

BUILDING THE FUTURE OF DRUG DISCOVERY

JEFFREY REID [TWITTER: @JGREID]
REGENERON GENETICS CENTER
SPARK_AI 2018

REGENERON
SCIENCE TO MEDICINE®



BIOLOGY IS COMPLEX AND (MOSTLY) UNPREDICTABLE

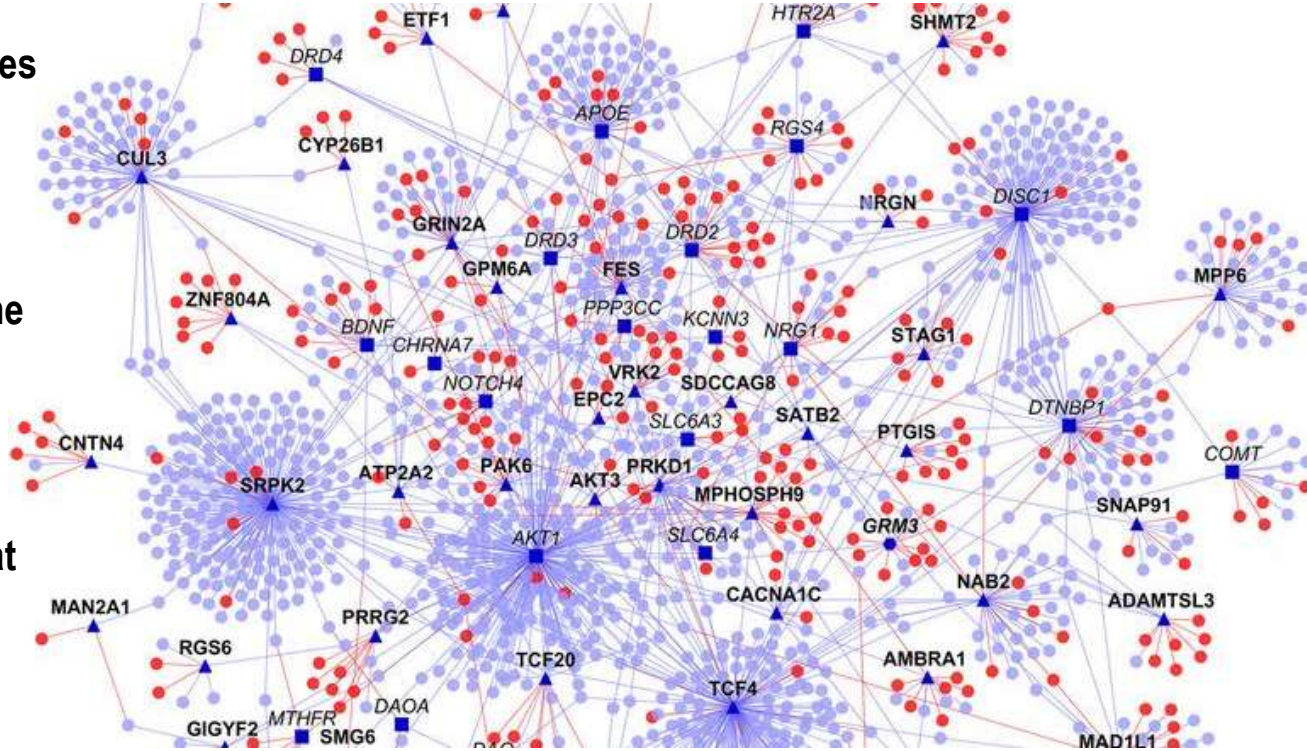
~20k human protein-coding genes

~400M possible PPIs

~2B seconds in a human lifetime

~30T cells in the avg human

...and we don't even know what
half the genes do!



BIOLOGY IS COMPLEX AND (MOSTLY) UNPREDICTABLE

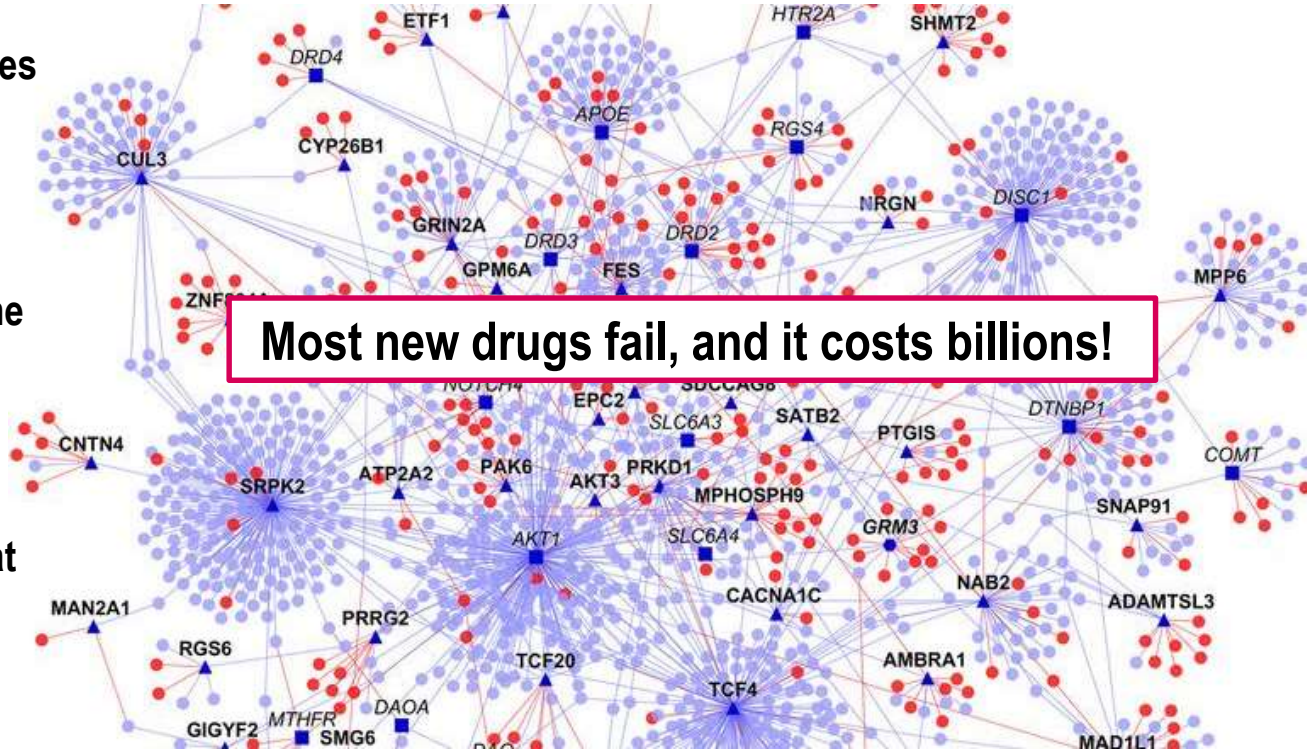
~20k human protein-coding genes

~400M possible PPIs

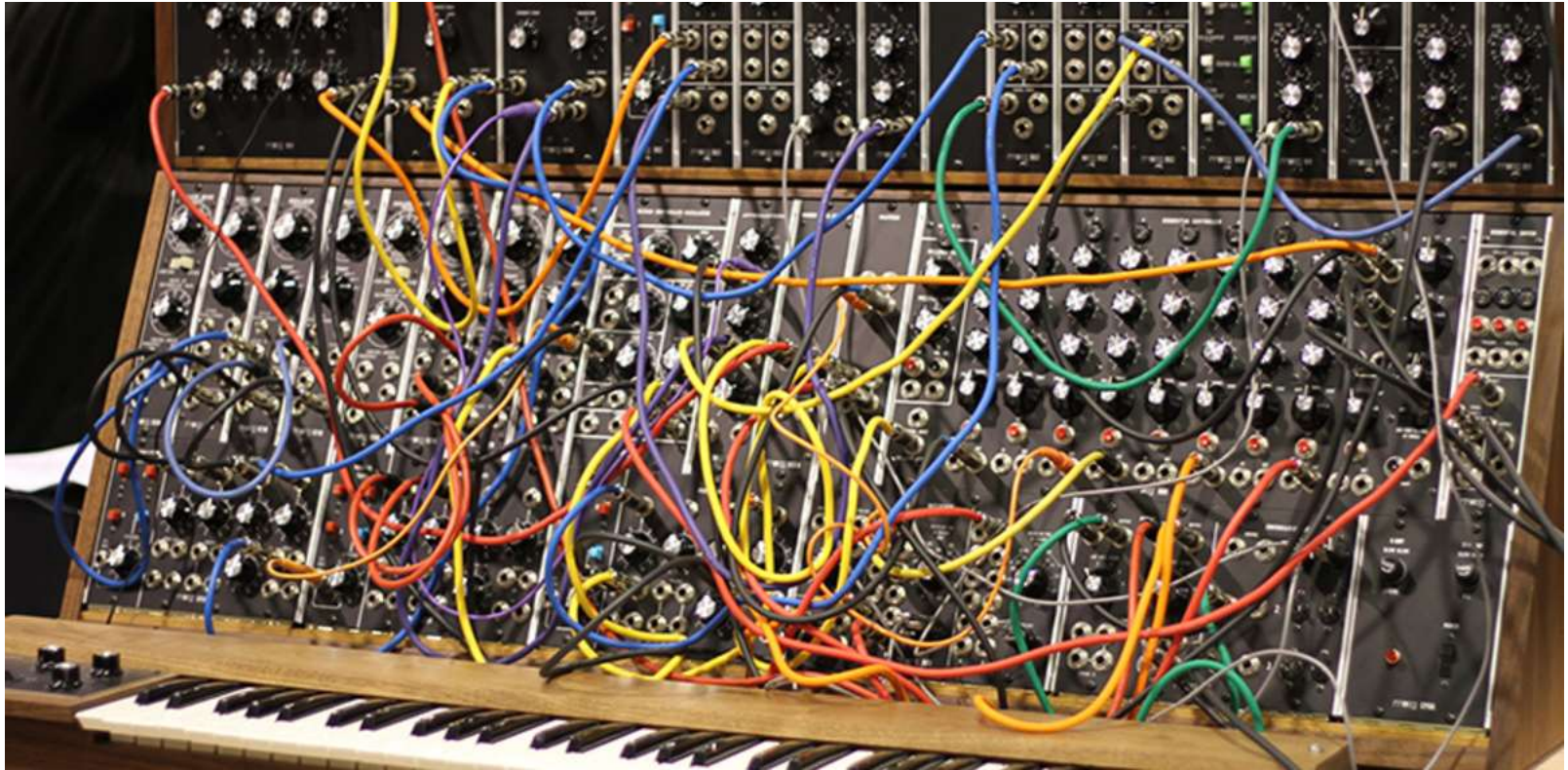
~2B seconds in a human lifetime

~30T cells in the avg human

...and we don't even know what
half the genes do!



LIKE AN ANALOG SYNTHESIZER... BUT FOR YOUR HEALTH



GENETICS PROVIDES A GLIMPSE AT THE INSTRUCTION MANUAL



“LOSS-OF-FUNCTION” VARIANTS CAN “KNOCK OUT” A PROTEIN



SOME MUTATIONS ARE 'PROTECTIVE'

THE NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

A Protein-Truncating *HSD17B13* Variant and Protection from Chronic Liver Disease

N.S. Abul-Husn, X. Cheng, A.H. Li, Y. Xin, C. Schurmann, P. Stevis, Y. Liu, J. Kozlitina, S. Stender, G.C. Wood, A.N. Stepanchick, M.D. Still, S. McCarthy, C. O'Dushlaine, J.S. Packer, S. Balasubramanian, N. Gosalia, D. Esopi, S.Y. Kim, S. Mukherjee, A.E. Lopez, E.D. Fuller, J. Penn, X. Chu, J.Z. Luo, U.L. Mirshahi, D.J. Carey, C.D. Still, M.D. Feldman, A. Small, S.M. Damrauer, D.J. Rader, B. Zambrowicz, W. Olson, A.J. Murphy, I.B. Borecki, A.R. Shuldiner, J.G. Reid, J.D. Overton, G.D. Yancopoulos, H.H. Hobbs, J.C. Cohen, O. Gottesman, T.M. Teslovich, A. Baras, T. Mirshahi, J. Gromada, and F.E. Dewey

SOME MUTATIONS ARE 'PROTECTIVE'

THE NEW ENGLAND JOURNAL OF MEDICINE

ORIGINAL ARTICLE

A Protein-Truncating *HSD17B13* Variant and Protection from Chronic Liver Disease

N.S. Abul-Husn, X. Cheng, A.H. Li, Y. Xin, C. Schurmann, P. Stevis, Y. Liu, J. Kozlitina, S. Stender, G.C. Wood, A.N. Stepanchick, M.D. Still, S. McCarthy, C. O'Dushlaine, J.S. Packer, S. Balasubramanian, N. Gosalia, D. Esopi, S.Y. Kim, S. Mukherjee, A.E. Lopez, E.D. Fuller, J. Penn, X. Chu, J.Z. Luo, U.L. Mirshahi, D.J. Carey, C.D. Still, M.D. Feldman, A. Small, S.M. Damrauer, D.J. Rader, B. Zambrowicz, W. Olson, A.J. Murphy, I.B. Borecki, A.R. Shuldiner, J.G. Reid, J.D. Overton, G.D. Yancopoulos, H.H. Hobbs, J.C. Cohen, O. Gottesman, T.M. Teslovich, A. Baras, T. Mirshahi, J. Gromada, and F.E. Dewey

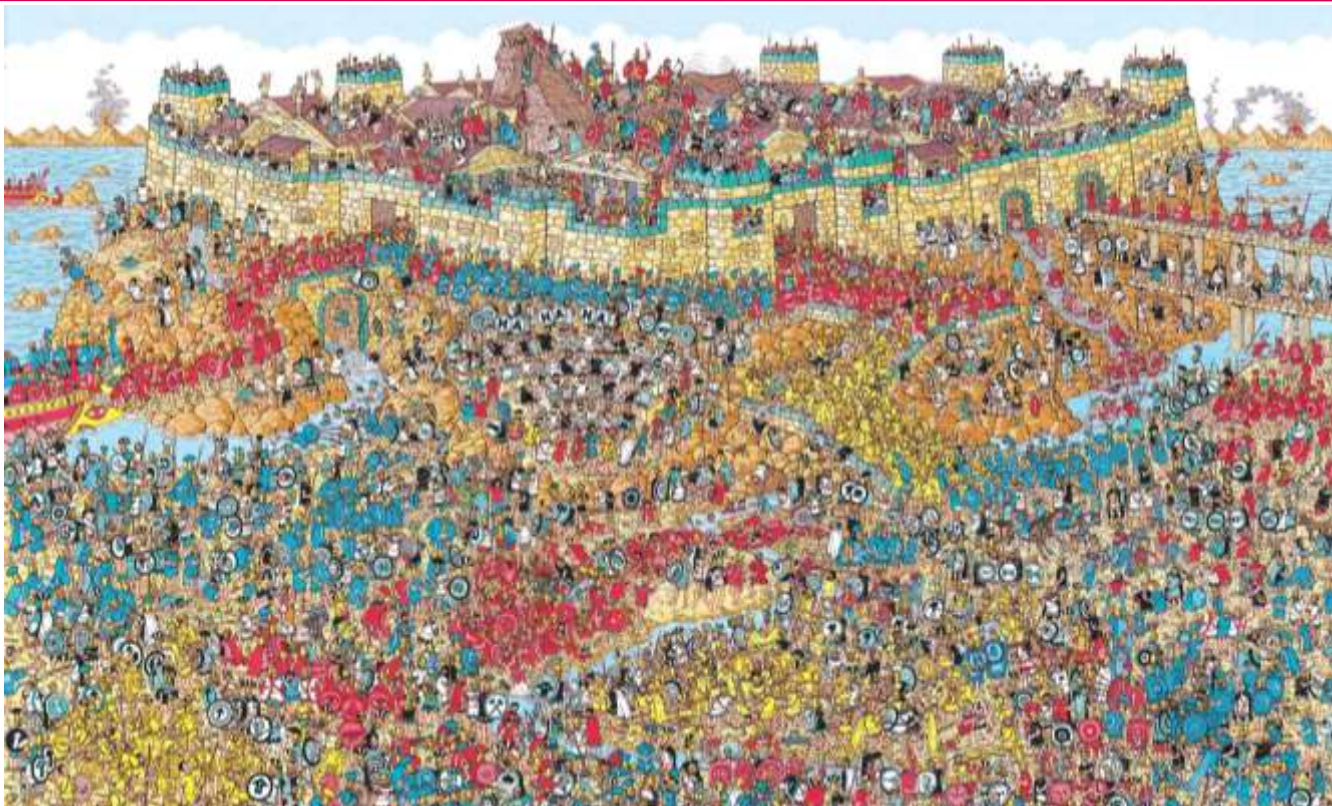
REGENERON



BUT HOW TO FIND ALL THE MUTANTS???



EXHAUSTIVE SEARCH SOLUTION TO 'WHERE'S WALDO (THE LOF CARRIER)'



+

biobank^{uk}
Improving the health of future generations

Electronic Health
Record data
On 500k people

AUTOMATION IS ESSENTIAL TO LARGE-SCALE DATA GENERATION



A UNIFIED, CENTRALIZED DATA & ANALYSIS RESOURCE

The screenshot displays the Databricks SparkSummitDemo workspace interface. On the left is a vertical sidebar with navigation icons for Databricks, Home, Workspace, Recent, Data, and Clusters. The main area shows a notebook with two commands:

Cmd 1

Load the association results and common functions for manipulating this data

```
%run "/Groups/GI R&D/dev/Includes/RB_tools_v1.2" ...
```

Cmd 2

Count the number of rows in the entire results data set

```
1 RB_unfiltered.count
```

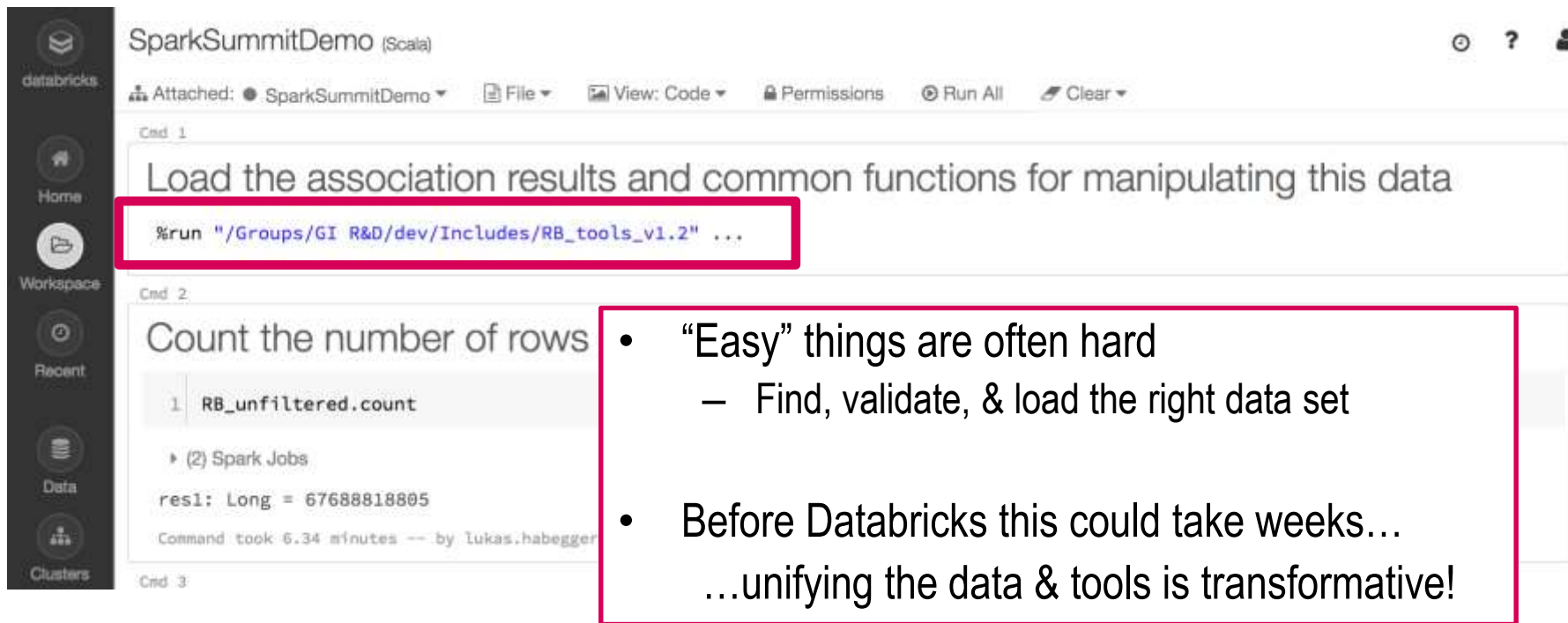
► (2) Spark Jobs

res1: Long = 67688818805

Command took 6.34 minutes -- by lukas.habegger@regeneron.com at 5/24/2018, 1:47:24 PM on SparkSummitDemo

Cmd 3

A UNIFIED, CENTRALIZED DATA & ANALYSIS RESOURCE



The screenshot shows the Databricks SparkSummitDemo workspace. The left sidebar contains navigation icons for Databricks, Home, Workspace, Recent, Data, and Clusters. The main area displays a code editor with the following content:

Cmd 1

Load the association results and common functions for manipulating this data

```
%run "/Groups/GI R&D/dev/Includes/RB_tools_v1.2" ...
```

Cmd 2

Count the number of rows

```
1 RB_unfiltered.count
```

▸ (2) Spark Jobs

res1: Long = 67688818805

Command took 6.34 minutes -- by lukas.habegger

Cmd 3


On the right, a pink-bordered box contains the following text:

- “Easy” things are often hard
 - Find, validate, & load the right data set
- Before Databricks this could take weeks...
...unifying the data & tools is transformative!

ENABLING STANDARD ANALYSES IN ONE LINE

Search the association results for "HSD17B13" in one line

```
1 val HSD17B13_results = RB_filtered.searchByGene("HSD17B13")
2   .filter($"cohort" like "GHS%")
3   .orderBy($"pvalue")
4   .reformat
5   .fullHierarchy
```

▶  HSD17B13_results: org.apache.spark.sql.DataFrame = [traitName: string, phenotypeName: string ... 20 more fields]
HSD17B13_results: org.apache.spark.sql.DataFrame = [traitName: string, phenotypeName: string ... 20 more fields]
Command took 0.80 seconds -- by lukas.habegger@regeneron.com at 5/31/2018, 4:30:35 PM on SparkSummitDemo

Cmd 6

Count the number of association results for "HSD17B13"

```
1 HSD17B13_results.count
```

▶ (2) Spark Jobs

res2: Long = 3774

Command took 4.84 seconds -- by lukas.habegger@regeneron.com at 5/31/2018, 4:30:56 PM on SparkSummitDemo

- Once the data and tools are in one place...
... "hard" things become easy.

AGGREGATING & MINING DATA IN NEW AND EXCITING WAYS

databricks

Home

Workspace

Recent

Data

Clusters

Jobs

Search

SparkSummitDemo (Scala)

Attached: SparkSummitDemo

File View: Code Permissions Run All Clear

Schedule Comments Revision history

Display the results for liver disease traits and the mutation of interest

```
1 display(  
2   HSD17B13_GHS_results  
3   .where("traitClass IN ('Liver measurements','Diseases of liver') AND mutation = '4:87310240:T:TA')  
4   .select("mutation", "trait", "traitClass", "geneName", "pValue", "effect")  
5   .orderBy("pValue")  
6 )
```

(2) Spark Jobs

mutation	trait	traitClass	geneName	pValue	effect
4:87310240:T:TA	Aspartate Aminotransferase (AST)	Liver measurements	HSD17B13	1.318e-17	decreased
4:87310240:T:TA	Alanine Aminotransferase (ALT)	Liver measurements	HSD17B13	7.005e-14	decreased
4:87310240:T:TA	ICD10 3D: Alcoholic liver disease	Diseases of liver	HSD17B13	0.000001252	protective
4:87310240:T:TA	ICD10 4D: Alcoholic cirrhosis of liver	Diseases of liver	HSD17B13	0.000001632	protective
4:87310240:T:TA	ICD9 4D: Alcoholic cirrhosis of liver	Diseases of liver	HSD17B13	0.00000176	protective
4:87310240:T:TA	Alcoholic Liver Disease RGC Composite Definition	Diseases of liver	HSD17B13	0.000001782	protective
4:87310240:T:TA	Liver Cirrhosis RGC Composite Definition	Diseases of liver	HSD17B13	0.00002237	protective

Command took 3.35 seconds -- by lukas.habegger@regeneron.com at 5/31/2018, 4:37:20 PM on SparkSummitDemo

AGGREGATING & MINING DATA IN NEW AND EXCITING WAYS

databricks

Home

Workspace

Recent

Data

Clusters

Jobs

Search

SparkSummitDemo (Scala)

Attached: SparkSummitDemo File View: Code Permissions Run All Clear

Display the results for liver disease traits and the mutation of interest

```
1 display(  
2   HSD17B13_GHS_results  
3   .where("traitClass IN ('Liver measurements','Diseases of liver') AND mutation = '4:87310240:T:TA')  
4   .select("mutation", "trait", "traitClass", "geneName", "pValue", "effect")  
5   .orderBy("pValue")  
6 )
```

(2) Spark Jobs

mutation	trait	traitClass	geneName	pValue	effect
4:87310240:T:TA	Aspartate Aminotransferase (AST)	Liver measurements	HSD17B13	1.318e-17	decreased
4:87310240:T:TA	Alanine Aminotransferase (ALT)	Liver measurements	HSD17B13	7.005e-14	decreased
4:87310240:T:TA	ICD10 3D: Alcoholic liver disease	Diseases of liver	HSD17B13	0.000001252	protective
4:87310240:T:TA	ICD10 4D: Alcoholic cirrhosis of liver	Diseases of liver	HSD17B13	0.000001632	protective
4:87310240:T:TA	ICD9 4D: Alcoholic cirrhosis of liver	Diseases of liver	HSD17B13	0.00000176	protective
4:87310240:T:TA	Alcoholic Liver Disease RGC Composite Definition	Diseases of liver	HSD17B13	0.000001782	protective
4:87310240:T:TA	Liver Cirrhosis RGC Composite Definition	Diseases of liver	HSD17B13	0.00002237	protective

Command took 3.35 seconds -- by lukas.habegger@regeneron.com at 5/31/2018, 4:37:20 PM on SparkSummitDemo



...AND ITS FAST!

SparkSummitDemo (Scala)

Attached: SparkSummitDemo File View: Code Permissions Run All Clear Schedule Comments Revision history

Display the results for liver disease traits and the mutation of interest

```
1 display(  
2   HSD17B13_GHS_results  
3   .where("traitClass IN ('Liver measurements','Diseases of liver') AND mutation = '4:87310240:T:TA')  
4   .select("mutation", "trait", "traitClass", "geneName", "pValue", "effect")  
5   .orderBy("pValue")  
6 )
```

(2) Spark Jobs

mutation	trait	traitClass	geneName	pValue	effect
4:87310240:T:TA	Aspartate Aminotransferase (AST)	Liver measurements	HSD17B13	1.318e-17	decreased
4:87310240:T:TA	Alanine Aminotransferase (ALT)	Liver measurements	HSD17B13	7.005e-14	decreased
4:87310240:T:TA	ICD10 3D: Alcoholic liver disease	Diseases of liver	HSD17B13	0.000001252	protective

- Previously this was manual spreadsheet “wrangling”, took untold time and effort...
...now it happens trivially in seconds.

Command took 3.35 seconds -- by lukas.habegger@regeneron.com at 5/31/2018, 4:37:26 PM on SparkSummitDemo

ACCELERATING INNOVATION THROUGH UNIFIED ANALYTICS

- At Regeneron our goal is to bring the power of science to medicine and improve the lives of patients
- We recognize the unique and transformative power of data as a means towards this end...
...in fact we think genomic data is so important we built a genome center from scratch 5 years ago!
- We have overcome important bottlenecks in the lab with robotics and automation...
....and are doing the same with our massive data set using the Databricks Unified Analytics Platform
- **Accelerated key steps: large-scale queries (30 min to 3 sec – 600x) & ETL (3 wks to 2 days – 10x)!**
More importantly, this platform brings our data and people together to accelerate innovation.

ACKNOWLEDGEMENTS

- RGC-LT
 - Alan Shuldiner
 - Aris Baras
 - Aris Economides
 - John Overton
- RGC-GI
 - Lukas Habegger
 - Alicia Hawes
 - Ashish Yadav
 - Claire Chai
 - Evan Maxwell
 - Gisu Eom
 - Jeff Staples
 - John Penn
 - Leland Barnard
 - Shareef Khalid
 - Sheldon Bai
 - Suganthi Balasubramanian
 - Young Hahn
- RGC
 - Alexander Li
 - Alexander Lopez
 - Amy Damask
 - Charlie Paulding
 - Claudia Schurmann
 - Colm O'Dushlaine
 - Cristopher Van Hout
 - Dylan Sun
 - Jan Freudenberg
 - Kavita Praveen
 - Kia Manoochchri
 - Lauren Gurski
 - Manasi Pradhan
 - Mike Norsen
 - Nehal Gosalia
 - Nila Banerjee
 - Rick Ulloa
 - Shane McCarthy
 - Tanya Teslovich Dostal
 - Tony Marcketta
- REGN-IT
 - Abdul Shaik
 - Allen Chiang
 - Brandon Fetch
 - Christopher McCabe
 - Dale Cochran
 - David Glosser
 - Long Le
 - Michael Phillips
 - Mohammad Saeed
 - Pat Leblanc
 - Sal Mineo
 - Shaw Nawaz
 - Shiva Ravi
 - Stephen Huvane
 - Vin Dahake
 - Weylin Preodor
- Databricks
 - Ali Ghodsi
 - Ali Hodroj
 - Allan Marcos
 - Ambareesh Kulkarni
 - Bavesch Patel
 - Christopher Hoshino-Fish
 - David Weaver
 - Francis Gerace
 - Hossein Falaki
 - Ion Stocia
 - Juliusz Sompolsk
 - Li Yu
 - Navid Bazzazzadeh
 - Paris Georgallis
 - Ram Sriharsha
 - Ronak Shah
 - Shiva Bhattacharjee
 - Vida Ha
 - Yongsheng Huang

ACKNOWLEDGEMENTS

- RGC-LT
 - Alan Shuldiner
 - Aris Baras
 - Aris Economides
 - John Overton
- RGC-GI
 - **Lukas Habegger**
- RGC
 - Alexander Li
 - Alexander Lopez
 - Amy Damask
 - Charlie Paulding
 - Claudia Schurmann
 - Colm O'Dushlaine
 - Cristopher Van Hout
 - Dylan Sun
- REGN-IT
 - Abdul Shaik
 - Allen Chiang
 - Brandon Fetch
 - Christopher McCabe
 - Dale Cochran
 - David Glosser
 - Long Le
 - Michael Phillips
- Databricks
 - Ali Ghodsi
 - Ali Hodroj
 - Allan Marcos
 - Ambareesh Kulkarni
 - Bavesch Patel
 - Christopher Hoshino-Fish
 - David Weaver
 - Francis Gerace

Learn More -- 2:40pm in Room 2000 (Enterprise Track)

- Gisu Eom
- Jeff Staples
- John Penn
- Leland Barnard
- Shareef Khalid
- Sheldon Bai
- Suganthi Balasubramanian
- Young Hahn
- Manasi Pradhan
- Mike Norsen
- Nehal Gosalia
- Nila Banerjee
- Rick Ulloa
- Shane McCarthy
- Tanya Teslovich Dostal
- Tony Marcketta
- Shiva Ravi
- Stephen Huvane
- Vin Dahake
- Weylin Preodor
- Navid Bazzazzadeh
- Paris Georgallis
- Ram Sriharsha
- Ronak Shah
- Shiva Bhattacharjee
- Vida Ha
- Yongsheng Huang