# An Optimal Algorithm for the Distinct Elements Problem

## Problem and Results

Sequence of integers:

1 17 2 4 17 9 2 5 1 1 4 6

- One pass over a stream of integers each between 1 and \( n \)
- Query() – Output the number of distinct integers seen thus far
- Goals – Use little memory, and process each integer quickly

## Applications

<table>
<thead>
<tr>
<th>track spread of Code Red worm</th>
<th>network intrusion detection</th>
<th>database query optimization</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
</tr>
</tbody>
</table>

## Algorithm Ideas

### Balls-and-bins approach

Inspired by [Bar-Yossef et al. 2002]

- Subsample the stream at geometrically decreasing rates
- Perform balls and bins at each level

### Balls-and-bins diagram

\[
h: \{1, \ldots, n\} \rightarrow \{1, \ldots, \log n\}
\]

\[
Pr[h(i) = j] = \frac{1}{2^j}
\]

\[
g: \{1, \ldots, n\} \rightarrow \{1, \ldots, 1/\varepsilon^2\}
\]

- When \( i \) appears in stream, put a ball in cell \((g(i), h(i))\)
- For each column, store the largest row containing a ball
- Identify the largest row \( j \) which is at least half full, and count the number of columns with at least \( j \) written. Base estimate on this count.

## Memory and Update Time

<table>
<thead>
<tr>
<th>Author, Year</th>
<th>Memory</th>
<th>Update Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flajolet, Martin 1983</td>
<td>( O(\log n) )</td>
<td>–</td>
</tr>
<tr>
<td>Ailon, Matias, Szegedy 1996</td>
<td>( O(\log n) )</td>
<td>( O(\log n) )</td>
</tr>
<tr>
<td>Gibbons, Tirthapura 2001</td>
<td>( O((\log n)/\varepsilon^2) )</td>
<td>( O(1/\varepsilon^2) )</td>
</tr>
<tr>
<td>Bar-Yossef, Jayram, Kumar, Sivakumar, Trevisan 2002</td>
<td>( O((\log n)/\varepsilon^2) )</td>
<td>( O(\log(1/\varepsilon)) )</td>
</tr>
<tr>
<td>Bar-Yossef, Jayram, Kumar, Sivakumar, Trevisan 2002</td>
<td>( O((\log \log n + \log(1/\varepsilon))/\varepsilon^2 + \log n) )</td>
<td>( O(1/\varepsilon^2) )</td>
</tr>
<tr>
<td>Durand, Flajolet 2003</td>
<td>( O((\log \log n)/\varepsilon^2 + \log n) )</td>
<td>–</td>
</tr>
<tr>
<td>Kane, Nelson, Woodruff 2010</td>
<td>( O(1/\varepsilon^2 + \log n) )</td>
<td>( O(1) )</td>
</tr>
</tbody>
</table>

## References