Detecting Related Message Traffic

David Skillicorn
School of Computing, Queen's University
Math and CS, Royal Military College
skill@cs.queensu.ca

The problem:

Connect intercepted messages into conversations when conventional markers (sender/receivers etc.) are missing.

The solution:

Look for correlated use of words that are used with the "wrong" frequency.

The technique:

Use singular value decomposition and independent component analysis applied to noun frequency profiles; suspicious related messages appear as outliers.

Why it's interesting:

Applications in counterterrorism (also fraud).

THE PROBLEM

Many governments collect and analyze message traffic (e.g. Echelon) – email, file traffic/web, cellphone traffic, radio.

There are 3 analysis strategies:

1. Match the content of individual messages against a watch list of words that suggest the message is suspicious.

German Federal Intelligence Service: nuclear proliferation (2000 terms), arms trade (1000), terrorism (500), drugs (400), as of 2000 (certainly changed now).

Countermeasures: use a speech code (hard in realtime) or use locutions ("the package is ready").

2. Look for sets of messages that are connected, that form a conversation, based on some of their properties: sender/receiver identities, time of transmission, specialized word use, etc..

(Social Network Analysis)

Countermeasures: conceal the connections between the messages by making sure they share no obvious attributes:

* use temporary email addresses, stolen cell phones
* decouple by using intermediaries
* smear time factors e.g. by using web sites

In general, *hide in the background noise* .

3. Look for sets of messages that are connected in more subtle ways because of *correlation* among their properties.

Workable countermeasures are hard to find because:

* conversations are about something, so that correlation in their content arises naturally
* sensitivity to watch list surveillance alters the way words are used

We hypothesize that related messages among a threat group in the context of watch list surveillance will be characterized by correlated word use; and that the words will be used with the "wrong" frequencies.

Common words will be used as if they were uncommon; uncommon words will be used as if they were common.

THE DATA

The frequency of words in English (and many other languages) is Zipf – frequent words are *very* frequent, and frequency drops off very quickly.

We restrict our attention to nouns.

In English
Most common noun – time
3262nd most common noun – quantum

We assume that messages are reduced to a frequency histogram of their nouns (this can be done reliably with a tagger).

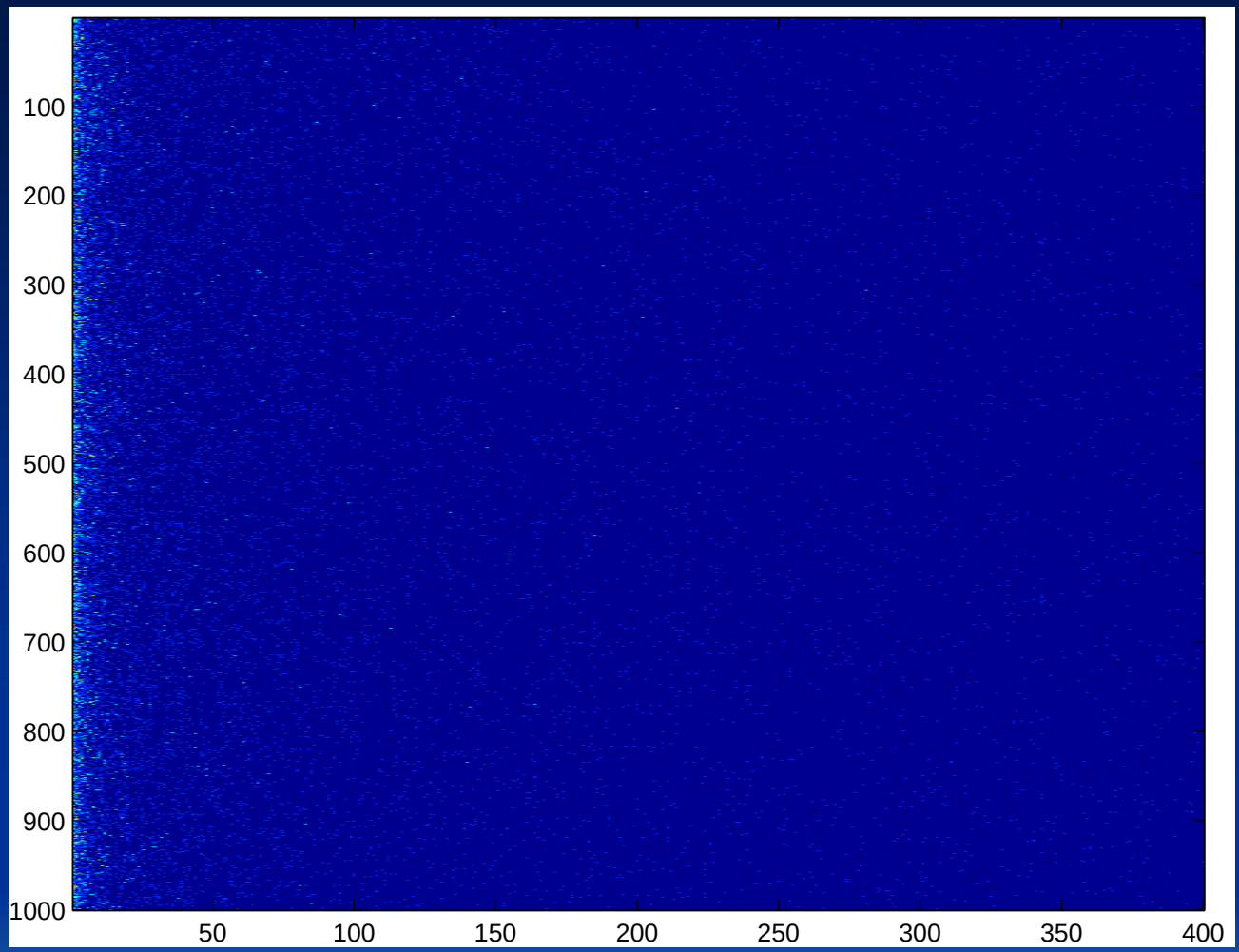
The dataset has a row corresponding to each message, and a column corresponding to each noun. The *ij* th entry is the frequency of noun *j* in message *i* .

The matrix is very sparse.

We generate artificial datasets using a Poisson distribution with mean *f * 1/j+1* , where f models the base frequency.

We add 10 extra rows representing the correlated threat messages, using a block of 6 columns, uniformly randomly 0s and 1s, added at columns 301—306.

messages



nouns

THE TECHNIQUES

Matrix decompositions.

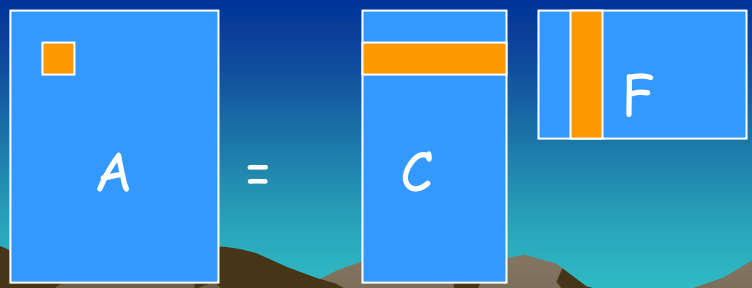
The basic idea:

* Treat the dataset as a matrix, A, with n rows and m columns;
* Factor A into the product of two matrices, C and F

$$A = C F$$

where C is n x r, F is r x m and r is smaller than m.

Think of F as a set of underlying `real' somethings and C as a way of `mixing' these somethings together to get the observed attribute values. Choosing r smaller than m forces the decomposition to somehow represent the data more compactly.



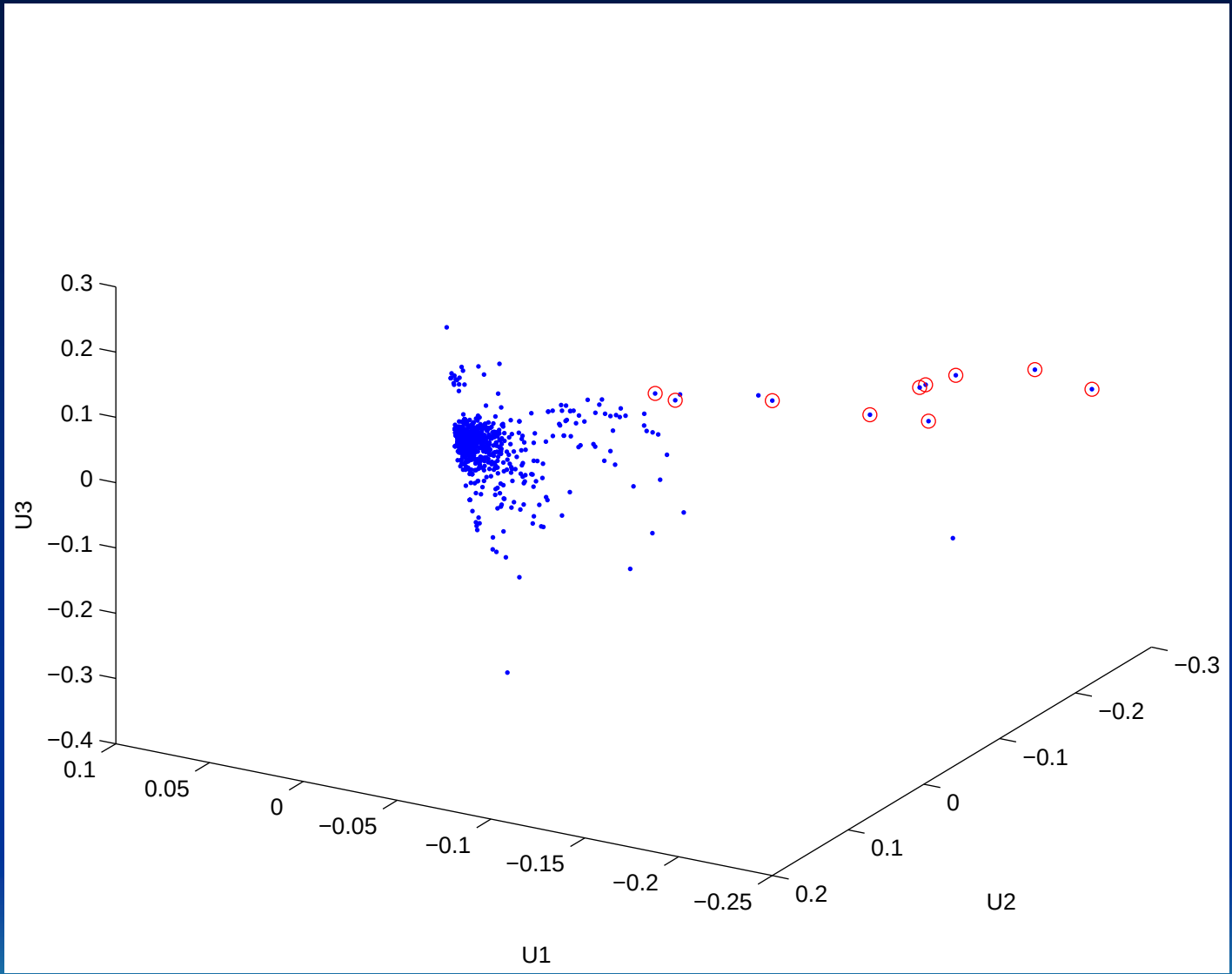
Two matrix decompositions are useful :

Singular value decomposition (SVD) – the rows of F are orthogonal axes such that the maximum possible variation in the data lies along the first axis; the maximum of what remains along the second, and so on. The rows of C are coordinates in this space.

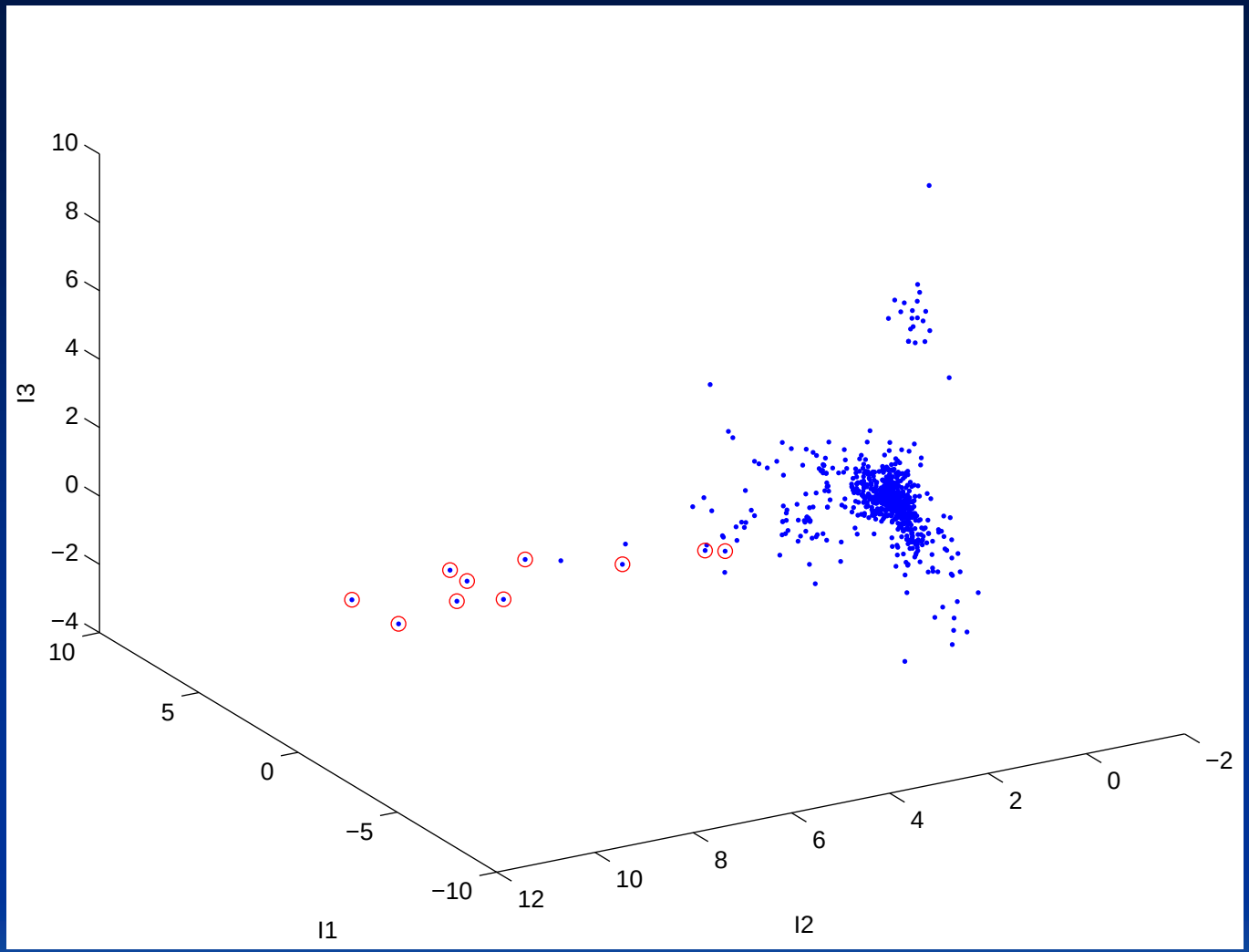
Independent component analysis (ICA) – the rows of F are statistically independent factors. The rows of C describe how to mix these factors to produce the original data.

Strictly speaking, the row of C are not coordinates, but we can plot them to get some idea of structure.

The messages with correlated unusual word usage are marked with red circles



First 3 dimensions – SVD

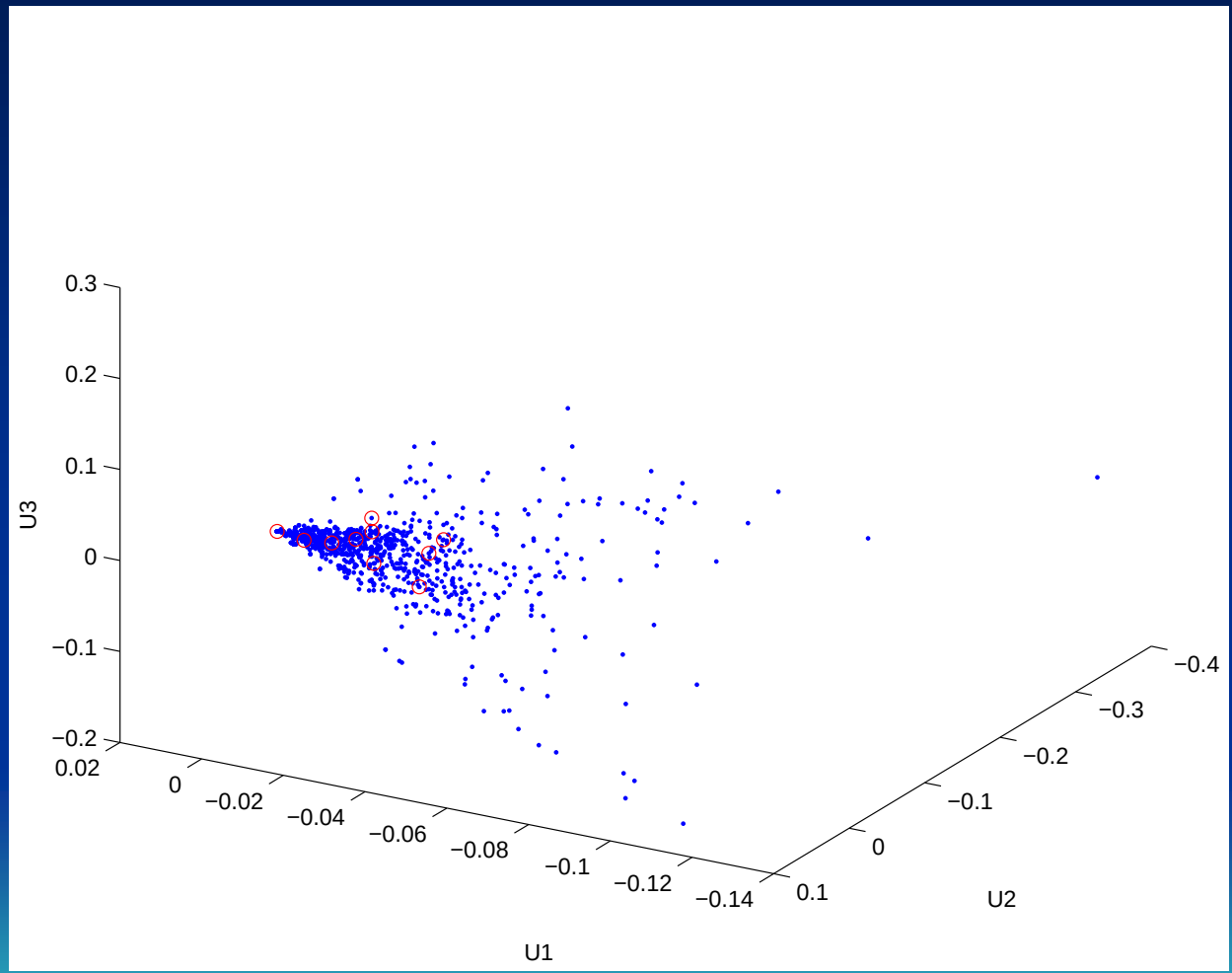


First 3 dimensions - ICA

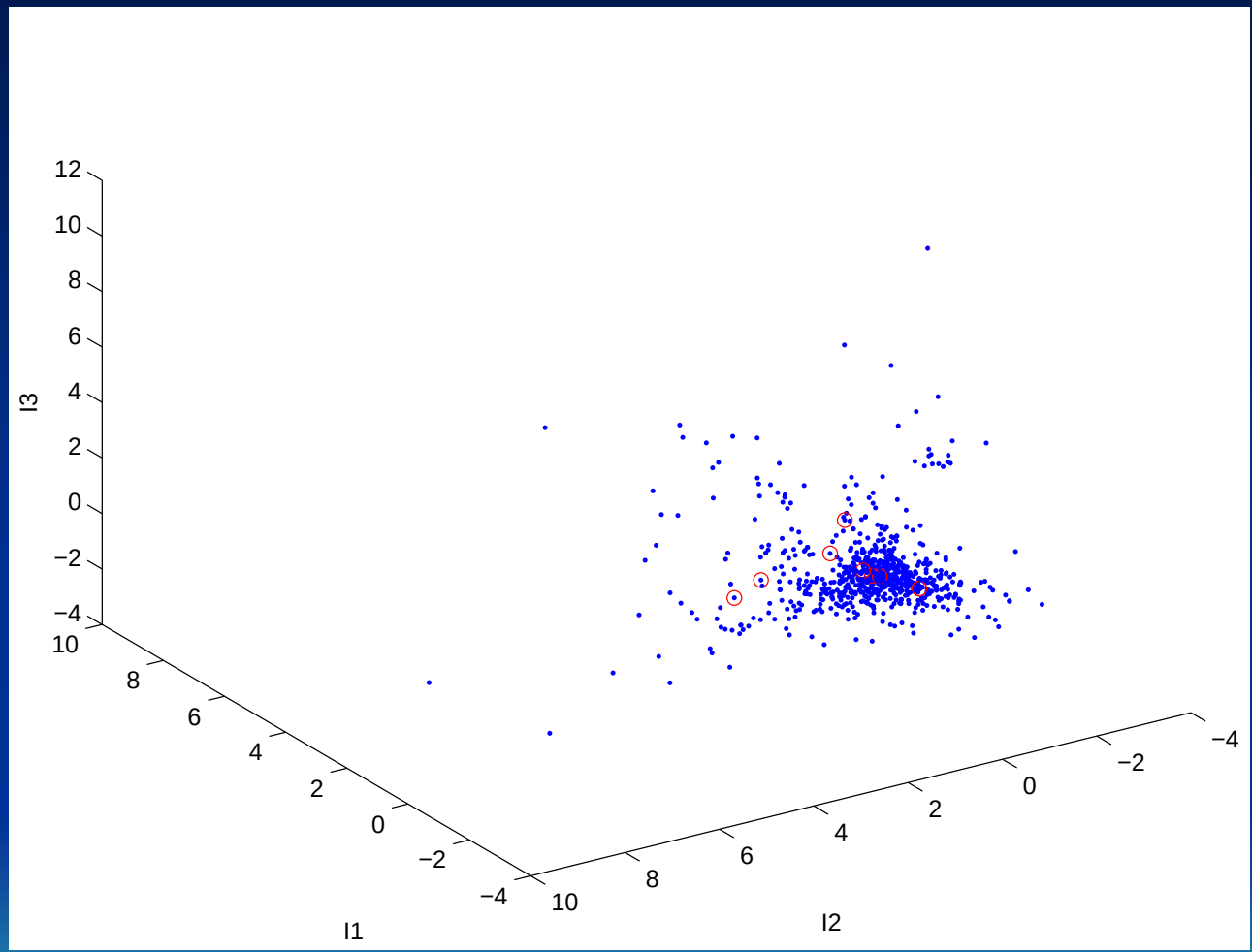
(Fortunately) both unusual word use and correlated word use are necessary to make such messages detectable.

Correlation with proper word frequencies (SVD)

So ordinary conversations don't show up as false positives!!

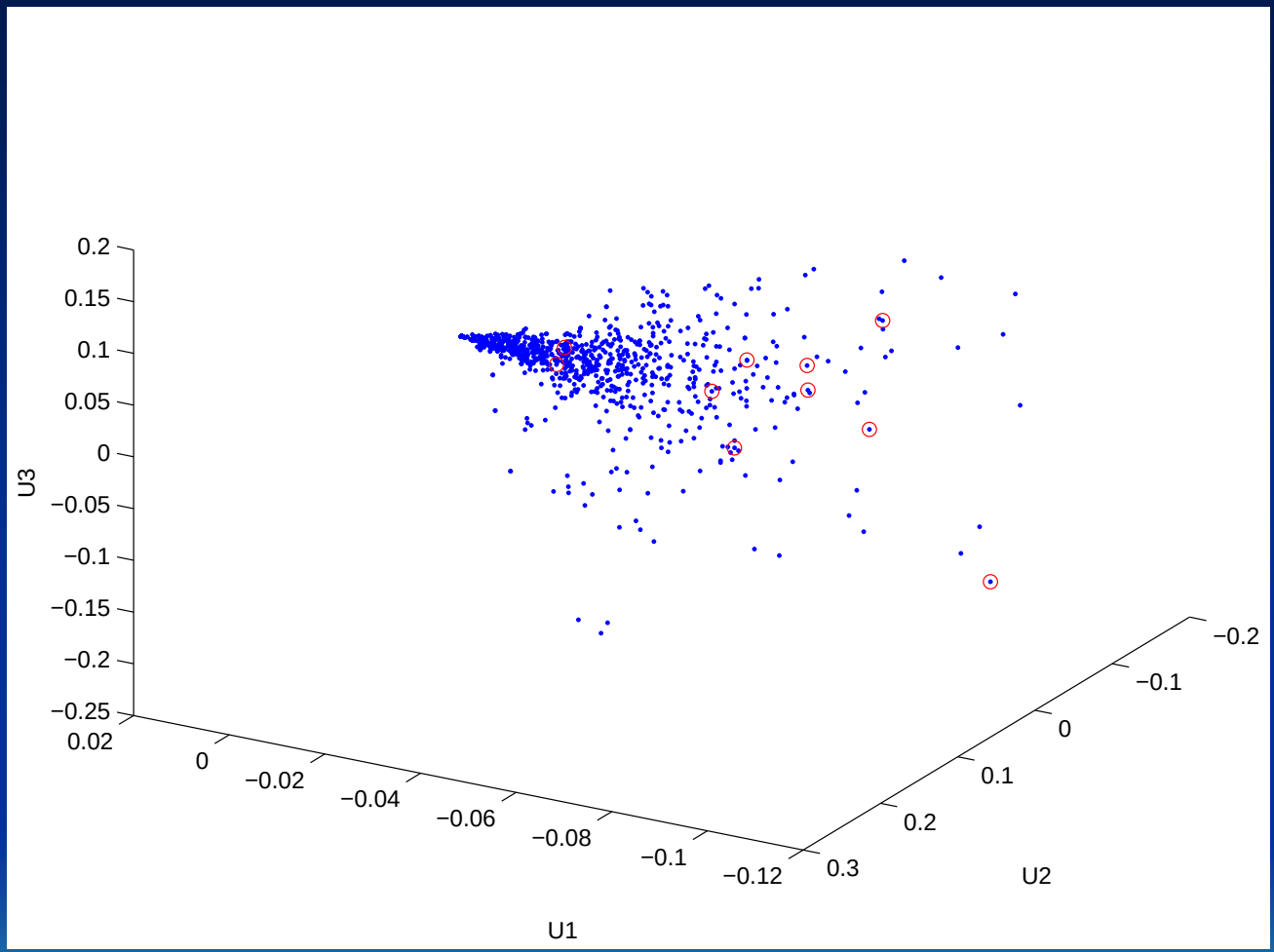


Correlation with proper word frequencies (ICA)

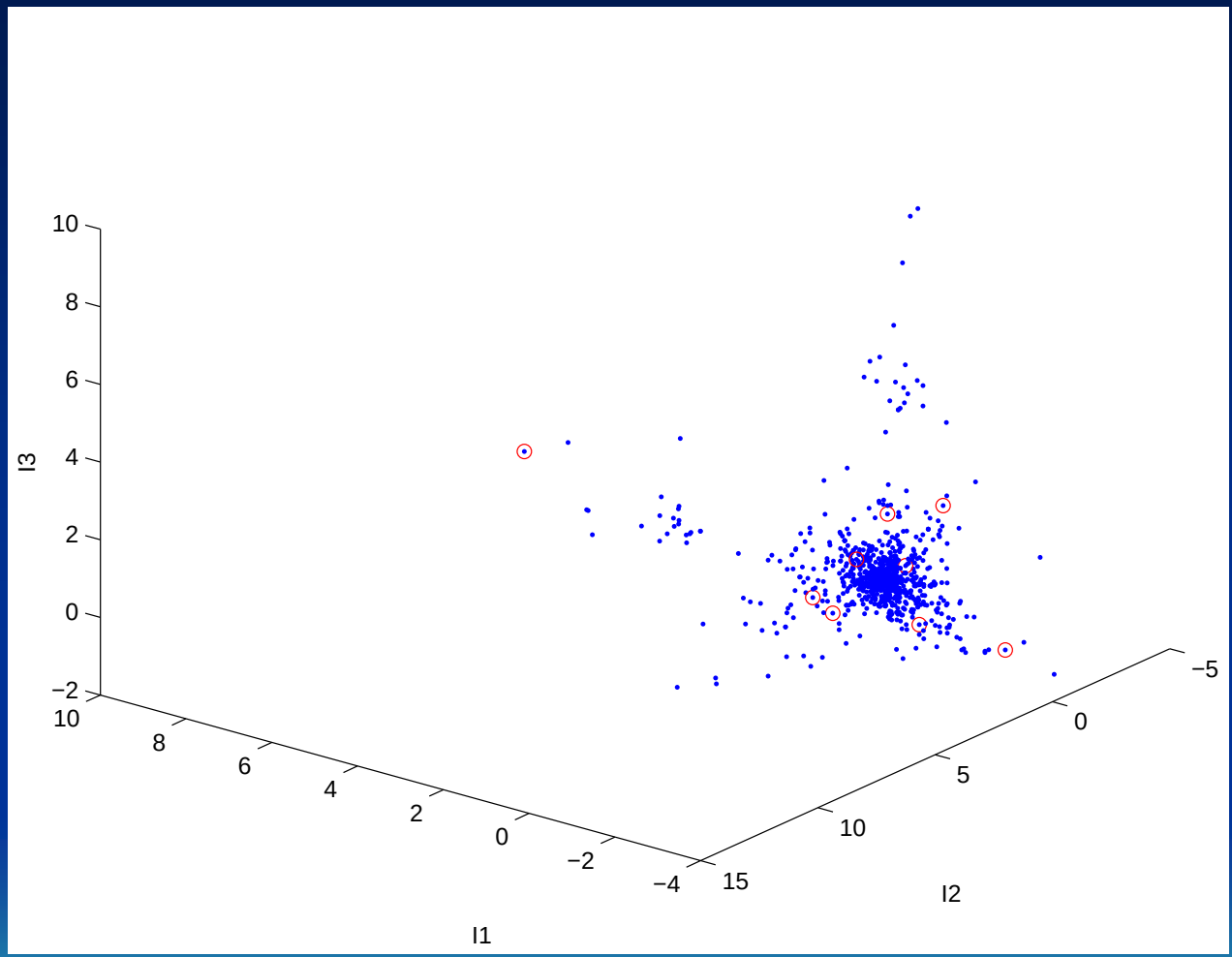


Uncorrelated with unusual word frequencies (SVD)

Conversations about unusual things don't show up as false positives either!!



Uncorrelated with unusual word frequencies (ICA)



This trick permits a new level of sophistication in connecting related messages into conversations when the usual indicators are not available.

It does exactly the right thing – ignoring conversations about ordinary topics, and conversations about unusual topics, but homing in on conversations about unusual topics using inappropriate words.

Because the dataset is sparse, SVD takes time linear in the number of messages. The complexity of ICA is less clear but there are direct hardware implementations.