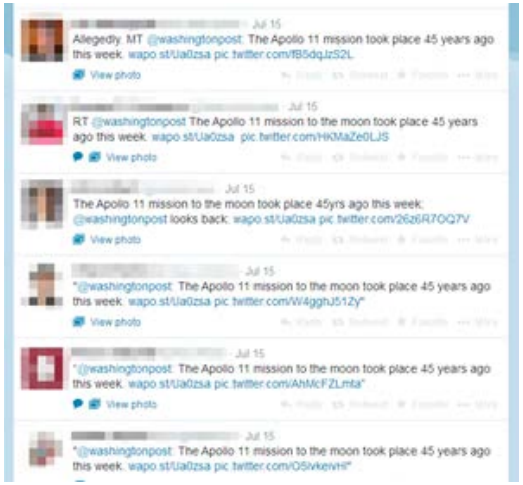
Expected Tensor Decomposition with Stochastic Gradient Descent

Takanori Maehara (Shizuoka Univ)
Kohei Hayashi (NII -> AIST)
Ken-ichi Kawarabayashi (NII)

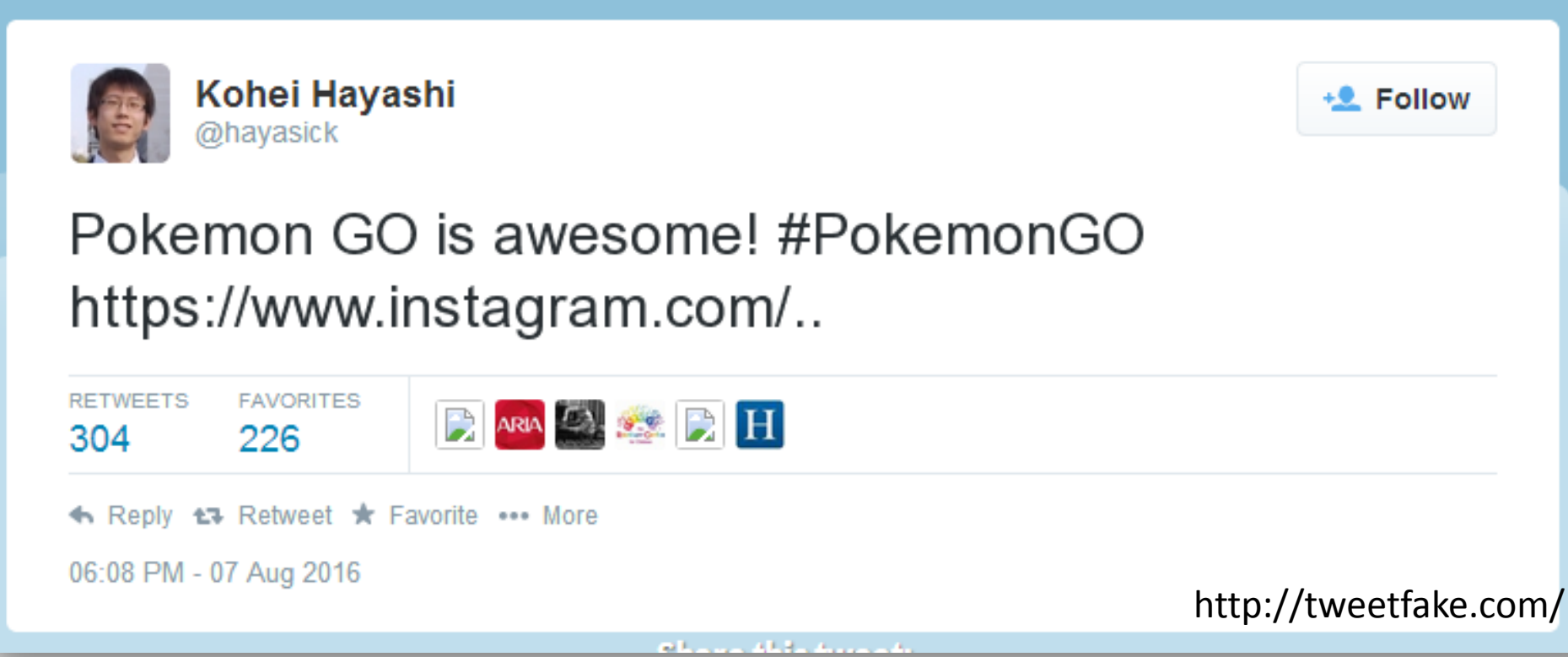
Presented at AAAI 2016

Information Overload

- Too many texts to catch up

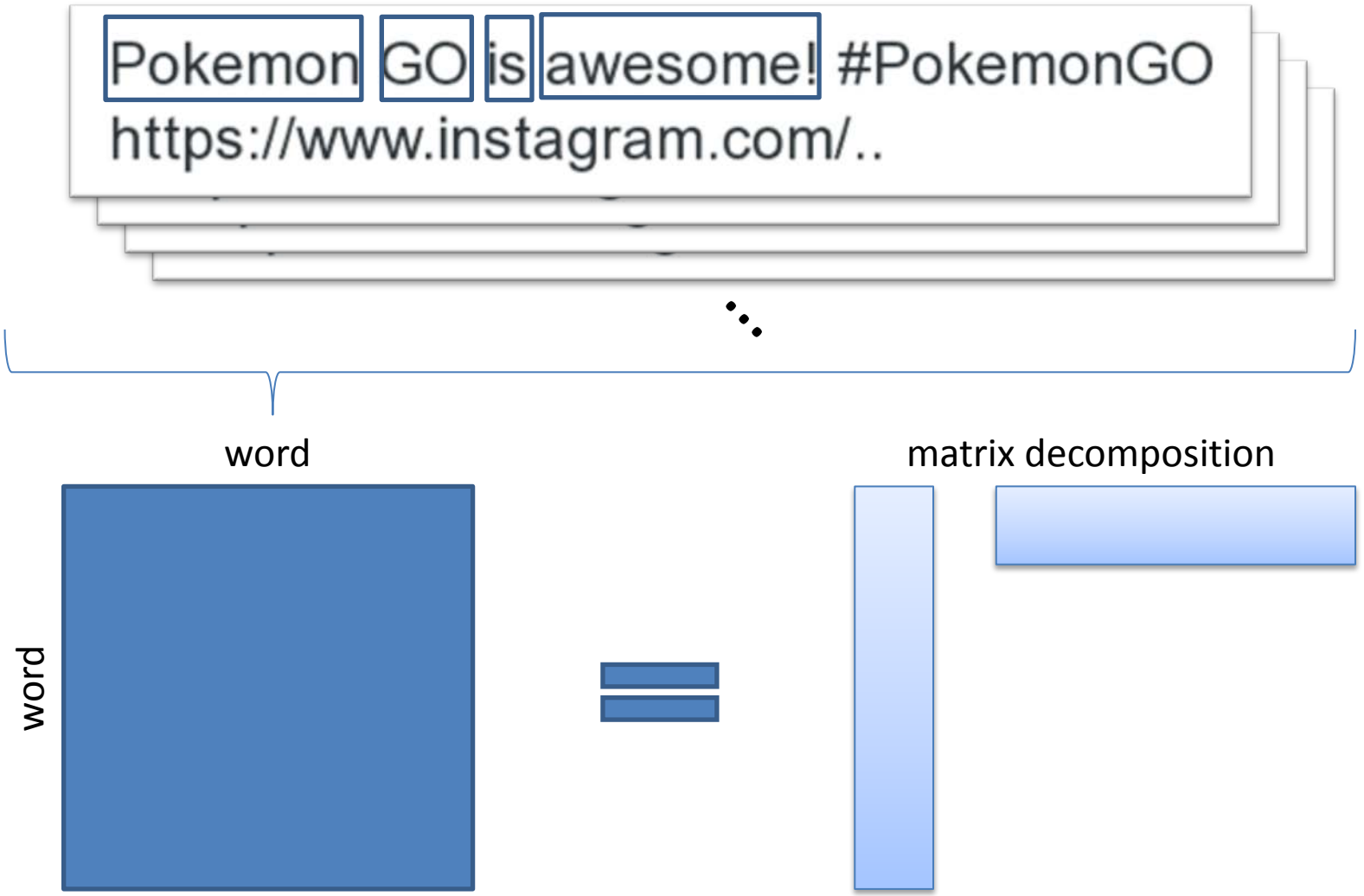


- How can we know what's going on?
  --- Machine learning!

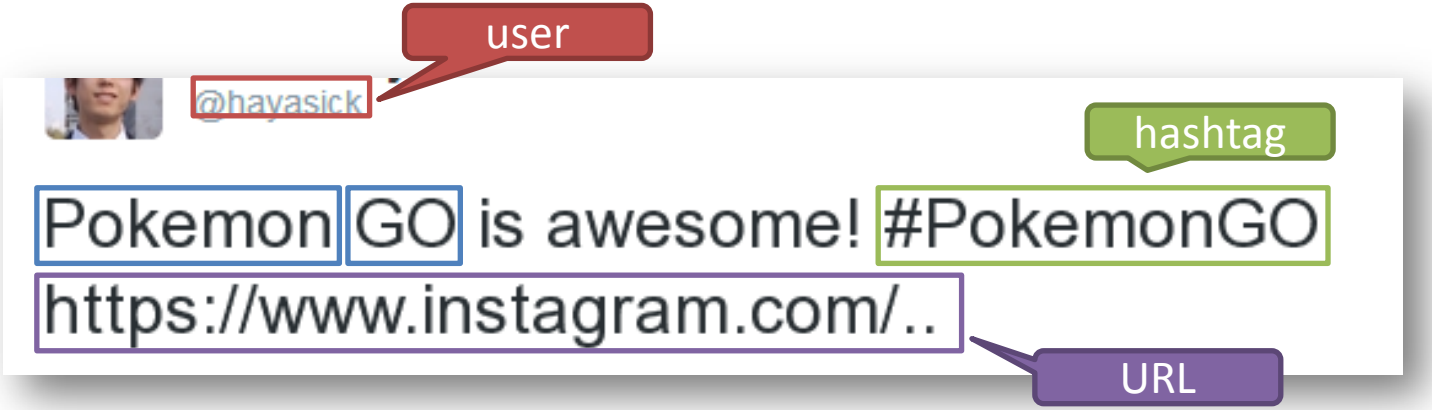


How can we analyze this?

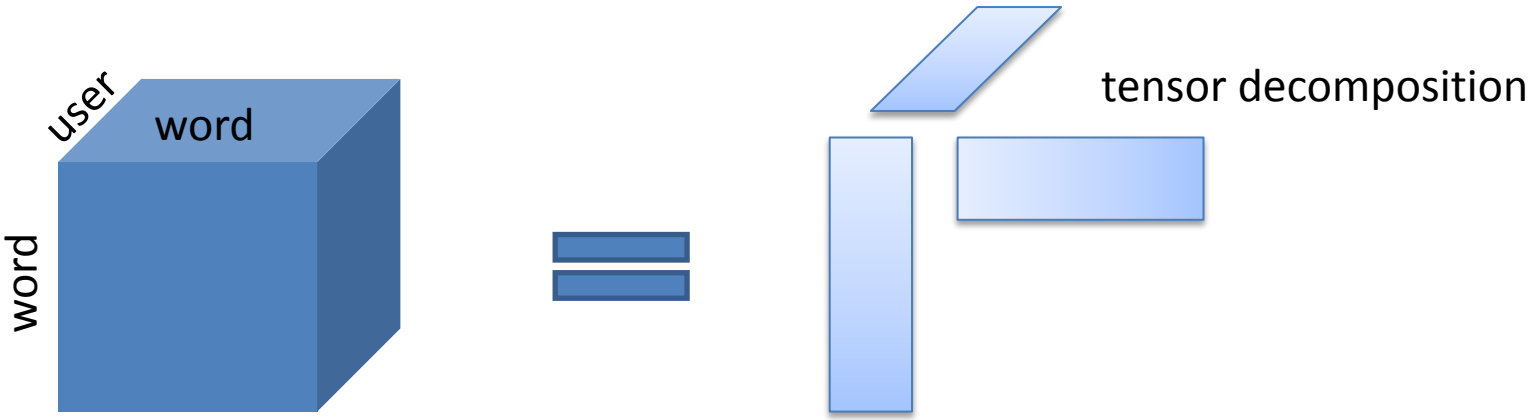
Word Co-occurance



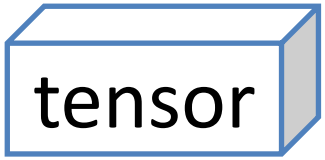
Higher-order Information



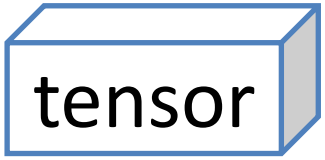
- Idea is generalized as tensor



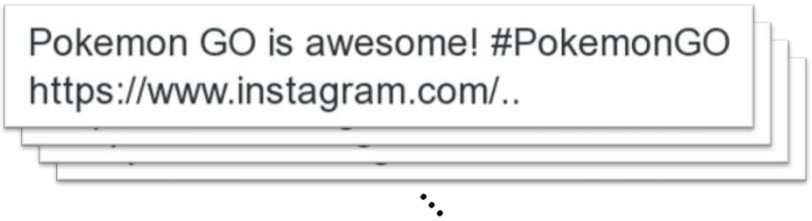
Scalability Issue

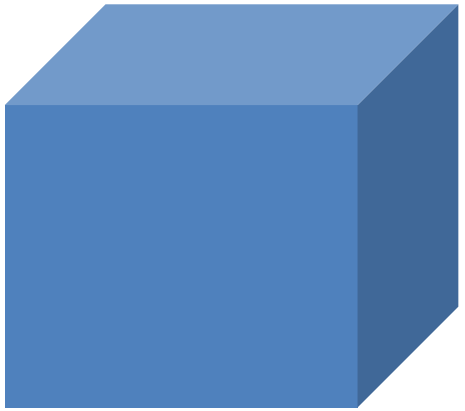
- To compute decomposition, we need 

beforehand.

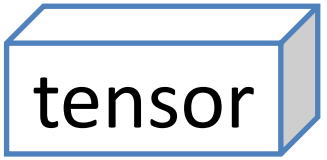
- Sometimes 

is too huge to put on memory.





Questions

- Can we get decomposition without 

?

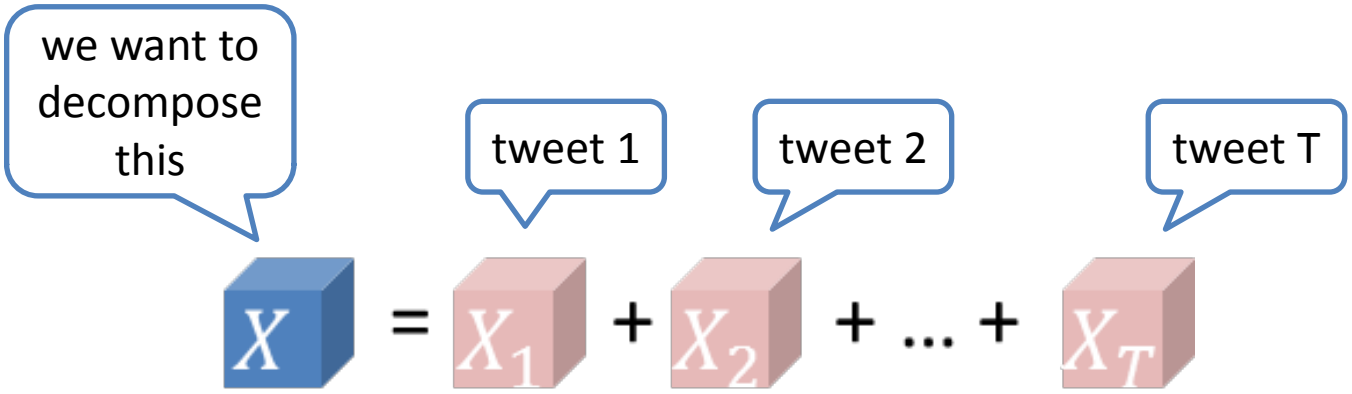
YES

- If possible, when and how?

- When: 

is given by sum
- How: stochastic optimization

Revisit: Word Co-occurance

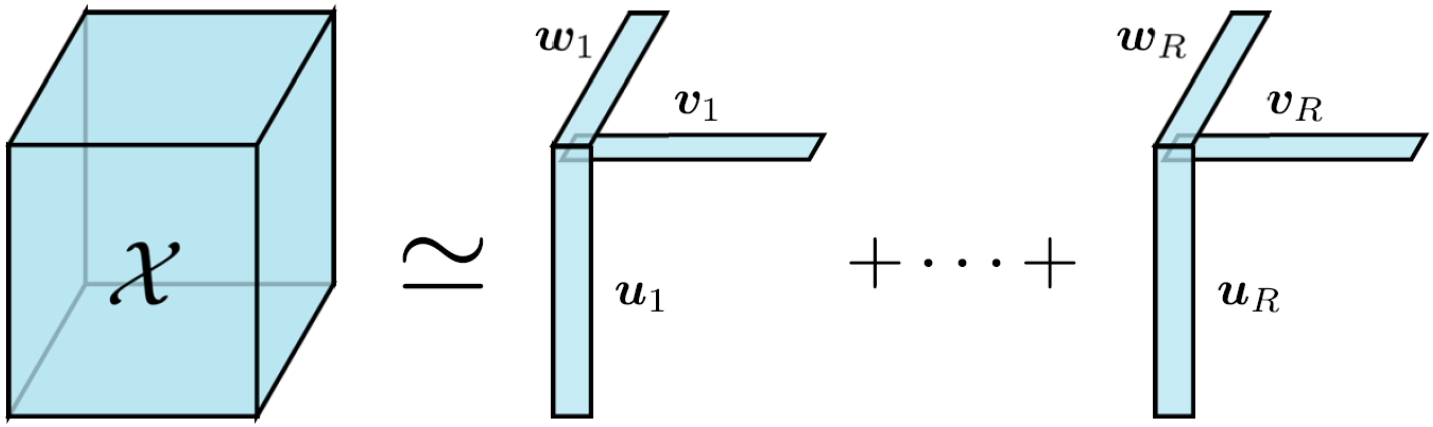


- X : “Heavy” tensor (e.g. dense)
- X_t : “Light” tensor (e.g. sparse)

CP Decomposition

$$\min_{U,V,W} L(\mathcal{X}; U, V, W)$$

- $L(\mathcal{X}; U, V, W) = \| \mathcal{X} - \sum_{r=1}^{R} u_r \circ v_r \circ w_r \|^2$



Batch Optimization

- CP decomposition is non-convex
- Use alternating least squares

ALS

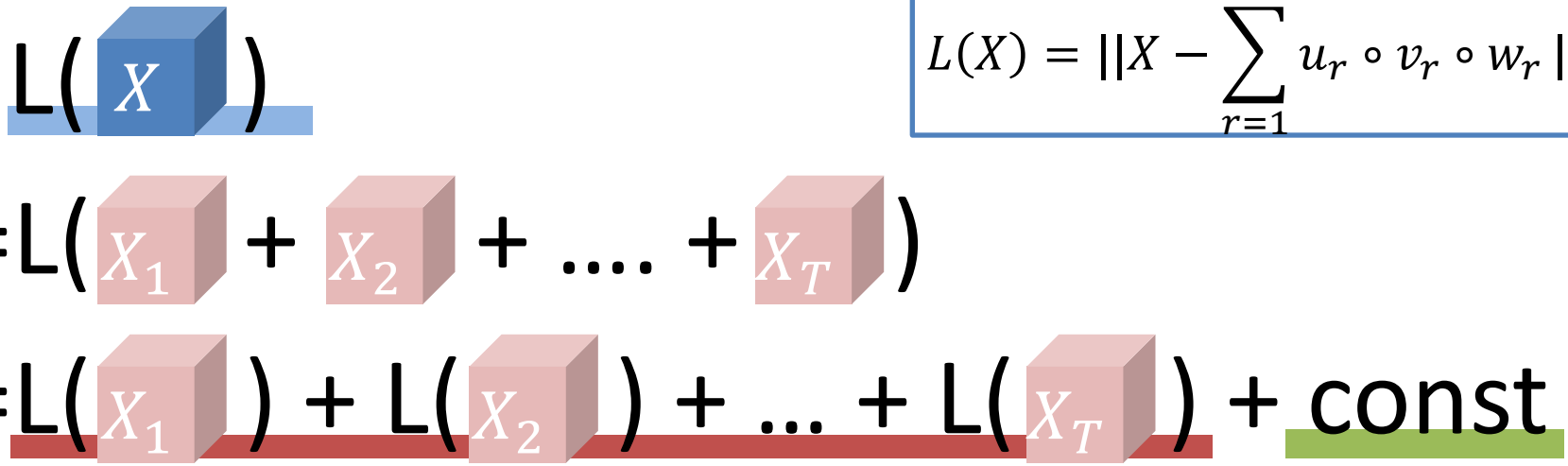
Repeat until convergence:

- $U \leftarrow U - \eta \nabla_U L($  $)$
- $V \leftarrow V - \eta \nabla_V L($  $)$
- $W \leftarrow W - \eta \nabla_W L($  $)$

step size

Key Fact: Loss is Decomposable

$$L(X) = ||X - \sum_{r=1}^{R} u_r \circ v_r \circ w_r ||^2$$



Intuition: for stochastic variable x and non-stochastic variable a,

$$(\mathrm{E}x - a)^2$$
$$= (\mathrm{E}x)^2 - 2a\mathrm{E}x + a^2$$
$$= \mathrm{E}[x^2] - 2a\mathrm{E}x + a^2 + (\mathrm{E}x)^2 - \mathrm{E}[x^2]$$
$$= \mathrm{E}[x - a]^2 + \mathrm{Var}[x]$$

Stochastic Optimization

- If $L = L_1 + L_2 + \cdots + L_T$,
  we can use ∇L_t instead of ∇L

⇨ Stochastic ALS

ALS

Repeat until convergence:

- $U \leftarrow U - \eta \nabla_U L(\mathcal{X})$
- $V \leftarrow V - \eta \nabla_V L(\mathcal{X})$
- $W \leftarrow W - \eta \nabla_W L(\mathcal{X})$

SALS

for $t = 1, \ldots, T$:

- $U \leftarrow U - \eta_t \nabla_U L(\mathcal{X}_t)$
- $V \leftarrow V - \eta_t \nabla_V L(\mathcal{X}_t)$
- $W \leftarrow W - \eta_t \nabla_W L(\mathcal{X}_t)$

Observations

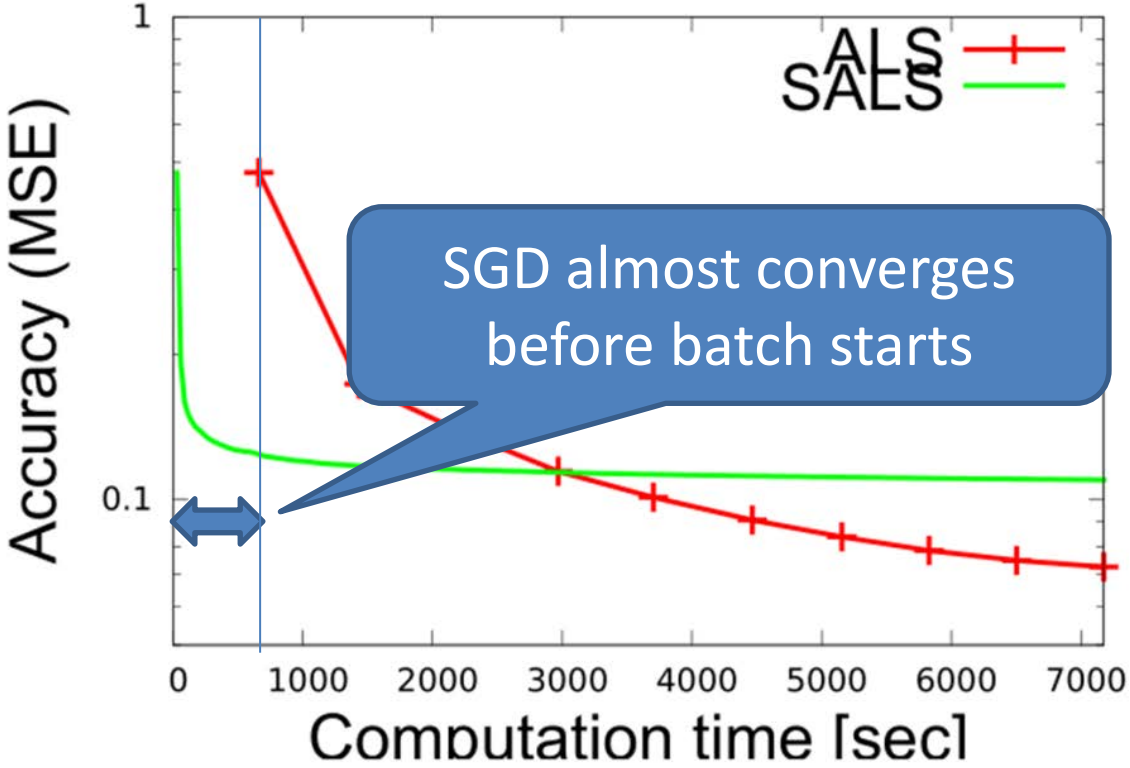
- ALS and SALS converge to the same solutions
  - $(L_{\mathrm{ALS}} - L_{\mathrm{SALS}}) = \mathrm{const}$
- SALS does not require to precompute 
  - Batch gradient: ∇L()
  - Stochastic gradient: ∇L()
- One step of SALS is lighter
  -  is sparse, whereas  is not
  - ∇L() is also sparse

Related Work

- Element-wise SGD

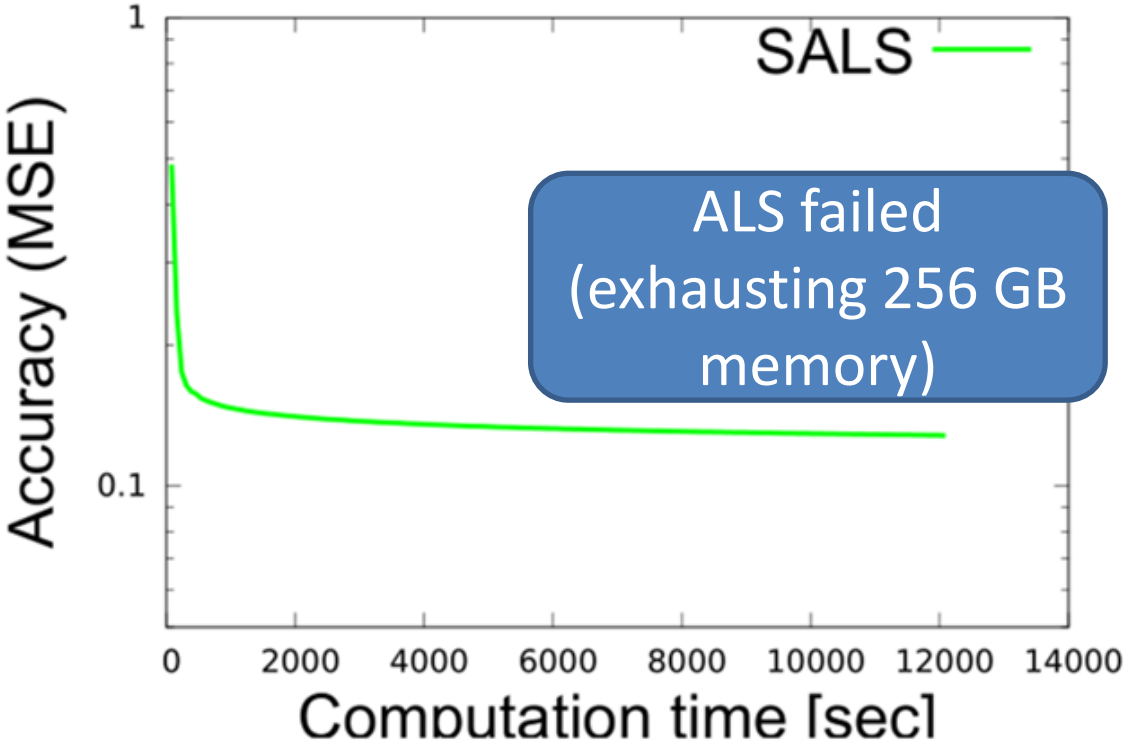
⇒ A special case of SALS

Experiment



- Amazon Review datasets (word triplets)
- 10M reviews, 280K vocabs, 500M non-zeros

Experiment (Cont'd)



- 34M reviews, 520K vocabs, 1G non-zeros

More on Paper

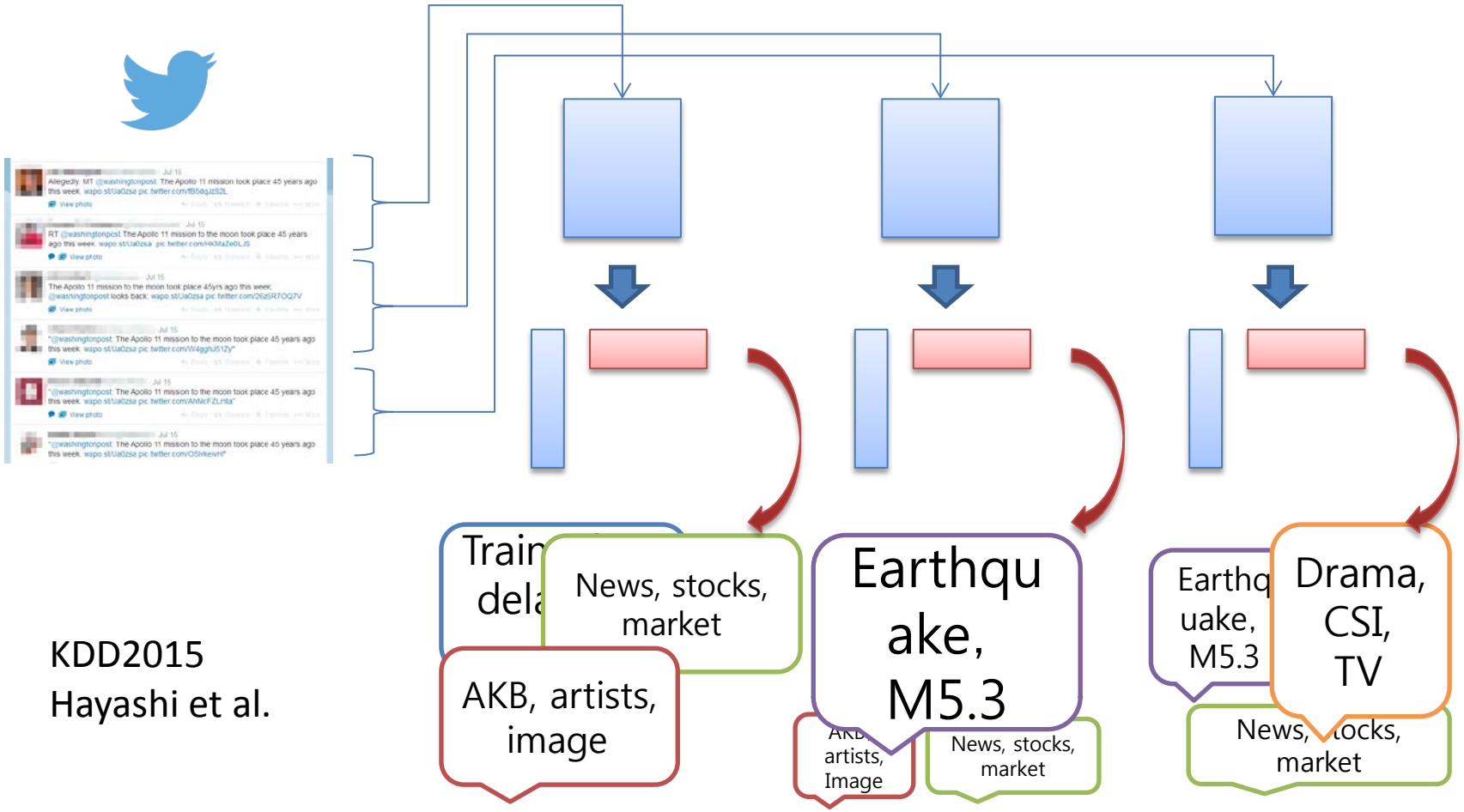
- 2nd-order SALS
  - Using block diagonal of Hessian
  - Faster convergence, low sensitivity of η

- With l2 norm regularizer:
  - min L(X; U, V, W) + R(U, V, W),
    $R(U,V,W) = \lambda(||U||^2 + ||V||^2 + ||W||^2)$

- Theoretical analysis of convergence

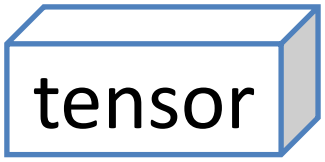
Further Applications

- Online Learning



KDD2015
Hayashi et al.

Take Home Message

If 

is given by sum …

- Don’t use batch gradient
- Use stochastic gradient!