A *curious method* for data from *curious experiments*:

Random projection with TOF-SIMS data from comet-relevant samples

Kurt Varmuza

Vienna University of Technology, Institute of Chemical Engineering,
Laboratory for **C**hemo**M**etrics

kvarmuza@email.tuwien.ac.at

www.lcm.tuwien.ac.at

Chemometrics Workshop 2010, TU Vienna, 28.6.2010

Special Issue Article

Journal of CHEMOMETRICS

Received: 15 December 2009, Revised: 13 January 2010, Accepted: 14 January 2010, Published online in Wiley InterScience: 2010

(www.interscience.wiley.com) DOI: 10.1002/cem.1295

Random projection experiments with chemometric data

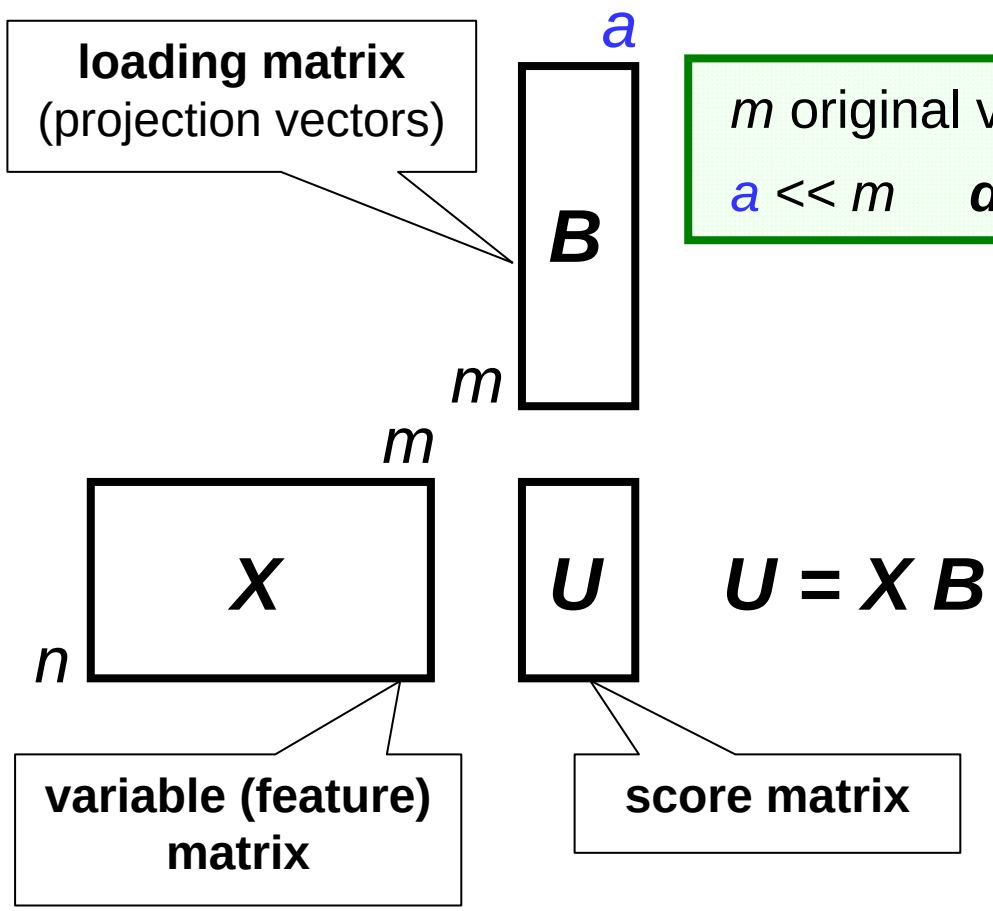
Kurt Varmuza[a]*, Peter Filzmoser[b] and Bettina Liebmann[a]

Random projection (RP) is a linear method for the projection of high-dimensional data onto a lower dimensional space. RP uses projection vectors (loading vectors) that consist of random numbers taken from a symmetric distribution with zero mean; many successful applications have been reported for high-dimensional data sets. The basic ideas of RP are presented, and tested with artificial data, data from chemoinformatics and from chemometrics. RP's potential in dimensionality reduction is investigated by a subsequent cluster analysis, classification or calibration, and is compared to PCA as a reference method. RP allowed drastic reduction in data size and computing time, while preserving the performance quality. Successful applications are shown in structure similarity searches (53 478 chemical structures characterized by 1233 binary substructure descriptors) and in classification of mutagenicity (6506 chemical structures characterized by 1455 molecular descriptors). Only in calibration tasks with low-dimensional data as in many chemical applications, RP showed limited performance. For special applications in chemometrics with very large data sets and/or severe restrictions for hardware and software resources, RP is a promising method.

Keywords: dimensionality reduction; PCA; similarity of chemical structures; KNN classification; PLS regression

Journal of Chemometrics **24** (2010) 209-217

Dimensionality Reduction by Projection



m original variables → *a* new variables

a << *m* ***defined objective for the scores***

- **PCA**
  maximum variance
- **LDA**
  maximum separation
- **OLS**
  maximum correlation
- **PLS**
  maximum covariance
- . . .

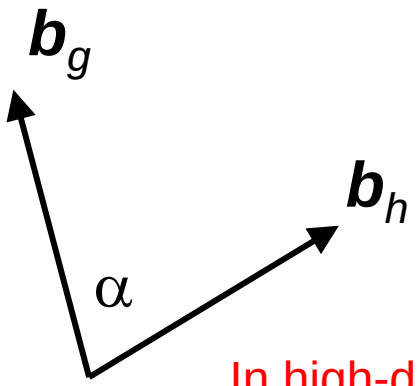
random projection vectors in *B* ☹

simple, fast,
independent from ***X*** and ***y***

Curious method: Random Projection (RP)

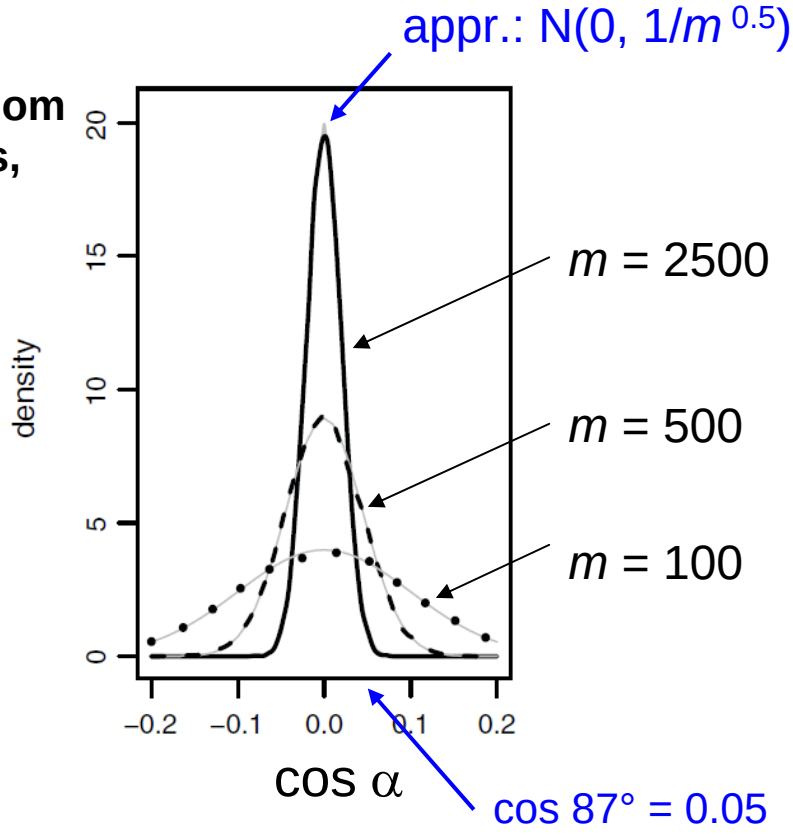
Random projection vector $\boldsymbol{b} = [b_1, b_2, \ldots, b_j, \ldots, b_m]$

from a distribution symmetrical to zero:
e.g. from N(0,1), U[-1, +1], {-1, +1}, ...



In high-dimensional space
two such random vectors
have a **high probability**
for being
almost orthogonal.

10,000 random vector pairs, U[-1, +1]



Random Projection (RP): Example 1

X n = 53,478 chemical structures (objects)

m = 1233 binary substructure descriptors (binary variables)

Similarity of objects (chemical structures) calculated from

- Tanimoto index from 1233 binary variables
- Euclidean distance from a = 30 RP scores

Summary: Both yielded very similar hitlists (nearest neighbors)

Varmuza K., Filzmoser P., Liebmann B.: J. Chemom. **24** (2010) 209-217

Random Projection (RP): Example 2 (QSAR)

X n = 6506 chemical structures (objects)

n_1 = 3502 mutagenic, n_2 = 3004 not mutagenic (AMES test)

m = 1455 molecular Dragon descriptors (variables)

KNN classification (Euclidean distance) of mutagenicity from

- m = all 1455 variables
- a = 100 PCA scores (from all data)
- a = 100 PCA scores (from 4% of the objects)
- a = 100 RP scores

Summary: Similar classification performance (ca 75% correct)

Varmuza K., Filzmoser P., Liebmann B.: J. Chemom. **24** (2010) 209-217

Curious experiment: TOF-SIMS Data from Comet

2 March 2004 Launch of **ROSETTA** (ESA)

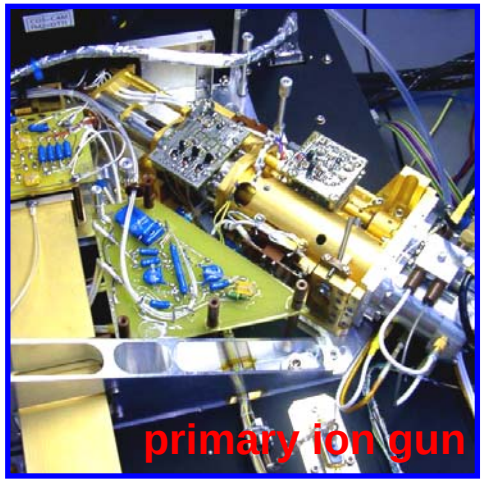
2014 Orbit around the comet
(experimental data expected)

2010 Instruments (still) working

TOF-SIMS instrument
Time-of-Flight Secondary Ion Mass Spectrometer

COSIMA

Cometary Secondary Ion Mass Spectrometer
measurements of cometary dust particles (orbit)



Kissel J., et al. (41 authors): *Space Science Reviews* **128**, 823-867 (2007)
COSIMA – High resolution time-of-flight secondary ion mass spectrometer for the analysis of cometary dust particles onboard ROSETTA.

Curious experiment: TOF-SIMS Data from Space

Restrictions

- **very limited data storage (onboard)**
  only one or a few full spectra
- **very limited data transfer**
  about 1 - 2 times a week,
  signal needs ca 20 minutes per way
- **very unexpected data possible**
  no PCA or calibration in advance

Curious experiment: TOF-SIMS Data from Lab

Cosima Research Module

Max-Planck Institute for Solar System Research
(Katlenburg-Lindau, Germany):

Hilchenbach Martin (P.I.)
Kissel Jochen (former P.I.)
Krüger Harald

Finnish Meteorological Institute (FMI, Helsinki):
Silen Johan

2005

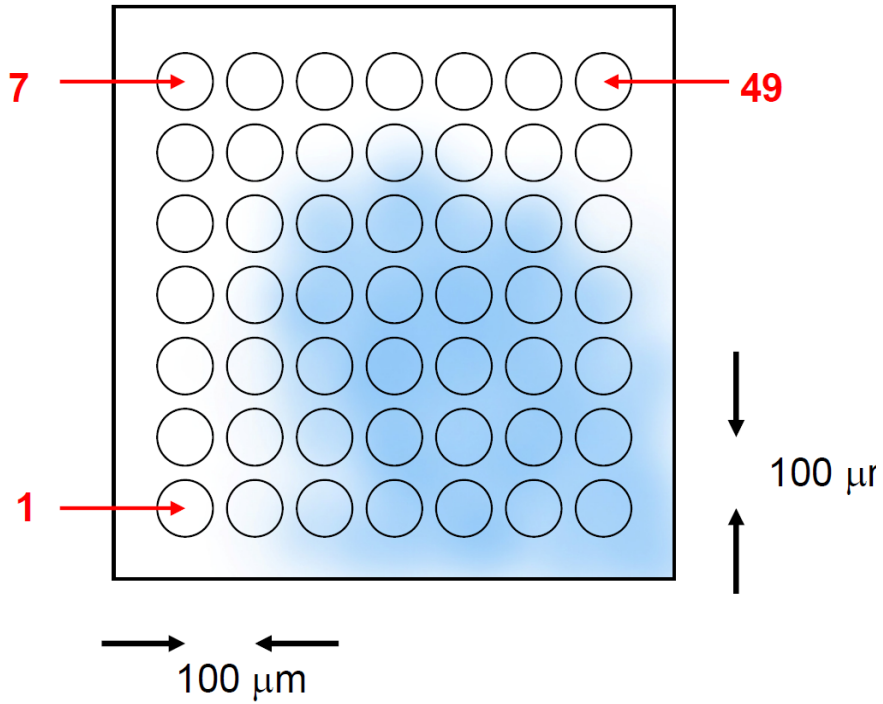
Curious experiment: TOF-SIMS Data from Lab

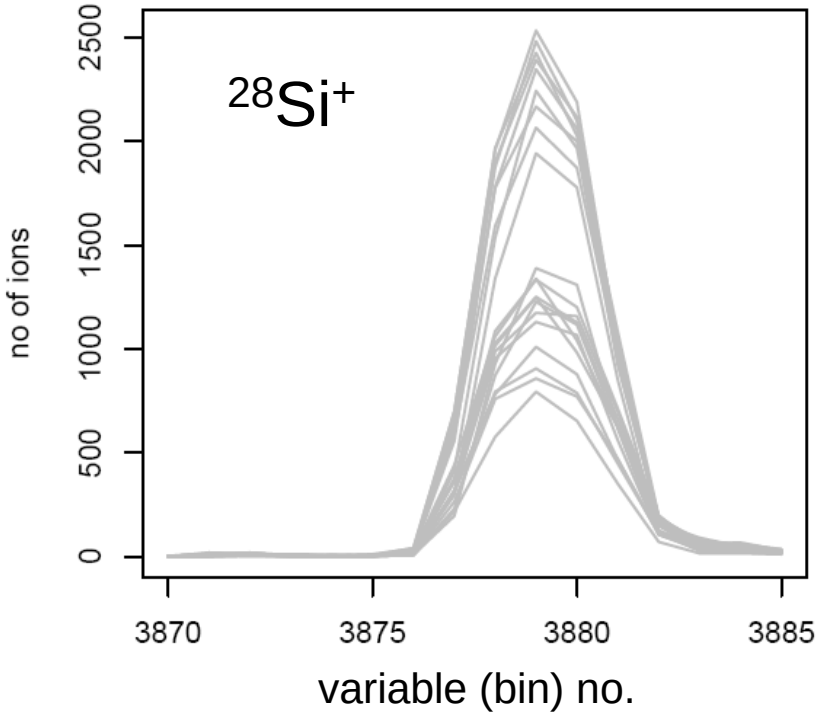
A typical **TOF-SIMS spectrum** (after some data reduction, $m/z < 114$, ^{115}In):

m = 8118 variables = no. of detected ions in 4 ns time intervals ("bins")

Experiment Ag target with a grain of pyroxene (Al, Mg, ... silicate), diameter ca 500 μm

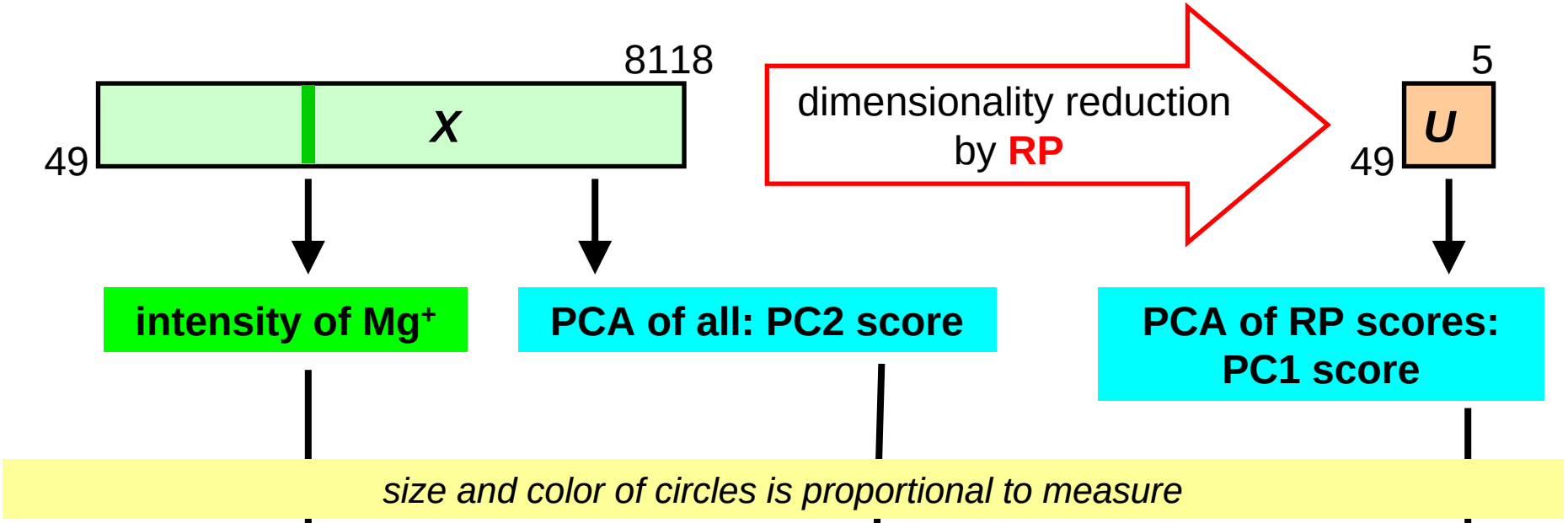
n = 49 spectra measured at the positions of a 7 × 7 grid

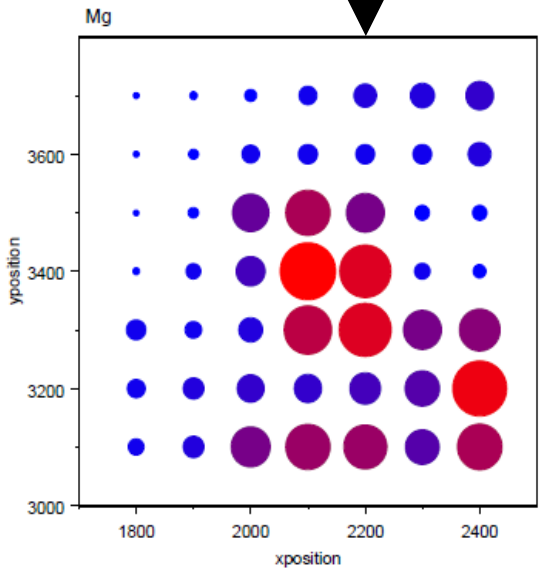


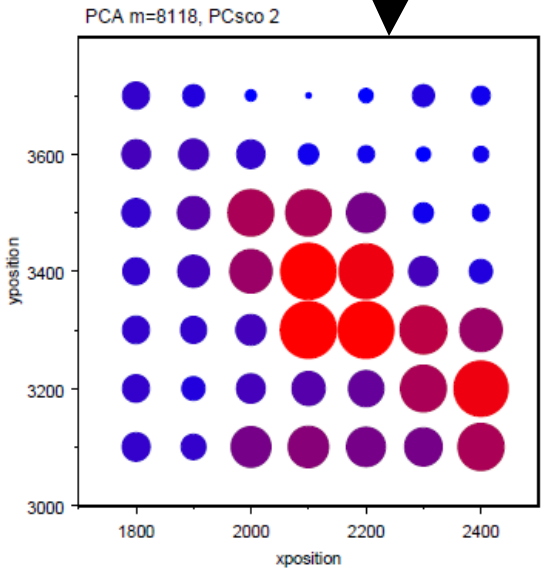


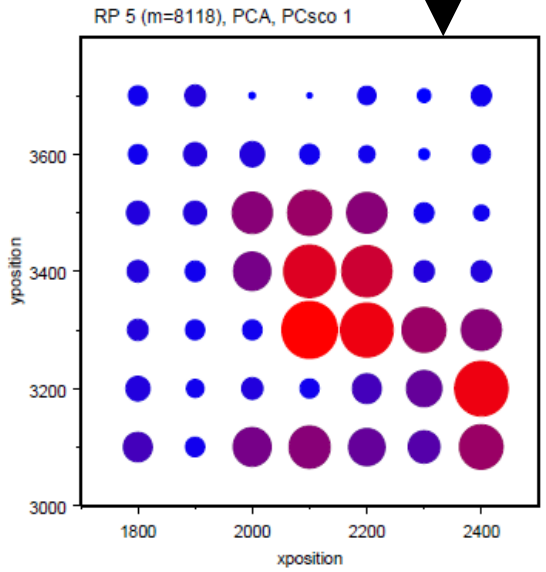
data from lab

TOF-SIMS Data from Lab: Random Projection



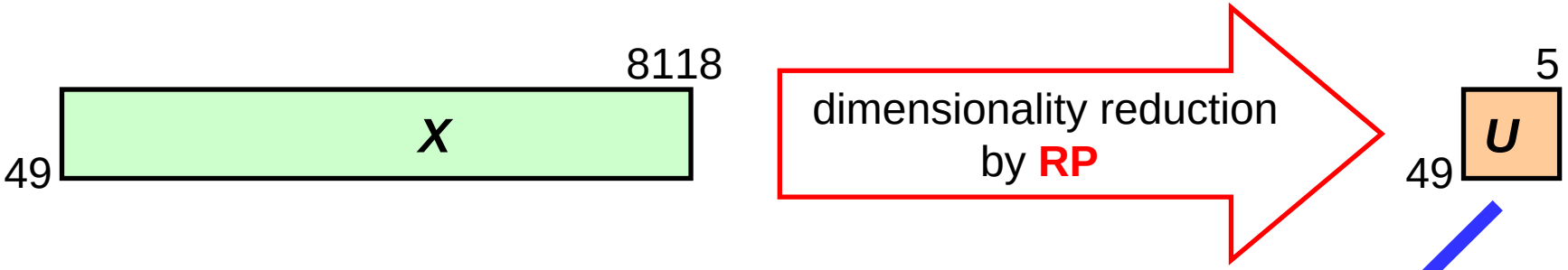






Mg, PCA, RP-PCA-scores

TOF-SIMS Data from Lab: RP / Spectra Similarity



Dissimilarity of spectra ~

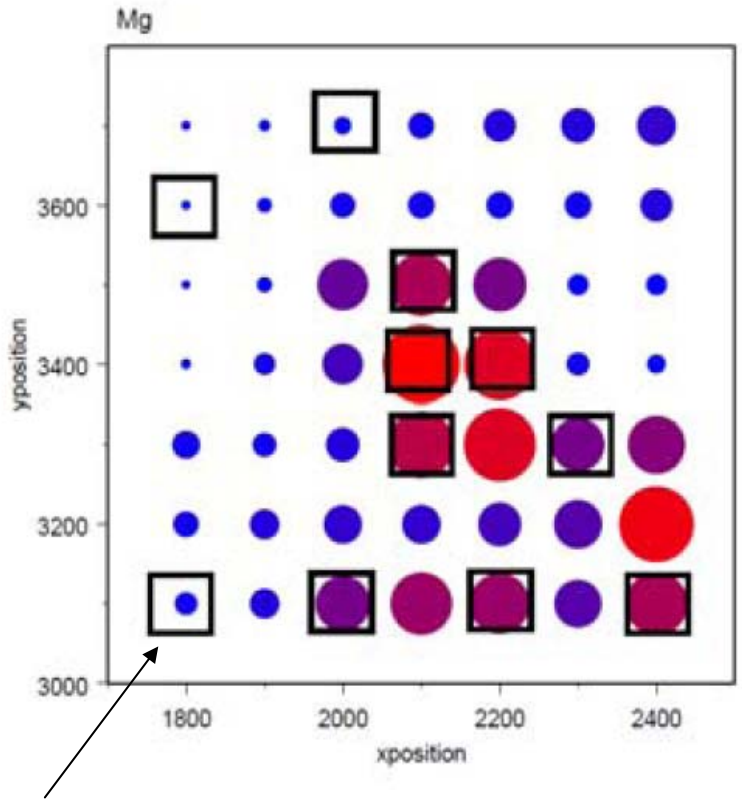
~ Euclidean distance of ***u***-vectors (5-dimensional)

Evaluate spectra sequentially

Subset with **characteristic spectra**

TOF-SIMS Data from Lab: RP / Spectra Selection

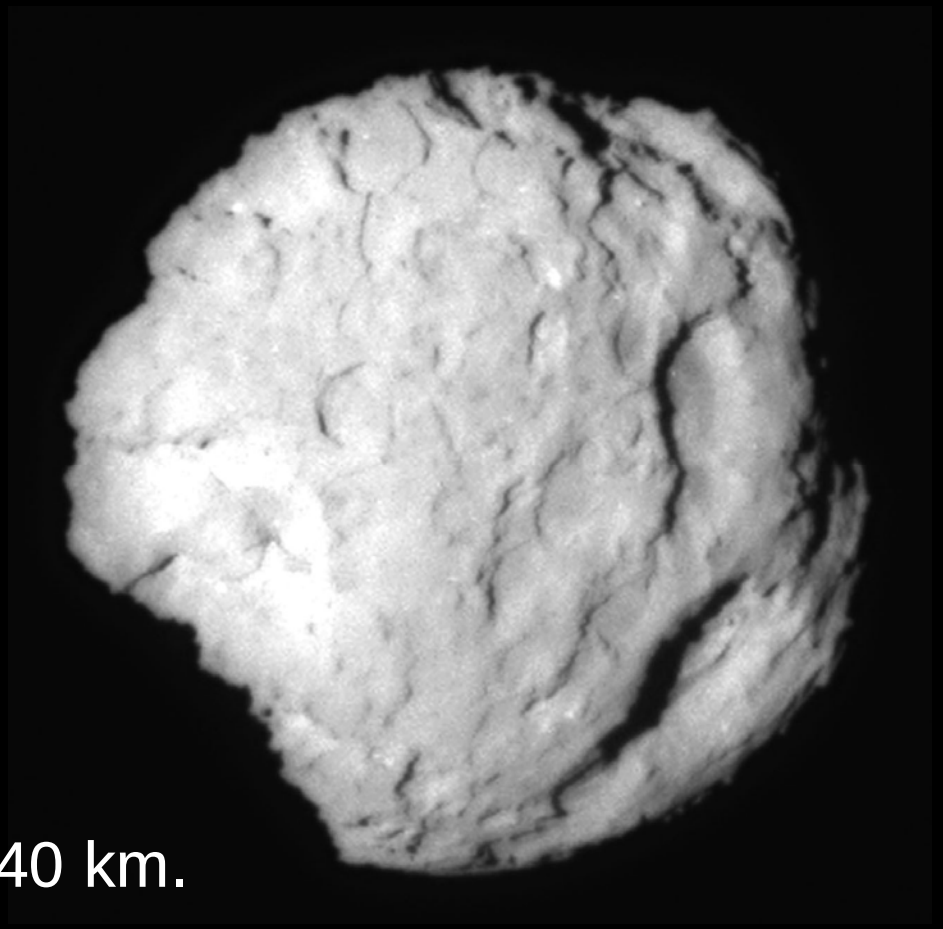
size and color of circles is proportional to the number of Mg^+ ions



11 marked spectra selected from 49

(3 from target, 8 from mineral)

selecti
on



Stardust (NASA)

Comet Wild-2

Ca 4.5 km diameter, from 240 km.

Collection of cometary dust (2002):
aerogel, fly-by at 240 km distance,
6.1 km/s relative speed (= 5-10 times a gun bullet).

Return: January 2006