SI6

Gestion des données

Eléments de Data Management

1. Tendances

* 2013 année du Big Data (2008) : data scientist
* 2014 année du Small Data : simplicité ?

* Pour comprendre votre Business :
  * Analyser vos données
  * Ne pas prendre celles des autres (paramètres différents ?)
  * Comparer ce qui est comparable

* Small Data compare ce qui comparable (données de même type)

2. Problématique du Big Data

* Expose à des problèmes éthiques
* Set de données colossale
* Bruits, parasites, silences, omissions dans les données
* Données non structurées
* Données à faibles densité
* Infrastructure matérielle, logicielle
* Exhaustivité possible (besoins réels ?)
* Seuil de 100000 clients

3. Réponse du Small Data

* Données locales :
  * Feuille Excel
  * Base de données d'une entreprise
  * Avis des collaborateurs
  * Données à fortes densité

* Le niveau social est à prendre en compte (écouter les gens !)
* Le signal faible est a écouter (veille nécessaire)
* Intelligence économique : plusieurs niveaux d'analyse (brevets, sécurité, droit d'auteur, piratage, partage de la données, …)

4. Domaine

* Impossible de tout surveiller avec moins de moyens
* Il faut définir le terrain d'observation !
* Quels réseaux intéressent l'entreprise ?
* Quelles informations cherche-t-on ?
* Quelles informations « critiques » attend-on ?
* Signaux forts ? Facile
* Signaux faibles ? Veille

5. Outils

* Outils permettant de faire de la veille :
  * Infomous : http://get.infomous.com (analyse de contenu)
  * Gephi : open source (prise en main) https://gephi.org
  * NodeXL : http://nodexl.codeplex.com (template Excel)
  * TAGSExplorer : Google Document (Twitter)
  * Les outils de requêtage de chaque réseau
  * Formatage Google : utiliser un autre moteur pour comparer.

* Outils d'analyse :
  * Google fusion Table (experimental Google Drive app)

6. Comparaisons

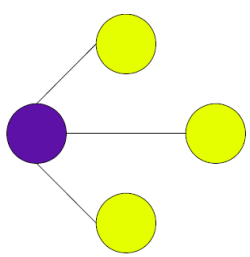
Category	Big Data	Small Data
Data Sources	Data generated outside the enterprise from nontraditional data sources. Include: • Social media • Sensor data • Log data • Device data • Video, Images, etc.	Traditional enterprise data. Includes: • Enterprise Resource Planning transactional data • Customer Relationship Management (CRM) systems • Web transactions • Financial data e.g. general ledger data
Volume	• Terabytes (10^{14}) • Petabytes (10^{15}) • Exabytes (10^{18}) • Zettabytes (10^{21})	• Gigabytes (10^{9}) • Terabytes (10^{12})
Velocity	• Often real-time • Requires immediate response	• Batch or near real-time • Does not always require immediate response
Variety	• Structured • Unstructured • Multi-structured	• Structured • Unstructured
Value	• Complex, advanced, predictive business analysis and insights	• Business Intelligence, analysis and reporting

7. Analyse des réseaux sociaux

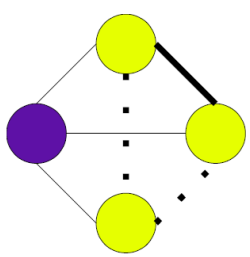
* Facebook
* Twitter
* Google +
* Analyse de la forme du réseau (Modelisation)
  * Critères + outils
* Collecte des données (Database = noSQL ou pas)
* Analyse du contenu (Data Mining)
  * Critères + outils
* Restitution des données (DataViz)
  * Critères + outils

8. Analyse de la forme du réseau

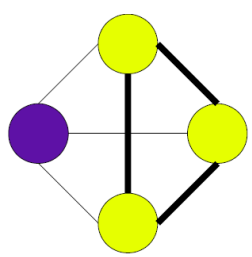
* Anthropologie, sociologie, sociométrie, ethnographie, netnographie.
* Réseau social : échantillon de société numérisée
* Échantillonnage :
  * Qualitatif : à la volée, set incomplet, faible densité
* Critères : niveaux de relations entre les membres
  * Recherche des leaders (hub)
  * Analyse des clusters (hub)
  * Coefficient clustering
  * Effet petit monde



(a) No pairs formed among neighbors: $C = 0$



(b) One pair formed among neighbors: $C = 1 / 3$



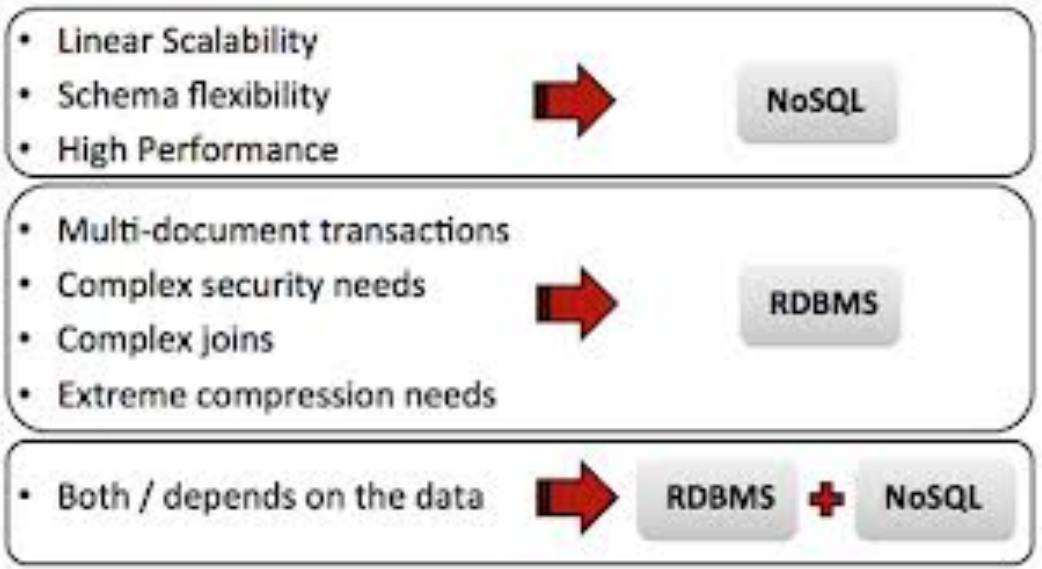
(c) Three pairs formed among neighbors: $C = 3 / 3$

8. Analyse de la forme du réseau

* Outils d'analyse des réseaux sociaux :
  * Automap : http://www.casos.cs.cmu.edu/projects/automap/
  * Cfinder : http://www.cfinder.org
  * Commetrix : http://www.commetrix.de
  * Dynet : http://www.casos.cs.cmu.edu/projects/DyNet/
  * Egonet : http://sourceforge.net/projects/egonet/
  * Idiro : http://www.idiro.com
  * KXEN : http://www.kxen.com/Products/Social+Network+Analysis

9. Collecte de la données

* Réseaux sociaux, moteurs de recherche
* Stockage :
  * Base de données (SGBDR) : SQL (Web : PDO, ORM : sqlite + CoreData, hibernate, nhibernate)
  * Base noSQL (clé/valeur, graphe, …) : FrameWork : POO
  * Sérialisation : XML, JSON



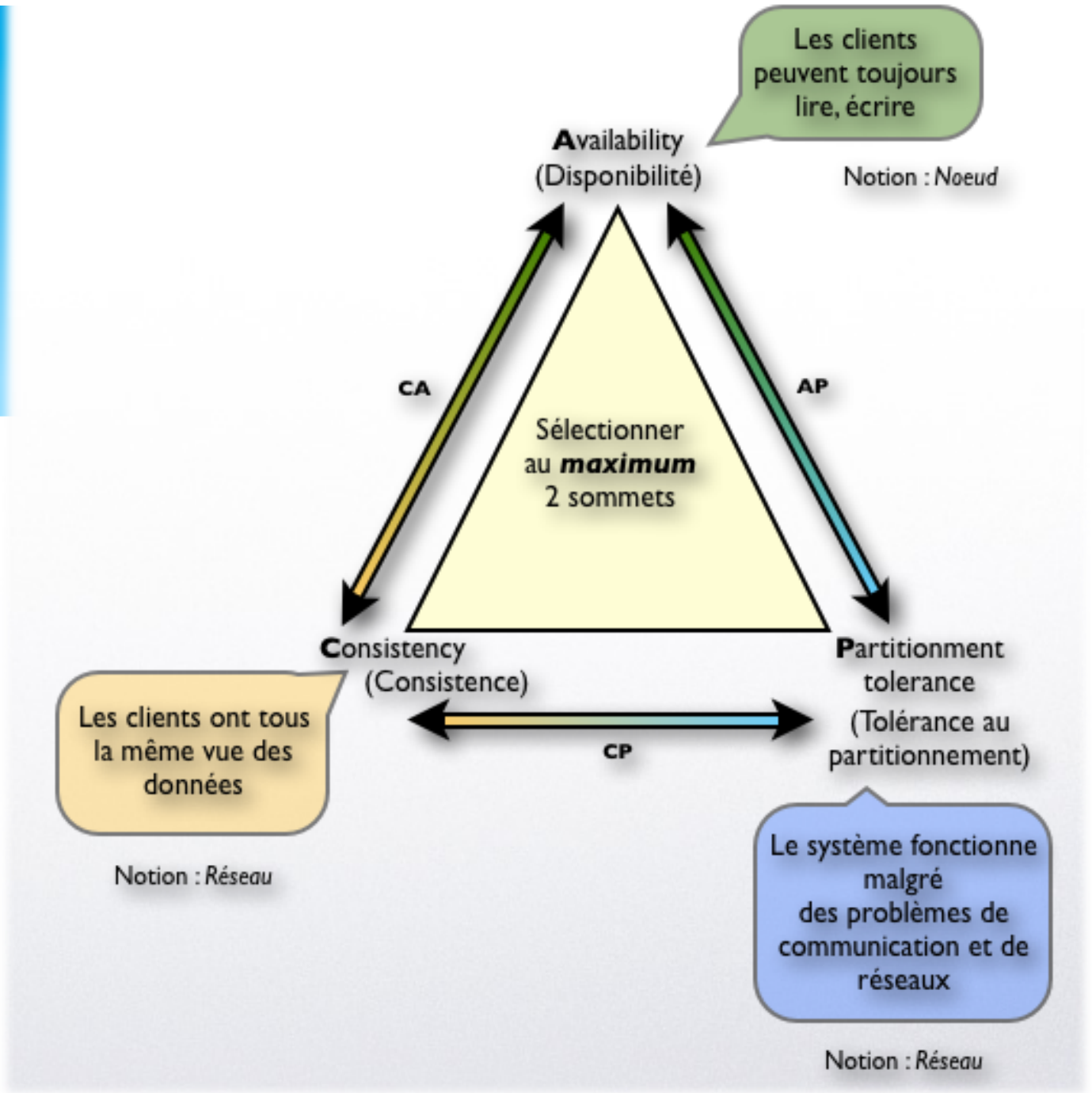
SGBDR vs noSQL

ACID

Théorème CAP

* **SGBDR :** AC (disponibilité, cohérence)

* **noSQL** : CP, AP (cohérence, résistance au partitionnement), (Disponibilité, résistance au



10. Analyse du contenu

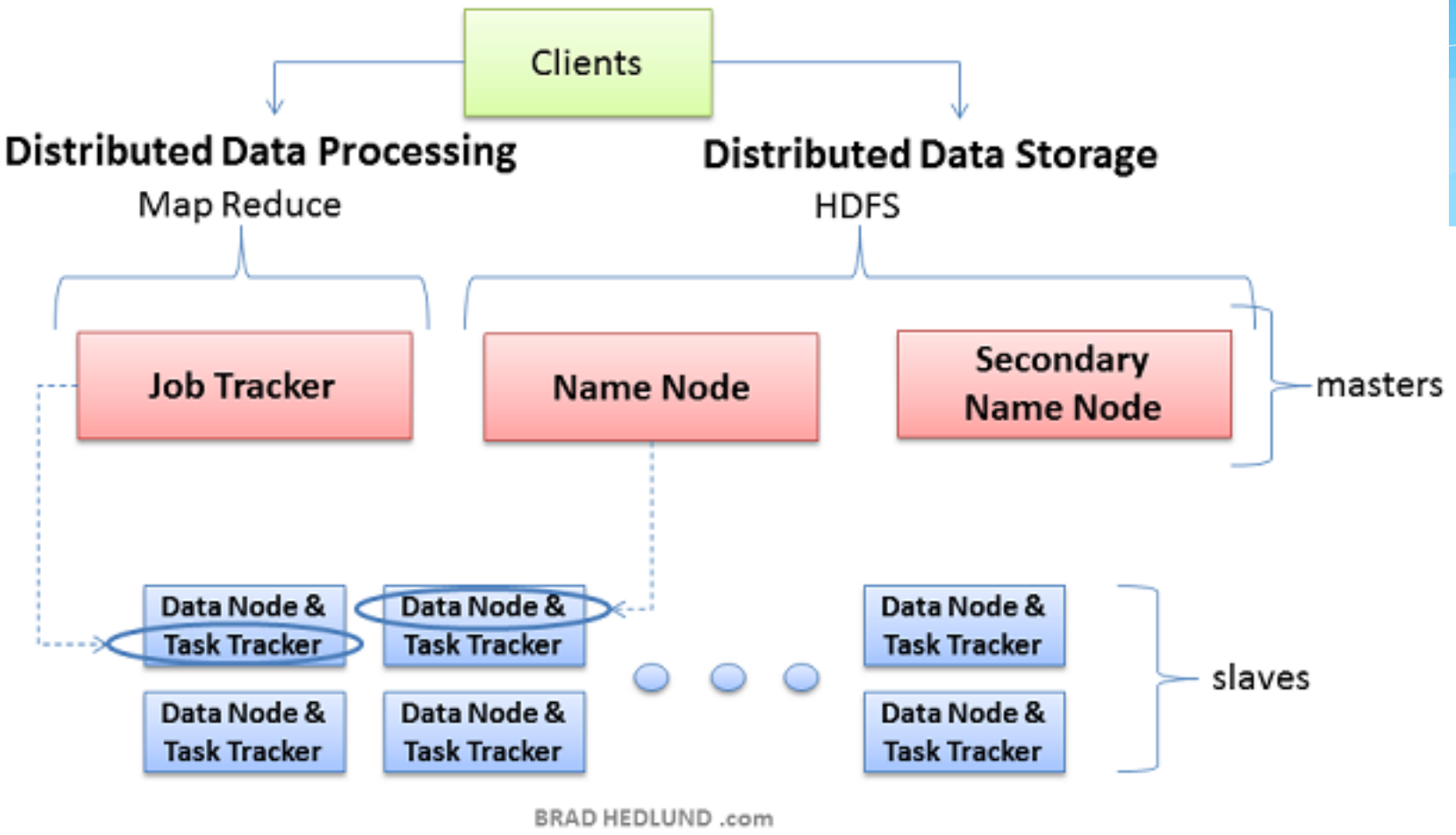
* Seuil recommandé : 100 000 unités
* Critères :
  * Suivant ce que l'on veut faire dire aux données
  * Suivant les accès autorisés sur les API
  * Exemple : FireHose Twitter

* Méthodes : (Big & Small Data)
  * Prepare : get and clean data
  * Analyse : Processed them (Big analytics)
  * Apply : Make the right decision (visualization)

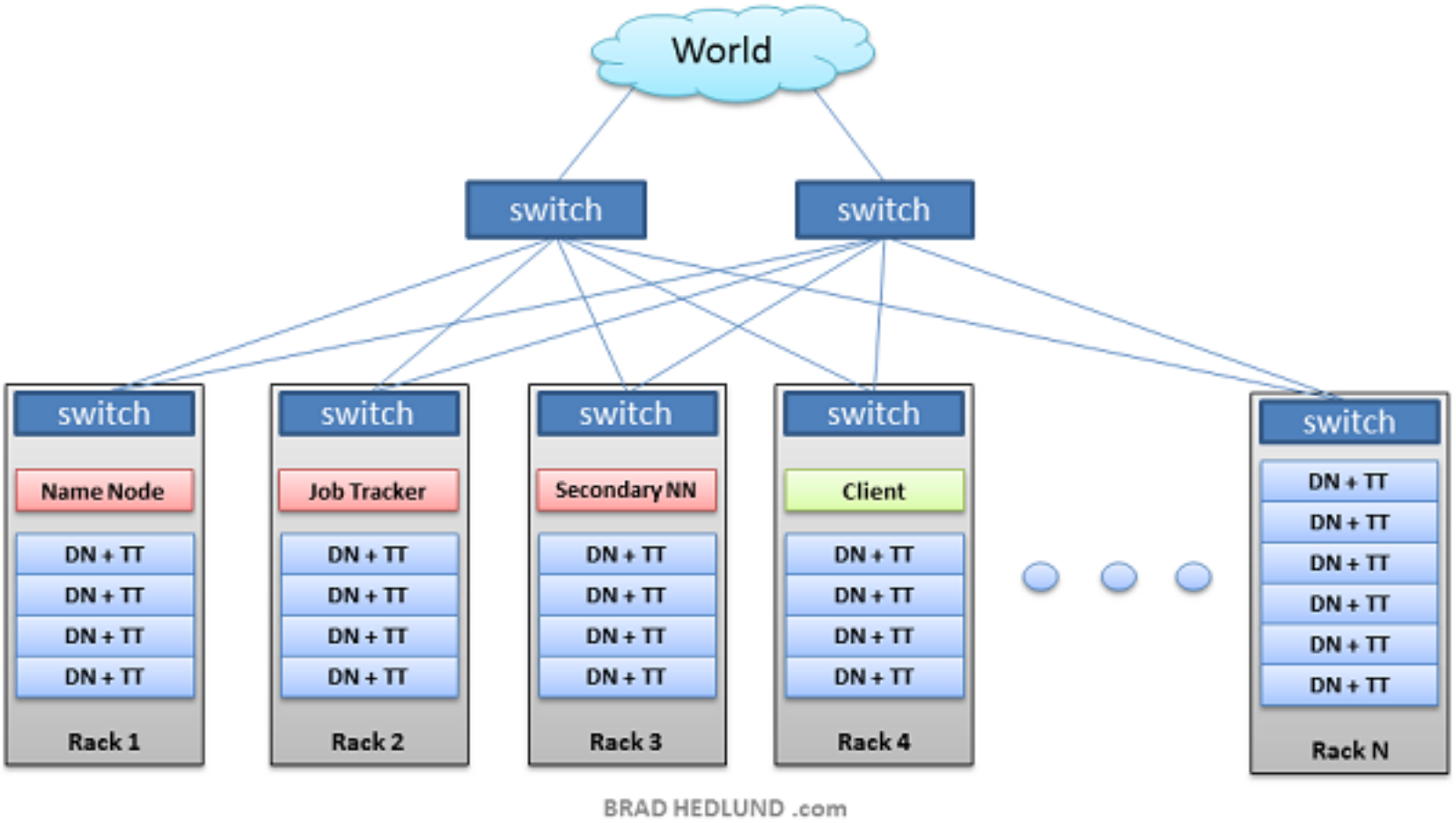
10. Analyse du contenu

* Outils (scalable) : Présence de l'openSource (Apache)
  * Hadoop, Mahout = Big Data (FrameWork Java)
  * Parallélisation (node + cluster = data Center)
  * Google Fusion Table = Small Data (summarize, chart, filter)
* Algorithmes (scalable) :
  * mapReduce
  * Machine learning :
    * Clustering : aggrégation (Google News)
    * Recommandation : Amazon
    * Classification de Spams : Messagerie

Hadoop Server Roles



Hadoop Cluster



Typical Workflow

- Load data into the cluster (HDFS writes)
- Analyze the data (Map Reduce)
- Store results in the cluster (HDFS writes)
- Read the results from the cluster (HDFS reads)

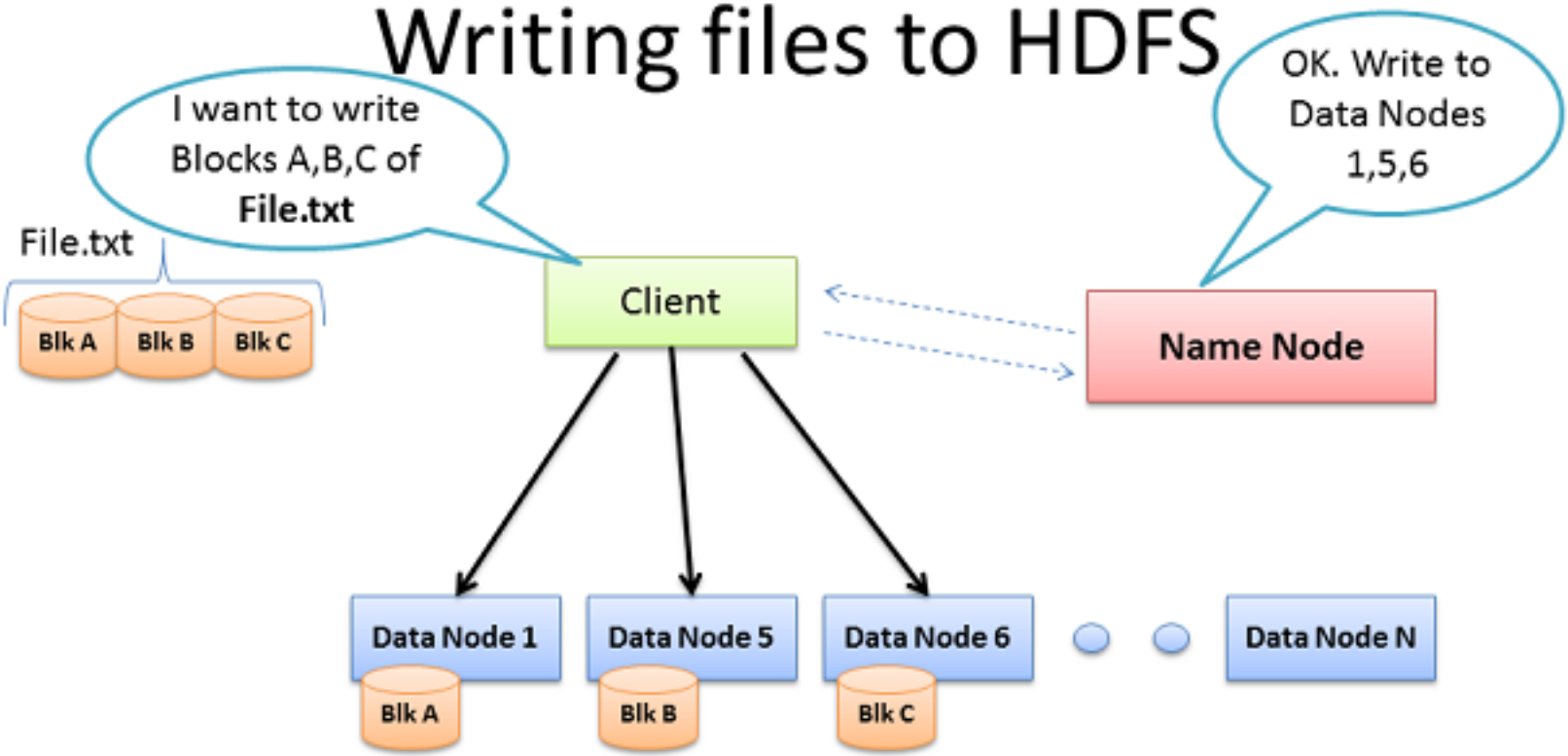
Sample Scenario:

How many times did our customers type the word "**Refund**" into emails sent to customer service?

Huge file containing all emails sent to customer service

BRAD HEDLUND .com

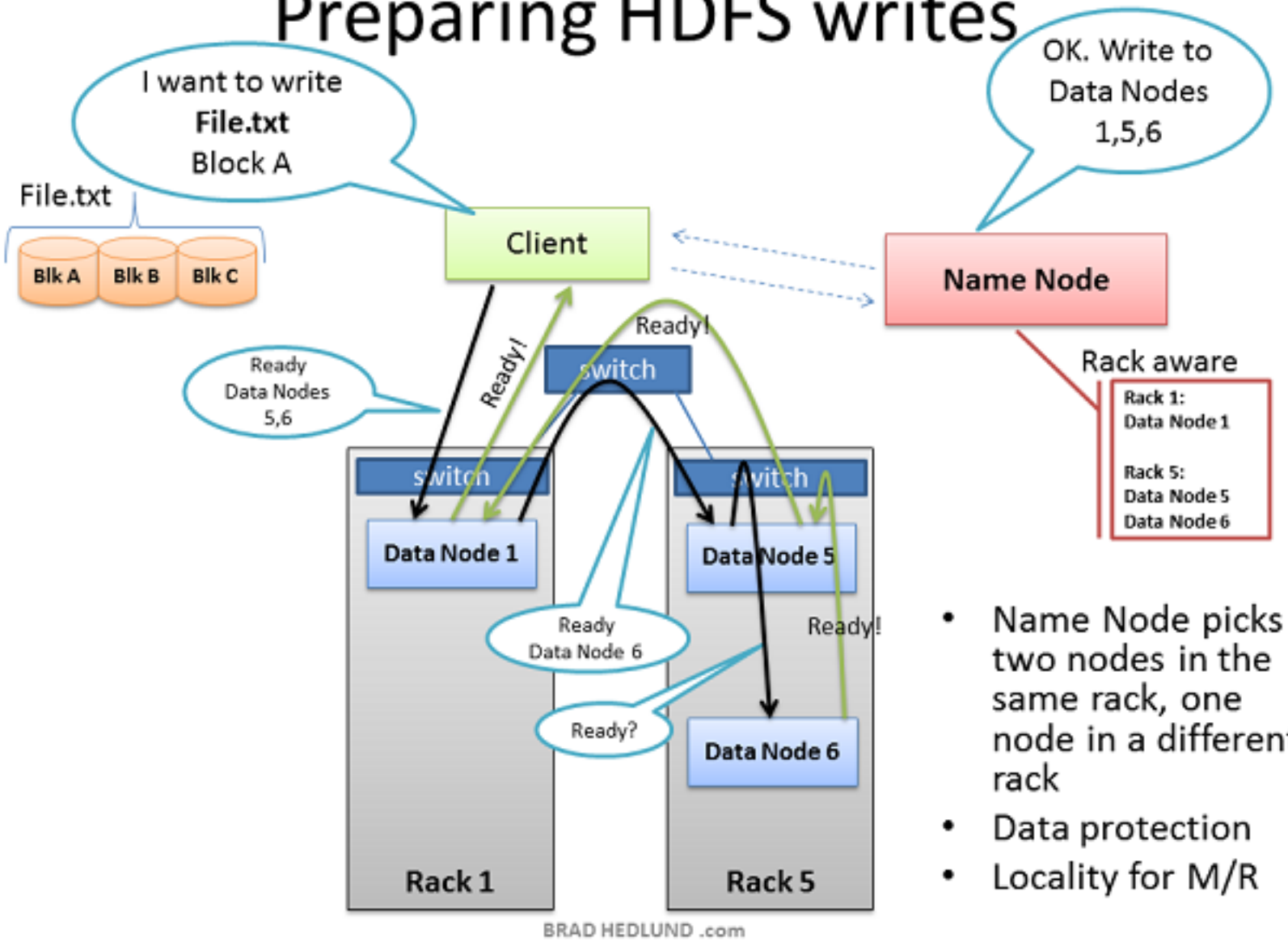
Writing files to HDFS



- Client consults Name Node
- Client writes block directly to one Data Node
- Data Nodes replicates block
- Cycle repeats for next block

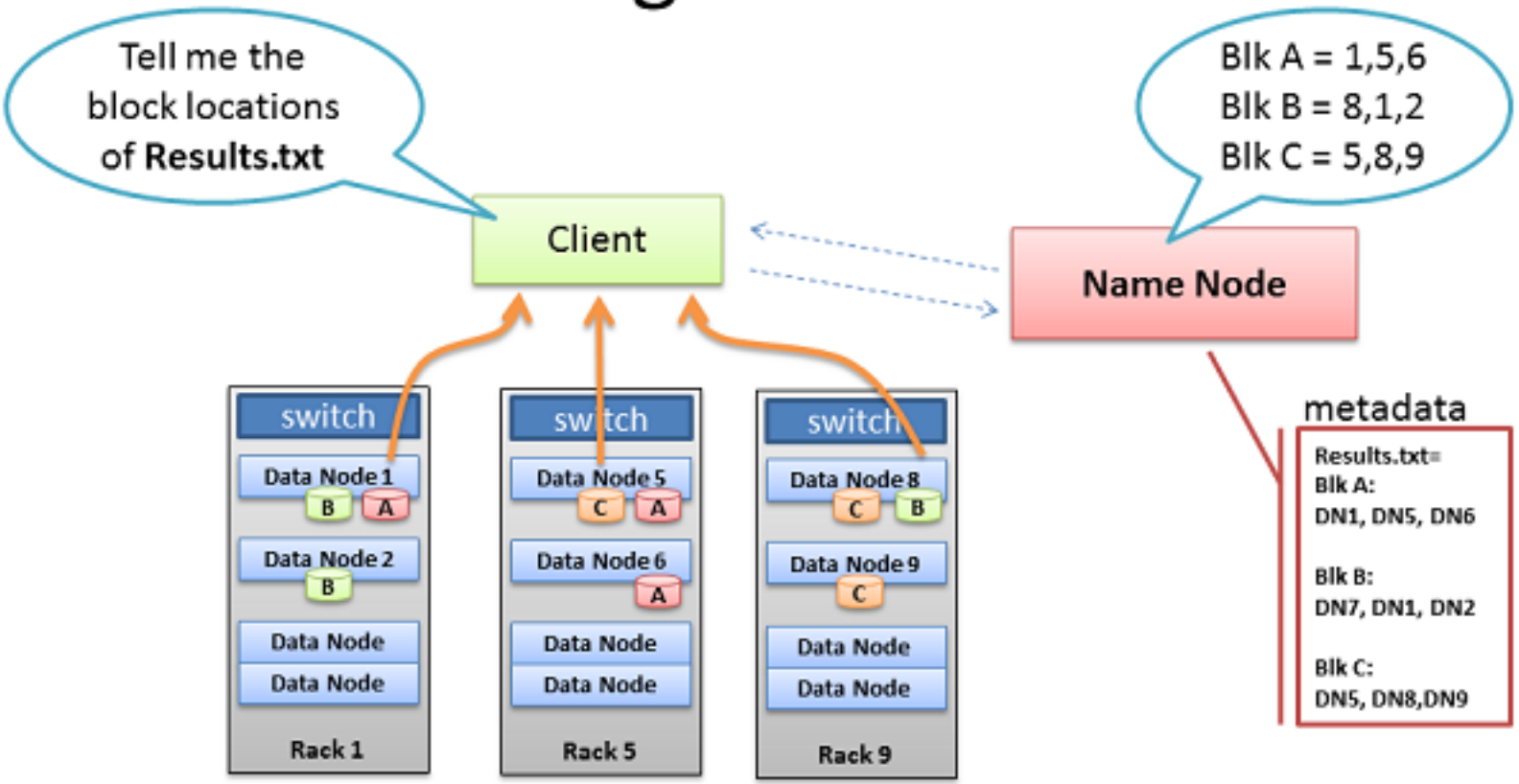
BRAD HEDLUND .com

Preparing HDFS writes



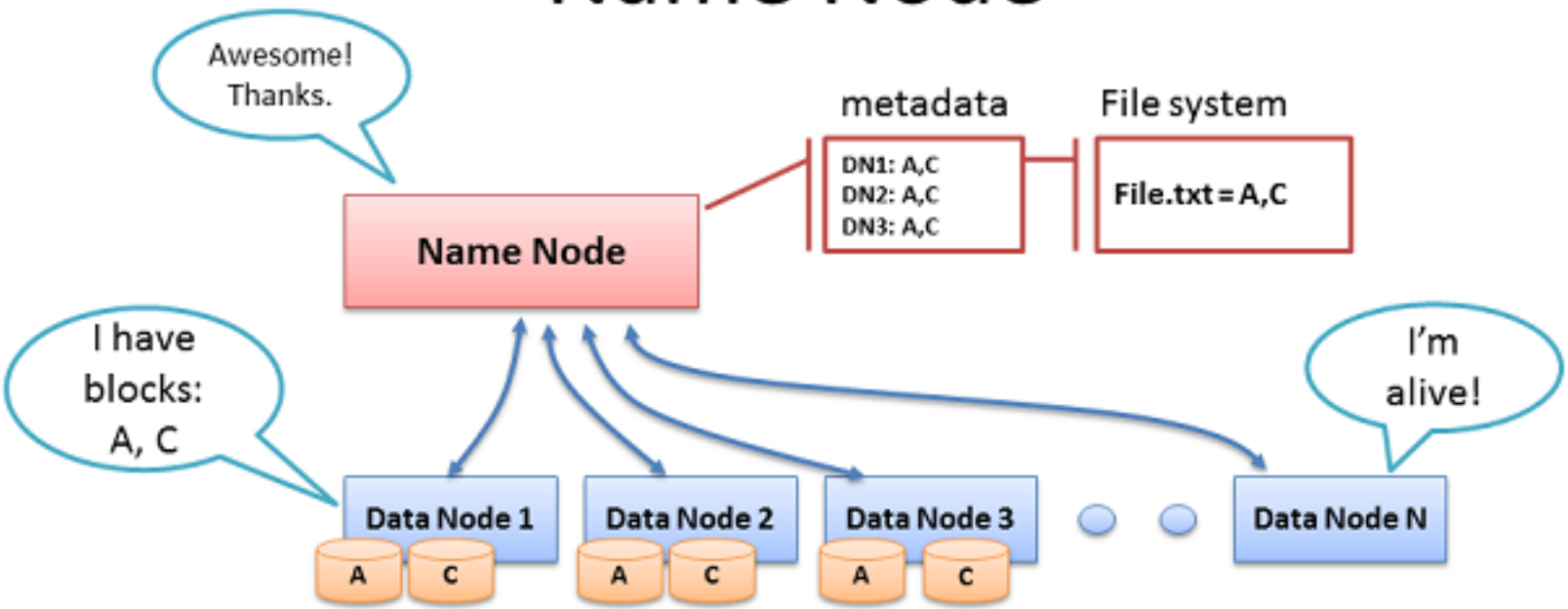
- Name Node picks two nodes in the same rack, one node in a different rack
- Data protection
- Locality for M/R

Client reading files from HDFS



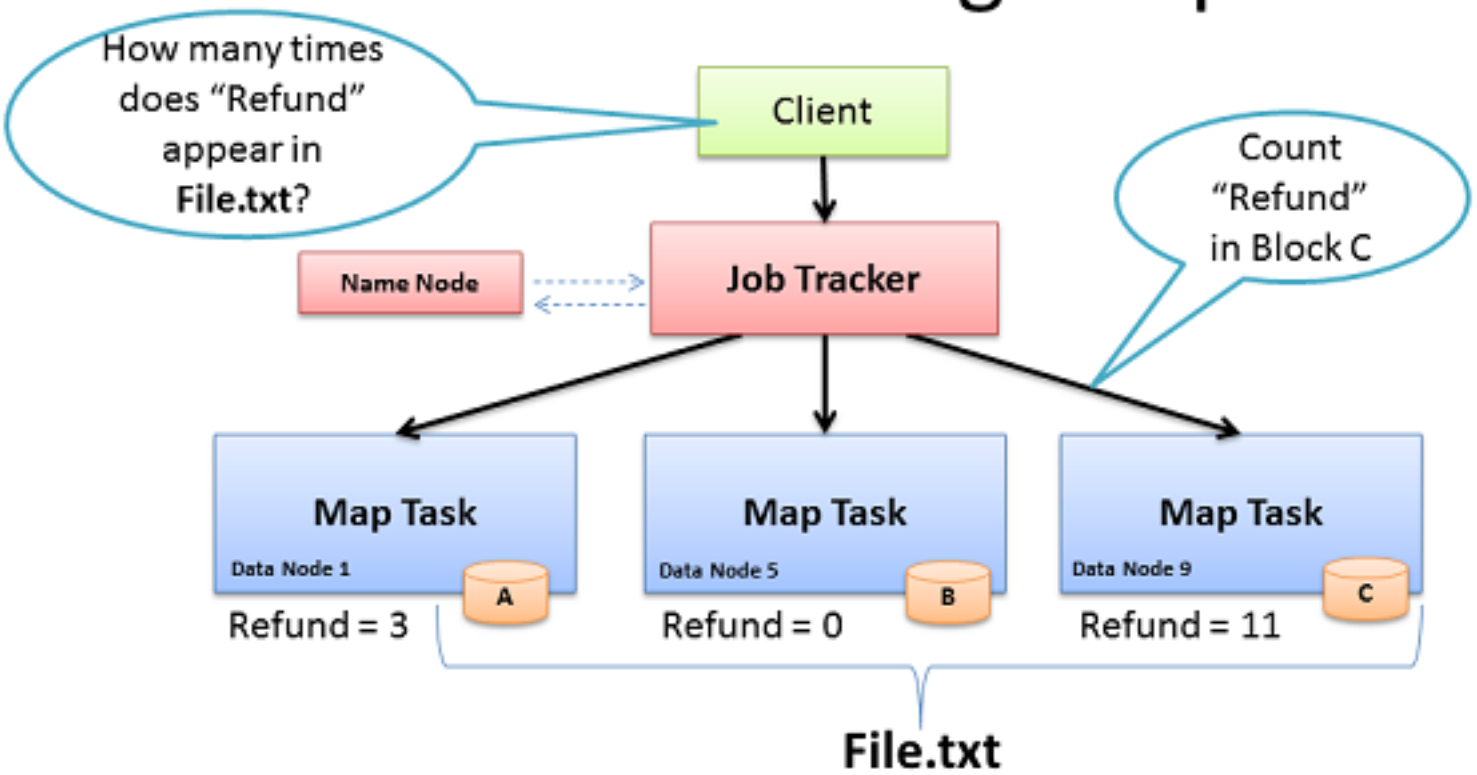
- Client receives Data Node list for each block
- Client picks first Data Node for each block
- Client reads blocks sequentially

Name Node



- Data Node sends Heartbeats
- Every 10th heartbeat is a Block report
- Name Node builds metadata from Block reports
- TCP – every 3 seconds
- If Name Node is down, HDFS is down

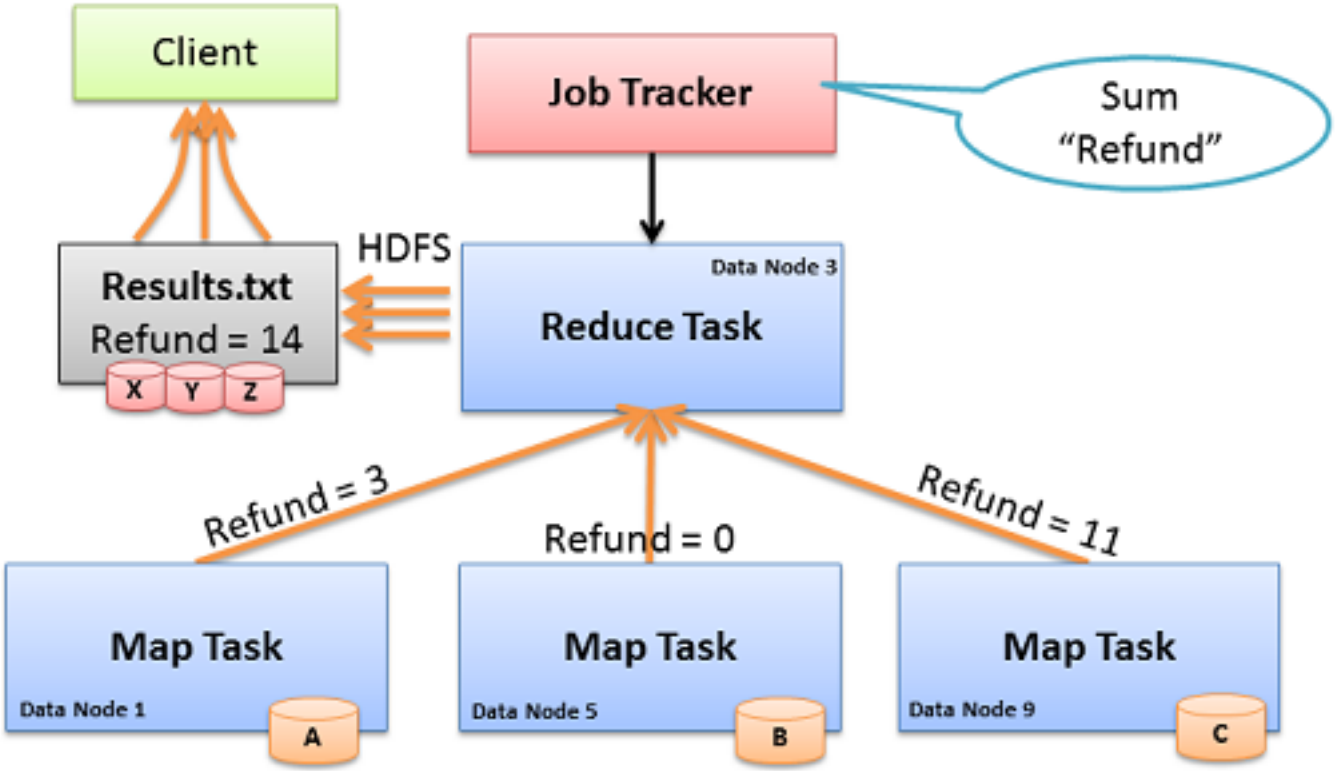
Data Processing: Map



- **Map:** "Run this computation on your local data"
- Job Tracker delivers Java code to Nodes with local data

BRAD HEDLUND .com

Data Processing: Reduce

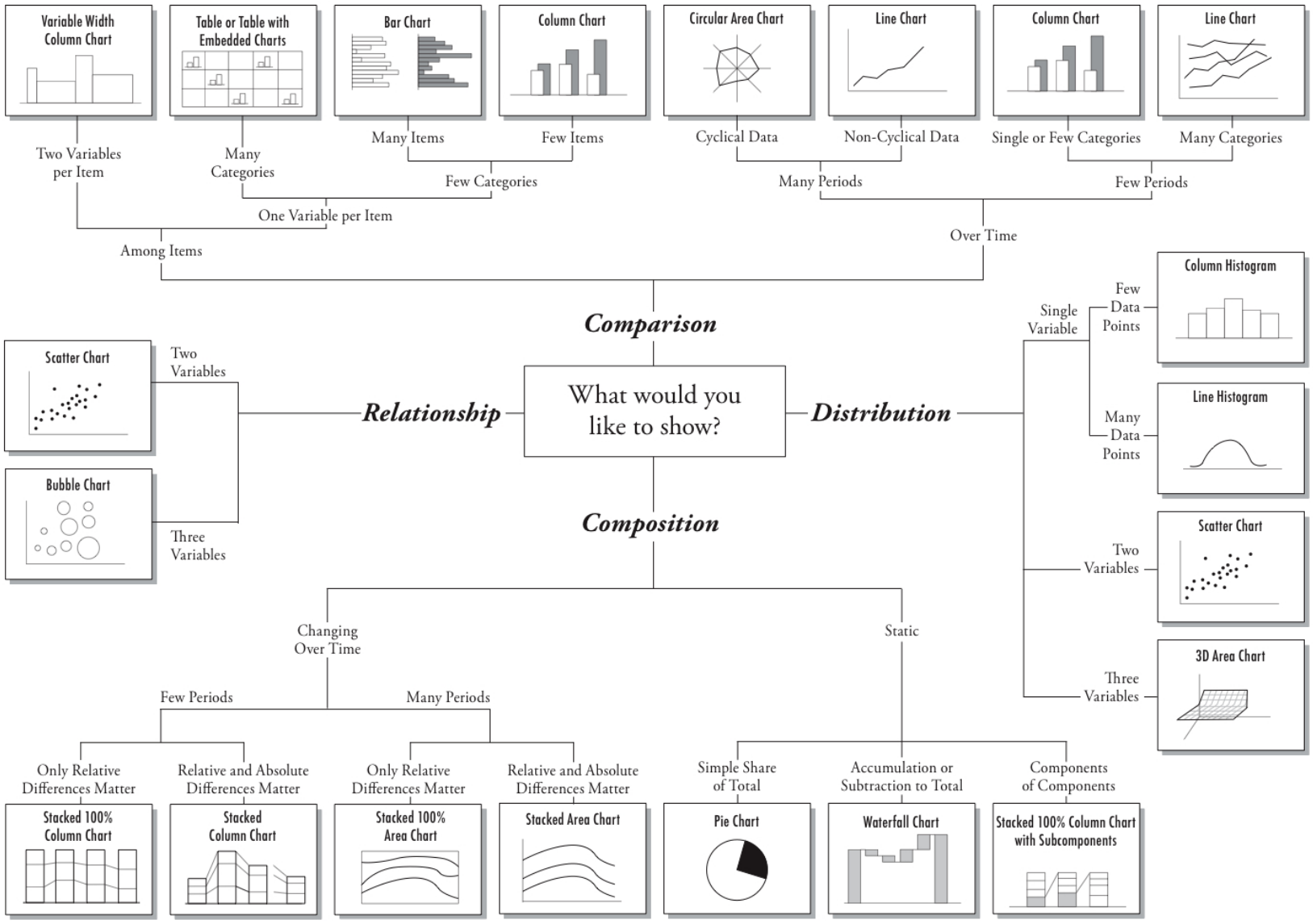


- **Reduce:** "Run this computation across Map results"
- Map Tasks send output data to Reducer over the network
- Reduce Task data output written to and read from HDFS

11. Restitution de données

* Data Viz, Data Designer, infographie
* Outils :
  * http://www.creativebloq.com/infographic/tools-2131971
  * http://www.creativebloq.com/design-tools/data-visualization-712402
  * Java Script, HMTL5, Flash,
* Critères :
  * Comparaisons :
  * Relations : relationships (scatter plots)
  * Composition :
  * Distribution :
  * Statique, dynamique (dans le temps)

Chart Suggestions—A Thought-Starter



12. Working with Data

* Exemple de programmation
  * mapReduce avec le langage C
* Exemple d'utilisation d'un produit basique :
  * Google Fusion Table (experimental google drive app)

* Les techniques présentées sont « scalable »
* Big Data et Small utilise les mêmes recettes

13. Working with langage

* Version Artisanale :
  * comprendre les concepts
  * Traiter les différentes étapes : prepare, analyse, apply
* mapReduce :
  * DIY : standalone app
  * Parallélisation ou pas
  * Utilisation d'API pour les requêtes (API REST)
* bibliothèques tiers :
  * libCurl : création de la commande curl
  * oAuth : gestion de l'authentification (tokens)
  * Jansson : parsage du fichier JSON

13. Working with langage

* Version industrielle
  * Pour travailler vite
  * Les concepts sont déjà acquis
* FrameWork : Java, C# (Hadoop, Mahout)
  * Utilisation d'une plateforme serveur (local, cloud)
  * Réutilisation d'algorithmes : clustering, mapReduce, …
  * Parallélisation automatique (node, clusters, data centers)

14. Working with Google Fusion table

* Solution cloud (Saas)
* Hérite de Google Spread Sheet
* Importation possible en CSV (Google docs, Excel)
* Géolocalisation des données sur Google Map
* Publication automatique sur site Web
* Le Set de données est raisonnable
* Les 3 étapes :
  * Prepare : clean data
  * Analyse : process data
  * Apply : data make decision

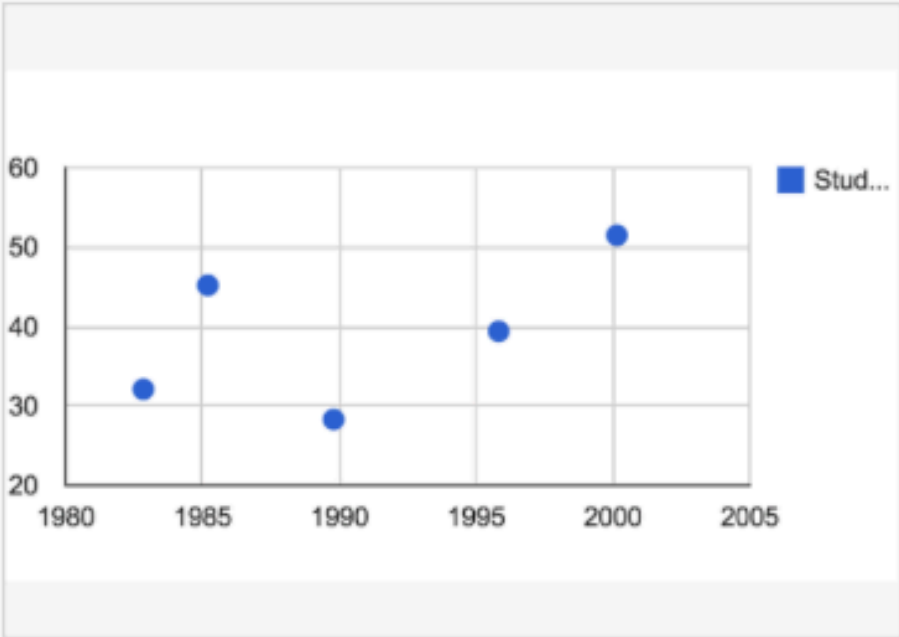
15. Cleansing Data

* Small data set can be cleaned :
  * Remove unnecessary rows and columns
  * Make variable names consistent and informative
  * Format values for readability
  * Handle missing values appropriately
  * Add descriptions and documentation
  * Applying data types to collected data
  * Identifying variables and records in the data
* Attention :
  * Cleansing data reduces noice

Research

* What ask to Data ?
* Searching trend ?

Continuous Variable Charts



When: Variable on the x-axis consists of numbers.

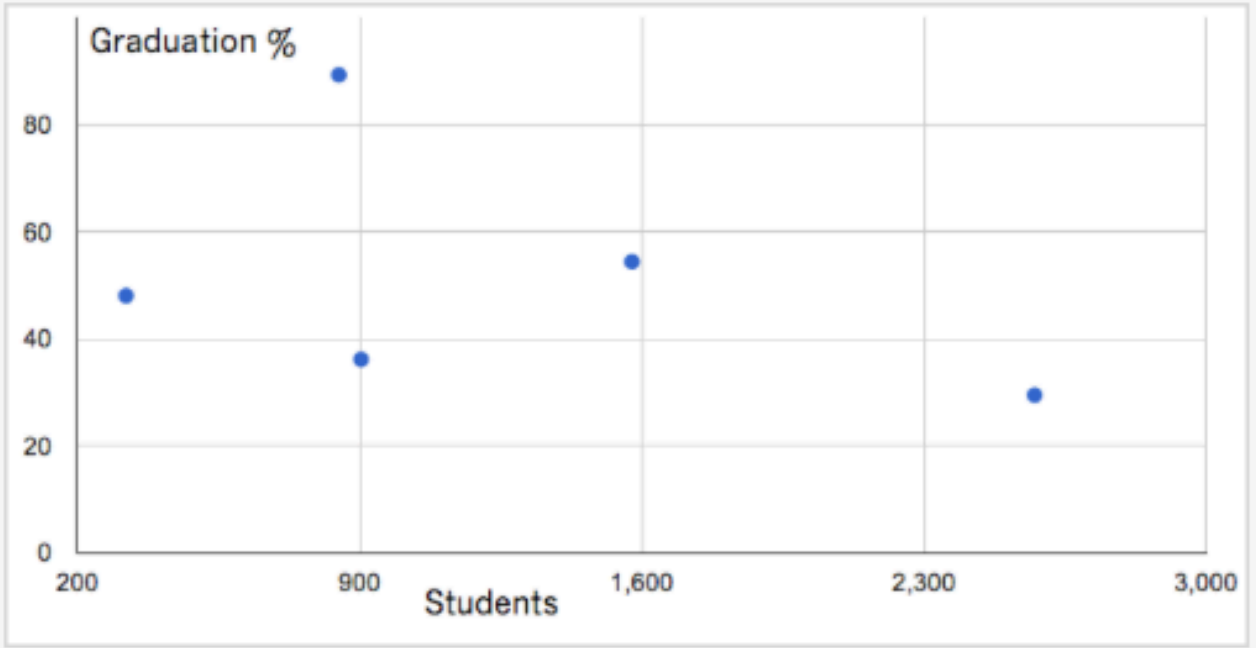
What: Finding mathematical relationships.

How: Looking for patterns between points:

- What is the relationship between variable A and variable B?
- As variable A increases, does variable B a) increase, b) decrease, c) stay the same, or d) change non-linearly?

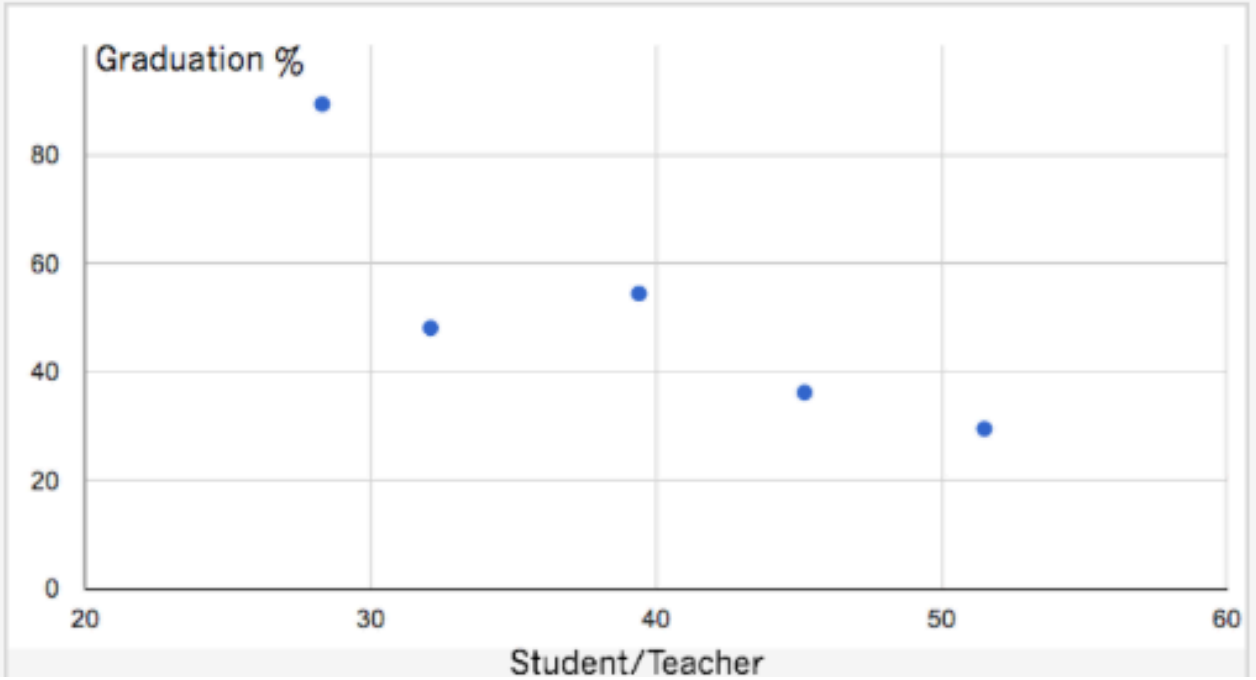
Is there a relationship between number of Students and Graduation percentage?

Looking at the chart, there is no clear relationship between Students and Graduation %.



Is there a relationship between Student/Teacher ratio and Graduation percentage?

In this chart there is a pattern - a line sloping downward.



16. What is a summary

* Summary is a smaller table that describes your data
* Summary = aggregated information (count, sum, maximum, minimum, average).

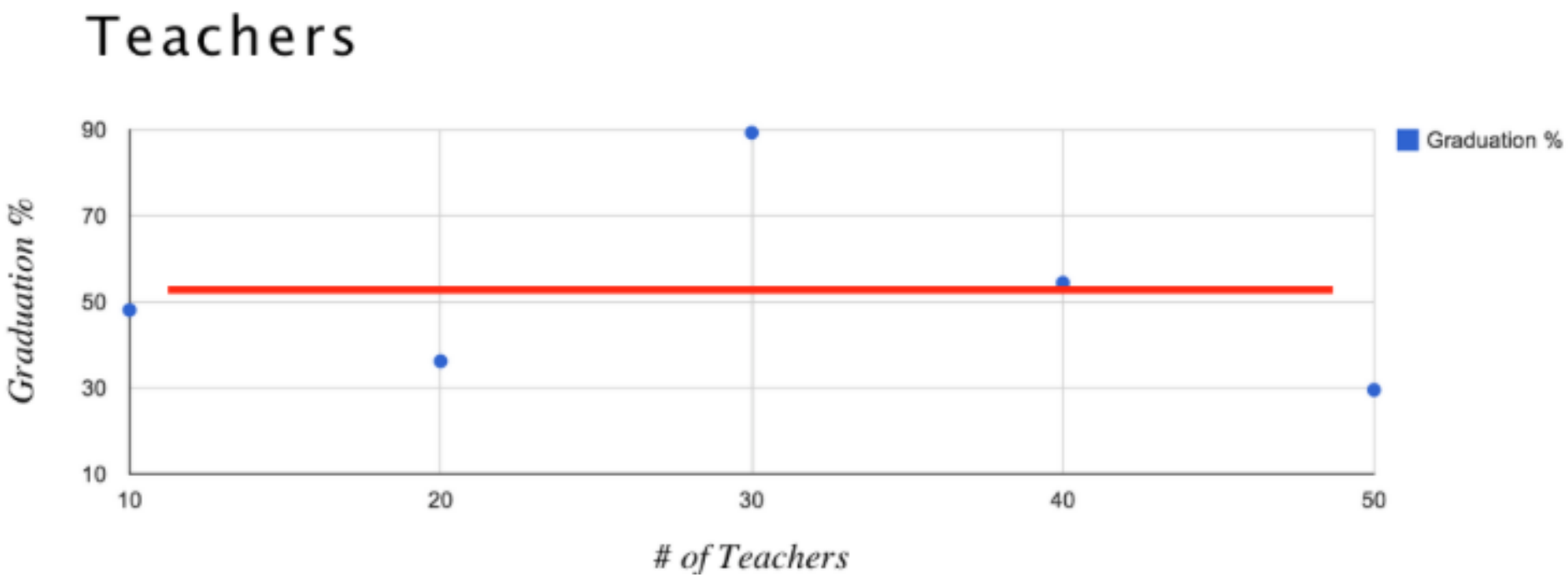
17. What is a filter

* A filter is a statement that is evaluated for each row using values of the columns.
* Filter is like a WHERE clause with SQL.

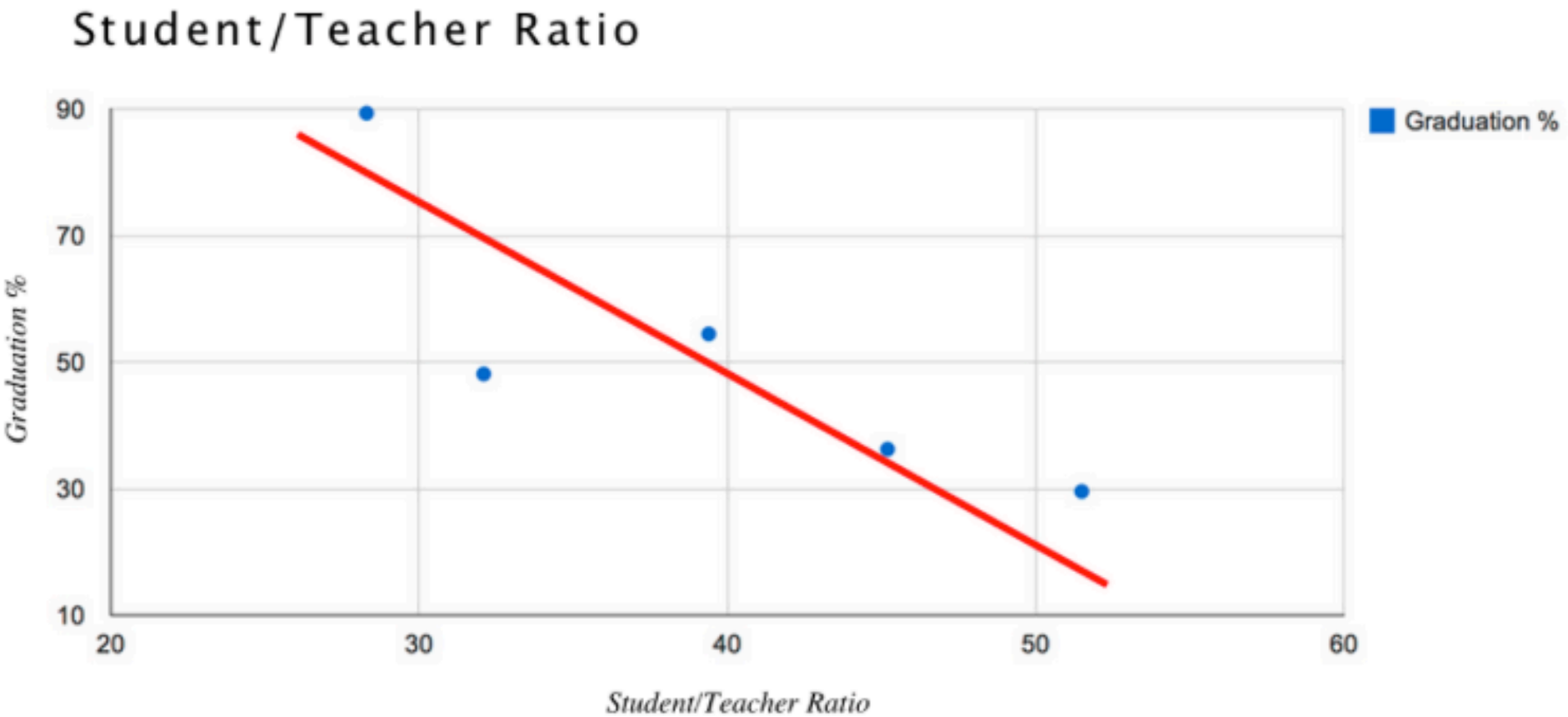
18. Why Merge Data?

* Patterns and relationships are often found in data by comparing variables to one another
* However, your initial table might not contain all of the variables that you want to use for your analysis
* Merging two tables requires that they have a common column
* A common column must have a same type

19. Finding Relationships with Drill-Down Analysis



19. Finding Relationships with Drill-Down Analysis



20. What is Prediction Analysis?

* Prediction analysis is a technique for estimating values that you do not have
* There are a number of reasons why you might not have certain data :
  * The data set is incomplete
  * The data set is a sample that represents a larger data set
  * The data are hypothetical or in the future
* The variable you are estimating is called the **prediction variable**
* the variables used to make the prediction are called **influencing variables**

20. What is Prediction Analysis?

* **Interpolation and Extrapolation**
* **Interpolation** involves estimating values that are within the range you already have
* **Extrapolation** involves estimating values that are outside of the range you already have
* one simple technique for interpolation called **nearest neighbors.**

20. What is Prediction Analysis?

Example

Suppose that a school district is considering opening a new high school called HS 6. The district is concerned about what the graduation percentage will be at HS 6. The district projects a student population of 2,520, and they are considering hiring 60 teachers for HS 6. This means that the **Student/Teacher** ratio would be 42. The district can use prediction analysis to estimate the **Grad. %** at HS 6.

Fictitious School District with High School 6

School	Address	Remodel Date	Student/ Teacher	Grad. %	Seas. of Remodel	Students	Teachers
HS 1	11 1st	11-03-82	32.1	48.11	Fall	321	10
HS 2	22 2nd	03-13-85	45.2	36.22	Winter	904	20
HS 3	33 3rd	10-18-89	28.3	89.33	Fall	849	30
HS 4	44 4th	10-30-95	39.4	54.44	Fall	1576	40
HS 5	55 5th	02-25-00	51.5	29.55	Winter	2575	50
HS 6 (projected)	66 6th		42	??.??		2520	60

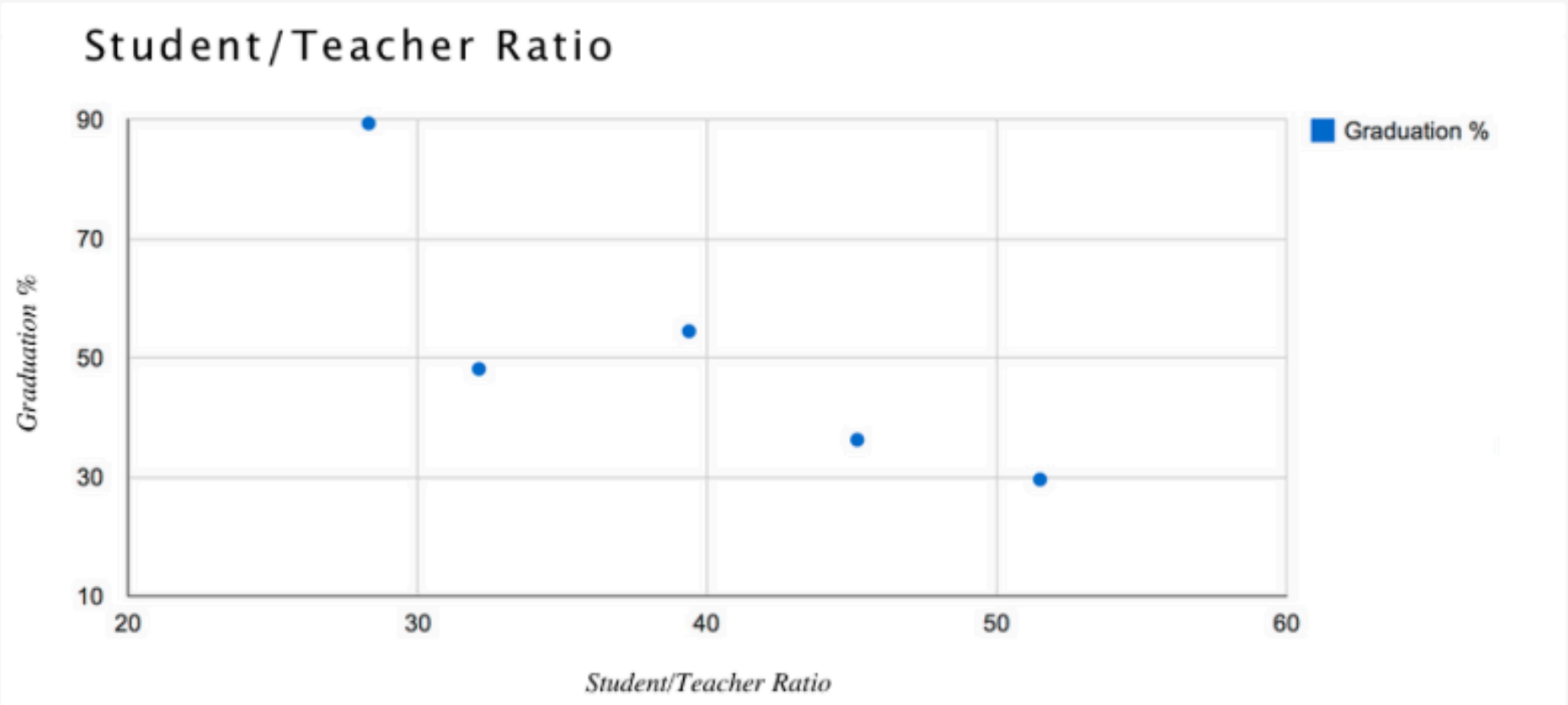
21. Summary of Steps in Prediction Analysis for Numeric Variables

* **Construct a scatter plot of the influencing and prediction variables.** There should be a pattern to the displayed relationship if you truly have an influencing variable for the prediction variable.
* **Mark the location of the value of the influencing variable for the prediction.** Since you can assume that this value lies within the range of measured data, there will be measured values to the left and to the right of the marked value.
* **Select measured values of the influencing variable near the marked value.** For each, find its associated value of the prediction variable.
* **Compute the average of the values of the prediction variable selected in Step 3.** This is the estimated value of the prediction value.

Step 1: Construct a Scatter Plot of the Influencing and Prediction Variables

In the first step of prediction analysis, you construct a scatter plot which provides a graphical display of the relationships between the influencing and prediction variables. The scatter plots should be constructed so that the x-axis is the influencing variable and the y-axis is the prediction variable.

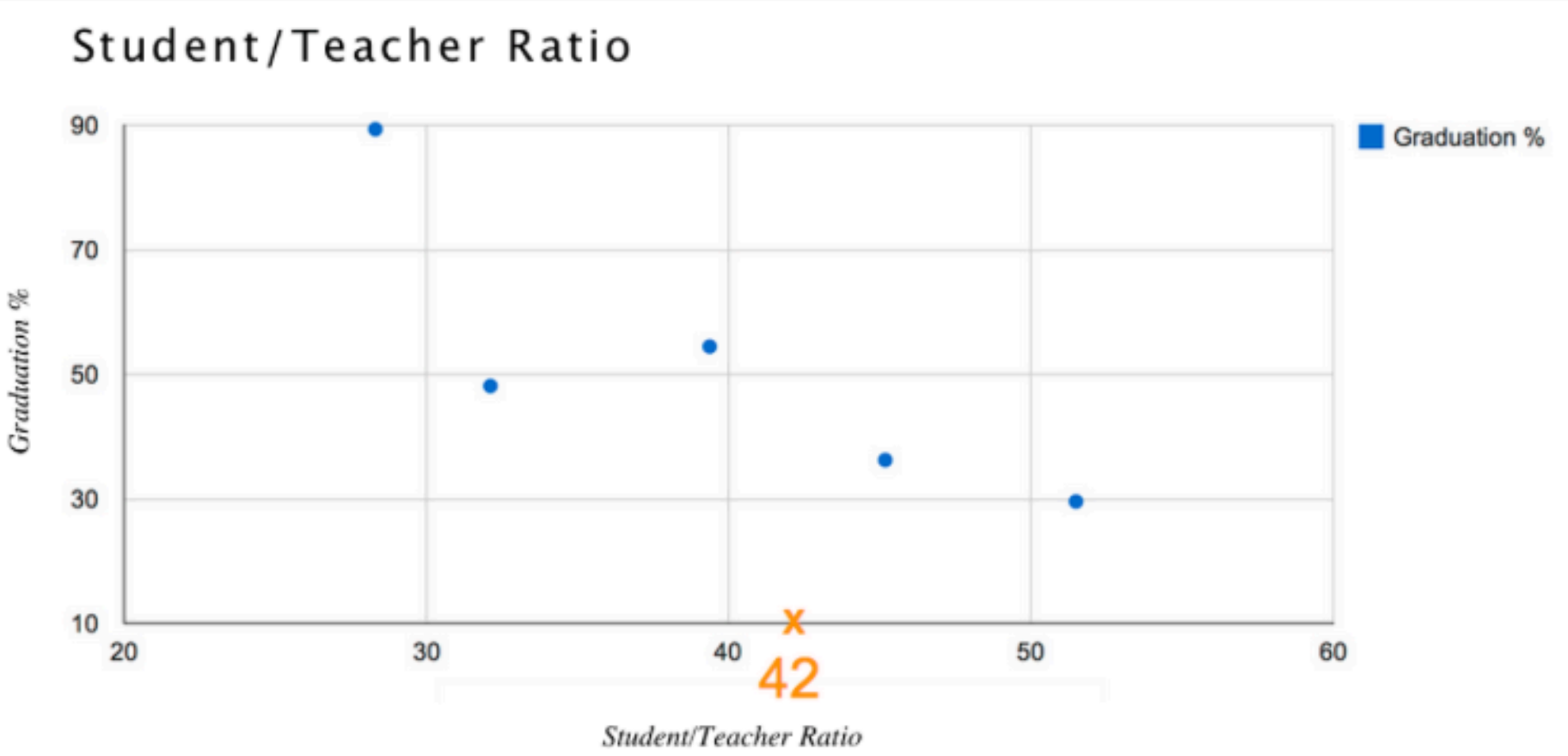
Here is a scatter plot for the data in the example.



Step 2: Mark the Location of the Value of the Influencing Variable for the Prediction

In the second step of prediction analysis, you mark the position where you want to make a prediction. This is done by marking on the x-axis the value of the influencing variable you want to predict.

To illustrate this step, the scatter plot below is marked at a value of 42 on the x-axis.

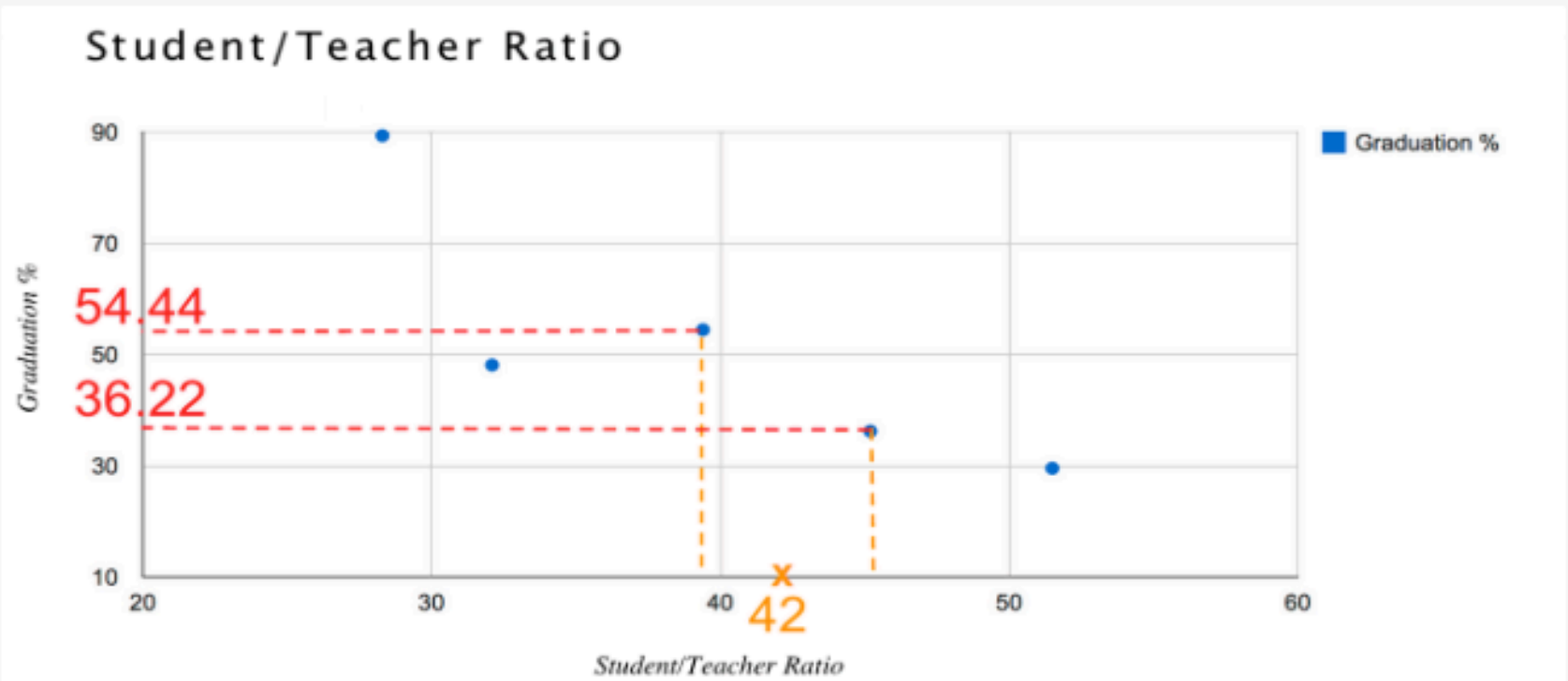


Step 3: Select Measured Values of the Influencing Variable Near the Marked Value

In the third step of prediction analysis, you find measured values of the prediction variable that you expect will be very close to the the unknown value to be estimated. In statistics, this is called a **nearest neighbors** technique.

Using the **nearest neighbors** technique, you find measurements where the value of the influencing variable is close to what you want to estimate. The values of the associated prediction variable should be close to the unknown value you want to estimate.

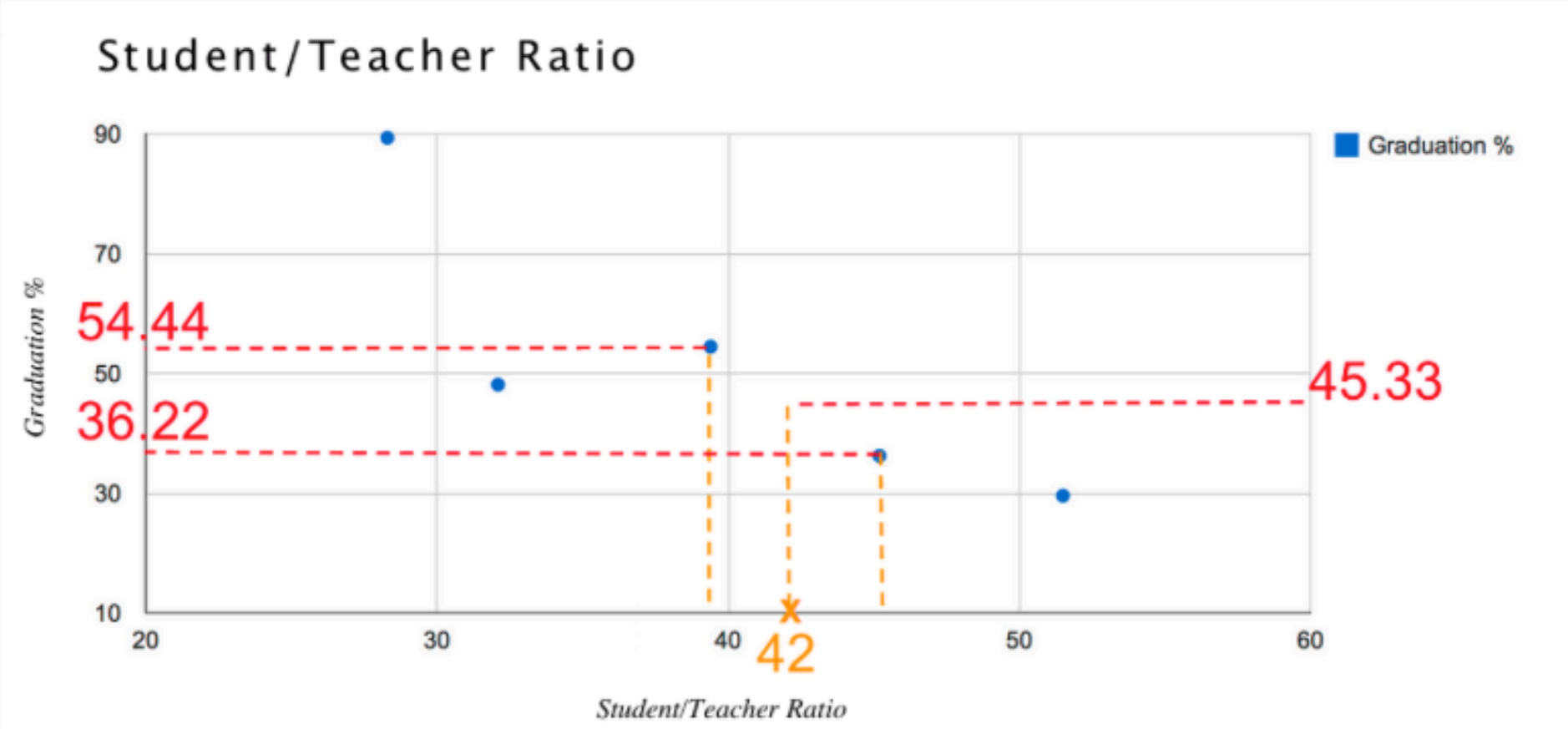
Locate points in the scatter plot that have values on the horizontal access that are close to the one that has been marked. In the example below, two points are selected. One point has a value just to the left of 42, and the other is just to the right. Next, find the values of the prediction variable associated with the selected points. In the example below, these values are 54.44 and 36.22.



Step 4: Compute The Average of the Selected Values of the Prediction Variable

In the fourth step of prediction analysis, you construct the estimate of the unknown value by finding the average of the values of the prediction variable that you obtained in Step 3.

In the example below, you compute the average of **54.44** and **36.22** to obtain the estimated value of **45.33** as the unknown value of the prediction variable.



22. Prediction with Non-Numeric Influencing Variables

* The **nearest neighbors** method is useful when your influencin t non-numeric ir

Grad % and Season of Remodel

School	Grad. %	Season of Remodel
HS 1	48.11	Fall
HS 2	36.22	Winter
HS 3	89.33	Fall
HS 4	54.44	Fall
HS 5	29.55	Winter
HS 6	??	Fall

22. Prediction with Non-Numeric Influencing Variables

* First step :

Average Grad % per Season of Remodel

Season of Remodel	AVG. Grad. %
Fall	63.96
Winter	32.88

* Since the average graduation percentage for Fall remodels is **63.96,** we can use this value to estimate the graduation percentage for remodeling HS 6. Note however that the number obtained in this analysis is quite different than the number obtained when using the **Student/Teacher** variable to estimate the **Grad** % for HS 6.
* Season of remodel is not a good criteria prefer numeric influencing variables

23. Locating Patterns with Geospatial Analysis

* Heatmaps are a common method for visualizing numeric data on a map
* Heatmaps use variations in color to convey a range of values
* Green is used for small values, yellow is used for medium values, and red is used for large values.
* heat maps help you visualize continuous changes in a variable
* buckets allow you to find patterns when ranges of values are grouped together.

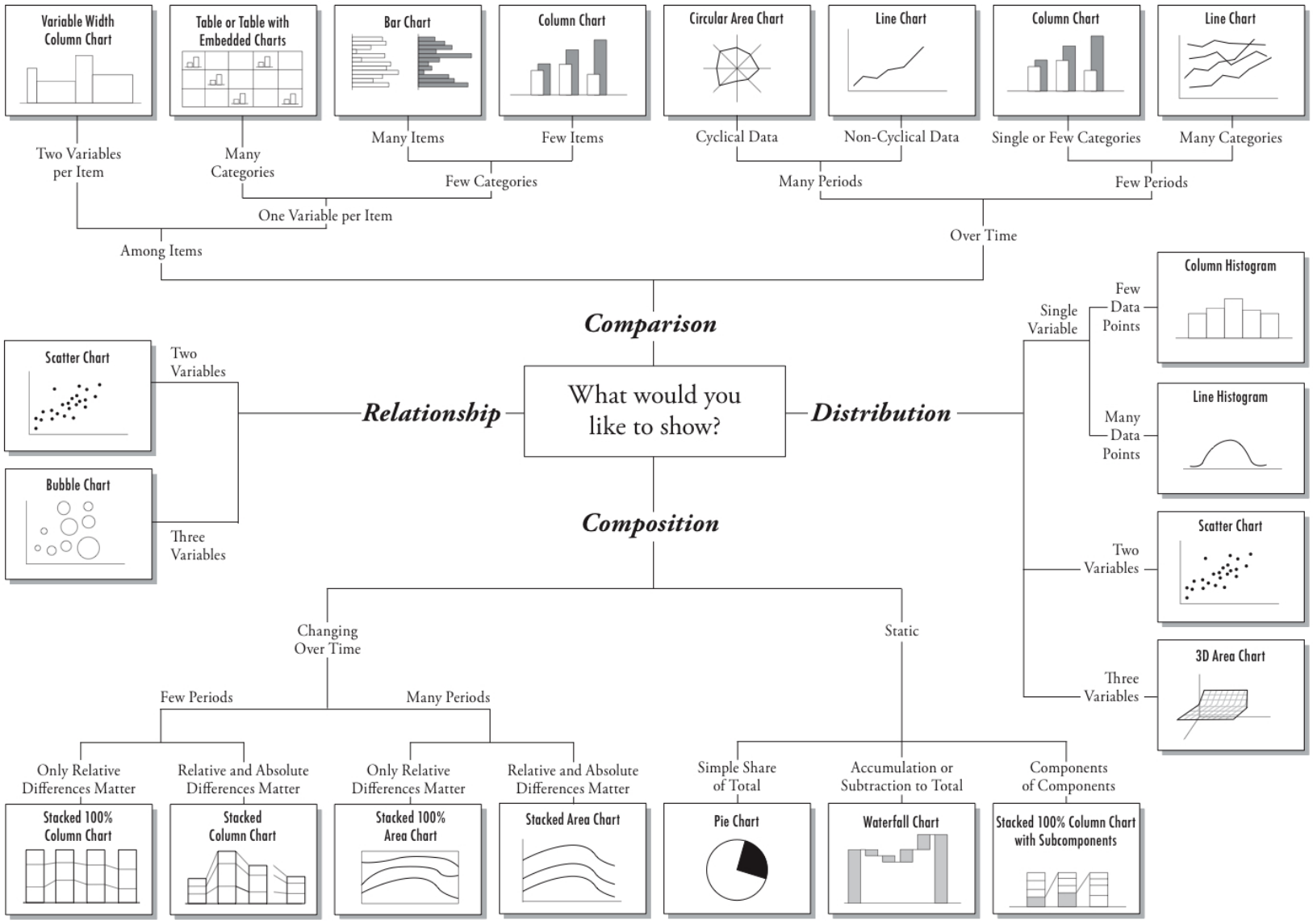
24. Sharing Analysis Results

* Choose a visualization that illustrates your analysis.
* Assess the effectiveness of visualizations.
* Create a view that allows you to share only a portion of your data.
* Point others to your table or view so that you can share data with them.

25. Choose a Format to Illustrate Your Analysis

* **Views**
  * You can create a view (a subset of the table) to share with different audiences who can examine and interact with the pieces of data they find most relevant
* **Charts**
  * Charts and graphs help illustrate key trends and findings, and can highlight the important parts of vast quantities of data.
* **Summaries**
  * Summaries can easily display sums and averages that are of interest to your target audience

Chart Suggestions—A Thought-Starter



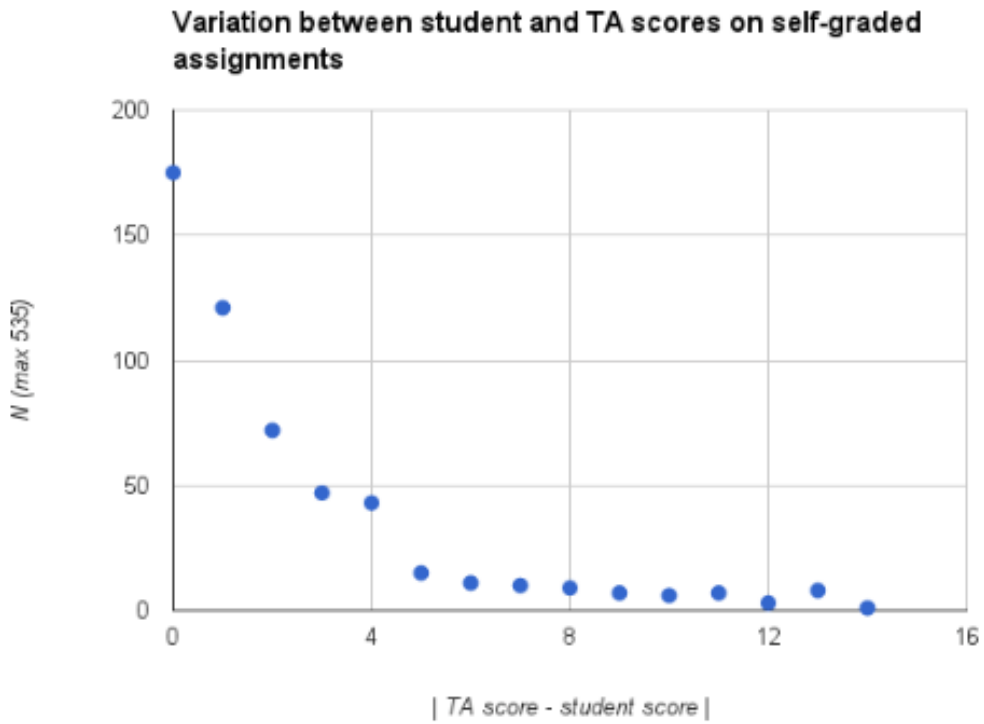
25. Assess the Effectiveness of a Visual

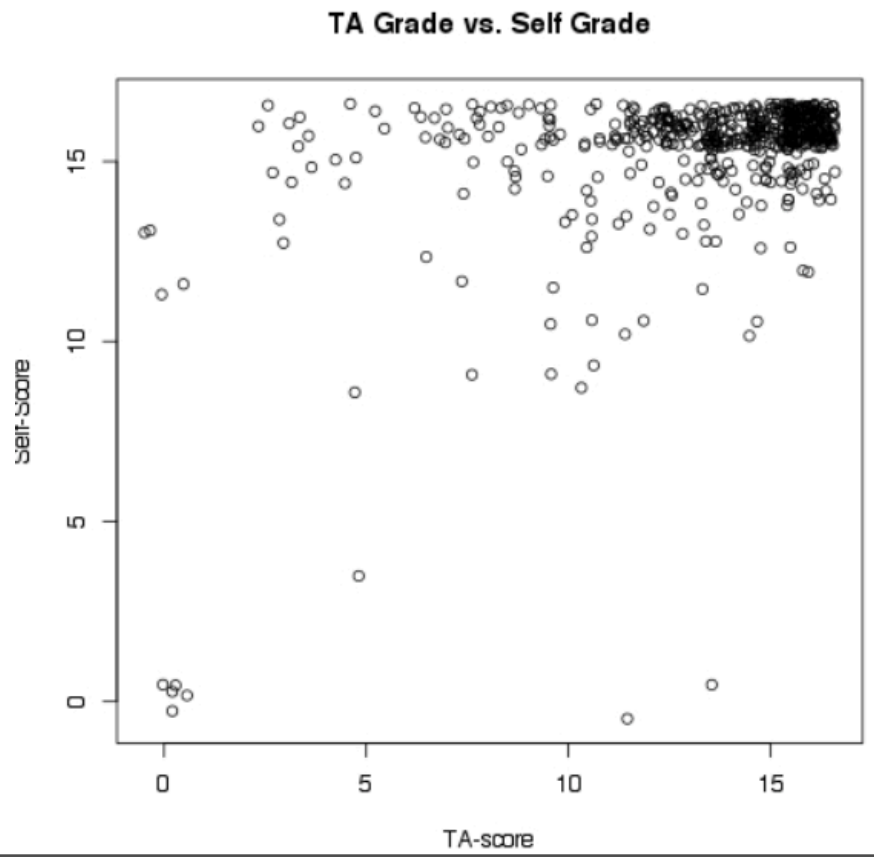
* Here are some important elements to consider when evaluating how well your visual communicates the story you are trying to tell:
* Does the visual (chart or table) have a title?
* Do the variables include labels?
* Are axes labeled?
* Are units clear?
* Do colors enhance the visual or detract from it?

25. Assess the Effectiveness of a Visual

GOOD

BAD





26. Assess the Effectiveness of a Visual

* Information Visualization MOOC offered by Indiana University : http://ivmooc.cns.iu.edu
* Google example: preattentive attributes by Cole Nussbaumer : http://www.storytellingwithdata.com/2011/10/google-example-preattentive-attributes.html
* Periodic Table of Visualization methods : http://www.visual-literacy.org/periodic_table/periodic_table.html
* Choosing the right kind of chart : http://apandre.wordpress.com/dataviews/choiceofchart/

27. 30+ free tools for dataViz and

Data visualization and analysis tools

Tool	Category	Multi-purpose visualization	Mapping	Platform	Skill level	Data stored or processed	Designed for Web publishing?
IBM Word-Cloud Generator	Word clouds	No	No	Desktops running Java	2	Local	As image
Google Fusion Tables	Visualization app/service	Yes	Yes	Browser	1	External server	Yes
Impure	Visualization app/service	Yes	No	Browser	3	Varies	Yes
Many Eyes	Visualization app/service	Yes	Limited	Browser	1	Public external server	Yes
Tableau Public	Visualization app/service	Yes	Yes	Windows	3	Public external server	Yes
VIDI	Visualization app/service	Yes	Yes	Browser	1	External server	Yes
Zoho Reports	Visualization app/service	Yes	No	Browser	2	External server	Yes
Weave	Visualization app/service	Yes	Yes	Flash-enabled browsers; Linux server on backend	4	Local or external server	Yes
Statwing	Visualization app/service	Yes	No	Browser	1	External server	Not yet
Infogr.am	Visualization app/service	Yes	Limited	Browser	1	External server	Yes
Datawrapper	Visualization app/service	Yes	No	Browser	1	Local or external server	Yes
Jolicharts	Visualization app/service	Yes	Yes	Browser	1	External server	Yes
Silk	Visualization app/service	Yes	Yes	Browser	1	External server	Yes
Chartbuilder	Visualization app/service	Yes	No	Browser	1	Local	Yes
Plotly	Visualization app/service	Yes	No	Browser	1	External server	Yes

30+ free tools for dataViz and

TimeFlow	Temporal data analysis	No	No	Desktops running Java	1	Local	No
R Project	Statistical analysis	Yes	With plugin	Linux, Mac OS X, Unix, Windows XP or later	4	Local	No
Gephi	Network analysis	No	No	Desktops running Java	4	Local	As image
NodeXL	Network analysis	No	No	Excel 2007 and 2010 on Windows	4	Local	As image
Google Chart Tools	Library and Visualization app/service	Yes	Yes	Code editor and browser	2	Local or external server	Yes
Exhibit	Library	Yes	Yes	Code editor and browser	4	Local or external server	Yes
JavaScript InfoVis Toolkit	Library	Yes	No	Code editor and browser	4	Local or external server	Yes
D3	Library	Yes	Yes	Code editor and browser	4	Local or external server	Yes
Highcharts*	Library	Yes	No	Code editor and browser	3	Local or external server	Yes
Cascading Tree Sheets	Library	Yes	Yes	Browser	1	Local or external server	Yes
Dataset	Library	No	No	Browser	4	Local or external server	Yes
Leaflet	Library	No	Yes	Browser	4	Local or external server	Yes
Searchable Fusion Table Map Template	Library	No	Yes	Browser	3	Local or external server	Yes
Tabletop	Library	No	No	Browser	3	Local or external server	Yes

30+ free tools for dataViz and

OpenLayers	GIS/mapping: Web, Library	No	Yes	Code editor and browser	4	local or external server	Yes
OpenHeatMap	GIS/mapping: Web	No	Yes	Browser	1	External server	Yes
OpenStreetMap	GIS/mapping: Web	No	Yes	Browser or desktops running Java	3	Local or external server	Yes
Quantum GIS (QGIS)	GIS/mapping: Desktop	No	Yes	Linux, Unix, Mac OS X, Windows	4	Local	With plugin
eSpatial	GIS/mapping	No	Yes	Browser	2	External	Yes
Choosel	Framework	Yes	Yes	Chrome, Firefox, Safari	4	Local or external server	Not yet
MicroStrategy Analytics Desktop	Desktop application	Yes	No	Windows	3	Local	Yes
Mr. Data Converter	Data reformatting	No	No	Browser	1	Local or external server	No
Data Wrangler	Data cleaning	No	No	Browser	2	External server	No
OpenRefine (formerly Google Refine)	Data cleaning	No	No	Browser	2	Local	No
Data Explorer**	Data acquisition, data reformatting	No	No	Excel 2010 and 2013 on Windows	2	Local	No
CSVKit	CSV file analysis	No	No	Linux, Mac OS X or Linux with Python installed	3	Local	No
DataTables	Create sortable, searchable tables	No	No	Code editor and browser	3	Local or external server	Yes
FreeDive	Create sortable, searchable tables	No	No	Browser	2	External server	Yes
Panda Project	Create searchable tables	No	No	Browser with Amazon EC2 or Ubuntu Linux	2	Local or external server	No
PowerPivot**	Analysis and charting	Yes	No	Excel 2010 and some 2013 versions on Windows	3	Local	No