Midterm Exam Review Notes

William Cohen

When and where the exam will be

- This room seats 217
- Enrollment as of this morning
  - 10-605A: 171
  - 10-605B: 63
  - 10-805: 10
  - total: 244
- So we have an **overflow room**: NSH 1507
  - It seats 63
  - Go there if you are **enrolled in 10-605B** and your last name **starts with D,E,…Z**
    - Margin is now 7 ☺
  - Otherwise come **to this room**

General hints in studying

- Understand what you've done and *why*
  - There will be questions that test your *understanding* of the techniques implemented
    - why will/won't this shortcut work?
    - what does the analysis say about the method?
- We're mostly about learning meets computation here
  - there won't be many "pure 601" questions
  - but 601 topics that are discussed in lecture are fair game for questions

General hints in studying

- Techniques covered in class but not assignments:
  - When/where/how to use them
  - That usually includes understanding the *analytic* results presented in class
  - Eg:
    - is the lazy regularization update an approximation or not? when does it help? when does it not help?

General hints in studying

- What about assignments you haven't done?
  - You should **read through the assignments** and be familiar with the algorithms being implemented
- There won't be questions about programming details that you could look up on line
  - but you should know how architectures like Hadoop work (eg, when and where they communicate, how they recover from failure, ...)
  - you should be able to sketch out simple map-reduce algorithms and answer questions about GuineaPig
  - you should be able to read programs in workflow operators and discuss how they'd be implemented and/or how to optimize them

General hints in studying

- There are not detailed questions on the guest speakers
- There might be high-level questions

General hints in exam taking

- You can bring in one 8 ½ by 11" sheet (front and back)
- Look over everything quickly and skip around
  - probably nobody will know everything on the test
- If you're not sure what we've got in mind: state your *assumptions* clearly in your answer.
  - This is ok even on true/false
- If you look at a question and don't know the answer:
  - we probably haven't told you the answer
  - but we've told you enough to work it out
  - imagine arguing for some answer and see if you like it

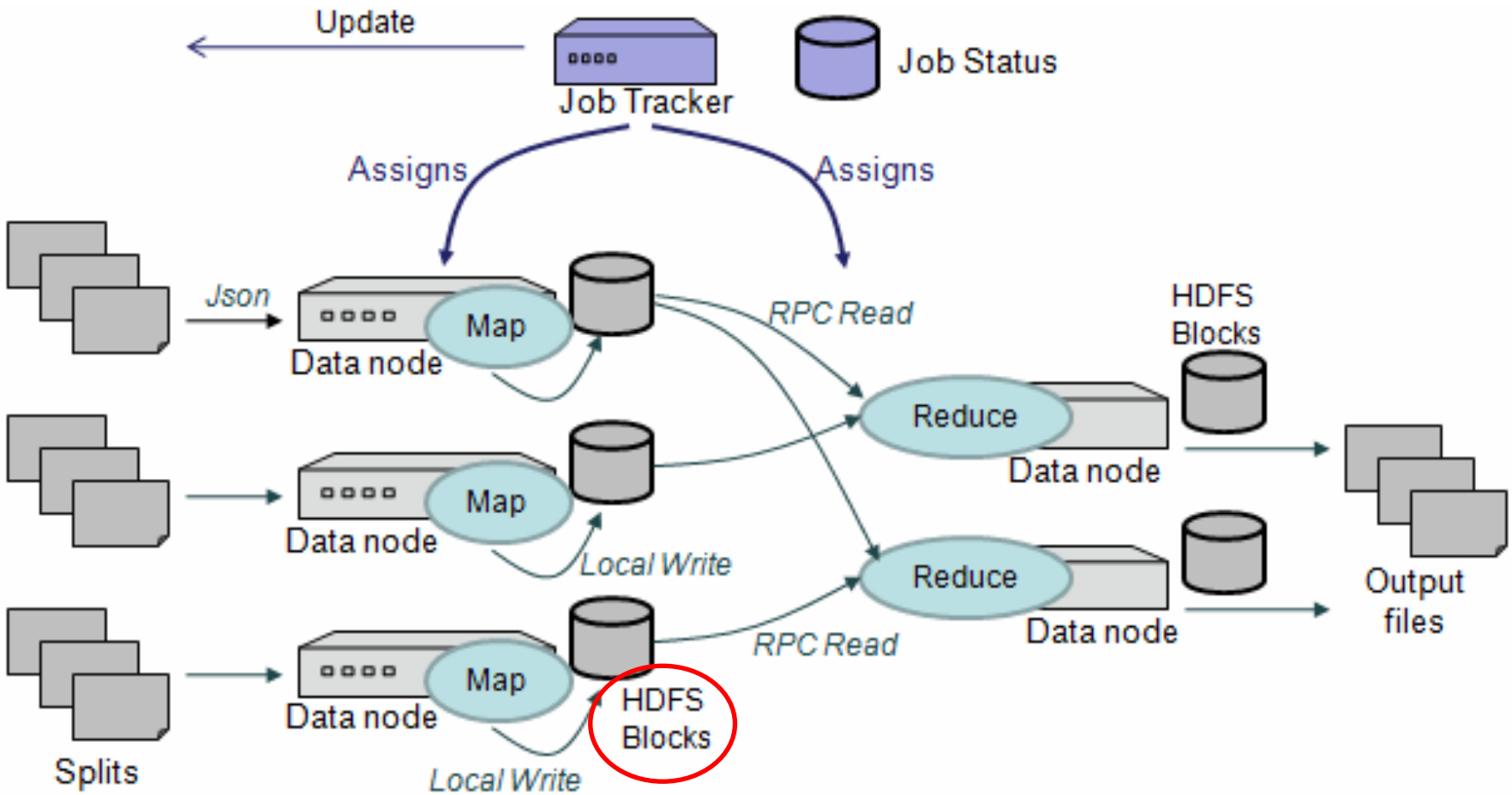
Outline – major topics before midterm

- Hadoop
  - stream-and-sort is how I ease you into that, not really a topic on its own
- ~~Phrase finding~~
- Similarity joins
- Parallelizing learners (perceptron, …)
- Hash kernels and streaming SGD
- Distributed SGD for Matrix Factorization

- Some of these are easier to ask questions about than others.

MORE REVIEW SLIDES

HADOOP



What data gets lost if the job tracker is rebooted?
If I have a 1Tb file and shard it 1000 ways will it take longer than sharding it 10 ways?
Where should a combiner run?

```
$ hadoop fs -ls rcv1/small/sharded
Found 10 items
-rw-r--r--   3 …  606405 2013-01-22 16:28  /user/wcohen/rcv1/small/sharded/part-00000
-rw-r--r--   3 … 1347611 2013-01-22 16:28  /user/wcohen/rcv1/small/sharded/part-00001
-rw-r--r--   3 …  939307 2013-01-22 16:28  /user/wcohen/rcv1/small/sharded/part-00002
-rw-r--r--   3 … 1284062 2013-01-22 16:28  /user/wcohen/rcv1/small/sharded/part-00003
-rw-r--r--   3 … 1009890 2013-01-22 16:28  /user/wcohen/rcv1/small/sharded/part-00004
-rw-r--r--   3 … 1206196 2013-01-22 16:28  /user/wcohen/rcv1/small/sharded/part-00005
-rw-r--r--   3 … 1384658 2013-01-22 16:28  /user/wcohen/rcv1/small/sharded/part-00006
-rw-r--r--   3 … 1299698 2013-01-22 16:28  /user/wcohen/rcv1/small/sharded/part-00007
-rw-r--r--   3 …   928752 2013-01-22 16:28  /user/wcohen/rcv1/small/sharded/part-00008
-rw-r--r--   3 …   806030 2013-01-22 16:28  /user/wcohen/rcv1/small/sharded/part-00009

$ hadoop fs -tail rcv1/small/sharded/part-00005
weak as the arrival of arbitraged cargoes from the West has put the local market under pressure…
M14,M143,MCAT    The Brent crude market on the Singapore International …
```

Where is this data? How many disks is it on?
If I set up a directory on /afs that looks the same will it work the same? what about a local disk?

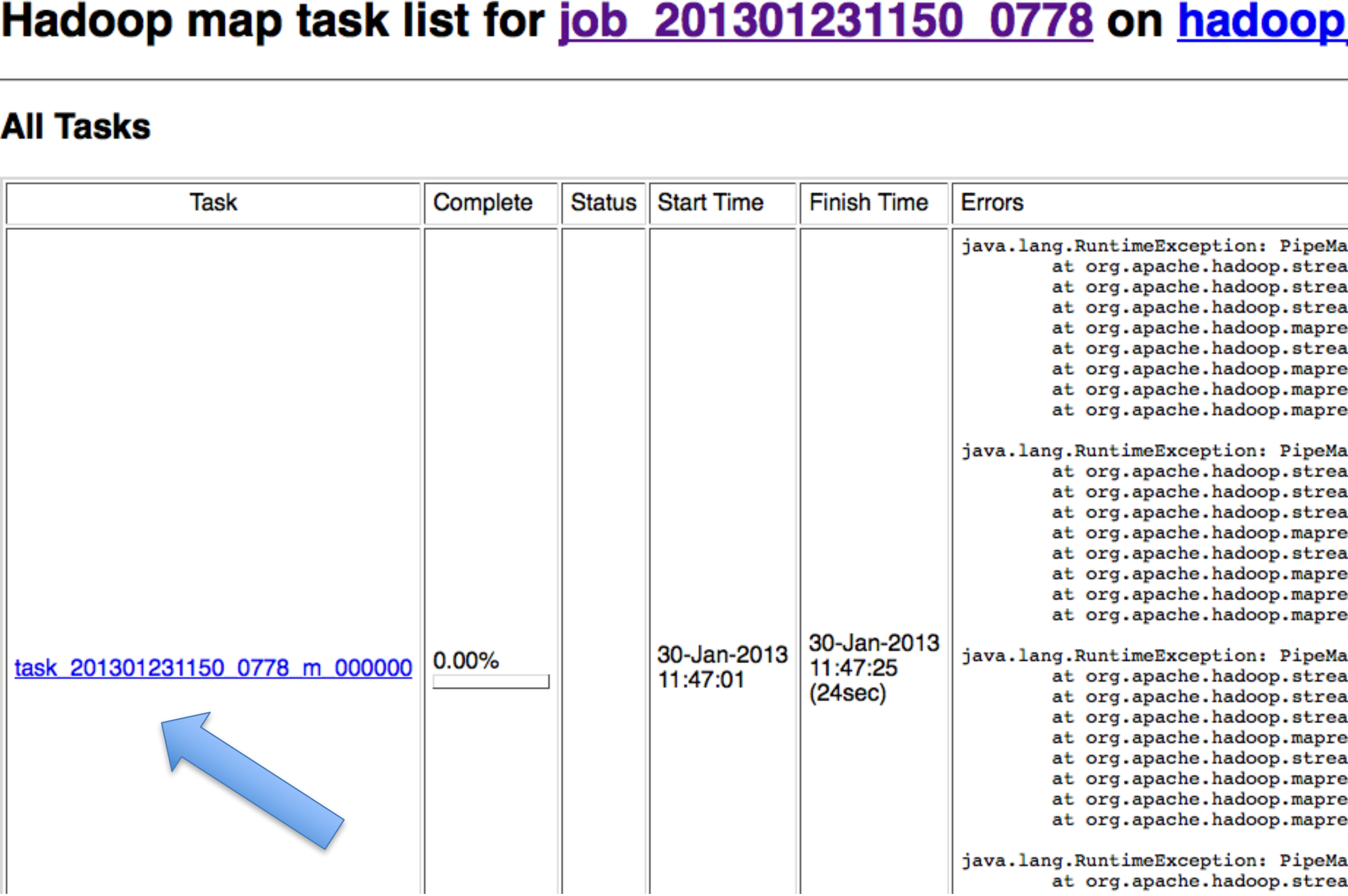
Hadoop job_201301231150_0778 on hadoopjt

User: wcohen
Job Name: streamjob6055532903853567038.jar
Job File: hdfs://hdfsname.opencloud/l/a2/scratch/hadoop-data/global/mapred/system/job_201301231150_0778/job.xml
Job Setup: Successful
Status: Failed
Started at: Wed Jan 30 11:46:47 EST 2013
Failed at: Wed Jan 30 11:47:28 EST 2013
Failed in: 41sec
Job Cleanup: Successful
Black-listed TaskTrackers: 2
Job Scheduling information: 5 running map tasks using 5 map slots, 0 running reduce tasks using 0 reduce slots.

Kind	% Complete	Num Tasks	Pending	Running	Complete	Killed	Failed/Killed Task Attempts
map	100.00%	10	0	0	0	10	35 / 5
reduce	00%	10	0	0	0	10	0 / 0

	Counter	Map	Reduce	Total
Job Counters	Rack-local map tasks	0	0	38
	Launched map tasks	0	0	40
	Data-local map tasks	0	0	2
	Failed map tasks	0	0	1

Map Completion Graph - close



Hadoop map task list for job_201301231150_0778 on hadoop

All Tasks

Task	Complete	Status	Start Time	Finish Time	Errors
task_201301231150_0778_m_000000	0.00%		30-Jan-2013 11:47:01	30-Jan-2013 11:47:25 (24sec)	

```
java.lang.RuntimeException: PipeMa
        at org.apache.hadoop.strea
        at org.apache.hadoop.strea
        at org.apache.hadoop.strea
        at org.apache.hadoop.mapre
        at org.apache.hadoop.strea
        at org.apache.hadoop.mapre
        at org.apache.hadoop.mapre
        at org.apache.hadoop.mapre

java.lang.RuntimeException: PipeMa
        at org.apache.hadoop.strea
        at org.apache.hadoop.strea
        at org.apache.hadoop.strea
        at org.apache.hadoop.mapre
        at org.apache.hadoop.strea
        at org.apache.hadoop.mapre
        at org.apache.hadoop.mapre
        at org.apache.hadoop.mapre

java.lang.RuntimeException: PipeMa
        at org.apache.hadoop.strea
        at org.apache.hadoop.strea
        at org.apache.hadoop.strea
        at org.apache.hadoop.mapre
        at org.apache.hadoop.strea
        at org.apache.hadoop.mapre
        at org.apache.hadoop.mapre
        at org.apache.hadoop.mapre

java.lang.RuntimeException: PipeMa
        at org.apache.hadoop.strea
```

Job job_201301231150_0778

All Task Attempts

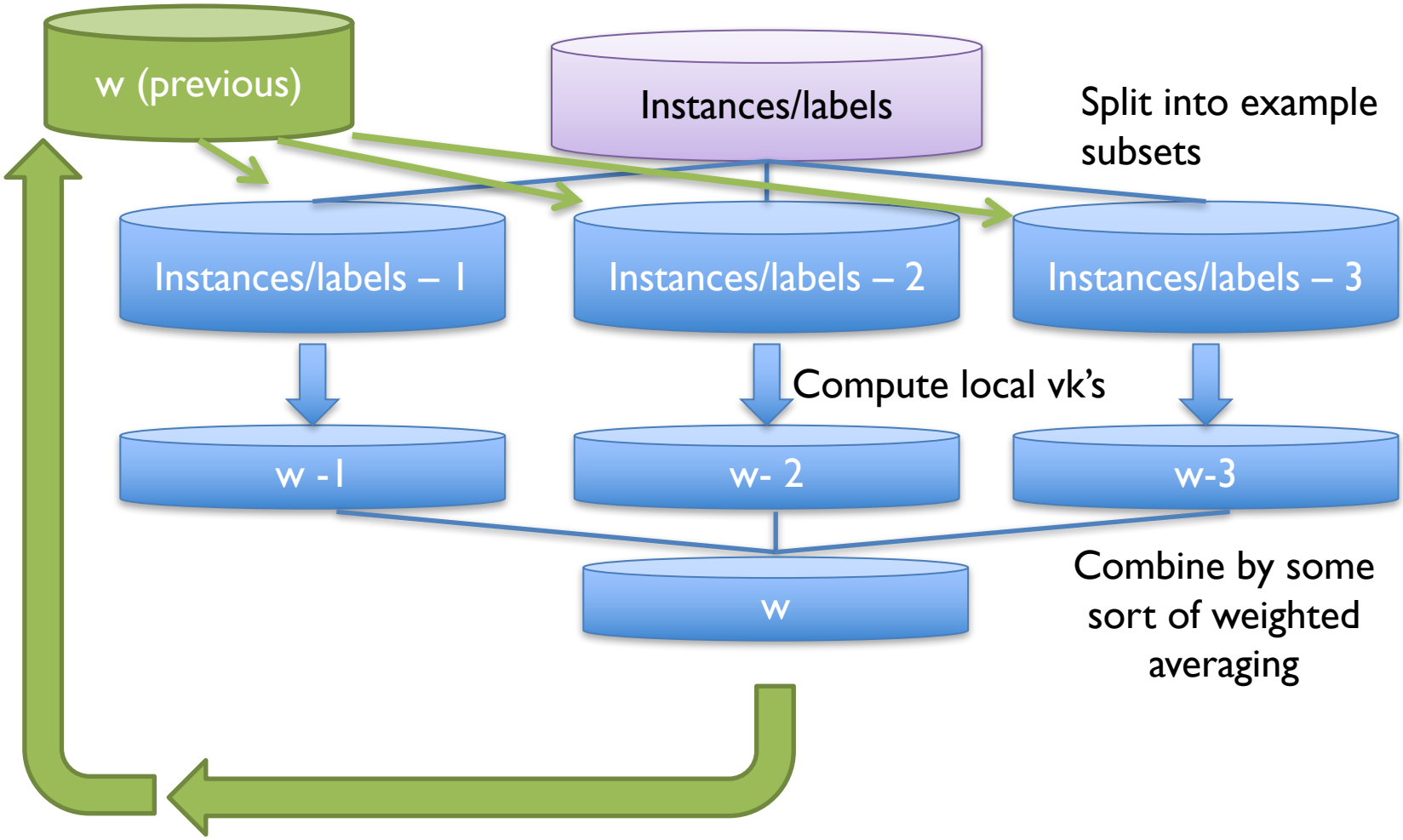
Task Attempts	Machine	Status	Progress	Start Time	Finish Time	Error
attempt_201301231150_0778_m_000000_0	/default-rack/cloud3u12.opencloud	FAILED	0.00%	30-Jan-2013 11:47:01	30-Jan-2013 11:47:06 (4sec)	java.
attempt_201301231150_0778_m_000000_1	/default-rack/cloud2u28.opencloud	FAILED	0.00%	30-Jan-2013 11:47:07	30-Jan-2013 11:47:11 (4sec)	java.

me	Errors	Task Logs	Counters	Actions
2013	java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed w de 1 at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.j 1) at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java: at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:132) at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:57) at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:36) at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:358) at org.apache.hadoop.mapred.MapTask.run(MapTask.java:307) at org.apache.hadoop.mapred.Child.main(Child.java:170)	Last 4KB Last 8KB All	1	
2013	java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 1 at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:311) at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:540) at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:132) at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:57) at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:36) at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:358) at org.apache.hadoop.mapred.MapTask.run(MapTask.java:307) at org.apache.hadoop.mapred.Child.main(Child.java:170)	Last 4KB Last 8KB All	1	
2013	java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 1 at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:311) at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:540) at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:132) at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:57) at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:36) at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:358) at org.apache.hadoop.mapred.MapTask.run(MapTask.java:307) at org.apache.hadoop.mapred.Child.main(Child.java:170)	Last 4KB Last 8KB All	1	

Why do I see this same error over and over again?

PARALLEL LEARNERS

Parallelizing perceptrons – take 2



A theorem

Theorem 3. *Assume a training set $\mathcal{T}$ is separable by margin γ. Let $k_{i,n}$ be the number of mistakes that occurred on shard i during the nth epoch of training. For any N, when training the perceptron with iterative parameter mixing (Figure 3),*

$$\sum_{n=1}^{N}\sum_{i=1}^{S}\mu_{i,n}k_{i,n} \leq \frac{R^2}{\gamma^2}$$

I probably won't ask about the proof, but I could definitely ask about the theorem.

Corollary: if we weight the vectors uniformly, then the number of mistakes is still bounded.

I.e., this is "enough communication" to guarantee convergence.

Distributed Training Strategies for the Structured Perceptron

Ryan McDonald **Keith Hall** **Gideon Mann**
Google, Inc., New York / Zurich
{ryanmcd|kbhall|gmann}@google.com

NAACL 2010



What does the word "structured" mean here? why is it important? would the results be better or worse with a regular perceptron?

STREAMING SGD

Learning as optimizatio regularized logistic regression

what if you had a different update for the regularizer?

- Algorithm:

$$w^j = w^j + \lambda(y - p)x^j - \lambda 2\mu w^j$$

Time goes from O(nT) to O(mVT) where

- n = number of non-zero entries,
- m = number of examples
- V = number of features
- T = number of passes over data

1. Initialize a hashtable W
2. For $t = 1, \ldots, T$
   - For each example $\mathbf{x}_i, y_i$:
     - Compute the prediction for $\mathbf{x}_i$:

$$p_i = \frac{1}{1 + \exp(-\sum_{j:x_i^j > 0} x_i^j w^j)}$$

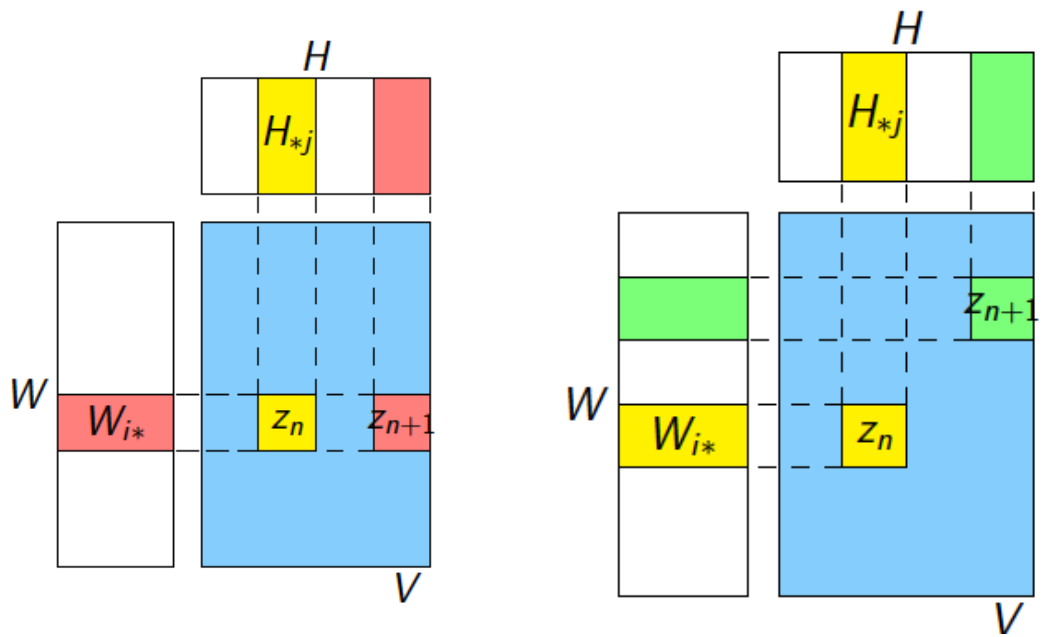
     - For each ~~non-zero~~ feature of $\mathbf{x_i}$ with index j and value x^j:
       * If j is not in W, set $W[j] = 0$.
       * Set $W[j] = W[j] + \lambda(y - p)x^j - \lambda 2\mu w^j$
3. Output the hash table W.

Formalization of the "Hash Trick":

First: Review of Kernels

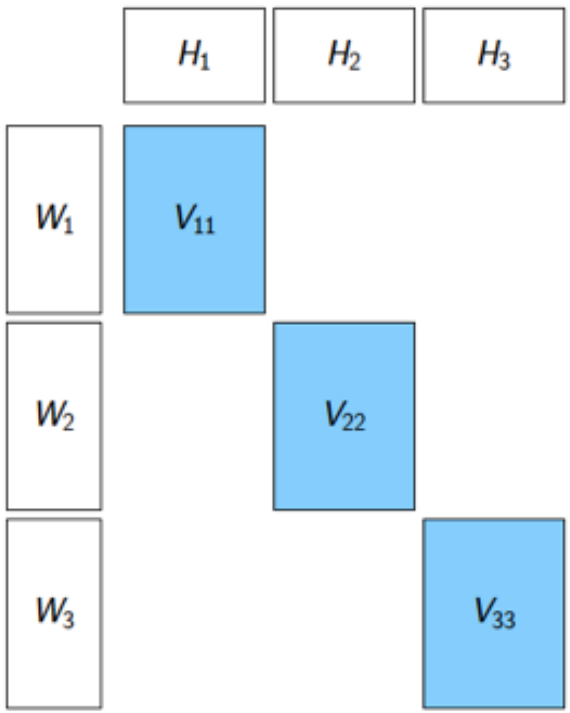
What is it? how does it work? what aspects of performance does it help? What did we say about it formally?

Matrix factorization



What is this picture trying to show?

Matrix factorization



What is this picture trying to show?
What would be another possible "stratum"?