Programming for Big Data

Lecture 3

Data Processing with Spark

Dr. Bojan Božić
Dublin institute of Technology

Agenda

Preliminaries

1. Use Case: Personalisation
   - Sales and Marketing
   - Customer Segmentation
2. Machine Learning with Spark
   - Basic Data Types
   - Basic Statistics
   - Supervised
   - Unsupervised
   - Evaluation
   - PMML - Publishing and Sharing
4. Summary

PRELIMINARIES

My Details

Name: Dr. Bojan Božić

Email: bojan.bozic@dit.ie

Details on SPARK labs:

https://ceadar.dit.ie/bojan.bozic/SPARK/

Final assignment will be published on webcourses.

Any Questions, Please email me!

Books and Resources

Mastering Spark

www.packtpub.com/big-data-and-business-intelligence/mastering-apache-spark

Apache Spark From Inception to Production

http://info.mapr.com/rs/mapr/images/Getting_Started_With_Apache_Spark.pdf

Spark Programming Guide

http://spark.apache.org/docs/latest/programming-guide.html

Course Details

Classes: Wednesday 18:30 - 21:30

- 4 labs with 3 lab assignments (voluntary)
- Final assignment starts in last (4^{th}) lab

Grading: Only final assignment is graded (100%)

Software

http://spark.apache.org/

USE CASE: PERSONALISATION

From then to now…

Pfizer

ups

Coca-Cola

COMCAST

P&G

RYANAIR

NOVARTIS

Customer Acquisition:

The process of acquiring new leads, prospects and customers.

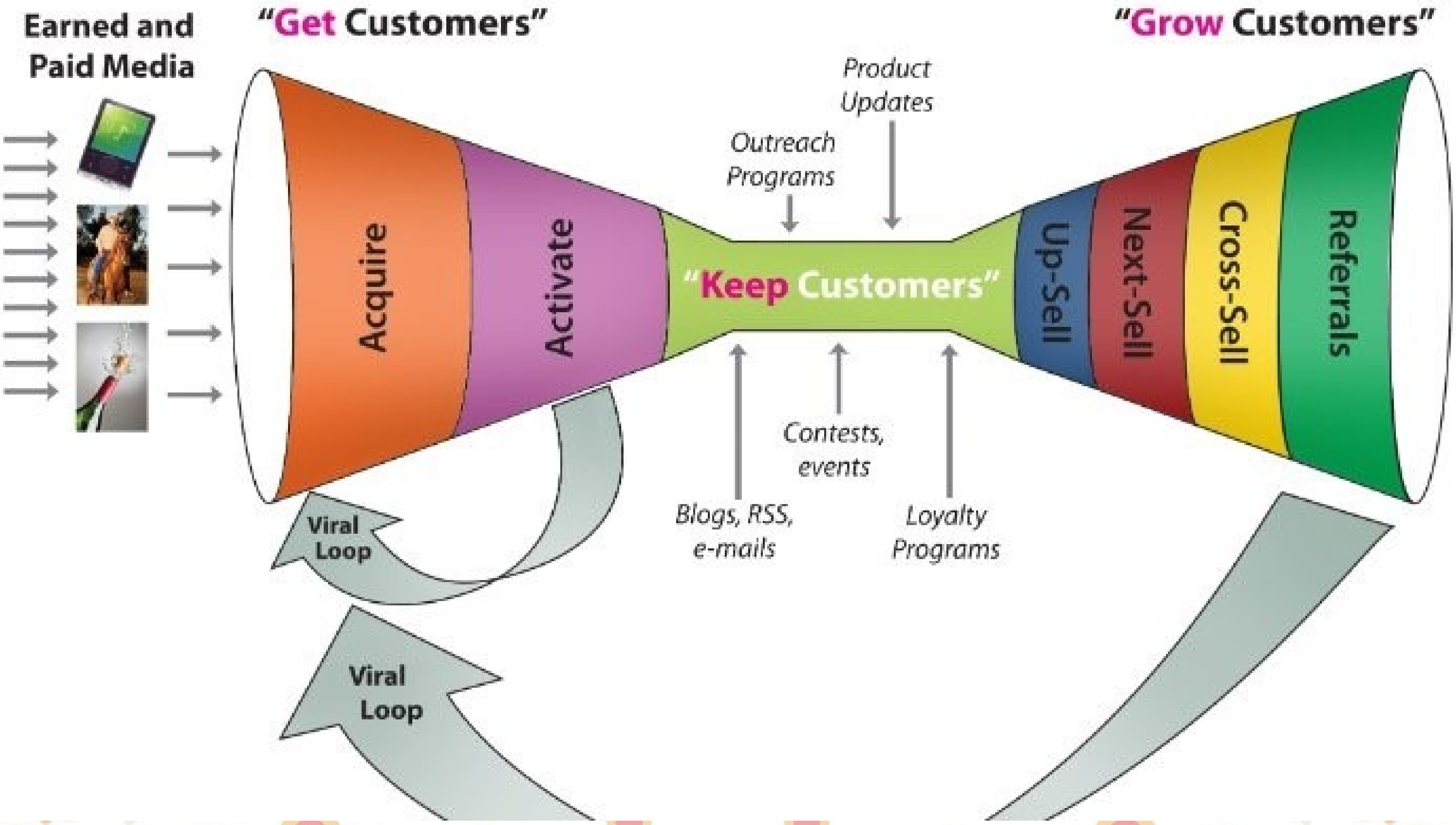
(CAQ - Customer Acquisition Cost)

Customer Retention:

Using CRM we can improve customer retention by building brand loyalty and develop the customer by increasing share of wallet through up-sell, cross-sell, cross-sell and referrals

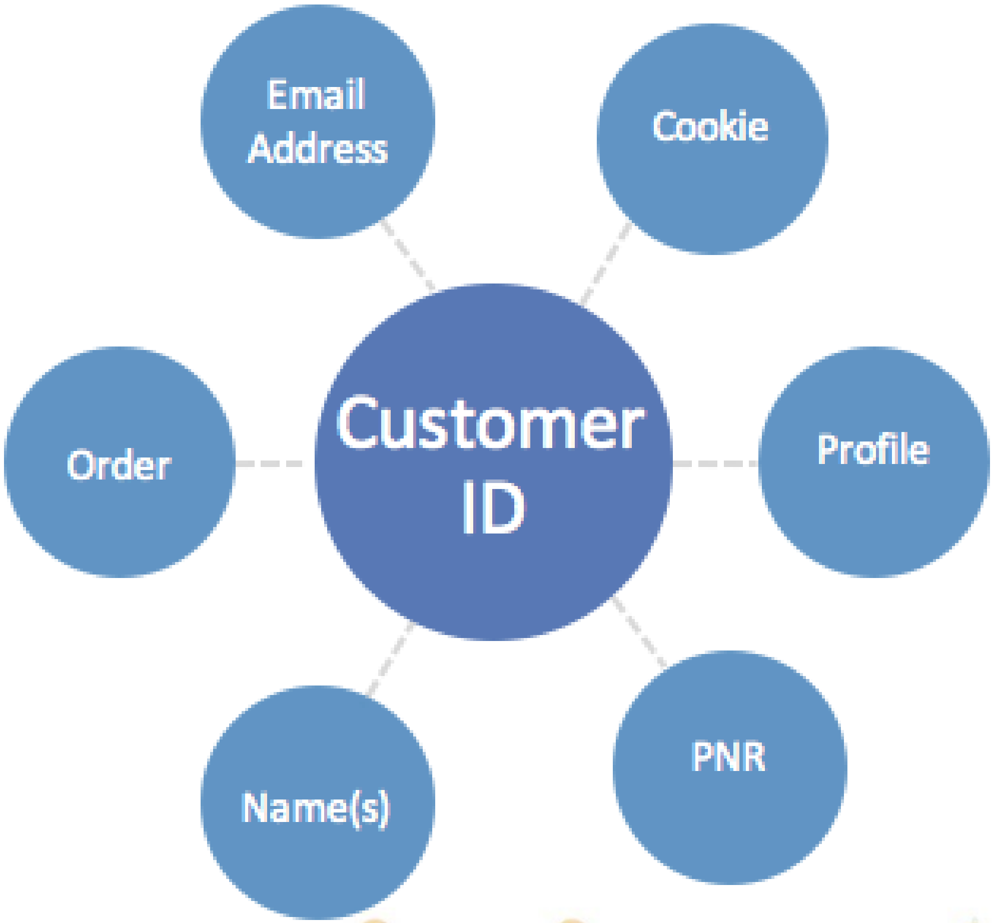
salesforce Infusionsoft. Microsoft

Sales and Marketing 2016



Sales and Marketing 1989

Sales and Marketing 2016



Identity Resolution

Sales and Marketing 2016

Brand Touchpoints

Sales and Marketing 2016

HOW IS THIS DONE?

Customer Segmentation

Customer segmentation is the practice of dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing, such as age, gender, interests and spending habits.

searchcrm.techtarget.com

Customer Segmentation is the subdivision of a market into discrete customer groups that share similar characteristics. Customer Segmentation can be a powerful means to identify unmet customer needs.

Bain and Company

Segmentation

Approaches:

1. RULE-BASED SEGMENTATION

Simple logic, user-focused, quick to define and execute

2. ALGORITHIMICALLY-BASED SEGMENTATION

More complex logic, typically different processing engine and analyst

Types:

1. BEHAVIOURAL SEGMENTATION

Demographic, geographic, technographic, psychographic

2. PREDICTIVE SEGMENTATION

Survival modelling, propensity modelling, Next Best Action

3. DYNAMIC SEGMENTATION

Mapping the migration of populations between segments

Rule-based Segmentation

- Usually SQL-based - Spark
- (Supposedly) Simple logic
- Business user-focused
- Application:
  - Quick to define and execute
  - Can be difficult to maintain
  - Need updating and can go stale

Rule-based Segmentation

```
CREATE  TABLE #WORKSHEETS (WorkId Int)

INSERT  #WORKSHEETS
SELECT  WorkId                       = WORS.WORK_ID
FROM    CARS.DBO.WORS_WorksheetsLeads           WORS
INNER   JOIN CARS.DBO.LEAD_Leads            LEADS ON LEADS.LEAD_ID = WORS.LEAD_ID AND WORS.WORK_ID IS NOT NULL
INNER   JOIN COM.DBO.CNTL_ContactCenterLeads        CNTL ON LEADS.LEAD_OldCode = CNTL.CNTL_ID

GO

SELECT
    LeadId                           = LEAD.LEAD_ID,
    AffinityGroupId                      = CU.AFFINITY_GROUP_ID,
    CntlId                           = CNTL.CNTL_ID,
    CntlPrimaryId                        = CNTL.CNTL_PrimaryID,
    CnttID                           = CONVERT(Int, ISNULL(CNTL_Type, 0)),
    CntcID                           = CNTC.CNTC_ID,
    Leads800                         = CASE WHEN (PHONE.CNTS_ID <> 7) THEN 1 ELSE 0 END,
    LeadsWarmTransfer                    = CASE WHEN (PHONE.CNRE_ID IS NOT NULL AND PHONE.CNTS_ID = 7) THEN 1 ELSE 0 END,
    LeadsWalkIn                      = CASE WHEN (WORK.WORK_ID NOT IN (SELECT WorkId FROM #WORKSHEETS) AND WORK.AGRA_ID = CU.AFFINITY_GROUP_ID) THEN 1 ELSE 0 END,
    LeadsDistributed                     =  CASE WHEN (CNTL.CNTL_Type IN ('0', '1', '2', '3', '4', '5', '6', '7', '8') AND CNTL.AGRP_ID = CU.OldId) THEN 1 ELSE 0 END,
    CntlDate                         = CNTL.CNTL_Datetime,
    WorkDate                         = WORK.WORK_CreatedDate,
    ResultYear                       = CONVERT(varchar, DATEPART(YY, WORK.WORK_CreatedDate), 101),
    ResultMonth                      = CONVERT(varchar, DATEPART(MM, WORK.WORK_CreatedDate), 101),
    ResultQuarter                        = CONVERT(varchar, DATEPART(QQ, WORK.WORK_CreatedDate), 101)

FROM    AIME.dbo.LEAD                       LEAD
LEFT    JOIN AIME.dbo.AFFINITY_GROUP                CU ON LEAD.AFFINITY_GROUP_ID = CU.AFFINITY_GROUP_ID
LEFT    JOIN CARS.dbo.WORK_Worksheets           WORK ON WORK.WORK_ID = LEAD.CARS#WORK_ID
LEFT    JOIN CARS.dbo.COWS_CodesWorksheetStatus     COWS ON WORK.COWS_ID = COWS.COWS_ID
LEFT    JOIN COM.dbo.CNTC_ContactCenterStatusCodes  CNTC ON LEAD.COM#CNTC_ID = CNTC.CNTC_ID
LEFT    JOIN AIME.DBO.DEALER                DEALER ON LEAD.DEALER_ID = DEALER.DEALER_ID
LEFT    JOIN CARS.DBO.EMPL_Employees            EMPL ON LEAD.CARS#EMPL_ID = EMPL.EMPL_ID
LEFT    JOIN COM.DBO.CNTL_ContactCenterLeads        CNTL ON LEAD.COM#CNTL_ID = CNTL.CNTL_ID
LEFT    JOIN COM.DBO.CNRE_ContactRequestEmailPhone  PHONE ON PHONE.CNTL_ID = CNTL.CNTL_ID AND CNTL.AGRP_ID = CU.OldId
LEFT    JOIN COM.DBO.CNRE_ContactRequestEmailPhone  EPHONE ON EPHONE.CNTL_ID = CNTL.CNTL_ID

WHERE WORK.WORK_CreatedDate <= '10/1/2014'
AND WORK.WORK_CreatedDate >= '1/1/' + CONVERT(varchar, DATEPART(yy, DATEADD(YY, -1, '10/1/2014')), 101)
AND CU.AFFINITY_GROUP_ID = 11765
```

Algorithmically-based Segmentation

- Unsupervised (Noun) and Supervised (Verb)
- (Supposedly) complex logic
- Analyst focused
- Application:
  - More precise
  - Requires processing framework - Spark

Behavioural Segmentation

- Demographic
  - Age, Background, Social Class, Economic Class...
- Geographic
  - Address, Current location
- Technographic (indicative of lifestyle)
  - Technology adoption - Mobile, Tablet, Browser ...
  - Orbitz found Apple users spend 30% more a night
- Psychographic
  - The 4-Cs - Cross Cultural Consumer Characterisation

Behavioural Segmentation

Demographic

The socio-economic scale

Social grade	Description of occupation	Example
A	higher managerial, administrative or professional	Company director
B	intermediate managerial, administrative or professional	Middle manager
C1	supervisory, clerical, junior administrative or professional	Bank clerk
C2	skilled manual workers	Plumber
D	semi- and unskilled manual workers	Labourer
E	state pensioners with no other income, widows, casual and lowest grade earners	Unemployed

Behavioural Segmentation

Geographic

ACORN

A. Agricultural areas	3% of UK population
B. Modern family housing, higher incomes	18% of UK population
C. Older housing of intermediate status	17% of UK population
D. Poor quality older terraced housing	4% of UK population
E. Better-off council estates	13% of UK population
F. Less well-off council estates	9% of UK population
G. Poorest council estates	7% of UK population
H. Multi-racial areas	4% of UK population
I. High status non-family areas	4% of UK population
J. Affluent suburban housing	16% of UK population
K. Better-off retirement areas	4% of UK population
U. Unclassified	1% of UK population

Behavioural Segmentation

Psychographic

The 4Cs Cross Cultural Consumer Characterisation

Resigned	Rigid, strict, authoritarian and chauvinist values, oriented to the past and to Resigned roles. Brand choice stresses safety, familiarity and economy. (Older)
Struggler	Alienated, Struggler, disorganised - with few resources apart from physical/mechanical skills (e.g. car repair). Heavy consumers of alcohol, junk food and lotteries, also trainers. Brand choice involves impact and sensation.
Mainstreamer	Domestic, conformist, conventional, sentimental, passive, habitual. Part of the mass, favouring big and well-known value for money 'family' brands. Almost invariably the largest 4Cs group.
Aspirer	Materialistic, acquisitive, affiliative, oriented to extrinsics ... image, appearance, charisma, persona and fashion. Attractive packaging more important than quality of contents. (Younger, clerical/sales type occupation)
Succeeder	Strong goal orientation, confidence, work ethic, organisation ... support status quo, stability. Brand choice based on reward, prestige - the very best . Also attracted to 'caring' and protective brands ... stress relief. (Top management)
Explorer	Energy - autonomy, experience, challenge, new frontiers. Brand choice highlights difference, sensation, adventure, indulgence and instant effect - the first to try new brands. (Younger - student)
Reformer	Freedom from restriction, personal growth, social awareness, value for time, independent judgement, tolerance of complexity, anti-materialistic but intolerant of bad taste. Curious and enquiring, support growth of new product categories. Select brands for intrinsic quality, favouring natural simplicity, small is beautiful.(Higher Education)

Behavioural Segmentation

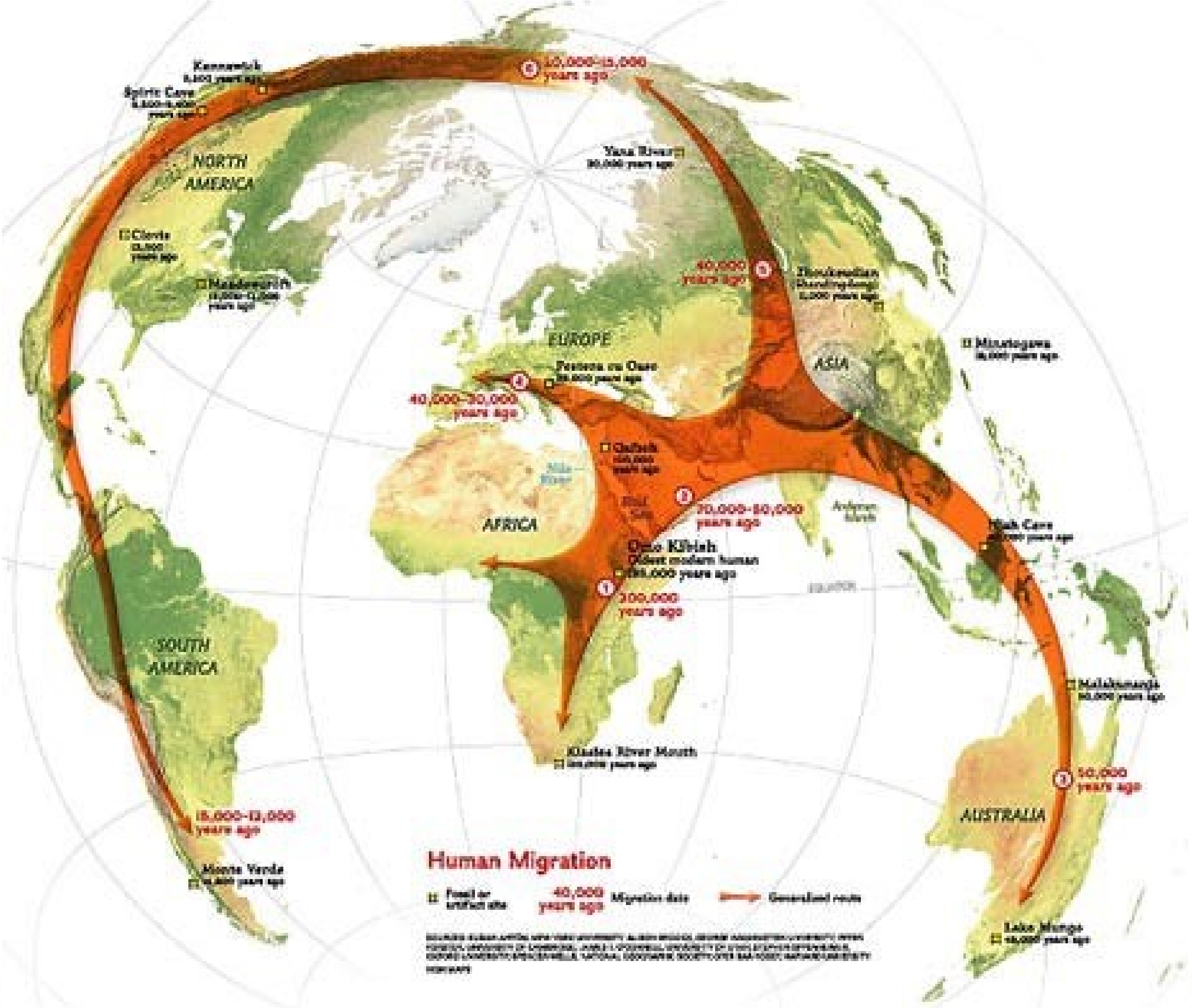
Technographic

Technology attitude		Primary motivation: Career (25%)	Primary motivation: Family (24%)	Primary motivation: Entertainment (22%)
Technology optimist (50%)	High income	**Fast Forwards** are time-strapped, driven, and top users of technology. 13%	**New Age Nurturers** are underserved believers in technology for family and education. 8%	**Mouse Potatoes** are dedicated to interactive entertainment, especially on the PC. 9%
Technology optimist (50%)	Low income	**Techno-Strivers** are up-and-coming believers in technology for career advancement. 6%	**Digital Hopefuls** are family-oriented technology lovers: a promising market for low-cost PCs. 7%	**Gadget Grabbers** are focused on low-cost, high-tech toys like MP3 players and portable game players. 8%
Technology pessimist (45%)	High income	**Handshakers** are successful professionals with a low technology tolerance. 6%	**Traditionalists** are suspicious of technology beyond the basics. 9%	**Media Junkies** are visual TV lovers and interested in TV features like video on demand. 5%
Technology pessimist (45%)	Low income	**Sidelined Citizens** are technophobes and technology laggards, the least receptive audience for any technology or digital channel. 25%		

Predictive Segmentation

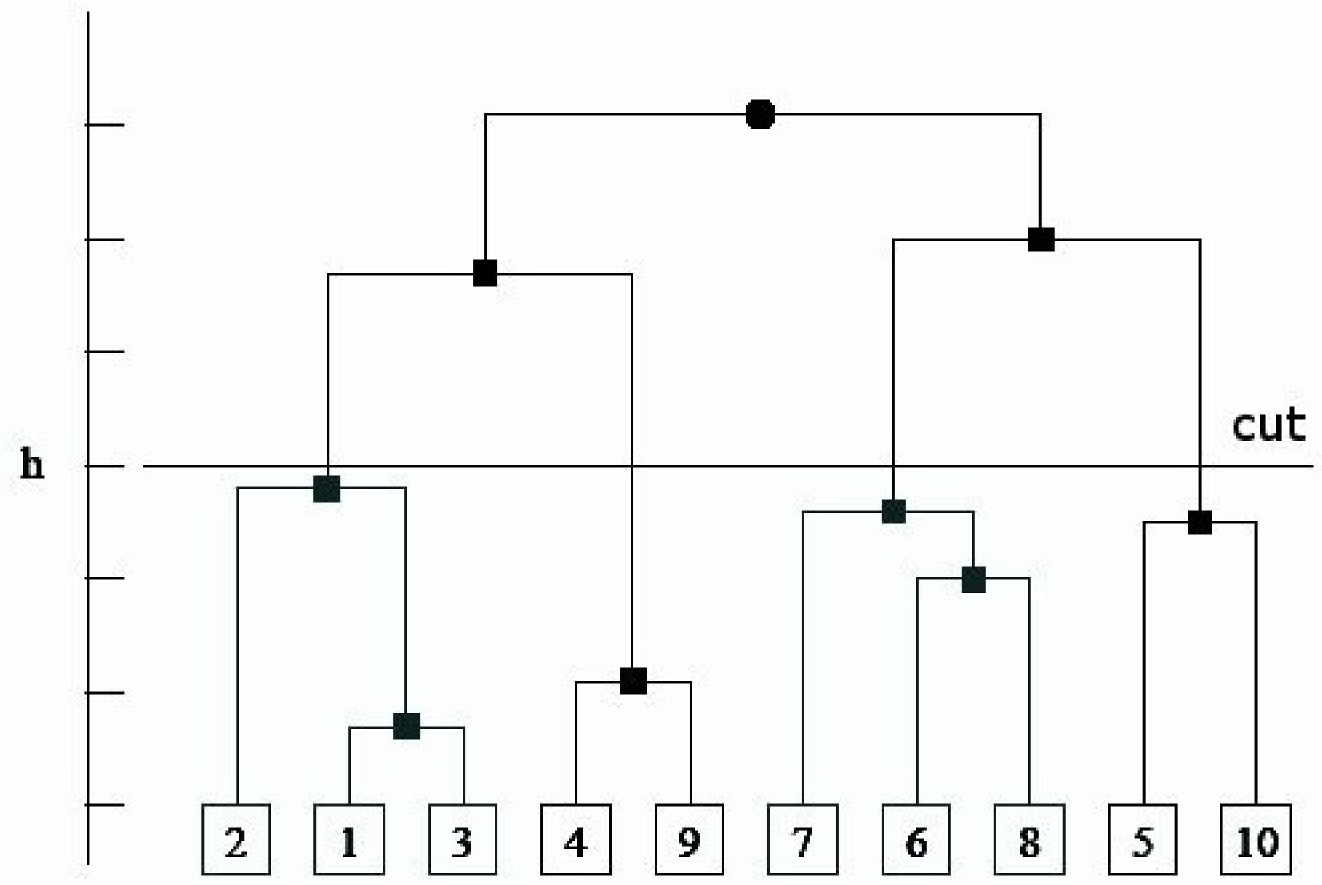
- Survival or Churn modelling
  - Mobile Phone Networks, Social Networks, Brands
  - Classification Algorithms - churn / not churn
- Propensity modelling
  - Propensity / Likelihood of someone to buy
- Next Best Action
  - Offers, Cross sell, Up sell, Service Recovery

Dynamic Segmentation



Segmentation Strategies

Segmentation Strategies



Personalisation

Automated Customer Engagement - NBA, Service Recovery

From the web - Amazon, Booking, Spotify, Hotels, RyanAir

To the real world - In Cabin, Airport Shopping

MACHINE LEARNING WITH SPARK

Machine Learning with Spark

- MLIB Version 2.10s
- Comprehensive suite of algorithms
- One infrastructure to address a wide range of problems

Machine Learning with Spark

- Basic Data Types
- Basic Statistics
- Supervised
  - Classification and Regression
- Unsupervised
  - Collaborative Filtering (Matrix Factorisation)
  - Clustering
  - Dimensionality Reduction
- Frequent Pattern Matching
- Evaluation
- PMML - Publishing and Sharing

Machine Learning with Spark

- Data Types
  - Local Vector
    - Vector on a single machine
    - Sparse and Dense Vector types
  - Labelled point
    - Vector associated with a label/response
  - Local Matrix
    - Dense and Sparse Matrices supported
  - Distributed Matrix
    - RowMatrix, IndexedRow Matrix, Coordinate Matrix

Machine Learning with Spark

- Data Statistics
  - Summary Statistics

```
val summary: MultivariateStatisticalSummary = Statistics.colStats(observations)
println(summary.mean)
```

  - Correlations

```
val correlation: Double = Statistics.corr(seriesX, seriesY, "pearson")
```

  - Stratified Sampling - splits automatically on a given level

```
val approxSample = data.sampleByKey(withReplacement = false, fractions)
```

  - Hypothesis Testing
  - Random data generation

```
val u = normalRDD(sc, 1000000L, 10)
```

Machine Learning with Spark

- Supervised
  - Classification (Binary and Multinomial)
    - Linear SVM
    - Logistic Regression
    - Decision trees (random forests)
  - Regression
    - Linear Regression (Linear least squares)
    - Lasso (Support regularisation and feature selection)
    - Ridge Regression (Addresses Co-linearity)
    - Streaming linear regression (A version of online learning)
    - Decision Tree Regression

Machine Learning with Spark

- Unsupervised
  - Collaborative Filtering (Matrix Factorisation)
    - Supports user-item recommendation
    - Supports Implicit and Explicit Feedback
  - Clustering
    - K-means
    - Gaussian Mixture Models
    - Latent Dirichlet allocation (Topic Modelling)
    - Bisecting k-means, Streaming K-means
  - Dimensionality Reduction
    - PCA
    - SVD

Machine Learning with Spark

- Evaluation
  - Classification
    - Confusion Matrix
    - Precision (Positive Predictive Value)
    - Recall (True Positive Rate)
    - F-measure
    - Receiver Operating Characteristic (ROC)
    - Area Under ROC Curve
  - Regression
    - Mean Squared Error (MSE)
    - Root Mean Squared Error & Mean Absolute Error
    - Coefficient of Determination (R2)

Machine Learning with Spark

- Publishing and Sharing
  - PMML Format
    - An XML-based predictive model interchange format
    - Enabled publishing and sharing models (zementis, r, Oracle)
  - Models
    - K-means
    - Linear Regression Model
    - Ridge Regression Model
    - Lasso Model
    - SVM Model
    - Logistic Regression Model

SUMMARY

Summary

Preliminaries
1. Use Case: Personalisation
   - Sales and Marketing
   - Customer Segmentation
2. Machine Learning with Spark
   - Basic Data Types
   - Basic Statistics
   - Supervised
   - Unsupervised
   - Evaluation
   - PMML - Publishing and Sharing
4. Summary

Further Reading

MLIB Release:
https://databricks.com/blog/2014/07/16/new-features-in-mllib-in-spark-1-0.html

Lasso Explained:
https://www.youtube.com/watch?v=qU1_cj4LfLY

Ridge Regression Explained:
http://www.ncss.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Ridge_Regression.pdf

Netflix Prise using Matrix Factorisation:
http://dx.doi.org/10.1007/978-3-540-68880-8_32

Gaussian Mixture Models
https://www.youtube.com/watch?v=Rkl30Fr2S38

Questions

?