### MyGene.info

- Making Elastic Gene API

Chunlei Wu, Ph.D.

The Scripps Research Institute La Jolla, CA, USA

> BOSC 2013 July 20, 2013

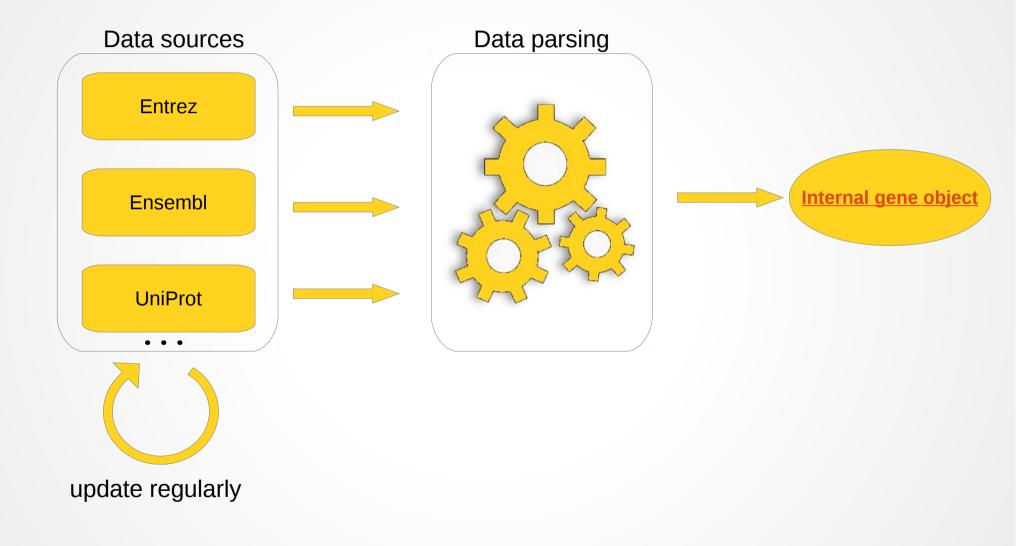




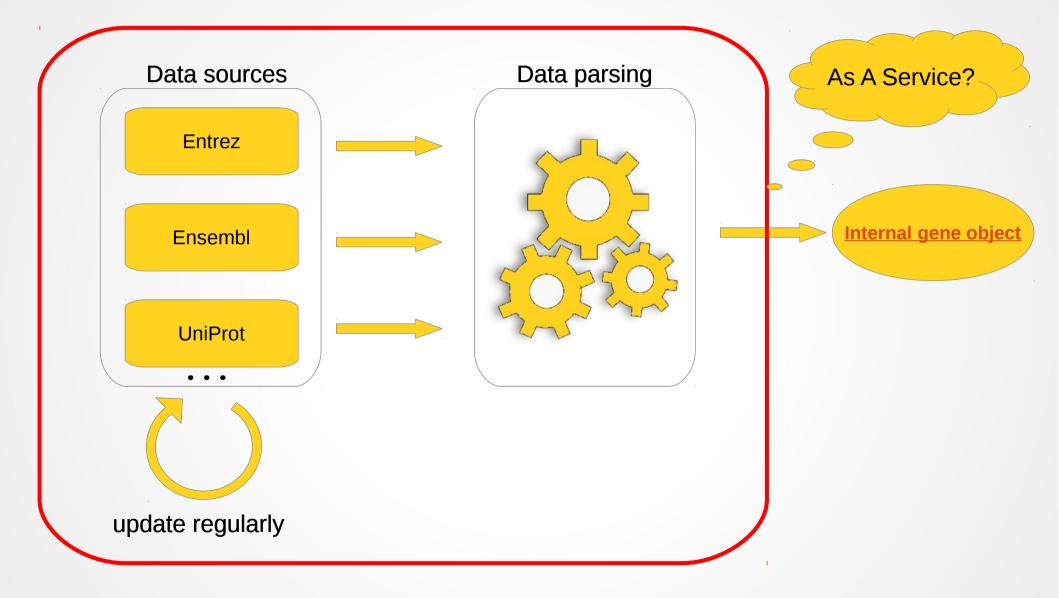
# Being "Elastic"

- Fast
- Always ON
- Up-to-date
- Scalable
- Extensible

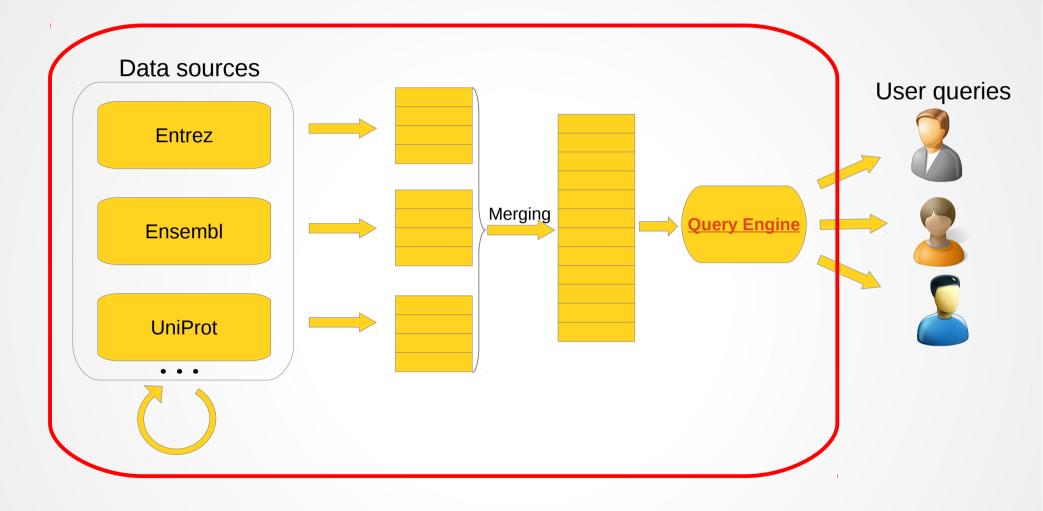
### Common procedure for gene data retrieval



### Common procedure for gene data retrieval

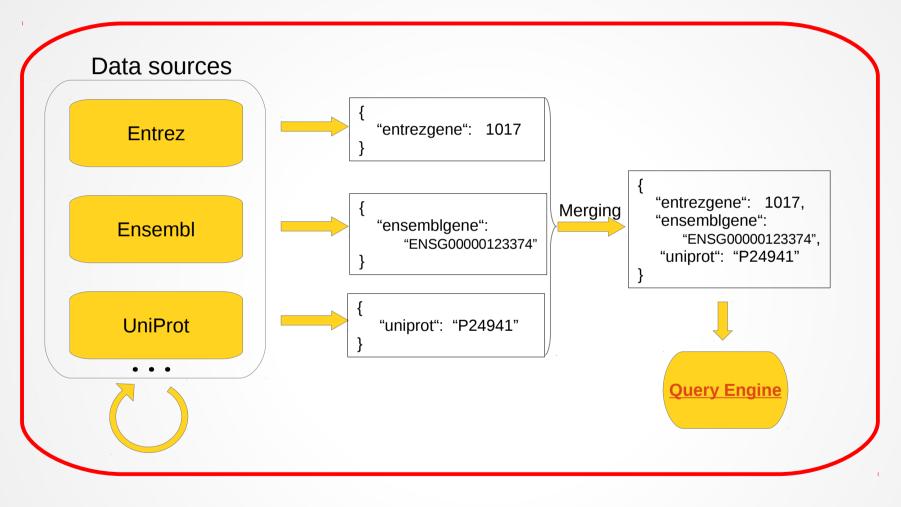


# Working model - 1



# Working model - 1

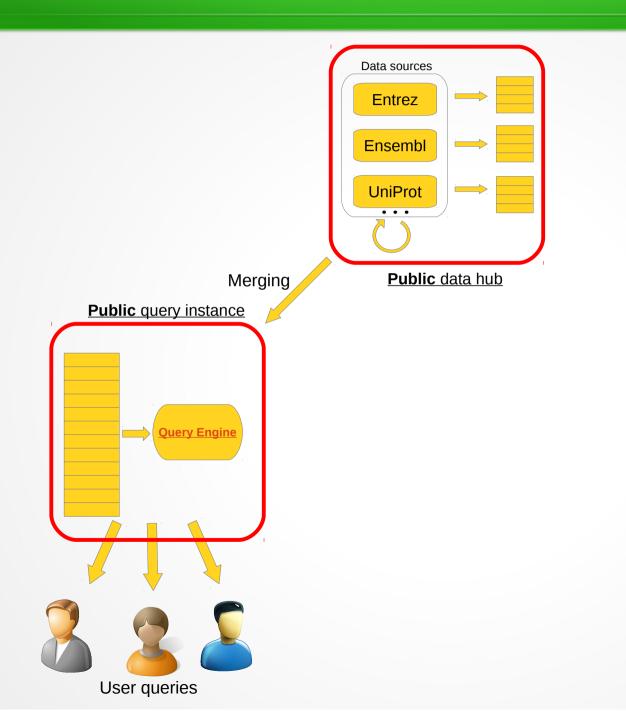
A dummy merging example:



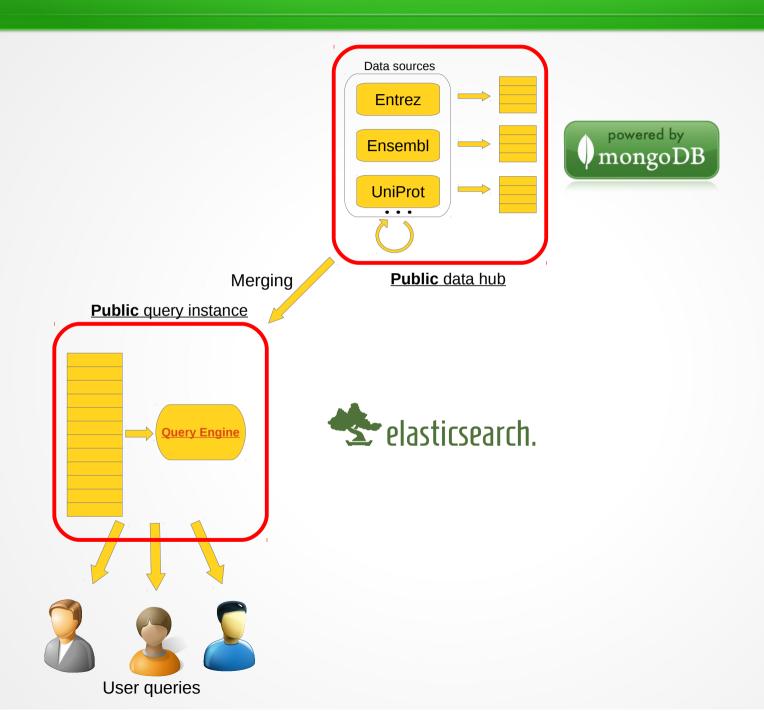
## Gene object in noSQL database



### Syncing from data-hub to query instance



### Syncing from data-hub to query instances



### http://MyGene.info

(currently v2 API, two endpoints)



### http://MyGene.info

- Support ALL species, from NCBI (>12K species, >13M genes)
- >40 annotation fields and expanding
- Weekly-updated
- Flexible query interface
  - Simple queries
  - Fielded queries
  - Wildcard queries
  - Genomic interval queries
  - Species filter
  - Returning fields filter
- Support batch queries, JSONP, CORS
- Committed for long-term availability

?q=cdk2
?q=symbol:cdk2
?q=cdk\*
?q=cdk1:1-100,000&species=human
?q=cdk2&species=mouse,rat
?q=cdk2&fields=symbol,homologene

### http://MyGene.info

#### High-performance host (serving ~500K requests/day)

Google Analytics					reillywu@gmail.com Settin	gs MyA	ccount Si	ign out
M S mygene.info - http://mygene.ir mygene.info [DEFAULT]	nfo	- Reporting	Customizatio	n		Adm	in	Help
Q, Find reports & more	Overview					Cr	reate Shor	tcut BET
MY STUFF		Distriction		Pa	igeviews			
Dashboards	_	Right now		Pe	er minute Per second			
Shortcuts	, L	5082		1,5	00 30			
Intelligence Events		active visitors on site	•	1,0	00 <b></b>			-
STANDARD REPORTS	NEW					ililiti	ahilip	ahihik
🕐 Real-Time		100%		500				
📥 Audience					-25 min -20 min -15 min -10 min -5 min -5 min -60 cco	-60	899	-15 830
Overview	Top Referrals	:		Тор	Active Pages:			
▶ Demographics	Source	Active Visitor	rs ↓		Active Page	Act	tive Visitors	s ↓
▶ Behavior	T	here is no data for this view.		1.	1	2,478	48.75%	
▶ Technology	Tan Casial Ta	- 661		2.	/query	96	1.89%	
▶ Mobile	Top Social Tra			3.	/v2/gene/1017	31	0.61%	
▶ Custom	Source	Active Visitor		4.	/v2/gene/16691	9	0.18%	
Visitors Flow	Ir	here is no data for this view.		5.	/v2/gene/2729	6	0.12%	
Traffic Sources	Top Keywords	s:		6.	/v2/gene/72114	6	0.12%	
- Ourland	Keyword	Active Visite	ors 🗸	7.	/v2/gene/2810	5	0.10%	
Content	Th	here is no data for this view.		8.	/v2/gene/3809	5	0.10%	
Conversions				9.	/v2/gene/4312	5	0.10%	
				10.	/v2/gene/4318	5	0.10%	

### Third-party packages

### MyGene.py - Python wrapper

https://pypi.python.org/pypi/mygene

pip install mygene

```
In [1]: import mygene
In [2]: mg = mygene.MyGeneInfo()
In [3]: mg.getgene(1017)
Out[3]:
{' id': '1017', 'entrezgene': 1017, 'name': 'cyclin-dependent kinase 2',...}
In [4]: mg.guery('cdk2')
Out[4]:
{'hits': [{' id': '1017',
   ' score': 373.24667,
   'entrezgene': 1017,
   'name': 'cyclin-dependent kinase 2',
   'symbol': 'CDK2',
   'taxid': 9606},
   ...],
 'max score': 373.24667,
 'took': 10.
 'total': 28}
```

### Third-party packages

### MyGene.autocomplete

- Gene query autocomplete widget

https://bitbucket.org/sulab/mygene.autocomplete

### Third-party packages

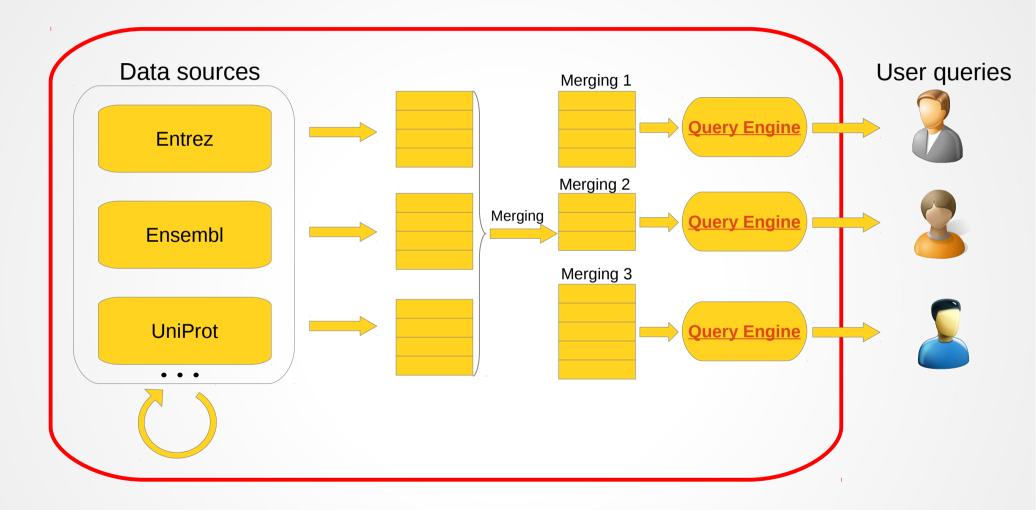
#### MyGene.autocomplete

- Gene query autocomplete widget

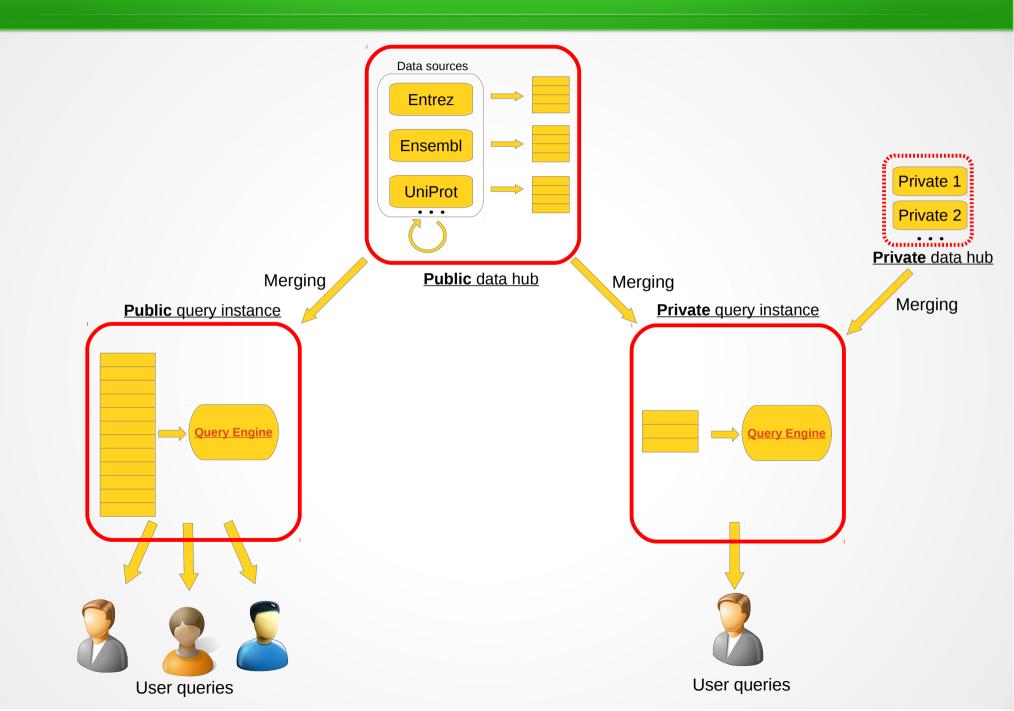
https://bitbucket.org/sulab/mygene.autocomplete

Enter a gene here:	cdk2						
	CDK2: cyclin-dependent kinase 2 (human)						
	Cdk2: cyclin-dependent kinase 2 (mouse)						
	Cdk2: cyclin dependent kinase 2 (rat)						
	CDK2AP1: cyclin-dependent kinase 2 associated protein 1 (human)						
	Cdk2ap1: CDK2 (cyclin-dependent kinase 2)-associated protein 1 (mouse)						
	Cdk2ap2: CDK2-associated	protein 2 (mouse)					
	Cdk2ap1-ps5: Cdk2ap1-ps	5 pseudogene (rat)					
	Cdk2ap1-ps1: Cdk2ap1-pseudogene 1 (rat)						
	Cdk2ap1-ps3: Cdk2ap1-ps3	3 pseudogene (rat)					
	Cdk2ap1-ps2: Cdk2ap1-ps2 pseudogene (rat)						

# Working model – 2



### Syncing from data-hub to query instances



## Private query instance

- Dedicated host
- Same powerful query interface
  - Third-party packages still work
- Public data still get sync-ed
- Allow to merge private data

### To reach us?

### Questions on public query instance or interested in setting up your own private query instance?

Please let us know: help@mygene.info

# Code repositories

• Web front-end

https://bitbucket.org/sulab/mygene.info Apache 2 licensed

• Data hub

https://bitbucket.org/sulab/mygene.hub GPL v3 licensed

## Acknowledgement

#### <u>Sulab</u>

#### Andrew Su

Benjamin Good Max Nannis Salvatore Loguercio Katie Fisch Tobias Meissner

