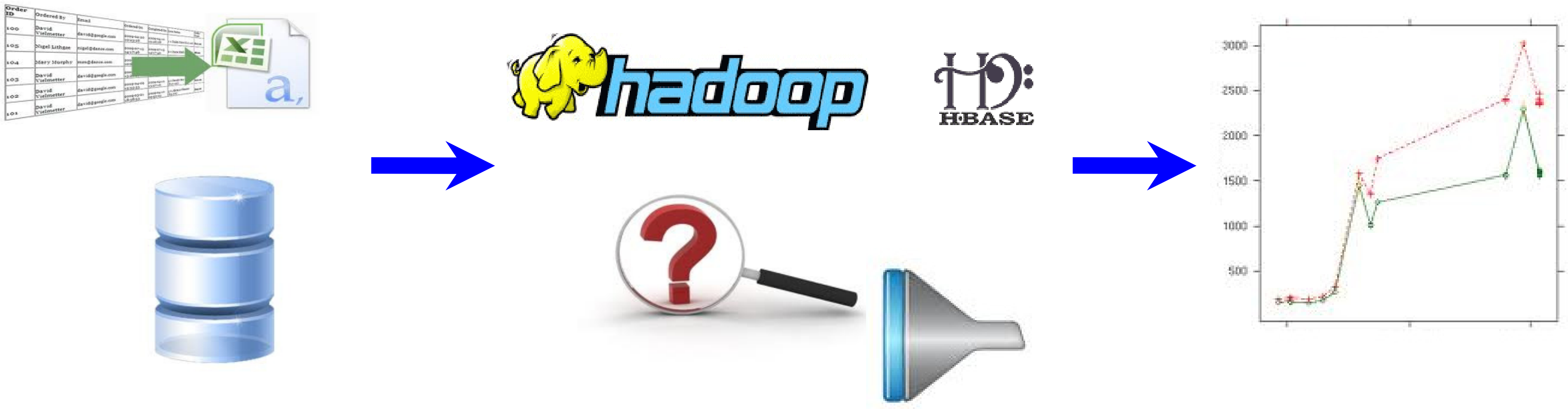
Distributed Correlation Discovery

CMPE 272 Final Project

San Jose State University

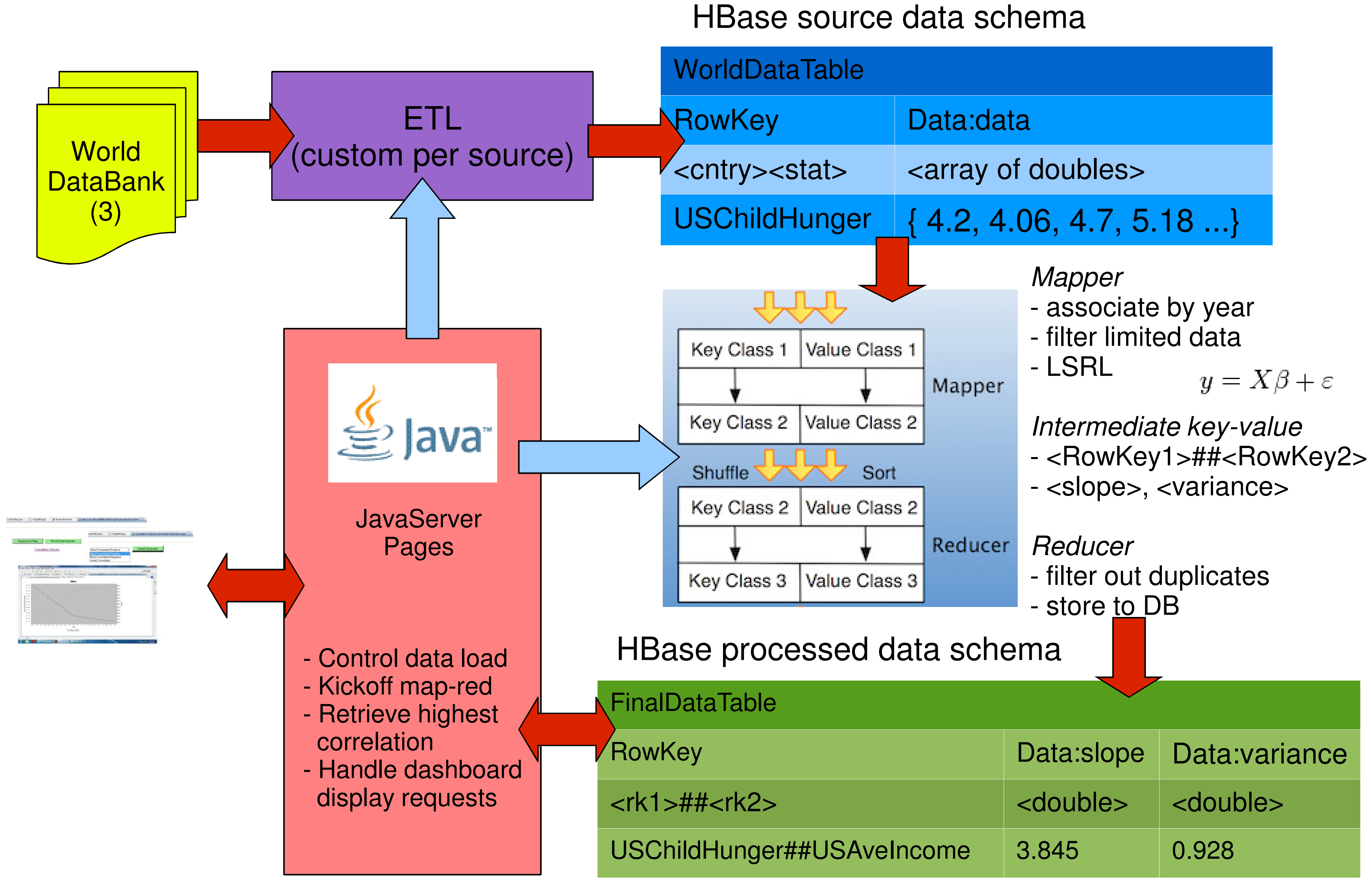
The RipVanWinkles
Daniel Honegger
Pallavi Sastry
Michael Shen

Purpose



- Automated correlation discovery over large, disparate data sources
  - Datasource: World Data Bank
- Distributed processing for speed and scalability
  - Hadoop with HBase
- Visualization dashboard
  - JfreeCharts, Java Server Pages

Architecture



World DataBank (3)

ETL (custom per source)

HBase source data schema

WorldDataTable	
RowKey	Data:data
<cntry><stat>	<array of doubles>
USChildHunger	{ 4.2, 4.06, 4.7, 5.18 ...}

Mapper
- associate by year
- filter limited data
- LSRL

$y = X\beta + \varepsilon$

Intermediate key-value
- <RowKey1>##<RowKey2>
- <slope>, <variance>

Reducer
- filter out duplicates
- store to DB

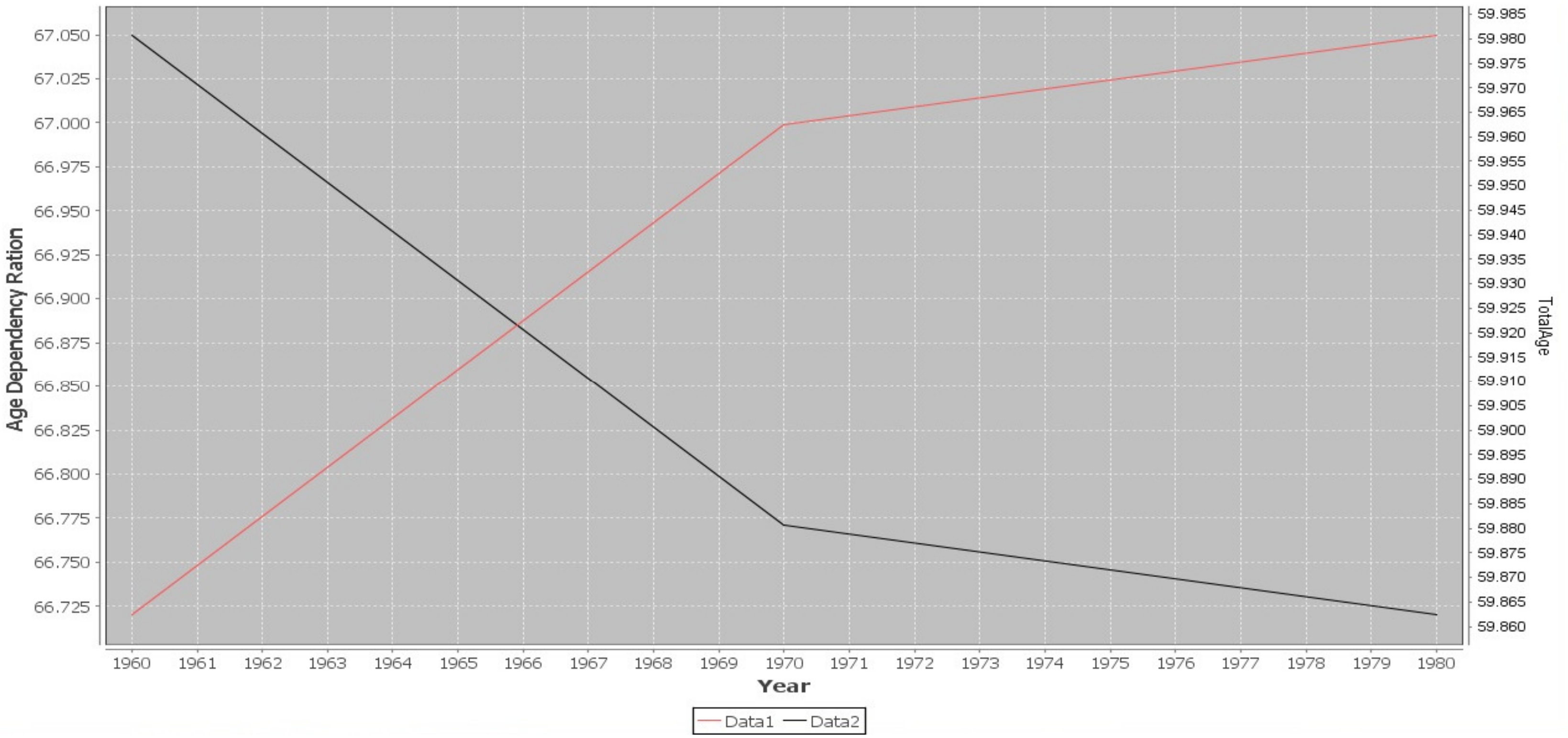
JavaServer Pages

- Control data load
- Kickoff map-red
- Retrieve highest correlation
- Handle dashboard display requests

HBase processed data schema

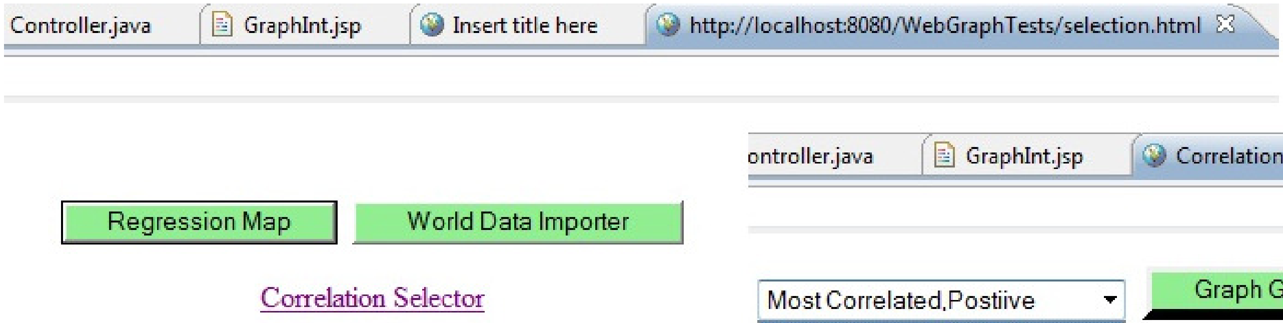
FinalDataTable		
RowKey	Data:slope	Data:variance
<rk1>##<rk2>	<double>	<double>
USChildHunger##USAveIncome	3.845	0.928

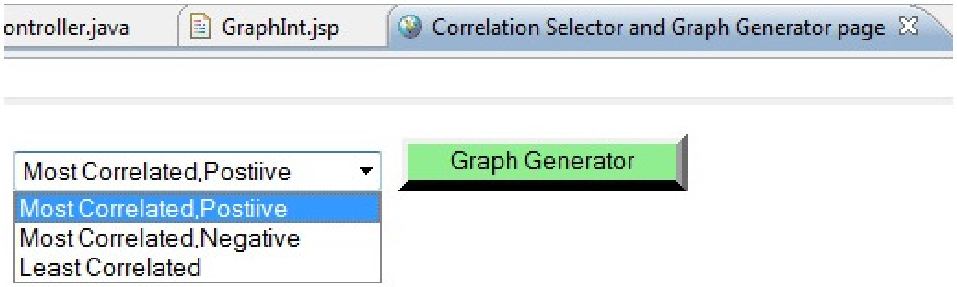
Demo

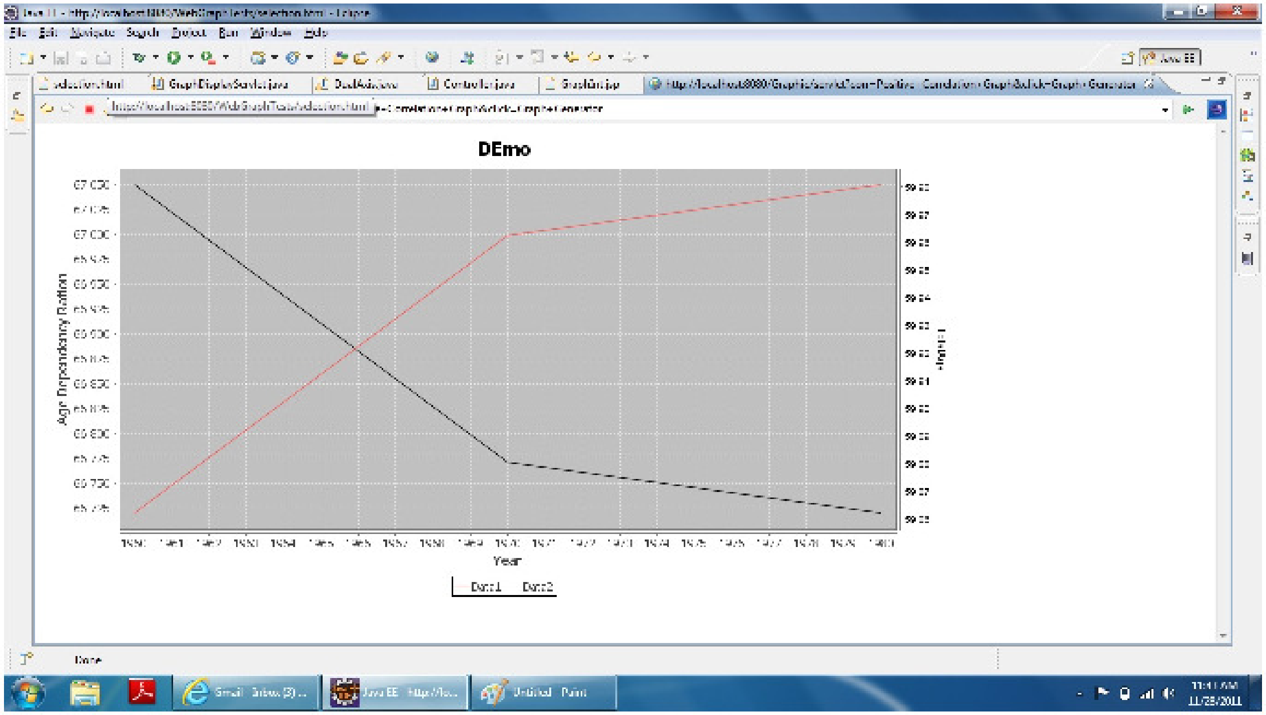


Demonstration of non-integrated solution

Integrated Demo (screenshot)

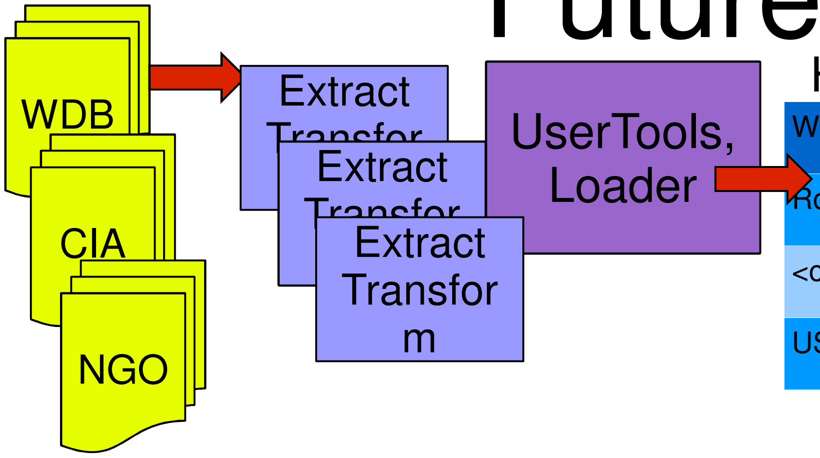






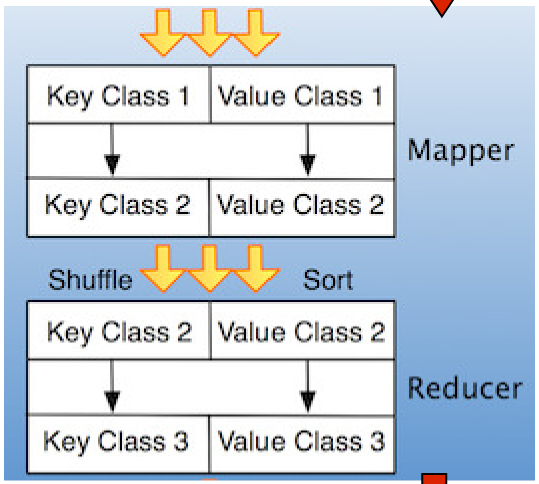
Integration issues didn’t allow us to get the full demo up and running

Future Development



HBase source data schema

WorldDataTable					
RowKey	Field:country	Field:cat1	Field:cat2	Field:stat	Data:data
<cntry><stat>	<cntry>	<type>	<type>	<fulltype>	<array of dbls>
USChildHunger	USA	hunger	child	hunger-age7	{ 4.2, 5.18 ...}



Mapper
- loosen association requirements
- allow temporal shifting, windowing
-filter on fields
- expand correlation type options

Intermediate key-value
- unique-keygen
- keys, time, fields, slope, variance

Reducer
- smart filters
-remove “junk”, “obvious” comparison based on fields

JavaServer Pages

- Direct control of filters
- Custom field filtering designed by tool
- More open browse
- Access to base data

HBase processed data schema

FinalDataTable							
UUID	Field: country	Field:cat1	Field:cat2	Field:stat	Field:time	Data:slope	Data:R2
uuid	<c1,c2>	<t1,t2>	<t1,t2>	<ft1,ft2>	<T,T>	<double>	<double>
7290347	USA,USA	hunger,obesity	child,child	H-a7 vs.O-a8	61-69, 65-73	-2.46	0.928