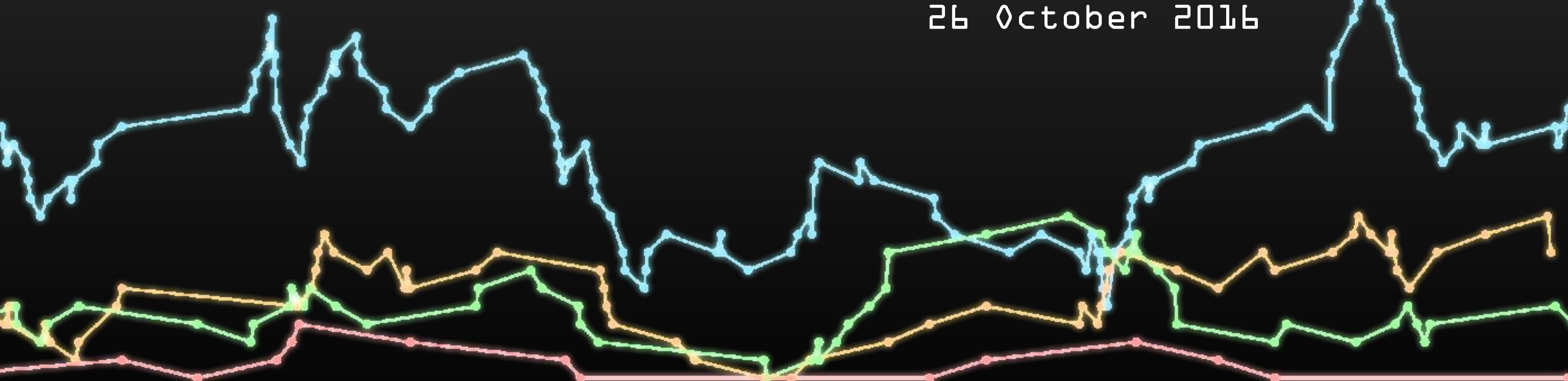
Supporting Data-Centered Software Development

Rob DeLine

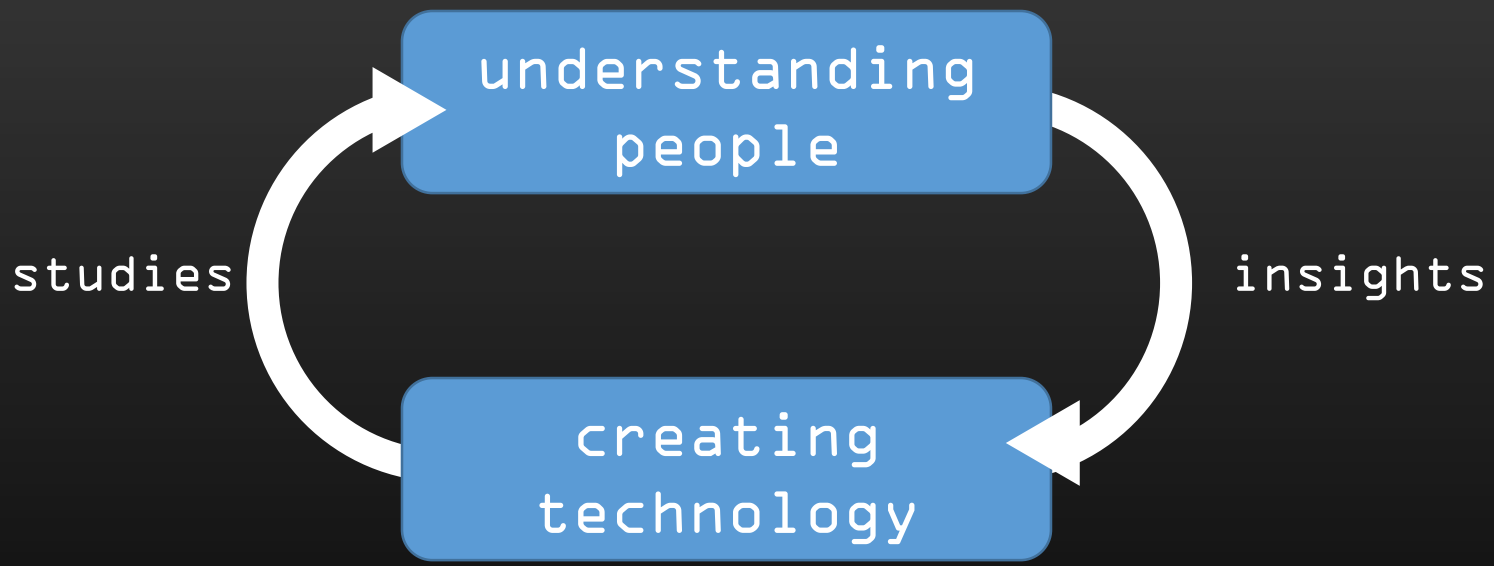
Microsoft Research

BSR Winter School

26 October 2016



User-centered design



What is a data scientist?

We interviewed 16 data scientists at Microsoft.

- 5 women, 11 men. 3 MS/MBA, 8 PhDs
- Ads, Azure, Bing, Exchange, Office, R&D, Skype, Windows, and Xbox
- Recruited through snowball sampling



Kim, Zimmermann, DeLine, and Begel,
"The Emerging Role of Data Scientists on Software Development Teams"
International Conf. on Software Engineering, 2016

Five working styles of data scientists

insight provider

team leader

platform builder

model specialist

polymath

Five working styles of data scientists

insight provider

team leader

platform builder

model specialist

polymath

Coordinate between managers and engineers within a product group

Generate insights and to guide their managers in decision making

Strong communication and coordination skills

P2 got a clear goal for managers and worked with engineers to get data:

I basically tried to eliminate from the vocabulary the notion of “You can just throw the data over the wall ... She'll figure it out.” [P2]

Five working styles of data scientists

insight provider

team leader

platform builder

model specialist

polymath

Senior data scientists who typically run their own data science teams

Act as data science “evangelists”

Work with senior company leaders to inform broad business decisions

P10 led a team to do bug estimation:

Sometimes people who are really good with numbers are not as good with words (laughs), so having an intermediary to handle the human interfaces between the data sources and the data scientists, I think, is a way to have a stronger influence.

Five working styles of data scientists

insight
provider
team leader

platform builder

model
specialist
polymath

Build reusable data engineering platforms

Make trade-offs between engineering and scientific concerns

Strong systems background

P4 makes crash data actionable:

You come up with something called a bucket feed. It is a name of a function most likely responsible for the crash in the small bucket. We found in the source code who touched this function last time. He gets the bug.

Five working styles of data scientists

insight
provider
team leader
platform builder
model specialist
polymath

Act as expert consultants

Build predictive models that can be instantiated as new software features and support other team's data-driven decision making

Strong background in machine learning and statistics.

P7 is an expert in time series analysis:

The [Program Managers] and the Dev Ops from that team... come up with a new set of time series data that they think has the most value and then they will point us to that, and we will try to come up with an algorithm or with a methodology to find the anomalies for that set of time series.

Five working styles of data scientists

insight provider
team leader
platform builder
model specialist
polymath

"Do it all", from forming a business goal, to data collection, to analysis, to communication

P13 thinks of new ideas for ads:

For months at a time I'll wear a dev hat and I actually really enjoy that, too. ... I spend maybe three months doing some analysis and maybe three months doing some coding that is to integrate whatever I did into the product. … I love the flexibility that I can go from being developer to being an analyst.

What do data scientists work on?

Performance Regression

Are we getting better in terms of crashes or worse? [P3]

Requirements Identification

If you see the repetitive pattern where people don't recognize, the feature is there. [P3]

Root Cause Analysis

What areas of the product are failing and why? [P3]

Bug Prioritization

Oh, cool. Now we know which bugs we should fix first. Then how can we reproduce this error? [P5]

Server Anomaly Detection

Is this application log abnormal w.r.t. the rest of the data? [P12]

Failure Rate Estimation

Is the beta ready to ship? [P8]

Customer Understanding

How long do our users use the app? [P1]

What are the most popular features? [P4]

Cost Benefit Analysis

How many customer service calls can we prevent if we detect this type of anomaly? [P9]

Follow-up Survey

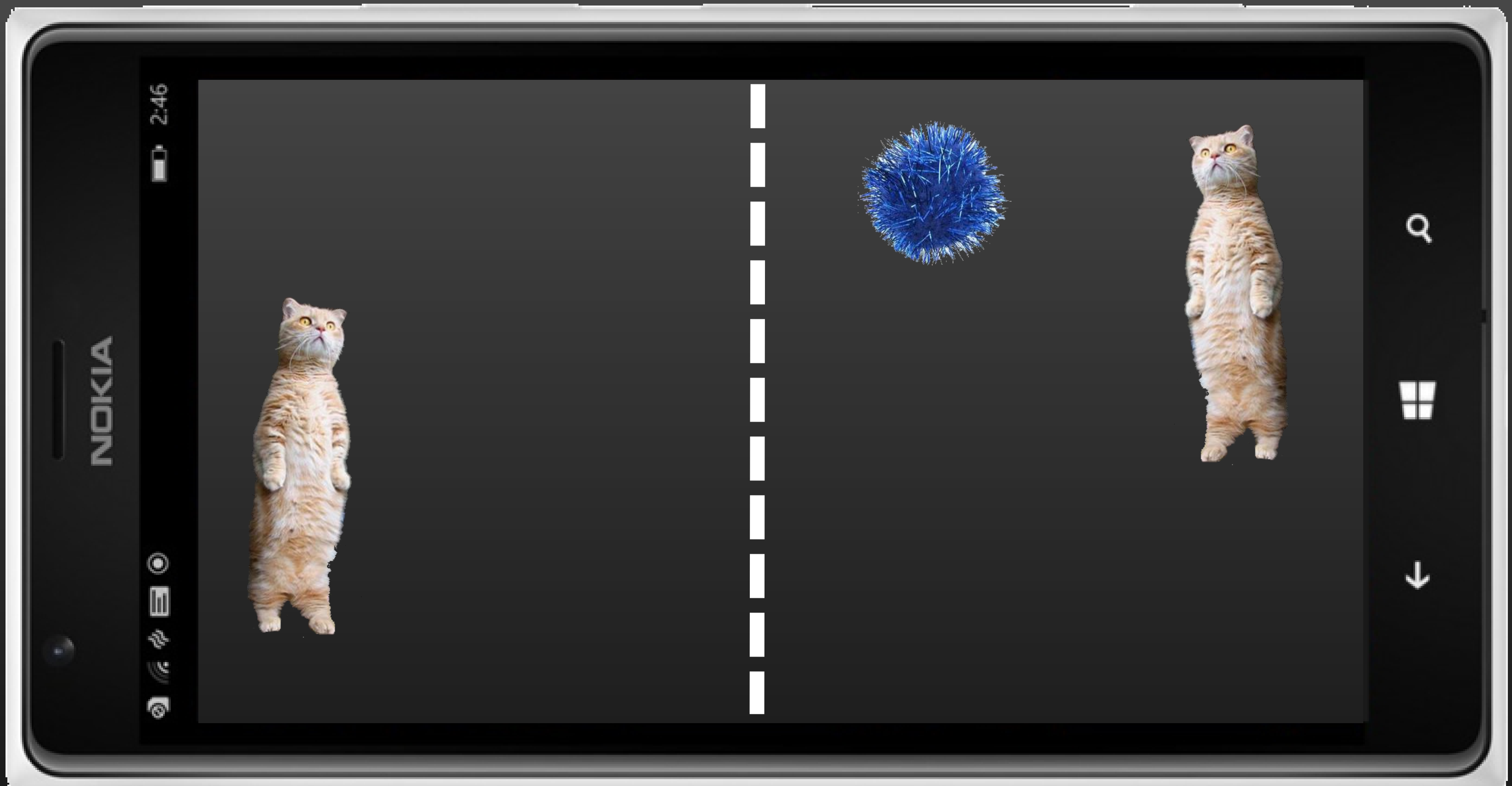
Questionnaire with 793 respondents

- Two populations: data science discipline (36% resp. rate); data science distribution list (32%)
- Demographics/education
- Work styles and activities
- Challenges and best practices
- Correctness/quality

Kim, Zimmermann, DeLine, and Begel,
"Everything You Wanted to Know About Data Scientists in Software Teams"
In submission, Trans. on Software Engineering

		Query existing data	Build platforms to gather data	Prepare data	Analyze data	Experiment	Validate insight	Disseminate insight	Engage with others	Operationalize insight	Act on insight	Other work related to DS	Other work not related to DS
	Entire population 532 people	12.0% 4.7h	7.2% 2.9h	11.7% 4.9h	12.5% 5.2h	4.8% 2.1h	6.9% 3.0h	8.5% 3.5h	9.2% 3.8h	2.4% 1.1h	5.5% 2.1h	4.1% 1.9h	15.1% 6.7h
polymath	Cluster 1 Polymath 156 people	10.4% 4.4h	8.5% 3.6h	11.5% 5.1h	15.1% 6.7h	9.1% 4.0h	7.7% 3.6h	7.4% 3.5h	7.9% 3.6h	3.2% 1.5h	5.2% 2.3h	4.0% 2.0h	10.1% 4.5h
insight provider team leader	Cluster 2 Data Evangelist 71 people	6.8% 2.2h	2.1% 1.0h	6.7% 2.5h	7.7% 2.9h	2.4% 1.2h	7.0% 2.6h	12.0% 4.5h	23.0% 8.6h	3.7% 1.3h	9.5% 3.3h	13.4% 6.0h	5.7% 2.6h
modeling specialist	Cluster 3 Data Preparer 122 people	24.5% 9.4h	4.9% 1.9h	19.6% 7.8h	10.0% 4.0h	3.0% 1.3h	9.0% 4.1h	11.6% 4.5h	8.8% 3.5h	1.5% 0.7h	3.9% 1.3h	1.5% 0.7h	1.8% 0.8h
	Cluster 4 Data Shaper 33 people	5.6% 2.5h	1.8% 0.7h	27.0% 11.5h	25.7% 10.9h	6.0% 2.6h	8.9% 3.8h	7.6% 3.3h	7.5% 3.2h	2.1% 1.0h	3.3% 1.4h	2.5% 1.1h	1.9% 0.9h
	Cluster 5 Data Analyzer 24 people	9.9% 3.7h	0.9% 0.3h	5.8% 2.4h	49.1% 18.4h	4.6% 2.2h	6.6% 2.7h	5.2% 2.2h	5.8% 2.4h	1.8% 0.9h	4.2% 1.6h	2.8% 1.3h	3.2% 1.3h
platform builder	Cluster 6 Platform Builder 27 people	12.5% 4.4h	48.5% 18.4h	6.1% 2.6h	4.3% 1.9h	3.8% 1.1h	2.7% 1.2h	4.4% 2.0h	4.1% 1.9h	2.1% 0.9h	3.0% 1.1h	1.4% 0.6h	6.9% 3.1h
moonlighter	Cluster 7 Moonlighter 50% 63 people	7.3% 3.1h	5.0% 2.2h	5.0% 2.1h	5.5% 2.4h	2.8% 1.2h	4.2% 2.0h	7.8% 3.3h	5.9% 2.4h	1.8% 0.8h	5.7% 2.3h	2.5% 1.1h	46.5% 20.0h
	Cluster 8 Moonlighter 10% 32 people	2.9% 1.2h	1.4% 0.6h	1.9% 0.9h	1.6% 0.7h	0.4% 0.2h	1.5% 0.7h	1.7% 0.8h	2.3% 1.0h	0.6% 0.3h	2.1% 1.0h	2.9% 1.3h	80.9% 36.1h
insight consumer	Cluster 9 Act on Insight 4 people	0.9% 0.1h	2.1% 1.0h	1.8% 0.2h		0.9% 0.1h	5.7% 1.5h	18.5% 4.8h	10.1% 1.6h	3.0% 1.1h	57.1% 11.8h		

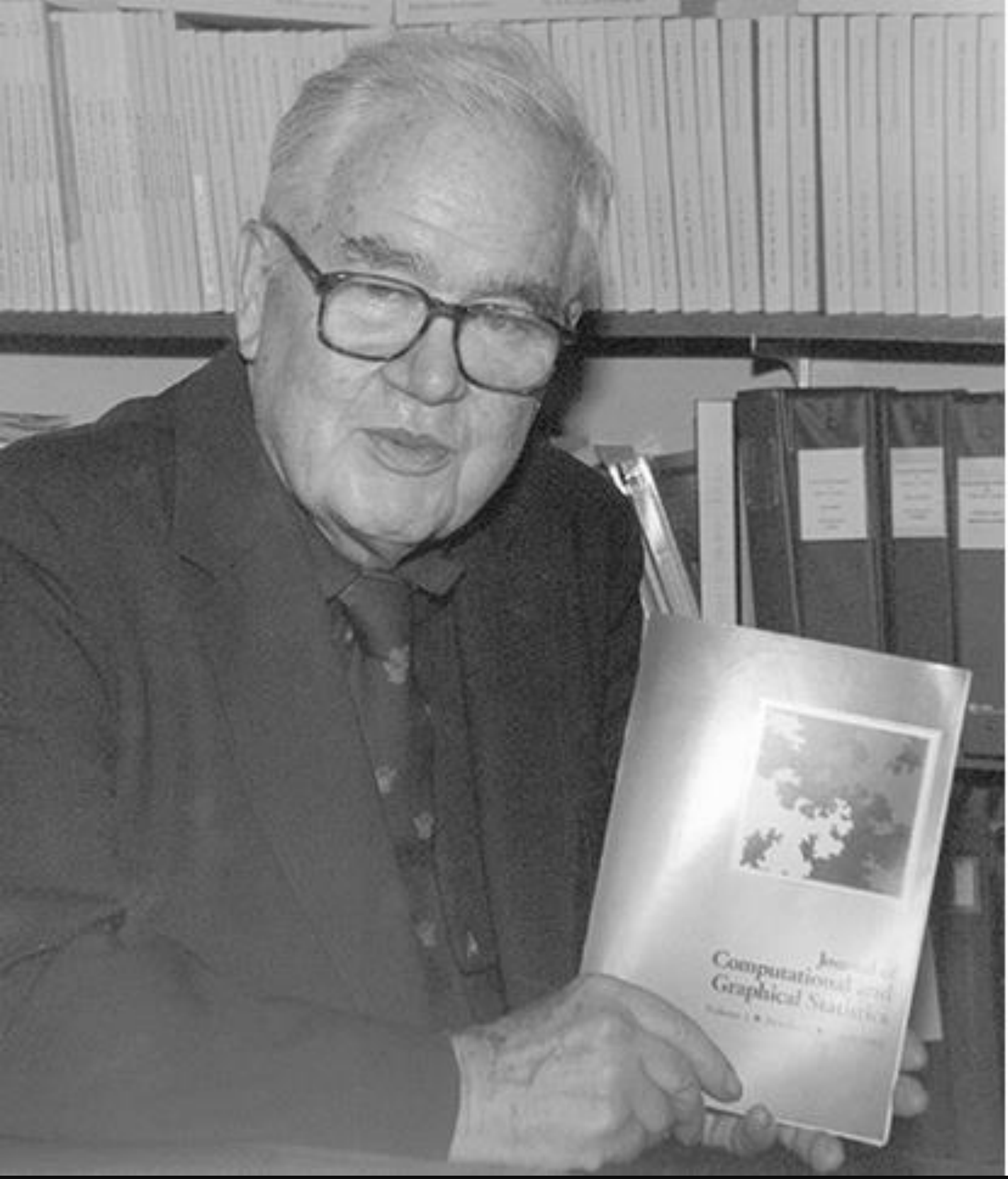
What's it like to be a data scientist?

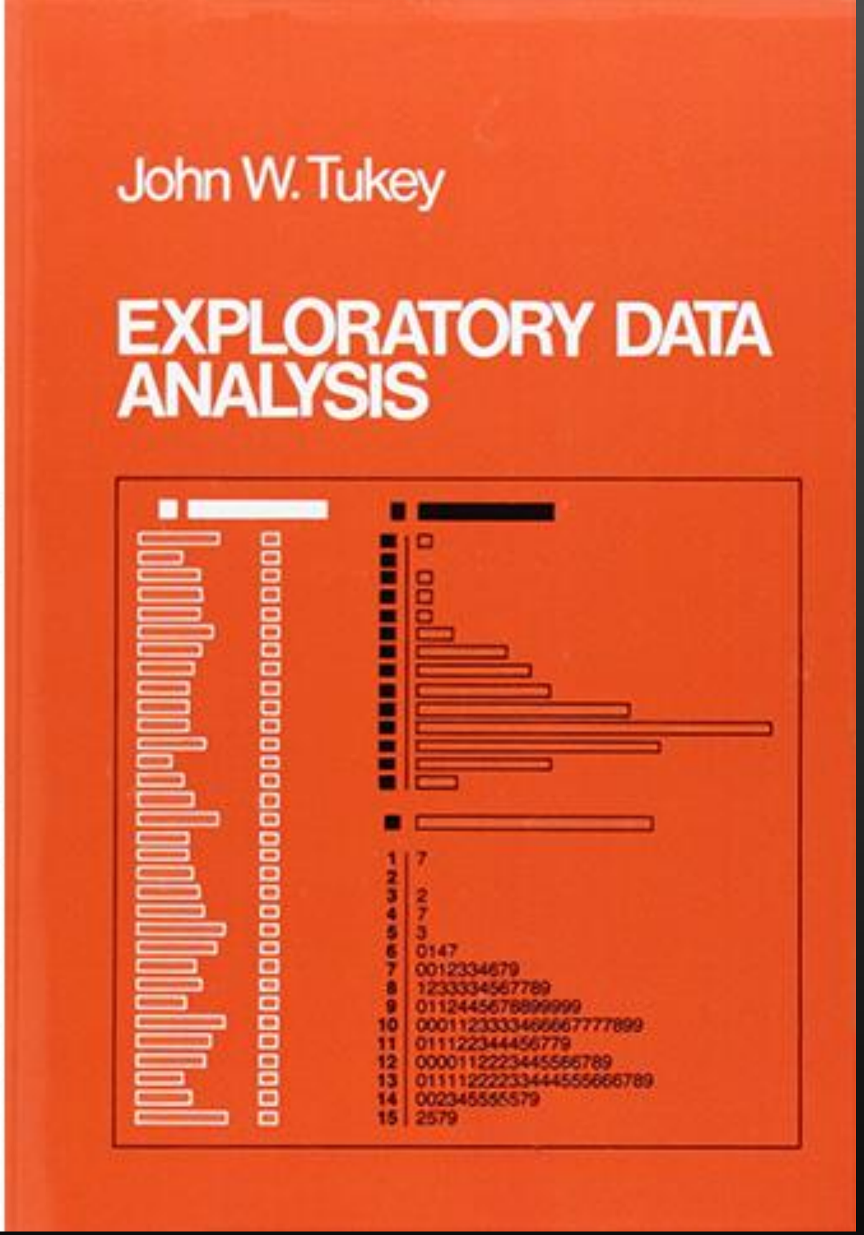


Kitteh Pong!

Matching players by skill

1. Help the team decide whether to implement this feature (analytics).
2. If so, help the team deploy the feature.
3. Measure player reaction to the feature (flighting).
4. Monitor customer usage of the feature.




John W. Tukey
EXPLORATORY DATA ANALYSIS

1a. Grab a manageable sample of the data.

```
Players = LOAD 'player_data';
PSample = SAMPLE Players 0.01;
STORE PSample INTO 'psample';

Games = LOAD 'games_data';
GSample = SAMPLE Games 0.0001;
STORE GSample INTO 'gsample';
```

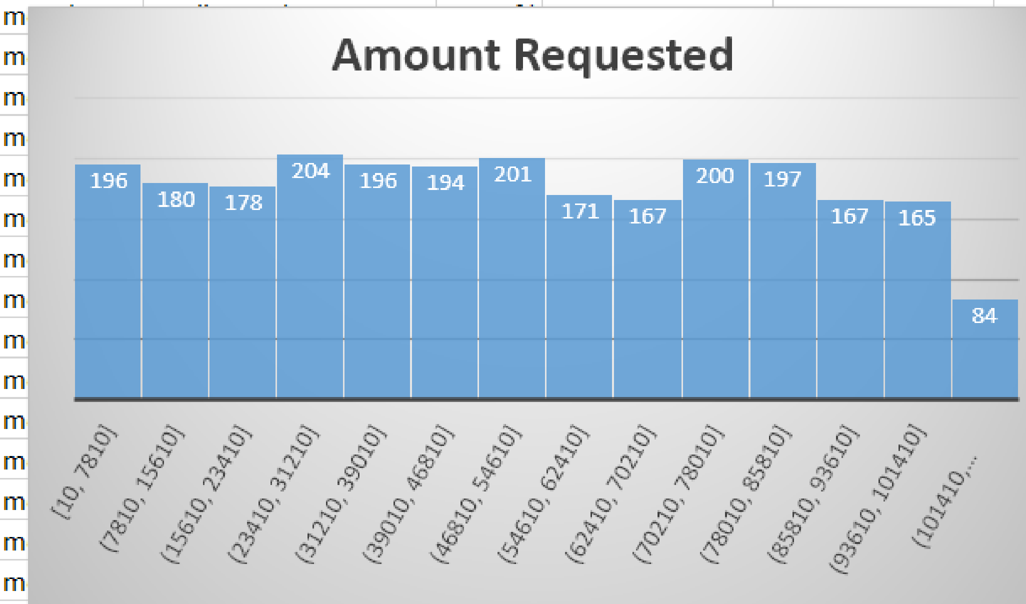
loansData.csv [Read-Only] - Excel

File Home Insert Page Layout Formulas Data Review View LOAD TEST FOXIT READER PDF Team Tell me Rob De... Share

F9 | credit_card

1b. Explore the data sample.

	A	B	C	D	E	F	G	H	I	J
1	Amount.Re	Amount.Fun	Interest Rate	Loan Length	Loan Purpose	Debt To Income Ratio	State	Home Ownership	Monthly.Income	FICO.Range
2	81174	20000							MORTGAGE	6541.6
3	99592	19200							MORTGAGE	4583.3
4	80059	35000							MORTGAGE	1150
5	15825	10000	9975	9.99%	36 months	debt_consolidation	14.30%	KS	MORTGAGE	3833.3
6	33182	12000	12000	11.71%	36 months	credit_card	18.78%	NJ	RENT	319
7	62403	6000	6000	15.31%	36 months	other	20.05%	CT	OWN	4891.6
8	48808	10000	10000	7.90%	36 months	debt_consolidation	26.09%	MA	RENT	2916.6
9	22090	33500	33450	17.14%	60 months	credit_card	14.70%	LA	MORTGAGE	13863.4
10	76404	14675	14675	14.33%	36 months	credit_card	26.92%	CA	RENT	315
11	15867	7000	7000	6.91%	36 m					500
12	94971	2000	2000	19.72%	36 m					357
13	36911	10625	10625	14.27%	36 m					425
14	41200	28000	27975	21.67%	60 m					14166.6
15	83869	35000	34950	8.90%	36 m					9166.6
16	53853	9600	9600	7.62%	36 m					1125
17	21399	25000	24975	15.65%	60 m					5416.6
18	62127	10000	10000	12.12%	36 m					900
19	23446	14000	13900.25	10.37%	60 m					4333.3
20	44987	10000	10000	9.76%	36 m					2733.3
21	17977	5200	5175	9.99%	60 m					375
22	86099	22000	21975	21.98%	36 m					6666.6
23	99483	30000	30000	19.05%	60 m					625
24	28798	6500	6500	17.99%	60 m					410
25	24168	17400	17400	11.99%	36 m					6833.3
26	10356	4000	4000	16.82%	60 months	vacation	13.71%	GA	MORTGAGE	450
27	46027	7200	7200	7.90%	36 months	debt_consolidation	24.82%	TX	RENT	5416.6



loansData

Ready 100%

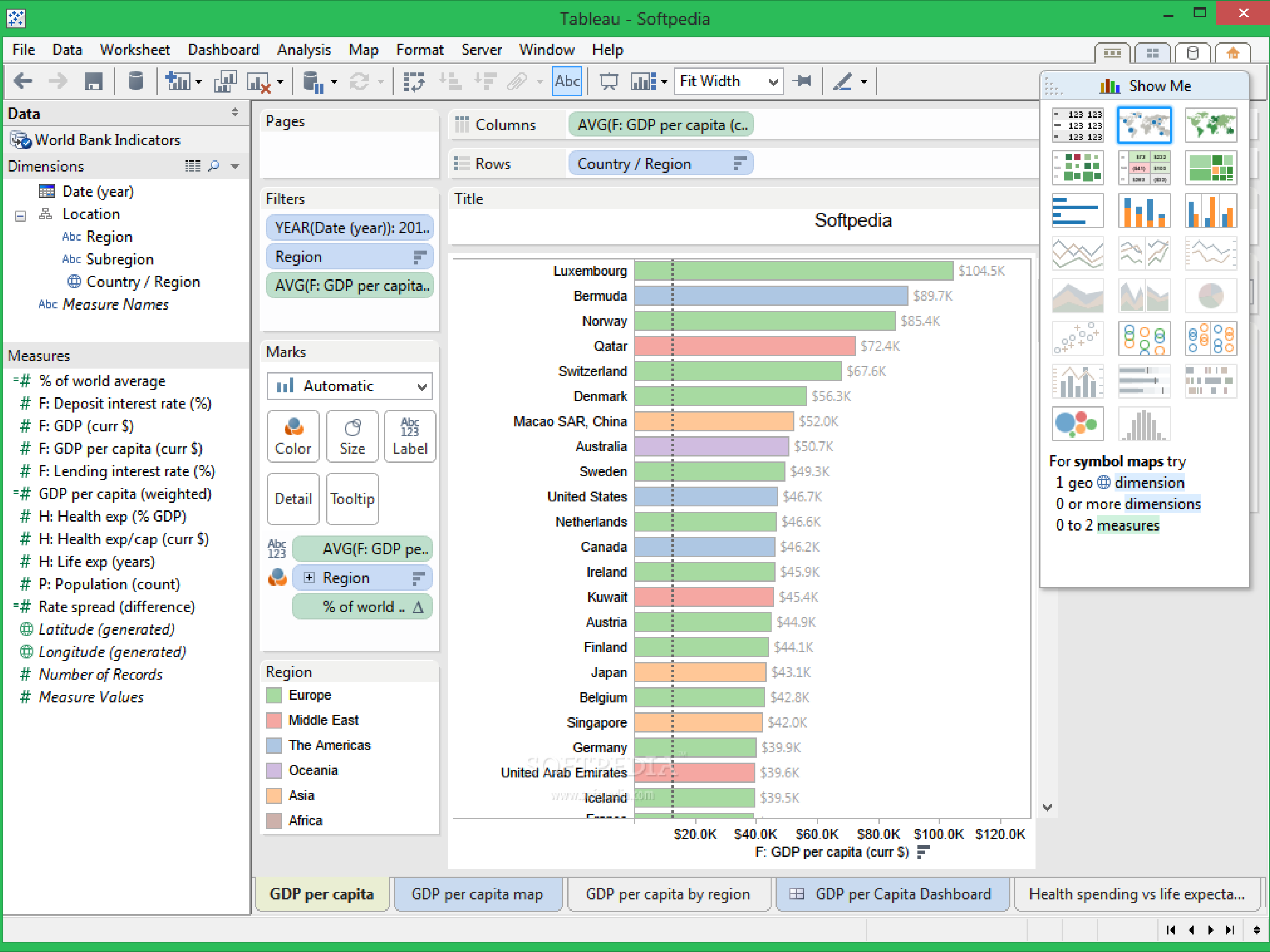
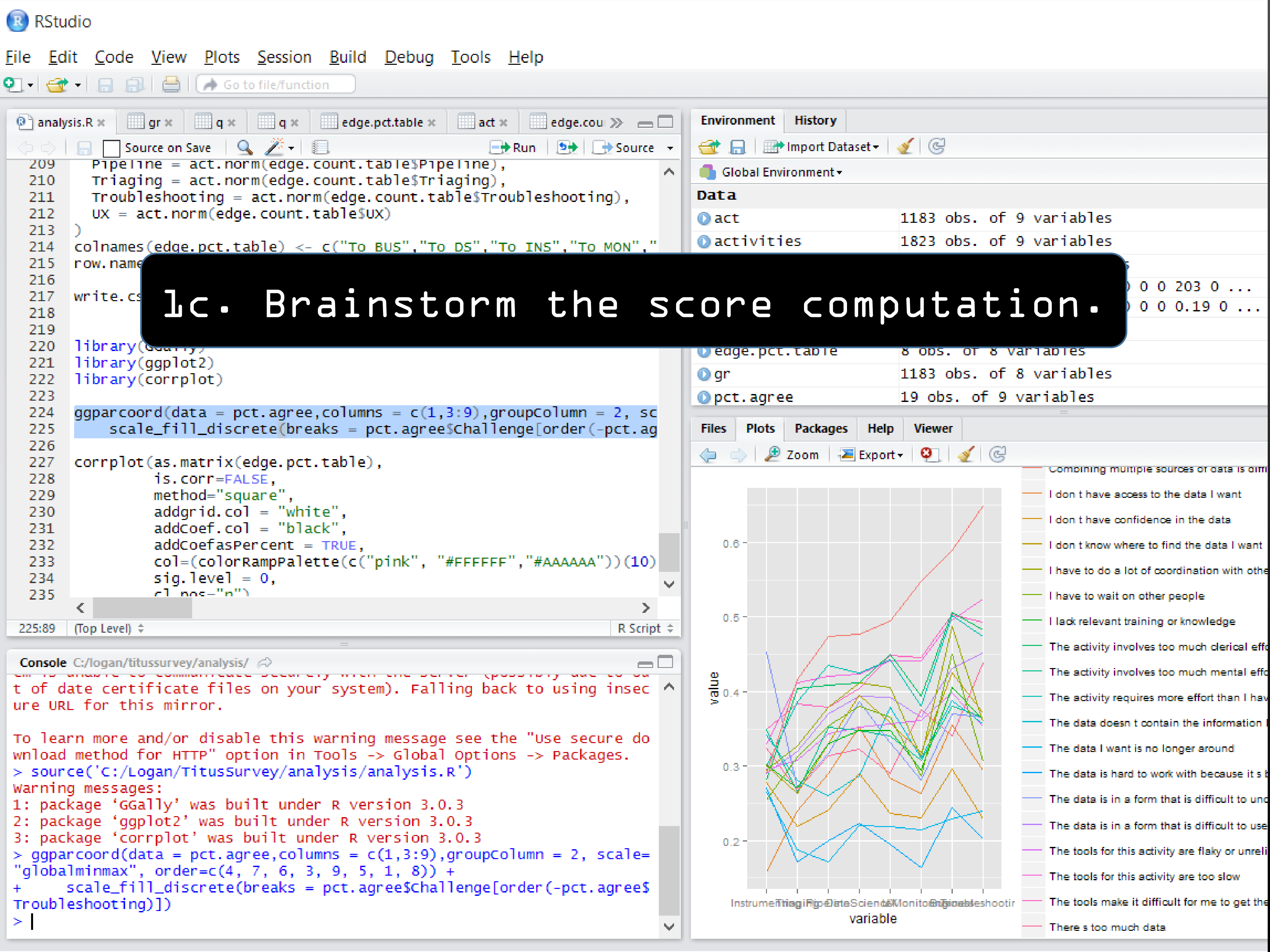


Tableau - Softpedia
File Data Worksheet Dashboard Analysis Map Format Server Window Help
Fit Width
Show Me
Data
World Bank Indicators
Dimensions
Date (year)
Location
Region
Subregion
Country / Region
Measure Names
Measures
% of world average
F: Deposit interest rate (%)
F: GDP (curr $)
F: GDP per capita (curr $)
F: Lending interest rate (%)
GDP per capita (weighted)
H: Health exp (% GDP)
H: Health exp/cap (curr $)
H: Life exp (years)
P: Population (count)
Rate spread (difference)
Latitude (generated)
Longitude (generated)
Number of Records
Measure Values
Pages
Filters
YEAR(Date (year)): 201..
Region
AVG(F: GDP per capita..
Marks
Automatic
Color
Size
Label
Detail
Tooltip
AVG(F: GDP pe..
Region
% of world ..
Region
Europe
Middle East
The Americas
Oceania
Asia
Africa
Columns
AVG(F: GDP per capita (c..
Rows
Country / Region
Title
Softpedia
Luxembourg $104.5K
Bermuda $89.7K
Norway $85.4K
Qatar $72.4K
Switzerland $67.6K
Denmark $56.3K
Macao SAR, China $52.0K
Australia $50.7K
Sweden $49.3K
United States $46.7K
Netherlands $46.6K
Canada $46.2K
Ireland $45.9K
Kuwait $45.4K
Austria $44.9K
Finland $44.1K
Japan $43.1K
Belgium $42.8K
Singapore $42.0K
Germany $39.9K
United Arab Emirates $39.6K
Iceland $39.5K
$20.0K $40.0K $60.0K $80.0K $100.0K $120.0K
F: GDP per capita (curr $)
For symbol maps try
1 geo dimension
0 or more dimensions
0 to 2 measures
GDP per capita
GDP per capita map
GDP per capita by region
GDP per Capita Dashboard
Health spending vs life expecta...

1c. Brainstorm the score computation.



jupyter spectrogram Last Checkpoint: an hour ago (autosaved)

File Edit View Insert Cell Kernel Help

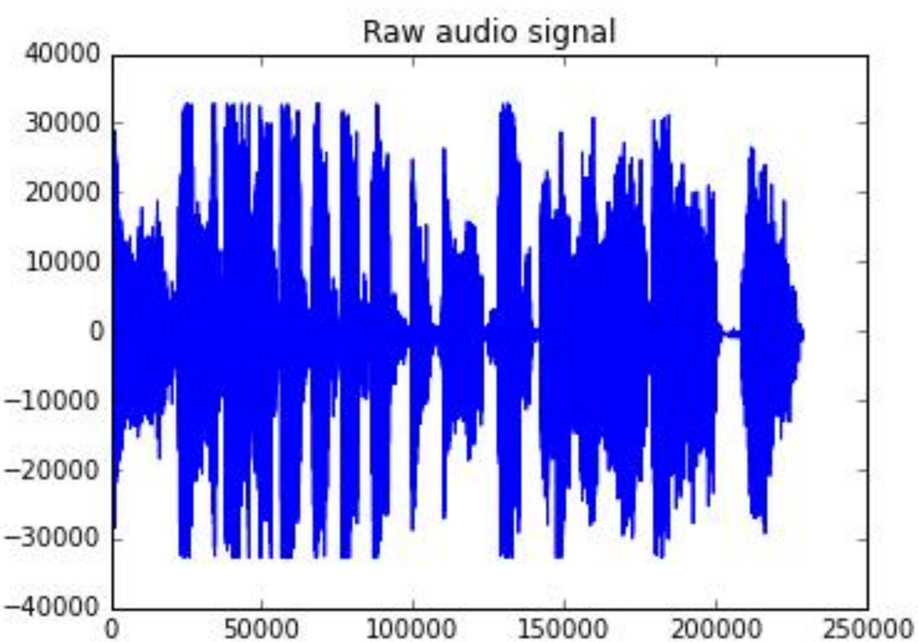
Code ▾ Cell Toolbar: None ▾

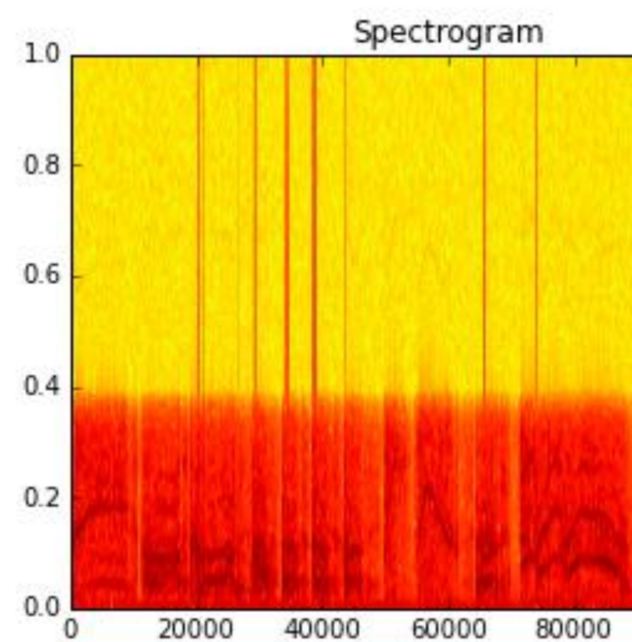
In [1]:

```
from scipy.io import wavfile
rate, x = wavfile.read('test_mono.wav')
```

In [2]:

```
import matplotlib.pyplot as plt
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 4))
ax1.plot(x); ax1.set_title('Raw audio signal')
ax2.specgram(x); ax2.set_title('Spectrogram')
plt.show()
```



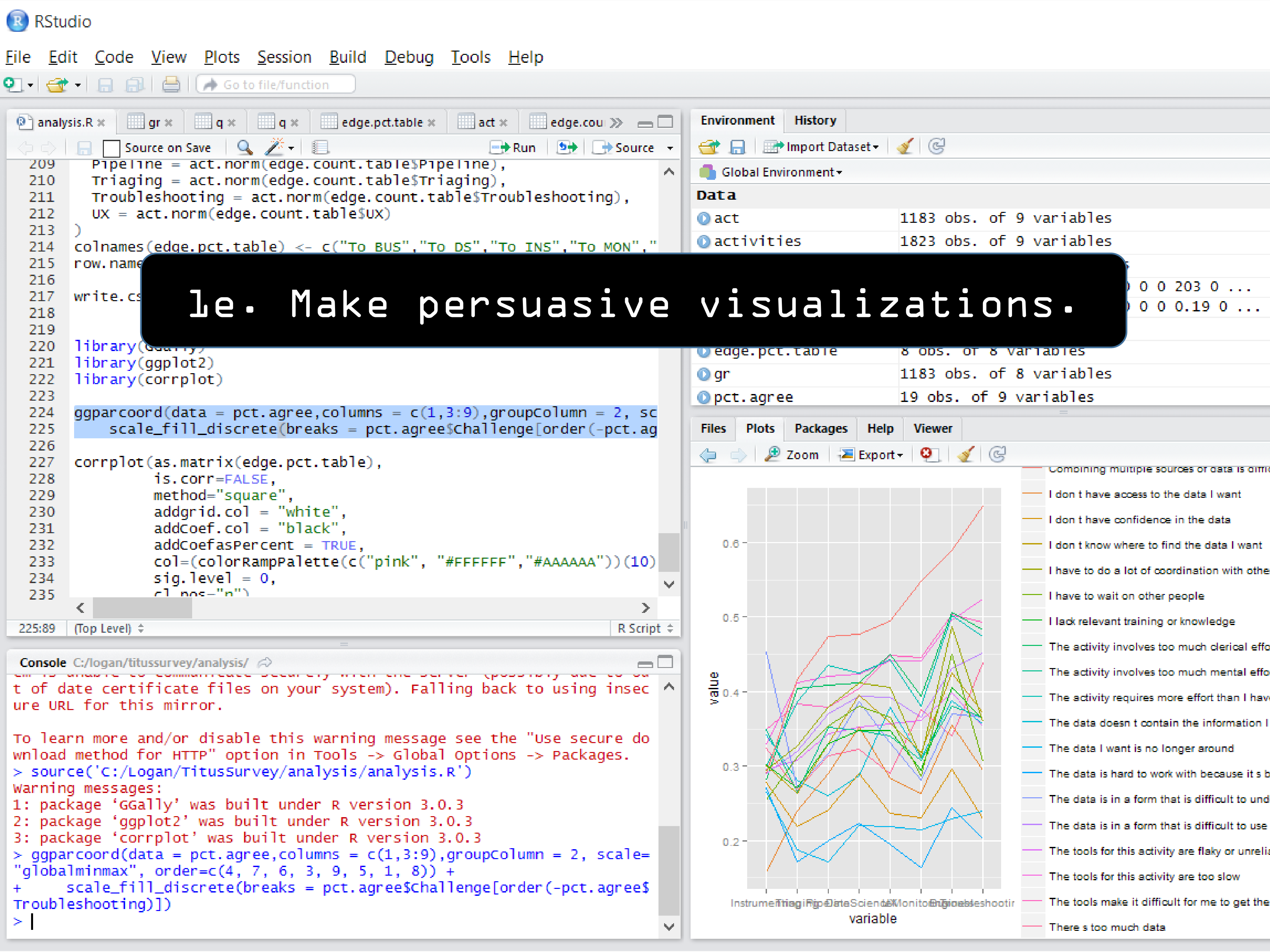


1d. Scale out the score computation.

```
REGISTER gamemetrics.jar;
Player = LOAD 'player_data' AS
    (pid: chararray, region: chararray, signupDate: chararray);
Games = LOAD 'game_data' AS
    (gid: chararray, winner: chararray, winningScore: int, timestamp: chararray);
WinPlayers = JOIN Player BY pid, Games by winner;
Wins = GROUP WinPlayers BY pid;
WinSummary = FOREACH Wins
    GENERATE pid, COUNT($1), MAX($1.timestamp) AS latest
Skills = FOREACH WinSummary GENERATE gamemetrics.SkillScore(*);
STORE Skills INTO 'skills';
```

```
public class SkillScore extends EvalFunc (String)
{
    public Double exec(Tuple input) throws IOException {
        if (input == null || input.size() == 0)
            return null;
        // COMPUTE SKILL SCORE
        return skillScore;
    }
}
```

1e. Make persuasive visualizations.



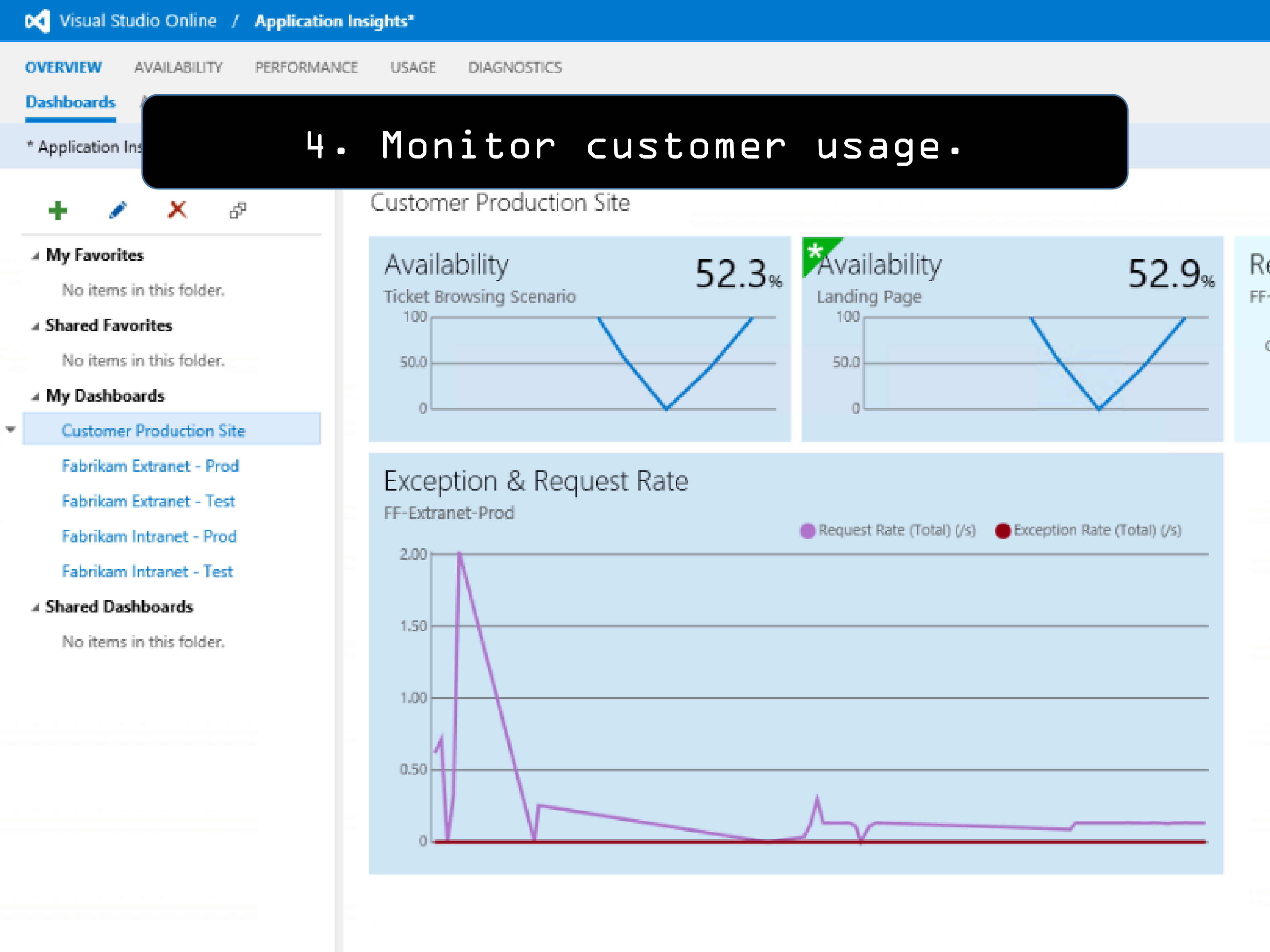
1. Help the team decide whether to implement player matching (analytics).

2. If so, help the team deploy the feature.

3. Measure player reaction to the feature (flighting).

4. Monitor customer usage of the feature.

4. Monitor customer usage.



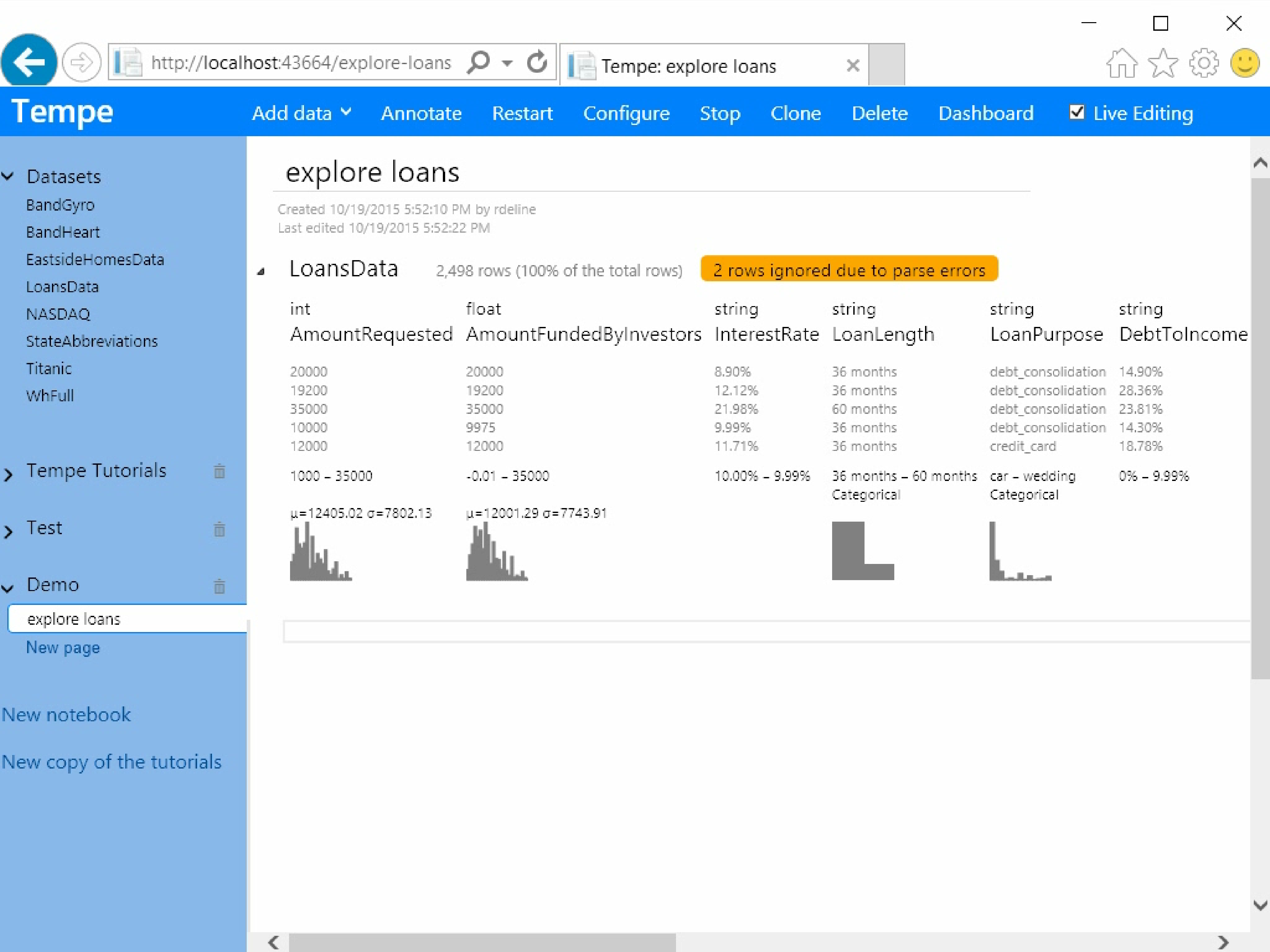
Lots of work to make the data fit the system

Many hours of waiting for batch processing of data

Clerical work to handle files and pass data between separate tools

Constant vigilance to keep data organized and not lose it

Interaction and iteration require working with small samples


http://localhost:43664/explore-loans
Tempe: explore loans
Tempe
Add data
Annotate
Restart
Configure
Stop
Clone
Delete
Dashboard
Live Editing
Datasets
BandGyro
BandHeart
EastsideHomesData
LoansData
NASDAQ
StateAbbreviations
Titanic
WhFull
Tempe Tutorials
Test
Demo
explore loans
New page
New notebook
New copy of the tutorials
explore loans
Created 10/19/2015 5:52:10 PM by rdeline
Last edited 10/19/2015 5:52:22 PM
LoansData
2,498 rows (100% of the total rows)
2 rows ignored due to parse errors
int
AmountRequested
float
AmountFundedByInvestors
string
InterestRate
string
LoanLength
string
LoanPurpose
string
DebtToIncome
20000 20000 8.90% 36 months debt_consolidation 14.90%
19200 19200 12.12% 36 months debt_consolidation 28.36%
35000 35000 21.98% 60 months debt_consolidation 23.81%
10000 9975 9.99% 36 months debt_consolidation 14.30%
12000 12000 11.71% 36 months credit_card 18.78%
1000 – 35000
-0.01 – 35000
10.00% – 9.99%
36 months – 60 months
Categorical
car – wedding
Categorical
0% – 9.99%
μ=12405.02 σ=7802.13
μ=12001.29 σ=7743.91

http://localhost:43664/twitter

Tempe: twitter

Tempe | Add data | Annotate | Restart | Configure | Stop | Clone | Delete | Dashboard | ☑ Live Editing

Datasets
- BandGyro
- EastsideHomesData
- NASDAQ
- StateAbbreviations
- Titanic
- WhFull

Tempe Tutorials
- Welcome!
- Writing scripts
- Working with data
- Visualizing data
- Query samples
- Machine learning: using TLC
- Sample: NASDAQ stock data
- Sample: White House visitors
- New page

Demo
- twitter
- stock
- loans
- easy
- signals
- heartbeat
- New page

twitter

Created 10/29/2015 3:32:38 PM by Rob DeLine
Last edited 10/29/2015 3:33:08 PM

```
var myConf = Badrish.GetTwitterConfig();
```

oauth_consumer_key	s5ocdNJ2JnKi87ayC0Fpwto1f
oauth_consumer_secret	D7OH70N1QUhCMUYdblQuqZFx6Dqc1mKiuHljT1Sv7lgh4scXqo
oauth_token	2547031843-C3Dt114eXNSYkZx83Xj1kBkL3AK4F4E76dWMmrD
oauth_token_secret	YIX6vSQoofKZp78Zf1j4hxGBiPSQbtmdVOCUjRiA5BDN2
twitter_keywords	Office Microsoft,Surface Microsoft,Phone Window,Windows 8,SQL Server,SharePoint,Bing,Skype,XBox,System Center
delay_ms	0

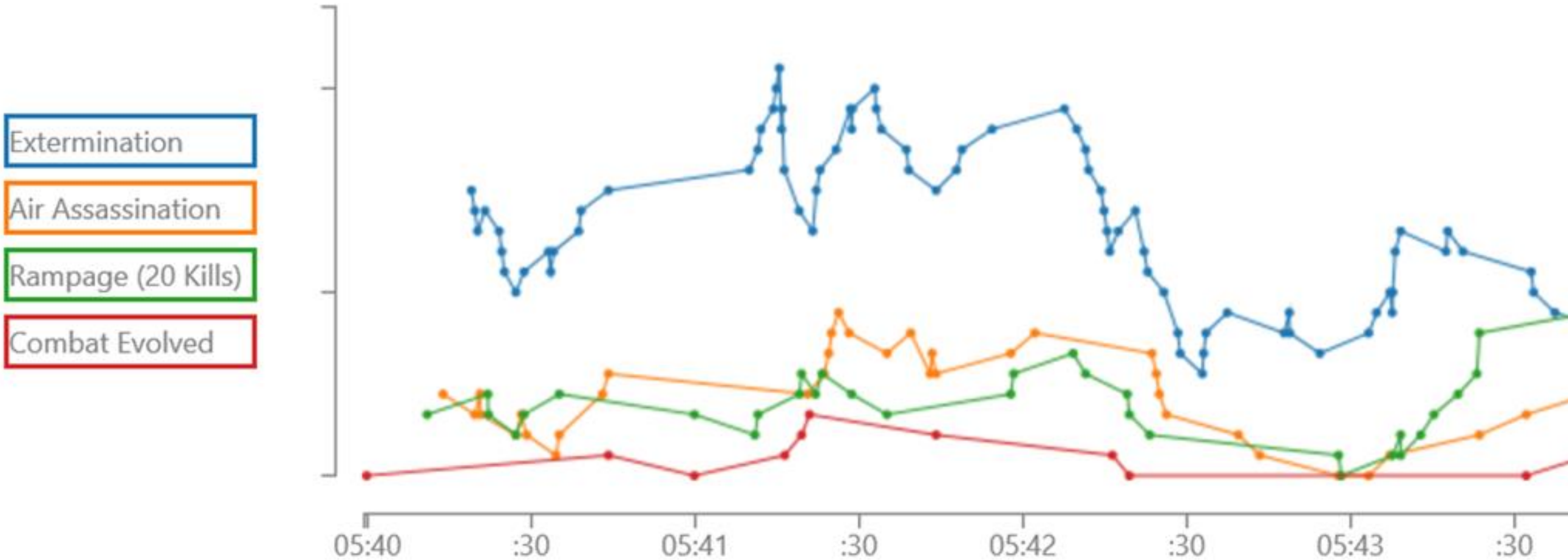
```
var inputObservable = TwitterObservable.TwitterStream.Create(myConf);

var twitterStreamable = inputObservable
    .ToStreamable(OnCompletedPolicy.EndOfStream(), PeriodicPunctuationPolicy.Count(1))
    .RepetitiveHoppingWindowLifetime(TimeSpan.FromSeconds(10).Ticks, TimeSpan.FromSeconds(1).Ticks);
```

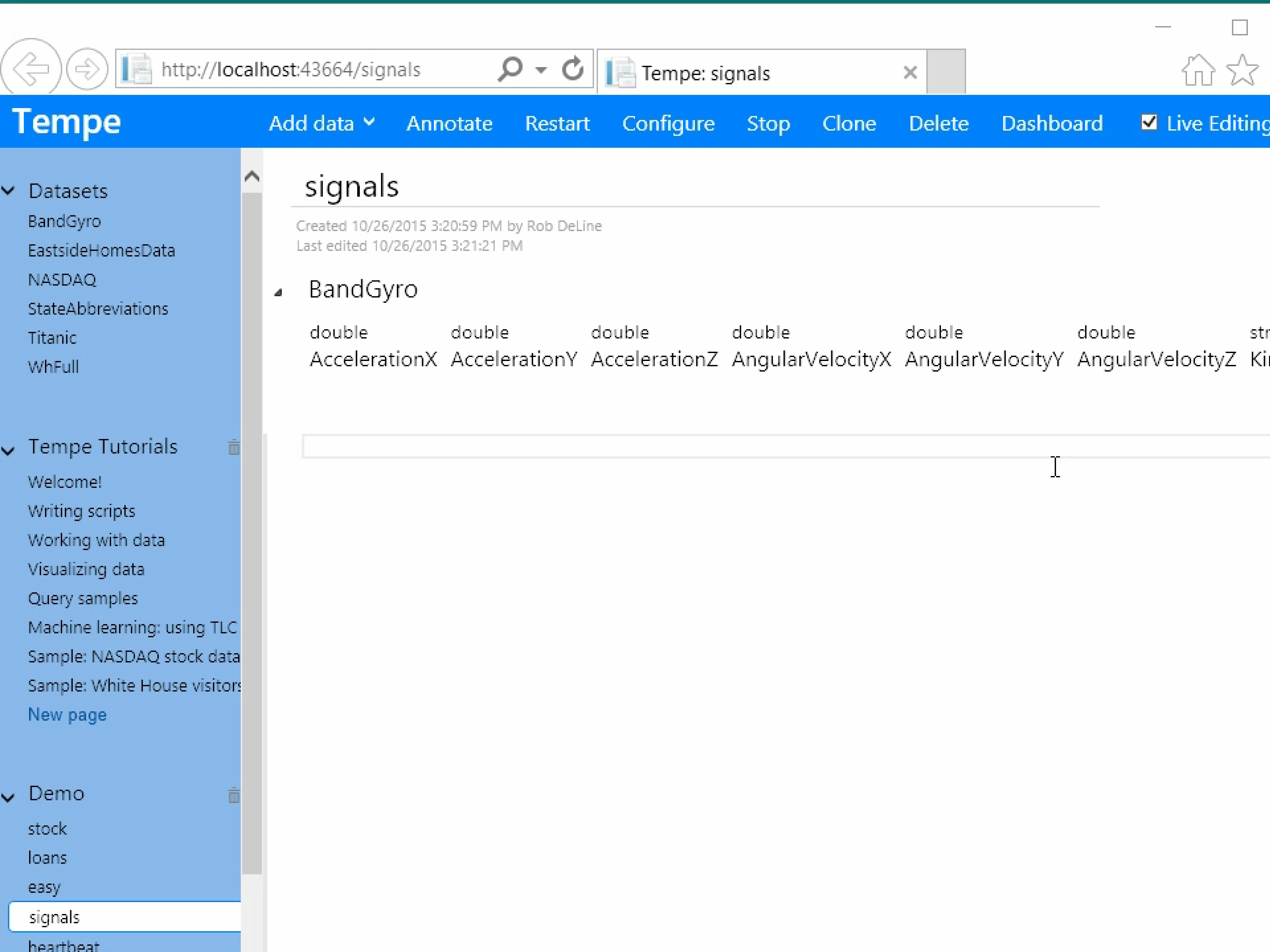
	ID	CreatedAt	UserName	ProfileImageUrl	Text
2015/10/29 16:14:28 – 16:14:29	659825640011644928	10/29/2015 8:14:26 PM	freewebcamsluts	http://pbs.twimg.com/profile_images/648288720387354624/RV4Udni3_normal.jpg	♥♥♥♥ https:/, ♥♥♥♥ chat? https:/, #kik #c #skype https:/,


HALO 5
GUARDIANS
MULTIPLAYER BETA
12.29.14 - 1.18.15

```
var medalInteresting = medals.Where(val => goodMedalIds.ContainsKey(val.MedalId))
        .AlterEventDuration(TimeSpan.FromMinutes(1).Ticks)
        .GroupApply(val => goodMedalIds[val.MedalId], e => e.Count(),
                (input, aggregateData) => Tuple.Create(input.Key, aggregateData))
        .Vis("Medals per minute Count", "", true, false)
        .GroupBy(val => val.Item1, val => val.Select( e=> e.Item2));
```




WOLF
32
flumutu OAJL
Full Scoreboard


http://localhost:43664/signals
Tempe: signals
Tempe
Add data
Annotate
Restart
Configure
Stop
Clone
Delete
Dashboard
Live Editing
Datasets
BandGyro
EastsideHomesData
NASDAQ
StateAbbreviations
Titanic
WhFull
Tempe Tutorials
Welcome!
Writing scripts
Working with data
Visualizing data
Query samples
Machine learning: using TLC
Sample: NASDAQ stock data
Sample: White House visitors
New page
Demo
stock
loans
easy
signals
heartbeat
signals
Created 10/26/2015 3:20:59 PM by Rob DeLine
Last edited 10/26/2015 3:21:21 PM
BandGyro
double AccelerationX
double AccelerationY
double AccelerationZ
double AngularVelocityX
double AngularVelocityY
double AngularVelocityZ

What Tempe gets right and wrong

- URL sharing and visualizations support team communication.
- A single query API supports many scenarios.
- The use of C# allows code to move between Tempe and the production system.
- Easy switching between monitoring and ad hoc queries supports "drilling in".
- The need for scripting turns developers into gatekeepers.

  Originally we took the PMs and said here's a quick way to do this. They sort of tried to use it, but they weren't able to, so it fell back to me.

Study of logging and telemetry

Stage 1: Interviews with 28 Microsoft engineers

- 10 devs, 9 PMs, 4 data scientists, 2 ops, 2 content devs, 1 service eng
- We learned that engineers do 8 activities, with several pain points

Stage 2: Internal survey with 1823 respondents

- Random selection from the address book (28% response rate)
- Confirmed activities and pain points

Barik, DeLine, Drucker, Fisher
"The Bones of the System: A Case Study of Logging and Telemetry at Microsoft"
ICSE 2016

Eight activities with logs/telemetry

Engineering the data pipeline
e.g. data collection, deploying data features

Instrumenting for logs/telemetry
e.g. adding log statements to the code

Troubleshooting problems
e.g. finding a bug's root cause

Monitoring services
e.g. looking for anomalies

Doing data science
e.g. analyzing data, running experiments

Improving the user experience
e.g. understanding feature usage

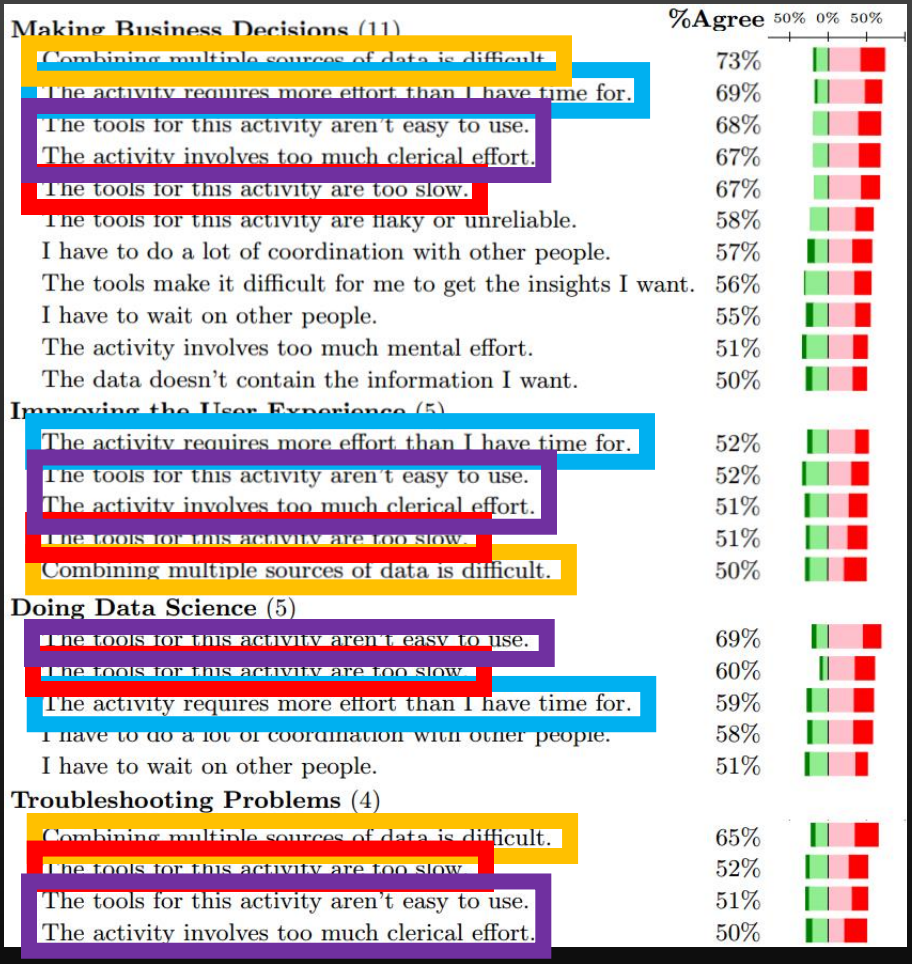
Triaging work items
e.g. assigning priority and responsibility

Making business decisions
e.g. product planning, marketing strategies

%respondents by discipline who analyze events at least weekly

There are several recurring pain points.

- Working with data is only part of the job. Tools require too much effort.
- Getting the whole picture means combining multiple logs.
- Our tools are too slow.
- Our tools are too hard to use and clerical.



	%Agree
Making Business Decisions (11)	
Combining multiple sources of data is difficult.	73%
The activity requires more effort than I have time for.	69%
The tools for this activity aren't easy to use.	68%
The activity involves too much clerical effort.	67%
The tools for this activity are too slow.	67%
The tools for this activity are flaky or unreliable.	58%
I have to do a lot of coordination with other people.	57%
The tools make it difficult for me to get the insights I want.	56%
I have to wait on other people.	55%
The activity involves too much mental effort.	51%
The data doesn't contain the information I want.	50%
Improving the User Experience (5)	
The activity requires more effort than I have time for.	52%
The tools for this activity aren't easy to use.	52%
The activity involves too much clerical effort.	51%
The tools for this activity are too slow.	51%
Combining multiple sources of data is difficult.	50%
Doing Data Science (5)	
The tools for this activity aren't easy to use.	69%
The tools for this activity are too slow.	60%
The activity requires more effort than I have time for.	59%
I have to do a lot of coordination with other people.	58%
I have to wait on other people.	51%
Troubleshooting Problems (4)	
Combining multiple sources of data is difficult.	65%
The tools for this activity are too slow.	52%
The tools for this activity aren't easy to use.	51%
The activity involves too much clerical effort.	50%

Logan

Event Sequence Exploration

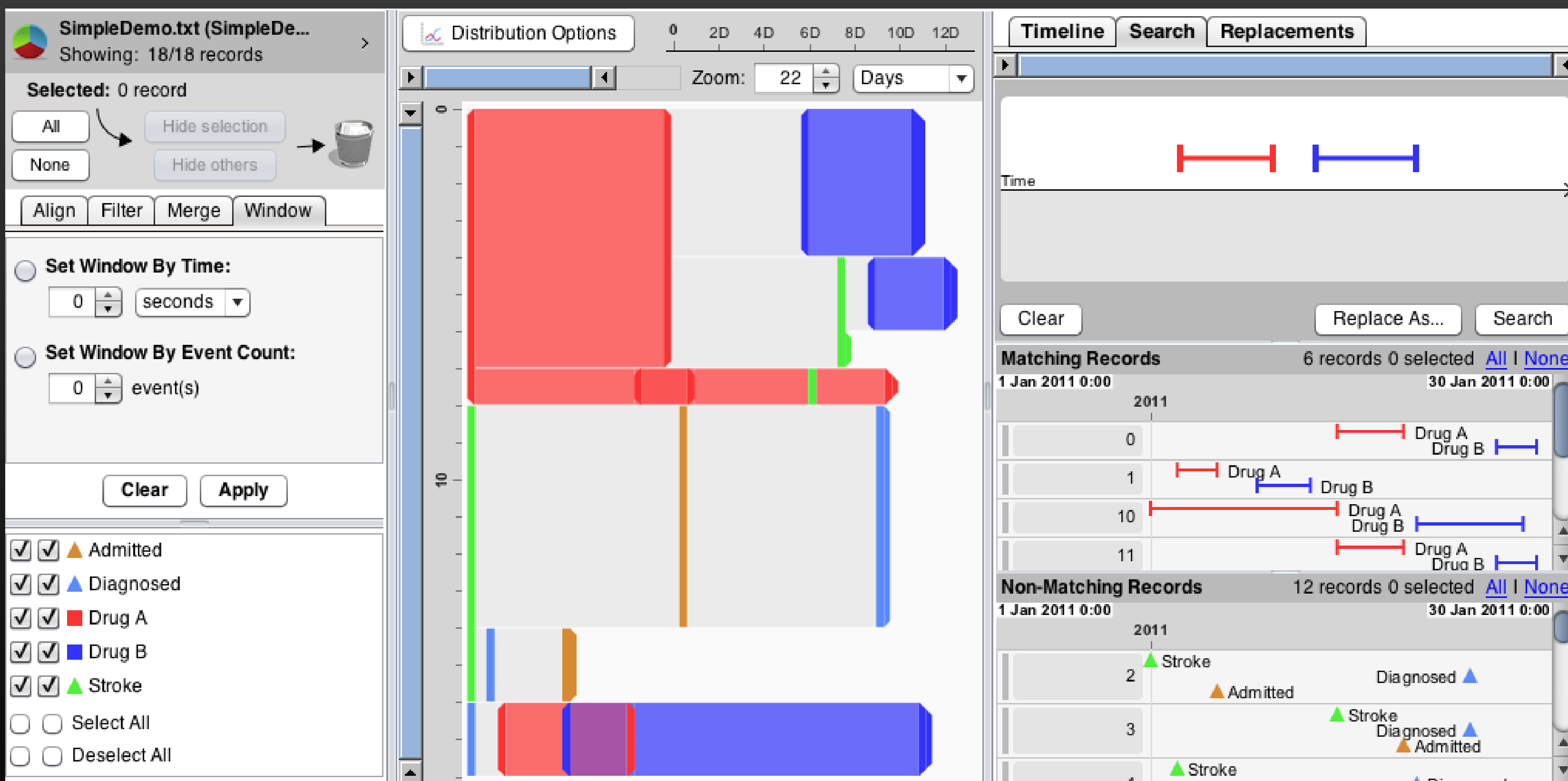
http://research.microsoft.com/logan

Mary Czerwinski, Steven Drucker, Rob DeLine, Danyel Fisher, Kael Rowan, Microsoft Research

Alper Sarikaya, University of Wisconsin-Madison

Emanuel Zraggen, Brown University

EventFlow - Univ. of Maryland



http://hcil.umd.edu/eventflow/

1. Originally, testing was its own discipline; today it is a skill.
   Today, data science is its own discipline; tomorrow will it be a skill?

2. Both experts and non-experts want to get answers from data.
   Conclusions we draw from data are having an increasingly large effect on the world.

We need tools for data science to make it *easy* to get answers from data with *high confidence*.