ARCADA

YRKESHÖGSKOLAN ARCADA

Introduction to Analytics

Magnus.Westerlund @arcada.fi

Researcher & Programme Director
01.04.2017

ARCADA

Intro to Analytics - Course Schedule

- Week 1
  - 6.9: Intro to Analytics, Machine Learning, and AI
  - 7.9: Feature engineering, Pandas
- Week 2
  - 20.9: Time series processing, linear modeling and setting targets/labels
  - 21.9: Time series data visualization and regression
- Week 3
  - 4.10: Understanding model output, and going from output to decision
  - 5.10: Open discussion, creating decisions, finalizing project

Todays Agenda

- Big Data Analytics (cont.)
- A note on Output Quality for forecasting
- Constructing software
- **Assignment, dl 19.4** (See the end slides)

ARCADA

Definitions

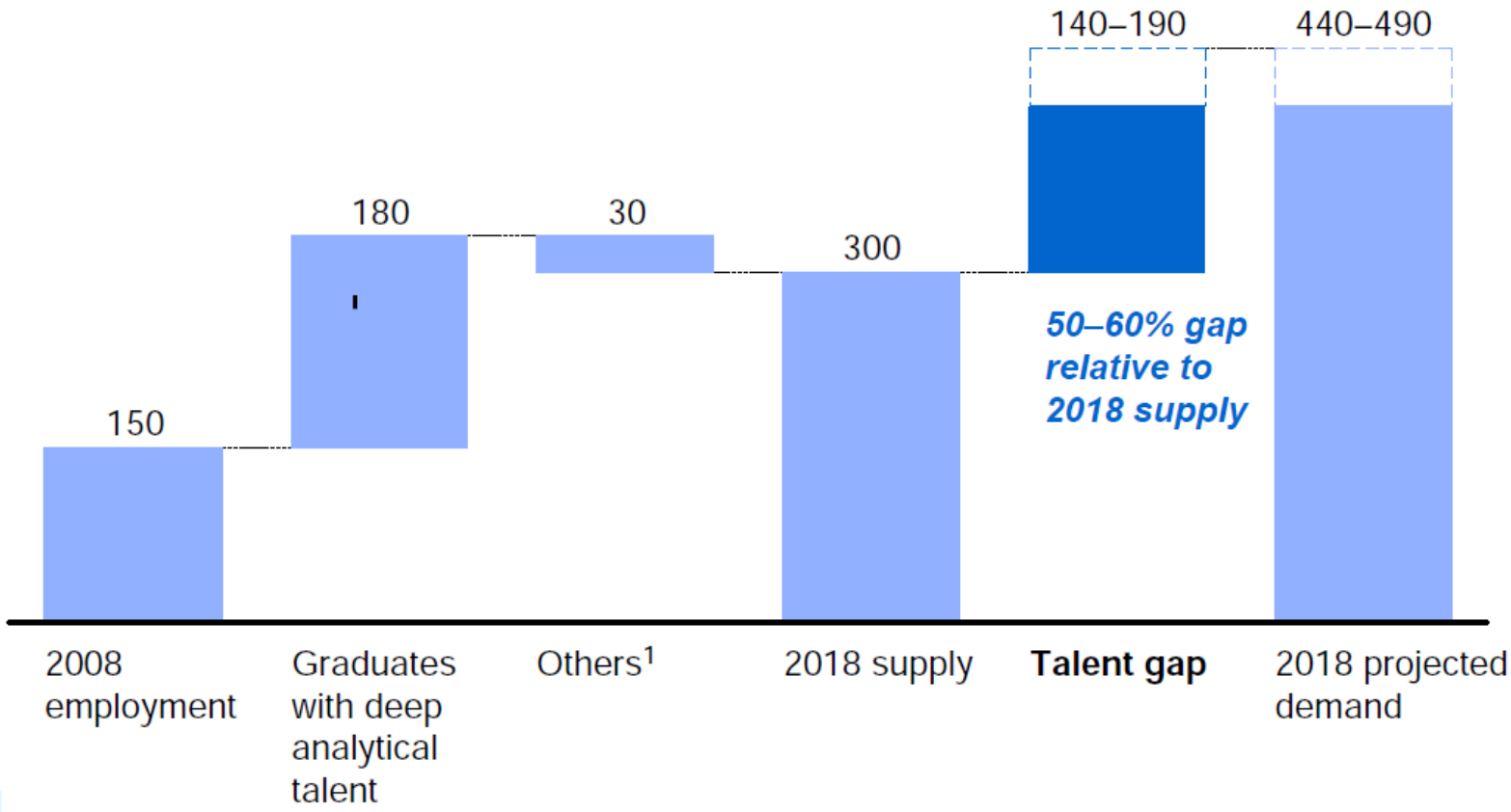
Short reflection on yesterday

Analytics Skills Gap

Demand for deep analytical talent in the United States could be 50 to 60 percent greater than its projected supply by 2018

Supply and demand of deep analytical talent by 2018

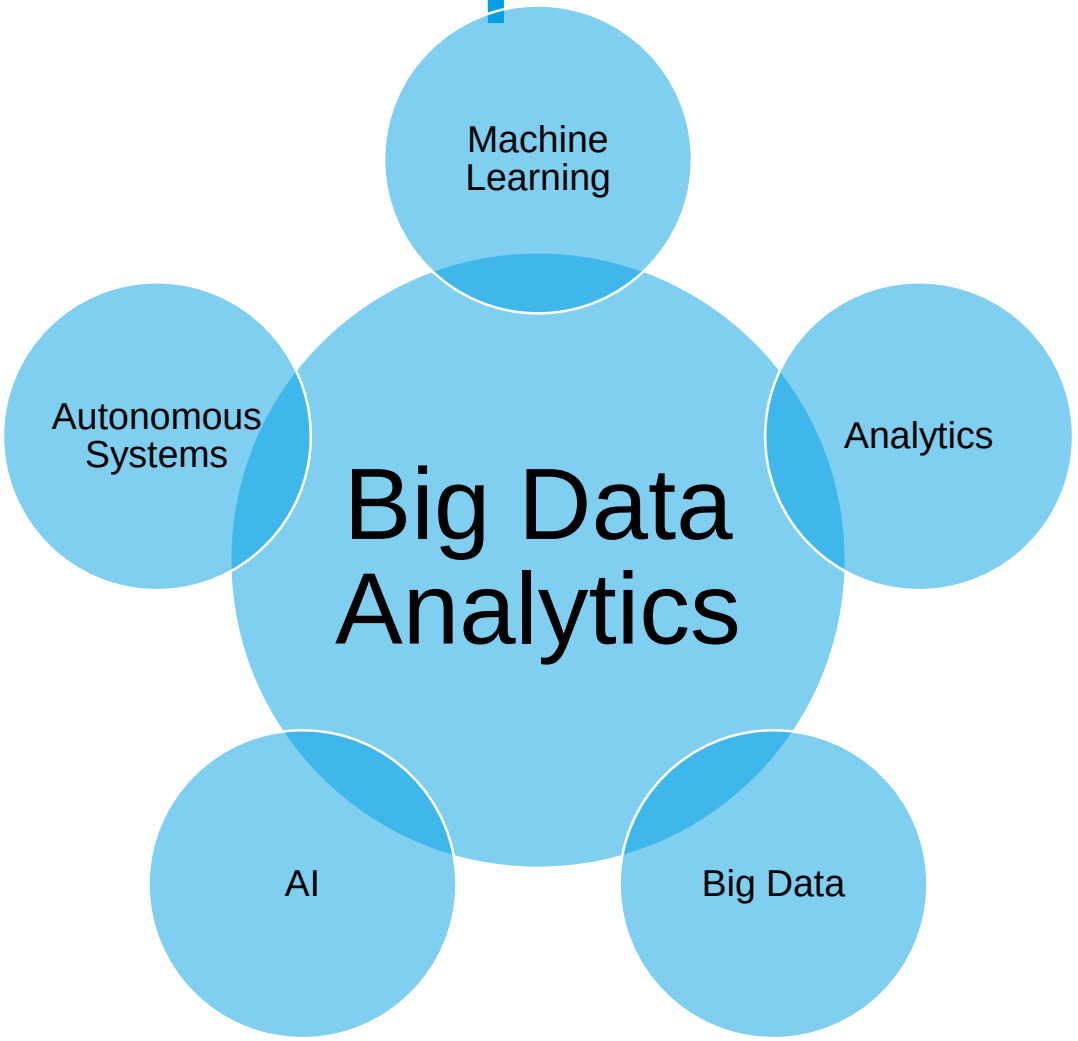
Thousand people



1 Other supply drivers include attrition (-), immigration (+), and reemploying previously unemployed deep analytical talent (+).

SOURCE: US Bureau of Labor Statistics; US Census; Dun & Bradstreet; company interviews; McKinsey Global Institute analysis

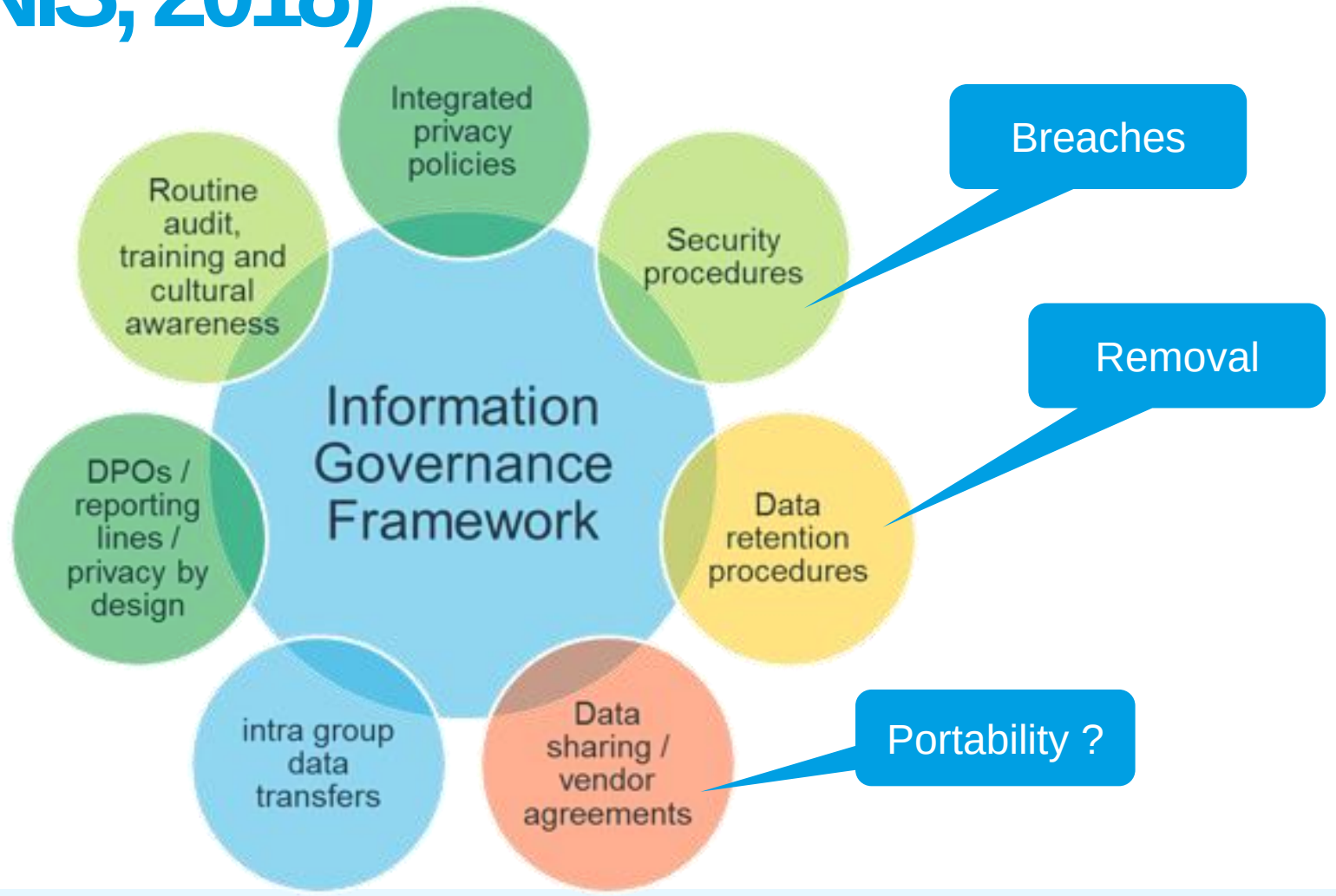
The field and some important terms



Understanding data through descriptive statistics: completeness and representativness

	#	Survived	Male Survived	Female Survied	Avg Age
Master	76	58%	58%		
Mr.	915	16%	16%		
Miss.	332			71%	21.8
Mrs	235			79%	36.9
Military	10	40%	40%		36.9
Clergy	12	0%	0%		41.3
Nobility	10	60%	33%	100%	41.2
Doctor	13	46%	36%	100%	43.6

EU Data Protection (GDPR, 2018) and Network Security (NIS, 2018)



GDPR design guidelines

- The GDPR defines two types of data, personal and special category (sensitive) data, any other data is not covered be the Regulation
- An identifiable natural person is one who can be **identified, directly or indirectly,** in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural, or social identity of that natural person (GDPR, art. 4(1)).

- Design requirements
  - requiring freely given consent or
  - by contract or for the performance of a contract
  - data minimization
  - data protection by design and default.
- Handling of personal data
  - access
  - rectification
  - portability
  - erasure.
- Limitations on processing
  - notification
  - restriction
  - security.

My Doctoral thesis:
http://www.doria.fi/handle/10024/155519

Automation of Society

- Hal Varian, chief economist for Google, envisions a future with fewer ‘jobs’ but a more equitable distribution of labor and leisure time. “If ‘displace more jobs’ means ‘eliminate dull, repetitive, and unpleasant work,’ the answer would be yes. How unhappy are you that your dishwasher has replaced washing dishes by hand, your washing machine has displaced washing clothes by hand, or your vacuum cleaner has replaced hand cleaning?

ARCADA

Evaluating and maintaining results; Do we trust the output?

How do we determine Quality from the Output of Analytical Services

Cause and Effect relationship

- Causality refers to a relationship that connects one process (the cause) with another process or state (the effect), where the first is understood to be partly responsible for the second, and the second is dependent on the first.
  - Causality is often considered to be temporally bound, as cause precede their effects
- Statistical effect size (a measure of the strength of a phenomenon)
  - E.g. the correlation between two variables

Validation of results

- How do you determine that you have found a cause-effect relationship?
  - Does it hold up when examining historic data?
    - If its partially true, does it provide any value?
  - If you do a test run for a week, are the results still valid?
  - What are the guarantees that the results will remain valid for system lifetime?
- System requirements
  - How are data faults handled?
  - How are effects of infrastructure failures mitigated?

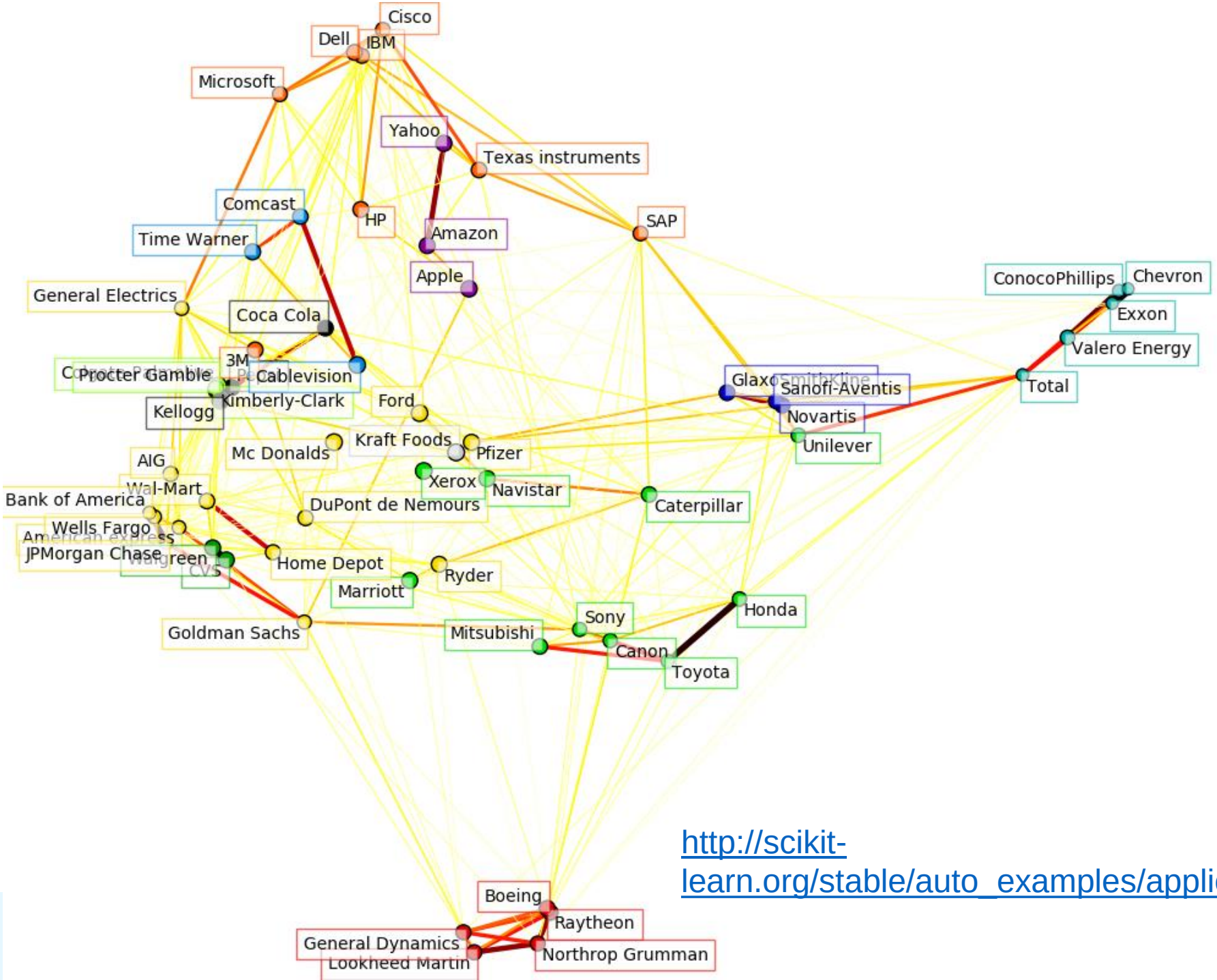
Variable Interaction

- Often we find that there is an interaction between input variables that e.g. measure a system
  - Consider the example of how to determine the amount of CO2 released by a car, for the purpose of taxation based on CO2 levels.
    - Can CO2 be measured correctly?
    - What are factors that influence the result, given that we can determine what to measure?
  - Consider genetics, how do we determine conclusively what relates to environmental changes and what is predestined in our DNA?
- Determining what inputs that capture the whole phenomena is crucial in analytics, and particularly so when predicting.

https://en.wikipedia.org/wiki/Interaction_(statistics)

Example: Stock Market Structure

- Financial markets are by many considered to be interconnected.
  - One example of this is that stocks in a certain industry tend to move together (see figure in next slide).
  - Another example is that nearby/related economies often tend to move in the same direction, although there might be a temporal delay.
- We often divide the relation into a spatial relation or a temporal relation.
  - Spatial, specifies how some object is located in space in relation to some reference object (e.g. a near-direct relation in stock price changes)
  - Temporal, separates the time dimension (relation over time)



http://scikit-learn.org/stable/auto_examples/applications/plot_stock_market.htm

ARCADA

Problem:
Prognosis based on Stock Market Structure

- There is a statistical effect, but can we say with certainty that the relation is based on a cause and effect relationship?
  - If any of the stocks in the energy cluster moves, will the others follow?
  - How was data selected, at what resolution, for what timeframe?
  - Is there a temporal shift that can be used for trading?
- What is the difference between the Japanese cluster and e.g. The energy cluster?
- What other type of clusters do you find?

ARCADA

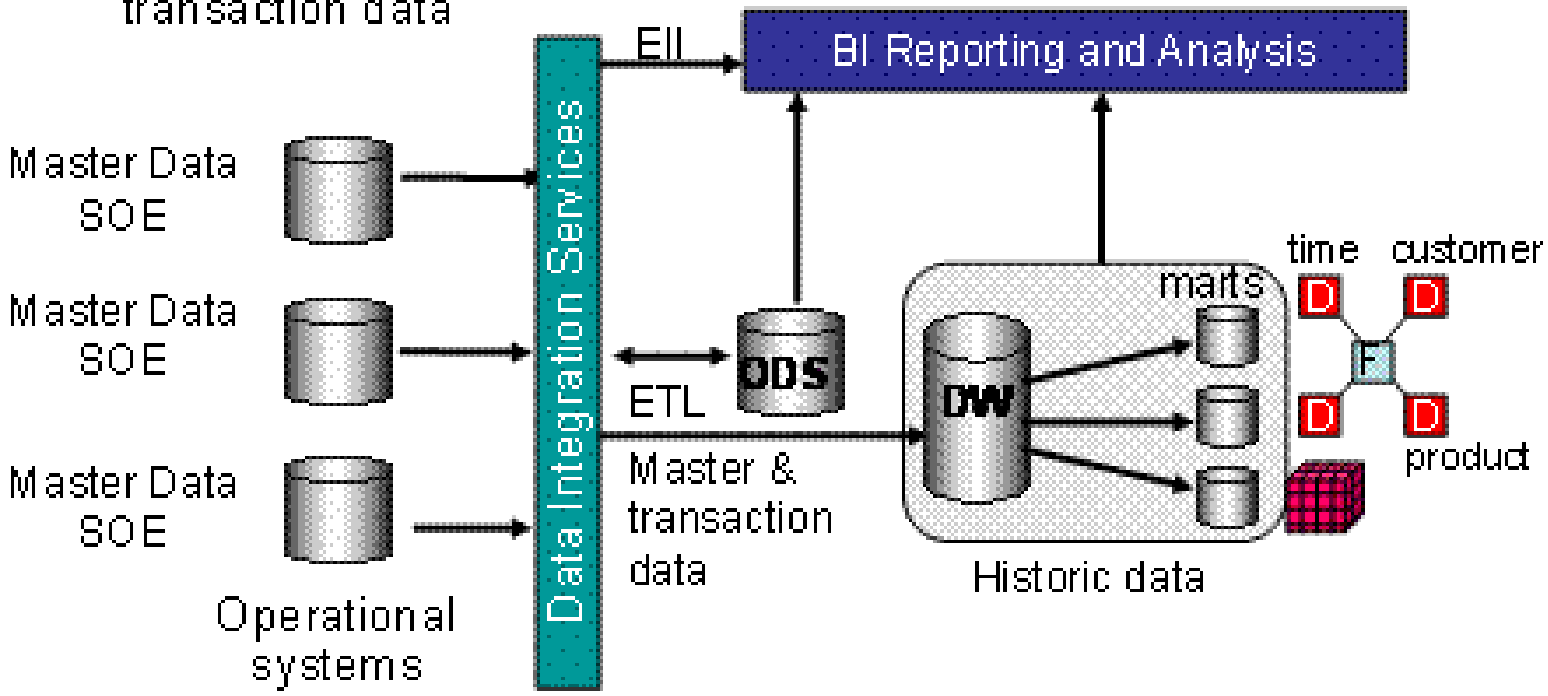
Constructing the software

Input data from a master source

- Data may be collated and distributed to/from other systems. This includes:
  - Data consolidation – capturing data from multiple sources and integrating
  - Data federation – a single virtual view of data from one or more sources
  - Data propagation – copying master data from one system to another

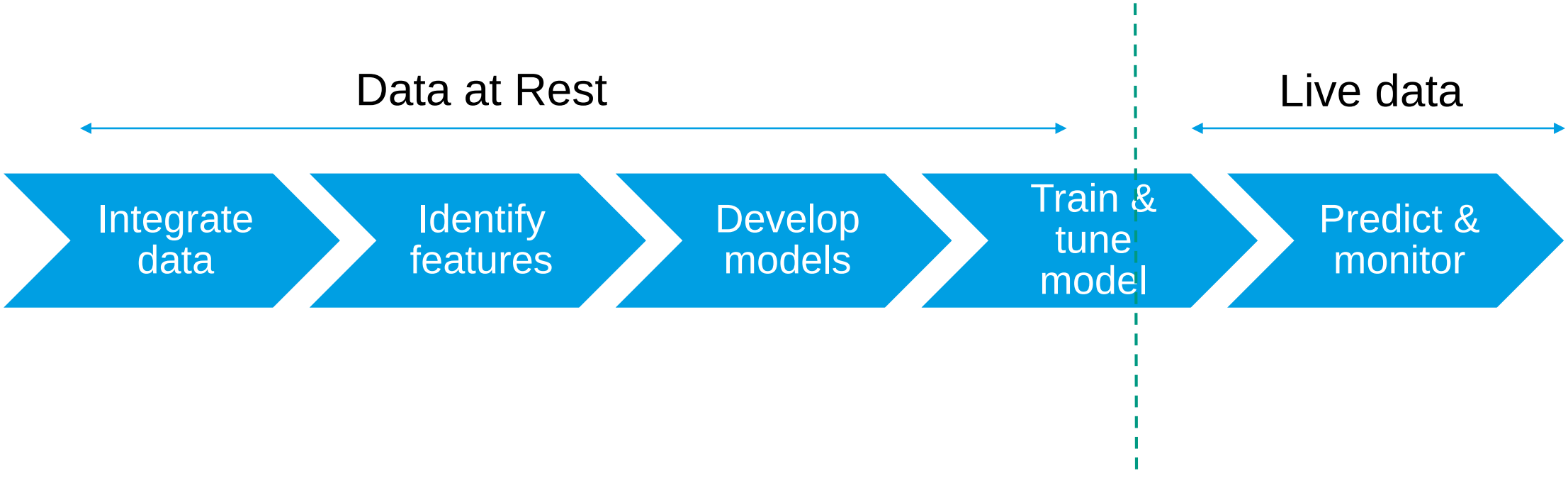
Master Data Integration in a Data Warehouse

- Data warehousing uses data integration tools to integrate disparate master data maintained in multiple operational systems to build DIMENSIONS
- Data warehousing also uses data integration to integrate transaction data

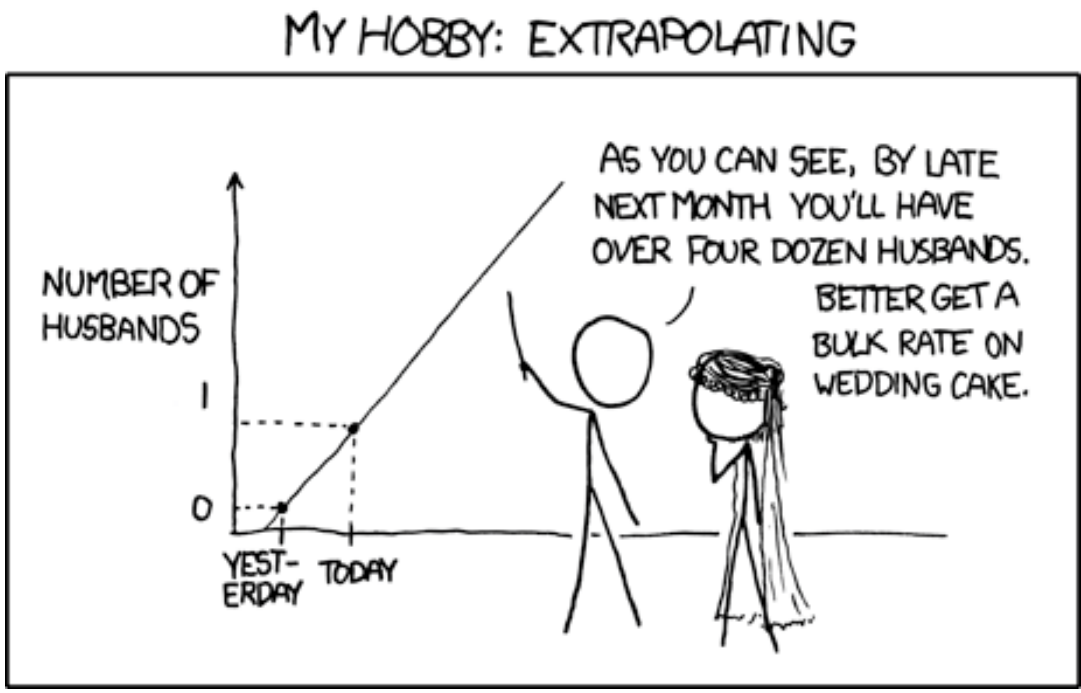


SOE = System of Entry

The analytics process, for prediction

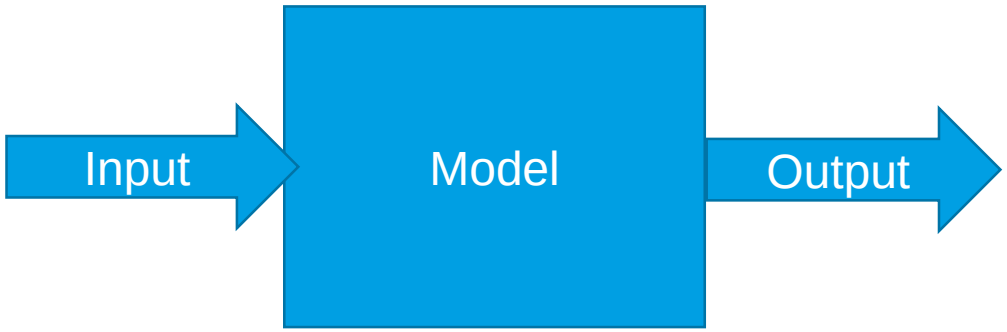


General idea of modeling



- Assume a distribution p(X,Y).
  - X : input
  - Y : output

- Given multiple features
  - X_i : one input feature
  - $X_{i,t}$: one input feature, at a time or index
- Given multiple outputs
  - Y_i : one output type
  - $Y_{i,t}$: one output, at a time or index



Feature Engineering

- Human concepts to model inputs
  - image → pixels, contours, textures, etc.
  - signal → samples, spectrograms, etc.
  - time series → ticks, trends, reversals, etc.
  - biological data → dna, marker sequences, genes, etc.
  - text data → words, grammatical classes and relations, etc.

Feature Relevance

- Some features hold more information than others, how to determine which to use.
  - Strongly relevant feature
    - Feature Xi brings information that no other feature contains.
  - Weakly relevant feature
    - Feature Xi brings information that also exists in other features.
    - Feature Xi brings information in conjunction with other features.
  - Irrelevant feature
    - Feature Xi is neither strongly relevant nor weakly relevant.

Feature Selection

- The selection is often a heuristic process (a discovery that employs a practical method).
- Can be a difficult and time consuming process, as you may have to test many combinations and variations.
- Feature relevance may in semi-chaotic processes change over time, e.g. price movements on a financial market.
- Feature selection should always be part of a validation and verification process, i.e. once the system is in use, you may need to maintain a certain measure for continued feature relevance.

Some techniques for Feature Selection

- Forward selection
  - Start with empty set of features.
  - Incrementally add features Xt.
  - Will find all strongly relevant features. May not find some weakly relevant features.
- Backward selection
  - Start with full set of features .
  - Incrementally remove features Xi.
  - Will keep all strongly relevant features. May eliminate some weakly relevant features (e.g. redundant).
- You may perform an exhaustive search through all the subsets of features, but finding all relevant features is NP-hard.

Feature selection is difficult to get right

- You may ask, why bother with feature selection if its so hard?
- Should we not give our model all the data that exist?

> *Feature selection is itself useful, but it mostly **acts as a filter**, muting out features that aren't useful in addition to your existing features.* — Robert Neuhaus

- In this guide you can find more info and links: http://machinelearningmastery.com/an-introduction-to-feature-selection/

Answer to why we only want relevant features as inputs

- A model will find it difficult to learn from noisy and/or irrelevant data.
- The more features we use, it will also make the learning process computationally more complex.
- We consider each input as its own dimension.

- However, we can also try to reduce dimensions.
  - This works for linear problems, but not really for non-linear problems.

ARCADA

Programming

Development requirements for the assignment

- Python, Pandas, NumPy, and a bit of statistics

- Here is a link to a Pandas cheat sheet, also this site is an excellent resource for practical applications of analytics:
  - http://www.datasciencecentral.com/profiles/blogs/data-science-in-python-pandas-cheat-sheet

Data structures

- Arrays:

  ['2013-01-01', '2013-01-02', '2013-01-03', '2013-01-04', '2013-01-05', '2013-01-06']

  – Similar to other languages, but there are something a lot better..

- Dataframes from Pandas:

  # We create a DataFrame with random values, the above dates as indexes and A,
  # B C an D as columns“

  import pandas as pd

  df = pd.DataFrame(np.random.randn(6,4), index=dates, columns=list('ABCD'))

```
df.head() # We can use .head() to show the n first columns, 5 by default
```

	A	B	C	D
2013-01-01	-0.305386	-0.113169	-0.453408	-0.108692
2013-01-02	-1.212049	2.071357	-0.326874	-0.206681
2013-01-03	-0.466048	0.384297	0.616386	0.385557
2013-01-04	-0.531123	-1.775777	1.113464	0.309822
2013-01-05	-2.055577	0.581261	-0.496861	-0.789066

Pandas

- Pandas is a software library written for the Python programming language. It is used for data manipulation and analysis. It provides special data structures and operations for the manipulation of numerical tables and time series. The two main data structures are DataFrames and Series.
- A Series is a one-dimensional labelled array-like object. The underlying idea of a DataFrame is based on spreadsheets. We can see the data structure of a DataFrame as tabular and spreadsheet-like. It contains an ordered collection of columns.
- Documentation: http://pandas.pydata.org/pandas-docs/stable/index.html
- Introduction: http://pandas.pydata.org/pandas-docs/stable/10min.html
- If you prefer a video tutorial: https://www.youtube.com/watch?v=Iqjy9UqKKuo&list=PLQVvvaa0QuDc-3szzjeP6N6b0aDrrKyL-

Data Loading

- Files, you can load datasets from different types of files, text (.csv, .txt), different kind of delimiters (comma, tab, pipe), and table based such as .xls
  - You use different methods (read_csv, read_table)

```
import pandas as pd
data = pd.read_csv('E:/data/prices.csv')
```

Pandas help

- Start out with our guide that will help you with the project

- Here is also another data wrangling cheet sheet:
- https://github.com/pandas-dev/pandas/blob/master/doc/cheatsheet/Pandas_Cheat_Sheet.pdf

Plotting graphs

```
import matplotlib
import matplotlib.pyplot as plt
```

```
# Select a column using df["Column name"] or df.column_name,
# .plot() automatically creates a plot of the data
```

```
df["Adj Close"].plot(figsize=(12,8))
plt.show()
```

NumPy

- NumPy is an acronym for "Numeric Python" or "Numerical Python". It is an open source extension module for Python, which provides fast precompiled functions for mathematical and numerical routines. Furthermore, NumPy enriches the programming language Python with powerful data structures for efficient computation of multi-dimensional arrays and matrices. The implementation is even aiming at huge matrices and arrays. Besides that the module supplies a large library of high-level mathematical functions to operate on these matrices and arrays.
- Introduction to NumPy: https://docs.scipy.org/doc/numpy/user/quickstart.html
- Full documentation: https://docs.scipy.org/doc/numpy/reference/

Calculations

- You perform calculations per column

```
import numpy as np
#df["DPC"] = np.log(df["Adj Close"].iloc[1:] / df["Adj Close"].iloc[:-1].values)
df['Log_Ret'] = np.log(df["Adj Close"] / df["Adj Close"].shift(1))
```

ARCADA

Assignment 1

Assignment 1

- See Itslearning for the assignment.
  - arcada.itslearning.com
- DL is 19.9.

Questions?

- The end..