AUTOMATING TASKS WITH THE REDCAP API

William Santo, UCSF School of Dentistry

Aug 29, 2018 CAPS Presentation

WHAT TO EXPECT TODAY

- Technical presentation, including
  - Intro to Python
  - Working with Redcap data and the Redcapy library
  - Sending SMS and email with Redcap data
- High-level discussions of:
  - Python
  - Redcap token acquisition and handling
  - Production environment
  - Job scheduling
  - Secure server communication

PRESENTATION MATERIALS

- All slides and notebooks are available at:
  - https://github.com/billsanto/caps_redcap_api

BACKGROUND

- Our team functions as a Data Coordinating Center for several institutions collaborating as part of: Multidisciplinary and Collaborative Research Consortium to Reduce Oral Health Disparities in Children: A Multilevel Approach (UH2/UH3)
- Announcement: https://www.nih.gov/news-events/news-releases/nih-funds-consortium-childhood-oral-health-disparities-research
- NIH Grant U01DE025507
- Services provided (NIH source):
  - clinical and scientific expertise, including optimizing study design, measurement, and data analysis
  - informatics and project management, including secure data collection and management systems, data quality assurance/control, project and safety monitoring, and progress reporting
  - coordination of distance-collaborating teams

BENEFITS OF AUTOMATION

- Consistent results
- Increased productivity
- 24x7 Availability
- Scalability

DRAWBACKS OF AUTOMATION

- More development effort and code complexity
- Broader technical skillset required
- After hours monitoring and troubleshooting
- Server costs
- Server maintenance
- Changes to APIs must be monitored

- https://www.python.org/

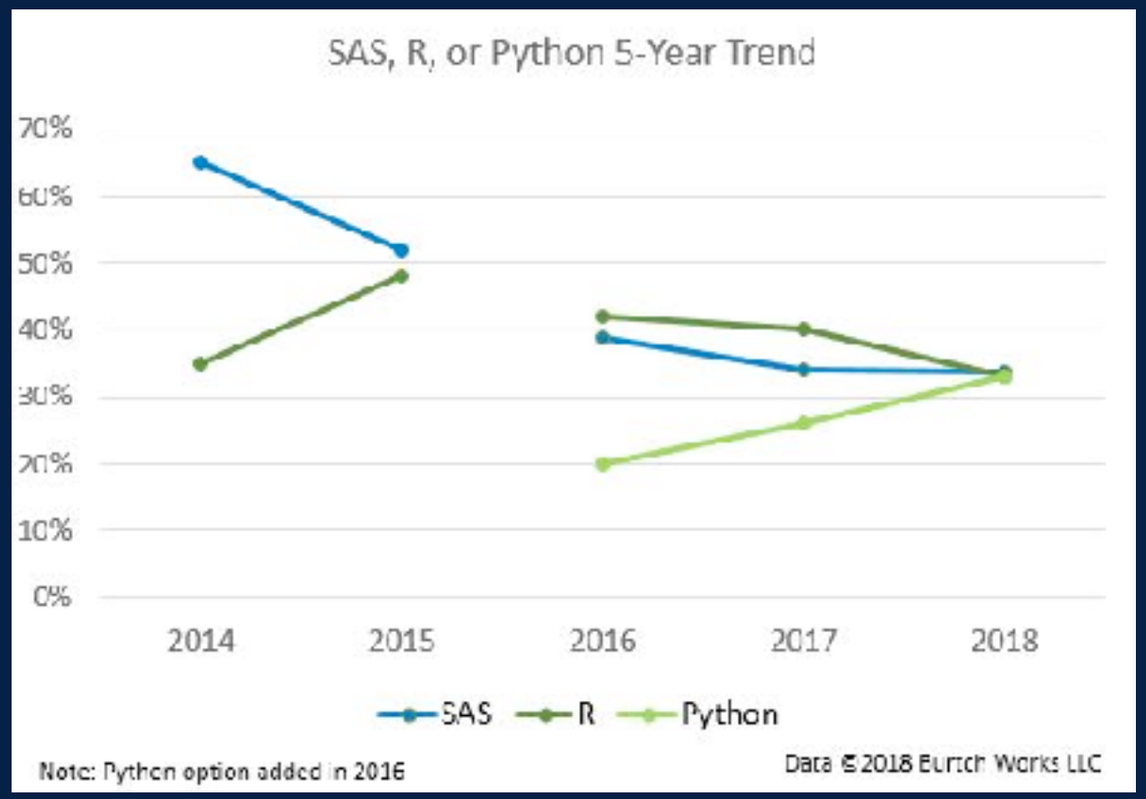
PYTHON BACKGROUND

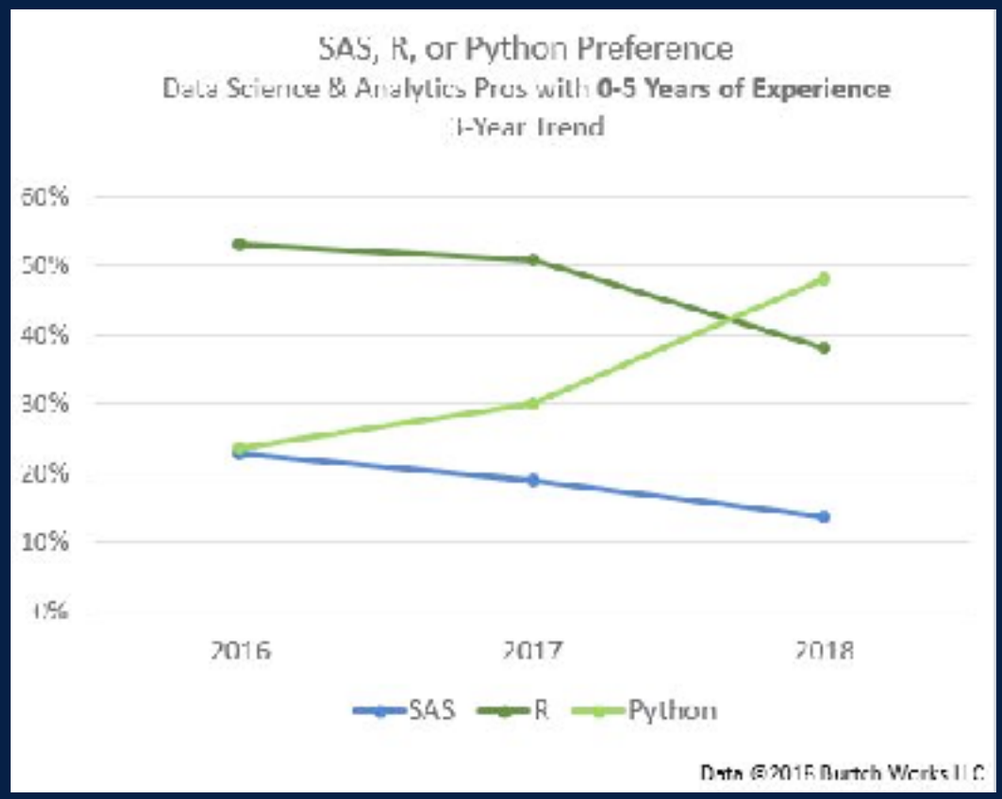
- General purpose, not a domain specific language
  - https://stackoverflow.com/a/39098766
- Mature with a wider user base
- Like R, offers a vast array of available libraries, with contributors ranging from individuals to companies like google, facebook, airbnb, twitter, etc.
- Cross platform, open source
- Scripting language, unlike Java or C
- “Relatively” easy to learn
- No annual licensing fees
- Ease of deploying code from prototype to production

PYTHON EDUCATION

- In 2014, Python displaced Java as the top intro language course at universities
- Recent evolution of online learning in MOOCs is contributing to a groundswell of people with Python experience
- Approximate count of available MOOCs (in English, filtered by CompSci, Math, DataSci, and Programming categories)
  - Python: 194
  - Java: 132
  - R: 124
  - SAS: 5

TRENDS: PREFERENCES

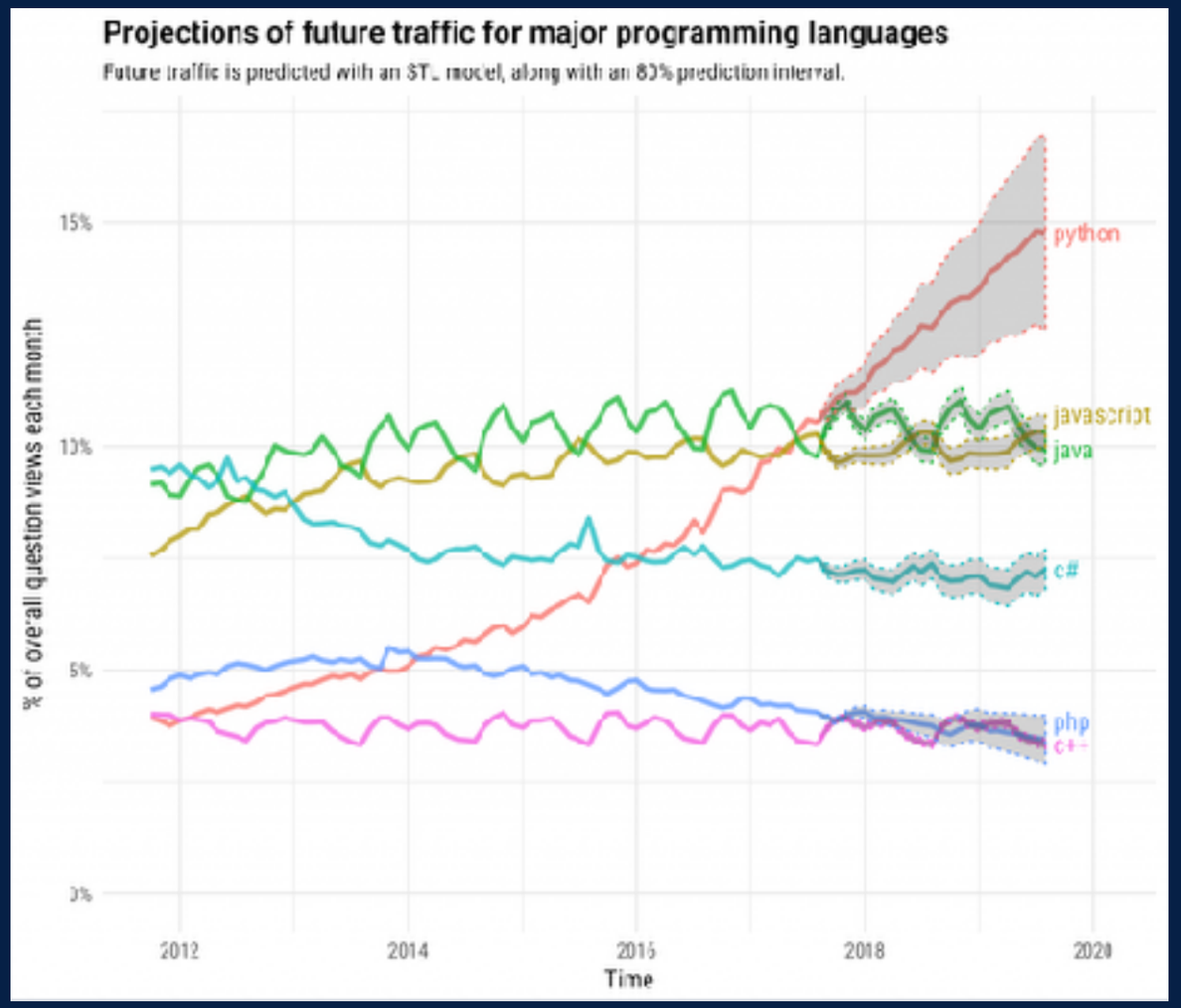




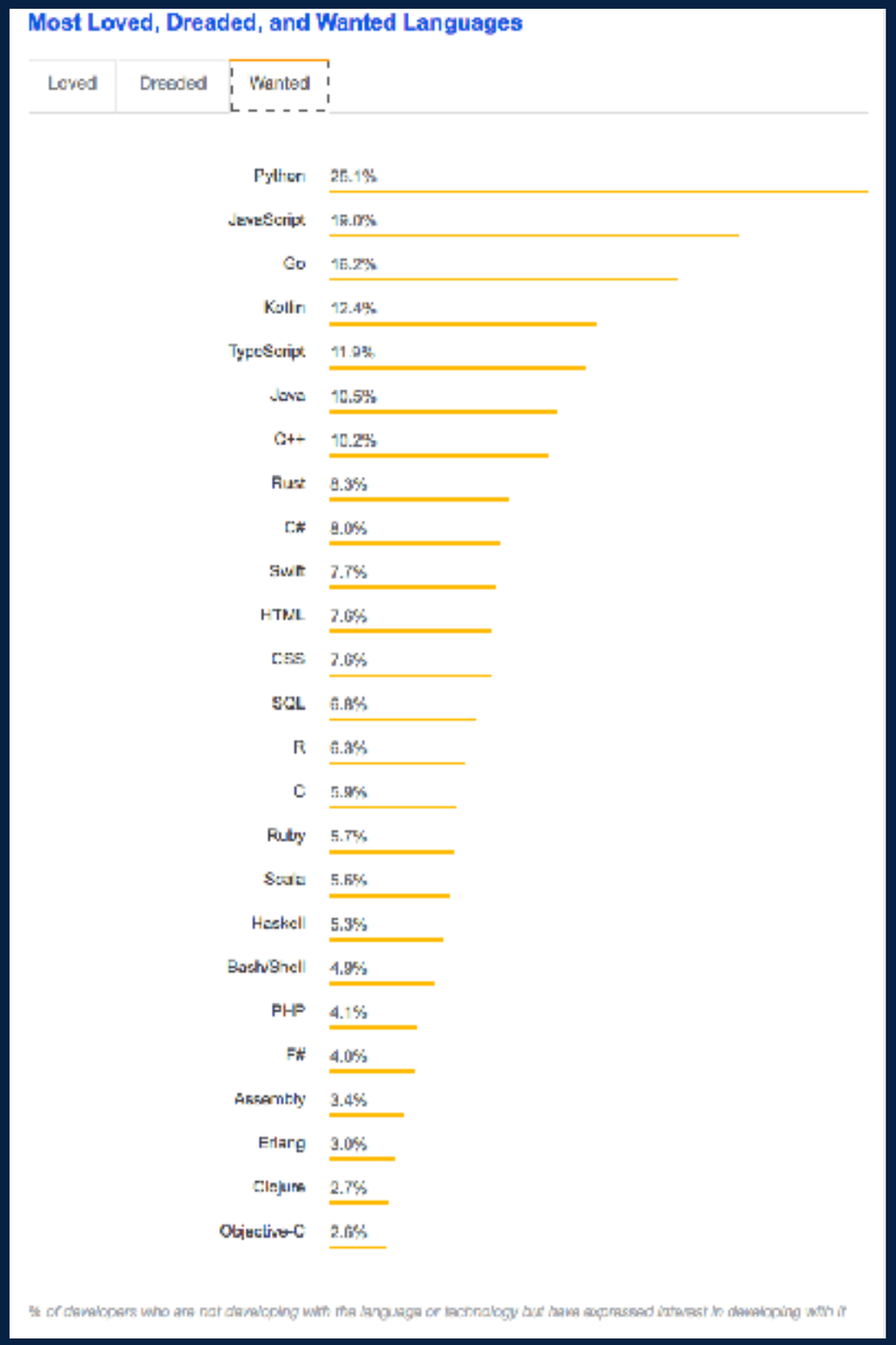
Question: Which do you prefer to use – SAS, R, or Python?

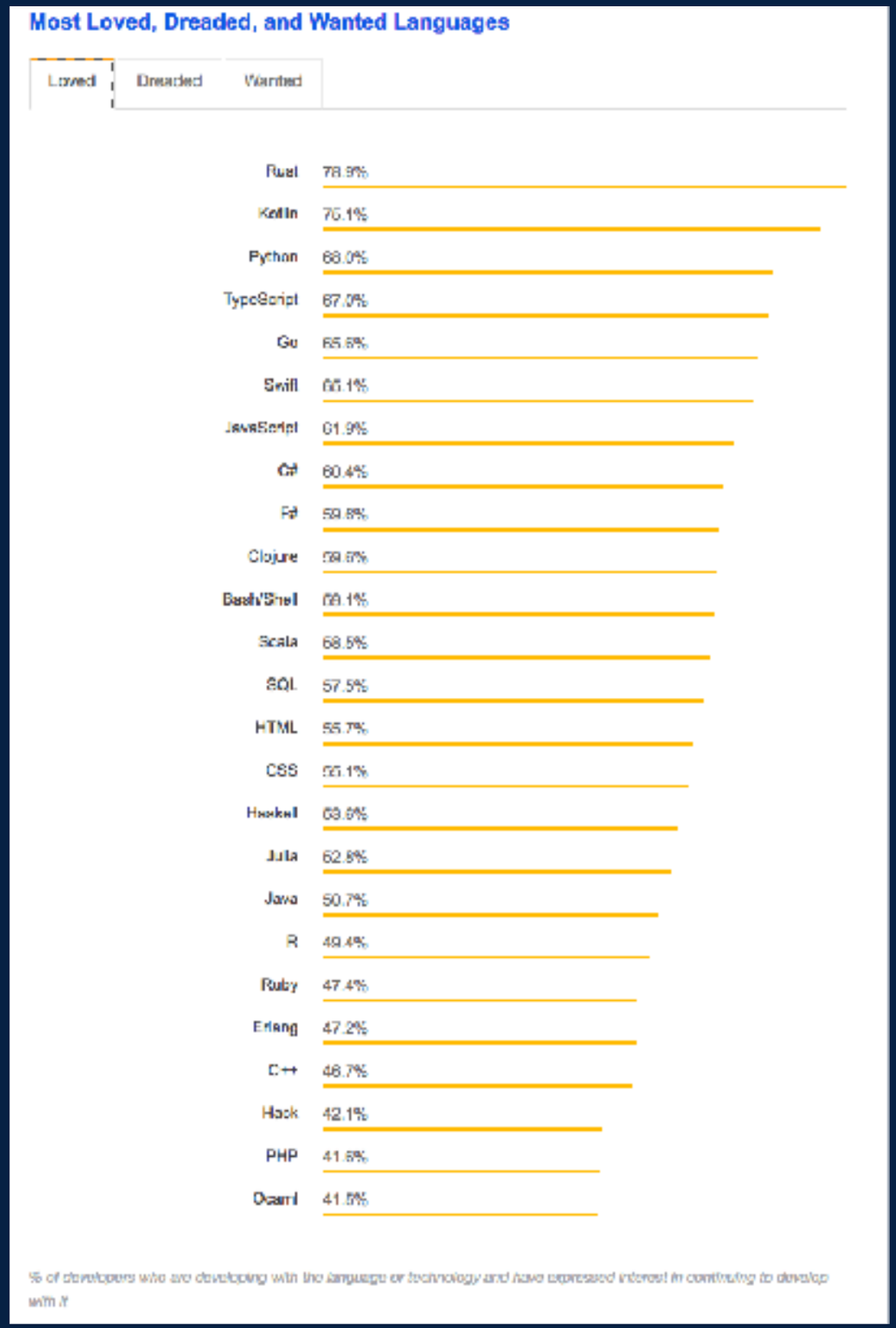
Source: https://www.burtchworks.com/2018/07/16/2018-sas-r-or-python-survey-results-which-do-data-scientists-analytics-pros-prefer/

TRENDS: STACK OVERFLOW QUESTION VIEWS



Source: https://stackoverflow.blog/2017/09/06/incredible-growth-python/





- Source: Stack Overflow 2018 Developer Survey

PYTHON VERSIONS

- 2.x vs. 3.x
- Python 2.7 support expires in 2020
- A few key differences

PYTHON TRAINING

- UCSF has a Data Science Initiative with regular workshops that cover various topics:
  - https://www.library.ucsf.edu/data-science/
- But consider online classes, with assignments and online forums for support
  - https://www.coursera.org/specializations/python
  - https://developers.google.com/edu/python/
  - https://www.codecademy.com/learn/learn-python
  - https://www.udemy.com/complete-python-bootcamp/

PYTHON ENVIRONMENTS

- Macs and Linux systems ship with built-in Python
- However, using a virtual environment is recommended
  - virtualenv (general purpose)
  - Enthought
  - Anaconda (recommended)
- Why use a virtual environment?
  - Permits each project (or group of projects) to use its own combination of libraries
  - Allows for testing of new libraries in their own environment

LIBRARY MANAGEMENT

- pip
  - Show installed libraries: pip freeze
  - Install new library: pip install pandas
  - Update existing library: pip install pandas —upgrade
- conda (for Anaconda users)
  - Show installed libraries: conda list
  - Install new library: conda install pandas
  - Update existing library: conda update pandas
  - https://www.quora.com/How-do-I-install-Python-packages-in-Anaconda

JUPYTER NOTEBOOKS

- Browser-based coding application
- Well suited for interactive work, data exploration, analysis, visualizations
- https://github.com/jupyter/jupyter/wiki/A-gallery-of-interesting-Jupyter-Notebooks#introductory-tutorials
- Revolutionary technology
  - https://www.forbes.com/sites/forbestechcouncil/2018/08/17/here-come-the-notebooks/
- Notebooks have .ipynb extension
- Jupyter Lab recently released
  - 10 minute youtube intro
  - Supports over 100 languages, including R and SAS

PANDAS LIBRARY

- DataFrames, similar to R DataFrames
- https://pandas.pydata.org/
- Widely adopted by the data science community
- Article: Python explosion blamed on pandas

API

- API = Application Programming Interface
- In general, it is a set of clearly defined methods of communication between components.
- APIs are applicable to web servers, operating systems, data structures, hardware, etc.
- In our context, we are concerned with how to interact with a web server using software that we write.
- Server developers have to tell us how what commands their system understands via their documentation or code.

NON-TECHNICAL EXAMPLE

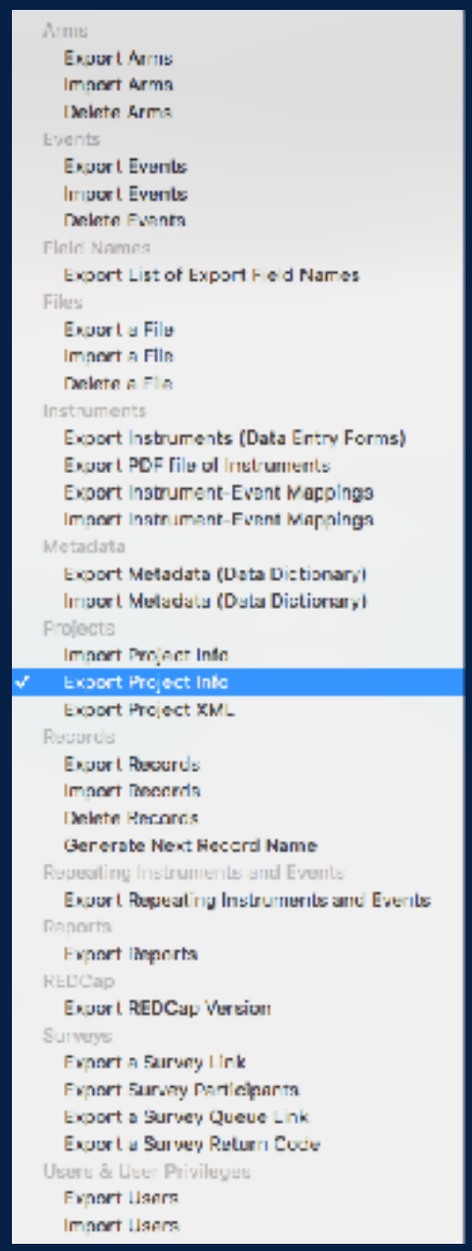
- Bob goes to Mel's Diner and orders ceviche
- Flo informs Bob that is not on the menu
- Flo instructs Bob to read the menu and choose a meal from it
- Bob studies the menu and orders a burger and fries
- Flo sends order to Mel
- Mel prepares food and informs Flo when order is ready
- Flo received meal from Mel and delivers prepared food to Bob
- Bob consumes the meal
- In API terms, Bob is the analyst hungry for data, Flo is the web server, Mel is the database, and the menu is the API documentation.
    - If it is not documented, the data typically doesn't get served

REDCAP API

- Is a RESTFUL API. In other words, data is exchanged via the HTTP protocol
  - Reference: https://restapitutorial.com/lessons/httpmethods.html
- API Documentation: https://redcap.ucsf.edu/api/help/?content=default
- SAS Example: https://redcap.ucsf.edu/announcement/SAS_API_workaround.pdf

REDCAP API FEATURES

- Permits importing and exporting of any item listed in the screenshot
- Not all data is available via the API, such as logging or calendar data



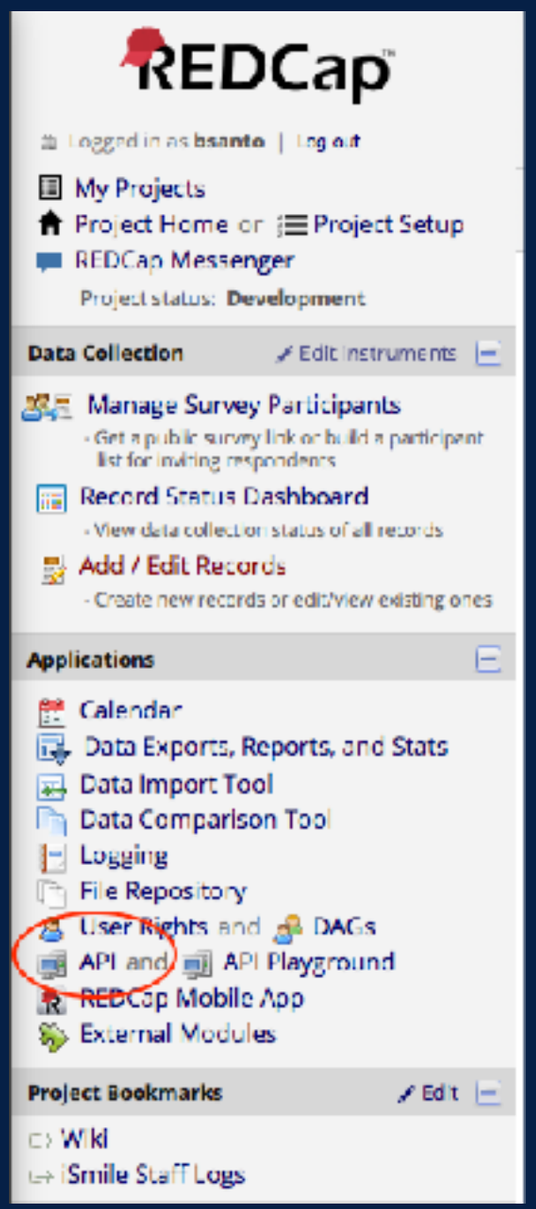
REDCAP API

- On a per-project basis, you request an API token from the Redcap support team.
  - A token (32 or 64 character string) is unique to a given project
  - It takes from a few minutes to the next business day to receive a token
  - Every copy of an existing project requires a new token if needed
- Your Redcap project manager can grant you either or both
  - import privileges
  - export privileges

ACQUIRING A TOKEN

- The project admin assigns API privileges to the user
- The user clicks the API hyperlink in the left frame (page refresh is required after the admin grants privilege)
- The user clicks the button to request a token
- The Redcap administrator receives the request and clicks something to make it happen
- Eventually, the user receives an email indicating the token was granted
- The user returns to the API page to see the token

ACQUIRING A TOKEN



- The project admin assigns API privileges to the user
- The user clicks the newly created API hyperlink in the left frame (page refresh is required after the admin grants privilege)
- The user clicks the button on the API page to request a token
- The Redcap administrator receives the request and clicks something to make it happen
- Eventually, the user receives an email indicating the token was granted
- The user returns to the API page to see the token

API LOGGING

08/07/2018 11:58am	bsanto	Data Export (API)	export_format: JSON, rawOrLabel: label
08/07/2018 11:58am	bsanto	Data Export (API)	export_format: JSON, rawOrLabel: raw
08/07/2018 10:19am	bsanto	Manage/Design	Edit project field
08/07/2018 10:17am	bsanto	Updated Record (API) 2216 (Final Data Collection)	vt_last_timestamp = '2018-08-07 10:10:27', vt_pending_visit = ''
08/07/2018 10:17am	bsanto	Updated Record (API) 2216 (6 Month Data Collection)	vt_last_timestamp = '2018-08-07 10:10:27', vt_pending_visit = ''
08/07/2018 10:17am	bsanto	Updated Record (API) 2190 (Final Data Collection)	vt_last_timestamp = '2018-08-07 10:10:27', vt_pending_visit = ''

- Every API call is logged
- Each imported field is logged if the field value changes
- Imported field values are not logged if the field values are not altered
- If you do not have access to the log, request it from your Redcap project admin

WHAT NOT TO DO WITH A TOKEN?

- Do NOT store a token in your source code
  - Eventually, you will forget that you placed it there
  - You cannot share your code without revealing your token
  - Once you have multiple programs for a given project, a token change requires multiple source code updates
- If source is under version control, the version control system now has the token
  - If the source gets published, then token gets published to the world

WHAT TO DO WITH A TOKEN?

- Store in an environment variable
  - https://www.twilio.com/blog/2017/01/how-to-set-environment-variables.html
- Store in a file which can be imported into code
  - Advisable to not store in the same folder as project source code under version control
- Either of the above stores in plain text, so special attention needs to be paid to your environment variables and those files to protect them
- R tips: http://happygitwithr.com/api-tokens.html
- Some secure storage methods: https://medium.freecodecamp.org/how-to-securely-store-api-keys-4ff3ea19ebda

API CAVEATS

- Tokens grant anyone in possession of the token the privileges accorded to that token, so it must be safeguarded and not shared
- Export privileges provide access to all API data
- Import privileges can instantly corrupt or destroy project data, so project admins should only grant this on an as-needed basis per project
- Redcap integrity constraints prevent some types of flawed import attempts from succeeding, but not all
  - e.g., you can import data for a record id into an incorrect arm
- Tip: Evaluate code which imports data on a recent copy of your production Redcap project until convinced it is working correctly
- Be PARANOID about which token is in use when testing import code

CURL

- curl facilitates transfer of data to/from a server
- https://curl.haxx.se/docs/manpage.html
- In a web server context, this allows you to see raw data that you would otherwise see rendered in a browser
- On a Mac, open Terminal and type: curl ‘http://www.google.com' to see an example
- PCs do not have curl built-in, but is available as part of other packages.
- Twilio API message status command example:

```
curl 'https://api.twilio.com/2010-04-01/Accounts/[Twilio Account SID]/Messages/[Message SID].json' -u[Twilio Account SID]:[AuthToken]
```

BASH SHELL & SCRIPTS

- On a Mac or Unix/Linux system, a terminal window operates in a shell environment, where commands and scripts can be executed interactively
  - A popular choice of shell is bash
  - Ref: http://www.tutorialspoint.com/unix/unix-shell.htm
- Windows command prompt: https://www.wikihow.com/Open-the-Command-Prompt-in-Windows
- Shell Scripts:
  - Programs used to execute a sequence of commands
    - Ability to write loops, conditionals, etc.
  - http://www.linuxcommand.org/lc3_writing_shell_scripts.php
  - Windows/DOS equivalent: https://www.howtogeek.com/263177/how-to-write-a-batch-script-on-windows/

USING NOTEBOOKS IN PRODUCTION

- Although it is intended for interactive use, under certain scenarios it can be useful for production, such as
  - Low frequency batch use
  - Complete reproducibility of inline execution is desired
  - Enhanced troubleshooting of executed code is needed

PRODUCTION NOTEBOOKS

- Launch notebook from a python script
- Pass key parameters from the script to the notebook's first cell using nbparameterise library
- https://github.com/takluyver/nbparameterise
- Requires organizing notebook code to accommodate
  - e.g., Parameter indicating whether to use a import data to a production or development project

NOTEBOOKS OR SCRIPTS?

- When using a notebook, there is overhead in creating a script to pass arguments to the notebook
- Scripts can leverage much more powerful debugging tools when using a modern IDE such as Visual Studio or PyCharm
- A notebook, if used for development, can be exported to a script
- Debugging scripts requires thoughtful implementation of logging to facilitate debugging of a failed job, but a notebook can be designed to display large amounts of useful data for troubleshooting
- Scripts are far more widely used for production purposes
- Notebooks would be inefficient or unsuitable for many applications
- Scripts are more amenable to use under version control. However, nbdiff-web can be used to visualize code changes.
- Markup documentation can make notebooks easier to comprehend
- Notebooks lack the refactoring capabilities that most IDEs possess

SCHEDULED JOB WORKFLOW

- Typically, a job is scheduled via cron
- cron executes python script and passes command line arguments (if any) to script
- If notebook is involved, python script launches notebook and passes arguments to notebook
  - Python script saves timestamped copy of notebook, complete with inline output
- Job results saved to new local mail message, including any printed output from python script

JOB SCHEDULING

- Unix/Linux:
  - crontab
    - Define a schedule to launch program
    - https://www.computerhope.com/unix/ucrontab.htm
  - at
    - Schedule a program to execute once
    - https://tecadmin.net/one-time-task-scheduling-using-at-commad-in-linux/
- Windows
  - https://www.dummies.com/computers/pcs/how-to-open-windows-task-scheduler/

CRONTAB SAMPLE JOBS

```
# ISMILE
00,15,30,45 05-22 * * 1,2,3,4,5 source $HOME/.bash_profile && cd $HOME/dev/python/ismile && python $HOME/dev/python/ismile/ismile_poll_participants.py 1 0 1
04 03,07,11,15,19,23 * * 0,6 source $HOME/.bash_profile && cd $HOME/dev/python/ismile && python $HOME/dev/python/ismile/ismile_poll_participants.py 1 0 1
00 00,04 * * * source $HOME/.bash_profile && cd $HOME/dev/python/ismile/ && python $HOME/dev/python/ismile/ismile_compute_visit_dates.py 0 1 1 1 1
05 00,04 * * * source $HOME/.bash_profile && cd $HOME/dev/python/ismile/ && python $HOME/dev/python/ismile/ismile_compute_visit_dates.py 1 0 1 1 1
16 00,04 * * * source $HOME/.bash_profile && cd $HOME/dev/python/ismile/ && python $HOME/dev/python/ismile/ismile_check_agile_status.py 1 0 0 1 1 0
21 00,04 * * * source $HOME/.bash_profile && cd $HOME/dev/python/ismile/ && python $HOME/dev/python/ismile/ismile_check_agile_status.py 0 1 0 1 1 0
50 13 * * * source $HOME/.bash_profile && cd $HOME/dev/python/ismile/ && python $HOME/dev/python/ismile/ismile_send_surveys.py "REDCAP_API_ISMILE_PROD" "REDCAP_API_ISM
LE_PROD" 1 1 "" 0 1 1 0 0
```

- Set schedule for minute, hour, days of week
- Source user environment and set working directory before launching code
- Run command and arguments
  - In these examples, run python and pass program location and its arguments as arguments to python

TWILIO API

- Permits sending and receiving SMS directly from code
- Alternately, why not send SMS using the Redcap/Twilio integration?
  - Redcap Twilio messages are associated with a specific survey mapped to an event, and some study requirements are not consistent with the linkage to a survey
- Redcap 8.5 reportedly stopped appending Survey URLs, but in practice standard text and survey URLs are still being appended to messages.
  - Namely, “To begin the survey, go to https://redcap.ucsf.edu/?abcdefg"
- No convenient means to extra extract message metadata from Redcap API
- No ability to re-evaluate delivery status of failed messages

TWILIO ACCOUNT

- Requires account setup and credit card on file
- Free trial number available
- Lease phone numbers for $1 a month
- The account is charged on demand when the account falls below $10, by default
- Twilio offers free test phone numbers, but they have limited ability to conduct full tests of code
- Each SMS costs $0.0075

TWILIO USAGE

- Each phone number must be provisioned before use
- Each Twilio account has an SID and token, which are used to access their API
- Each SMS generates a unique SID, which can be used to check status of a message
- Twilio offers a Python library, installed by the command: pip install twilio
- Note that the twilio library upgrade from version 5 to 6 made breaking changes, so testing of any major upgrade is highly advised
- If you have a Twilio number already assigned to a Redcap project (configured by a Redcap admin), it is advisable to purchase and use a second phone number for communicating directly with Redcap participants

TWILIO SENDING/RECEIVING

- Sending a SMS is relatively simple, but most of work is overhead in preparing data, checking delivery status, handling edge cases and missing data, and recording results.
- Receiving SMS replies requires significantly more effort to setup and maintain a web server that can receive and process those replies
- Such a web server would have to be installed on a server on the UCSF external network, or perhaps in the future on a secure AWS instance. The code in turn needs to parse replies and react accordingly, contingent on study requirements
- Alternately, you can configure auto-replies to refer replies to an external contact using TwiML: https://www.twilio.com/docs/sms/twiml
- Ensure your code does not permit runaway loops, which would not only be costly, but also annoy study participants

TWILIO LONG CODE LIMITATIONS

- Limited to no more than 1 message per second to not be considered a high volume user
  - Otherwise, you may need to apply for a short code number, a process which requires auditing, special arrangements, and can take up to several months for approval
- For use with Redcap studies, most probably don't require high volume, time-sensitive messaging
- Occasionally, message deliveries may fail with a 30007 error due to triggering a carrier's spam detection when using a long code number
  - There is no good recourse when this occurs
  - It is probably helpful to configure messages to send at a frequency lower than 1 per second, and maybe to customize content in each message
  - However, carriers, for obvious reasons, do not publicize their spam detection rules
- Currently, our highest volume project routinely sends between 50-100 messages to study participants during code execution, several times per week.
  - Only one participant persistently had their carrier reject messages, but this participant was not randomized and was otherwise not reachable after the screening visit.
- https://www.twilio.com/docs/sms/send-messages#a-note-on-rate-limiting

TWILIO SHORT CODES

- Approval process is 2-3 months
- Short codes can send up to 100 messages per second
- https://support.twilio.com/hc/en-us/articles/223182068-What-is-a-Messaging-Short-Code-
- https://usshortcodes.com/faqs
- Handbook: https://www.wmcglobal.com/ctia-short-code-handbook
- Note that some minor carriers will block short codes, which may require intervention on the part of the message recipient to unblock. This should affect only a small percentage of users, but numbers should be tested to receive them, and it may require time and special handling that may be beyond the scope of a research assistant's abilities
- In one of our projects, the messaging partner (who handles direct participant intervention via SMS) uses short codes. The number validation workflow looks like: https://www.lucidchart.com/documents/view/d80abb39-fe4f-4232-ad34-01f959ab33b0/0
- Short codes should not be subject to spam detection

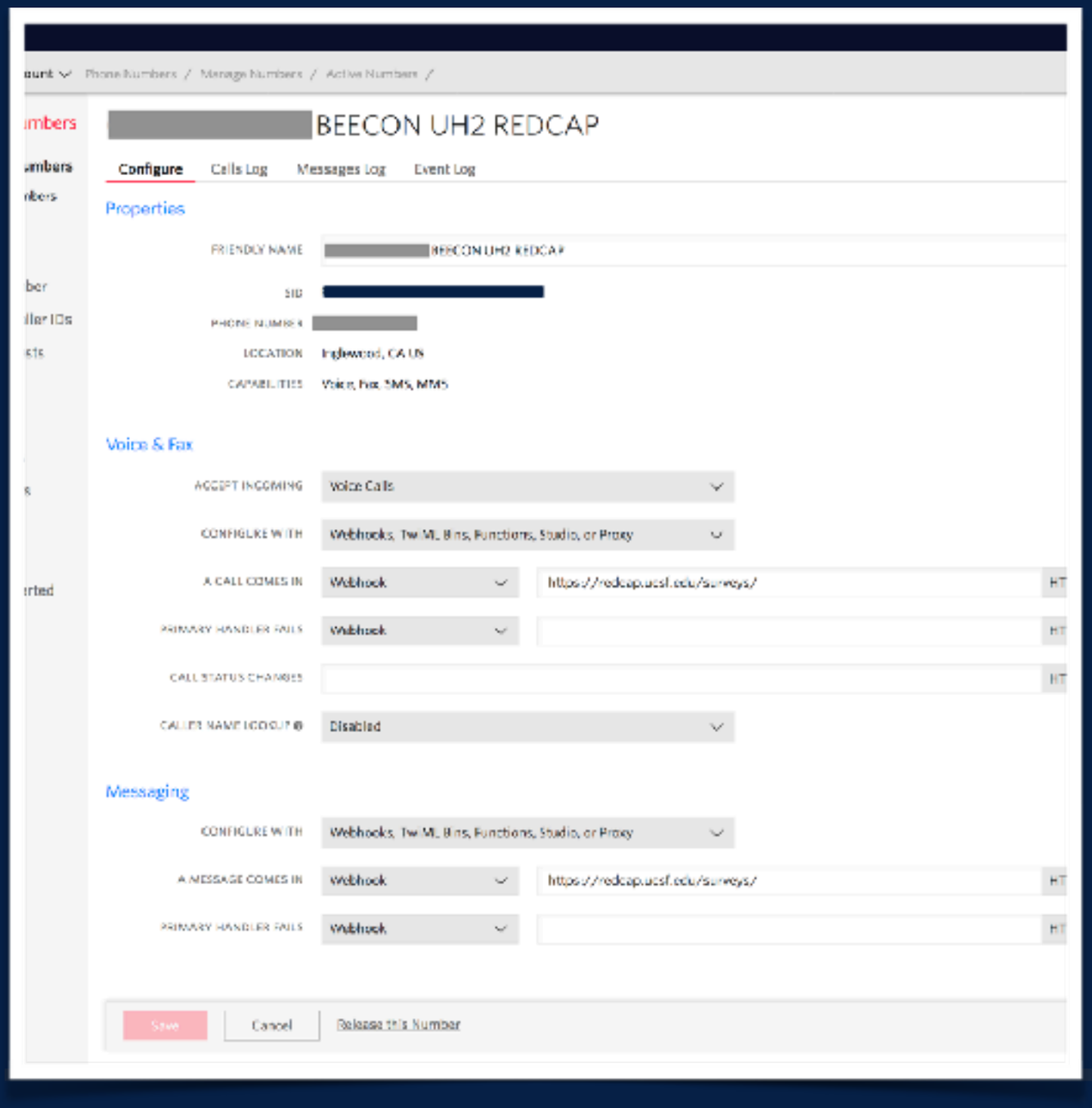
CONFIGURING TWILIO REPLIES

- To configure a number with an auto-reply, select TwiML under the Messaging section, click + to create a new message, and compose an appropriate message in a properly formed XML document

- Once provisioned, it is easy to test how a number responds to SMS by initiating a text message to the number from your own device

- Reference: https://support.twilio.com/hc/en-us/articles/230878368-How-to-use-templates-with-TwiML-Bins

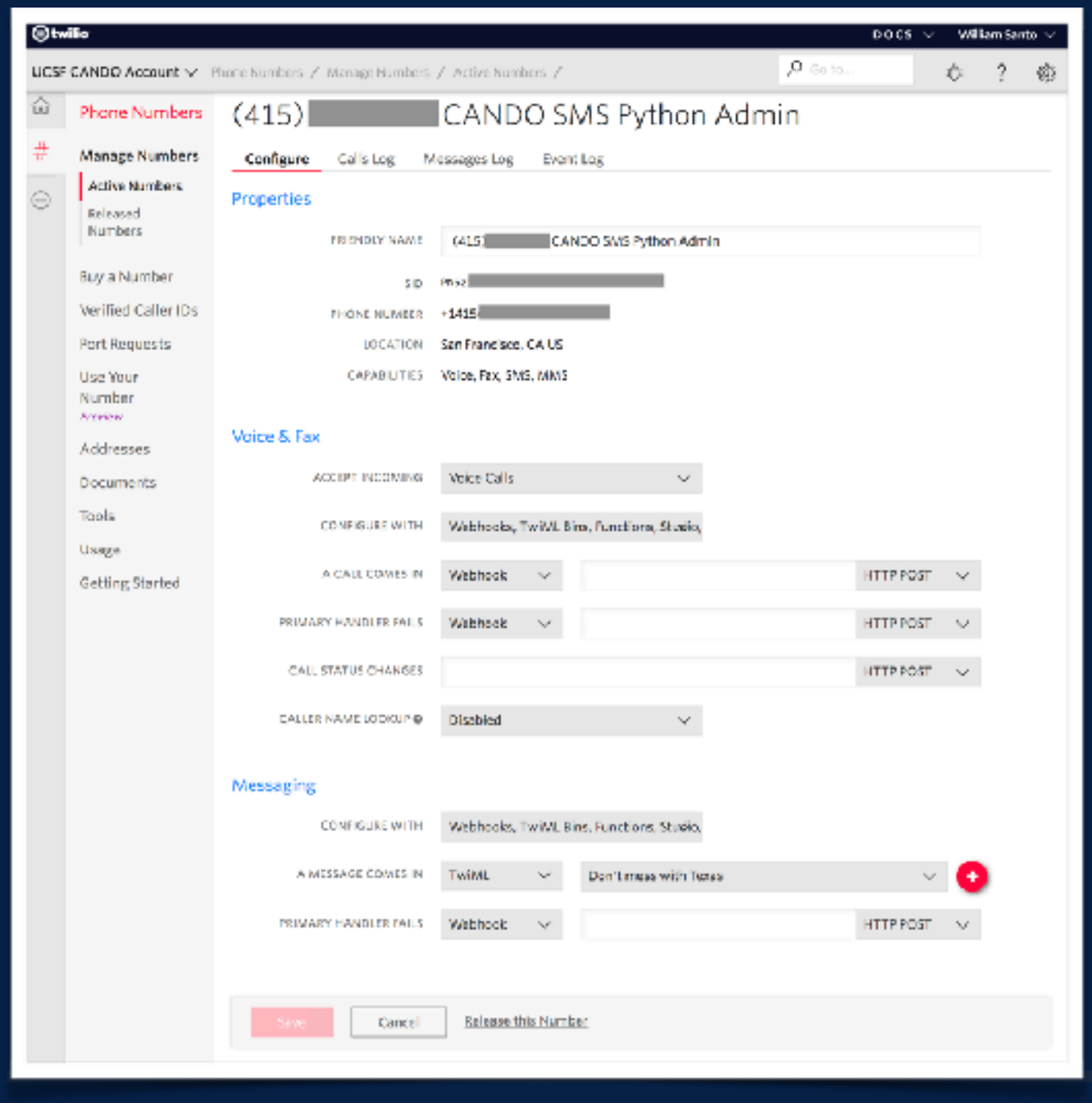
PROVISIONING FOR REDCAP



Simply add the Redcap URL to these two rows

Tip: Change the title of key numbers to make it easy to identify their purpose

PROVISIONING FOR CODE



This particular number is used for communicating code execution status updates

TWILIO API TYPICAL USAGE

- Typical use case would involve:
  - Export participant contact data from Redcap API
  - Determine which participants qualify for receiving SMS
  - Establish a connection to Twilio using credentials
  - Send SMS
  - Check status of each message using Twilio SID
  - Format status data for import to Redcap API
  - Import status data

SSH

- Secure shell (ssh) allows encrypted cross-platform communication between a client and a server using a combination of public and private keys
  - https://www.wikihow.com/Use-SSH
- scp (secure copy) uses a secure ssh tunnel to transmit/receive files
- sshpass allows a client to store credentials securely to bypass a manual login
  - Also prevents credentials from being exposed in command history
  - https://www.cyberciti.biz/faq/noninteractive-shell-script-ssh-password-provider/

DEPLOYING CODE TO PRODUCTION

- One approach:
  - Use a shell script to specify files in current directory for scp to server
  - Before copy, shell script makes a timestamped copy of production files on server to backup server directory
  - Invoke sshpass to scp updated files to server

OBTAINING A VIRTUAL SERVER

- Start: https://datacenter.ucsf.edu/services/virtual-hosting
- Benefits of a virtual server: https://datacenter.ucsf.edu/services/benefits-virtual-hosting-services
- FAQ: https://datacenter.ucsf.edu/faq-page#t2n91

- Running on bare minimum hardware may be possible, but the marginal cost of additional resources (CPU/RAM) is low
- Ordering system admin services for $100/month is worthwhile (do you want to be a sys admin?)
- Determine if needed on internal or external network
  - External network requests require additional justification/approval
  - You must declare which inbound TCP ports are required and provide source IP addresses
- Caveat—could be challenging if you have never managed a server

NOTEBOOK DEMOS

- https://github.com/billsanto/caps_redcap_api

FINAL CAVEATS

- While the published code works for our circumstances, you are responsible for validating it to ensure it meets your expectations
- Protect your tokens
- Be careful using the import and delete methods
- Be mindful of server resources when interacting with the Redcap API
- Ensure you have necessary resources to deploy and maintain a production server and code base

REVIEW OF MAIN TOPICS

- Introduction to Python
- Redcap and Twilio APIs
- Redcap Data Characteristics
- Production environment