FÖRBÄTTRA DIN PREDIKTIVA MODELLERING MED MACHINE LEARNING I SAS ENTERPRISE MINER

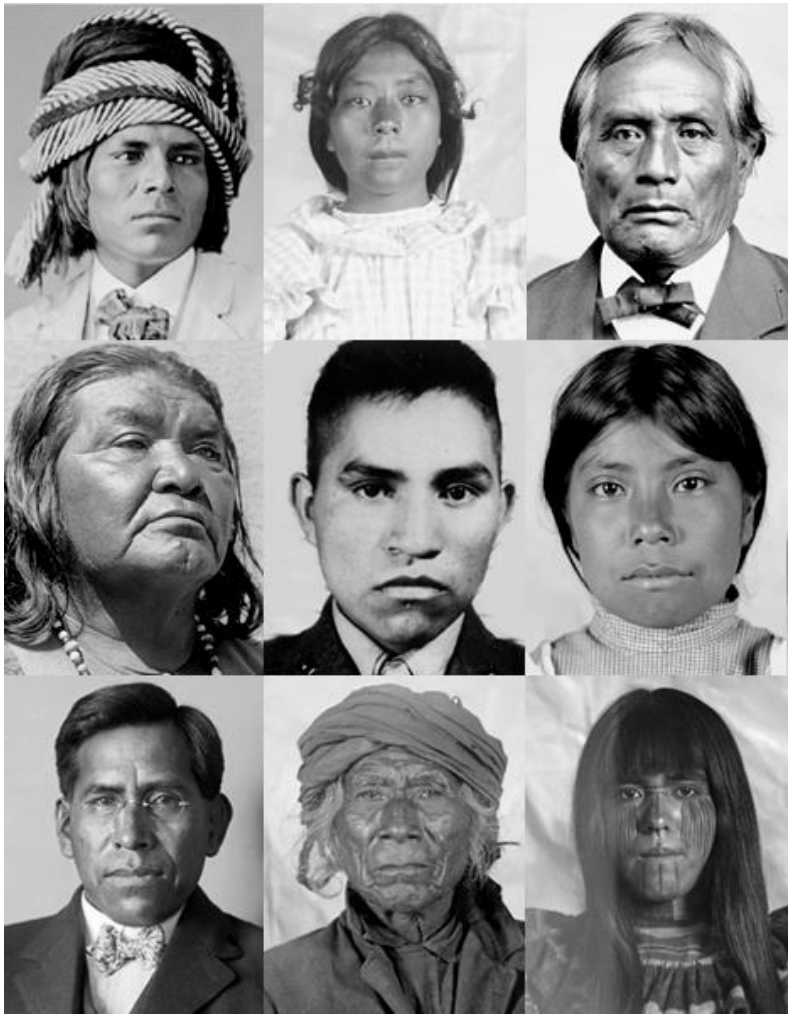
OSKAR ERIKSSON - ANALYSKONSULT

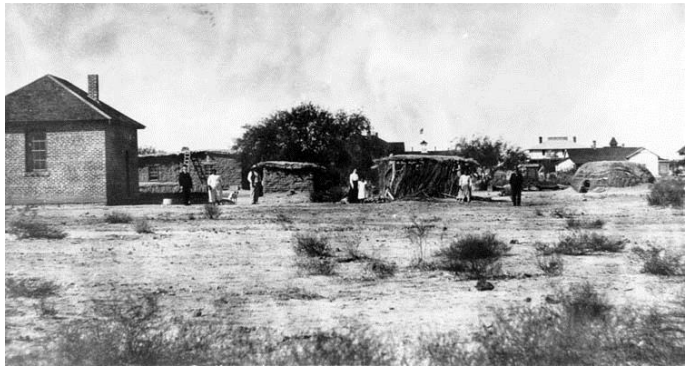
VEM ÄR JAG?

VAD SKA VI GÖRA?

- Pimafolket
  - Vilka då?
- Diabetes
  - Typ 2
- Regressionsanalys
- Machine Learning

PIMAFOLKET | AKIMEL O'OTHAM – "FLODFOLKET"





PIMAFOLKET

AKIMEL O'OTHAM – "FLODFOLKET"

- Bor i södra och centrala Arizona
- Antal: 15 – 25 000 enligt 2010 års skattning
- Pima ~ "pi mac" till européerna
- Kända för Pimarevolten 1751, under ledning av Luis Oacpicagigua
- Källa: Wikipedia

PIMAFOLKET | KÄNDISAR

PIMAFOLKET KÄNDISAR

PIMAFOLKET

AKIMEL O'OTHAM – "FLODFOLKET"

- Högst prevalens av typ 2-diabetes i världen
- Sannolikt kostrelaterat
- Deras kamrater av samma folk i Mexico har inte samma prevalens
- Kuriosa: Från 10 års ålder till äktenskap får de inte uttala sina egna namn.
  - Genomsnittlig ålder för första bröllop i Sverige (SCB 2014)
    - Kvinnor: 33,2 år (samkönat: 34,6)
    - Män 35,7 (samkönat: 42,3)

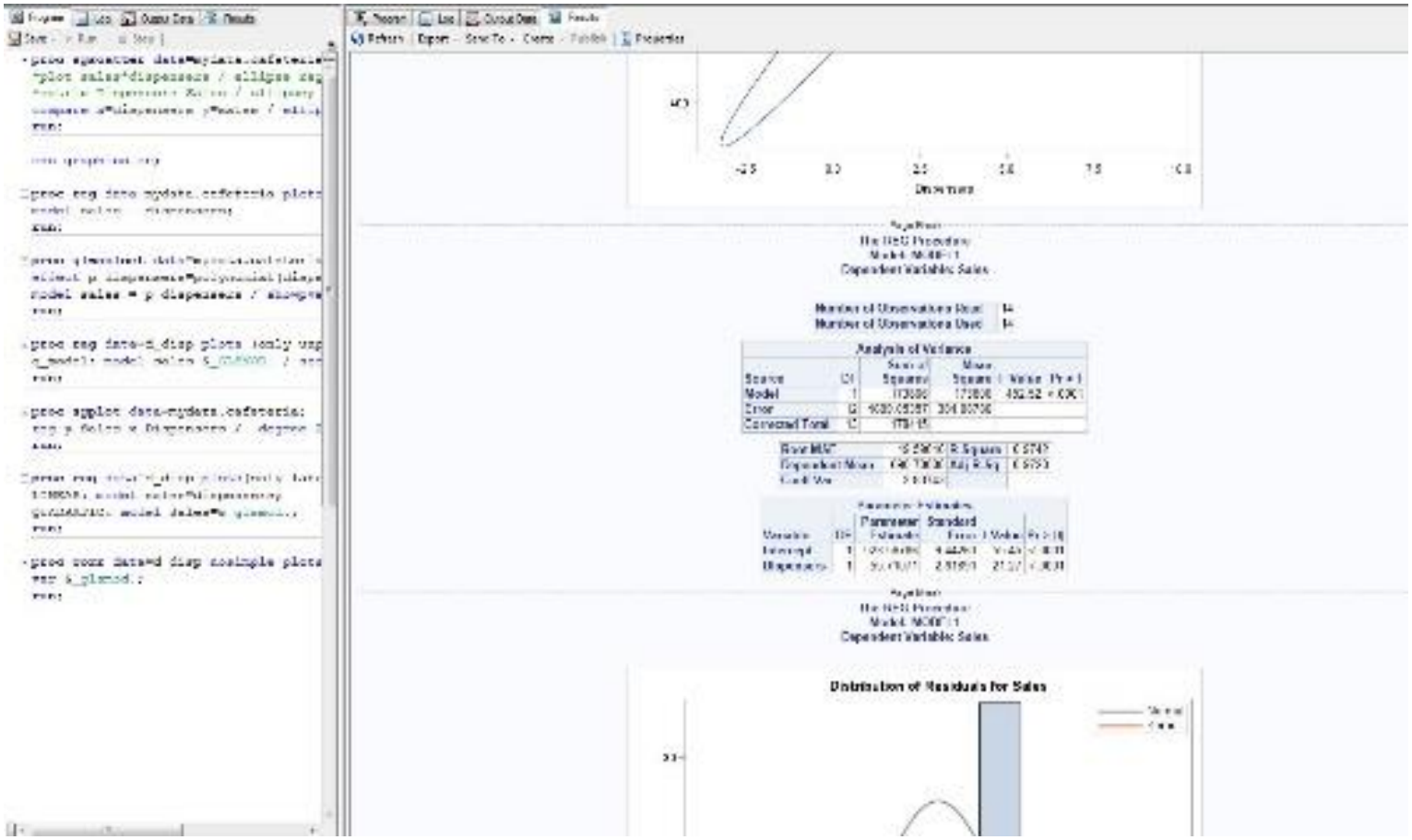
PIMAFOLKET | AKIMEL O'OTHAM – ”FLODFOLKET”

- Detta har vi data på!
- Tack till UCI Machine Learning Repository, ”Pima Indians Diabetes Data Set”, donerat av National Institute of Diabetes and Digestive and Kidney Diseases och Applied Physics-laboratoriet vid Johns Hopkins University
- 8 x-variabler, 1 utfallsvariabel
- 768 observationer
- Visst missing data…
- Kvinnor över 21
- (Från 1990…)

PIMAFOLKET | VARIABLER I DATA

- Number of times pregnant
- Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- Diastolic blood pressure (mm Hg)
- Triceps skin fold thickness (mm)
- 2-Hour serum insulin (mu U/ml)
- Body mass index (weight in kg/(height in m)^2)
  - Nollvärden?!
- Diabetes pedigree function
- Age (years)
- Class variable (0 or 1)

MACHINE LEARNING OCH PREDIKTIV MODELLERING



MACHINE LEARNING OCH PREDIKTIV MODELLERING

- Min erfarenhet
- Bra på frågan ”Varför?”
- ”Black box” och skrämselhicka
- Kombinationen!

Svenska	Machine Learning
Y-värde	Label
Variabel	Feature eller attribute
Transformering	Feature creation
Intercept	Bias

MACHINE LEARNING - AN INTRODUCTION

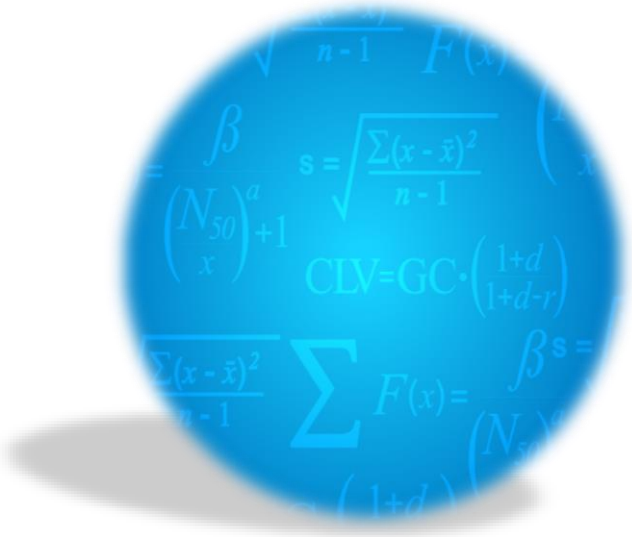
WHAT IS MACHINE LEARNING?

Wikipedia: Machine learning, a branch of artificial intelligence, concerns the construction and study of systems that can learn from data.

SAS: Machine learning is a method of data analysis that automates analytical model building. Using algorithms that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed where to look.

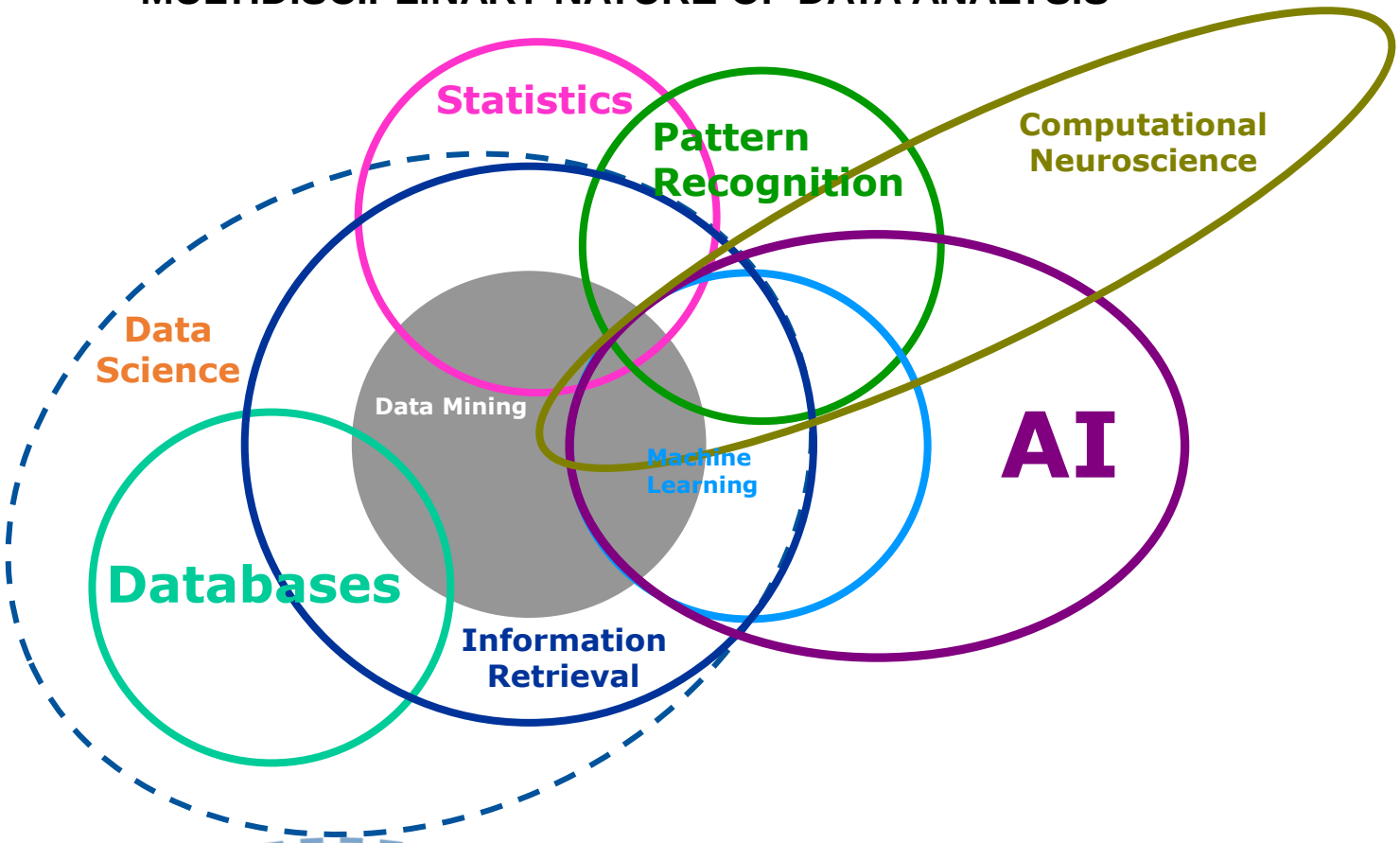
MACHINE LEARNING - AN INTRODUCTION

” Complicated methods, consumable results”



MACHINE LEARNING - AN INTRODUCTION

MULTIDISCIPLINARY NATURE OF DATA ANALYSIS



WHEN TO USE MACHINE LEARNING*?*

When the predictive accuracy of a model is more important than the interpretability of a model.

When traditional approaches are inappropriate, e.g. when you have:

- more variables than observations
- many correlated variables
- unstructured data
- fundamentally nonlinear or unusual phenomena

WHERE IS MACHINE LEARNING USED*?*

A few examples:

- Recommendation applications
- Fraud detection
- Predictive maintenance
- Text analytics
- Pattern recognition
- Self driving cars

- DEMO!

PIMAFOLKET

PARAMETERESTIMAT OCH ODDSKVOTER FÖR STEPWISE LOGISTISK REGRESSION

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	1	-8.9782	1.0078	79.36	<.0001		0.000
IMP_bmi	1	0.0902	0.0212	18.01	<.0001	0.3324	1.094
IMP_glucose	1	0.0321	0.00463	47.91	<.0001	0.5351	1.033
LOG_pedigree	1	1.8236	0.6526	7.81	0.0052	0.1990	6.194
LOG_times_pregnant	1	0.4947	0.1693	8.54	0.0035	0.2140	1.640

Odds Ratio Estimates

Effect	Point Estimate
IMP_bmi	1.094
IMP_glucose	1.033
LOG_pedigree	6.194
LOG_times_pregnant	1.640

NOTE: No (additional) effects met the 0.05 significance level for entry into the model.