

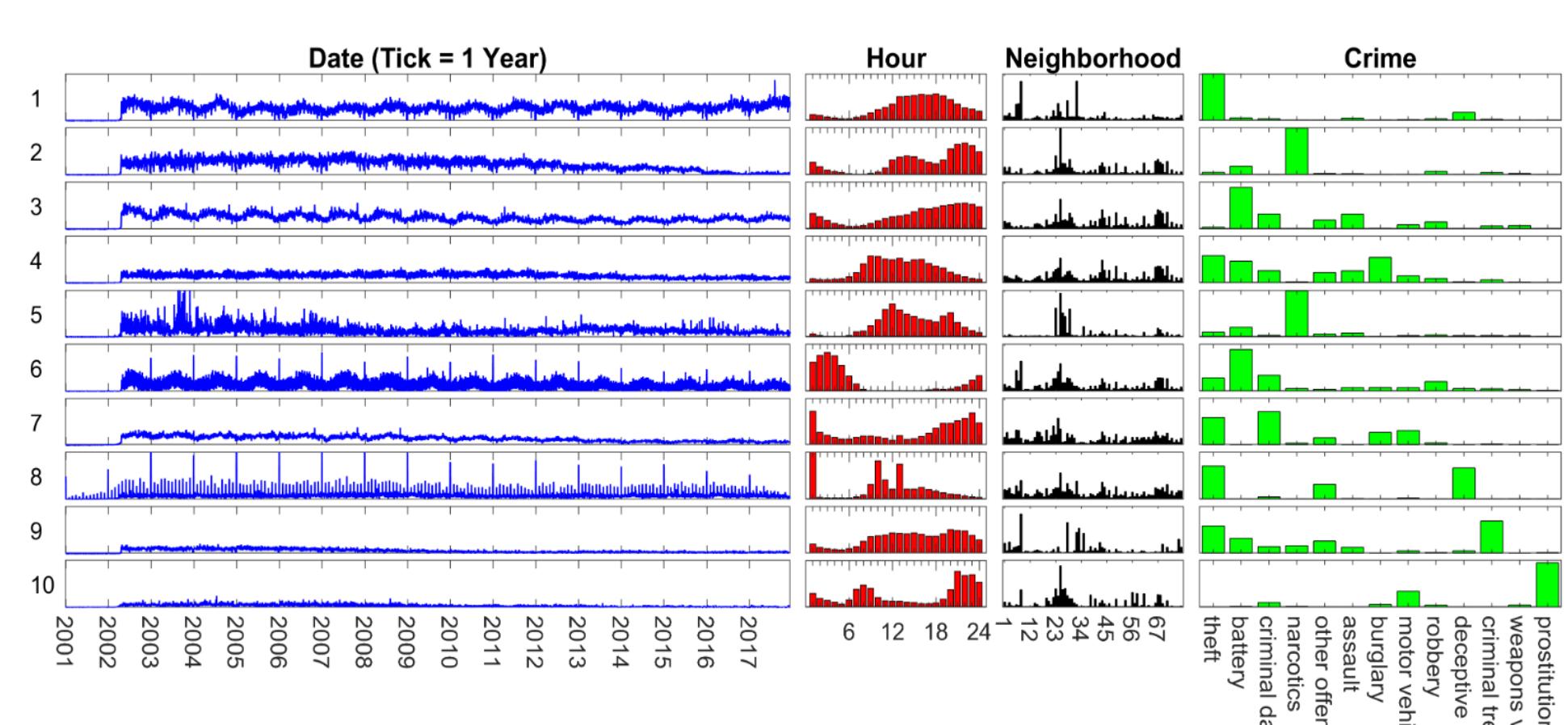
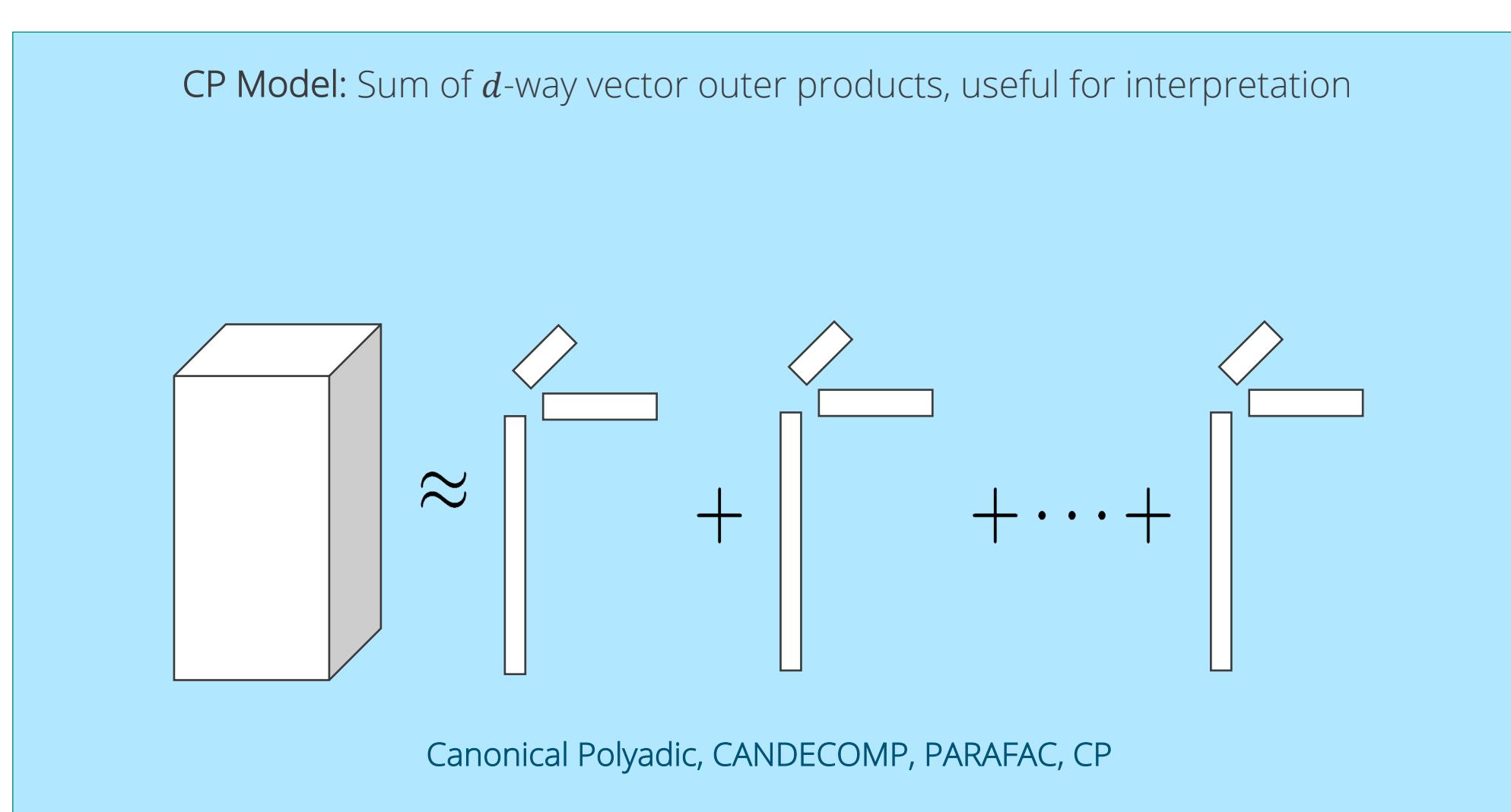


**Problem:** A well-known problem with count data is the corruption by false zeros not representing observed counts.

**Goal:** Develop a low-rank tensor approximation unaffected by false zeros for sparse data.

**Challenge:** Sparse count data contains many, many zero counts; which are which?

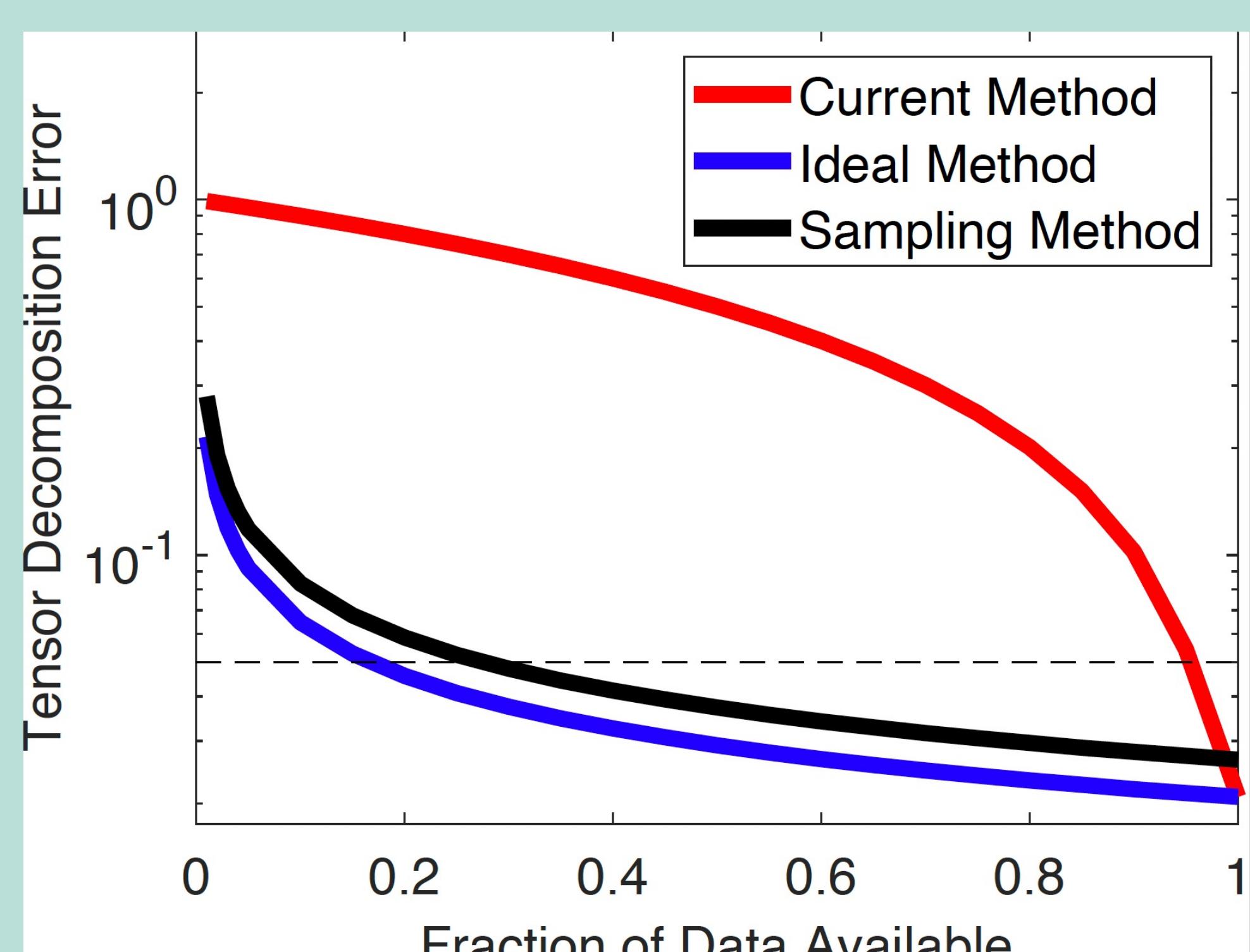
**Our solution:** Ignore all zeros and then quantify the effect of ignorance.



Crime reports in the city of Chicago, 2001-2017 (Kolda and Hong (2020), Stochastic Gradients for Large-Scale Tensor Decomposition)

- Why do we need a low-rank approximation?
  - Tensors contain  $I_1 I_2 \cdots I_d$  elements (product of the coordinate directions)
  - Canonical Polyadic (CP) rank  $R$  decomposition contains  $R (I_1 + I_2 + \cdots + I_d)$  elements
- Statistical model for the count data
  - Each tensor element is Poisson distributed
  - Other count models can be used (e.g. negative binomial)
- CP rank  $R$  decomposition and Poisson model lead to a max likelihood estimator
  - Resulting inference problem is corrupted by false zeros. Sigh.

- **Our solution:** ignore zeros and fit tensor decompositions using only a random sample of the sparse tensor elements
  - No need to access all the elements.
- **Model:** Zero-truncated Poisson model
  - We'll just ignore all the zeros since we don't know which are correct
- **Benefits:**
  - No *a-priori* knowledge of zeros needed
  - Better than assuming all zeros are true values (current approach)
- **Theorem summary**
  - Probabilistic bound that only a small constant multiple of error will be incurred given the sample size when ignoring zeros



For all the math, see

### Zero-Truncated Poisson Regression for Sparse Multiway Count Data Corrupted by False Zeros

OSCAR F. LÓPEZ<sup>✉</sup>

Harbor Branch Oceanographic Institute, Florida Atlantic University, Street, Postcode, State, Country

DANIEL M. DUNLAVY<sup>✉</sup>

Machine Intelligence and Visualization, Sandia National Laboratories, Street, Postcode, State, Country

AND

[doi.org/10.48550/arXiv.2201.10014](https://doi.org/10.48550/arXiv.2201.10014)

RICHARD B. LEHOUCQ<sup>✉</sup>

Discrete Math and Optimization, Sandia National Laboratories, Street, Postcode, State, Country