La ricerca si fa impresa



Filippo Biscarini

PhD, Statistical geneticist, Dipartimento di bioinformatica PTP - Lodi

A graph database to store and manage phenotypic, pedigree and genotypic data of livestock

EAAP 2013 – Nantes (France)



Premise:

One of the coauthors left (the main expert ... argh!)

"Salvage-whatever-possible"

- ✤ ☺ tell you what this is about
- ✤ ☺ tell you how it basically works
- ✤ ☺ tell you what we have so far set up
- Solution won't be able to show any comparison in performance (it was planned but ...)
- Doesn't mean this can't be taken up again (maybe someone is interested!)



Text files & spreadsheets

Parco Tecnologico Padano

Strumenti per costruire valore

A1	Calibri	* 12								
		· 12			1 % 00	.00) 0+	🗄 • 🔳	• 놀 • 🔳		
1	v	$f_{iii} \Sigma = \Box$	sample							
	A	В	С	D	E	F	G	н	-	J
	ID sample	Name	Phenotype		SNP1			SNP2		
z				vv	vf	ff	vv	vf	ff	w
3	1	LOW	Low root vigor	0	0	1	0	1	0	
1	2	LOW	Low root vigor	0	0	1	0	1	0	
5 7 3 9 0	3	LOW	Low root vigor	0	0	1	0	1	0	
5	4	LOW	Low root vigor	0	0	1	0	1	0	
	5	LOW	Low root vigor	0	0	1	0	1	0	
	6	LOW	Low root vigor	0	0	1	0	1	0	
)	7	LOW	Low root vigor	0	0	1	0	1	0	
0	8	LOW	Low root vigor	0	0	1	0	1	0	
1	1	0DM78	Low root vigor	1	0	0	0	1	0	
2	2	0DM78	Low root vigor	1	0	0	0	1	0	
3	3	0DM78	Low root vigor	1	0	0	1	0	0	
2 3 4	4	0DM78	Low root vigor	1	0	0	0	1	0	
5	5	0DM78	Low root vigor	1	0	0	0	1	0	
6	6	0DM78	Low root vigor	1	0	0	0	1	0	
7	7	0DM78	Low root vigor	1	0	0	0	1	0	
8	8	0DM78	Low root vigor	1	0	0				
9	1	7A	Low root vigor	1	0	0	0	0	1	
)	2	7A	Low root vigor	1	0	0	0	0	1	
1	3	7A	Low root vigor	1	0	0	0	0	1	
1	4	7A	Low root vigor	1	0	0	0	0	1	
3	5	7A	Low root vigor	1	0	0	0	0	1	
4	6	7A	Low root vigor	1	0	0	0	0	1	
	7	7.6	Low root vigor	1	0	0	0	0	1	

																																									-				
		SNP_IGA_112																																											
M	1.	SNP_IGA_200	5 A	С	А	С	А	C	A (C A	C	А	С	А	C /	4 (C /	A C	A	C	Α	С	А	С	А	C J	A (C A	A C	Α	С	AC	: A	С	А	С	А	C	AC	: A	L C	A	C /	A C	1
M	1	SNP_IGA_205	5 C	Α	С	А	С	А	c /	A C	A	С	Α	С	A (: A	4 (C A	C	A	С	Α	С	Α	C	A	C A	۹ (: A	С	Α	C A	۱C	Α	С	Α	C	Α	C A	C	: A	C	A (C A	۱.
M	1	SNP_IGA_227	2 G	Α	G	Α	G	Α	G /	A C	A	G	Α	G	A (G A	A (G A	G	A	G	Α	G	А	G	A	G A	4 0	G A	G	Α	G A	G	Α	G	Α	G	Α	G A	G	A	G	A	G A	<u>،</u>
M	1	SNP_IGA_233	1 G	Α	G	А	G	А	G /	A C	A	G	Α	G	A (i A	A (G A	G	A	G	Α	G	А	G	A	G A	A (ΓA	G	А	G A	G	Α	G	Α	G	А	G A	G	A	G	A	G A	<u>،</u>
M	1	SNP IGA 265	1 A	G	Α	G	А	G	A (G A	G	Α	G	Α	G /	4 0	5 A	A G	A	G	Α	G	А	G	Α	G	A (S A	۹ G	Α	G	AC	; A	G	Α	G	A	G	AC	A	G	A	G /	A C	i .
м	1	SNP IGA 371	7 C	Α	с	Α	с	A	c /	A C	A	с	Α	с	A	: A	4 0	C A	c	A	с	Α	С	Α	с	A	C A	4 0	C A	С	Α	C A	۱ c	Α	с	Α	с	Α	C A	C	: A	С	A	C A	`
М	1	SNP IGA 377	ЭА	G	Α	G	А	G	A	G A	G	Α	G	А	G /	A (G A	A G	A	G	Α	G	А	G	А	G	A	G A	A G	Α	G	A	i A	G	Α	G	A	G	A	A	G	A	G /	A C	
		SNP IGA 380																																											
м	1	SNP IGA 380	5 G	Α	G	Α	G	Α	G /	A (A	G	Α	G	A	5 A	4 0	G A	G	A	G	Α	G	Α	G	A	G A	4 0	G A	G	Α	G A	G	Α	G	Α	G	Α	G A	G	A	G	A	G A	`
М	1	SNP IGA 385	9 G	Α	G	А	G	А	G /	A C	A	G	А	G	A	i A	A (G A	G	A	G	Α	G	А	G	A	G A	4 0	G A	G	Α	G A	G	Α	G	А	G	А	G A	G	A	G	A (G A	.
М	1	SNP IGA 386	5 G	Α	G	А	G	А	G /	A C	A	G	А	G	A (i A	4 0	5 A	G	A	G	А	G	А	G	A	G A	4 0	5 A	G	А	G A	G	Α	G	А	G	А	G A	G	i A	G	A (G A	۱.
M	1	SNP_IGA_454	4 G	Α	G	Α	G	Α	G /	A (A	G	А	G	A	G A	4 (G A	G	A	G	Α	G	А	G	A	G A	4 0	5 A	G	Α	G A	G	Α	G	Α	G	Α	G A	G	A	G	A	G A	`
M	1	SNP_IGA_480	ΒA	G	Α	G	Α	G	A	G A	G	Α	G	Α	G /	4 0	G A	A G	A	G	Α	G	Α	G	Α	G	A (G A	A G	Α	G	AC	i A	G	Α	G	А	G	A	A	G	A	G /	A C	i
M	1	SNP_IGA_490	5 G	Α	G	А	G	А	G /	A C	A	G	А	G	A (i A	A (G A	G	A	G	А	G	А	G	A	G A	A (5 A	G	А	G A	G	Α	G	А	G	А	G A	G	A	G	A	G A	.
M	1	SNP_IGA_525	B G	Α	G	А	G	Α	G /	A C	A	G	Α	G	A (5 A	4 0	G Α	G	A	G	Α	G	Α	G	A	G A	4 0	G A	G	Α	G A	G	Α	G	Α	G	Α	G A	G	i A	G	A (G A	۱.
M	1	SNP_IGA_527	B C	Α	С	Α	С	Α	C /	A C	A	С	Α	С	A (C A	A (C A	C	A	С	Α	С	Α	C	A	C A	4 0	C A	С	Α	C A	I C	Α	С	Α	C	Α	C A	C	: A	C	A	C A	۱.
M	1	SNP_IGA_567	5 G	Α	G	А	G	А	G /	A C	A	G	А	G	A (i A	A (G A	G	A	G	Α	G	А	G	A	G A	A (G A	G	Α	G A	G	Α	G	А	G	Α	G A	G	i A	G	A (G A	۱.
M	1	SNP IGA 624	7 G	Α	G	А	G	A	G /	A C	A	G	Α	G	A (5 A	4 (5 A	G	A	G	Α	G	А	G	A	G A	4 0	5 A	G	Α	G A	G	Α	G	Α	G	Α	G A	G	A	G	A (G A	`
M	1	SNP_IGA_649	3 C	Α	С	Α	С	Α	C /	A C	A	С	Α	С	A	C A	4 (C A	C	A	С	Α	С	Α	C	A	C A	4 0	C A	С	Α	C A	۱C	Α	С	Α	C	Α	C A	C	: A	С	A (C A	۱.
M	1	SNP_IGA_661	5 G	Α	G	А	G	А	G /	A C	A	G	Α	G	A (i A	A (G A	G	A	G	Α	G	А	G	A	G A	4 0	i A	G	Α	G A	G	Α	G	Α	G	А	G A	G	A	G	A	G A	<u>،</u>
М	1	SNP_IGA_662	ЭС	С	С	С	С	С	С	СС	C	С	С	С	C (c (сс	C	C	С	С	С	С	С	C (С	c (C	С	С	С	C	С	С	С	С	С	с	c	C	C	C (СС	1
M	1	SNP_IGA_663	3 G	Α	G	Α	G	Α	G /	4 0	A	G	Α	G	A (G A	4 (GΑ	G	A	G	Α	G	Α	G	A	G A	4 0	G A	G	Α	G A	G	Α	G	Α	G	Α	G A	G	A	G	A	G A	۱.
M	1	SNP IGA 679	5 G	Α	G	А	G	Α	G A	A C	A	G	Α	G	A	i A	A (G A	G	A	G	Α	G	А	G	A	G A	4 0	ΓA	G	Α	G A	G	Α	G	А	G	А	G A	G	A	G	A (G A	۱.
М	1	SNP IGA 697	ZA	G	А	G	А	G	A (G A	G	А	G	А	G /	4 0	5 A	A G	A	G	Α	G	А	G	А	G	A	S A	A G	А	G	AC	i A	G	А	G	А	G	AC	i A	G	A	G /	A C	í i
		SNP IGA 706																																											

country	cows	Pedigree	records
Germany	~ 4 million	~ 20 million	> 300 million

- parse each time (e.g. by chromosome, subset of animals etc ...)
- less stable and less safe
- increasingly impractical with big data (e.g. HD SNP-chips, whole-genome sequences etc ...)





Relational databases

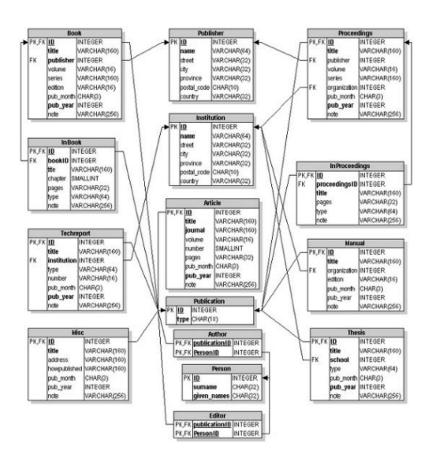
Strumenti per costruire valore

- Relational databases
 - Tables
 - Relations (between tables)
- SQL (queries instead of writing a script)

SELECT snp_genotype FROM
genotypes_table WHERE
chromosome=1;

1. more practical

- 2. faster
- 3. safer and more stable



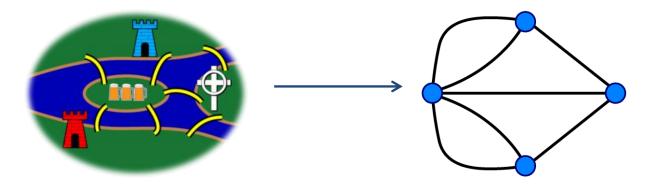


Graph theory



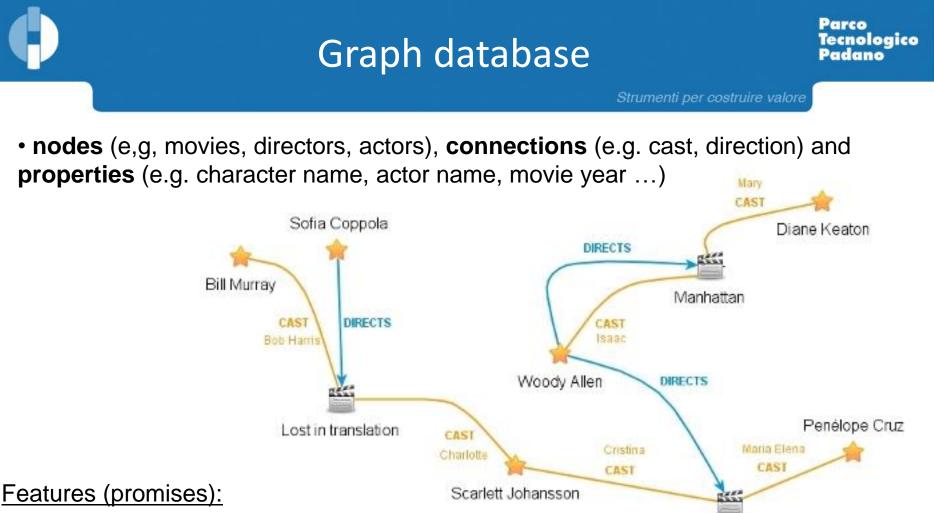
• Euler and the 7 bridges of Königsberg

• *"Solutio problematis ad geometriam situs pertinentis".* Commentarii Academiae Scientiarum Imperialis Petropolitanae 8 (1736) 128-140.



- Graphs: nodes and connections
- common models for natural and human structures
 - computer networks
 - molecules (atoms and chemical bonds)
 - habitats and migration paths (breeding patterns, spread of diseases or parasites ...)





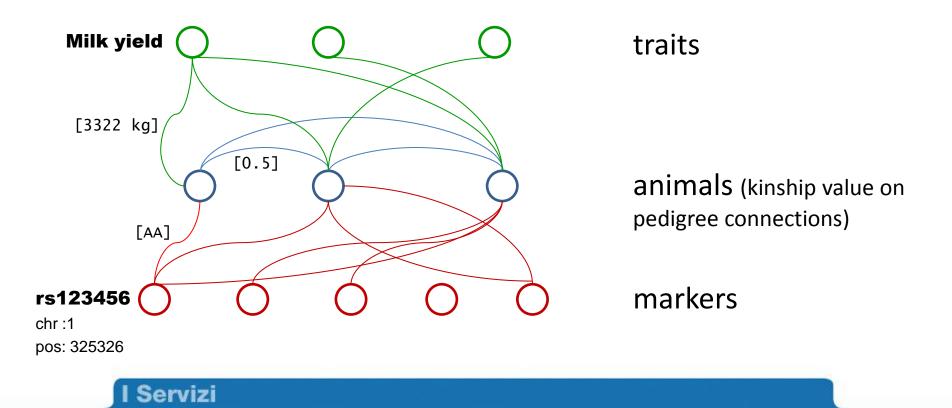
Vicky Cristina Barcelona

- no indexes
- suited for associative data
- faster (at least for graph-like queries)
- naturally translates into object-oriented programming
- scales better to large datasets
- accommodates better changing data

Our graph database

Working example

- animal dataset (buffaloes): phenotypes, genotypes and pedigree
- 3 layers: traits, animals, markers



Strumenti per costruire valore

- Graph DBs are still an active field of research
- Not yet a standard query language (like SQL for relational databases)

Setting up the graph database

- Lack of mature commercial products and user-friendly interfaces
- Several different ongoing projects
 - neo4j (<u>http://www.neo4j.org/</u>)
 - Ruby (create and populate the graph database)
 - Json (format for data interchange -associative array or hash)





Web Interface

Strumenti per costruire valore

l7 des		179 properties	206 relationships Year	One month	rela One week	7 ationship type One day	es 6 hours	30 minutes
			Year	One month	One week	One day	6 hours	30 minutes
ides operties lationships								
	telationships	1						





Strumenti per costruire valore

Ruby gem to traverse the graph

```
n1 = Neography::Node.load(15)
```

q1= n1.outgoing(:contain_maker).depth(2).include_start_node

#<Neography::NodeTraverser:0x686baa51 @order="depth first", @filter={"language"=>"builtin",
"name"=>"all"}, @relationships=[{"type"=>"contain_maker", "direction"=>"out"}], @depth=2,
@uniqueness="none", @from=#<Neography::Node position="85031448", name="AX85040742", chromosome="15">>

[parse the hash with Ruby]

Marker name: AX-85040742

Marker chromosome: 15

Marker position: 85031448







Strumenti per costruire valore

 declarative query language to search and update the graph (no need to traverse the graph structure writing a script)

- still growing and maturing.
- some keywords (e.g.WHERE, ORDER BY) are inspired by SQL

START animal=node:node_auto_index(ID = 'ITM123456789') MATCH animal[:related]->()-[:related]->related_to RETURN animal, related_to

START a=node(4) RETURN a

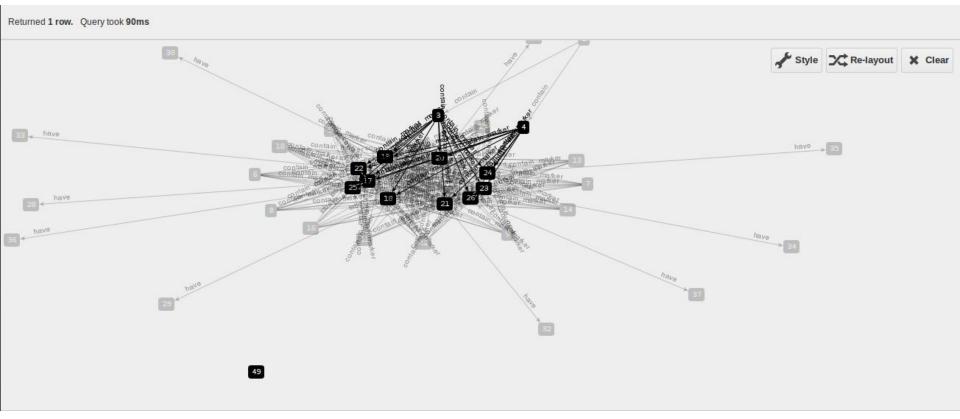
Node 4 http://127.0.0.1:7474/db/data/node/4	Show relations	ships Saved Delete
position	"54264005"	Remove
name	"AX-85040751"	Remove
chromosome	"13"	Remove
Add property		



Graph visualization



Strumenti per costruire valore





Conclusions

Strumenti per costruire valore

Breeding and genomics data are associative

Scaling is an issue (big data)

Graph databases may offer a convenient alternative to store and manage data in animal breeding and genomics

Relatively recent area of research

R&D needed before a "commercial" product can be obtained

Compare with relational DBs in terms of performance, ease of use etc ...

