



Interactive Discovery in Large Data Sets

UCLA Statistics Seminar

Kiri L. Wagstaff
Jet Propulsion Laboratory
kiri.wagstaff@jpl.nasa.gov

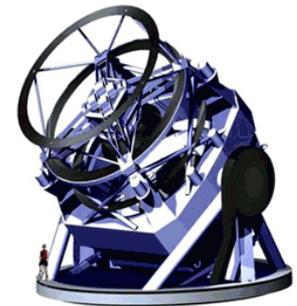
October 16, 2012

Joint work with
David R. Thompson (JPL),
Nina Lanza (Los Alamos National Lab),
Thomas G. Dietterich (Oregon State University), and
Diana Blaney (JPL)

This work was carried out in part at the Jet Propulsion Laboratory, California Institute of Technology, © 2012. Government sponsorship acknowledged. It was also supported by the Defense Advanced Research Projects Agency (DARPA) under Contract W911NF-11-C-0088. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author's and do not necessarily reflect the views of the DARPA, the Army Research Office, or the US government.

Interactive Discovery in Large Data Sets

- Discovery
 - What is interesting? Novel?
 - Big NASA data sets
 - LSST: 28 TB/day
 - SKA: 86 TB/day
- Explanations
 - AI: actions + reasons for them
- Why “interactive”?
 - No general definition of “interesting”



Large Synoptic Survey
Telescope (LSST)



Square Kilometre Array (SKA)

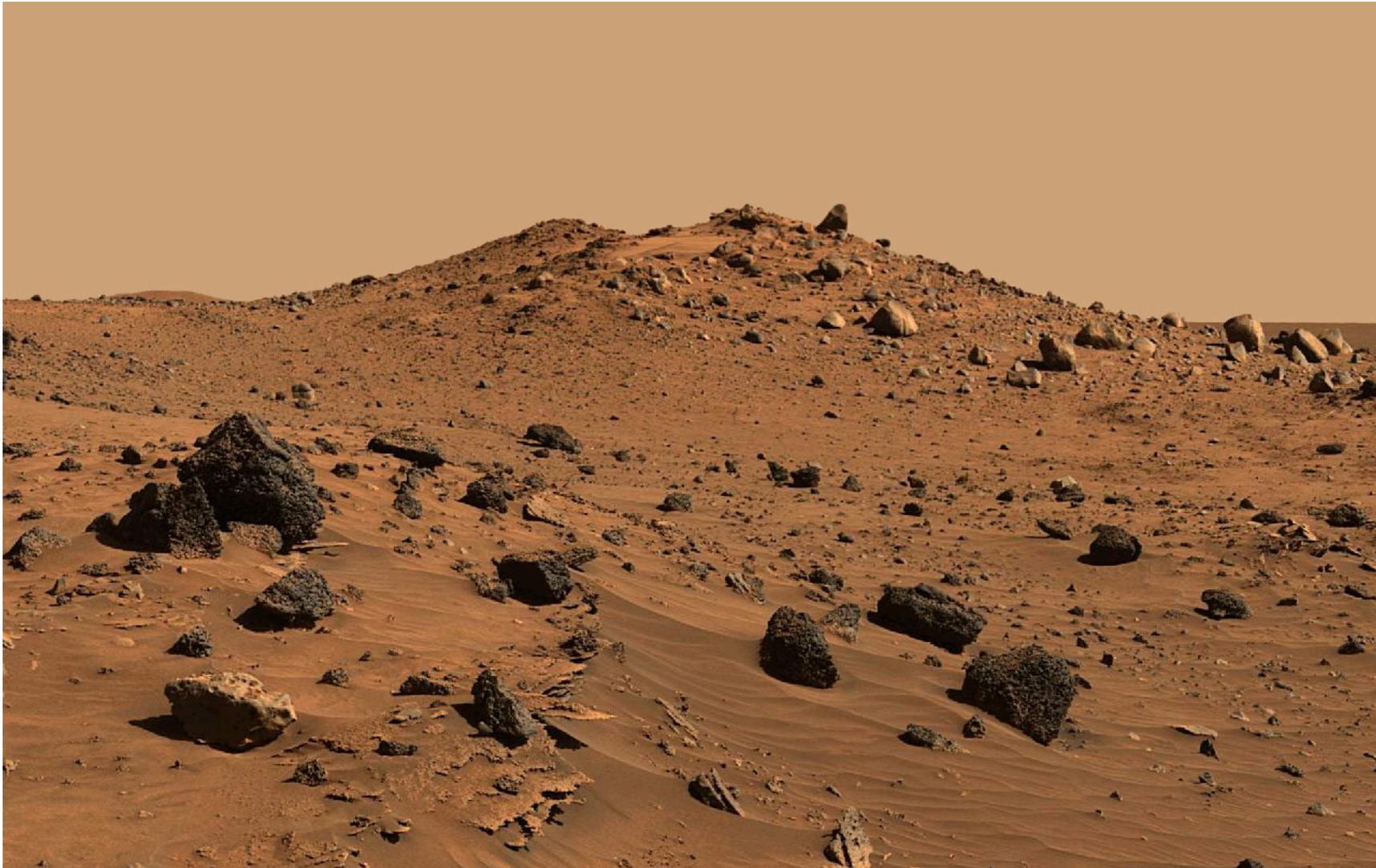
Example: Mars Rover Panorama



Spirit's McMurdo Panorama, 1000 sols, October 2006 (NASA/JPL/Cornell)
22,348 x 5771 pixels = 386 MB

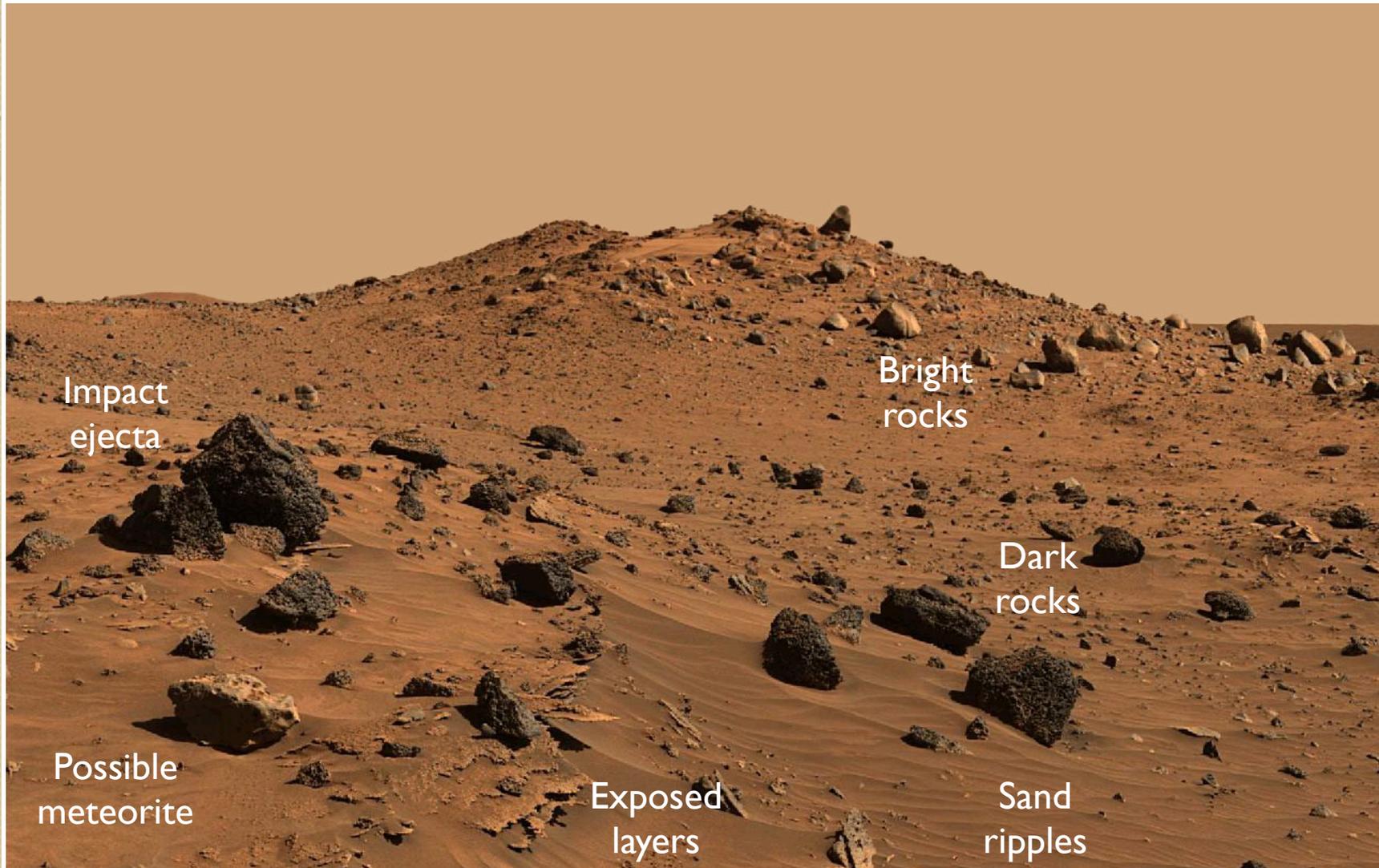
What's most interesting here?

Zooming in



Portion of Spirit's McMurdo Panorama,
1000 sols, October 2006 (NASA/JPL/Cornell)

Zooming in



Impact
ejecta

Bright
rocks

Dark
rocks

Possible
meteorite

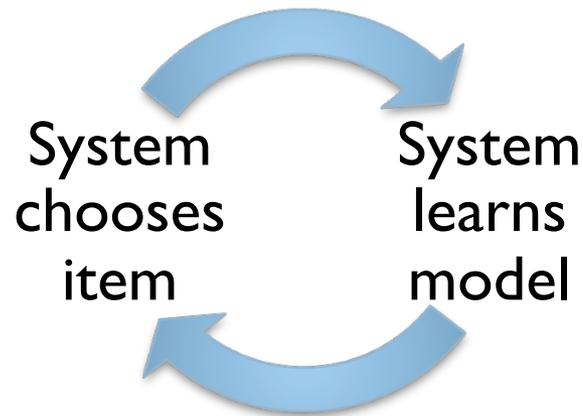
Exposed
layers

Sand
ripples

Portion of Spirit's McMurdo Panorama,
1000 sols, October 2006 (NASA/JPL/Cornell)

Discovery

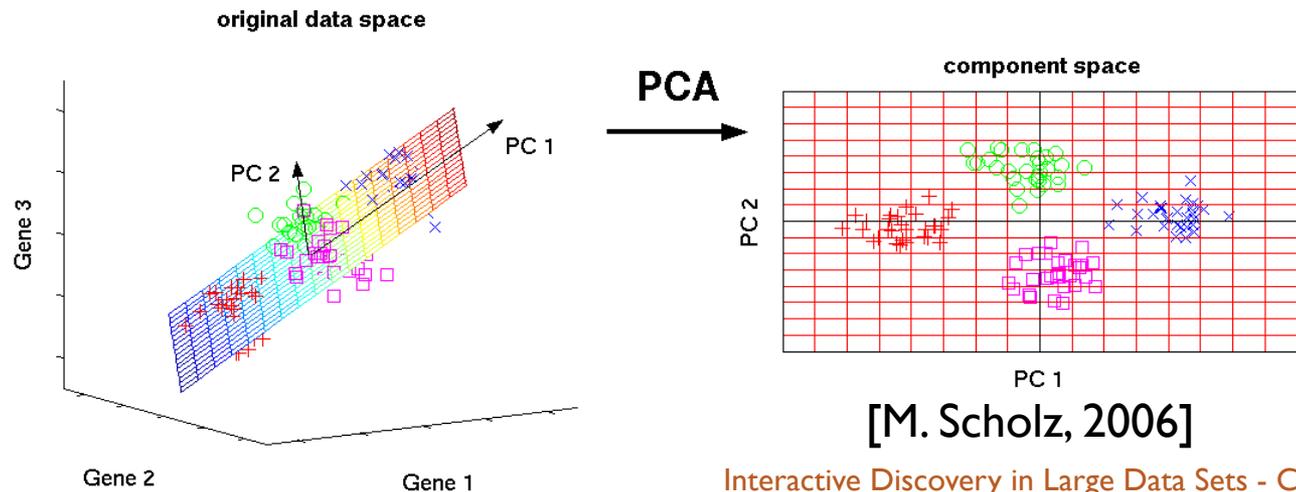
- Exploration of large data sets



- Desiderata
 - Diverse sampling of data set

What to select?

- Items that differ from those previously seen
 - Principal Components Model
 - Approximate model of data set variation
- Known items — $X = U\Sigma V^T$
- Keep only the top K vectors from U



What to select?

- Items that differ from those previously seen

- Principal Components Model

- Approximate model of data set variation

Known items — $X = U\Sigma V^T$

- Keep only the top K vectors from U
- Select items in D that are difficult to represent with model U

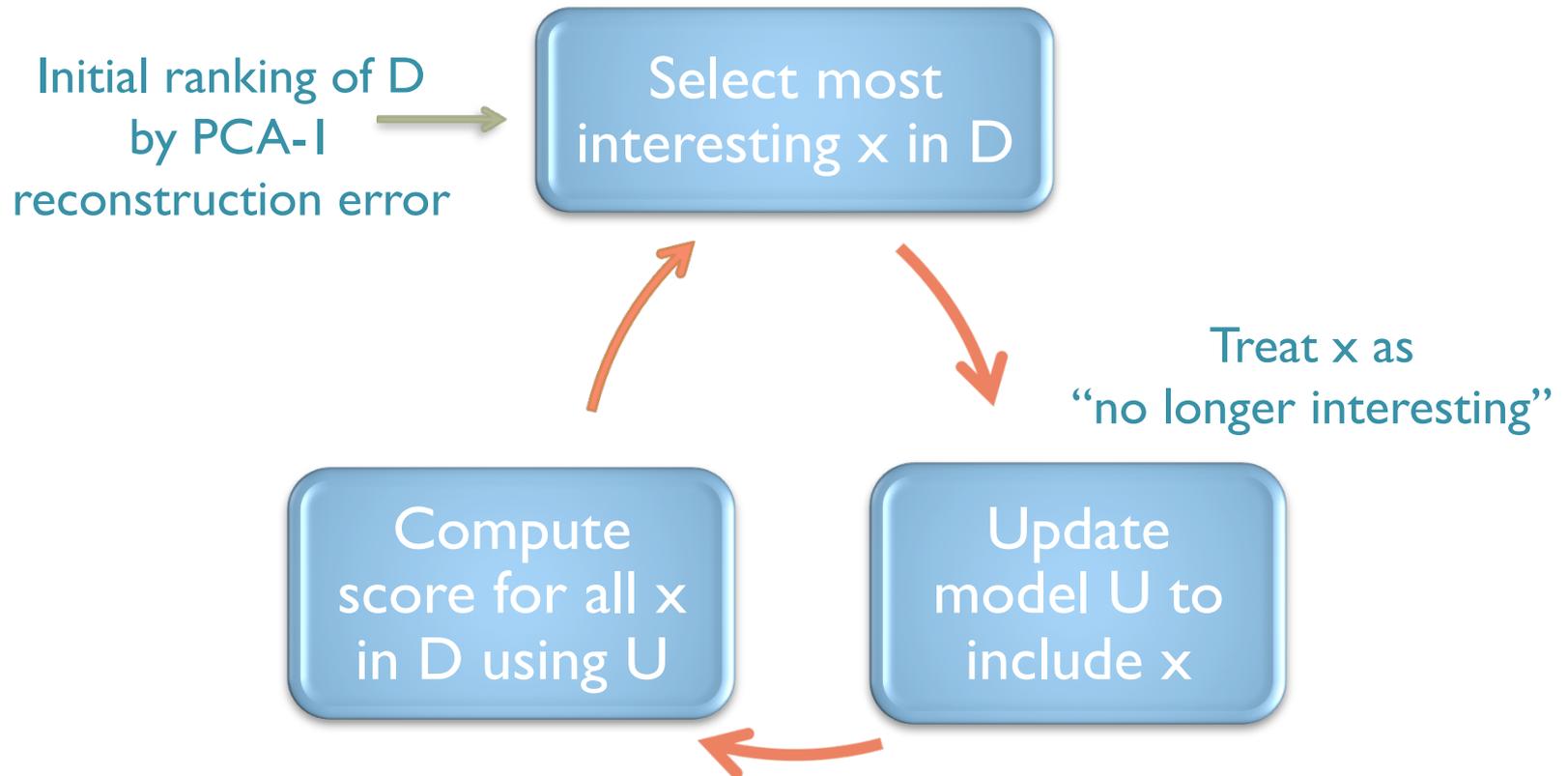
- Reconstruction error

$$R(x) = \|x - \underbrace{(UU^T(x - \mu) + \mu)}_{\text{Reconstruction of } x}\|_2$$

Mean of X

For x in D

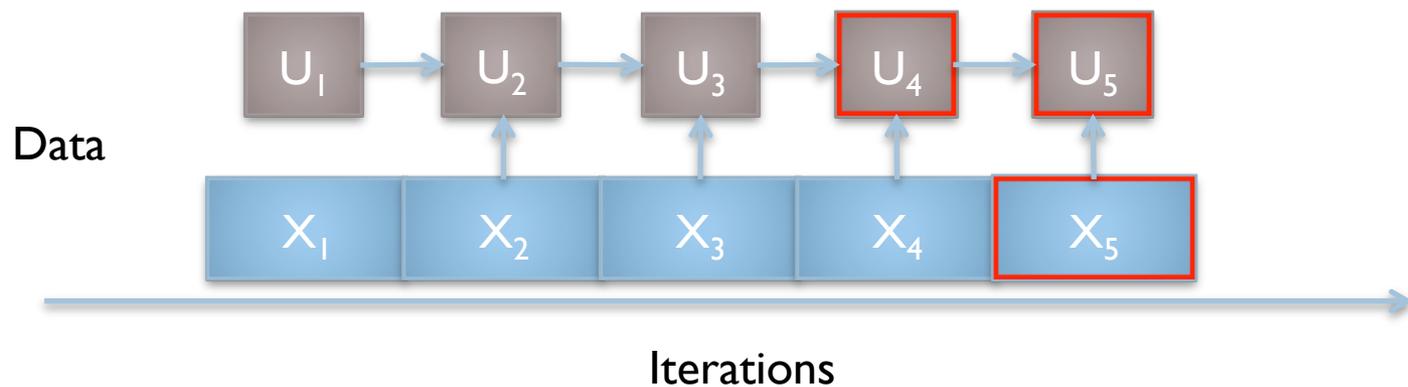
DEMUD: Discovery through Eigenbasis Modeling of Uninteresting Data



Updating model U with new x

- Redo PCA from scratch: expensive
- Incrementally update U : fast!
 - U depends only on previous U and new x
 - [Ross et al., 2008]

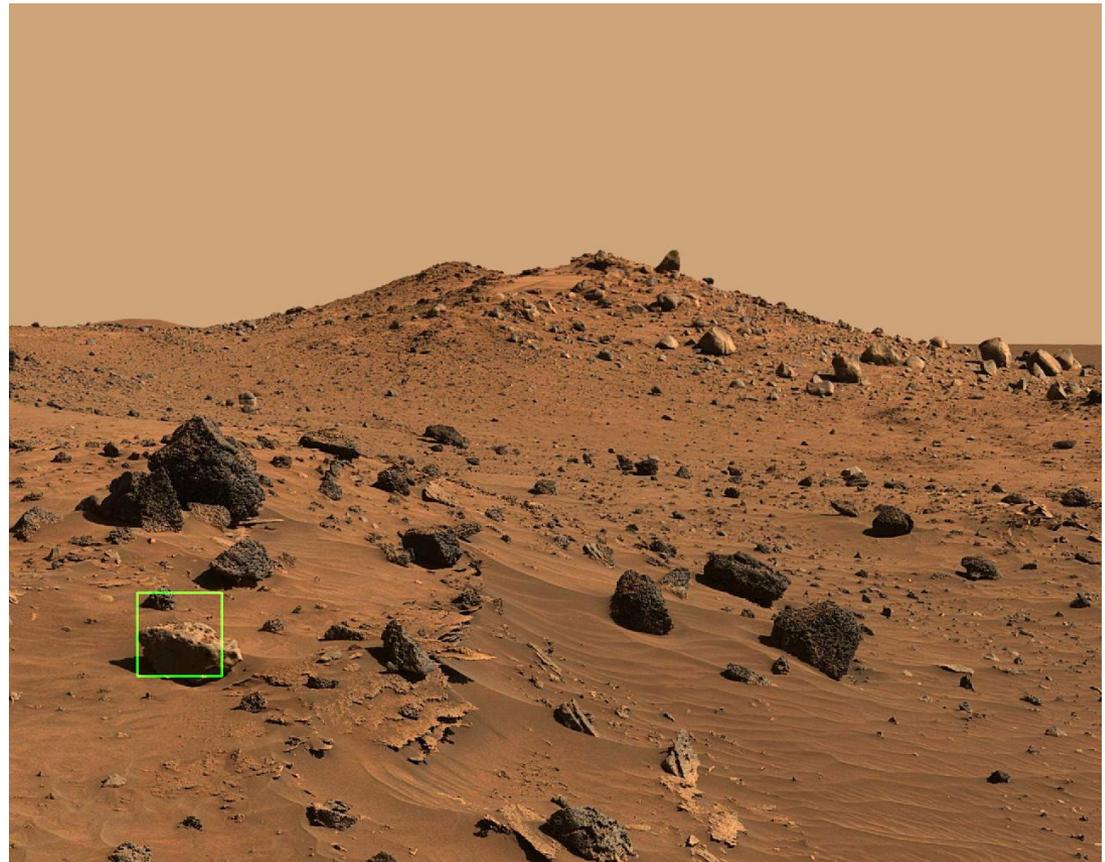
Principal Components



McMurdo selections

- 1200 features: 100x100 RGB, downsamp. 5x
- $K=20$

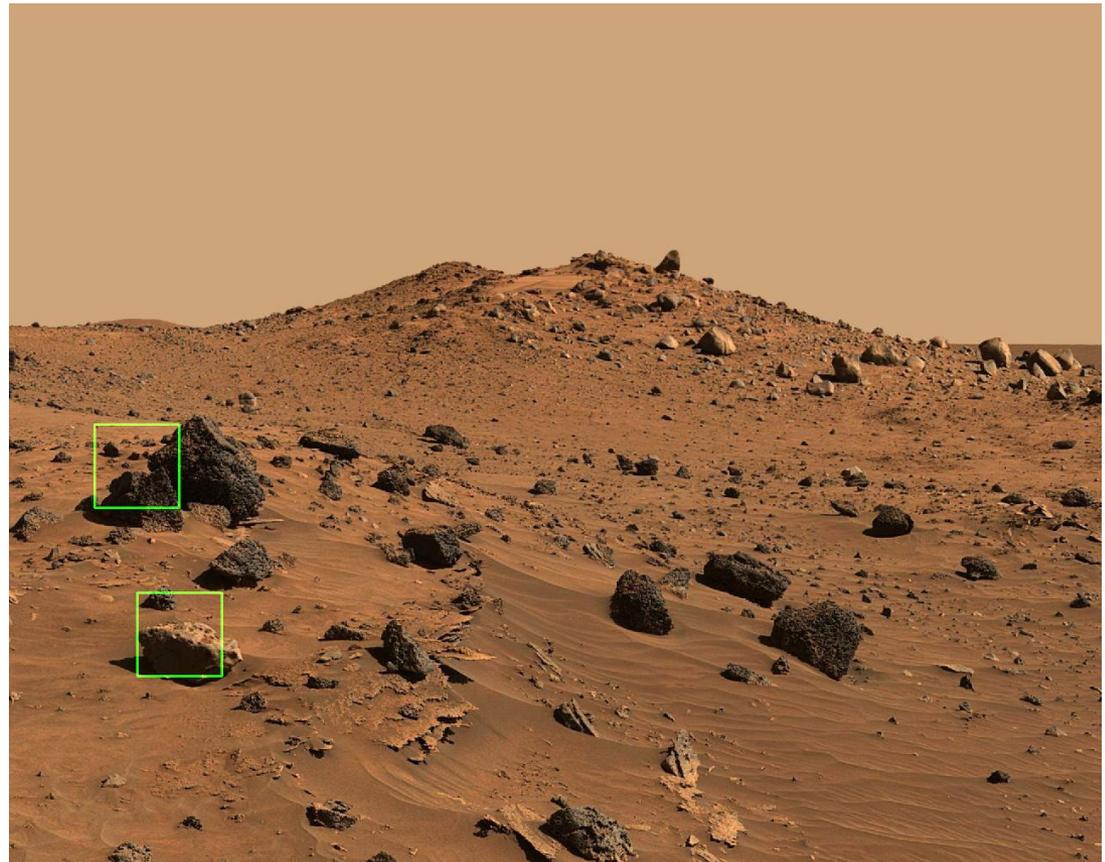
Selection #1



McMurdo selections

- 1200 features: 100x100 RGB, downsamp. 5x
- $K=20$

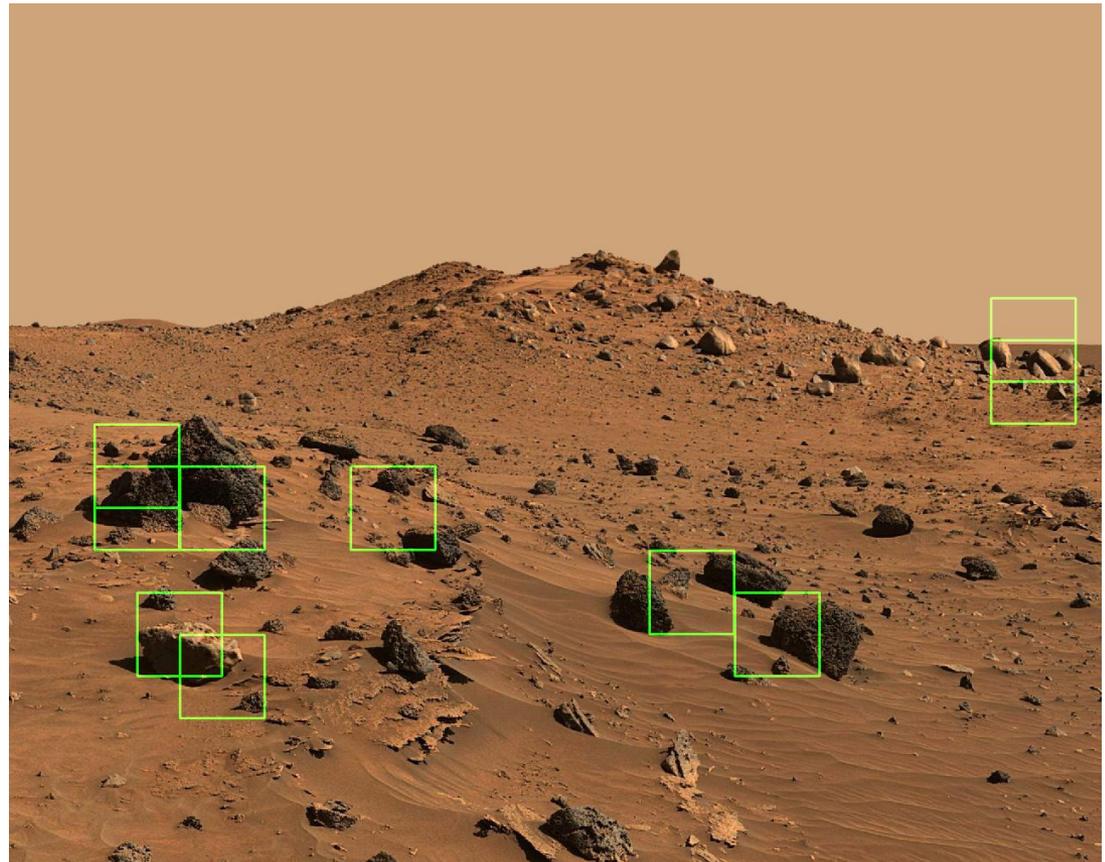
Selection #2



McMurdo selections

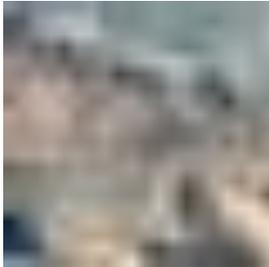
- 1200 features: 100x100 RGB, downsamp. 5x
- $K=20$

Selection #10



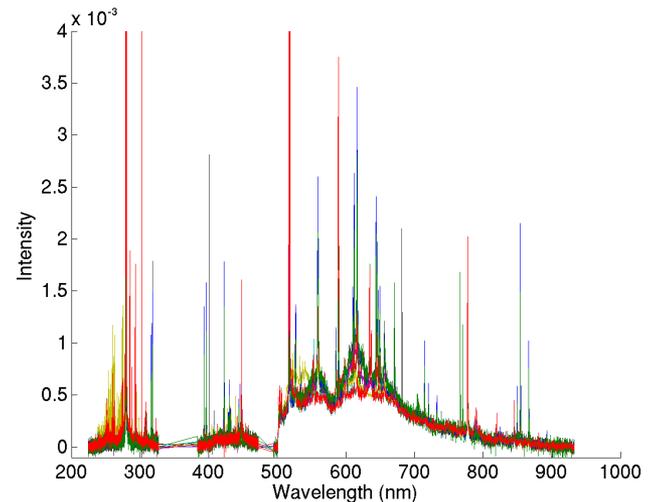
Explanations

- Reconstruction residuals

	Original	Features	Residuals	
1				Dark: lower intensity than expected Bright: higher intensity than expected
2				Dark area at bottom has small residual → learning happened!
3				

ChemCam: Carbonates

- ChemCam: LIBS instrument on MSL
- Data set: 60 lab samples + 40 carbonates
 - 6143 features (bands)
 - K (8) chosen to capture 90% variance



Regular PCA



ISVM-int



DEMUD

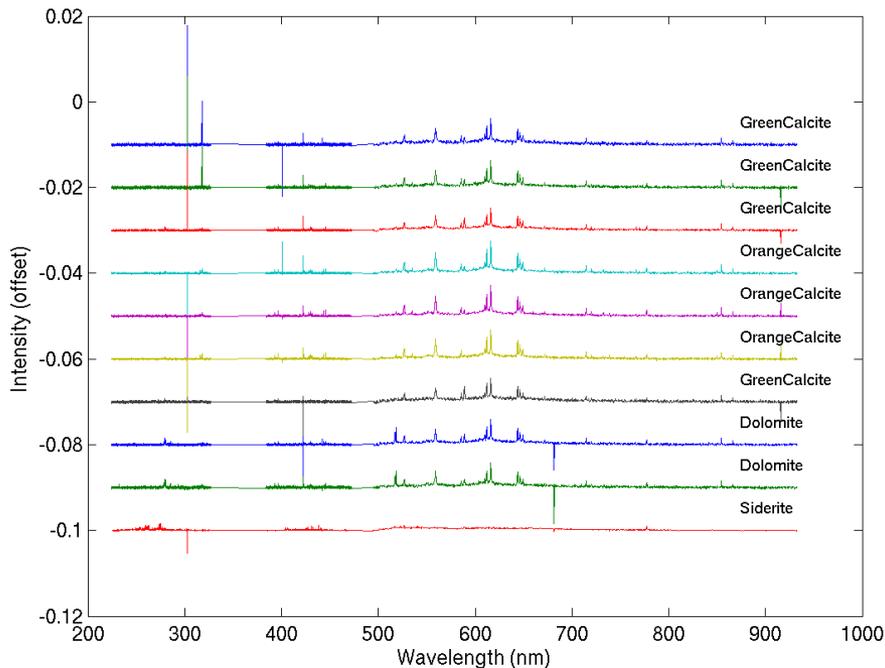


Colored items are carbonates; white are non-carbonates

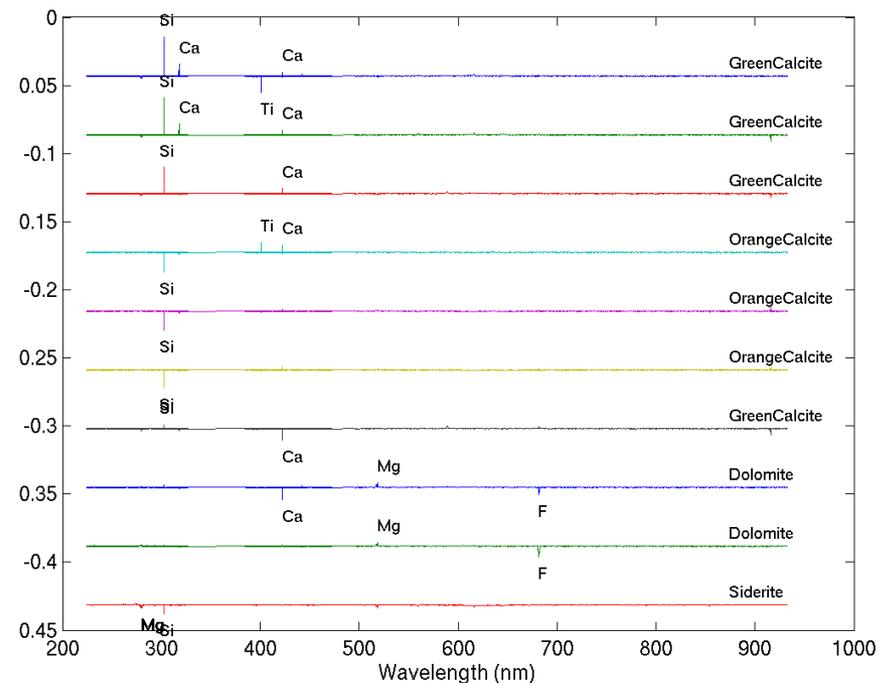
ChemCam: Explanations

- Top 10 items chosen by DEMUD

Ranked by “interestingness”

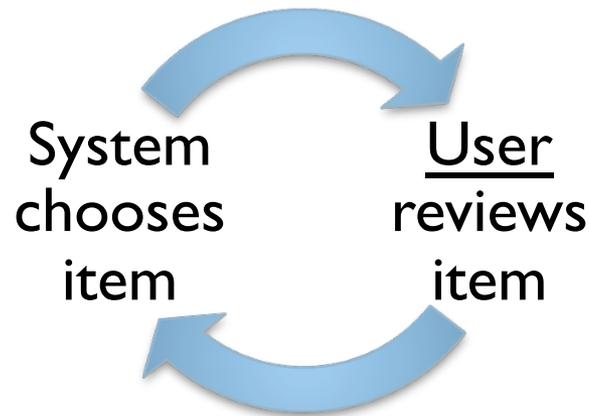


Explanations (residuals)



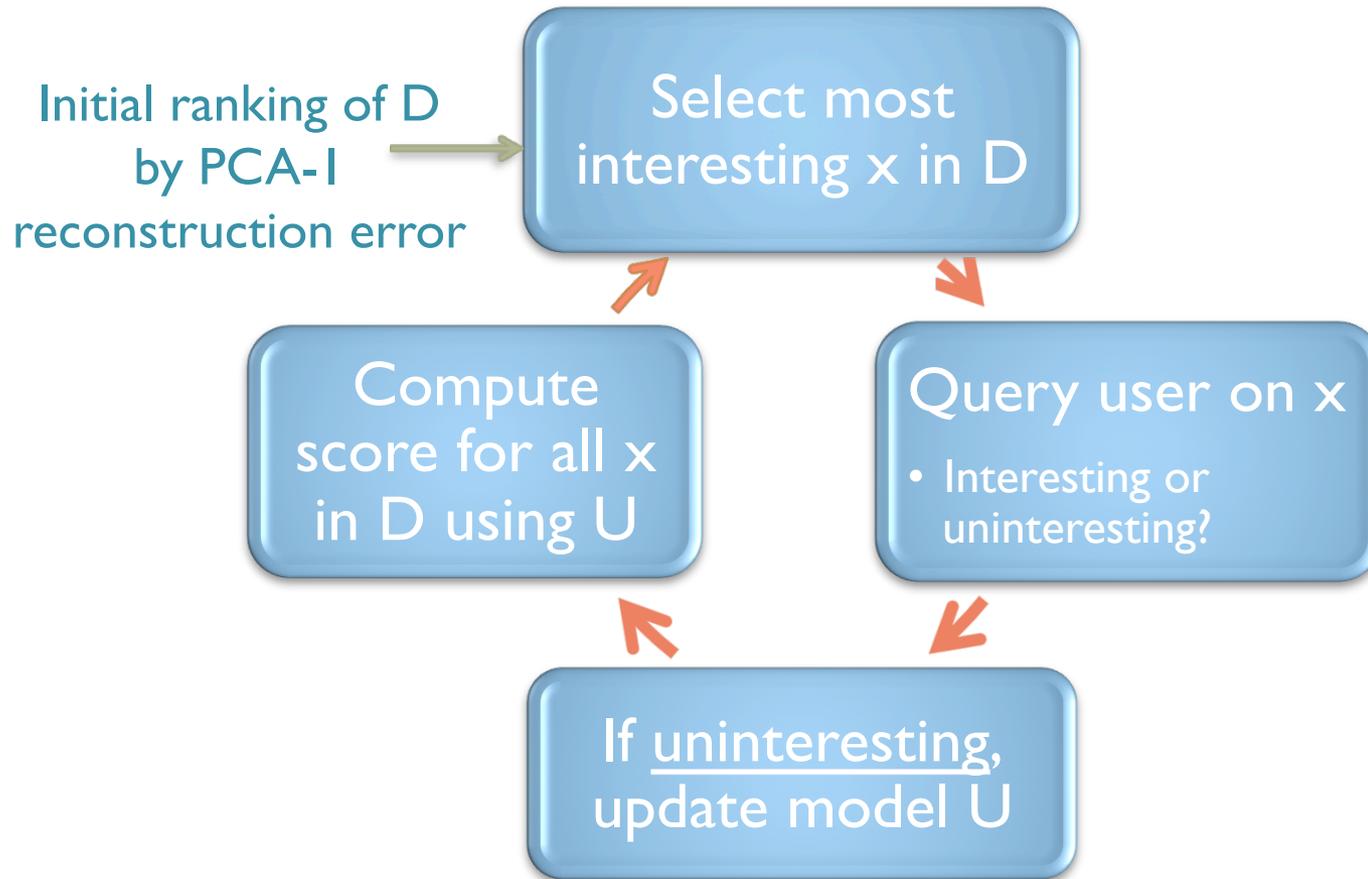
Interactive Discovery

- Guided exploration of large data sets



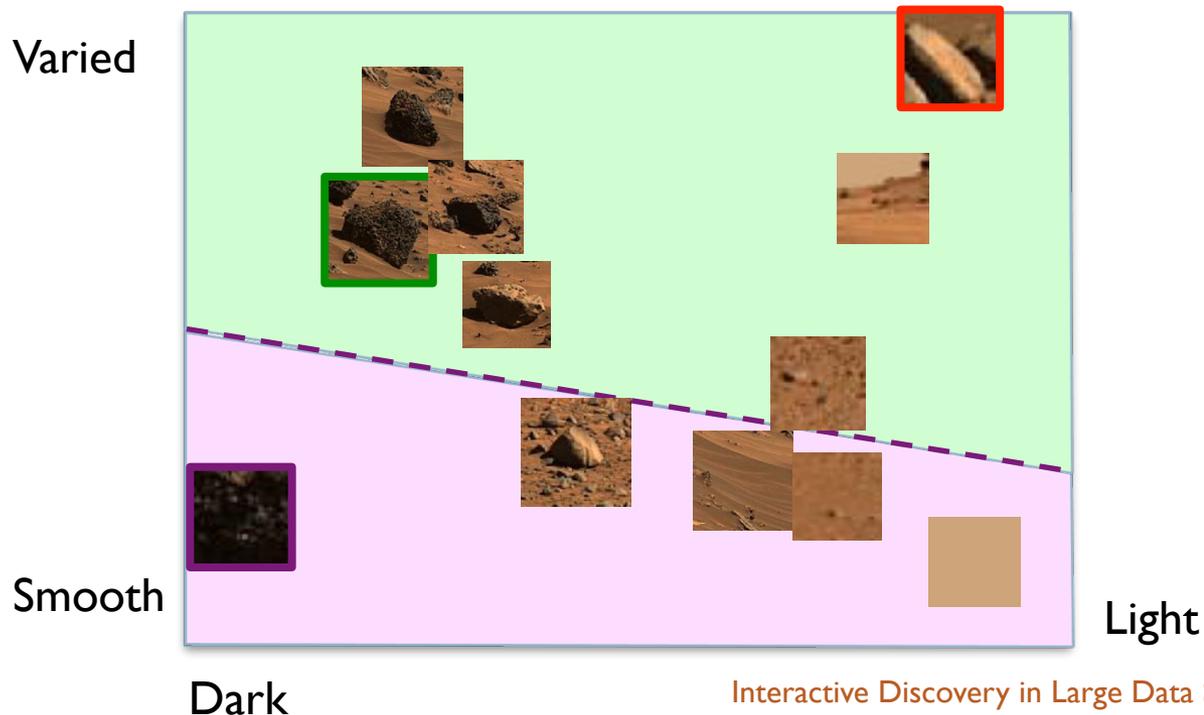
- Desiderata
 - Quickly find items of interest, even if rare
 - Don't miss anything!

Interactive DEMUD



Alternative: Two-class SVM

- Model interesting and uninteresting classes
- Select **most interesting** item

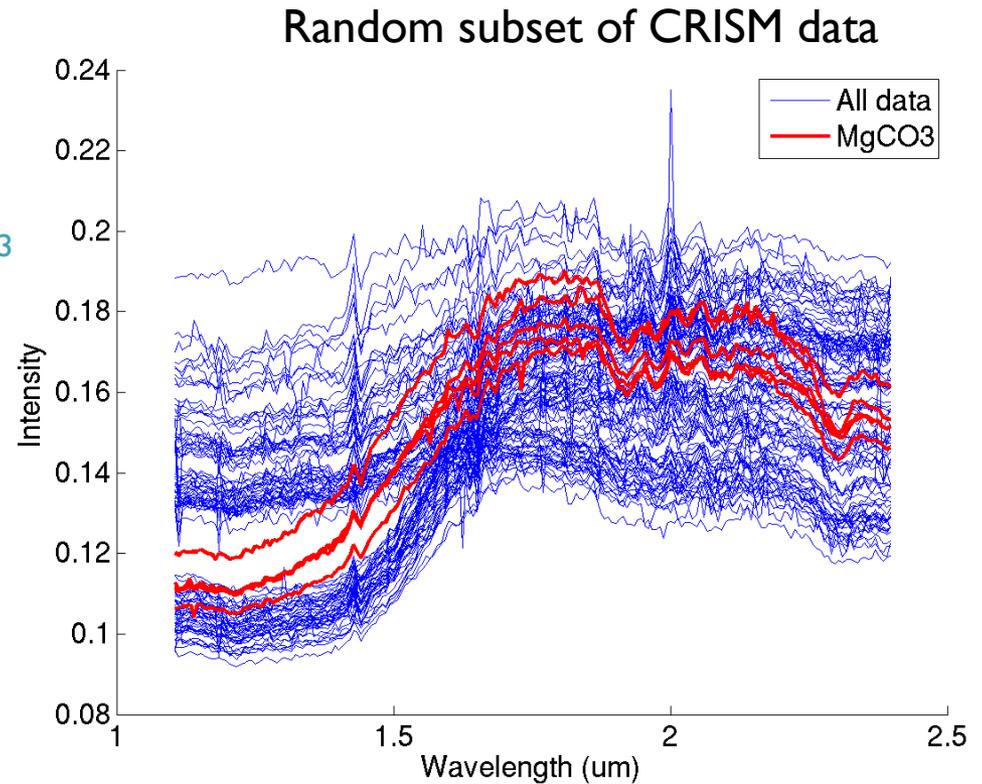
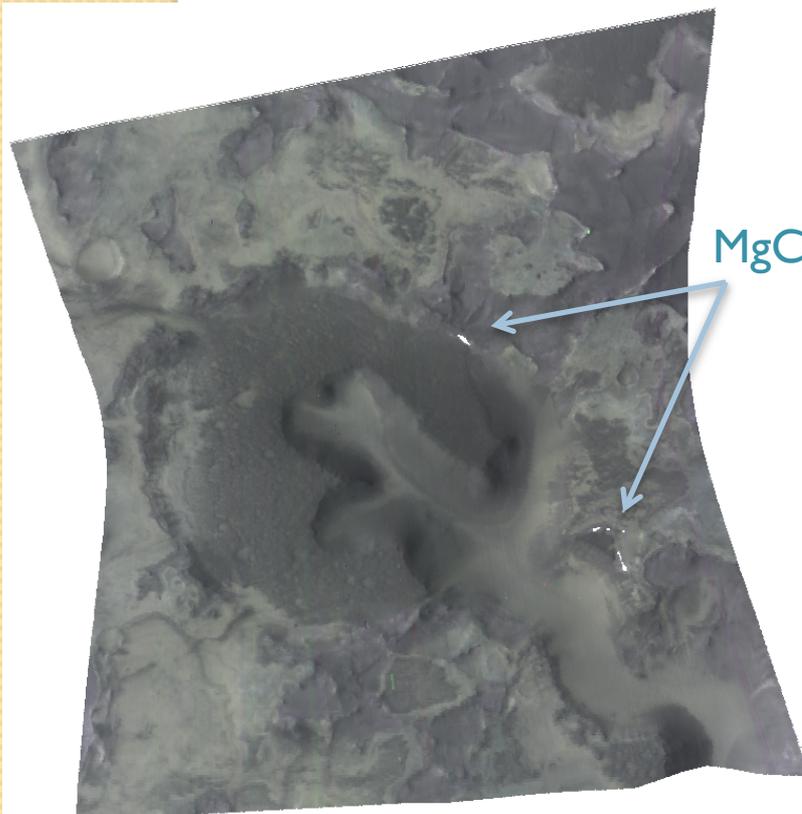


Alternatives: Static Baseline

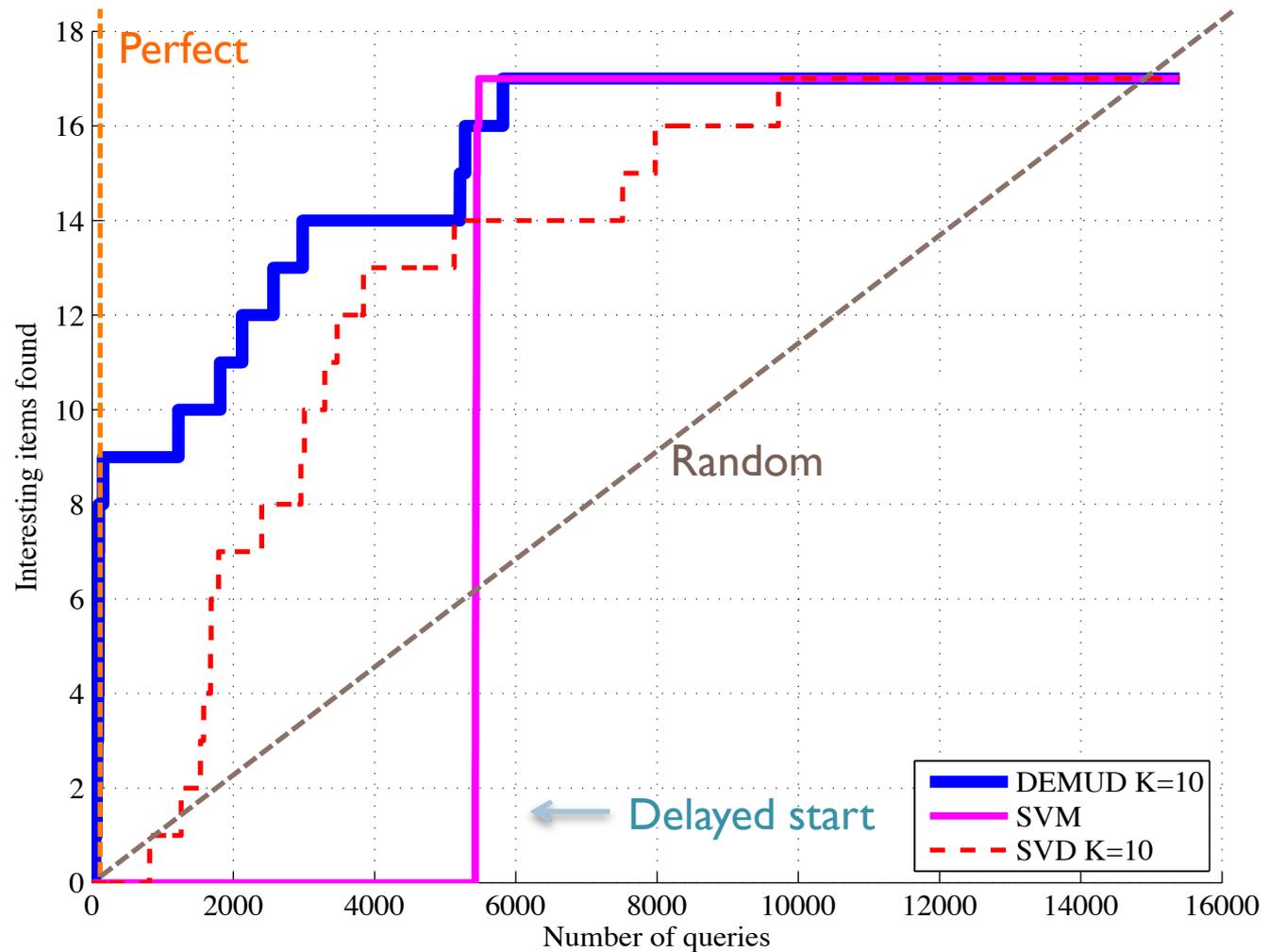
- Select by PCA-K ordering
 - Same initial model as DEMUD
- No feedback

CRISM: Magnesite Discovery

- Magnesite (MgCO_3): possible groundwater deposit
- CRISM data: 0.364 to 3.92 μm , 197 bands
- Only 17 of 15,400 items match



CRISM: Magnesite Discovery



Other Applications

- Text
 - Long Wavelength Array system log files
 - Detect anomalous system behavior
- Onboard prioritization
 - Imaging spectrometers
 - Hyperion on EO-1: 256x6000x242
 - Assign priorities for input to onboard compression: ROI-ICER

Summary

- Discovery
 - PCA-based model + reconstruction error
- Explanations
 - Why was it chosen?
- Interactive discovery
 - Model the uninteresting to avoid it
- Next challenge: evolving class of interest

Thank you!

Contact: kiri.wagstaff@jpl.nasa.gov



Faces Data Set

- 40 people, 10 poses each
- High dimensionality: 10,304
- Goal: Discover 3 women
 - Data set is mostly men
 - Challenge: disjunction



Faces: Three Women Discovery

