Selling Ideas

Examining the Relationships Between Startups Patents and Fundraising

Jason Knobloch
GADC Data Science
03.11.2015

What VC’s Look For

- Management
- Market Size
- Original Idea

Do a startup’s patents have any correlation with the amount of venture funding it receives?

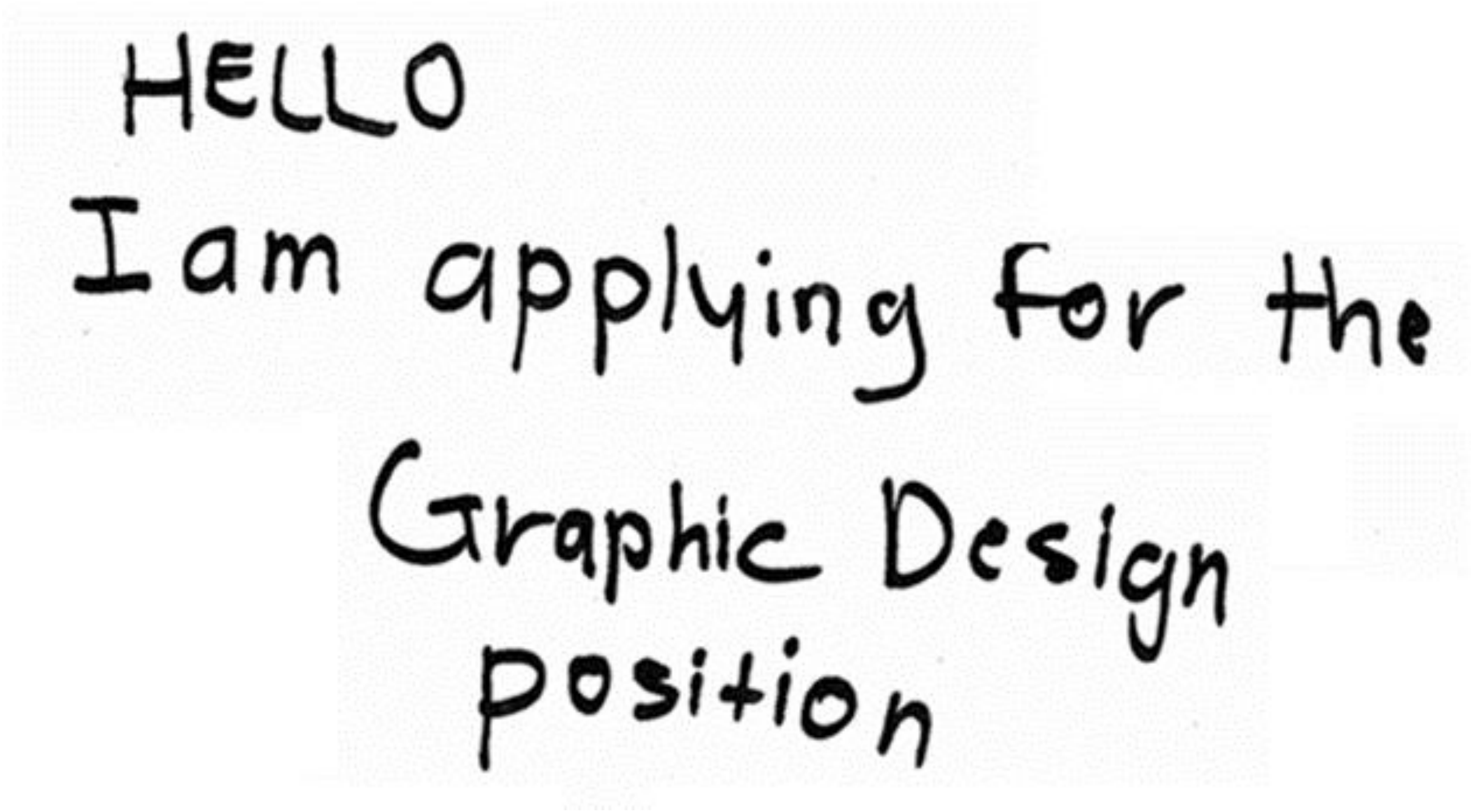
The Data

uspto UNITED STATES
PATENT AND TRADEMARK OFFICE

CrunchBase

Issue: File Size

Issue: Poor Formatting



HELLO

I am applying for the
Graphic Design
position

Issue: Dirty Data

Data Alignment

Patent & Funding Round Selection

- Funding Rounds:

  Earliest Angel, Seed, or Venture Round

- Patents:

  Patent with application date closest to first funding round date

$n = 2832$

$p = 25$

Determining Patent ‘Uniqueness’

Term Frequency-Inverse Document Frequency
(TF-IDF) Score for each patent claim
x
# of non-stop-word tokens in claim

	Values
mean	734.498034
std	569.970717
min	7.321309
25%	419.510578
50%	616.545718
75%	924.421893
max	9790.160784

‘Uniqueness’ Example 1

Claim:

The design for the sunglasses, as shown and described.

Score:

7.3213089344126026

‘Uniqueness’ Example 2

Claim:

1. An end-to-end publish/subscription messaging system with a middleware architecture, comprising: at least one messaging appliance configured to receive and route messages; and an interconnect utilizing channel-based messaging that routes messages over a first messaging layer based on at least one channel, each channel mapped to a subscription topic, each channel assigned to a communication pathway of a second messaging layer, wherein each messaging appliance is further configured to execute the routing of messages by dynamically selecting, in real time, a message transmission protocol and a message routing path. 2. The system of claim 1, wherein the messaging appliances include one or more of an edge messaging appliance and a core messaging appliance. 3. The system of claim 1, wherein each edge messaging appliance is linked to a message transformation engine for transforming incoming messages from an external protocol to a native protocol and for transforming routed messages from the native protocol to the external protocol. 4. The system of claim 1, wherein the message transmission protocol is selected to be one of a unicast, multicast or broadcast protocol. 5. The system of claim 1, further including one or more application programming interfaces configured for interfacing between one or more applications and one of the messaging appliances. 6. The system of claim 5, in which the messaging appliances and the one or more application programming interfaces are operative to communicate with each other by incorporating one or more messages in a single frame. 7. The system of claim 1, wherein each of the applications is configured to send requests, including registration and subscription requests, to a respective one of the messaging appliances. 8. The system of claim 1, wherein each of the application programming interfaces is logically linked to a messaging appliance having been registered to it via a topic-based subscription. 9. The system of claim 8, wherein the messaging appliances include one or more core messaging appliances to which the application programming interfaces register. 10. The system of claim 1, wherein the subscriptions are topic-based each being established via a subscription request, and wherein a single subscription request is capable of establishing subscriptions to a group of related topics. 11. The system of claim 1, in which the interconnect is an interconnect network over which the messaging appliances is deployed, the network being configured with any number of routers, a switches and a subnets. 12. The system of claim 1, wherein the interconnect is a channel-based, fabric agnostic physical medium. 13. A system as in claim 1, wherein the messaging appliances and interconnect incorporate transport logic. 14. The system of claim 13, configured for transport transparent channel-based messaging where messages are communicated in native protocol format independent of the transport logic. 15. The system of claim 1, further comprising one or more external sources and external destinations, wherein the messaging appliances include one or more edge messaging appliances, each associated to a protocol translation engine and translating messages between external and native protocols, and one or more core messaging appliances, each conducting communications of messages with the native protocol, the external sources and destinations communicating with the edge messaging appliances which, in turn, communicate with the core messaging appliances using neighbor-based message routing. 16. The system of claim 15, in which an edge MA is operative to route an ingress message simultaneously to both a native protocol consumer and an external protocol consumer. 17. The system of claim 1, wherein the messages are constructed with schema and payload which are separated from each other when messages enter the system, and which are combined when messages leave the system. 18. The system of claim 1, wherein, with the namespace management, consumers or external destinations who are subscribed to topics associated with a particular namespace are entitled to publish and subscribe messages identified with such topics. 19. The system of claim 1, wherein each messaging appliance has a routing table with routing of messages between messaging appliances being neighbor-based such that each messaging appliance is configured to route messages via one of its channels to a neighbor that has subscribed to all or a subset of messages transmitted via that channel, and wherein each messaging appliance is further configured to optimize the mapping between its subscriptions and channels in order to reduce the discard rate at neighbor that subscribe only to a subset of the messages. 20. The system of claim 1, wherein the routing table has one of a plurality of structures two of which are a tree structure and a dynamic has map structure. 21. The system of claim 1, wherein the messages and administrative messages have a topic based format, each message having a header and a payload, the header including a topic field in addition to source and destination namespace identification fields. 22. The system of claim 21, wherein the topic field includes a variable-length string or a key, the key being a unique value. 23. The system of claim 21, wherein the messages include a subscription message with a topic field that has a variable-length string with any number of wild card characters for matching it with any topic substring provided that such topic and the subscription message have the same number of topic substrings. 24. The system of claim 1, wherein the dynamic selection of transmission protocol and message routing path is based on system topology, health and performance reports and it involves one or both of dynamic resource allocation and dynamic channel creation and/or selection. 25. The system of claim 1, having boundaries that transcend regional, national or continental borders, with subsystems in each region, country or continent, wherein the subsystems are linked via a networking infrastructure and each subsystem includes an interconnect and one or more messaging appliances. 26. The system of claim 1, in which each messaging appliance includes: a network management stack linked to a physical interface management functional block; a services block containing a system management services functional block and a time stamping service functional block, both linked via a network management internal communications logical bus to the network management stack; and a native message layer in communication with a messaging transport layer, both of which linked via a messaging internal communications logical bus to the services block. 27. The system of claim 26, wherein the native message layer includes an administrative message layer, a message routing engine, message transmit and message receive logic and a master protocol optimization service. 28. The system of claim 26, wherein the messaging transport layer include channel management, and wherein the dynamic selection of a message routing path includes selection and/or creation of a channel. 29. The system of claim 28, wherein each channel is configured for network-based, node-based or memory-based transmission protocol and is associated with a physical interface to a physical medium which is fabric agnostic. 30. The system of claim 29, wherein the physical medium is configured as Ethernet, memory-based direct connect or Infiniband. 31. The system of claim 28, wherein the interconnect includes the transport channel and physical medium through which the messaging appliance communicates with the provisioning and management system. 32. The system of claim 1, further comprising one or more caching engines each operatively connected to a respective messaging appliance for providing quality of service functionality including message data store and forward functionality. 33. The system of claim 32, wherein each caching engine includes a caching layer connected with a native message layer which is, in turn, connected to a message transport layer, wherein the caching layer includes storage, a storage service and an indexing service. 34. The system of claim 1 wherein one or more of the messaging appliances is operatively connected to an application via an application programming interface that is registered to such messaging appliance and delivers messages between the application and the messaging appliance. 35. The system of claim 34, wherein each messaging appliance includes a master protocol optimization service and each application programming interface includes a slave protocol optimization service responsive to its respective master protocol optimization service. 36. The system of claim 34, wherein each application programming interface includes, a communication engine and one or more application stubs linked thereto. 37. The system of claim 36, wherein the communication engine is a daemon. 38. The system of claim 34, wherein each application programming interface is deployed on top of an operating system in the client application host. 39. The system of claim 34, wherein each application programming interface includes: an application delivery engine for transmitting messages to the application; and an administrative message layer for handling administrative messages. 40. The system of claim 1, wherein each of the messaging appliances are configured for fault tolerance. 41. The system of claim 40, in which the messaging appliances are arranged in fault tolerant pairs each pair including a primary and secondary messaging appliance, the secondary messaging appliance taking over a session from the primary messaging appliance upon failure of such session but without interfering with other active sessions on the primary messaging appliance. 42. A system with a publish/subscribe middleware architecture, comprising: a plurality of namespace domains; and a physical domain interconnect medium for connecting at least two of the namespace domains, wherein each namespace domains includes: at least one messaging appliance configured to receive and route messages; and an interconnect utilizing channel-based messaging that routes messages over a first messaging layer based on at least one channel, each channel mapped to a subscription topic, each channel assigned to a communication pathway of a second messaging layer, wherein each messaging appliance is further configured to execute the routing of messages by dynamically selecting a message transmission protocol and a message routing path. 43. An enterprise system with a publish/subscribe middleware architecture, comprising: a market data delivery infrastructure having at least one messaging appliance for receiving and routing market data messages; a market order routing infrastructure having at least one messaging appliance to receive and route transaction order messages; and an intermediate infrastructure in communication with the market data delivery and market order routing infrastructures, respectively, wherein the intermediary infrastructure includes: at least one messaging appliance configured for receiving and routing the market data and transaction order messages; and an interconnect utilizing channel-based messaging that routes messages over a first messaging layer based on at least one channel, each channel mapped to a subscription topic, each channel assigned to a communication pathway of a second messaging layer, wherein each of the messaging appliances is further configured to execute the routing of messages it receives by dynamically selecting a message transmission protocol and a message routing path. 44. The enterprise system of claim 43, further comprising: market data sources for publishing the market data messages; and market data consumers for receiving the market data messages and for publishing the transaction order messages, the market data consumers including at least one application,

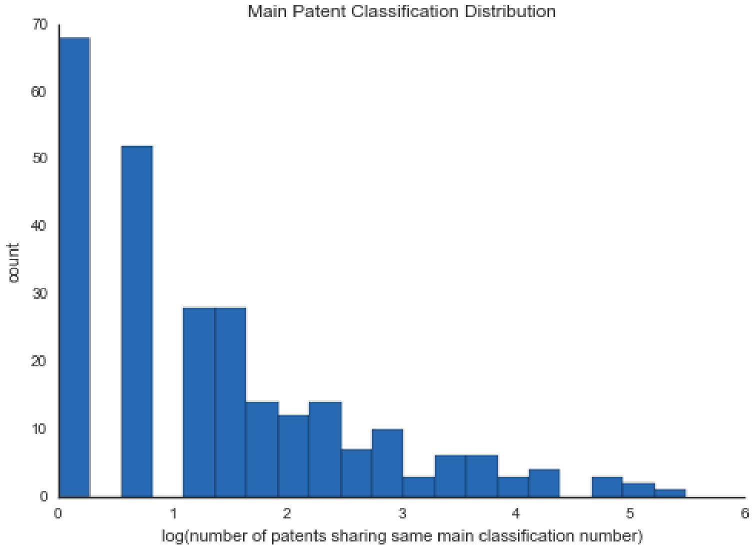
Score:

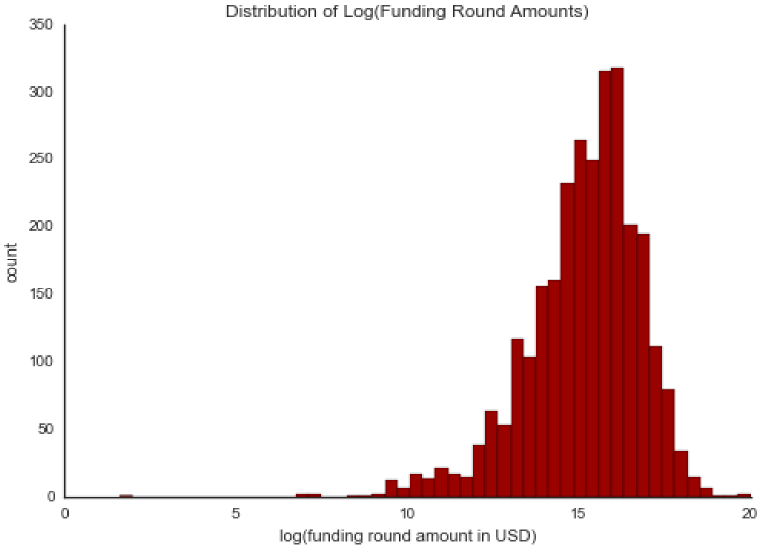
2605.5875380426073

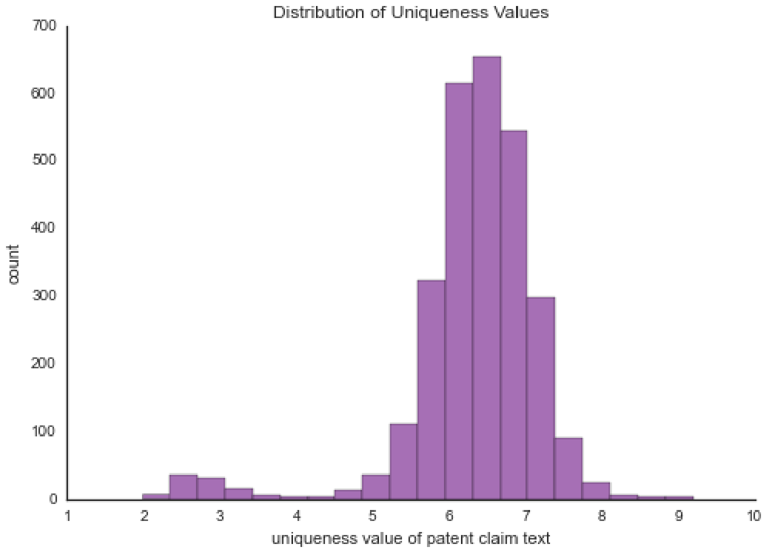
Determining Market Size

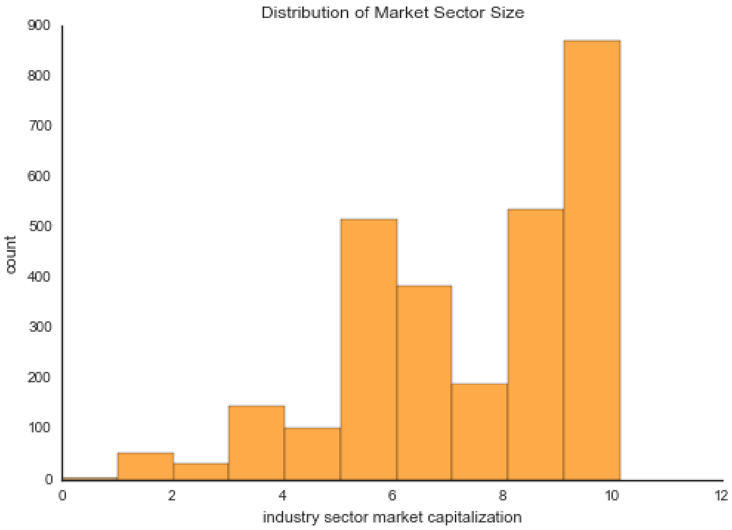
Hand coded the main patent classification to the most relevant industry sector market cap as listed on Yahoo! Finance

Dataset


Main Patent Classification Distribution


Distribution of Log(Funding Round Amounts)


Distribution of Uniqueness Values


Distribution of Market Sector Size

Analysis

First Approach

- Response: log(funding amount)

- All the Features

- Random Forest Regression
  - Continuous Response
  - Nonparametric

First Approach Feature Importances

Features	Importance
Uniqueness	0.274470
Number of References	0.067614
Market Sector Size	0.060545
Number of Applicants	0.037062
From California	0.018003
Patent Classification 424: Drug, bio-affecting and body treating compositions	0.011930
From New York	0.009787
From the US	0.009556
Patent Classification 707: Data Processing: database and file management or data structures	0.009417
From Texas	0.009222

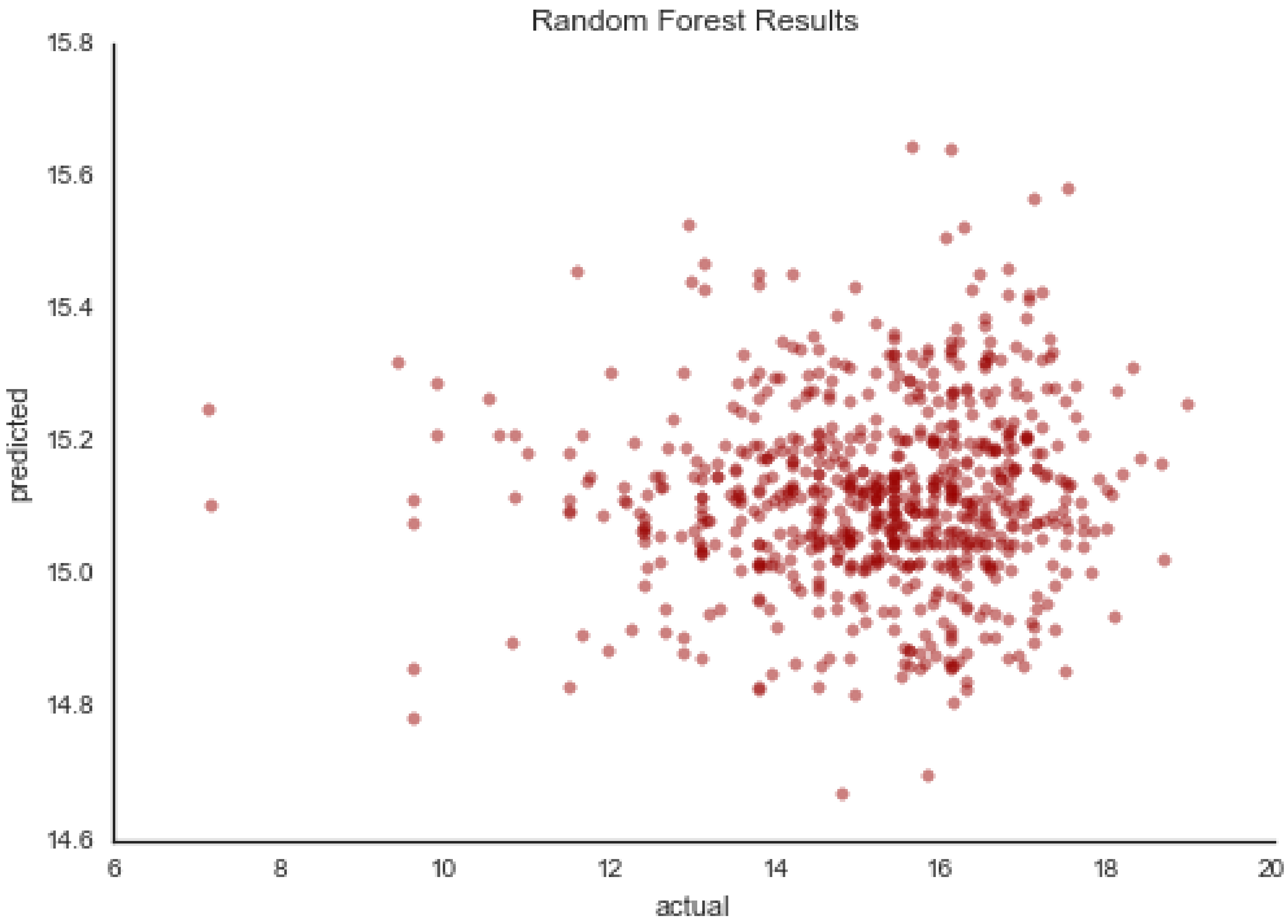
Second Approach Results

- n = 2832, p = 4
- Random Forest n_estimators, max_leaf_nodes = 800, 6
- Out-of-Bag Score = 0.0021809859111234786
- Root Mean Squared Error (RMSE) = 1.6303398440226933
  - Log(funding amount) standard deviation = 1.668907

Second Approach Feature Importances

Features	Importance
Uniqueness	0.374133
Market Sector Size	0.323855
Number of References	0.183901
Number of Applicants	0.118112

Second Approach Results



Third Approach

- Response: log(funding amount)

- All the Features for Series A Funding Rounds

- Random Forest Regression
  - Continuous Response
  - Nonparametric

Third Approach Results

- n = 925, p = 252
- Random Forest n_estimators = 700
- Out-of-Bag Score = -0.12759478764657217
- Root Mean Squared Error (RMSE) = 1.0070616514513329
  - Log(funding amount) standard deviation = 0.978134

Third Approach Feature Importances

Features	Importance
Uniqueness	0.269596
Market Sector Size	0.068458
Number of References	0.058733
Number of Applicants	0.040417
From California	0.020101
Patent Classification 290: Prime-mover dynamo plants	0.016295
Patent Classification 707: Data Processing: database and file management or data structures	0.015874
Patent Classification 435: Chemistry: molecular biology and microbiology	0.015526
From Colorado	0.014197
From India	0.012590

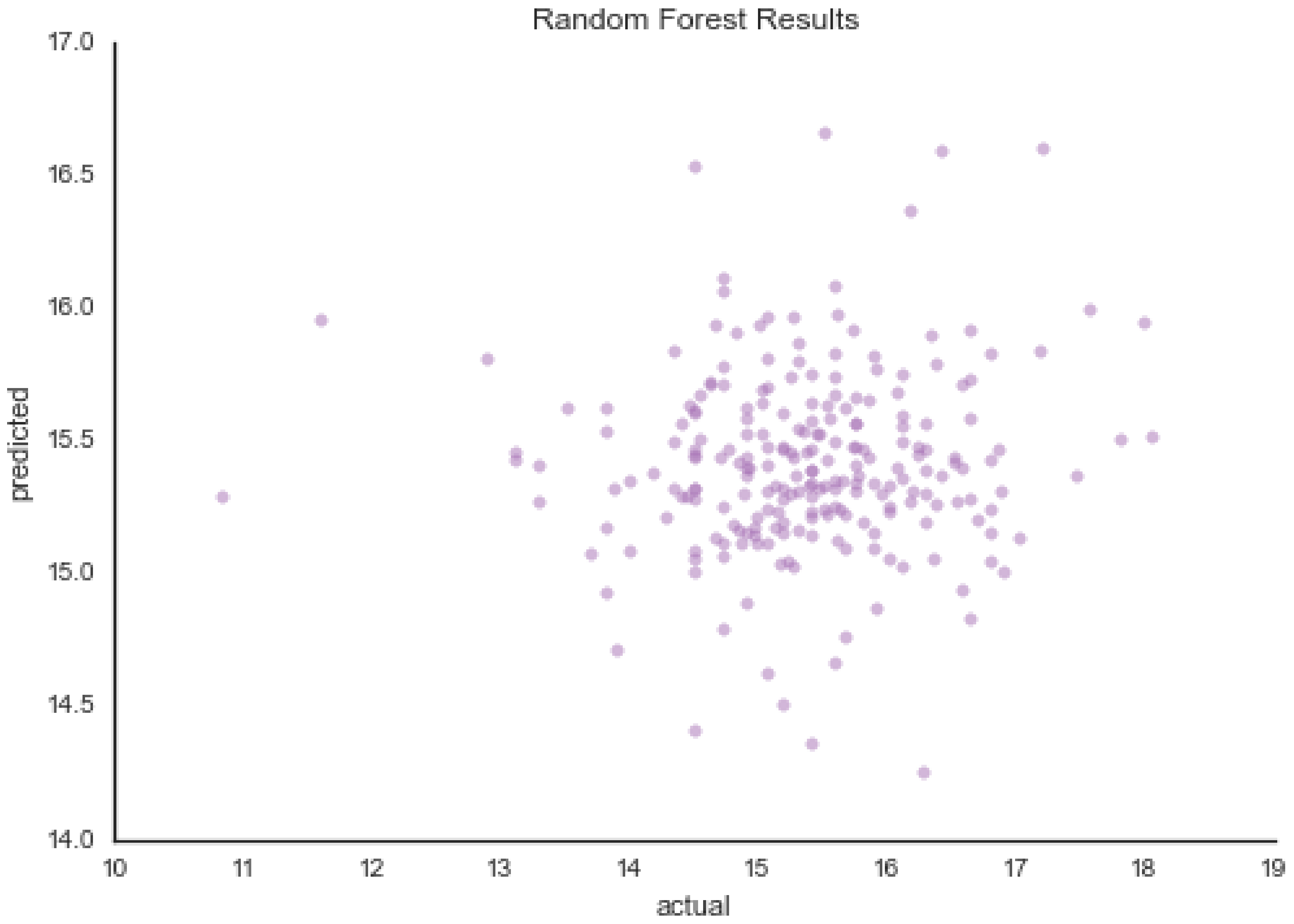
Final Approach Results

- n = 2832, p = 4
- Random Forest n_estimators, max_leaf_nodes = 100, 2
- Out-of-Bag Score = -0.002899632678329489
- Root Mean Squared Error (RMSE) = 0.9626713483785625
  - Log(funding amount) standard deviation = 0.978134

Second Approach Feature Importances

Features	Importance
Market Sector Size	0.47
Uniqueness	0.39
Number of References	0.13
Number of Applicants	0.01

Final Approach Results



Insights

Opportunities