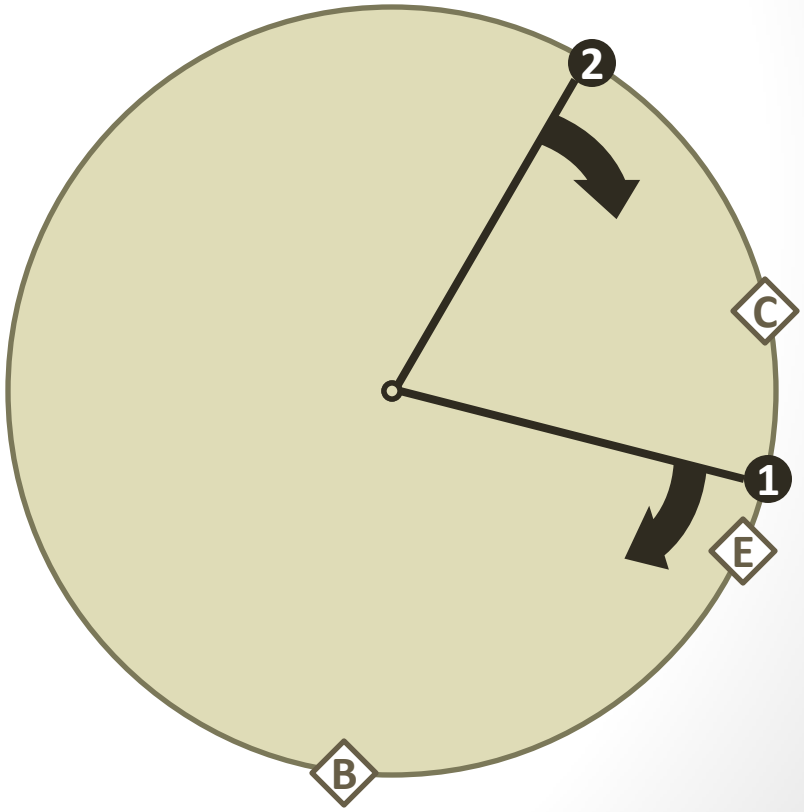
Symbol	Object Type	Hash	Relation
B	Data Object	32	1
C	Data Object	14	1
E	Data Object	18	1
1	Machine 1	17	E C B



No SQL: Scale Out

Consistent Hashing

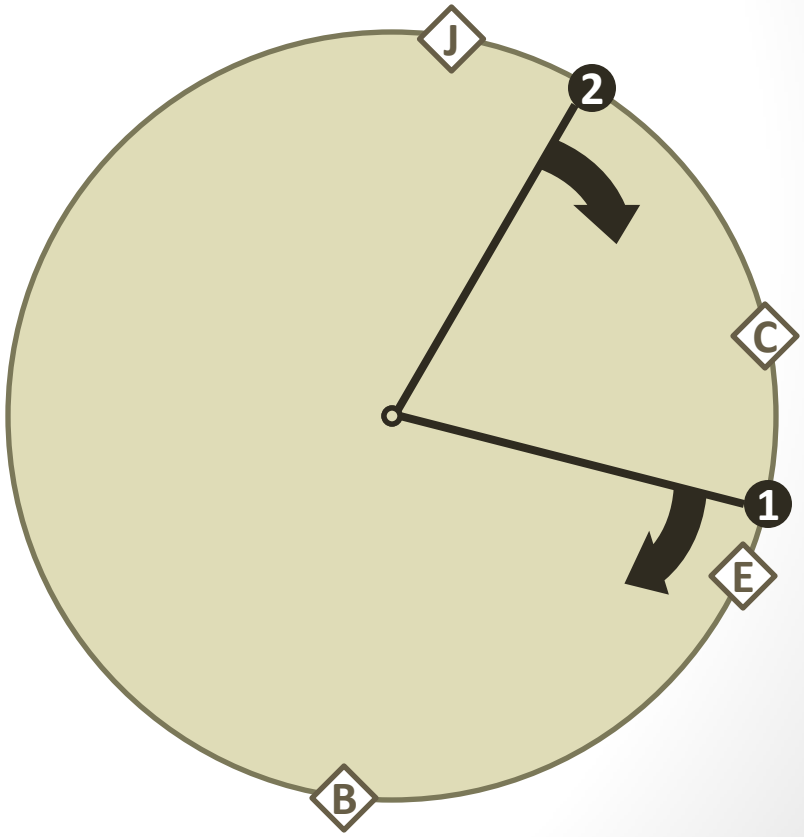
Symbol	Object Type	Hash	Relation
B	Data Object	32	1
C	Data Object	14	2
E	Data Object	18	1
1	Machine 1	17	E B
2	Machine 2	5	C



No SQL: Scale Out

Consistent Hashing

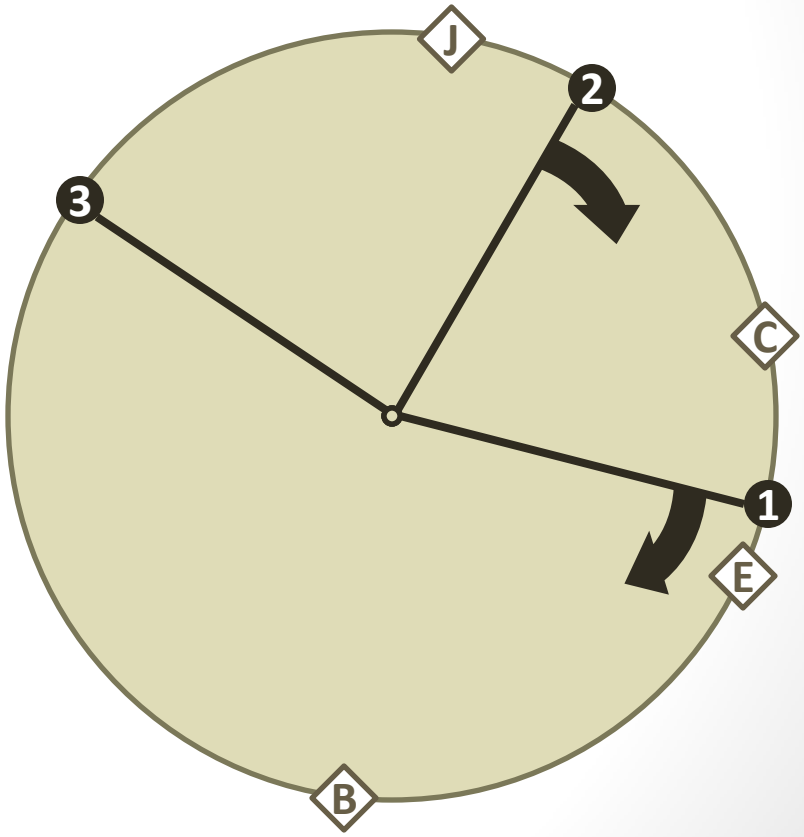
Symbol	Object Type	Hash	Relation
B	Data Object	32	1
C	Data Object	14	2
E	Data Object	18	1
J	Data Object	2	1
1	Machine 1	17	E J B
2	Machine 2	5	C



No SQL: Scale Out

Consistent Hashing

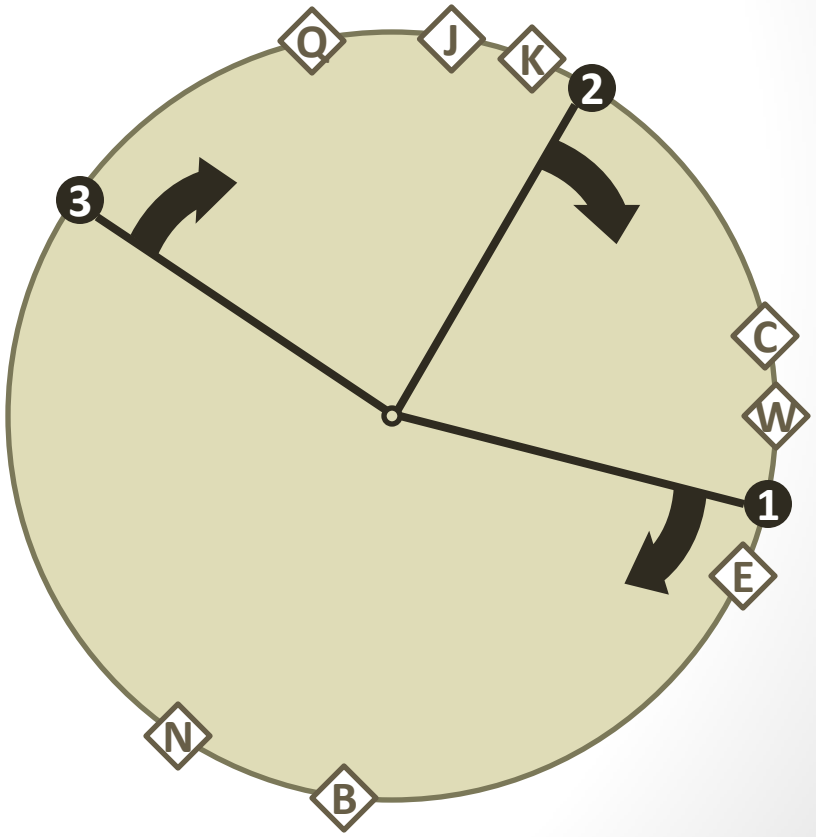
Symbol	Object Type	Hash	Relation
B	Data Object	32	1
C	Data Object	14	2
E	Data Object	18	1
J	Data Object	2	3
1	Machine 1	17	E B
2	Machine 2	5	C
3	Machine 3	51	J



No SQL: Scale Out

Consistent Hashing

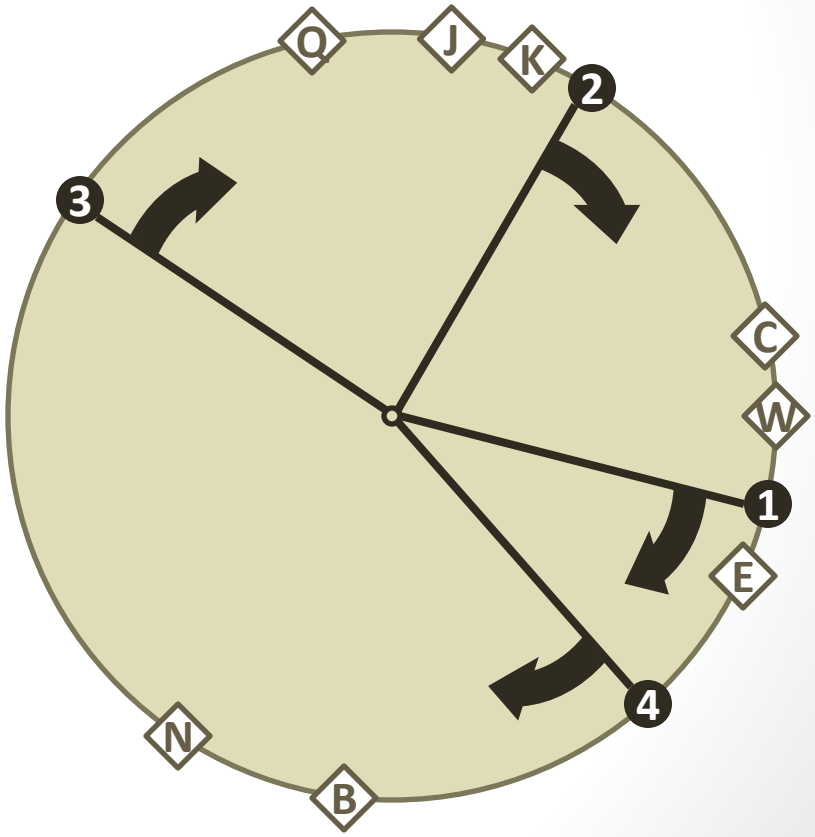
Symbol	Object Type	Hash	Relation
B	Data Object	32	1
C	Data Object	14	2
E	Data Object	18	1
J	Data Object	2	3
K	Data Object	4	3
N	Data Object	35	1
Q	Data Object	57	3
W	Data Object	15	2
1	Machine 1	17	E B N
2	Machine 2	5	C W
3	Machine 3	51	J K Q



No SQL: Scale Out

Consistent Hashing

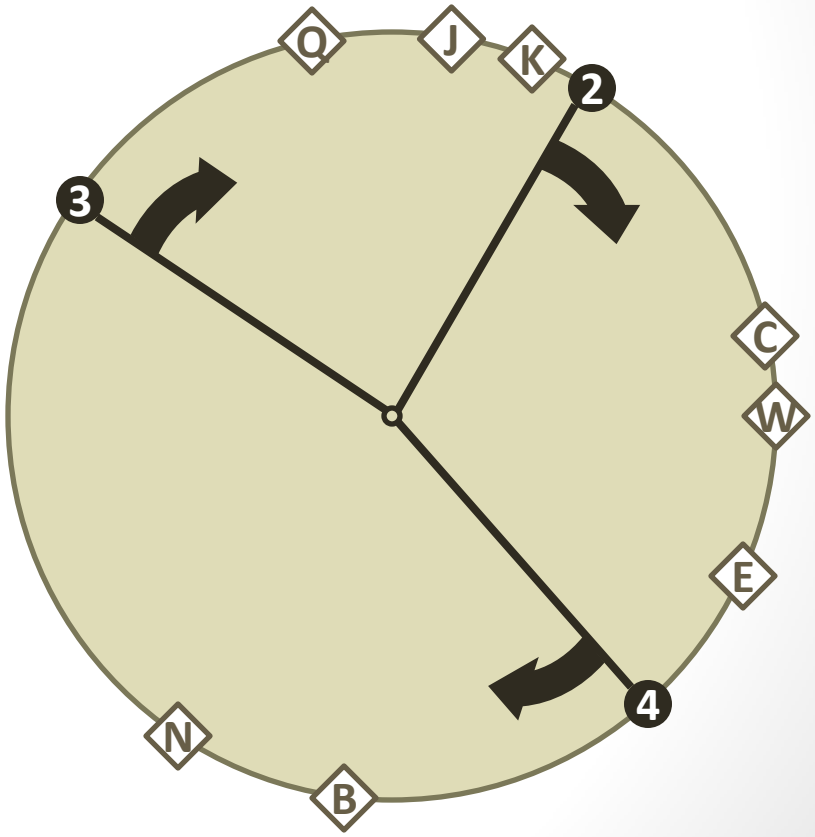
Symbol	Object Type	Hash	Relation
B	Data Object	32	4
C	Data Object	14	2
E	Data Object	18	1
J	Data Object	2	3
K	Data Object	4	3
N	Data Object	35	4
Q	Data Object	57	3
W	Data Object	15	2
1	Machine 1	17	E
2	Machine 2	5	C W
3	Machine 3	51	J K Q
4	Machine 4	23	B N



No SQL: Scale Out

Consistent Hashing

Symbol	Object Type	Hash	Relation
B	Data Object	32	4
C	Data Object	14	2
E	Data Object	18	2
J	Data Object	2	3
K	Data Object	4	3
N	Data Object	35	4
Q	Data Object	57	3
W	Data Object	15	2
2	Machine 2	5	C W E
3	Machine 3	51	J K Q
4	Machine 4	23	B N



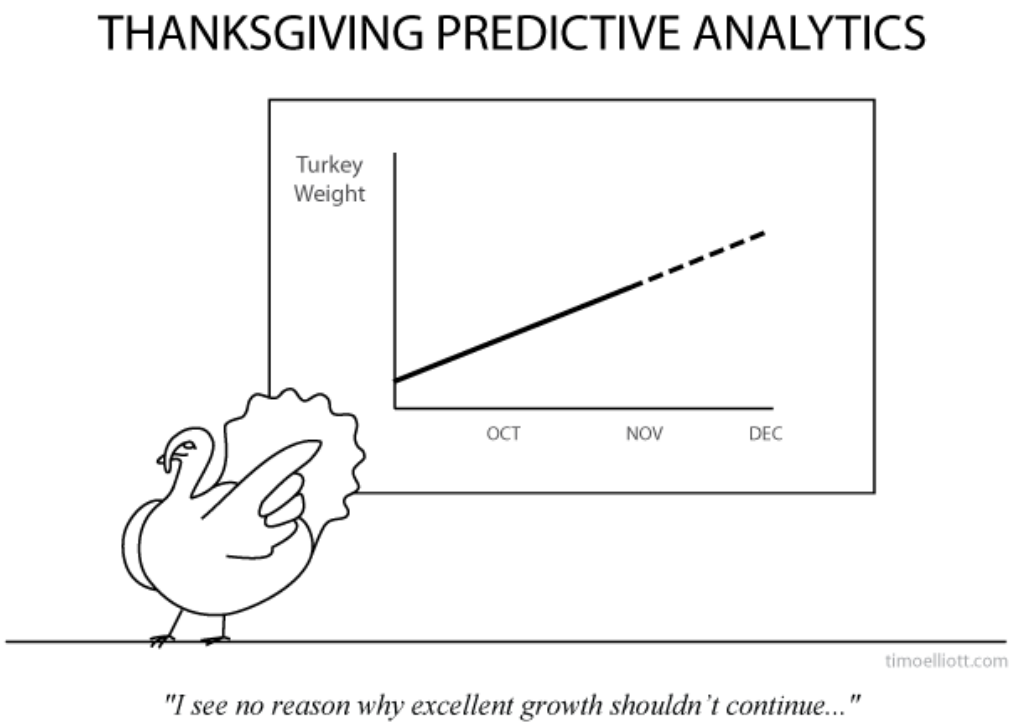
No SQL: Scale Out

What does Scale-Out have to do with NOSQL?

- Traditional Relational Database Management Systems (RDBMS) have problems with scale-out.
- Therefore, new data base management schemes were desired.

NOSQL: Scale-out

Break



"I see no reason why excellent growth shouldn't continue..."

NOSQL: Defined in contrast to RDBMS

NOSQL *vs.* RDBMS (1)

- NOSQL
  - **NO**SQL may stand for: **NO**T-SQL, **N**ot-**O**nly-SQL, K**NO**W-SQL
  - There is no consensus definition of NOSQL. NOSQL is a misnomer. NOSQL has little to do with SQL or an alternative to SQL. NOSQL has more to do with new database strategies and data structures.

NOSQL *vs.* RDBMS (2)

- NOSQL
  - **NO**SQL may stand for: **NO**T-SQL, **N**ot-**O**nly-SQL, K**NO**W-SQL
  - There is no consensus definition of NOSQL. NOSQL is a misnomer. NOSQL has little to do with SQL or an alternative to SQL. NOSQL has more to do with new database strategies and data structures.

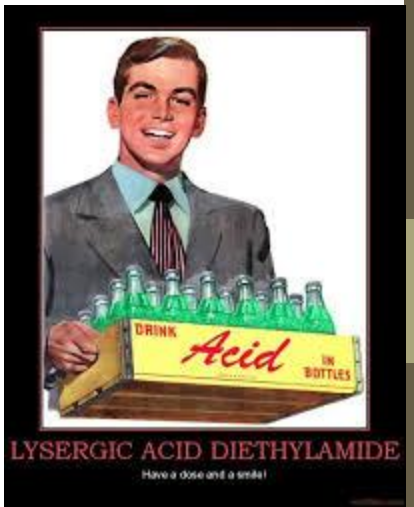
- RDBMS vs. NOSQL
  - NOSQL has to do with databases that do not follow the pattern of a relational database management system (RDBMS)
  - Therefore we need to define NOSQL in contrast to RDBMS. The hallmark of RDBMS is the relational model and **ACID**.

- Quick and Simple Overview of NOSQL (watch at home): http://www.youtube.com/watch?v=sh1YACOK_bo

NOSQL *vs.* RDBMS (3)

ACID

- ACID and the relational model are the hallmarks of RDBMS. ACID stands for:
  - **Atomic**
  - **Consistent**
  - **Isolation**
  - **Durability**



NOSQL *vs.* RDBMS (4)

ACID: Atomic

- Atomic is Greek for unsplitable
- All or nothing
- All the changes of a transaction will happen or none of them will happen.
- Aborted transactions are rolled back.

NOSQL *vs.* RDBMS (5)

ACID: Consistent

- Database is consistent before transaction. Database is consistent after transaction.
- Database will adhere to all the consistency rules before and after every transaction.
- Database constraints and column relations to other data are maintained. In other words, data written to the database must abide by integrity constraints. For Example:
  - A column which requires a unique identifier will not tolerate a duplicate value.
  - A column that requires no NULL values will not accept a NULL value.
  - The database will verify that each value is a valid foreign key in a column that demands that each value is a valid foreign key.

Beware! The word consistency is overloaded. The C in ACID is for a different consistency than the well-known consistency problem in a NOSQL distributed system. When we talk about consistency in NOSQL we mean that replicated data are the same after updates.

NOSQL *vs.* RDBMS (6)

ACID: Isolation

- Transactions are isolated from one another.
- During a transaction, other processes cannot see the affected parts of the database until the transaction has completed. The other processes have to wait. The result is as if the transactions occurred in sequentially
- Isolation is achieved by concurrency control. When two transactions execute at the same time, each attempting to modify the same data, one of the two must wait until the other completes.

NOSQL *vs.* RDBMS (7)

ACID: Durability

- What is written is readable until explicitly deleted
- Data doesn't evaporate

Durability is the hallmark of databases in general. NOSQL and RDBMS both succeed equally well in durability.

NOSQL *vs.* RDBMS (8)

- The atomic, consistent, and isolated aspects of an RDBMS are the basis of what is called a transaction shell or bubble.
- Durability is just as important in NOSQL as it is in an RDBMS
- Base
  - Basic Availability: Basic Availability means that the system is available most of the time. (Availability means that a database request receives a response about success or failure.)
  - Soft-state
  - Eventual consistency

Beware! The word consistency is overloaded. The C in ACID is for a different consistency than the well-known consistency problem in a NOSQL distributed system. When we talk about consistency in NOSQL we mean that replicated data are the same after updates.

NOSQL *vs.* RDBMS (9)

NOSQL

- NOSQL databases are distributed databases that split up data into manageable blocks and replicate data to prevent data loss
- NOSQL databases allow scale-out using many cheap servers and, typically, do not fully use scaled-up servers
- NOSQL databases may have a relaxed schema and can dynamically add new attributes to records
- NOSQL databases have a relaxed transaction shell and do not abide by ACID
- NOSQL databases do not need to be immediately consistent after every transaction. They can be eventually Consistent.

Beware! The word consistency is overloaded. The C in ACID is for a different consistency than the well-known consistency problem in a NOSQL distributed system. When we talk about consistency in NOSQL we mean that replicated data are the same after updates.

NOSQL: Defined in contrast to RDBMS

Break

NOSQL: CAP Theorem

CAP Theorem (1)

CAP Theorem (2)

Distributed system with Shared Data: Vasanti Bhat-Nayak and Grace Hopper need a package from R to do a naïve Bayes classification. If there were only one server that contained this package, then consistency would be easy. But, availability would be restricted. When multiple R users want to download a package, the server gets clogged. Therefore, the cran packages are replicated on multiple servers around the world. When a package needs to be updated, then the master node asks all servers to update simultaneously. So when Vasanti and Grace download a package from different servers they will get the same version of the Naive Bayes package.

CAP Theorem (3)

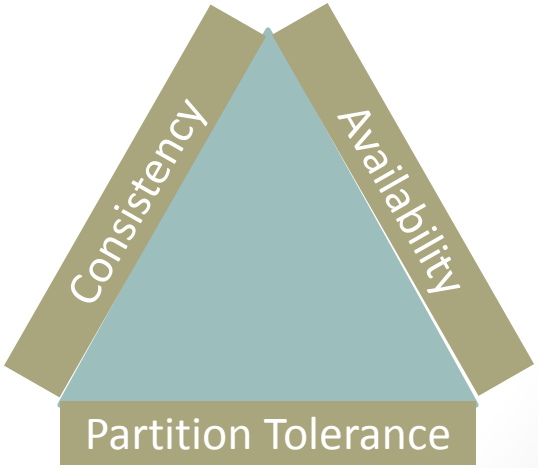
Distributed system with Shared Data: Vasanti Bhat-Nayak and Grace Hopper need a package from R to do a naïve Bayes classification. If there were only one server that contained this package, then consistency would be easy. But, availability would be restricted. When multiple R users want to download a package, the server gets clogged. Therefore, the cran packages are replicated on multiple servers around the world. When a package needs to be updated, then the master node asks all servers to update simultaneously. So when Vasanti and Grace download a package from different servers they will get the same version of the Naive Bayes package.

Partition of the Distributed System: But, what happens if on that day the Andorran server that Vasanti uses, can't be updated because of a communication error. The database has two choices: (1) It can wait until the Andorran server is fixed and then do the update. (2) Or, it updates all the other servers that allow the update.

In the first case we forgo availability and nobody has access to the most recent Naive Bayes package. In the second case Vasanti and Grace will have different results because the packages are different.

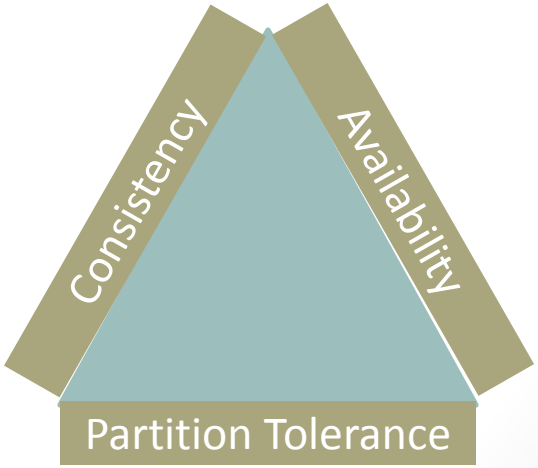
CAP Theorem (4)

- CAP stands for:
  - **C**onsistency
  - **A**vailability
  - **P**artition Tolerance



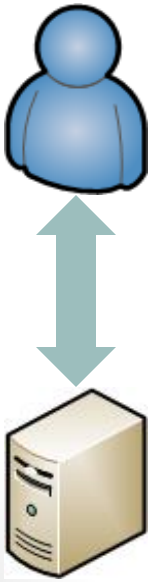
CAP Theorem (5)

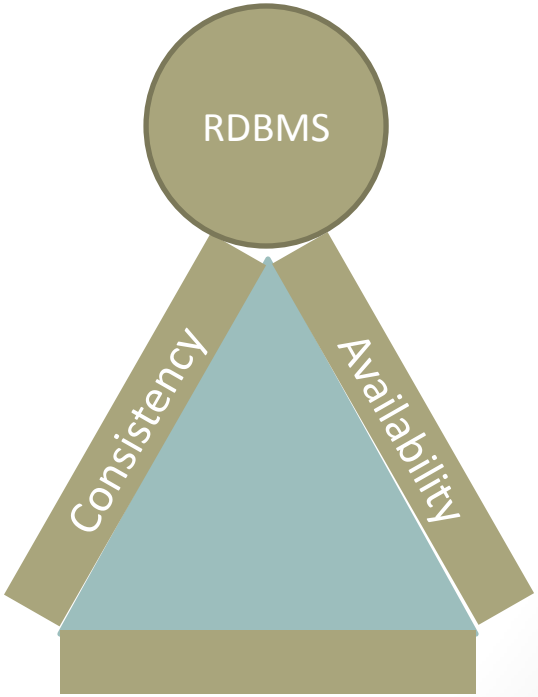
- CAP stands for:
  - **C**onsistency: All nodes see the same data at the same time
  - **A**vailability: Nodes are available for updates and reads
  - **P**artition Tolerance: Arbitrary message loss or partial failure does not bring down the system



CAP Theorem (6)

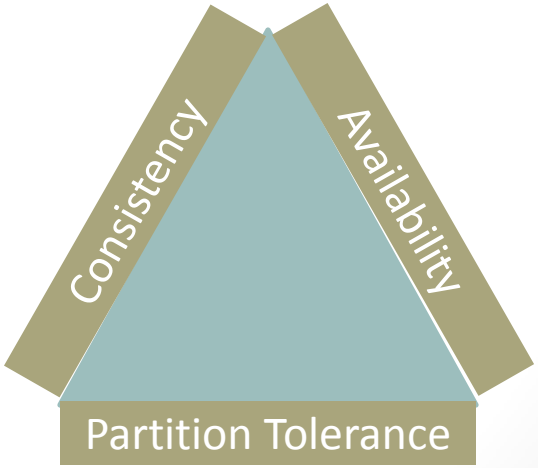
- Assume a single node with one set of data.
- This simple system resembles a typical RDBMS.
- Partition tolerance is irrelevant, because we only have one node.





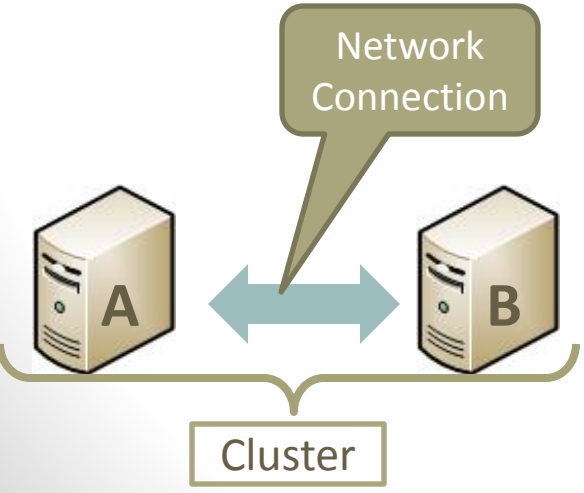
CAP Theorem (7)

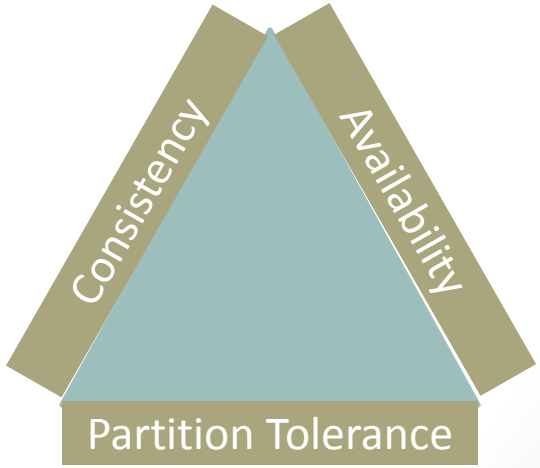
- The CAP theorem was formulated by Eric Brewer http://en.wikipedia/wiki/CAP_theorem
- Two formulations of the CAP theorem:
  - You can have at most two of the CAP properties for any shared data system.
  - During a network partition, a distributed system must choose either Consistency or Availability.



CAP Theorem (8)

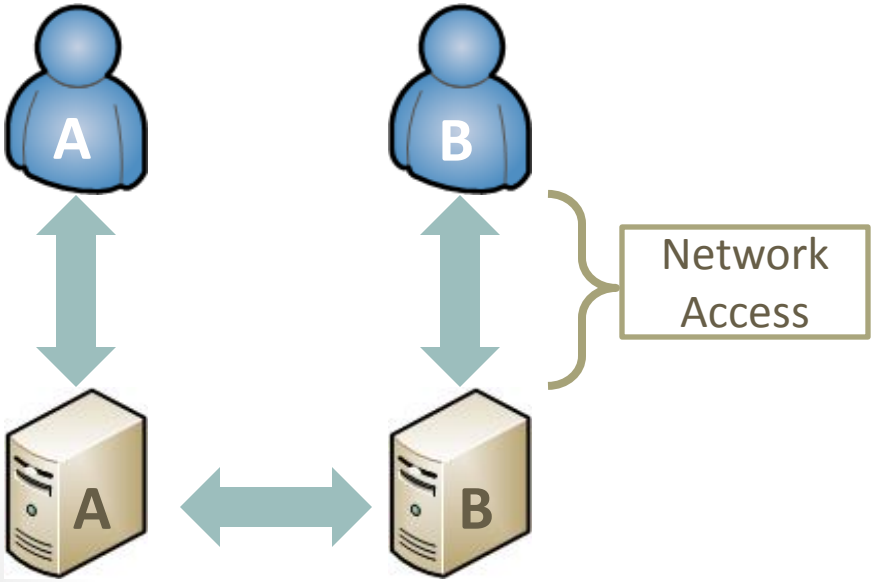
- Assume a cluster with shared and replicated data.
- The cluster consists of two connected nodes called A and B.

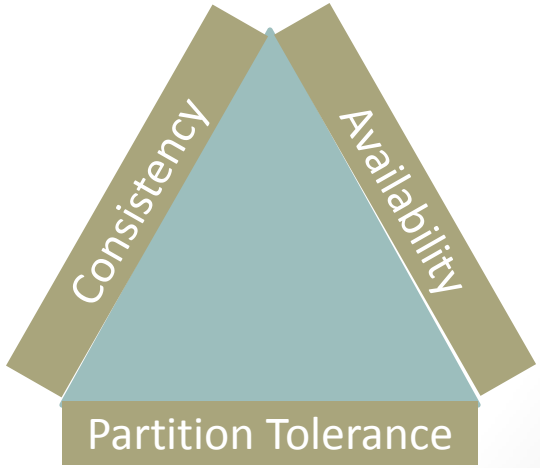




CAP Theorem (9)

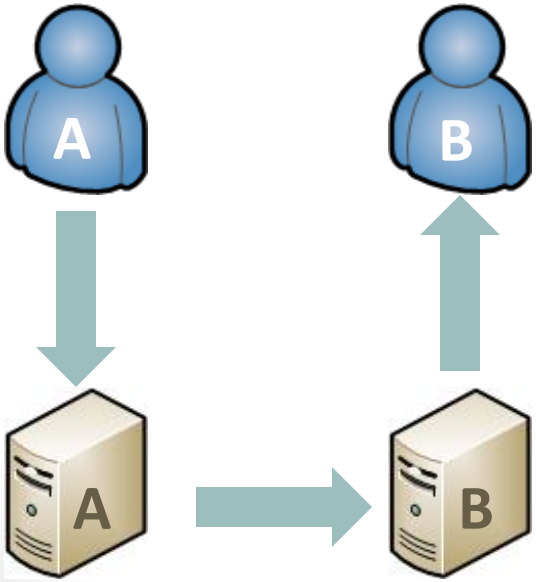
- Assume a cluster with shared and replicated data.
- The cluster consists of two connected nodes called A and B.
- The cluster is used by two users, called A and B. Each user has network access to a separate node

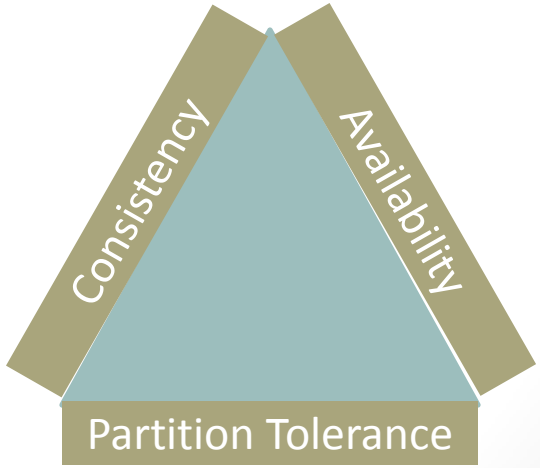




CAP Theorem (10)

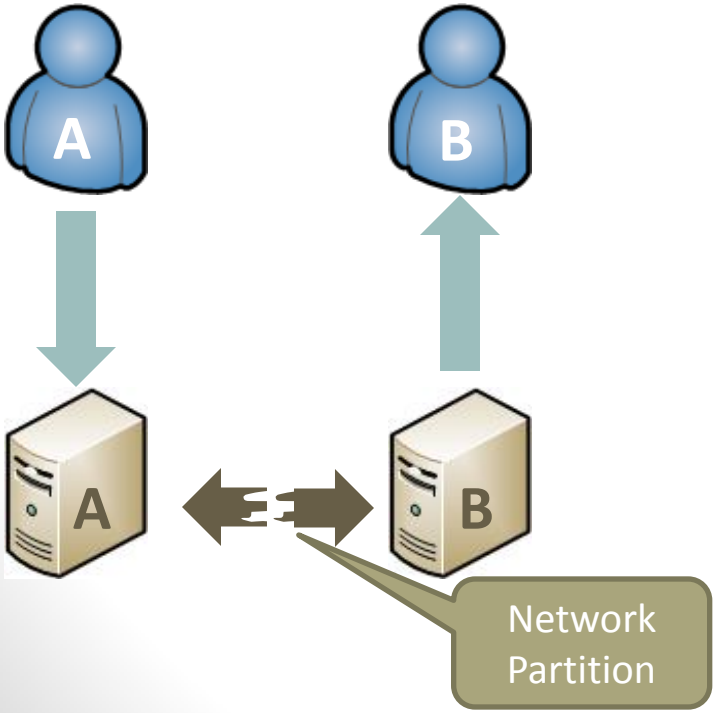
- Scenario 1: Network is available and Data are Consistent
  1. User A updates node A
  2. Update is communicated to node B
  3. User B reads the update from node B

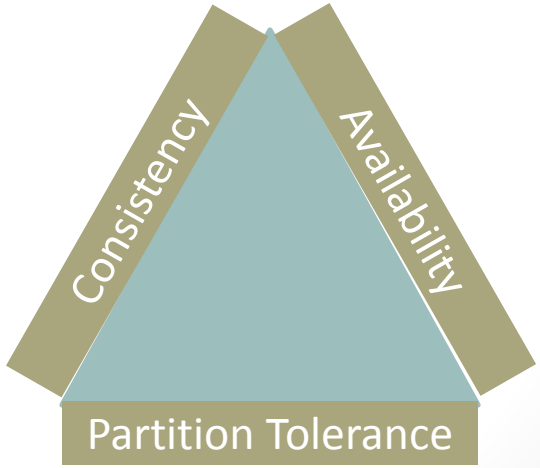




CAP Theorem (11)

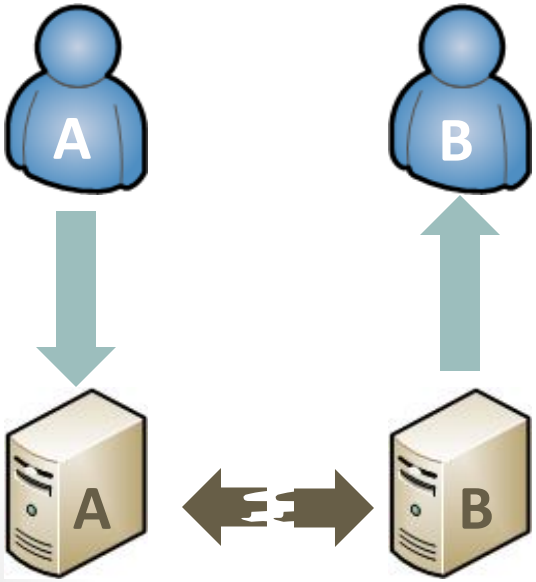
- Scenario 2: A network failure occurred.
  1. User A attempts to update node A
  2. Any Update cannot be communicated
  3. User B attempts to read the update

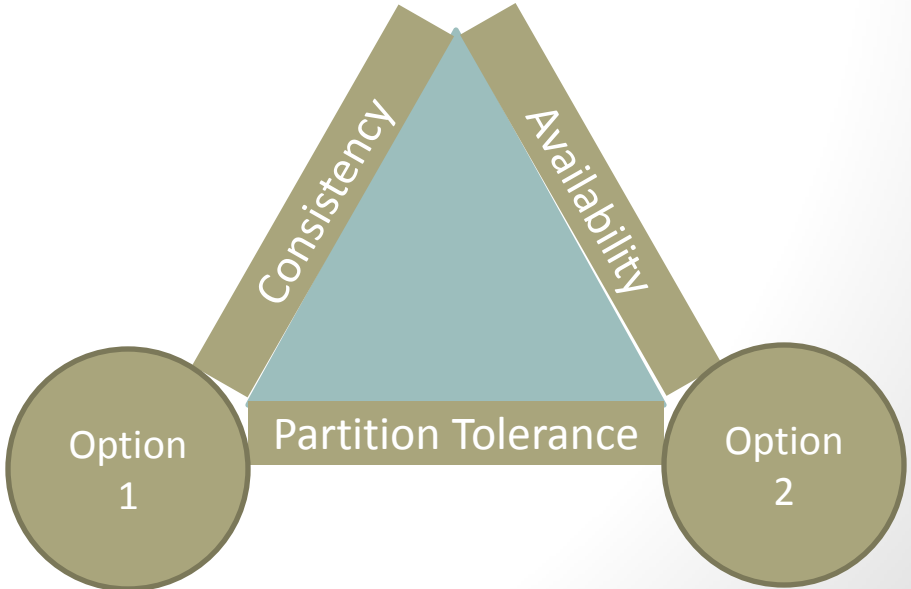




CAP Theorem (12)

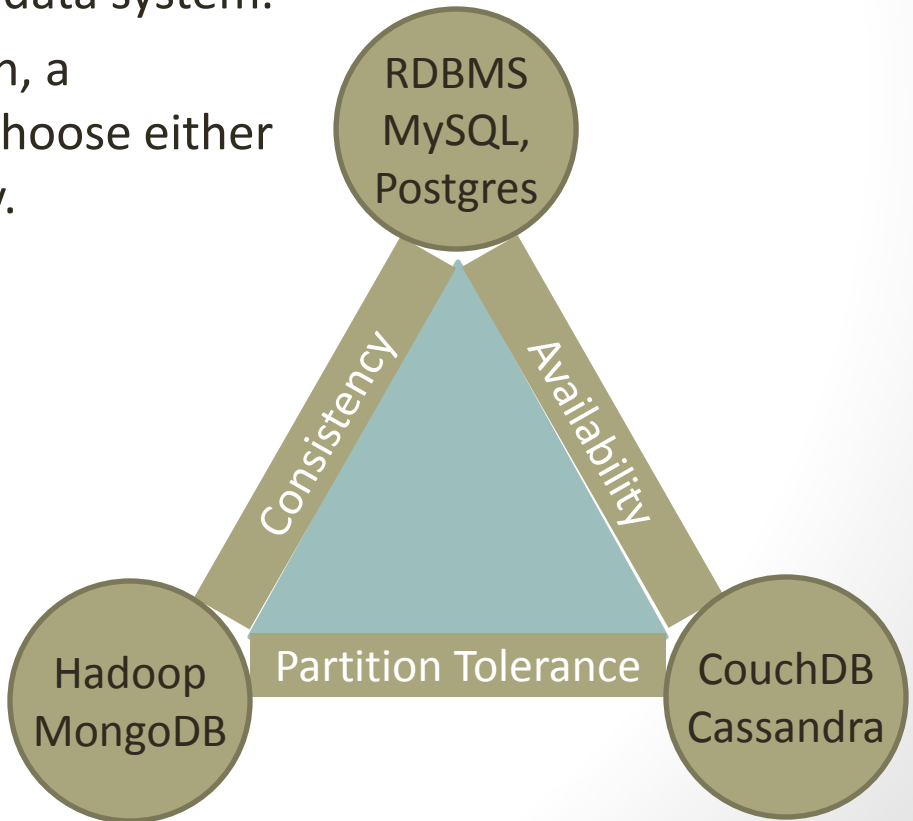
- Scenario 2: A network failure occurred. Two options:
  1. Make the database unavailable to avoid inconsistency
  2. Keep the database available and tolerate inconsistency





CAP Theorem (13)

- http://en.wikipedia/wiki/CAP_theorem
- Two formulations of the CAP theorem:
  1. You can have at most two of the CAP properties for any shared data system.
  2. During a network partition, a distributed system must choose either Consistency or Availability.



NOSQL: CAP Theorem

Quiz 07b NoSQL Introduction

Predictive Faux Pas

Predictive Faux Pas (1)

Facts vs. Hypotheses

"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."

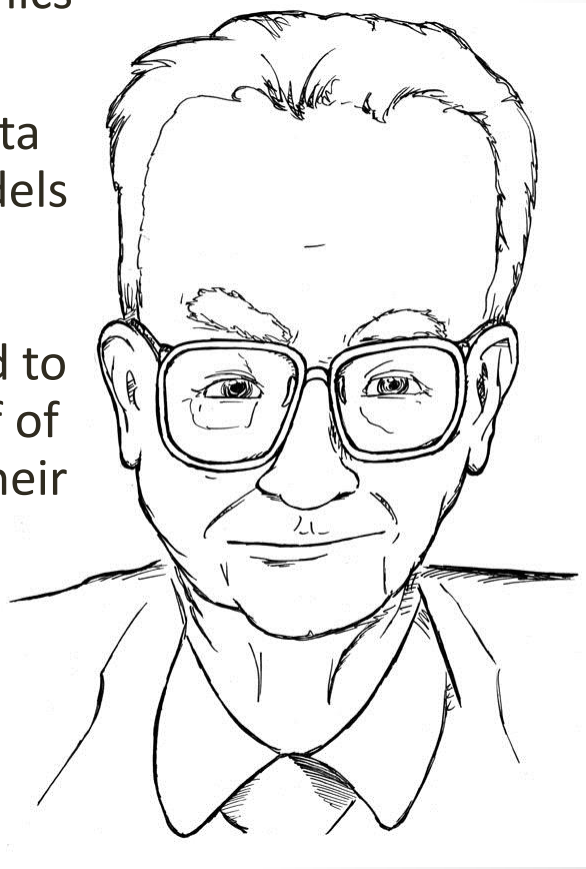
Sir Arthur Conan Doyle as the character of Sherlock Holmes

Predictive Faux Pas (2)

Facts vs. Hypotheses

- "If you torture the data long enough, it will confess," Ronald Coase, Nobel Prize in Economics

- Scientists are not interested in the data for data sake. Scientists want to use data to build models of nature. In other words, the results are secondary to the theories that help us understand nature. Good hypotheses are hard to come by and in my experience, easily one half of all scientists have twisted interpretations of their experimental results to support their pet hypotheses.



Predictive Faux Pas (3)

Facts vs. Hypotheses

In a nursing school we found that if the student's 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 overall student population and those students whose race was categorized as "Missing".

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 Faux Pas (4)

Facts vs. Hypotheses

- One might conclude: Facts before Hypotheses!

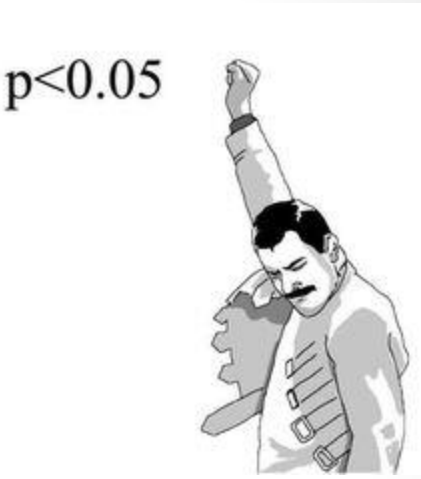
- On the other hand, there are arguments for Hypotheses before Facts:
  - We need hypotheses to guide research. Without a hypothesis we wouldn't know what to data to collect.
  - Facts before hypothesis may also lead to cherry picking or shot-gunning of hypotheses until a hypothesis fits. Then, we can use the p-value to determine if a hypothesis is good. The next slides explain the problem of looking for a hypothesis to fit your data.

Predictive Faux Pas (5)

Misuse of p-Value

- How is p-value misused?
  - Shot gunning or cherry picking hypotheses
  - Misunderstanding the nature of a p-value.

Predictive Faux Pas (6)



Misuse of p-Value

- How is p-value misused?
  - Do Jelly Beans Cause Acne with p < 0.05?
  - http://xkcd.com/882/
  - The null hypothesis states that the observed variations do no follow the hypothesis. If you choose enough hypotheses the there is an increasing chance that you will find a null hypothesis that has a low 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 assume that there is a 95% chance that the hypothesis accounts for the observations.
  - How many hypotheses (n) should we test if we want a more than even (50%) chance to find 1 or more p-values (p) at less than 5% from random data?
    - $0.5 < 1 - (1 - p)^n$; for p = 0.05 we find: n ≥ 14

Predictive Faux Pas (7)

Misuse of p-Value

How is p-value misused?

Misuse of p-values and experimental design

- After a large, epidemiological study failed to support a hypothesis, the researchers wanted to justify their grant. They looked for any pattern in their data. They transformed their data in as many ways as they could to find a pattern.
- When they found a pattern they retrospectively formulated a hypothesis and then they determined if that hypothesis had a p-value < 0.05, as is common in such studies. A a p-value < 0.05 means that there is only a 5% probability that the null hypothesis accounts for the patterns.
- The researchers announced many (50) hypotheses that were "verified" by this method. 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 about 5% of the null hypotheses are valid.

Predictive Faux Pas (8)

Misuse of p-Value

- How can we fix the problem?
- Statisticians need to explain the p-value to us.

- The following slides are adapted from: Statisticians Found One Thing They Can Agree On: It's Time To Stop Misusing p-values By Christie Ashwanden http://fivethirtyeight.com/features/statisticians-found-one-thing-they-can-agree-on-its-time-to-stop-misusing-p-values/

Predictive Faux Pas (9)

Misuse of p-Value

- How can we fix the problem?
- Statisticians need to explain the p-value to us.

Definition by Statisticians to layman:
Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data (for example, the sample mean difference between two compared groups) would be equal to or more extreme than its observed value.

Predictive Faux Pas (10)

Misuse of p-Value

- How can we fix the problem?
- Statisticians need to explain the p-value to us.

"That definition is about as clear as mud"

Definition by Statisticians to layman:
Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data (for example, the sample mean difference between two compared groups) would be equal to or more extreme than its observed value.

"... even scientists can't easily explain p-values"

Predictive Faux Pas (11)

Misuse of p-Value

More links:

- http://www.nature.com/news/statisticians-issue-warning-over-misuse-of-p-values-1.19503
- http://www.nature.com/news/how-scientists-fool-themselves-and-how-they-can-stop-1.18517
- http://www.nature.com/news/scientific-method-statistical-errors-1.14700
- http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124
- http://www.nature.com/news/statisticians-issue-warning-over-misuse-of-p-values-1.19503

Predictive Faux Pas (12)

Observational Studies

- If we have a problem with collecting facts before developing a hypothesis and we have a problem with developing a hypothesis before collecting facts, then we may want to observe and draw conclusions from observational studies.
- In Search of Excellence (1982) one of the most popular business books of all time. Studied 44 successful "excellent" companies.
- 5 years later 65% of the companies did worse than the S&P 500 Index
- The case study method makes for good drama / story-telling, but it's not good science. Small sample sizes, subject to numerous biases (survivorship, extreme cases,…)
- Other examples: One-off medical studies,

Predictive Faux Pas (13)

Observational Studies

- "Any claim coming from an observational study is most likely to be wrong" (Stanley Young)
- Young and Karr looked at 52 similar published epidemiological findings that were followed by a clinical trial testing the result.
- NONE of the 52 claims replicated in the clinical trials! (5 were significant in the opposite direction.)

Predictive Faux Pas (14)

Observational Studies

- Wrong results from an observational study could be
  - Innocent
  - Not so innocent – sort through the data to find evidence to prove your case and ignore all the other signals
- How to determine if it's a real insight?
  - Test it – conduct a valid experiment to see if the presumed cause and effect relationship holds (e.g. clinical trial, design of experiments (DOE))
  - Get additional, independent data sets and see if the relationship is still present
  - Caution: Most analyses only validate the presence of a relationship. Most analyses do not even show that a cause and effect relationship exist. And, even if a cause and effect relationship exists, we might know which is cause and which is effect.
  - Never allow the same dataset to suggest a relationship AND validate it.

Predictive Faux Pas (15)

Reliance on Descriptive Measures

- We need to understand our data better before we make conclusions.
- We can use descriptive measures to help us understand our data

Predictive Faux Pas (16)

Reliance on Descriptive Measures

Francis determined 6 measures of a dataset.

Property	Value
mean(y)	9
var(x)	11
mean(y)	7.50
var(y)	4.125
cor(x,y)	0.816
lm(y~x)	y = 3 + 0.5x

Predictive Faux Pas (17)

Reliance on Descriptive Measures

Francis determined 6 measures of a dataset.

Property	Value
mean(y)	9
var(x)	11
mean(y)	7.50
var(y)	4.125
cor(x,y)	0.816
lm(y~x)	y = 3 + 0.5x

What does the dataset look like?
Do the data have outliers?
Do the data form a linear relationship?
Can he extrapolate from this relationship?

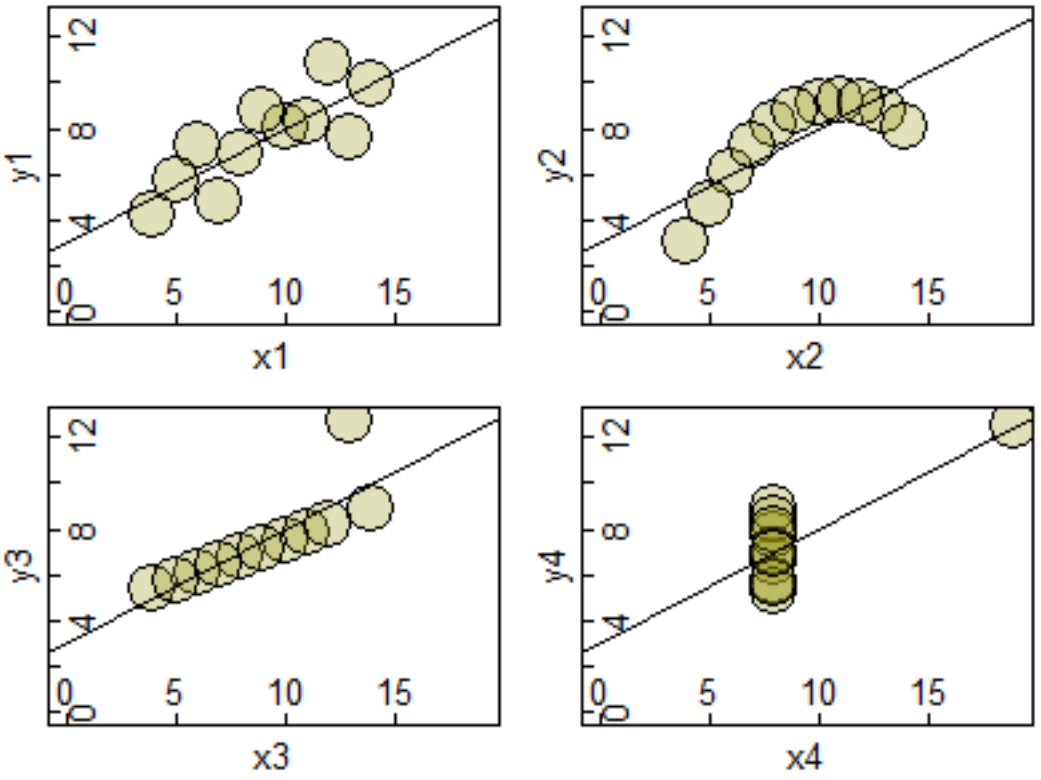
Predictive Faux Pas (18)

Reliance on Descriptive Measures

Francis determined 6 measures of a dataset.

Property	Value
mean(y)	9
var(x)	11
mean(y)	7.50
var(y)	4.125
cor(x,y)	0.816
lm(y~x)	y = 3 + 0.5x

What does the dataset look like?
Do the data have outliers?
Do the data form a linear relationship?
Can he extrapolate from this relationship?



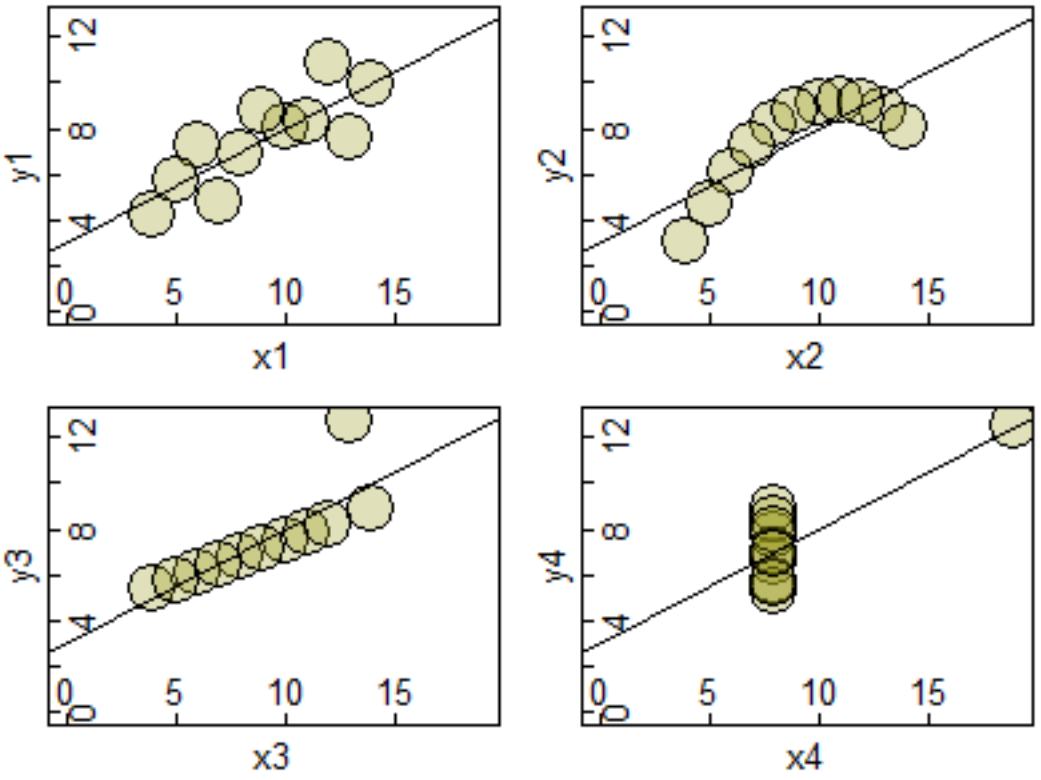
Predictive Faux Pas (19)

Reliance on Descriptive Measures

Francis determined 6 measures of a dataset.

Property	Value
mean(y)	9
var(x)	11
mean(y)	7.50
var(y)	4.125
cor(x,y)	0.816
lm(y~x)	y = 3 + 0.5x

What does the dataset look like?
Do the data have outliers?
Do the data form a linear relationship?
Can he extrapolate from this relationship?



Which data set belongs to these measurements?

Predictive Faux Pas (20)

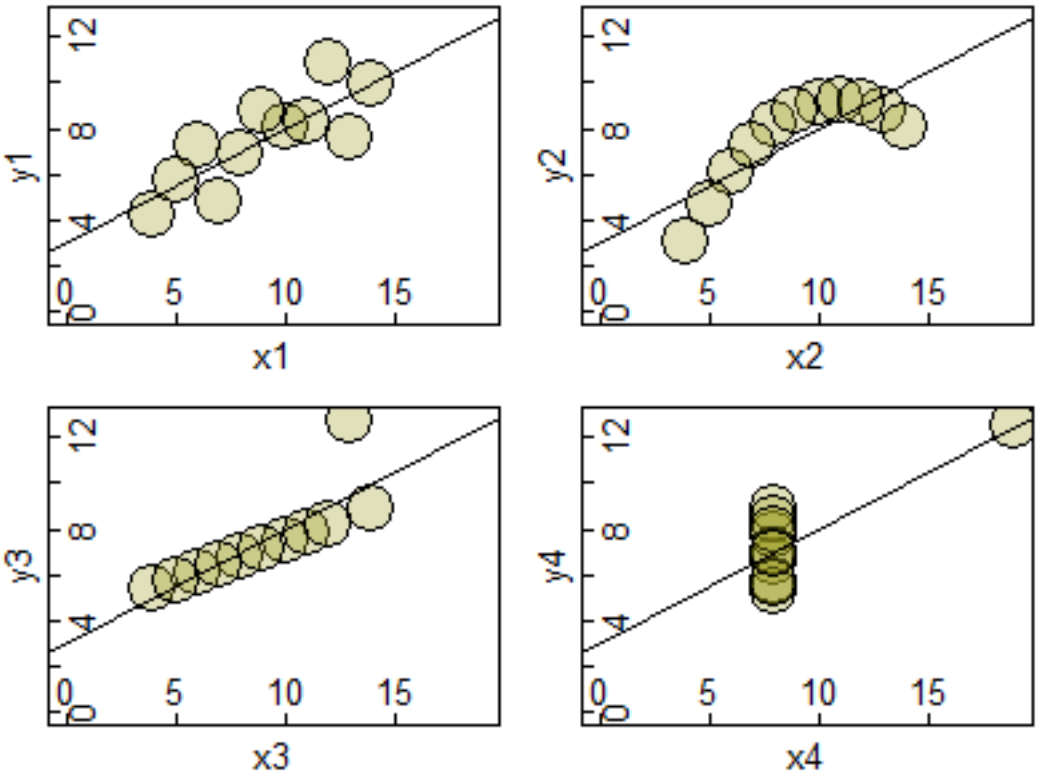
Reliance on Descriptive Measures

Francis determined 6 measures of a dataset.

Property	Value
mean(y)	9
var(x)	11
mean(y)	7.50
var(y)	4.125
cor(x,y)	0.816
lm(y~x)	y = 3 + 0.5x

All these data sets have these measurements!

Anscombe's Quartet
https://en.wikipedia.org/wiki/Anscombe%27s_quartet



Predictive Faux Pas (21)

- Other problems may arise due to expectations of future performance.

Predictive Faux Pas (22)

Expectations on Performance

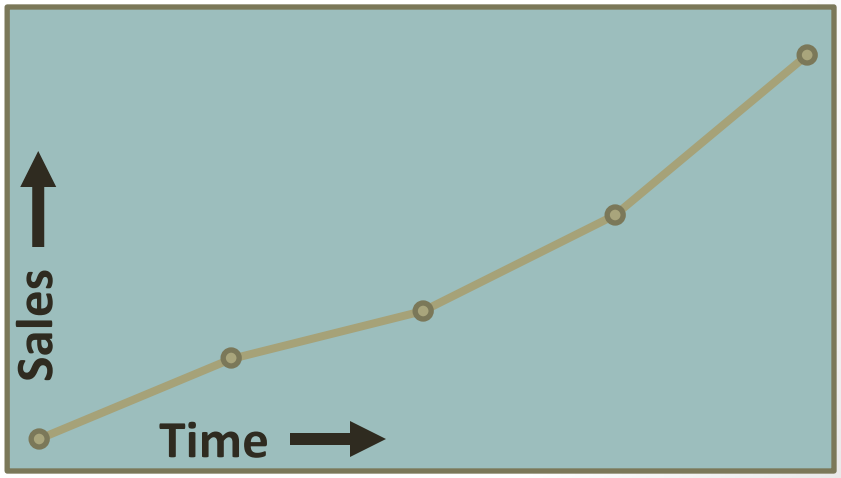
- Management complains: "Our top 1000 customers in 2014 bought 20% less in 2015"
- Management assumes that these customers were disappointed. But, a reduction is expected. The phenomenon is known as "regression to the mean"
- If a measurement of a variable is observed to be extreme, and there is no trend, it will tend to be closer to the average on the next measurement
- Examples:
  - Performance Reviews
  - Sales by Account Managers
  - Sports Illustrated Jinx

- ("Regression to the mean" is the origin of the word regression as in linear regression.)

Predictive Faux Pas (23)

Interpreting Recent Trends

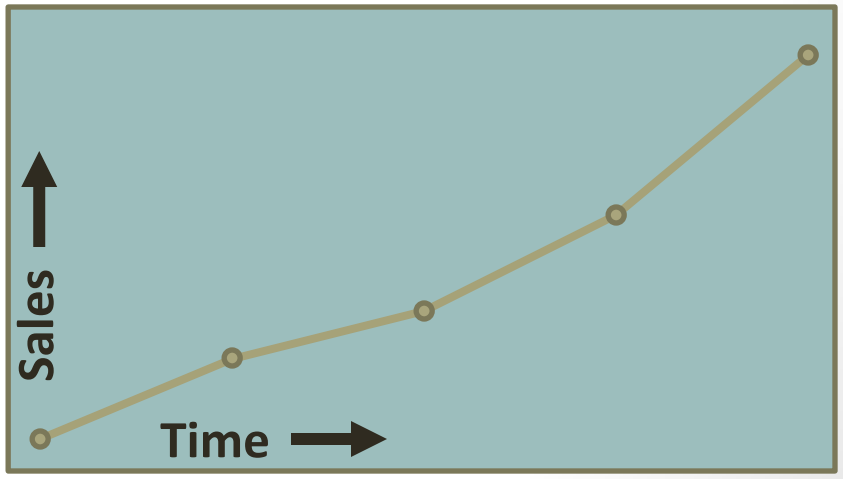
- Sales have increased for the past four months in a row. Are we on a meaningful trend? What are the chances?



Predictive Faux Pas (24)

Interpreting Recent Trends

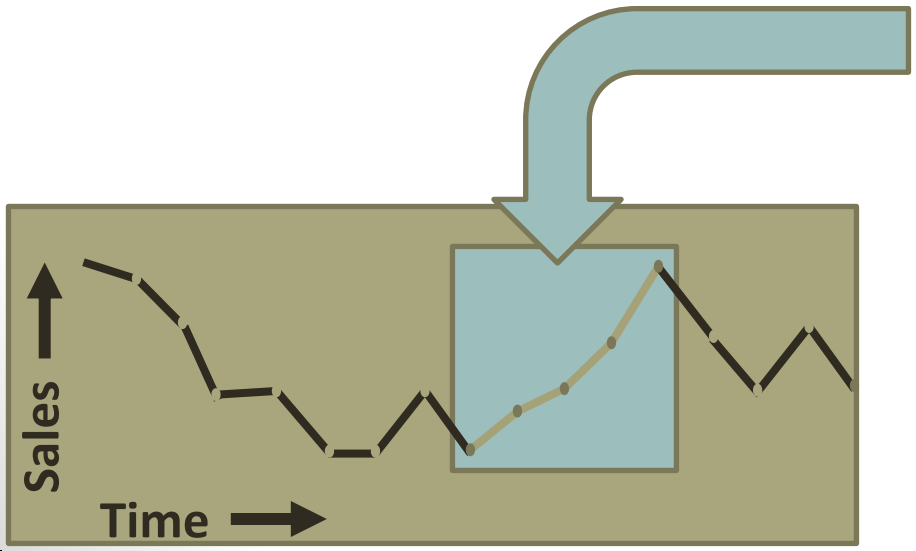
- Sales have increased for the past four months in a row. Are we on a meaningful trend? What are the chances?
- 5 monthly measurements where each successive measurement increased:
- ½ * ½ * ½ * ½ = 6%

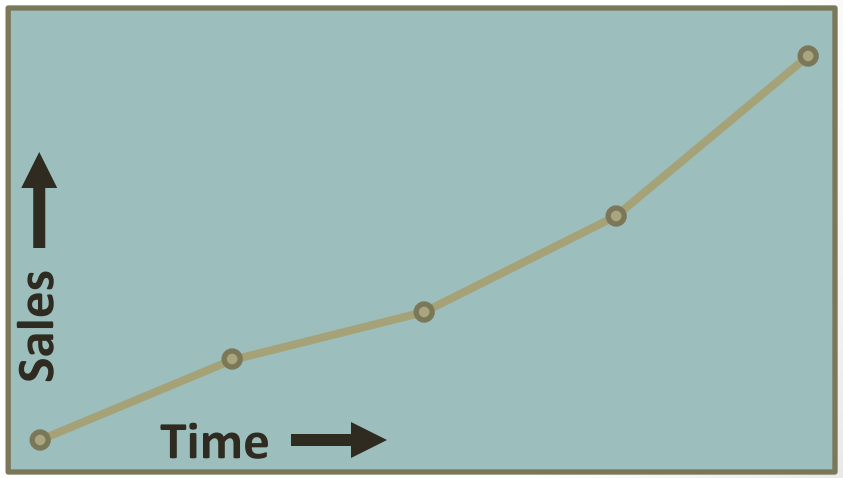


Predictive Faux Pas (25)

Interpreting Recent Trends

- Sales have increased for the past four months in a row. Are we on a meaningful trend? What are the chances in a year?

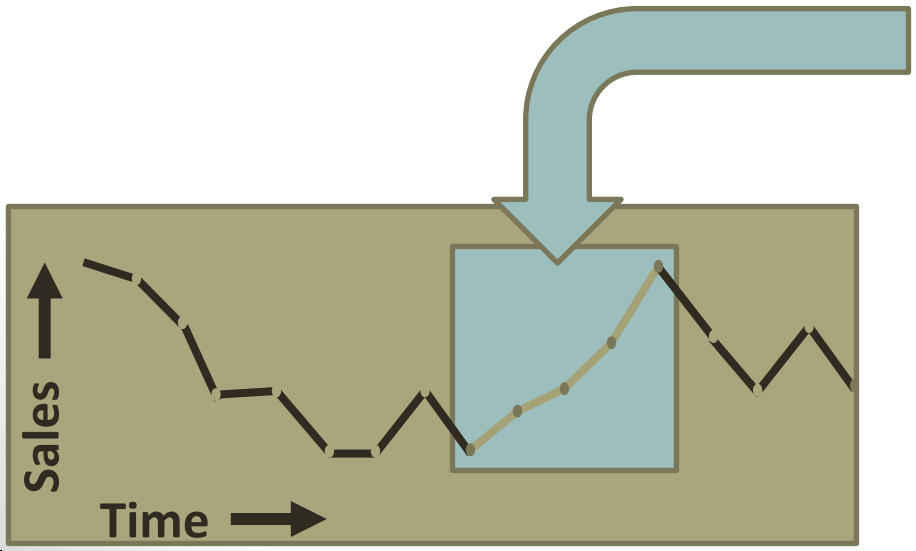


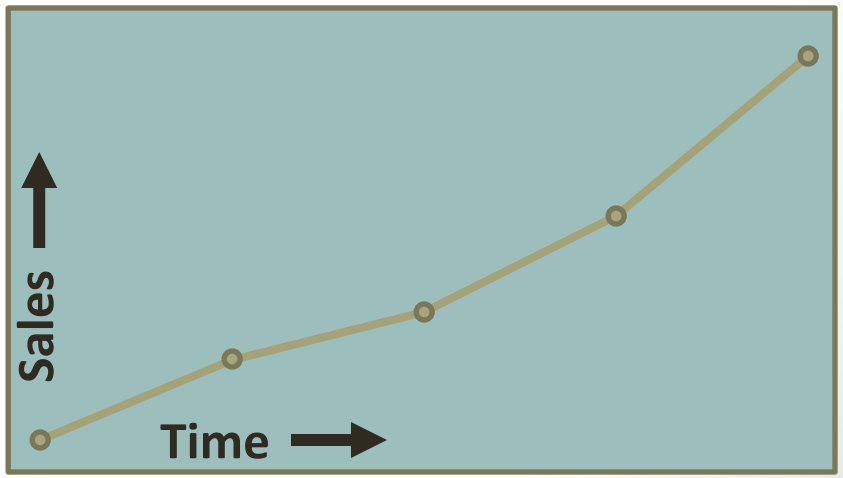


Predictive Faux Pas (26)

Interpreting Recent Trends

- Sales have increased or decreased for the past four months in a row. Are we on a meaningful trend? What are the chances in a year?
  - 9 sequences of 4 changes: 0-4, 1-5, …, 8-12
  - 4 sequential increases
  - $1 - (1 - 2^{-4})^{9} = 0.44$
- Assume a monthly random measurement. In a year there is a 44% chance of 4 sequential increases.





Predictive Faux Pas (27)

Correlation *vs.* Causation

- A common human trait is to observe two things occurring together and assume one is causing the other
- Examples:
  - Leading Economic Indicators
  - Bad Breath and Heart Disease

- An observed (statistically significant) relationship may be due to
  - Happenstance (i.e. chance or co-incidence)
  - Statistical significance helps, but among 100 relationships with p=0.05, odds are that about 5 will be by chance.
  - Common hidden factor
  - True cause-effect relationship but which direction?

Predictive Faux Pas (28)

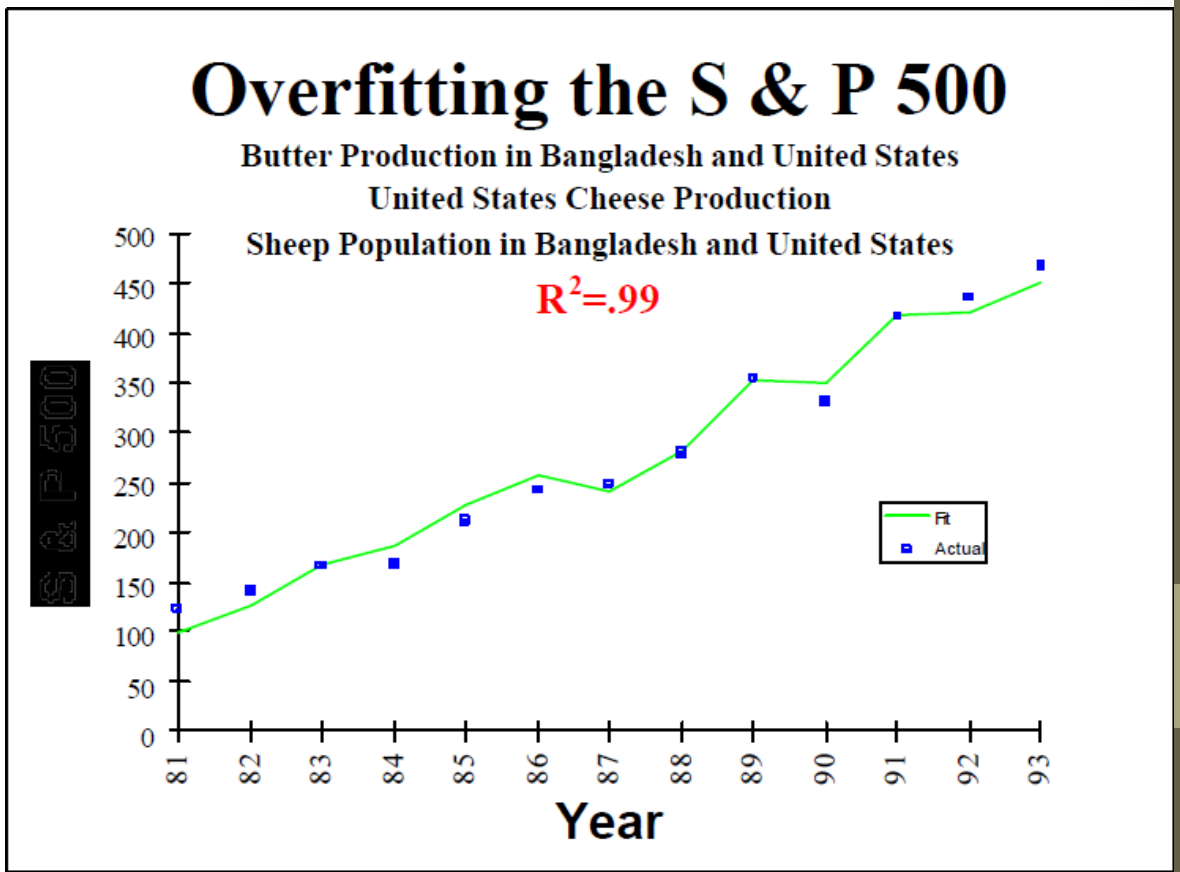
Spurious Relationship

https://en.wikipedia.org/wiki/Spurious_relationship

Predictive Faux Pas (29)

Spurious Relationship

Exact prediction of S&P 500 returns by Ivan O. Kitov, Oleg I. Kitov (See: SSRN-id1045281.pdf)



Predictive Faux Pas (30)

Spurious Relationship

- 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 Faux Pas (31)

Hidden Proxies

- 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 Faux Pas (32)

Hidden Proxies

- Proxies 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 by getting 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.

Predictive Faux Pas (33)

Selective Presentation of Outcomes

- In the 1950's, a convict in Italy, wrote to 80 stockbrokers from prison. He claimed to have insider information from a fellow convict who had been an executive at a local company.
- 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.

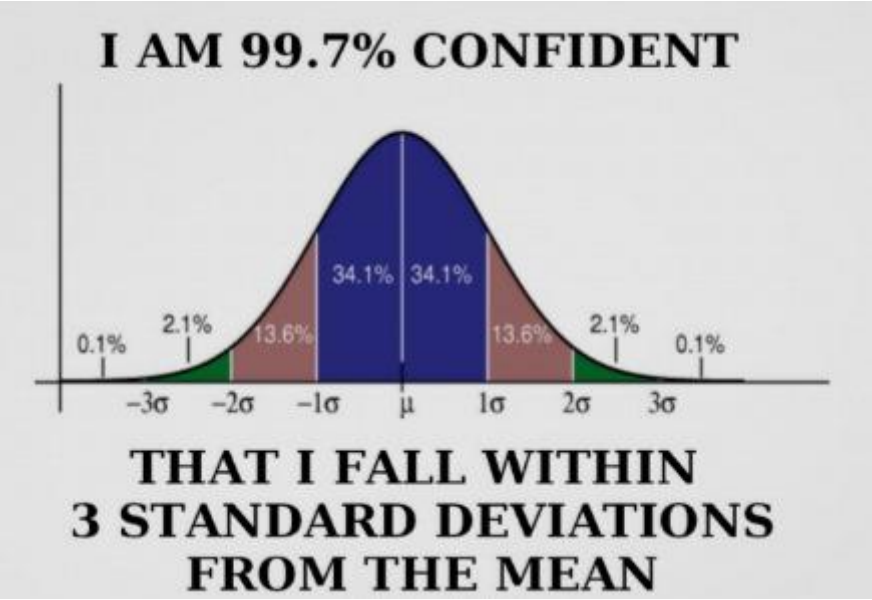
Predictive Faux Pas (34)

Superstition

- HBR Superstitious Learning
- "Superstitious learning takes place when the connection between the cause of an action and the outcomes experienced aren't clear, or are misattributed."
- Some Causes:
  - Expecting high/low performance to remain at that level
  - Interpreting trends that could be due to randomness
  - One-off occurrences
  - Causation inferred from correlation

Predictive Faux Pas (35)

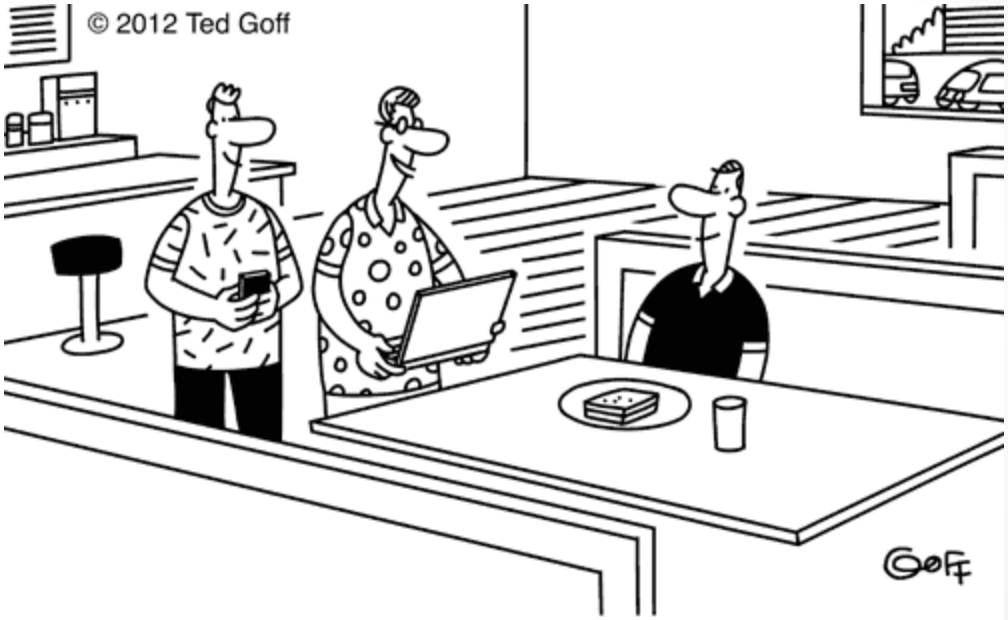
Tautology



Predictive Faux Pas (36)

- Some more links
  - http://skepdic.com/perfectprediction.html
  - http://www.investorhome.com/scam.htm
  - http://www.forbes.com/sites/davidleinweber/2012/07/24/stupid-data-miner-tricks-quants-fooling-themselves-the-economic-indicator-in-your-pants/
  - Leo Breiman, Statistical Modeling: The Two Cultures, Statistical Science, 2001, Vol. 16, No. 3, 199–231 http://projecteuclid.org/download/pdf_1/euclid.ss/1009213726

Predictive Faux Pas (37)



"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 Faux Pas

Assignment

1. Start a new discussion on statistics before Sunday Nov 20th 11:57 PM. The discussion should be on a topic that interests you. Write the specific topic of that discussion in a text file called discussion.txt.
2. Submit discussion.txt before Nov 20th 11:57 PM to Canvas.
3. Comment on a fellow student's discussion sometime after Nov 20th but before Sunday Nov 27th 11:57 PM. Write that same comment into a text file called comment.txt.
4. Submit comment.txt before Nov 27th 11:57 PM to Canvas.
5. Start up your VM.
   a) In the console verify that the following two commands work:
      i. $ hadoop fs -ls /
      ii. $ hadoop fs -ls /user
   b) Test that you can list the file "shakespeare.tar.gz"
      i. $ cd ~/training_materials/developer/data
      ii. $ ls
6. Be prepared to use your VM during class starting Nov 30th 2016.
7. If you were a late admit who did not hand in your first assignment, then you can do an optional assignment for half-credit. The optional assignment is twice as much work as the original assignment. Please contact me and download DataPreparation.pdf from Canvas on Nov 17th 2016. Submission due date is Nov 20th 11:57 PM.

Introduction to Data Science