In [16]:

```
# Data analysis packages:
import pandas as pd
import numpy as np

# Visualization packages:
import seaborn as sns
import matplotlib.pyplot as plt
plt.rcParams['figure.figsize'] = [6, 2]

%matplotlib inline

#Utility packages:
import multiprocessing
import email
```

In [17]:

```
df = pd.read_csv('../data/emails.csv')
df.shape
```

Out[17]:

```
(517401, 2)
```

In [18]:

```
print(df.loc[1]['message'])
```

```
Message-ID: <15464986.1075855378456.JavaMail.evans@thyme>
Date: Fri, 4 May 2001 13:51:00 -0700 (PDT)
From: phillip.allen@enron.com
To: john.lavorato@enron.com
Subject: Re:
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Phillip K Allen
X-To: John J Lavorato <John J Lavorato/ENRON@enronXgate@ENRON>
X-cc:
X-bcc:
X-Folder: \Phillip_Allen_Jan2002_1\Allen, Phillip K.\'Sent Mail
X-Origin: Allen-P
X-FileName: pallen (Non-Privileged).pst

Traveling to have a business meeting takes the fun out of the trip.  Especi
ally if you have to prepare a presentation.  I would suggest holding the bu
siness plan meetings here then take a trip without any formal business meet
ings.  I would even try and get some honest opinions on whether a trip is e
ven desired or necessary.

As far as the business meetings, I think it would be more productive to try
and stimulate discussions across the different groups about what is working
and what is not.  Too often the presenter speaks and the others are quiet j
ust waiting for their turn.   The meetings might be better if held in a rou
nd table discussion format.

My suggestion for where to go is Austin.  Play golf and rent a ski boat and
jet ski's.  Flying somewhere takes too much time.
```

In [19]:

```
message = df.loc[1]['message']
em = email.message_from_string(message)
em.items()
```

Out[19]:

```
[('Message-ID', '<15464986.1075855378456.JavaMail.evans@thyme>'),
 ('Date', 'Fri, 4 May 2001 13:51:00 -0700 (PDT)'),
 ('From', 'phillip.allen@enron.com'),
 ('To', 'john.lavorato@enron.com'),
 ('Subject', 'Re:'),
 ('Mime-Version', '1.0'),
 ('Content-Type', 'text/plain; charset=us-ascii'),
 ('Content-Transfer-Encoding', '7bit'),
 ('X-From', 'Phillip K Allen'),
 ('X-To', 'John J Lavorato <John J Lavorato/ENRON@enronXgate@ENRON>'),
 ('X-cc', ''),
 ('X-bcc', ''),
 ('X-Folder', "\\Phillip_Allen_Jan2002_1\\Allen, Phillip K.\\'Sent Mail"),
 ('X-Origin', 'Allen-P'),
 ('X-FileName', 'pallen (Non-Privileged).pst')]
```

In [20]:

```python
def get_field(field, messages):
    column = []
    for message in messages:
        em = email.message_from_string(message)
        column.append(em.get(field))
    return column
```

In [21]:

```python
df['X-From'] = get_field("X-From", df['message'])
df['X-To'] = get_field("X-To", df['message'])
df['X-cc'] = get_field("X-cc", df['message'])
df['X-Subject'] = get_field("Subject", df['message'])
```

In [22]:

```python
print(df.head(10))
```

```
                       file  \
0     allen-p/_sent_mail/1.   
1    allen-p/_sent_mail/10.   
2   allen-p/_sent_mail/100.   
3  allen-p/_sent_mail/1000.   
4  allen-p/_sent_mail/1001.   
5  allen-p/_sent_mail/1002.   
6  allen-p/_sent_mail/1003.   
7  allen-p/_sent_mail/1004.   
8   allen-p/_sent_mail/101.   
9   allen-p/_sent_mail/102.   

                                             message           X-From  \
0  Message-ID: <18782981.1075855378110.JavaMail.e...  Phillip K Allen   
1  Message-ID: <15464986.1075855378456.JavaMail.e...  Phillip K Allen   
2  Message-ID: <24216240.1075855687451.JavaMail.e...  Phillip K Allen   
3  Message-ID: <13505866.1075863688222.JavaMail.e...  Phillip K Allen   
4  Message-ID: <30922949.1075863688243.JavaMail.e...  Phillip K Allen   
5  Message-ID: <30965995.1075863688265.JavaMail.e...  Phillip K Allen   
6  Message-ID: <16254169.1075863688286.JavaMail.e...  Phillip K Allen   
7  Message-ID: <17189699.1075863688308.JavaMail.e...  Phillip K Allen   
8  Message-ID: <20641191.1075855687472.JavaMail.e...  Phillip K Allen   
9  Message-ID: <30795301.1075855687494.JavaMail.e...  Phillip K Allen   

                                                X-To X-cc  \
0           Tim Belden <Tim Belden/Enron@EnronXGate>        
1  John J Lavorato <John J Lavorato/ENRON@enronXg...        
2                                   Leah Van Arsdall        
3                                      Randall L Gay        
4                                         Greg Piper        
5                                         Greg Piper        
6             david.l.johnson@enron.com, John Shafer        
7                                     Joyce Teixeira        
8                                         Mark Scott        
9                                    zimam@enron.com        

                                           X-Subject  
0                                                     
1                                                Re:  
2                                           Re: test  
3                                                     
4                                          Re: Hello  
5                                          Re: Hello  
6                                                     
7                       Re: PRC review - phone calls  
8                     Re: High Speed Internet Access  
9  FW: fixed forward or other Collar floor gas pr...  
```

In [24]:

```
# Drop rows with any empty cells
df.dropna(
    axis=0,
    how='any',
    subset=None,
    inplace=True
)
```

In [ ]:

```
```

In [ ]:

```
```

In [25]:

```
df.nunique()
```

Out[25]:

```
file         517372
message      517372
X-From        27980
X-To          73552
X-cc          33701
X-Subject    159277
dtype: int64
```

In [26]:

```
df['X-From'] = pd.factorize(df['X-From'])[0]
df['X-To'] = pd.factorize(df['X-To'])[0]
df['X-cc'] = pd.factorize(df['X-cc'])[0]
print(df.head(10))
```

```
                       file  \
0     allen-p/_sent_mail/1.
1    allen-p/_sent_mail/10.
2   allen-p/_sent_mail/100.
3  allen-p/_sent_mail/1000.
4  allen-p/_sent_mail/1001.
5  allen-p/_sent_mail/1002.
6  allen-p/_sent_mail/1003.
7  allen-p/_sent_mail/1004.
8   allen-p/_sent_mail/101.
9   allen-p/_sent_mail/102.

                                             message  X-From  X-To  X-cc  \
0  Message-ID: <18782981.1075855378110.JavaMail.e...       0     0     0
1  Message-ID: <15464986.1075855378456.JavaMail.e...       0     1     0
2  Message-ID: <24216240.1075855687451.JavaMail.e...       0     2     0
3  Message-ID: <13505866.1075863688222.JavaMail.e...       0     3     0
4  Message-ID: <30922949.1075863688243.JavaMail.e...       0     4     0
5  Message-ID: <30965995.1075863688265.JavaMail.e...       0     4     0
6  Message-ID: <16254169.1075863688286.JavaMail.e...       0     5     0
7  Message-ID: <17189699.1075863688308.JavaMail.e...       0     6     0
8  Message-ID: <20641191.1075855687472.JavaMail.e...       0     7     0
9  Message-ID: <30795301.1075855687494.JavaMail.e...       0     8     0

                                           X-Subject
0
1                                                Re:
2                                           Re: test
3
4                                          Re: Hello
5                                          Re: Hello
6
7                       Re: PRC review - phone calls
8                     Re: High Speed Internet Access
9  FW: fixed forward or other Collar floor gas pr...
```

In [28]:

```
df.nunique()
```

Out[28]:

```
file         517372
message      517372
X-From        27980
X-To          73552
X-cc          33701
X-Subject    159277
dtype: int64
```

In [11]:
```
with pd.option_context('display.max_rows', 5,
                       'display.max_columns', None,
                       'display.width', 1000,
                       'display.precision', 3,
                       'display.colheader_justify', 'center'):
    display(df)
```

	file	message	X-From	X-To	X-cc	X-Su
0	allen-p/_sent_mail/1.	Message-ID: <18782981.1075855378110.JavaMail.e...	0	0	0	
1	allen-p/_sent_mail/10.	Message-ID: <15464986.1075855378456.JavaMail.e...	0	1	0	
...	...	...	...	...	...	
517399	zufferli-j/sent_items/98.	Message-ID: <22052556.1075842030013.JavaMail.e...	7114	414	0	C Analyst/Ass
517400	zufferli-j/sent_items/99.	Message-ID: <28618979.1075842030037.JavaMail.e...	7114	73483	0	RE: ali's

517372 rows × 6 columns

In [14]:
```
df.drop(['file','message','X-Subject'], axis=1, inplace=True)
df.astype({'X-From':'int'})
df.astype({'X-To':'int'})
df.astype({'X-cc':'int'})
```

Out[14]:

	X-From	X-To	X-cc
0	0	0	0
1	0	1	0
2	0	2	0
3	0	3	0
4	0	4	0
...	...	...	...
517396	7114	2777	0
517397	7114	422	0
517398	7114	9428	0
517399	7114	414	0
517400	7114	73483	0

517372 rows × 3 columns

In [15]:
```
df.to_csv('../data/clean_emails.csv',index=False, header=False)
```

In [ ]:

In [ ]: