The future of Statistics: challenges for understanding new phenomena in a rapidly changing world

Giorgio Alleva

Università Sapienza di Roma
Dipartimento di Metodi e modelli per l'economia, il territorio e la finanza (MEMOTEF)

ITACOSM 2019 - Survey and Data Science

Firenze, 5 June 2019

Outline

1. **BACKGROUND**
   - Big changes are expected in our societies
   - Big uncertainty for the future and the need to design new strategies
   - Statistics and Official statistics are called on to offer an important contribution
2. **THE FUTURE CHALLENGES FOR OFFICIAL STATISTICS AND STATISTICAL SCIENCE.** FOUR MAIN AXES:
   - Data
   - Capabilities to manage data
   - Methods
   - Data governance
3. **CONCLUSIONS**

Background

Big change are expected in our societies, related to the impact of:

- ❒ Globalization (on businesses, migrations, competencies….);
- ❒ Digital technologies (on the nature of work, quantity & quality; on the perimeter, organization and performances of businesses)
- ❒ Climate change (on ecosystems and biodiversity).

In Europe we are still facing the effects of the big financial-real crisis.
Big uncertainty for the future, fears, hopes, anxieties.

Need to design new strategies for future of prosperity for people and the planet

- ❒ Inequalities
- ❒ Work and social protection
- ❒ Sustainability
- ❒ Human rights.

Statistics is called to offer an important contribution!

Background

The purpose of Statistics is to shed light on problems and to support decisions on how to solve them (independently by the field).

The great challenge of the coming years is an increase in the use of statistics to understand current trends in Society and outline possible strategies, providing answers both at the individual and the collective level.
(There are so many questions to answer!)

Providing a clear picture of the Society represents the traditional mission of Official statistics (OS), the mandate entrusted to them and for which public resources are destined.

In this era of great uncertainty about the future OS have to play an important role, therefore it is necessary to have qualified and independent National Statistical Institutes (NSIs) and international statistical agencies.

Background

Some examples of recent questions from government and parliamentary commissions

- Size of the floor and costs of the measure of citizenship income; effects on income inequality; with territorial detail, and by size of families;
- Effects of investment incentives (Industry Plan 4.0 and others);
- Evolution of the BES indicators included in the planning documents;
- Tax and contribution evasion;
- Spread of digital technologies by companies and the corresponding investments in human capital;
- Winning and losing professions in the different sector of economic activities;
- Price variability for the purchase of goods and services of the PA.

To measure facts & perceptions is increasingly important to understand opinions, attitudes and behaviors (on safety, health, life satisfaction and trust, ...).

For OS, nowadays, the strategic challenge is to produce high quality data guaranteeing ever more timeliness, details (territorial, social & economics sub-groups), absolute privacy protection ;
also providing analysis and assessments on determinants and impacts of phenoena.

Background

Nobody can do it alone!

Partnerships among OS, Academia, public and private researches, users, are fundamental, as is promoting the interest of young people in relation to OS and the methods needed to improve their quality.

Naturally, in addition to understanding our Society,
Statistics must continue to make their contribution to support the development of knowledge in all fields,
thanks to data, methods and skills (life sciences, social and hard sciences, ..).

Background

Over time, the NSIs have progressively updated their methods , technologies and capabilities of production and dissemination of data with respect to the new data sources.

This a long story: from the advent of different calculation tools, to the establishment of the probabilistic sample, to the availability of public registers and administrative archives.

Thanks to the integration of the international statistical system, experiences and standards have been improved and shared.

However, the maintenance of these capabilities and more generally of their institutional mandate constitutes nowadays a considerable challenge for the OS and NSIs, and indirectly for the whole statistical community.

Why?

Background

What are the new elements with respect to a natural evolution ?

① **The growth of the interest in data** by industries, governments and the media, that entails a responsibility to provide the right answers to the right questions.

② **The extraordinary opportunities and great risks associated with the availability of large amounts of data**, in part related to unexplored phenomena and in part held by large private companies that are not part of the National statistical systems.

③ **New competitors** that invest large resources in the organization and treatment of the sources they hold, in a legal framework that is sometimes with fewer constraints than those of the NSIs; with implications on the trust of citizens regarding the quality and objectivity of such data and on the protection of the confidentiality of the information they themselves produce.

Background

In Istat, the structured strategy to respond to these new challenges was the design and implementation of a Modernization Program , based on a new production model and an organization fully aligned with this model, in line with the Vision 2020 of the ESS, and the international vision (Istat, 2016).

Future challenges and developments

Future challenges and developments of official statistics and of statistical science can be presented along four main axes:

- Data,
- Capabilities to manage data,
- Methods,
- Data governance.

Future challenges and developments. DATA

The first point is the new role and value of data.

There is a new awareness on the utility of data and of the ability of their treatment.

- Data to understand needs and preferences of consumers and users,
- Data for searching new ways to produce and sell goods and services,
- Data to be more competitive,
- Data to take decisions on strategies, on material and immaterial investments,
- Data for monitoring regional & global strategies (Europe2020, SDGs).

Future challenges and developments. DATA

There is simultaneous enormous growth in the volume of data, thanks to digital technologies, which has led to perceiving them as new oil, gold, the source of great gains thanks to their possession.

Not only for companies and private purposes, but also for research and public decisions.

Future challenges and developments. DATA

A first challenge for the statistical community to have a central role in the data issue, is the exploitation of Big Data.

The challenge *"starting from the data"* is certainly not a novelty,
Big Data represent a natural evolution, accelerated by a greater capacity for acquisition, storage and new analytics.

Over time, the community of statisticians has confronted itself:
- in the transition from experimental data ("the cult of the single study", Nelder, 1986) to large databases that allowed multidimensional analysis,
- the exploratory data analysis (EDA) of John Tukey,
- the exploratory approaches of the *l'analyse des données* of Jean-Paul Benzécri,
- Data mining and prediction methods based on black box advocated by Leo Breiman.

With heated debates between different cultures.

The dualism of data-driven and model-driven approaches to science and the need to move to a more diverse set of tools favouring the aim of using data to solve problems.

More recently the different approaches based on stochastic models or algorithmic procedures within the domain of artificial intelligence (AI) and knowledge engineering (KE).

Future challenges and developments. DATA

The need for innovations starting from new data and sources is not a novelty, and Big Data represent a natural evolution.

Which are the novelties?

- The enormous interest on the part of industries and governments to be able to exploit these sources,

- The prominent role of computer science and data architecture in this first phase,

- The risk of loss of the centrality of some fundamental principles for drawing scientific conclusions from data, such as providing a measure of the uncertainty for scientific statements based on data, however, unavoidable.

The core task still remains the inductive inference from data to models and scientific conclusions, also in the advent of massive data sets.

This is central point to clarify and to strengthen, with all our energies and commitment.

Future challenges and developments. DATA

The challenge for our community is to contribute to the analysis of these new type of data,

- assuming a leadership role in producing high quality public information in the new data ecosystem, also leveraging the opportunity to work with other scientific communities interested in the use of Big Data;
- showing successes and innovations, in the deepening of new phenomena and in methods to manage the new types of data; but there should also be more room for publishing negative findings;
- denouncing the incorrect use by data holders or the media; this means to be able to communicate and to use communication tools;
- promoting these sources as public goods, firstly for NSIs & NSSs, but also for researchers.

Future challenges and developments. DATA

As statisticians we must be aware of:

- responsibility for explaining the inescapable role of uncertainty in the findings from Big Data: uncertainty is certainly not diminished by the size of the data and by the processes of their acquisition and integration;

- the need to go beyond , moving from the probability sample survey paradigm of the past 75years to a mixed data source paradigm for the future (Citro, 2104).

The data and the problems guide the solutions: to solve a wider range of data problems a larger set of tools is needed.
If we do not recognize this, others will take over.

Future challenges and developments. DATA

How do we reinforce the instrumental role of statistics for the advancement of knowledge in the most diverse fields?

How do we overcome the so often reported among us “little recognized role of statistics” by society and in the educational system?

How do we assume a guiding role in the Big Data issue?

In my opinion the critical point is always the same: the lack of interaction with other cultures and communities.

Full cooperation with substantive disciplines and users are decisive.

Future challenges and developments. DATA

For OS the exploitation of these new data is inescapable but the challenge is even more complex:

It concerns the design of their *joint treatment together with traditional sources*, those from surveys and from administrative sources, with a rethinking of their role in the process of production of official statistics.

Images, sounds, communications through social media and digital platforms, mobile telephony, movements of people and goods, economic transactions and many other types of signals acquired by sensors and devices require reading and analysis through the development of new methods.

Future challenges and developments. DATA

Trusted Smart Statistics is the ESS strategy to face with Big Data (*Bucharest Memorandum*, adopted in October 2018 by the DGINS).

As indicated by Eurostat, the NSIs will have to go through a **'playground' phase**, which is necessary to understand the potential and limits of Big Data sources, as well as the methods necessary to treat them, **to a phase of mature use** of these sources, called Trusted Smart Statistics.

What is the concept of Trusted Smart Statistics (TSS) ?

TSS is the result of a production process with the following features:

Input: Smart Data Sources

Processing: Smart Methods Design and Execution

Output: Trust

Future challenges and developments. DATA

Smart Data Sources: include Big Data sources, mainly belonging to the category of “Internet of Things/machine generated data”, in charge of data providers out of NSSs (business\partnership models are needed).

Smart Methods Design and Execution

The design of processing will result from both the traditional statistical processing framework (design- and model-based frameworks) and the algorithmic-based processing framework (mainly machine-learning framework). A specific feature for smart statistical processing is the geographically distributed computation.

The design of methods/algorithms is the responsibility of the NSIs.

The execution of methods/algorithms can take place either at the data provider side or at the premises of the NSIs.

Output: Trust. Statistical products need to have *specific trust guarantees* established in the design phase.

Future challenges and developments. CAPABILITIES

The Data Scientists is considered increasingly as a key competence and profession with a great future.

Data science represents a multi-disciplinary field.

There is a broad consensus to include mathematics, computer science and statistics, also considering visualization and communication.

However, there is still no clear awareness that a data scientist should be able to work in consultation with social scientists and humanists who are gathering this data, on the importance of deep engagement in the subject matter.

Future challenges and developments. CAPABILITIES

How to build the new skills, abilities and professionalism?

The University system is experimenting with new paths in this direction.

Is there any need to train statisticians with new skills and abilities, or build a new professional profile?

Is it possible for a single researcher to maintain sufficient expertise in both statistics/computer science to model complex problems alone?

In my opinion, alongside the figure of data scientists, the decisive element is knowing how to dialogue and interact between different communities, between experts with different skills.

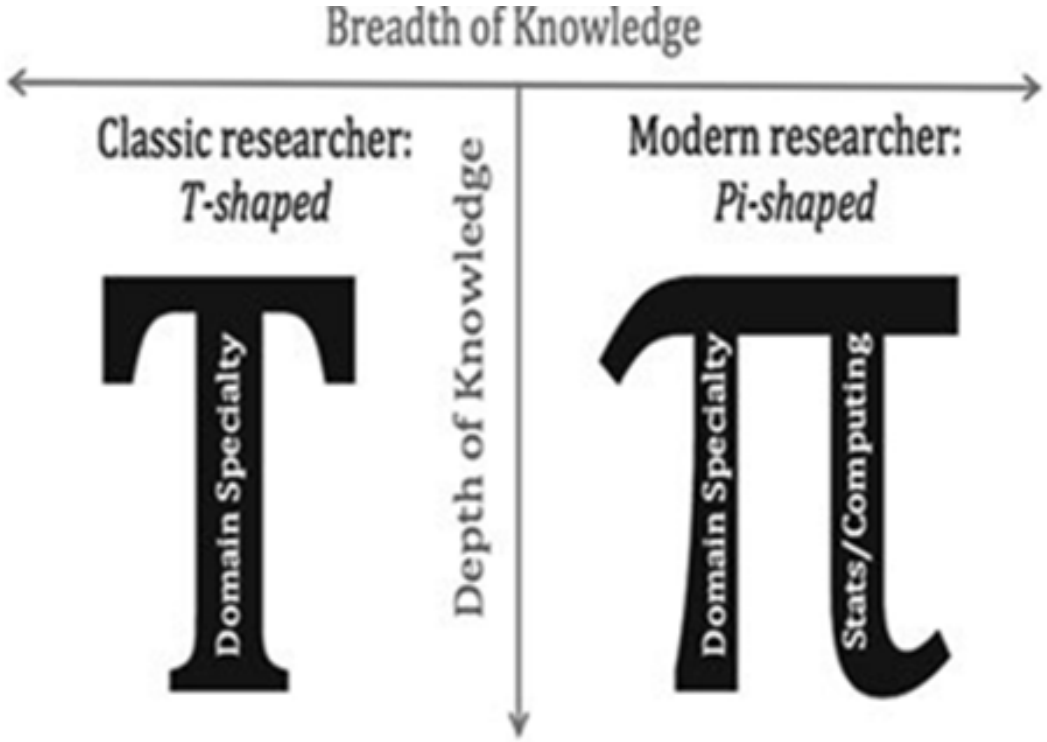
Not only in the design of new training courses but also in the search for new methods of analysis.

Therefore not only data integration but also integration among competencies and different professional communities.

Future challenges and developments. CAPABILITIES

Modern education around the world is advocating the evolution from T-shaped to Pi-shaped models of education (Ceri, 2018; Secchi, 2018).

The new model adds another vertical competence, relative to the statistical and computational abilities which are required in order to deal with data analysis.



Data science research expertise: T vs. Pi-shaped education (DrewConway, 2015)

Professional life requires in any domain the ability to draw information from large datasets, using a broader set of skills.

This is (would be) an important recognition of our discipline as useful in all fields.

Future challenges and developments. CAPABILITIES

To work in an interdisciplinary team motivates and encourages us to keep moving forward, raising additional questions.

We have a competitive advantage, offering our natural critical eye and scepticism with current understanding.

Our openness and commitment in this direction are fundamental and need to be encouraged and pursued.

Essential for attracting young people into the field, to survive as an energetic, open and creative discipline.

Future challenges and developments. CAPABILITIES

Is our community of statisticians convinced on how strategic this point is for the future?

Does our evaluation system encourage young researchers of different disciplines to work together in interdisciplinary teams?

Does it promote mutual advantages from working in large teams with different cultures and communities?

Or does it encourage us to work in stable little specialized groups, of pure statisticians, experts in some limited aspect?

Is the value of starting from real data sufficiently recognized for understanding the sense of data.

Has a bias in the choice of research themes been introduced by the evaluation system and how is it used in our community?

Future challenges and developments. METHODS

The essential role of statistical research is to develop new tools for use at the frontiers of science.

What are the research challenges facing the core of statistics, i.e. regardless of the needs of statistics in particular scientific domains?

Some of the commonly mentioned strategic research lines at the beginning of the new century are still valid: scale of data, data reduction and compression, data analysis outside statistics (machine learning and neural network), multivariate analysis for large p small n, combination of Bayesian and frequentist methodologies, middle ground between proof and computational experiments (Lindsay et al, 2005).

But other issues are emerging.
Mostly connected with the new types of data.

Future challenges and developments. METHODS

How to deal with the replicability and stability of conclusions and findings, with:

- the high degree of heterogeneity and, not-IID data,
- highly unstructured data (including images, text, sound, other new forms) to be understood together,
- data observed through different types of measurement devices,
- not necessarily resulting from designed experiments, but corresponding to a mixture of many heterogeneous populations,
- with missing and biased observations (making valid inferences from non probabilistic sampling).

The data complexities call for substantial pre-processing (Reid, 2018) and new approaches to theoretical concepts and methodological developments.

The point is to find the ways to make valid inferences with very complex models (high-dimensional inference) on networks, shapes, images, spatial data evolving over time, also with multidimensional variable of interest (Iaccarino, 2019).

Developing new theories and methods but above all knowing how to interact with other scientists and stakeholders.
If not we risk rapidly becoming irrelevant to the future of data science.

Future challenges and developments. METHODS

For Official Statistics the more relevant expected developments concern the issue of data integration in the new statistical ecosystem.

In particular the following connected themes.

- The design of a new role for sample survey in the new multi-source ecosystem (Alleva, 2017);

- How to deal with the uncertainty of information produced exploiting the new data ecosystem (for example the Italian Integrated system of statistical registers, ISSR), created by a massive integration of administrative archives and survey data (Alleva, Falorsi, Petrarca, Righi, 2019; Iaccarino, 2019);

- How to exploit Big Data for producing OS , a path starting from the experimentation of their treatment and assessment with respect to statistical quality (Alleva, 2019, Falorsi, 2019);

- How to bridge the gap between statistical models and algorithmic inference.

Future challenges and developments. METHODS

How to deal with uncertainty : a strategic issue for NSIs , for trust and transparency.

The main strategic choice is if either
"*make the use of ISSR limited and allow the dissemination of only planned* outputs *having a certified level of accuracy*" or
"*make the system more flexible, allowing the users to produce their own statistics from the ISSR*".

Istat opted for the second option, that makes the NSI more relevant for users but which obliges the need for a policy for reducing the risk of an erroneous use of the data.

This leads to new methodological and technological challenges (Alleva et al. cit., 2019):
(*i*) how to ensure the confidentiality of the results,
(*ii*) how to measure the accuracy of the register estimates, and
(*iii*) how to make the users aware of the accuracy.

Coherently Istat decided that "*users should be informed of the accuracy of the estimates*" *(*instead of the traditional approach "*users should ignore it*").

This choice is positive for trust, transparency and for a correct use of register data, but it is computationally complex.

Future challenges and developments. DATA GOVERNANCE

We must adapt the statistical theory and methods to the new types of data, otherwise we serially risk to become irrelevant.

The point is not only advancements in methods, technologies and competences but also in data governance, in terms of policy, openness, privacy, and trust: the fourth challenge for the future of statistics is the governance of data, their production, processing and communication.

The rapid increasing of availability and utility of data poses not only great opportunities and advantages, but also threats.

The most important cross-cutting theme related to big data is privacy, which covers all aspects of the data life-cycle.

New regulations look not only to protect our privacy, and how we store information about ourselves, but also to include all the process and also what we are allowed to do with that data (EU-GDPR, 2016).

Future challenges and developments. DATA GOVERNANCE

The NSIs have studied for many decades the disclosure limitation and innovation in combining statistics with encryption to ensure privacy are expected, as well in terms of anonymization schemes and methods to analyse anonymized data.

The novelty is the advent of big companies and global digital platforms, out of the NSSs, that have the monopoly of huge volumes of data about people and businesses.

This is a big risk, not only for the relevance of NSIs but also for the asymmetry in the information market and the risk of illegal use and disclosure of data.

Risk of mistrust for statistics and data science.

We must be aware of these risks and we must promote and act vigilance over the correct use of data and its transformation into a public good through access by NSSs and researchers.

Future challenges and developments. DATA GOVERNANCE

More recently privacy concerns have expanded to include concerns about two crucial statistical tools/products:

- the informative value of the integration of a multiplicity of administrative archives (considered a sort of police filing of people, a big brother) and

- the algorithmic decision fairness (particularly with deep learning algorithms), which can be quite opaque even to their developers (Shah, 2017; O'Neil, 2016).

The first one, considering the Istat's investment in building the Integrated system of statistical registers (ISSR) as the core of the new production model, represents e a great risk. It is strategic to clarify with the Privacy Authorities its utility and absolute safety.

As for the second one, the new European Regulation (GDPR) details the rights of citizens, if affected by a particular algorithmic decision, to an explanation of why that decision was reached.

Conclusions

The big increase in public interest in data and data science is very promising:

- for the growth of statistical science, and
- for the active engagement of statisticians with new emerging fields.

Society can largely benefit from new informative infrastructures and tools for decision making, thanks to the power of statistics and data science, but the need for greater scrutiny and transparency is crucial.

The challenge for our community is to adapt and develop theory and methods to analyse the new types of data, assuming a leadership role in producing high quality public information, also leveraging the opportunity to work with other scientific communities interested in the use of Big Data.

The core task is still the inductive inference from data to models and scientific conclusions in the new data ecosystem.

NSIs should play an important role, and the choice of *Trust Smart Statistics* as the ESS strategy seems to be powerful and responsible.

Conclusions

Integration and partnership are the two key points:

- data sources integration,
- integration between stochastic modelling and algorithmic procedures,
- combination of frequentist and bayesian approaches,
- integration among competencies and professional profiles for education and training,
- partnership between OS and researchers from Academia and other public and private institution and companies,
- partnership between NSIs and private holders of data,
- partnership between NSIs and Privacy Authorities.

It is not easy, but this is the route we must take to move forward.

References (in text)

Alleva G. (2017) The new role of sample surveys in official statistics, ITACOSM 2017, The 5th Italian Conference on Survey Methodology,14 giugno 2017, Bologna.

Alleva G., Falorsi P.D., Petrarca F., Righi P. (2019) Measuring the accuracy of aggregates computed from a statistical register, Submitted to the Journal of Official Statistics.

Alleva G. (2019) The path for using Big Data sources in Istat, Convegno SIS-2019, 20 giugno 2019, Milano,

Ceri S. (2018) On the role of statistics in the era of big data: A computer science perspective, Statistics and Probability Letters 136 (2018) 68–72

Citro C.F.(2014) From multiple modes for surveys to multiple data sources for estimates.Survey Methodology, Statistics Canada

Eurostat (2018) *Bucharest Memorandum* on *Trusted Smart Statistics*, DGINS 2018).

EU GDPR (2016) Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, Official Journal of the European Union, Volume 59, 4 May 2016, ISSN 1977-0677

Falorsi P.D. (2019) Istat's Experimental Statistics based on Big Data, SIS-2019, 20 giugno 2019, Milano.

Iaccarino G. (2019) Metrics and Methods for Uncertainty Quantification, New Techniques and Technologies for Statistics (NTTS), 12-14 March 2019, Brussels.

Istat, (2016), Istat's Modernisation Programme,

Lindsay B.G., Kettenring J., Siegmund D.O. (2004) A Report on the Future of Statistics, *Statistical Science,* 2004, Vol. 19, No. 3, 387–413.

Nelder J.A. (1986) Statistics, Science and Technology ,*Journal of the Royal Statistical Society. Series A (General)*, Vol. 149, No. 2 (1986), pp. 109-121

O'Neil C. (2016) Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, Crown Random House

Reid N. (2018) Statistical science in of the world of big data, Statistics and Probability Letters 136 (2018) 42–45

Shah H. (2017) The DeepMind debacle demands dialogue on data. Nature 547, 259. http://dx.doi.org/10.1038/547259a.

References (other)

Benzécri J.P. (1979) L'analyse des données, Tome I Taxinomie, Tome II Correspondances, Dunod.
Breiman L. (1996) Heuristics of instability and stabilization in model selection. Ann. Statist. 24, 2350–2383.
Breiman L. (2001) Statistical Modeling: The Two Cultures, *Statistical Science,* 2001, Vol. 16, No. 3, 199–231
Bühlmann P., van de Geer S. (2018) Statistics for big data: A perspective, Statistics and Probability Letters 136 (2018) 37–41
Cox D.R. (2015) Big dat a and precision, Biometrika 102, 712-716.
Eurostat (2013) *Scheveningen Memorandum* on *Big Data for Official Statistics*, DGINS 2013.
Friedman J.H. (2001) The Role of Statistics in the Data Revolution?, International Stat Rev., Vol. 69, No. 1, pp. 5-10
Gini C. (1930) Present Condition and Future Progress of Statistics, Journal of the American Statistical Association, Vol. 25, No. 171 (Sep., 1930), pp.295-304
Harford T. (2014) Big data: are we making a big mistake? Financial Times, March 28.
Mandello J.G. (1905) The Future of Statistics, Journal of the Royal Statistical Society, Vol. 68, No. 4, pp. 725-732
Nelder J.A. (1999) From Statistics to Statistical Science, *Journal of the Royal Statistical Society. Series D*, Vol. 48, No. 2 (1999), pp. 257-269
Olhede S.C., Wolfe P.J. (2018) The future of statistics and data science, Statistics and Probability Letters 136 (2018) 46–50
Pearson K. (1920) The fundamental problem of practical statistics, Biometrika 13, 1-16.
Rao C.R. (2001) Statistics: Reflections on the past and visions for the future, Comm. Statist. Theory Methods 30, 2235-2257.
Savage L.J. (1954) The Foundation of Statistics, Wiley, New York.
Scott E.M. (2018) The role of Statistics in the era of big data: Crucial, critical and under-valued, Statistics and Probability Letters 136 (2018) 20–24
Secchi P. (2018) On the role of statistics in the era of big data: a call for a debate. Statist. Probab. Lett. 136, 10–14.
Shah H. (2017) The DeepMind debacle demands dialogue on data. Nature 547, 259. .
Tukey J.W. (1962) The future of data analysis, Ann. Statist. 33, 1-67
Tukey J.W. (1977) Exploratory data analysis. Reading, PA: Addison-Wesley.
Watts D.G. (1968) Conference on the Future of Statistics, Academic Press, New York.
Westergaard H. (1918) On the Future of Statistics, Journal of the Royal Statistical Society, Vol. 81, No. 3, pp. 499-520

Thank you for your attention!