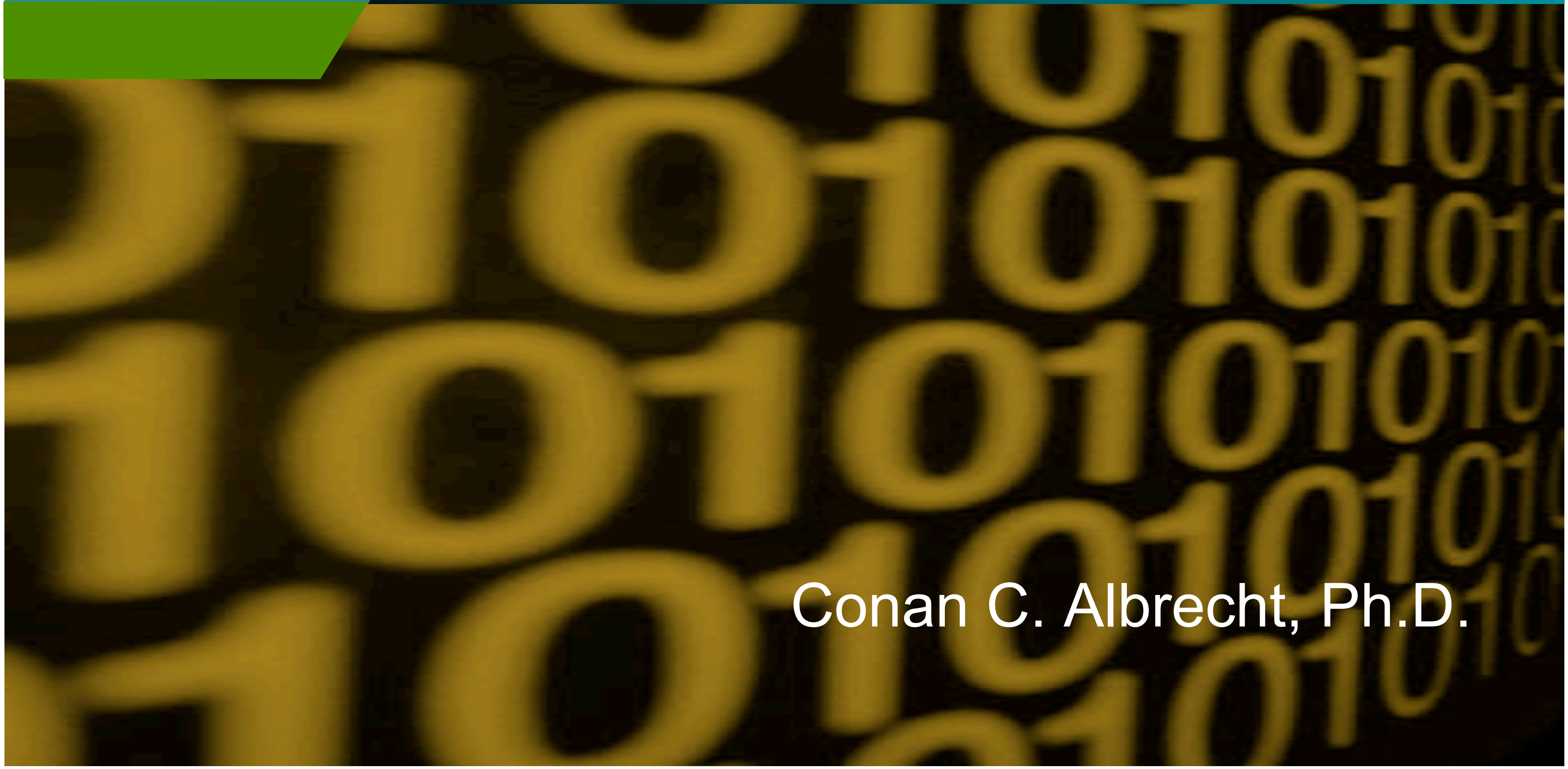
Data Analysis



Conan C. Albrecht, Ph.D.

Module: Descriptives

Simple Descriptives

- Record counts
- Field (column) totals
- Range
- Average
- Standard Deviation
- Histogram (stem and leaf)

Descriptives in Picalo, ACL

Generating Norms

- Norms are generated in two ways
  - *Theory*: the fraud test being run determines the norm
    - Example: No overtime in a company. Anyone over 40 hours per week is a concern
  - *Data*: let the data speak for itself
    - Calculate norms from the entire population of data, then compare each transaction/group with the norm
    - Example: Average price of a painter

Where’s My Outliers?

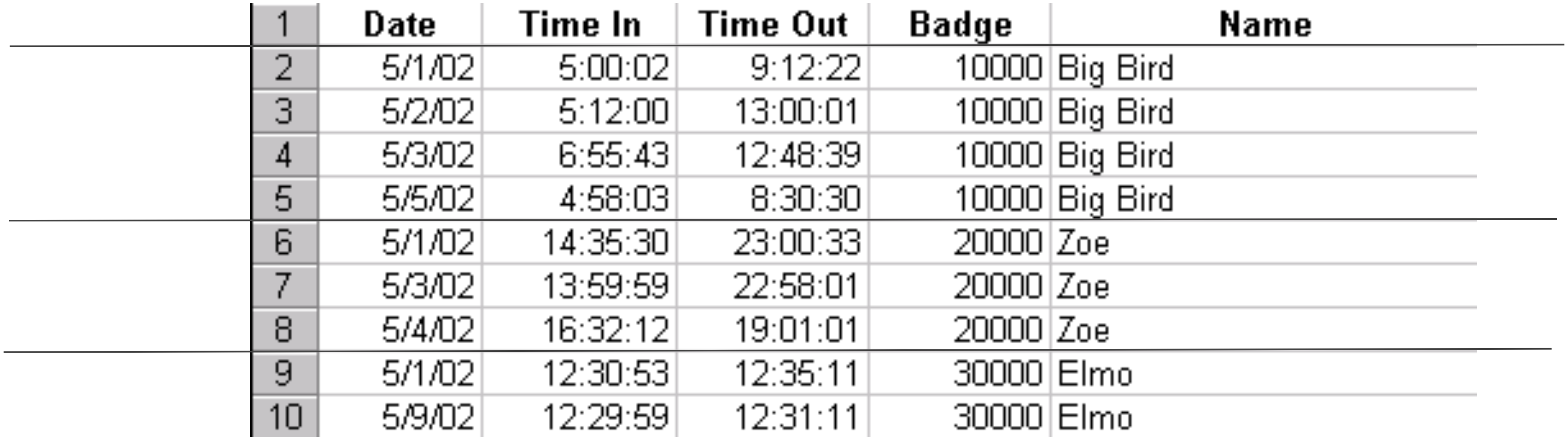
- Some analyses focus on outliers
  - Overtime, pay rates
- Some analyses exclude outliers
  - Average product prices
- Z-Score - measure of distance for each point

```
        (value - mean) / std dev
```

  - 68% is between -1 and 1
  - 95% is between -2 and 2
  - 99.7% is between -3 and 3

Table: chargessmall
Add z-score column

Grouping (stratification)



1	Date	Time In	Time Out	Badge	Name
2	5/1/02	5:00:02	9:12:22	10000	Big Bird
3	5/2/02	5:12:00	13:00:01	10000	Big Bird
4	5/3/02	6:55:43	12:48:39	10000	Big Bird
5	5/5/02	4:58:03	8:30:30	10000	Big Bird
6	5/1/02	14:35:30	23:00:33	20000	Zoe
7	5/3/02	13:59:59	22:58:01	20000	Zoe
8	5/4/02	16:32:12	19:01:01	20000	Zoe
9	5/1/02	12:30:53	12:35:11	30000	Elmo
10	5/9/02	12:29:59	12:31:11	30000	Elmo

By Badge ID or Name

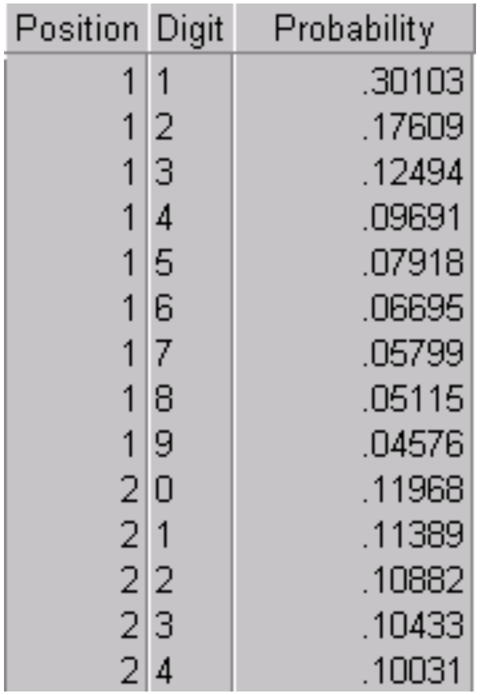
Grouping (stratification)

- Grouping data is a basic analysis technique
  - Column values
  - Ranges
  - Dates and Aging
- Most tables are thousands of tables in one
  - How you split it depends upon the analysis

Table: chargessmall
Software: IDEA, Picalo
Stratify by vendor, purchaser

Benford’s Law

- Invoice numbers are not truly random
- Invoice numbers follow a predictable pattern
- Human-generated (fraudulent) numbers do not follow the pattern



Position	Digit	Probability
1	1	.30103
1	2	.17609
1	3	.12494
1	4	.09691
1	5	.07918
1	6	.06695
1	7	.05799
1	8	.05115
1	9	.04576
2	0	.11968
2	1	.11389
2	2	.10882
2	3	.10433
2	4	.10031

Picalo: Benford’s law detectlets

Benford’s Law

- How does Benford’s Law help fraud investigators?

- When is it useful?

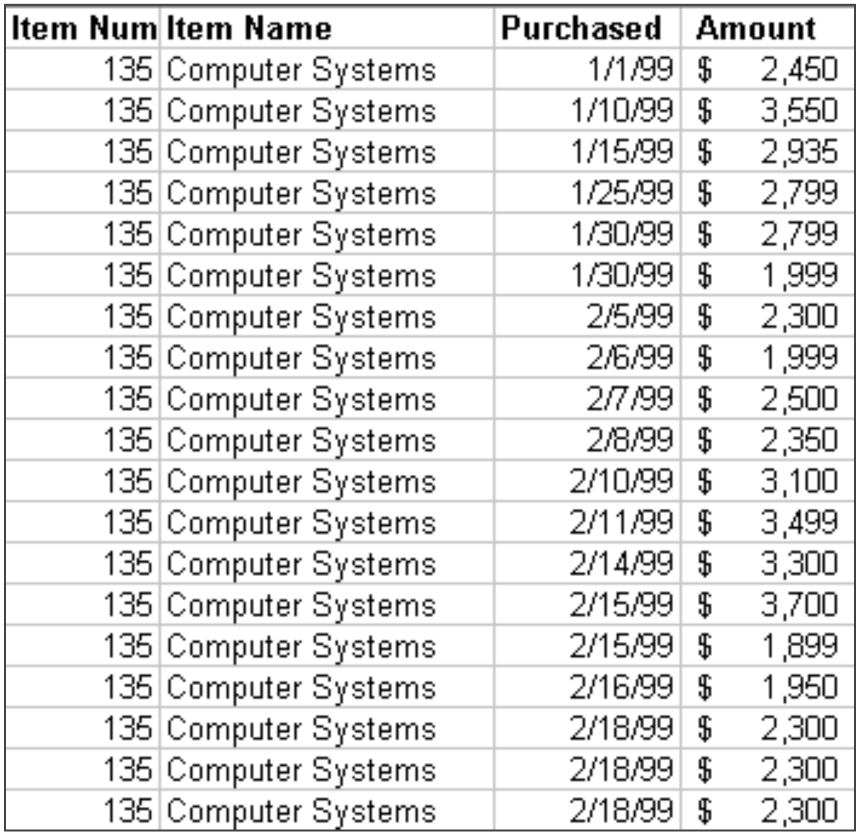
- What are its limitations?

Module: Trending

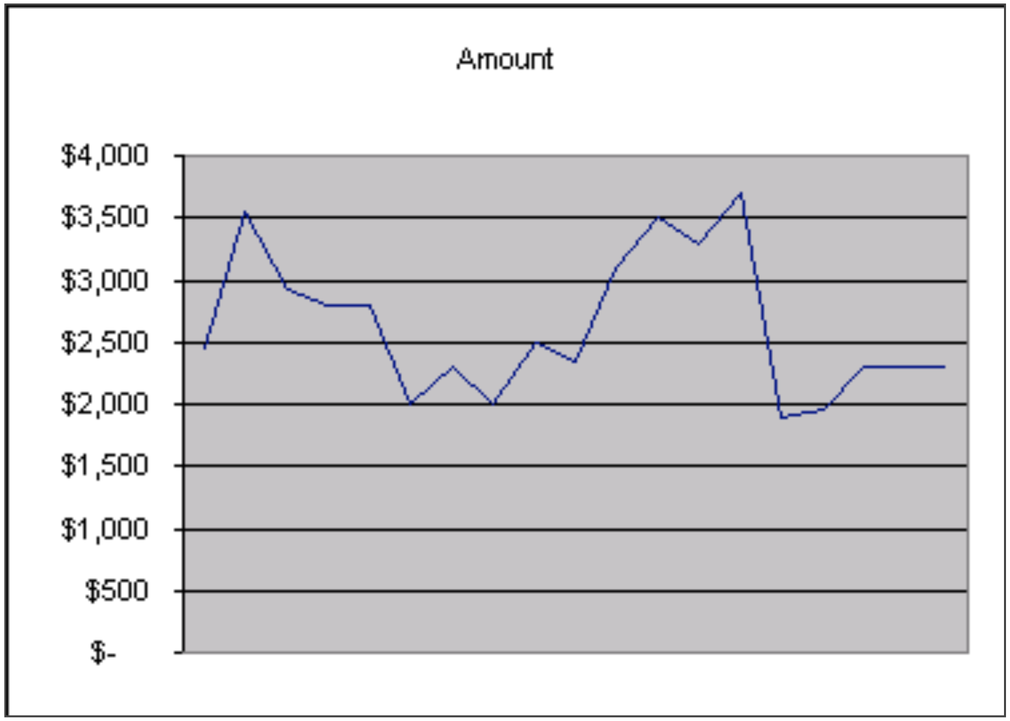
Analyzing Time Trends

- Most fraud is found by analyzing changes over time
- Databases are good to keep transactions, but not so good at standardizing over time

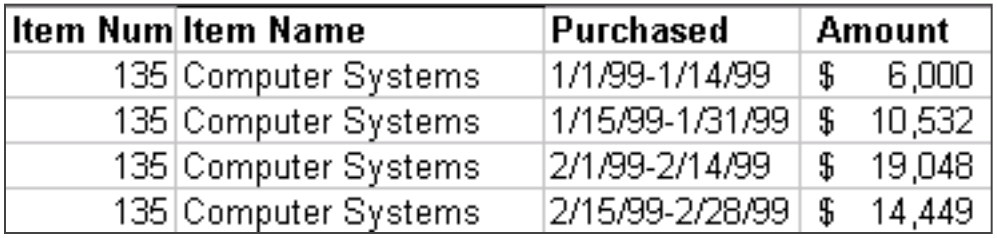
Unstandardized Time Axis



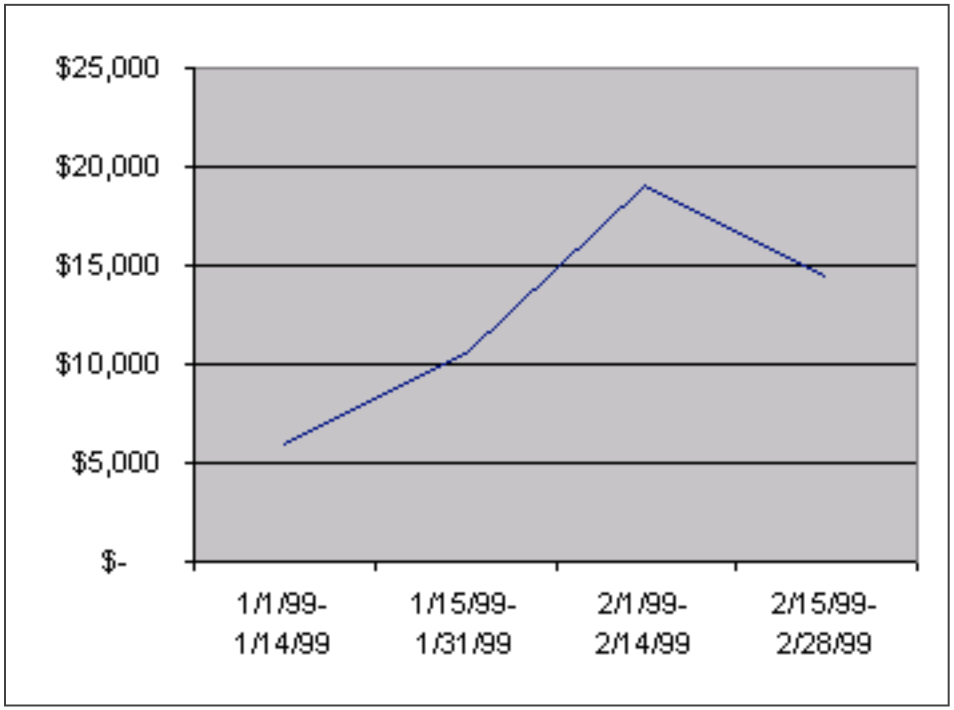
Item Num	Item Name	Purchased	Amount
135	Computer Systems	1/1/99	$ 2,450
135	Computer Systems	1/10/99	$ 3,550
135	Computer Systems	1/15/99	$ 2,935
135	Computer Systems	1/25/99	$ 2,799
135	Computer Systems	1/30/99	$ 2,799
135	Computer Systems	1/30/99	$ 1,999
135	Computer Systems	2/5/99	$ 2,300
135	Computer Systems	2/6/99	$ 1,999
135	Computer Systems	2/7/99	$ 2,500
135	Computer Systems	2/8/99	$ 2,350
135	Computer Systems	2/10/99	$ 3,100
135	Computer Systems	2/11/99	$ 3,499
135	Computer Systems	2/14/99	$ 3,300
135	Computer Systems	2/15/99	$ 3,700
135	Computer Systems	2/15/99	$ 1,899
135	Computer Systems	2/16/99	$ 1,950
135	Computer Systems	2/18/99	$ 2,300
135	Computer Systems	2/18/99	$ 2,300
135	Computer Systems	2/18/99	$ 2,300



Standardized Time Axis



Item Num	Item Name	Purchased	Amount
135	Computer Systems	1/1/99-1/14/99	$ 6,000
135	Computer Systems	1/15/99-1/31/99	$ 10,532
135	Computer Systems	2/1/99-2/14/99	$ 19,048
135	Computer Systems	2/15/99-2/28/99	$ 14,449



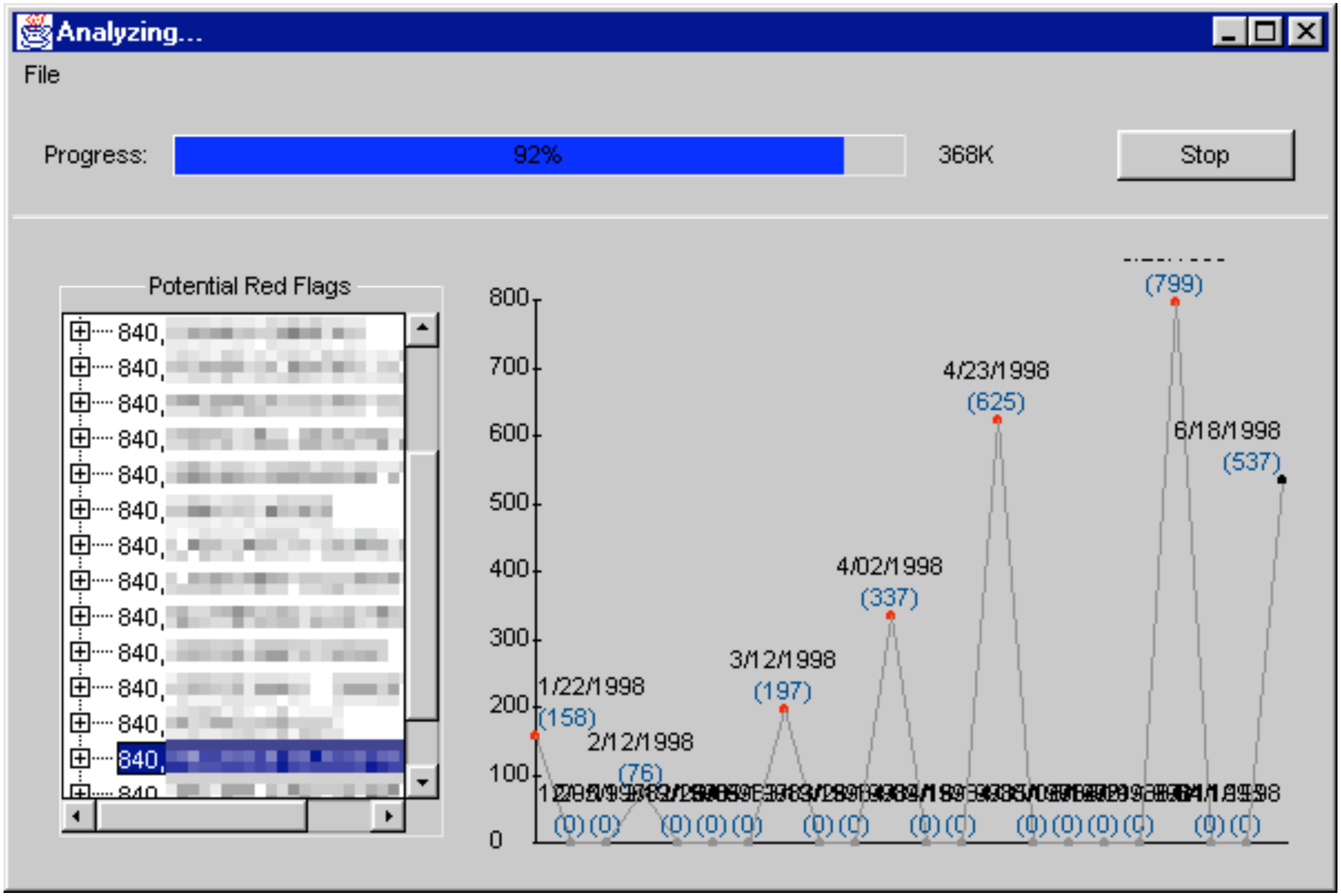
Regression

- A regression fits a straight line to a trend
  - y = a + bx
  - A positive slope (b) indicates an increasing trend
- Simple regressions are easy to calculate in Excel and other application
- Custom scripts can usually us statistical libraries to calculate regressions

QuickTime™ and a TIFF (Uncompressed) decompressor are needed to see this pictur

Picalo: Trend by regression slope

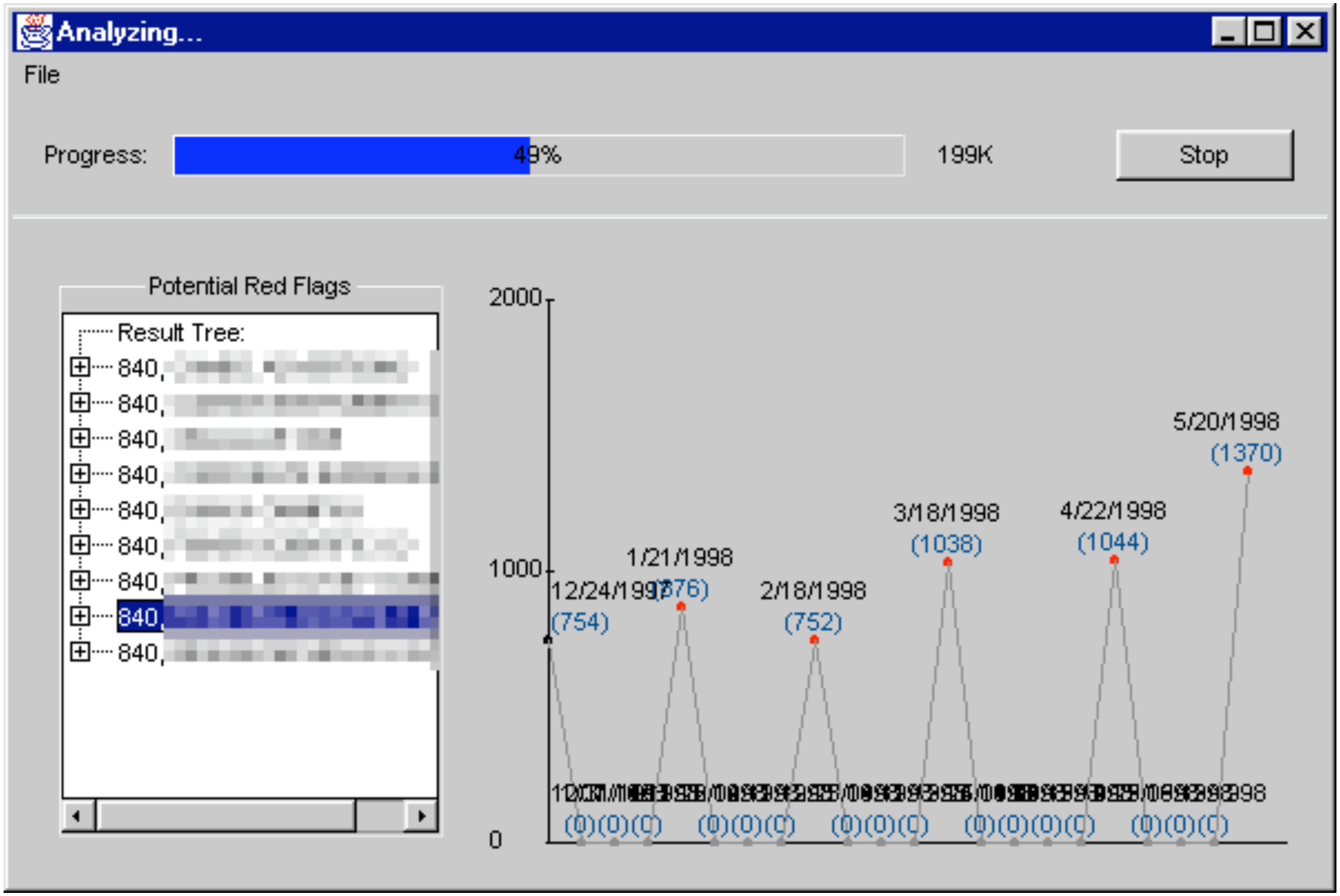
Systematic Increases In Spending



Progress: 92% 368K

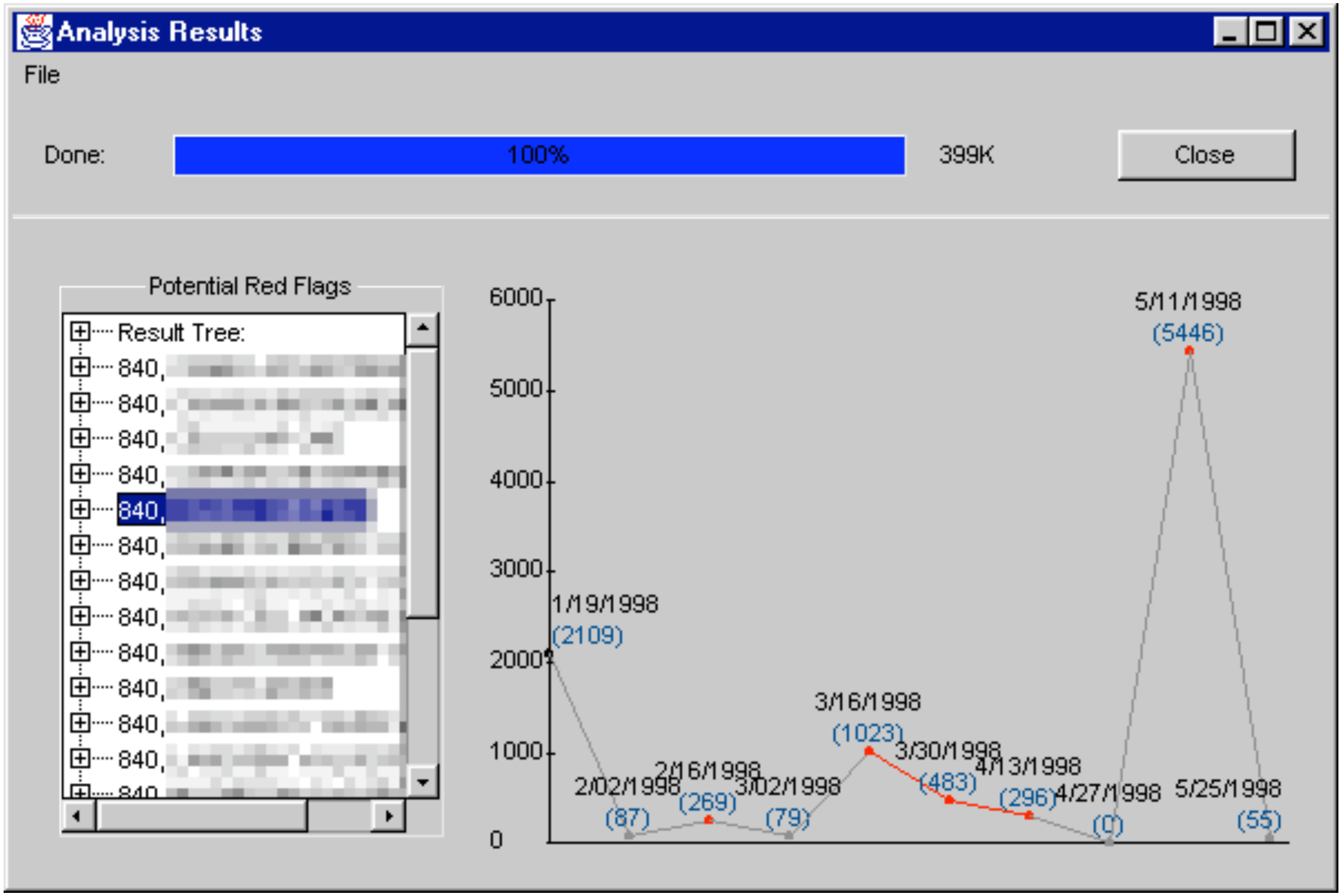
840, [redacted]
840, [redacted]
840, [redacted]
840, [redacted]
840, [redacted]
840, [redacted]
840, [redacted]
840, [redacted]
840, [redacted]
840, [redacted]
840, [redacted]
840, [redacted]
840, [redacted]
840, [redacted]

Systematic Increases In Spending



Result Tree:
- 840, [redacted]
- 840, [redacted]
- 840, [redacted]
- 840, [redacted]
- 840, [redacted]
- 840, [redacted]
- 840, [redacted]
- 840, [redacted]
- 840, [redacted]

Unexpected Peaks In Spending



Analysis Results

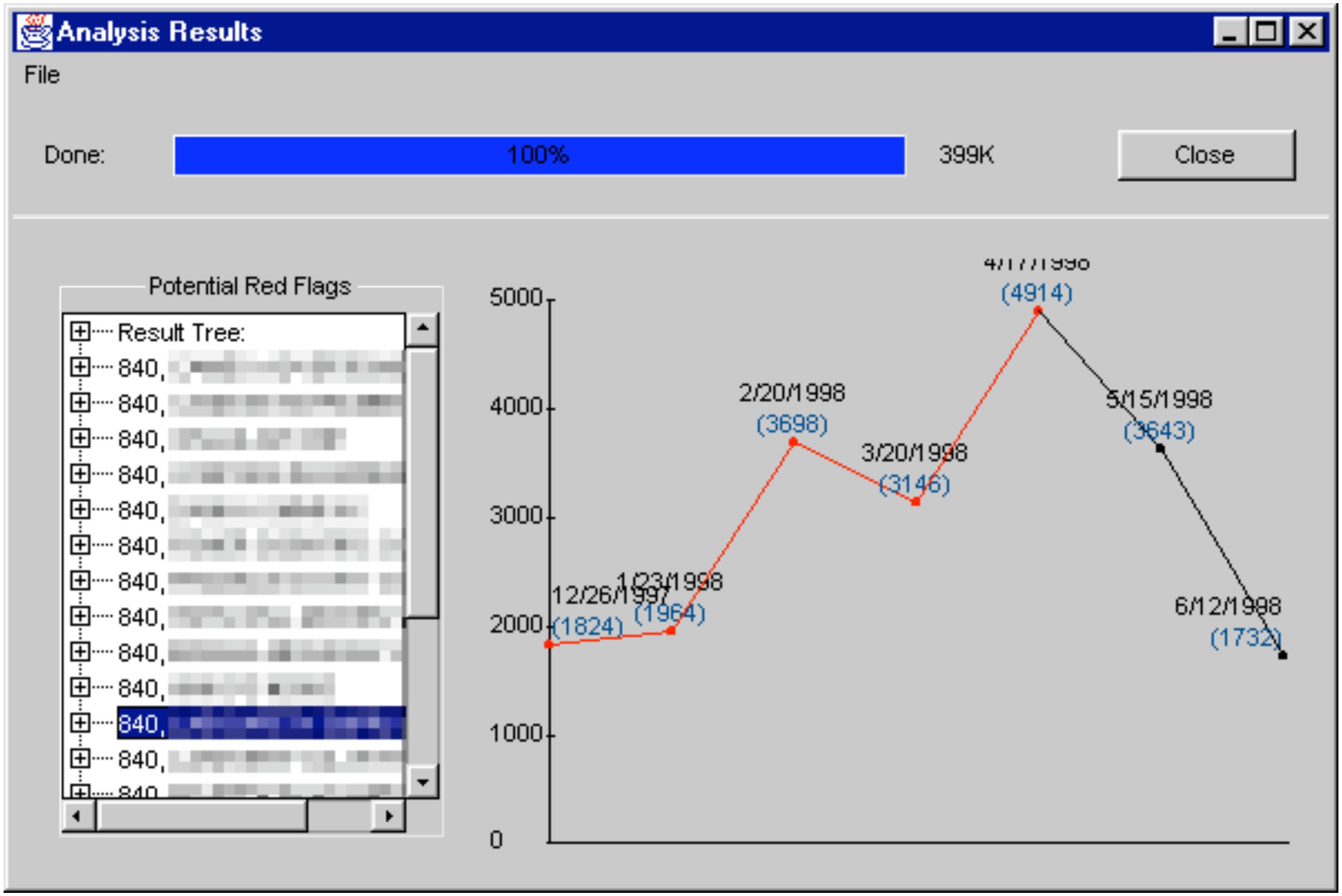
File

Done: 100% 399K Close

Potential Red Flags

- Result Tree:
- 840, [redacted]
- 840, [redacted]
- 840, [redacted]
- 840, [redacted]
- 840, [redacted]
- 840, [redacted]
- 840, [redacted]
- 840, [redacted]
- 840, [redacted]
- 840, [redacted]
- 840, [redacted]
- 840, [redacted]
- 840, [redacted]

Increases In Only Part Of A Trend



Analysis Results

File

Done: 100% 399K Close

Potential Red Flags

Result Tree:
840, [redacted]
840, [redacted]
840, [redacted]
840, [redacted]
840, [redacted]
840, [redacted]
840, [redacted]
840, [redacted]
840, [redacted]
840, [redacted]
840, [redacted]
840, [redacted]
840, [redacted]

Module: Searching Text

Searching Text

- LIKE queries

  SELECT * FROM Vendor WHERE name LIKE “%Dane%”

- Regular Expressions
  - One of the oldest and most powerful methods of searching for patterns in text

Search for “Dane” anywhere	.*Dane.*
Search for dates in format mm/dd/yyyy	\d{1,2}/\d{1,2}/\d{4}

regex.py

Simple Methods for Fuzzy Matching

- Number of common characters
- Order of characters
- Remove the vowels

Soundex Algorithm

- Identify matches based upon sounds
- Need to specify the number of significant sounds
- Vowels are ignored
- Soundex patterns are different for English, Spanish, etc.
  - Accepted patterns for most languages and cultures are on the Internet

Picalo: Simple.soundex
ACL: Soundslike

N-Grams

- A method of comparing N-number of letters in two texts:

The fat cat sat in the hat

“at” appears in 4 of the 27 two-letter grams

Employee.txt (tsv)
Vendor.txt (tsv)
Fuzzy join by city, address