Regression based classification of leukemia patients

Yaron Orenstein

Scientific writing - research presentation

21st December, 2011

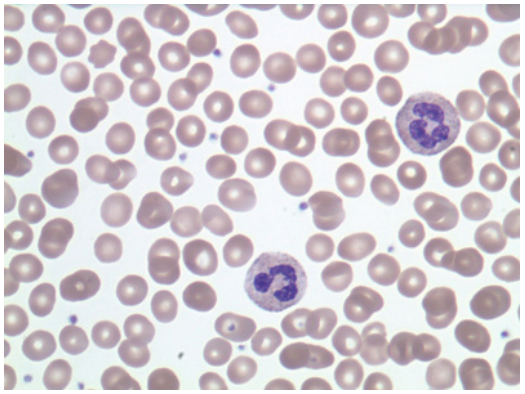
Talk overview

1. Introduction
2. The challenge
3. The solution
4. Results
5. Summary

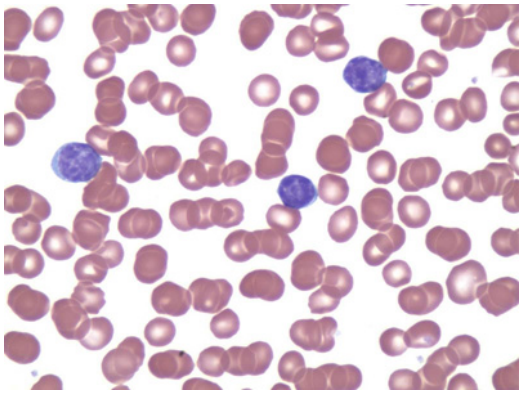
PART 1:
INTRODUCTION

Acute myeloid leukemia

- Acute myeloid leukemia (AML) is a malignancy that arises in white blood cells.
- These cells normally battle infectious agents throughout the body.



Normal blood cells

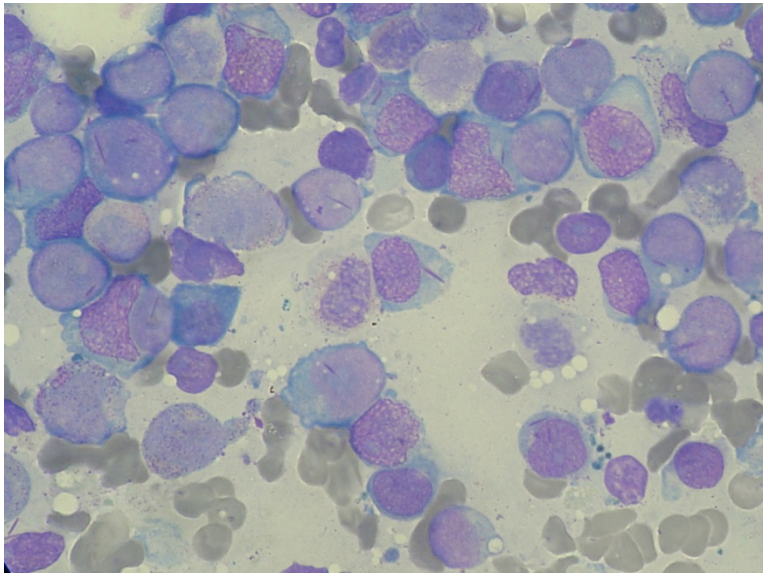


Leukemia blood cells

Acute myeloid leukemia – cont'd

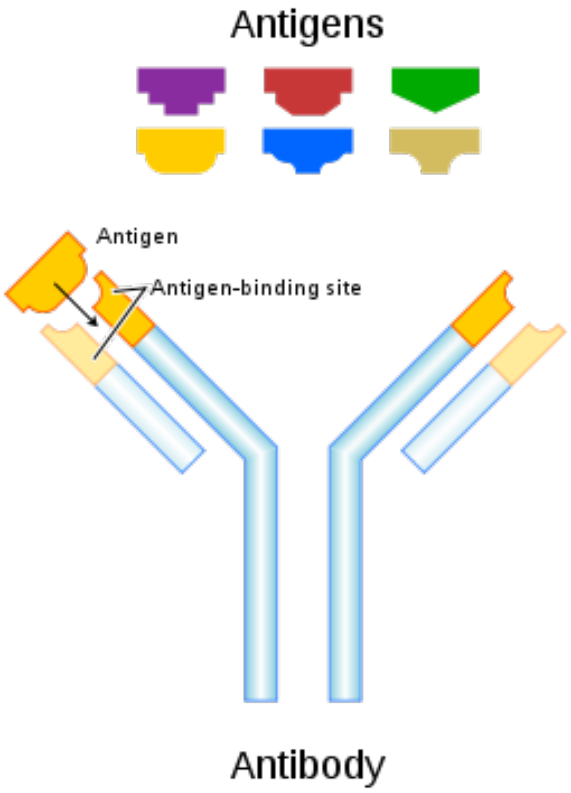
- AML is the most common type of leukemia.
- First diagnoses by an abnormal blood count.
- Full diagnoses by marrow or blood sample examined in cell resolution.

Bone marrow: myeloblasts with Auer rods seen in AML



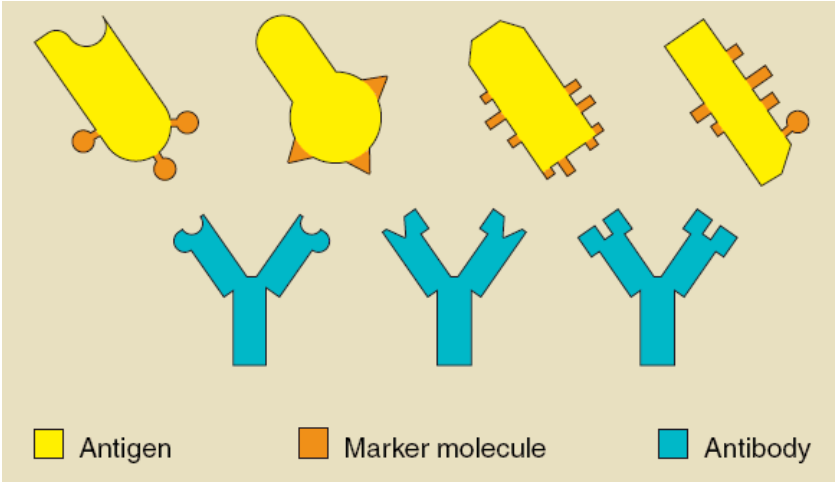
Antibodies

- Antibodies are proteins used by the immune system to neutralize foreign invaders.
- They recognize, through specific binding, molecules called antigens.



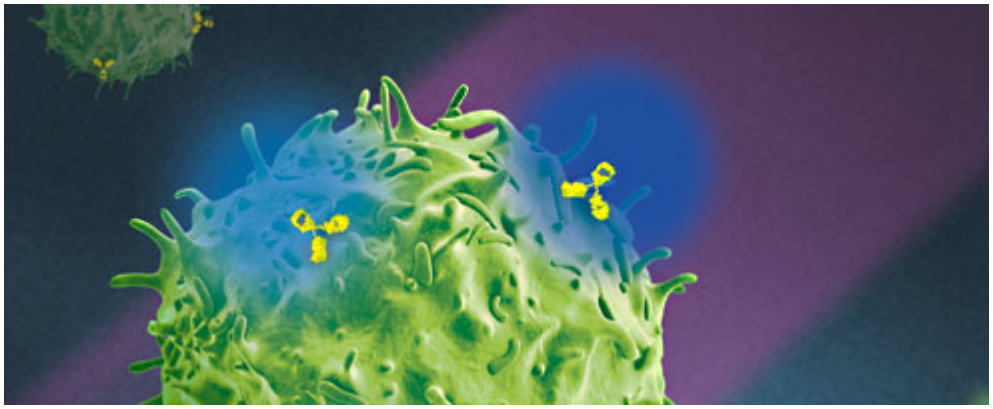
Antigens

- An antigen is a molecule that triggers the production of an antibody.
- The cell express antigens.
- Antigens expression can help in distinguishing cell populations.



Flourochromes

- Antibodies can be ‘colored’ using fluorochromes.
- These molecules can be bound to antibodies and emit fluorescence light.
- This is a reliable way to mark antibodies.



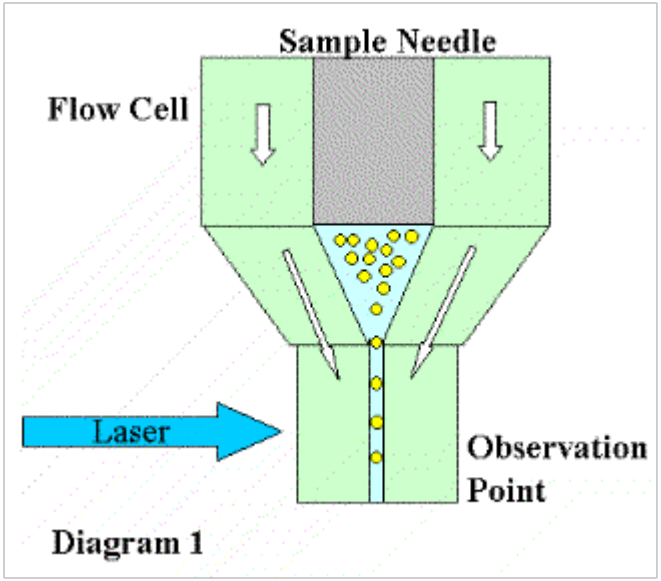
Flow cytometry

- Flow cytometry is a technique used to measure the properties of cells.
- Cells are measured individually, but in large numbers.
- These measurements can help in the diagnosis of AML.

http://www.usuhs.mil/bic/
http://www.abdserotec.com/

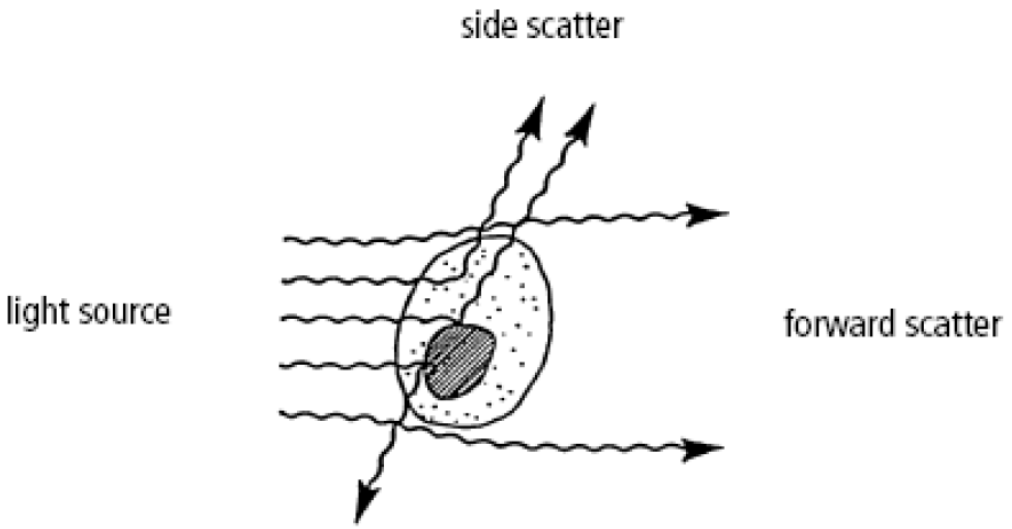
Flow cytometry overview

- The cell sample is injected into a stream.
- The cells in the sample are accelerated and individually pass through a laser beam for interrogation.



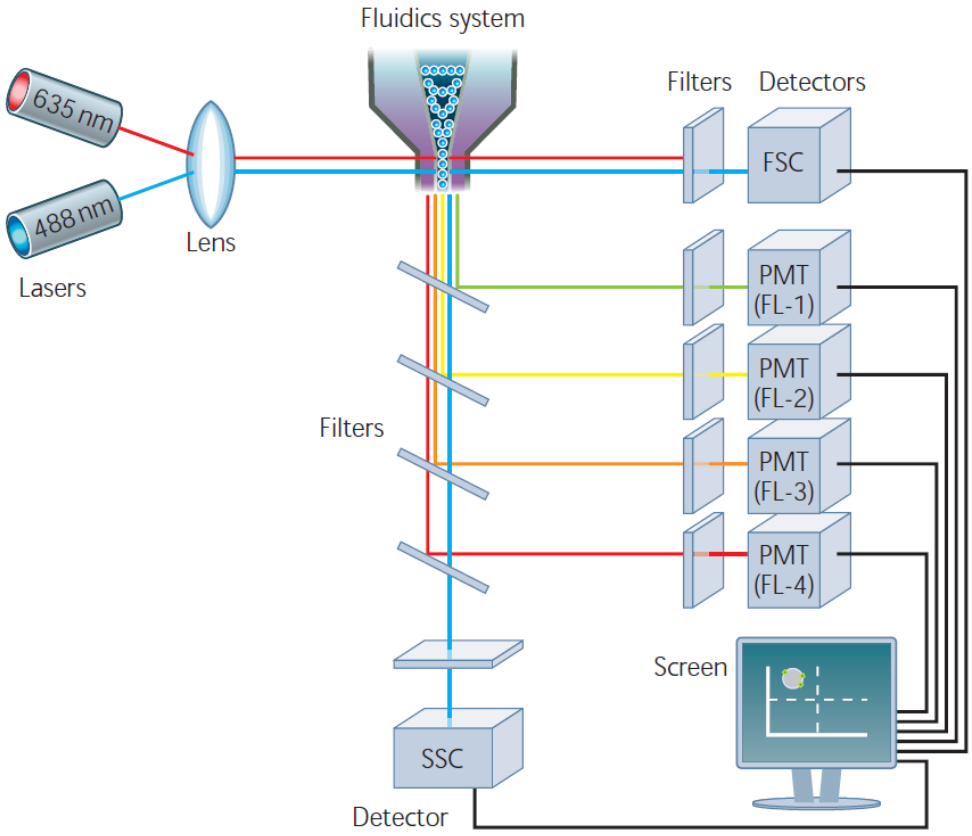
Forward and side scatters

- When a cell passes through the laser beam, it deflects incident light.
- Forward-scattered light (FSC) is proportional to the surface area or size of a cell.
- Side-scattered light (SSC) is proportional to the granularity or internal complexity of a cell.



Flourchromes measurement

- ‘Colored’ antibodies are bound to antigens on the surface of the cells.
- The flourscence light they emit can be detected by a system of optical filters and mirrors.



Regression

- Given a set of points (x_i, y_i), calculate function f, that best predicts y_i given x_i.

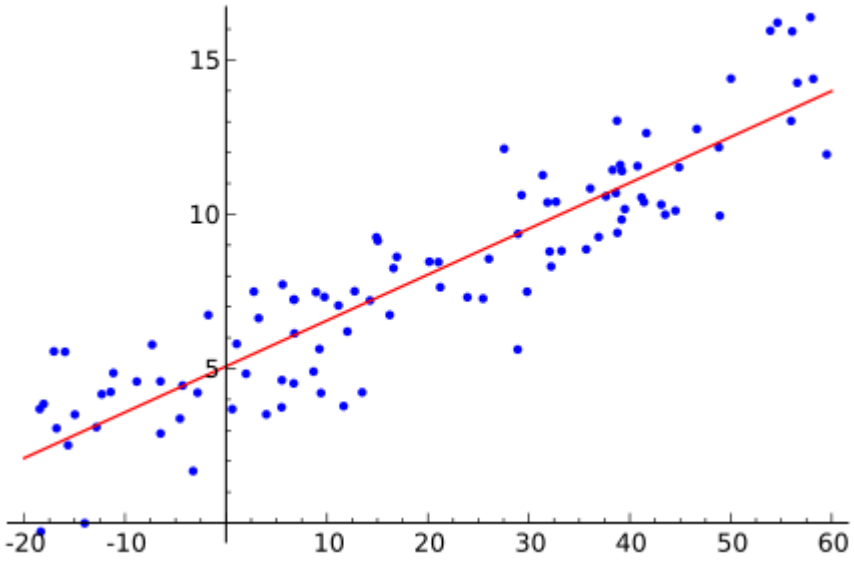
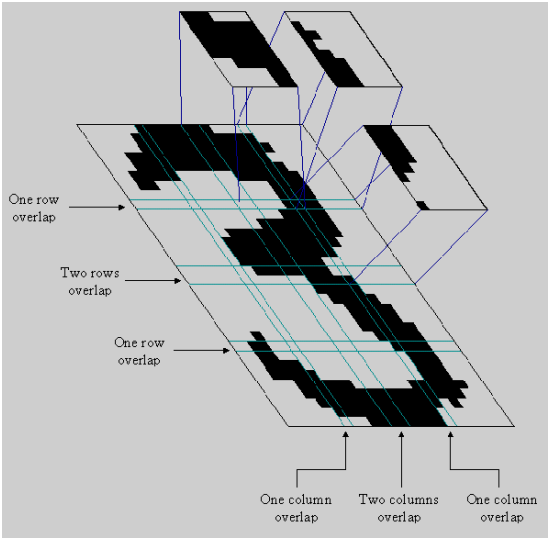


Illustration of linear regression on a data set

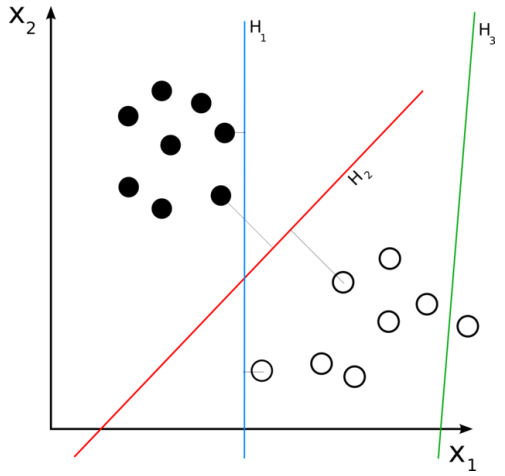
Classification

- Given a set paired values and classes (x_i, z_i), predict *h*, that best predicts z_i given x_i.
- Some solvers use regression based methods.
- Example: optical character recognition.

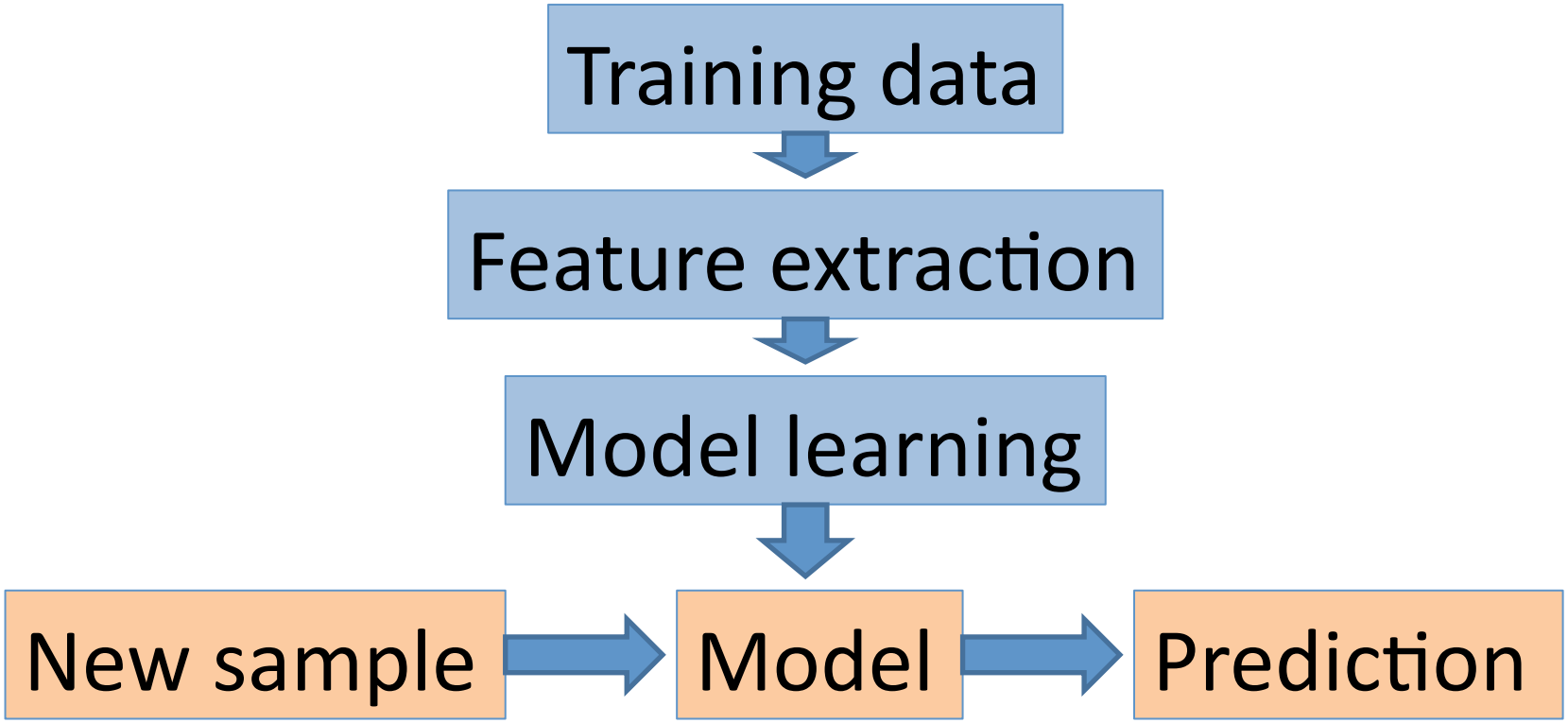


Support vector machines

- A classifier is learned to maximize the gap between the categories.
- The gap is the minimal distance between points of different categories to the divider.
- The same variation exists for regression.



General pipeline

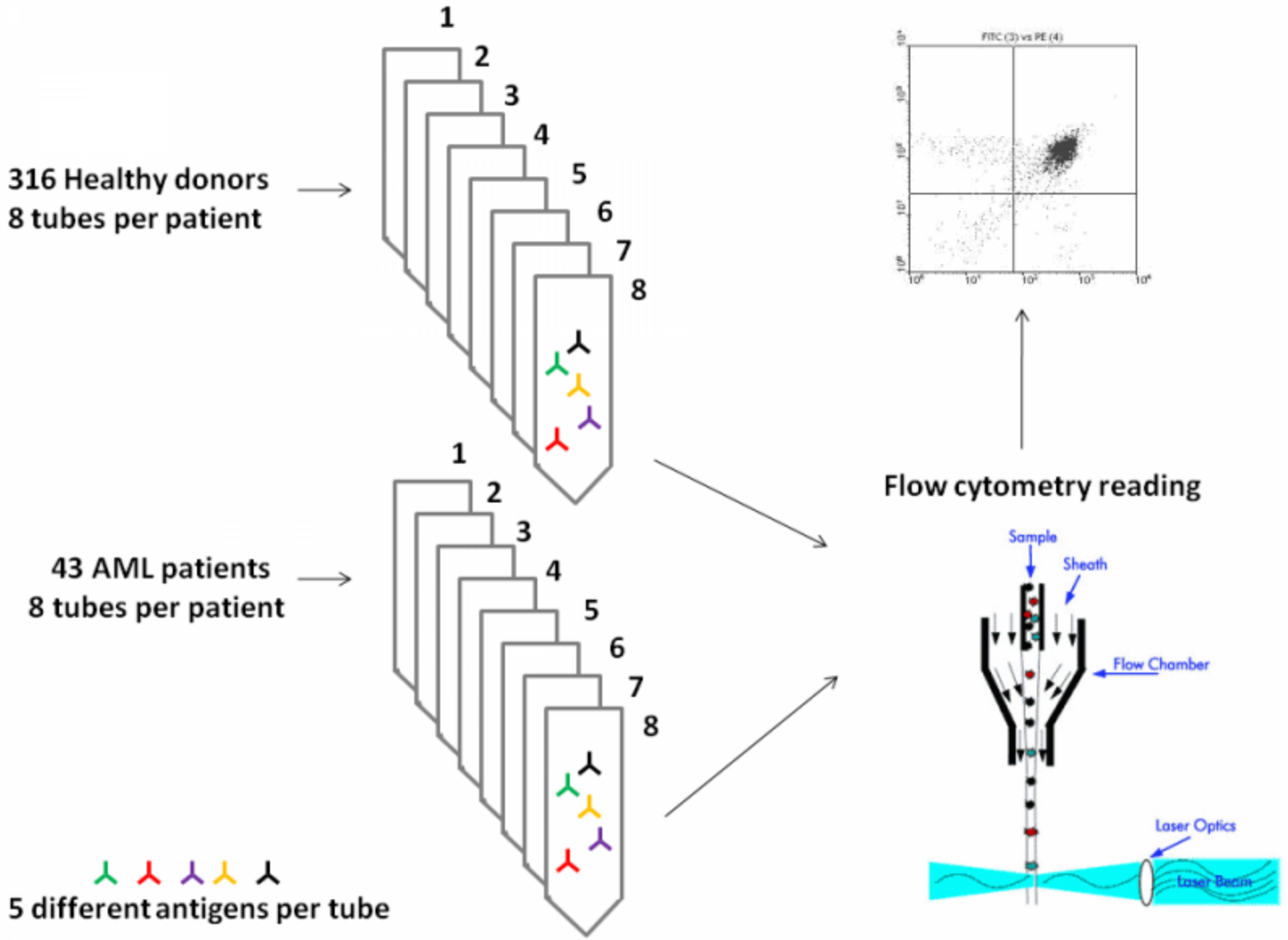


PART 2:
THE CHALLENGE

The challenge

- The samples consist of 43 AML positive patients and 316 healthy donors.
- Samples from peripheral blood or bone marrow aspirate.
- The samples were subsequently studied with flow cytometry.

The challenge



The challenge – cont'd

- The training set: healthy / AML classification for half of the patients.
- The test set: the other half.
- The challenge: predict healthy / AML for each sample in the test set.

The challenge – cont'd

	FL1	FL2	FL3	FL4	FL5
Tube 1	IgG1-FITC	IgG1-PE	CD45-ECD	IgG1-PC5	IgG1-PC7
Tube 2	Kappa-FITC	Lambda-PE	CD45-ECD	CD19-PC5	CD20-PC7
Tube 3	CD7-FITC	CD4-PE	CD45-ECD	CD8-PC5	CD2-PC7
Tube 4	CD15-FITC	CD13-PE	CD45-ECD	CD16-PC5	CD56-PC7
Tube 5	CD14-FITC	CD11c-PE	CD45-ECD	CD64-PC5	CD33-PC7
Tube 6	HLA-DR-FITC	CD117-PE	CD45-ECD	CD34-PC5	CD38-PC7
Tube 7	CD5-FITC	CD19-PE	CD45-ECD	CD3-PC5	CD10-PC7
Tube 8	Non Specific	Non Specific	Non Specific	Non Specific	Non Specific

The challenge – cont'd

- Expression of the CD45 molecule correlates with the stage of differentiation of the cells studied.
- Its expression is weak in the case of acute myeloid leukaemia.
- The 8th tube is an isotype control tube, with non-specific-binding antibodies (i.e., mouse antibodies).

The data

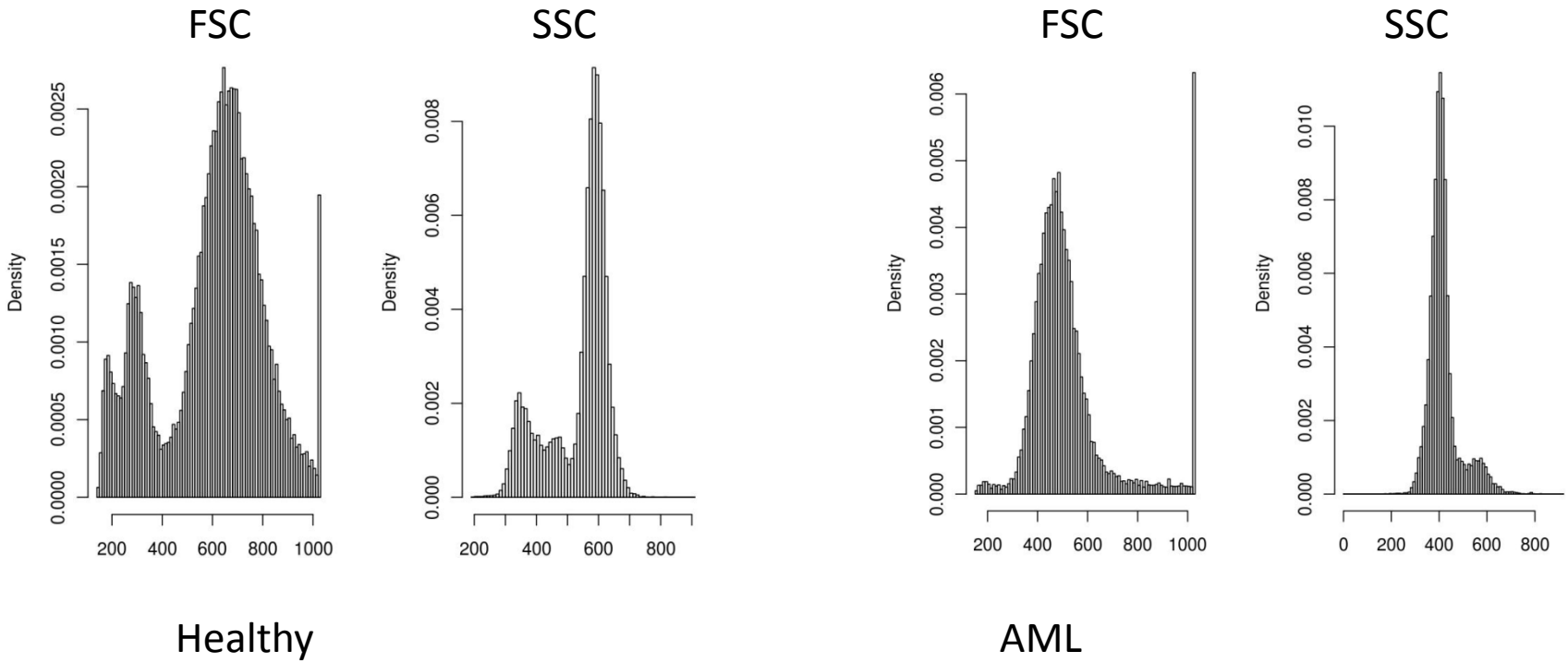
- Raw data in Flow Cytometry Standard (FCS).
- Preprocessed data (transformed/compensated) in CSV format.
- Each subsequent row is an event (a cell) detected by the flow cytometer (~ 30,000)

```
"FS Lin","SS Log","FL1 Log","FL2 Log","FL3 Log","FL4 Log","FL5 Log"
273,    0.545,  0.219,  0.210,  0.181,  0.163,  0.144
......
793,    0.649,  0.457,  0.377,  0.344,  0.1889, 0.149
```

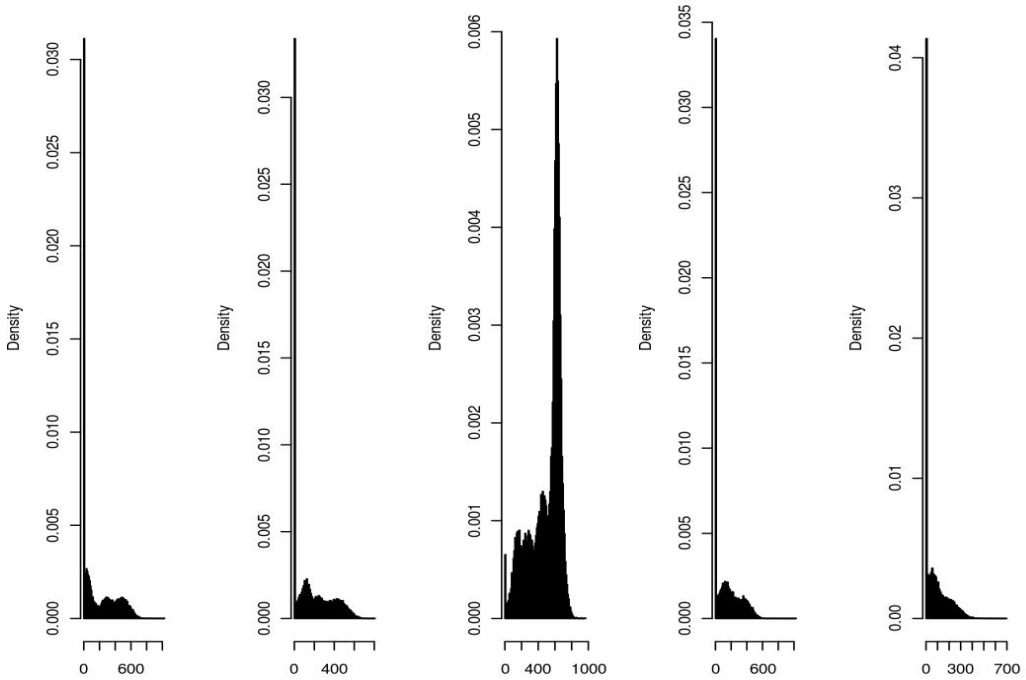
PART 3:
THE SOLUTION

Forward and side scatter distribution differences

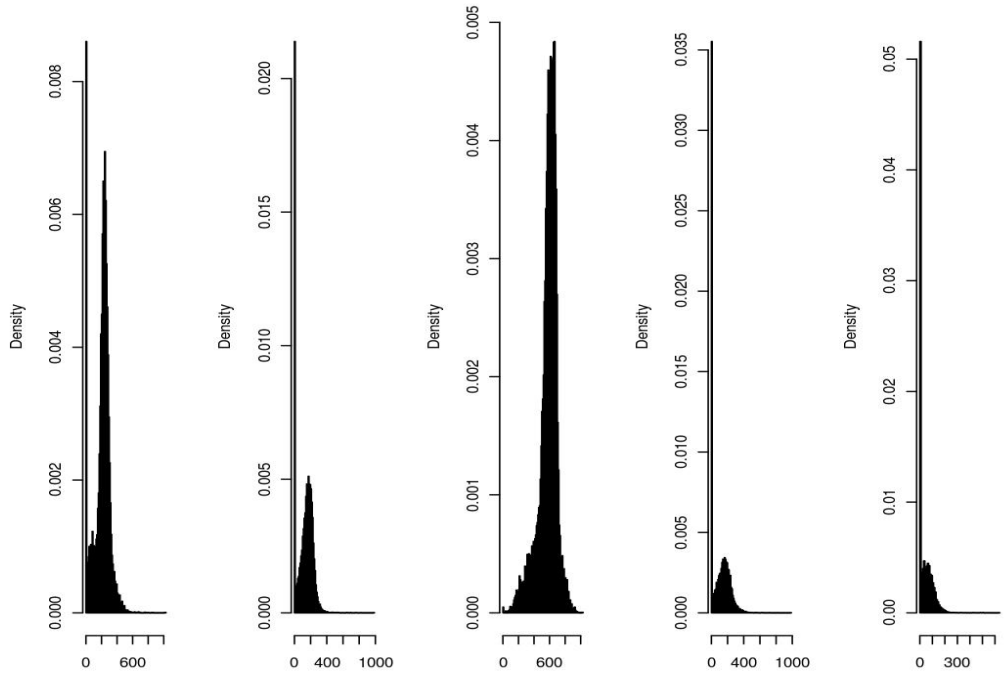
- Typical histograms of FSC and SSC of healthy and AML patients:



5 markers distribution differences



Healthy



AML

FL1	FL2	FL3	FL4	FL5
Kappa-FITC	Lambda-PE	CD45-ECD	CD19-PC5	CD20-PC7

Feature extraction

- In order to utilize the difference in distribution we extract 4 features for each marker:
  1. Mean.
  2. Variance.
  3. Skewness (3^{rd} moment).
  4. Kurtosis (4^{th} moment).

- 7 markers for 7 tubes (omitted 8^{th} tube).
- Total of 7×7×4=196 features for each sample.

Model learning

- We used a meta-classifier that utilizes regression for classification tasks (Frank *et al*. 98).

- One regression model is built for each class, and the probability (confidence level) to belong to each class is estimated using the models.

Model learning – cont'd

- The class is chosen as the one attaining the higher probability.

- The regression method used was SVM-regression (Drucker *et al.* 96).

- Implementation in Weka (Hall *et al.* 09).

PART 4:
RESULTS

Results

- The DREAM6/Flowcap Challenge test set consisted of 180 samples (20 AML).
- We achieved a perfect prediction (together with 7 other teams).
- In a leave-one-out cross validation test our method achieved an area under the ROC curve of 0.992 and area under the precision recall curve of 0.984.

The top selected features

Marker designation	Marker	Moment	Weight
stem cell marker, adhesion, found on hematopoietic precursors, capillary endothelium, and embryonic fibroblasts	CD34	2	0.5032
found on thymocytes, some T cells, monocytes, natural killer cells, and hemopoietic stem cells	CD7	1	0.3475
a marker of unknown function found on immature myeloid cells	CD33	4	0.3374
found naturally on myelomonocytic cells	CD13	4	0.3301
found naturally on myelomonocytic cells	CD13	3	0.3159
found on B cells that forms a calcium channel in the cell membrane	CD20	1	0.3001
c-kit, the receptor for Stem Cell Factor, a glycoprotein that regulates cellular differentiation, particularly inhematopoiesis	CD117	3	0.2675
B-lymphocyte surface antigen B4	CD19	1	0.2586
	CD64	2	0.2176
found on thymocytes, T cells, and some natural killer cells that acts as a ligand for CD58 and CD59 and is involved in signal transduction and cell adhesion	CD2	3	0.2174
	CD117	2	0.2149
a membrane protein found on macrophages which binds to bacterial lipopolysaccharide.	CD14	4	0.2114
leucocyte common antigen, a type I transmembrane protein present on all hemopoietic cells except erythrocytes that assists in cell activation	CD45	1	0.2018

PART 5:
SUMMARY

Summary

- We developed a classifier for diagnosing AML patients.
- It is based on classical machine learning methods.
- Features are based on distribution differences.
- Our method achieved perfect prediction of the test set.
- It is applicable to other flow cytometry data.

Credit

- This was part of DREAM6/FlowCap2 challenge

http://www.the-dream-project.org/challanges/dream6flowcap2-molecular-classification-acute-myeloid-leukaemia-challenge

- Word done together with: David Amar and Ron Zeira, supervised by Prof. Ron Shamir.