

# Thinking in MapReduce

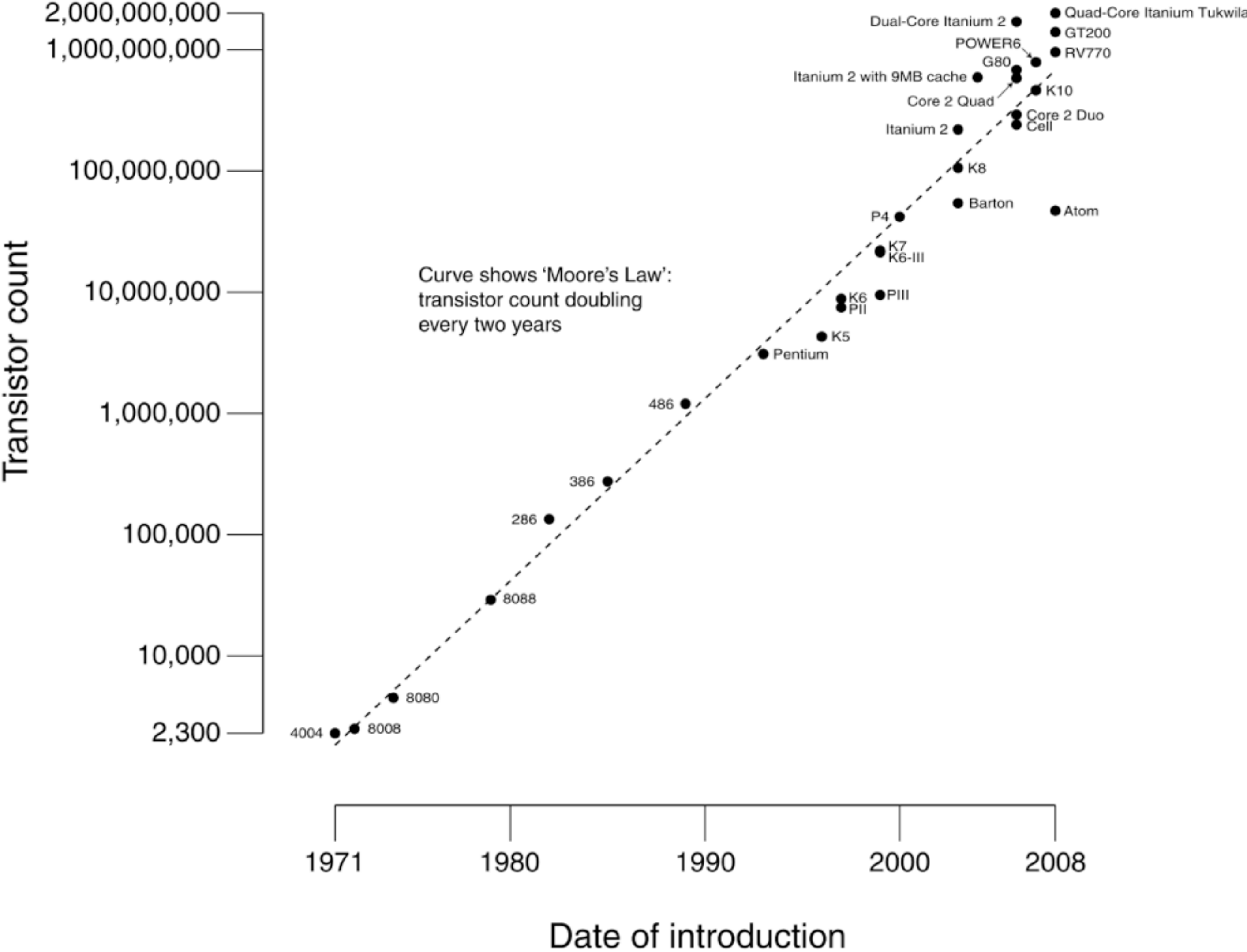
Ryan Brush



We programmers have had  
it pretty good

Hardware has scaled up faster  
than our problem sets

# CPU Transistor Counts 1971-2008 & Moore's Law





**Software  
Engineers**

**Moore's  
Law**

But the party is ending  
(or at least changing)



Data is growing faster than  
we can scale individual machines

So we have to spread our work across many  
machines



# This is a big deal in health care

Fragmented Information

Spread across many systems

No one has the complete picture

We need to put the picture back  
together again

Better-informed decisions

Reduce systematic friction

Understand and improve the  
health of populations

# Chart Search

Chart Search

Print 0 minutes ago

100%

Everything

Documents

Results

Newest documents

Newest results

Any time

Past 24 hours

Past week

Past month

Past year

Older than 1 year

Specific date range

Most relevant first

Newest first

Oldest first

Filter this search:

diabetes

Search

Cerner's Chart Search provides a delayed index to a subset of the patient's medical record. [Learn more.](#) Matches 1 - 10 of 16

Search again using: "diabetes"

21.7 days ago	Glucose, Random	90 mg/dL	Normal	65 - 110	
4.9 months ago	Estimated Average Glucose	100 mg/dL	Normal	65 - 109	
4.9 months ago	Hgb A1c	5.1 %	Normal	4.0 - 6.0	Interpretation ...
7.6 months ago	Glucose, Random	80 mg/dL	Normal	65 - 110	
1.1 years ago	Glucose, Random	140 mg/dL	High	65 - 110	
1.1 years ago	Glucose, Random	140 mg/dL	High	65 - 110	
2.1 years ago	Hgb A1c	6.5 %	High	4.0 - 6.0	Interpretation ...
2.1 years ago	Estimated Average Glucose	140 mg/dL	High	65 - 109	
2.1 years ago	Glucose Level	170 mg/dL	High	65 - 110	
2.2 years ago	Glucose, Random	190 mg/dL	High	65 - 110	

[...16 more](#)

Office/Clinic Note-Physician: "Ophthalmic Examination"

Madison MD, Edward

Reason for Visit: This 49 year old woman presents today for a complete ophthalmic examination for Diabetes. ... Visual Fields: Confrontation visual field exam reveals full to finger confrontation OU. ... Macula with normal reflex and color bilaterally. Impression: Examination of eye and vision normal.

2.2 years ago Mar 1, 2010 11:00 AM CST East Clinic

Office/Clinic Note-Physician: "Diabetes, Type 2"

Feldman MD, Mark

Report Summary ☐ Endocrine: Excessive thirst. Chief Complaint 2/17/2010 3:00 PM Diabetes management ... Interval History Diabetes, Type 2 Glucose results elevated. ... Hemoglobin A1c results > 6% and within target range. ... mg/dL Normal Impression and Plan Diagnosis Type 2 diabetes (ICD9 250.00).

2.2 years ago Feb 17, 2010 4:00 PM CST BW Cardiology Clinic - Baseline West Medical Center

Office/Clinic Note-Physician: "Diabetes, Type 2"

Feldman MD, Mark

Interval History Diabetes, Type 2 Glucose results within target range. ... Glucose, Fasting (Ordered): Blood, Collected, Lab Collect, ST collect, type 2 diabetes, for 1 day(s) Hemoglobin A1c (Ordered): Blood, Collected, Lab Collect, RT collect, type 2 diabetes, for 1 day(s)

3.1 years ago Apr 5, 2009 1:00 PM CDT BWMP Family Practice Clinic - Baseline West Medical Pavillion

Office/Clinic Note-Physician: "Chest Pain, COPD, Diabetes, Type 2, Hypertension"

Feldman MD, Mark

The course is progressing as expected. Diabetes, Type 2 Glucose results elevated. ... Glucose Level 170 mg/dL HI ... HgbA1C (Ordered): Blood, Collected, Lab Collect, Routine collect, Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled

2.1 years ago Mar 15, 2010 11:00 AM CDT BW Cardiology Clinic - Baseline West Medical Center

Assessment Forms-Text: "Adult Ambulatory Care Intake"

Feldman MD, Mark

Date: 4/7/2011 ; Life Cycle Status: Active ; Responsible Provider: Feldman MD, Mark; Vocabulary: ICD-9-CM Diabetes mellitus type 2 Name of Problem: Diabetes mellitus type 2 ; Onset Date: 11/30/2003 ; Recorder: Snyder RN, Kara; Confirmation: Confirmed ; Classification: Medical ;

# Chart Search

- Information extraction
- Semantic markup of documents
- Related concepts in search results

The screenshot displays the Cerner Chart Search application. The search term 'diabetes' is entered in the top search bar, which indicates 10 matches (1 of 16 shown). The left sidebar contains navigation options: 'Everything', 'Documents', 'Results', 'Newest documents', and 'Newest results'. Under 'Results', there are filters for 'Any time' (Past 24 hours, Past week, Past month, Past year, Older than 1 year, Specific date range) and 'Most relevant first' (Newest first, Oldest first). The main content area shows a table of search results for 'diabetes'.

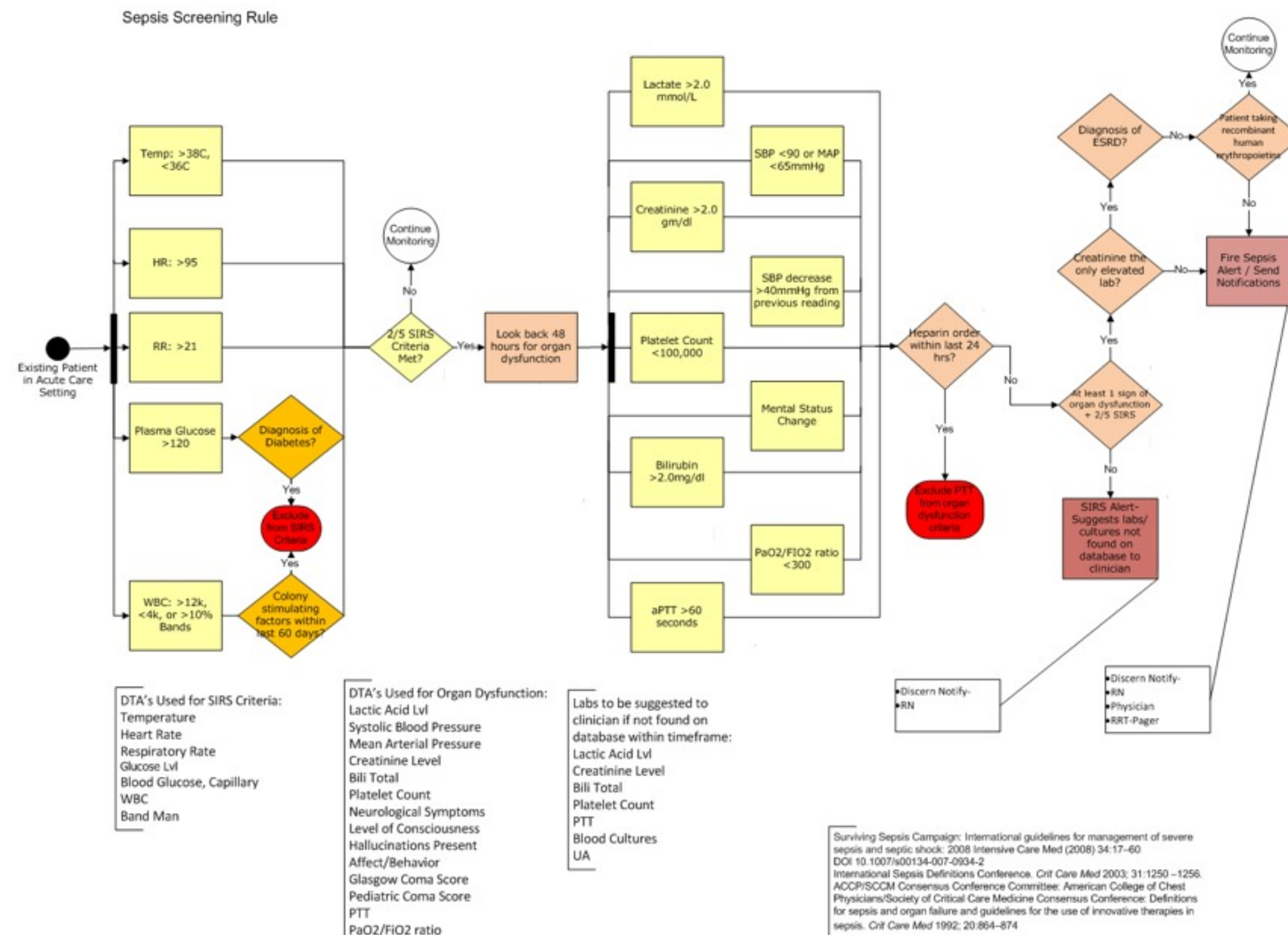
Time	Test	Value	Normal	Range	Interpretation
21.7 days ago	Glucose, Random	90 mg/dL	Normal	65 - 110	
4.9 months ago	Estimated Average Glucose	100 mg/dL	Normal	65 - 109	
4.9 months ago	Hgb A1c	5.1 %	Normal	4.0 - 6.0	Interpretation ...
7.6 months ago	Glucose, Random	80 mg/dL	Normal	65 - 110	
1.1 years ago	Glucose, Random	140 mg/dL	High	65 - 110	
1.1 years ago	Glucose, Random	140 mg/dL	High	65 - 110	
2.1 years ago	Hgb A1c	6.5 %	High	4.0 - 6.0	Interpretation ...
2.1 years ago	Estimated Average Glucose	140 mg/dL	High	65 - 109	
2.1 years ago	Glucose Level	170 mg/dL	High	65 - 110	
2.2 years ago	Glucose, Random	190 mg/dL	High	65 - 110	

Below the table, there are several clinical notes and reports:

- Office/Clinic Note-Physician: "Ophthalmic Examination"** - Madison MD, Edward. Reason for Visit: This 49 year old woman presents today for a complete ophthalmic examination for Diabetes. ... Visual Fields: Confrontation visual field exam reveals full to finger confrontation OU. ... Macula with normal reflex and color bilaterally. Impression: Examination of eye and vision normal.
- Office/Clinic Note-Physician: "Diabetes, Type 2"** - Feldman MD, Mark. Report Summary: Endocrine: Excessive thirst. Chief Complaint 2/17/2010 3:00 PM Diabetes management ... Interval History Diabetes, Type 2 Glucose results elevated. ... Hemoglobin A1c results > 6% and within target range. ... mg/dL Normal Impression and Plan Diagnosis Type 2 diabetes (ICD9 250.00).
- Office/Clinic Note-Physician: "Diabetes, Type 2"** - Feldman MD, Mark. Interval History Diabetes, Type 2 Glucose results within target range. ... Glucose, Fasting (Ordered): Blood, Collected, Lab Collect, ST collect, type 2 diabetes, for 1 day(s) Hemoglobin A1c (Ordered): Blood, Collected, Lab Collect, RT collect, type 2 diabetes, for 1 day(s).
- Office/Clinic Note-Physician: "Chest Pain, COPD, Diabetes, Type 2, Hypertension"** - Feldman MD, Mark. The course is progressing as expected. Diabetes, Type 2 Glucose results elevated. ... Glucose Level 170 mg/dL, HI ... HgbA1C (Ordered): Blood, Collected, Lab Collect, Routine collect, Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled.
- Assessment Forms-Text: "Adult Ambulatory Care Intake"** - Feldman MD, Mark. Date: 4/7/2011 ; Life Cycle Status: Active ; Responsible Provider: Feldman MD, Mark; Vocabulary: ICD-9-CM Diabetes mellitus type 2 Name of Problem: Diabetes mellitus type 2 - Onset Date: 11/30/2003 ; Recorder: Snyder RN ; Kaza; Confirmation: Confirmed ; Classification: Medical ;

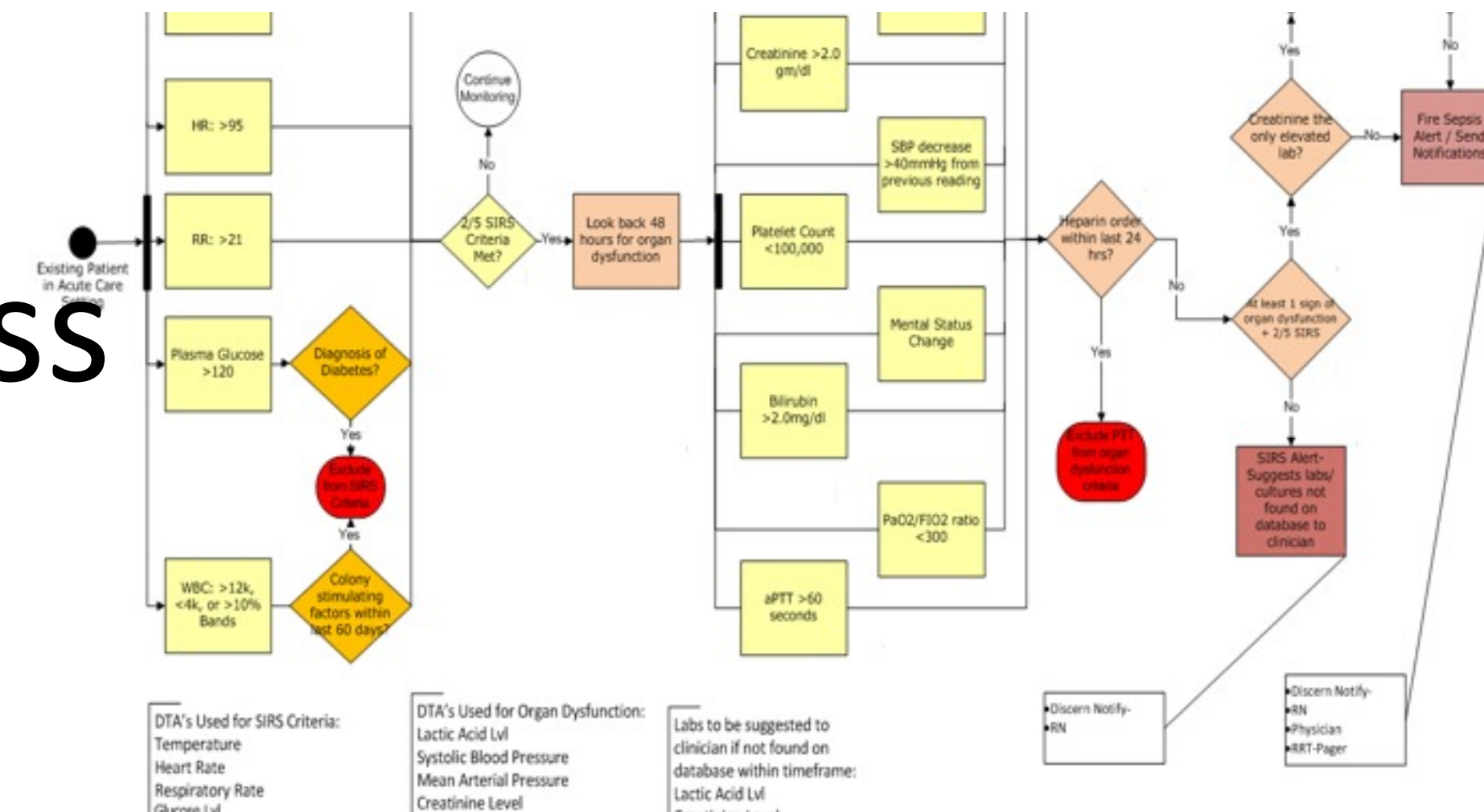


# Medical Alerts

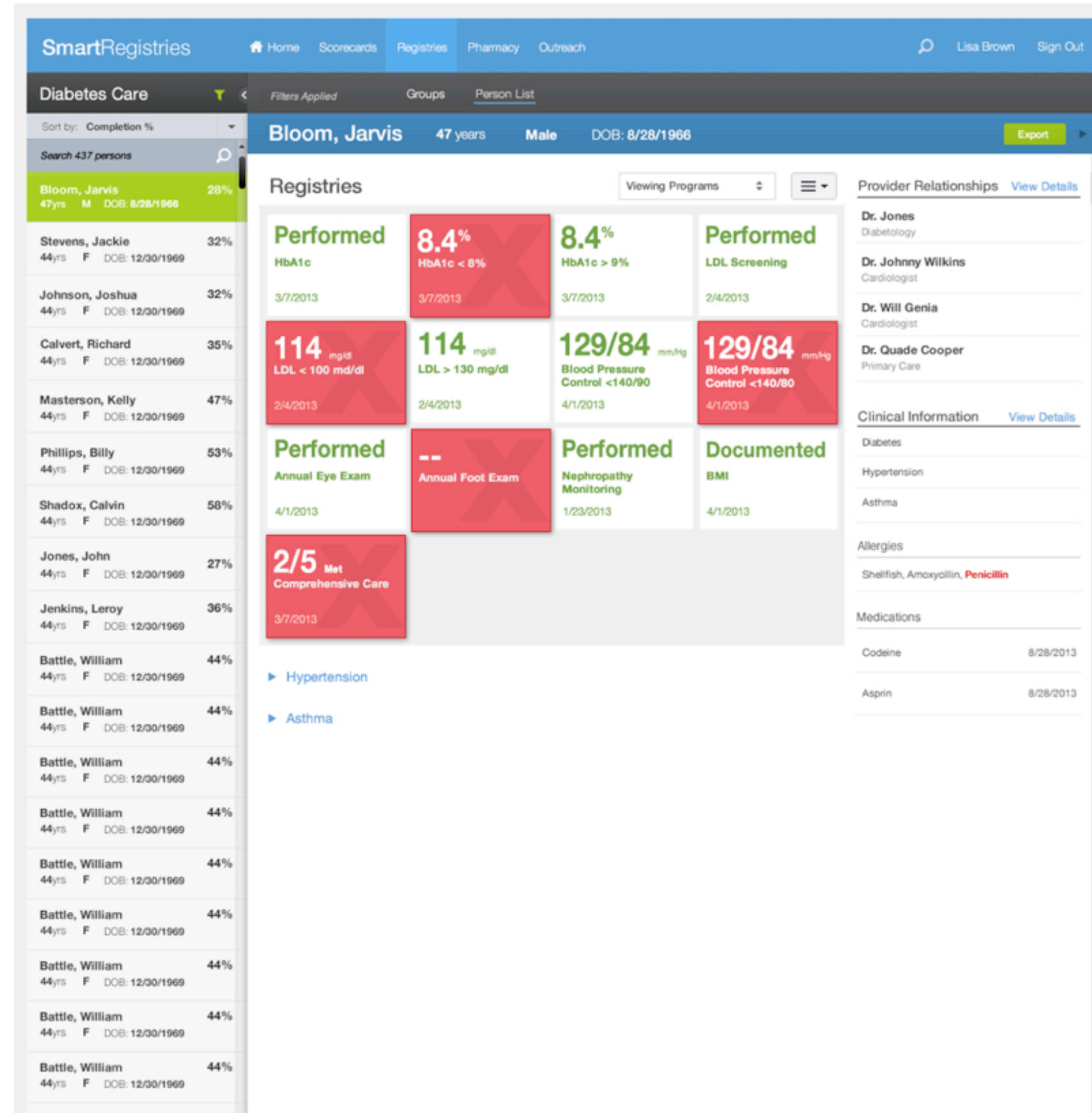


# Medical Alerts

- Detect health risks in incoming data
- Notify clinicians to address those risks
- Quickly include new knowledge



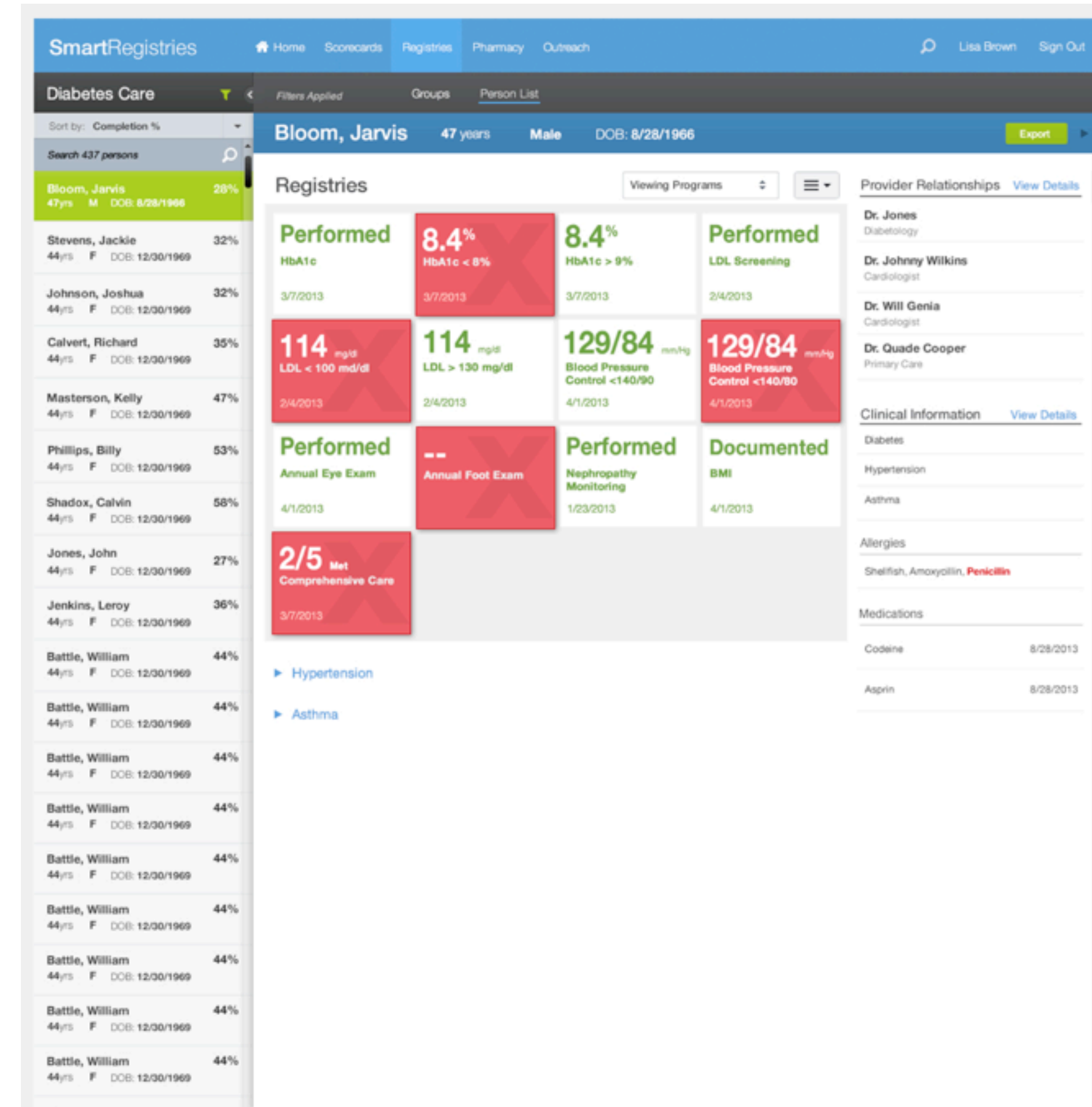
# Population Health





# Population Health

- Securely bring together health data
- Identify opportunities to improve care
- Support application of improvements
- Close the loop





# The Unreasonable Effectiveness of Data

Simple models with lots of data almost always outperform  
complex models with less data

So how can we tackle  
such large data sets?

Can we adapt what has  
worked historically?

After all,

***Relational Databases are Awesome***

Atomic, transactional updates

Guaranteed consistency

***Relational Databases are Awesome***

Declarative queries

Easy to reason about

Long track record of success

***Relational Databases are Awesome***

***...so use them!***

***Relational Databases are Awesome***

***...so use them!***

***But...***

# Those advantages have a cost

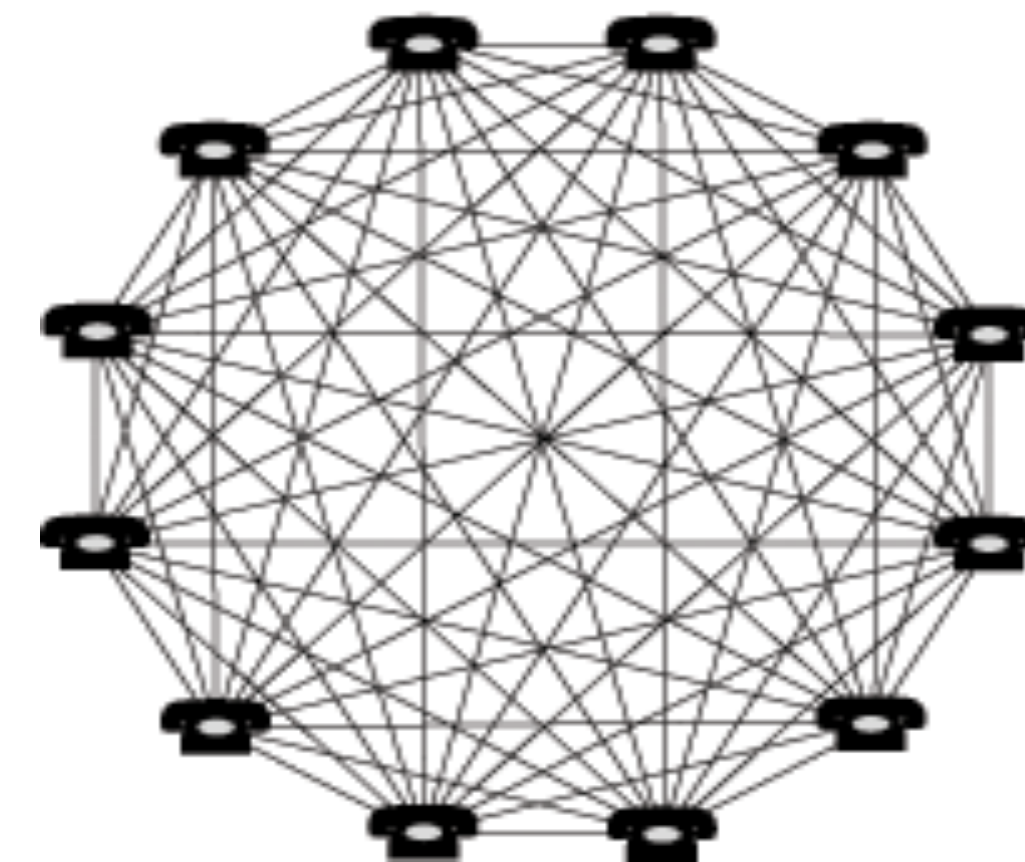
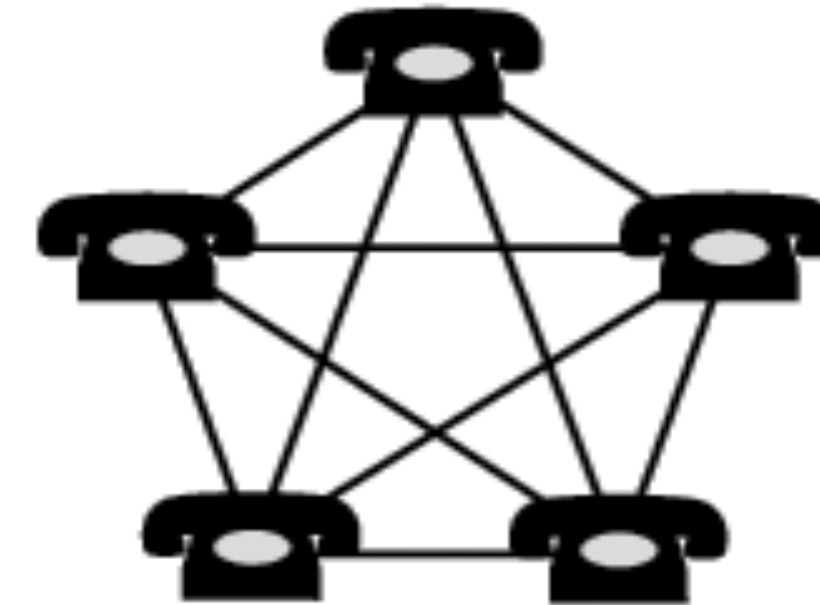
Global, atomic, consistent state means  
global coordination

Coordination does not scale linearly



# The costs of coordination

Remember the  
network effect?



# The costs of coordination

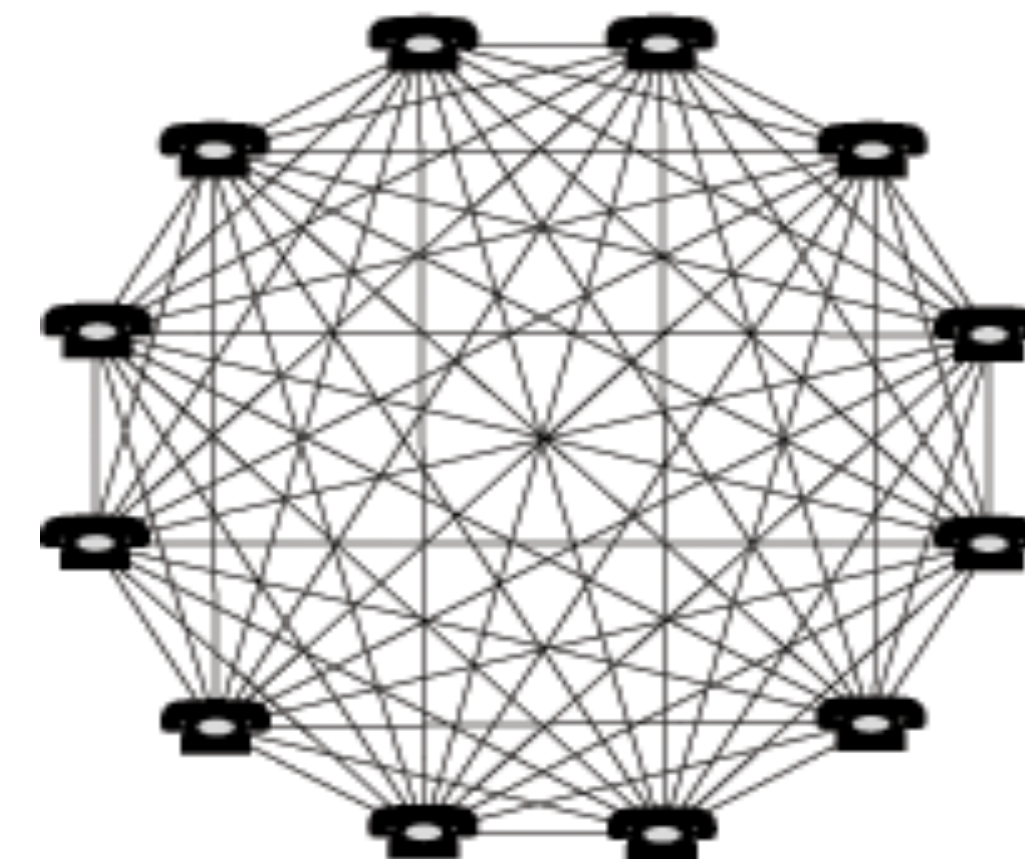
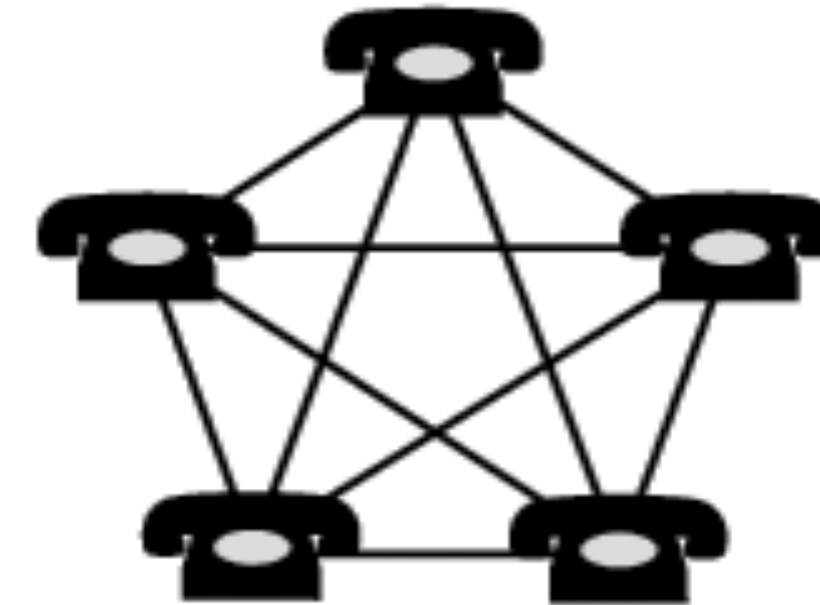
$$\text{channels} = \frac{n(n-1)}{2}$$

2 nodes = 1 channel

5 nodes = 10 channels


12 nodes = 66 channels


25 nodes = 300 channels



The result is we don't scale linearly as we  
add nodes

**Independence**  **Parallelizable**

**Independence**  **Parallelizable**

**Parallelizable**  **Scalable**

“Shared Nothing” architectures are the most scalable...

...but most real-world problems require us to share *something*...

...so our designs usually have a *parallel* part and a *serial* part

The key is to make sure the vast majority of our work in the cloud is *independent* and *parallelizable*.

# Amdahl's Law

$$S(N) = \frac{1}{(1-P) + \frac{P}{N}}$$

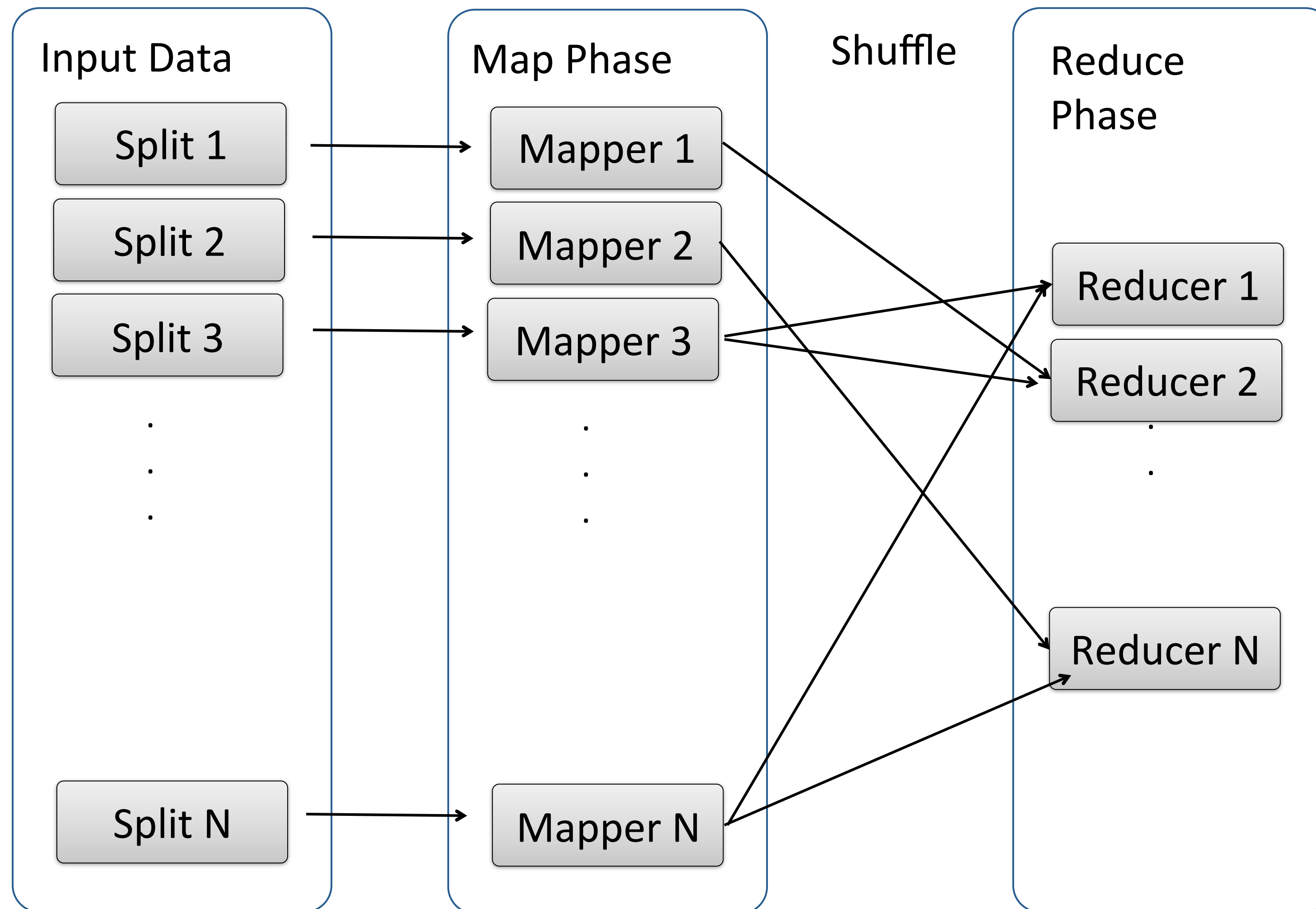
S : speed improvement

P : ratio of the problem that  
can be parallelized

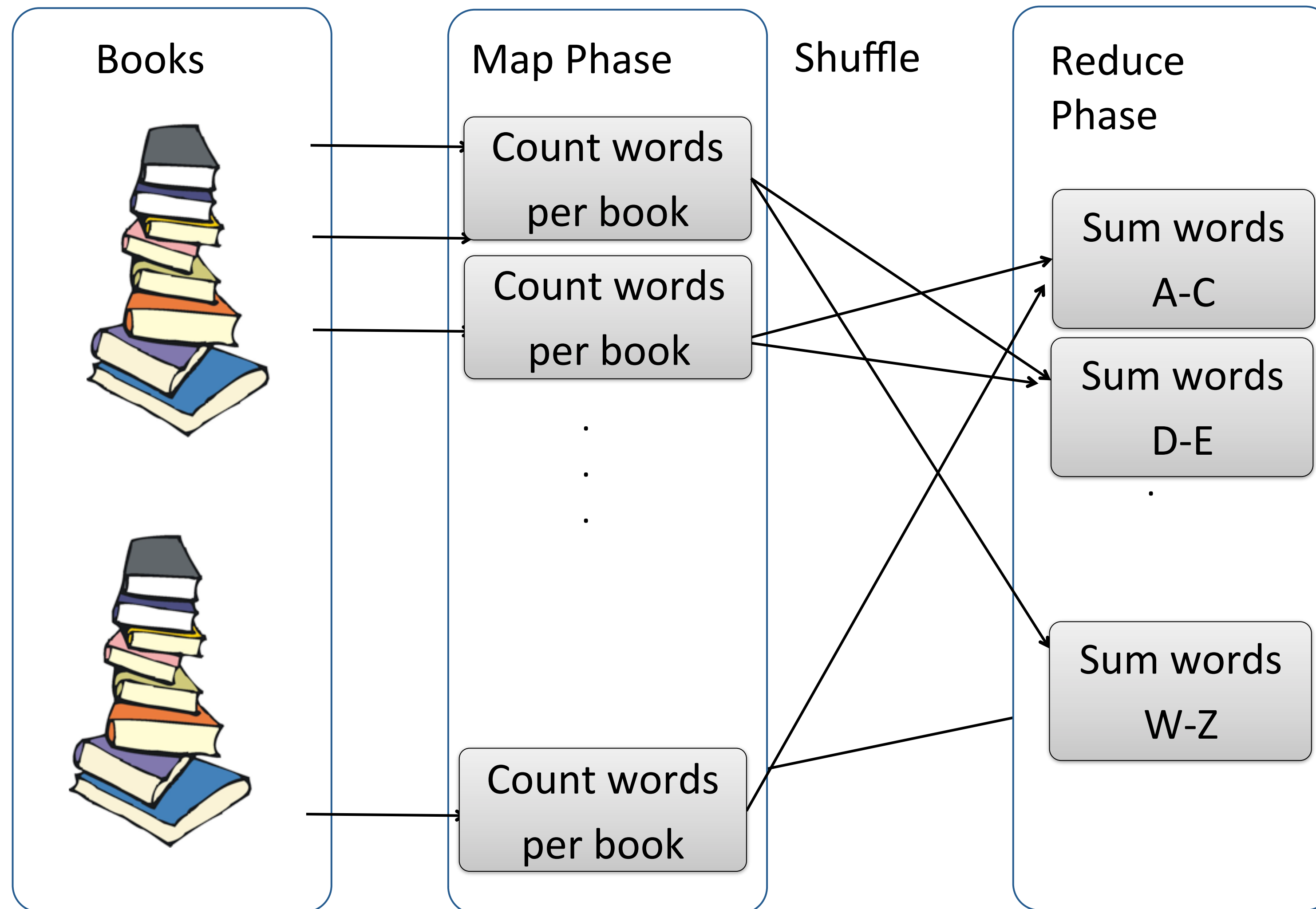
N: number of processors



# MapReduce Primer



# MapReduce Example: Word Count

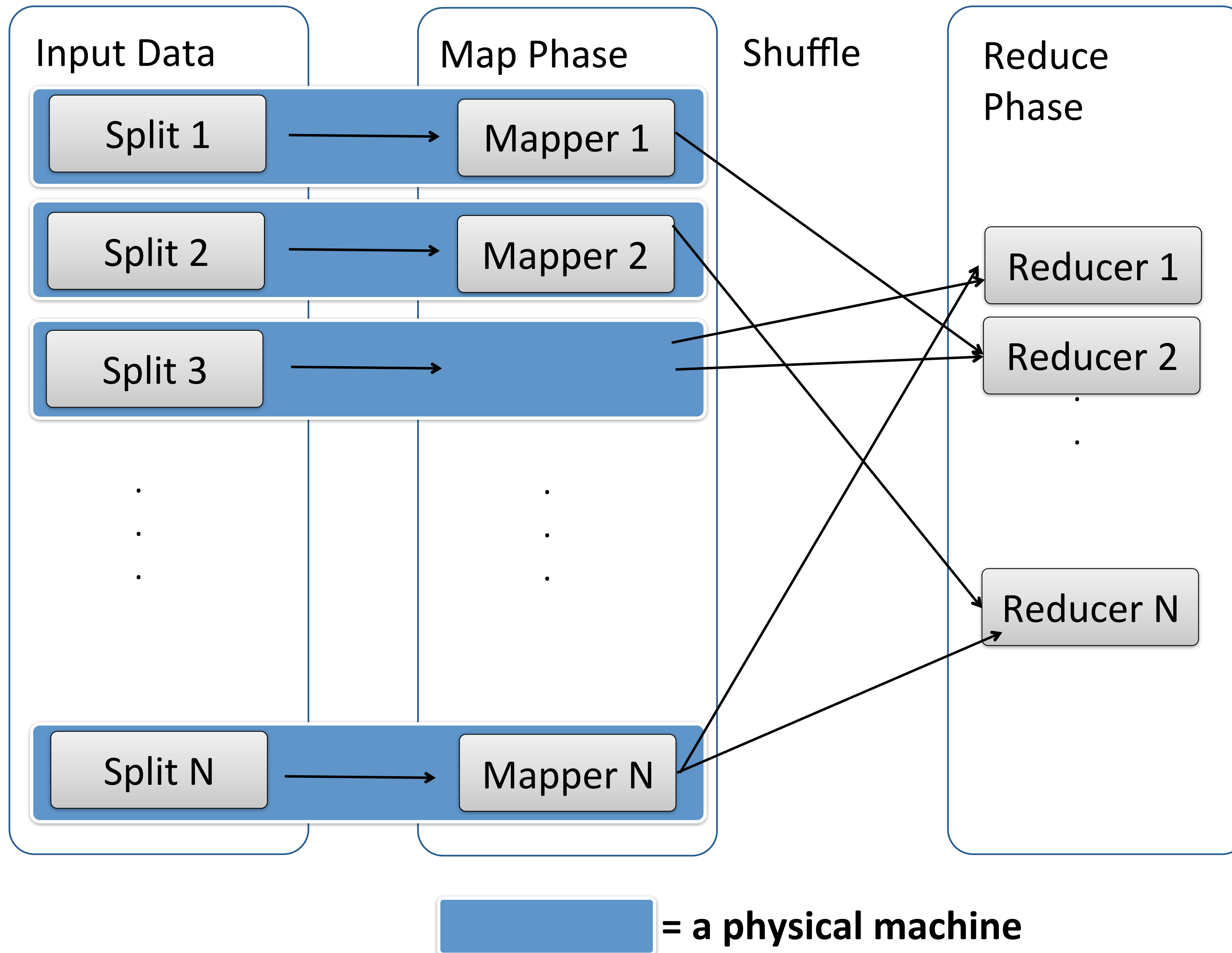


The network is a shared resource

Too much data to move to computation

So move *computation* to *data*

# MapReduce Data Locality



Data locality only guaranteed in  
the Map phase

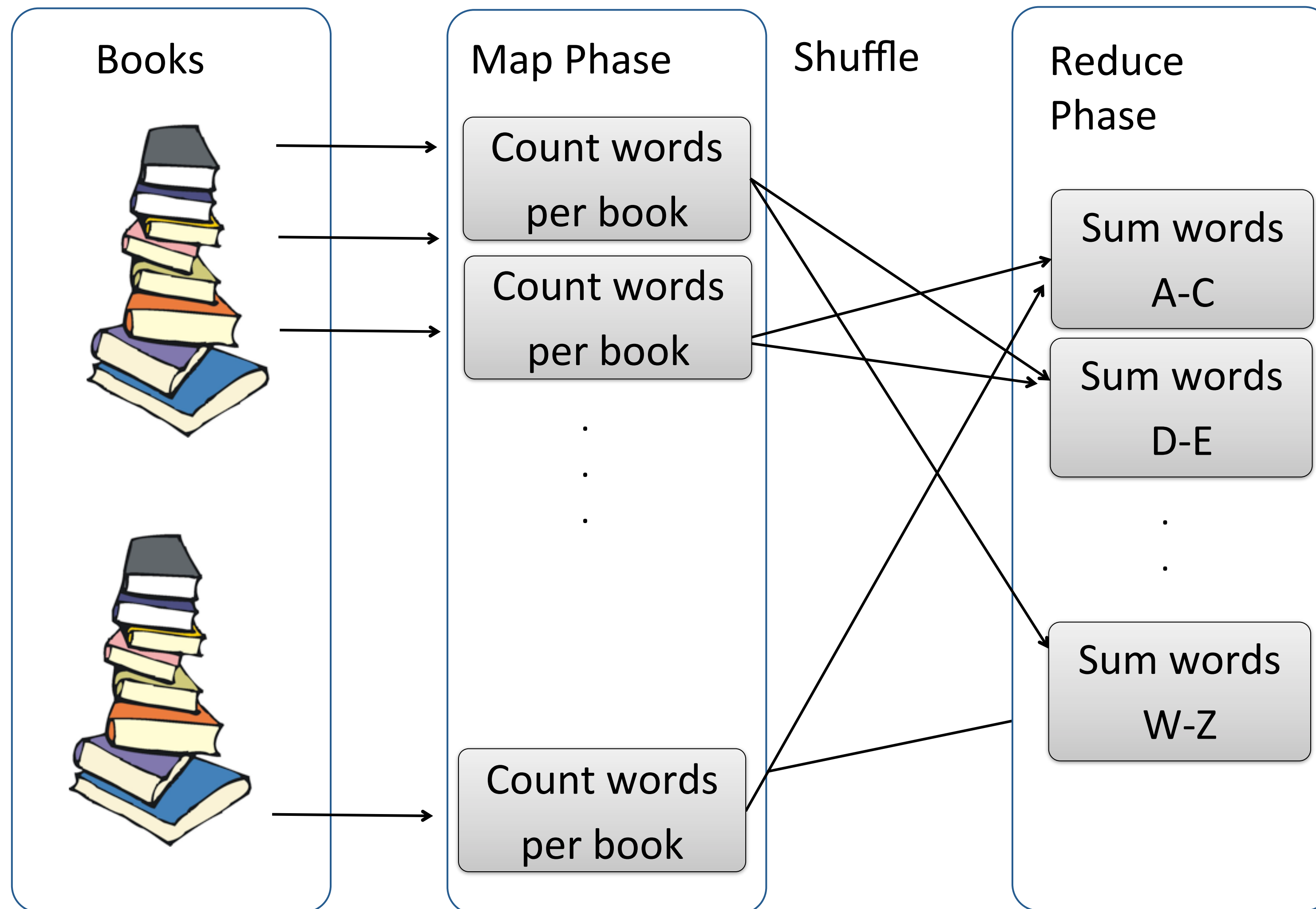
So do as much work as possible there

Some jobs have no reducer at all!

MapReduce is a building block

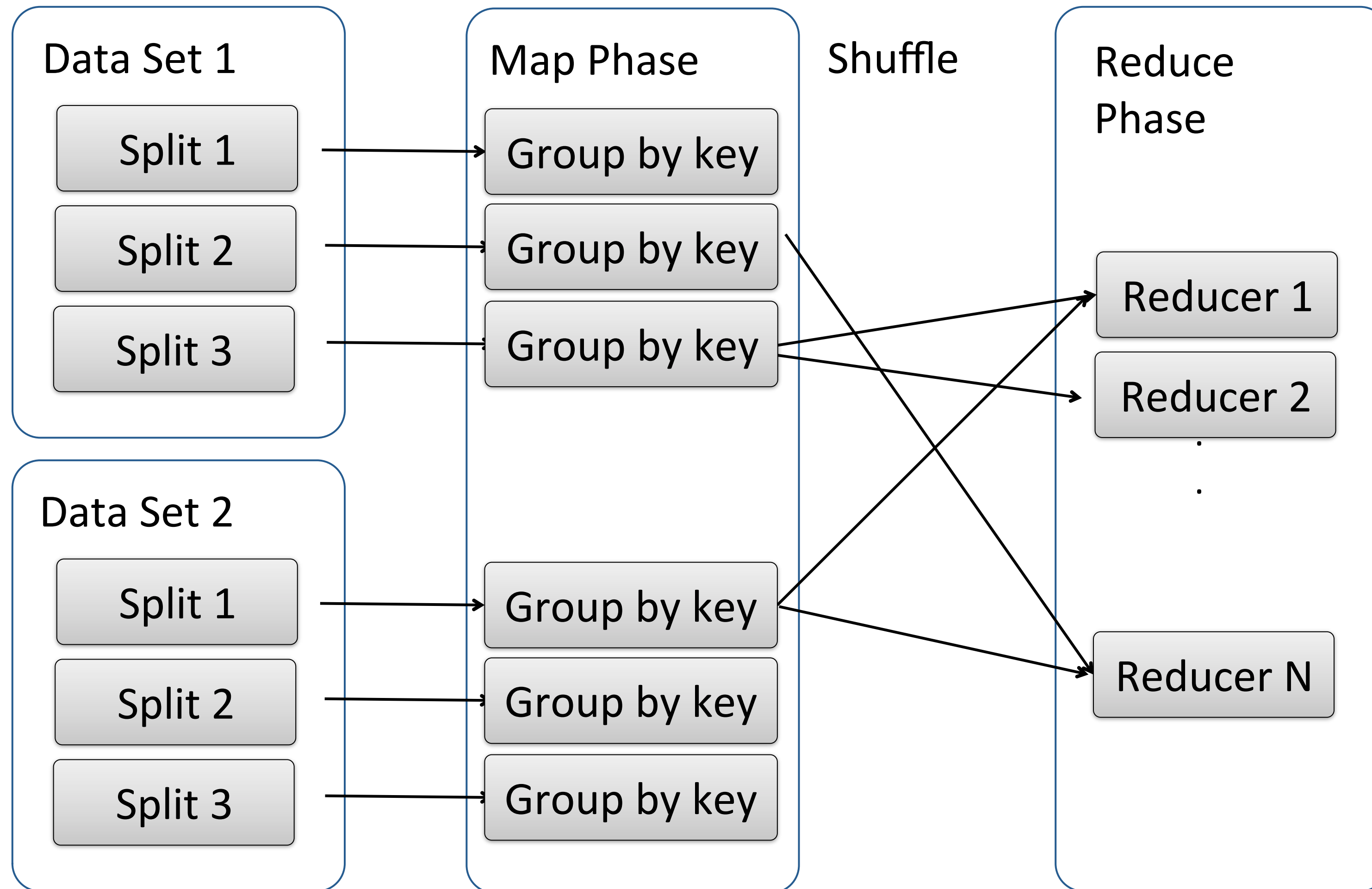
So let's build higher-level functions

# Grouping and Aggregating

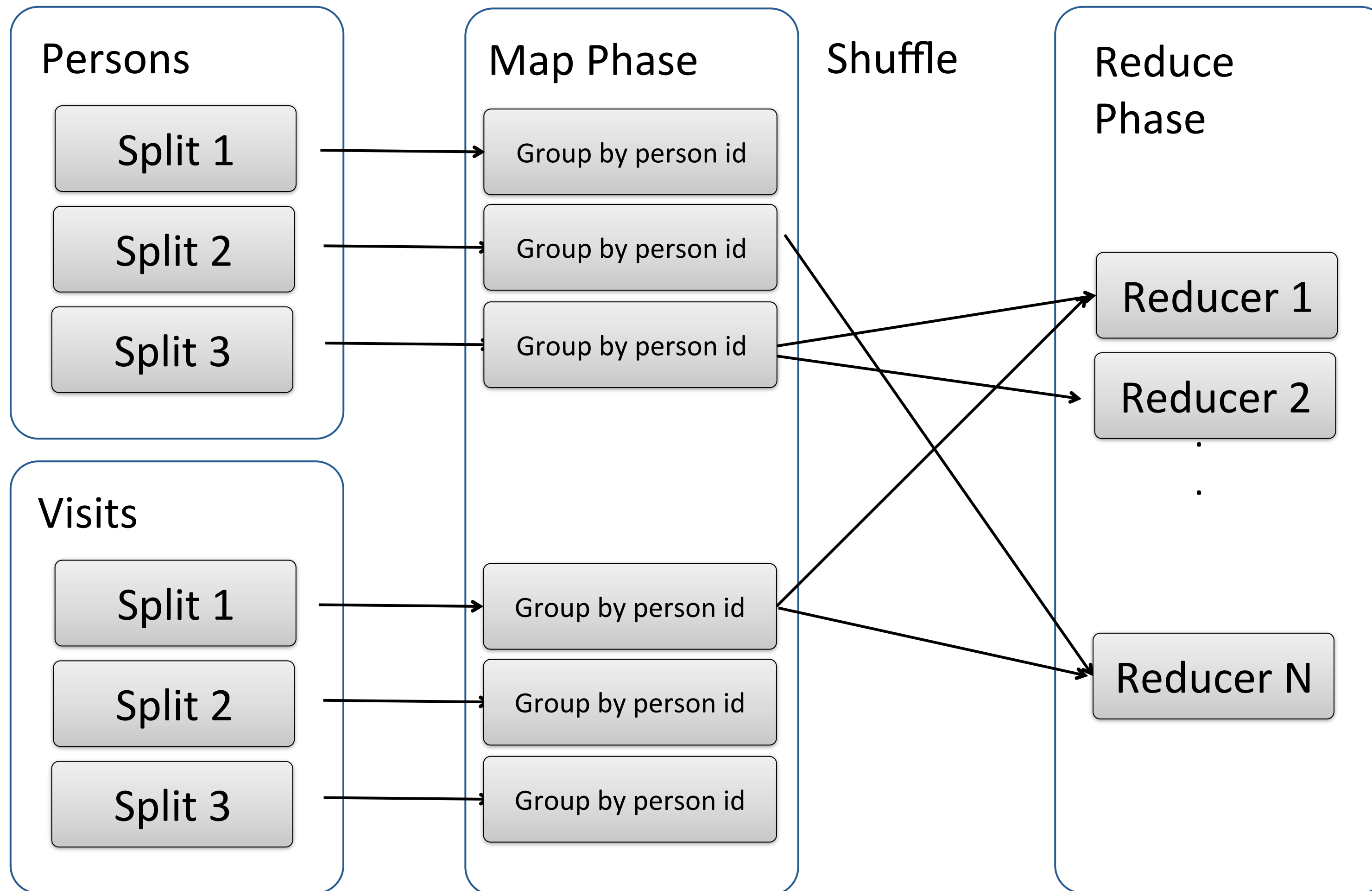




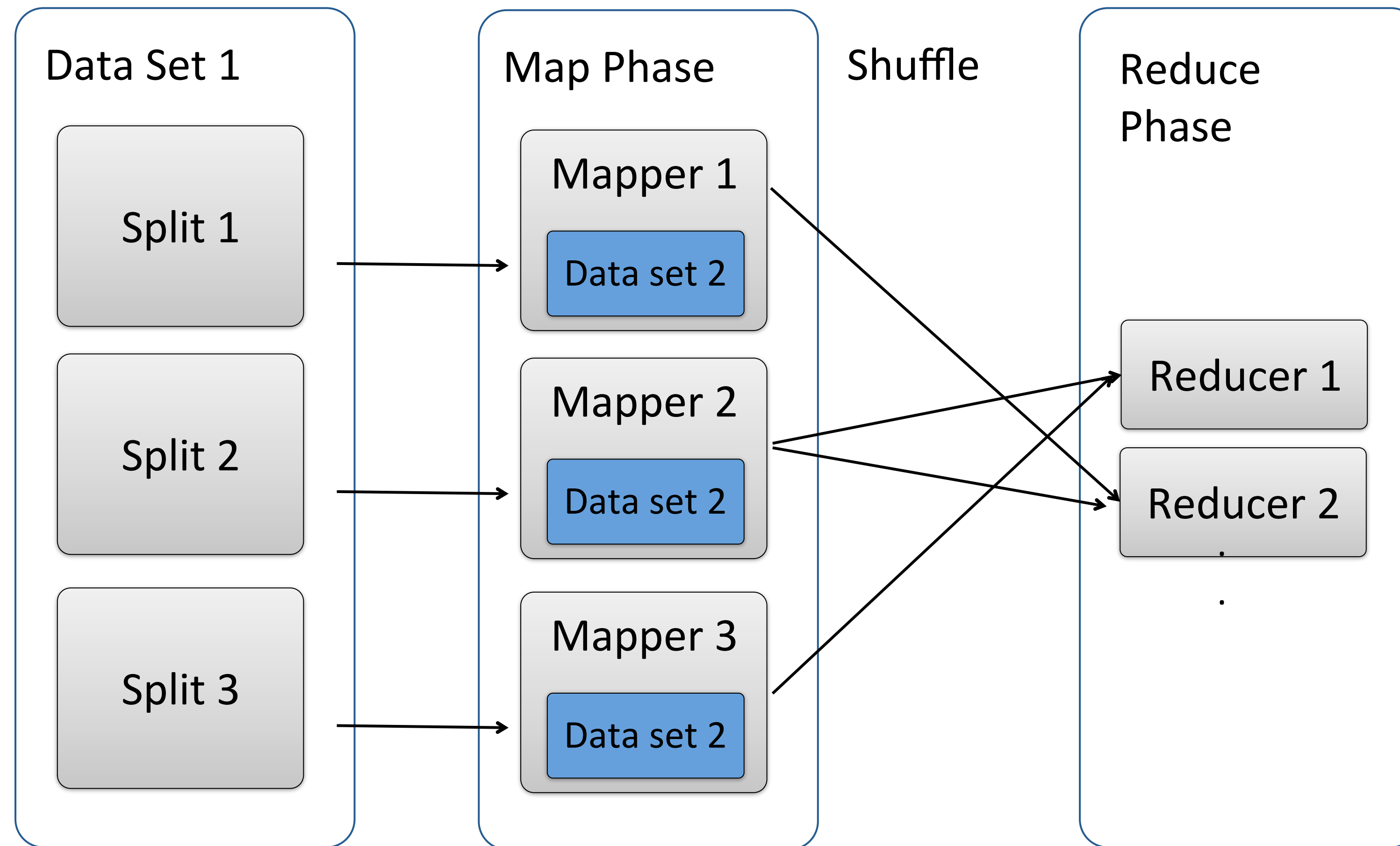
# Joins



# Joins



# Map-Side Joins



# Filtering

Map or reduce functions can simply discard data we're not interested in

# And Others

Distinct

Sort

Binning

Top N

...

More sophisticated patterns  
composable

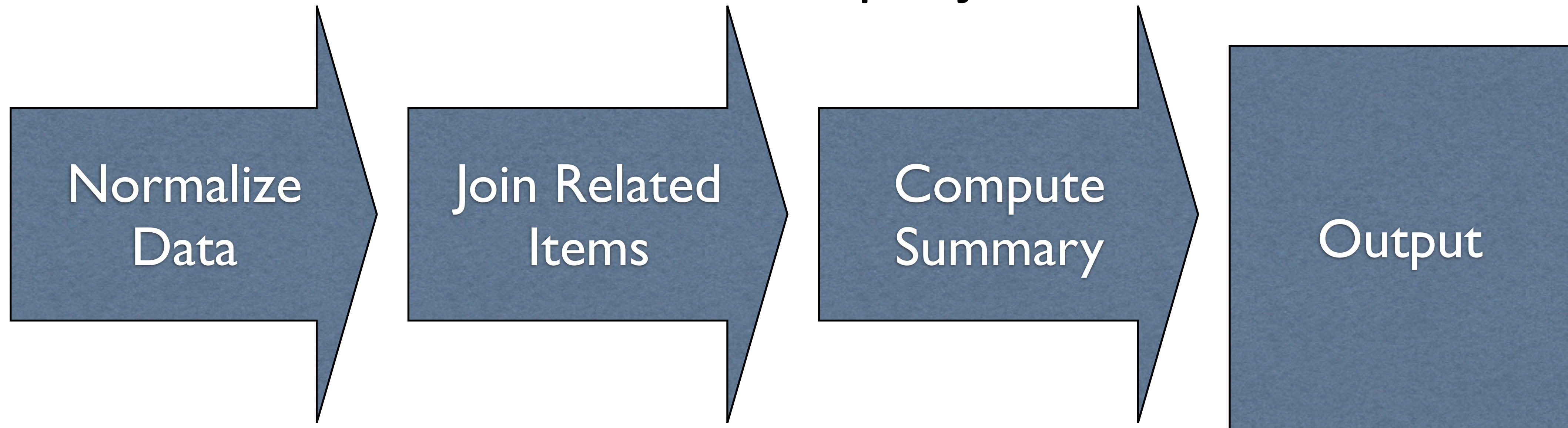




# Chain Jobs Together

Large-scale joins must have a reduce phase

Multiple joins or group-by operations  
mean multiple jobs



# Codified in High-Level Libraries

Hive, Pig, Cascading, and Crunch provide simple means to use these patterns

The era of writing MapReduce by hand is over



Apache  
Crunch

How do we use these tools?



*Start with the question you  
want to ask, then transform the  
data to answer it.*

*output = transform (input)*

Functional Programming over  
Place-Oriented Programming

Work with data holistically

Re-running functions simpler to reason about  
than updating state

Hadoop makes this possible at scale

Don't be afraid to re-process  
the world

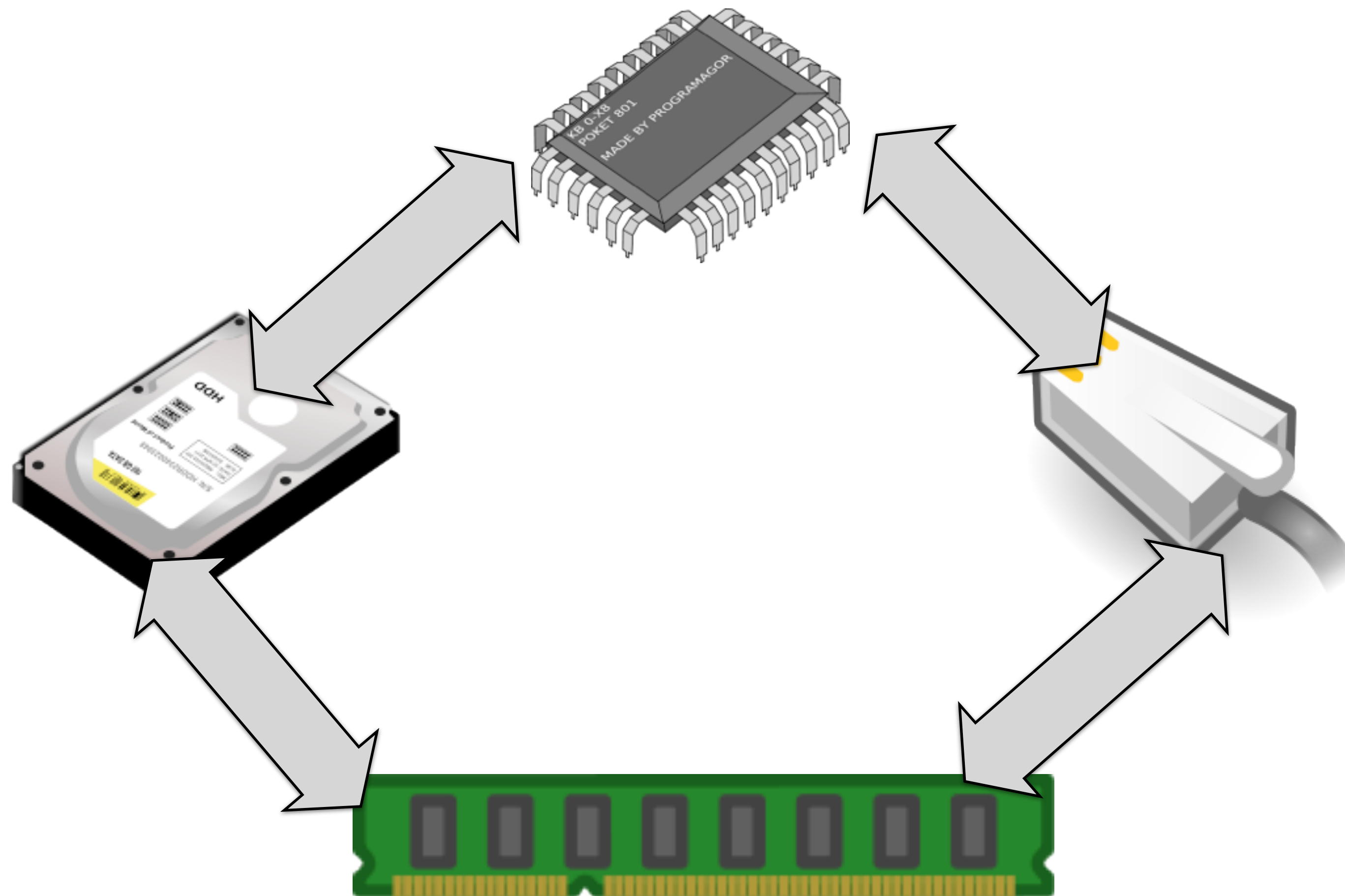
*Something's wrong, we're above 95% usage!*

-Traditional System Administrator

*Something's wrong, we're below 95% usage!*

-Hadoop System Administrator

# Maximize Resource Usage



# From Databases to Dataspaces

(Also referred to as Data Lakes)

Bring all of your data together...  
..structured or unstructured...  
...transform it with unlimited  
computation...  
...at any time for any new need.

And offer a variety of interactive  
access patterns.

SQL, Search, Domain-Specific Apps



Hadoop is becoming an adaptive,  
multi-purpose platform.

The gap between asking novel questions  
and our ability to answer them  
is closing.

Questions?