Applications of Data Analytics

Vincent Granville, Ph.D., Co-founder

Data Science Central

vincentg@datasciencecentral.com

www.DataScienceCentral.com

September 27, 2019

General Overview (I)

- Traditional Applications
  - Marketing analytics
  - Retail (pricing, cross-selling, sales forecasting)
  - Supply chain optimization (OR)
  - Segmentation, customer profiling, clustering (Insurance)
  - Advertising (attribution modeling – NBCi example, ad matching / relevancy)
  - Fintech, healthcare, clinical trials
  - Fraud/spam detection, transaction scoring

General Overview (II)

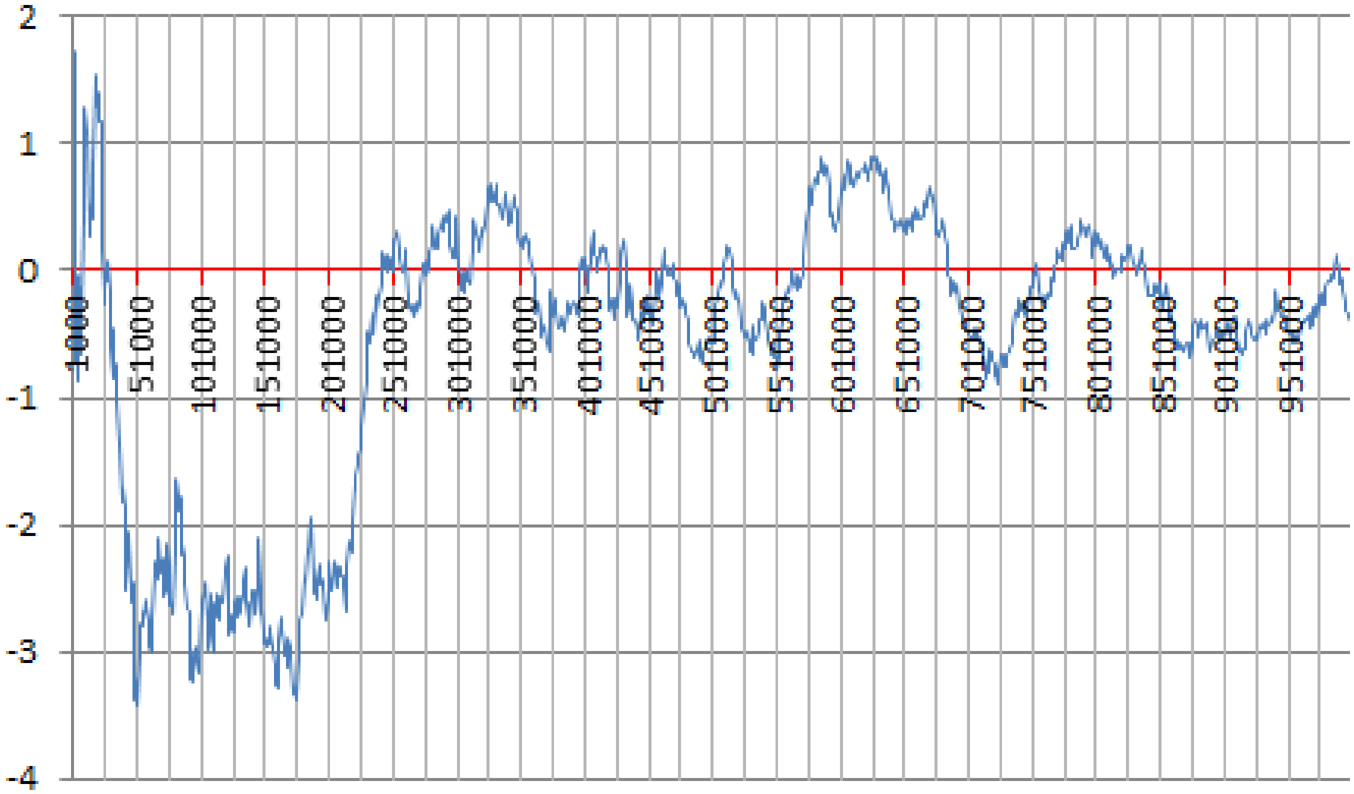
- New Applications
  - NLP, sentiment analysis
  - Taxonomy building (keyword associations, product taxonomy, Yellow pages: improved Restaurant category)
  - Automated vision (Google cars), automated translation, Chatbots
  - Recommendation engines, detection of fake news, scoring users
  - IoT: sensor data, machine-to-machine communications, smart farming, smart cities
  - Scale-invariant techniques
- Applications of Theoretical Data Science
  - Generalized resampling, new foundational theorems
  - New data **science** conjecture (causation vs. correlation problem)
  - From number theory to new models (Fintech/gaming applications)

New Models

- Ensembles: model blending, HDT
  - 2 blended approximate models better than 1 more accurate
- Deep neural networks (several layers)
  - Hierarchical Bayesian models
  - Nested mixture models
- Optimization techniques
  - Stochastic gradient
  - Swarm optimization
- Improved cross-validation / walk-forward techniques
- Non-standard Brownian motions (Fintech)
- Visualization tools and HPC (high performance computing)
  - Automated elbow rule, hexagonal binning

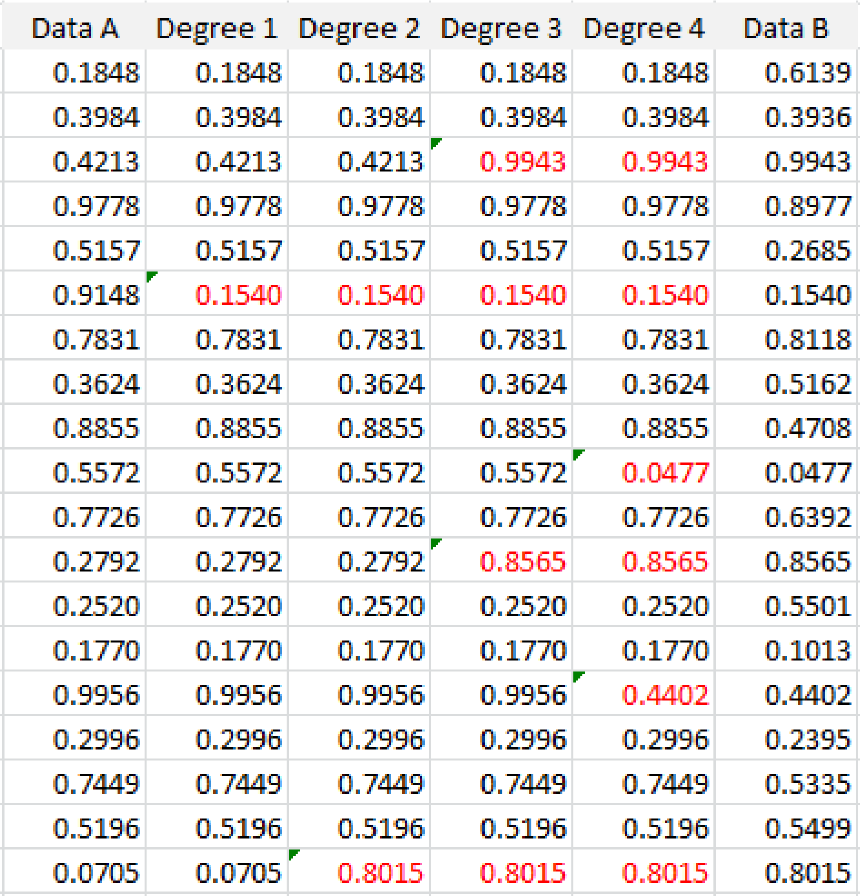
Theoretical Data Science (I)

- Bounded Brownian motion based on digits of SQRT(2)



Theoretical Data Science (IIa)

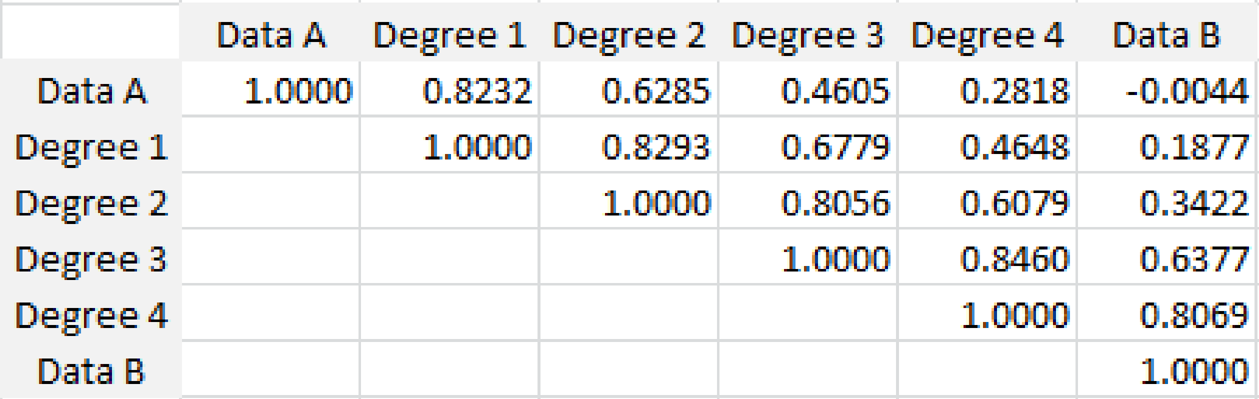
- Conjecture: Each data set is 6 degrees of separation away from any other data sets (how to lie with data!)



Data A	Degree 1	Degree 2	Degree 3	Degree 4	Data B
0.1848	0.1848	0.1848	0.1848	0.1848	0.6139
0.3984	0.3984	0.3984	0.3984	0.3984	0.3936
0.4213	0.4213	0.4213	0.9943	0.9943	0.9943
0.9778	0.9778	0.9778	0.9778	0.9778	0.8977
0.5157	0.5157	0.5157	0.5157	0.5157	0.2685
0.9148	0.1540	0.1540	0.1540	0.1540	0.1540
0.7831	0.7831	0.7831	0.7831	0.7831	0.8118
0.3624	0.3624	0.3624	0.3624	0.3624	0.5162
0.8855	0.8855	0.8855	0.8855	0.8855	0.4708
0.5572	0.5572	0.5572	0.5572	0.0477	0.0477
0.7726	0.7726	0.7726	0.7726	0.7726	0.6392
0.2792	0.2792	0.2792	0.8565	0.8565	0.8565
0.2520	0.2520	0.2520	0.2520	0.2520	0.5501
0.1770	0.1770	0.1770	0.1770	0.1770	0.1013
0.9956	0.9956	0.9956	0.9956	0.4402	0.4402
0.2996	0.2996	0.2996	0.2996	0.2996	0.2395
0.7449	0.7449	0.7449	0.7449	0.7449	0.5335
0.5196	0.5196	0.5196	0.5196	0.5196	0.5499
0.0705	0.0705	0.8015	0.8015	0.8015	0.8015

Theoretical Data Science (IIb)

- We have the following correlations:
  - □ Data A / Data B: -0.0044
  - □ Degree 1 / Data A: 0.8232
  - □ Degree 2 / Degree 1: 0.8293
  - □ Degree 3 / Degree 2: 0.8056
  - □ Degree 4 / Degree 3: 0.8460
  - □ Data B / Degree 4: 0.8069

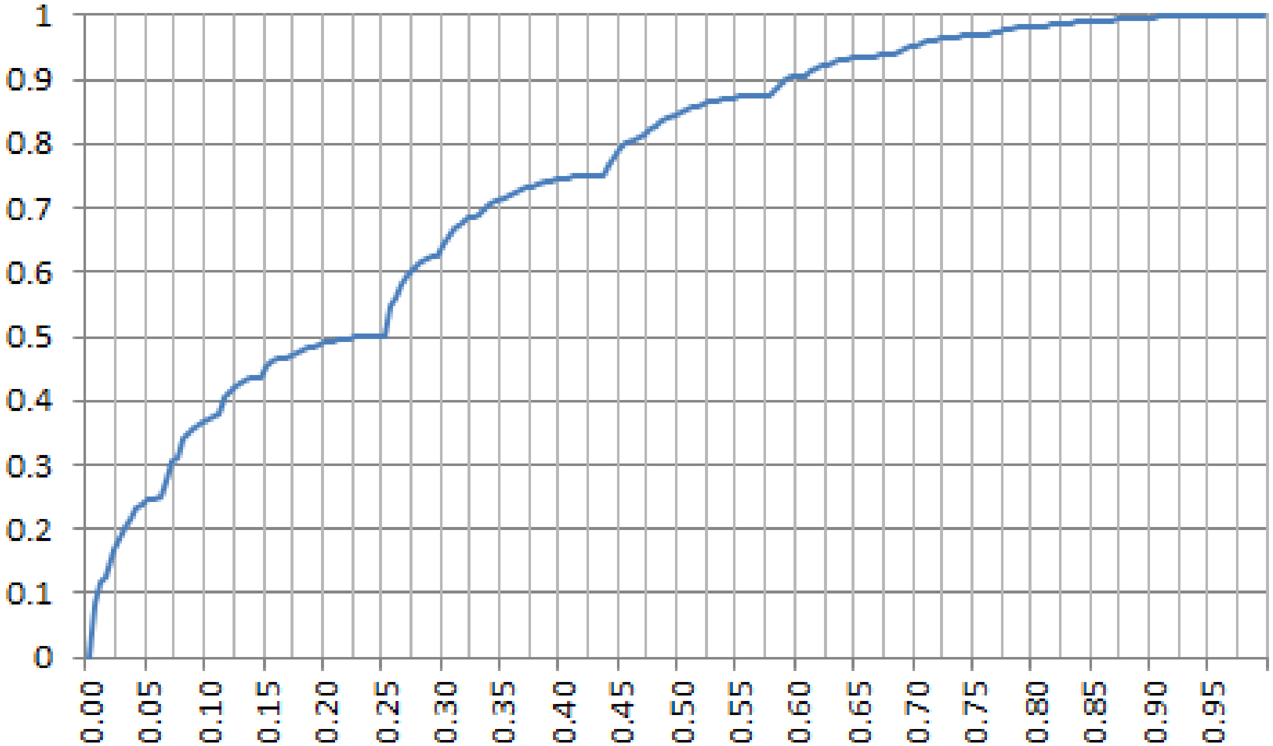


	Data A	Degree 1	Degree 2	Degree 3	Degree 4	Data B
Data A	1.0000	0.8232	0.6285	0.4605	0.2818	-0.0044
Degree 1		1.0000	0.8293	0.6779	0.4648	0.1877
Degree 2			1.0000	0.8056	0.6079	0.3422
Degree 3				1.0000	0.8460	0.6377
Degree 4					1.0000	0.8069
Data B						1.0000

Cross-correlations between the 6 data sets

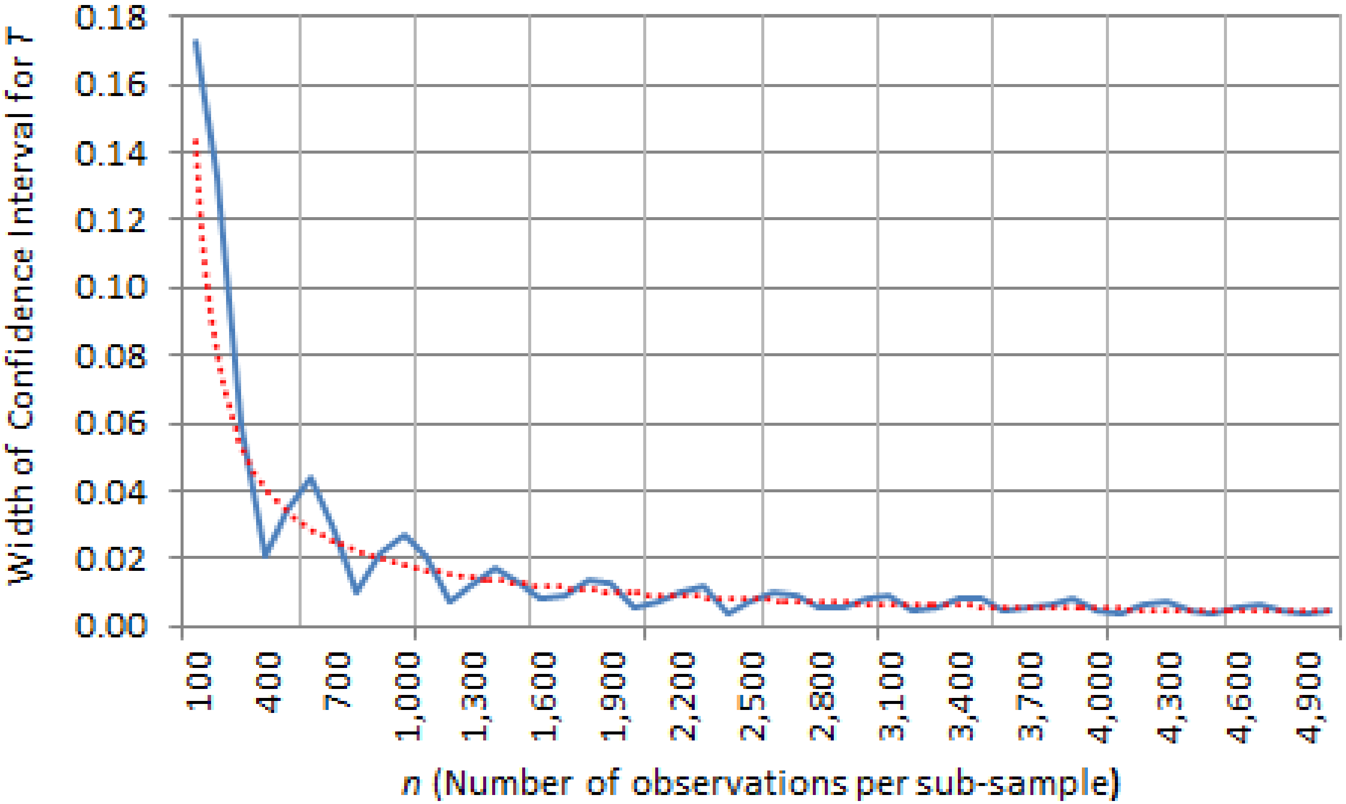
Theoretical Data Science (III)

- New family type of (bumpy) statistical distributions (Fintech)



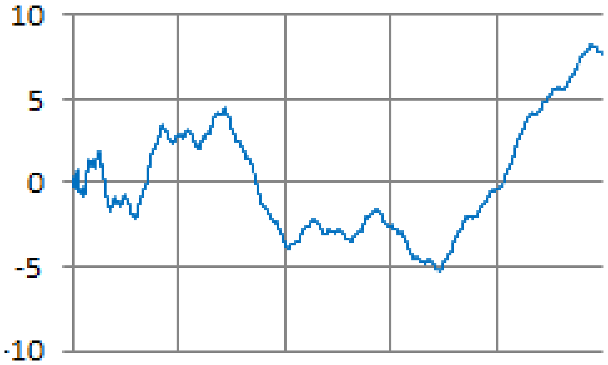
Theoretical Data Science (IV)

- Model-free confidence intervals based on general resampling: width ~ A / n^B (0 < B < 1; 0.5 = Central Limit Theorem)

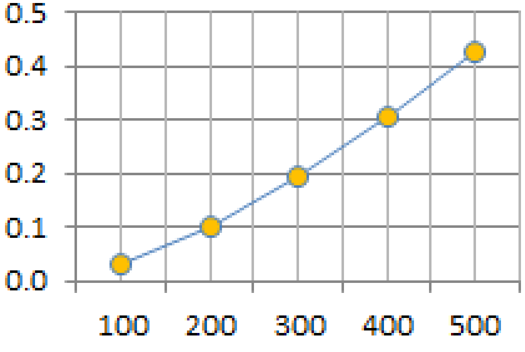


Theoretical Data Science (V)

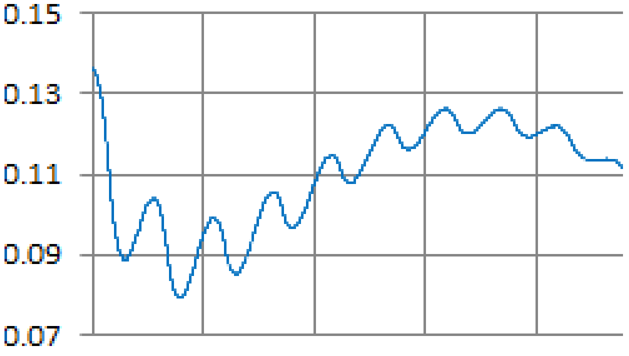
- Time series with strong, long-range auto-correlations, or non-ergodic, and modified Hurst exponent



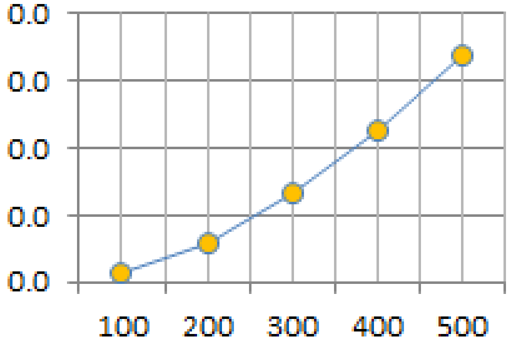
Series D: Sample path, X-axis is time



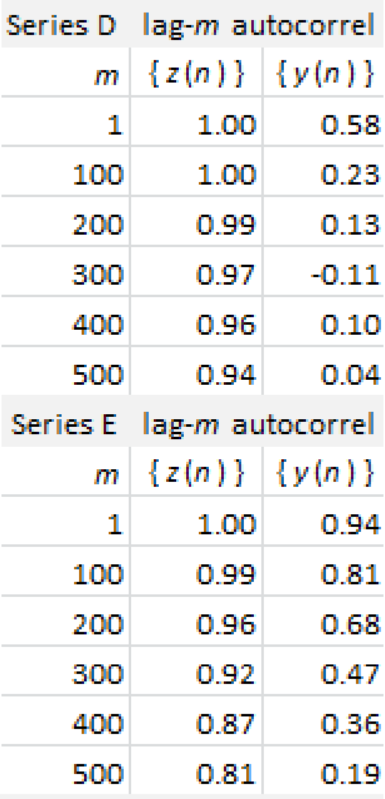
Series D: $S(N, m)$, X-axis is m



Series E: Sample path, X-axis is time



Series E : $S(N, m)$, X-axis is m

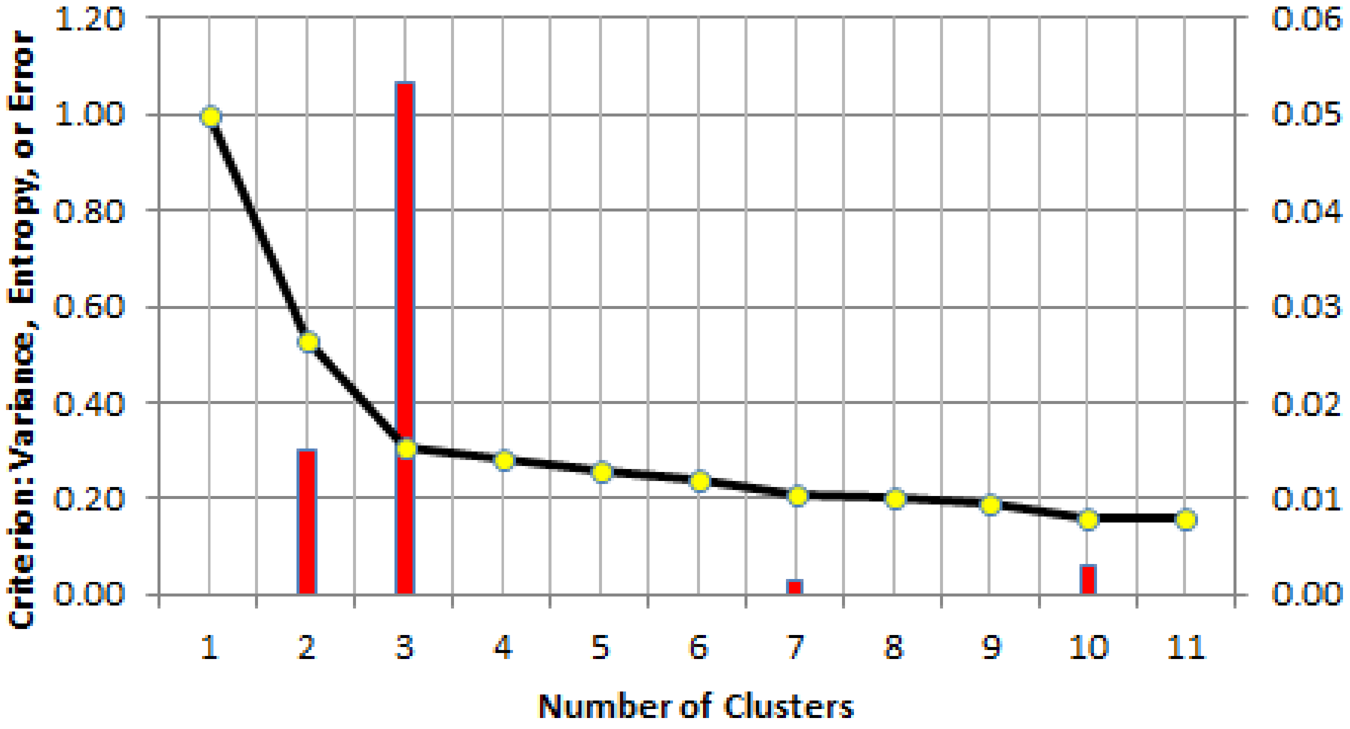


Series D	lag-m autocorrel	
m	$\{z(n)\}$	$\{y(n)\}$
1	1.00	0.58
100	1.00	0.23
200	0.99	0.13
300	0.97	-0.11
400	0.96	0.10
500	0.94	0.04

Series E	lag-m autocorrel	
m	$\{z(n)\}$	$\{y(n)\}$
1	1.00	0.94
100	0.99	0.81
200	0.96	0.68
300	0.92	0.47
400	0.87	0.36
500	0.81	0.19

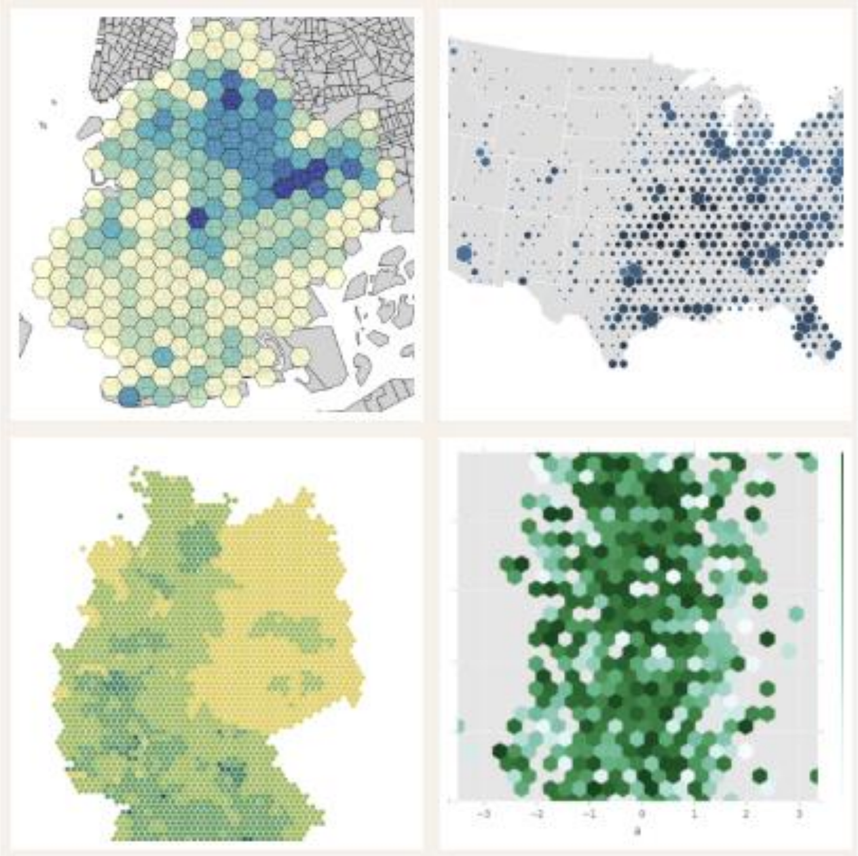
Visualization Tools (I)

- Automated elbow rule for black-box clustering or optimal binning



Visualization Tools (II)

- Hexagonal binning (for better density estimates), 3-D tessellations, data videos with R, and more

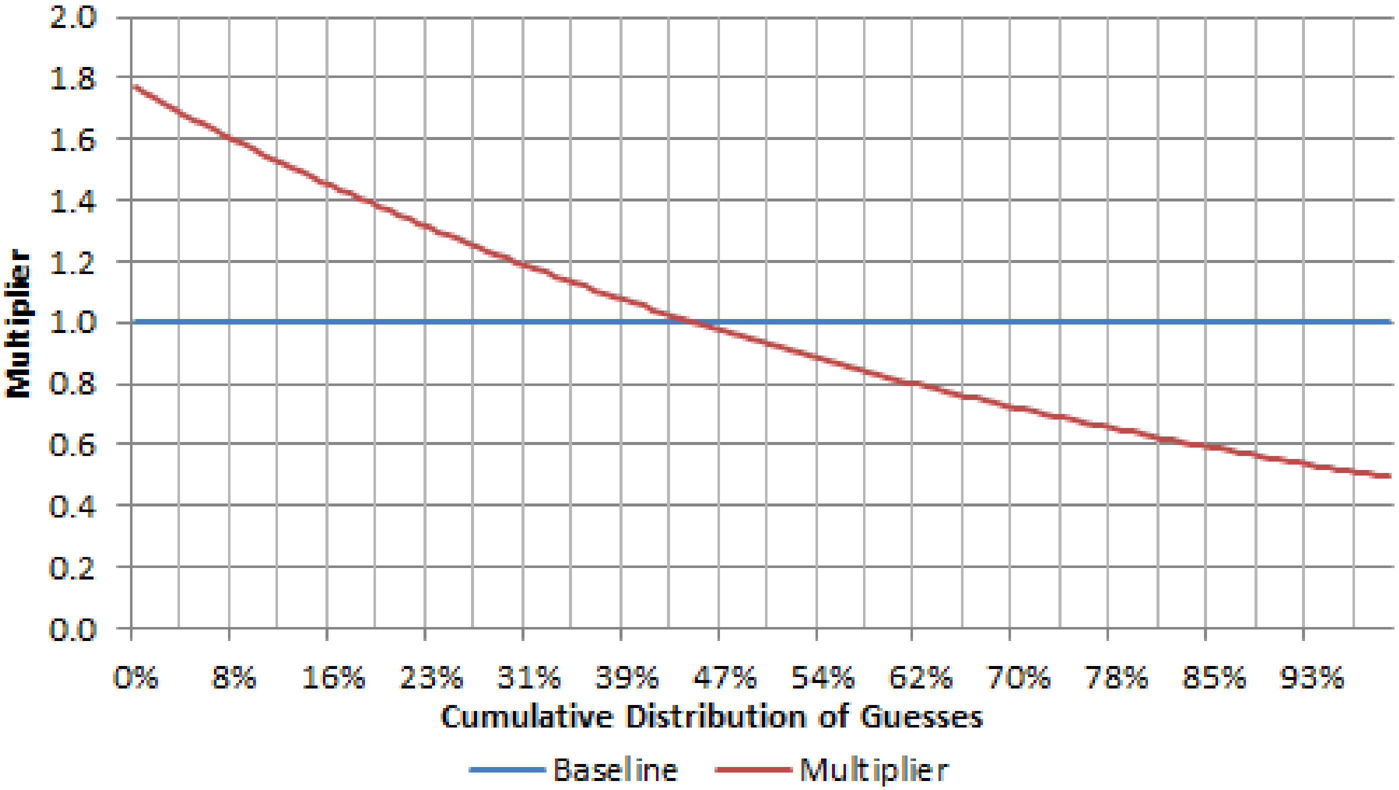


Gaming Industry App (I)

- Blending number theory, chaos theory, cryptography, HPC, computer science, statistical science, stochastic processes
- Number guessing game
  - Public algorithm to compute next winning numbers
  - Design your own ROI table
  - Gains depend on distance between your bet and the winning number
  - Public algorithm requires billions of operations on numbers with 250,000 digits: not practical
  - Private algorithm kept secret, patent-pending
  - Impossible to reverse engineer (in theory)
  - Blockchain to process financial transactions (victual currency)

Gaming Industry Application (II)

- User-generated neutral ROI table (geometric distribution; (multiplier > 1 means gain; chance of winning > 40%)



20 Apps Used by Amazon (I)

- Supply chain optimization (I). Sites selection for warehouses to minimize distribution costs (proximity to vendors, customers).
- Supply chain optimization (II). Selection of optimal routes, schedules, and products groupings, to minimize delivery costs.
- Pricing and profit optimization (price elasticity).
- Fraud detection for credit card transactions. Detect criminal activity on AWS. Detect system intrusions.
- Fake reviews detection.
- Taxonomy creation to categorize products, using tagging and indexing algorithms.
- Smart search engine technology (based also on taxonomy discussed above) to help users find what they want.

20 Apps Used by Amazon (II)

- Multivariate testing, for instance to find out which version of a search engine increases sales.
- Recommendation engine (and detection of artificial purchases aimed at fooling these algorithms.)
- Customer segmentation, churn analysis, using survival analysis models, to increase marketing and advertising efficiency.
- Advertising optimization, including automated bidding on Google Adwords for millions of keywords in real time, most having no historical data (use bucketasition techniques to group keywords in buckets that have real predictive power); algorithms to identify millions of keywords worth purchasing, based on expected yield. Advertising mix optimization and attribution modeling. SEO and SEM.

20 Apps Used by Amazon (III)

- Inventory forecasting.
- Sales forecasting broken down by category / location .
- HR analytics: who to hire, how to score candidates to better predict who will succeed; detect employees at risk of leaving or committing fraud; optimize purchase of office supplies; optimize employee compensation; optimize travel expenses
- Real estate analytics.
- Software/hardware system analytics: minimizing/predicting server crashes, optimizing redundancy with budget constraints, optimizing load balance and bandwidth usage; Also, create email alert systems, automatically prioritize messages and select recipients. Also manage external email campaigns (delivery rate, open and click rate optimization).
- Monitoring system, dashboard metrics and visus for managers.

20 Apps Used by Amazon (IV)

- Payments analytics. Optimization of payments: to authors, vendors, publishers, while maximizing profits and minimizing publisher / author / vendor churn; vendor and publisher selection algorithms.
- Competitive analysis: automatically process billions of comments posted by users on social networks about Amazon, its competitors, and new trends; summarize this data, take actions based on the insights derived from this daily / hourly / real-time, automated analyses.
- Tax engineering.
- Ad Relevancy Algorithm to select and rank Ads to be displayed on a particular webpage to a particular visitor, to maximize some yield metrics (click through rate.

Selected References

1. Deep mixture models : https://dsc.news/2GEPcFj
2. Model-free confidence intervals - foundational theorem: https://dsc.news/2PUhCNh
3. New family of bumpy statistical distributions: https://dsc.news/2PelcoS
4. Automated elbow rule and clustering: https://dsc.news/2EYkh3E
5. Time series: Long-range auto-correlation, modified Hurst exponent: https://dsc.news/2uGqBYC
6. Bounded Brownian motions: https://dsc.news/2m0eUed
7. Six degrees of separation between any two data sets: https://dsc.news/2knXsA9
8. Gaming industry application: https://dsc.news/2Ujw56b