Tutorial, IEEE SERVICE 2014 Anchorage, Alaska

Big Data Science: Fundamental, Techniques, and Challenges (Data Science on Big Data)

2014. 6. 27.

By Ryoji Sawa

Presented by Incheon Paik

University of Aizu

Japan

Contents

- What is Big Data and Data Science?
- Why Big Data?
- Potential Applications
- Data Science Process

What is Big Data?

Dataset which is too large for traditional data processing systems.

3Vs: three main differences from `data` or `data analytics`.

- Volume

  As of 2012, about 2.5 exabytes (2.5 billion gigabytes) of data are created each day. The number is doubling about every 40 years. (Harvard Business Review Oct. 2012)

  90% of the data in the world today has been created in the last two years. (IBM ``What is big data?'')

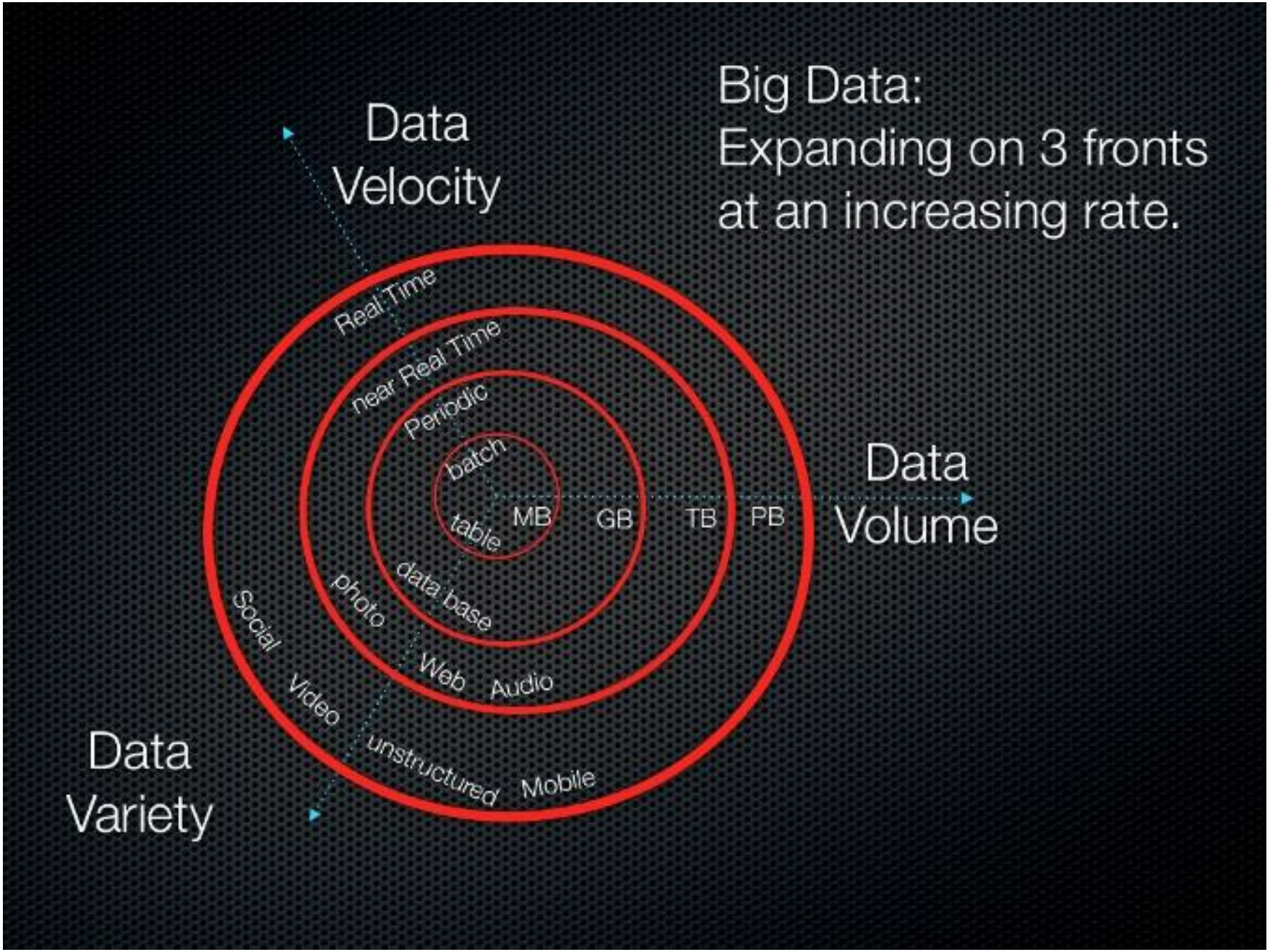
- Velocity

  Real-time information makes it possible for a company to be more agile than its competitors.

- Variety

  Big data takes the form of:

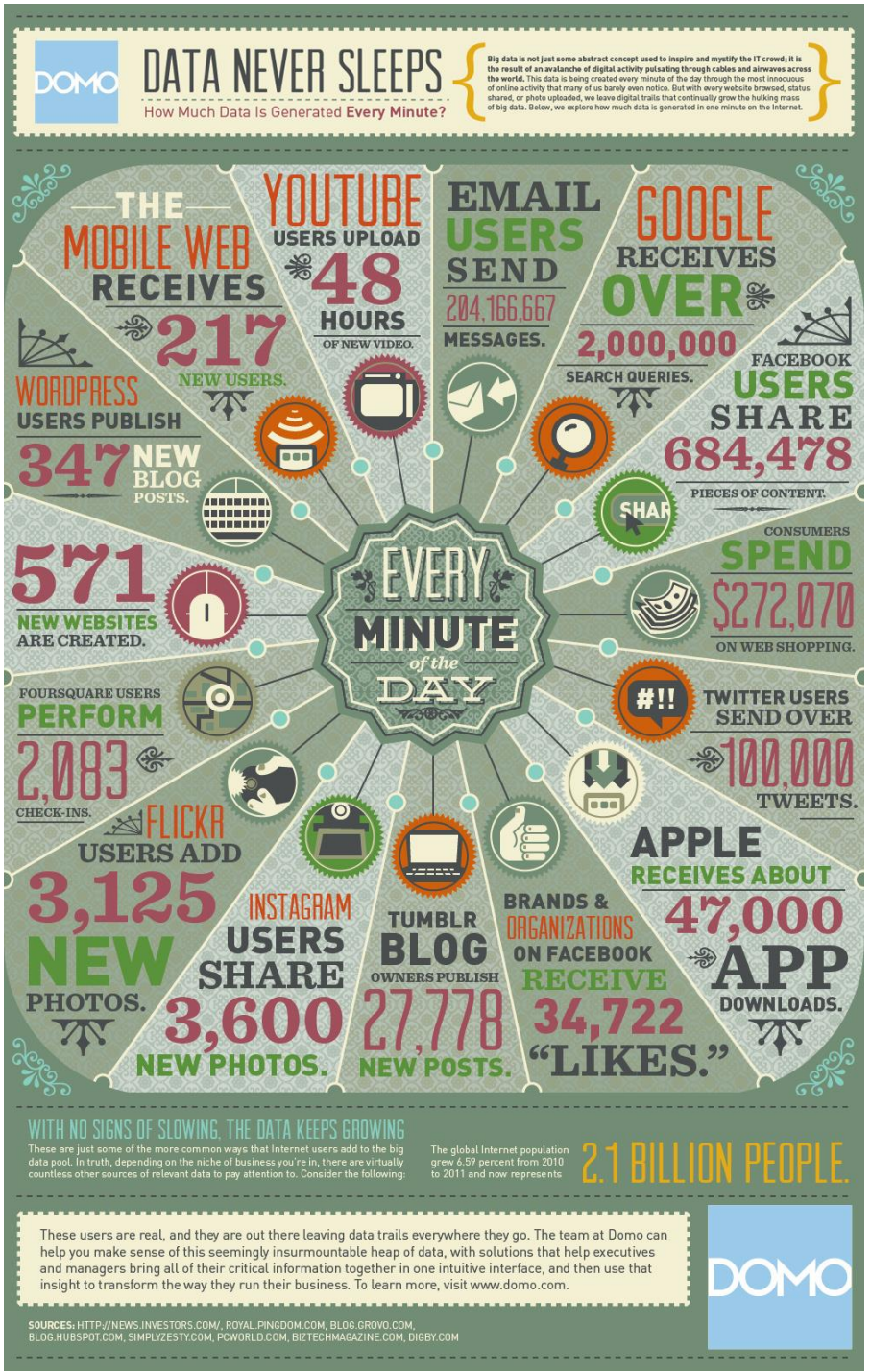
  - Messages, updates and images posted to social networks,
  - Readings from sensors,
  - GPS signals from cell phones, and etc.


Big Data:
Expanding on 3 fronts
at an increasing rate.
Data
Velocity
Real-Time
near Real Time
Periodic
batch
table
MB
GB
TB
PB
Data
Volume
data base
photo
Social
Web
Audio
Video
unstructured
Mobile
Data
Variety

How much data
is created?

An estimation by DOMO, Inc.

Data variety
Text: google, twitter, facebook
Images: Facebook, instagram
Movies: youtube



http://www.domo.com/blog/2012/06/how-much-data-is-created-every-minute/

◆ Why Big Data?

Make use of Big Data to make better decisions.

The analysis of data will enable us to

- Predict whether something will happen.
- Predict how much something will happen. ⇒ **Decisions**
- Group items by their similarity.

If data helps us, Big Data should help us more.

Some evidence that it actually helps.

Companies in the data-driven decision making:

5% more productive, and 6% more profitable than competitors.

(HBR Oct. 2012)

How it helps?

- Provide accurate information. → We can make an optimal choice.
- Grouping similar items. → We can use the same strategy for similar items.

Example (Predicting time of arrival):

Accurate information about flight arrival times matters. The ground staff needs to be ready for a plane landing. Inaccurate information is very costly.

Their duties include the handling of the baggage, stocking the plane with the refreshments, and cleaning the plane between flights.

Ground staff scheduling is important. It can be a topic of a Ph.D. thesis…

("Airport Ground Staff Scheduling", T. Clausen, DTU Management Engineering, 2011.)

Inaccurate estimations on arrival times will waste your perfect scheduling.

Example (Predicting time of arrival):

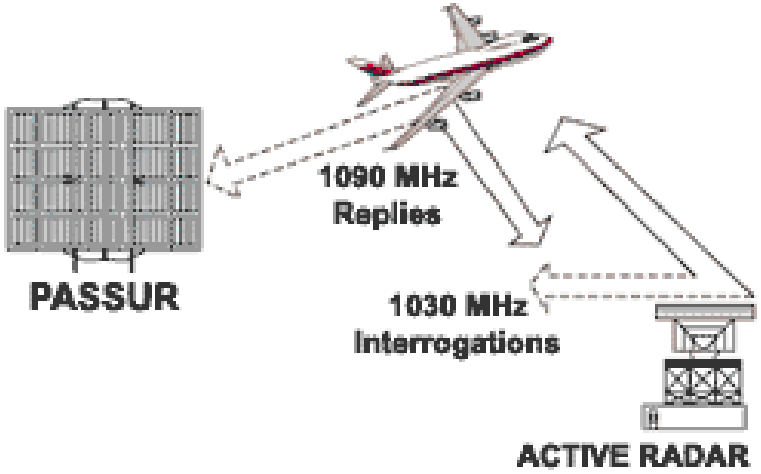
Accurate information about flight arrival times matters. The ground staff needs to be ready for a plane landing. Inaccurate information is very costly.

PASSUR (www.passur.com) offers its own arrival estimates as a service.

They use …

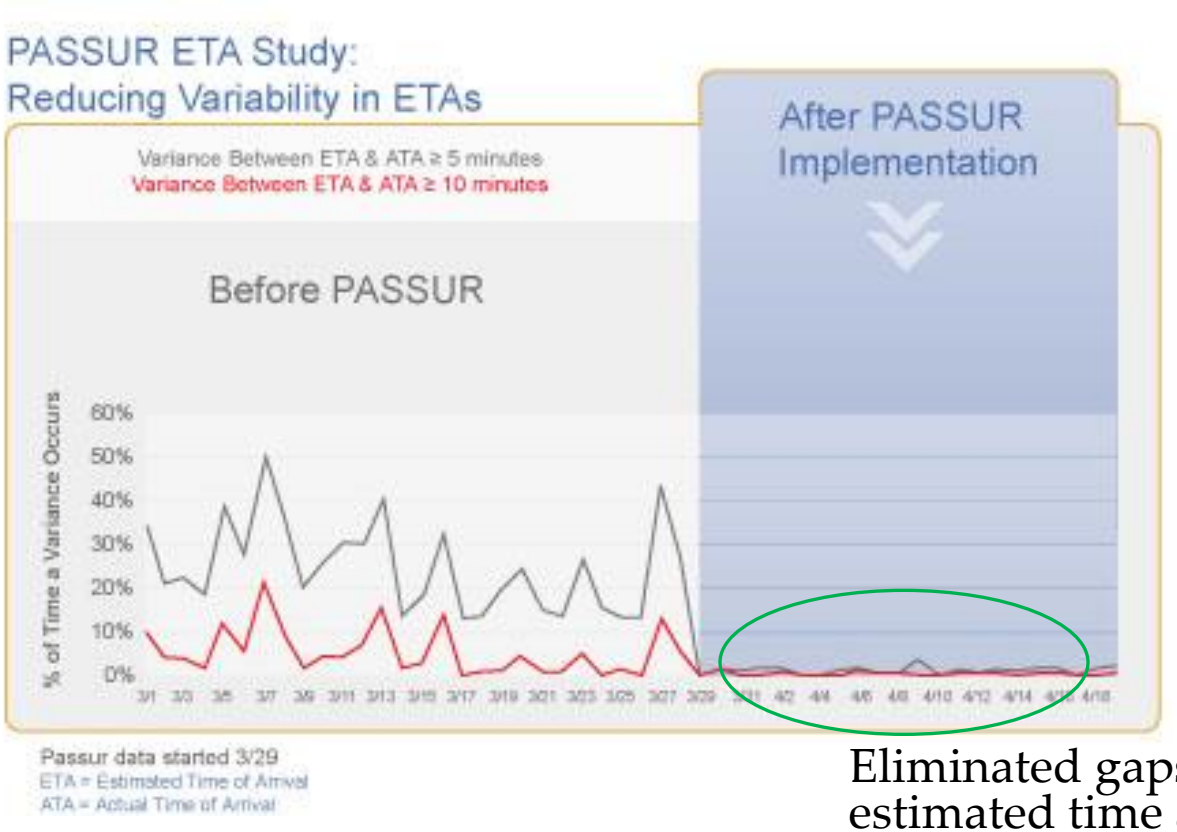
- Public data, e.g. weather,
- Proprietary sensor data (from a network of passive radar stations).

PASSUR piggybacks off the FAA's radar by borrowing the timing signal from 1030MHz interrogations. It shares the omnidirectional 1090MHz transponder replies and calculates aircraft position by triangulation based upon the infinitesimal time delay when the replies are received at the antenna.



http://www4.passur.com/What_is_Passur_How_does_it_work.html

% of time misestimating occurs



Eliminated gaps between estimated time and actual time

http://www.passur.com/righteta-product-detail.htm

Example (Grouping customers):

Ira Haimowitz and Henry Schwartz (1997) show an example of how clustering was used to improve decisions about how to set credit lines for new credit customers.

Data: existing GE Capital customers' use of their cards, payment of their bills, and profitability to the company.

> About GE Capital:
> GE Capital is a financial service unit of General Electric.
> One of their services is offering credit cards.
> In Japan, it is known as a former owner of Lake. (It already sold Lake.)

Haimowitz and Schwartz clustered those GE Capital customers based on similarity.

They settled on five clusters that represented very different consumer credit behavior (e.g., those who spend a lot but pay off their cards in full each month versus those who spend a lot and keep their balance near their credit limit).

These different sorts of customers can tolerate very different credit lines (in the two examples, extra care must be taken with the latter to avoid default).

Five clusters (groups of customers) they found;

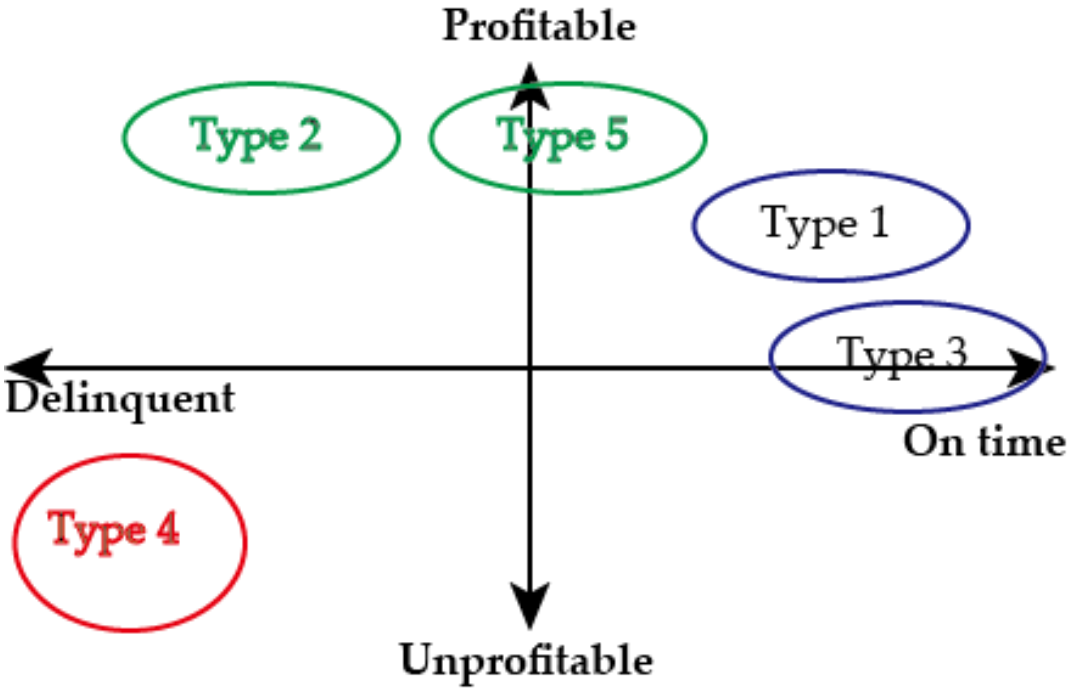
1. Usually on time with payments, pay most of their monthly balance, use some of their credit line, fairly high sales, and fairly profitable.

2. Fairly delinquent accounts, pay some of their monthly balance, high sales, and very profitable. Should be treated with caution in times of recession.

3. On time with payments, but very little sales activity. Not very profitable.

4. Very delinquent; all of these are write-offs. Generate fairly high sales but are unprofitable. Creditors lose money on these.

5. Mixture of on-time and delinquent accounts, generate high sales, and are very profitable, especially at lower credit lines.



The problem with using this clustering immediately for decision making is that the data are not available when the initial credit line is set. Haimowitz and Schwarz took this new knowledge and cycled back to the beginning of the data mining process. They used the knowledge to define a precise predictive modeling problem: using data that are available at the time of credit approval, predict the probability that a customer will fall into each of these clusters. This predictive model then can be used to improve initial credit line decisions.

What is Data Science?

Data analysis: help improving decision making.

Data Science: provide principles, processes, techniques for understanding phenomena via data analysis.

Principles: Incorporated into processes/techniques in the course.
Processes: CRISP-DM (Lec. #2), SEMMA and etc.
Techniques: Regression, Classification, Association and etc.

What is Data Science?

Data analysis: help improving decision making.

Data Science: provide principles, processes, techniques for understanding phenomena via data analysis.

Principles: Incorporated into processes/techniques in the course.
Processes: CRISP-DM (Lec. #2), SEMMA and etc.
Techniques: Regression, Classification, Association and etc.

An example of principles:

Extracting useful knowledge from data to solve problems should be treated systematically. → Incorporated into CRISP-DM.

What is Data Science?

Data analysis: help improving decision making.

Data Science: provide principles, processes, techniques for understanding phenomena via data analysis.

Principles: Incorporated into processes/techniques in the course.
Processes: CRISP-DM (Lec. #2), SEMMA and etc.
Techniques: Regression, Classification, Association and etc.

An example of principles:

Extracting useful knowledge from data to solve problems should be treated systematically. → Incorporated into CRISP-DM.

You will find something from any set of data – but it might not generalize.
→ Evaluation of each technique (overfitting)

What is Data Science?

Data analysis: help improving decision making.

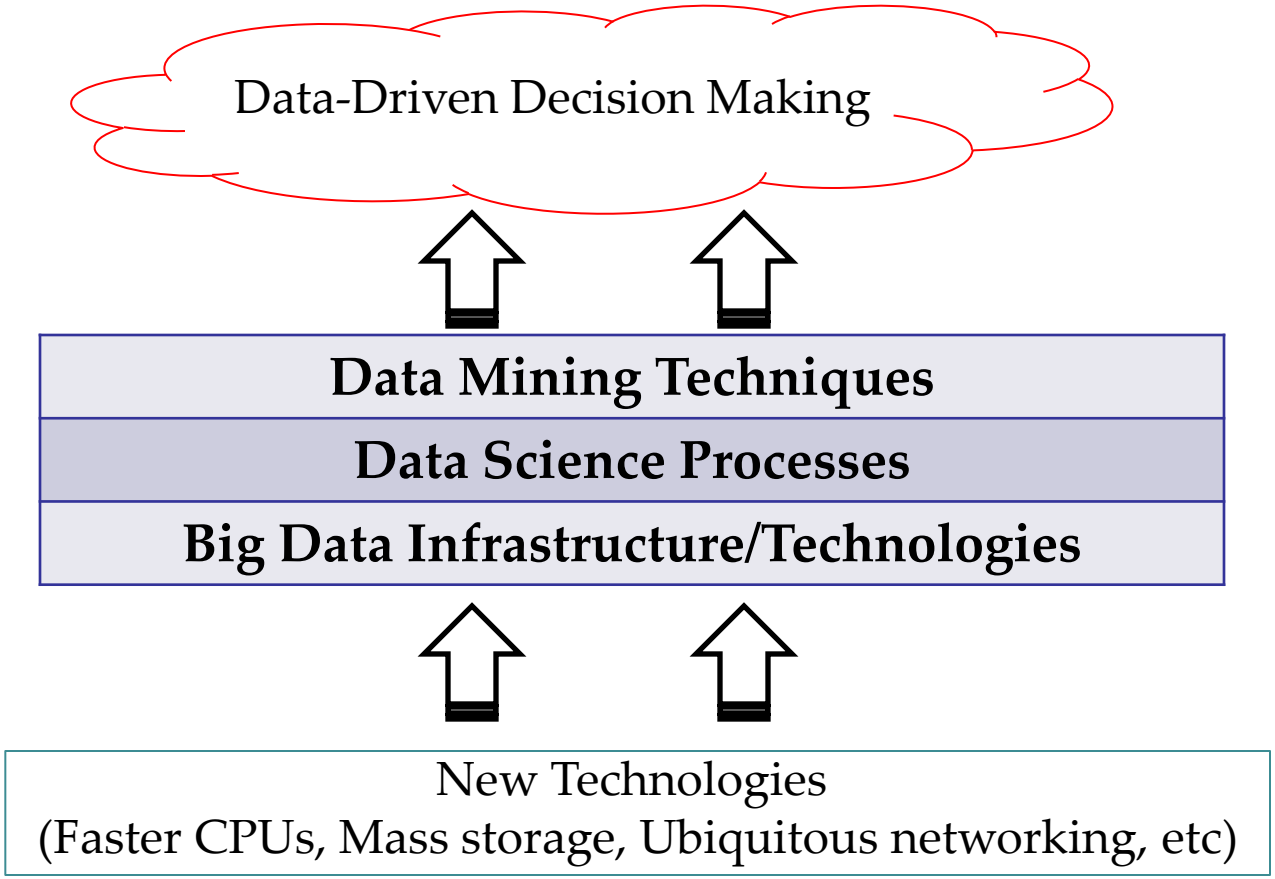
Data Science: provide principles, processes, techniques for understanding phenomena via data analysis.

Principles: Incorporated into processes/techniques in the course.
Processes: CRISP-DM (Lec. #2), SEMMA and etc.
Techniques: Regression, Classification, Association and etc.

Big Data Science: provide principles, processes, techniques for understanding phenomena via big data analysis.

+ Big Data Infrastructure (Hadoop etc.)

Big Data Science



Potential Applications (1/2)

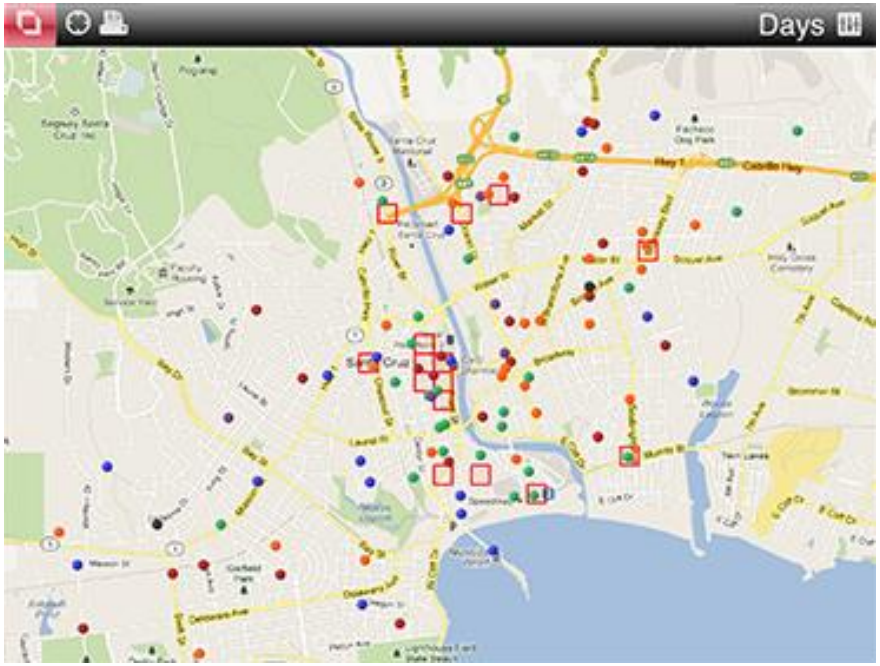
Many Big Data applications are in marketing.

But, its applications are not just in marketing…

- Predict crimes

LAPD (L.A. Police Dept.) uses Big data to predict crimes.

PredPol (*http://www.predpol.com/gun-violence*)



Red boxes indicate where the patrols should watch.

- Potential Applications (2/2)
  - Weather Forecast

    WeatherNews has made a weather communication community among users of its phone application. It collects users' weather observations and use them to forecast the weather together with publicly available data.

Some users' weather reports (*http://weathernews.jp*)

Weathernews collects weather reports from their community.

Weather reports around Fukushima Prefecture, Japan (*http://weathernews.jp*)

In some cases, they provided more accurate predictions than the public weather forecast, JMA (Japan Meteorological Agency).

For a forecast for Feb. 6 2013, JMA predicted (relatively heavy) snow, but WNews predicted rain. And it rained. (*http://diamond.jp/articles/32435*)

- Example 1

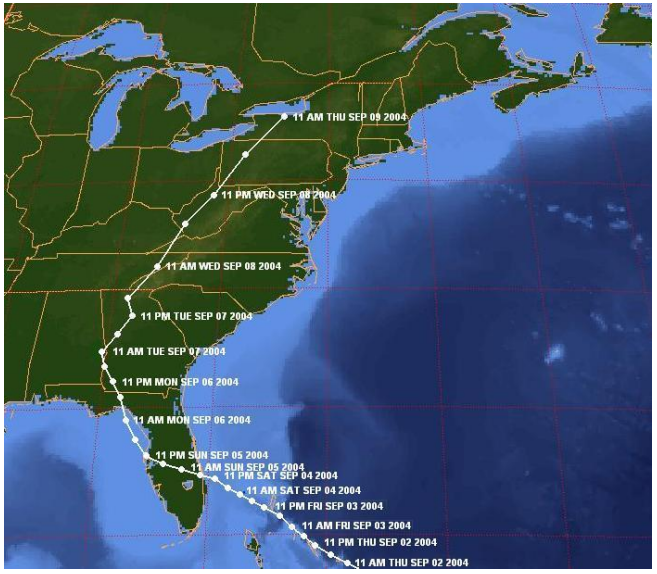
A *New York Times* story from 2004:

"Hurricane Frances was on its way, barreling across the Caribbean, threatening a direct hit on Florida's Atlantic coast. Residents made for higher ground, but far away, in Bentonville, Ark., executives at Wal-Mart Stores decided that the situation offered a great opportunity for one of their newest data-driven weapons … predictive technology.

A week ahead of the storm's landfill, Linda M. Dillman, Wal-Mart's chief information officer, pressed her staff to come up with forecasts based on what had happened when Hurricane Charley struck several weeks earlier. Backed by the trillions of bytes' worth of shopper history that is stored in Wal-Mart's data warehouse, she felt that the company could 'start predicting what's going to happen, instead of waiting for it to happen,' as she put it. (Hays, 2004)"

(See *http://www.nytimes.com/2004/11/14/business/yourmoney/14wal.html?_r=0*)

Hurricane Frances



http://www.noaanews.noaa.gov/stories2004/s2312.htm

http://www4.ncsu.edu/~nwsfo/storage/cases/20040908/

Would data analysis be useful in this scenario?

- Find that people in the path of the hurricane would buy more bottled water.

Maybe, but this point seems obvious.

- Discover patterns due to the hurricane that were not obvious.

(More valuable)

The New York Times (Hays, 2004) reported that:

"… the experts mined the data and found that the stores would indeed need certain products —and not just the usual flashlights. 'We didn't know in the past that **strawberry Pop-Tarts** increase in sales, like **seven times** their normal sales rate, ahead of a hurricane,' Ms. Dillman said in a recent interview."

Remark:

Your data can tell you that the sales will increase.

Your data cannot tell you why the sales will increase.

You often need to infer the implication of your analysis if you need to convince people to follow your suggestions.

Saying "Hey, according to data, your store should stock up pop-tarts" may not be enough.

So, let's infer the implication…

What will people stock up on before hurricanes?

One of such items is non-perishable food that can be eaten easily and without heat.

Why? After hurricanes electricity will often be out for multiple days.

Pop-Tarts require **no heating** to eat.

They do **not need a fridge or freezer** to be stored.

Hurricanes often hit the Southeast US during summer where it is very hot and humid. A great fruit to cool off with, next to watermelon, is the strawberry.

→ Strawberry pop-tarts seem to perfectly fit.

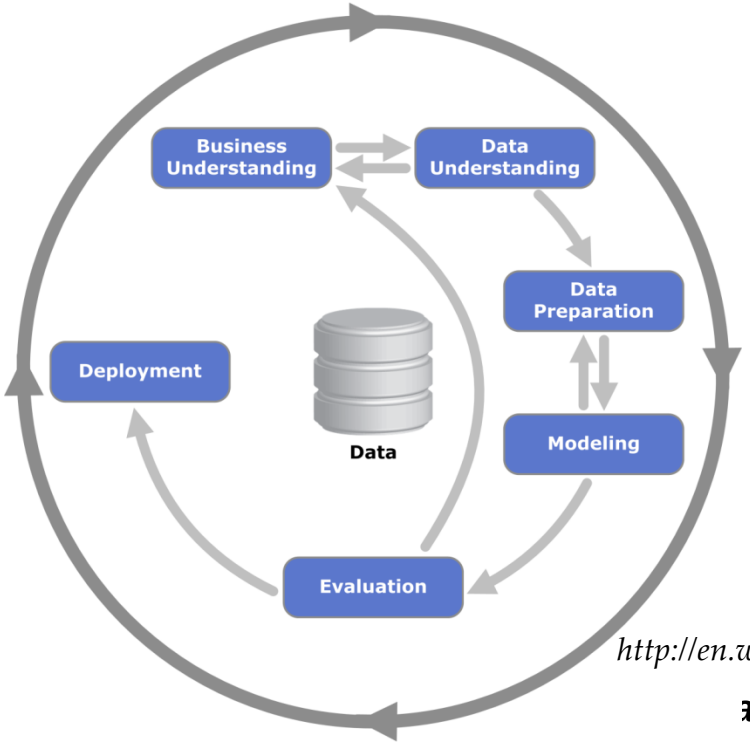
Sneak preview: Data Science Process

In this course, you will…

- Learn how to approach business/research problems data-analytically.
- Be able to assess whether and how data can solve problems.

How to approach problems data-analytically??

There exists a standard process, CRISP-DM.



Data Science Process consists of ..

1. Business understanding
2. Data understanding
3. Data preparation
4. Modeling
5. Evaluation
6. Deployment

http://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

Recap: CRISP-DM

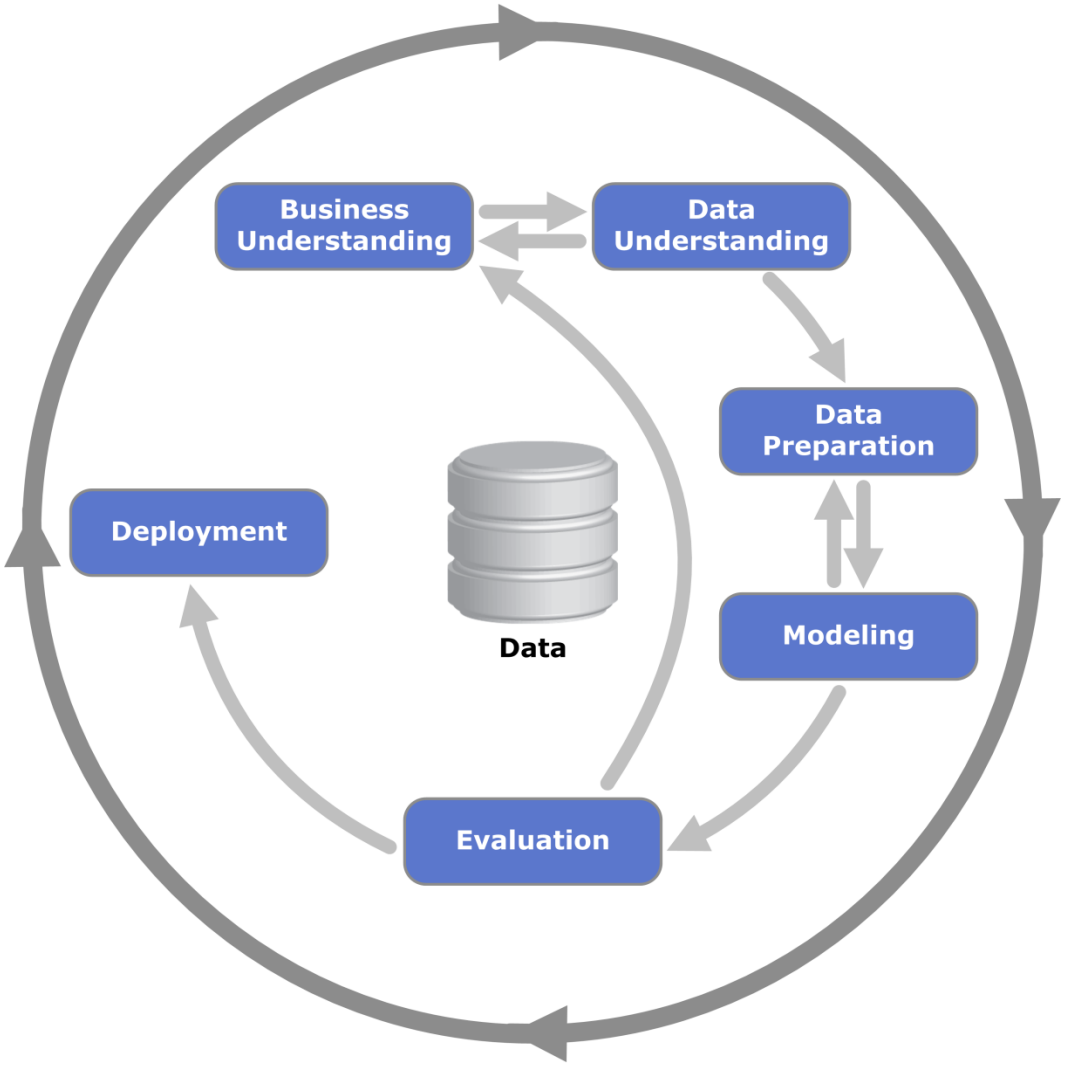
Cross-Industry Standard Process for Data Mining (CRISP-DM)

European Community funded effort to develop framework for data mining tasks

Goals:

- Encourage interoperable tools across entire data mining process
- Take the mystery/high-priced expertise out of simple data mining tasks

CRISP-DM overview



Note: Iteration is the rule rather than the exception.

CRISP-DM figure (wikipedia.org)

http://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

A Scenario

Example: Predicting Customer Churn

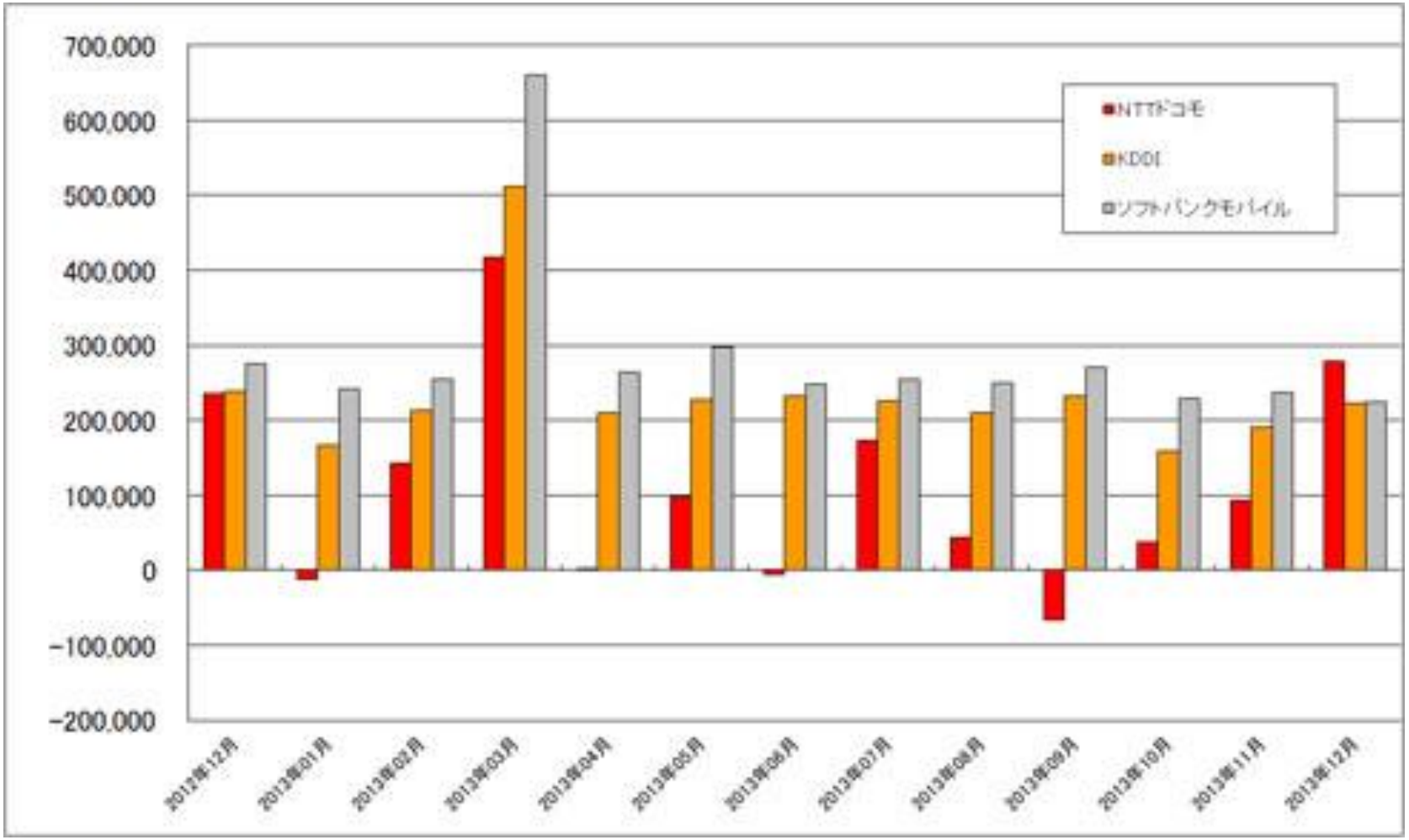
You just landed a great analytical job with MegaTelCo, one of the largest telecommunication firms in the United States.

Since the cell phone market is now saturated, the huge growth in the wireless market has tapered off. Communications companies are now engaged in battles to attract each other's customers while retaining their own.

They have allocated some marketing budget to spend customers. You have been called in to help figure out its marketing strategy.

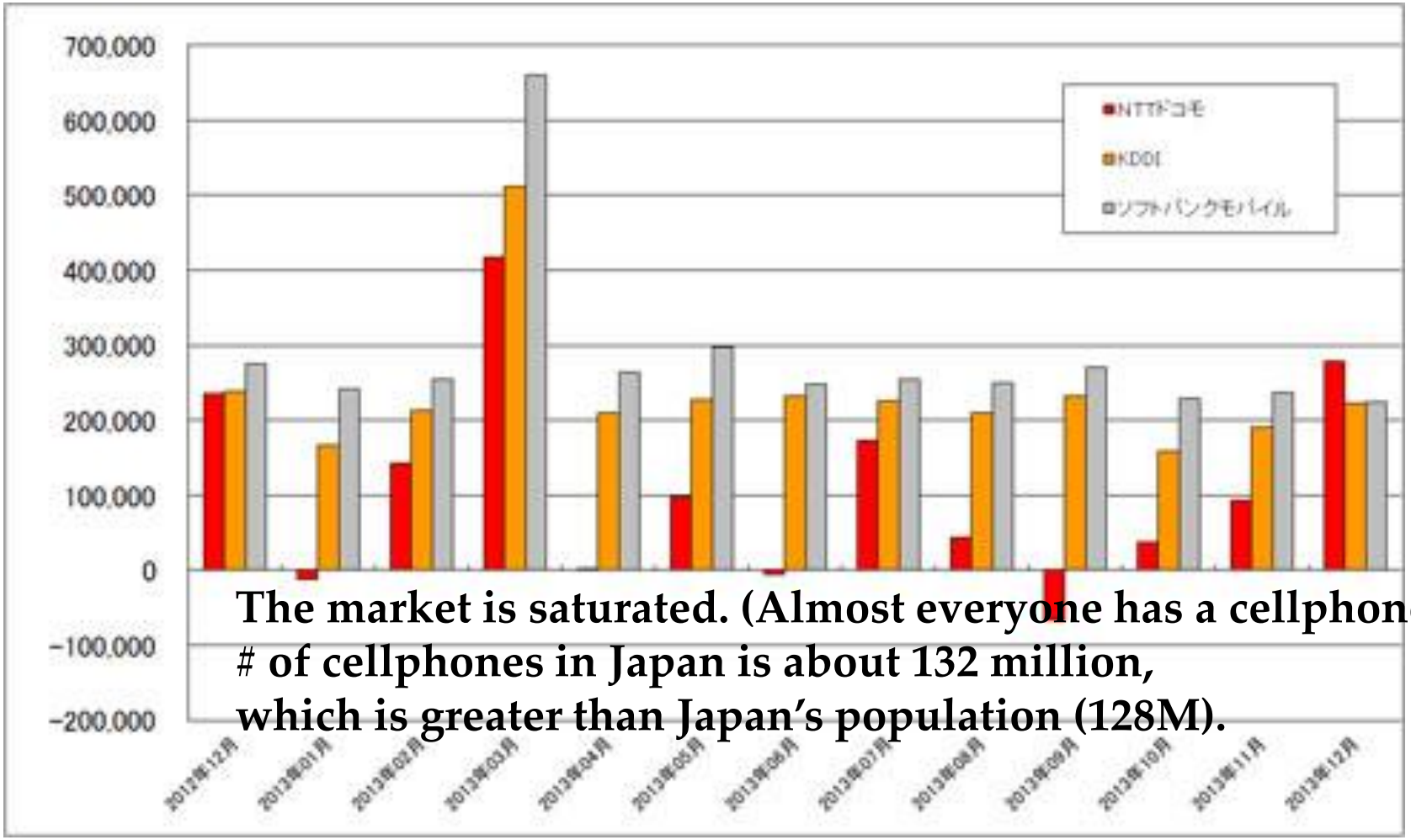
Think carefully about what data you might use and how they would be used. Specifically, they need your help to choose a set of customers to receive their offer in order to best increase its profits.

Figure 1: Fierce competition in mobile telecommunication industry (Japan).
Firms are trying to attract more customers..
The figure shows # of monthly increase in the carriers' contracts.



http://www.itmedia.co.jp/mobile/articles/1401/10/news106.html

Figure 1: Fierce competition in mobile telecommunication industry (Japan).
Firms are trying to attract more customers..
The figure shows # of monthly increase in the carriers' contracts.



The market is saturated. (Almost everyone has a cellphone.)
# of cellphones in Japan is about 132 million,
which is greater than Japan's population (128M).

http://www.itmedia.co.jp/mobile/articles/1401/10/news106.html

1. Business Understanding

- Determine business objectives
- Solve a specific problem
- Assess the current situation
- Convert the above into a data mining project
- Develop a project plan

1. Business Understanding

What is the best way to spend the marketing money?

Consider which group of people we should focus.

- Attract _new_ customers.
- Retain _existing_ customers.

1. Business Understanding

What is the best way to spend the marketing money?

Consider which group of people we should focus.

- Attract _new_ customers.

  Who should we target?

  - People who currently don't use any cell phone?
  - People who have some contract with another carrier?

- Retain _existing_ customers.

  Who should we target? (Who will likely leave/switch from us?)

  - Age?
  - Gender?
  - Income?
  - Depending on their plans?

1. Business Understanding

What is the best way to spend the marketing money?

Consider which group of people we should focus.

- Attract _new_ customers.
- Retain _existing_ customers.

After discussions with the marketing team, you found that …

1. Business Understanding

What is the best way to spend the marketing money?

Consider which group of people we should focus.

- Attract _new_ customers.
- Retain _existing_ customers.

After discussions with the marketing team, you found that …

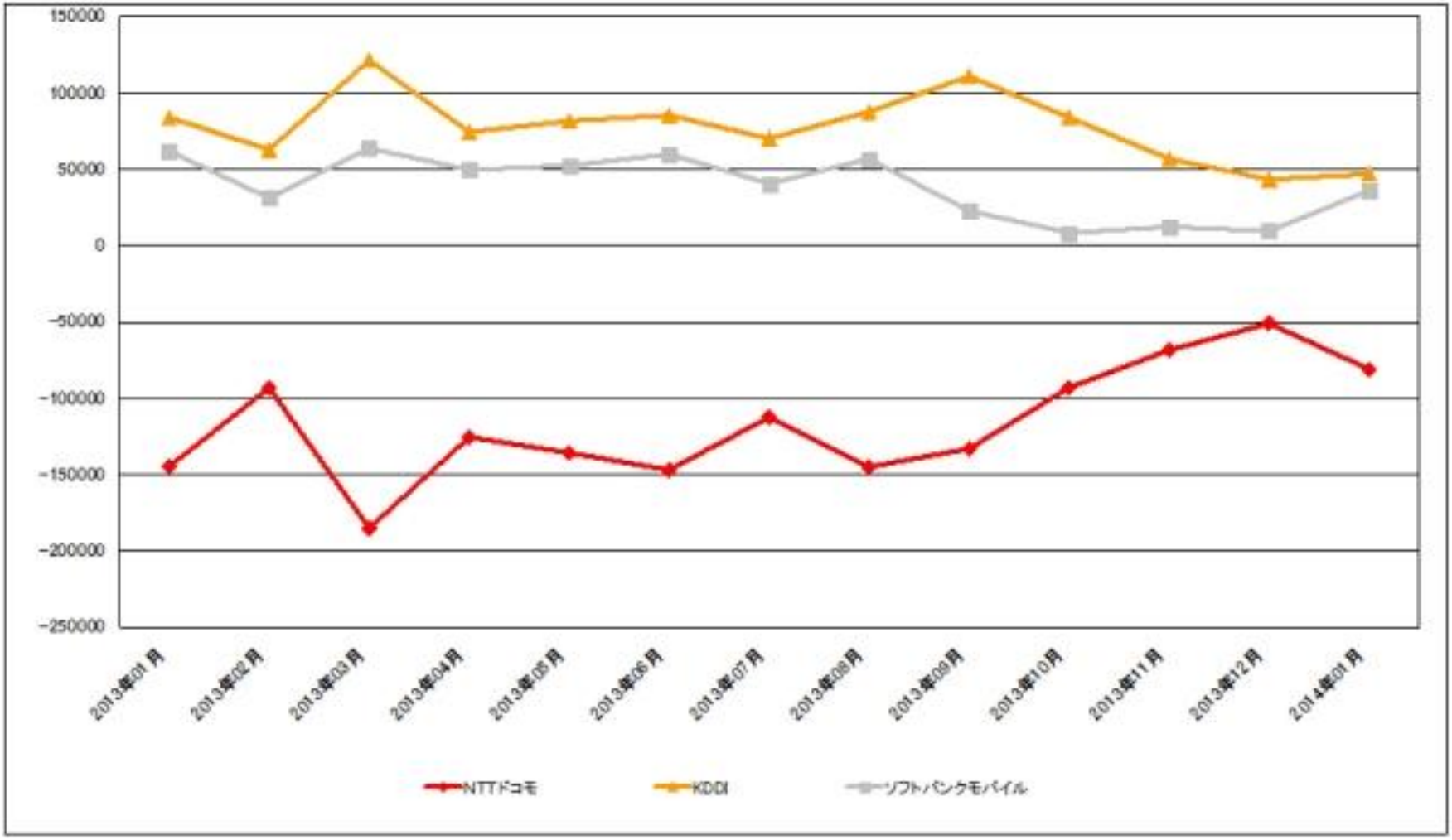
Attracting new customers is much more expensive than retaining existing ones, so a good deal of marketing budget is allocated to retain existing customers.

In fact, MegaTelCo is having a major problem with customer retention in their wireless business. In the mid-Atlantic region, 20% of cell phone customers leave when their contracts expire.

Terminology: Customers switching from one company to another is called ***churn***, and it is expensive all around: one company must spend on incentives to attract a customer while another company loses revenue when the customer departs.

Figure 2: Fierce competition in mobile telecommunication industry (Japan).

Churn is a severe problem. The figure shows increase/decrease in contracts via MNP (Mobile Number Portability). Docomo is losing 50,000 to 150,000 customers monthly..



http://www.itmedia.co.jp/mobile/articles/1402/07/news116.html

- Transform the business problem into a data mining one

Marketing has already designed a special retention offer. Your task is to devise a precise, step-by-step plan for how the data science team should use MegaTelCo's vast data resources to decide which customers should be offered the special retention deal prior to the expiration of their contracts.

How should MegaTelCo choose a set of customers to receive their offer in order to best reduce churn for a particular incentive budget?

Your model should predict which customers will likely churn.

2. Data Understanding

- Initial Data Collection
- Data Description
- Data Exploration
- Data Quality Verification
- Data Selection
- Related data can come from many sources

2. Data Understanding

We have a historical data set of 20,000 customers. At the point of collecting the data, each customer either had stayed with the company or had left (churned). Each customer is described by the variables listed below.

Variable	Explanation
COLLEGE	Is the customer college educated?
INCOME	Annual income
OVERAGE	Average overcharges per month
LEFTOVER	Average number of leftover minutes per month
HOUSE	Estimated value of dwelling (from census tract)
HANDSET_PRICE	Cost of phone
LONG_CALLS_PER_MONTH	Average number of long calls (15 mins or over) per month
AVERAGE_CALL_DURATION	Average duration of a call
REPORTED_SATISFACTION	Reported level of satisfaction
REPORTED_USAGE_LEVEL	Self-reported usage level
LEAVE (Target variable)	Did the customer stay or leave (churn)?

Data looks like …

ID	COLLEGE	INCOME	OVERAGE	LEFTOVER	HOUSE	HANDSET_PRICE	LONG_CALLS_PER_MONTH	AVERAGE_CALL_DURATION	REPORTED_SATISFACTION	REPORTED_USAGE_LEVEL	LEAVE (Target variable)
0001	1	50000	30	0	1M	150	3	15			0
0002	1	60000	10	20	.5M	50	.8	5			1
0003	0	30000	15	5		75	1.2	6			1
0004	0		7	0	0	30	.5	7			0
:	:	:	:	:	:	:	:	:	:	:	:

Data looks like …

ID	COLLEGE	INCOME	OVERAGE	LEFTOVER	HOUSE	HANDSET_PRICE	LONG_CALLS_PER_MONTH	AVERAGE_CALL_DURATION	REPORTED_SATISFACTION	REPORTED_USAGE_LEVEL	LEAVE (Target variable)
0001	1	50000	30	0	1M	150	3	15			0
0002	1	60000	10	20	.5M	50	.8	5			1
0003	0	30000	15	5		75	1.2	6			1
0004	0		7	0	0	30	.5	7			0
:	:	:	:	:	:	:	:	:	:	:	:

Available data to predict churn.

This is what we want to predict.

3. Data Preparation

- Clean selected data for better quality
  - Treat missing values
  - Identify or remove outliers
  - Resolve redundancy caused by data integration
  - Correct inconsistent data
- Transform data
  - Convert different measurements of data into a unified numerical scale by using simple mathematical formulations

For further reference: “Introduction to Data Mining” by Tan, Steinbach and Kumar.

ID	COLLEGE	INCOME	OVERAGE	LEFTOVER	HOUSE	HANDSET_PRICE	LONG_CALLS_PER_MONTH	AVERAGE_CALL_DURATION	REPORTED_SATISFACTION	REPORTED_USAGE_LEVEL	LEAVE (Target variable)
0001	1	50000	30	0	1M	150	3	15			0
0002	1	60000	10	20	.5M	50	.8	5			1
0003	0	30000	15	5		75	1.2	6			1
0004	0		7	0	0	30	.5	7			0
:	:	:	:	:	:	:	:	:	:	:	:

Example: Treat missing values

- Eliminate data objects with missing values.
- Ignore attributes with missing values.
- Fill in missing values.

ID	COLLEGE	INCOME	OVERAGE	LEFTOVER	HOUSE	HANDSET_PRICE	LONG_CALLS_PER_MONTH	AVERAGE_CALL_DURATION	REPORTED_SATISFACTION	REPORTED_USAGE_LEVEL	LEAVE (Target variable)
0001	1	50000	30	0	1M	150	3	15			0
0002	1	60000	10	20	.5M	50	.8	5			1
0003	0	30000	15	5		75	1.2	6			1
0004	0		7	0	0	30	.5	7			0
:	:	:	:	:	:	:	:	:	:	:	:

Example: Treat missing values

- (E) Eliminate data objects with missing values.
- (I) Ignore attributes with missing values.
- (F) Fill in missing values.

Consider "REPORTED_USAGE_LEVEL" field.

(E) Most customers don't provide "REPORTED_USAGE_LEVEL" info.

If we do (E), we eliminate almost all records. → Not appropriate

(I) Most customers don't provide "REPORTED_USAGE_LEVEL" info.

If we do (I), almost all records will remain. → Looks good.

ID	COLLEGE	INCOME	OVERAGE	LEFTOVER	HOUSE	HANDSET_PRICE	LONG_CALLS_PER_MONTH	AVERAGE_CALL_DURATION	REPORTED_SATISFACTION	REPORTED_USAGE_LEVEL	LEAVE (Target variable)
0001	1	50000	30	0	1M	150	3	15			0
0002	1	60000	10	20	.5M	50	.8	5			1
0003	0	30000	15	5		75	1.2	6			1
0004	0		7	0	0	30	.5	7			0
:	:	:	:	:	:	:	:	:	:	:	:

Example: Treat missing values

- (E) Eliminate data objects with missing values.
- (I) Ignore attributes with missing values.
- (F) Fill in missing values.

Consider "INCOME" field.

(E) Some customers don't provide "INCOME" info.

If we do (E), we eliminate some fraction of records. → Okay.

(I) Many customers provide "INCOME" info.

If we do (I), we eliminate many. → Not appropriate.

(F) We may be able to reasonably estimate income from other data, e.g. COLLEGE.

Fill in a blank with average income of college graduate/non-graduate. → Okay.

4. Modeling

- Data Treatment
  - Training set
  - Test set
- Data Mining Techniques
  - Regression (prediction)
  - Association
  - Classification
  - Clustering

4. Modeling

We have clean data and want to predict churn.

Which variables we should use to predict churn?

Restate: Which of the variables would be best to segment these people into groups, "churn" and "non-churn"?

Information gain

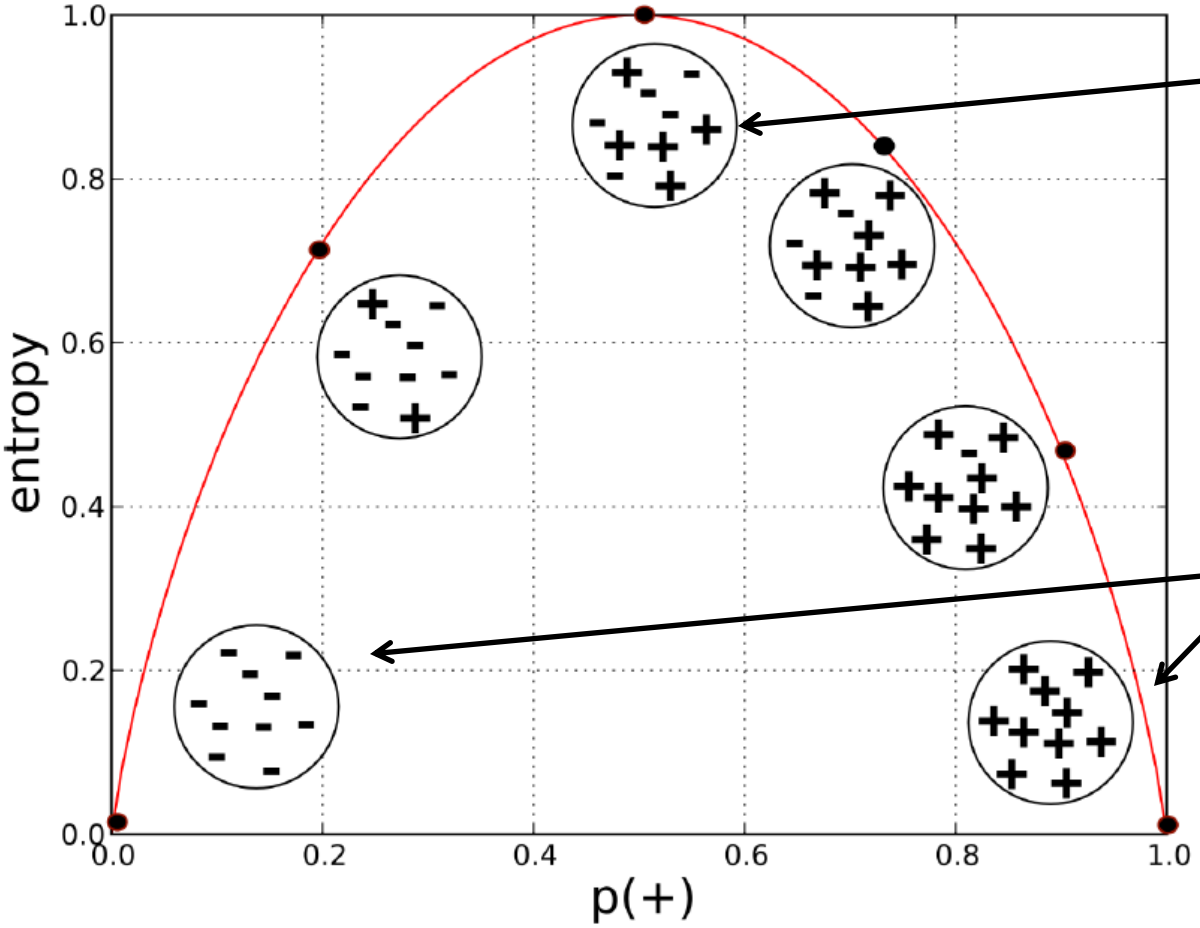
Information gain will tell us how informative an attribute is.

Entropy

$$entropy = -p1 \log p1 - p2 \log p2 \cdots$$

Each pX is the probability of property X within the set.

(pX=1 implies that all observations have property X.)

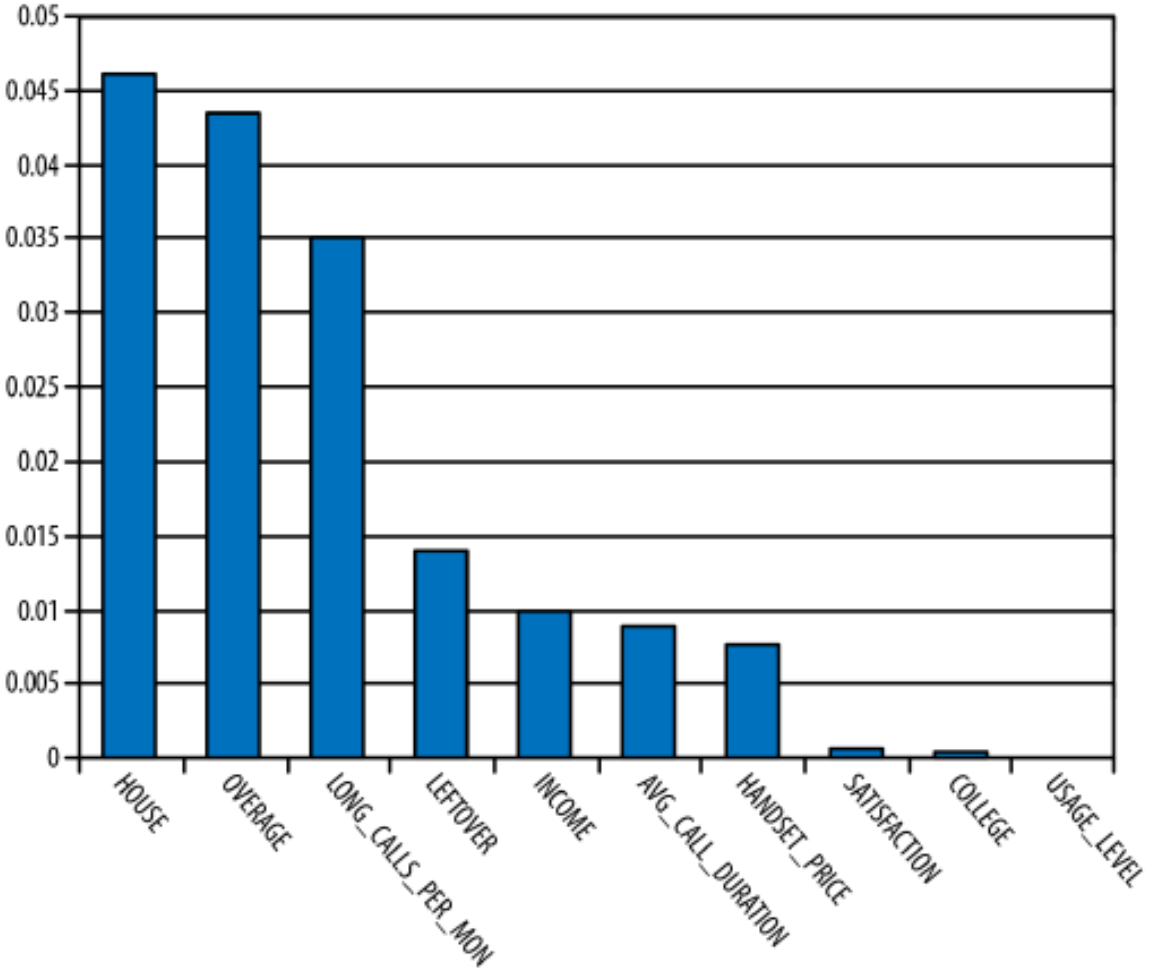


Not useful for prediction

Useful for prediction

p(+) is probability that a customer churned.

The model is more predictive if it achieves lower entropy.

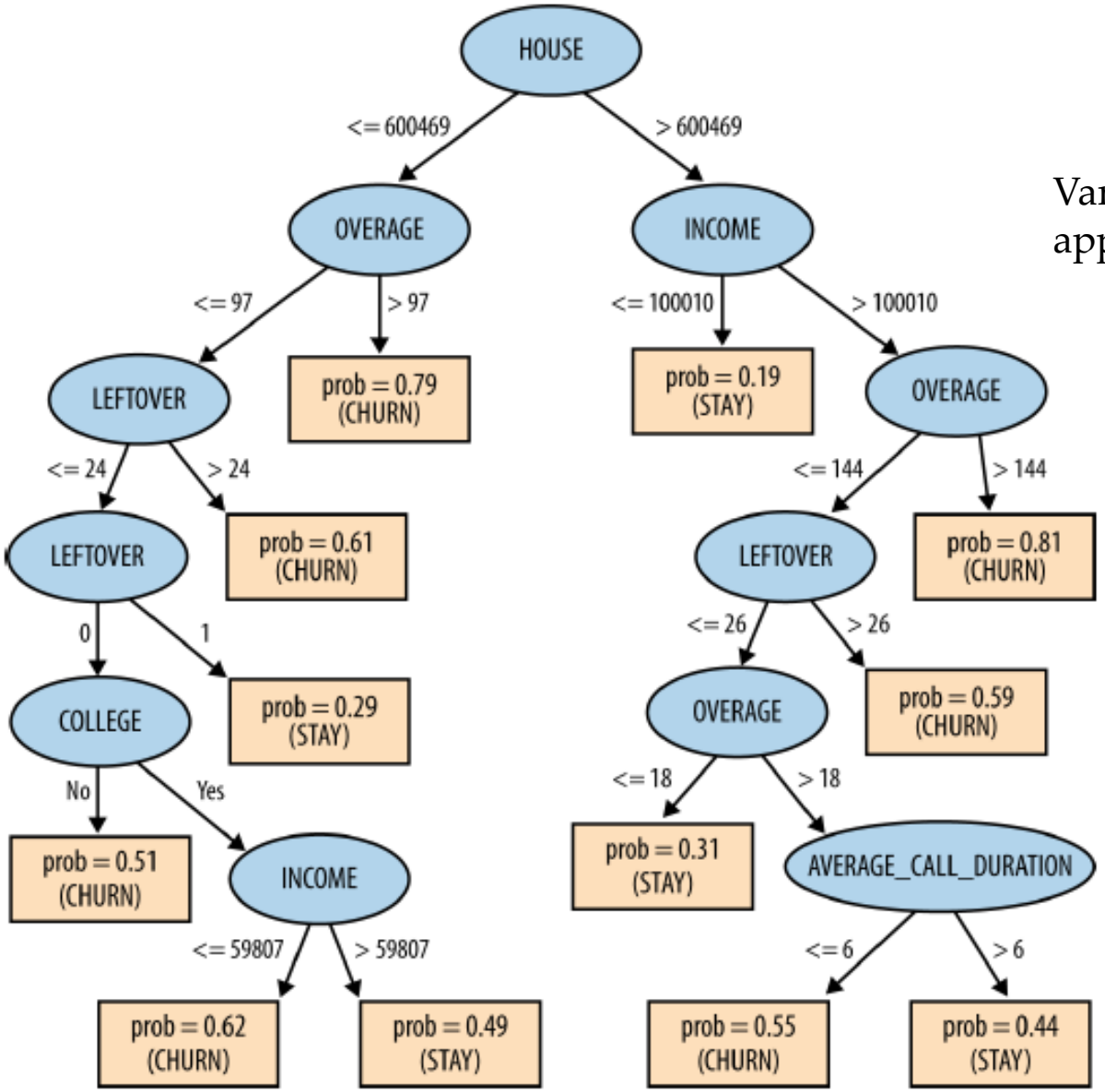


Rank	Info. gain	Attribute name
1	0.0461	HOUSE
2	0.0436	OVERAGE
3	0.0350	LONG_CALLS_PER_MON
4	0.0136	LEFTOVER
5	0.0101	INCOME
6	0.0089	AVG_CALL_DURATION
7	0.0076	HANDSET_PRICE
8	0.0003	SATISFACTION
9	0.000	COLLEGE
10	0.000	USAGE_LEVEL

Information gain for variables in churn example.

It would be good to include into the model variables which have relatively high IG.

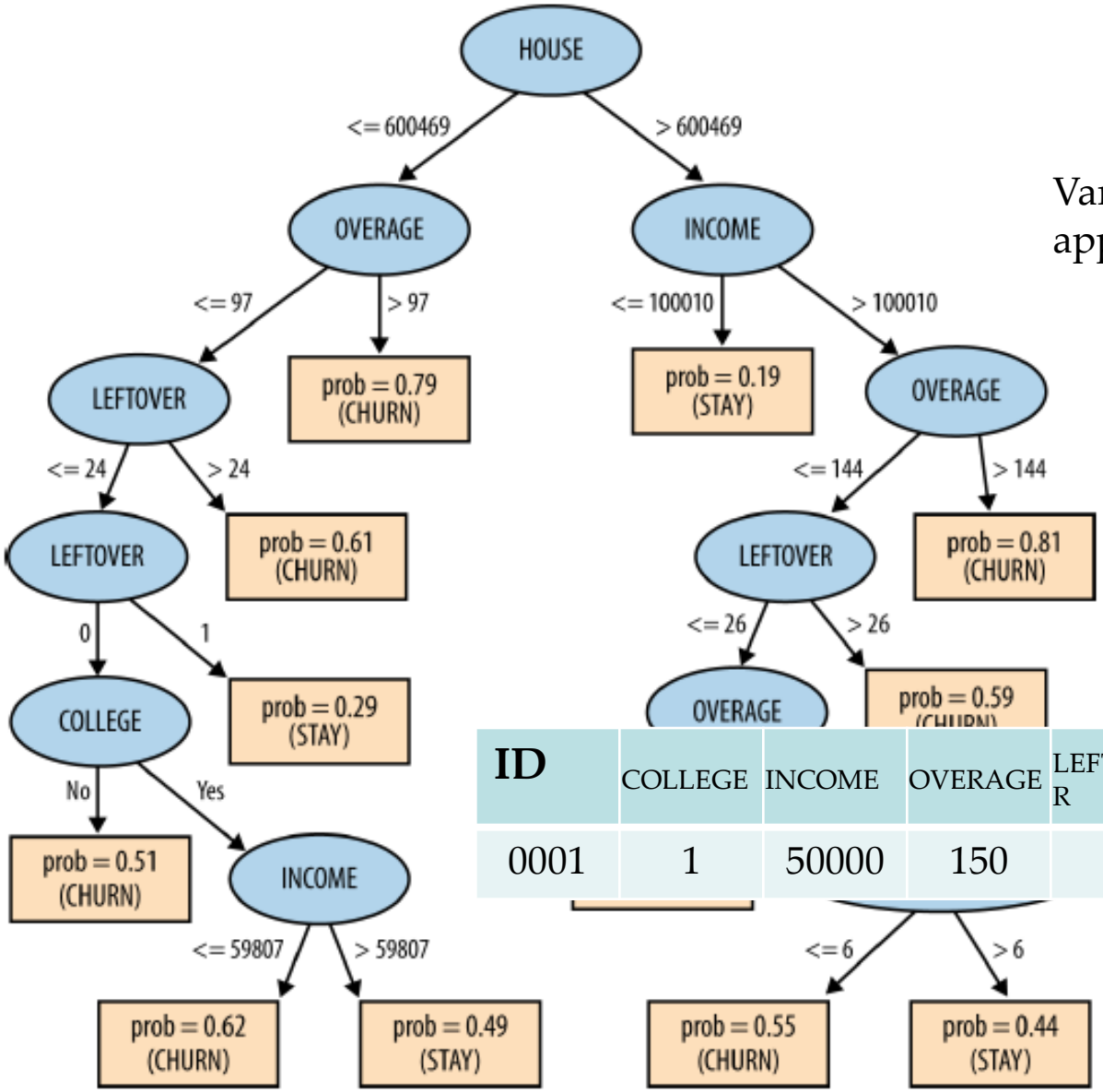
Modeling Example: choice of tree classification



Variables which have higher IG appear first. But not always..

("Data Science for Business by Foster Provost and Tom Fawcett (O'Reilly). Copyright 2013 Foster Provost and Tom Fawcett, 978-1-449-36132-7.")

Modeling Example: choice of tree classification

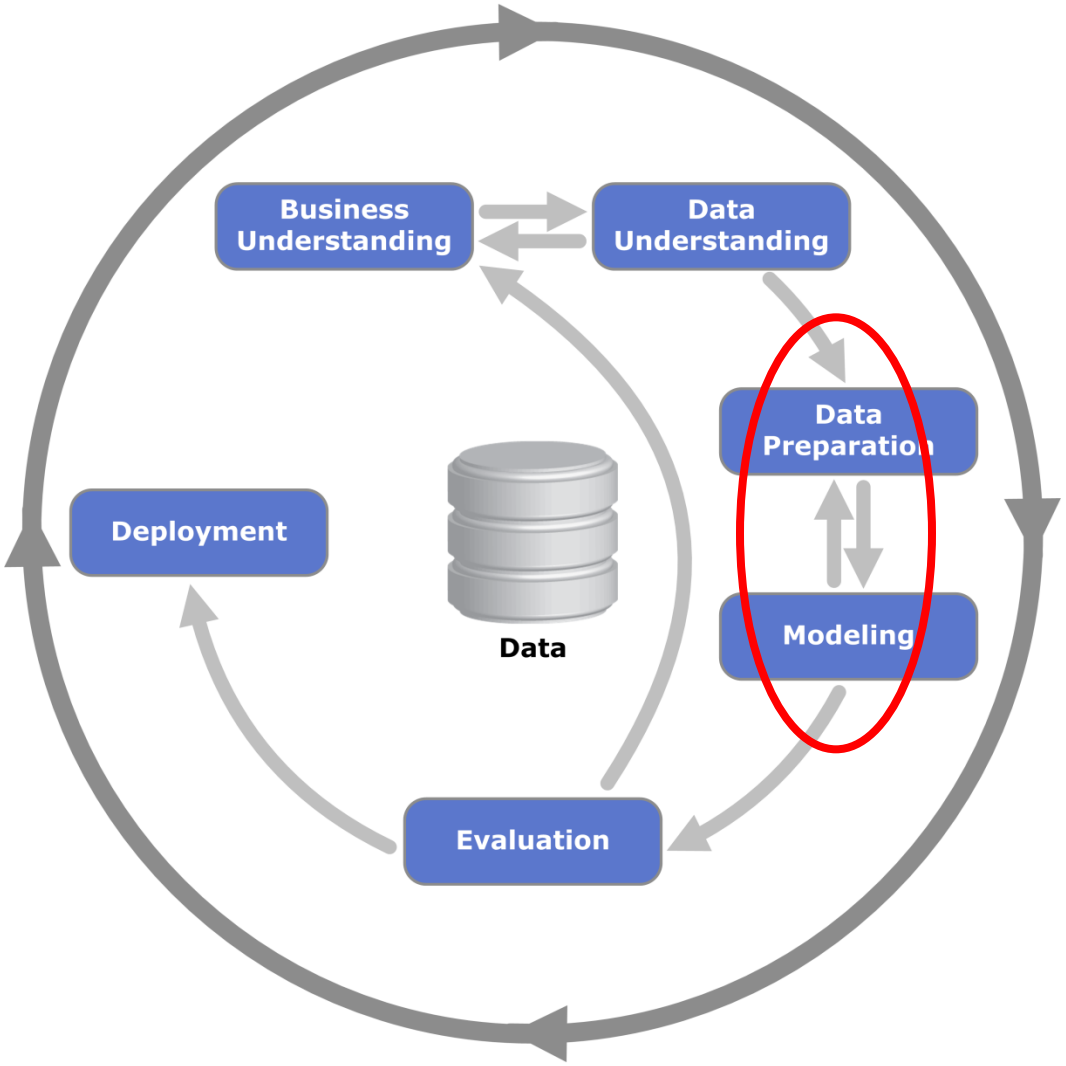


Variables which have higher IG appear first. But not always..

ID	COLLEGE	INCOME	OVERAGE	LEFTOVER	HOUSE	…..	LEAVE (Target variable)
0001	1	50000	150	0	1M	….	0

("Data Science for Business by Foster Provost and Tom Fawcett (O'Reilly). Copyright 2013 Foster Provost and Tom Fawcett, 978-1-449-36132-7.")

CRISP-DM overview



Note:
After Modeling stage, we may need to go back to Data Preparation.

Your model may require a specific form of data as input.

CRISP-DM figure (wikipedia.org)

http://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

5. Evaluation

- Does model meet business objectives?
- Any business objectives not addressed?
- Does model make sense?
- Is model actionable?
- It should be possible to make business decisions after this step.
- All important objectives should be achieved.

5. Evaluation

To evaluate the model, one of the most important fundamental notions:

Avoid Overfitting.

Overfitting is the tendency of data mining procedures to tailor models to the training data, at the expense of generalization.

An extreme example:

Consider the following *table* model.

We stored the feature vector for each customer who has churned in a database table. Let's call that Tc. In use, when the model is presented with a customer to determine the likelihood of churning, it takes the customer's feature vector, looks her up in Tc, and reports "100% likelihood of churning" if she is in Tc and "0% likelihood of churning" if she is not in Tc.

When the tech team applies our model to the historical dataset, the model predicts perfectly. The model 100% fits with the historical dataset.

ID	COLLEGE	INCOME	OVERAGE	LEFTOVER	HOUSE	HANDSET_PRICE	LONG_CALLS_PER_MONTH	AVERAGE_CALL_DURATION	REPORTED_SATISFACTION	REPORTED_USAGE_LEVEL	LEAVE (Target variable)
0001	1	50000	30	0	1M	150	3	15			0
0002	1	60000	10	20	.5M	50	.8	5			1
0003	0	30000	15	5		75	1.2	6			1
0004	0		7	0	0	30	.5	7			0
:	:	:	:	:	:	:	:	:	:	:	:

Use ID to predict LEAVE.
Does it work?

table model (illustration):

Your prediction system stores customer IDs who left MegaTelCo.

If you input a customer ID and it matches some stored ID, then your system predicts churn.

If you apply the existing data to your prediction system, you will obtain 100% accuracy.

What is the problem of the *table* model?

Consider how we'll use the model in practice.

When a previously unseen customer's contract is about to expire, we'll apply the model. This customer was not part of the historical dataset, and there will be no exact match in the table. Thus, the lookup will fail, and the model will predict "0% likelihood of churning" for this customer. In fact, the model will predict this for every customer (who is not in the training data).

The model looked perfect, but it is completely useless in practice!

The table model `overfits` to the training data and looses generalization.

Generalization is the property of a model or modeling process, whereby the model applies to data that were not used to build the model. In this example, the model does not generalize at all beyond the data that were used to build it.

The issue is more complex than it looks. The answer to overfitting is _not_ to use a data mining procedure that doesn't overfit. All of procedures do overfit.

How do we recognize overfitting?

A simple analytic tool: Fitting graph

It shows the accuracy of a model as a function of complexity.

First, we need to "hold out" some data.

"hold out" data: has the value of the target variable, but will not be used to build the model.

The issue is more complex than it looks. The answer to overfitting is _not_ to use a data mining procedure that doesn't overfit. All of procedures do overfit.

How do we recognize overfitting?

A simple analytic tool: Fitting graph

It shows the accuracy of a model as a function of complexity.

First, we need to "hold out" some data.

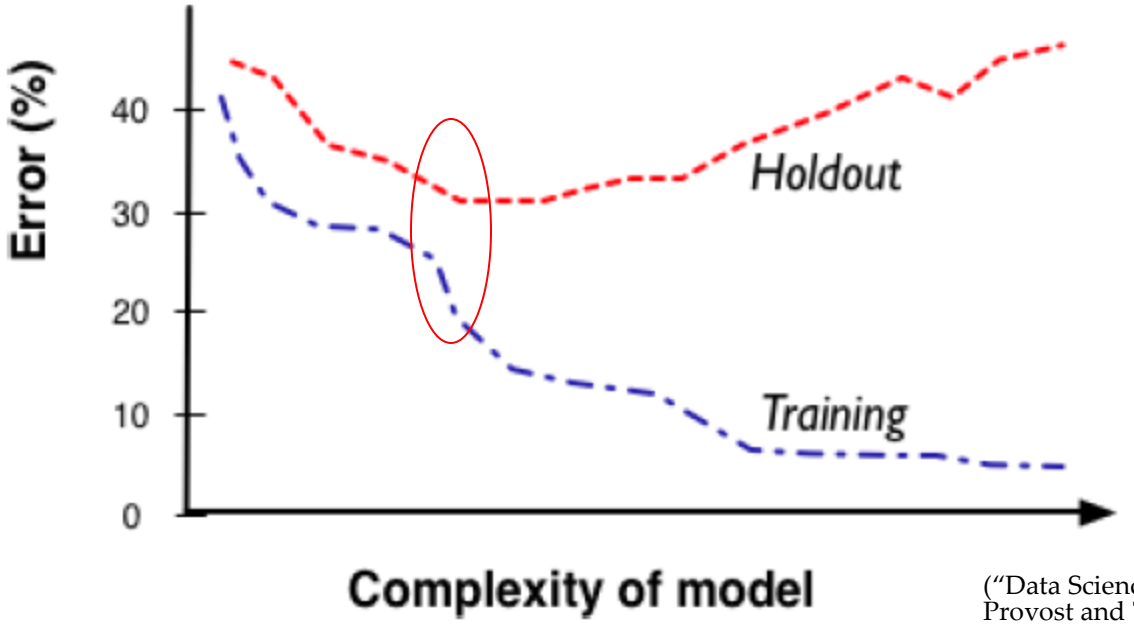
"hold out" data: has the value of the target variable, but will not be used to build the model.

Example:

You have data of 100,000 customers. Then,

50,000 customers: You apply a DM technique to build a model.

50,000 : You "hold out" for evaluation of your model.



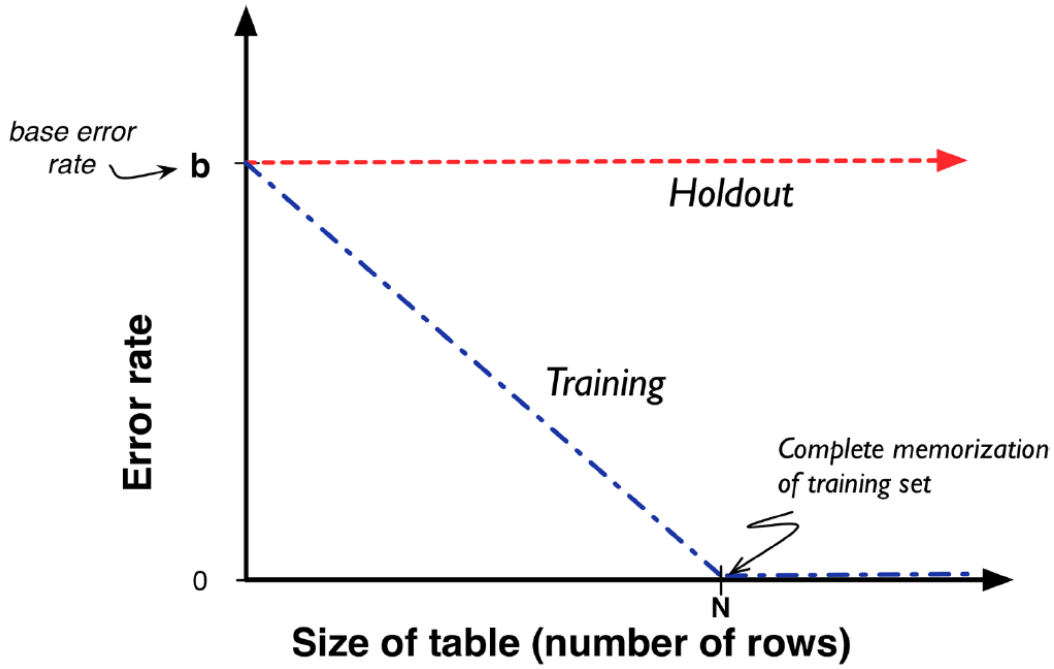
("Data Science for Business by Foster Provost and Tom Fawcett (O'Reilly). Copyright 2013 Foster Provost and Tom Fawcett, 978-1-449-36132-7.")

Fitting graph

Complexity: the number of variables we use in the model.

Each point on a curve represents an accuracy estimation of a model with a specified complexity (as indicated on the horizontal axis).

When the model is not allowed to be complex enough, it is not very accurate. As the models get too complex, they look very accurate on the training data, but in fact are overfitting—the training accuracy diverges from the holdout (generalization) accuracy.



("Data Science for Business by Foster Provost and Tom Fawcett (O'Reilly). Copyright 2013 Foster Provost and Tom Fawcett, 978-1-449-36132-7.")

Fitting graph (the *table* model)
Complexity: Size of table

What would '**b**' be?

The table model always predicts no churn for every new case.

It will get every no churn case right and every churn case wrong.

'**b**' in the figure will be the percentage of churn cases in the population.

This is known as the base rate, and a classifier that always selects the majority class is called a base rate classifier.

6. Deployment

- Ongoing monitoring and maintenance
  - Evaluate performance against success criteria
  - Market reaction & competitor changes

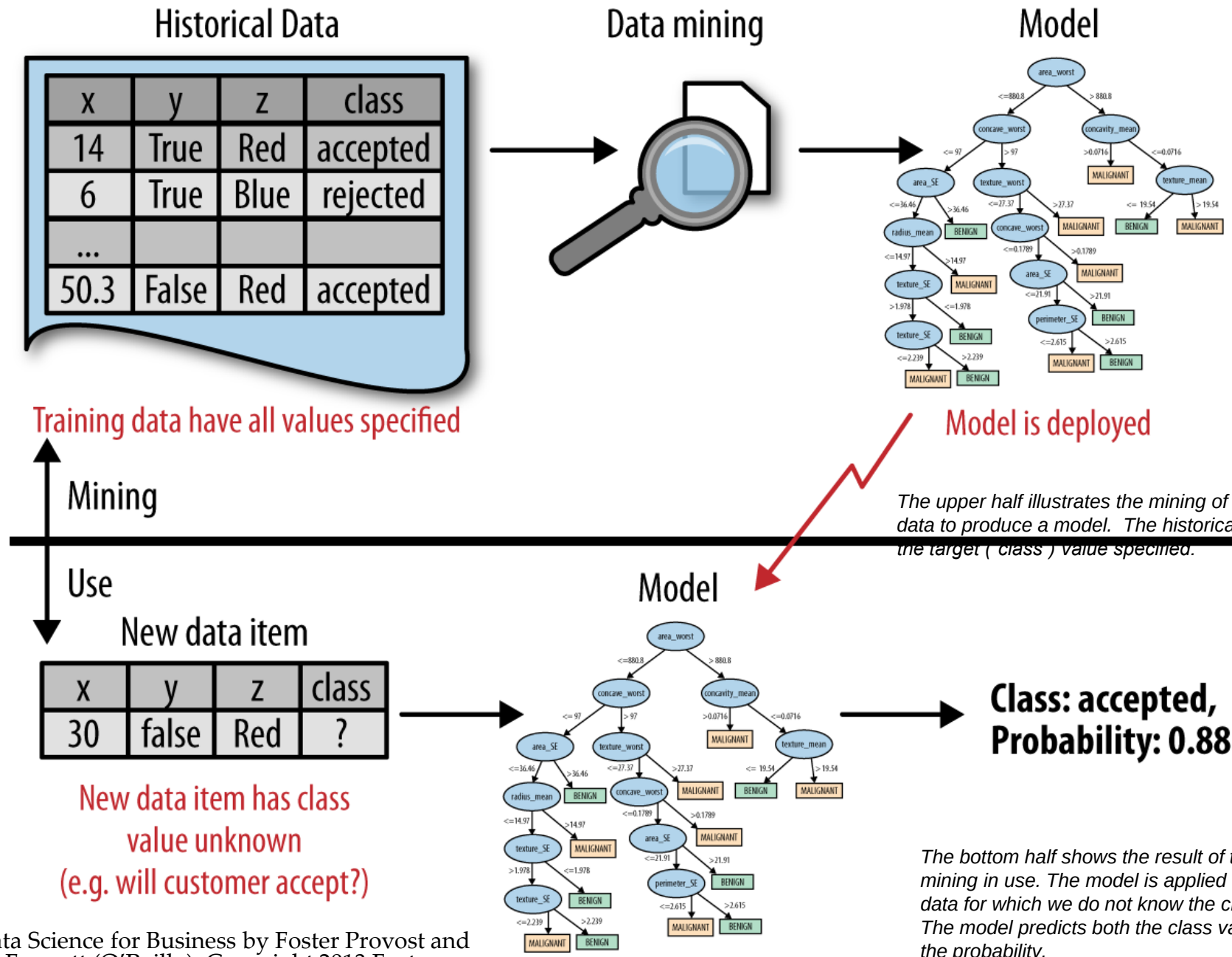
6. Deployment

A deployment scenario

We want to use the model to predict which of our customers will leave. Specifically, assume that data mining has created a class probability estimation model M.

Given each existing customer, M takes her characteristics as input and produces a score or probability estimate of attrition. This is the use of the results of data mining. The data mining produces the model M from some other data, and the model M provides a prediction.

The next figure illustrates these two phases. Data mining produces the probability estimation model, as shown in the top half of the figure. In the use phase (bottom half), the model is applied to a new, unseen case and it generates a probability estimate for it.



Pitfalls

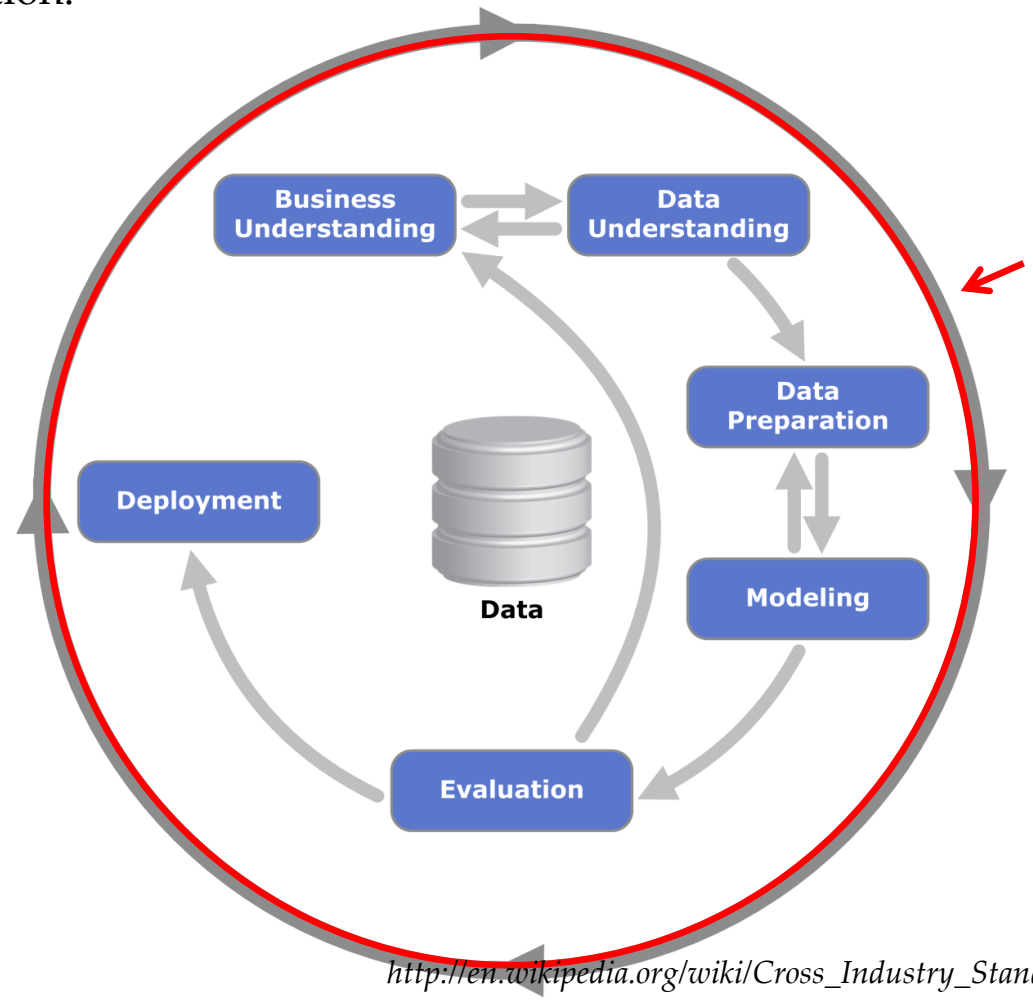
Deploying a model into a production system typically requires that the model be adjusted for the production environment, usually for greater speed or compatibility with an existing system.

There are risks with transfers from data science to development. Remember that: "A deployed model is not what the data scientists design, it's what the engineers build."

It is advisable to have members of the development team involved early on in the data science project. They can begin as advisors, providing critical insight to the data science team.

The process never ends…

Regardless of whether deployment is successful, the process often returns to the Business Understanding phase. The process of mining data produces a great deal of insight into the business problem and the difficulties of its solution. A second iteration can yield an improved solution.



Note: Iteration is the rule rather than the exception.

http://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining