A-Brain: Using the Cloud to Understand the Impact of Genetic Variability on the Brain

Joint work with
Alexandru Costan, Benoit Da Mota, Gabriel Antoniu, Bertrand Thirion

Radu Tudoran
KerData Team
Inria Rennes
ENS Cachan

10 April 2012

The A-Brain Project

Application

- Large-scale joint genetic and neuroimaging data analysis

Goal

- Assess and understand the variability between individuals

Approach

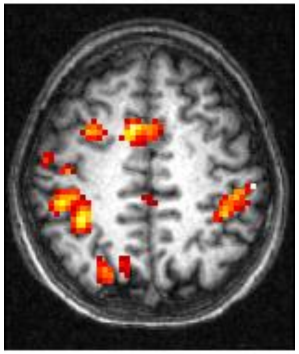
- Optimized data processing on Microsoft's Azure clouds

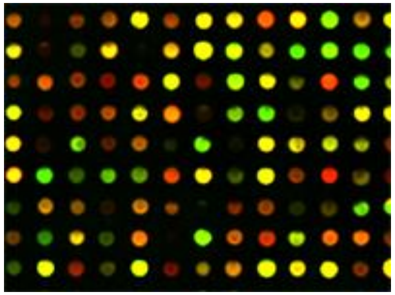
Inria teams involved

- KerData (Rennes)
- Parietal (Saclay)

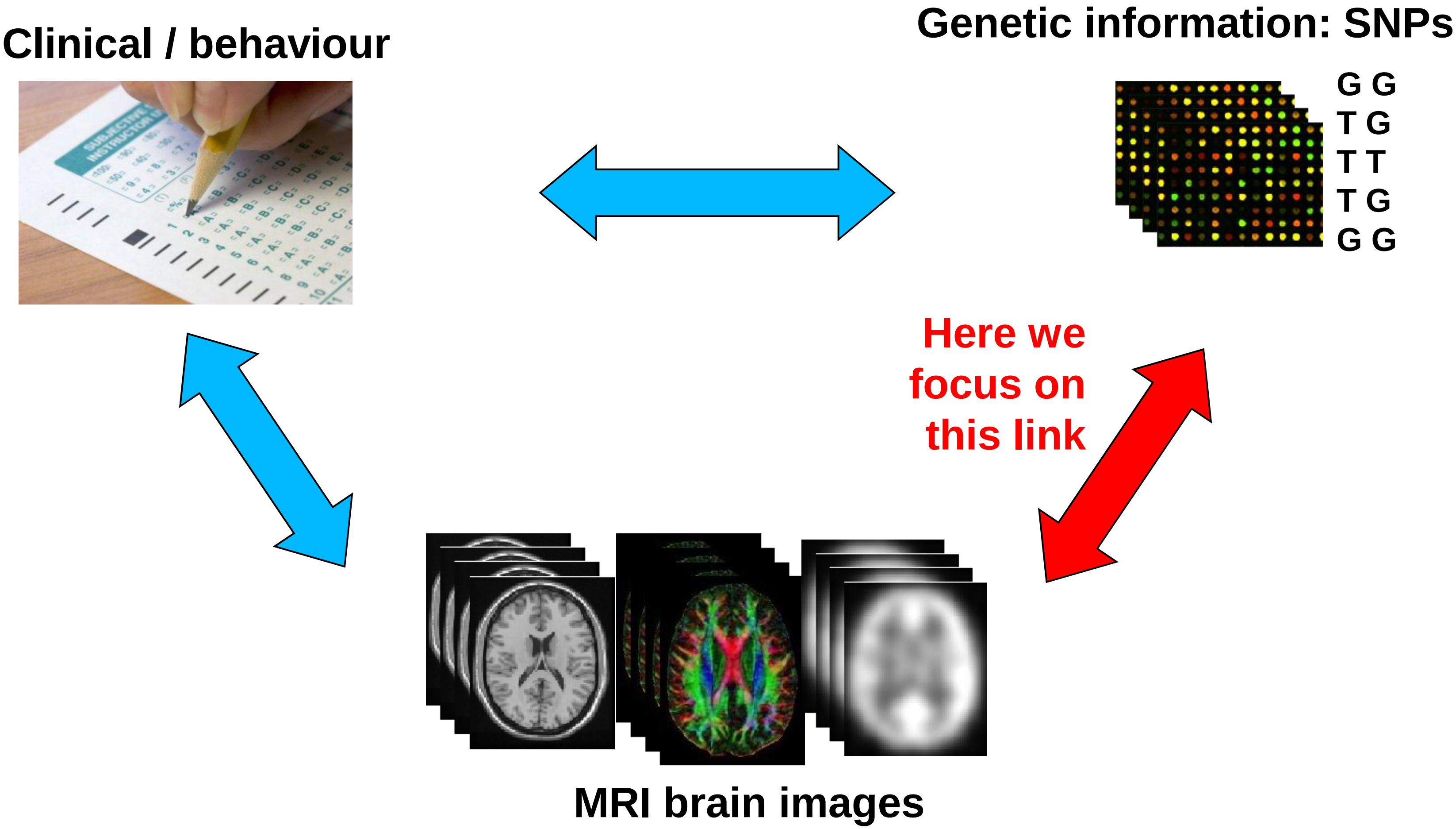
Framework

- Joint MSR-Inria Research Center
- MS involvement: Azure teams, EMIC





The Imaging Genetics Challenge: Comparing Heterogeneous Information



Clinical / behaviour

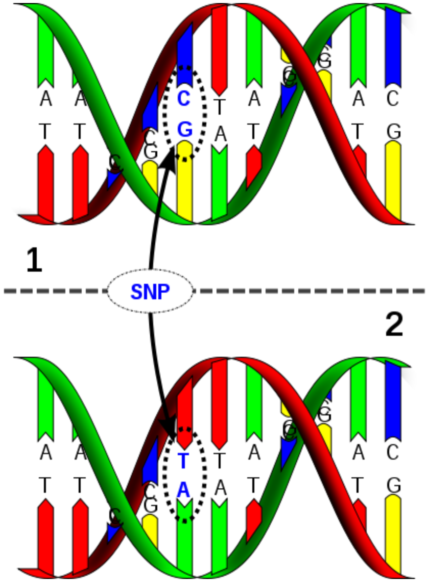
Genetic information: SNPs

G G
T G
T T
T G
G G

Here we focus on this link

MRI brain images

Neuroimaging-genetics: The Problem



- Several brain diseases have a genetic origin, or their occurrence/severity related to genetic factors
- Genetics is important to understand & predict response to treatment
  - identify risk and protective factors for brain diseases
  - Brain: Huntington's disease, autism…

- Currently: large-scale studies to assess the relationships between diseases and genes: typically 10^4 patients per study + control groups
- Genetic variability captured in DNA microarray data

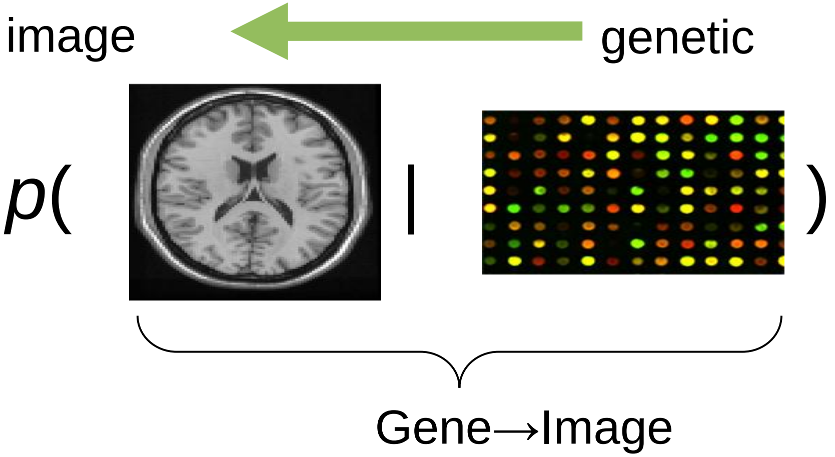


image ← genetic

p(|)

Gene→Image

A-Brain

Brain image ⟷ Genetic data

finding associations: $p($, $)$

$q \sim 10^{5\text{-}6}$

$p \sim 10^{6}$

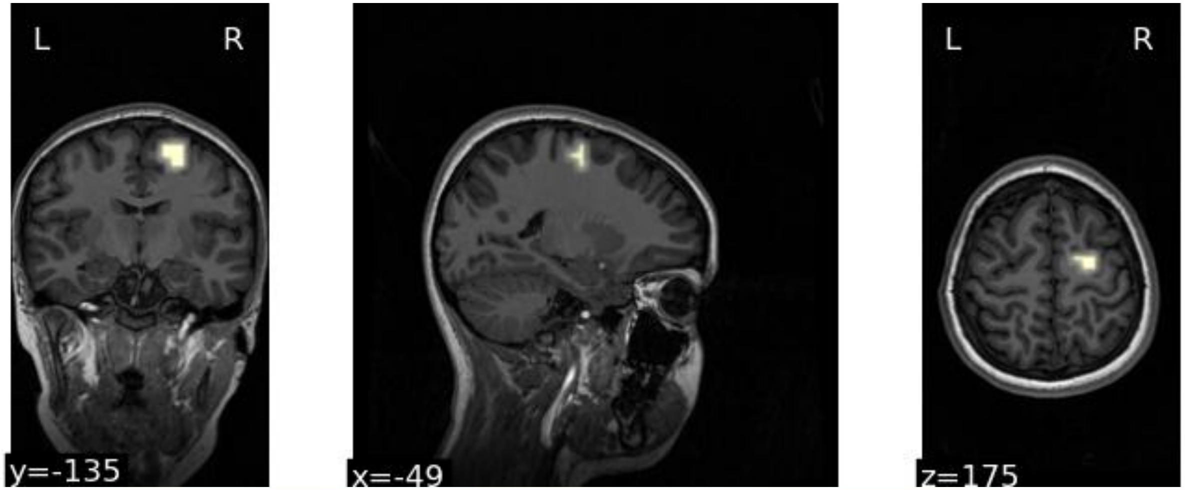
$N \sim 2000$

Y

– Anatomical MRI
– Functional MRI
– Diffusion MRI

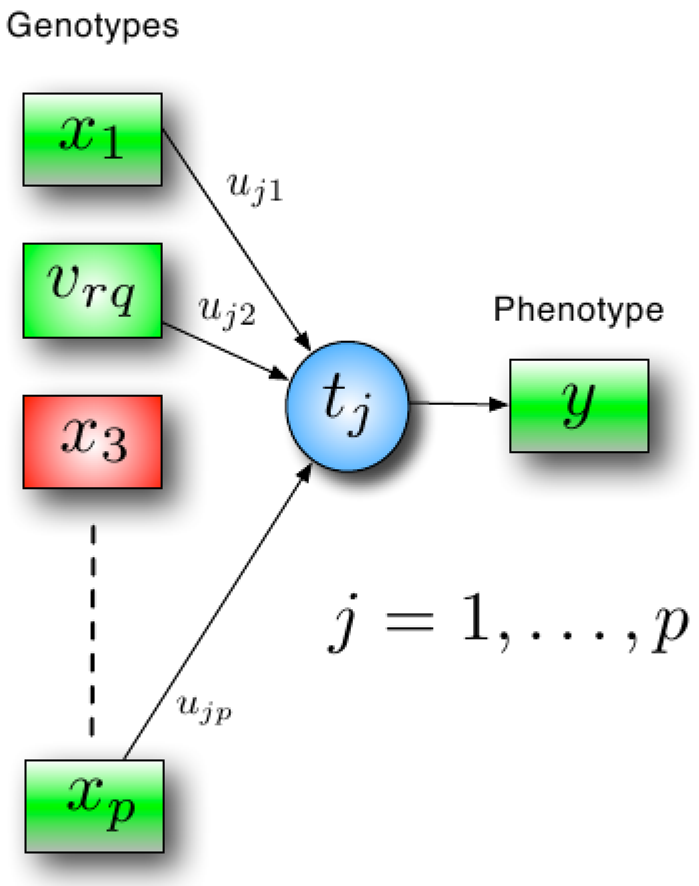
X

– DNA array (SNP/CNV)
– gene expression data
– others...



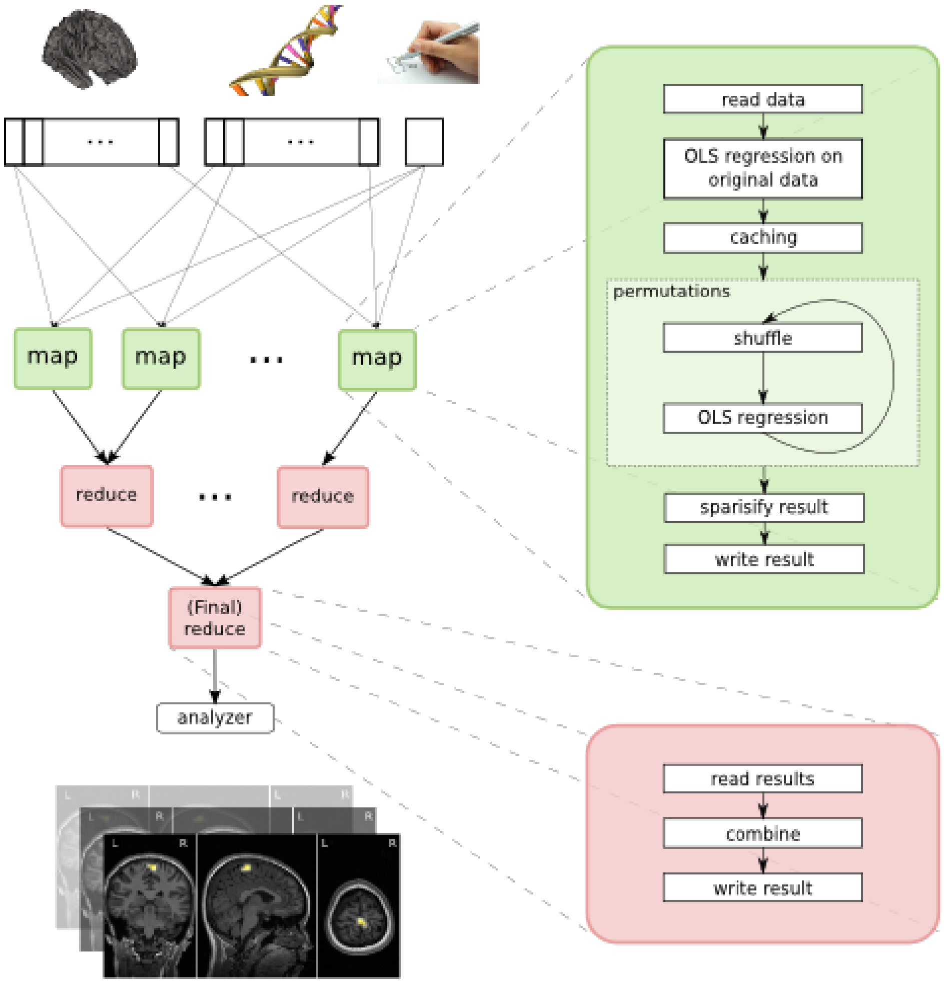
Imaging Genetics Methodological Issues

- Multivariate methods: predict brain characteristic with many genetic variables
- Elastic net regularization: combination of ℓ_1 and ℓ_2 penalties → sparse loadings
- O(p^3 complexity)
- parameters setting: internal cross-validation/bootstrap
- Performance evaluated using permutations



$$\hat{\beta}^{enet} = \text{argmin}_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} (y_i - \sum_{k=1}^{p} x_{ik}\beta_k)^2 + \lambda_1 \sum_{k=1}^{p} |\beta_k| + \lambda_2 \sum_{k=1}^{p} \beta_k^2 \right\}$$

A-Brain as MapReduce process



map

map

...

map

reduce

...

reduce

(Final) reduce

analyzer

read data

OLS regression on original data

caching

permutations

shuffle

OLS regression

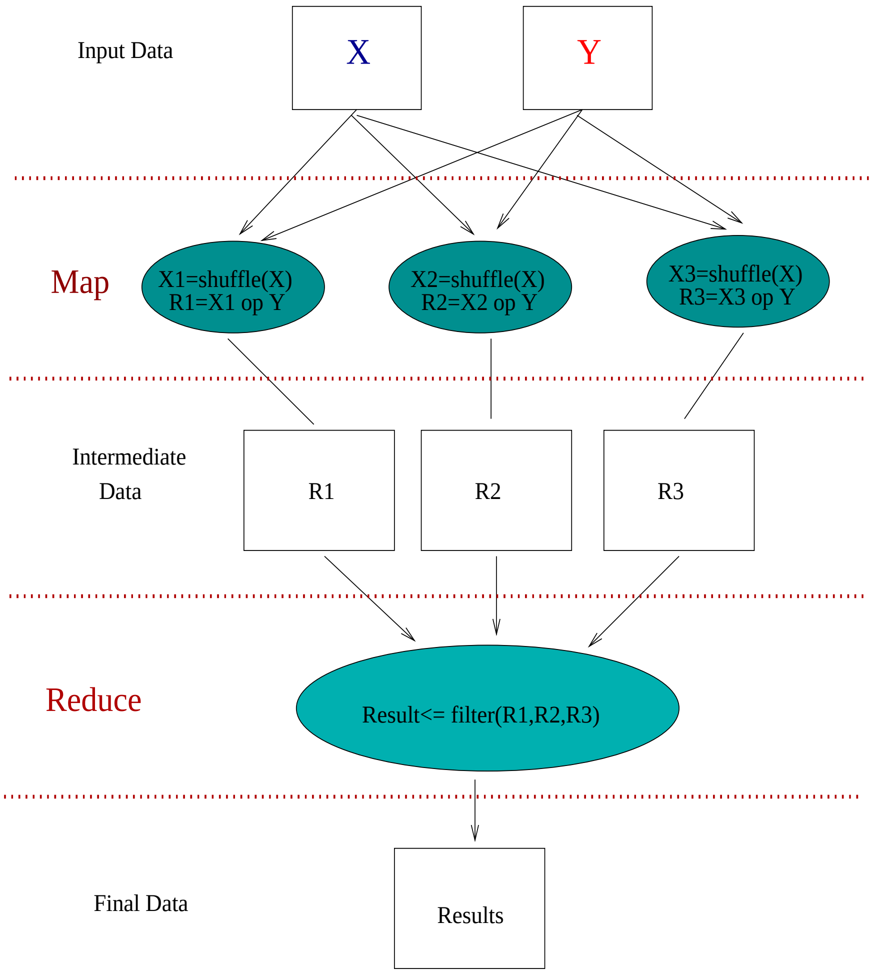
sparisify result

write result

read results

combine

write result



Input Data

X

Y

Map

X1=shuffle(X)
R1=X1 op Y

X2=shuffle(X)
R2=X2 op Y

X3=shuffle(X)
R3=X3 op Y

Intermediate Data

R1

R2

R3

Reduce

Result<= filter(R1,R2,R3)

Final Data

Results

Challenges …

Data: $\underbrace{8}_{\text{double}} * \underbrace{10^4}_{\text{permutation}} * \underbrace{5 * 10^4}_{\text{voxels}} * \underbrace{5 * 10^5}_{\text{SNPs}} \Rightarrow 1.77\ PB$

5%-10% useful

Computation: $\overbrace{10^4} * \overbrace{5 * 10^4} * \overbrace{5 * 10^5} \Rightarrow\ 2.5 * 10^{14}\ \ associations$

Initial Algorithm: $1.67 * 10^4\ \ associations/seconds$

Current Algorithm: $1.5 * 10^6\ \ associations/seconds$

Estimate timespan on single machine $1.67 * 10^8\ \ seconds\ \ \Rightarrow\ \ 5.3\ years$

Azure can help…

Evaluation of the algorithm on Azure : $1.47 * 10^{6}\ associations/second$

Estimation for A-Brain on Azure (350 cores)

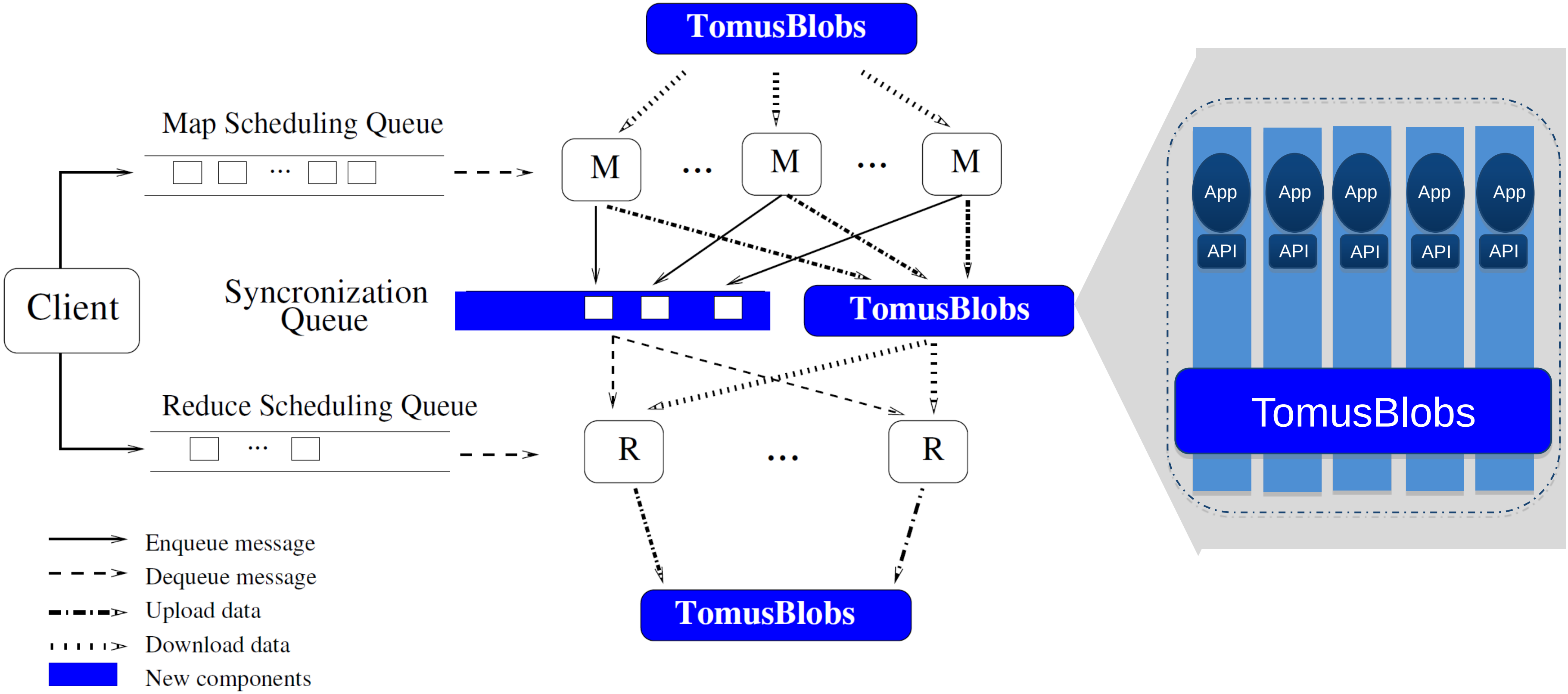
$$\frac{2.5 * 10^{14}}{350 * 1.47 * 10^{6}} \text{ seconds } \approx 485 * 10^{3}\ seconds$$

$$5.3\ years \Rightarrow 5.6\ days$$

Storage capacity estimations (350 cores) $255GB * 350 \approx 87TB$

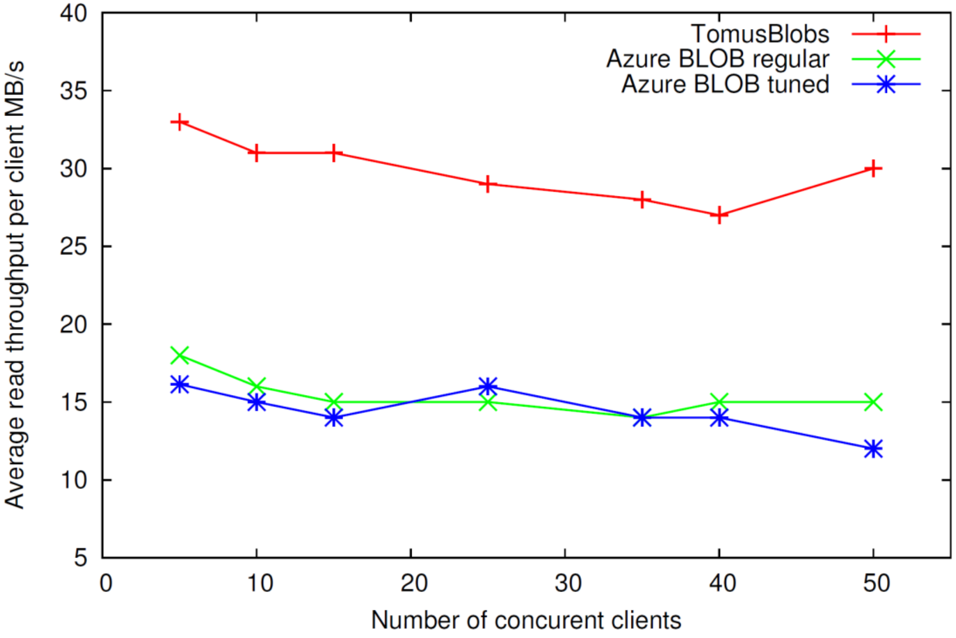
- Feats the 5% threshold of useful data
- We can always do several iterations

TomusBlobs as a Storage Backend for Sharing Application Data in MapReduce



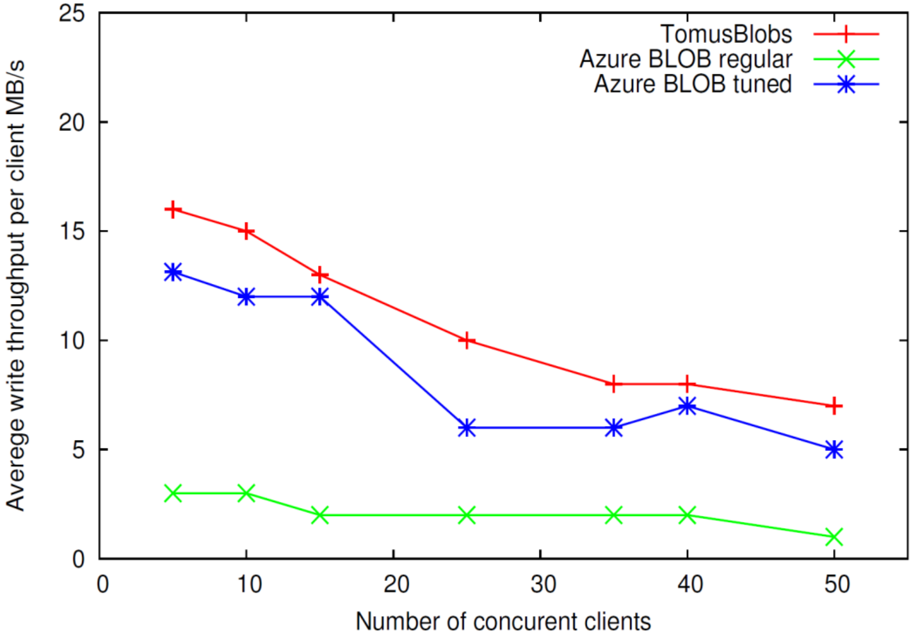
TomusBlobs: Application’s Throughput

Application pattern read throughput



read: **2.5x**

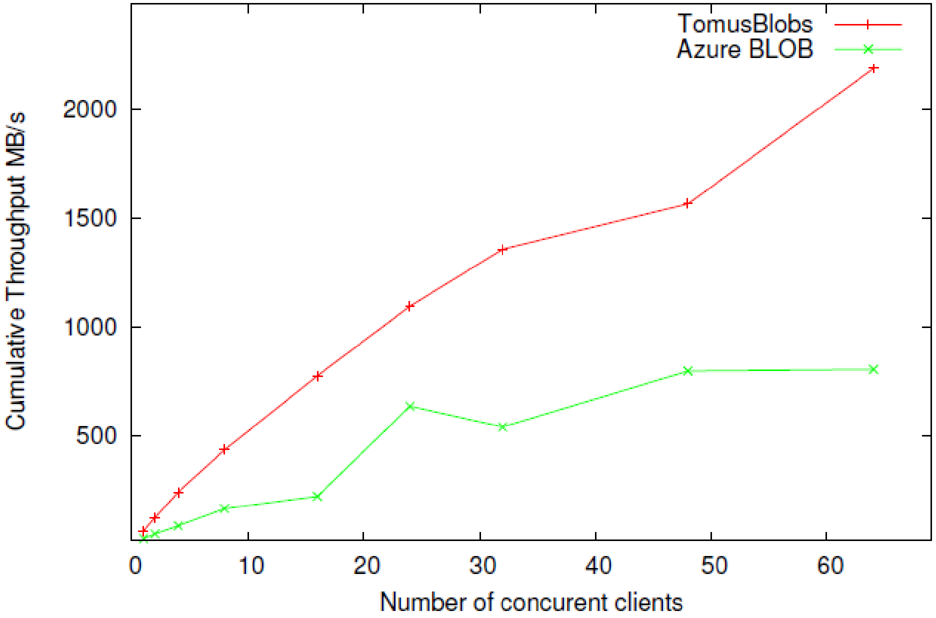
Application pattern write throughput



write: **3x**

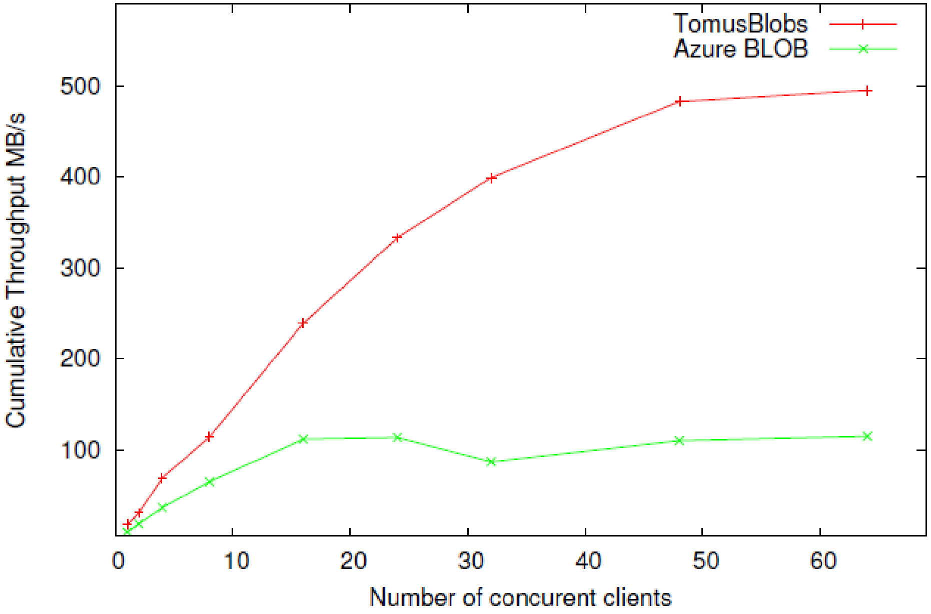
TomusBlobs: Cumulative Throughput

Cumulative read throughput



read: **4x**

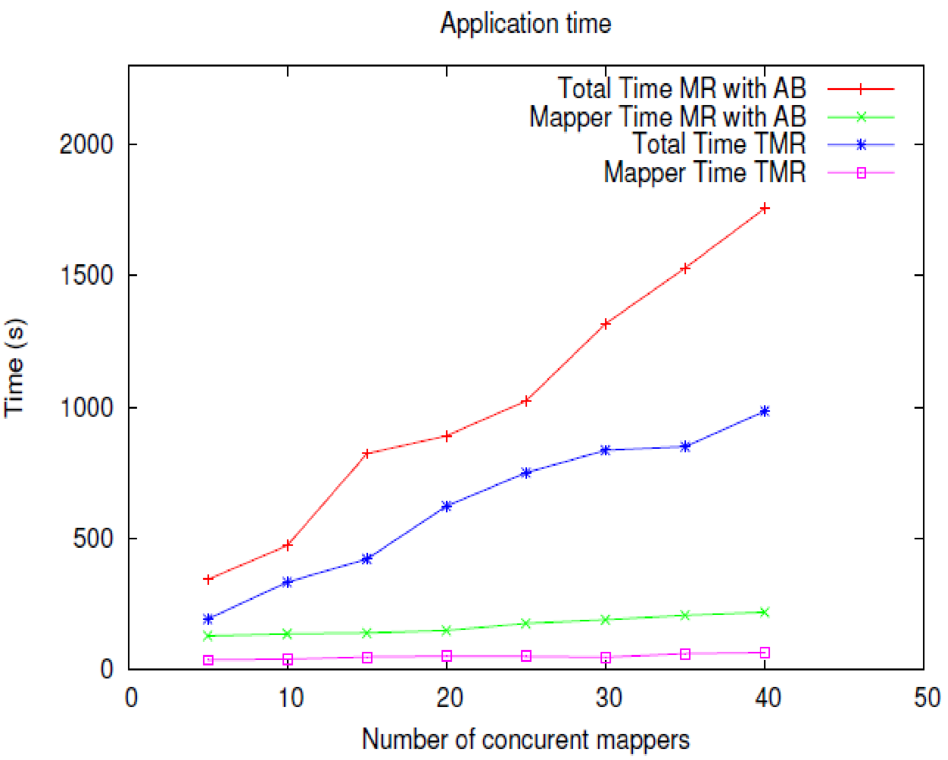
Cumulative write throughput



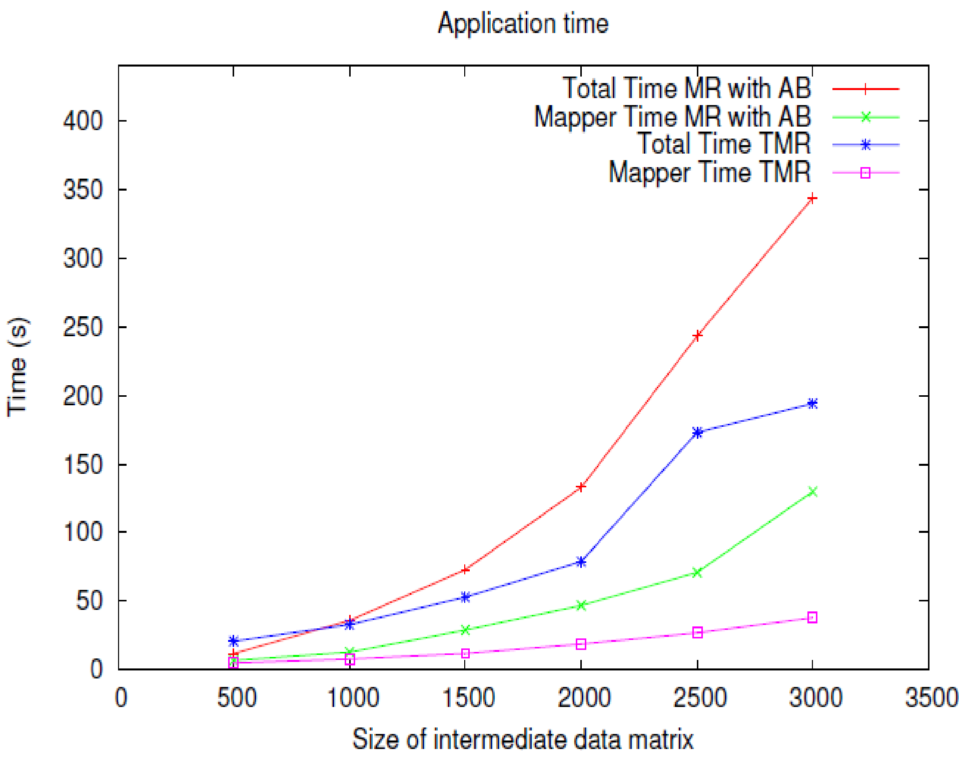
write: **5x**

A-Brain’s timespan

Increase precision



Increase data size



Our experience on Azure in the A-Brain project

- Scale up to 350 cores
- Memory/CPUs tradeoff for the VM selection
- Planning soon to launch “the big experiments”
- Continuous running time so far 1-2 days
- ≈ 60K hours of computation used so far